Econometric Theory, **26**, 2010, 1363–1397.
doi:10.1017/S0266466609990612

TESTS OF THE MARTINGALE DIFFERENCE HYPOTHESIS USING BOOSTING AND RBF NEURAL NETWORK APPROXIMATIONS

GEORGE KAPETANIOS
Queen Mary, University of London

ANDREW P. BLAKE
Bank of England

The martingale difference restriction is an outcome of many theoretical analyses in economics and finance. A large body of econometric literature deals with tests of that restriction. We provide new tests based on radial basis function (RBF) neural networks. Our work is based on the test design of Blake and Kapetanios (2000, 2003a, 2003b). However, unlike that work we provide a formal theoretical justification for the validity of these tests and present some new general theoretical results. These results take advantage of the link between the algorithms of Blake and Kapetanios (2000, 2003a, 2003b) and boosting. We carry out a Monte Carlo study of the properties of the new tests and find that they have very good power performance. A simplified implementation of boosting is found to have desirable properties and small computational cost. An empirical application to the S&P 500 constituents illustrates the usefulness of our new test.

1. INTRODUCTION

The martingale or martingale difference restriction arises repeatedly in finance and economics. Models using rational expectations, market efficiency, and similar theoretical frameworks impose this restriction on economic variables such as consumption and stock returns. From an econometric point of view, the martingale difference hypothesis (MDH) amounts to the statement that the best predictor of a covariance stationary stochastic process at any point in time, conditional on the currently available information set, is equal to the unconditional expectation. Tests of this restriction are useful, as tools for falsifying economic theories.

A number of such tests have been proposed in the literature. Bierens and Ploberger (1991) provided a test based on the fact that under the MDH the spectral

We would like to thank the Bruce Hansen, co-editor, and two anonymous referees for very valuable comments that substantially improved the paper. Address correspondence to George Kapetanios, Department of Economics, Queen Mary, University of London, Mile End Road, London E1 4NS, UK; e-mail: G.Kapetanios@qmul.ac.uk.

© Cambridge University Press 2010 0266-4666/10 $15.00

https://doi.org/10.1017/S0266466609990612 Published online by Cambridge University Press

distribution function is a straight line. Deo (2000) has provided extensions of this test to conditional heteroskedasticity. As noted by Escanciano and Velasco(2006a), the test based on the spectral distribution function is not consistent against all deviations from the MDH and, in particular, it cannot detect deviations that imply lack of autocorrelation. Escanciano and Velasco (2006a) propose a new test based on the fact that, for a process y_t that satisfies the MDH,

$$\mathrm{E}(y_t|I_{t-1}) = \delta \quad \text{a.s.} \Leftrightarrow \mathrm{E}((y_t - \delta)\mu(I_{t-1})) = 0 \tag{1}$$

for some constant δ and any $\mathcal{F}_{t-1}$ measurable function $\mu(.)$, where $I_{t-1} = (y_{t-1}, y_{t-2}, ...)'$, $\mathcal{F}_t$ is the σ-field generated by I_{t-1}. Noting the equivalence in (1) links the MDH testing problem to a large specification testing literature that aims to capture deviations from some parametric null hypothesis and uses tests based on particular forms for $\mu(.)$ to do so. The most popular forms for $\mu(.)$ are the exponential function used by Bierens (1984, 1990), Bierens and Ploberger (1997), Hong (1999a, 1999b), and de Jong (1996) and the indicator function used by Stute (1997), Koul and Stute (1999), Park and Whang (1999), Whang (2000), and Dominguez and Lobato (2003). The recent test of the MDH by Escanciano and Velasco (2006a) is based on the work of Hong (1999a,1999b) and also uses the exponential function. Finally, we note the work of Escanciano (2006), which combines exponential and indicator weights.

The strong focus on the exponential function as a tool for deriving specification tests for deviations from parametric null models has been questioned by Stinchcombe and White (1998), who argue that there is nothing special about the exponential function (or indeed the indicator function) that makes it capable of detecting arbitrary deviations from parametric null models. They show that most bases of functions are capable of this, with the exception of polynomials. In particular, they note that neural network specifications are powerful approximators whose approximation properties have been established formally in the literature (see, e.g., Hornik, Stinchcombe, and White, 1989).

In specification testing, the focus on the exponential and indicator functions can be partly explained by the lack of robust and efficient estimation algorithms for flexible nonlinear specifications that could play the role of μ. As a result, focus has been placed on the exponential function or, more generally, specifications that are restricted to involve linear combinations of basis functions, such as trigonometric functions. Such basis functions do not involve unknown parameters and, therefore, estimation boils down to linear least squares estimation of the linear combination coefficients. Such restrictions, however, have considerable costs in the sense that many classes of powerful flexible nonlinear specifications are excluded.

In a series of papers, Blake and Kapetanios (2000, 2003a, 2003b) introduce a new class of neural networks in the context of a diverse set of testing problems in econometrics. These neural network specifications, based on radial basis function

https://doi.org/10.1017/S0266466609990612 Published online by Cambridge University Press

neural networks (RBFNN), provide a novel way to alleviate the aforementioned estimation (and in some cases identification) problem. This earlier work focuses on small sample performance, but in the current paper we provide a rigorous justification for their setup using ideas from boosting.[1]

This paper uses the equivalence in (1) to propose regression-based tests for the MDH based on the neural network testing procedures of Blake and Kapetanios (2000, 2003a, 2003b). In particular, (1) immediately implies that for all $\mu(I_{t-1})$, the null hypothesis that $\alpha = 0$ in

$$y_t = \delta + \alpha \mu(I_{t-1}) + \epsilon_t, \tag{2}$$

where ϵ_t is assumed to be a martingale difference process, holds. This fact can be used to construct valid tests for the MDH. Unlike previous work on MDH tests, we use neural network approximations to choose μ. We provide novel theoretical results for our testing procedure and carry out a Monte Carlo study that suggests that the new procedures provide superior power performance compared to the most powerful existing MDH tests in the literature.

The structure of the paper is as follows: Section 2 presents the new testing procedures. Section 3 presents some theoretical results for a general neural network algorithm. Section 4 provides some theoretical results for the MDH testing procedures. Section 5 presents a Monte Carlo study. Section 6 provides an empirical application. Finally, Section 7 concludes.

2. SETUP

Consider a stochastic process $y_t, t = 1, ..., T$. We wish to test the MDH that

$$\mathrm{E}(y_t | I_{t-1}) = \delta \ \ \forall t \ \text{ a.s.} \tag{3}$$

By the equivalence stated in (1) it follows that there is no $\mathcal{F}_{t-1}$ measurable function $\mu(I_{t-1})$ such that $\alpha \neq 0$ in a regression model of the form

$$y_t = \delta + \alpha' \mu(x_t) + \epsilon_t, \tag{4}$$

where $x_t = (y_{t-1}, ..., y_{t-q})'$. Therefore the problem of testing the MDH becomes one of testing $\alpha = 0$, for some appropriate function $\mu(.)$ where both $\mu(.)$ and α can be either scalar or, more generally, vectors of functions and coefficients, respectively. A standard Wald test can be applied. The main issue is to construct $\mu(.)$ so as to have appropriate performance both under the null MDH and under a wide variety of alternative hypotheses. We wish to devise a portmanteau test and so need to provide a method that is nonparametric in the sense that it can capture any function for which $\alpha \neq 0$.

Given the work of Blake and Kapetanios (2000, 2003a, 2003b, 2007), who show that RBFNN specifications can be used to construct powerful tests for a wide variety of alternative hypotheses in different regression contexts, our aim is

https://doi.org/10.1017/S0266466609990612 Published online by Cambridge University Press

to estimate the unknown regression function by an RBFNN series expansion of the form

$$\hat{\mu}(x_t) = \sum_{i=1}^{m} c_i \psi(x_t, t_i, \sigma_T), \tag{5}$$

where the RBF nodes, $\psi(x_t, t_i, \sigma_T)$, are radially symmetrical, integrable, bounded functions; t_i are referred to as the centers of the RBFs; and σ_T is a sequence of constants. Examples include the Gaussian function of the form $\exp\left(-\left(\frac{||x-t_i||}{\sigma_T}\right)^2\right)$ or the multiquadric function $\left(1+\left(\frac{||x-t_i||}{\sigma_T}\right)^2\right)^{-1}$, $\sigma_T > 0$, where $||.||$ denotes Euclidean distance. Obviously, estimation of (5) is challenging since, unlike standard series expansions, there are two problems that need attention. The first is that $\psi(x, t_i, \sigma_T)$ contain unknown parameters, in particular the centers, and the second is that the nodes are not ranked, so that the choice of the nodes in the series expansion is not obvious. Once the order of the nodes and the centers are determined, the series expansion can be estimated by least squares.

A popular approach to the solution of both of the above problems was suggested by Orr (1995), who suggested a form of forward selection procedure using every data point as a potential center. In a series of papers, Blake and Kapetanios (2000, 2003a, 2003b, 2007) have modified that algorithm for specifically econometric applications with some success. In this paper we modify it further to bring it more in line with the regression-based boosting algorithm of Buhlmann (2006) and to provide the first theoretical results for this algorithm. We define this as Algorithm 1 and label it as the (RBF) MDH Boosting Algorithm.

Algorithm 1: (RBF) MDH Boosting Algorithm

1. Let σ_T be some sequence such that $\sigma_T = o(1)$. We construct the initial set of T RBF nodes given by $\Psi^{(1,\ldots,T)} = \{\psi(x, x_1, \sigma_T), \psi(x, x_2, \sigma_T), \ldots, \psi(x, x_T, \sigma_T)\}$.
2. These are ranked according to their ability to reduce the residual variance, when each $\psi(x_t, x_i, \sigma_T)$, $i = 1, \ldots, T$, is entered individually in (4).
3. The node that minimizes the residual variance becomes the first node in the ranking of the nodes. Denote this node by $\psi(x, x_{\mathcal{S}_1}, \sigma_T)$. Denote the residual from the regression of y_t on $\psi(x_t, x_{\mathcal{S}_1}, \sigma_T)$, by $y_t^{(1)}$. Let $\tilde{\mathcal{S}}_1 = \{\mathcal{S}_1\}$. Let $\Psi^{(1,\ldots,T)/\tilde{\mathcal{S}}_1}$ be the set of nodes in $\Psi^{(1,\ldots,T)}$ apart from the nodes indexed by the elements of $\tilde{\mathcal{S}}_1$.
4. Set $i = 1$.
5. The nodes in $\Psi^{(1,\ldots,T)/\tilde{\mathcal{S}}_1}$ are ranked according to their ability to reduce the residual variance of $y_t^{(i)}$, when $y_t^{(i)}$ is regressed on each $\psi(x_t, x_i, \sigma_T)$, $i \in \tilde{\mathcal{S}}_1$.
6. The node that minimizes the residual variance becomes the $i+1$th node in the ranking of the nodes. Denote this node by $\psi(x, x_{\mathcal{S}_{i+1}}, \sigma_T)$. Denote the residual from the regression of $y_t^{(i)}$ on $\psi(x_t, x_{\mathcal{S}_{i+1}}, \sigma_T)$, by $y_t^{(i+1)}$. Let

https://doi.org/10.1017/S0266466609990612 Published online by Cambridge University Press

$\tilde{\mathcal{S}}_{i+1} = \tilde{\mathcal{S}}_i \cup \{\mathcal{S}_{i+1}\}$. Let $\Psi^{(1,\ldots,T)/\tilde{\mathcal{S}}_{i+1}}$ be the set of nodes in $\Psi^{(1,\ldots,T)}$ apart from the nodes indexed by the elements of $\tilde{\mathcal{S}}_{i+1}$.

7. If $i = m$ for some $m = m_T \to \infty$, stop; otherwise, set $i = i + 1$ and go to Step 5.

Some remarks are in order for this algorithm.

Remark 1. The choice for m is not discussed in Algorithm 1 apart from noting that $m \to \infty$. Theorem 1 below suggests that the maximum possible rate is logarithmic in T.

Remark 2. The sequence σ_T is left unspecified in Algorithm 1. The proof of Theorem 1 below suggests that the choice $\sigma_T = O\left((\ln \ln T)^{-1}\right)$ is acceptable. It is worth noting here that the issue of selecting σ_T has not been addressed from a theoretical perspective in the literature. For example, Moody and Darken (1989) and Orr (1995), who discuss this issue in some detail, suggest a variety of ad hoc data-based rules of thumb but do not provide any theoretical discussion. Accordingly, and following from the work of Orr (1995), Blake and Kapetanios (2000, 2003a, 2003b, 2007) find that setting $\sigma_T = \sigma$, where $\sigma = 2 \max_t |x_t - x_{t-1}|$, provides good small sample results in a variety of econometric problems. Since to the best of our knowledge there is no theoretical work in the neural network literature on data-dependent methods to choose σ, and given this is primarily an issue for the RBFNN specification, we do not consider possible new solutions but stick to existing practices. Therefore, the above setting for σ_T is our choice for the Monte Carlo study.

Remark 3. The choice of the initial set of RBF nodes given by

$$\Psi^{(1,\ldots,T)} = \{\psi(x, x_1, \sigma_T), \psi(x, x_2, \sigma_T), \ldots, \psi(x, x_T, \sigma_T)\}$$

may be straightforwardly generalized to $\Psi^{(1,\ldots,p_T)}$, where p_T is chosen to reflect a subset of the observations or possibly be of a larger order than T where candidate values other than actual observations are, in principle, possibly used. More generally, we introduce the sequence τ_T to index the bounds of the grid used to construct candidate values for the centers. In particular, we specify the candidate centers to be chosen so as to belong in $[-\tau_T, \tau_T]^q$. This sequence will be used in the ensuing theoretical analysis.

Remark 4. Algorithm 1 is more computationally demanding than that used in Blake and Kapetanios (2000, 2003a, 2003b, 2007). There the nodes are ranked only once, according to their ability to reduce the residual variance, when entered individually in (5). Although the two algorithms are very similar, the additional cost is an increase in computational effort of the order of $T(T+1)/2$ for Algorithm 1.

Remark 5. Although the discussion in this paper is couched in terms of RBFNNs, it is worth noting that extensions to other neural network specifications

https://doi.org/10.1017/S0266466609990612 Published online by Cambridge University Press

such as neural networks based on logistic function nodes are possible once a grid of possible parameter values is constructed. One such specification is considered in White (2006), where a similar algorithm is constructed but no formal theoretical justification for it is given. The advantage of RBFNNs, in the context of Algorithm 1, is the fact that the construction of the grid is obtained by using the actual sample observations, thus ensuring an appropriate coverage of the relevant state space for the processes under consideration.

