Boosting Non-linear Predictability of Macroeconomic Time Series

Heikki Kauppi Timo Virtanen

heikki.kauppi@utu.fi timo.virtanen@utu.fi

Department of Economics

Turku School of Economics

FI-20014 University of Turku

Finland

November 15, 2019

Abstract

We apply the boosting estimation method to investigate to what extent and at what horizons macroeconomic time series have nonlinear predictability coming from their own history. Our results indicate that the U.S. macroeconomic time series have more exploitable nonlinear predictability than previous studies have found. On average, the most favorable out-of-sample performance is obtained by a two-stage procedure, where a conventional linear prediction model is fitted first and the boosting technique is applied to build a nonlinear model for its residuals.

Keywords: boosting, forecasting, linear autoregression, macroeconomic time series, mean squared error, non-linearity

1 Introduction

There has been a longlasting debate about whether a linear or a nonlinear modeling approach should be applied in the forecasting of macroeconomic time series. While it is often argued that nonlinearity is an inherent feature of macroeconomic time series, linear forecasts have been found to perform mostly better than forecasts based on various nonlinear models. There are cases where nonlinear models have yielded more accurate forecasts than linear models, but generally it has remained quite unclear to what extent and when nonlinear forecasts are likely to be useful in macroeconomic forecasting. The literature from which these observations arise includes Stock and Watson (1999), Marcellino (2005), Teräsvirta, van Dijk and Medeiros (2005), and Kock and Teräsvirta (2016).

In this paper, we use a novel approach to examine how common it is that macroeconomic time series have exploitable nonlinear predictability in their own history. The target of our empirical analysis is the FRED-MD data set introduced by McCracken and Ng (2016). The data set contains 128 monthly macroeconomic time series with observations (mostly) from January 1959 to December 2017. The variables in the data set cover sectors and markets that are central to the development of the U.S. economy. By analyzing this data set we get a comprehensive view of what to expect about typical macroeconomic time series in terms of nonlinear predictability.

Our analysis makes use of the so called boosting estimator from the machine learning literature. This technique has been found to have superior performance in estimating complex nonlinear regression functions (Friedman (2001)). As this method is nonparametric and can be made very flexible, it offers us a device for examining whether the data can reveal any nonlinear predictability without need to make specific assumptions on the form of the underlying nonlinear model. This is important, because the target set of time series is so large that there is no chance that any given class of parametric models is able to fit to all of them. The boosting method has also some unique strengths by which it can potentially outperform the artificial neural network (ANN) technique that has been the dominant nonparametric alternative in the previous studies closest to the present one

(especially the papers mentioned in the first paragraph).[1]

The starting question in the present study is whether the boosting estimation technique or a conventional linear prediction strategy yields more accurate 1 to 12 month ahead predictions for a given series in the FRED data set. Our empirical results indicate that at least every fourth series in the FRED-MD data set has nonlinear predictability exploitable by the boosting method. On the other hand, we find that for some series the boosting estimator has much worse out-of-sample performance than the linear prediction procedure. This observation is similar to what related previous studies have made on other nonlinear prediction techniques. Hence, we would wish to have a testing procedure that could diagnose which of the two techniques is more accurate out-of-sample. Unfortunately, although the econometrics literature has developed various tests for comparing the out-of-sample accuracy of alternative forecasts, none of them seems to work reliably enough in our application.[2]

As we have found no satisfactory pretest for predetermining whether the linear or the boosting technique provides more accurate out-of-sample predictions in our application, we include in our analysis a two-stage estimation procedure, where the first step involves

[1] While the boosting method applies a similar type of additive function approximation strategy as the ANN method, its unique step-wise fitting algorithm is likely to give it an advantage over the ANN method. In particular, the boosting method has the so called slow overfitting property (to be discussed in Section 3.1) by which it is more resistant to overfitting than alternative nonparametric methods (including ANN).

[2] In an ongoing study, we use simulation techniques and the present data set to examine how well various existing testing techniques (e.g., those proposed in Clark and McCracken (2001), Clark and West (2006, 2007), Corradi and Swanson (2002), Diebold and Mariano (1995), and Giacomini and White (2006)) are able to determine in advance whether the boosting or the linear prediction procedure has the best out-of-sample performance. So far, we have found that none of the examined testing methods is sufficiently reliable for this purpose. One problem is that it is not clear whether the boosting technique is consistent for a target model that nests or does not nest the linear null model. Moreover, prediction accuracy tests (such as the one of Giacomini and White (2006)) that do not make specific assumptions on the nesting structure of the null and the alternative models entail that the underlying simulated out-of-sample predictions are generated by small enough sample sizes (to prevent estimation errors from vanishing), which is not an advantage to the flexible boosting technique.

estimating the conventional linear prediction model and the second step uses the boosting technique to build a nonlinear model for its residuals. We call this technique the two-stage boosting method. Altogether, our analysis examines the performance of three prediction procedures: (i) the conventional linear method, (ii) the "pure" or "direct" boosting method, and (iii) the two-stage boosting method.

We start the analysis by conducting a brief simulation study where we compare the performance of the three methods. The simulation demonstrates that the boosting estimation method can successfully capture various forms of nonlinearities in samples that are similar in size to what we have in the empirical application. We also find no clear differences in the out-of-sample accuracy between the pure and the two-stage boosting methods. The simulations confirm that both boosting procedures have the capacity to outperform the conventional linear procedure in situations where there is noticeable nonlinear predictability beyond the best linear approximation. Importantly, we find that the two-stage boosting method has a better out-of-sample accuracy than the pure boosting method when the true optimal model is linear. As a whole, the detailed simulation results help us to interpret the results that we obtain in the empirical analysis.

The most interesting result from our empirical analysis concerns the two-stage boosting procedure. In a nutshell, we find that it is almost always at least as accurate as the "pure" boosting procedure, while it is rarely less accurate than the linear procedure. Hence, among the three examined procedures the two-stage boosting technique is the most reliable one. This finding holds, often even more clearly, in alternative settings, where the original data is restricted to the time period before the 2008 financial crisis, and when we use conventional techniques to handle "outlier" observations. Our analysis suggests that the safest approach for exploiting nonlinear predictability in the forecasting of macroeconomic variables is to apply a flexible nonlinear procedure not as such, but rather as a device to add a nonlinear component to the linear autoregressive model.

In terms of the main question and the volume of the analyzed macroeconomic variables, the present paper is closest to the four papers already cited in the first paragraph above, although the present data set contains more time series observations than used in the mentioned papers. Studies that consider more specific groups of macroeconomic variables

or that otherwise discuss the problem of choosing between a linear and a non-linear prediction model include Swanson and White (1997a, 1997b), De Gooijer and Kumar (2002) and Marcellino (2004). For reviews of related literature see Clements, Franses and Swanson (2004) and Teräsvirta (2006).

Our paper is not the first one that applies the boosting method to predict macroeconomic variables. Robinzonov, Tutz and Hothorn (2012) use the boosting technique to predict the German industrial production. They find that the boosting method is competitive to the other methods they consider, but does not dominate them in all cases. Wohlrabe and Buchen (2014) use the boosting estimation technique to fit prediction models for a range of macroeconomic time series. However, they consider only a linear autoregressive model with additional regressors and focus on the question how well (in terms of out-of-sample accuracy) the boosting method selects the regressors from a large pool of variables. Kim and Swanson (2014) conduct a prediction horse race between various machine learning techniques including the boosting method, but their analysis focuses on comparing advanced estimation techniques rather than on analyzing how these compare with the conventional linear prediction procedure.

The two-stage boosting procedure applied in this paper is related to a boosting-based prediction technique presented in Taieb and Hyndman (2014). In the latter paper, the first stage involves estimating a "one-step ahead" linear autoregressive (AR) model once and then iterating it to obtain the desired multiperiod ahead prediction. Our two-stage approach differs from this strategy in that we estimate the linear AR model for each horizon separately. While both approaches have the same target model, we favor the direct one, because it is known to be more robust against misspecification of the one-step ahead model (e.g., Findley (1983, 1985), Bhansali (1997, 1999), and Ing (2003)). Nevertheless, we have repeated our similations and empirical analyses by using the indirect AR prediction in the first stage. In the simulations, we find no significant difference between the direct and the indirect procedures, while in the empirical comparisons the direct method is on average more accurate than the indirect approach. These results are reported in an appendix.

The rest of the paper is organized as follows. Section 2 presents the basic prediction

problem at the general level. Section 3 introduces the boosting estimation technique and the two alternative ways we apply it to forecasting. The simulation study is presented in Section 4. Section 5 conducts the empirical analysis. Section 6 gives concluding remarks.

2 The Starting Point

Let Y_t be a time series of interest. At time t, we seek to use the current and past observations $Y_t, Y_{t-1}, Y_{t-2}, ...$ to predict the future value Y_{t+h}, where h is the forecast horizon.

In our application, Y_t is often a logarithmic value of an original series. Moreover, Y_t may be I(0), I(1), or I(2), where I(k) signifies integrated of order k. As, e.g., in Stock and Watson (1999), we form the h-period ahead prediction in one of three ways depending on the order of integration of the series. In each case, the predicted variable is denoted by x_{t+h} and the predictors by $y_t, y_{t-1}, ...$ If Y_t is I(0) [I(1)] {I(2)}, then $x_{t+h} = Y_{t+h}$ and $y_t = Y_t$ [$x_{t+h} = Y_{t+h} - Y_t$ and $y_t = \Delta Y_t$] {$x_{t+h} = Y_{t+h} - Y_t - h\Delta Y_t$ and $y_t = \Delta^2 Y_t$}. Given a prediction $\widehat{x}_{t+h}$ made at time t for x_{t+h}, the prediction for Y_{t+h} is $\widehat{x}_{t+h}$ [$Y_t + \widehat{x}_{t+h}$] {$Y_t + h\Delta Y_t + \widehat{x}_{t+h}$}, if Y_t is I(0) [I(1)] {I(2)}.

In the theoretical analysis, we assume that y_t is a strictly stationary and ergodic time series. This also means that the same applies to x_t (in all three cases). We will let $y_t(k)$ signify the vector $(y_t, y_{t-1}, ..., y_{t-k+1})' \in \mathbb{R}^k$.

The h-step ahead prediction is obtained by using available data to estimate the model

$$x_{t+h} = g_h(y_t(p)) + e_{t+h} \tag{1}$$

where $g_h(y_t(p))$ is the prediction function and $e_{t+h} = x_{t+h} - g_h(y_t(p))$ is the prediction error. In our empirical analysis, we have monthly data and choose $p = 12$ as is common in macroeconomic applications.

As is usual in economic time series forecasting, the accuracy of the prediction $g_h(y_t(p))$ is assessed by the mean squared error (MSE)

$$MSE(g_h(y_t(p))) = E(x_{t+h} - g_h(y_t(p)))^2 \tag{2}$$

The optimal prediction minimizing (2) is the conditional mean

$$g_h^*(y_t(p)) = E(x_{t+h}|y_t(p))$$

We write $MMSE(h;p)$ for $MSE(g_h^*(y_t(p)))$.

A common approach is to assume a parametric prediction function $g_h(y_t(p);\theta_h)$, where the parameter θ_h lies in a (closed) space $\Theta_h \subset \mathbb{R}^{d_h}$, with d_h being finite. The unknown parameter θ_h is typically estimated by minimizing the empirical (in-sample) mean squared error and the estimation procedure may include intermediate specification steps, where the original dimension of Θ_h is reduced by using a statistical information criterion (like AIC or BIC).

Let $\widehat{g}_h(y_T(p))$ denote an estimate for $g_h(y_t(p))$ in (1) based on the sample $y_1, ..., y_T$.[3] The predictive performance of $\widehat{g}_h(y_T(p))$ is measured by the out-of-sample MSE

$$MSE(\widehat{g}_h(y_T(p))) = E(x_{T+h} - \widehat{g}_h(y_T(p)))^2 \tag{3}$$

We mostly interpret that the expectation in (3) is with respect to the joint distribution of the sample $y_1, ..., y_T$ and x_{T+h}. This means that $MSE(\widehat{g}_h(y_T(p)))$ depends on the sample size T, but not on a particular realization of observations.

Let $\widehat{g}_h^L(y_T(p)))$ and $\widehat{g}_h^N(y_T(p)))$, respectively, denote an estimated prediction function based on a linear and a nonlinear modeling approach. As in Stock and Watson (1999), Marcellino (2005), Teräsvirta, van Dijk and Medeiros (2005), Kock and Teräsvirta (2016)), we address the question which approach is better.[4] That is, do we have $MSE(\widehat{g}_h^L(y_T(p))) < MSE(\widehat{g}_h^N(y_T(p))))$ or $MSE(\widehat{g}_h^N(y_T(p)))) < MSE(\widehat{g}_h^L(y_T(p)))$?

It is often thought that the optimal prediction function is nonlinear, which suggests that if $\widehat{g}_h^N$ is based on a sufficiently general nonlinear model then $MSE(\widehat{g}_h^N(y_T(p)))) < MSE(\widehat{g}_h^L(y_T(p)))$. Nevertheless, the majority of the past studies indicates that the linear modeling approach is generally better, that is, mostly $MSE(\widehat{g}_h^L(y_T(p))) < MSE(\widehat{g}_h^N(y_T(p)))$. This widely shared finding may have several potential explanations.

[3]Here and in what follows, the notation $\widehat{g}_h(y_T(p))$ refers to an estimate of the prediction function based on a sample of size T, while $\widehat{g}_h(y_t(p))$ may refer to the corresponding prediction at time point t.

[4]On the part of the nonlinear modeling, the mentioned studies focus mostly on the smooth transition autoregressive, the exponential smoothing, and the artificial neural network models.

First, the true optimal prediction function may be linear. This is something we cannot know for certainty, but it is generally hard to believe this could be the case. Yet, it may often be that the best linear approximation is very close to the optimal nonlinear prediction. If there is only a small difference, the linear approach is likely to win, because with finite data one can estimate the best linear approximation more accurately than various nonlinear models. Also, the relative difference in the MSE of an optimal nonlinear model and the best linear approximation is necessarily smaller with weaker overall degree of predictability. Due to stationarity and ergodicity, the degree of predictability decreases eventually as the prediction horizon increases. Hence, as h becomes larger, the linear model, in fact eventually a constant prediction, will yield the best prediction.

Second, it may often be that the applied competitive nonlinear parametric model does not capture sufficiently accurately the optimal function $g_h^*(y_t(p))$. That is, the estimate $\widehat{g}_h^N(y_T(p)))$ may be consistent for a function $g_h^N(y_t(p)))$ such that $MSE(g_h^N(y_t(p))) > MMSE(h;p)$. Various nonlinear models do not nest the best linear model as a special case. We may then have a situation where the nonlinear target function is less accurate than the best linear approximation, that is, $MSE(g_h^N(y_t(p))) > LMSE(h;p)$, where $LMSE(h;p)$ is the MSE for the linear projection of x_{T+h} on $y_t(p)$. It is in general very difficult to specify $g_h^N(y_t(p)))$ so that it nests the optimal function $g_h^*(y_t(p))$. By assuming a more general parametric family for $g_h^N(y_t(p))$ is going to result in more estimation and specification uncertainty so that often $MSE(\widehat{g}_h^N(y_T(p))) > MSE(\widehat{g}_h^L(y_T(p)))$ even if $MSE(g_h^N(y_t(p))) < LMSE(h;p)$. Finally, it is not practically possible to apply all existing nonlinear parametric models in any study.

