A Bootstrap-Based Non-Parametric Forecast Density

Sebastiano Manzan[†] and Dawit Zerom[‡]
[†] Department of Economics, University of Leicester, United Kingdom
[‡] School of Business, University of Alberta, Canada

Abstract

The interest in density forecasts (as opposed to solely modelling the conditional mean) arises from the possibility of dynamics in higher moments of a time series as well as, in some applications, the interest in forecasting the probability of future events. By combining the idea of Markov bootstrapping with kernel density estimation, this paper presents a simple nonparametric method for estimating out-of-sample multi-step density forecasts. The paper also cosiders a host of evaluation tests to examine dynamical misspecification of estimated density forecasts by targeting autocorrelation, heteroskedasticity and neglected nonlinearity. These tests are useful as rejections of the tests give insights into ways to improve a particular forecasting model. In an extensive Monte Carlo analysis involving a range of commonly used linear and nonlinear time series processes, the nonparametric method is shown to work reasonably well across the simulated models for a suitable choice of bandwidth (smoothing parameter). Furthermore, the application of the method to the US Industrial Production series provides multi-step density forecasts that show no sign of dynamic misspecification.

Keywords: Dynamic misspecification; Evaluation; Kernel smoothing; Markov bootstrap; Multi-step density forecasts

1 Introduction

Recently, density forecasting has become an important area of research in the analysis of economic and financial time series. A density forecast of a realization of a random variable at some future time is an estimate of the probability distribution of the possible future values of that variable. In financial applications, the shift in attention toward density forecasts is motivated mainly by the increased diffusion of risk management practice among financial institutions. Calculating the capital at risk deriving from an asset or a portfolio position requires to predict the complete conditional density of returns with a particular interest on the left tail of the distribution. In the area of macroeconomic forecasting, policy analysts are not only interested in the mean evolution of macroeconomic variables in the future, but also in predicting the probability of certain events, such as whether inflation will surpass a certain level a year ahead. Granger and Pesaran (2000a, 2000b) argue that when the interest of the decision-maker is in this type of events (e.g., an interval forecast) and the loss function is non-quadratic, the use of a conditional mean forecast (and its evaluation using mean squared errors) is inappropriate. Instead, they suggest the use of density forecasts along with a nonlinear and/or asymmetric loss function that captures the preferences of the decision maker.

A common approach to modelling the conditional density is to assume a distributional form for the error term in a conditional mean model. In this case, the conditional density is fully characterized by the conditional mean and the distribution of the innovations. However, the resulting density forecasts do not account for possible time variation in the uncertainty of the forecast. In finance for example, there is overwhelming evidence of the relevant variation over time of the second moment of the distribution. An often used approach to deal with this situation is to assume that the conditional variance follows a GARCH-type specification. In this case, the conditional density is completely specified by the conditional mean and variance in addition to a distributional assumption for the error term. Instead, GARCH-type volatility is less of a concern for macroeconomic variables, where a more relevant role is played by nonlinearities in the conditional mean and possible heteroskedasticity due to the business cycle condition. Since the work of Hamilton (1989), there is now extensive evidence of regime-switching behaviour in many macroeconomic time series. This type of models assume that the economic variable switches (in a deterministic or stochastic way) between regimes characterized by different dynamics and/or different properties of the shocks. Some recent surveys of the extensive literature on modelling and forecasting using nonlinear time series models are provided by van Dijk *et al.* (2002), Tsay (2002), Clements *et al.* (2004) and Teräsvirta (2006).

One of the striking findings on the application of nonlinear models to economic and financial time series is that they outperform linear models in-sample but fail to improve in out-of-sample forecast (see de Gooijer and Kumar, 1992, and Clements *et al.*, 2003). The evidence on the forecast failure of nonlinear models typically arises when conditional mean forecasts are evaluated based on mean square errors. However, Pesaran and Potter (1997) argue that nonlinear models are better in capturing features of higher moments dynamics that are not evaluated when the comparison criteria narrowly focus on the conditional mean. Now it has become routine to assess the goodness of density forecast estimates generated by various linear and nonlinear time series models. Some recent examples are Clements and Smith (2000), Clements and Smith (2001), Boero and Marrocu (2002), Clements *et al.* (2003), Siliverstovs and van Dijk (2003), and Hong *et al.* (2004). The evaluation and comparison of density forecasts has been made possible due to the development of appropriate criteria. The most popular of the evaluation methods is the one first introduced to forecasting by Diebold *et al.* (1998). Their approach is very convenient because it transforms the problem of evaluating the forecast density into the problem of testing the distributional and independence properties of the corresponding probability integral transform. Various refinements of the proposals of Diebold *et al.* (1998) have also been developed and Corradi and Swanson (2006) survey the growing literature on density forecasts evaluation.

In this paper, we introduce a simple bootstrap-based nonparametric approach to estimate density forecasts for Markovian time series processes. The approach is an adaptation of nonparametric bootstrap methods designed for dependent processes (see Rajarshi, 1990, Paparoditis and Politis, 2001 and 2002, Horowitz, 2003, and Chan and Tong, 2004) to the multistep forecasting context. The assumption of Markov dependence is not restrictive as it encompasses a wide range of relevant structures implied by various commonly used linear and nonlinear models (e.g., AR and SETAR models). To examine the forecast performance of the proposed method, we perform an extensive Monte Carlo analysis by simulating a range of linear and nonlinear time series models. We evaluate the goodness of the density forecasts using probability integral transform-based tests. Clements *et al.* (2003) find that commonly used probability integral transform tests have low power in detecting the misspecification of linear-model based density forecasts when the true data generating process is nonlinear. Thus, we supplement the standard testing approaches with a test for neglected nonlinearity proposed by Teräsvirta *et al.* (1993). The simulation results indicate that considering this additional test has high power in uncovering neglected nonlinearity. On balance, the simulation experiment indicates that the proposed nonparametric method works reasonably well across the simulated models, provided a suitable bandwidth is chosen. The application to the US Industrial Production series indicates that the method provides correctly specified forecasts along the dimensions tested using the PIT approach.

The paper is structured as follows: in Section (2) we describe the method and discuss some companion issues including bandwidth selection. Section (3) presents the Monte Carlo evidence of the performance of the method in finite samples. We also discuss the evaluation strategy of the density forecasts based on probability integral transforms. An empirical application of the method to a macroeconomic series is discussed in Section (4) and Section (5) concludes.

2 Description of the forecasting procedure

In this section, we introduce a simple non-parametric procedure to estimate out-of-sample multi-step density forecasts. To help motivate our proposal, we first briefly describe the model-based bootstrap method to generate density forecasts (see, for example, Clements and Smith, 1997).

2.1 A parametric-bootstrap density forecast

Let $\{Y_t; t = 1, \ldots, N\}$ be a strictly stationary time series process. A commonly used parametric specification for modeling Y_t is the conditionally heteroskedastic autoregressive model, i.e.

$$Y_{t+1} = \mu(\mathbf{X}_t, \theta) + \sigma(\mathbf{X}_t, \beta)\epsilon_{t+1} \tag{1}$$

where $\mathbf{X}_t = (Y_t, \ldots, Y_{t-p+1})'$ is a p-dimensional vector, $\mu(\cdot)$ is the conditional mean of the process, $\sigma(\cdot)$ denotes the conditional variance, and ϵ_{t+1} is a disturbance term with zero mean. The vectors θ and β denote the parameters in the conditional mean and the conditional variance of the process, respectively. Model (1) includes several familiar time series models such as linear AR, ARCH, and SETAR.

In the context of model (1), the use of bootstrapping to generate forecasts has received considerable interest in recent years. At least two main motivations can be cited for the popularity of bootstrapping. First, it does not require a priori assumption about the distribution of the error term ϵ_{t+1}. In addition, for many nonlinear specifications, there are no exact formula for multi-step ahead forecasts and the bootstrap represents a natural way to approximate the forecast distributions.

Under the model specification in (1) and conditional on current time period, N, the bootstrap realization one-step ahead is given by

$$Y^*_{N+1,b} = \mu(\mathbf{X}_N, \hat{\theta}) + \sigma(\mathbf{X}_N, \hat{\beta})\hat{\epsilon}^*_{t+1,b}$$

where $\hat{\epsilon}^*_{t+1,b}$ are resampled (bootstrap) residuals, and $\hat{\theta}$ and $\hat{\beta}$ are consistent estimators of θ and β, respectively. The one-step ahead forecast density, denoted by $f_{N+1}(\cdot|\mathbf{X}_N)$, at time N is then given by the empirical density function of the bootstrap realizations, $Y^*_{N+1,b}$, $b = 1, \ldots, B$ where B is the desired number of replications. Denote the forecast horizon by τ. For $\tau \geq 2$, the forecasting density can be obtained by applying an iterative scheme. First, the two-step ahead conditioning vector is updated to $\mathbf{X}^*_{N+1,b} = (Y^*_{N+1,b}, Y_N, \cdot\cdot\cdot, Y_{N-p+2})'$, and then the forecast density, $f_{N+2}(\cdot|X_N)$, is simply the empirical density of the bootstrap realizations

$$Y^*_{N+2,b} = \mu(\mathbf{X}^*_{N+1,b}, \hat{\theta}) + \sigma(\mathbf{X}^*_{N+1,b}, \hat{\beta})\hat{\epsilon}^*_{t+1,b},$$

for $b = 1, \ldots, B$.

Most approximate parametric methods in general, and the above parametric bootstrap in particular, have at least two main shortcomings. First, the parametric specifications for $\mu(\cdot)$ and $\sigma(\cdot)$ may be inappropriate for the time series data of interest. Second, even when the specifications for the mean and variance may be appropriate as in (1), it might still be the case that the conditional distribution of the error processes $\epsilon_{t+\tau}$, is not independent of the conditioning information $\mathbf{X}_t$. In other words, there is no a priori reason to assume, in general, that the only features of the conditional distribution which depend on $\mathbf{X}_t$ are the mean and variance. Indeed, it seems quite reasonable that other features of the distribution (such as skewness and kurtosis) of $\epsilon_{t+\tau}$ might depend on $\mathbf{X}_t$. In general, under the model set-up of (1), the accuracy of forecast densities is critically dependent upon knowledge of the correct conditional distribution for ϵ_{t+1} that allows the whole conditional density of $\epsilon_{t+\tau}$ to depend on $\mathbf{X}_t$.

