JOURNAL OF TIME SERIES ANALYSIS
J. Time Ser. Anal. (2022)
Published online in Wiley Online Library
(wileyonlinelibrary.com) DOI: 10.1111/jtsa.12668

ORIGINAL ARTICLE

A NONPARAMETRIC PREDICTIVE REGRESSION MODEL USING PARTITIONING ESTIMATORS BASED ON TAYLOR EXPANSIONS

JOSE OLMO[a,b]

[a]*Departamento de Análisis Económico, Universidad de Zaragoza, Zaragoza, Spain*
[b]*Department of Economics, University of Southampton, Southampton, UK*

This article proposes a nonparametric predictive regression model. The unknown function modeling the predictive relationship is approximated using polynomial Taylor expansions applied over disjoint intervals covering the support of the predictor variable. The model is estimated using the theory on partitioning estimators that is extended to a stationary time series setting. We show pointwise and uniform convergence of the proposed estimator and derive its asymptotic normality. These asymptotic results are applied to test for the presence of predictive ability. We develop an asymptotic pointwise test of predictive ability using the critical values of a Normal distribution, and a uniform test with asymptotic distribution that is approximated using a *p*-value transformation and Wild bootstrap methods. These theoretical insights are illustrated in an extensive simulation exercise and also in an empirical application to forecasting high-frequency based realized volatility measures. Our results provide empirical support to the presence of nonlinear autoregressive predictability of these measures for the constituents of the Dow Jones index.

Received 28 January 2022; Revised 10 October 2022; Accepted 10 October 2022

Keywords: Series estimators; Taylor expansions; asymptotic theory; realized volatility; time series predictability.

JEL. C12; C13; C14.

1. INTRODUCTION

Forecasting is one of the main objectives of time series econometrics. Although the time series forecasting literature has been dominated by parametric models there has also been progress on nonparametric models. Kernel-based estimators characterize an important class of nonparametric time series models. In this group, local polynomial kernel estimators, Nadaraya (1965), Watson (1964) and Fan and Gijbels (1996), have been the main workhorse in nonparametric time series analysis. These models have been also extended to semiparametric settings such as the partially linear model with dependent data, see Andrews (1994), additive time series regression models as in Kim *et al.* (1999), and varying coefficient models as in Cai *et al.* (2000a), Cai *et al.* (2000b) and Fan *et al.* (2003) for local linear estimators.

Another important class of nonparametric models is characterized by series regressions. Work on these models was initiated by Tukey, 1947; Tukey, 1961 and developed further by Stone (1985), Chen (1988), Andrews (1991) and Newey (1997), among others. These methods improve the fit of standard linear regression models by approximating an unknown smooth function by an increasing number of regressors characterized by a set of basis functions. Representatives of this class are wavelets, power series and splines. These methods have gained popularity due to their tractability, flexibility and conceptual simplicity. Recent work by Cattaneo and Farrell (2013) have specialized this class of nonparametric models by considering partitioning estimators in cross-sectional settings. These authors derive optimal uniform convergence rates and asymptotic normality of these estimators in independent and identically distributed (i.i.d.) settings. Cattaneo *et al.* (2020) formalize these results further by

* **Correspondence to:** Departamento de Análisis Económico, Universidad de Zaragoza, Gran Vía 2, 50005 Zaragoza, Spain.
Email: joseolmo@unizar.es; J.B.Olmo@soton.ac.uk

studying their large sample properties. These authors develop pointwise inference methods based on undersmoothing and robust bias correction, and present uniform distributional approximations for the corresponding t-statistic processes.

One major advantage of parametric time series models compared to nonparametric methods is the possibility of testing for forecast ability using standard parametric tests. This is particularly the case in traditional ARIMA type models in which the predictive ability of a model is reflected in the statistical significance of a set of autoregressive parameters. More generally, for time series regression models involving a set of predictive regressors, testing for forecast ability is technically challenging when the regressors are persistent. This has been covered in the work of Campbell and Yogo (2006), Jansson and Moreira (2006), Lewellen (2004), and Stambaugh (1999), among others. The statistical estimation and testing of such models requires a treatment beyond the traditional normal approximations for the regression parameters. The nonparametric literature has also made some progress on this direction. Juhl (2014) develops a nonparametric test of predictive ability based on a kernel estimator of the predictive regression model that works in general time series settings. This author develops a test for the significance of a regressor without specifying a functional form. The results are used to test the null hypothesis that the entire predictive function takes the value of zero.

The aim of the current article is to propose a nonparametric predictive regression model based on partitioning estimators in a stationary setting. The theory on partitioning estimators accommodates power series and different types of splines as basis functions. We propose an alternative approach in which the unknown function modeling the predictive relationship between the variables is approximated using polynomial Taylor expansions applied over disjoint intervals covering the support of the predictor variable. This choice of basis function is conceptually and empirically superior to power series and splines. Conceptually, for analytic functions (a function that is locally given by a convergent power series) the approximation offered by the Taylor expansion converges to the true predictive function as the order of the Taylor polynomial increases. For the remaining class of functions, the Taylor expansion provides a reliable local approximation that improves as we consider thinner partitions of the support of the predictor. Importantly, the regression coefficients associated to the Taylor expansion are interpreted as derivatives of the unknown predictive function evaluated at different knots of the partition. This is not the case for spline methods. In doing so, we obtain a sample of estimates of high-order derivatives of the unknown function and not only of the predictive function of interest. Empirically, our choice of partitioning estimator requires a lower polynomial order to achieve the same fit than spline methods. This is by construction of the Taylor polynomial as a local expansion around the knots.[1]

We derive the asymptotic theory for our class of partitioning estimators in predictive regression models. We adapt the results in Cattaneo and Farrell (2013) and present the uniform convergence of the partitioning estimator under persistence of the predictor variable. We also obtain the asymptotic normality of the estimator that is extended to the functional space, where we derive the weak convergence of the estimator to a centered Gaussian process, see also Cattaneo *et al.* (2020) for a recent contribution on this area. These asymptotic results allow us to propose pointwise and uniform tests of predictive ability. The main advantage of the proposed procedure is its flexibility to test for the predictability of a regressor over the entire support of the random variable. This method provides an alternative to the kernel-based method proposed in Juhl (2014) for predictive regression models. In contrast to this author, the asymptotic theory of our approach does not rely on U-statistics but on the distribution of the supremum of a functional process. This distribution is nonstandard and critical values cannot be tabulated. Nevertheless, we approximate these critical values using the simulation methods in Hansen (1996) and Cattaneo *et al.* (2020) adapted to our setting.

These results are illustrated empirically in a controlled simulation experiment and also in an empirical application. The simulation experiment studies the predictive ability of our partitioning estimator that is compared against the predictive ability of the OLS estimator of a parametric AR(1) model for several data generating processes (DGPs) in terms of mean square prediction error and predictive accuracy. The model comparison

[1] Penalized spline methods can provide better fit than our partitioning estimator at the expense of imposing further regularity conditions on the penalty function, more algebra and more convoluted asymptotic properties. As an extension of the current method, we could propose penalized partitioning estimators. This is, however, left for future research.

DOI: 10.1111/jtsa.12668

is done in sample and out of sample. For the in-sample exercise, our empirical results confirm the suitability of the partitioning estimator in general settings. This estimation strategy outperforms the parametric OLS estimator when the DGPs are nonlinear and is comparable to the OLS estimator when the DGP is a parametric AR(1) process. For the out-of-sample evaluation period the only instance in which the OLS estimator outperforms the nonparametric partitioning estimator is when the DGP is indeed a linear autoregressive process.

The simulation exercise also studies the empirical coverage of the forecast intervals associated to the partitioning estimator. We obtain empirical coverage rates close to the nominal ones across sample sizes and DGPs providing further support to the asymptotic normality of the partitioning estimator. Our simulation results also provide empirical support to the p-value transformation for obtaining critical values for the uniform predictive ability test that is applied over the entire compact support of the predictor variable but also over compact subsets. The results provide satisfactory values of the empirical size and power estimates that are close to one in many instances.

The empirical section applies this methodology to model the dynamics of the realized volatility and bipower variation measures constructed from five-minute returns for the constituents of the Dow Jones Industrial Average index. Using the dataset in Bollerslev *et al.* (2016), we find strong empirical evidence of predictability for both measures in sample and out of sample. In both settings, the partitioning estimator provides superior one-period-ahead forecasts of the conditional volatility than the linear AR(1) model for most stocks. The estimates of the partitioning estimator also reveal strong nonlinearities on the autoregressive function over the support of the outcome variable.

The rest of the article is structured as follows. Section 2 introduces the model assumptions and the partitioning estimator. Section 3 presents results on asymptotic convergence and normality of the estimators, and the extension to the functional space. In Section 4, we derive pointwise and uniform predictive ability tests and an algorithm for the practical implementation of the latter. Section 5 presents a simulation exercise to evaluate different features of the partitioning estimator and predictability tests for autoregressive processes in finite samples. Section 6 contains the empirical application assessing the forecast ability of our methodology for different realized volatility measures for the constituents of the Dow Jones index, and Section 7 concludes. A separate online appendix contains additional simulations for general nonlinear predictive regression models and the mathematical proofs with the main results of the article. Tables and figures are collected at the end of this document.

Throughout the text, we use $\|A\| = \left(\sum_{r=1}^{q}\sum_{s=1}^{q} a_{rs}^2\right)^{1/2}$ to denote the Frobenious norm of a $q \times q$ matrix A, and $\|a\|_2 = \left(\sum_{r=1}^{q} a_r^2\right)^{1/2}$ to denote the L_2 norm for a vector a of dimension q. Similarly, $\|a\|_\infty = \max_{r=1,\dots,q} |a_r|$ to denote the corresponding L_∞ norm. For a function $h(\cdot)$, let $\|h\|_p^p = E[|h(y)|^p]$ and $\|h\|_\infty = \sup_{y\in\chi} |h(y)|$ denote the L_p and L_∞ norms respectively. $\xrightarrow{p}$ denotes convergence in probability, $\xrightarrow{d}$ denotes convergence in distribution and $\xrightarrow{w}$ denotes weak convergence.

2. ECONOMETRIC THEORY

2.1. The Model

To motivate our partitioning estimator we introduce the following nonparametric predictive regression model:

$$y_t = g(x_{t-1}) + \varepsilon_t, \tag{2.1}$$

where x_{t-1} is a predictor variable observed at time $t-1$ and ε_t an error term. The function $g(x_{t-1})$ captures the relationship between y_t and x_{t-1}. We impose the following assumptions on the DGP.

DOI: 10.1111/jtsa.12668

Assumption 1. (a) The predictor variable $\{x_t\}_{t=-\infty}^{\infty}$ is a β-mixing stationary process with mixing coefficient $\beta(n)$ satisfying $\sum_{n=1}^{\infty} n^2 \beta^{\delta/(1+\delta)}(n) < \infty$, for some $\delta > 0$. (b) $E[|x_t|^{2+\delta}] < \infty$ for some $\delta > 0$. (c) The random variable $x_t \in \chi$ is continuously distributed with Lebesgue density $f(x)$, that is bounded and bounded away from zero on $\chi \subset \mathbb{R}$, with χ a compact interval in the real line. (d) The random variable ε_t is a martingale difference sequence with respect to x_{t-1} such that $E[\varepsilon_t \mid x_{t-1} = x] = 0$ and $E[\varepsilon_t^{2+\eta} \mid x_{t-1} = x] < \infty$ for $x \in \chi$ and some constant $\eta > 0$. (e) The function $g(x)$ is $(q+1)$-times continuously differentiable on (and extension of) χ, with $q > 0$ fixed.

Assumption 1(a), (b) and (d) allows us to extend the results in Cattaneo and Farrell (2013) from an i.i.d. context to a time series setting. The β-mixing condition on x_t is necessary in addition to the martingale difference assumption on the error term to accommodate time series models with persistent predictors. This assumption guarantees the consistency of the sample estimator $\frac{1}{T}\sum_{s,t=1}^{T} x_t x_s$ to $E[x_t^2]$ as $T \to \infty$. This assumption can be relaxed in nonparametric autoregressive processes in which x_{t-1} is replaced by lags of the dependent variable. In these cases, Assumption 1(d) and (e) are sufficient for the time series model to be correctly specified. Assumption 1(c) guarantees the existence of nonempty intervals across the partition of the support of the predictor variable. The second part of the assumption is required for tractability purposes. This condition requires the support of the predictor variable to be defined on a compact space. This assumption can be relaxed in empirical applications by assuming that the probability outside the compact support is negligible and only affects the estimation of the unknown function $g(x)$ in the far tails of the distribution of the predictor variable. Under Assumption 1(d), the error term of the predictive regression is a martingale difference sequence with respect to x_{t-1}, which implies that x_{t-1} and ε_t are uncorrelated and that $E[\varepsilon_t] = 0$. This condition also entails the identifiability condition $g(x) = E[y_t \mid x_{t-1} = x]$. Assumption 1(e) extends classical smoothness conditions on nonparametric models by imposing that $g(x)$ is differentiable up to order $q+1$. This assumption allows us to approximate the unknown function $g(x)$ using Taylor expansions of order q over disjoint intervals covering the support of the predictor variable.