The interesting question is then whether there exists a sufficiently general and robust nonparametric estimation procedure that could find whether there is exploitable nonlinear predictability beyond the linear prediction approach. So far, this question is almost solely addressed by the ANN estimation strategy. In what follows, we examine whether the so called boosting estimator from machine learning literature could do the job.

3 The Boosting Approach

3.1 Basics

The boosting estimator is implicitly estimating an additive prediction function

$$g_h^{B(M)}(y_t(p)) = \sum_{m=1}^{M} b^m(y_t(p); \theta_m) \tag{4}$$

where $b^m(\cdot; \theta_m)$ are parametric functions. That is, the boosting estimator assumes that the function $g_h(y_t(p))$ in (1) is of the form (4). For simplicity, we have dropped the horizon index h from the component functions $b^m(\cdot; \theta_m)$ and from their number M, even if it is natural to assume that these differ by horizon.

A rationale behind the model in (4) is that with large enough M any conceivable function (from $\mathbb{R}^p$ to $\mathbb{R}$) can be expressed arbitrarily accurately in this form. A similar function approximation strategy is employed in the artificial neural network (ANN) method (see Friedman (2001, p. 1190)). For example, the single hidden layer feed forward model, the most common ANN model in econometric prediction applications, is obtained from (4) by adding a constant (μ) and specifying $b^m(y_t(p); \theta_m)$ by the model $\beta_m \sigma(\alpha_m + \gamma_m' y_t(p))$, where σ is a bounded "activation function," typically the sigmoid function. The ANN method tackles with two major problems: (i) how to estimate the $(1 + 2M + Mp)$-vector of parameters $(\mu, \alpha_1, ..., \alpha_M, \beta_1, ..., \beta_M, \gamma_1', ..., \gamma_M')'$ for a given M (the number of "hidden units" in the ANN terminology) and (ii) how to find how large M is needed for the approximation to be accurate enough. Despite several advanced ANN techniques have been applied, the success of the ANN method has been rather weak in the forecasting of macroeconomic time series (e.g., Stock and Watson (1999), Kock and Teräsvirta (2016)). The boosting method is a unique strategy to overcome a host of estimation and specification problems present in the ANN method and conventional parametric modeling approaches.[5]

In the boosting approach, the model (4) is "learned" from the available data gradually, term by term. The underlying algorithm can be shown to conduct numerical optimization

[5] While our intention is not to compare the boosting and the ANN methods thoroughly, our experiments indicate that the performance that we obtain with the boosting method in this paper cannot be reached by the ANN method. These results are reported in Appendix A.

in function space and is hence nonparametric. In the present setting, where the loss function ($MSE(g_h(y_t(p)))$) is quadratic, the population level target function is the conditional mean $g_h^*(y_t(p))$, the optimal prediction function we are after.

In the boosting estimation, one first chooses a parametric model $b(y_t(p);\theta)$ called the base learner. The boosting estimation algorithm for a sample of observations $(x_{t+h}, y_t(p)), t = 1, ..., T$, is as follows.[6]

Choose a positive integer M. Initialize the algorithm by solving the least squares (LS) problem

$$\widehat{\theta}_1 = \arg\min_{\theta\in\Theta} \sum_{t=1}^{T} (x_{t+h} - b(y_t(p);\theta))^2$$

and setting $\widehat{g}_h^{B(1)}(y_t(p)) = b(y_t(p);\widehat{\theta}_1)$ and $u_{1,t+h} = x_{t+h} - \widehat{g}_h^{B(1)}(y_t(p))$. Then compute the recursion

$$\widehat{g}_h^{B(m)}(y_t(p)) = \widehat{g}_h^{B(m-1)}(y_t(p)) + b(y_t(p);\widehat{\theta}_m),\ m = 2, 3, ..., M \tag{5}$$

where

$$\widehat{\theta}_m = \arg\min_{\theta\in\Theta} \sum_{t=1}^{T} (u_{m,t+h} - b(y_t(p);\theta))^2 \tag{6}$$

with

$$u_{m,t+h} = u_{m-1,t+h} - \widehat{g}_h^{B(m-1)}(y_t(p)) \tag{7}$$

The final boosting estimate can be written as

$$\widehat{g}_h^{B(M)}(y_t(p)) = \sum_{m=1}^{M} b(y_t(p);\widehat{\theta}_m)$$

where $b(y_t(p);\widehat{\theta}_m)$ represents an estimate for $b^m(y_t(p);\theta_m)$ in (4).

The accuracy of the "intermediate fit" $\widehat{g}_h^{B(m)}(y_t(p))$ in (5) improves, "boosts," by each iteration step m in the sense that $\sum_t \left(x_{t+h} - \widehat{g}_h^{B(m+1)}(y_t(p))\right)^2 \leq \sum_t \left(x_{t+h} - \widehat{g}_h^{B(m)}(y_t(p))\right)^2$ for all $m \geq 1$. This explains in part why the estimation method can potentially track the optimal function g_h^* even if it is very "complex" in form. Yet, the fact that the fit can only improve as M grows implies that the method can "overfit" such that the fit predicts x_{t+h} within the estimation sample better than the optimal prediction function g_h^* does. As $\widehat{g}_h^{B(M)}(y_t(p))$ deviates from g_h^*, the true out-of-sample prediction accuracy is worse than

[6] For a more general presentation of the algorithm see Bühlmann and Yu (2003).

in the case of the optimal prediction function, $MSE(\widehat{g}_h^{B(M)}) \geq MSE(g_h^*)$. This problem is basically present in all methods including those discussed above. The success of the boosting estimator rests on techniques that prevent it from overfitting.

For the boosting estimate not to overfit it is central that M, the number of boosting iterations, is selected properly. Most commonly, this is handled by a cross-validation (CV) procedure. In a k-fold CV, the available (estimation) data are divided into k portions of equal size. One of the folds is put aside at the time and the remaining $k-1$ folds are used to obtain the boosting estimate for a given M. For each of the k such boosting estimate, a measure of out-of-sample accuracy is computed by using data on the single fold that was put aside. The resulting k measures are averaged so as to obtain the final estimate for the out-of-sample accuracy of the boosting estimator based on M iterations. This is run for a given range of values of M, and the one yielding the best out-of-sample accuracy will serve as the estimate for the optimal number of iterations.

The performance of the above described procedure for estimating the optimal M is better, if the true out-of-sample accuracy of the boosting estimator changes only slowly as a function of M. This is sometimes called the "slow overfitting behavior" (see Bühlmann and Hothorn (2007)). Under slow overfitting, it is more likely that the estimated M will yield a prediction function that is close to the optimal one.

In general, slow overfitting is more likely when the base learner, $b(y;\theta)$, is "simple." Sufficient simplicity is usually attained by specifying the learner so that it belongs to a narrow class of functions and that it applies a small number of predictors at an iteration. As to a simple functional form, the commonly applied regression tree is a good option as it amounts to a piecewise constant function with a chosen maximum number of discrete jumps. A linear base model is also simple, but it has the potential disadvantage that the final boosting estimate is also restricted to be linear.

Among nonlinear parametric functions, smoothing splines have turned out to be particularly well performing as they allow one to restrict the base learner to be smooth of a given degree. Bühlmann and Yu (2003) find that a spline learner of a given degree of smoothness (and with fixed smoothing parameter) can be boosted to adapt to a higher-order smoothness and that the corresponding boosting estimate achieves the min-

imax optimal MSE rate of convergence. For a given degree of smoothness, the ordinary smoothing spline also achieves the optimal rate, but the boosting approach has the advantage that the near optimal region of boosting iterations is wider than the set of near optimal smoothing parameter of the ordinary smoothing spline estimator. We find the spline learner particularly well suited for the present application, where it is natural to assume that the underlying predictive effects are smooth rather than discrete.

In addition to having a simple functional form of the base learner, it is important that it is kept parsimonious in the sense that it uses only a few predictors (in our case, a subset of the lags $y_t, y_{t-1}, ..., y_{t-p+1}$) at the time. In this regard the most efficient and in fact the most common choice is the "componentwise" base learner, where only a single predictor is applied at an iteration. In this case, the baseline learner $b(y;\theta)$ is fitted for each of the lags y_{t+1-j} at a time. The fit (and the corresponding lag) that yields the smallest value of the criterion function in (6) is selected at an iteration m. The final boosting estimate can be written as

$$\widehat{g}_h^{B(M)}(y_t(p)) = \sum_{j=1}^{p} c_j(y_{t-j+1};\widehat{\gamma}_j) \tag{8}$$

where

$$c_j(y_{t+1-j};\widehat{\gamma}_j) = \sum_{m=1}^{M} b(y;\widehat{\theta}_{j,m}) I(y = y_{t+1-j})$$

where $I(y = y_{t+1-j})$ is one, if $b(y;\widehat{\theta}_{j,m})$ uses $y = y_{t+1-j}$, and zero otherwise.

With componentwise boosting some predictors may be left out of the model altogether. This automatic predictor selection property of the componentwise boosting estimator has been found to perform particularly well even if the number of applied predictors is very large (see Bühlmann and Yu (2003), Bühlmann (2006)). A restriction in the componentwise boosting estimator is that it does not allow the marginal predictive content of any given predictor to depend on the other predictors. This feature may sometimes rule out the true optimal prediction function. If this possibility is a concern, one can allow for "interactions" between the predictors such that the learner $b(y;\theta)$ is a function of at most k $(< p)$ predictors at the time. One can also impose sparsity to $b(y;\theta)$ by using penalized or regularized least squares estimation in place of the plain least squares (in (6)) or by applying a statistical information criterion to select predictors at each boost iteration.

To keep our analysis simple enough, we only use the componentwise spline learners in this paper, and leave it for later study to examine whether other base learners make a difference to our results.

Friedman (2001) recommends an additional regularization strategy where the base learner is multiplied by a shrinkage parameter ν, $0 < \nu \leq 1$, at each iteration step. Bühlmann and Yu (2003) adapt this strategy and conclude that boosting with a small value of ν is "safe" in the sense that the out-of-sample MSE increases very slowly if the optimal number of iterations is exceeded. While it would be possible to find optimal values for both M and ν, a more common strategy is to fix ν to a pre-chosen, "small enough" value and optimize only the number of iterations. Friedman (2001) conducts an empirical study of the performance of boosting with different values of ν and concludes that values smaller than $\nu = 0.125$ have diminishing returns in terms of the out-of-sample predictive performance of the model. The optimized value of M is usually larger the smaller the value of ν is.

3.2 Direct and Two-stage Boosting

As pointed out in the introduction, we consider two strategies for obtaining the boosting estimate in this paper. In the first of these, which we call the "direct boosting strategy," we apply the boosting estimator using x_{t+h} as the dependent variable and the lags $y_t(p)$ as the predictors. This procedure corresponds to the one used, e.g., in Robinzonov, Tutz and Hothorn (2012).

In theory, the direct boosting estimate, which we denote by $\widehat{g}_h^B(y_T(p))$, should agree with the linear autoregressive estimate, if the true optimal prediction function is linear. However, as the boosting method is nonparametric, it cannot reach the same estimation efficiency as the linear parametric procedure when the true model is linear. Thus, we expect that in finite samples the direct boosting method may yield less accurate forecasts than the linear method when the optimal prediction function is linear. Moreover, the linear modeling approach may often be more robust than the flexible nonparametric procedure when the true optimal model is sufficiently close to linear.

To keep up with the robustness of the linear estimation procedure, our second estimation strategy conducts the estimation in two stages. We first estimate the linear prediction model and then add a nonlinear component to it by applying the boosting estimator to its residuals. We call this strategy "two-stage boosting." Formally, the two-stage boosting estimate is

$$\widehat{g}_h^{LB}(y_T(p)) = \widehat{g}_h^{L}(y_T(p)) + \widehat{r}_h^{B}(y_T(p))$$

where $\widehat{g}_h^{L}(y_T(p))$ is the linear estimate described above and $\widehat{r}_h^{B}(y_T(p))$ is a boosting prediction for its residuals.

Effectively, the two-stage boosting and the direct boosting estimators are the same except that the former applies a linear autoregressive model at the first step ($m = 1$) of the boosting algorithm (see Section 3.1). The fact that the two-stage boosting estimator specifies $b^1(y_t(p); \theta_1)$ in (4) by a linear function makes no difference to the approximation capacity of the estimator. Moreover, the linear model (likewise the smoothing spline function) is regarded as a simple base learner (see Bühlmann and Yu (2003), Friedman (2001)). Hence, the two-stage estimator retains the slow overfitting property, the key strength of the boosting estimation technique. Based on these facts, the two-stage boosting estimator should be at least as good as the direct version in estimating nonlinear prediction models.

The advantage of the two-stage boosting procedure is its potential robustness when the true optimal model is linear. When the true model is linear, the linear learner in the first step of the algorithm assumes the right parametric model and hence enables estimating the optimal prediction model more efficiently than any alternative base learner. As a result, the first stage residuals should be merely noise with no significant serial dependence. One then expects that the subsequent boosting iterations do not drive the final fit too much away from the first-stage estimate. Here, the overall resistance of the boosting method against overfitting is particularly useful. We think that it is also likely that the same effect applies when the optimal model is so close to linear that one gets more accurate predictions by using a linear parametric model rather than a more general nonparametric approach. On a part, our conjecture gets support from simulations where we find that when the optimal model is linear or nearly linear, then the final number of boosting iterations is much smaller for the two-stage method than for the direct method.

We, however, also conjecture that the direct boosting procedure may be better than the two-stage method at estimating certain types of nonlinear models. In particular, if the true optimal nonlinear model uses just a single or a few lags, it may happen that the first stage linear model uses more lags than the optimal model. Such extra lags may be a burden and cause bias to the boosting estimation at the second stage. We obtain some evidence on this effect in our simulations even if our overall conclusion from the simulations and the empirical analysis will be that the two-stage method is generally more robust than the direct method.

A naturally arising alternative for the two-stage boosting approach described above would be a procedure, where the first stage estimation is conducted by the boosting estimator using linear componentwise base learners. Such a boosting estimate in the first stage is necessarily linear and hence could capture the true optimal linear model before the second stage estimation. However, when the true model is linear, the conventional linear approach conducts its estimation more efficiently than the iterative boosting method. The linear boosting method might still gain some advantage from its good performance at selecting the right predictors (lags). Nevertheless, in our simulation experiments, the linear boosting estimator is outperformed by the linear autoregressive estimation approach where the lag order is selected using BIC. Moreover, we find that the two-stage procedure that uses the conventional linear method at the first stage is more accurate out-of-sample than a "double boosting procedure," where the first stage estimation is conducted by the boosting estimator using linear base learners. In view of these results, we proceed using the two-stage boosting strategy. The results of the simulations with the double boosting approach are available upon request.

