NÁRODNÁ BANKA SLOVENSKA
EUROSYSTÉM

On a Bootstrap Test for Forecast Evaluations

Marián Vávra

Working Paper 5/2015

www.nbs.sk
Imricha Karvaša 1
813 25 Bratislva

reserarch@nbs.sk

June 2015

ISSN 1337-5830

The views and results presented in this paper are those of the authors and do not necessarily represent the official opinions of the National Bank of Slovakia.

NÁRODNÁ BANKA SLOVENSKA
EUROSYSTÉM

On a Bootstrap Test for Forecast Evaluations[1]

Working paper NBS

Marián Vávra[2]

Abstract

This paper is concerned with the problem of testing for the equal forecast accuracy of competing models using a bootstrap-based Diebold-Mariano test statistic. The finite-sample properties of the test are assessed via Monte Carlo experiments. As an illustration, the forecast accuracy of the US Survey of Professional Forecasters is compared to that of an autoregressive model. The empirical results indicate that professionals beat AR models systematically only for a single economic variable – the unemployment rate.

JEL classification: C12, C15, C32, C53
Key words: Forecast evaluation; Diebold-Mariano test; Sieve bootstrap

[1] I would like to thank Lubos Pastor, Zacharias Psaradakis, Ron Smith, and participants in the Research Seminar at the NBS for useful comments and interesting suggestions. All remaining errors are only mine.
[2] Marián Vávra, Research Department of the NBS.

1. INTRODUCTION

It is nowadays well understood that monetary policy should be forward-looking in order to efficiently impact the economy and achieve its policy goals. For this purpose, central banks use a suite of models to forecast key economic variables (see, e.g., Kapetanios, Labhard, and Price (2008) for a survey). The evaluation of the forecast accuracy of competing models is thus of the fundamental importance for central banks not only for selecting the "best" forecasting tools but also for further development of the progressive forecasting methods.

Two classes of statistics for testing predictive ability have become particularly popular in the literature: (i) equal predictive ability (EPA) tests (see, e.g., West (1996) and Diebold and Mariano (1995)); and (ii) superior predictive ability (SPA) tests (see, e.g., White (2000) and Hansen (2005)). We focus on the EPA family of tests for the following reasons: (i) Unlike the EPA statistics, the SPA ones (e.g. the White's reality check) require the specification of the benchmark model for comparison. Forecast inference is thus conditional on the benchmark. In practice, however, one can face situations when it might be difficult to find a natural benchmark (see Section 4 for an example); (ii) Moreover, inference from the SPA tests seems to be more suitable when comparing a large number (i.e. hundreds or even more) of competing models[3], whereas inference from the EPA tests is very convenient for comparison of several competing models (usually the case in empirical macroeconomics).

In this paper, special attention is paid to a test statistic proposed by Diebold and Mariano (1995) which is a prominent member of the EPA family. Although the Diebold-Mariano (DM) test is conceptually simple, easy to calculate, and very popular in empirical macroeconomics and finance, its finite sample properties are not satisfactory – in fact, the magnitude of a size distortion makes the DM test unreliable for empirical applications (see Tables 1– 2 in this paper). This paper contributes to the literature by considering a bootstrap-based Diebold-Mariano (BDM) statistic. It is shown that the BDM test is very easy to compute, it has very good size and reasonable power properties in finite samples, which makes the BDM test conservative, yet reliable for empirical applications.

The paper is organized as follows. The original and bootstrap-based Diebold-Mariano test statistics are discussed in Section 2. Section 3 examines the finite-sample properties of both statistics by means of Monte Carlo experiments. Section 4 presents an application of the tests to the selected indicators from the US Survey of Professional Forecasters. Section 5 summarizes and concludes.

[3]For instance, White (2000) uses the reality check approach when comparing 3,654 competing forecast strategies for asset returns.

2. ORIGINAL AND BOOTSTRAP DIEBOLD-MARIANO TESTS

Consider $\{(X_{1,t}, X_{2,t}) : t \in \mathbb{Z}\}$ of being a pair of the covariance stationary and pairwise correlated forecast errors coming from two alternative (non-nested) models. Then, Diebold and Mariano (1995) proposed a conceptually simple statistic for testing the equal forecast accuracy of two competing models based on testing that the population mean of the loss differential $\mathbb{E}(d_t) = \mathbb{E}[G(X_{1,t}) - G(X_{2,t})]$ is zero, where $G(X)$ represents a loss function. The authors consider a mean squared error (MSE) measure for which $G(X) = X^2$.[4] It follows from the CLT of dependent observations (see White (2001, Theorem 5.20)) that

$$\sqrt{n}(\bar{d} - \mathbb{E}(d_t)) \xrightarrow{d} N(0, \sigma^2), \tag{1}$$

where $\bar{d} = n^{-1} \sum_{t=1}^{n} (X_{1,t}^2 - X_{2,t}^2)$ stands for a sample analogy of the loss differential and $\sigma^2 = \sum_{j=-\infty}^{\infty} \gamma_j$ denotes the long-run variance and γ_j stands for the autocovariance at lag j.[5] The authors propose to use a standard t-test for testing the null hypothesis that $\mathbb{E}(d) = 0$, that is

$$\mathcal{D} = \sqrt{n} \left(\frac{\bar{d}}{\hat{\sigma}} \right) \xrightarrow{d} N(0, 1), \quad \text{as} \quad n \to \infty. \tag{2}$$

The null hypothesis of the equal forecast accuracy is rejected at the nominal level $0 < \alpha < 1$ once $|\mathcal{D}| > q_{1-\alpha/2}$, where $q_{1-\alpha/2}$ is the $1 - \alpha/2$ quantile of the $N(0, 1)$ distribution.[6] Motivated by the literature on estimation of the asymptotic variance in the presence of weak dependence, the following estimator is used

$$\hat{\sigma}^2 = \sum_{j=-m}^{m} w(j/m) \hat{\gamma}_j, \quad \text{where} \quad w(j/m) = \left(1 - \frac{|j|}{m+1} \right), \tag{3}$$

where $w(\cdot)$ are the Bartlett weights, m is a real-valued bandwidth such that $m \to \infty$ and $m/n \to 0$ as $n \to \infty$, and $\hat{\gamma}_j = n^{-1} \sum_{t=j+1}^{n} (d_t - \bar{d})(d_{t-j} - \bar{d})$ denotes a sample autocovariance at lag j. The lag order m is usually determined using an automatic lag selection procedure proposed by Newey and West (1994).[7]

[4] Note that other loss functions can be considered as well. See Remark 4 in this paper.

[5] Note that the Gaussian distribution asymptotically holds also for nested forecasting models provided that the forecast horizon is fixed and the sample size increases (see West (2006)).

[6] Note that the DM test can be used for testing the one-sided hypothesis as well.

[7] Note that the lag order m is sometimes linked to the forecast horizon h (see Harvey, Leybourne, and Newbold (1997, p. 282)). This argument is motivated by the fact that the h-step ahead forecast errors from dynamic time series models (e.g. ARMA models) follow an MA($h-1$) process. However, this assumption holds under rather unrealistic assumptions such that the model is correctly identified (i.e. the error terms are white noise) and the model parameters are known. Therefore, we hold the view that a Wold representation provides a more realistic approximation to the true data generating process of empirically observed forecast errors. See also Appendix A for further details.