The partitioning scheme is as follows. The choice of a single predictive regressor allows us to operate in the real line such that the compact set χ is given by a closed interval $[a, b] \subset \mathbb{R}$. This interval is partitioned into K disjoint intervals $[a, z_1 + h_1)$, $[z_k - h_k, z_k + h_k)$ and $[z_K - h_K, b]$, such that $z_{k-1} + h_{k-1} = z_k - h_k$ for $k = 2, \ldots, K-1$. This partition is characterized by the vector $(h_1, \ldots, h_K)$ that determines the width of the intervals and the knots $\{z_1, \ldots, z_K\}$. These knots are a sequence of nondecreasing real numbers $(z_k \leq z_{k+1})$ that are constructed as the quantiles from the empirical distribution of the predictive regressor. Quantile knots guarantee that an equal number of sample observations lie in each interval while the intervals have different lengths. In practice, the partition is constructed as follows; let $x_{[t]}$ denote the increasing order statistics of the stationary sequence x_t, for $t = 1, \ldots, T$ such that $x_{[1]} < \cdots < x_{[T]}$. The compact set $[a, b]$ is proxied by $[x_{[1]}, x_{[T]}]$ and the set of disjoint intervals is given by $[x_{[1]}, x_{[n]}) \bigcup_{k=2}^{K-1} [x_{[(k-1)*n]}, x_{[k*n]}) \cup [x_{[(K-1)*n]}, x_{[T]}]$, with $T = nK$. The knots of the partition are defined as $z_1 = (x_{[1]} + x_{[n]})/2$ and $z_k = (x_{[(k-1)*n]} + x_{[k*n]})/2$, and the adaptive tuning parameters are $h_1 = (x_{[n]} - x_{[1]})/2$ and $h_k = (x_{[k*n]} - x_{[(k-1)*n]})/2$, for $k = 2, \ldots, K$.

Let $[z_k - h_k, z_k + h_k)$ be a generic interval of the partition and, for $x \in [a, b]$, let $d_k(x)$ be the indicator function such that $d_k(x) = 1$ if x belongs to the interval and zero, otherwise. Our modeling strategy is to apply a Taylor expansion of order $q \geq 0$ to $g(x_{t-1})$ around the different knots of the partition. Let $z_{\bar{k}}$ be such that $d_{\bar{k}}(x_{t-1}) = 1$, then

$$g(x_{t-1}) = \sum_{m=0}^{q} \frac{1}{m!} g^{(m)}(z_{\bar{k}})(x_{t-1} - z_{\bar{k}})^m + R(x_{t-1}, z_{\bar{k}}), \tag{2.2}$$

where $g^{(m)}(z_{\bar{k}})$ denotes the mth-derivative of $g(\cdot)$ evaluated at $z_{\bar{k}}$; $g^{(0)}(z_{\bar{k}}) = g(z_{\bar{k}})$ and $R(x_{t-1}, z_{\bar{k}}) = g^{(q+1)}(c_{\bar{k}})(x_{t-1} - z_{\bar{k}})^{q+1}$ is the remainder of the Taylor expansion, with $c_{\bar{k}} \in (z_{\bar{k}} - h_{\bar{k}}, z_{\bar{k}} + h_{\bar{k}})$. Using the Taylor expansion in (2.2), we denote the coefficients of the polynomial expansion as $\gamma_{km} = \frac{1}{m!} g^{(m)}(z_k)$, for $k = 1, \ldots, K$

DOI: 10.1111/jtsa.12668

and $m = 0, 1, \ldots, q$ such that model (2.1) can be expressed as

$$y_t = \sum_{k=1}^{K}\sum_{m=0}^{q}\gamma_{km}(x_{t-1} - z_k)^m d_k(x_{t-1}) + \tilde{\varepsilon}_t(x_{t-1}), \tag{2.3}$$

where $\tilde{\varepsilon}_t(x_{t-1}) = \varepsilon_t + \overline{R}(x_{t-1})$ is the error term obtained from aggregating the remainder of the Taylor expansions evaluated at the different intervals; $\overline{R}(x_{t-1}) = \sum_{k=1}^{K} R(x_{t-1}, z_k)d_k(x_{t-1})$.

Alternative approximations of the predictive regression model (2.1) can be obtained by applying power series expansions and spline methods over the intervals of the partition. A simple example given by a q-order spline is

$$y_t = \sum_{k=1}^{K}\sum_{m=0}^{q}\gamma_{km}^s(x_{t-1} - z_k)_+^m + \varepsilon_t^s, \tag{2.4}$$

where $(x_{t-1} - z_k)_+ = \max(x_{t-1} - z_k, 0)$, γ_{km}^s is a set of regression coefficients associated to the basis functions and ε_t^s is the corresponding error term. More sophisticated methods such as B-splines and penalized splines are also available in the literature. We, nevertheless, focus on the basis functions obtained from the Taylor approximation in (2.2). These functions are theoretically motivated and provide additional information, compared to power series and spline methods, on the derivatives of the unknown function $g(x)$ evaluated at the knots of the partition.

We discard the residual term $\overline{R}(x_{t-1})$ in (2.3) for estimation purposes, and consider the predictive regression model

$$Y = \sum_{k=1}^{K}\mathbb{X}_k\Gamma_k + \varepsilon, \tag{2.5}$$

where $Y = (y_1, \ldots, y_T)'$, $\mathbb{X}_k$ is a $T \times (q+1)$ matrix with rows $\mathbb{X}_{k,t} = (1, x_t - z_k, \ldots, (x_t - z_k)^q)d_k(x_t)$, for $t = 0, \ldots, T-1$, $\Gamma_k = (\gamma_{k0}, \ldots, \gamma_{kq})'$ and $\varepsilon = (\varepsilon_1, \ldots, \varepsilon_T)'$.

2.2. Estimation

Our partitioning estimator is obtained from applying ordinary least squares to the local polynomials forming the Taylor expansion in each interval of the partition of the compact set. As mentioned above, the partitioning estimator may be recast as a series estimator such that

$$\widehat{\Gamma}_k = \widehat{Q}_k^{-1}\left(\frac{1}{Tp_k}\sum_{t=1}^{T}\mathbb{X}_{k,t-1}'y_t\right) \tag{2.6}$$

with $\widehat{Q}_k = \frac{1}{Tp_k}\sum_{t=1}^{T}\mathbb{X}_{k,t-1}'\mathbb{X}_{k,t-1}$ and $p_k = E[d_k(x)] = P\{x \in [z_k - h_k, z_k + h_k)\}$. For partitions given by equivalent blocks, this probability is constant across intervals and can be estimated as n/T, with n the number of observations in each interval and such that $n/T \to 0$ as $n, T \to \infty$. In practice, estimation of the probability p_k is not required as this quantity cancels out in the definition of the estimator of Γ_k.

The model predictions $\widehat{y}_t$ obtained from a realization x_{t-1} of the predictor variable are constructed as $\widehat{g}(x_{t-1}) = \sum_{k=1}^{K}\mathbb{X}_{k,t-1}\widehat{\Gamma}_k = \mathbb{X}_{\overline{k},t-1}\widehat{\Gamma}_{\overline{k}}$, with $\overline{k}$ denoting the knot $z_{\overline{k}}$ such that $d_{\overline{k}}(x_{t-1}) = 1$. This result can be extended to obtain

DOI: 10.1111/jtsa.12668

predictions for any value $x \in \chi$ as

$$\widehat{g}(x) = \sum_{k=1}^{K}\sum_{m=0}^{q}\widehat{\gamma}_{km}(x - z_k)^m d_k(x) = \sum_{k=1}^{K} v_k(x)'\widehat{\Gamma}_k \tag{2.7}$$

with $v_k(x) = d_k(x)(1, (x - z_k), (x - z_k)^2, \ldots, (x - z_k)^q)'$. Estimates of the derivatives of $g(x)$ are obtained from this expression as

$$\widehat{g}^{(m)}(x) = \sum_{k=1}^{K} v_k^{(m)}(x)'\widehat{\Gamma}_k,$$

where $v_k^{(m)}(x)$ is the mth-order derivative of the vector $v_k(x)$.

To be able to make inference about the model predictions $\widehat{g}(x)$, for $x \in \chi$, we need to obtain reliable measures of the standard error of the parameter estimates. Let

$$\widehat{V}(\widehat{\Gamma}_k) = \frac{1}{Tp_k}\widehat{Q}_k^{-1}\widehat{\Psi}_k\widehat{Q}_k^{-1} \tag{2.8}$$

with $\widehat{\Psi}_k = \frac{1}{Tp_k}\sum_{t=1}^{T}\mathbb{X}'_{k,t-1}\mathbb{X}_{k,t-1}e_t^2$, be an estimator of the variance of the parameter estimator $\widehat{\Gamma}_k$ that accommodates the presence of conditional heteroscedasticity. The corresponding population counterparts of the above sample covariance matrices are $Q_k = E[\mathbb{X}'_{k,t-1}\mathbb{X}_{k,t-1}]/p_k$ and $\Psi_k = E[\mathbb{X}'_{k,t-1}\mathbb{X}_{k,t-1}\epsilon_t^2]/p_k$ respectively.

2.3. Model Selection

An important aspect of the partitioning estimator developed herein is the choice of the intervals defining the partition of the compact set. There is a bias-variance trade-off between the number of observations in each interval and the number of intervals covering the compact set χ. Increasing the number of intervals reduces the bias of the approximation at the expense of increasing the variance of the parameter estimators. The constant n denoting the number of observations in each interval is considered a tuning parameter that is endogenously selected within the model.

The practical choice of the order of the Taylor expansion q is also of relevance. As discussed above, the approximation offered by the Taylor expansion improves for higher orders if the Taylor polynomial is convergent (analytic functions). Otherwise, low orders of the Taylor expansion can perform as well as higher orders, in fact, high orders may lead to the overfit of the unknown function and may not be desirable for forecasting purposes. Similar findings are observed for power series and splines; for instance, for splines, the literature usually suggests the choice of a cubic model beyond which the in-sample fit does not generally improve performance. For these reasons, we consider the choice of the number of terms in the Taylor expansion as another tuning parameter. To take explicit account of this choice, we will perform the model selection exercise in sample and out of sample.

We propose several model selection mechanisms given by the mean square error (MSE), time series methods such as the Akaike and Bayesian information criteria (BIC), and nonparametric methods for bandwidth selection, see Mallows (1973), Li (1987) and Wahba (1985). We adapt these criteria to choose the pair $(\widehat{K}, \widehat{q})$ that determines the number of regressors $\tilde{K} = \widehat{K}(\widehat{q}+1)$ in the regression model (2.5). By doing so, our procedure implicitly selects n optimally, as $n_{opt} = T/\widehat{K}$. Thus, the AIC procedure to optimally select K and q is

$$\{\widehat{K}_{AIC}, q_{AIC}\} = \arg\min_{\{K,q\}}\left\{\ln\widehat{\sigma}_\varepsilon^2 + 2\frac{(q+1)K+1}{T}\right\}, \tag{2.9}$$

DOI: 10.1111/jtsa.12668

where $\widehat{\sigma}_{\epsilon}^2$ is the standard error of the nonparametric regression model under homoscedasticity of the error term. The BIC is

$$\{\widehat{K}_{BIC}, q_{BIC}\} = \underset{\{K,q\}}{\arg\min} \left\{ \ln \widehat{\sigma}_{\epsilon}^2 + \frac{((q+1)K+1)ln\ T}{T} \right\}. \tag{2.10}$$

Similarly, we implement Mallows (1973) procedure to select K and q such that

$$\{\widehat{K}_{M}, q_{M}\} = \underset{\{K,q\}}{\arg\min} \left\{ \widehat{\sigma}_{e}^2 (1 + 2K/T) \right\}. \tag{2.11}$$

Craven and Wahba (1978) propose a generalized cross-validation method[2] that we apply here to select K and q:

$$\{\widehat{K}_{CV}, q_{CV}\} = \underset{\{K,q\}}{\arg\min} \left\{ \frac{\widehat{\sigma}_{e}^2}{(1 - 2K/T)^2} \right\}. \tag{2.12}$$

The performance of these methods will be explored in a simulation exercise below.

