NÁRODNÁ BANKA SLOVENSKA
EUROSYSTÉM

Testing for Normality with Applications

Marián Vávra

Working Paper 1/2015

www.nbs.sk
Imricha Karvaša 1
813 25 Bratislva

reserarch@nbs.sk

March 2015

ISSN 1337-5830

Testing for Normality with Applications[1]

Working paper NBS

Marián Vávra[2]

Abstract

This paper considers the problem of testing for normality of the marginal law of univariate and multivariate stationary and weakly dependent random processes using a bootstrap-based Anderson-Darling test statistic. The finite-sample properties of the test are assessed via Monte Carlo experiments. An application to the inflation forecast errors is also presented.

JEL classification: C12, C15, C32
Key words: testing for normality; Anderson-Darling statistic; sieve bootstrap; weak dependence

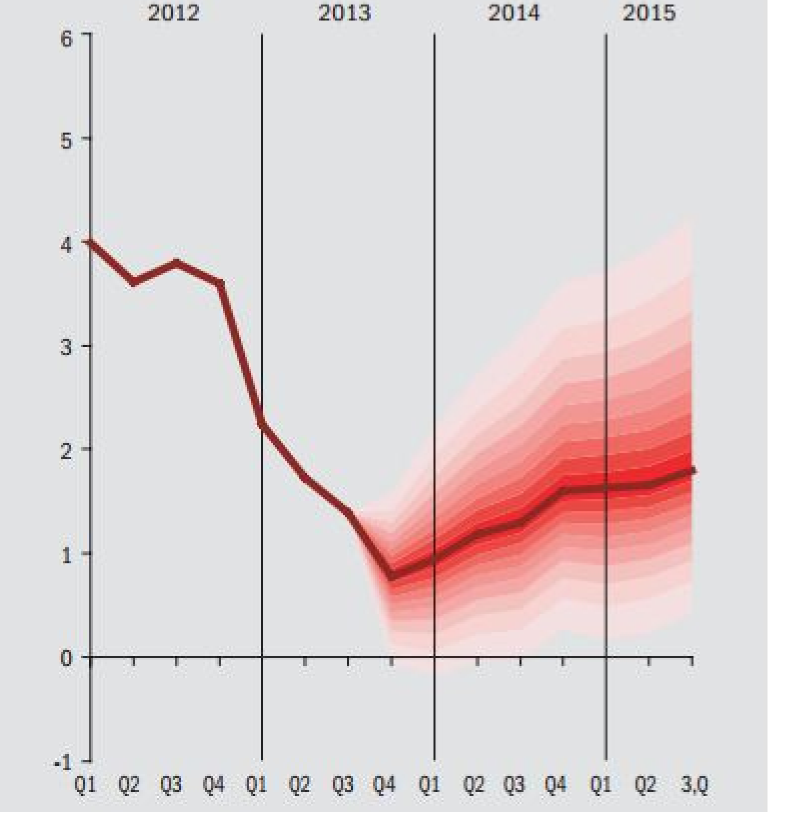
Downloadable at http://www.nbs.sk/en/publications-issued-by-the-nbs/working-papers

[1] I would like to thank 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

Monetary policy decisions are based on the expected development of key economic variables such as the inflation rate and the output gap. Since point forecasts of macroeconomic indicators provide only limited information to policy makers about the future development of the economy, some "risk" assessment of key variables in the form of (marginal) prediction bands has been found useful (see Clements and Hendry (2008, Chap. 2, 3)). Two types of prediction bands are routinely used in practice: (i) Gaussian intervals[3]; (ii) two-piece (asymmetric) Gaussian intervals[4]. Following Britton and Fisher (1998), the National Bank of Slovakia (NBS) currently relies on using a two-piece Gaussian distribution when calculating prediction bands of the inflation rate, see Figure 1. The figure depicts a system of prediction bands (the so called fan-chart) with probability coverage ranging from 10 % to 90 %.

Figure 1: NBS Prediction Bands of the Inflation Rate



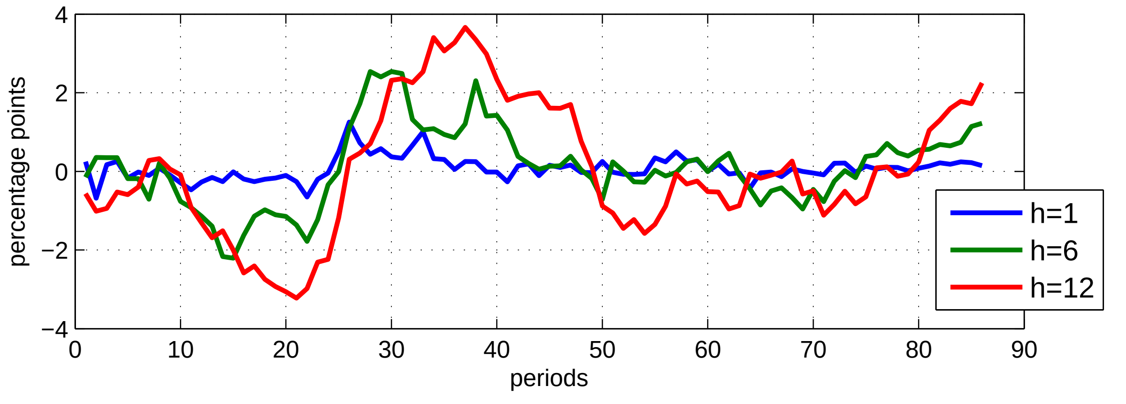
Although we do fully agree that an appropriate risk assessment can provide useful information for policy makers, the problem is that a decision about using Gaussian or asymmetric prediction bands is based on ad-hoc (and often odd) arguments rather than on formal statistical evidence.[5]