It is worth remarking that the finite sample properties of the DM test are poor – the test suffers from a serious size distortion which increases with the persistence of the forecast errors and with the forecast horizon (see, e.g., Harvey, Leybourne, and Newbold (1997) for Monte Carlo evidence). The small-sample correction of the DM test discussed in Harvey, Leybourne, and Newbold (1997) altogether with using the Student t distribution as the limiting distribution in (2) provide only a marginal improvement to a size distortion. The main problem with the DM test is that although consistency of the estimator in (3) is well established in the literature (see Andrews (1991, Theorem 1)), it is highly inaccurate (downward biased) for persistent stochastic processes in finite samples (see Andrews (1991)). As a result, convergence of the empirical distribution of the DM statistic to its limiting Gaussian (or Student) distribution is rather slow. Andrews and Monahan (1992) and Müller (2014) suggest to pre-white the observed data with a fixed-order AR (AR(1) respectively) model and use the estimated parameters to calculated the target quantity (i.e. the long-run variance in our case).[8] Although the authors demonstrate some improvements in estimates of the long-run variance in finite samples, the problem here is that this approach requires the existence of the closed-form expression of the target quantity. This fact limits the use of this approach in practice. Another, and a more general solution, is to apply a bootstrap technique to calculate the long-run variance (see Goncalves and Vogelsang (2011)). This approach is particularly useful in cases where the limiting distribution of a given statistic is an asymptotically pivotal quantity (as in our case – see (2)) (see Davison and Hinkley (1997, p. 40)).

Without loss of generality we consider a real-valued Wold representation for the bivariate forecast errors[9] $\boldsymbol{x}_t = (X_{1,t}, X_{2,t})'$ given by

$$\boldsymbol{x}_t = \boldsymbol{\mu} + \sum_{j=1}^{\infty} \boldsymbol{\psi}_j \boldsymbol{\epsilon}_{t-j} + \boldsymbol{\epsilon}_t, \quad t \in \mathbb{Z}, \tag{4}$$

where $\boldsymbol{\mu} \in \mathbb{R}^2$ and the error sequence $\{\boldsymbol{\epsilon}_t : t \in \mathbb{Z}\}$ is assumed to be a strictly stationary and ergodic vector of innovations such that $\mathbb{E}(\boldsymbol{\epsilon}_t) = \mathbf{0}$, $\mathbb{E}(\boldsymbol{\epsilon}_t \boldsymbol{\epsilon}_t') = \boldsymbol{\Sigma}$, which is a symmetric and positive definite matrix, $\mathbb{E}(||\boldsymbol{\epsilon}_t||^8) < \infty$ and the density function $f(\boldsymbol{\epsilon}_t)$ is absolutely continuous on $\mathbb{R}^2$.[10] Additionally, we assume the spectral density matrix of $\boldsymbol{x}_t$ fulfils the boundedness condition – eigenvalues of the density matrix are uniformly bounded away from zero at all frequencies $\lambda \in [-\pi, \pi]$. Under additional mild assumption about invertibility, it is easy to show that the

[8] For instance, the long-run variance of an AR(1) process is $\gamma_0 = \sigma^2/(1-\phi^2)$, where ϕ denotes the AR coefficient and σ^2 is the residual variance.

[9] It is worth remarking that a Wold representation is often considered as a representative stochastic process for forecast errors in the literature (see, e.g., Diebold and Lopez (1996) or Mariano and Preve (2012)). See also Footnote 7 and Appendix A for further details.

[10] Note that a rather strict moment condition can be weaken using a different loss function. For instance, when considering the mean absolute error loss function, only the first four moments are required to be finite.

process in (4) can be rewritten into the form of a bivariate VAR(∞) model

$$\boldsymbol{x}_t = \boldsymbol{c} + \sum_{j=1}^{\infty} \boldsymbol{\phi}_j \boldsymbol{x}_{t-j} + \boldsymbol{\epsilon}_t, \quad t \in \mathbb{Z}, \tag{5}$$

where the roots of the lag polynomial $\det(\boldsymbol{I} - \sum_{j=1}^{\infty} \boldsymbol{\phi}_j z^j)$ lie outside the unit disk and $\boldsymbol{I}$ denotes a (2×2) identity matrix. We also assume that the summability condition $\sum_{j=-\infty}^{\infty}(1+|j|)||\Gamma_j|| < \infty$ holds, where Γ_j is the vector autocovariance of $\boldsymbol{x}_t$ at lag j. These conditions are necessary to ensure to validity of a bootstrap procedure discussed in the next paragraph (see condition A in Meyer and Kreiss (2014)).

The functional form of weakly dependent forecast errors in (5) immediately suggests the use of a VAR-sieve bootstrap of Paparoditis (1996) and Meyer and Kreiss (2014) to calculate more accurate critical values of the DM statistic in finite samples. The VAR-sieve bootstrap is particularly attractive because VAR modelling is a well-studied problem in the literature and, therefore, the procedure can be implemented in a straightforward way (see, e.g., Choi and Hall (2000), Psaradakis (2003a,b), Alonso, Peña, and Romo (2002, 2003), Chang and Park (2003), Kapetanios and Psaradakis (2006), Gonçalves and Kilian (2007), Poskitt (2008), Fuertes (2008), Palm, Smeekes, and Urbain (2010), Psaradakis (2015) for other applications of a sieve bootstrap).

The procedure applied to generate bootstrap based critical values of the Diebold-Mariano test statistic (BDM) is summarized in the following algorithm.

Algorithm 1 (i) Select an appropriate lag order p of a VAR model for a bivariate forecast error vector $\{\boldsymbol{x}_t : t = 1, \ldots, n\}$ using the Akaike information criterion (AIC), where the lag order is restricted by $0 \leq p \leq 5\log_{10}(n)$, where n denotes the sample size.[11]

(ii) Estimate the unknown VAR parameters by the multivariate least-squares (LS) method (see Lütkepohl (2005, Ch. 3). Note that the setup of the the upper lag order is sufficient to ensure consistency of the estimated VAR parameters at the desired rate (see condition B in Meyer and Kreiss (2014)).[12]

(iii) Construct a sequence of the estimated residuals $\{\hat{\boldsymbol{\epsilon}}_t : t = p+1, \ldots, n\}$ by the recursion

$$\hat{\boldsymbol{\epsilon}}_t = \boldsymbol{x}_t - \hat{\boldsymbol{c}} - \sum_{j=1}^{p} \hat{\boldsymbol{\phi}}_j \boldsymbol{x}_{t-j}.$$

[11] Note that other lag order selection criterion such as the Bayesian information criterion (BIC) could be used, but since the processes are not assumed to be of finite dimensions, the AIC is asymptotically efficient (see Shibata (1980))

[12] In contrast to Bühlmann (1997) and Meyer and Kreiss (2014), who implemented the Yule-Walker (YW) estimator, we rely on the multivariate LS estimator. The main reason for doing so is that the LS estimator produces superior results as compared to the YW estimator (see Tjøstheim and Paulsen (1983)).

(iv) Draw a random vector $\{\hat{\boldsymbol{\epsilon}}_t^* : t = 1, \ldots, n + 100\}$ from a bivariate empirical distribution function given by

$$\hat{F}_n(\boldsymbol{u}) = \frac{1}{n - 2p - 1} \sum_{t=p+1}^{n} \mathbb{I}(\hat{\boldsymbol{\epsilon}}_t \leq \boldsymbol{u}),$$

where $\mathbb{I}(\cdot)$ denotes a standard indicator function and $\boldsymbol{u} \in \mathbb{R}^2$.

(v) Generate bootstrap replicates $\{\boldsymbol{x}_t^* : t = 1, \ldots, n + 100\}$ by the recursion

$$\boldsymbol{x}_t^* = \hat{\boldsymbol{c}} + \sum_{j=1}^{p} \hat{\boldsymbol{\phi}}_j \boldsymbol{x}_{t-j}^* + \hat{\boldsymbol{\epsilon}}_t^*.$$

where the process is initiated by a vector of sample averages $(\boldsymbol{x}_{-p+1}^*, \ldots, \boldsymbol{x}_0^*) = (\bar{\boldsymbol{x}}, \ldots, \bar{\boldsymbol{x}})$ where $\bar{\boldsymbol{x}} = n^{-1} \sum_{t=1}^{n} \boldsymbol{x}_t$. The first 100 data points are then discarded in order to eliminate start-up effects and the remaining n data points are used. Consistently with the null hypothesis (i.e. the equality of mean squared forecast errors: $\mathbb{E}(X_{1,t}^2) = \mathbb{E}(X_{2,t}^2)$), generate the normalized bootstrap vector $\boldsymbol{z}_t^* = (Z_{1,t}^*, Z_{2,t}^*)'$ according to

$$Z_{1,t}^* = X_{1,t}^* \sqrt{(\omega_1^2 + \omega_2^2)/2\omega_1^2},$$
$$Z_{2,t}^* = X_{2,t}^* \sqrt{(\omega_1^2 + \omega_2^2)/2\omega_2^2},$$

where $\omega_i^2 = n^{-1} \sum_{t=1}^{n} X_{i,t}^2$ denotes the sample second raw moment.