Once an ordered set of m RBFNN nodes is available via Algorithm 1, a data-dependent method can be used to determine the final number of nodes to enter in the testing regression. Guay and Guerre (2006) provide a theoretical analysis of tests based on similar series expansions[2] and suggest the use of a data-dependent method to determine the final number of nodes to enter in the testing regression. Their method depends on a penalty term of order $(\ln \ln T)^{1/2}$ to counterbalance the increase in fit from the use of more nodes in the testing regression. In particular, they suggest that the number of nodes, k^*, finally used in the testing regression be given by

$$k^* = \text{argmax}_{k=k_{\min},...,m} \left\{ R_k - \zeta_{T,k} \right\} \qquad \text{for some } k_{\min},$$

where $\zeta_{T,k} = k - \sqrt{2\gamma_T k}$ is a penalty term, $\gamma_T = \ln \ln T$, $R_k = y' \Psi_k \left(\Psi_k' \Psi_k \right)^{-1} \Psi_k' y$, $\Psi_k = (\psi_{1,k}, ..., \psi_{k,k})$, and $\psi_{i,k} = (\psi(x_1, x_{\mathcal{S}_i}, \sigma_T), ..., \psi(x_T, x_{\mathcal{S}_i}, \sigma_T)'$, $i = 1, ..., k$. This is similar to the method adopted in Blake and Kapetanios (2000, 2003a, 2003b, 2007). The penalty terms used in Blake and Kapetanios (2003b) are the ones associated with either the Akaike or the Bayesian information criteria. These penalties are not optimal in the sense of Guay and Guerre (2006), since the Akaike penalty term, given by a finite constant, results in a test that does not have an asymptotic χ^2 approximation, whereas the Bayesian criterion, with a penalty term of order $\ln T$, is too parsimonious. In the context of the information criterion-based work of Blake and Kapetanios (2003b), the Hannan-Quinn criterion with a penalty term of order $\ln \ln T$ seems a more appropriate choice. All of these choices are explored in the Monte Carlo study, in Section 4. Finally, the joint significance of the coefficients of the chosen set of nodes and β are tested via a Wald test in the following regression:

$$y_t = \delta + \beta' x_t + \sum_{i=1}^{k^*} \tilde{\alpha}_i \psi(x_t, x_{\mathcal{S}_i}, \sigma_T) + \epsilon_t. \tag{6}$$

We refer to the Wald test that $\beta = 0$ and $\tilde{\alpha}_1 = \tilde{\alpha}_2 = ... = 0$ as the RBFNN-BOOST test.

Remark 6. In the above discussion we have not addressed the issue of choosing the number of lags of the process y_t to be used in the construction of the neural network nodes. Although the above analysis, and the theory of the next section, allows either for a fixed q or a q that is a growing function of T, it is

https://doi.org/10.1017/S0266466609990612 Published online by Cambridge University Press

straightforward to envisage the possibility of choosing q via any of the criteria discussed above. In this more general case, the analysis would consider two iterations to fully construct the final regression model in (6). First, one would consider all possible values for $q = 1, ..., q_{max}$, and for each q a value of k^* would be chosen. Then the criterion would be used to jointly select a (q, k^*) pair over all permutations.

3. SOME GENERAL THEORETICAL RESULTS

In this section we present some results for the validity of a general neural network specification algorithm that will be useful for our MDH testing problem, but also of independent interest. We consider a regression model of the form (4) for the stochastic process y_t. The aim is to estimate the unknown regression function, $\mu(x_t)$, by an RBFNN series expansion of the form (5). Since we are now interested in estimation rather than testing, as in the MDH testing case, we set $a = 1$ in (4). For the general case we consider in this section, the q-dimensional vector x_t, need not be restricted to be lags of y_t. Algorithm 1 can be used to provide $\hat{\mu}(x_t)$. Next we present our main theoretical result. The following assumptions will be needed:

Assumption 1. $\mathrm{E}|\epsilon_t^s| < \infty$ for some $s > 4$.

Assumption 2. Assume $\mu(.)$ is L_2-bounded, q is either finite or $q = q_T = o\left(\frac{\ln(\log_a T)}{\ln\left(\frac{2\tau_T}{p_T \sigma_T}\right)}\right)$, where a is defined by $m_T \leq \log_a T$.

Assumption 3. Let $\mathcal{F}_t$ be the Borel field generated by $(x_1, \epsilon_0), \ldots, (x_t, \epsilon_{t-1})$. The sequence $\{\epsilon_t\}_{t=-\infty}^{\infty}$ is a stationary martingale difference sequence with $\mathrm{E}(\epsilon_t|\mathcal{F}_t) = 0$, $\mathrm{E}(\epsilon_t^2|\mathcal{F}_t) = \varpi^2(x_{1t})$, where $\varpi(.)$ is continuous and bounded away from zero and infinity and x_{1t} is a subset of x_t of finite dimension q_ϖ. Here x_t and ϵ_t are uncorrelated.

Assumption 4. The sequences of constants σ_T and τ_T tend to zero and infinity, respectively, with T.

Assumption 5. Assume x_t is stationary vector L_2-*NED* (near epoque dependent) process of size -3 on some α mixing process η_{1t} of size $-C$, $C > 1$; $p_T = o(T^{1/4})$.

Assumption 6. Assume x_t and ϵ_t are stationary vector and stationary scalar α-mixing processes with α-mixing coefficients given by $\alpha(k) = C_0 C_2^k$, $C_0 > 0$, $0 < C_2 < 1$; $p_T = O(T^{C_3})$ for some $C_3 > 0$.

Remark 7. Assumption 2 allows for the dimension of x_t to grow with T, which is of interest in MDH testing, as it allows the lag order used to tend to infinity. Note, however, the low allowable rate specified in Appendix B at (B.19). Assumption 3 allows for conditional heteroskedasticity. Note that although the dimension of x_t can grow with T, that of x_{1t} cannot. Assumptions 5 and 6 provide

https://doi.org/10.1017/S0266466609990612 Published online by Cambridge University Press

alternative dependence structures for x_t and ϵ_t. Note the dependence of the rate for p_T on these dependence assumptions. Assumptions 5 is much weaker: first because it does not assume mixing, and second because the mixing process on which x_t depends has α-mixing coefficients, which decline at a polynomial rate rather that the exponential rate of Assumption 6. The stronger dependence Assumption 6, however, allows for a much faster rate of increase in p_T.

Then the following theorem, proved in Appendix B, holds.

THEOREM 1. *Let Assumptions 1–4 and Assumption 5 or Assumption 6 hold. The estimate of the regression function $\mu(x_t)$, obtained using the iterative boosting Algorithm 1 and denoted $\hat{\mu}(x_t)$, satisfies $\hat{\mu}(x_t) - \mu(x_t) = o_p\left(m^{-1/C_1}\right)$, for all $C_1 > 6$, if $m \equiv m_T \leq \log_a T$, for all a that satisfy $\log_a e < \frac{\ln(5/2)}{4}$ if Assumption 5 or Assumption 6 holds. Further, if Assumption 6 holds and*

$$\Pr(|\epsilon_t| > a) \leq C_1 \exp\left(-C_2 a^p\right)$$

for some $C_1 > 0$, $C_2 > 0$, and $p > 1$, then $m \leq \log_a T$, for all a that satisfy $\log_a e < \frac{\ln(5/2)}{2}$.

To the best of our knowledge, this theorem provides the first consistency and rate result for a boosting algorithm in the context of neural networks for dynamic models.

Remark 8. The rate of convergence to the true unknown regression function μ, given in Theorem 1, is rather sharp. Not all logarithmic rates are accommodated. The nature of the logarithmic rates allowed depends crucially on the dependence assumption made about x_t and ϵ_t as well as the tail behavior of ϵ_t.

4. THEORETICAL RESULTS FOR THE MHD TESTS

In this section we present the main theoretical results for the RBFNN MDH test. The following extra assumptions will be needed.

Assumption 7. Assume y_t has a density $f(.)$ that is bounded away from zero and infinity.

Assumption 8. Under MDH, there exists an estimator of ϖ, denoted $\hat{\varpi}$, such that $\sup_x |\hat{\varpi}(x) - \varpi(x)| = O_p(\vartheta_T)$ and some integer $\ell > q_\varpi/2$, and all $(\ell_1, \ldots, \ell_{q_1})$ with $\ell_1 + \cdots + \ell_{q_1} = \ell$, such that $\sup_x \left| \frac{\partial^\ell \hat{\varpi}(x)}{\partial^{\ell_1} x_1 \ldots \partial^{\ell_{q_1}} x} \right| = O_p(\vartheta_T)$, where $\vartheta_T = o(m^{-3q_\varpi/2} (\ln T)^{-1})$ and $\liminf_{T\to\infty} T^{1/2}\vartheta_T > 0$.

Remark 9. Assumption 8 provides the details on the properties of estimator of conditional heteroskedasticity under the null hypothesis. As noted in Section 3.1 of Guay and Guerre (2006), consistency of $\hat{\varpi}$ is not required under the alternative hypothesis, which is important, as the rate required for Assumption 8 may not be obtainable under the alternative using boosting estimation of μ. Estimation of ϖ

https://doi.org/10.1017/S0266466609990612 Published online by Cambridge University Press

may be undertaken using nonparametric variance estimation techniques such as kernel, sieves, or series expansion.

Then the following theorems proved in Appendix A hold.

THEOREM 2. *Let Assumptions 1–4 and 6–8 hold with* $s > 8$ *and* $C_3 < 1/4$, *for Assumptions 1 and 6, respectively. Then, under the MDH, the RBFNN-BOOST test based on the Wald test for the null hypothesis that* $\tilde{\alpha}_1 = \tilde{\alpha}_2 = \cdots = 0$ *in (6) is asymptotically of level* α *when* k^* *is chosen by maximizing* $R_k - \zeta_{T,k}$ *over* k, *and the penalty term,* $\zeta_{T,k}$, *is either that of Guay and Guerre (2006) or that associated with either the Bayesian or the Hannan-Quinn information criterion. The asymptotic distribution is* $\chi^2_{k_{\min}}$.

THEOREM 3. *Let Assumptions 1–4 and 6–7 hold with* $s > 8$ *and* $C_3 < 1/4$, *for Assumptions 1 and 6, respectively. Consider the sequence of alternatives* Δ_T *in* $H_1(C\rho_T, L, s)$, *where*

$$H_1(C\rho_T, L, s) = \left\{ \Delta_T \in \mathcal{C}(L,s), \mathrm{E}\left(\Delta_T(x_t) I(x_t \in \Lambda)^2 \geq \rho_T \right) \right., \tag{7}$$

$$\mathcal{C}(L,s) = \left\{ \Delta(.) : \sup_{x,x' \in \Lambda} \frac{\Delta(x) - \Delta(x')}{||x - x'||^s} \leq L \right\} \quad \text{if } s \in (0,1],$$

and

$$\mathcal{C}(L,s) = \{\Delta(.) : \text{ the } [s]\text{th partial derivatives of } \Delta(.) \text{ are in } \mathcal{C}(L, s - [s])\},$$

for some $\Lambda \in \mathbb{R}^q$, $\rho_T = \log_a T$, *where* a *is defined in Theorem 4 and for some unknown finite* L *and some unknown finite and sufficiently large* s, *where* [.] *denotes integer part. Then, if* $\zeta_{T,k} = k - \sqrt{2\gamma_T k}$ *and* $\gamma_T = O(\ln \ln T)$, *the RBFNN-BOOST test based on the Wald test of the null hypothesis that* $\tilde{\alpha}_1 = \tilde{\alpha}_2 = \cdots = 0$ *in (6) is consistent.*

THEOREM 4. *Let Assumptions 1–4 and 6–7 hold with* $s > 8$ *and* $C_3 < 1/4$, *for Assumptions 1 and 6, respectively. Consider the sequence of alternatives* $r_T \Delta_T$ *in* $\mathcal{C}(L,s)$ *for some unknown* L *and some unknown and sufficiently large* s, *where*

$$\mathrm{E}\left(\frac{\Delta_T(x_t) I(x_t \in \Lambda)}{\varpi(x_t)} \right)^2 \geq 1, \qquad \sup_{x \in \Lambda} |\Delta_T(x_t)| = O(1),$$

for some $\Lambda \in \mathbb{R}^q$. *Then, if* $\zeta_{T,k} = k - \sqrt{2\gamma_T k}$ *and* $\gamma_T = O(\ln \ln T)$, *the RBFNN-BOOST test based on the Wald test of the null hypothesis that* $\tilde{\alpha}_1 = \tilde{\alpha}_2 = \cdots = 0$ *in (6) is consistent, provided that* $r_T = o(T^{-1/2})$.

Remark 10. Theorem 3 relates to alternatives with varying smoothness characteristics as evidenced by the family of functions in (7). Note that for this family of functions the MHD boosting algorithm can only detect alternatives that tend to zero at a logarithmic rate, unlike trigonometric approximations, which can

https://doi.org/10.1017/S0266466609990612 Published online by Cambridge University Press

detect alternatives that tend to zero at a polynomial rate. However, it is worth noting two things: First, the small sample performance of the RBFNN test discussed in the next section suggests that the ability of RBFNN specifications to adapt in a data-dependent fashion, not only in terms of the number of nodes but also in terms of the shape of nodes, gives it a distinct advantage in terms of power performance. Second, the logarithmic rate is only the consequence of the use of boosting. RBFNN specifications have polynomial approximation rates, and if nonlinear estimation of the RBFNN specification was practical, a polynomial rate would be obtained. Theorem 4 relates to smooth alternative hypotheses. In this case, the RBFNN test can achieve a detection rate arbitrarily close to the parametric one.