3. ASYMPTOTIC CONVERGENCE

This section presents convergence results for the approximating function $\widehat{g}(x)$ in (2.7) as well as results necessary to make asymptotic inference on the pointwise predictions. The section also explores uniform approximations and convergence results when the estimator is considered a process in $x \in \chi$. First, we introduce the following assumption that introduces some smoothness conditions and suitable convergence rates between the tuning parameters h_k, p_k, K, and T.

Assumption 2. (a) $\sigma^2(x) = V(y_t \mid x_{t-1} = x)$ is continuous and bounded away from zero for $x \in \chi$.

(b) Q_k and Ψ_k are $(q+1) \times (q+1)$ positive definite matrices, for q fixed and $k = 1, \ldots, K$.

(c) Let $\overline{h} = \max_{\{k=1,\ldots,K\}} \{h_k\}$, then $\overline{h} \to 0$ and $T\overline{h} \to \infty$. Furthermore, we assume $h_k \asymp K^{-1}$ and $p_k \asymp K^{-1}$, with K the number of intervals of the partition and such that for scalars a and b, $a \asymp b$ denotes that $C_* b \leq a \leq C^* b$ for positive constants C_* and C^*. Similarly, we assume $TK^{-2(q+1)} \to 0$ as $K, T \to \infty$.

The following auxiliary result shows the asymptotic convergence of the sample covariance estimators introduced above.

Lemma 1. Under Assumptions 1 and 2, we have $\|Q_k\| = O(1)$ and $\|\widehat{Q}_k - Q_k\| = o_p(1)$, for every $k = 1, \ldots, K$, as $T \to \infty$.

The following result studies the asymptotic convergence of the partitioning estimator (2.6).

Proposition 1. Under Assumptions 1 and 2, it follows that $\|\widehat{\Gamma}_k - \Gamma_k\| = O_p(\sqrt{K/T})$, for $k = 1, \ldots, K$, as $K, T \to \infty$.

This result illustrates the nonparametric character of the partitioning estimator. The convergence of the estimator is at a nonparametric rate due to the partition of the compact set χ into K disjoint intervals. These results also allow us to derive the uniform convergence of the estimator of the functional coefficient.

[2] Other more sophisticated model selection procedures for series estimators can be found in the literature, for example, the leave-one-out cross-validation method of Stone (1974). More recently, Györfi *et al.* (2002), Cattaneo and Farrell (2013) and Cattaneo *et al.* (2020) explore cross-validation and plug-in methods for partitioning estimators obtained from minimizing integrated mean square error measures.

DOI: 10.1111/jtsa.12668

Proposition 2. If Assumption 1 holds, with $0/0 = 0$, then

$$\sup_{y\in\chi} |\widehat{g}(x) - g(x)| = O_p\left(\sqrt{K/T} + K^{-(q+1)}\right). \tag{3.1}$$

The order $O_p\left(\sqrt{K/T}\right)$ is due to the estimation of the vector Γ_k and the order $O_p\left(K^{-(q+1)}\right)$ is due to the approximation to the true function by the Taylor expansion over intervals of width $O(K^{-1})$.

The following result establishes a Bahadur type representation for $\widehat{g}(x) - g(x)$. The estimator of the unknown function may be represented as an average of serially uncorrelated, zero-mean random variables forming a triangular array based on certain smoothing weights plus a remainder $v_T(x)$ that is a function of the aggregate residual term $\overline{R}(x)$. This is possible by assumption 1(e) that guarantees that the error term ε_t is a martingale difference sequence.

Lemma 2. Under Assumptions 1 and 2, the Bahadur representation of the partitioning estimator is

$$\widehat{g}(x) - g(x) = \frac{1}{T}\sum_{t=1}^{T} \Phi_T(x, x_{t-1})\varepsilon_t + v_T(x) + o_p(1), \tag{3.2}$$

where $\Phi_T(x, x_{t-1}) = \sum_{k=1}^{K} v_k(x)' Q_k^{-1} \mathbb{X}'_{k,t-1}/p_k$ and $v_T(x) = \frac{1}{Tp_k}\sum_{t=1}^{T}\sum_{k=1}^{K} v_k(x)' Q_k^{-1} \mathbb{X}'_{k,t-1} R(x_{t-1}, z_k)$ is the remainder term. Furthermore, under the above assumptions, it follows that $|v_T(x)| = o_p(1)$, for $x \in \chi$.

We introduce further notation to formulate the asymptotic distribution of the Bahadur representation. Let $V_0(x) = E[\Phi^2(x)\sigma^2(x)]$ and $V_T(x) = E[\Phi_T^2(x, x_{t-1})\varepsilon_t^2]$ that, under assumption 2(a), can be written as $V_T(x) = E[\Phi_T^2(x, x_{t-1})\sigma^2(x)] = \sum_{k=1}^{K} v_k(x)' Q_k^{-1} \Psi_k Q_k^{-1} v_k(x)/p_k$. The empirical counterpart of V_T is $\widehat{V}_T(x) = \frac{1}{T}\sum_{t=1}^{T} \widehat{\Phi}_T^2(x, x_{t-1}) e_t^2 = \sum_{k=1}^{K} v_k(x)' \widehat{Q}_k^{-1} \widehat{\Psi}_k \widehat{Q}_k^{-1} v_k(x)/p_k$. The estimator $\widehat{V}_T(x)$ does not require knowledge of the probability p_k. This expression cancels out when combined with the estimators $\widehat{Q}_k$ and $\widehat{\Psi}_k$ such that $\widehat{V}_T(x)$ is a feasible estimator.

Lemma 3. Under Assumptions 1 and 2, we have $\|\widehat{\Psi}_k - \Psi_k\| = o_p(1)$, for every $k = 1, \ldots, K$, as $T \to \infty$.

Proposition 3. If Assumptions 1 and 2 hold, for any $x \in \chi$ fixed, $V_T(x) - V_0(x) \xrightarrow{p} 0$ and $\widehat{V}_T(x) - V_T(x) \xrightarrow{p} 0$ as $T \to \infty$, with $V_T(x) = O(K)$.

The following result states the asymptotic normality of the model forecasts.

Theorem 1. Under Assumptions 1 and 2, and any $x \in \chi$ fixed, it follows that

$$\sqrt{T}\frac{\widehat{g}(x) - g(x)}{V_T^{1/2}(x)} \xrightarrow{d} N(0, 1). \tag{3.3}$$

These results can be extended to the functional space if $\widehat{g}(x)$ is considered a process in $x \in \chi$. Unfortunately, the stochastic process $\widehat{g}(x)$ is not asymptotically tight and, therefore, does not converge weakly in $\mathcal{L}^\infty$, where $\mathcal{L}^\infty$ denotes the set of all uniformly bounded real functions on χ equipped with the uniform norm. Nevertheless, we adapt the results in Cattaneo *et al.* (2020) to our context and construct Gaussian processes that approximate the finite-sample distribution of $\sqrt{T}\frac{\widehat{g}(x)-g(x)}{V_T^{1/2}(x)}$. More formally,

DOI: 10.1111/jtsa.12668

Lemma 4. Let Assumptions 1 and 2 hold, and assume that (*i*) $T\bar{h} = o(r_T^{-2})$, with r_T some nonvanishing positive sequence and (*ii*) $\sup_{x\in\chi} E[|\varepsilon_t|^{2+\eta} \mid x_{t-1} = x] < \infty$ and $\frac{T^{2/(2+\eta)}\ (logT)^{\frac{2+2\eta}{2+\eta}}}{T\bar{h}} = o(r_T^{-2})$. Then,

$$\sup_{x\in\chi}\left|\sqrt{T}\frac{\widehat{g}(x) - g(x)}{V_T^{1/2}(x)} - \mathbb{G}_T(x)\right| = o_P(r_T^{-1}) \tag{3.4}$$

with $\mathbb{G}_T(x) = \frac{\sum_{t=1}^{T}\Phi_T(x,x_{t-1})\varepsilon_t}{\sqrt{TV_T(x)}}$.

The proof of this result follows from a direct application of Lemma 6.1 in Cattaneo *et al.* (2020) and is omitted for space constraints. The definition of $\mathbb{G}_T(x)$ allows us to obtain a distributional approximation for these stochastic processes by a sequence of centered Gaussian processes $\mathbb{G}(x)$ in $\mathcal{L}^\infty$. To do this, we introduce the following assumption.

Assumption 3. In a sufficiently rich probability space, for each $x \in \chi$, there exists a copy $\mathbb{G}_T^*(x)$ of $\mathbb{G}_T(x)$ and a Normal random variable ε_t^* following a $N(0, 1)$ distribution such that

$$\sup_{x\in\chi}|\mathbb{G}_T^*(x) - \mathbb{G}(x)| = o_P(r_T^{-1}), \tag{3.5}$$

where $\mathbb{G}_T^*(x) = \frac{\sum_{t=1}^{T}\Phi_T(x,x_{t-1})\varepsilon_t^*}{\sqrt{TV_T(x)}}$ and $\mathbb{G}(x)$ is a sequence of centered Gaussian processes in $\mathcal{L}^\infty$.

Under Assumption 3, the approximation satisfies that

$$\sqrt{T}\frac{\widehat{g}(x) - g(x)}{V_T^{1/2}(x)} \xrightarrow{w} \mathcal{G}(x) \tag{3.6}$$

in $\mathcal{L}^\infty$, with $\mathcal{G}(x)$ denoting the probability law of the Gaussian process $\mathbb{G}(x)$. A formal proof of this result follows from applying the novel two-step coupling approach in Cattaneo *et al.* (2020) and is beyond the scope of this article. This result can be applied for developing hypothesis tests over the support of the predictor variable or compact subsets of it. To do this, we derive first the asymptotic distribution of the supremum functional.

Theorem 2. Under Assumptions 1–3 and the conditions of Lemma 4, for $r_T \to \infty$, it follows that

$$\sqrt{T}\sup_{x\in\chi}\left|\frac{\widehat{g}(x) - g(x)}{V_T^{1/2}(x)}\right| \xrightarrow{d} \sup_{x\in\chi}|\mathcal{G}(x)|\,. \tag{3.7}$$

The proof of this result follows from applying the continuous mapping theorem to the supremum of the process on the left hand side of expression (3.6).

4. PREDICTIVE ABILITY TEST

The presence of predictive ability at specific points of the compact support of the predictor variable can be tested using *t*-tests for pointwise predictions. The predictive ability at a given point $x_{t-1} = x$ is given by a value of $g(x)$ different from zero. Therefore, for $x_{t-1} \in \chi$, we define the null hypothesis of absence of predictive ability as $H_0 : g(x_{t-1}) = 0$ and the alternative hypothesis as $H_A : g(x_{t-1}) \neq 0$. A suitable test for this hypothesis is

DOI: 10.1111/jtsa.12668

$$t_\alpha = \sqrt{T}\frac{\widehat{g}(x_{t-1})}{\widehat{V}_T^{1/2}(x_{t-1})}, \tag{4.1}$$

that, under the null hypothesis, converges in distribution to $N(0,1)$. This result is immediate from the application of Theorem 1. Similarly, we obtain asymptotically valid pointwise prediction intervals for $g(x)$ as

$$g(x_{t-1}) \in \widehat{g}(x_{t-1}) \pm z_{1-\alpha/2}\widehat{V}_T^{1/2}(x_{t-1})/\sqrt{T}, \tag{4.2}$$

where $z_{1-\alpha/2}$ is the critical value obtained from the asymptotic approximation to a Normal distribution.

More interesting is the extension of the test to the functional space. The hypothesis of interest is $H_0 : g(x) = 0$, almost everywhere in χ, against the alternative $H_A : g(x) \neq 0$, for some $x \in \chi$. A suitable test statistic is

$$D_T = \sqrt{T}\sup_{x\in\chi}\left|\frac{\widehat{g}(x)}{\widehat{V}_T^{1/2}(x)}\right|. \tag{4.3}$$

The asymptotic distribution of this statistic, obtained as a direct application of Theorem 2 and given by

$$D_T \xrightarrow{d} \sup_{x\in\chi}|\mathcal{G}(x)|, \tag{4.4}$$

depends on nuisance parameters given by the covariance kernel of the Gaussian processes and cannot be universally tabulated. This is the well known Davies (1977, 1987) problem of hypothesis tests under the presence of nuisance parameters. Hypothesis tests involving nuisance parameters under the null have been widely investigated in the time series literature and, in particular, in threshold models and structural break testing.