(vi) Construct a bootstrap analogy of the DM test statistic $\mathcal{B}^*$ using (2)-(3) but calculated from the normalized bootstrap samples $\{\boldsymbol{z}_t^* : t = 1, \ldots, n\}$.

(vii) Repeat steps (iv)–(vi) independently B times to get a sample of the bootstrap DM statistics $\{\mathcal{B}_j^* : j = 1, \ldots, B\}$. Then, the sampling distribution of the $\mathcal{B}$ test statistic is approximated by the empirical distribution function associated with $\{\mathcal{B}_j^* : j = 1, \ldots, B\}$: $H^*(u) = B^{-1} \sum_{j=1}^{B} I(\mathcal{B}_j^* \leq u)$, where $u \in \mathbb{R}$. Finally, a bootstrap test of the nominal level α rejects the null hypothesis if

$$|\mathcal{B}| > \inf\{u : H^*(u) \geq (1 - \alpha/2)\},$$

where $\mathcal{B}$ is the DM test statistic obtained from the observed samples $\{\boldsymbol{x}_t : t = 1, \ldots, n\}$.

Remark 1: Since the DM test in (2) is constructed by rewriting (1), it is fully sufficient to focus on the quantity $\bar{d}$, the sample loss differential, when proving the validity of the VAR-sieve bootstrap. Note that $\bar{d}$ can be also written as follows

$$\bar{d} = \frac{1}{n} \sum_{t=1}^{n} \left(X_{1,t}^2 - X_{2,t}^2\right) = \frac{1}{n} \sum_{t=1}^{n} \boldsymbol{\delta}'(\boldsymbol{x}_t \odot \boldsymbol{x}_t) = \frac{1}{n} \sum_{t=1}^{n} g(\boldsymbol{x}_t), \tag{6}$$

where $\boldsymbol{x}_t = (X_{1,t}, X_{2,t})'$, $\odot$ denotes element-by-element multiplication and $\boldsymbol{\delta} = (1, -1)'$. Since $g(\boldsymbol{x}_t)$ function in (6) is continuously differentiable with bounded (second) partial derivatives, condition C in Meyer and Kreiss (2014) is satisfied. The validity of the VAR-sieve bootstrap thus follows directly from Theorem 4.1 in Meyer and Kreiss (2014).

Remark 2: An important question is to what extent the VAR-sieve bootstrap works for stochastic processes not stemming from the representation in (5). One can argue that the closure of a VAR(∞) representation in (5) is fairly large which means that for any non-linear stochastic process there exist another process in the closure of linear processes (see Bickel and Bühlmann (1997)). This finding implies that the VAR-sieve bootstrap is very likely to give satisfactory results even for stochastic processes deviating from a Wold representation.

Remark 3: It is well known statistical fact that standard estimators of VAR models suffer from a small sample bias (see, e.g., Yamamoto and Kunitomo (1984) and Engsted and Pedersen (2014)). However, Kim and Durmaz (2012) show that a bootstrap bias correction does not necessarily improve the forecast performance. The reason is that although a bootstrap procedure reduces a bias, it tends to increase a variance, and thus, the impact on the MSE is ambiguous. Therefore, the bootstrap bias correction of the estimated VAR parameters in Step (ii) of Algorithm 1 is not implemented here.

Remark 4: It is worth remarking that the bootstrap procedure can be used also for other forecast accuracy measures such as the mean absolute error (MAE) (i.e. $G(X) = |X|$). It can be shown that the MAE is continuously differentiable almost everywhere for $X \in \mathbb{R}$.[13] Our additional simulation experiments show the BDM test based on the MAE works works as well as for the MSE measure.[14]

3. Monte Carlo Simulations

In this section, the size and power properties of the proposed BDM and DM test statistics are assessed by means of Monte Carlo experiments.

3.1 Experimental Design

Following Christensen, Diebold, Rudebusch, and Strasser (2007), the experiments are based on artificial data generated according to various configurations of AR(1) models given by[15]

$$X_{i,t} = c_i + \phi_i X_{i,t-1} + \kappa_i \epsilon_{i,t}, \quad \text{for} \quad i \in \{1, 2\}. \tag{7}$$

[13] Note that the fact that the MAE is not continuously differentiable at a single point $X = 0$ does not represent a problem for forecast errors drawn from a continuous distribution.

[14] The results are available from the author upon request.

[15] It can be shown that an AR(1) model is a good approximation to the empirically observed forecasts errors (see Appendix A for further details).

The configuration of individual parameters is as follows:

M1: $c_1 = c_2 = 0.2$, $\phi_1 = \phi_2 = 0.5$, $\kappa_1 = \kappa_2 = 1.0$;

M2: $c_1 = c_2 = 0.2$, $\phi_1 = \phi_2 = 0.8$, $\kappa_1 = \kappa_2 = 1.0$;

M3: $c_1 = 0.4$, $c_2 = 0.2$, $\phi_1 = \phi_2 = 0.8$, $\kappa_1 = \kappa_2 = 1.0$;

M4: $c_1 = c_2 = 0.2$, $\phi_1 = 0.8$, $\phi_2 = 0.5$, $\kappa_1 = \kappa_2 = 1.0$;

M5: $c_1 = c_2 = 0.2$, $\phi_1 = \phi_2 = 0.8$, $\kappa_1 = \sqrt{2.0}$, $\kappa_2 = 1.0$;

Note that M1 and M2 models are used to asses the size properties of the tests, whereas M3 – M5 models assess the power properties. Experiments proceed by generating 1,000 independent artificial time series $\{(X_{1,t}, X_{2,t}) : t = 1, \ldots, 100 + n\}$ with $n \in \{50, 100, 200\}$ for each design point. The first 100 data points of each series are then discarded in order to eliminate start-up effects and the remaining n data points are used to compute the value of the BDM and the DM test statistics. In each case, the bivariate vector of correlated errors $\boldsymbol{\epsilon}_t = (\epsilon_{1,t}, \epsilon_{2,t})'$ is calculated as $\boldsymbol{\epsilon}_t = \boldsymbol{P}\boldsymbol{v}_t$, where $\boldsymbol{v}_t = (v_{1,t}, v_{2,t})'$ are bivariate independent random varieties drawn from either $N(0, 1)$ or $t(10)$ distributions and $\boldsymbol{P}$ is the lower triangular matrix of the Choleski decomposition of the correlation matrix $\boldsymbol{R}$ with the pairwise correlation coefficient ρ such that

$$\boldsymbol{R} = \begin{bmatrix} 1 & \rho \\ \rho & 1 \end{bmatrix} = \boldsymbol{P}\boldsymbol{P}', \quad \text{where} \quad \boldsymbol{P} = \begin{bmatrix} 1 & 0 \\ \rho & \sqrt{1-\rho^2} \end{bmatrix}.$$

Two values of the pairwise correlation $\rho \in \{0.25, 0.75\}$ between model innovations are considered for the Monte Carlo experiments.

3.2 Simulation Results

The Monte Carlo rejection frequencies of the DM test based on (2)-(3) and the proposed BDM test based on Algorithm 1 of nominal level 0.10 (the level usually used in the forecasting literature) are reported in Tables 1 – 2. The results suggest the following:

(i) For all relevant data points (see models M1 and M2), the proposed BDM test has empirical levels very close to the nominal level 0.10, regardless of the sample size n and the pairwise correlation ρ of model innovations, whereas the original DM test exhibits a serious size distortion even in a relatively large sample (e.g. $n = 200$). The size distortion of the DM test is negatively affected mainly by the persistence of data (compare the results for M1 and M2 models), whereas the cross-correlation between forecast errors plays a minor role.