5. MONTE CARLO STUDY

Having provided a thorough analysis of the theoretical properties of the newly proposed MDH tests, we provide a Monte Carlo study of their small sample properties in this section. Comparability with results of Monte Carlo studies of other MDH tests is very important. Therefore, we follow very closely (and in the case of power experiments, exactly) the Monte Carlo study of Escanciano and Velasco (2006a). As discussed in the previous section, the RBFNN-BOOST test is similar but more computationally intensive than the tests proposed in Blake and Kapetanios (2000, 2003a, 2003b). Further, RBFNN-BOOST is likely to be more powerful than those tests. Since we feel that the premium on computational ease is considerable and since the test of Blake and Kapetanios (2003a) will provide a lower bound in terms of power properties for the power of RBFNN-BOOST, we choose to also use the RBFNN specification of Blake and Kapetanios (2003a) in our Monte Carlo study. In particular, we use the following algorithm for the RBFNN test.

Algorithm 2: (RBF) MDH Algorithm

1. We construct T initial RBF terms given by $\Psi^{(1,\ldots,T)}$, where $\Psi^{(1,\ldots,T)}$ is defined in Algorithm 1.
2. These are ranked according to their ability to reduce the residual variance when entered individually in (6) (i.e., when x_t and only one nonlinear regressor are included in (6)).
3. Penalized likelihood criteria are used to determine how many of the T sorted RBF terms will eventually enter (6).

We then test for the significance of the included hidden units together with the linear part of the specification using a Wald test. We refer to this as the RBFNN test. The question of which penalized likelihood criteria to use is very important. In particular, the choice of the penalty term is very important. The theory of the preceding section has used the criterion suggested by Guay and Guerre (2006) denoted for our purposes as GG, which, as noted before, is of the

https://doi.org/10.1017/S0266466609990612 Published online by Cambridge University Press

form $k + \sqrt{2 \ln \ln Tk}$ and has some theoretical optimality property. However, this property is less appropriate for this analysis, since the order of the approximation has to be logarithmic for the boosting algorithm to be operational, as discussed in the previous section.

On the other hand, Blake and Kapetanios (2000, 2003a, 2003b) have used standard information criteria, which are of the form $\gamma_T k$, where γ_T takes the value 2, $\ln T$, and $2 \ln \ln T$ for the Akaike (AIC), Bayesian (BIC), and Hannan-Quinn (HQ) information criteria, respectively. The AIC is theoretically inappropriate, since its associated penalty term is too small, and so the resulting test does not have an asymptotic χ^2 distribution. It is straightforward to see that the other two criteria have an asymptotic $\chi^2_{k_{\min}}$ distribution, as shown in Theorem 1. In fact, their penalty terms are too parsimonious, with the one associated with the BIC being the most parsimonious. However, the work of Blake and Kapetanios (2000, 2003a, 2003b) suggests that the RFB approximations are efficient, and so few approximation terms are sufficient for obtaining powerful tests. Hence, the BIC was found to be best, since it minimized overrejection under the null to perfectly acceptable levels. To investigate all these issues, we consider all four penalty terms (AIC, BIC, HQ, and GG) in our Monte Carlo study.

We also wish to consider tests based on polynomial approximations and on logistic neural networks along the lines of Teräsvirta, Lin, and Granger (1993) and Lee, White, and Granger (1993). Both approximation classes were considered by Blake and Kapetanios (2000, 2003a, 2003b) as well and were found to be reasonable alternatives to RBFNN approximations. In the case of the polynomial approximations, we again use a data-dependent method for determining the order of the approximation using the penalized likelihood criteria discussed above. Approximations of orders 2, 3, and 4 are considered. For the logistic neural network approximations we use the approach of Lee et al. (1993) and simply augment the set of regressors whose significance we test with the linear part of the regression. Thus, this test becomes one for the MDH rather than of neglected nonlinearity as in Lee et al. Their approach suggests that the specification of each neural network node is given by $\phi(\gamma' x_t)$, where $\phi(\lambda)$ is the logistic function $\{1 + \exp(-\lambda)\}^{-1}$. This is a monotonic function, with output bounded between 0 and 1. The coefficients γ_j are randomly generated from a uniform distribution over (γ_l, γ_h). For given k^*, the constructed regressors $\phi(\gamma_j' x_t)$, $j = 1, \ldots, k^*$ may suffer from multicollinearity. Lee et al. suggest that $\tilde{k}^*$ largest principle components of the constructed regressors excluding the largest one be used as extra regressors. We set $\gamma_l = -2$, $\gamma_h = 2$, $k^* = 10$, and $\tilde{k}^* = 2$ as in the original paper. Recent work by White (2006) suggests that a similar boosting approach can be used with the logistic neural network specifications. However, we note here that the construction of the parameter grid is much less intuitive than that for the RBFNN. The tests based on the approaches of Teräsvirta et al. (1993) and Lee et al. are denoted by TLG and LWG, respectively.

A final point to note is that while the theoretical part of the paper suggests the use of nonparametric methods to correct for the possible presence of conditional

https://doi.org/10.1017/S0266466609990612 Published online by Cambridge University Press

heteroskedasticity, we chose not to consider such corrections in the Monte Carlo, following the approach adopted in the Monte Carlo study of Guay and Guerre (2006). There are a number of reasons for this choice. The first is that nonparametric methods for conditional variances are not well behaved in many cases. Initial experimentation for our test suggests that this may be the case here. Further, considerable work is needed to conclude which nonparametric method is appropriate, in small samples, for modeling conditional variances, when some sieve approach is used to model the conditional mean. This is a general issue, beyond the scope of the current paper, which focuses on the theoretical justification of RBFs as a powerful basis for sieve approximations used in MDH testing. Additionally, it is of interest to explore how problematic the behavior of the tests is under the null hypothesis when conditional heteroskedasticity occurs but is not corrected for.

5.1. Experiment Design

Following the Monte Carlo study of Escanciano and Velasco (2006a), we consider the following experiments. The first three experiments involve processes that satisfy the MDH and therefore provide information about the size properties of the tests, whereas the rest of the experiments involve processes that are not martingale difference sequences and therefore provide information about the power properties of the tests. All experiments are exactly the same as in Escanciano and Velasco (2006a) so as to enable valid comparisons with their power results. We have not considered the long memory power experiment of Escanciano and Velasco (2006a), since it is an experiment involving a linear model, and by the AR representation of long memory processes (see, e.g., Beran, 1994; Poskitt, 2005) it is obvious that our methods will work extremely well. Throughout the experiment description, $\epsilon_t, u_t \sim NID(0,1)$.

Experiments.

1. $y_t = \epsilon_t$, (IID)
2. $y_t = \epsilon_t \sigma_t$, with $\sigma_t^2 = 0.001 + 0.01 y_{t-1}^2 + 0.97\sigma_{t-1}^2$, (GARCH1)
3. $y_t = \epsilon_t \sigma_t$, with $\sigma_t^2 = 0.001 + 0.09 y_{t-1}^2 + 0.89\sigma_{t-1}^2$, (GARCH2)
4. $y_t = \epsilon_t \sigma_t$, with $\sigma_t^2 = 0.001 + 0.09 y_{t-1}^2 + 0.90\sigma_{t-1}^2$, (GARCH3)
5. $y_t = \epsilon_t exp(\sigma_t)$, with $\sigma_t = 0.932\sigma_{t-1} + 0.32u_t$, (SV)
6. $y_t = \epsilon_{t-1}\epsilon_{t-2}(\epsilon_{t-2} + \epsilon_t + 1)$, (NLMA)
7. $y_t = \epsilon_t + 0.15\epsilon_{t-1}y_{t-1} + 0.05\epsilon_{t-1}y_{t-2}$, (BILIN1)
8. $y_t = \epsilon_t + 0.25\epsilon_{t-1}y_{t-1} + 0.15\epsilon_{t-1}y_{t-2}$, (BILIN2)
9. $y_t = \epsilon_t + x_t - x_{t-1}$, $x_t = 0.85x_{t-1} + u_t$, (NDAR)
10. $y_t = -0.5y_{t-1}I(y_{t-1} \geq 1) + 0.4y_{t-1}I(y_{t-1} < 1) + \epsilon_t$, (SETAR)
11. $y_t = 0.6y_{t-1}exp(-0.5y_{t-1}^2) + \epsilon_t$, (EXP).

It is worth observing that, as noted by Escanciano and Velasco (2006a), the DGPs underlying Experiments 3 and 4 generate processes that do not possess eighth and sixth moments, respectively. Since we assume finite eighth moments

https://doi.org/10.1017/S0266466609990612 Published online by Cambridge University Press

in the theoretical part of the paper, some problematic behavior under the null hypothesis can be expected. Therefore we consider these experiments to gauge the extent of the problem caused by the lack of higher moment for these processes. All tests use the first lag of the process to construct the RBFNN nodes. The first lag is also used for the TLG and LWG tests.[3] Rejection probabilities for a nominal significance level of 95%, produced using 1,000 replications, are reported in Tables 1–4.

5.2. Size Results

Looking at the performance of the RBFNN and RBFNN-BOOST tests under the null hypothesis, it is clear that depending on the penalty term used, there is some overrejection. This is most severe for the AIC, followed by GG and HQ, as expected given their relative parsimony. The BIC performs best in this respect with less noticeable overrejection. Similar patterns occur for the TLG and LWG tests. The results accord with our experience in related applications, where searching for any form of significant neglected structure tends to induce overrejection. The reasonably good performance of the BIC in this respect removes any need to resort to bootstrap size correction. Finally, it is worth noting that all tests overreject more for Experiments 3 and 4, which, as noted earlier, do not satisfy the moment conditions needed for the theoretical analysis of the paper. On the other hand, for the case of GARCH 1 where the moment conditions definitely hold but the GARCH part of the process is not modeled, the performance of the tests is reasonable and clearly improves with sample size. In order to compare fairly the power properties of the new tests with existing tests, we report size-corrected rejection probabilities under the alternative hypothesis, as well as rejection probabilities based on nominal critical values. For the size correction we use the worse performing model under the null hypothesis, which in most cases is Experiment 4, even though this may be viewed as unfair to our tests, since Experiment 4 does not satisfy the required moment conditions.

5.3. Power Results

The tests based on BIC have slightly lower power, both when using nominal critical values and size-corrected ones. Since these are the tests that overreject least and the loss of power compared to the other tests is minimal, these tests seem preferable. In terms of relative power performance, tests have more power against the SETAR alternative, followed by the BILIN1, NLMA, EXP, BILIN2, and NDAR alternatives. The NDAR alternative seems to be extremely difficult to detect. This is corroborated by the Monte Carlo results of Escanciano and Velasco (2006a). The RBFNN test seems to be more powerful that the TLG and LWG tests for all experiments considered. The RBFNN-BOOST test appears only slightly more powerful than the RBFNN test, and that superiority only appears in some experiments. We feel that use of the RBFNN-BOOST may not be necessary given the extra computational cost.

https://doi.org/10.1017/S0266466609990612 Published online by Cambridge University Press

https://doi.org/10.1017/S0266466609990612 Published online by Cambridge University Press

TABLE 1. Results for RBFNN-BOOST test (nominal level: 5%)

	AIC			BIC			HQ			GG		
Exp/T	100	200	300	100	200	300	100	200	300	100	200	300
Rejection probabilities under the null												
IID	0.086	0.085	0.087	0.061	0.050	0.054	0.076	0.069	0.075	0.085	0.084	0.087
GARCH1	0.156	0.139	0.107	0.086	0.078	0.054	0.134	0.123	0.092	0.156	0.133	0.105
GARCH2	0.134	0.126	0.123	0.080	0.072	0.068	0.117	0.108	0.108	0.129	0.123	0.120
GARCH3	0.147	0.144	0.144	0.092	0.084	0.077	0.125	0.121	0.120	0.144	0.141	0.142
SV	0.111	0.100	0.106	0.075	0.069	0.065	0.094	0.086	0.084	0.107	0.098	0.102
Rejection probabilities under the alternative												
BILIN1	0.518	0.806	0.925	0.471	0.776	0.911	0.496	0.789	0.916	0.512	0.804	0.924
BILIN2	0.861	0.991	1.000	0.836	0.991	1.000	0.848	0.991	1.000	0.857	0.991	1.000
NDAR	0.140	0.133	0.142	0.107	0.092	0.104	0.125	0.117	0.129	0.137	0.131	0.141
NLMA	0.626	0.737	0.780	0.530	0.622	0.682	0.601	0.704	0.750	0.624	0.731	0.778
SETAR	0.878	0.996	1.000	0.835	0.995	1.000	0.864	0.996	1.000	0.873	0.996	1.000
EXP	0.622	0.898	0.979	0.540	0.849	0.947	0.595	0.872	0.965	0.619	0.892	0.975
Size-corrected rejection probabilities												
BILIN1	0.205	0.468	0.655	0.314	0.660	0.856	0.218	0.480	0.704	0.199	0.475	0.656
BILIN2	0.572	0.919	0.990	0.694	0.972	1.000	0.583	0.925	0.994	0.560	0.924	0.990
NDAR	0.035	0.033	0.025	0.050	0.052	0.055	0.037	0.026	0.032	0.035	0.036	0.026
NLMA	0.483	0.589	0.657	0.467	0.568	0.630	0.478	0.569	0.647	0.475	0.595	0.657
SETAR	0.614	0.948	0.996	0.703	0.981	0.998	0.619	0.942	0.996	0.599	0.948	0.996
EXP	0.336	0.706	0.870	0.406	0.762	0.912	0.344	0.667	0.867	0.327	0.706	0.872

https://doi.org/10.1017/S0266466609990612 Published online by Cambridge University Press

TABLE 2. Results for RBFNN test (nominal level: 5%)