Obtaining asymptotic critical values for this test is difficult. Fortunately, simulation and resampling methods can be applied to approximate the critical values in finite samples, see Andrews (1993) and Hansen (1996). More recently, Cattaneo *et al.* (2020) propose a simulation-based method to implement uniform inference on the partitioning estimator for i.i.d. data. In what follows, we apply a simulation procedure in the same spirit of these authors. We operate conditionally on a realization of $\{y_t, x_{t-1}\}_{t=1}^T$ and present a simple plug-in approach to approximate the infeasible Gaussian processes $\mathbb{G}(x)$. Lemma 4 implies that

$$\sup_{x\in\chi}\left|\sqrt{T}\frac{\widehat{g}(x)}{\widehat{V}_T^{1/2}(x)} - \widehat{\mathbb{G}}_T(x)\right| = o_P(r_T^{-1}),$$

where $\widehat{\mathbb{G}}_T(x) = \frac{\sum_{t=1}^T \widehat{\Phi}_T(x,x_{t-1})e_t^{(0)}}{\sqrt{T\widehat{V}_T(x)}}$ and $e_t^{(0)}$ is the vector of residuals of the regression model (2.5) under the null hypothesis H_0 of no predictability such that $e_t^{(0)} = y_t$. The feasible stochastic processes $\widehat{\mathbb{G}}_T(x)$ replace the unfeasible stochastic processes $\mathbb{G}_T(x)$. Similarly, $\widehat{\mathbb{G}}_T^\star(x)$ denote the corresponding i.i.d. replicas.

Let $D_T^\star = \sqrt{T}\sup_{y\in\chi}|\widehat{\mathbb{G}}_T^\star(x)|$ be an independent replica of the test statistic D_T. Under the null hypothesis H_0, the distribution of $D_T^\star$ conditional on $\{y_t, x_{t-1}\}_{t=1}^T$ converges to $\sup_{x\in\chi}|\mathcal{G}(x)|$. This asymptotic distribution can be approximated by generating independent replicas of $\widehat{\mathbb{G}}_T^\star(x) = \frac{\sum_{t=1}^T \widehat{\Phi}_T(x,x_{t-1})e_t^\star}{\sqrt{T\widehat{V}_T(x)}}$. To do this, we generate a vector of i.i.d. $N(0,1)$ random variables $\{u_t\}_{t=1}^T$ to construct the simulated residuals $e_t^\star = e_t^{(0)}u_t$. These residuals have the

DOI: 10.1111/jtsa.12668

same variance of ϵ_t conditional on $x_{t-1} = x$. From assumption 3 above, we have

$$\sup_{x\in\chi}|\widehat{\mathbb{G}}_T^*(x) - \mathbb{G}(x)| = o_P(r_T^{-1}).$$

Similarly, the p-value of the test given by $P_{H_0}\left\{D_T > \sup_{x\in\chi}|\mathcal{G}(x)|\right\}$ can be approximated in finite samples by $P\left\{D_T^\star > D_T \mid \{y_t, x_{t-1}\}_{t=1}^T\right\}$. Although the distribution of $D_T^\star$ is not directly observed, it can be approximated to any degree of accuracy by operating conditionally on $\{y_t, x_{t-1}\}_{t=1}^T$. The algorithm to compute the p-value of the test is described below.

Algorithm.

1. Construct a partition of the compact set $\chi \equiv [a, b]$ using the intermediate statistics of the stationary sequence $\{x_t\}_{t=1}^T$ as discussed above. In particular, choose $a = x_{(1)} < x_{(2)} < \cdots < x_{(T)} = b$ and a value for n such that $K = \lceil T/n \rceil$. The set of disjoint intervals is given by $[x_{[1]}, x_{[n]}) \bigcup_{k=2}^{K-1} [x_{[(k-1)*n]}, x_{[k*n]}) \cup [x_{[(K-1)*n]}, x_{[T]}]$. The Taylor expansions will be evaluated at the center of these intervals with $z_1 = (x_{[1]} + x_{[n]})/2$ and $z_k = (x_{[(k-1)*n]} + x_{[k*n]})/2$. The adaptive tuning parameters are $h_1 = (x_{[n]} - x_{[1]})/2$ and $h_k = (x_{[k*n]} - x_{[(k-1)*n]})/2$, for $k = 2, \ldots, K$.
2. Estimate Γ_k, for $k = 1, \ldots, K$, using the partitioning estimator (2.6) and construct the function $\widehat{g}(x)$ as in (2.7) for $x \in [a, b]$.
3. Construct an equidistant grid of points $\mathbb{Z}_k = \{x_{k1}, \ldots, x_{ki}\}$ covering each interval $[z_k - h_k, z_k + h_k)$, and let $\mathbb{Z} = \{\mathbb{Z}_1, \ldots, \mathbb{Z}_K\}$ denote the full grid covering the interval $[a, b]$. Compute the test statistic $D_T = \sqrt{T}\ \sup_{x\in\mathbb{Z}}\left|\frac{\widehat{g}(x)}{\widehat{V}_T^{1/2}(x)}\right|$.
4. For a given realization $\{y_t, x_{t-1}\}_{t=1}^T$, execute the following steps for $b = 1, \ldots, B$:
   (a) Generate the sequence $\{u_t^{(b)}\}_{t=1}^T$ of i.i.d. $(0, 1)$ random variables independent of the data and construct the simulated residuals $e_t^{\star(b)} = e_t^{(0)} u_t^{(b)}$, with $e_t^{(0)}$ the vector of residuals of the partitioning estimators obtained under the null hypothesis H_0.
   (b) Compute the simulated process $\widehat{\mathbb{G}}_T^{\star(b)}(x) = \frac{\sum_{t=1}^T \widehat{\Phi}_T(x, x_{t-1}) e_t^{\star(b)}}{\sqrt{T\widehat{V}_T(x)}}$, for $x \in \mathbb{Z}$, with $\widehat{\Phi}_T(x, x_{t-1}) = \sum_{k=1}^K v_k(x)' \widehat{Q}_k^{-1} \mathbb{X}'_{k,t-1}/p_k$ and $\widehat{V}_T(x) = \sum_{k=1}^K v_k(x)' \widehat{Q}_k^{-1} \widehat{\Psi}_k \widehat{Q}_k^{-1} v_k(x)/p_k$.
   (c) Store the bootstrap test statistic

$$D_T^{\star(b)} = \sup_{y\in\mathbb{Z}}|\widehat{\mathbb{G}}_T^{\star(b)}(x)|.$$

This algorithm yields a random sample of B observations from the distribution of $\sup_{x\in\chi}|\widehat{\mathbb{G}}_T^\star(x)|$. Using the Glivenko–Cantelli theorem and previous assumptions, the empirical p-value conditional on $\{y_t, x_{t-1}\}_{t=1}^T$ defined by $\widehat{p}_{T,B}^\star = \frac{1}{B}\sum_{b=1}^B 1(D_T^{\star(b)} > D_T)$ converges, in probability, to the bootstrap distribution $P\left\{D_T^\star > D_T \mid \{y_t, x_{t-1}\}_{t=1}^T\right\}$, as $B \to \infty$. As mentioned above, this conditional probability converges to the p-value obtained from the asymptotic distribution of the test statistic D_T as $T \to \infty$.

5. MONTE CARLO SIMULATIONS

This section studies the predictive ability of our partitioning estimator for linear autoregressive processes using different evaluation criteria. The section also studies the coverage probability of the asymptotic interval forecasts in (4.2), the empirical size and power of the t-tests for pointwise predictability and the uniform test for predictability

DOI: 10.1111/jtsa.12668

over the compact support of the predictor variable. The section finishes discussing model selection procedures to optimally determine in sample and out of sample the number of regressors in the nonparametric regression model characterized by the partitioning estimator. The online appendix extends this analysis to the case of linear predictive regression models.

5.1. Simulation Design

The simulation exercise focuses on assessing the performance of the partitioning estimator for autoregressive processes since these models are studied in more detail in the empirical application. The choice of this DGP implies minor adjustments to the above set of assumptions. In particular, it is sufficient to impose the martingale difference assumption to the sequence of errors for the model to be correctly specified. In contrast, the compactness of the support of the predictor variable no longer holds in this context if the distribution of the error is defined over the real line. In this case, we consider an extension of the compact set χ given by $(-\infty, a) \bigcup [a, b] \bigcup (b, \infty)$, such that $P\left\{x \in (-\infty, a) \bigcup (b, \infty)\right\} \leq tol$, with tol some tolerance level very close to zero.

We consider the following DGPs:

$$\text{(i) } y_t = \rho y_{t-1} + \varepsilon_t, \tag{5.1}$$

$$\text{(ii) } y_t = sin(y_{t-1}) y_{t-1} + \varepsilon_t, \tag{5.2}$$

$$\text{(iii) } y_t = cos(y_{t-1}) y_{t-1} + \varepsilon_t, \tag{5.3}$$

where ε_t is a $N(0, 1)$ distribution independent of y_t. The autoregressive coefficient is constant in the first model and given by $\rho = 0.5$ throughout experiments. This coefficient is time varying in the remaining two models and given by the sine and cosine functions. These DGPs are formulated for simulation purposes but knowledge of these parametric forms is not required for estimation or forecasting purposes as this is done nonparametrically using the partitioning estimator. The number of periods T varies between $T = 200, 500, 1000$ depending on the experiment, and the number of replications of the DGPs is $B = 500$ throughout.[3]

We study the performance of the approximation in (2.5) for different choices of n and q. The choice of n not only determines the value of h_k in the intervals $[z_k - h_k, z_k + h_k)$ but also the value of K since $T = nK$. Therefore, by choosing n and q we select the number of regressors in the nonparametric model (2.5).

5.2. Predictive Accuracy of the Partitioning Estimator

We study the predictive accuracy of the approximation $\widehat{g}(y)$ for (i) $g(y_{t-1}) = \rho y_{t-1}$, (ii) $g(y_{t-1}) = sin(y_{t-1}) y_{t-1}$, and (iii) $g(y_{t-1}) = cos(y_{t-1}) y_{t-1}$. The competing models are (a) the parametric AR(1) process given by $\widehat{g}(y_{t-1}) = \widehat{\rho} y_{t-1}$, with $\widehat{\rho}$ estimated using OLS methods, and (b) the partitioning estimator $\widehat{g}(y_{t-1}) = \sum_{k=1}^{K} \mathbb{X}_{k,t-1} \widehat{\Gamma}_k$ introduced in (2.5), with $\mathbb{X}_{k,t-1} = (1, (y_{t-1} - z_k), \ldots, (y_{t-1} - z_k)^q)$.

To compute the root mean square prediction error (RMSPE) of the model over the support of the outcome variable y_t, we create a grid of points $\{y_1, \ldots, y_m\} \in [a, b]$, with a the left end point of the sample and b the right

[3] In an online appendix, we study the predictive performance of the partitioning estimator for similar DGPs. In these models we replace the predictor variable y_{t-1} by an exogenous autoregressive process x_t. We explore the performance of the model for different values of the correlation between the error term and different degrees of persistence of the predictor x_t.

DOI: 10.1111/jtsa.12668

end point. For each y_j in the grid, the RMSPE is calculated as

$$RMSPE(y_j) = \frac{1}{\sqrt{B}} \sqrt{\sum_{b=1}^{B} (\widehat{g}_b(y_j) - g(y_j))^2}, \tag{5.4}$$

where $\widehat{g}_b(y_j) = \widehat{\rho}_b y_j$ for model (a) and $\widehat{g}_b(y_j) = \sum_{k=1}^{K} \mathbb{X}_{k,t-1} \widehat{\Gamma}_{k(b)}$ for model (b); $\widehat{\rho}_b$ and $\widehat{\Gamma}_{k(b)}$ are the parameter estimates obtained from the MC simulations $b = 1, \dots, B$. This loss function can be divided into a component $\sum_{b=1}^{B} (\widehat{g}_b(y_j) - g_b(y_j))^2$ capturing parameter uncertainty (estimation error) and a second component $\sum_{b=1}^{B} (g_b(y_j) - g(y_j))^2$ that reflects the error between the model specification $g_b(y)$ and the true DGP given by $g(y)$. Importantly, if the DGP is known, the functions $g_b(\cdot)$ and $g(\cdot)$ are the same and the second component of the RMSPE vanishes. In our simulation exercise, this case is represented by DGP (i) and model (*a*) above. The latter model is misspecified, though, if the DGP is generated as in (ii) or (iii). However, the nonparametric approach in (*b*) is robust to the choice of DGP.

Figure 1 presents the RMSPE for DGPs (i)–(iii) using the parametric OLS estimator and the partitioning estimator, respectively, and for sample sizes equal to $T = 500$ and $T = 1000$. We report the RMSPE for a grid of points covering 90% of the support of the random variable y_t.[4] To do this, we construct a grid of 100 points covering the compact set $[\overline{y} - 1.64 * \widehat{\sigma}_Y, \overline{y} + 1.64 * \widehat{\sigma}_Y]$, with $\overline{y}$ and $\widehat{\sigma}_Y$ the sample mean and standard deviation of $\{y_t\}_{t=1}^{T_0}$ obtained from a realization of $T_0 = 2000$ observations from the DGP.