(ii) As might be expected, some power loss is observed in the case of the BDM test as compared to the original DM test. Nevertheless, the power results improve quickly with the increasing

sample size n. In any case, the observed power loss is not of a magnitude that makes the BDM test unattractive for applications.

(iii) No significant differences in the size and power results of the BDM and DM tests are observed for Gaussian and heavy-tailed innovations.

Table 1: Empirical Rejection Frequencies of the BDM and DM Tests: $N(0,1)$ innovations

	$n=50$				$n=100$				$n=200$			
	$\rho=0.25$		$\rho=0.75$		$\rho=0.25$		$\rho=0.75$		$\rho=0.25$		$\rho=0.75$	
DGP	$\mathcal{B}$	$\mathcal{D}$	$\mathcal{B}$	$\mathcal{D}$	$\mathcal{B}$	$\mathcal{D}$	$\mathcal{B}$	$\mathcal{D}$	$\mathcal{B}$	$\mathcal{D}$	$\mathcal{B}$	$\mathcal{D}$
M1	0.08	0.19	0.07	0.16	0.10	0.16	0.09	0.15	0.09	0.11	0.10	0.14
M2	0.10	0.32	0.09	0.29	0.09	0.24	0.10	0.25	0.09	0.19	0.11	0.21
M3	0.22	0.53	0.39	0.68	0.38	0.60	0.61	0.82	0.59	0.76	0.88	0.95
M4	0.30	0.64	0.46	0.80	0.64	0.87	0.78	0.96	0.93	0.98	0.99	1.00
M5	0.20	0.47	0.32	0.61	0.32	0.54	0.57	0.76	0.53	0.70	0.82	0.92

Table 2: Empirical Rejection Frequencies of the BDM and DM Tests: $t(10)$ innovations

	$n=50$				$n=100$				$n=200$			
	$\rho=0.25$		$\rho=0.75$		$\rho=0.25$		$\rho=0.75$		$\rho=0.25$		$\rho=0.75$	
DGP	$\mathcal{B}$	$\mathcal{D}$	$\mathcal{B}$	$\mathcal{D}$	$\mathcal{B}$	$\mathcal{D}$	$\mathcal{B}$	$\mathcal{D}$	$\mathcal{B}$	$\mathcal{D}$	$\mathcal{B}$	$\mathcal{D}$
M1	0.08	0.18	0.06	0.17	0.08	0.14	0.09	0.15	0.09	0.13	0.09	0.13
M2	0.09	0.28	0.11	0.30	0.09	0.24	0.09	0.24	0.10	0.20	0.10	0.20
M3	0.20	0.46	0.33	0.59	0.31	0.53	0.52	0.73	0.49	0.66	0.80	0.88
M4	0.31	0.64	0.44	0.79	0.54	0.81	0.75	0.94	0.91	0.97	0.99	1.00
M5	0.21	0.49	0.31	0.63	0.31	0.53	0.55	0.77	0.52	0.68	0.84	0.92

4. Do Private Forecasters Beat Time Series Models?

As mentioned earlier, accurate forecasts of economic variables are of the fundamental importance not only for the measures of central banks but also for economic decisions of firms and households. In this section, a question whether professionals can beat forecasts from simple (autoregressive) time series models is briefly discussed.[16] Due to long history and a large number of economic variables, we evaluate forecasts from the US Survey of Professional Forecasters (SPF) and linear autoregressive models using the proposed BDM test. The null hypothesis

$$H_0 : MSE(SPF) = MSE(AR) \quad \text{against} \quad H_1 : MSE(SPF) \neq MSE(AR)$$

[16] The interested reader is referred to the Federal Reserve Bank of Philadelphia website for a list of approximately 80 related research papers.

Table 3: P-values of the BDM and DM Test Statistics

		1983 Q1 – 2012 Q3		1991 Q1 – 2012 Q3	
variables	horizon	$\mathcal{B}$	$\mathcal{D}$	$\mathcal{B}$	$\mathcal{D}$
AAA	1	0.389	**0.098**	**0.071**	**0.023**
	2	0.586	0.168	0.220	0.108
	3	0.300	0.212	0.324	0.147
	4	0.307	0.192	0.398	0.198
TBILL	1	**0.080**	**0.034**	0.121	**0.051**
	2	0.282	0.133	0.265	0.154
	3	0.385	0.221	0.436	0.317
	4	0.463	0.318	0.541	0.425
PGDP	1	**0.097**	**0.062**	0.485	0.454
	2	0.109	**0.033**	0.557	0.401
	3	0.187	**0.034**	0.455	0.315
	4	**0.056**	**0.009**	0.207	**0.048**
GDP	1	0.130	**0.069**	0.265	0.177
	2	0.417	0.331	0.994	0.993
	3	0.693	0.681	0.825	0.806
	4	0.734	0.681	0.900	0.877
IP	1	0.182	0.114	0.291	0.198
	2	0.305	0.217	0.363	0.222
	3	0.258	0.147	0.270	0.161
	4	0.343	0.246	0.244	0.147
UR	1	**0.019**	**0.003**	**0.027**	**0.014**
	2	**0.029**	**0.006**	**0.079**	**0.030**
	3	**0.064**	**0.013**	0.101	**0.037**
	4	**0.025**	**0.006**	**0.022**	**0.012**
HOUS	1	**0.059**	**0.013**	0.158	**0.045**
	2	0.109	**0.034**	0.245	**0.093**
	3	0.141	**0.043**	0.235	**0.078**
	4	0.138	**0.044**	0.221	**0.050**

is tested for the following set of forecast errors with the forecast horizon from 1 up to 4 quarters: the 3-month Treasury Bill rate (TBILL), the AAA Corporate Bond yield (AAA), the real Gross Domestic Product growth rate (RGDP), the GDP deflator growth rate (PGDP), the Industrial Production growth rate (IP), the Unemployment rate (UR), and the Housing Starts (HOUS).[17] Due to an institutional break in the survey – the Federal Reserve Bank of Philadelphia took over the survey from the National Bureau of Economic Research in 1990 – we conduct our analysis over two different sub-samples: (i) 1983 Q1 – 2012 Q4 (i.e 120 observations); and (ii) 1991 Q1 – 2012 Q4 (i.e. 88 observations). Note that the SPF and AR model forecasts of the selected variables, including the forecast errors, are publicly available at the Federal Reserve Bank of Philadelphia website.[18]

The p-values of the BDM (based on $B = 1,000$ bootstrap replications) and DM tests are pre-

[17] For all selected variables, the first data releases are considered. Similar results are obtained for later data vintages. Results are available from the second author upon request.

[18] www.philadelphiafed.org

sented in Table 3. The results reveal dramatic differences in rejecting the null hypothesis. In particular, the BDM test rejects the null only in about 29% (14%) of cases at the nominal level 0.10, whereas the original DM in 54% (39%) using the long- (short-) sample. Put differently, the original DM test overrejects the null of the equal forecast accuracy hypothesis by factor 2 – 3 as compared to the BDM test. According to the reliable BDM test, the results indicate that the professionals beat the AR model forecasts systematically (over all forecast horizons) only for a single economic variable – the unemployment rate. This conclusion is perhaps not very surprising since the unemployment rate exhibits a high degree of business cycle asymmetry which might be predicted by professionals (based on previous experience) but not by linear AR models. What is more, we cannot find evidence of the systematic superior forecast performance of professionals in the case of interest rates (i.e. TBILL and AAA variables). This result is somewhat surprising keeping in mind the importance of the interest rates for the financial industry.

5. Conclusion

This paper has considered a new bootstrap-based Diebold-Mariano test statistic for testing for the equal forecast accuracy between two alternative forecast models/methods. We have shown that the BDM test has very good size and reasonable power properties in finite samples, making the BDM test conservative, yet reliable. Empirical results reveal dramatic differences in inference between the DM and BDM statistics.

REFERENCES

ALONSO, A., D. PEÑA, AND J. ROMO (2002): "Forecasting time series with sieve bootstrap," *Journal of Statistical Planning and Inference*, 100, 1–11.

