JOURNAL OF APPLIED ECONOMETRICS
J. Appl. Econ. **24**: 607–629 (2009)
Published online 16 March 2009 in Wiley InterScience
(www.interscience.wiley.com) DOI: 10.1002/jae.1063

BOOSTING DIFFUSION INDICES

JUSHAN BAI[a] AND SERENA NG[b]*

[a] *Department of Economics, New York University, New York, USA*
[b] *Department of Economics, Columbia University, New York, USA*

SUMMARY

In forecasting and regression analysis, it is often necessary to select predictors from a large feasible set. When the predictors have no natural ordering, an exhaustive evaluation of all possible combinations of the predictors can be computationally costly. This paper considers 'boosting' as a methodology of selecting the predictors in factor-augmented autoregressions. As some of the predictors are being estimated, we propose a stopping rule for boosting to prevent the model from being overfitted with estimated predictors. We also consider two ways of handling lags of variables: a componentwise approach and a block-wise approach. The best forecasting method will necessarily depend on the data-generating process. Simulations show that for each data type there is one form of boosting that performs quite well. When applied to four key economic variables, some form of boosting is found to outperform the standard factor-augmented forecasts and is far superior to an autoregressive forecast. Copyright © 2009 John Wiley & Sons, Ltd.

1. INTRODUCTION

This paper studies boosting as a variable selection device within the factor-augmented regression (FAR) framework, which takes the form

$$y_{t+h} = \kappa_0 + \kappa(L)W_t + \alpha(L)y_t + \beta(L)\tilde{f}_t + \tilde{\varepsilon}_{t+h} \tag{1}$$

where $W_t = (w'_t, \dots w'_{t-p})'$, w_t is a vector of predetermined variables, $\tilde{f}_t \subset \tilde{F}_t$ is an $r_y \times 1$ vector, and $\tilde{F}_t$ is a $r \times 1$ vector of principal component estimates of F_t extracted from N observed predictors. The predictors are assumed to have an underlying factor structure given by

$$x_{it} = \lambda'_i F_t + e_{it}, \quad i = 1, \dots N, t = 1, \dots T \tag{2}$$

The primary appeal of FAR is that the factors embody information in many variables. It has been successfully used in providing 'diffusion index' (DI) forecasts, a methodology proposed by Stock and Watson (2002) and Forni *et al.* (2005). The framework is also useful for estimation of conditional moments such as the mean and volatility of asset returns. In these applications, the results are known to be sensitive to the choice of conditioning variables.

In practice, a FAR analysis is obtained as follows. After obtaining the r estimated factors, $\tilde{F}_t$, let $\tilde{z}_t = (1, W'_t, y_t, y_{t-1}, \dots, y_{t-p}, \tilde{F}_{t1} \dots \tilde{F}_{t-p,1}, \dots, \tilde{F}_{t,r}, \dots, \tilde{F}_{t-p,r})'$ be the potential set of predictors. The next step is to determine p^* and r_y, where p^* is the optimal lags of y_t, and r_y is

[*] Correspondence to: Serena Ng, Department of Economics, Columbia University, 420 W 118 Street, MC 3308, New York, NY 10027, USA. E-mail: serena.ng@columbia.edu

the number of estimated factors that enter the forecasting equation, to yield $\tilde{f}_t = (\tilde{F}_{t1}, \ldots \tilde{F}_{tr_y})'$. Information criteria such the AIC or BIC are used to determine these two auxiliary parameters. These parameters then determine the dimension of $\beta(L), \kappa(L), \alpha(L)$ in (1).[1]

There are several aspects of the DI methodology that remain unsatisfactory. First, information criteria assume that the components in $\tilde{F}_t$ are ordered. In consequence, the order of the factors chosen to explain y are determined by the order in which the factors are important for x_{it}. But there is no reason to think, for example, that the factors that best explain the conditional mean of asset returns need to be the same as those that best forecast employment. In Bai and Ng (2008), we introduced the concept of 'targeted predictors' to highlight the point that which factors to use as predictors should depend on what is the variable to be explained.