[3]See, e.g., Bank of Canada (2013, p. 23), Norges Bank (2013, p. 13), or Sveriges Riksbank (2013, p. 6)
[4]See, e.g., Bank of England (2013, p. 6)
[5]Rather surprisingly one cannot find any formal statistical evidence about normality or otherwise of the forecast

Note that testing for normality may be useful in other forecast-based applications as well. For example, Adolfson, Lindé, and Villani (2007) implicitly assume normally distributed forecast errors when evaluating density forecasts.[6] Clearly, possible density misspecification can, in turn, give rise to erroneous monetary policy decisions. Therefore, it is desirable to establish the adequacy or otherwise of normality of forecast errors before exploring more complicated asymmetric density structures.[7]

Although many statistics for testing normality have been developed in the literature (see, e.g., Mardia (1980), D'Agostino and Stephens (1986), or Thode (2002) for more details)[8], they are inappropriate for weakly dependent processes (e.g. forecast errors and other economic variables) since the presence of dependence in observations invalidates critical values of standard tests derived under an i.i.d assumption. Our point is illustrated in Figure 2 where the NBS inflation forecast errors are depicted for the selected horizons. Apart from obvious dependence in observations, it is clear that persistence of the forecast errors increases significantly with the forecast horizon (see Section 4 for a discussion).

Figure 2: NBS Inflation Forecast Errors



Only recently, a modified Jarque-Bera normality test has been developed for weakly dependent (w.d.) stochastic processes as well (see Lobato and Velasco (2004) and Bai and Ng (2005), henceforth BN). However, the finite sample performance of the BN test statistic is not convincing. First, the power of the BN test fails for persistent stochastic processes (see Table 3 in this paper). Second, the BN test requires the existence of the first eight moments to provide valid (asymptotic) inference. This requirement is, however, in sharp contrast with empirical findings

errors in the above cited Banks' reports.

[6]Some other examples about the usefulness of testing for normality can be found in Kilian and Demiroglu (2000).

[7]It is worth remarking that even if the historical forecast errors are (approximately) Gaussian, the Bank may still want to use some asymmetric probability distribution allowing for signaling the direction of potential risks to market participants. An example of this type might be uncertainty about the oil and gas prices caused by the current Russian–Ukraine conflict. Nevertheless, it is necessary the present the results in a convincing way.

[8]The following three main classes of test statistics for checking normality of independently and identically distributed (i.i.d.) random variables have become popular in the literature: (i) tests based on the empirical distribution function (see, e.g., Anderson and Darling (1952)) or the empirical characteristic function (see, e.g., Hall and Welsh (1983)); (ii) order statistics (see, e.g., Shapiro and Wilk (1965)); and (iii) tests based on skewness and kurtosis (see, e.g., Jarque and Bera (1980)).

about economic time series (see, e.g., Jansen and Vries (1991), Loretan and Phillips (1994), or Runde (1997)). Moment condition failure of the test is very likely to lead to misleading inference (see Psaradakis and Vávra (2014) for a related example).

This paper contributes to the literature by considering a bootstrap-based Anderson-Darling (BAD) statistic for testing for normality of scalar and vector stationary and weakly dependent stochastic processes. The proposed test considered here has several features that make it attractive for applications. First, in addition to having an intuitive interpretation and being easy to compute, the test does produce very good size and power results which are superior to the BN test. Second, the proposed BAD test requires only the first four moments to be finite which is in line with stylized facts about economic time series. Third, following the Cramér-Wold device, it is shown that the BAD test can be easily used to test for normality of vector stochastic processes as well.

The paper is organized as follows. A bootstrap-based Anderson-Darling type statistic is discussed in Section 2. Section 3 examines the finite-sample properties of the proposed test by means of Monte Carlo experiments. Section 4 summarizes and concludes.

2. Anderson-Darling Test

2.1 Univariate Approach

It is assumed through out the paper that the underlying stochastic process is a real-valued stationary and weakly dependent process allowing for a Wold representation given by

$$X_t = \mu + \sum_{j=1}^{\infty} \psi_j \epsilon_{t-j} + \epsilon_t, \quad t \in \mathbb{Z}, \tag{1}$$

where $\mu \in \mathbb{R}$, the roots of the lag polynomial $\psi(q) = 1 - \sum_{j=1}^{\infty} \psi_j q^j$ lie outside the unit disk and $\sum_{j=1}^{\infty} j|\psi_j| < \infty$, the error sequence $\{\epsilon_t : t \in \mathbb{Z}\}$ is assumed to be stationary and ergodic such that $\mathbb{E}(\epsilon_t|\mathcal{F}_{t-1}) = 0$, $\mathbb{E}(\epsilon_t^2|\mathcal{F}_{t-1}) = s^2 < \infty$, where $\mathcal{F}_t = \{\epsilon_t, \epsilon_{t-1}, \ldots\}$ is the sigma-field, $\mathbb{E}(\epsilon_t^4) < \infty$ and the density function $f(\epsilon_t)$ is absolutely continuous. A Wold decomposition represents a fairly large class of stochastic processes including, for instance, causal and invertible linear ARMA models often applied in economics and finance. Note also that under an additional mild assumption about invertibility, it is easy to show that the process in (1) can be rewritten into the form of an AR(∞) model

$$X_t = c + \sum_{j=1}^{\infty} \phi_j X_{t-j} + \epsilon_t, \quad t \in \mathbb{Z}. \tag{2}$$