——— (2003): "On sieve bootstrap prediciton intervals," *Statistics & Probability Letters*, 65, 13–20.

ANDREWS, D. (1991): "Heteroskedasticity and autocorrelation consistent covariance matrix estimation," *Econometrica*, 59, 817–858.

ANDREWS, D., AND J. MONAHAN (1992): "An improved heteroskedasticity and autocorrelation consistent covariance matrix estimator," *Econometrica*, pp. 953–966.

BICKEL, P., AND P. BÜHLMANN (1997): "Closure of linear processes," *Journal of Theoretical Probability*, 10, 445–479.

BOX, G., G. JENKINS, AND G. REINSEL (2008): *Time Series Analysis: forecasting and control*. Wiley.

BÜHLMANN, P. (1997): "Sieve bootstrap for time series," *Bernoulli*, 3, 123–148.

CHANG, Y., AND J. PARK (2003): "A sieve bootstrap for the test of a unit root," *Journal of Time Series Analysis*, 24, 379–400.

CHOI, E., AND P. HALL (2000): "Bootstrap confidence regions computed from autoregressions of arbitrary order," *Journal of the Royal Statistical Society*, 62, 461–477.

CHRISTENSEN, J., F. DIEBOLD, G. RUDEBUSCH, AND G. STRASSER (2007): "Multivariate Comparisons of Predictive Accuracy," *University of Pennsylvania*, manuscript.

DAVISON, A., AND D. HINKLEY (1997): *Bootstrap Methods and Their Application*. Cambridge University Press.

DIEBOLD, F., AND J. LOPEZ (1996): "Forecast evaluation and combination," .

DIEBOLD, F., AND R. MARIANO (1995): "Comparing predictive accuracy," *Journal of Business & economic statistics*, 13, 253–263.

ENGSTED, T., AND T. PEDERSEN (2014): "Bias-correction in vector autoregressive models: A simulation study," *Econometrics*, 2, 45–71.

FUERTES, A. (2008): "Sieve bootstrap t-tests on long-run average parameters," *Computational Statistics & Data Analysis*, 52, 3354–3370.

GONCALVES, S., AND T. VOGELSANG (2011): "Block bootstrap HAC robust tests: the sophistication of the naive bootstrap," *Econometric Theory*, 27, 745–791.

GONÇALVES, S., AND L. KILIAN (2007): "Asymptotic and bootstrap inference for AR (∞) processes with conditional heteroskedasticity," *Econometric Reviews*, 26, 609–641.

HANSEN, P. (2005): "A test for superior predictive ability," *Journal of Business & Economic Statistics*, 23, 365–380.

HARVEY, D., S. LEYBOURNE, AND P. NEWBOLD (1997): "Testing the equality of prediction mean squared errors," *International Journal of forecasting*, 13, 281–291.

KAPETANIOS, G., V. LABHARD, AND S. PRICE (2008): "Forecast combination and the Bank of England's suite of statistical forecasting models," *Economic Modelling*, 25, 772–792.

KAPETANIOS, G., AND Z. PSARADAKIS (2006): "Sieve bootstrap for strongly dependent stationary processes," Discussion paper.

KIM, H., AND N. DURMAZ (2012): "Bias correction and out-of-sample forecast accuracy," *International Journal of Forecasting*, 28, 575–586.

LÜTKEPOHL, H. (2005): *New Introduction to Multiple Time Series Analysis*. Springer.

MARIANO, R., AND D. PREVE (2012): "Statistical tests for multiple forecast comparison," *Journal of Econometrics*, 169, 123–130.

MEYER, M., AND J. KREISS (2014): "On the vector autoregressive sieve bootstrap," *Journal of Time Series Analysis*, fortcoming.

MÜLLER, U. K. (2014): "HAC corrections for strongly autocorrelated time series," *Journal of Business & Economic Statistics*, 32, 311–322.

NEWEY, W., AND K. WEST (1994): "Automatic lag selection in covariance matrix estimation," *The Review of Economic Studies*, 61, 631–653.

PALM, F., S. SMEEKES, AND J. URBAIN (2010): "A sieve bootstrap test for cointegration in a conditional error correction model," *Econometric Theory*, 26, 647–681.

PAPARODITIS, E. (1996): "Bootstrapping autoregressive and moving average parameter estimates of infinite order vector autoregressive processes," *Journal of Multivariate Analysis*, 57, 277–296.

POSKITT, D. (2008): "Properties of the Sieve Bootstrap for Fractionally Integrated and Non-Invertible Processes," *Journal of Time Series Analysis*, 29, 224–250.

PSARADAKIS, Z. (2003a): "A bootstrap test for symmetry of dependent data based on a Kolmogorov–Smirnov type statistic," *Communications in Statistics*, 32, 113–126.

——— (2003b): "A sieve bootstrap test for stationarity," *Statistics & Probability Letters*, 62, 263–274.

——— (2015): "Using the bootstrap to test for symmetry under unknown dependence," *Journal of Business & Economic Statistics*, fortcoming.

SHIBATA, R. (1980): "Asymptotically efficient selection of the order of the model for estimating parameters of a linear process," *The Annals of Statistics*, 8, 147–164.

STARK, T. (2010): "Realistic evaluation of real-time forecasts in the Survey of Professional Forecasters," *Federal Reserve Bank of Philadelphia Research Report*, 1.

TJØSTHEIM, D., AND J. PAULSEN (1983): "Bias of some commonly-used time series estimates," *Biometrika*, 70, 389–399.

WEST, K. (1996): "Asymptotic inference about predictive ability," *Econometrica*, 64, 1067–1084.

——— (2006): "Forecast evaluation," *Handbook of Economic Forecasting*, 1, 99–134.

WHITE, H. (2000): "A reality check for data snooping," *Econometrica*, 68, 1097–1126.

——— (2001): *Asymptotic Theory for Econometricians*. Academic Press.

YAMAMOTO, T., AND N. KUNITOMO (1984): "Asymptotic bias of the least squares estimator for multivariate autoregressive models," *Annals of the Institute of Statistical Mathematics*, 36, 419–430.

A. Do Empirically Observed Forecast Errors Follow an MA Process?

In this section, we analyze the empirically observed forecast errors in order to shed some light on the question whether the h-step ahead forecast errors follow an MA($h-1$) process or rather a more complex ARMA model, for which a Wold representation might be a reasonable approximation. This question is important not only theoretically to ensure validity of the VAR-sieve bootstrap (see Section 2) but also for the appropriate setup of the Monte Carlo experiments and, thus, reliability of their results (see Section 3).

Provided that the h-quarter ahead forecast errors follow a pure MA($h-1$) process, then only the first $h-1$ autocorrelations should be statistically significantly different from zero, whereas the partial autocorrelations should decay gradually towards zero (see Box, Jenkins, and Reinsel (2008)). The estimated sample autocorrelations and partial autocorrelations of the 4-quarter ahead SPF and AR forecast errors for all selected economic variables are depicted in Figures 1-14 (including the asymptotic 95 $\%$ confidence intervals). The results indicate that an assumption about MA processes of forecasts errors is clearly violated for all economic variables under consideration. We can also conclude that an AR representation seems to be a reasonable approximation for the forecasting errors.[19] If this hypothesis is about to be correct, then the estimated residuals from an identified AR model should be white noise. For this reason, the AR models (with automatically determined lag order) are fitted to observed forecast errors and the estimated residuals are then inspected using both the Ljung-Box portmanteau (LBQ) test and McLeod and Li portmanteau (MLQ) test with the lag order set to 10 (which is a reasonably small number as compared to the sample). The asymptotic p-values of the diagnostic tests are presented in Table 4. Although some differences between the SPF and AR models are observed, neither serial correlation nor heteroscedasticity is a serious problem for the estimated residuals. In particular, the null hypotheses about serial correlation and conditional heteroscedasticity are rejected only in less then 1/3 of the cases at the nominal level 0.10.[20]

[19] Note that similar results are obtained for the forecast errors with a different forecast horizon $h \in \{1, 2, 3\}$.
[20] Note that very similar results are obtained for other lag configurations of the diagnostic tests.