The top panels of Figure 1 consider the autoregressive DGP (5.1). The RMSPE obtained from the parametric estimator is represented with a dashed red line whereas the RMSPE of the partitioning estimator is reported with a black dashed line. Unsurprisingly, the results in both panels show considerably smaller values of the RMSPE for the OLS estimator across values of $y \in [a, b]$. The figure shows some perturbations of the partitioning estimator in the left and right end points of the partition. The middle panels report the RMSPE for the DGP (5.2). In this case the OLS estimator is applied to a misspecified AR(1) model whereas the nonparametric partitioning estimator is robust to the functional form of the autoregressive process. The OLS estimator reports very low values of the RMSPE for y in a neighborhood of zero. Outside this neighborhood, the RMSPE of the partitioning estimator is significantly superior. The results for DGP (iii) in (5.3), reported in the bottom panels of Figure 1, are qualitatively similar to those for DGP (ii): the parametric OLS estimator reports low values of the RMSPE in a small neighborhood around zero. This is discussed in more detail below.

A related exercise is to compare the predictive ability of both methods. A popular strategy in the forecasting literature is to implement Diebold and Mariano (1995) (DM) test. In this case we compare the predictions of both models and assess statistically the differences in forecast performance across models using the RMSPE loss function. To do this, we specialize the definition of the RMSPE in (5.4) and differentiate between in-sample and out-of-sample measures that are applied to the forecast errors of the competing models. Figure 2 reports the DM test statistics for the in-sample and out-of-sample periods. To attach a statistical significance to the values in the figures, we should compare these values against 1.96 and -1.96 to determine if the forecasts of the partitioning estimator outperform those of the linear autoregressive model or the other way around respectively. The top panels of Figure 2 present the results for the linear autoregressive model in (5.1). The black solid line takes values around 2, which suggests the outperformance of the partitioning estimator in sample. In contrast, the red dashed line takes negative values around -2, suggesting the opposite for the out-of-sample data when the number of observations is $M = 500$. The latter analysis illustrates the superior forecast performance of the linear AR(1) model out of sample. This result is unsurprising given that the DGP is exactly an AR(1) model. However, in sample, the additional flexibility of the partitioning estimator implies a better fit and a more favorable DM statistic.

[4] The unbounded support of the standard Normal random variable does not allow to extend the grid to the entire support of the outcome variable.

DOI: 10.1111/jtsa.12668

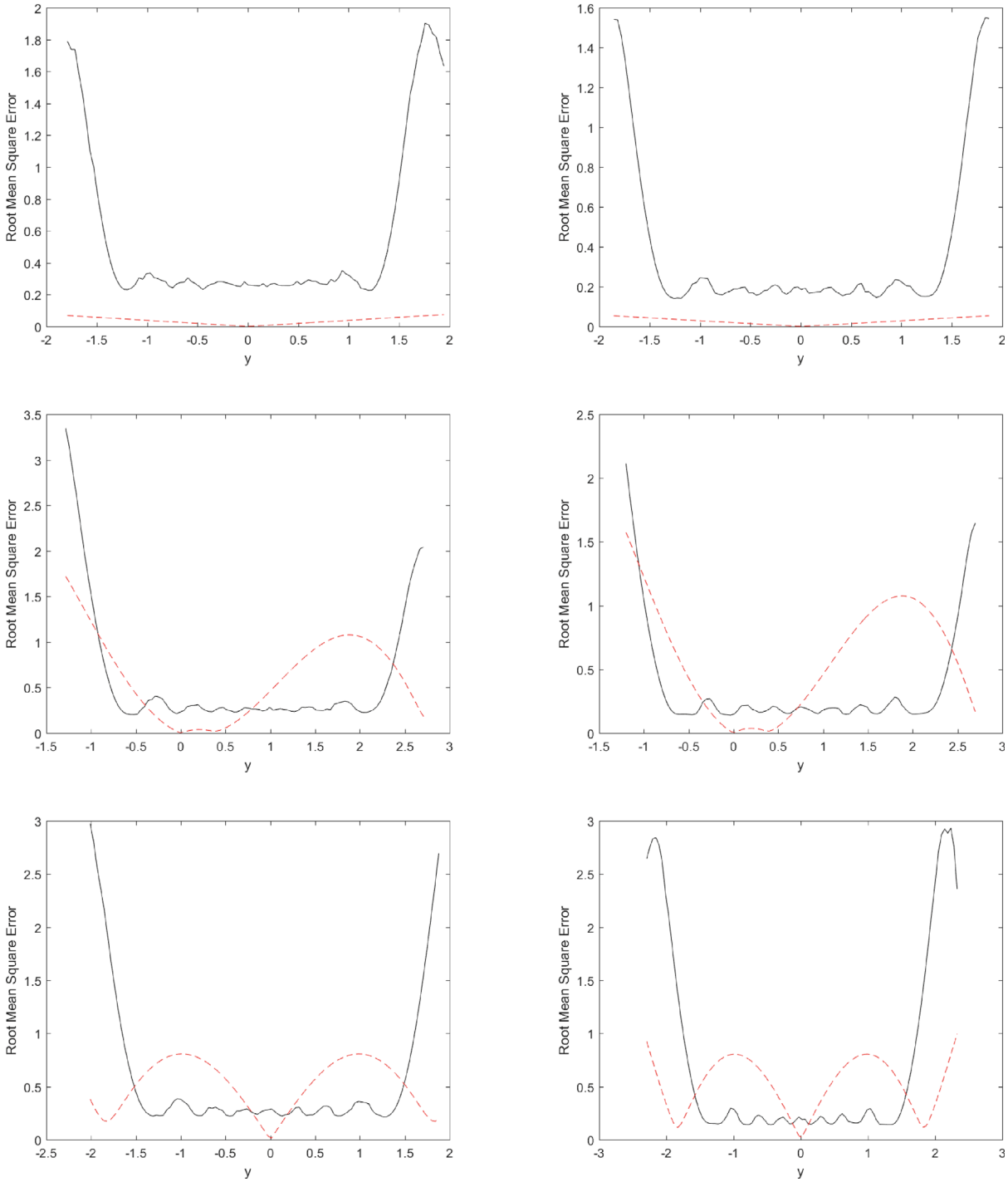

Figure 1. Root mean square prediction error for DGPs (5.1)–(5.3). Left panels correspond to $T = 500$ and right panels to $T = 1000$. The number of observations in each interval $[z_k - h_k, z_k + h_k)$ is $n = 0.1 \times T$. Black solid line for the partitioning estimator and red dashed line for the parametric autoregressive model

DOI: 10.1111/jtsa.12668

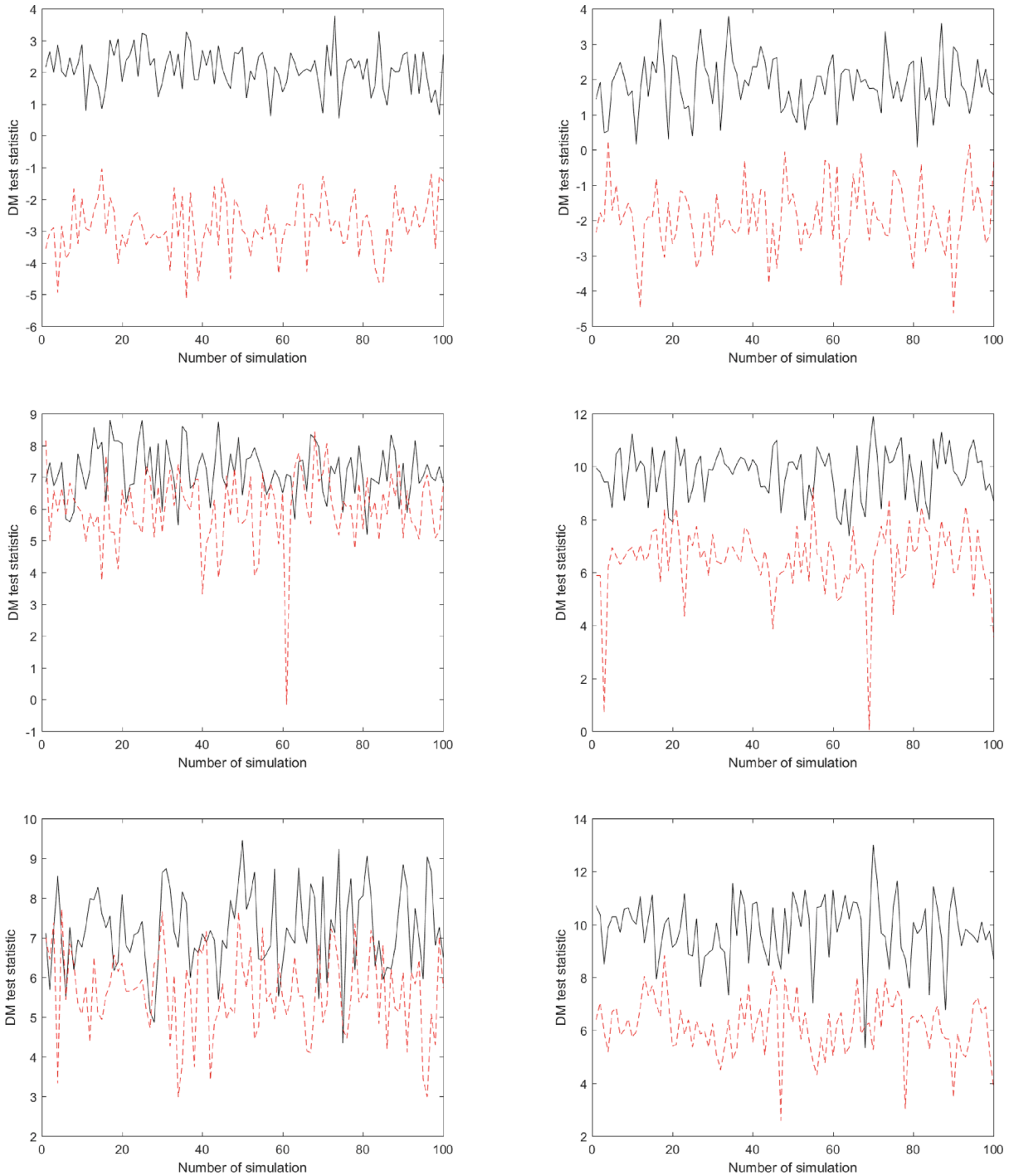

Figure 2. Diebold–Mariano (DM) test statistics for $B = 100$ simulated series for DGPs (5.1)–(5.3). Left panels correspond to $T = 500$ and right panels to $T = 1000$. The number of observations in each interval $[z_k - h_k, z_k + h_k)$ is $n = 0.1 \times T$. Black solid line for the DM test with in-sample data and red dashed line for the DM test with out-of-sample data

DOI: 10.1111/jtsa.12668

The analysis of the DGPs in (5.2) (middle panels) and (5.3) (bottom panels) yields very different findings. The DM test statistics in Figure 2 indicate a better fit of the partitioning estimator in sample and out of sample. This result is highly statistically significant for both sample sizes and across the two DGPs in (ii) and (iii). These exercises highlight the advantages of the nonparametric partitioning estimator when the functional form is nonlinear.

5.3. Empirical Rejection Rates and Power Analysis of Pointwise Predictability Tests

This section investigates the empirical rejection rates associated to the asymptotic prediction interval (4.2). The section also studies the empirical power of the pointwise predictability test in (4.1) for different DGPs and sample sizes. The empirical rejection rates ($\widehat{\alpha}_B(y_j)$) are computed as the fraction of times the true observation $g(y_j)$ is outside the forecast interval (4.2) such that

$$\widehat{\alpha}_B(y_j) = \frac{1}{B}\sum_{b=1}^{B} 1\left(|\widehat{g}_b(y_j) - g(y_j)| > z_{1-\alpha/2}\widehat{V}^{1/2}_{T(b)}(y_j)/\sqrt{T}\right), \tag{5.5}$$

where $\widehat{g}_b(y_j)$ and $\widehat{V}_{T(b)}(y_j)$ are the pointwise forecast and associated variance estimator of the DGP for simulation $b = 1, \ldots, B$; $z_{1-\alpha/2}$ is the $1-\alpha/2$ quantile of a standard Normal distribution function and $1-\alpha$ denotes the nominal coverage probability of the two-sided asymptotic confidence interval. The grid $\{y_1, \ldots, y_m\}$ is constructed as for the mean square error exercise above. These estimates are reported in the figures as a solid black line. The results in Figure 3 show excellent empirical rejection rates at $\alpha = 0.05$ across the three DGPs. The asymptotic forecast interval (4.2) provides an accurate description of the uncertainty about the point forecasts $\widehat{g}(y)$ returned by the partitioning estimator.