The problem of interest is to test the null hypothesis that one-dimensional marginal distribution function $F(z) = \mathbb{P}((X_t - \mu)/\sigma \leq z)$, $z \in \mathbb{R}$, is Gaussian. A popular test based on a (weighted) quadratic distance between the empirical and hypothetical distribution function is the Anderson-Darling test given by

$$\mathcal{A}_n = n \int_{\mathbb{R}} \frac{(F_n(z) - \Phi(z))^2}{\Phi(z)(1 - \Phi(z))} \, \mathrm{d}\Phi(z), \quad z \in \mathbb{R}, \tag{3}$$

where Φ denotes a standard normal distribution and F_n is the empirical distribution function (EDF) associated with $\{X_t : t = 1, \ldots, n\}$

$$F_n(z) = \frac{1}{n} \sum_{t=1}^{n} I\left(\frac{X_t - \mu}{\sigma} \leq z\right), \quad z \in \mathbb{R}, \tag{4}$$

where $I(\cdot)$ is a standard indicator function and $\mu = \mathbb{E}(X_t)$ and $\sigma = \sqrt{\mathrm{var}(X_t)}$. In finite sample, the following formula is used to calculate the Anderson-Darling test

$$\hat{\mathcal{A}}_n = -n - \frac{1}{n} \sum_{i=1}^{n} (2i - 1)[\log(\hat{Y}_{(i)}) - \log(1 - \hat{Y}_{(n-i+1)})], \tag{5}$$

where $\hat{Y}_{(i)} = \Phi((X_{(i)} - \hat{\mu})/\hat{\sigma})$, $X_{(i)}$ denotes the i-th order statistic and $\hat{\mu}$ and $\hat{\sigma}$ are the estimated location and scale parameters. Note that consistent estimators of the location μ and scale σ can be obtained by a sample average and a standard deviation (the results follow from Theorem 3.34 in White (2001, p. 44)).

Our choice of using the Anderson-Darling test statistic is motivated by the following three arguments. First, Shapiro, Wilk, and Chen (1968) show that the EDF-based test statistics (such as Kolmogorov-Smirnov, Cramér-von Mises, or Anderson-Darling tests) are more powerful as compared to moment-based test statistics (based on coefficients of skewness and kurtosis) when testing for normality. Second, Noceti, Smith, and Hodges (2003) show that the Anderson-Darling test is one of the most powerful tests among the EDF-based tests. Third, the Anderson-Darling test (and other EDF-based tests as well) has a pleasant property of being consistent against any alternative (continuous) distribution (see DasGupta (2008, Sec. 26.5)).

In large samples as $n \to \infty$ and under assumptions that $\{X_t : t = 1, \ldots, n\}$ is a sequence of i.i.d random variables and location and scale parameters (i.e. μ and σ) are known, the asymptotic null distribution of $\mathcal{A}_n$ can be derived (see Anderson and Darling (1952)).[9] Complications arise when the location and scale parameters in (4) are unknown and must be estimated. Then not only the exact distribution of the Anderson-Darling statistic but also its asymptotic distribution

[9] In particular, using a change of variable ($\omega = \Phi(z)$), $\mathcal{A}_n$ takes the following form

$$\mathcal{A}_n = \int_0^1 \frac{U_n^2(\omega)}{\omega(1 - \omega)} \, \mathrm{d}\omega,$$

where $U_n(\omega) = \sqrt{n}(G_n(\omega) - \omega)$ and $G_n(\omega) = n^{-1} \sum_{t=1}^{n} I(\Phi(z_t) \leq \omega)$ for each $\omega \in [0, 1]$. Then, it holds that $U_n \xrightarrow{d} U$, where U is the Brownian bridge which immediately implies statistic $\mathcal{A}_n \xrightarrow{d} \int_0^1 \frac{U^2(\omega)}{\omega(1-\omega)} \, \mathrm{d}\omega$.

depends on the shape of the null distribution (see de Wet and Randles (1987) for details). Since the test statistic in (3) depends on unknown parameters governing the stochastic process $\{X_t : t \in \mathbb{Z}\}$, an appropriate bootstrap method should be used to calculate reliable critical values of the test. Although there exist various bootstrap methods for weakly dependent data in the literature (see Lahiri (2003) for a comprehensive treatment), the AR-sieve bootstrap seems to be a preferred method. The main advantage of the AR-sieve is that it allows us to impose the normality condition under the null and it is able to replicate the (linear) dependence structure in data sufficiently well.[10] Moreover, due to its simplicity and good finite sample properties, the AR-sieve bootstrap has become popular in the time series literature (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), Palm, Smeekes, and Urbain (2010)).

The procedure applied to generate bootstrap based critical values of the Anderson-Darling test statistic (BAD) is summarized in the following algorithm.

Algorithm 1 (i) Select an appropriate lag order p of an AR model using the Akaike information criterion, where the lag order is restricted by $0 \le p \le 10\log_{10}(n)$, where n denotes the sample size. Note that other lag order selection criteria such as BIC or HQ could be used, but since the process $\{X_t\}$ in (2) is not assumed to be of finite dimension, the AIC is asymptotically efficient (see Shibata (1980)).