Finally, our two-stage boosting strategy resembles an estimation procedure proposed in Taieb and Hyndman (2014). In the procedure of Taieb and Hyndman (2014), the first stage involves estimating a conventional AR(p) model (the "first-step ahead model") and iterating it to obtain $\widehat{g}_h^L(y_T(p))$ for each h. Otherwise, the second stage part is essentially the same. The "iterative" linear prediction used in Taieb and Hyndman (2014) and the "direct" linear prediction used in the present paper attempt to estimate the same target model. Which one of the two approaches is better as such has been investigated

theoretically e.g. by Shibata (1980), Findley (1983, 1985), Weiss (1991), Bhansali (1996, 1997), Clements and Hendry (1996), and Ing (2003). The general conclusion from these studies is that the iterative approach is more efficient, if one manages to specify the first-step ahead model correctly, while if not, then the direct approach is more robust. The aforementioned papers emphasize the robustness of the direct approach. As our initial hypothesis is that macroeconomic time series are inherently nonlinear, it seems likely that the linear model is deemed to be misspecified in most cases. Hence, we find it natural to apply the direct rather than the indirect method in the first stage estimation. Nevertheless, as both methods have their theoretical pros and cons and as we do not know what conditions apply to a given series, we have also examined the performance of a "recursive two-stage boosting" procedure that applies the iterative linear AR prediction in the first stage. The results are reported in Appendix A and amount to the following main conclusions. In the simulations, we find no significant difference between the direct and the indirect procedures, while in the empirical comparisons the direct method is on average more accurate than the indirect approach.

4 Simulation Study

We conduct a simulation study to examine how the boosting prediction approach compares with the conventional linear prediction approach and whether the direct or the two-stage boosting approach has some advantage in specific settings. We start by describing the simulation set-up.

4.1 The Set-up

We simulate independent samples (sequences of observations) from the q-th order Markov process

$$y_t = g(y_{t-1}(q)) + \varepsilon_t \tag{9}$$

where ε_t is a zero mean iid series. We report results based on the models shown in Table 1. These models are sufficient for demonstrating our main findings that we obtained from

a larger set of non-linear and linear models.

In the simulation experiment, we consider a situation where the h-period ahead prediction is based on the estimation of the model (1). The choice of p, the number of applied predictors, is taken as given. In the simulations, we let $p = 12$ so that the examined methods could, at least in principle, find the best possible prediction model given by

$$x_{t+h} = g_h^*(y_t(q)) + u_{t+h}$$

where $x_{t+h} = y_{t+h}$, $g_h^*(y_t(q)) = E(x_{t+h}|y_t(q))$ and $u_{t+h} = x_{t+h} - g_h^*(y_t(q))$. When the regression function g in (9) is nonlinear, the corresponding optimal h-period ahead prediction function $g_h^*(y_t(q))$ tends be complicated and more so with larger h. However, as g and the distribution of ε_t are known to us, we can approximate $g_h^*(y_t(q))$ arbitrarily accurately by numerical and simulation methods (see Tong (1990)).

We analyze and compare the performance of the linear estimation approach described in Section 2 and the boosting approach described in Section 3. In setting the hyperparameters of the boosting estimators, we follow the recommendations of Schmid and Hothorn (2008), among others. The shrinkage parameter (learning rate) is fixed to $\nu = 0.1$. The base learners are component-wise cubic P-splines with 20 knots and degrees of freedom $df = 4$. A maximum of 300 training iterations is done for each model. The optimal number of iterations is determined through 10-fold CV. All boosting calculations are done with the mboost R package.

The linear approach yields for each simulated sample the estimate $\widehat{g}_h^L(y_t(p))$ of the best linear prediction function denoted by $g_h^+(y_t(p))$. The boosting estimation approaches yield the estimates $\widehat{g}_h^{B(M)}(y_t(p))$ and $\widehat{g}_h^{LB(M)}(y_t(p))$. We regard the optimal prediction function $g_h^*(y_t(q)) = g_h^*(y_t(p))$ $(q \leq p)$ as their asymptotic target function.

For each estimation approach, we compute the (out-of-sample) coefficient of determination of the estimated prediction model. This measure is defined by

$$R^2(\widehat{g}_h(y_T(p))) = 1 - \frac{MSE(\widehat{g}_h(y_T(p)))}{\text{var}(x_{T+h})}$$

where $MSE(\widehat{g}_h(y_T(p)))$ is the (out-of-sample) MSE of the estimated prediction for x_{T+h} and $\text{var}(x_{T+h})$ is the variance of x_{T+h}. Clearly, we have $R^2(\widehat{g}_h(y_T(p))) \leq 1$, but unlike for

a conventional R-squared, $R^2(\widehat{g}_h(y_T(p)))$ may be negative. The latter possibility arises, because $\widehat{g}_h(y_T(p))$ may have poorer predictive performance than the optimal constant prediction $E(x_{T+h})$. We report negative values of $R^2(\widehat{g}_h(y_T(p)))$ as zeros. Note that $R^2(\widehat{g}_h(y_T(p)))$ depends on the sample size T.

A natural reference point for $R^2(\widehat{g}_h^L(y_T(p)))$ is the coefficient of determination of the linear projection model

$$R^2(g_h^+(p)) = 1 - \frac{LMSE(h;p)}{\text{var}(x_t)}$$

We have $0 \leq R^2(g_h^+(p)) \leq 1$ and can interpret $R^2(g_h^+(p))$ as the degree of linear predictability at horizon h. Clearly, $R^2(\widehat{g}_h^L(y_T(p))) \leq R^2(g_h^+(p))$.

As the measure of overall predictability we use the coefficient of determination of the optimal prediction model

$$R^2(g_h^*(p)) = 1 - \frac{MMSE(h;p)}{\text{var}(x_t)}$$

This measure is also within the interval $[0, 1]$ and tells for a given horizon the maximum predictability that can be attained from the predictors $y_t(p)$. Given that the boosting strategies have the capacity (at least asymptotically) to yield the optimal prediction function $g_h^*(y_t(p))$, $R^2(g_h^*(p))$ is a natural yard stick for $R^2(\widehat{g}_h^{B(M)}(y_T(p)))$ and $R^2(\widehat{g}_h^{LB(M)}(y_T(p)))$.[7]

Finally, note that due to stationarity and ergodicity of the process, $R^2(g_h^*(p)), R^2(g_h^+(p)) \to 0$, as $h \to \infty$. Despite this, we may have $R^2(g_{h_1}^*(p)) < R^2(g_{h_2}^*(p))$ or $R^2(g_{h_1}^+(p)) < R^2(g_{h_2}^+(p))$ for some horizons $h_1 < h_2$.

4.2 Results

We divide the simulation results into two sub-sections. In Section 4.2.1 we assess the overall performance of the boosting estimation procedure and compare it to that of the

[7] Earlier we noted that when the optimal model is nonlinear using $k < p$ lags its best linear approximation may use more than k lags. In this situation, one expects that the linear estimation part, the first step, of the two-stage estimator $\widehat{g}_h^{BL(M)}(y_T(p))$ selects too many lags and hence might not be consistent for the optimal model. However, at least asymptotically, the second stage boosting estimation part of the estimator $\widehat{g}_h^{BL(M)}(y_T(p))$ has the capacity to adjust the estimated function so as to off-set the extra lags resulting from the first step estimation.

linear estimation approach. The comparisons are conducted in terms of the R^2 measures. In Section 4.2.2, we compare the performances of the two boosting strategies in terms of their relative MSEs.

All of our reported simulations assume the sample size $T = 500$ that represents roughly the length of the estimation samples in our empirical analysis. We have run the same simulations using sample sizes $T = 200$, $T = 300$ and $T = 1000$. The result from these additional simulations (available upon request) are qualitatively similar to those reported here.

4.2.1 Boosting vs. Linear Procedure

Figure 1 illustrates the performance of the boosting estimation procedure and the linear estimation approach when the prediction is made for $x_{t+h} = y_{t+h}$ and the sample size is 500. The estimation procedures are allowed to use the twelve lags, $y_t, ..., y_{t-11}$ as predictors (that is, $p = 12$). The lag order $\widehat{p}$ ($\widehat{p} \in \{0, 1, ..., 12\}$) of the linear model is selected by BIC. The 5-fold CV is applied to select the number of boosting iterations.

The illustration in Figure 1 is in terms of the value of the coefficient of determination of the estimated prediction functions for horizons $h = 1, 2, ..., 12$. On the part of the boosting estimator, the figure depicts only graphs for the direct boosting estimator, because the corresponding results for the two-stage boosting estimator are virtually the same in all cases. That is, in these simulations, we can hardly see a difference in the R^2's of the two boosting strategies. Nevertheless, when we compare the two boosting strategies by the MSE ratio, we will find that there are some recognizable differences and we will analyze these in the subsequent section.

The red (blue) line is R^2 for the boosting (the linear) prediction procedure. The black dotted line with star markers indicate the coefficient of determination of the optimal prediction, or the degree of predictability of the process, at each horizon. This is the yard stick against which we can assess the performance of the boosting prediction. The black dotted line with plus markers indicate the coefficient of determination of the best linear prediction based on all twelve lags. For the simulated models, this measure is virtually the same as the one based on the whole past of the series ($p = \infty$). Hence, the line with

plus markers tells what we can expect at best from a linear prediction procedure.

The results of panel (a) of Figure 1 are for a simple first order threshold autoregression. The performance of the boosting prediction is pretty close to the optimal performance for horizons 1 to 4. Also, the performance of the linear prediction is hardly distinguishable from its reference point (the dotted line with plus markers) for the whole range of horizons. When the horizon is longer than 4, the boosting and the linear methods perform equally well. This is as expected, because the corresponding population level reference points are very close to each other and the overall degree of predictability is low from horizon 5 on.

In the case of the second model (Figure 1, panel (b)), the two prediction approaches attain well their reference points for horizons 1 and 2. However, both procedures have increasing trouble in catching up with their maximal performances for longer prediction horizons. For $h \geq 6$, the performance of both predictions is worse than that of the optimal constant prediction. In this example, the boosting estimator is better option for horizons 1 to 5, but the fact that both procedures go astray at longer horizons suggests that one should somehow try to detect whether the estimated prediction function in hand is indeed any better than the simple constant prediction.

The process behind the results of panel (c) of Figure 1 is more predictable than the previous cases. The performance of the boosting estimator is again not far from the optimal performance at horizons 1 to 6 and it remains at least as good as the linear prediction for longer horizons. In the case of this process, one loses nothing by using the boosting estimator throughout.

The results of panels (d) and (e) are for models where the optimal prediction is a nonlinear function of two predictors (lags). In panel (d), both estimators perform nearly as accurately as their reference points. The advantage of the boosting estimator is considerable for horizons 1 to 6. Thereafter, for $h \geq 7$, the difference between the performances of the two methods is less marked. The process behind panel (e) is again very predictable with the optimal coefficient of determination being close to 1 at horizons 1 and 2. The difference between the optimal and the best linear performance is also very large for all horizons. The boosting estimator performs again better than the linear approach although it gradually drops from its reference point level towards the performance of the

linear predictor.

Panel (f) in Figure 1 concerns a simple AR(3) process.[8] As is expected, now the linear prediction approach is better than the boosting estimator. The flexibility of the direct boosting estimator causes it to overfit although in this case the method does not lose much compared to the linear approach. When $h \geq 4$, there is very little predictability and both prediction approaches are worse than the optimal constant predictor.

Figure 2 shows results that correspond to those of Figure 1 with the exception that the underlying predictions are for $x_{t+h} = \sum_{j=1}^{h} y_{t+j}$ rather than for $x_{t+h} = y_{t+h}$. Hence, in Figure 2, we can interpret that there is an underlying original time series $Y_t = \sum_{s=1}^{t} y_s$, an I(1) process, and that the predictions are for the h period ahead change $Y_{t+h} - Y_t = \sum_{j=1}^{h} y_{t+j}$. The results in Figure 2 are particularly interesting from the point of view of our empirical application, where the majority of the original series are classified as I(1).

The profiles of the graphs of the R^2 measures in Figures 1 and 2 are quite different in most cases. For example, for the threshold autoregression (panel (a) in the figures), the degree of overall predictability is stronger, and the gap between the optimal and the linear predictability is larger, for all horizons ($h > 1$) when the predicted outcome is the h-period change ($x_{t+h} = Y_{t+h} - Y_t$) of an underlying original series. A similar observation applies to the nonlinear model of panel (c). There we also see that for the h-period change the degree of predictability is stronger for horizons 2 to 4 than for horizon 1. The model behind panel (b) is more extraordinary. In Figure 1, the degree of predictability dampens geometrically, while in Figure 2 it declines with an oscillating pattern as a function of h. For this case, the gap between the linear and the nonlinear (optimal) predictability is not large for any horizon, while there is more overall predictability at longer horizons, when the prediction is made for the h-period change. For the model of panel (d), the degree of predictability is almost uniform over the horizons when the prediction is for the h-period change. For panel (e), the wave-like pattern in Figure 2 is quite different compared to the monotone decreasing pattern in Figure 1. Finally, for the linear model (panel (f)) the profile of the R^2's are fairly similar, though there is more to predict in terms of the

[8] We obtained the underlying coefficients from an AR(3) model that we specified and estimated for a series in our empirical data set, "All Employees: Total Nonfarm."

average difference for any horizon $h > 1$.

Overall the results of Figures 1 and 2 demonstrate that the boosting estimator can beat the linear prediction approach very well in various types of settings where there is nonlinear predictability beyond the best linear approximation. Sometimes the gain in relative terms from using the boosting approach is considerable and sometimes it is minor compared to the linear estimate. In general, the linear estimate is closer to its reference point than the boosting estimator is to its reference point, that is, the optimal prediction. A natural explanation for this difference in the two techniques is that as a very flexible nonparametric modeling approach the boosting estimator can overfit more easily than the linear estimator.

The relative performance of the two boosting estimators remains the same with varying estimation data lengths. As one may expect, the relative performance of the boosting procedures to the linear prediction approach weakens when the sample size is smaller. When the true model is linear (model (f)) and the sample size is 200, the MSE of the direct boosting estimator is about 10% larger than that of the linear modeling approach. For the nonlinear models, for which the MSE ratio of the optimal prediction to that of the linear approximation is less than 0.9, the boosting procedures are more accurate than the linear approach for all experimented estimation data lengths.