Figure 3 also reports, as dashed red lines, the empirical power of the pointwise predictability tests $H_0(y_j) : g(y_j) = 0$ against $H_A(y_j) : g(y_j) \neq 0$, for $j = 1, \ldots, m$. The estimates of the power of the test are obtained from the following expression:

$$\widehat{p}_B(y_j) = \frac{1}{B}\sum_{b=1}^{B} 1\left(\sqrt{T}\left|\frac{\widehat{g}_b(y_j)}{\widehat{V}^{1/2}_{T(b)}(y_j)}\right| > z_{1-\alpha/2}\right). \tag{5.6}$$

The different panels in the three figures show, in general, high power to reject the null hypothesis of no forecast ability for the DGPs (5.1) to (5.3). Interestingly, the empirical power significantly drops in regions of the support of the predictor variable that contain zeros of the function $g(y)$. A zero of the predictor function entails the condition $g(y) = 0$, that is interpreted as lack of predictability in nonparametric settings, see Juhl (2014). In these regions the empirical power converges to the size (0.05) as y approaches the zero of the function. This is observed in the top panels of Figure 3 that report the power of the test (dashed red line) for the DGP in (5.1). In this model, $g(y) = 0$ for $y = 0$. For the DGP in (5.2), the middle panels of Figure 3 show how the empirical power converges to the nominal size for $y = 0$. This is so because the function $g(y) = sin(y)y$ is equal to zero in the interval $[-2, 2]$ at $y = 0$. Similarly, the power analysis for the DGP in (5.3) exhibits high power to reject the null hypothesis of no predictability outside the zeros of the function, obtained at $y = \{-\pi/2, 0, \pi/2\}$, and power values close to the nominal size when the predictability is evaluated at the zeros. The null hypothesis of absence of predictability is not rejected at these points. An alternative interpretation is that the pointwise model prediction at these points is zero.

5.4. Finite-sample Properties of Uniform Test

This section concludes with the analysis of the finite-sample properties of the uniform test of predictive ability. Table I reports the empirical size and power of the test for Models 1–3 corresponding to processes (5.1) to (5.3)

DOI: 10.1111/jtsa.12668

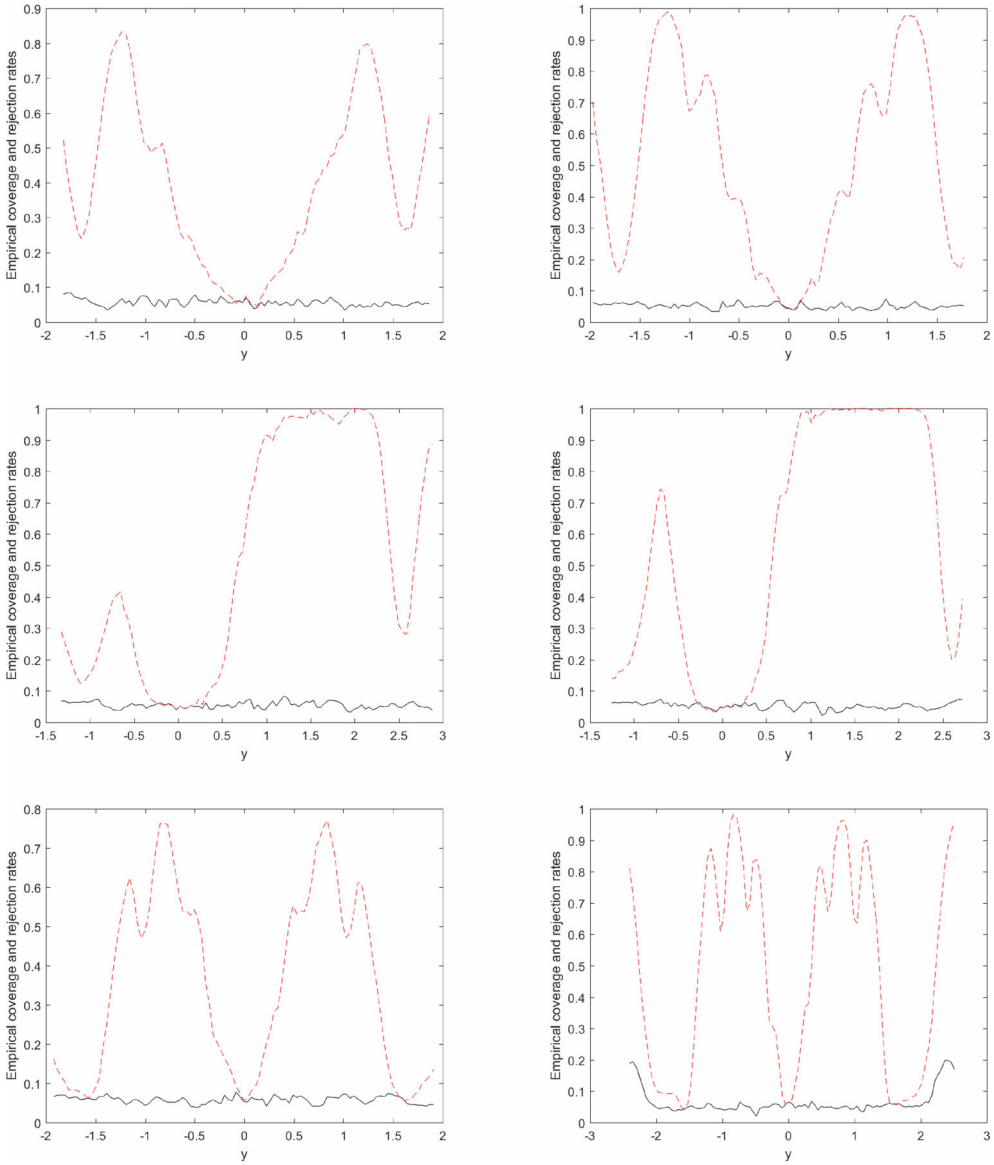

Figure 3. Empirical rejection rates and power analysis of pointwise predictability tests for the DGPs (5.1) to (5.3). The null hypothesis is $H_0 : g(y_j) = 0$ against $H_A : g(y_j) \neq 0$, for y_j a grid of points covering 90% of the support. Black solid line for the empirical rejection rates of the confidence interval (5.2) and red dashed line for the empirical power of the predictability test (4.2). Left panels corresponds to $T = 500$ and right panels to $T = 1000$. The number of observations in each interval $[z_k - h_k, z_k + h_k)$ is $n = 0.1 \times T$

respectively. Model 4 corresponds to the null hypothesis of no predictability. The DGP of the latter model is generated as $y_t = \varepsilon_t$, with $\varepsilon_t \sim N(0, 1)$. The top panel considers $T = 200$, the middle panel is for $T = 500$ and the bottom panel of the table considers the case $T = 1000$.

The results illustrate the empirical validity of the Wild bootstrap procedure proposed above to approximate the asymptotic p-value of the uniform test. The test has very strong power to reject the null hypothesis under the alternative hypothesis of predictive ability and also reports accurate empirical sizes under the absence of predictive

DOI: 10.1111/jtsa.12668

Table I. Empirical size and power of uniform test

$T = 200$	n	Model 1	Model 2	Model 3	Model 4
	10	0.73	0.34	0.23	0.15
	15	0.94	0.46	0.50	0.11
	20	0.98	0.58	0.64	0.08
	25	0.99	0.73	0.76	0.06
	30	1.00	0.75	0.87	0.06
	35	1.00	0.79	0.88	0.08
	40	1.00	0.86	0.91	0.05
$T = 500$					
	25	1.00	0.72	0.84	0.05
	38	1.00	0.89	0.98	0.04
	50	1.00	0.96	0.98	0.04
	63	1.00	0.94	0.99	0.05
	75	1.00	0.95	0.99	0.05
	88	1.00	0.90	0.98	0.05
	100	1.00	0.94	0.98	0.04
$T = 1000$					
	50	1.00	0.96	0.99	0.04
	75	1.00	0.96	0.99	0.02
	100	1.00	0.95	1.00	0.02
	125	1.00	0.95	0.99	0.02
	150	1.00	0.94	0.99	0.04
	175	1.00	0.94	0.99	0.06
	200	1.00	0.94	0.99	0.04

Note: This table reports the empirical size and power of uniform test (4.4) for sample sizes $T = 200{,}500, 1000$ and different values of n. Number of simulations is $B = 500$.

ability (Model 4). The results do not show great improvement as the sample size increases. Larger values of n for $T = 200, 1000$ seem to favor accurate test size estimates. On the other hand, for small values of n, we observe minor size distortions.

As an additional exercise, we also test for the presence of predictive ability for compact subsets of the support of the random variable. We consider the DGP (*ii*) given by $g(y) = sin(y)y$ as an illustrative example. This process takes values close to zero for values of y in a neighborhood of zero and exhibits predictability for values of y different from zero. To measure the power of the uniform test in this setting we consider different closed intervals $[-a, a]$, for $a = 0, 0.2, 0.4, 0.6, 0.8, 1$. Figure 4 reports the rejection probabilities of the uniform test for each compact subset. The x axis corresponds to the value a that characterizes the interval $[-a, a]$ and the y axis reports the empirical rejection probability. We obtain a curve that is increasing with a, reflecting the increase in predictive ability of the function $g(y)$ as the compact subset includes values more distant from zero.

5.5. Optimal Choice of Tuning Parameters

The tuning parameters characterizing the partitioning estimator are the number of intervals K and the order of the Taylor expansion q. In Section 2.3, we discussed five alternative methods to optimally choose these quantities. We implement these methods using the three DGPs discussed above and differentiate between an in-sample and an out-of-sample analysis, with particular emphasis on the out-of-sample case. The simulation setup is analogous to previous examples. The out-of-sample set is given by $M = 500$ observations and we consider $n = \lceil cT \rceil$, with c ranging between 5% and 20% and $T = 100{,}500, 1000$ such that $K = \lfloor T/n \rfloor \approx 1/c$. Potential orders for the Taylor expansion are $q = 1, 2, 3$.

DOI: 10.1111/jtsa.12668

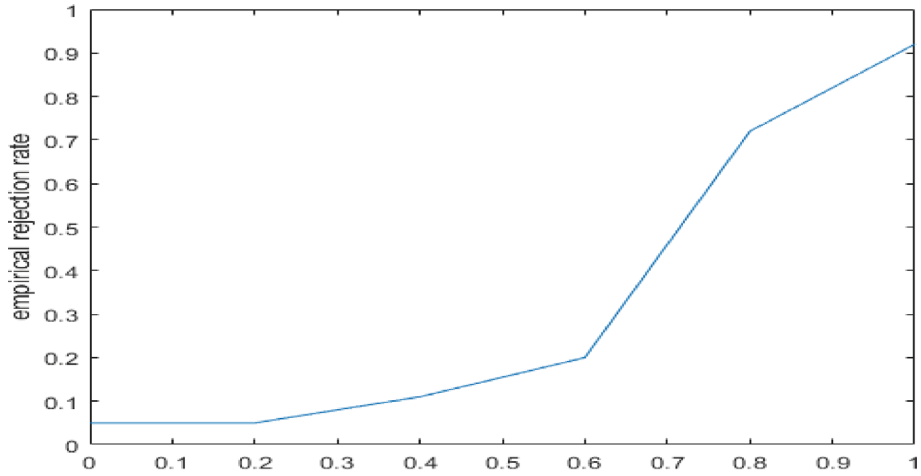

Figure 4. Empirical power of the uniform test (4.4) restricted to the compact subsets $[-a, a]$, with $a = 0, 0.2, 0.4, 0.6, 0.8, 1$. The sample size is $T = 1000$ and number of observations in each interval $[z_k - h_k, z_k + h_k)$ is $n = 100$

Table II. Out-of-sample optimal choice of the tuning parameter n

Model	T	RMSE	*M*	CV	AIC	BIC
AR(1)	100	17.842	18.886	19.032	18.696	18.954
		(2.302)	(1.224)	(1.000)	(1.533)	(1.127)
	500	85.840	92.758	93.736	91.466	92.902
		(13.58)	(7.114)	(6.000)	(8.733)	(7.053)
	1000	165.8	184.7	188.1	181.15	184.8
		(35.28)	(16.09)	(12.59)	(19.08)	(15.79)
sine(1)	100	18.042	19.218	19.354	18.936	19.124
		(2.586)	(1.218)	(0.983)	(1.543)	(1.178)
	500	85.436	92.612	95.646	89.74	91.596
		(14.83)	(9.129)	(6.246)	(10.97)	(9.471)
	1000	156.2	177.8	184.85	169.2	176.5
		(37.307)	(22.375)	(15.643)	(28.42)	(22.61)
cos(1)	100	17.846	19.024	19.174	18.686	19.036
		(2.431)	(1.250)	(1.000)	(1.666)	(1.141)
	500	85.588	93.304	95.738	91.012	92.742
		(16.30)	(8.612)	(6.132)	(10.65)	(8.421)
	1000	164.5	184.15	191.15	175.6	182.45
		(35.31)	(19.83)	(13.35)	(26.71)	(20.59)

Note: This table reports the optimal value of the tuning parameter n under the different criteria described in Section 2.3. The out-of-sample size is $M = 500$. Standard errors are shown in parentheses. The number of simulations is 500.