(ii) Estimate the unknown AR(p) model parameters by the OLS method. In contrast to Bühlmann (1997), who implemented the Yule-Walker (YW) estimator, we rely on the standard OLS estimator. The main reason for doing so is that the OLS estimator produces superior results as compared to the YW estimator (see Tjøstheim and Paulsen (1983)).

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

$$\hat{\epsilon}_t = X_t - \hat{c} - \sum_{i=1}^{p} \hat{\phi}_i X_{t-i}.$$

(iv) Under the null hypothesis of marginal normality, the hypothesized distribution equals to a standard normal distribution Φ. Therefore, consistently with the null, draw independent random errors $\epsilon_t^* \sim N(0, \hat{s}^2)$, for $t = 1, \ldots, n+100$, where $\hat{s}^2 = (n-2p-1)^{-1} \sum_{t=p+1}^{n} \hat{\epsilon}_t^2$.

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

$$X_t^* = \hat{c} + \sum_{i=1}^{p} \hat{\phi}_i X_{t-i}^* + \epsilon_t^*,$$

[10] Note that Gaussian nonlinear processes appear to be the exception rather than the rule which implies that a linear (auto-regressive) filter is fully sufficient under the null hypothesis of marginal normality.

where the process is initiated by a vector of sample averages: $(X^*_{-p+1}, \ldots, X^*_0) = (\bar{X}, \ldots, \bar{X})$. The first 100 data points are then discarded in order to eliminate start-up effects and the remaining n data points are used.

(vi) Construct a bootstrap analogy of the BAD test statistic $\mathcal{A}^*_n$ calculated from a bootstrap sample $\{X^*_t : t = 1, \ldots, n\}$. It is worth noting that using Gaussian replicates in Step (v) reduces the simulation error of the bootstrap procedure and, thus makes the procedure more efficient as compared to relying on bootstrap replicates (see Davison and Hinkley (1997, pp. 31-37) for details).

(vii) Repeat steps (iv)–(vi) independently B times to get a sample of the BAD statistics $\{\mathcal{A}^*_{n,i} : i = 1, \ldots, B\}$. Then, the sampling distributions of the BAD test statistic is approximated by the empirical distribution functions associated with $\{\mathcal{A}^*_{n,i} : i = 1, \ldots, B\}$: $H^*_n(u) = B^{-1} \sum_{i=1}^{B} I(\mathcal{A}^*_{n,i} \leq u)$. Finally, a bootstrap test of the nominal level α rejects the null hypothesis of normality if

$$\hat{\mathcal{A}}_n > \inf\{u : H^*_n(u) \geq (1-\alpha)\},$$

where $\hat{\mathcal{A}}_n$ is the BAD test statistic obtained form the observed sample $\{X_t : t = 1, \ldots, n\}$.

Remark 1: Kreiss, Paparoditis, and Politis (2011) and Jentsch and Politis (2013) prove that the distribution of a relevant statistic must only depend on the first and second moments of the process to ensure asymptotic validity of the AR-sieve bootstrap (see (4)).

Remark 2: An important question is to what extent the AR-sieve bootstrap works for stochastic processing not stemming from the Wold representation. Bickel and Bühlmann (1997) explain that the closure of the Wold representation is fairly large. It means that for any non-linear stochastic process there exist another process in the closure of linear processes having identical sample paths with probability exceeding $1/e \approx 0.37$. This finding implies that the AR-sieve bootstrap is very likely to give satisfactory results even for stochastic processes deviating from the Wold representation.

2.2 Multivariate Approach

Although the above approach could be used for testing joint normality of vector stochastic processes as well, the estimation of a multivariate version of the EDF-based tests is computationally intensive (see Justel, Peña, and Zamar (1997) for details). Therefore, some dimensionality reduction technique is desirable. A natural solution utilizes the fact that if a $(k \times 1)$ random vector $\boldsymbol{x}_t$ is distributed as $N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ under the null hypothesis, then $y_t = (\boldsymbol{x}_t - \boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}_t - \boldsymbol{\mu})$ is distributed as χ^2 with k degrees of freedom (see, e.g., Malkovich and Afifi (1973), Koziol (1982), or Paulson, Roohan, and Sullo (1987)). This dimensionality reduction method is not, however, convenient in our case since the quadratic-form transformation, which is a non-linear transformation, precludes the use of the AR-sieve bootstrap on $\{y_t : t \in \mathbb{Z}\}$ when calculating the critical values of the BAD test. Instead, a VAR-sieve bootstrap of Paparoditis (1996) on $\{\boldsymbol{x}_t : t \in \mathbb{Z}\}$ should be implemented. This approach, however, imposes a restriction on the relationship between a number of observations and a number of variables. In order to bypass this limitation, we incorporate a simple, yet flexible, method of dimensionality reduction which is in spirit similar to a quadratic-form transformation. In particular, our method relies on the well-known Cramér-Wold device and skewness-based linear transformation.