2.2 Markov forecast density (MFD)

One way to address the above-mentioned weaknesses of the parametric-bootstrap forecast density is through the use of nonparametric bootstrapping. In particular, we adapt the recently popularized local bootstrap approach of Paparoditis and Politis (2001, 2002) to the context of out-of-sample forecast density estimation. We assume that the time series Y_t is the outcome of a p-th order Markov process, i.e.

$$\Pr(Y_{t+1} \leq y_{t+1} | Y_t = y_t, \ldots) = \Pr(Y_{t+1} \leq y_{t+1} | Y_t = y_t, \ldots, Y_{t-p+1} = y_{t-p+1})$$

almost surely for $y_{t+1}, y_t, y_{t-1}, \ldots$ and some finite integer $p \geq 1$. The assumption of Markov dependence is satisfied by a large class of linear and nonlinear models that are of interest in time series analysis and forecasting. The conditionally heteroskedastic autoregressive model is one special case of a Markov process.

Let $Y_1, \cdots, Y_N$ denotes a realization from a Markovian time series process of order p. The goal is to estimate the out-of-sample τ-step forecast density, i.e. the conditional density of $Y_{N+\tau}$ given $Y_N, Y_{N-1}, \cdots, Y_1$ where $\tau \geq 1$. Because of the recursive nature of the proposed procedure, we introduce a time-varying forecast base, T, which takes values in $\{N, N+1, \ldots, N+\tau-1\}$. Now define,

$$\mathbf{X}_T = (Y_T, Y_{T-1}, \ldots, Y_{T-p+1})' \quad \text{and} \quad \mathbf{X}_t = (Y_t, Y_{t-1}, \ldots, Y_{t-p+1})'$$

where $t \in S_{p,T}$ and $S_{p,T} = \{p, p+1, \ldots, T-1\}$. Further, let J be a discrete random variable taking its values in the set $S_{p,T}$ with the following probability mass function

$$P(J = t) = K_1\Big(\frac{\mathbf{X}_T - \mathbf{X}_t}{h_{1,N}}\Big) / \sum_{m \in S_{p,T}} K_1\Big(\frac{\mathbf{X}_T - \mathbf{X}_m}{h_{1,N}}\Big). \tag{2}$$

where $h_{1,N} > 0$ is a bandwidth (or smoothing) parameter and $K_1(\cdot)$ is a non-negative symmetric density function (called kernel function).

We suggest to compute the τ-step MFD estimator, $\hat{f}_{N+\tau}(\cdot|\mathbf{X}_N)$ using the following iterative algorithm.

Step 1 Begin with $T = N$. Compute $P(\cdot)$ via the formula in (2). Then, use $P(\cdot)$ to resample with replacement from the successors of $\mathbf{X}_t$, i.e., $Y^*_{N+1} = Y_{J+1}$ (we attach asterics to indicate resampled data). If $\tau \geq 2$, Go to Step (2.1). Otherwise, proceed to Step (3).

Step 2

> **Step 2.1** Move T by one period forward, i.e. $T = N + 1$. Accordingly, update the conditioning vector $\mathbf{X}_T$ to $\mathbf{X}^*_T = (Y^*_T, Y_{T-1}, \ldots, Y_{T-p+1})'$. Once the update is done, compute $P(\cdot)$ via the formula in (2) and resample with replacement from the successors of $\mathbf{X}^*_t$, i.e., $Y^*_{N+2} = Y_{J+1}$. If $\tau \geq 3$, Go to Step (2.2). Otherwise, proceed to Step (3).
>
> **Step 2.2** Keep moving T forward one step at a time and repeat Step 2.1 by updating $\mathbf{X}_T$. This will be done until $T = N - \tau - 1$.

Step 3 Repeat Step (1) (and (2) if $\tau \geq 2$) B times resulting in $Y^{*,1}_{N+\tau}, Y^{*,2}_{N+\tau}, \ldots, Y^{*,B}_{N+\tau}$.

Step 4 Using another bandwidth $h_{2,N} > 0$ and kernel function $K_2(\cdot)$, compute the kernel density from the B-bootstrap observations in Step (3). This will provide the required τ-step forecast density estimate, i.e $\hat{f}_{N+\tau}(\cdot|\mathbf{X}_N)$.

In essence the strategy above is to assign probability weights to each vector $\mathbf{X}_p, \cdots, \mathbf{X}_{T-1}$, and use those probabilities to resample from their successors. The values of these probabilities will depend on the "closeness" of the vectors $\mathbf{X}_t$ to $\mathbf{X}_T$. Those states that are "close" to $\mathbf{X}_T$ receive larger probability weights (compared to those that are further away) and are thus more likely to be sampled in the bootstrap procedure. Thus, by suitably choosing these probability values, the Markov dependence in the data is maintained. On the other hand, the parametric-bootstrap approach is residual-based in the sense that it begins by estimating θ and β and subsequently uses independent resampling of the fitted residuals to generate bootstrap realizations. Thus, it assumes that dependence is fully captured by the the conditional mean and conditional variance. Instead, the MFD estimator does not attempt to reduce the problem to independent residuals, but the resampling procedure is applied directly to the time series realizations. This resampling procedure is done in such a way that the resulting forecast density is able to capture the dependence structure of the time series, beyond the first two conditional moments.

Note that when $\tau = 1$, $\hat{f}_{N+1}(\cdot|\mathbf{X}_N)$ as given in Step (4) of the above algorithm can also be written as follows,

$$\begin{aligned}\hat{f}_{N+1}(y|\mathbf{X}_N) &= \frac{1}{h_{2,N}} \sum_{J \in S_{p,N}} K_2\Big(\frac{y - Y_{J+1}}{h_{2,N}}\Big) P(J) \\ &= \frac{1}{h_{2,N}} \sum_{J \in S_{p,N}} K_2\Big(\frac{y - Y_{J+1}}{h_{2,N}}\Big) K_1\Big(\frac{\mathbf{X}_N - \mathbf{X}_J}{h_{1,N}}\Big) / \sum_{J \in S_{p,N}} K_1\Big(\frac{\mathbf{X}_N - \mathbf{X}_J}{h_{1,N}}\Big) \qquad (3)\end{aligned}$$

where $y \in \mathbb{R}$ and the kernel $K_2(\cdot)$ is a nonnegative symmetric density function. Lets create an associated process $(\mathbf{X}_J, Z_J)$ where $Z_J = Y_{J+1}$ for $J \in S_{p,N}$. Then, (in retrospect) $\hat{f}_{N+1}(y|\mathbf{X}_N)$ reduces to the well-known conditional density estimator of Z_J conditional on $\mathbf{X}_J = \mathbf{X}_N$, see for example, Hyndman *et al.* (1996), Fan *et al.* (1996) and de Gooijer and Zerom (2003), among others. Under some regularity conditions about $(\mathbf{X}_J, Z_J)$ (that are also satisfied by our strictly stationary markovian time series Y_t), de Gooijer and Zerom (2003) has shown the asymptotic consistency of another conditional density estimator where $\hat{f}_{N+1}(y|\mathbf{X}_N)$ is one special case. So, this shows that our bootstrap procedure for estimating the forecast density is valid when $\tau = 1$. In the Appendix, we show asymptotic validity of MFD when $\tau \geq 2$.

2.2.1 Choice of bandwidth

To apply the MFD estimator, choices need to be made with regard to the form of the kernels, $K_1(\cdot)$ and $K_2(\cdot)$, as well as the values for the smoothing parameters, b_T and a_T. Both theoretical and empirical studies regarding kernel-based nonparametric smoothers have confirmed that the choice of kernel functions does not have a relevant influence on the accuracy of the estimators. This has been shown to be the case for density and regression-type estimators by Fan and Yao (2003). In both the simulation and the empirical part of this paper, we use the standard normal density for both K_1 and K_2. To the contrary, appropriate choice of the smoothing parameters is crucial for the accuracy of all kernel-based methods.

Note that the MFD estimator does not require both $h_{1,N}$ and $h_{2,N}$ to be chosen simultaneously. By construction, the estimator is implemented in two stages. Detailed simulation experiments suggest (using the time series models in Section 3.1) that the MFD estimator is not sensitive to the choice of $h_{2,N}$ as long as $h_{2,N} \sim N^{-1/5}$. So, we use $h_{2,N} = \hat{\sigma} N^{-1/5}$ where $\hat{\sigma}$ is the standard deviation of the time series $\{Y_1, \ldots, Y_N\}$. On the other hand, the choice of $h_{1,N}$ seems to be critical to the quality of the forecast density estimator. This might suggest that accounting for the correct dependence (dynamics) in time series is crucially important for the accuracy of the forecast density.

For fixed N, when $h_{1,N} \to 0$, the τ-step forecast density will tend to accurately capture the dependence structure or dynamical properties of the data. The problem is that the forecast density becomes excessively peaked compared to the true forecast density (see the Monte Carlo simulation experiment Section for more on this situation). On the other hand, when $h_{1,N} \to \infty$, the τ-step forecast density does not reflect the dependence structure of the data. The latter case represents a situation where the data are in fact nearly independent. For a particular τ, notice from Equation (2) that for $h_{1,N} \to \infty$, the probability weight $P(J = t) \to 1/(T - p)$ such that the information contained in $\mathbf{X}_N$ becomes irrelevant to forecast $Y_{N+\tau}$. Therefore, $h_{1,N}$ should lie between the above two extremes to obtain a forecast density estimate that better reflects the shape of the true conditional density and accurately mimicks the dependent characteristics of the data.

Observe that the probabilities $P(\cdot)$ used to resample the data are simply kernel estimates scaled by kernel densities. Thus, we adopt a simple two-step procedure to select the bandwidth $h_{1,N}$. First, estimate a pilot density estimate for X_t,

$$\hat{\pi}(X_t) = \frac{1}{T-p} \sum_{t \in S_{p,T}} \frac{1}{h_{1,N}} K_1\Big(\frac{X_T - X_t}{h_{1,N}}\Big)$$

using a preliminary bandwidth estimate $h_{1,N} = \hat{\sigma} N^{-\frac{1}{p+4}}$, where $\hat{\sigma}$ is the standard deviation of the time series $\{Y_1, \ldots, Y_N\}$. This bandwidth estimate is not adaptive to the data configuration of X_t. In the second step, we use the pilot density estimate $\hat{\pi}(X_t)$ to adjust the preliminary bandwidth (hereinafter called "fixed bandwidth") in such a way that areas of high density use a smaller bandwidth and areas of low density use a larger bandwidth. Following Silverman (1986), we define a local bandwidth factor, λ_t, by $\lambda_t = \left\{\frac{\hat{\pi}(X_t)}{\tilde{\pi}}\right\}^{-\alpha}$ where $\tilde{\pi}$ is the geometric mean of $\hat{\pi}(X_t)$ and α $(0 \le \alpha \le 1)$ denotes the sensitivity parameter that regulates the amount of weight that is attributed to the observations in the low density regions. We consider $\alpha = 0.5$. Using λ_t, we define an adaptive bandwidth as $h_{1,N}(X_t) = \lambda_t \hat{\sigma} N^{-\frac{1}{p+4}}$. In Section (3), we compare the performances in finite samples of both fixed and adaptive bandwidths.