	AIC			BIC			HQ			GG		
Exp/T	100	200	300	100	200	300	100	200	300	100	200	300
Rejection probabilities under the null												
IID	0.081	0.074	0.076	0.058	0.050	0.058	0.072	0.065	0.069	0.079	0.072	0.075
GARCH1	0.141	0.120	0.101	0.092	0.080	0.064	0.128	0.110	0.091	0.139	0.118	0.098
GARCH2	0.122	0.117	0.112	0.086	0.076	0.074	0.112	0.111	0.102	0.122	0.114	0.111
GARCH3	0.126	0.121	0.140	0.088	0.072	0.075	0.113	0.108	0.124	0.122	0.119	0.135
SV	0.097	0.089	0.084	0.076	0.066	0.065	0.089	0.083	0.078	0.095	0.087	0.084
Rejection probabilities under the alternative												
BILIN1	0.499	0.792	0.921	0.470	0.765	0.912	0.488	0.782	0.915	0.498	0.788	0.921
BILIN2	0.862	0.991	1.000	0.845	0.991	1.000	0.853	0.991	1.000	0.859	0.991	1.000
NDAR	0.128	0.120	0.131	0.103	0.096	0.103	0.120	0.111	0.122	0.127	0.116	0.130
NLMA	0.588	0.716	0.761	0.523	0.612	0.671	0.570	0.685	0.729	0.586	0.711	0.758
SETAR	0.851	0.997	1.000	0.808	0.992	0.999	0.839	0.995	1.000	0.849	0.996	1.000
EXP	0.611	0.883	0.977	0.547	0.842	0.947	0.592	0.872	0.962	0.604	0.881	0.975
Size-corrected rejection probabilities												
BILIN1	0.239	0.531	0.715	0.312	0.680	0.852	0.234	0.570	0.746	0.240	0.526	0.721
BILIN2	0.615	0.939	0.994	0.703	0.975	0.999	0.616	0.950	0.995	0.619	0.936	0.994
NDAR	0.036	0.030	0.030	0.047	0.058	0.053	0.034	0.035	0.034	0.035	0.029	0.032
NLMA	0.479	0.587	0.647	0.453	0.565	0.626	0.462	0.585	0.641	0.482	0.583	0.647
SETAR	0.660	0.958	0.996	0.697	0.975	0.998	0.646	0.968	0.997	0.660	0.957	0.997
EXP	0.363	0.718	0.894	0.414	0.774	0.912	0.371	0.727	0.883	0.373	0.711	0.890

https://doi.org/10.1017/S0266466609990612 Published online by Cambridge University Press

TABLE 3. Results for TLG test (nominal level: 5%)

	AIC			BIC			HQ			GG		
Exp/T	100	200	300	100	200	300	100	200	300	100	200	300
Rejection probabilities under the null												
IID	0.063	0.063	0.070	0.053	0.041	0.048	0.060	0.057	0.062	0.063	0.063	0.070
GARCH1	0.110	0.114	0.086	0.089	0.083	0.064	0.101	0.109	0.082	0.110	0.114	0.086
GARCH2	0.092	0.108	0.105	0.080	0.074	0.080	0.091	0.098	0.098	0.092	0.108	0.105
GARCH3	0.102	0.113	0.125	0.086	0.084	0.092	0.098	0.104	0.117	0.104	0.113	0.125
SV	0.075	0.077	0.080	0.063	0.058	0.065	0.074	0.069	0.073	0.075	0.077	0.080
Rejection probabilities under the alternative												
BILIN1	0.472	0.782	0.920	0.449	0.760	0.910	0.464	0.774	0.916	0.473	0.782	0.920
BILIN2	0.838	0.990	1.000	0.822	0.990	1.000	0.827	0.990	1.000	0.838	0.990	1.000
NDAR	0.104	0.108	0.123	0.092	0.087	0.095	0.100	0.103	0.114	0.104	0.108	0.123
NLMA	0.540	0.672	0.720	0.500	0.603	0.651	0.531	0.654	0.700	0.540	0.672	0.720
SETAR	0.832	0.996	1.000	0.792	0.989	0.999	0.822	0.993	0.999	0.833	0.996	1.000
EXP	0.561	0.875	0.968	0.501	0.834	0.941	0.550	0.861	0.960	0.562	0.875	0.968
Size-corrected rejection probabilities												
BILIN1	0.341	0.683	0.866	0.333	0.695	0.860	0.334	0.690	0.861	0.340	0.682	0.869
BILIN2	0.744	0.972	1.000	0.741	0.977	1.000	0.742	0.975	1.000	0.742	0.972	1.000
NDAR	0.046	0.051	0.049	0.039	0.054	0.046	0.041	0.052	0.049	0.046	0.051	0.054
NLMA	0.418	0.526	0.601	0.416	0.533	0.591	0.420	0.538	0.598	0.417	0.525	0.602
SETAR	0.707	0.971	0.998	0.708	0.973	0.997	0.716	0.974	0.998	0.703	0.969	0.998
EXP	0.401	0.739	0.901	0.396	0.748	0.901	0.399	0.751	0.908	0.398	0.738	0.904

TABLE 4. Results for LWG test (nominal level: 5%)

Exp/T	100	200	300
Rejection probabilities under the null			
IID	0.092	0.090	0.094
GARCH1	0.146	0.129	0.117
GARCH2	0.118	0.127	0.144
GARCH3	0.134	0.134	0.163
SV	0.104	0.103	0.113
Rejection probabilities under the alternative			
BILIN1	0.510	0.824	0.927
BILIN2	0.868	0.993	1.000
NDAR	0.135	0.148	0.161
NLMA	0.544	0.649	0.703
SETAR	0.867	0.997	1.000
EXP	0.648	0.920	0.983
Size-corrected rejection probabilities			
BILIN1	0.323	0.641	0.823
BILIN2	0.718	0.973	0.997
NDAR	0.049	0.043	0.047
NLMA	0.422	0.502	0.562
SETAR	0.735	0.977	0.999
EXP	0.449	0.786	0.933

5.4. Comparison with Escanciano and Velasco (2006a, 2006b)

We chose our experiments to facilitate comparison of the rejection probabilities under the alternative of our new tests with those proposed in Escanciano and Velasco (2006a). Since we use exactly the same experimental design, a comparison is possible without replicating their results. Note that the bootstrap $\mathbf{D}_n^2$ test of Escanciano and Velasco (2006a) has more power than all the alternative tests considered in that paper. The new tests we propose compare favorably with that test, both when using nominal critical values and size-corrected ones, with a single exception.

The one exception relates to experiment EXP. However, it is only to be expected as the data generation process for that experiment has an exponential structure. The $\mathbf{D}_n^2$ uses the exponential function and considers the covariance between y_t and $\exp(y_{t-j})$, $j > 0$. Hence it is essentially a parametric test for this sort of deviation from the MDH. Overall, it is pretty clear that the newly proposed tests have comparable small sample power performance to $\mathbf{D}_n^2$ and (by the Monte Carlo results of Escanciano and Velasco (2006a)) to many other MDH tests.

Finally, note that three of the experiments of Escanciano and Velasco (2006b) we consider (NLMA, BILIN1, BILIN2) are also used by Escanciano and Velasco (2006b) for an alternative MDH test, which is similar to that of Escanciano and Velasco (2006a) but uses the indicator function rather than the

https://doi.org/10.1017/S0266466609990612 Published online by Cambridge University Press

exponential one. From the Monte Carlo study results of Escanciano and Velasco (2006b) we see that the RBFNN test has comparable power properties to those of the test of Escanciano and Velasco (2006b).

5.5. The Benefits of the RBF Approach

As a final comment, it is worth noting that in small samples the nature and shape of the nodes that form a series approximation are more important than the number of nodes. We offer two pieces of evidence to support this. First, we note that the power performance depends only slightly on the choice of the penalty terms in the penalized likelihood criteria. Adding additional nodes does not improve the performance, and a more parsimonious test can safely be used. Second, as a check we considered (but have not reported) a trigonometric approximation of the form considered in Guay and Guerre (2006). However, the power performance of the tests based on this approximation was significantly inferior to that of the RBFNN based test.

6. EMPIRICAL APPLICATION TO STOCK RETURNS

In this section we provide an empirical application that illustrates the potential of the new test to evaluate the martingale difference hypothesis. It is widely thought that stock returns are 'close' to unpredictable, and empirical analysis to predict future returns based on past return data has little or no explanatory power (see, for example, Cochrane, 2005, Ch. 20). We apply our test to stock returns to see if this can be supported. As it is sometimes difficult to draw meaningful conclusions from the empirical analysis of a single series for the performance of a new statistical test, we consider the large S&P 500 data set, which allows us to draw wider conclusions based on the proportion of the series that reject.

Weekly returns data were obtained from Datastream spanning the period 01/01/1993–20/01/2004 and comprising 575 weekly observations. We consider only companies for which data are available throughout the period, a total of 412 series. We normalize the returns series to have a mean equal to zero and variance equal to one prior to testing. We report the probability values for the test of the martingale difference hypothesis, carried out on the 412 company return series in Tables 5–8. Probability values below 0.05 and the company names to which they correspond are reported in bold typescript for easy identification.

As we can see, for these tables a large minority of the series (165 stock returns in total) are in fact found to reject the martingale difference null hypothesis at the 95% significance level. This is almost exactly 40% of the series tested, far higher than what we would expect if returns were indeed unpredictable. While the use of macroeconomic factors can improve forecastability (see, for example, Lettau and Ludvigson, 2001), our analysis suggests that more general nonlinear specifications also may have more forecasting power than traditional linear specifications.

https://doi.org/10.1017/S0266466609990612 Published online by Cambridge University Press

TABLE 5. Probability values for S&P 500 series (ABBOTT LABS.– COMPUTER SCIS)

Company name	P-value	Company name	P-value
ABBOTT LABS.	0.079	ADC TELECOM.	0.232
ADOBE SYS.	0.650	ADVD. MICRO DEVC.	0.535
AES (1)	**0.000**	**AFLAC (2)**	**0.000**
AIR PRDS.& CHEMS.	0.228	ALBERTO CULVER 'B'	0.495
ALBERTSONS (3)	**0.044**	ALCOA	0.760
ALLEGHENY EN. (4)	**0.000**	**ALLEGHENY TECHS. (5)**	**0.008**
ALLERGAN	0.214	**ALLIED WASTE INDS. (6)**	**0.030**
ALLTEL	0.277	**ALTERA (7)**	**0.027**
ALTRIA GP.	0.469	**AMBAC FINANCIAL (8)**	**0.003**
AMERADA HESS	0.777	AMER.ELEC.PWR.	0.594
AMERICAN EXPRESS (9)	**0.000**	AMER. GREETINGS 'A'	0.308
AMERICAN INTL. GP. (10)	**0.000**	AMER. POWER CONV.	0.199
AMGEN	0.688	AMSOUTH BANC.	0.070
ANADARKO PETROLEUM	0.234	**ANALOG DEVICES (11)**	**0.023**
ANDREW	0.227	**ANHEUSER - BUSCH COS. (12)**	**0.000**
AON (13)	**0.010**	**APACHE (14)**	**0.017**
APPLE COMPUTERS (15)	**0.032**	**APPLERA APPD.BIOS. (16)**	**0.000**
APPLIED MATS.	0.255	ARCHER - DANLS.	0.068
ASHLAND	0.358	AT&T	0.799
AUTODESK	0.655	**AUTOMATIC DATA PROC. (17)**	**0.000**
AUTONATION (18)	**0.001**	AUTOZONE	0.522
AVERY DENNISON (19)	**0.005**	AVON PRODUCTS	0.698
BAKER HUGHES (20)	**0.010**	BALL	0.305
BANK OF AMERICA	0.144	**BANK OF NEW YORK (21)**	**0.011**
BANK ONE	0.145	BARD C R	0.058
BAUSCH & LOMB	0.258	**BAXTER INTL. (22)**	**0.017**
BB & T	0.390	**BEAR STEARNS (23)**	**0.001**
BECTON DICKINSON & CO.	0.097	**BED BATH & BEYOND (24)**	**0.000**
BELLSOUTH (25)	**0.026**	**BEMIS (26)**	**0.011**
BEST BUY CO.	0.970	BIG LOTS	0.403
BIOGEN IDEC	0.115	**BIOMET (27)**	**0.007**
BJ SVS. (28)	**0.008**	BLACK & DECKER	0.325
H & R BLOCK	0.959	BMC SOFTWARE	0.266
BOEING	0.108	BOISE CASCADE	0.583
BOSTON SCIENTIFIC	0.343	**BRISTOL MYERS SQUIBB (29)**	**0.000**
BROWN - FORMAN 'B'	0.535	**BRUNSWICK (30)**	**0.002**
BURL.NTHN.SANTA FE C (31)	**0.002**	**BURLINGTON RES. (32)**	**0.002**
CAMPBELL SOUP	0.133	CARDINAL HEALTH	0.189
CARNIVAL	0.319	CATERPILLAR	0.119
CENDANT	0.569	**CENTERPOINT EN. (33)**	**0.000**
CENTEX	0.052	CENTURYTEL	0.152
CHARLES SCHWAB (34)	**0.019**	**CHARTER ONE FINL. (35)**	**0.000**
CHEVRONTEXACO	0.054	CHIRON CORP	0.861
CHUBB (36)	**0.011**	CIGNA	0.728
CINCINNATI FIN. (37)	**0.000**	CINTAS	0.070
CIRCUIT CITY STORES	0.334	CISCO SYSTEMS	0.143
CITIGROUP (38)	**0.019**	**CITIZENS COMMS. (39)**	**0.000**
CLEAR CHL. COMMS.	0.889	**CLOROX (40)**	**0.000**
CMS ENERGY	0.096	COCA COLA	0.224
COCA COLA ENTS.	0.239	**COLGATE - PALM. (41)**	**0.003**
COMCAST 'A' (42)	**0.039**	COMERICA	0.054
COMPUTER ASSOCS. INTL.	0.158	COMPUTER SCIS.	0.902

https://doi.org/10.1017/S0266466609990612 Published online by Cambridge University Press

TABLE 6. Probability values for S&P 500 series (COMPUWARE – ITT INDUSTRIES)