Theorem 1 *For a* $(k \times 1)$ *random vectors* $\boldsymbol{x}_t = (X_{1t}, \ldots, X_{kt})'$ *and* $\boldsymbol{x} = (X_1, \ldots, X_k)'$*, a necessary and sufficient condition for* $\boldsymbol{x}_t \xrightarrow{d} \boldsymbol{x}$ *with a joint distribution* $F(\mathbf{x})$ *as* $t \to \infty$ *is that* $\boldsymbol{\lambda}'\boldsymbol{x}_t \xrightarrow{d} \boldsymbol{\lambda}'\boldsymbol{x}$ *with a marginal distribution function* $F(\boldsymbol{\lambda}'\boldsymbol{x})$ *for each* $\boldsymbol{\lambda} \in \boldsymbol{R}^k$*.*

Proof. *See Billingsley (1995, p. 383) for a proof.*

The main conclusion of Theorem 1 is that, after an appropriate linear transformation, one can use the advantage of the test statistic defined in (3) when testing for joint normality of vector processes as well. An ultimate question here is how to determine the weighting vector $\boldsymbol{\lambda}$ in practice. Our reasoning is as follows: Since components in the vector $\boldsymbol{x}_t$ are assumed to be cross-correlated, a natural solution is to orthogonalize and then aggregate them. The orthogonalization can be done, for instance, by the eigenvalue decomposition. When doing so, a consistent estimator of the long-run variance-covariance matrix of $\boldsymbol{x}_t$ is required. Motivated by the literature on estimation of asymptotic covariance matrices in the presence of weak dependence, we consider an estimator

$$\hat{\boldsymbol{\Sigma}} = \hat{\boldsymbol{\Gamma}}_0 + \sum_{h=1}^{m} W(h/m)(\hat{\boldsymbol{\Gamma}}_h + \hat{\boldsymbol{\Gamma}}_h'), \tag{6}$$

where $W(\cdot)$ are Bartlett weights, m is a real-valued bandwidth such that $m \to \infty$ and $m/T \to 0$ as $T \to \infty$, and $\hat{\boldsymbol{\Gamma}}_h = T^{-1} \sum_{t=h+1}^{T} (\boldsymbol{x}_t - \bar{\boldsymbol{x}})(\boldsymbol{x}_{t-h} - \bar{\boldsymbol{x}})'$, where $\bar{\boldsymbol{x}}$ is a sample average. Finally, using the eigenvalue decomposition one can estimate a $(k \times k)$ matrix $\hat{\boldsymbol{P}}$ such that $\hat{\boldsymbol{\Sigma}} = \hat{\boldsymbol{P}}\hat{\boldsymbol{P}}'$ (see Schott (2005, Chap. 4) for details).[11] Consistency of the estimator in (6) follows from well-known results on covariance matrix estimation (see Andrews (1991, Theorem 1)).

Orthogonalized components $\boldsymbol{z}_t = \hat{\boldsymbol{P}}^{-1}\boldsymbol{x}_t = (Z_{1t}, \ldots, Z_{kt})'$ are then aggregated using the skewness-based weighting function defined as $\boldsymbol{w} = [w_i]$, where $w_i = 1$ if $skew(Z_{it}) \geq 0$ and $w_i = -1$ if $skew(Z_{it}) < 0$, for $i \in \{1, \ldots, k\}$, and $skew(\cdot)$ can by any measure of skewness (e.g. the coefficient of skewness). The proposed linear dimensionality reduction preserves a departure from normality when aggregating positively and negatively skewed random variables. Note that if all components of $\boldsymbol{z}_t$ are, for example, positively skewed, then the weighting func-

[11] Note that the Cholesky factorization may be used as well.

tion $\boldsymbol{w}$ becomes a $(k \times 1)$ vector of ones, which means that the aggregation is equivalent to a summation of individual components in $\boldsymbol{z}_t$.[12] Finally, one can then apply the testing procedure described in Algorithm 1 to the transformed scalar process $X_t = \boldsymbol{\lambda}'\boldsymbol{x}_t$, where $\boldsymbol{\lambda}' = \boldsymbol{w}'\hat{\boldsymbol{P}}^{-1}$.

3. Monte Carlo Simulations

3.1 Experimental Design

In this section, the size and power properties of the proposed BAD test statistic are assessed by means of Monte Carlo experiments. The experiments are based on artificial data generated according to the models:

M0: $X_t = \epsilon_t$

M1: $X_t = 0.5X_{t-1} + \epsilon_t$

M2: $X_t = 0.8X_{t-1} + \epsilon_t$

M3: $X_t = 0.8X_{t-1} - 0.4X_{t-2} - 0.5\epsilon_{t-1} + \epsilon_t$

M4: $X_t = 0.5X_{t-1} - 0.3X_{t-1}\epsilon_{t-1} + \epsilon_t$

M5: $X_t = 1.5S_t - 0.5(1 - S_t) + 0.5X_{t-1} + \epsilon_t$

M6: $\boldsymbol{x}_t = \begin{pmatrix} 0.4 & 0.3 \\ 0.3 & 0.4 \end{pmatrix} \boldsymbol{x}_{t-1} + \begin{pmatrix} a_t \\ \varepsilon_t \end{pmatrix}$

In each case, $\{\epsilon_t\}$ are i.i.d. random variables having zero mean and unit variance. Models M1 – M3 are linear ARMA models with different persistence, model M4 is a bilinear model, model M5 is a Markov-switching AR model, and model M6 is a bivariate linear VAR model. It is important to point out that models M4 and M5 do not fulfill the regularity assumptions related to the Wold representation (see Section 2 for details). These two models are considered as a form of a robustness check - to examine how the AR-sieve bootstrap works under non-standard circumstances. This property is important especially for empirical applications where it might be difficult to check in advance the validity of the test assumptions. The DGPs are set in such a way that the marginal distribution of M4 and M5 models is non-normal even under Gaussian innovations.