3 Evaluating the forecast accuracy of the MFD estimator

In this section we examine the forecast performance of the MFD estimator $\hat{f}_{N+\tau}(y|\mathbf{X}_N)$ in a simulation experiment. The evaluation of the density forecasts is based on testing the probability integral transform. We also introduce a simple useful extension of these tests to detect neglected nonlinearity in the estimated forecast densities.

3.1 Simulation set-up

We consider 6 models of markov order 1 and 2 that belong to either of the following time series processes: the linear AutoRegressive (AR) model, the AutoRegressive Conditional Heteroscedastic (ARCH) model and the Self-Exciting Threshold AutoRegressive (SETAR) model.

(i) First-Order Markov Processes ($p = 1$):

AR(1): $Y_t = 0.6Y_{t-1} + \epsilon_t$

ARCH(1): $Y_t = \sigma_t \epsilon_t$ and $\sigma_t = 0.7 + 0.3Y_{t-1}^2$

SETAR(1): $Y_t = [-1.25 - 0.7Y_{t-1} + \sigma_1 \epsilon_t] I(Y_{t-1} \leq r) + [0.3Y_{t-1} + \sigma_2 \epsilon_t] I(Y_{t-1} > r)$

where the error ϵ_t is $N(0, 1)$ distributed and $I(A)$ represents the indicator function that assumes value 1 when the event A is true and 0 otherwise. For the SETAR(1) model, we follow Clements *et al.* (2003) and consider two parameterizations: one in which the two regimes have the same variance $\sigma_1 = \sigma_2 = 1$ ($r = -0.2$) and the other in which they are heteroskedastic, i.e. $\sigma_1 = 1$ and $\sigma_2 = 2$ (and the threshold parameter is set equal to $r = -0.1$).

(ii) Second-Order Markov Processes ($p = 2$):

SETAR(2): $Y_t = [-1.25 - 0.7Y_{t-2}] I(Y_{t-2} \leq r) + 0.3Y_{t-2} I(Y_{t-2} > r) + \sigma \epsilon_t$

SETAR(1-2): $Y_t = [-1.25 - 0.7Y_{t-1}] I(Y_{t-2} \leq r) + 0.3Y_{t-2} I(Y_{t-2} > r) + \sigma \epsilon_t$

For the SETAR(2) model, the dependence occurs only at the second lag whereas in the SETAR(1-2) model, the dependence is both at the first and second lags. For both second-order models, we set $\sigma = 1$ and $r = -0.20$.

We adopt a rolling approach to generate density forecasts. Let M be the total number of observations and n be the first forecast origin. This means that there are n observations up to and including the nth observation. By "rolling" we mean that the forecast base N extends as far as $M - \tau^*$ where τ^* is the maximum forecast horizon. Hence, $N = n, n+1, \ldots, M - \tau^*$. In the simulation exercise, we consider $\tau = 1, 2, 3$ (thus, $\tau^* = 3$) and $n = 300$ ($M = 600$), $n = 600$ ($M = 900$) and $n = 900$ ($M = 1200$). The number of bootstrap replications B is always fixed at 1000 and the number of simulations to 2000.

3.2 Probability integral transform based tests

When the object of interest is the prediction of the conditional mean, a straightforward evaluation criteria consists of comparing the predicted and realized values according to a chosen loss function (e.g., quadratic). However, the evaluation of density forecasts is complicated by the fact that we do not observe the realization of the forecast density. To overcome this problem, evaluation methods based on the Probability Integral Transform (PIT) have been proposed in the literature.

First, consider the case of testing the density forecasts for $\tau = 1$. The approach of testing for the case of $\tau \geq 2$ will be discussed later in this Section. Because PIT tests involve a sequence of density forecast estimates, the rolling approach is an ideal context. The PIT, here denoted by z_N, for $N = n, n+1, \cdots, M - \tau^*$, is defined as

$$z_N = \int_{-\infty}^{Y_{N+1}} \widehat{f}_{N+1}(u|\mathbf{X}_N)du \tag{4}$$

where $\widehat{f}_{N+1}(\cdot|\mathbf{X}_N)$ is the one-step forecast density based at the forecast origin N. Let $f_{N+1}(u|\mathbf{X}_N)$ denote the true forecast density. If the forecasting model is correctly specified, $f_{N+1}(u|\mathbf{X}_N)$ and $\hat{f}_{N+1}(u|\mathbf{X}_N)$ coincide and the sequence $\{z_N\}$ is *i.i.d.* $U(0,1)$. Thus, evaluating the goodness of the estimated forecast densities is equivalent to evaluate the independence and uniformity properties of $\{z_N\}$. This result forms the foundation for the PIT-based family of tests. Diebold *et al.* (1998) first introduced this idea of evaluating density forecasts by testing whether the empirical CDF of the $\{z_N\}$ is significantly different from the theoretical uniform CDF. They employed mainly qualitative graphical tools to assess uniformity and independence. The more recent literature has introduced formal tests to evaluate the uniformity of the PIT, such as the Kolmogorov-Smirnov (**KS**) test.

However, testing the uniformity of the PIT evaluates only the distributional part of the hypothesis on $\{z_N\}$ and does not capture violations of independence. Berkowitz (2001) proposes to transform the PIT using the inverse of the standard normal distribution. Under the assumption that the PIT is *i.i.d.* U(0,1) the transformed random variable is distributed as an *i.i.d.* standard normal. He then uses Likelihood Ratio (LR) tests of the hypothesis of independence (versus the alternative of autoregressive structure in the PIT) and of the joint hypothesis of *i.i.d* and standard normality. Recently, Hong *et al.* (2004) and Hong and Li (2005) propose an omnibus statistics that tests jointly the hypothesis of independence and uniformity and is also robust to parameter estimation uncertainty. The relevance of assessing violations of the independence assumption is clear from the results of Clements *et al.* (2003). They investigate the power of the KS and LR tests to distinguish between a linear-based and nonlinear-based forecast densities when the true data generating process is nonlinear. The linear density forecasts are misspecified and this should be indicated by testing the PIT. However, Clements *et al.* (2003) showed in a Monte Carlo exercise that these tests have negligible power to indicate the misspecification of the linear forecast density. This motivates us to introduce an additional test of the PIT in order to detect neglected nonlinearity in the conditional mean of density forecasts.

We test separately the assumptions of uniformity and independence of the PIT series. In particular, we consider specific directions in which independence might be violated, such as serial correlation, heteroskedasticity of the ARCH-type, and neglected nonlinearity. Rejections of the null hypothesis of these tests indicate misspecification of the density forecasts and suggest directions in which the conditional model could be improved. For uniformity, the **KS** test is used. To test for the presence of serial correlation in the PIT, we assume that the z_N follows the process

$$(z_N - \bar{z}) = \alpha_1(z_{N-1} - \bar{z}) + \cdots + \alpha_q(z_{N-q} - \bar{z}) + \epsilon_N. \tag{5}$$

A test for linear independence is equivalent to testing the hypothesis that all the α_i's (for $i = 1, \cdots, q$) are equal to zero. An LM-type test is carried out with a statistic equal to $(M - \tau^* - n + 1)$ times the R^2 of Equation (5) that is distributed as a $\chi^2(q)$. Rejection of the null hypothesis suggests the presence of linear dependence unaccounted by the forecasting model. We denote the test for serial correlation as **SC**[1]. To test whether the density forecasts correctly account for the possibility of ARCH structure in the residuals, we perform an ARCH LM test, i.e. we regress the squared residuals of Equation (5) on r lags

$$\epsilon_N^2 = \beta_0 + \beta_1 \epsilon_{N-1}^2 + \cdots + \beta_r \epsilon_{N-r}^2 + \eta_N \tag{6}$$

and test the null hypothesis that the β_j (for $j = 1, \cdots, r$) are jointly equal to zero. The test statistic is $(M - \tau^* - n + 1)$ times the R^2 of Equation (6) and it is distributed as a $\chi^2(r)$. We denote this test as **HET**. As mentioned earlier, an interesting alternative to evaluate is the ability of the forecast densities to account for possible nonlinearity in the underlying generating process. We adopt the **V23** test proposed by Terasvirta et al. (1993) to test the hypothesis of neglected nonlinearity in the conditional mean of the z_N. This is done by estimating the following regression

$$(z_N - \bar{z}) = \alpha_1(z_{N-1} - \bar{z}) + \cdots + \alpha_q(z_{N-q} - \bar{z}) + \sum_{i=1}^{q}\sum_{j=1}^{q} p\delta_{i,j}(z_{N-i} - \bar{z})(z_{N-j} - \bar{z}) + \tag{7}$$

$$+ \sum_{i=1}^{q}\sum_{j=1}^{q}\sum_{k=1}^{q} \delta_{i,j,k}(z_{N-i} - \bar{z})(z_{N-j} - \bar{z})(z_{N-k} - \bar{z}) + \epsilon_N$$

Under the null hypothesis of linearity all the $\delta_{i,j}$ and $\delta_{i,j,k}$ are equal to 0. A standard F-test can be used to test this hypothesis. We interpret the rejection of the null hypothesis as evidence that the forecast density does not account for nonlinearity of the time series.

We described the testing approach for one-step ahead density forecasts. The above procedures can be used for the evaluation of multi-step density forecasts, i.e. $\tau \geq 2$, provided the following simple provisions for the autocorrelation in $\{z_N\}$ are being made. That is, for τ-step ahead forecasts, each of the sub-series $(z_1, z_{1+\tau}, z_{1+2\tau}, \cdots)$, $(z_2, z_{2+\tau}, z_{2+2\tau}, \cdots)$ and $(z_\tau, z_{2\tau}, z_{3\tau}, \cdots)$ should be $i.i.d.$ $U(0,1)$. The same battery of tests described above can be applied to the sub-series of the PIT using a significance level equal to $\frac{\alpha}{\tau}$, where α is the size of the test and τ the forecasting step. The null hypothesis is rejected if any of the τ tests is rejected.