Company Name	P-value	Company Name	P-value
COMPUWARE	0.596	COMVERSE TECH.	0.258
CONAGRA	0.327	**CONCORD EFS (43)**	**0.042**
CONOCOPHILLIPS (44)	**0.021**	CONS. EDISON	0.495
CONSTELLATION EN.	0.565	**COOPER INDS. (45)**	**0.001**
COOPER TIRE RUB.	0.137	**ADOLPH COORS 'B' (46)**	**0.024**
CORNING	0.152	COUNTRYWIDE FINL.	0.554
CRANE	0.347	**CSX (47)**	**0.006**
CUMMINS	0.766	CVS	0.529
DANA (48)	**0.025**	DANAHER	0.722
DEERE & CO.	0.561	DELL	0.793
DELTA AIR LINES (49)	**0.000**	DELUXE	0.276
DILLARDS 'A'	0.414	DOLLAR GENERAL	0.415
DOMINION RES.	0.103	DONNELLEY R R	0.876
DOVER	0.687	DOW CHEMICALS	0.329
DOW JONES & CO.	0.059	**DTE ENERGY (50)**	**0.014**
DU PONT E I DE NEMOURS (51)	**0.000**	DUKE ENERGY	0.170
DYNEGY 'A' (52)	**0.000**	EASTMAN KODAK	0.667
EATON	0.366	**ECOLAB (53)**	**0.000**
EDISON INTL. (54)	**0.000**	**EL PASO (55)**	**0.000**
ELECTRONIC ARTS (56)	**0.011**	**ELECTRONIC DATA SYSTEMS(57)**	**0.000**
EMC (58)	**0.002**	**EMERSON ELECTRIC (59)**	**0.002**
ENGELHARD (60)	**0.001**	ENTERGY	0.882
EOG RES.	0.278	EQUIFAX	0.455
EXELON	0.865	**EXPRESS SCRIPTS 'A' (61)**	**0.023**
EXXON MOBIL (62)	**0.000**	FAMILY $ STRS.	0.144
FANNIE MAE (63)	**0.000**	**FREDDIE MAC (64)**	**0.006**
FEDERATED DEPT.STRS.	0.428	FEDEX	0.333
FIFTH THIRD BANCORP	0.159	**FIRST DATA (65)**	**0.008**
FIRST TEN. NAT. (66)	**0.043**	FIRSTENERGY	0.487
FISERV (67)	**0.030**	**FLEETBOSTON FINL. (68)**	**0.000**
FORD MOTOR	0.393	FOREST LABS.	0.154
FORTUNE BRANDS	0.700	FPL GROUP	0.835
FRANK. RES.	0.209	GANNETT	0.079
GAP	0.368	GEN. DYNAMICS	0.346
GENERAL ELECTRIC (69)	**0.000**	GEN. MILLS	0.547
GENERAL MOTORS	0.840	GENUINE PARTS	0.065
GENZYME	0.213	GEORGIA PACIFIC	0.619
GILLETTE	0.341	**GOLDEN WEST FINL. (70)**	**0.039**
GOODRICH	0.063	GOODYEAR TIRE	0.384
GRAINGER W W	0.229	GT. LAKES CHM.	0.313
HALLIBURTON (71)	**0.000**	**HARLEY - DAVIDSON (72)**	**0.035**
HARRAHS ENTM.	0.675	**HASBRO (73)**	**0.001**
HCA (74)	**0.003**	**HEALTH MAN. AS. A (75)**	**0.034**
HEINZ HJ (76)	**0.019**	HERCULES	0.052
HERSHEY FOODS (77)	**0.004**	HEWLETT - PACKARD	0.268
HILTON HOTELS	0.940	**HOME DEPOT (78)**	**0.007**
HONEYWELL INTL.	0.920	**HUMANA (79)**	**0.007**
HUNTINGTON BCSH.	0.065	**ILLINOIS TOOL WKS. (80)**	**0.009**
INGERSOLL - RAND	0.641	INTEL	0.829
INTL. BUS. MACH.	0.174	INTL. FLAV. & FRAG.	0.681
INTL. GAME TECH.	0.218	INTL. PAPER	0.401
INTERPUBLIC GP.	0.099	ITT INDUSTRIES	0.182

https://doi.org/10.1017/S0266466609990612 Published online by Cambridge University Press

TABLE 7. Probability values for S&P 500 series (JP MORGAN CHASE – PULTE HOMES)

Company name	P-value	Company name	P-value
JP MORGAN CHASE & CO.	0.471	JEFFERSON PILOT	0.365
JOHNSON & JOHNSON (81)	**0.000**	JOHNSON CONTROLS	0.056
JONES APPAREL GROUP	0.867	**KB HOME (82)**	**0.017**
KELLOGG	0.318	KERR - MCGEE	0.060
KEYCORP	0.243	KEYSPAN	0.779
KIMBERLY - CLARK (83)	**0.000**	**KINDER MORGAN KANS (84)**	**0.000**
KLA TENCOR (85)	**0.001**	KNIGHT - RIDDER	0.374
KOHLS (86)	**0.021**	**KROGER (87)**	**0.027**
LEGGETT & PLATT	0.134	LILLY ELI	0.584
LIMITED BRANDS	0.489	LINCOLN NAT.	0.686
LINEAR TECH. (88)	**0.025**	LIZ CLAIBORNE	0.689
LOEWS (89)	**0.042**	LNA. PACIFIC	0.500
LOWE'S COMPANIES	0.138	LSI LOGIC	0.325
MANOR CARE	0.225	**MARATHON OIL (90)**	**0.029**
MARSH & MCLENNAN (91)	**0.000**	MARSHALL & ILSLEY	0.785
MASCO (92)	**0.043**	**MATTEL (93)**	**0.000**
MAXIM INTEGRATED PRDS. (94)	**0.000**	MAY DEPT. STORES	0.413
MAYTAG	0.188	**MBIA (95)**	**0.000**
MBNA (96)	**0.000**	**MCCORMICK & CO. NV. (97)**	**0.000**
MCDONALDS	0.563	**MCGRAW - HILL CO. (98)**	**0.000**
MEADWESTVACO	0.171	MEDIMMUNE	0.541
MEDTRONIC (99)	**0.001**	MELLON FINL.	0.226
MERCK & CO.	0.321	**MEREDITH (100)**	**0.044**
MERRILL LYNCH & CO.	0.227	MGIC INVT	0.279
MICRON TECH.	0.813	MICROSOFT	0.201
MILLIPORE (101)	**0.032**	**MOLEX (102)**	**0.003**
MOTOROLA (103)	**0.006**	**NABORS INDS. (104)**	**0.011**
NAT. CITY	0.203	NATIONAL SEMICON.	0.368
NAVISTAR INTL.	0.853	NEW YORK TIMES 'A'	0.678
NEWELL RUBBERMAID (105)	**0.011**	NEWMONT MINING	0.602
NEXTEL COMMS. A	0.188	**NICOR (106)**	**0.000**
NIKE 'B' (107)	**0.010**	**NISOURCE (108)**	**0.040**
NOBLE (109)	**0.000**	NORDSTROM	0.612
NORFOLK SOUTHERN	0.203	**NORTH FORK BANCORP. (110)**	**0.018**
NTHN. TRUST (111)	**0.001**	NORTHROP GRUMMAN	0.525
NOVELL	0.836	**NOVELLUS SYSTEMS (112)**	**0.001**
NUCOR	0.230	OCCIDENTAL PTL.	0.721
OFFICE DEPOT	0.091	**OMNICOM GP. (113)**	**0.036**
ORACLE	0.177	**PACCAR (114)**	**0.027**
PALL	0.653	**PARAMETRIC TECH. (115)**	**0.009**
PARKER - HANNIFIN	0.091	**PAYCHEX (116)**	**0.014**
PENNEY JC (117)	**0.007**	PEOPLES ENERGY	0.856
PEOPLESOFT (118)	**0.021**	**PEPSICO (119)**	**0.032**
PERKINELMER (120)	**0.003**	**PFIZER (121)**	**0.046**
PG & E (122)	**0.038**	**PHELPS DODGE (123)**	**0.023**
PINNACLE WEST CAP.	0.304	**PITNEY - BOWES (124)**	**0.032**
PLUM CREEK TIMBER (125)	**0.001**	PMC - SIERRA	0.150
PNC FINL. SVS. GP.	0.063	PPG INDUSTRIES	0.126
PPL	0.245	**PRAXAIR (126)**	**0.044**
PROCTER & GAMBLE (127)	**0.003**	PROGRESS EN.	0.784
PROGRESSIVE OHIO	0.086	**PROVIDIAN FINL. (128)**	**0.001**
PUB. SER. ENTER. GP.	0.245	PULTE HOMES	0.238

https://doi.org/10.1017/S0266466609990612 Published online by Cambridge University Press

TABLE 8. Probability values for S&P 500 series (QUALCOMM – 3M)

Company name	P-value	Company name	P-value
QUALCOMM	0.133	RADIOSHACK	0.092
RAYTHEON 'B'	0.491	REEBOK INTL.	0.424
REGIONS FINL. (129)	**0.023**	ROBERT HALF INTL.	0.443
ROCKWELL AUTOMATION (130)	**0.007**	ROHM & HAAS	0.183
ROWAN COS.	0.174	**RYDER SYSTEM (131)**	**0.033**
SAFECO	0.871	SAFEWAY	0.642
SARA LEE	0.802	**SBC COMMUNICATIONS (132)**	**0.013**
SCHERING - PLOUGH (133)	**0.004**	**SCHLUMBERGER (134)**	**0.000**
SCIENTIFIC ATLANTA	0.349	SEALED AIR	0.154
SEARS ROEBUCK & .CO.	0.838	SEMPRA EN.	0.076
SHERWIN - WILLIAMS	0.244	SIGMA ALDRICH	0.490
SLM	0.314	SNAP - ON	0.825
SOLECTRON	0.665	SOUTHERN	0.219
SOUTHTRUST	0.253	SOUTHWEST AIRLINES	0.297
SPRINT (135)	**0.000**	**ST. JUDE MED. (136)**	**0.000**
ST.PAUL	0.343	STANLEY WORKS	0.270
STAPLES (137)	**0.014**	STARBUCKS	0.056
STARWOOD HTLS. & RESORTS	0.059	STATE STREET	0.333
STRYKER	0.092	SUN MICROSYSTEMS	0.533
SUNGARD DATA SYSTEMS (138)	**0.019**	SUNOCO	0.712
SUNTRUST BANKS	0.385	SUPERVALU	0.488
SYMANTEC (139)	**0.023**	SYMBOL TECHS.	0.980
SYNOVUS FINL. (140)	**0.004**	**SYSCO (141)**	**0.000**
T ROWE PRICE GP. (142)	**0.001**	**TARGET (143)**	**0.004**
TECO ENERGY (144)	**0.000**	**TEKTRONIX (145)**	**0.001**
TELLABS	0.177	TEMPLE INLAND	0.307
TENET HLTHCR. (146)	**0.000**	TERADYNE	0.059
TEXAS INSTS. (147)	**0.003**	**TEXTRON (148)**	**0.001**
THERMO ELECTRON	0.145	THOMAS & BETTS	0.063
TIFFANY & CO	0.936	TIME WARNER	0.855
TJX COS.	0.545	TORCHMARK	0.337
TOYS R US HOLDINGS CO.	0.676	TRIBUNE	0.942
TXU (149)	**0.000**	**TYCO INTL. (150)**	**0.000**
US BANCORP	0.220	**UNION PACIFIC (151)**	**0.013**
UNION PLANTERS (152)	**0.017**	UNISYS	0.988
UNITEDHEALTH GP. (153)	**0.037**	US. STEEL	0.485
UNITED TECHNOLOGIES (154)	**0.006**	**UNOCAL (155)**	**0.007**
UNUMPROVIDENT	0.397	UST	0.375
V F	0.534	**VERIZON COMMS. (156)**	**0.004**
VIACOM 'B' (157)	**0.021**	**VULCAN MATERIALS (158)**	**0.003**
WACHOVIA	0.079	**WALGREEN (159)**	**0.001**
WAL MART STORES (160)	**0.001**	WALT DISNEY	0.491
WASHINGTON MUTUAL	0.212	WASTE MAN.	0.076
WELLS FARGO & CO. (161)	**0.035**	WENDY'S INTL.	0.120
WEYERHAEUSER	0.355	WHIRLPOOL	0.530
WILLIAMS COS. (162)	**0.000**	**WINN - DIXIE STRS. (163)**	**0.000**
WORTHINGTON INDS.	0.089	WRIGLEY WILLIAM JR.	0.190
WYETH (164)	**0.002**	XCEL ENERGY	0.275
XEROX	0.809	XILINX	0.120
ZIONS BANCORP.	0.272	**3M (165)**	**0.001**

https://doi.org/10.1017/S0266466609990612 Published online by Cambridge University Press

7. CONCLUSIONS

The martingale difference restriction is an outcome of many theoretical analyses in economics and finance. A large body of econometric literature deals with tests of that restriction. We provide new tests based on radial basis function neural networks. Our work is based on the test design of Blake and Kapetanios (2000, 2003a, 2003b). However, unlike that work, we can provide a formal theoretical justification for the validity of these tests using a general approximation result for RBFNN provided in the current paper. This result takes advantage of the link between the algorithms of Blake and Kapetanios (2000, 2003a, 2003b) and boosting. We carry out a Monte Carlo study of the properties of the new tests and find that they have superior power performance to all existing tests of the martingale difference hypothesis we consider. An empirical application to the S&P 500 constituents further illustrates the usefulness of our new tests.

NOTES

1. Boosting refers to a set of algorithms that have become very popular in disciplines such as machine learning and, more recently, statistics, in the context of classification and prediction (see, e.g., Freund and Schapire (1996), Friedman, Hastie, and Tibshirani (2000), Schapire (2002), Friedman (2001), and Buhlmann (2006)).

2. Guay and Guerre (2006) advocate a trigonometric expansion.

3. We have considered using two lags and found that the results we report do not change appreciably in this case.

REFERENCES

Beran, J. (1994) *Statistics for Long-Memory Processes*. Chapman and Hall.

Bierens, H.J. (1984) Model specification testing of time series regression. *Journal of Econometrics* 26, 323–353.

Bierens, H.J. (1990) A consistent conditional moment test of functional form. *Econometrica* 58, 1443–1458.

Bierens, H.J. & W. Ploberger (1991) Spectral-based test for the martingale hypothesis. *Journal of Econometrics* 50, 1–19.

Bierens, H.J. & W. Ploberger (1997) Asymptotic theory of integrated conditional moment tests. *Econometrica* 65, 1129–1151.

Blake, A.P. & G. Kapetanios (2000) A radial basis function artificial neural network test for ARCH. *Economics Letters* 69, 15–23.

Blake, A.P. & G. Kapetanios (2003a) Pure significance tests of the unit root hypothesis against nonlinear alternatives. *Journal of Time Series Analysis* 24(3), 253–267.

Blake, A.P. & G. Kapetanios (2003b) A radial basis function artificial neural network test for neglected nonlinearity. *The Econometrics Journal* 6(2), 357–373.

Blake, A.P. & G. Kapetanios (2007) Testing for ARCH in the presence of nonlinearity of unknown form in the conditional mean. *Journal of Econometrics* 137(2), 472–488.

Buhlmann, P. (2006) Boosting for high-dimensional linear models. *Annals of Statistics* 34, 559–583.

Cochrane, J.H. (2005) *Asset Pricing*, revised ed. Princeton University Press.