The distribution of ϵ_t is either Gaussian or belongs to the family of generalized lambda distributions. The latter may be specified via its quantile function, which is $Q(\nu) = \lambda_1 + \{\nu^{\lambda_3} - (1 - \nu)^{\lambda_4}\}/\lambda_2$, $\nu \in [0, 1]$ (see Ramberg and Schmeiser (1974)); the parameter values used in the experiments are taken from Bai and Ng (2005) and can be found in Table 2. In particular, apart from a Gaussian distribution of innovations (denoted as "N"), we consider three symmetric but

[12] Note that this type of aggregation is used, for instance, in Newey and West (1994, pp. 633–634) when calculating the long-run covariance matrices using the automatic lag order procedure.

heavy-tailed distributions (denoted as "S1", "S2", "S3") and three asymmetric distributions (denoted as "A1", "A2", "A3"). The distribution of a_t in model M6 is Gaussian unless otherwise stated.

Experiments proceed by generating 1,000 independent artificial time series $\{X_t\}$ and $\{\boldsymbol{x}_t\}$ of length $100 + n$, with $n \in \{100, 200, 500\}$, 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 BAD test statistic defined in (3). For the selected DGPs (i.e. models M0, M1 and M2), the BAD test is compared with the modified Jarque-Bera test discussed in Bai and Ng (2005, p. 52) (henceforth the BN test). The BN test is based on the coefficients of skewness and kurtosis with an appropriately estimated variance-covariance matrix. The rejection frequencies of the BN test for models M0, M1 and M2 reported in Table 3 are reproduced from Bai and Ng (2005, p. 57).

3.2 Simulation Results

The Monte Carlo rejection frequencies of tests of nominal level 0.05 are reported in Tables 3 – 4. The results suggest the following:

(i) For all relevant data points (see models M0, M1, M2, M3 with Gaussian innovations), the proposed BAD test has empirical levels very close to the nominal level 0.05, regardless of the sample size n, whereas the BN test exhibits a size distortion especially for strongly persistent processes (see M2 model results).

(ii) The BAD test has very good power properties (see models M0, M1, M2, M3 with non-normal innovations S1–A3), even for the smallest of the sample sizes considered in the paper. For example, in the case of M1, the rejection frequency of the BAD test is 0.81 under A1 when $n = 100$, whereas only 0.21 for the BN test. Surprisingly, we find that significant differences between the BAD and BN tests are observed even in large samples (i.e. $n = 500$). For instance, in the case of M1, the rejection frequency of the BAD test is 1.00 under S3 when $n = 500$, whereas only 0.33 for the BN test.

(iii) The BAD test produces very encouraging results when testing for joint normality of vector random variables. In particular, the BAD test has very good size and power properties for a bivariate VAR model (model M6) even if only one component of model innovations is drawn from a non-normal distribution, whereas the other random component is drawn from a standard normal distribution (see the Monte Carlo setup).

(iv) The BAD test does perform very well even for DGPs deviating from the Wold representation. For example, in the case of M4 (a bilinear model), the rejection frequency of the BAD test exceeds 0.68 for non-normal innovations, regardless of the sample size.

4. Are the NBS Inflation Forecast Errors Gaussian?

Following Britton and Fisher (1998), the National Bank of Slovakia (NBS) currently relies on using a two-piece Gaussian distribution when calculating the prediction bands of the inflation rate (see National Bank of Slovakia (2012, p. 17)). The main task of this exercise is to investigate whether marginal and joint normality (or lack of them) are common characteristic features of the NBS inflation forecast errors. Put differently, we investigate whether or not Gaussian prediction bands of the NBS inflation forecasts might be appropriate or not.

Let us define the forecast error as $X_t(h) = \pi_{t+h} - \pi_t(h)$, for the horizon $h \in \{1, \ldots, 12\}$, where $\pi_t(h)$ stands for the h-step ahead forecast of the (year-on-year) CPI inflation rate and π_{t+h} denotes the actual inflation rate. The series span the period January 1996 – January 2014 (i.e. $n = 86$) for each forecast horizon h. The NBS inflation forecasts for the selected horizons $h \in \{1, 6, 12\}$ are depicted in Figure 2. It can be easily concluded from the figure that persistence of the forecast errors $X_t(h)$ increases significantly with the forecast horizon h (the first-order autocorrelation coefficient of $X_t(h)$ increases from 0.59 for $h = 1$ to 0.97 for $h = 12$.)

Table 1: P-values of the BAD Test for Marginal and Joint Normality of the Inflation Forecast Errors

hypothesis	horizon	$B = 1000$	$B = 5000$	$B = 10000$
marginal	$h = 1$	0.04	0.04	0.04
	$h = 2$	0.36	0.35	0.35
	$h = 3$	0.24	0.24	0.24
	$h = 4$	0.40	0.41	0.40
	$h = 5$	0.49	0.50	0.50
	$h = 6$	0.63	0.61	0.63
	$h = 7$	0.91	0.89	0.87
	$h = 8$	0.84	0.86	0.86
	$h = 9$	0.87	0.86	0.87
	$h = 10$	0.73	0.74	0.73
	$h = 11$	0.57	0.59	0.57
	$h = 12$	0.56	0.59	0.58
joint	$h = 1, \ldots, 12$	0.65	0.62	0.62

We focus on testing for both marginal and joint normality of the forecast errors. In particular, the null hypotheses are set as follows:

(i) marginal: $H_0 : F(X_t(h)) = N(0, \sigma_h^2)$ against $H_1 : F(X_t(h)) \neq N(0, \sigma_h^2)$ for $h \in \{1, \ldots, 12\}$, where $0 < \sigma_h^2 < \infty$.