4.2.2 Direct vs. Two-stage Boosting Procedure

In this section, we compare the direct and the two-stage boosting strategies. We find it most informative to conduct the comparison in terms of their MSEs in relation to the MSE of the linear method

$$\frac{MSE(\widehat{g}_h^P(y_T(p)))}{MSE(\widehat{g}_h^L(y_T(p)))}$$

where $P \in \{B, LB\}$. The results are presented in Tables 2 and 3. The tables also include the MSE ratio between the optimal prediction and the one of the best linear approximation (i.e., $MMSE(h;p)/LMSE(h;p)$).

The results in Tables 2 and 3 demonstrate a small but consistent difference in the behavior of the two boosting strategies. The direct boosting procedure is able to model

a nonlinear prediction function slightly more accurately than the two-stage procedure. On the other hand, the direct boosting produces a slightly worse estimate for a linear prediction function. In cases, where the optimal prediction is linear or "close to linear" (i.e. the ratio $MMSE(h;p)/LMSE(h;p)$ is almost one), both boosting procedures perform slightly worse on average than the pure linear prediction procedure.

The following factors seem to drive this result. First, in a close examination of the estimated models, we find that the direct boosting procedure is more parsimonious (in terms of the number of predictors chosen for the model) than the other methods, when the optimal prediction function is non-linear, but tends to overfit, when the optimal model is linear. Second, in the case of the direct boosting strategy, the estimated spline functions tend to be non-linear even if the underlying population level target function is linear. On the other hand, when the optimal prediction is linear, the first stage of the two-stage boosting strategy tends to capture the most of it and the second stage boosting part does not usually alter the first stage estimation result much.

In summary, the choice between the two boosting strategies boils down to a trade-off between flexibility and robustness. On the one hand, the two-stage approach appears to be more resistant to overfitting when the true optimal prediction is linear or when the degree of predictability is low (the optimal R^2 is close to zero). On the other hand, the direct boosting strategy can sometimes model non-linear prediction functions slightly more accurately than the two-stage procedure. In what follows, we examine the relative importance of these trade-offs in the forecasting of macroeconomic variables.

5 Empirical Analysis

In this section, we analyze how well the boosting and the linear methods predict the series of the FRED-MD data set.

5.1 Procedures

The empirical prediction problem is similar to what we have studied above. For a given monthly series Y_t from the FRED-MD data set we wish to estimate the optimal prediction

function for horizons $h = 1, ..., 12$. We treat the original time series in the same way as in the analysis of McCracken and Ng (2016). Hence, Y_t is most often the logarithmic form of an original time series in the data set. Moreover, based on the evidence of McCracken and Ng (2016), we regard Y_t as either I(0), I(1) or I(2) process and accordingly formulate the h-period ahead prediction model (with response x_{t+h} and predictors $y_t, y_{t-1}, ..., y_{t-p+1}$) as was described in Section 2.

For the present analysis, we use a vintage of the FRED-MD data set that was available in early 2018. The data set contains 128 series and all of these are named in the data appendix of McCracken and Ng (2016).[9] Concerning each series in the data set, we apply observations from the first month (typically January 1959) until December 2016 (available for all series in the data set). The aforementioned data appendix indicates the first observation month for each series as well as the applied transformation (whether the original series is used as such or in logarithmic form) and whether the series is regarded as I(0), I(1) or I(2).

We compare the three methods by a "simulated" out-of-sample (SOOS) analysis (similarly as in many previous studies, e.g., Stock and Watson (1999). Denote by T_1 (T_2) the number of observations at the first (the last) month at which the simulated out-of-sample prediction is made. In our baseline analysis, the corresponding months are December 1998 (T_1) and December 2015 (T_2). For most series, $T_1 = 480$ and $T_2 = 684$. We interpret that our SOOS results are representative for a situation where the sample size, T, is roughly the average of T_1 and T_2, or casually, $T = 500$. With this idea in mind, we denote the corresponding estimates by $\widehat{g}_h^B(y_T(p))$, $\widehat{g}_h^{BL}(y_T(p))$ and $\widehat{g}_h^L(y_T(p))$.

In the SOOS analysis, we estimate $MSE(\widehat{g}_h^P(y_T(p))) = E(x_{T+h} - \widehat{g}_h^P(y_T(p)))^2$ for $P \in \{B, LB, L\}$ by the simulated out-of-sample MSEs

$$\widehat{MSE}(\widehat{g}_h^P(y_T(p))) = \frac{1}{T_2 - T_1} \sum_{t=T_1}^{T_2} \left(x_{t+h} - \widehat{g}_h^P(y_t(p))\right)^2$$

Here, the estimates $\widehat{g}_h^P(y_t(p))$ are generated recursively in the standard fashion with the exception that we update the estimation only once per year. Hence, the first estimation

[9] We note that six of the series in the original FRED-MD data set are not in this vintage of the data set.

is run using observations until T_1 and the corresponding estimates are applied to make predictions at periods $t = T_1, T_1+1, ..., T_1+12$ for each $h = 1, ..., 12$. The second estimation round is made using a sample that ends at period $T_1 + 12$, and so on.

As in the simulation study, we find it informative to consider the empirical (or simulated) out-of-sample coefficient of determination

$$\widehat{R}^2(\widehat{g}_h^P(y_T(p))) = 1 - \frac{\widehat{MSE}(\widehat{g}_h^P(y_T(p)))}{\widehat{\text{var}}(x_{T+h})}$$

where $\widehat{\text{var}}(x_{T+h})$ is the sample variance based on the out-of-sample observations x_{t+h}, $t = T_1, ..., T_2$. Again, $\widehat{R}^2(\widehat{g}_h^P(y_T(p)))$ is regarded as an estimate for the population level counterpart $R^2(\widehat{g}_h^P(y_T(p)))$, where $\widehat{MSE}(\widehat{g}_h^P(y_T(p)))$ and $\widehat{\text{var}}(x_{T+h})$, respectively, is replaced with $MSE(\widehat{g}_h^P(y_T(p)))$ and $\text{var}(x_{T+h})$.

Finally, to compare the methods in relative terms we use the empirical MSE ratio

$$\frac{\widehat{MSE}(\widehat{g}_h^{P_1}(y_T(p)))}{\widehat{MSE}(\widehat{g}_h^{P_2}(y_T(p)))} \tag{10}$$

where $P_1, P_2 \in \{B, LB, L\}$, $P_1 \neq P_2$ indicate the compared methods (e.g., $P_1 = B$, $P_2 = L$). When the ratio in (10) is less than 1, the procedure P_1 is estimated to be more accurate than the procedure P_2.[10]

[10] A standard practice would be to use the Diebold-Mariano test to determine whether the differences in the forecast accuracy are statistically significant. However, we refrain from doing this as the DM test assumes that the tested models are non-nested. In our simulations, for example, the size of the DM test is less than 1% for the nominal size of 10% and the power for the non-linear alternatives is rather poor. In Appendix B, we report results on the unconditional version of the Giacomini and White (2006) test that does not entail knowing whether the alternative model nests or does not nest the null model. A problem with the GW test is that its validity rests on the assumption of "nonvanishing estimation errors" and hence it is designed for situations where the underlying simulated out-of-sample prediction errors are obtained by using a fixed (or a rolling) window rather than an expanding window estimation scheme applied here. Hence, the results in Appendix B should be interpreted with caution. Moreover, if we applied a fixed window estimation scheme, it might not reveal the true potential of the boosting estimator that as a nonparametric method calls for using as much estimation data as possible.

5.2 Main Results

We start by considering how the three procedures compare in terms of their empirical out-of-sample R^2's, or $\widehat{R}^2$'s, for all 12 horizons. Table 4 presents the average of the $\widehat{R}^2$'s for all 128 series (panel a), for series where the boosting method is more accurate (panel b), and for ones where the linear approach is more accurate (panel c). Within each panel separate results are given for the subsets where the maximum of $\widehat{R}^2$ over the three methods is at least 0.1. This is motivated by the fact that the optimal prediction is necessarily close to linear when there is little predictability (i.e., when the true out-of-sample R-squared is small).

The general observation from panel (a) of Table 4 is that on average the linear and the two-stage boosting procedures perform similarly, while the direct boosting procedure is clearly inferior to both. This result holds whether the averages are for all 128 series or for the subset of series where we require that $\widehat{R}^2$ is at least 0.1 for one of the methods. The observation is largely explained by the fact that in many cases where the linear approximation seems to yield sufficient accuracy the direct boosting procedure tends to overfit, while the two-stage procedure does not.

The results in panel (b) of Table 4 show how much more the boosting method is more accurate when it is more accurate than the linear procedure. An interesting observation is that in these cases the two-stage boosting procedure is on average more accurate than the direct boosting procedure with the exception of the one step ahead prediction $h = 1$ where the two methods are essentially equally accurate. In relative terms, the advantage of the two-stage boosting over the linear procedure is quite marked for horizons $h > 1$. The overall picture is quite similar in the baseline case and in the subset where the maximum of $\widehat{R}^2$ over the three methods is at least 0.1.

The results in panel (c) of Table 4 show how much more the conventional linear procedure is more accurate when it is more accurate than the boosting procedure. Note that for horizons from 1 to 7 the boosting method yields on average more accurate predictions than the linear procedure. Moreover, when the linear prediction is more accurate its relative advantage over the boosting method is not as large as the corresponding relative

gain of the boosting method (shown in panel (b)). However, for horizons longer than 7 the gain of the linear procedure over the boosting method is about as marked as the gain of the boosting method at its best cases (in panel (b)).

In summary, the results of Table 4 indicate that the two-stage boosting method is the preferred method among the three procedures for horizons from 1 to 7, while for longer horizons it is more up to the series whether the two-stage or the linear procedure is better. That is, for horizons from 1 to 7, the two-stage boosting method seems fairly robust in that it is more accurate for most series and when it is less accurate than the linear procedure it is so only slightly. It seems that the two-stage boosting method can somehow utilize the best parts of the boosting and the linear procedure. For horizons longer than 7, there are more cases where the linear procedure is more accurate than the two-stage procedure.

Overall, we find that there are 14 (28) series where the two-stage boosting procedure is more accurate than the linear forecast for all 12 (for at least 10) forecasting horizons. Conversely, there are 0 (10) series for which the linear method is more accurate for all 12 (for at least 10) horizons. These observations support the view that for a given macroeconomic series it is more likely that the two-stage boosting procedure yields more accurate predictions than the linear procedure. A general conclusion is also that it is beneficial to apply the boosting prediction procedure to 30-40 percent of the U.S. macroeconomic series that have at least some predictability in their own history. For the linear procedure, the corresponding percentage is between 10 and 25. For the rest of the series the two methods are essentially equally accurate.

Figures 3 and 4 plot $\widehat{R}^2(\widehat{g}_h^L(y_T(p)))$, $\widehat{R}^2(\widehat{g}_h^B(y_T(p)))$, and $\widehat{R}^2(\widehat{g}_h^{BL}(y_T(p)))$ over $h = 1, ..., 12$ for 12 series from the FRED-MD data base. The series in the figures represent cases, where at least one of the boosting methods performs better than the linear procedure for at least 10 horizons. Interestingly, the graphs in the figures are somewhat similar to those that we obtained in our simulations (see Figures 1 and 2). For these series, there is no clear difference between the boosting strategies. An exception is the series in Figure 4(e) for which the performance of the two-stage procedure start to decline when $h > 3$. This suggests that the direct boosting can sometimes capture the underlying nonlinear-

ity better than the two-stage procedure. In 59% of the cases where the direct boosting forecast is more accurate than the linear one, it is also more accurate than the two-stage boosting forecast. However, the direct boosting forecast is more accurate than the linear forecast only in 37% of all cases. Thus, to safely reap the potential benefit from the direct boosting strategy, we should have a testing procedure to assess in advance whether its out-of-sample performance is indeed superior to that of the linear method.

5.3 The Impact of Financial Crisis

We are aware that in the present sample the U.S. economy experienced a very turbulent period at and after the financial crisis in 2008. It is quite likely that a large share (if not all) of the series in the FRED-MD data set are subject to some structural breaks during these times and the return to "normal times" may have taken a long time depending on the series. To see whether such prolonged, but perhaps temporary, deviations from normal times make a difference to our conclusions we present in Table 5 results that are as in Table 4 except that the out-of-sample period is restricted to the pre-crisis years 1999-2007 (that is, now the last estimation sample ends in December 2006 so that for most series $T_2 = 564$).

The results of Table 5 are generally similar to those in Table 4 except that now the two-stage boosting method is on average more accurate at all horizons (panel a). Among cases, where the two-stage boosting is more accurate than the linear procedure (panel b), it is more so than it is in Table 4. Moreover, for horizons from 1 to 11 the two-stage boosting predictions are most accurate for the majority of the series. For horizon 1 (7) [11], the boosting based prediction is more accurate for 28 (43) [40] series and the linear method is more accurate for 9 (17) [20] series. Overall, the results of Table 5 are even more in favor of the two-stage boosting method than the results of Table 4.

For the pre-crisis out-of-sample period we also find that there are 30 series for which the two-stage boosting procedure is more accurate than the linear procedure for at least 10 horizons, while in the case of the linear procedure corresponding count is only 4. These observations give additional support to the view that the two-stage boosting procedure is

the preferred method for predicting the majority of macroeconomic series at least in the absence of major structural breaks.

The above findings suggest that the boosting method may lose at least a part of its relative advantage when the underlying series is subject to a temporary break. In a closer examination, we find a natural mechanism through which the boosting based predictions tend to fail during the period of the 2008 financial crisis. The key observation is that as a result of the financial crisis some series have taken on values that have hardly never realized in the past. As these outliers are not in the "training data," the boosting method has to extrapolate them. When the boosting procedure extrapolates the estimates of the underlying nonlinear spline functions the resulting prediction tends to "overshoot" in that its value is even further away from the historical area of the variation of the series. The conventional linear method is more moderate in this type of occasions and hence does not overshoot so easily.

To examine our conjecture on the spline extrapolation issue, we device "hybrid forecasts" that are equal to the boosting forecast (either direct or two-stage) with the exception that we replace their predictions with the conventional linear forecast at every point where spline extrapolation would be needed. In the case of the two-stage procedure this means that the second stage refinement is skipped when it requires extrapolation. The results using the hybrid versions of the boosting procedures are presented in Table 6 for the whole out-of-sample period 1999-2016.

When we compare the results in Tables 4 and 6 we make the following main observations. First, in Table 6, there are more cases where the boosting method is more accurate than the conventional linear procedure. Secondly, the differences in accuracy between the two boosting strategies are smaller in Table 6 than in Table 4. Finally, in cases where the linear procedure is better than the two boosting procedures it is much less so in Table 6. In particular, the hybrid two-stage boosting procedure is on average at least as (or only slightly less) accurate as (than) the conventional linear procedure in Table 6. We also find that there are only 3 series for which the pure linear procedure is the most accurate for at least 10 horizons. On the other hand, even in these cases, the hybrid two-stage procedure is essentially equally accurate. Hence, overall, based on the results of Table 6,

we may conclude that the hybrid two-stage boosting procedure is the most robust method for predicting the series in the FRED-MD data set.