Figure 1: SPF (AAA)

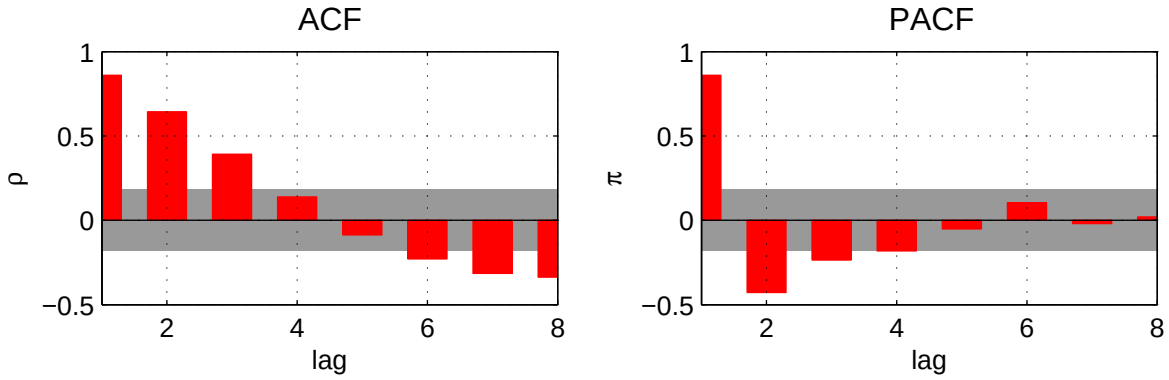

Figure 2: SPF (TBILL)

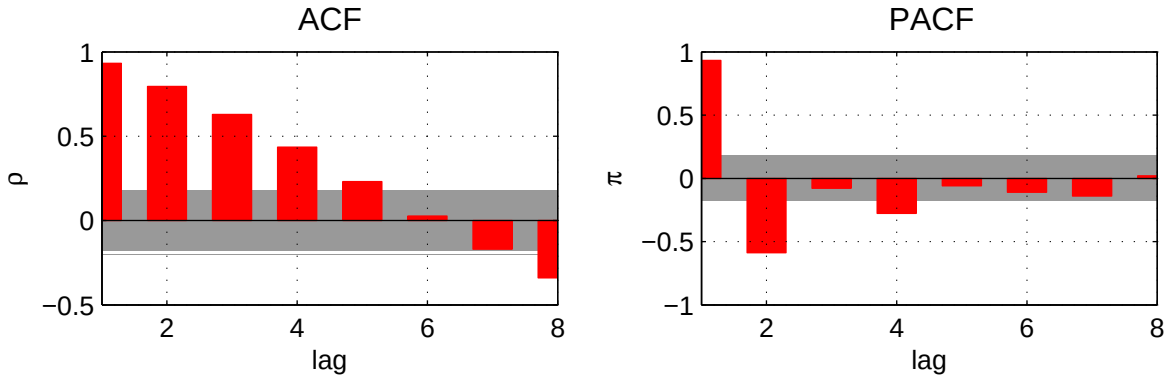

Figure 3: SPF (PGDP)

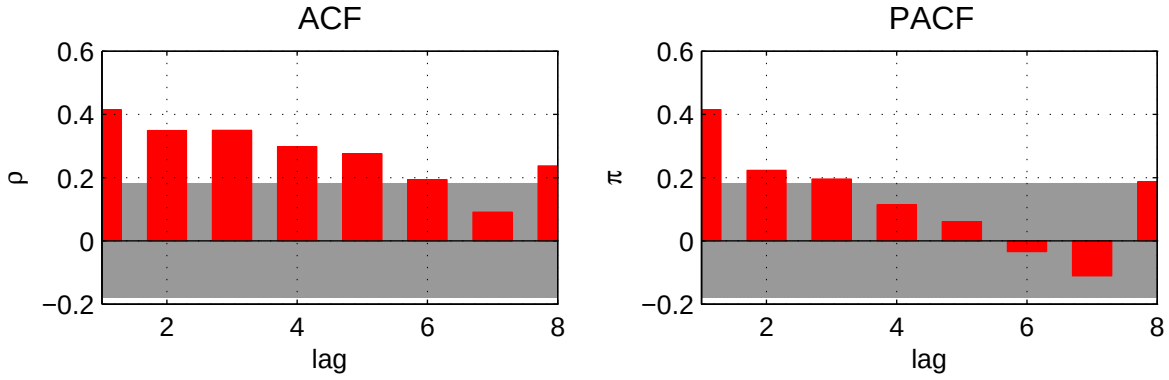

Figure 4: SPF (GDP)

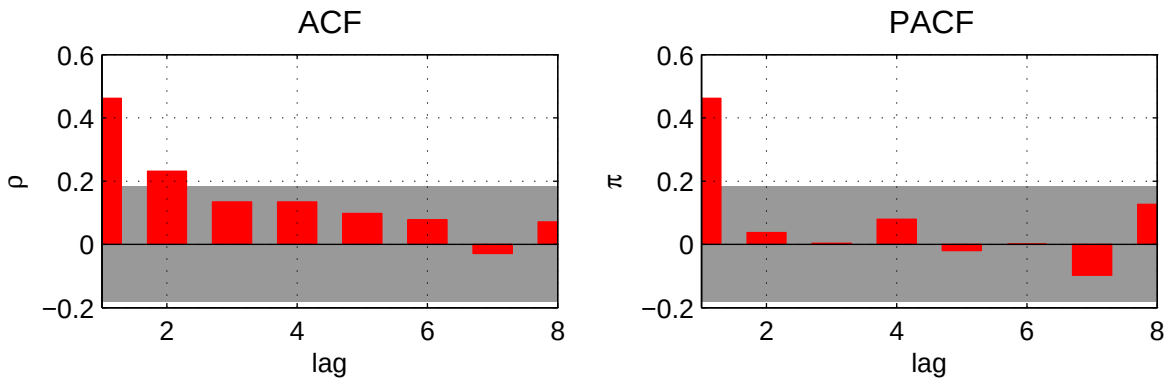

Figure 5: SPF (IP)

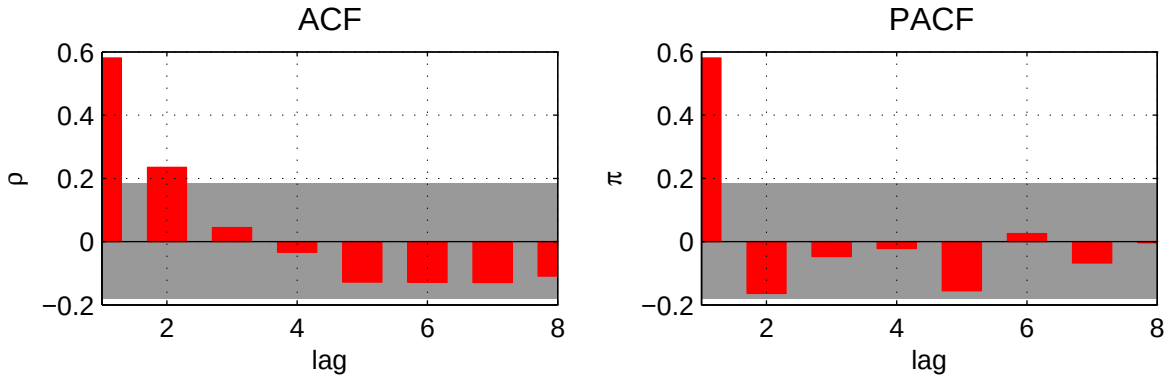

Figure 6: SPF (UR)