3.3 Performance of the MFD estimator

We now evaluate the performance of the MFD for some simulated (linear and nonlinear) time series processes. To benchmark the performance of the MFD, we also evaluate the forecast

[1] Siliverstovs and van Dijk (2003) consider a similar approach that consist of testing $(z_N - \bar{z})$ (and power transformations) using the Ljung-Box test.

accuracies of two alternative methods to generate density forecasts. The first method is to simply resample the data under the assumption of independence. In this case we destroy the dependence in the data and hence the forecasting densities are misspecified for all six simulated models. Evaluating these density forecasts gives an indication of the power of the tests, that is, their ability to signal the misspecification of the forecast density estimate (that are generated under the null of independence while the true underlying process is dependent). The second method that we use for comparison is a linear AR model (with bootstrap errors as discussed in Section 2). The linear forecast densities are appropriate when we consider the AR(1) process but are misspecified for the nonlinear time series models. In this case, it is interesting to evaluate whether testing the PIT with the V23 test is a useful tool in detecting the misspecification of the forecast densities. We indicate the results for the density forecasts under the null of independence as **IND** and for the linear forecast densities as **LIN**.

The MFD is implemented both for the fixed and adaptive bandwidth rules as discussed in Section (2.2.1). In particular, we use $h_{1,N} = c\hat{\sigma}N^{-\frac{1}{p+4}}$ in the case of the fixed bandwidth and $h_{1,N}(X_t) = c\lambda_t\hat{\sigma}N^{-\frac{1}{p+4}}$ in the case of the adaptive bandwidth where c=(0.5, 0.75, 1, 1.25). The aim for introducing c is to evaluate the effect of under- and over-smoothing on the performance of the MFD.

Simulated model: AR(1) Table (1) reports the frequency of rejections (at 5% significance level) of the KS, SC, HET, and V23 tests when the simulated series are generated from the AR(1) model. We first consider the case of $\tau = 1$. The results show that the test for serial correlation (SC) has very high power in detecting the misspecification of the IND density forecasts. For all simulated series the SC test rejects the null hypothesis of no serial correlation in the PIT. This result is expected since resampling the data destroys the AR(1) structure of the series. However, testing the uniformity of the PIT with the KS test rejects in only 27% of the simulations. This provides evidence of the low power of this test in detecting the dynamic misspecification of the density forecasts. When the density forecasts are generated by the linear AR model (to predict an AR(1) series) the rejection frequencies are, as expected, very close to 5% and all tests are correctly sized.

Table (1) here

Considering the case of n equal to 300, the results for the MFD offer the following insights. Small bandwidth values (e.g., $c = 0.5$) are associated with rejection frequencies of the SC test close to the nominal level. However, increasing the value toward $c = 1.25$ worsens the performance and the rejections rise to 23% (for both fixed and adaptive rules). The results improve when we consider larger sample sizes (n equal to 600 and 900), although with a similar pattern. This seems to suggest that under-smoothing is required to correctly capture the dynamics of the process.

On the other hand, the HET test indicates that for small values of c the test over-rejects, that is, suggests evidence of misspecification in the conditional variance of the process. However, this result is spurious because the true underlying DGP is an homoskedatic AR(1) model. Increasing the bandwidth value, the rejections decrease and for $c = 1.25$ the HET test is close to the 5% level. Increasing the sample size reduces significantly the evidence of spurious heteroskedasticity. The KS and V23 tests are, for all bandwidth values, correctly sized. Therefore, these results show that the method provides appropriate density forecasts for values of the bandwidth (c between 0.75 and 1) that balance the trade-off between small (correct dynamics) and large (spurious heteroskedasticity) values. Comparing the rejection frequencies for the HET test it is evident that the adaptive scheme is affected by the spurious heteroskedasticity effect to a less extent.

Increasing the forecasting step τ to two shows that the SC test has still high power while the KS test reject in only 13% of the cases. However, the power decreases to 58% when three-step ahead forecasts are considered.

Simulated model: ARCH(1) The ARCH(1) model is a markovian model where the dependence occurs in the conditional second moment of the process. The results of the simulations are reported in Table (2). The evaluation of the IND (and LIN) density forecasts shows that the HET test rejects in 82% of the simulations, while the other tests have frequencies very close to the nominal 5% level.

Table (2) here

The density forecasts generated by the MFD have correct size for the KS, SC, and V23 test. However, there are significant over-rejections for the HET test, in particular for small values of the bandwidth. For this model the adaptive rule performs better in terms of frequency of rejections of the HET test. This can be explained as follows. Time series generated from the ARCH model are characterized by few and short-lived clusters of large observations that make difficult for the nonparametric estimator (using the fixed bandwidth) to smooth in the tails of distribution (where the data are sparse). As it is clear from the Table, using the adaptive bandwidth and increasing the sample size have beneficial effects on the performance of the MFD estimator. For $n = 900$ and $c = 0.75$ the HET test rejects in 8.7% of the simulations, slightly over-rejecting compared to the 5% nominal level. Increasing the forecasting step τ to 2 and 3 shows that all the density forecasts provide similar results, in the sense that they have rejection frequencies close to 5%.

Simulated model: homoskedastic SETAR(1) Table (3) reports the results for the homoskedastic SETAR model with dependence in the first lag. For this parametrization, the model shows both linear and nonlinear dependence. This is clearly captured by the SC and V23 tests when evaluating the IND density forecasts. Both tests reject in more than 96% of the simulations, clearly indicating the misspecification of these forecasts to account for linear and nonlinear dependence in the conditional mean. The evaluation of the linear density forecasts (LIN) indicates slight over-rejection for the SC test (13%) and 94% of rejections for the V23 test. However, the KS test is rejecting for only 6.5% of the simulations. These results confirm the findings of Clements *et al.* (2003) on the inability of the KS test to signal the dynamic misspecification of the conditional densities. A possible solution to the difficulty of distinguishing linear and nonlinear density forecast when the true DGP is nonlinear is addressed here by introducing the V23 test. As expected, the test provides high power in detecting the neglected nonlinearity (in the conditional mean) of the forecast densities.

Table (3) here

For $\tau = 1$, the performance of the MFD follows a similar pattern to the case of the AR(1) series. Under-smoothing achieves correct size for the V23 test (and already for sample size 300) while over-smoothing increases the frequency of rejections to over 10%. However, also in this case the spurious heteroskedasticity effect is present and disappears only when larger bandwidth values are used. A balance between the two effects is obtained when the constant c is between c =0.75 and 1, in particular for larger sample sizes. For all cases considered, the KS and SC tests are quite close to the nominal level of the test. For $\tau > 1$, the frequency of rejections of SC and V23 tests decrease (compared to $\tau = 1$) when used to evaluate the IND and LIN density forecasts. This is not due to a significant loss of power of the test[2] rather to the fact that the nonlinear dependence becomes weaker and (almost) indistinguishable from independent observations. Thus, it is not surprising that the MFD has rejections very close to the nominal level in many cases, with the exception of the HET test where under-smoothing contributes to reject too often.

Simulated model: heteroskedastic SETAR(1) In Table (4) we report the results of the density forecasts for the SETAR(1) specification with error variances different in the regimes. Neglecting the dependence in the data and simply forecasting the unconditional distribution of the observations (IND) leads to significant rejections of the SC and V23 tests (56 and 99%, respectively). The KS and HET tests have rejection frequencies between 11 and 16%. Similarly, the misspecification of linear density forecasts is indicated by the V23 test (99% of rejections).

Table (4) here

Considering the one-step ahead density forecasts, the MFD method correctly accounts for the dynamics of the SETAR specification (both SC and V23 are close to the nominal rejection level), but it is affected by the spurious evidence of heteroskedasticity (HET rejects in 21% of

[2] At least in the case of $\tau = 2$, the tests are calculated on the PIT's with sample size 150 that is large enough to guarantee reasonably high power.

the simulations for $c = 0.5$) that disappears only at large values of the bandwidth (for $c = 1.25$ the rejection frequency is 6%). The performance of the MFD improves significantly when larger samples are considered and for values of the constant c between 0.75 and 1 in the bandwidth formula. When $\tau = 2$ is considered the IND (and LIN) show significant rejections for the SC and HET tests, but the evidence of nonlinearity is less pronounced. The V23 test rejects in about 8% of the simulations.

Simulated model: SETAR(2) In this specification the dependence (linear and nonlinear) occurs in the second lag. Hence, we set $p = 2$ for both the AR forecasting method and MFD. Table (5) shows the results for this model. For $\tau = 1$, the IND density forecasts show high frequency of rejections for the SC and V23 tests (97 and 85%, respectively) while the LIN density forecasts appear (in 79% of the simulations) to reject often, as expected, the null of linearity. Contrary to the SETAR(1) case, when $\tau = 2$ the V23 test shows high levels of rejections for the IND and LIN density forecasts indicating their misspecification in accounting for nonlinear dependence. However, this evidence disappears when $\tau = 3$ is considered. The two-step ahead IND forecasts show also significant deviations in the SC test, while LIN correctly accounts for the linear dependence.

Table (5) here

Simulated model: SETAR(1-2) This specification has linear dependence in the first and second lag and the switching between the two regimes is determined by the second lag. We set p equal to 2 for both the LIN density forecasts and MFD. The evaluation of the IND density forecasts (see Table (6)) indicates they are misspecified according to the SC, HET and V23 (100, 22 and 98%, respectively) at $\tau = 1$, while for $\tau = 2$ the SC has still high power but V23 rejects in only 12% of the simulations. This is probably due to the weak form of nonlinearity that is built-in the model.

Table (6) here

The results for the MFD are similar to the SETAR(2) case. Mid-values of c achieve reasonable results, in particular when the adaptive bandwidth is used. Increasing the sample size is also useful to correct the over-rejections of the tests compared to the 5% significance level.