A comparison of the individual series where two-stage boosting is more accurate among different subsamples reveals an important fact. If one looks at the list of individual series where the two-stage boosting method is more accurate for the majority of forecasting horizons, the pre-crisis list contains about 75% of the series in the full sample list. Or alternatively, for about 60% of the series where the boosting forecast is more accurate for the post-crisis period (2009-2015), the same is true for the pre-crisis period as well. This suggests that nonlinear predictability is in some cases an inherent feature of the time series that is preserved over a major shock.

5.4 Category-specific Findings

It may also be of interest to drill down deeper and look at the average empirical R^2 values by variable category. The average per-category $\widehat{R}^2$ values for the "hybrid two-stage boosting" method are presented in Table 7. It appears that the forecasting accuracy improvement from boosting is concentrated on designated variable categories. On average, the variables in the *Output and Income* category benefit the most. *Labor Market* variables benefit from boosting especially in the short-term forecasts ($h = 1, ..., 6$), while there is also some improvement in the *Orders and Inventories* and *Stock Market* categories for the longest forecasting horizons. In *Output and Income* and *Orders and Inventories* categories many variables are just weakly predictable by the linear procedure, while boosting improves the accuracy significantly. Similar cases can also be found among *Labor Market* variables. Figures 3 and 4 illustrate some examples in more detail.

The variables in *Interest and Exchange Rates* category are hardly predictable by their own history, but interest rate spreads can be predicted rather well using the linear method. For some of the spread variables the boosting approach improves the forecast by 10-20%. Also in *Money and Credit* category, some individual variables benefit significantly from the use of boosting, but on average, there is no clear improvement.

The housing variables in *Consumption and Orders* category suffer the most from the

aforementioned spline extrapolation issue. If we look at the pre-2007 forecasts, the boosting estimators can reduce the forecasting MSE up to 25% for the longer forecasting horizons. After 2008, the boosted forecasts are inferior to the linear forecast, while the MSEs of the hybrid forecasts are equal to that of the linear forecast.

Grouping the variables by the order of integration of the original variable, the $I(1)$ series usually benefit most from the boosting approach, and this observation applies to both boosting strategies. On the contrary, $I(2)$ series (most of which are in *Prices* category) seem not to benefit from boosting and the direct boosting approach produces forecasts that are less accurate than the linear forecasts.

Hypothesizing that some of the $I(2)$ variables may actually be over differenced, we experimented with treating them as $I(1)$ instead. To compare the different forecasts, we reversed the transforms to obtain the forecasts for the original variable (Y_{t+h}) and calculated the forecasting MSE for the different variations. We found that when using the $I(1)$ transform, direct boosting produces on average the most accurate forecasts for the transformed variable while two-stage boosting is better than the linear model by a smaller margin. In 65% of the cases, the most accurate forecast for Y_{t+h} can be obtained by using the $I(1)$ transform and direct boosting. In 17% of the cases, the most accurate model is the linear model with the $I(1)$ transform. Altogether modeling the variables as $I(2)$ produced the most accurate forecast for Y_{t+h} in less than 10% of all cases. Thus it seems that in many cases the double differencing may be unnecessary and the single differenced time series has non-linear predictability exploitable by the boosting procedure. The detailed results for this exercise are available upon request.

6 Conclusion

We have applied the boosting estimation technique to examine whether macroeconomic time series have exploitable nonlinear predictability beyond what is obtained by using a conventional linear prediction method. We motivated the boosting approach by the following points. The method is nonparametric and hence not restricted to a specific parametric model family. It has been shown to have advantages over alternative non-

parametric techniques and it has proved to be superior in various previous studies in the machine learning literature. It is handy and reliable to use in practice.

We first conducted a simulation study to get an initial assessment of the performance of the boosting method and how it compares with the conventional linear method. We attempted to design the simulation set-up so that it could resemble our empirical setting. The simulations demonstrated that the boosting method indeed has capacity to exploit nonlinear predictability when it is present. On the other hand, we also learned that for the boosting method to be useful it is often required that the true optimal prediction model is clearly more accurate than its best linear approximation.

Our empirical analysis concerned 128 monthly macroeconomic time series available in the FRED-MD data set, originally provided and introduced in McCracken and Ng (2016). We showed that for a significant share of variables in this data set the boosting estimation method can improve the accuracy of forecasts over a conventional linear prediction approach. We also identified variable categories where nonlinear modeling is likely to produce the largest benefits. An issue is however that despite the boosting estimator is known of its resistance to overfitting this cannot be avoided with certainty. As a result, there are series for which the "pure" or "direct" boosting prediction is inferior to the one based on the linear method. As there is currently no reliable method for diagnosing when this happens, we considered a "two-stage" boosting procedure, where the conventional linear prediction model is fine-tuned by applying the boosting technique to its residuals.

The two-stage boosting method turned out to have the best out-of-sample performance in our empirical analysis. We found that it is almost always at least as accurate as the direct boosting procedure, while it is rarely less accurate than the linear method. When the two-stage boosting method is combined with a conventional procedure for handling outliers, it is virtually never inferior to the linear method when we consider 1 to 12 month ahead predictions for the variables in the FRED data set. We are willing to conclude that the most robust approach to exploiting nonlinear predictability in the forecasting of macroeconomic variables is to apply a flexible nonlinear procedure not as such, but as a device to add a nonlinear component to the linear autoregressive model.

Our results suggest that while the direct boosting procedure is more sensitive to over-

fitting it can sometimes reap nonlinear predictability (in excess of the linear model) better than the two-stage procedure. Hence, an important topic for future research is to develop a reliable method for testing whether the direct boosting has better out-of-sample accuracy than the linear method. Other interesting topics for further research include the use of additional predictor variables and more flexible base learners allowing for interactions between the predictors. It would naturally also be of interest to compare the accuracy of the boosting forecasts with conventional nonlinear parametric models and with alternative nonparametric methods.

References

Bhansali, R.J. (1996) "Asymptotically efficient autogressive model selection for multistep prediction," *Annals of the Institute of Statistical Mathematics*, 48, 577–602.

Bhansali, R.J. (1997) "Direct autoregressive predictions for multistep prediction: order selection and performance relative to the plug in predictors," *Statistica Sinca*, 7, 425–449.

Bhansali, R.J. (1999) "Parameter estimation and model selection for multistep prediction of a time series: a review, *In* Ghosh, S. (Ed.), *Asymptotics, Nonparametrics, and Time Series*, Marcel Dekker, New York, pp. 201–225.

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

Bühlmann, P., and Hothorn, T. (2007) "Boosting algorithms: Regularization, prediction and model fitting," *Statistical Science*, 22, 477–505.

Bühlmann, P., and Yu, B. (2003) "Boosting with the L2 loss: Regression and classification," *Journal of American Statistical Association*, 98, 324–339.

Clark, T. E, and McCracken, M. W. (2001) "Tests of equal forecast accuracy and encompassing for nested models," *Journal of Econometrics*, 105, 85–110.

Clark, T. E, and West, K. D. (2006) "Using out-of-sample mean squared prediction errors to test the martingale difference hypothesis," *Journal of Econometrics*, 135, 155–186.

Clark, T. E, and West, K. D. (2007) "Approximately normal tests for equal predictive accuracy in nested models," *Journal of Econometrics*, 138, 291–311.

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

Clements, M.P. and Hendry, D.F. (1996) "Multi-step estimation for forecasting," *Oxford Bulletin of Economics and Statistics*, 58, 657–684.

Corradi, V., and Swanson, N. R. (2002) "A consistent test for nonlinear out of sample predictive accuracy," *Journal of Econometrics*, 110, 353–381.

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

Diebold, F. X., and Mariano, R. S. (1995) "Comparing predictive accuracy," *Journal of Business & Economic Statistics*, 13, 253–263.

Findley, D.F. (1983) "On the use of multiple models for multi-period forecasting," *Proceedings of the Business and Statistics Section, American Statistical Association*, 528–531.

Findley, D.F. (1985) "Model selection for multi-step-ahead forecasting," In: Baker, H.A., Young, P.C. (Eds.), Proceedings of the Seventh Symposium on Identification and System Parameter Estimation. Pergamon, Oxford, pp. 1039–1044.

Foresee, F. D., and Hagan, M. T. (1997) "Gauss-Newton approximation to Bayesian regularization," *Proceedings of the 1997 International Joint Conference on Neural Networks.*

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

Giacomini, R., and White, H. (2006) "Tests of conditional predictive ability," *Econometrica*, 74, 1545–1578.

Hansen, M., and Yu, B. (2001) "Model selection and minimum description length principle," *Journal of the American Statistical Association*, 96, 746–774.

Hwang, J. T. G., and Ding, A. A. (1997) "Prediction intervals for artificial neural networks," *Journal of the American Statistical Association*, 92, 748–757.

Ing, C.-K. (2003) "Multistep prediction in autoregressive processes," *Econometric Theory*, 19, 254–279.

Kim, H. H., and Swanson, N. R. (2014) "Forecasting financial and macroeconomic variables using data reduction methods: New empirical evidence," *Journal of Econometrics*, 178, 352–367.

Kock, A. B., and Teräsvirta, T. (2016) "Forecasting macroeconomic variables using neural network models and three automated model selection techniques," *Econometric Reviews*, 35, 1753–1779.

MacKay, D. J. C. (1992) "Bayesian interpolation," *Neural Computation*, 4(3), 415–447.

Marcellino, M. (2004) "Forecasting EMU macroeconomic variables," *International Journal of Forecasting*, 20, 359–372.

Marcellino, M. (2005). Instability and non-linearity in the EMU. In C. Milas, P. Rothman, & D. van Dijk (Eds.), *Nonlinear Time Series Analysis of Business Cycles*. Amsterdam, Elsevier.

McCracken, M. W., and Ng, S. (2016) "FRED-MD: A monthly database for macroeconomic research," *Journal of Business & Economic Statistics*, 34, 574–589.

Robinzonov, N., Tutz, G., and Hothorn, T. (2012) "Boosting techniques for nonlinear time series models," *Advances in Statistical Analysis*, 96, 99–122.

Schmid, M., and Hothorn, T. (2008) "Boosting additive models using component-wise P-Splines," *Computational Statistics and Data Analysis*, 53, 298–311.

Shibata, R. (1980) "Asymptotically efficient selection of the order of the model for estimating parameters of a linear process," *Annals of Statistics*, 8, 1464–1470.

Stock, J. H., and Watson, M. W. (1999) "A comparison of linear and nonlinear univariate models for forecasting macroeconomic time series," *In* Engle, R. F., and White, H. (eds.) *Cointegration, causality and forecasting. Festschrift in Honour of Clive W. J. Granger*, Oxford University Press, Oxford, 1–44.

Swanson, N. R., and White, H. (1997a) "A model selection approach to real-time macroeconomic forecasting using linear models and artificial neural networks," *Review of Economics and Statistics*, 79, 540–550.

Swanson, N. R., and White, H. (1997b) "Forecasting economic time series using flexible versus fixed specification and linear versus nonlinear econometric models," *Journal of International Forecasting*, 13, 439–461.

Taieb, S. B., and Hyndman, R. J. (2014) "Boosting multi-step autoregressive forecasts," Proceedings of the International Conference on Machine Learning (ICML).

Teräsvirta, T. (2006) "Forecasting economic variables with nonlinear models," *In*: Elliott, G., Granger, C. W. J., and Timmermann, A. (eds.) *Handbook of Economic Forecasting*, Elsevier, Amsterdam, 459–512.

Teräsvirta, T., Tjøstheim, D., and Granger, C. W. J. (1990) "Modeling nonlinear economics time series," Oxford University Press, Oxford.

Teräsvirta, T., van Dijk, D., and Medeiros, M. C. (2005) "Linear models, smooth transition autoregressions, and neural networks for forecasting macroeconomic time series: A re-examination," *International Journal of Forecasting*, 21, 755–774.

Tong, H. (1990) "Non-linear time series: A dynamical system approach," Oxford University Press, USA.

Weiss, A. A. (1991) "Multi-step estimation and forecasting in dynamic models," *Journal of Econometrics*, 48, 135–149.

Wohlrabe, K., and Buchen, T. (2014) "Assessing the Macroeconomic Forecasting Performance of Boosting: Evidence for the United States, the Euro Area and Germany," *Journal of Forecasting*, 33, 231–242.

Table 1. Simulated Models

(a)	$y_t = -0.5y_{t-1}I(y_{t-1} \leq 0.5) + 0.9y_{t-1}I(y_{t-1} > 0.5) + 0.5\varepsilon_t,\ \varepsilon_t \sim NID(0, 0.25)$
(b)	$y_t = 0.4\frac{5-y_{t-1}^2}{1+y_{t-1}^2} + \varepsilon_t,\ \varepsilon_t \sim NID(0, 0.25)$
(c)	$y_t = (0.5 + 2\exp(-y_{t-1}^2))y_{t-1} + \varepsilon_t,\ \varepsilon_t \sim NID(0, 0.25)$
(d)	$y_t = (0.4 - 2\exp(-50y_{t-6}^2))y_{t-6} + (0.5 - 0.5\exp(-50y_{t-10}^2))y_{t-10} + \varepsilon_t,\ \varepsilon_t \sim NID(0, 0.01)$
(e)	$y_t = -0.4\frac{3-y_{t-1}^2}{1+y_{t-1}^2} + 0.6\frac{3-(y_{t-2}-0.5)^3}{1+(y_{t-2}-0.5)^4} + \varepsilon_t,\ \varepsilon_t \sim NID(0, 0.01)$
(f)	$y_t = 0.21y_{t-1} + 0.35y_{t-2} + 0.17y_{t-3} + \varepsilon_t,\ \varepsilon_t \sim NID(0, 0.01)$

Notes: $NID(0, \sigma^2)$ indicates an iid normal series with mean zero and variance σ^2.