Davidson, J. (1994) *Stochastic Limit Theory*. Oxford University Press.

de Jong, R.M. (1996) The Bierens tests under data dependence. *Journal of Econometrics* 72, 1–32.

Deo, R.S. (2000) Spectral tests of the martingale hypothesis under conditional heteroskedasticity. *Journal of Econometrics* 99, 291–315.

https://doi.org/10.1017/S0266466609990612 Published online by Cambridge University Press

Dominguez, M. & I.N. Lobato (2003) A consistent test for the martingale difference hypothesis. *Econometric Reviews* 22, 351–377.

Escanciano, J.C. (2006) Consistent diagnostic test for regression models using projections. *Econometric Theory* 22(1), 1030–1051.

Escanciano, J.C. & C. Velasco (2006a) Generalised spectral tests for the martingale difference hypothesis. *Journal of Econometrics* 134(1), 151–185.

Escanciano, J.C. & C. Velasco (2006b) Testing the martingale difference hypothesis using integrated regression functions. *Computational Statistics and Data Analysis* 51(1), 2278–2294.

Freund, Y. & R. Schapire (1996) Experiments with a new boosting algorithm. *Machine Learning: Proceedings of the 13th International Conference* 148–156.

Friedman, J. (2001) Greedy function approximation: A gradient boosting machine. *Annals of Statistics* 29, 1189–1232.

Friedman, J., T. Hastie, & R. Tibshirani (2000) Additive logistic regression: A statistical view of boosting. *Annals of Statistics* 28, 337–374.

Girosi, F. & G. Anzelloti (1993) Rates of convergence for radial basis functions and neural networks. In R.J. Mammone (ed.), *Artificial Neural Networks for Speech and Vision*. Chapman and Hall.

Guay, A. & E. Guerre (2006) A data-driven nonparametric specification test for dynamic regression models. *Econometric Theory* 22, 543–586.

Hong, Y. (1999a) Hypothesis testing in time series via the empirical characteristic function: A generalized spectral density approach. *Journal of the American Statistical Association* 84, 1201–1220.

Hong, Y. (1999b) Testing serial independence via the empirical characteristic function. Working paper.

Hornik, K., M. Stinchcombe, & H. White (1989) Multi-layer feedforward networks and universal approximators. *Neural Network* 2, 359–366.

Koul, H.L. & W. Stute (1999) Nonparametric model checks for time series. *Annals of Statistics* 27, 204–236.

Lee, T.H., H. White, & C.W.J. Granger (1993) Testing for neglected nonlinearity in time series models: A comparison of neural network methods and alternative tests. *Journal of Econometrics* 56, 269–290.

Lettau, M. & S. Ludvigson (2001) Consumption, aggregate wealth and expected stock returns. *Journal of Finance* 56, 815–849.

Moody, J. & C. Darken (1989) Fast learning in networks of locally-tuned processing units. *Neural Computation* 1(2), 289–303.

Orr, M.J. (1995) Regularisation in the selection of radial basis function centers. *Neural Computation* 7(3), 606–623.

Park, J. & I.W. Sandberg (1991) Universal approximation using radial-basis-function networks. *Neural Computation* 3(4), 246–257.

Park, J.Y. & Y.J. Whang (1999) Testing for the martingale hypothesis. Working paper.

Poskitt, D.S. (2005) Autoregressive Approximation in Nonstandard Situations: The Non-Invertible and Fractionally Integrated Cases. Working paper 16/05, Monash University.

Schapire, R. (2002) The boosting approach to machine learning: An overview. In D. Denison, M. Hansen, C. Holmes, B. Mallick, & B. Yu. Springer (eds.), *MSRI Workshop on Nonlinear Estimation and Classification*. Mathematical Sciences Research Institute.

Stinchcombe, M.B. & H. White (1998) Consistent specification testing with nuisance parameters present only under the alternative. *Econometric Theory* 14, 295–325.

Stute, W. (1997) Nonparametric model checks for regression. *Annals of Statistics* 25, 613–641.

Temlyakov, V.N. (2000) Weak greedy algorithms. *Advances in Computational Mathematics* 12, 213–227.

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

Whang, Y.J. (2000) Consistent bootstrap tests of parametric regression functions. *Journal of Econometrics* 98, 27–46.

White, H. (1999) *Asymptotic Theory for Econometricians*. Academic Press.

https://doi.org/10.1017/S0266466609990612 Published online by Cambridge University Press

White, H. (2006) Approximate nonlinear forecasting methods. In G. Elliott, C.W.J. Granger, & A. Timmermann (eds.), *Handbook of Economics Forecasting*. Elsevier.

White, H. & J. Wooldridge (1991) Some results on sieve estimation with dependent observations. In W. Barnett, J. Powell, & G. Tauchen (eds.), *Nonparametric and Semiparametric Methods in Econometrics and Statistics*. Cambridge University Press.

APPENDIX A: Theorems for the MDH Tests

The proofs of Theorems 2–4 consist of showing that all conditions used in Theorems 1–3 of Guay and Guerre (2006), and therefore by extension in the relevant parts of Propositions 1 and 2 and Lemmas 1 and A.1–A3 of the same paper, for the trigonometric series expansion, hold for the neural network expansion apart from the different polynomial approximation rate. These conditions are (1) uniform boundedness and orthonormality of the basis functions used to construct the approximation to the unknown regression function (Guay and Guerre, Lems. A.1–A.3); (2) the cardinality of the set of the possible number of nodes for the approximation should be $\ln T$ (Guay and Guerre, Lem. A.2); and (3) The series expansion approximates the unknown regression function at a polynomial rate (Guay and Guerre, Lem. 1). Conditions (2) and (3) follow immediately from Theorem 1 and Algorithm 1.

We investigate condition (1). The set of radial basis functions is uniformly bounded by definition for any radial basis function. However, the ordered set of functions arising out of the boosting algorithm is not orthonormal. Nevertheless, it can be made orthonormal using a number of possible orthonormalization algorithms. We consider the Gram-Schmidt orthonormalization algorithm. Let $\Psi_m = \{\psi(x, t_1, \sigma_T), \ldots, \psi(x, t_m, \sigma_T)\}$ denote a set of radial basis functions used in a regression to approximate μ_1. Let the transformed set of functions be denoted $\breve{\Psi}_m = \left\{\breve{\psi}(x, t_1, \sigma_T), \ldots, \breve{\psi}(x, t_m, \sigma_T)\right\}$, where $\breve{\Psi}_m$ has been obtained from Ψ_m by Gram-Schmidt orthonormalization as follows:

$$\breve{\psi}(x, t_1, \sigma_T) = \frac{\psi(x, t_1, \sigma_T)}{\|\psi(x, t_1, \sigma_T)\|}, \tag{A.1}$$

$$\breve{\psi}(x, t_2, \sigma_T) = \frac{\psi(x, t_2, \sigma_T) - \left\langle \psi(x, t_2, \sigma_T), \breve{\psi}(x, t_1, \sigma_T) \right\rangle \breve{\psi}(x, t_1, \sigma_T)}{\left\| \psi(x, t_2, \sigma_T) - \left\langle \psi(x, t_2, \sigma_T), \breve{\psi}(x, t_1, \sigma_T) \right\rangle \breve{\psi}(x, t_1, \sigma_T) \right\|}, \tag{A.2}$$

$\ldots$

$$\breve{\psi}(x, t_m, \sigma_T) = \frac{\psi(x, t_m, \sigma_T) - \sum_{i=1}^{m-1} \left\langle \psi(x, t_m, \sigma_T), \breve{\psi}(x, t_i, \sigma_T) \right\rangle \breve{\psi}(x, t_i, \sigma_T)}{\left\| \psi(x, t_m, \sigma_T) - \sum_{i=1}^{m-1} \left\langle \psi(x, t_m, \sigma_T), \breve{\psi}(x, t_i, \sigma_T) \right\rangle \breve{\psi}(x, t_i, \sigma_T) \right\|}, \tag{A.3}$$

where $\langle ., . \rangle$ and $\|.\|$ denote some functional space inner product and norm, respectively. In order to prove the equivalence of using either Ψ_m or $\breve{\Psi}_m$ in a regression to approximate μ_1, we simply note that, for all i,

$$\breve{\psi}(x, t_i, \sigma_T) = \sum_{j=1}^{i} \breve{c}_{ji} \psi(x, t_j, \sigma_T),$$

https://doi.org/10.1017/S0266466609990612 Published online by Cambridge University Press

where the $\breve{c}_{ji}$'s are determined in the recursions (A.1)–(A.3). Therefore,

$$\psi(x;m) = \sum_{i=1}^{m} \breve{c}_i \breve{\psi}(x, t_i, \sigma_T) = \sum_{i=1}^{m} \breve{c}_i \left(\sum_{j=1}^{i} \breve{c}_{ji} \psi(x, t_j, \sigma_T) \right)$$

$$= \sum_{i=1}^{m} \sum_{j=1}^{i} \breve{c}_i \breve{c}_{ji} \psi(x, t_j, \sigma_T) = \sum_{i=1}^{m} c_i \psi(x, t_i, \sigma_T),$$

where by grouping appropriate terms,

$$c_i = \sum_{\ell=i}^{m} \breve{c}_\ell \breve{c}_{i\ell}.$$

To complete the proof of Theorems 2–4 we need to establish two more facts. The first relates to the validity of using the penalty terms associated with the Bayesian and Hannan-Quinn information criteria for Theorem 1. But given that these penalty terms are of a higher order than $(\ln \ln T)^{1/2}$, the result follows immediately. The second fact relates to the relaxation of the assumption that the minimum possible order k_{min} over which to search for k^* has to tend to infinity, which was made in Guay and Guerre (2006), since they consider the case where a preliminary estimation leads to a set of residuals, which are then tested for lack of structure (in our case, the MDH hypothesis). The assumption is needed to make the estimation error of the preliminary estimation negligible. Since we do not consider any preliminary estimation, this assumption is not needed. This completes the proof. ■

APPENDIX B: Proof of Theorem 1

The following two lemmas are needed for the main result.

LEMMA 1. *Under Assumptions 1–3 and Assumption 5 or Assumption 6, for all* $0 < \xi < 1/2$ *and assuming further that* $\mathrm{E}|\epsilon_t^s| < \infty$ *for some* $s > \max(2/\xi, 4)$, *we have*

$$\sup_{1 \le j,k \le p_T} \left| T^{-1} \sum_{t=1}^{T} g_j(x_t) g_k(x_t) - \mathrm{E}\left(g_j(x_t) g_k(x_t)\right) \right| = o_p(T^{-\xi/2}), \tag{B.1}$$

$$\sup_{1 \le j \le p_T} \left| T^{-1} \sum_{i=1}^{T} g_j(x_t) \epsilon_t \right| = O_p(T^{-\xi/2}), \tag{B.2}$$

$$\sup_{1 \le j \le p_T} \left| T^{-1} \sum_{t=1}^{T} f(x_t) g_j(x_t) - \mathrm{E}\left(f(x_t) g_j(x_t)\right) \right| = O_p(T^{-\xi/2}) \tag{B.3}$$

and

$$\sup_{1 \le j \le p_T} \left| T^{-1} \sum_{t=1}^{T} g_j(x_t) y_t - \mathrm{E}\left(g_j(x_t) y_t\right) \right| = O_p(T^{-\xi/2}), \tag{B.4}$$

where g_j, $j = 1, \ldots, p_T$ *are, uniformly over* j, *bounded continuous functions,* $f(.) = \sum_{j=1}^{p_T} \beta_j g_j(.)$ *and* $\sum_{j=1}^{p_T} |\beta_j| = o(T^{1/s})$ *for all* $s > 0$.

https://doi.org/10.1017/S0266466609990612 Published online by Cambridge University Press

Proof of Lemma 1. For (B.1) we need to consider the different implications of Assumptions 5 and 6 for the bounded quantity $g_j(x_t)$, $j = 1, \ldots, p_T$. We start with Assumption 5, which allows a greater extent of temporal dependence in x_t at the expense of a slower rate of increase in p_T. Let the autocovariance function of $g_j(x_t)g_k(x_t)$ be denoted by $c_{jk,\tau}$. By Markov's inequality, it easily follows that

$$\sup_j \Pr\left(\left|\sum_{t=1}^{T} g_j(x_t)g_k(x_t) - \mathrm{E}\left(g_j(x_t)g_k(x_t)\right)\right| > \epsilon\right) \leq \frac{2T}{\epsilon^2} \sup_j \sum_{\tau=0}^{T} c_{jk,\tau}. \tag{B.5}$$

We examine the behavior of the right-hand side (RHS) of (B.5). By Assumption 5, x_t is an $L_2 - NED$ process of size -3. By the fact that $g_j(.)$ is, uniformly over j, a bounded function for all j, $g_j(.)$ satisfies the following uniform Lipschitz condition for all finite constant vectors a and b and some finite constant scalar C:

$$\sup_j |g_j(a) - g_j(b)| \leq C\rho(a, b),$$

where $\rho(a, b) = \sum_{i=1}^{d} |a_i - b_i|$. Then, by Theorem 17.12 of Davidson (1994), it follows that $g_j(x_t)$ is an $L_2 - NED$ process of size -3. Note that $\sup_j \left\|g_j(x_t)\right\|_r \leq \infty$ for all finite r. Then, by Example 17.17 of Davidson (1994), $g_j(x_t)g_k(x_t)$ is an $L_2 - NED$ process of size $-3(r-2)/2(r-1)$ for all finite r, uniformly over j, which implies that it is an $L_2 - NED$ process of size $-3/2$. Then, by Theorem 17.7 of Davidson and the mixing assumption in Assumption 5, $\sup_j \sum_{\tau=0}^{T} c_{jk,\tau} < \infty$ since the theorem requires a *NED* size of -1. Setting $\epsilon = T\epsilon$ in (B.5) gives a rate of convergence to zero for the RHS of (B.5) of T^{-1}. Now (B.1) holds if

$$\Pr\left(T^{\xi/2} \sup_{1 \leq j,k \leq p_T} \left|T^{-1} \sum_{t=1}^{T} g_j(x_t)g_k(x_t) - \mathrm{E}\left(g_j(x_t)g_k(x_t)\right)\right| > \epsilon\right) = o(1), \tag{B.6}$$

for all $0 < \xi < 1/2$. But

$$\Pr\left(T^{\xi/2} \sup_{1 \leq j,k \leq p_T} \left|T^{-1} \sum_{t=1}^{T} g_j(x_t)g_k(x_t) - \mathrm{E}\left(g_j(x_t)g_k(x_t)\right)\right| > \epsilon\right)$$

$$\leq p_T^2 \sup_j \Pr\left(T^{\xi/2} \left|T^{-1} \sum_{t=1}^{T} g_j(x_t)g_k(x_t) - \mathrm{E}\left(g_j(x_t)g_k(x_t)\right)\right| > \epsilon\right). \tag{B.7}$$