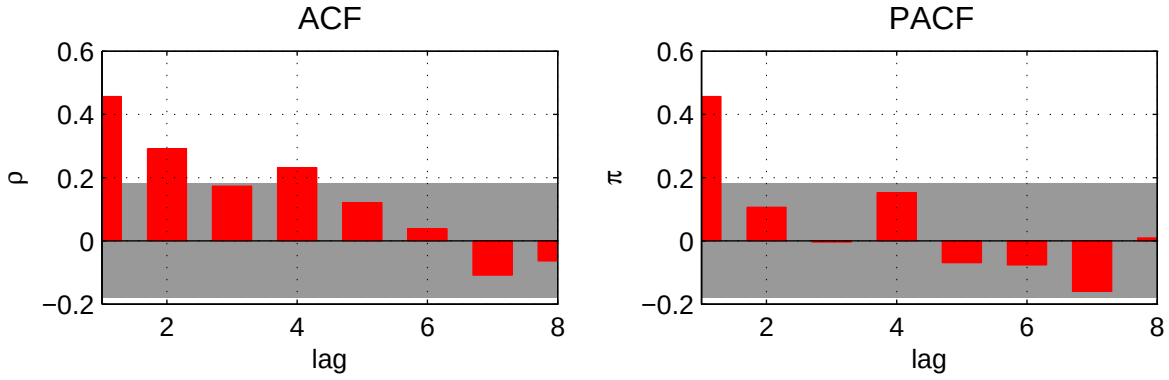

Figure 7: SPF (HOUS)

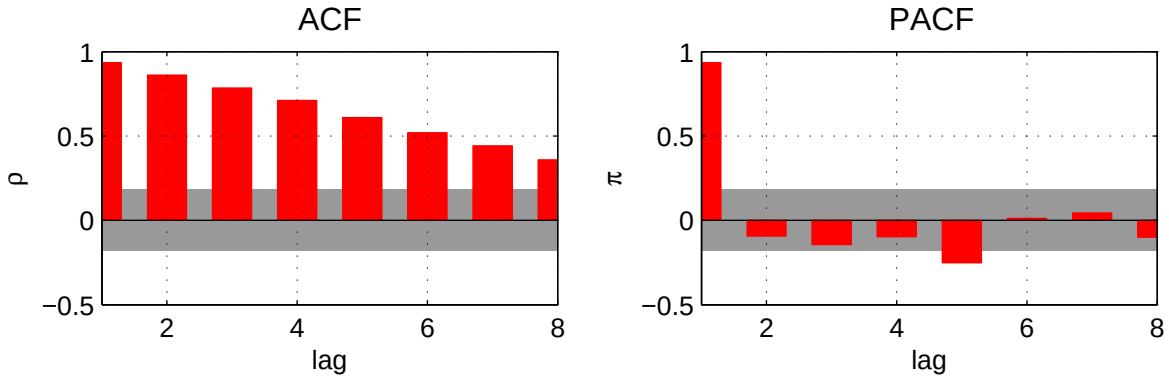

Figure 8: AR (AAA)

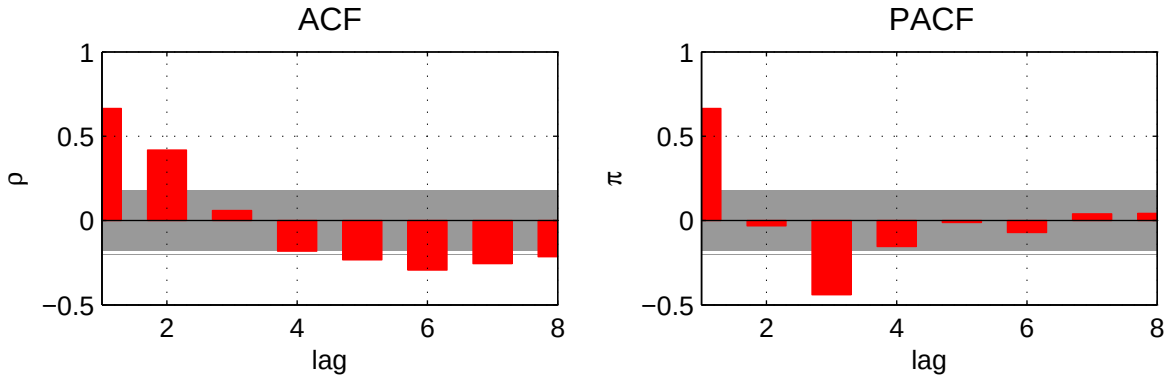

Figure 9: AR (TBILL)

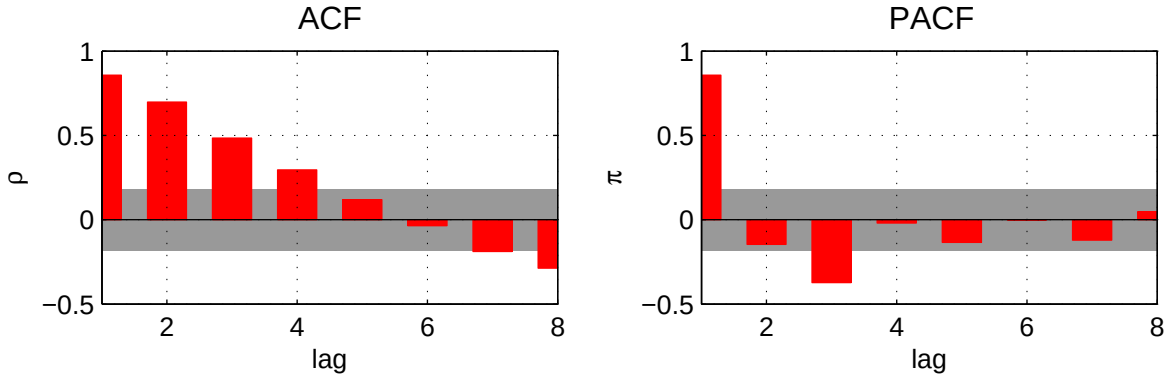

Figure 10: AR (PGDP)

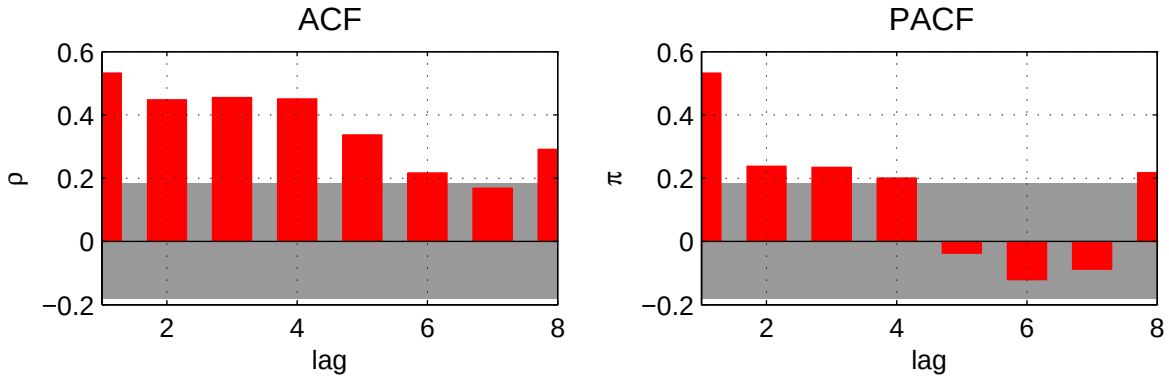

Figure 11: AR (GDP)

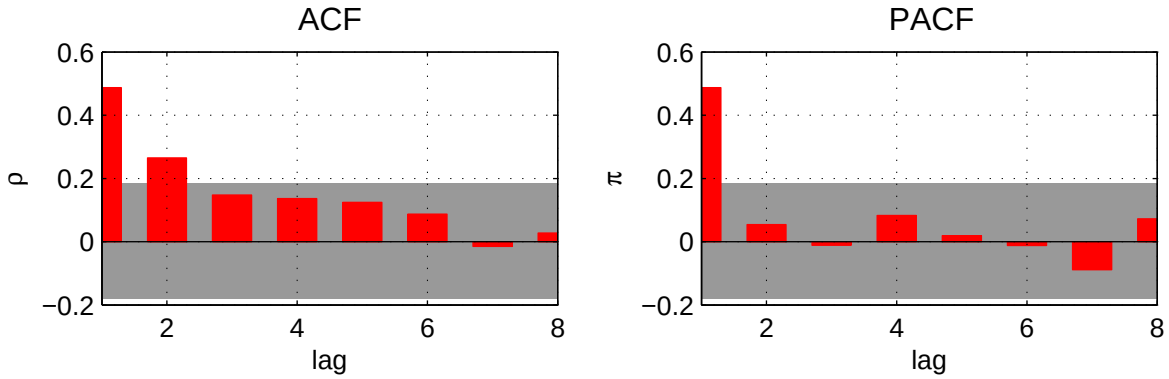

Figure 12: AR (IP)