3.3.1 Summary of the simulation results

We investigate the finite sample performance of the MFD by simulating a range of linear and nonlinear time series models. The main result that emerges from this analysis is that the method works reasonably well across the simulated models, provided a suitable bandwidth is chosen. In particular, under-smoothing leads to density forecasts that better account for the dynamics of the process. But, the forecast density estimates becomes narrower (in the center) compared to the true forecast density. This is reflected in over-rejection of the null in the HET test. On the other hand, over-smoothing has the opposite effect: the spurious heteroskedasticity disappears as well as the ability of the method to correctly model the dynamics in the density forecasts. The simulation results suggest that the trade-off between the above two extremes can be minimized by choosing the constant c in the range of 0.75 and 1. As to sample size requirement of the MFD, the rejection frequencies of the PIT tests, for some models, are slightly higher compared to the 5% nominal value) when the sample size for estimation is equal to 300, but improves significantly when larger samples are considered.

4 Real data example

In this section, we estimate and evaluate density forecasts for different models (parametric and MFD) using a macroeconomic time series. In particular, we consider the seasonally adjusted time series of monthly growth rate of US Industrial Production (IP). The data period starts in January 1960 and ends in April 2004 (532 observations). We use observations until December 1985 as the in-sample period and forecast out-of-sample from January 1986 up to the end of the sample (312 and 220 observations for in- and out-of-sample, respectively). We forecast one- to three-step ahead ($\tau = 1, 2, 3$) with the in-sample set expanding to include the new observation available (a rolling framework as in the simulations). For the MFD we select the Markov order based on the $\delta(p_{\text{MFD}})$ test proposed in Diks and Manzan (2002) for the first available in-sample period and keep it fixed in the rest of the sample. The results strongly indicate that $p_{\text{MFD}} = 3$ is the best choice. As to the bandwidth, both adaptive and fixed bandwidth rules are implemented. Inferring from the simulation evidence, a constant c=0.75 is used.

The MFD will be compared against the following three methods: 1) resample the data under the null of independence and we denote it as IND, 2) assuming a linear AR specification (indicated as *LIN*) and 3) a two-regime $SETAR(p_{\text{TAR}}, d)$ model defined as

$$Y_{t+1} = [\mathbf{X}_t\theta_1 + \sigma_1\epsilon_{t+1}]I\left\{\left(\sum_{i=0}^{d-1} Y_{t-i}\right) \leq r\right\} + [\mathbf{X}_t\theta_2 + \sigma_2\epsilon_{t+1}]I\left\{\left(\sum_{i=0}^{d-1} Y_{t-i}\right) > r\right\} \tag{8}$$

where Y_{t+1} denotes the growth rate of the series and $\mathbf{X}_t$ is a p_{TAR}-dimensional vector of lagged values of Y_{t+1}. The switching in the model depends on the cumulative growth rate of the last d months. The vectors θ_1 and θ_2 represent the parameters governing the dynamics in the two regimes. We allow for heteroskedastic regimes and denote the variances of the innovations by σ_1 and σ_2, respectively. We followed the approach of Siliverstovs and van Dijk (2003) and select recursively the lags p_{AR}, p_{TAR} and d based on a search up to the 6^{th} lag and using the AIC criterion. The forecasting densities (for methods 2 and 3) are obtained by drawing with replacement from the standardized residuals as discussed in Section (2.1).

The forecast density evaluation is based on SC, HET, and $V23$ tests where we use up to a maximum order of 5 (see the simulation section). Table (7) shows the p-values of the evaluation tests for the different methods used to forecast IP in the case of one up to three steps ahead. The results show that, for $\tau = 1$, the IND density forecasts are misspecified according to the KS, SC and V23 test (using a 5% significance level). This is expected since a vast literature found evidence of linear and nonlinear dependence in the IP growth rate. The evaluation of the PIT deriving from the LIN density forecasts indicates signs of misspecification, in particular, they are not able to account for the nonlinear dependence in the data. The V23 test rejects the null hypothesis of linearity providing further evidence to the literature that models the IP growth rate using regime switching models. The evaluation of the SETAR density forecasts shows that they are correctly specified, although the p-value of the KS test rejects the null of uniformity. The results also suggest that the addition of the V23 test seems to provide a relevant tool in testing density forecasts. If we would have only considered KS, SC, and HET tests the LIN and SETAR forecasts would both seem correctly specified. However, the V23 test indicates that the LIN density forecasts neglect to account for the nonlinearity in the data. The MFD method seems to provide appropriate density forecasts in all the dimensions we are evaluating: the tests do not reject the respective null hypotheses for both the fixed and adaptive bandwidth rules.

The two-steps ahead density forecasts indicate that IND provides misspecified forecasts (KS and SC reject the null), while for the LIN predictive densities we reject the null of uniformity of the PIT (at the 1% level). The results in the Table also suggest that the PIT of the SETAR forecasts reject the null for the KS, SC and HET tests at the 5% significance level. Instead, the V23 test for neglected nonlinearity does not indicate signs of misspecification. The MFD perform reasonably well and none of the tests rejects at the 5% significance level. When considering three-step ahead, all forecasts seem to correctly account for the dynamical and distributional properties of the conditional density. The only rejection (at 5% significance level) occurs for the KS when testing the SETAR forecasts.

5 Conclusion

This paper proposes a simple bootstrap-based nonparametric approach to forecast the density in a time series context. The main feature of the Markov Forecast Density (MFD) method is that it does not require the researcher to make a priori assumptions about the moments in which the dynamics occurs, and the specification of a parametric form for the conditional moments. We investigate the finite sample performance of the method by simulating a range of linear and nonlinear time series models. The main result that emerges from this analysis is that the method works reasonably well across the simulated models, provided a suitable bandwidth is chosen. Furthermore, the application of the MFD to the US Industrial Production series provides forecasts that show no sign of misspecification along the dimensions tested using the PIT approach (distribution, linear dependence, heteroskedasticity, and neglected nonlinearity).

Concluding, further work is required to improve the applicability of the method. One interesting extension of the method is to consider a semi-parametric approach, where some of the moments are modelled parametrically while using a Markov bootstrap on the residuals. This would probably allow more flexibility (compared to a purely parametric model) and provide better performance in small samples (compared to the fully nonparametric case). We will explore this extension in a future work.

References

Berkowitz, J. (2001). Testing density forecasts, with applications to risk management. *Journal of Business and Economic Statistics*, **19,** 465–474.

Boero, G. and Marrocu, E. (2002). The performance of nonlinear exchange rate models: a forecasting comparison. *Journal of Forecasting*, **21,** 513–542.

Chan, K.S. and Tong, H. (2004). Testing for multimodality with dependent data. *Biometrika*, **91,** 113–123.

Clements, M.P., Franses, P.H., Smith, J. and van Dijk, D. (2003). On SETAR non-linearity and forecasting. *Journal of Forecasting*, **22,** 359–375.

Clements, M.P., Franses, P.H. and Swanson, N.R. (2004). Forecasting economic and financial time-series with non-linear models. *International Journal of Forecasting*, **20,** 169–183.

Clements, M.P. and Smith, J. (1997). The performance of alternative forecasting methods for SETAR models. *International Journal of Forecasting*, **13,** 463–475.

Clements, M.P. and Smith, J. (2000). Evaluating the forecast densities of linear and non-linear models: Application to output growth and unemployment. *Journal of Forecasting*, **19,** 255–276.

Clements, M.P. and Smith, J. (2001). Evaluating forecasts from SETAR models of exchange rates. *Journal of International Money and Finance*, **20,** 133–148.

Corradi, V. and Swanson, N.R. (2006). Predictive density evaluation. In *Handbook of Economic Forecasting* (eds G. Elliott, C.W.J. Granger and A. Timmermann). Elsevier.

de Gooijer, J. and Zerom, D. (2003). On conditional density estimation. *Statistica Neerlandica*, **57,** 159–176.

de Gooijer, J.G. and Kumar, K. (1992). Some recent developments in non-linear time series modelling, testing and forecasting. *International Journal of Forecasting*, **8,** 135–156.

Diebold, F.X., Gunther, T.A. and Tay, A.S. (1998). Evaluating density forecasts. *International Economic Review*, **39,** 863–883.

Diks, C. and Manzan, S. (2002). Tests for serial independence and linearity using the correlation integrals. *Studies in Nonlinear Dynamics and Econometrics*, **6,** number 2.

Fan, J. and Yao, Q. (2003). *Nonlinear time series: Nonparametric and Parametric Methods.* Springer.

Fan, J., Yao, Q. and Tong, H. (1996). Estimation of conditional densities and sensitivity measures in nonlinear dynamical systems. *Biometrika*, **83,** 189–206.

Granger, C.W.J. and Pesaran, M.H. (2000a). A decision theoretic approach to forecast evaluation. In *Statistics and Finance: An Interface* (eds W.S. Chan, W.K. Li and H. Tong), pp. 261–278. Imperial College Press.

Granger, C.W.J. and Pesaran, M.H. (2000b). Economic and statistical measures of forecast accuracy. *Journal of Forecasting*, **19,** 537–560.

Hamilton, J. (1989). A new approach to the economic analysis of nonstationary time series and the business cycle. *Econometrica*, **57,** 357–384.

Hong, Y. and Li, H. (2005). Nonparametric specification testing for continuous-time models with applications to term structure of interest rates. *Review of Financial Studies*, **18,** 37–84.

Hong, Y., Li, H. and Zhao, F. (2004). Out-of-sample performance of discrete-time spot interest rate models. *Journal of Business and Economic Statistics*, **22,** 457–473.

Horowitz, J.L. (2003). Bootstrap methods for markov processes. *Econometrica*, **71,** 1049–1082.

Hyndman, R.J., Bashtannyk, D.M. and Grunwald, G.W. (1996). Estimating and visualizing conditional densities. *Journal of Computational and Graphical Statistics*, **5,** 315–336.

Paparoditis, E. and Politis, D.N. (2001). A markovian local resampling scheme for nonparametric estimators in time series analysis. *Econometric Theory*, **17,** 540–566.

Paparoditis, E. and Politis, D.N. (2002). The local bootstrap for markov processes. *Journal of Statistical Planning and Inference*, **108,** 301–328.

Pesaran, M.H. and Potter, S.M. (1997). A floor and ceiling model of US output. *Journal of Economic Dynamics and Control*, **21,** 661–695.

Rajarshi, M.B. (1990). Bootstrap in markov sequences based on estimates of transition density. *Annals of the Institute of Statistical Mathematics*, **40,** 565–586.

Siliverstovs, B. and van Dijk, D. (2003). Forecasting industrial production with linear, nonlinear, and structural change models. *Econometric Institute Report 2003-16.*

Silverman, B.W. (1986). *Density Estimation*. Chapman & Hall.