Using (B.5), the RHS of (B.7) is majorized by $Cp_T^2 T^{\xi-1}$. But by Assumption 5, $p_T = o(T^{1/4})$. Hence, $Cp_T^2 T^{\xi-1} = o(1)$ and (B.6) follows. We now prove (B.1) under Assumption 6. Now p_T is allowed to grow at a faster polynomial rate than $1/4$, but this implies that the rate obtained in (B.5) is too slow. We therefore make use of Bernstein's inequality. Theorem 3.3 of White and Wooldridge (1991) gives a Bernstein inequality allowing for α-mixing stationary x_t. Using Theorem 3.49 of White (1999), we note that if x_t is α-mixing of a given size then $g_j(x_t)g_k(x_t)$ is also α-mixing of the same size. Then, noting that $g_j(x_t)g_k(x_t)$ has a finite upper bound, we get from Theorem 3.3 of White and Wooldridge (1991) that, for some finite constants C_1 and C_2 and for all $0 < \beta < 1$,

$$\sup_j \Pr\left(\left|\sum_{t=1}^{T} g_j(x_t)g_k(x_t) - \mathrm{E}\left(g_j(x_t)g_k(x_t)\right)\right| > \epsilon\right) \leq C_1 \exp\left(-C_2 \epsilon T^{-\frac{1}{2}}\right). \tag{B.8}$$

https://doi.org/10.1017/S0266466609990612 Published online by Cambridge University Press

Setting $\epsilon = T^{1-\xi/2}\epsilon$, we get

$$p_T^2 \sup_j \Pr\left(T^{\xi/2}\left|T^{-1}\sum_{t=1}^{T} g_j(x_t)g_k(x_t) - \mathrm{E}\left(g_j(x_t)g_k(x_t)\right)\right| > \epsilon\right)$$

$$\leq C_1 p_T^2 \exp\left(-C_2\epsilon T^{1/2-\xi/2}\right).$$

Thus,

$$C_1 p_T^2 \exp\left(-C_2\epsilon T^{1/2-\xi/2}\right) = o(1),$$

for all $\epsilon > 0$, as long as $p_T = O(T^q)$ for all $q > 0$.

We now consider (B.2). Once again we consider the different implications of Assumptions 5 and 6. Starting with Assumption 5 we note that by the Markov inequality we get

$$\sup_j \Pr\left(\left|\sum_{t=1}^{T} g_j(x_t)\epsilon_t\right| > \epsilon\right) \leq \frac{2T}{\epsilon^2}\sup_j \sum_{\tau=0}^{T} c_{j,\tau}, \tag{B.9}$$

where $c_{j,\tau}$ denotes the autocovariance function of $g_j(x_t)\epsilon_t$. If Assumption 3 holds, then all autocovariances are trivially zero and so $\sup_j \sum_{\tau=0}^{T} c_{j,\tau} < \infty$. Thus

$$\Pr\left(T^{\xi/2}\sup_{1\leq j\leq p_T}\left|T^{-1}\sum_{t=1}^{T} g_j(x_t)\epsilon_t\right| > \epsilon\right) \leq p_T \sup_j \Pr\left(T^{\xi/2}\left|T^{-1}\sum_{t=1}^{T} g_j(x_t)\epsilon_t\right| > \epsilon\right). \tag{B.10}$$

Then the RHS of (B.10) is majorized by $Cp_T T^{\xi-1}$ which, by Assumption 5, that $p_T = o(T^{1/4})$, is $o(1)$. Next we establish (B.2) under Assumption 6. Direct use of Bernstein's inequality for dependent processes is not possible, as it applies to bounded random variables. So we use a truncation argument to get the inequality we need. Let

$$\tilde{\epsilon}_t = \begin{cases} \epsilon_t, & \text{if } |\epsilon_t| \leq C_T \\ \mathrm{sign}(C_T)C_T, & \text{if } |\epsilon_t| > C_T \end{cases},$$

where C_T is a sequence to be defined below. We have that

$$\Pr\left(T^{\xi/2}\sup_{1\leq j\leq p_T}\left|T^{-1}\sum_{t=1}^{T} g_j(x_t)\epsilon_t\right| > \epsilon\right)$$

$$\leq \Pr\left(T^{\xi/2}\sup_{1\leq j\leq p_T}\left|T^{-1}\sum_{i=1}^{T} g_j(x_t)\tilde{\epsilon}_t - \mathrm{E}\left(g_j(x_t)\tilde{\epsilon}_t\right)\right| > \epsilon/3\right)$$

$$+\Pr\left(T^{\xi/2}\sup_{1\leq j\leq p_T}\left|T^{-1}\sum_{t=1}^{T} g_j(x_t)\left(\epsilon_t - \tilde{\epsilon}_t\right)\right| > \epsilon/3\right)$$

$$+I\left(T^{\xi/2}\sup_{1\leq j\leq p_T}\left|T^{-1}\sum_{t=1}^{T}\mathrm{E}\left(g_j(x_t)\left(\epsilon_t - \tilde{\epsilon}_t\right)\right)\right| > \epsilon/3\right), \tag{B.11}$$

https://doi.org/10.1017/S0266466609990612 Published online by Cambridge University Press

where $I(.)$ denotes the indicator function. We look, in turn, at the three terms on the RHS of (B.11). For the first term we can use the Bernstein inequality of (B.8). However, it takes the slightly different form below, since $\tilde{\epsilon}_t$ is not bounded by a constant but by C_T:

$$\sup_j \Pr\left(\left|\sum_{t=1}^{T} g_j(x_t)\tilde{\epsilon}_t - \mathrm{E}\left(g_j(x_t)\tilde{\epsilon}_t\right)\right| > \epsilon\right) \leq C_1 \exp\left(\frac{-C_2\epsilon T^{-\frac{1}{2}}}{C_T}\right).$$

Then, setting $\epsilon = T^{1-\xi/2}\epsilon$ and $C_T = T^{\xi/2}$, we have that

$$\Pr\left(T^{\xi/2} \sup_{1\leq j\leq p_T}\left|T^{-1}\sum_{t=1}^{T} g_j(x_t)\tilde{\epsilon}_t - \mathrm{E}\left(g_j(x_t)\tilde{\epsilon}_t\right)\right| > \epsilon/3\right)$$

$$\leq C_1 p_T \exp\left(-C_2\epsilon T^{1/2-\xi}\right) = o(1).$$

We next look at the second term of the RHS of (B.11). We have that

$$\Pr\left(T^{\xi/2} \sup_{1\leq j\leq p_T}\left|T^{-1}\sum_{t=1}^{T} g_j(x_t)\left(\epsilon_t - \tilde{\epsilon}_t\right)\right| > \epsilon/3\right)$$

$$\leq \Pr(\exists t \text{ such that } |\epsilon_t| > C_T) \leq T\Pr(|\epsilon_t| > C_T) \leq T\frac{\mathrm{E}|\epsilon_t|^s}{C_T^s}.$$

But

$$T\frac{\mathrm{E}|\epsilon_t|^s}{C_T^s} = O(T^{1-s\xi/2}) = o(1)$$

since, by Assumption 1, $s > 2/\xi$. Finally, we consider the third term of the RHS of (B.11). By the uncorrelatedness of $g(x_t)$ and ϵ_t and the boundedness of $g_j(x_t)$, the term can be bounded by

$$T^{\xi/2}\mathrm{E}\left(g_j(x_t)\left(\epsilon_t - \tilde{\epsilon}_t\right)\right) \leq T^{\xi/2}\mathrm{E}(g_j(x_t))\mathrm{E}\left(\epsilon_t - \tilde{\epsilon}_t\right) \leq C\left|\mathrm{E}\left(\epsilon_t - \tilde{\epsilon}_t\right)\right|.$$

This can be bounded by

$$\left|\mathrm{E}\left(\epsilon_t - \tilde{\epsilon}_t\right)\right| \leq \int I(|x| > C_T)\left(C_T + |x|\right) dP_\epsilon(x)$$

$$= C_T\Pr(|\epsilon_t| > C_T) + \int |x| I(|x| > C_T) dP_\epsilon(x)$$

$$\leq C_T^{1-s}\mathrm{E}|\epsilon_t|^s + \left(\mathrm{E}|\epsilon_t|^2\right)^{1/2}\Pr(|\epsilon_t| > C_T)^{1/2} = o\left(C_T^{-2}\right) = o\left(T^{-\xi}\right).$$

This proves (B.2) under both Assumptions 5 and 6. We next consider (B.3),

$$\sup_{1\leq j\leq p_T}\left|T^{-1}\sum_{t=1}^{T} f(x_t)g_j(x_t) - \mathrm{E}\left(f(x_t)g_j(x_t)\right)\right|$$

$$\leq \sum_{j=1}^{p_T}\left|\beta_j\right| \sup_{1\leq j,k\leq p_T}\left|T^{-1}\sum_{t=1}^{T} g_j(x_t)g_k(x_t) - \mathrm{E}\left(g_j(x_t)g_k(x_t)\right)\right|$$

https://doi.org/10.1017/S0266466609990612 Published online by Cambridge University Press

$$\leq \left(\sum_{j=1}^{p_T} |\beta_j| \right) O_p(T^{-\xi/2}).$$

But $\sum_{j=1}^{p_T} |\beta_j| = o(T^{1/s})$ for all $s > 0$, hence giving the result. Finally, (B.4) easily follows from (B.2) and (B.3). ■

LEMMA 2. *Let Assumptions 1–3 and Assumption 6 hold and assume that* $0 < \xi < b$ *for any* $b < 1$. *Further, assume that*

$$\Pr(|\epsilon_t| > a) \leq C_1 \exp(-C_2 a^p)$$

for some $C_1 > 0$, $C_2 > 0$, *and* $p > 1$, *and that* $\mathrm{E}|\epsilon_t^s| < \infty$ *for some* $s > \max(2/\xi, 4)$. *Then Lemma 1 holds.*

Proof of Lemma 2. The result of the lemma will be established if we show that (B.2) holds under the conditions of the lemma, since it is easy to see from the proof of Lemma 1 that (B.1) and (B.3) follow under the conditions of the current lemma. We revisit the analysis of the Bernstein inequality for unbounded random variables used in Lemma 1. A different truncation argument is then used. Let

$$\tilde{\epsilon}_t = \begin{cases} \epsilon_t, & \text{if } |\epsilon_t| \leq C_T \\ 0, & \text{if } |\epsilon_t| > C_T \end{cases} \quad \text{and} \quad \bar{\epsilon}_t = \begin{cases} 0, & \text{if } |\epsilon_t| \leq C_T \\ \epsilon_t, & \text{if } |\epsilon_t| > C_T \end{cases},$$

where C_T is a sequence to be defined below. Then,

$$\begin{aligned}
&\Pr\left(T^{\xi/2} \sup_{1\leq j\leq p_T} \left| T^{-1} \sum_{t=1}^{T} g_j(x_t)\epsilon_t \right| > \epsilon \right) \\
&\quad \leq \Pr\left(T^{\xi/2} \sup_{1\leq j\leq p_T} \left| T^{-1} \sum_{t=1}^{T} g_j(x_t)\tilde{\epsilon}_t - \mathrm{E}\left(g_j(x_t)\tilde{\epsilon}_t\right) \right.\right. \\
&\qquad \left.\left. + T^{-1} \sum_{t=1}^{T} g_j(x_t)\bar{\epsilon}_t - \mathrm{E}\left(g_j(x_t)\bar{\epsilon}_t\right) \right| > \epsilon \right) \\
&\quad \leq \Pr\left(T^{\xi/2} \sup_{1\leq j\leq p_T} \left| T^{-1} \sum_{t=1}^{T} g_j(x_t)\tilde{\epsilon}_t - \mathrm{E}\left(g_j(x_t)\tilde{\epsilon}_t\right) \right| > \epsilon/2 \right) \\
&\qquad + \Pr\left(T^{\xi/2} \sup_{1\leq j\leq p_T} \left| T^{-1} \sum_{t=1}^{T} g_j(x_t)\bar{\epsilon}_t - \mathrm{E}\left(g_j(x_t)\bar{\epsilon}_t\right) \right| > \epsilon/2 \right). \qquad \textbf{(B.12)}
\end{aligned}$$

For the first term of the RHS of (B.12), we can using Theorem 3.3 of White and Wooldridge (1991) to get

$$\sup_j \Pr\left(\left| T^{-1} \sum_{t=1}^{T} g_j(x_t)\tilde{\epsilon}_t - \mathrm{E}\left(g_j(x_t)\tilde{\epsilon}_t\right) \right| > \epsilon/2 \right) \leq C_1 p_T \exp\left(-C_2 \epsilon \frac{T^{1/2}}{C_T} \right).$$

https://doi.org/10.1017/S0266466609990612 Published online by Cambridge University Press

Then,

$$\Pr\left(T^{\xi/2}\sup_{1\le j\le p_T}\left|T^{-1}\sum_{t=1}^{T}g_j(x_t)\bar{\epsilon}_t-\mathrm{E}\left(g_j(x_t)\bar{\epsilon}_t\right)\right|>\epsilon/2\right)$$

$$\le C_1 p_T\exp\left(-C_2\epsilon\frac{T^{1/2-\xi/2}}{C_T}\right)\le C_1 p_T\exp\left(-C_2\epsilon\frac{T^{(1-b)/2}}{C_T}\right).$$