Tables II and III report the optimal values of n and q respectively, for the out-of-sample setting. The in-sample setting is available from the author on request. The in-sample case is more conservative than the out-of-sample exercise yielding a larger order of the Taylor expansion and a larger value of K, implying more regressors and a better fit of the data. The out-of-sample case avoids the overfit of the nonparametric regression. Table II suggests that the optimal value of n is about 15%–20% of the sample size T, implying values between 15 and 20 for $T = 100$; between 75 and 100 for $T = 500$, and between 150 and 200 for $T = 1000$. The specific optimal choice

DOI: 10.1111/jtsa.12668

Table III. Out-of-sample optimal choice of the Taylor expansion q

Model	T	RMSE	M	GCV	AIC	BIC
AR(1)	100	1.032	1.000	1.000	1.042	1.042
		(0.187)	(0.000)	(0.000)	(0.219)	(0.219)
	500	1.154	1.000	1.000	1.204	1.218
		(0.417)	(0.000)	(0.000)	(0.493)	(0.508)
	1000	1.316	1.008	1.000	1.364	1.374
		(0.617)	(0.089)	(0.000)	(0.651)	(0.653)
sine(1)	100	1.408	1.082	1.01	1.446	1.452
		(0.538)	(0.274)	(0.099)	(0.558)	(0.562)
	500	2.25	1.898	1.584	2.286	2.286
		(0.562)	(0.447)	(0.505)	(0.533)	(0.533)
	1000	2.424	2.102	1.922	2.472	2.484
		(0.563)	(0.429)	(0.363)	(0.534)	(0.523)
cos(1)	100	1.194	1.022	1.002	1.196	1.206
		(0.465)	(0.159)	(0.045)	(0.475)	(0.489)
	500	2.546	1.744	1.148	2.580	2.590
		(0.759)	(0.916)	(0.467)	(0.738)	(0.739)
	1000	2.744	2.576	1.816	2.820	2.866
		(0.546)	(0.759)	(0.914)	(0.494)	(0.452)

Note: This table reports the optimal order q of the Taylor expansion under the different criteria described in Section 2.3. The out-of-sample size is $M = 500$. Standard errors are shown in parentheses. The number of simulations is 500.

of n also depends on the loss function but there is ample agreement across methods. The RMSE loss function reports the smallest values of n and the CV the largest. The underlying theory supports these choices. This is so because the RMSE does not penalize the number of regressors in the model and, therefore, entails a smaller n (a larger number of regressors) than for the other loss functions. In contrast, the other measures penalize, to different extents, regression models with many regressors. The standard errors provide further validity to the optimality results.

Table III provides empirical evidence on the optimal order of the local Taylor expansions. In general, the results are quite robust across loss functions and DGPs. As the sample size increases, the model can accommodate more regressors and the optimal choice of q is given by higher order expansions. More specifically, for $T = 100$ the optimal choice is $q = 1$, however, as T increases the optimal choice of q is close to 3.

6. EMPIRICAL APPLICATION

We focus our empirical investigations on the high-frequency based volatility measures dataset analyzed in Bollerslev *et al.* (2016). These authors consider the 27 Dow Jones constituents as of September 20, 2013 that are traded continuously from the start of the sample until the end. Data on these individual stocks comes from the TAQ database. The sample starts on April 21, 1997 and ends on December 31, 2013, yielding a total of 4,202 observations for the DJIA constituents. Table 2 in Bollerslev *et al.* (2016) provides the summary statistics for the daily realized volatilities (RV). We use a subset of the realized measures provided in Bollerslev *et al.* (2016), in particular, we focus on the realized volatility measures (RV_t) initially explored in Andersen *et al.* (2003) and the bipower variation (BPV_t) measures of Barndorff-Nielsen and Shephard (2004), both obtained from five-minute intraday squared returns.

Figure 5 reports the scatter plot for both RV_t and BPV_t measures for the first firm of the dataset given by American Express (AXP). This figure shows clear evidence of nonlinearity for the raw measures RV_t and BPV_t, whereas the log transformation suggests a linear relationship between the variables. Both panels in Figure 5 provide empirical

DOI: 10.1111/jtsa.12668

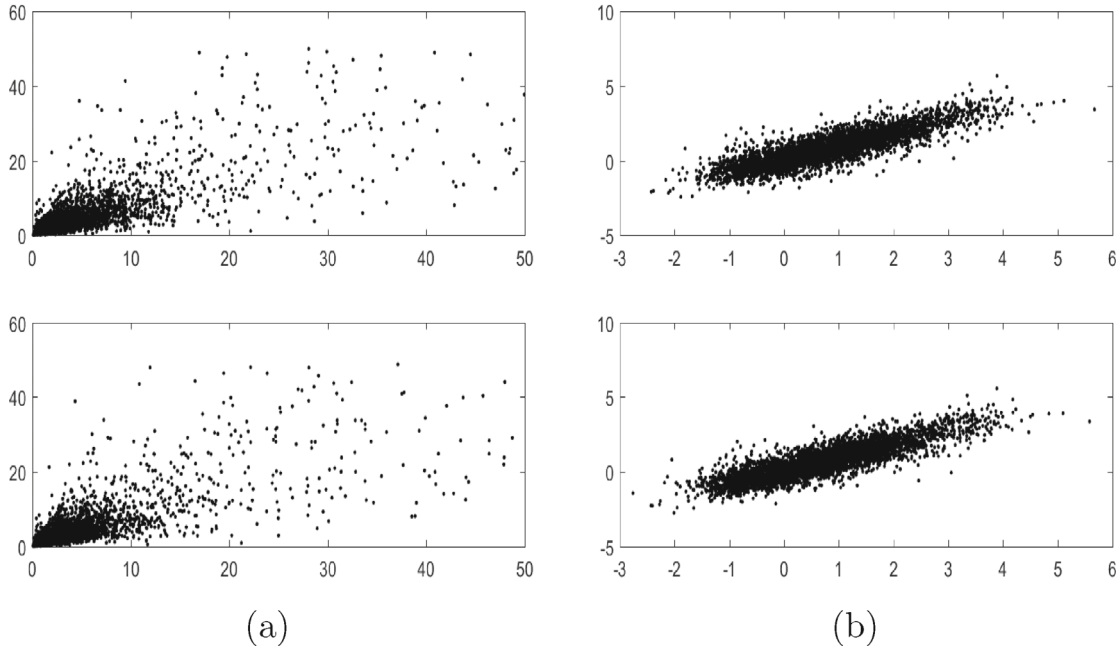

Figure 5. Panel (a) presents the scatter plot of (V_{t-1}, V_t) for AXP firm with V_t the realized volatility measures constructed at 5 minute frequencies. Panel (b) presents the scatter plot for $(\ln\ V_{t-1}, \ln\ V_t)$. Top panels for $V_t = RV_t$ and bottom panels for $V_t = BPV_t$

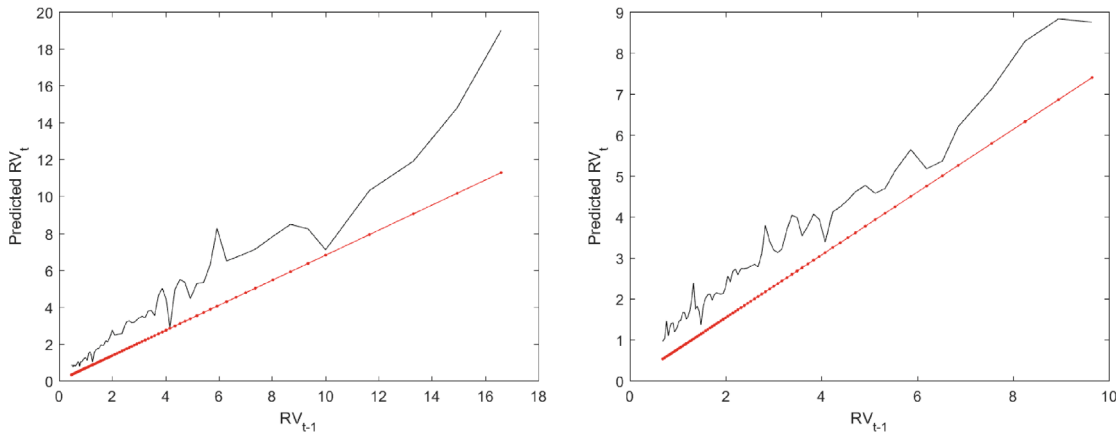

Figure 6. Predictions of RV_t constructed from 5-minute returns given by the linear AR(1) model (solid red line) and the partitioning estimator (solid black line). Panel (a) corresponds to AXP firm and panel (b) to BA firm

evidence of the existence of predictability for the realized volatility measures over time. Importantly, whereas the linear AR(1) model is a suitable model for the log realized volatility measures it is not for the raw volatility measures. In this section, we investigate the suitability of the partitioning estimator and compare its predictive ability against the AR(1) model.[5]

[5] We focus on the raw data for RV_t and BPV_t. Unreported results show the excellent performance of the AR(1) model for the log transformation of the volatility measures. We should note, however, that the predictions of the log realized volatility measures need to be transformed to obtain meaningful predictions of the raw realized volatility, which is generally the object of interest.

DOI: 10.1111/jtsa.12668

Figures 6 reports the predicted value of $g(y)$, for y in the support of the random variable RV_t, for both models. The black solid line corresponds to the forecasts of the partitioning estimator and the red dashed line to the forecasts of the linear autoregressive model of order one. Figure 7 reports the out-of-sample predictions over the evaluation period given by the last $M = 2000$ observations of the sample. The nonparametric model is constructed with $K(q+1) = 33$ regressors, with $K = 11$ corresponding to $n = 200$ observations inside each interval $[z_k - h_k, z_k + h_k)$, for an in-sample period of $T = 2202$ observations, and $q = 2$, as suggested by the model selection simulation exercise. The top panels of Figure 7 plot together the in-sample realized volatility (RV_t) forecasts and the actual observations for the AXP firm. The bottom panels report the out-of-sample forecasts and actual out-of-sample realized measures. The forecasts of the nonlinear model are less volatile than the forecasts of the linear AR(1) process.

The performance of both methods is assessed through the comparison of the adjusted R^2 coefficients for the in-sample and out-of-sample evaluation periods. $R1$ corresponds to the R^2 measure of the parametric AR(1) model and $R2$ corresponds to the partitioning estimator. Theoretically, the in-sample period should favor the partitioning estimator because the number of regressors is substantially larger and the out-of-sample period should report fairer comparisons. The in-sample results in Table IV confirm, for most firms, the outperformance of the partitioning estimator. This is reflected in a larger R^2 statistic. Importantly, such improvements of the partitioning estimator are also observed over the out-of-sample evaluation period. This is clearly the case for MCD, BA, HD, MMM, IBM, MRK, CSCO, INTC, MSFT, JNJ, WMT, DIS, KO, and PG. For a few firms such as AXP, UTX, and VZ the improvement is for the RV measure but not for the BPV. For the PFE stock return volatility, the improvement in out-of-sample goodness of fit is for the BPV measure only. Finally, for the remaining firms, the AR(1) model provides superior out-of-sample forecasts. The differences in forecast ability between models are also confirmed in most cases by the Diebold-Mariano test. Results on these tests for individual stocks are available from the author on request.