Teräsvirta, T. (2006). Forecasting economic variables with nonlinear models. In *Handbook of Economic Forecasting* (eds C. W. J. Granger G. Elliott and A. Timmermann). Elsevier.

Teräsvirta, T., Lin, C-F. and Granger, C.W.J. (1993). Power of the neural network linearity test. *Journal of Time Series Analysis*, **14,** 209–220.

Tsay, R.S. (2002). Nonlinear models and forecasting. In *A Companion to Economic Forecasting* (eds M.P. Clements and D.F. Hendry), pp. 453–484. Blackwell.

van Dijk, D., Teräsvirta, T. and Franses, P.H. (2002). Smooth transition autoregressive models, a survey of recent developments. *Econometric Reviews*, **21,** 1–47.

Appendix: Asymptotic validty of the MFD for $\tau \geq 2$

As outlined in the algorithm, when $\tau \geq 2$, the MFD estimator can also be defined by repeating one-step ahead predictions τ times, treating the bootstrap value from the last round as the true value. In this way, the τ-step MFD estimator can also be viewed as a *one-step plug-in estimator*. Thus, the asymptotic consistency result for the 1-step MFD also holds for the τ-step iterative MFD estimator. We only need to show that replacing actual values by bootstrap replicates is valid. For example, for $\tau = 2$, it suffices to show the validity of replacing Y_{N+1} by the bootstrap counterpart Y^*_{N+1}.

Denote by $F(y|\mathbf{x})$ the one-step transition distribution function of Y_t, i.e.

$$F(y|\mathbf{x}) = \Pr(Y_{t+1} \leq y|\mathbf{X}_t = \mathbf{x}).$$

Let $\mathcal{C}$ denotes a fixed compact subset of $\mathbb{R}^p$ on which the marginal density of x is lower bounded by some positive constant. Under a set of regularity conditions, Paparoditis and Politis (2001, 2002) show that the one-step transition distribution function $F^*(y|x)$ that governs the law of the Markov bootstrap process satisfies the following uniform convergence property (see Theorem 3 in Paparoditis and Politis (2002)): $\sup_{y \in R} \sup_{\mathbf{x} \in \mathcal{C}} |F^*(y|\mathbf{x}) - F(y|\mathbf{x})| \to 0$ (*a.s.* - almost surely).

Now, to adapt this result to the case of out-of-sample forecasting, we replace the conditioning vector $\mathbf{x}$ by $\mathbf{X}_T$. Then, it is easy to see that

$$\sup_{y \in \mathbb{R}} |F^*(y|\mathbf{X}_T) - F(y|\mathbf{X}_T)| \mathbf{1}(\mathbf{X}_T \in \mathcal{C}) \leq \sup_{y \in R} \sup_{\mathbf{x} \in \mathcal{C}} |F^*(y|\mathbf{x}) - F(y|\mathbf{x})|$$

where $\mathbf{1}(A)$ denotes the indicator function for set A. Therefore,

$$\sup_{y \in \mathbb{R}} |F^*(y|\mathbf{X}_T) - F(y|\mathbf{X}_T)| \to 0 \quad a.s.$$

Because $F^*(y|\mathbf{X}_T)$ is the law that generates Y^*_{T+1} and $F(y|\mathbf{X}_T)$ the law that generates Y_{T+1}, we can replace Y_{T+1} by Y^*_{T+1}. For $\tau > 3$, the same argument holds by induction.

Table 1: **AR(1) model (P=300)**

n	Test	IND	LIN	MFD (fixed)				MFD (adaptive)			
				c = 0.5	c = 0.75	c = 1	c = 1.25	c = 0.5	c = 0.75	c = 1	c = 1.25
						$\tau = 1$					
300	KS	0.27	0.043	0.061	0.066	0.056	0.068	0.063	0.060	0.057	0.072
	SC	1.000	0.056	0.079	0.089	0.15	0.230	0.082	0.090	0.160	0.240
	HET	0.150	0.051	0.14	0.094	0.075	0.061	0.110	0.072	0.063	0.053
	V23	0.065	0.049	0.056	0.055	0.053	0.055	0.054	0.052	0.051	0.053
600	KS			0.056	0.054	0.054	0.064	0.053	0.052	0.056	0.069
	SC			0.070	0.075	0.110	0.160	0.074	0.079	0.110	0.170
	HET			0.110	0.069	0.063	0.053	0.076	0.056	0.052	0.048
	V23			0.055	0.067	0.059	0.047	0.058	0.062	0.064	0.053
900	KS			0.038	0.047	0.059	0.060	0.043	0.051	0.057	0.068
	SC			0.060	0.070	0.087	0.140	0.060	0.077	0.095	0.160
	HET			0.088	0.059	0.056	0.064	0.068	0.053	0.044	0.055
	V23			0.041	0.053	0.053	0.053	0.039	0.059	0.052	0.053
						$\tau = 2$					
300	KS	0.130	0.036	0.043	0.045	0.041	0.045	0.036	0.048	0.037	0.045
	SC	0.990	0.043	0.053	0.053	0.071	0.077	0.053	0.056	0.070	0.080
	HET	0.083	0.053	0.100	0.066	0.070	0.053	0.071	0.055	0.052	0.048
	V23	0.055	0.051	0.059	0.053	0.052	0.052	0.056	0.050	0.047	0.053
600	KS			0.045	0.035	0.038	0.047	0.040	0.036	0.040	0.045
	SC			0.060	0.051	0.055	0.067	0.057	0.053	0.059	0.071
	HET			0.080	0.065	0.049	0.047	0.067	0.056	0.041	0.047
	V23			0.056	0.051	0.046	0.043	0.049	0.051	0.048	0.044
900	KS			0.036	0.036	0.045	0.048	0.036	0.035	0.042	0.046
	SC			0.043	0.041	0.045	0.065	0.049	0.044	0.047	0.066
	HET			0.070	0.060	0.059	0.055	0.058	0.051	0.050	0.052
	V23			0.059	0.048	0.043	0.050	0.057	0.051	0.049	0.053
						$\tau = 3$					
300	KS	0.087	0.033	0.043	0.038	0.037	0.035	0.042	0.037	0.040	0.038
	SC	0.580	0.043	0.044	0.055	0.051	0.041	0.042	0.055	0.046	0.039
	HET	0.051	0.047	0.072	0.052	0.045	0.044	0.058	0.049	0.049	0.047
	V23	0.050	0.052	0.055	0.046	0.054	0.053	0.053	0.048	0.053	0.055
600	KS			0.042	0.044	0.035	0.039	0.039	0.037	0.037	0.037
	SC			0.046	0.052	0.043	0.044	0.044	0.050	0.045	0.047
	HET			0.070	0.056	0.045	0.040	0.060	0.049	0.047	0.040
	V23			0.055	0.056	0.044	0.045	0.053	0.052	0.039	0.043
900	KS			0.032	0.035	0.039	0.040	0.031	0.034	0.037	0.043
	SC			0.041	0.049	0.044	0.045	0.038	0.047	0.048	0.048
	HET			0.056	0.050	0.044	0.048	0.049	0.042	0.043	0.044
	V23			0.053	0.043	0.048	0.050	0.053	0.043	0.050	0.047

Percentage of rejections (at the 5% significance level) of the null hypothesis of the tests based on 2000 simulations. The forecasting methods are: **IND** resampling under independence, **LIN** linear AR model with bootstrap residuals, and **MFD** the method described in Section (2.2). The number of bootstraps is equal to 1000 for all methods. The lag in the MFD and the tests is equal to 1. The test are: **KS** = Kolmogorov-Smirnov test for uniformity, **SC** = LM test of no serial correlation of order 1, **HET** = LM (ARCH) test of no serial correlation of the squared residuals of order 1, **V23** = test for linearity.

Table 2: **ARCH(1)**

n	Test	IND	LIN	MFD (fixed)				MFD (adaptive)			
				c = 0.5	c = 0.75	c = 1	c = 1.25	c = 0.5	c = 0.75	c = 1	c = 1.25
$\boldsymbol{\tau = 1}$											
300	KS	0.043	0.044	0.059	0.059	0.051	0.049	0.057	0.055	0.055	0.051
	SC	0.075	0.058	0.053	0.051	0.056	0.044	0.051	0.049	0.048	0.045
	HET	0.820	0.820	0.260	0.190	0.170	0.180	0.190	0.160	0.150	0.170
	V23	0.073	0.073	0.051	0.050	0.047	0.049	0.052	0.051	0.047	0.047
600	KS			0.048	0.051	0.047	0.042	0.042	0.050	0.049	0.042
	SC			0.058	0.051	0.050	0.047	0.056	0.051	0.050	0.045
	HET			0.180	0.140	0.120	0.130	0.140	0.110	0.110	0.120
	V23			0.066	0.06	0.056	0.045	0.056	0.061	0.052	0.049
900	KS			0.051	0.050	0.040	0.047	0.046	0.045	0.043	0.051
	SC			0.051	0.058	0.047	0.050	0.053	0.056	0.048	0.051
	HET			0.150	0.120	0.110	0.120	0.120	0.087	0.091	0.110
	V23			0.058	0.060	0.049	0.051	0.057	0.060	0.048	0.050
$\boldsymbol{\tau = 2}$											
300	KS	0.038	0.035	0.050	0.041	0.043	0.043	0.047	0.042	0.043	0.041
	SC	0.053	0.056	0.050	0.056	0.048	0.051	0.049	0.053	0.050	0.049
	HET	0.075	0.083	0.130	0.096	0.073	0.061	0.100	0.074	0.054	0.056
	V23	0.057	0.057	0.055	0.056	0.045	0.043	0.057	0.051	0.047	0.039
600	KS			0.041	0.037	0.047	0.037	0.039	0.035	0.045	0.036
	SC			0.049	0.051	0.040	0.047	0.047	0.047	0.043	0.043
	HET			0.110	0.081	0.056	0.062	0.073	0.066	0.045	0.048
	V23			0.058	0.051	0.050	0.050	0.055	0.052	0.047	0.045
900	KS			0.045	0.041	0.036	0.041	0.044	0.036	0.030	0.044
	SC			0.055	0.055	0.049	0.045	0.059	0.055	0.045	0.047
	HET			0.099	0.073	0.060	0.060	0.075	0.064	0.052	0.051
	V23			0.063	0.053	0.044	0.057	0.058	0.054	0.040	0.054
$\boldsymbol{\tau = 3}$											
300	KS	0.036	0.038	0.047	0.042	0.044	0.045	0.043	0.043	0.045	0.043
	SC	0.041	0.039	0.058	0.058	0.053	0.041	0.054	0.057	0.053	0.040
	HET	0.043	0.045	0.086	0.070	0.056	0.050	0.073	0.066	0.047	0.046
	V23	0.052	0.057	0.056	0.064	0.054	0.044	0.052	0.061	0.051	0.041
600	KS			0.038	0.040	0.038	0.035	0.036	0.042	0.037	0.037
	SC			0.046	0.050	0.051	0.051	0.040	0.043	0.054	0.051
	HET			0.070	0.061	0.057	0.049	0.053	0.053	0.052	0.047
	V23			0.063	0.065	0.058	0.048	0.053	0.052	0.058	0.045
900	KS			0.044	0.043	0.042	0.044	0.042	0.042	0.040	0.045
	SC			0.050	0.040	0.061	0.050	0.047	0.036	0.058	0.048
	HET			0.070	0.059	0.053	0.047	0.061	0.052	0.050	0.048
	V23			0.055	0.053	0.047	0.048	0.055	0.048	0.044	0.045