(ii) joint: $H_0 : F(X_t(1), \ldots, X_t(12)) = N(\mathbf{0}, \mathbf{\Sigma})$ against $H_1 : F(X_t(1), \ldots, X_t(12)) \neq N(\mathbf{0}, \mathbf{\Sigma})$, where $\mathbf{\Sigma}$ is a symmetric and positive-definite matrix. A dimensionality reduction of the forecast errors proceeds as described in Section 2.2 when testing for joint normality.

The bootstrap-based p-values of the BAD test, based on a different number of bootstrap replications $B \in \{1000, 5000, 10000\}$, are presented in Table 1.[13] The results reveal that apart from the 1-month ahead forecast error (i.e. $X_t(1)$), neither marginal nor joint hypothesis of normality can be rejected at the usual significance level 0.05, regardless of a number of bootstrap replications B.[14] Put differently, the statistical analysis of the historical NBS inflation forecast errors does not support the use of two-piece (asymmetric) Gaussian prediction bands.

5. CONCLUSION

This paper has considered the bootstrap-based Anderson-Darling test statistic for testing for normality of the marginal law of strictly stationary and weakly dependent scalar and vector stochastic processes. We have shown that the BAD test is intuitive, easy to implement, and requires only the first four moments to be finite to provide valid statistical inference. Monte Carlo results have revealed that the BAD test has very good size and power properties in finite samples which unambiguously outperform those obtained from the BN test discussed in Bai and Ng (2005).

[13] Note that since the forecast errors are assumed to be zero mean processes, the constant term is omitted in (2) when testing for normality.

[14] It can be concluded that no gain is achieved from using more than 1000 bootstrap replications.

REFERENCES

ADOLFSON, M., J. LINDÉ, AND M. VILLANI (2007): "Forecasting performance of an open economy DSGE model," *Econometric Reviews*, 26, 289–328.

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.

ANDERSON, T., AND D. DARLING (1952): "Asymptotic theory of certain "goodness of fit" criteria based on stochastic processes," *The Annals of Mathematical Statistics*, 23, 193–212.

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

BAI, J., AND S. NG (2005): "Tests for skewness, kurtosis, and normality for time series data," *Journal of Business and Economic Statistics*, 23, 49–60.

Bank of Canada (2013): "Monetary Policy Report," 4/2013.

Bank of England (2013): "Inflation Report," 4/2013.

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

BILLINGSLEY, P. (1995): *Probability and Measure*. Wiley.

BRITTON, E., AND P. FISHER (1998): "The Inflation Report projections: understanding the fan chart," *Bank of England Quarterly Bulletin*, 1, 30–37.

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.

CLEMENTS, M., AND D. HENDRY (2008): *A Companion to Economic Forecasting*. Wiley.

D'AGOSTINO, R., AND M. STEPHENS (1986): *Goodness-of-Fit Techniques*. CRC.

DASGUPTA, A. (2008): *Asymptotic Theory of Statistics and Probability*. Springer.

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

DE WET, T., AND R. RANDLES (1987): "On the effect of substituting parameter estimators in limiting χ^2 U and V statistics," *The Annals of Statistics*, 15, 398–412.

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

HALL, P., AND A. WELSH (1983): "A test for normality based on the empirical characteristic function," *Biometrika*, 70, 485–489.

JANSEN, D., AND C. D. VRIES (1991): "On the frequency of large stock returns: Putting booms and busts into perspective," *The Review of Economics and Statistics*, 73, 18–24.

JARQUE, C., AND A. BERA (1980): "Efficient tests for normality, homoscedasticity and serial independence of regression residuals," *Economics Letters*, 6, 255–259.

JENTSCH, C., AND D. POLITIS (2013): "Valid Resampling of Higher-Order Statistics Using the Linear Process Bootstrap and Autoregressive Sieve Bootstrap," *Communications in Statistics-Theory and Methods*, 42, 1277–1293.

JUSTEL, A., D. PEÑA, AND R. ZAMAR (1997): "A multivariate Kolmogorov-Smirnov test of goodness of fit," *Statistics & Probability Letters*, 35, 251–259.

KAPETANIOS, G., AND Z. PSARADAKIS (2006): "Sieve bootstrap for strongly dependent stationary processes," Discussion paper, Working Paper, Department of Economics, Queen Mary, University of London.

KILIAN, L., AND U. DEMIROGLU (2000): "Residual-based tests for normality in autoregressions: Asymptotic theory and simulation evidence," *Journal of Business & Economic Statistics*, 18, 40–50.

KOZIOL, J. (1982): "A class of invariant procedures for assessing multivariate normality," *Biometrika*, 69, 423–427.

KREISS, J., E. PAPARODITIS, AND D. POLITIS (2011): "On the range of validity of the autoregressive sieve bootstrap," *The Annals of Statistics*, 39, 2103–2130.

LAHIRI, S. (2003): *Resampling Methods for Dependent Data*. Springer.

LOBATO, I., AND C. VELASCO (2004): "A simple test of normality for time series," *Econometric Theory*, 20, 671–689.

LORETAN, M., AND P. PHILLIPS (1994): "Testing the covariance stationarity of heavy-tailed time series," *Journal of Empirical Finance*, 1, 211–248.