Table 2: Simulated MSE for the Direct and the Two-stage Boosting Strategies in Ratio to the Linear Method

Model	Prediction	$h=1$	$h=2$	$h=3$	$h=4$	$h=5$	$h=6$	$h=7$	$h=8$	$h=9$	$h=10$	$h=11$	$h=12$
	Optimal	68.0	90.2	93.4	95.5	96.5	98.7	98.3	98.3	99.3	99.4	98.4	99.3
(a)	Direct	74.3	91.5	95.7	98.5	100.0	101.0	101.4	101.6	101.7	101.7	101.6	101.8
	Two-stage	75.1	91.9	95.7	98.3	99.7	100.6	101.1	101.5	101.6	101.7	101.6	101.8
	Optimal	72.6	85.7	92.3	95.5	96.9	98.4	99.0	99.2	99.2	99.9	100.0	100.0
(b)	Direct	74.1	87.0	94.2	97.4	99.2	100.1	100.4	100.4	100.4	100.4	100.5	100.5
	Two-stage	74.5	87.3	94.4	97.7	99.4	100.1	100.4	100.5	100.6	100.6	100.7	100.7
	Optimal	68.8	77.4	81.3	85.2	87.6	88.6	89.3	91.1	91.7	93.2	93.8	93.4
(c)	Direct	71.2	81.6	85.9	89.1	91.4	93.4	95.0	95.9	96.9	97.6	98.3	98.8
	Two-stage	74.0	83.8	88.0	91.1	93.1	94.8	96.0	96.8	97.5	98.0	98.6	99.0
	Optimal	70.9	70.9	70.9	70.9	70.9	70.9	96.5	96.5	96.5	96.5	93.9	93.9
(d)	Direct	73.4	73.8	74.2	74.4	74.3	74.4	99.6	99.8	100.0	100.2	96.5	96.5
	Two-stage	75.0	75.2	75.3	75.4	75.1	75.0	99.9	99.9	100.0	100.0	96.6	96.6
	Optimal	11.4	16.5	26.5	42.6	53.3	61.7	62.9	68.6	75.1	77.8	80.2	83.8
(e)	Direct	14.6	25.4	56.7	62.8	66.7	79.3	77.6	78.6	86.2	87.8	87.6	92.9
	Two-stage	15.0	27.1	57.5	63.9	67.5	79.4	78.0	80.0	86.7	88.0	88.1	93.1
	Optimal	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0
(f)	Direct	101.9	102.2	102.0	102.1	101.8	101.5	101.5	101.2	101.0	100.9	100.9	100.9
	Two-stage	100.2	100.2	100.4	100.5	100.5	100.6	100.6	100.6	100.6	100.7	100.8	100.9

Notes: "Optimal" refers to the population level ratio $MMSE(h;p)/LMSE(h;p)$, while "Direct" and "Two-stage," respectively, refers to the simulated ratio $MSE(\widehat{g}_h^B(y_T(p)))/MSE(\widehat{g}_h^L(y_T(p)))$ and $MSE(\widehat{g}_h^{LB}(y_T(p)))/MSE(\widehat{g}_h^L(y_T(p)))$ (see the text). The simulated models, as indicated by the letters in the first column, are shown in Table 1. The predictions are for $x_{T+h} = y_{T+h}$ and the sampe size $T = 500$.

Table 3. Simulated MSE for the Direct and the Two-stage Boosting Strategies in Ratio to the Linear Method

Process	Prediction	$h=1$	$h=2$	$h=3$	$h=4$	$h=5$	$h=6$	$h=7$	$h=8$	$h=9$	$h=10$	$h=11$	$h=12$
	Optimal	68.1	76.4	81.0	84.6	86.8	88.9	90.6	91.7	92.5	93.0	93.5	94.5
(a)	Direct	74.2	79.5	83.9	87.2	89.8	91.8	93.4	94.8	95.9	96.8	97.6	98.2
	Two-stage	75.0	80.7	84.9	88.1	90.4	92.3	93.7	95.0	96.1	96.9	97.5	98.3
	Optimal	72.8	99.0	89.8	98.8	94.9	99.7	96.8	98.9	97.8	98.8	98.7	99.2
(b)	Direct	74.4	100.3	92.1	100.3	96.8	100.2	98.7	100.3	99.5	100.4	100.0	100.4
	Two-stage	74.8	100.0	92.2	100.1	97.0	100.2	98.9	100.2	99.7	100.3	100.1	100.4
	Optimal	68.6	63.6	62.4	64.6	67.5	69.3	71.2	73.4	76.0	76.9	78.1	80.1
(c)	Direct	70.6	66.1	67.0	69.5	72.2	74.7	77.1	79.2	81.1	82.9	84.3	85.6
	Two-stage	73.3	69.0	70.0	72.4	74.8	77.3	79.5	81.5	83.4	84.9	86.4	87.6
	Optimal	71.1	71.1	71.3	71.4	69.4	68.2	71.2	78.7	84.6	88.6	86.3	86.0
(d)	Direct	72.8	73.7	74.9	75.9	74.2	73.4	75.4	76.9	78.1	79.1	80.6	81.6
	Two-stage	74.4	74.6	75.3	75.6	73.7	72.7	75.0	76.7	78.1	79.3	80.8	81.9
	Optimal	11.3	16.3	22.1	36.2	49.4	64.4	70.8	64.9	64.0	68.4	75.1	80.2
(e)	Direct	14.8	21.2	35.0	56.0	67.1	77.9	82.3	71.9	70.8	79.1	85.5	90.2
	Two-stage	15.2	24.2	36.2	56.0	67.2	78.1	83.1	73.0	71.4	79.0	85.6	90.4
	Optimal	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0
(f)	Direct	102.0	102.1	102.5	102.6	102.9	103.1	103.2	103.4	103.5	103.5	103.6	103.6
	Two-stage	100.2	100.3	100.4	100.6	100.8	101.0	101.3	101.5	101.5	101.6	101.6	101.8

Notes: The notes of Table 2 apply with the exception that the predictions are for $x_{T+h} = \sum_{j=1}^{h} y_{T+j}$.

Table 4: Average Empirical R^2 for Out-of-sample Forecasts for Period 1999-2016

$\widehat{R}^2_{\min}$	Method	$h=1$	$h=2$	$h=3$	$h=4$	$h=5$	$h=6$	$h=7$	$h=8$	$h=9$	$h=10$	$h=11$	$h=12$
							(a) All cases						
0.0	Linear	0.33	0.37	0.39	0.40	0.40	0.40	0.39	0.39	0.38	0.37	0.36	0.36
	Direct Boosting	0.32	0.36	0.37	0.37	0.37	0.37	0.36	0.35	0.34	0.32	0.31	0.30
	Two-stage Boosting	0.34	0.38	0.40	0.41	0.41	0.41	0.40	0.39	0.38	0.37	0.36	0.35
	N	128	128	128	128	128	128	128	128	128	128	128	128
0.1	Linear	0.43	0.48	0.51	0.51	0.51	0.53	0.53	0.53	0.52	0.50	0.48	0.47
	Direct Boosting	0.42	0.46	0.48	0.48	0.47	0.49	0.48	0.47	0.46	0.43	0.41	0.40
	Two-stage Boosting	0.44	0.50	0.52	0.52	0.52	0.55	0.54	0.53	0.52	0.50	0.48	0.47
	N	95	97	97	99	100	94	94	93	93	95	96	95
							(b) Cases where boosting is more accurate						
0.0	Linear	0.31	0.32	0.32	0.32	0.34	0.32	0.34	0.31	0.30	0.30	0.28	0.29
	Direct Boosting	0.33	0.34	0.34	0.33	0.36	0.34	0.36	0.34	0.31	0.31	0.30	0.31
	Two-stage Boosting	0.33	0.34	0.35	0.35	0.37	0.35	0.37	0.35	0.34	0.34	0.33	0.34
	N	77	70	65	70	69	72	71	62	61	65	61	61
0.1	Linear	0.40	0.42	0.41	0.41	0.40	0.44	0.43	0.39	0.38	0.37	0.33	0.34
	Direct Boosting	0.41	0.44	0.43	0.41	0.43	0.46	0.44	0.43	0.40	0.38	0.35	0.36
	Two-stage Boosting	0.41	0.45	0.44	0.44	0.44	0.48	0.47	0.44	0.42	0.42	0.38	0.38
	N	60	52	51	55	57	52	56	49	48	52	52	53
							(c) Cases where linear procedure is more accurate						
0.0	Linear	0.35	0.44	0.46	0.49	0.47	0.49	0.46	0.46	0.46	0.44	0.44	0.41
	Direct Boosting	0.31	0.38	0.39	0.42	0.38	0.40	0.36	0.35	0.35	0.34	0.32	0.28
	Two-stage Boosting	0.35	0.43	0.45	0.47	0.45	0.47	0.43	0.43	0.43	0.41	0.39	0.36
	N	51	58	63	58	59	56	57	66	67	63	67	67
0.1	AR(BIC)	0.27	0.34	0.38	0.39	0.39	0.37	0.37	0.38	0.38	0.37	0.39	0.37
	Direct Boosting	0.24	0.30	0.33	0.34	0.32	0.30	0.29	0.30	0.30	0.28	0.28	0.26
	Two-stage Boosting	0.27	0.34	0.37	0.38	0.38	0.36	0.35	0.36	0.36	0.34	0.35	0.33
	N	35	45	46	44	43	42	38	44	45	43	44	42

Notes: N is the number of series in the FRED data set that meet (for given horizon h) the criterion $\widehat{R}^2 \geq \widehat{R}^2_{\min}$ for at least one of the three forecasting methods.

Table 5: Average Empirical R^2 for Out-of-sample Forecasts for Period 1999-2007

$\widehat{R}^2_{\min}$	Method	$h=1$	$h=2$	$h=3$	$h=4$	$h=5$	$h=6$	$h=7$	$h=8$	$h=9$	$h=10$	$h=11$	$h=12$
							(a) All cases						
0.0	Linear	0.32	0.37	0.39	0.39	0.39	0.38	0.37	0.37	0.35	0.34	0.32	0.31
	Direct Boosting	0.31	0.35	0.37	0.37	0.38	0.37	0.36	0.35	0.34	0.32	0.30	0.29
	Two-stage Boosting	0.33	0.38	0.41	0.41	0.41	0.41	0.39	0.38	0.37	0.36	0.34	0.33
	N	128	128	128	128	128	128	128	128	128	128	128	128
0.1	Linear	0.44	0.50	0.55	0.52	0.53	0.52	0.50	0.50	0.49	0.47	0.46	0.45
	Direct Boosting	0.41	0.47	0.51	0.49	0.50	0.50	0.49	0.48	0.47	0.45	0.43	0.42
	Two-stage Boosting	0.45	0.52	0.57	0.55	0.56	0.55	0.53	0.53	0.51	0.50	0.49	0.47
	N	92	94	90	96	94	93	94	92	92	91	89	87
							(b) Cases where boosting is more accurate						
0.0	Linear	0.33	0.37	0.38	0.35	0.34	0.32	0.32	0.33	0.27	0.27	0.23	0.23
	Direct Boosting	0.33	0.38	0.40	0.38	0.39	0.36	0.37	0.37	0.32	0.32	0.28	0.29
	Two-stage Boosting	0.34	0.39	0.41	0.38	0.39	0.36	0.37	0.38	0.32	0.33	0.30	0.30
	N	88	84	77	73	68	72	70	72	68	69	65	58
0.1	Linear	0.45	0.48	0.49	0.42	0.43	0.42	0.39	0.41	0.36	0.35	0.31	0.29
	Direct Boosting	0.45	0.48	0.51	0.46	0.49	0.47	0.44	0.45	0.42	0.40	0.37	0.36
	Two-stage Boosting	0.47	0.50	0.52	0.46	0.49	0.48	0.45	0.47	0.42	0.42	0.39	0.37
	N	63	65	59	60	53	54	57	57	51	53	48	45
							(c) Cases where linear procedure is more accurate						
0.0	Linear	0.30	0.37	0.40	0.46	0.45	0.47	0.44	0.41	0.45	0.42	0.42	0.37
	Direct Boosting	0.24	0.30	0.32	0.36	0.36	0.38	0.35	0.33	0.36	0.33	0.33	0.29
	Two-stage Boosting	0.30	0.37	0.40	0.45	0.44	0.46	0.42	0.40	0.43	0.39	0.39	0.36
	N	40	44	51	55	60	56	58	56	60	59	63	70
0.1	AR(BIC)	0.19	0.26	0.30	0.37	0.36	0.36	0.36	0.33	0.35	0.33	0.33	0.32
	Direct Boosting	0.16	0.22	0.25	0.29	0.30	0.29	0.30	0.27	0.29	0.27	0.26	0.25
	Two-stage Boosting	0.19	0.26	0.30	0.37	0.36	0.35	0.35	0.32	0.34	0.31	0.32	0.31
	N	29	29	31	36	41	39	37	35	41	38	41	42

Notes: See Table 4.

Table 6: Average Empirical R^2 for "Hybrid" Out-of-sample Forecasts for Period 1999-2016

$\widehat{R}^2_{\text{min}}$	Method	$h=1$	$h=2$	$h=3$	$h=4$	$h=5$	$h=6$	$h=7$	$h=8$	$h=9$	$h=10$	$h=11$	$h=12$
							(a) All cases						
0.0	AR(BIC)	0.33	0.37	0.39	0.40	0.40	0.40	0.39	0.39	0.38	0.37	0.36	0.36
	Direct Boosting	0.33	0.37	0.38	0.39	0.39	0.39	0.38	0.37	0.36	0.35	0.34	0.33
	Two-stage Boosting	0.34	0.38	0.40	0.41	0.41	0.41	0.41	0.40	0.39	0.39	0.38	0.37
	N	128	128	128	128	128	128	128	128	128	128	128	128
0.1	AR(BIC)	0.43	0.49	0.51	0.52	0.52	0.53	0.53	0.53	0.52	0.50	0.49	0.47
	Direct Boosting	0.43	0.48	0.50	0.51	0.50	0.52	0.51	0.51	0.50	0.48	0.46	0.44
	Two-stage Boosting	0.44	0.51	0.53	0.54	0.53	0.55	0.54	0.55	0.54	0.52	0.50	0.49
	N	95	95	96	97	98	94	95	93	93	94	95	95
							(b) Cases where boosting is more accurate						
0.0	AR(BIC)	0.35	0.34	0.37	0.35	0.36	0.33	0.34	0.32	0.30	0.33	0.32	0.33
	Direct Boosting	0.37	0.36	0.39	0.36	0.38	0.35	0.36	0.35	0.32	0.35	0.34	0.34
	Two-stage Boosting	0.37	0.37	0.40	0.37	0.39	0.36	0.38	0.36	0.34	0.37	0.36	0.37
	N	83	79	77	76	73	74	73	63	65	68	63	62
0.1	AR(BIC)	0.44	0.46	0.46	0.44	0.43	0.43	0.43	0.40	0.40	0.41	0.37	0.36
	Direct Boosting	0.46	0.47	0.48	0.46	0.46	0.46	0.44	0.43	0.42	0.43	0.39	0.38
	Two-stage Boosting	0.46	0.48	0.49	0.47	0.47	0.47	0.47	0.45	0.44	0.46	0.41	0.40
	N	65	59	62	59	60	55	58	50	49	54	54	56
							(c) Cases where linear procedure is more accurate						
0.0	AR(BIC)	0.29	0.41	0.41	0.47	0.45	0.49	0.46	0.45	0.46	0.42	0.41	0.38
	Direct Boosting	0.25	0.37	0.37	0.43	0.40	0.44	0.41	0.40	0.41	0.36	0.35	0.32
	Two-stage Boosting	0.28	0.41	0.41	0.47	0.45	0.49	0.45	0.44	0.45	0.41	0.40	0.36
	N	45	49	51	52	55	54	55	65	63	60	65	66
0.1	AR(BIC)	0.21	0.30	0.32	0.36	0.37	0.37	0.36	0.38	0.37	0.34	0.36	0.35
	Direct Boosting	0.19	0.28	0.29	0.33	0.33	0.34	0.33	0.34	0.33	0.30	0.31	0.30
	Two-stage Boosting	0.21	0.30	0.32	0.36	0.37	0.37	0.36	0.37	0.37	0.33	0.35	0.34
	N	30	36	34	38	38	39	37	43	44	40	41	39

Notes: See Table 4.