Percentage of rejections (at the 5% significance level) of the null hypothesis of the tests based on 2000 simulations. The forecasting methods are: **IND** resampling under independence, **LIN** linear AR model with bootstrap residuals, and **MFD** the method described in Section (2.2). The number of bootstraps is equal to 1000 for all methods. The lag in the MFD and the tests is equal to 1. The test are: **KS** = Kolmogorov-Smirnov test for uniformity, **SC** = LM test of no serial correlation of order 1, **HET** = LM (ARCH) test of no serial correlation of the squared residuals of order 1, **V23** = test for linearity.

Table 3: **SETAR(1), homoskedastic regimes (P = 300)**

n	Test	IND	LIN	MFD (fixed)				MFD (adaptive)			
				c = 0.5	c = 0.75	c = 1	c = 1.25	c = 0.5	c = 0.75	c = 1	c = 1.25
						$\tau = 1$					
300	KS	0.110	0.065	0.059	0.049	0.054	0.047	0.060	0.047	0.055	0.049
	SC	0.980	0.130	0.060	0.058	0.049	0.045	0.060	0.055	0.050	0.053
	HET	0.066	0.088	0.150	0.120	0.088	0.070	0.140	0.110	0.086	0.064
	V23	0.960	0.940	0.061	0.059	0.085	0.110	0.061	0.059	0.097	0.120
600	KS			0.048	0.043	0.039	0.047	0.045	0.042	0.042	0.051
	SC			0.045	0.045	0.051	0.040	0.045	0.049	0.055	0.046
	HET			0.110	0.089	0.077	0.069	0.110	0.086	0.072	0.064
	V23			0.056	0.070	0.067	0.081	0.053	0.073	0.077	0.092
900	KS			0.044	0.048	0.042	0.042	0.047	0.051	0.051	0.043
	SC			0.048	0.048	0.049	0.047	0.048	0.049	0.053	0.051
	HET			0.094	0.075	0.064	0.044	0.090	0.069	0.059	0.045
	V23			0.056	0.056	0.064	0.081	0.056	0.060	0.070	0.100
						$\tau = 2$					
300	KS	0.056	0.049	0.050	0.041	0.035	0.037	0.052	0.043	0.037	0.034
	SC	0.085	0.036	0.058	0.043	0.041	0.047	0.059	0.040	0.036	0.043
	HET	0.056	0.057	0.100	0.069	0.049	0.049	0.096	0.061	0.048	0.047
	V23	0.120	0.110	0.056	0.049	0.051	0.048	0.054	0.045	0.050	0.050
600	KS			0.036	0.035	0.034	0.043	0.042	0.036	0.035	0.039
	SC			0.048	0.045	0.043	0.050	0.046	0.044	0.045	0.050
	HET			0.076	0.066	0.056	0.042	0.061	0.067	0.051	0.047
	V23			0.055	0.051	0.052	0.050	0.051	0.051	0.053	0.051
900	KS			0.042	0.048	0.048	0.043	0.041	0.049	0.038	0.036
	SC			0.047	0.049	0.064	0.041	0.047	0.045	0.061	0.040
	HET			0.075	0.059	0.045	0.047	0.069	0.054	0.049	0.047
	V23			0.056	0.047	0.054	0.054	0.058	0.050	0.057	0.055
						$\tau = 3$					
300	KS	0.042	0.035	0.047	0.042	0.041	0.040	0.051	0.043	0.042	0.042
	SC	0.043	0.049	0.055	0.049	0.047	0.048	0.057	0.050	0.045	0.050
	HET	0.048	0.044	0.083	0.050	0.051	0.051	0.079	0.048	0.047	0.053
	V23	0.055	0.057	0.060	0.057	0.053	0.046	0.054	0.056	0.055	0.048
600	KS			0.034	0.034	0.043	0.041	0.036	0.035	0.043	0.043
	SC			0.050	0.036	0.050	0.044	0.048	0.037	0.050	0.043
	HET			0.063	0.055	0.046	0.048	0.061	0.052	0.051	0.047
	V23			0.062	0.063	0.047	0.043	0.061	0.061	0.043	0.043
900	KS			0.044	0.044	0.039	0.037	0.039	0.047	0.037	0.034
	SC			0.051	0.043	0.048	0.047	0.048	0.045	0.046	0.046
	HET			0.061	0.049	0.042	0.044	0.053	0.052	0.041	0.045
	V23			0.053	0.055	0.052	0.051	0.055	0.050	0.055	0.051

Percentage of rejections (at the 5% significance level) of the null hypothesis of the tests based on 2000 simulations. The forecasting methods are: **IND** resampling under independence, **LIN** linear AR model with bootstrap residuals, and **MFD** the method described in Section (2.2). The number of bootstraps is equal to 1000 for all methods. The lag in the MFD and the tests is equal to 1. The test are: **KS** = Kolmogorov-Smirnov test for uniformity, **SC** = LM test of no serial correlation of order 1, **HET** = LM (ARCH) test of no serial correlation of the squared residuals of order 1, **V23** = test for linearity.

Table 4: **SETAR(1), heteroskedastic regimes (P = 300)**

n	Test	IND	LIN	MFD (fixed)				MFD (adaptive)			
				c = 0.5	c = 0.75	c = 1	c = 1.25	c = 0.5	c = 0.75	c = 1	c = 1.25
$\boldsymbol{\tau} = \mathbf{1}$											
300	KS	0.110	0.077	0.052	0.045	0.066	0.064	0.053	0.050	0.060	0.061
	SC	0.560	0.100	0.057	0.058	0.063	0.070	0.056	0.055	0.056	0.062
	HET	0.160	0.190	0.210	0.160	0.110	0.060	0.180	0.130	0.087	0.057
	V23	0.990	0.990	0.073	0.088	0.120	0.190	0.073	0.097	0.140	0.220
600	KS			0.051	0.036	0.048	0.060	0.050	0.043	0.049	0.060
	SC			0.052	0.067	0.067	0.065	0.051	0.067	0.060	0.057
	HET			0.130	0.090	0.077	0.075	0.110	0.083	0.069	0.067
	V23			0.071	0.079	0.099	0.150	0.070	0.083	0.110	0.160
900	KS			0.044	0.043	0.045	0.048	0.048	0.041	0.041	0.055
	SC			0.055	0.061	0.057	0.067	0.051	0.058	0.056	0.063
	HET			0.120	0.082	0.070	0.058	0.097	0.072	0.061	0.058
	V23			0.059	0.069	0.079	0.140	0.061	0.078	0.092	0.160
$\boldsymbol{\tau} = \mathbf{2}$											
300	KS	0.072	0.087	0.051	0.048	0.051	0.047	0.053	0.051	0.051	0.047
	SC	0.250	0.170	0.053	0.036	0.046	0.047	0.058	0.036	0.046	0.049
	HET	0.180	0.190	0.140	0.110	0.082	0.069	0.120	0.098	0.067	0.059
	V23	0.089	0.084	0.069	0.051	0.043	0.049	0.060	0.048	0.048	0.045
600	KS			0.047	0.043	0.039	0.049	0.048	0.042	0.042	0.053
	SC			0.055	0.052	0.048	0.052	0.055	0.049	0.049	0.052
	HET			0.110	0.083	0.069	0.068	0.099	0.075	0.057	0.060
	V23			0.049	0.055	0.059	0.056	0.045	0.054	0.059	0.056
900	KS			0.040	0.046	0.039	0.040	0.038	0.042	0.042	0.036
	SC			0.050	0.052	0.061	0.053	0.052	0.049	0.064	0.051
	HET			0.092	0.076	0.058	0.056	0.081	0.072	0.056	0.056
	V23			0.056	0.054	0.053	0.060	0.051	0.051	0.047	0.060
$\boldsymbol{\tau} = \mathbf{3}$											
300	KS	0.048	0.060	0.043	0.045	0.051	0.032	0.048	0.046	0.051	0.032
	SC	0.052	0.054	0.052	0.045	0.057	0.056	0.056	0.049	0.059	0.059
	HET	0.043	0.043	0.110	0.064	0.063	0.053	0.097	0.052	0.053	0.050
	V23	0.064	0.063	0.061	0.051	0.051	0.050	0.066	0.051	0.054	0.043
600	KS			0.048	0.033	0.042	0.043	0.047	0.033	0.039	0.042
	SC			0.060	0.051	0.052	0.047	0.060	0.049	0.050	0.049
	HET			0.069	0.059	0.053	0.047	0.064	0.056	0.051	0.047
	V23			0.055	0.056	0.057	0.058	0.051	0.054	0.054	0.059
900	KS			0.036	0.052	0.037	0.033	0.040	0.048	0.037	0.041
	SC			0.044	0.045	0.057	0.049	0.050	0.046	0.053	0.047
	HET			0.066	0.053	0.050	0.047	0.066	0.051	0.050	0.046
	V23			0.057	0.049	0.055	0.043	0.056	0.063	0.051	0.043

Percentage of rejections (at the 5% significance level) of the null hypothesis of the tests based on 2000 simulations. The forecasting methods are: **IND** resampling under independence, **LIN** linear AR model with bootstrap residuals, and **MFD** the method described in Section (2.2). The number of bootstraps is equal to 1000 for all methods. The lag in the MFD and the tests is equal to 1. The test are: **KS** = Kolmogorov-Smirnov test for uniformity, **SC** = LM test of no serial correlation of order 1, **HET** = LM (ARCH) test of no serial correlation of the squared residuals of order 1, **V23** = test for linearity.