MALKOVICH, J., AND A. AFIFI (1973): "On tests for multivariate normality," *Journal of the american statistical association*, 68, 176–179.

MARDIA, K. (1980): "Tests of Univariate and Multivariate Normality," *Handbook of Statistics*, 1, 279–319.

National Bank of Slovakia (2012): "Medium-term Forecast," 4/2012.

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

NOCETI, P., J. SMITH, AND S. HODGES (2003): "An evaluation of tests of distributional forecasts," *Journal of Forecasting*, 22, 447–455.

Norges Bank (2013): "Monetary Policy Report," 4/2013.

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.

PAULSON, A., P. ROOHAN, AND P. SULLO (1987): "Some empirical distribution function tests for multivariate normality," *Journal of Statistical Computation and Simulation*, 28, 15–30.

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.

PSARADAKIS, Z., AND M. VÁVRA (2014): "Testing for marginal asymemetry of weakly dependent processes," *Journal of Time Series Analysis*, forthcoming.

RAMBERG, J., AND B. SCHMEISER (1974): "An approximate method for generating asymmetric random variables," *Communications of the ACM*, 17, 78–82.

RUNDE, R. (1997): "The asymptotic null distribution of the Box-Pierce Q-statistic for random variables with infinite variance an application to German stock returns," *Journal of Econometrics*, 78, 205–216.

SCHOTT, J. (2005): *Matrix Analysis for Statistics*. Wiley.

SHAPIRO, S., AND M. WILK (1965): "An analysis of variance test for normality (complete samples)," *Biometrika*, 52, 591–611.

SHAPIRO, S., M. WILK, AND H. CHEN (1968): "A comparative study of various tests for normality," *Journal of the American Statistical Association*, 63, 1343–1372.

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.

Sveriges Riksbank (2013): "Monetary Policy Report," 4/2013.

THODE, H. (2002): *Testing for Normality*. Marcel Dekker New York.

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

WHITE, H. (2001): *Asymptotic Theory for Econometricians*. Academic Press.

A. Tables

Table 2: Parameters of Generalized Lambda Distribution

	λ_1	λ_2	λ_3	λ_4	skewness	kurtosis
S1	0.000000	-1.000000	-0.080000	-0.080000	0.0	6.0
S2	0.000000	-0.397912	-0.160000	-0.160000	0.0	11.6
S3	0.000000	-1.000000	-0.240000	-0.240000	0.0	126.0
A1	0.000000	-1.000000	-0.007500	-0.030000	1.5	7.5
A2	0.000000	-1.000000	-0.100900	-0.180200	2.0	21.1
A3	0.000000	-1.000000	-0.001000	-0.130000	3.2	23.8

Table 3: Empirical Rejection Frequencies of the BAD and BN Tests for Normality

		$n = 100$		$n = 200$		$n = 500$	
DGP	distr.	BN	BAD	BN	BAD	BN	BAD
M0	N	0.05	0.05	0.09	0.05	0.08	0.06
	S1	0.06	0.48	0.13	0.77	0.53	0.99
	S2	0.07	0.72	0.22	0.95	0.50	1.00
	S3	0.09	0.87	0.20	0.99	0.37	1.00
	A1	0.81	0.97	1.00	1.00	1.00	1.00
	A2	0.22	0.89	0.52	0.99	0.87	1.00
	A3	0.95	1.00	0.99	1.00	1.00	1.00
M1	N	0.03	0.05	0.05	0.05	0.09	0.04
	S1	0.01	0.23	0.04	0.40	0.22	0.78
	S2	0.04	0.44	0.09	0.69	0.34	0.97
	S3	0.04	0.63	0.11	0.88	0.33	1.00
	A1	0.21	0.81	0.83	0.97	1.00	1.00
	A2	0.10	0.67	0.35	0.89	0.79	1.00
	A3	0.45	1.00	0.97	1.00	1.00	1.00
M2	N	0.01	0.06	0.02	0.06	0.04	0.05
	S1	0.00	0.11	0.00	0.14	0.02	0.17
	S2	0.01	0.19	0.02	0.24	0.06	0.37
	S3	0.01	0.28	0.03	0.39	0.09	0.65
	A1	0.00	0.25	0.03	0.43	0.46	0.80
	A2	0.01	0.32	0.06	0.44	0.36	0.77
	A3	0.00	0.59	0.03	0.88	0.73	1.00

Table 4: Empirical Rejection Frequencies of the BAD tests for Normality

	$n = 100$				$n = 200$				$n = 500$			
distr.	M3	M4	M5	M6	M3	M4	M5	M6	M3	M4	M5	M6
N	0.07	0.39	0.41	0.07	0.05	0.61	0.75	0.07	0.05	0.93	0.99	0.05
S1	0.31	0.68	0.55	0.15	0.49	0.92	0.86	0.20	0.87	1.00	1.00	0.37
S2	0.49	0.77	0.63	0.26	0.80	0.98	0.91	0.39	1.00	1.00	1.00	0.60
S3	0.71	0.88	0.71	0.39	0.95	0.99	0.93	0.56	1.00	1.00	1.00	0.76
A1	0.86	0.62	0.82	0.53	0.99	0.84	0.99	0.83	1.00	0.99	1.00	0.98
A2	0.74	0.71	0.81	0.48	0.96	0.92	0.99	0.71	1.00	1.00	1.00	0.96
A3	1.00	0.98	0.96	0.81	1.00	1.00	1.00	0.96	1.00	1.00	1.00	1.00