Table 7: Series by Category, Average Empirical R^2 for the Linear and the Hybrid Two-stage Boosting Methods, Forecasting Period 1999-2016

Category	Method	$h=1$	$h=2$	$h=3$	$h=4$	$h=5$	$h=6$	$h=7$	$h=8$	$h=9$	$h=10$	$h=11$	$h=12$
Output and Income	Linear	0.16	0.22	0.20	0.22	0.21	0.20	0.19	0.18	0.17	0.15	0.14	0.13
	Two-stage Boosting	0.20	0.26	0.24	0.26	0.27	0.27	0.27	0.26	0.25	0.24	0.22	0.21
	Change (%)	29.8	20.2	19.2	17.7	26.0	37.4	43.8	44.2	49.0	56.8	60.0	55.6
	N	10	11	14	13	13	14	14	14	14	14	14	14
Orders and Inventories	Linear	0.31	0.28	0.46	0.40	0.39	0.47	0.36	0.42	0.31	0.23	0.21	0.17
	Two-stage Boosting	0.31	0.31	0.48	0.40	0.40	0.47	0.37	0.42	0.33	0.27	0.26	0.23
	Change (%)	1.1	11.3	3.9	1.5	0.7	0.7	2.1	0.5	4.8	14.1	21.6	31.8
	N	4	5	3	4	4	3	4	3	4	5	5	5
Labor Market	Linear	0.49	0.50	0.55	0.55	0.53	0.58	0.55	0.53	0.53	0.49	0.49	0.47
	Two-stage Boosting	0.50	0.52	0.58	0.58	0.56	0.61	0.56	0.54	0.54	0.50	0.50	0.48
	Change (%)	1.8	4.7	5.3	5.4	5.9	4.8	3.4	2.7	1.9	1.4	1.9	1.7
	N	26	29	27	27	28	25	26	26	25	26	25	25
Consumption and Orders	Linear	0.95	0.94	0.92	0.91	0.89	0.87	0.84	0.82	0.79	0.76	0.73	0.69
	Two-stage Boosting	0.95	0.94	0.92	0.91	0.88	0.86	0.83	0.81	0.78	0.75	0.71	0.67
	Change (%)	−0.0	−0.1	−0.1	−0.2	−0.4	−0.8	−1.3	−1.4	−1.5	−1.5	−2.1	−3.4
	N	10	10	10	10	10	10	10	10	10	10	10	10
Money and Credit	Linear	0.28	0.42	0.43	0.44	0.48	0.50	0.53	0.54	0.56	0.61	0.61	0.63
	Two-stage Boosting	0.28	0.42	0.43	0.44	0.48	0.50	0.52	0.54	0.56	0.60	0.61	0.63
	Change (%)	−0.2	−1.5	−0.4	0.6	−0.5	−0.4	−0.8	−0.3	0.3	−0.5	−0.3	0.3
	N	9	9	11	12	12	12	12	12	12	11	11	11
Stock Market	Linear	0.41	0.42	0.26	0.21	0.18	0.12	0.09	0.08	0.07	0.06	0.06	0.05
	Two-stage Boosting	0.42	0.44	0.30	0.28	0.25	0.19	0.18	0.22	0.18	0.17	0.16	0.14
	Change (%)	2.0	5.0	14.2	31.3	38.9	58.6	92.9	162.4	162.4	175.4	178.2	174.4
	N	3	2	3	3	3	3	3	2	2	2	2	2
Interest Rate and Exchange Rates	Linear	0.56	0.73	0.73	0.66	0.61	0.63	0.66	0.61	0.57	0.51	0.35	0.31
	Two-stage Boosting	0.56	0.74	0.74	0.67	0.61	0.63	0.67	0.60	0.56	0.50	0.36	0.31
	Change (%)	0.1	0.6	0.5	0.9	0.7	0.5	1.1	−1.1	−0.3	−2.3	5.4	−1.1
	N	14	9	8	8	8	7	6	6	6	6	8	8
Prices	Linear	0.24	0.40	0.49	0.54	0.58	0.61	0.63	0.65	0.67	0.69	0.70	0.71
	Two-stage Boosting	0.25	0.40	0.49	0.54	0.58	0.61	0.63	0.65	0.67	0.68	0.69	0.71
	Change (%)	2.6	−0.5	−0.7	−0.4	−0.3	0.0	−0.2	−0.1	−0.3	−0.2	−0.4	−0.2
	N	19	20	20	20	20	20	20	20	20	20	20	20

Notes: Only series where $\widehat{R}^2 \geq 0.1$ for the linear or the two-stage boosting forecast are included.

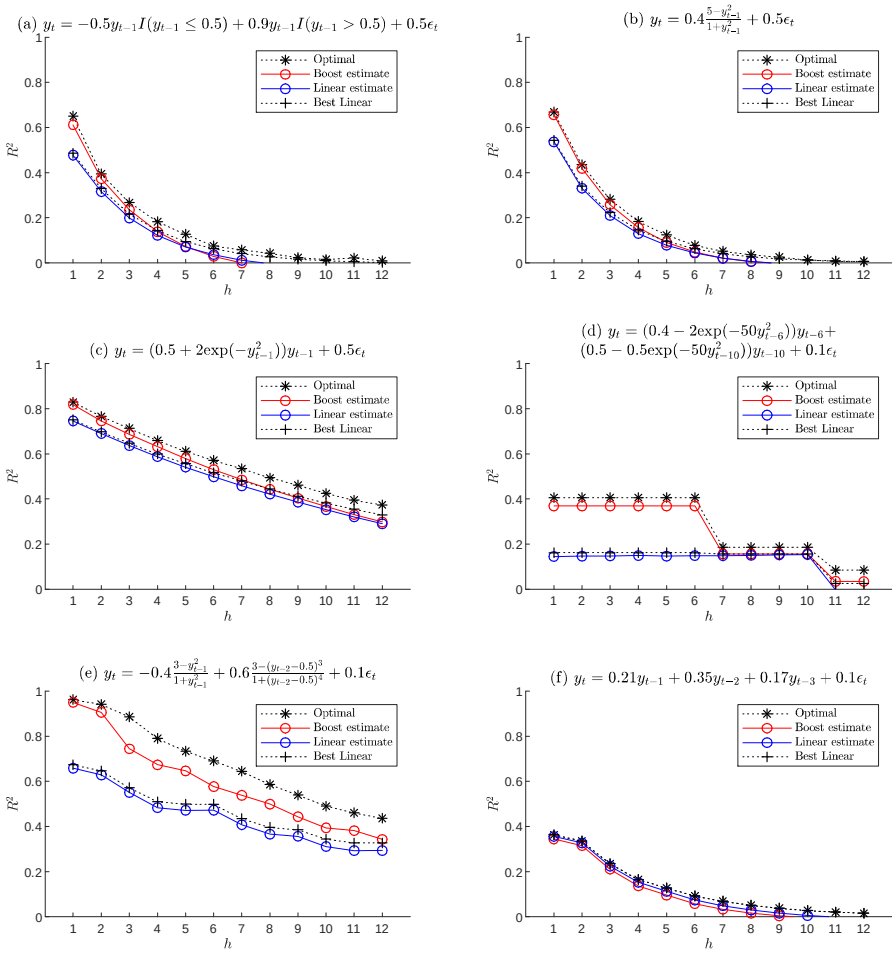

Figure 1: Performance of Simulated Predictions for $x_{T+h} = y_{T+h}$ and $T = 500$

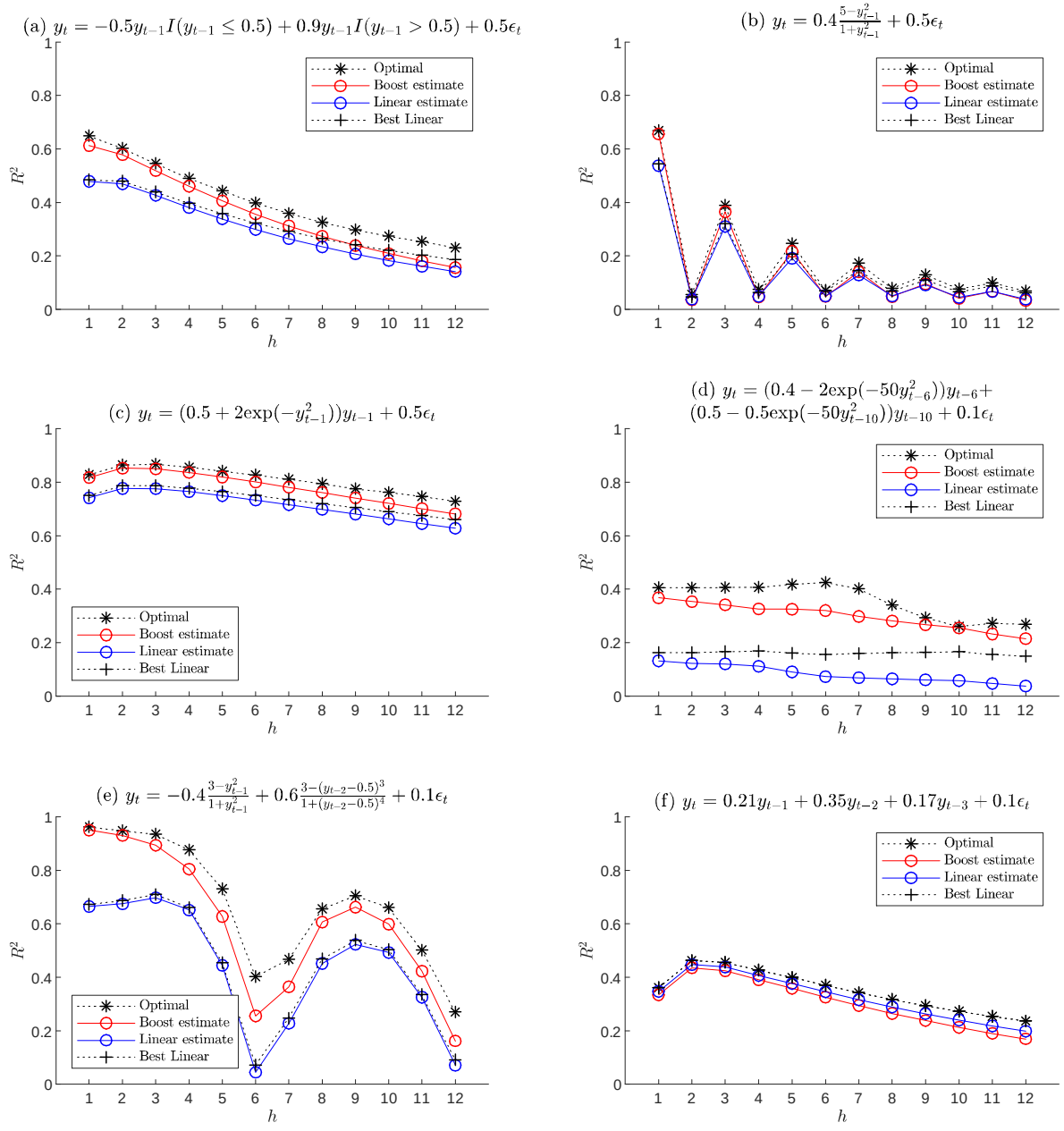

Figure 2: Performance of Simulated Predictions for $x_{T+h} = \sum_{j=1}^{h} y_{T+j}$ and $T = 500$