We let $C_T=T^{\upsilon}$. It is clear that we need $(1-b)/2>\upsilon$. Then,

$$C_1 p_T\exp\left(-C_2\epsilon\frac{T^{(1-b)/2}}{C_T}\right)=o(1)$$

for all polynomial rates of growth for p_T. We next examine the second term of the RHS of (B.12). Using

$$\Pr\left(T^{\xi/2}\sup_{1\le j\le p_T}\left|T^{-1}\sum_{t=1}^{T}g_j(x_t)\bar{\epsilon}_t-\mathrm{E}\left(g_j(x_t)\bar{\epsilon}_t\right)\right|>\epsilon/2\right)$$

$$\le p_T\sup_j\Pr\left(T^{\xi/2}\left|T^{-1}\sum_{t=1}^{T}g_j(x_t)\bar{\epsilon}_t-\mathrm{E}\left(g_j(x_t)\bar{\epsilon}_t\right)\right|>\epsilon/2\right), \tag{B.13}$$

we focus on the RHS of (B.13). Then we have, using Markov's inequality,

$$\sup_j\Pr\left(T^{\xi/2}\left|T^{-1}\sum_{t=1}^{T}g_j(x_t)\bar{\epsilon}_t-\mathrm{E}\left(g_j(x_t)\bar{\epsilon}_t\right)\right|>\epsilon/2\right)$$

$$\le\sup_j\Pr\left(T^{\xi/2-1}\sum_{t=1}^{T}\left|g_j(x_t)\bar{\epsilon}_t-\mathrm{E}\left(g_j(x_t)\bar{\epsilon}_t\right)\right|>\epsilon/2\right)$$

$$\le CT^{\xi/2-1}\sup_j\sum_{t=1}^{T}\mathrm{E}\left|g_j(x_t)\bar{\epsilon}_t\right|.$$

By the uniform boundedness of $g_j(.)$ and Holder's inequality,

$$\sup_j\mathrm{E}\left|g_j(x_t)\bar{\epsilon}_t\right|=\sup_j\mathrm{E}\left|g_j(x_t)\epsilon_t I\left(|\epsilon_t|>C_T\right)\right|$$

$$\le C\left(\mathrm{E}\left(|\epsilon_t|^p\right)\right)^{1/p}\left(\mathrm{E}\left(I\left(|\epsilon_t|>C_T\right)^u\right)\right)^{1/u}$$

$$=C\left(\mathrm{E}\left(|\epsilon_t|^p\right)\right)^{1/p}\Pr\left(|\epsilon_t|>C_T\right)^{1/u},$$

where $p^{-1}+u^{-1}=1$. Since $\left(\mathrm{E}\left(|\epsilon_t|^p\right)\right)^{1/p}<\infty$ and

$$\Pr\left(|\epsilon_t|>C_T\right)\le C_1\exp\left(-C_2C_T^p\right),$$

https://doi.org/10.1017/S0266466609990612 Published online by Cambridge University Press

for some $p > 1$, it follows that

$$\sup_j \Pr\left(T^{\xi/2-1}\sum_{t=1}^{T}\left|g_j(x_t)\bar{\epsilon}_t - \mathrm{E}\left(g_j(x_t)\bar{\epsilon}_t\right)\right| > \epsilon/2\right) \leq CT^{\xi/2-1}\sum_{t=1}^{T}\exp\left(-C_2C_T^p\right)^{1/u}$$

$$\leq CT^{\xi/2-1}\sum_{t=1}^{T}\exp\left(-C_2C_T^p\right) \leq CT^{\xi/2}\exp\left(-C_2C_T^p\right)$$

$$= CT^{\xi/2}\exp\left(-C_2T^{\upsilon p}\right) = o(1),$$

for all $p > 0$ and $\upsilon > 0$. Hence, the result follows. ■

Proof of Theorem 1. We split the problem into two parts: the approximation part and the estimation part. The approximation part relates to approximating $\mu(x)$ by an approximating function of the form

$$\psi(x; p_T) = \sum_{i=1}^{p_T} c_i\psi(x, t_i, \sigma_T), \tag{B.14}$$

where $\psi(x, t_i, \sigma_T)$ is a radial basis function node with center t_i and radius σ_T. The first part of the approximation proof relates to the ability of sums of the form (B.14) to approximate L_2-bounded functions and the conditions required for such an approximation. For that we consider the work of Park and Sandberg (1991). Let

$$\tilde{\psi}(x; p_T) = \sum_{i=1}^{p_T}\bar{\psi}(x, t_i, \sigma_T)\mu^c(x)\left(\frac{2\tau_T}{p_T\sigma_T}\right)^q, \tag{B.15}$$

where

$$\bar{\psi}(x, t_i, \sigma_T) = \frac{\psi(x, t_i, \sigma_T)}{\int_{\mathbb{R}^r}\bar{\psi}(x, t_i, \sigma_T)}, \tag{B.16}$$

t_i is a partitioning of $[-\tau_T, \tau_T]^q$ such that all partition intervals are $o(p_T^{-C})$ for some $0 < C < 1$, and $\mu^c(x)$ is some continuous function that approximates arbitrarily well $\mu(x)$. This latter fact is possible since the space of continuous functions is dense in the space of L_2 bounded functions. Then Park and Sandberg (1991) show that

$$\tilde{\psi}(x; p_T) - \mu(x) = o(1),$$

for all x not belonging to some set on $\mathbb{R}$ of measure zero, as $p_T \to \infty$, $\sigma_T \to 0$, and $\tau_T \to \infty$. The latter two limits can have arbitrarily slow rates with respect to T. It is clear that (B.15) is of the form (B.14), with

$$c_i = \frac{\mu^c(x)}{\int_{\mathbb{R}^r}\bar{\psi}(x, t_i, \sigma_T)}\left(\frac{2\tau_T}{p_T\sigma_T}\right)^q. \tag{B.17}$$

So

$$\sup_{T\in\mathbb{R}}\sum_{i=1}^{p_T}|c_i| = o\left(\left(\log_a T\right)^C\right), \tag{B.18}$$

https://doi.org/10.1017/S0266466609990612 Published online by Cambridge University Press

for all $C > 0$, if σ_T and τ_T converge to zero and ∞ at slow enough rates; e.g., they behave as $(\ln\ln T)^{-1}$ and $\ln\ln T$, respectively. Finally, Girosi and Anzelloti (1993) show that the approximation has a rate of $p_T^{1/2}$. This concludes the first part of the proof. The above argument holds for finite q. The theorem allows for $q \equiv q_T$, albeit at a slow rate given by

$$q_T = o\left(\frac{\ln\left(\log_a T\right)}{\ln\left(\frac{2\tau_T}{p_T\sigma_T}\right)}\right), \tag{B.19}$$

as can be easily seen from (B.17) and (B.18). So, for example, if $\sigma_T = (\ln\ln T)^{-1}$ and $\tau_T = o(\ln\ln T)$, $q_T = \ln\left(\left(\log_\kappa T\right)\right)$, for all $\kappa > a$, is compatible with (B.19). Since the argument of Girosi and Anzelloti (1993) concerning the rate of convergence, $p_T^{1/2}$, is explicitly independent of the dimension q, (see Girosi and Anzelloti, 1993 Sec. 8.2), the result follows.

The second part relates to the estimation. Given the above approximation argument, we now assume the existence of a representation of the form (B.14) for the regression function $\mu(x_t)$. We wish to estimate a representation of the form

$$\psi(x;m) = \sum_{i=1}^{m} c_i\,\psi(x, t_i, \sigma_T), \tag{B.20}$$

where t_i, $i = 1,\ldots,m$ are centers that are obtained by some partition of $[-\tau_T, \tau_T]^q$; τ_T and σ_T are defined in Remarks 3 and 2 ; $m \to \infty$, $m = o(p_T)$, and, more importantly, order the centers, t_i, via the boosting algorithm. To do that we use the framework of Buhlmann (2006). That framework is not directly applicable to our setting because it deals with independent observations. We therefore extend a number of results there to accommodate our needs. Let

$$\hat{\psi}(x;m) = \sum_{i=1}^{m} \hat{c}_{\mathcal{S}_i}\,\psi(x, t_{\mathcal{S}_i}, \sigma_T)$$

denote the estimated regression function after m iterations of the boosting algorithm, where $(\mathcal{S}_1, \mathcal{S}_2 \ldots)$ denotes the reordering of the centers $(1, 2, \ldots,)$ obtained by the boosting algorithm. Then Theorem 1 of Buhlmann (2006) states that $\hat{\psi}(x;m)$ converges to $\psi(x;p_T)$ as $m \to \infty$ at a slow enough rate, i.e., $m = o(\ln T)$ and $p_T = O(e^{T^{1-\xi}})$. In order to use this result in our framework, we need to (i) accommodate dependence in the data, (ii) allow for $\sup_{T\in\mathbb{R}} \sum_{i=1}^{p_T} |c_i| \to \infty$, and (iii) determine a rate at which $\hat{\psi}(x;m)$ converges to $\psi(x;p_T)$. We deal with each issue in turn. First we substitute Lemma 1 of Buhlmann (2006) with our Lemmas 1 and 2, which deal with dependent data. Second, we need to deal with the unboundedness of $\sup_{T\in\mathbb{R}} \sum_{i=1}^{p_T} |c_i|$. Note that it is sufficient for our results to only allow for a rate of growth of $\sup_{T\in\mathbb{R}} \sum_{i=1}^{p_T} |c_i|$ that is arbitrarily slow with respect to T. This unboundedness can be accommodated by examining Theorem 5.1 of Temlyakov (2000) which is used in (6.5) of Buhlmann (2006). Let the remainder function at the ith step of the boosting algorithm for some original regression function f be denoted by $R^i f$. Further, let b be defined by

$$\left|\left\langle R^{i-1}\psi(x;p_T), \psi(x, t_{\mathcal{S}_i}, \sigma_T)\right\rangle\right| \geq b \sup_{1\leq j\leq p_T} \left|\left\langle R^{i-1}\psi(x;p_T), \psi(x, t_j, \sigma_T)\right\rangle\right|,$$

https://doi.org/10.1017/S0266466609990612 Published online by Cambridge University Press

where $t_{\mathcal{S}_i}$ is the center selected at the ith step of the boosting algorithm. Then, by Theorem 5.1 of Temlyakov (2000),

$$\left\| R^i \psi(x; p_T) \right\| \leq \left(\sup_{T \in \mathbb{R}} \sum_{i=1}^{p_T} |c_i| \right) (1+m)^{-b/2(2+b)}.$$

Then, if

$$(1+m)^{-b/2(2+b)} = O\left(\left(\log_a T \right)^{-b/2(2+b)} \right),$$

it follows that by letting $\sup_{T \in \mathbb{R}} \sum_{i=1}^{p_T} |c_i|$ grow slowly enough, as in, e.g., (B.18) ,

$$\left(\sup_{T \in \mathbb{R}} \sum_{i=1}^{p_T} |c_i| \right) (1+m)^{-b/2(2+b)} = O\left(\left(\log_a T \right)^{-b/2(2+b)} \right)$$

as well. Finally, we need to determine a rate at which $\hat{\psi}(x; m)$ converges to $\psi(x; p_T)$. But

$$\left| \hat{\psi}(x; m) - \psi(x; p_T) \right| \leq \left\| R^m \mu \right\|. \tag{B.21}$$

Then, using (6.17), (6.19) of Buhlmann (2006) and considering sharper bounds in the analysis preceding (6.19) of Buhlmann,

$$\left\| R^m \psi(x; p_T) \right\| \leq C_1 (1+m)^{-b/2(2+b)} + C_2 m (5/2)^m T^{-\xi} + C_3 C_4^m T^{-\xi}, \tag{B.22}$$

for all $C_4 > 5/2$, on the set A_T, where A_T denotes the set of events where (B.1)–(B.4) simultaneously occur. The sharper bounds referred to above relate to the following: The third term of the RHS of (B.22) arises out of bounding $\left\| R^m \psi(x; p_T) - \tilde{R}^m \psi(x; p_T) \right\|$, where $\tilde{R}^m \psi(x; p_T)$ is a "semi"-population version of the remainder function R that uses population covariances rather than sample covariances. Buhlmann (2006) shows that

$$\left\| R^m \psi(x; p_T) - \tilde{R}^m \psi(x; p_T) \right\| \leq \left\| R^{m-1} \psi(x; p_T) - \tilde{R}^{m-1} \psi(x; p_T) \right\| + C_3 (5/2)^{m-1} T^{-\xi}.$$

Then

$$\left\| R^m \psi(x; p_T) - \tilde{R}^m \psi(x; p_T) \right\| \leq C_6 T^{-\xi} \sum_{i=1}^{m} (5/2)^{m-1} \leq C_6 T^{-\xi} (5/2)^m,$$

rather than

$$\left\| R^m \psi(x; p_T) - \tilde{R}^m \psi(x; p_T) \right\| \leq C_6 T^{-\xi} 3^m,$$

given in (6.19) of Buhlmann (2006). We note that in Algorithm 1, b can be arbitrarily close to 1. Since by Lemmas 1 and 2, $\Pr(A_T) = 1 - O(T^{-\xi/8})$, it follows that

$$\left\| R^m \psi(x; p_T) \right\| \leq C_1 m^{-C_5} + C_3 C_4^m T^{-\xi}, \tag{B.23}$$

for all $C_5 > 6$ and all $C_4 > 5/2$. If Lemma 1 holds, then if $m < \log_a T$, for all a that satisfies $\log_a e < \frac{\ln(5/2)}{4}$, it follows that there exists $C_4 > 5/2$ such that $C_4^m T^{-\xi/2} < T^{\log_a C_4 - \xi/2}$,

https://doi.org/10.1017/S0266466609990612 Published online by Cambridge University Press

and since $\xi < 1/2$, $\log_a C_4 - \xi/2 < 0$. If Lemma 2 holds, then $\xi < 1$, and so a only needs to satisfy $\log_a e < \frac{\ln(5/2)}{2}$. Under these conditions, the second term of the RHS of (B.23) declines polynomially in T, whereas the first term declines at a slower logarithmic rate in T, which therefore dominates. Overall

$$\left\|\hat{\mu}(x_t) - \mu(x_t)\right\| \leq \left\|R^m \psi(x; p_T)\right\| + \|\mu - \psi(x; p_T)\| = o_p\left(m^{-1/C_1}\right), \tag{B.24}$$

for all x_t and for all $C_1 > 6$, proving the theorem. ■

Note that the above proof does not explicitly consider the possible conditional heteroskedasticity of ϵ_t and that we do not make any conditional homoskedasticity assumption anywhere in the above proofs, thereby allowing for conditional heteroskedasticity. Further, note that even unconditional heteroskedasticity can be allowed for Theorem 1. The extension to this case follows easily upon noting Remark 1 of Buhlmann (2006).

https://doi.org/10.1017/S0266466609990612 Published online by Cambridge University Press