Table 5: **SETAR(2)**

n	Test	IND	LIN	MFD (fixed)				MFD (adaptive)			
				c = 0.5	c = 0.75	c = 1	c = 1.25	c = 0.5	c = 0.75	c = 1	c = 1.25
						$\boldsymbol{\tau = 1}$					
300	KS	0.120	0.065	0.550	0.086	0.060	0.066	0.460	0.081	0.061	0.062
	SC	0.970	0.090	0.059	0.058	0.052	0.059	0.058	0.059	0.051	0.054
	HET	0.070	0.071	0.590	0.310	0.160	0.089	0.350	0.170	0.100	0.074
	V23	0.850	0.790	0.077	0.071	0.086	0.110	0.069	0.073	0.088	0.120
600	KS			0.290	0.064	0.052	0.070	0.230	0.057	0.053	0.070
	SC			0.055	0.055	0.051	0.051	0.058	0.051	0.048	0.050
	HET			0.520	0.250	0.120	0.081	0.280	0.130	0.086	0.070
	V23			0.071	0.070	0.067	0.093	0.067	0.066	0.074	0.100
900	KS			0.170	0.059	0.043	0.054	0.140	0.053	0.044	0.050
	SC			0.064	0.053	0.057	0.059	0.061	0.050	0.056	0.055
	HET			0.440	0.200	0.110	0.060	0.230	0.100	0.069	0.047
	V23			0.073	0.077	0.080	0.096	0.064	0.068	0.079	0.100
						$\boldsymbol{\tau = 2}$					
300	KS	0.120	0.059	0.260	0.072	0.051	0.058	0.220	0.073	0.052	0.061
	SC	0.850	0.065	0.055	0.050	0.053	0.051	0.048	0.048	0.051	0.056
	HET	0.071	0.077	0.170	0.096	0.055	0.062	0.150	0.083	0.051	0.060
	V23	0.770	0.730	0.063	0.071	0.073	0.110	0.061	0.076	0.083	0.110
600	KS			0.130	0.043	0.041	0.051	0.120	0.049	0.043	0.051
	SC			0.057	0.044	0.049	0.050	0.056	0.052	0.044	0.052
	HET			0.130	0.086	0.059	0.058	0.130	0.075	0.059	0.058
	V23			0.063	0.066	0.064	0.086	0.056	0.070	0.068	0.091
900	KS			0.092	0.050	0.044	0.042	0.082	0.044	0.046	0.043
	SC			0.058	0.054	0.049	0.044	0.057	0.052	0.049	0.043
	HET			0.130	0.083	0.060	0.055	0.100	0.080	0.056	0.055
	V23			0.055	0.063	0.052	0.077	0.057	0.059	0.054	0.080
						$\boldsymbol{\tau = 3}$					
300	KS	0.044	0.044	0.190	0.063	0.040	0.044	0.180	0.057	0.033	0.048
	SC	0.047	0.047	0.050	0.050	0.051	0.055	0.051	0.049	0.052	0.052
	HET	0.044	0.043	0.110	0.072	0.056	0.035	0.100	0.063	0.053	0.037
	V23	0.039	0.037	0.049	0.045	0.042	0.051	0.047	0.043	0.046	0.050
600	KS			0.110	0.050	0.041	0.036	0.098	0.052	0.043	0.040
	SC			0.052	0.052	0.045	0.044	0.061	0.055	0.045	0.047
	HET			0.090	0.063	0.051	0.048	0.088	0.062	0.051	0.047
	V23			0.051	0.051	0.049	0.039	0.051	0.052	0.039	0.044
900	KS			0.095	0.047	0.040	0.043	0.081	0.047	0.042	0.039
	SC			0.053	0.048	0.045	0.041	0.050	0.050	0.043	0.045
	HET			0.096	0.070	0.044	0.042	0.083	0.062	0.043	0.040
	V23			0.067	0.056	0.051	0.043	0.058	0.053	0.048	0.048

Percentage of rejections (at the 5% significance level) of the null hypothesis of the tests based on 2000 simulations. The forecasting methods are: **IND** resampling under independence, **LIN** linear AR model with bootstrap residuals, and **MFD** the method described in Section (2.2). The number of bootstraps is equal to 1000 for all methods. The lag in the MFD and the tests is equal to 1. The test are: **KS** = Kolmogorov-Smirnov test for uniformity, **SC** = LM test of no serial correlation of order 1, **HET** = LM (ARCH) test of no serial correlation of the squared residuals of order 1, **V23** = test for linearity.

Table 6: **SETAR(1-2)**

n	Test	IND	LIN	MFD (fixed)				MFD (adaptive)			
				c = 0.5	c = 0.75	c = 1	c = 1.25	c = 0.5	c = 0.75	c = 1	c = 1.25
$\tau = 1$											
300	KS	0.040	0.060	0.330	0.065	0.036	0.051	0.280	0.063	0.042	0.056
	SC	1.000	0.130	0.070	0.100	0.160	0.240	0.075	0.110	0.180	0.250
	HET	0.220	0.280	0.510	0.220	0.110	0.076	0.290	0.130	0.071	0.061
	V23	0.980	0.930	0.077	0.081	0.120	0.190	0.072	0.085	0.140	0.220
600	KS			0.150	0.053	0.043	0.042	0.130	0.051	0.041	0.049
	SC			0.071	0.086	0.120	0.180	0.079	0.097	0.140	0.200
	HET			0.410	0.150	0.097	0.070	0.230	0.090	0.081	0.051
	V23			0.061	0.077	0.099	0.170	0.063	0.070	0.110	0.180
900	KS			0.120	0.055	0.048	0.048	0.096	0.049	0.050	0.051
	SC			0.063	0.072	0.110	0.160	0.068	0.085	0.110	0.170
	HET			0.340	0.140	0.076	0.062	0.180	0.084	0.076	0.055
	V23			0.069	0.068	0.085	0.140	0.063	0.072	0.085	0.150
$\tau = 2$											
300	KS	0.280	0.044	0.170	0.070	0.077	0.065	0.170	0.069	0.072	0.071
	SC	1.000	0.038	0.058	0.063	0.091	0.110	0.055	0.065	0.110	0.120
	HET	0.080	0.045	0.140	0.089	0.055	0.056	0.110	0.064	0.048	0.048
	V23	0.120	0.130	0.060	0.045	0.051	0.060	0.054	0.056	0.055	0.063
600	KS			0.100	0.051	0.047	0.068	0.091	0.055	0.055	0.071
	SC			0.061	0.054	0.068	0.095	0.061	0.058	0.079	0.110
	HET			0.120	0.079	0.060	0.051	0.083	0.061	0.048	0.050
	V23			0.055	0.047	0.063	0.062	0.058	0.050	0.063	0.057
900	KS			0.075	0.054	0.051	0.054	0.071	0.057	0.054	0.057
	SC			0.059	0.055	0.069	0.072	0.057	0.057	0.072	0.083
	HET			0.110	0.073	0.057	0.044	0.089	0.063	0.055	0.040
	V23			0.058	0.052	0.055	0.050	0.051	0.051	0.049	0.056
$\tau = 3$											
300	KS	0.130	0.120	0.130	0.041	0.039	0.030	0.120	0.041	0.039	0.032
	SC	0.052	0.058	0.052	0.058	0.059	0.069	0.058	0.057	0.063	0.070
	HET	0.092	0.065	0.092	0.052	0.050	0.039	0.065	0.056	0.048	0.043
	V23	0.060	0.054	0.060	0.046	0.050	0.046	0.054	0.048	0.052	0.047
600	KS			0.080	0.041	0.039	0.032	0.076	0.036	0.035	0.029
	SC			0.056	0.048	0.051	0.064	0.059	0.047	0.047	0.067
	HET			0.074	0.060	0.050	0.046	0.059	0.056	0.045	0.051
	V23			0.059	0.045	0.041	0.049	0.059	0.043	0.043	0.052
900	KS			0.057	0.045	0.041	0.030	0.057	0.043	0.040	0.026
	SC			0.059	0.047	0.055	0.058	0.057	0.050	0.059	0.057
	HET			0.071	0.051	0.055	0.053	0.072	0.048	0.051	0.051
	V23			0.055	0.059	0.048	0.048	0.051	0.058	0.050	0.051

Percentage of rejections (at the 5% significance level) of the null hypothesis of the tests based on 2000 simulations. The forecasting methods are: **IND** resampling under independence, **LIN** linear AR model with bootstrap residuals, and **MFD** the method described in Section (2.2). The number of bootstraps is equal to 1000 for all methods. The lag in the MFD and the tests is equal to 1. The test are: **KS** = Kolmogorov-Smirnov test for uniformity, **SC** = LM test of no serial correlation of order 1, **HET** = LM (ARCH) test of no serial correlation of the squared residuals of order 1, **V23** = test for linearity.

Table 7: **Comparison of Density Forecast Models**

Forecasting Method	KS	SC	HET	V23
	τ = 1			
IND	0.006	0.000	0.353	0.041
LIN	0.129	0.067	0.493	0.026
SETAR	0.046	0.842	0.564	0.225
MFD (fixed)	0.918	0.119	0.825	0.149
MFD (adaptive)	0.676	0.192	0.824	0.279
	τ = 2			
IND	0.000	0.025	0.156	0.294
LIN	0.007	0.594	0.189	0.169
SETAR	0.014	0.038	0.022	0.158
MFD (fixed)	0.372	0.071	0.765	0.386
MFD (adaptive)	0.441	0.151	0.255	0.401
	τ = 3			
IND	0.068	0.059	0.858	0.278
LIN	0.155	0.110	0.904	0.316
SETAR	0.021	0.110	0.839	0.330
MFD (fixed)	0.336	0.154	0.282	0.239
MFD (adaptive)	0.512	0.193	0.640	0.236

Density forecasts for monthly US Industrial Production from January 1960 until April 2004. The out-of-sample forecast period starts in January 1986 (total of 220 forecasts). The forecasting methods used are: **IND** bootstrap under independence, **LIN** linear AR, **SETAR** indicates the SETAR specification in Equation (8), **MFD** using both the fixed and adaptive bandwidth rule. Reported are the p-values of the tests described in Section (3.2). For SC, HET and V23 we set the lag of the tests to 5. For the MFD we set the order p to 3. The lag order for the AR and SETAR methods are chosen performing a search up to 6 lags using the AIC criterion.