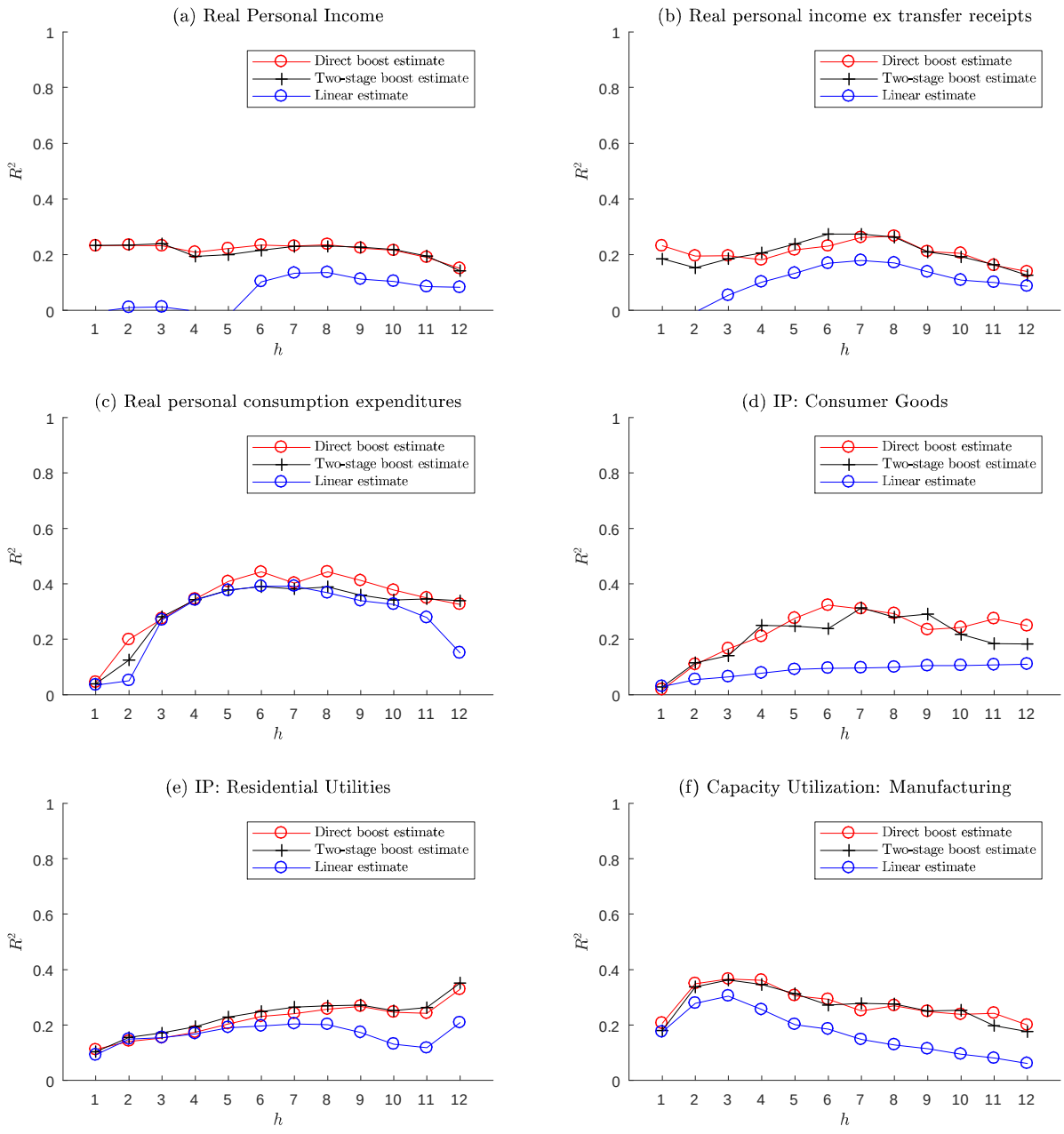

Figure 3: Empirical Out-of-sample R^2 for Selected FRED-MD Series

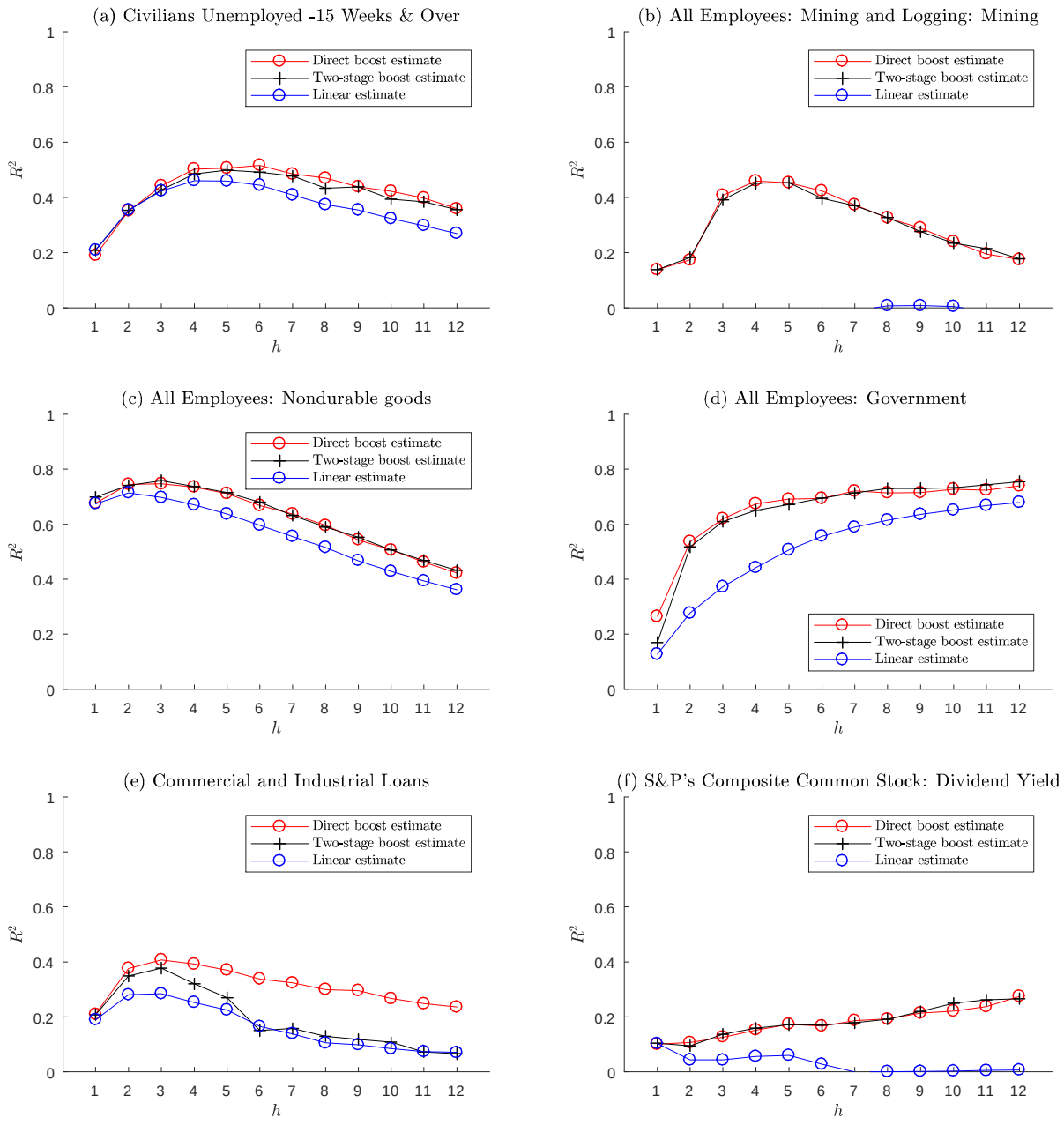

Figure 4: Empirical Out-of-sample R^2 for Selected FRED-MD Series

Appendix A: Results for Alternative Prediction Strategies

In this appendix, we present simulation and empirical results for two alternative strategies to the boosting strategies applied in the paper.

The first alternative is similar to our two-stage boosting strategy except that it replaces the first stage direct multistep linear AR prediction by its indirect version as is applied in the boosting procedure of Taieb and Hyndman (2014). We call this strategy the recursive two-stage boosting strategy. The corresponding prediction function is

$$\widehat{g}_h^{RLB}(y_T(p)) = \widehat{g}_h^{RL}(y_T(p)) + \widehat{r}_h^{B}(y_T(p))$$

where $\widehat{g}_h^{RL}(y_T(p))$ is a linear forecast that is obtained by iterating an AR(p) model (the one-step head model) h times and $\widehat{r}_h^{B}(y_T(p))$ is the boosting model fitted to its residuals.

In the second alternative, we replace our boosting estimator by an artificial neural network (ANN) model. More specifically, we apply a single layer feed forward neural network model and fit it to the estimation data by using Bayesian regularization as described in MacKay (1992) and Foresee and Hagan (1997). Additionally, we estimate multiple candidate models where the number of hidden units, q, and the number of inputs (lagged values of y_t), p, take on values from the set $\{3, 6, 9, 12\}$. The best combination of p and q is selected for each estimation sample by using a grid search with a split sample procedure where 70% of the estimation sample is used for training and the rest 30% for validation. All ANN estimations are conducted with the brnn R package.

A direct and a two-stage ANN versions are considered. In the "direct ANN," we estimate the model

$$y_{t+h} = \mu + \sum_{j=1}^{q} \beta_j \sigma(\alpha_j + \gamma_j' y_t(p)) + e_{t+h}$$

where $\sigma(\cdot)$ is the "activation function." In the "two-stage ANN" version, the estimated prediction function is

$$\widehat{g}_h^{L,ANN}(y_T(p)) = \widehat{g}_h^{L}(y_T(p)) + \widehat{r}_h^{ANN}(y_T(p))$$

where $\widehat{g}_h^L(y_T(p))$ refers to the direct linear AR prediction (as in our two-stage boosting) and $\widehat{r}_h^{ANN}(y_T(p))$ refers to an ANN model fitted to the first stage residuals. In stead of the two-stage model one might also consider an ANN model that incorporates a linear unit. However, linear units are not usually applied with the Bayesian regularization strategy.

The simulation results for the recursive two-stage boosting and the ANN models are shown in Table A.1. The numbers represent percentage ratios of the average MSE of a prediction method in ratio to that of the linear direct multistep method. Thus, they correspond to values given for the boosting methods in Table 3 of the paper.

The empirical results for the recursive two-stage boosting and the ANN models are shown in Table A.2. The table shows the average $\widehat{R}^2$ values calculated over the series in the FRED data set. The results correspond to those reported in panel (a) of Table 4 in the paper (for the direct boosting and the (direct) two-stage boosting methods).

The recursive and the direct two-stage procedures are essentially equally accurate in all simulation models. The direct linear AR forecast tends to select less lags for the longest forecasting horizons than what is used for the recursive AR forecast. For this reason, the recursive AR forecast may in some cases be more accurate. The second stage boosting procedure almost fully compensates for this difference.

In the empirical application, we find that the recursive two-stage boosting procedure is on average less accurate than the direct two-stage procedure. The difference originates from the first stage, where the recursive linear forecast is on average notably less accurate than the direct linear forecast. The second stage is not able to fully compensate for this difference, unlike in the simulations.

In the simulations of the nonlinear processes (models (a) to (e) in Table 1 of the paper), we find that the ANN forecasts are less accurate than the boosting forecasts. For the two-stage ANN model, the difference is notable. In case of the linear simulation model (f), ANN behaves analogously to the boosting procedures; The direct ANN model overfits more than the two-stage model, resulting in higher out-of-sample MSE.

With the empirical data set, the direct ANN model is on average less accurate than the linear AR model or the direct boosting model. The two-stage ANN model is less accurate than the two-stage boosting model. These results are in line with our expectations and

suggest that the ANN method cannot beat the boosting method in our application and that the boosting method is sometimes markedly superior to the ANN.

We note, however, that our experiments with the ANN model are fairly limited. There might, for example, be alternative regularization strategies that would yield better results, although the Bayesian regularization used in this comparison has been very succesful in recent studies (see Teräsvirta et al. (2005)).

Table A.1: Simulated MSE for the Recursive Two-stage and ANN Strategies in Ratio to the Linear Method

Process	Prediction	$h = 1$	$h = 2$	$h = 3$	$h = 4$	$h = 5$	$h = 6$	$h = 7$	$h = 8$	$h = 9$	$h = 10$	$h = 11$	$h = 12$
	Recursive two-stage	74.5	80.0	84.3	87.6	90.1	92.0	93.5	94.6	95.6	96.6	97.3	97.8
(a)	ANN direct	80.4	85.2	89.2	90.5	93.6	95.7	95.8	97.7	98.2	99.7	99.5	99.3
	ANN two-stage	93.1	94.4	94.7	96.3	97.7	97.4	96.6	98.5	98.7	99.9	100.6	100.0
	Recursive two-stage	74.7	100.2	92.0	100.1	96.8	100.0	98.8	100.1	99.5	100.1	99.9	100.1
(b)	ANN direct	77.2	100.7	95.8	101.6	101.0	100.7	100.6	100.6	100.8	101.0	101.3	101.3
	ANN two-stage	94.0	100.8	99.9	100.8	100.2	101.0	100.4	100.6	100.7	100.4	100.5	100.4
	Recursive two-stage	74.3	70.2	70.7	72.8	75.1	77.6	79.6	81.7	83.3	84.8	86.2	87.3
(c)	ANN direct	75.1	68.7	70.4	73.3	77.6	80.9	83.1	85.3	85.7	87.4	88.9	88.4
	ANN two-stage	95.3	94.8	95.3	95.9	94.6	96.3	96.1	96.8	98.1	96.6	97.6	97.4
	Recursive two-stage	74.9	75.3	75.2	75.1	71.8	70.7	73.6	74.8	76.9	78.2	79.8	80.7
(d)	ANN direct	100.6	99.7	99.8	98.7	96.4	96.1	96.0	95.4	95.5	96.2	95.4	95.0
	ANN two-stage	100.6	99.4	98.6	97.7	95.5	95.3	95.8	94.6	94.6	94.6	94.1	94.0
	Recursive two-stage	14.8	23.6	35.6	55.8	67.2	78.1	82.8	72.5	71.0	78.7	85.2	90.2
(e)	ANN direct	15.0	22.8	38.7	61.9	81.9	97.8	90.6	76.2	78.6	86.3	96.4	99.4
	ANN two-stage	41.1	46.5	61.1	79.7	95.1	97.9	97.9	90.6	91.5	97.3	99.5	100.2
	Recursive two-stage	100.1	99.8	99.8	99.9	100.4	100.4	100.4	100.4	100.6	100.8	100.6	100.6
(f)	ANN direct	101.3	101.6	101.6	101.8	101.7	102.3	102.1	101.9	102.1	102.5	101.8	102.4
	ANN two-stage	100.2	99.9	100.4	100.3	100.3	100.4	100.2	100.4	100.4	100.3	100.4	100.4

Notes: The notes of Table 3 apply.

Table A.2: Empirical series, average $\widehat{R}^2$ values of forecasts, forecasting period 1999-2016

$\widehat{R}^2_{min}$	Method	$h=1$	$h=2$	$h=3$	$h=4$	$h=5$	$h=6$	$h=7$	$h=8$	$h=9$	$h=10$	$h=11$	$h=12$
0.0	Recursive two-stage	0.34	0.36	0.37	0.38	0.37	0.37	0.36	0.35	0.34	0.33	0.32	0.30
	Direct ANN	0.31	0.33	0.35	0.35	0.34	0.34	0.33	0.32	0.30	0.29	0.29	0.28
	Two-stage ANN	0.33	0.37	0.38	0.39	0.40	0.39	0.39	0.39	0.38	0.37	0.36	0.35
0.1	Recursive two-stage	0.44	0.47	0.49	0.48	0.48	0.50	0.48	0.48	0.46	0.44	0.42	0.41
	Direct ANN	0.41	0.42	0.44	0.44	0.43	0.45	0.44	0.43	0.40	0.39	0.38	0.38
	Two-stage ANN	0.43	0.48	0.50	0.51	0.51	0.53	0.53	0.53	0.51	0.50	0.47	0.46

Notes: The filtering criteria applies to the boosting models, i.e., exactly the same series are included as in the Table 4.

Appendix B: Giacomini-White test

This appendix presents the results of the unconditional Giacomini-White (GW) test for the direct and two-stage boosting forecasts. Table B.1 lists the number of series for which the boosting forecast is significantly more accurate than the linear direct multistep AR forecast at 5% and 10% significance levels.

A caveat needs to be noted with these results. The test procedure of Giacomini and White (2006) entails the assumption of asymptotically non-vanishing estimation error. This is achieved in practice by requiring that the estimation window size is fixed, i.e., using a rolling estimation window. In our application, we have used the expanding estimation window since we expect that the nonparametric boosting estimator will benefit from having as much data as possible for the estimation. Although we expect that the test will work in these finite samples despite the increasing estimation window length, the setup does not fully comply with the underlying assumptions of the GW test. It is not clear how this might affect the result.

Table B.1: Number of series where the forecast accuracy difference is statistically significant

p	Method	$h=1$	$h=2$	$h=3$	$h=4$	$h=5$	$h=6$	$h=7$	$h=8$	$h=9$	$h=10$	$h=11$	$h=12$
0.95	Direct boosting	16	14	10	7	9	14	11	6	6	7	9	9
	Two-stage boosting	17	10	10	6	7	10	8	7	9	9	9	8
0.9	Direct boosting	21	21	20	15	16	22	18	16	13	14	19	16
	Two-stage boosting	21	21	21	14	14	18	19	17	18	18	19	14

Notes: The table lists the number of series for which the boosting forecast is significantly more accurate than the linear direct multistep AR forecast according to the unconditional Giacomini-White test.