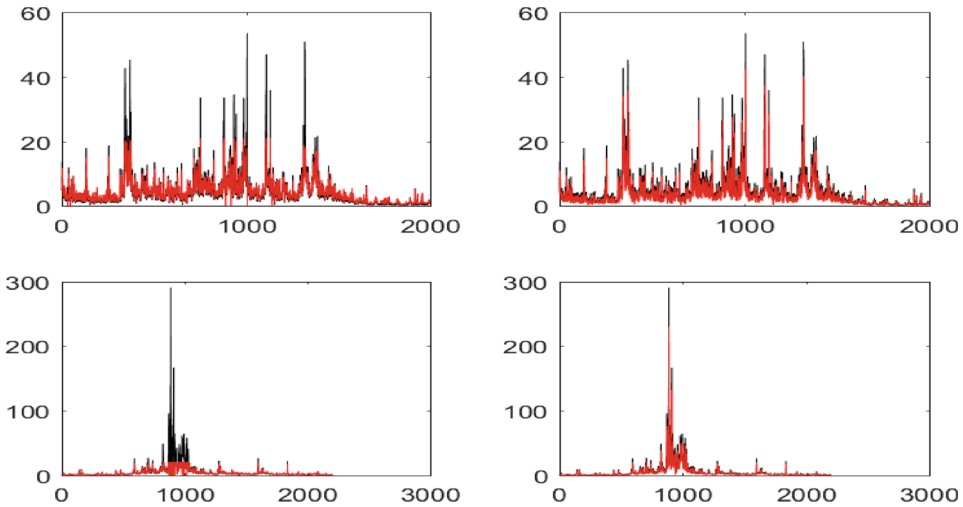

Figure 7. Top left panel reports the observed 5-minute returns realized volatility for AXP firm. The in-sample evaluation period is plotted as a black line and the predicted values obtained from the partitioning estimator as a red line. Top right panel reports the same time series using the AR(1) model. Bottom panels report the same figures for the out-of-sample evaluation period. In-sample period given by $T = 2202$ observations and out-of-sample period given by $M = 2000$ observations. Number of observations in each cell is $n = 200$ for the partitioning estimator

DOI: 10.1111/jtsa.12668

Table IV. Adjusted R^2 coefficients for the constituents of S&P 500 index

Firm	RV	BPV	Firm	RV	BPV	Firm	RV	BPV	Firm	RV	BPV
AXP			**GE**			**MCD**			**TRV**		
R1 in	0.356	0.397		0.224	0.273		−0.109	0.004		0.348	0.342
R2 in	**0.458**	**0.458**		**0.389**	**0.411**		**0.179**	**0.213**		**0.444**	**0.373**
R1 out	0.308	0.334		0.484	0.472		0.083	0.061		0.388	0.348
R2 out	**0.312**	0.272		0.275	0.271		**0.217**	**0.257**		0.153	0.134
BA			**HD**			**MMM**			**UNH**		
R1 in	0.131	0.197		0.318	0.305		0.230	0.222		0.084	0.070
R2 in	**0.327**	**0.338**		**0.447**	**0.440**		**0.371**	**0.363**		**0.233**	**0.226**
R1 out	0.514	0.519		0.367	0.350		0.106	0.057		0.419	0.409
R2 out	**0.535**	**0.534**		**0.497**	**0.490**		**0.158**	**0.158**		0.362	0.382
CAT			**IBM**			**MRK**			**UTX**		
R1 in	0.274	0.219		0.291	0.330		0.092	0.201		0.303	0.380
R2 in	**0.390**	**0.355**		**0.411**	**0.429**		**0.243**	**0.319**		**0.404**	**0.470**
R1 out	0.561	0.608		0.427	0.396		−0.020	−0.034		0.401	0.386
R2 out	0.208	0.234		**0.469**	**0.438**		**0.192**	**0.207**		**0.468**	−0.028
CSCO			**INTC**			**MSFT**			**VZ**		
R1 in	0.406	0.391		0.491	0.460		0.409	0.445		0.243	0.393
R2 in	**0.498**	**0.475**		**0.587**	**0.562**		**0.502**	**0.503**		**0.374**	**0.422**
R1 out	0.455	0.463		0.358	0.428		0.499	0.468		0.429	0.492
R2 out	**0.539**	**0.521**		**0.500**	**0.526**		**0.522**	**0.498**		**0.506**	0.307
CVX			**JNJ**			**NKE**			**WMT**		
R1 in	0.295	0.309		0.295	0.237		0.025	0.046		0.218	0.289
R2 in	**0.399**	**0.393**		**0.356**	**0.338**		**0.278**	**0.267**		**0.377**	**0.411**
R1 out	0.384	0.263		0.247	0.273		0.478	0.440		0.337	0.335
R2 out	0.040	0.012		**0.406**	**0.425**		0.427	0.427		**0.434**	**0.461**
DD			**JPM**			**PFE**			**XOM**		
R1 in	0.376	0.363		0.429	0.439		0.015	0.049		0.352	0.343
R2 in	**0.445**	**0.438**		**0.572**	**0.535**		**0.247**	**0.259**		**0.454**	**0.447**
R1 out	0.492	0.491		0.519	0.510		0.457	0.426		0.421	0.304
R2 out	0.288	0.307		0.405	0.445		0.444	**0.442**		0.211	0.182
DIS			**KO**			**PG**					
R1 in	0.202	0.290		0.278	0.335		0.247	0.364			
R2 in	**0.294**	**0.357**		**0.405**	**0.424**		**0.358**	**0.440**			
R1 out	0.463	0.497		0.273	0.384		0.272	0.258			
R2 out	**0.511**	**0.518**		**0.398**	**0.506**		**0.400**	**0.431**			

Note: This table reports the adjusted R^2 for the in-sample and out-of-sample exercises. R1 denotes the coefficient associated to the AR(1) process and R2 the coefficient associated to the partitioning estimator. The number of in-sample observations is 2202 and the number of out-of-sample observations is $M = 2000$. The number of observations in each interval $[z_k - h_k, z_k + h_k)$ is $n = 200$. This implies $K = 33$ regressors in the partitioning estimator regression model.

7. CONCLUSIONS

This article proposes a nonparametric predictive regression model that is approximated using Taylor expansions of low order ($q \leq 2$) applied over disjoint intervals covering the support of the predictor variable. The model is estimated using the theory on partitioning estimators developed in Cattaneo and Farrell (2013) that we extend to a predictive framework with stationary, β-mixing predictors and a model error that is a martingale difference sequence. We derive the asymptotic properties of the partitioning estimator that are applied to test for the presence of predictive ability. We develop an asymptotic pointwise test of predictive ability using the critical values of a Normal distribution, and a uniform test of predictability over the compact support of the predictor variable, with asymptotic distribution that is approximated, in finite samples, using Wild bootstrap methods.

DOI: 10.1111/jtsa.12668

The application of these results for modeling and forecasting nonparametrically different realized volatility measures for the 27 constituents of the Dow Jones highlights the strong predictive ability of this model that outperforms the standard AR(1) model for most stocks. Importantly, we also find overwhelming evidence on the existence of nonlinearities on the dynamics of the raw realized volatility measures. Whereas log transformations of these measures are linear, the raw measures are highly nonlinear, in particular for the upper tails of the distribution.

ACKNOWLEDGMENT

Jose Olmo acknowledges financial support from project PID2019-104326GB-I00 from Ministerio de Ciencia e Innovación and from Fundación Agencia Aragonesa para la Investigación y el Desarrollo (ARAID).

DATA AVAILABILITY STATEMENT

We focus our empirical investigations on the high-frequency based volatility measures dataset analyzed in Bollerslev *et al.* (2016). These authors consider the 27 constituents of the Dow Jones Industrial Average Index as of September 20, 2013. These stocks are traded continuously from the start of the sample until the end. Data on these individual stocks comes from the TAQ database.

SUPPORTING INFORMATION

Additional Supporting Information may be found online in the supporting information tab for this article.

REFERENCES

Andersen TG, Bollerslev T, Diebold FX, Labys P. 2003. Modeling and forecasting realized volatility. *Econometrica* **71**:579–625.

Andrews DWK. 1991. Asymptotic normality of series estimators for nonparametric and semiparametric regression models. *Econometrica* **59**(5):307–345.

Andrews DWK. 1993. Tests for parameter instability and structural change with unknown change point. *Econometrica* **61**(4):821–856.

Andrews D. 1994. Asymptotics for semiparametric models via stochastic equicontinuity. *Econometrica* **62**(4):43–72.

Barndorff-Nielsen OE, Shephard N. 2004. Power and bipower variation with stochastic volatility and jumps. *Journal of Financial Econometrics* **1**(2):1–37.

Bollerslev T, Patton A, Quaedvlieg R. 2016. Comparing predictive accuracy. *Journal of Econometrics* **192**(1):1–18.

Cai Z, Fan J, Li R. 2000a. Efficient estimation and inference for varying coefficient models. *Journal of the American Statistical Association* **95**(4):888–902.

Cai Z, Fan J, Yao Q. 2000b. Functional coefficient regression models for nonlinear time series. *Journal of the American Statistical Association* **95**(4):941–956.

Campbell J, Yogo M. 2006. Efficient tests of stock return predictability. *Journal of Financial Economics* **81**(4):27–60.

Cattaneo M, Farrell M. 2013. Optimal convergence rates, bahadur representation, and asymptotic normality of partitioning estimators. *Journal of Econometrics* **174**(1):127–143.

Cattaneo M, Farrell M, Feng Y. 2020. Large sample properties of partitioning-based series estimators. *The Annals of Statistics* **48**(3):1718–1741.

Chen H. 1988. Convergence rates for parametric components in a partly linear model. *Annals of Statistics* **16**(5):136–146.

Craven P, Wahba G. 1978. Smoothing noisy data with spline functions. estimating the correct degree of smoothing by the method of generalized cross-validation. *Numerische Mathematik* **31**:377–404.

Davies RB. 1977. Hypothesis testing when a nuisance parameter is present only under the alternative. *Biometrika* **64**(2):247–254.

Davies RB. 1987. Hypothesis testing when a nuisance parameter is present only under the alternatives. *Biometrika* **74**(1):33–43.

Diebold FX, Mariano RS. 1995. Comparing predictive accuracy. *Journal of Business & Economic Statistics* **13**(3):253–263.

Fan J, Gijbels I. 1996. *Local polynomial modelling and its applications*. In *Monographs on Statistics and Applied Probability*, Chapman & Hall/CRC Press, London, UK.

Fan J, Yao Q, Cai Z. 2003. Adaptive varying-coefficient linear models. *Journal of the Royal Statistical Society, Series B* **65**(4):57–80.

DOI: 10.1111/jtsa.12668

Györfi L, Kohler M, Krzyzak A, Walk H. 2002. *A Distribution-Free Theory of Nonparametric Regression* Springer-Verlag, New York.
Hansen BE. 1996. Inference when a nuisance parameter is not identified under the null hypothesis. *Econometrica* **64**(2):413–430.
Jansson M, Moreira M. 2006. Optimal inference in regression models with nearly integrated regressors. *Econometrica* **74**(4):681–714.
Juhl T. 2014. A nonparametric test of the predictive regression model. *Journal of Business and Economic Statistics* **32**(3):387–394.
Kim W, Linton O, Hengartner N. 1999. A computationally efficient oracle estimator for additive nonparametric regression with bootstrap confidence intervals. *Journal of Computational and Graphical Statistics* **8**(4):278–297.
Lewellen J. 2004. Predicting returns with financial ratios. *Journal of Financial Economics* **74**(4):209–235.
Li K-C. 1987. Asymptotic optimality for c_p, c_l, cross-validation and generalized cross-validation: discrete index set. *Annals of Statistics* **15**(3):958–975.
Mallows CL. 1973. Some comments on c_p. *Technometrics* **15**(4):661–675.
Nadaraya EA. 1965. On nonparametric estimates of density functions and regression curves. *Theory of Applied Probability* **10**(1):186–190.
Newey WK. 1997. Convergence rates and asymptotic normality for series estimators. *Journal of Econometrics* **79**(1):147–168.
Stambaugh R. 1999. Predictive regressions. *Journal of Financial Economics* **54**(4):375–421.
Stone M. 1974. Cross-validatory choice and assessment of statistical predictions. *Journal of the Royal Statistical Society. Series B (Methodological)* **36**(2):111–147.
Stone CJ. 1985. Additive regression and other nonparametric models. *The Annals of Statistics* **13**(2):689–705.
Tukey JW. 1947. Nonparametric estimation ii. statistically equivalent blocks and tolerance regions. *Annals of Mathematical Statistics* **18**(5):529–539.
Tukey JW. 1961. *Curves as parameters and touch estimation. Proceedings of the Fourth Berkeley Symposium* United States, Cambridge, MA.
Wahba G. 1985. A comparison of gcv and gml for choosing the smoothing parameter in the generalized spline smoothing problem. *The Annals of Statistics* **13**(4):1378–1402.
Watson GS. 1964. Smooth regression analysis. *Sankhya* **26**(1):359–372.

DOI: 10.1111/jtsa.12668