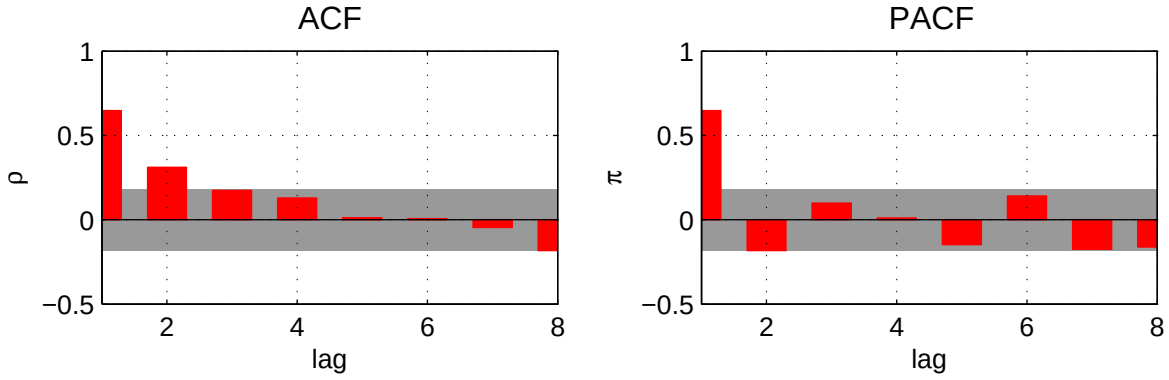

Figure 13: AR (UR)

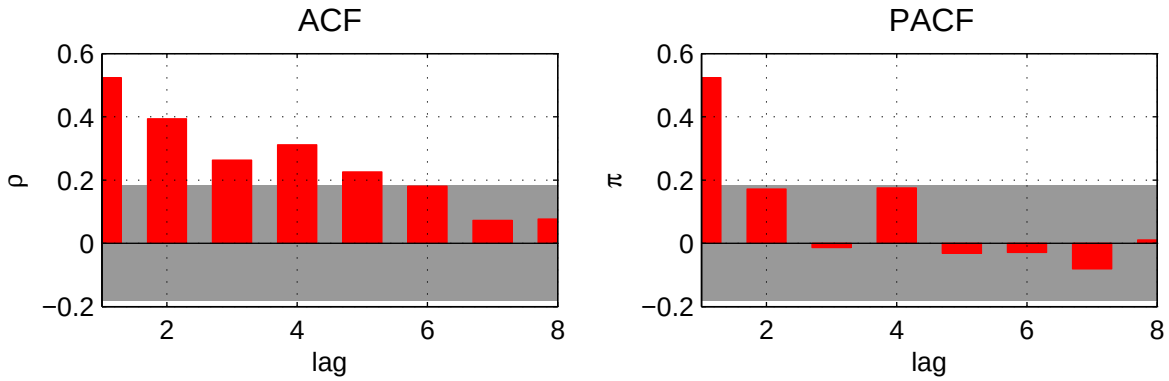

Figure 14: AR (HOUS)

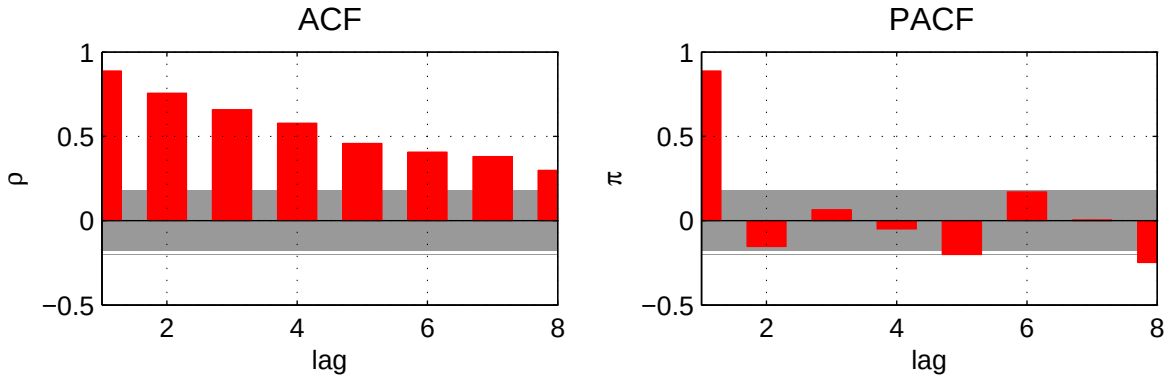

Table 4: P-values of the Diagnostic Tests

		SPF				AR			
		1983 Q1 – 2012 Q3		1991 Q1 – 2012 Q3		1983 Q1 – 2012 Q3		1991 Q1 – 2012 Q3	
variables	horizon	LBQ	MLQ	LBQ	MLQ	LBQ	MLQ	LBQ	MLQ
AAA	1	0.74	0.10	0.25	0.11	0.10	0.18	0.08	0.45
	2	0.36	0.02	0.82	0.97	0.14	0.42	0.01	0.11
	3	0.13	0.72	0.55	0.46	0.00	0.00	0.01	0.02
	4	0.33	0.85	0.38	0.66	0.00	0.03	0.23	0.16
TBILL	1	0.20	0.05	0.76	0.03	0.34	0.11	0.62	0.02
	2	0.08	0.64	0.17	0.25	0.39	0.01	0.47	0.01
	3	0.07	0.19	0.12	0.11	0.23	0.02	0.09	0.05
	4	0.14	0.83	0.03	0.35	0.28	0.61	0.38	0.10
PGDP	1	0.15	0.57	0.04	0.89	0.22	0.08	0.09	0.81
	2	0.67	0.79	0.03	0.68	0.40	0.91	0.53	0.68
	3	0.05	0.78	0.03	0.60	0.05	0.92	0.02	0.54
	4	0.43	0.56	0.05	0.82	0.14	0.38	0.08	0.48
GDP	1	0.65	0.01	0.97	0.07	0.44	0.06	0.46	0.12
	2	0.59	0.22	0.83	0.17	0.44	0.65	0.59	0.38
	3	0.63	0.69	0.89	0.42	0.14	0.55	0.30	0.65
	4	0.35	0.24	0.70	0.34	0.56	0.32	0.81	0.36
IP	1	0.72	0.26	0.57	0.12	0.21	0.00	0.33	0.00
	2	0.30	0.05	0.40	0.21	0.43	0.00	0.53	0.00
	3	0.69	0.04	0.75	0.03	0.02	0.00	0.35	0.11
	4	0.80	0.01	0.59	0.25	0.04	0.07	0.08	0.01
UR	1	0.95	0.79	0.73	0.47	0.96	0.18	0.31	0.50
	2	0.86	0.80	0.74	0.84	0.81	0.80	0.15	0.84
	3	0.37	0.82	0.15	0.94	0.56	0.74	0.04	0.97
	4	0.18	0.94	0.34	0.93	0.60	0.84	0.20	0.95
HOUS	1	0.16	0.08	0.22	0.36	0.24	0.86	0.18	0.47
	2	0.18	0.86	0.09	0.75	0.09	0.80	0.64	0.98
	3	0.57	0.25	0.18	0.28	0.82	0.69	0.64	0.64
	4	0.60	0.31	0.23	0.05	0.36	0.29	0.20	0.19