The second problem concerns the specification of the forecasting equation. If the p^*th lag of y_t or $\tilde{F}_t$ has strong predictive power, typical model selection procedures require that lags one through $p^* - 1$ also enter the model even though they may have no predictive power for y. Similarly, the r_yth factor cannot enter in the absence of the preceding $r_y - 1$ factors since $\tilde{F}_t$ is ordered by the importance in explaining x_t. Furthermore, in most applications $\tilde{f}_t$ is simply a subvector of $\tilde{F}_t$. But, in principle, functions of $\tilde{F}_t$ (such as the quadratic terms) should be allowed. There is limited flexibility in choosing which factors and which lags to enter the model. The consequence is that the FAR regression is as susceptible to overfitting as it is to underfitting.

These issues arise because there is no easy way to select the best predictors using a small number of regressions without imposing some structure on the predictors, or go through an exhaustive search that can be computationally costly. The problem is not specific to the DI methodology because instead of taking $\tilde{F}_t$ as the first r (instead of all N) principal components of x_{it}, we can also take $\tilde{F}_t$ to be the N observed predictors that underlie the principal components. One is still faced with the problem of which variables to choose and which lags to select.

To handle these problems, we need a model-fitting device that performs subset variable selection when the set of candidate predictors is large, but without having to rely on some a priori ordering of the variables or of the lags. Such methods have been developed in the statistics literature for use in genes and spam data analysis. The question is whether these methods are also useful in analysis of economic data. In Bai and Ng (2008), we considered three such methods—LASSO (least absolute shrinkage and selection operator), LARS (least angle regression), and the elastic net—with special focus on how to select functions of the factors as predictors. LASSO and LARS have also been considered by DeMol *et al.* (2006) as alternatives to diffusion index forecasting. In this paper, we consider another alternative, boosting, with focus on the selection of lags.

We consider variations of boosting not previously considered in the literature: a componentwise approach that treats each lag as a separate variable, and a block-wise approach that treats lags of the same variable jointly. Boosting necessitates a stopping rule. In the case when the predictors are the estimated factors, we suggest a new boosting-stopping rule to prevent the model from being overfitted with estimated predictors. Our stopping rule consists of two penalties—one is of order T^{-1} and one of order N^{-1}—in contrast to the usual penalty that depends on either the number of time series observations or the number of cross-section units used in the analysis.

Our paper evaluates the effectiveness of boosting diffusion indices and observed predictors. Since factor analysis and boosting can both handle large-dimension data, one might question our motivation for combining the two procedures. Factor analysis is an approach that summarizes

[1] Since in general $\tilde{f}_t$ is a rotation of f_t, the corresponding coefficients $\beta(L)$ are also rotated. But for notational simplicity we still use $\beta(L)$ to denote the coefficients.

DOI: 10.1002/jae

information in a large number of variables into a small number of variables, irrespective of which variable is to be explained. Boosting is a methodology that isolates which, amongst a large number of variables, are most helpful in predicting a variable of interest. In the boosting framework, these variables can be the observed raw data, or they can be orthonormal transformations of the raw data. In the latter case, they can further be interpreted as factors if the primitive assumptions of factor model holds. Combining them is natural because factor analysis provides a dimension reduction in the predictors, while boosting allows us to pick out the most relevant factors with a target variable y in mind. The diffusion-forecasting methodology as it stands does not have this capability.

The rest of the paper is structured as follows. We begin in Section 2 with a brief overview of boosting. Section 3 considers boosting of factor-augmented regressions, while Section 4 discusses the problem of estimated predictors. Simulations and an application are considered in Section 5. It is worth emphasizing that our objective is to better understand implementation issues that are specific to economic applications, with the understanding that which method is best necessarily depends on the design of the experiments.

2. BOOSTING

Boosting is a procedure that estimates an unknown function, especially the conditional mean, using M stage-wise regressions. Suppose we have observations on y_t and on each of n observed predictors, $z_t = (z_{t1}, \ldots z_{tn})'(t = 1, 2, \ldots, T)$. Let $\Phi(z)$ be a function defined on R^n, and let $C(y_t, \Phi(z_t))$ be the loss function that penalizes the deviation of $\Phi(z_t)$ from y_t. The objective is to estimate the function $\Phi(\cdot)$ that minimizes the expected loss $E[C(y_t, \Phi(z_t))]$. Under the quadratic loss function $C(y_t, \Phi(z_t)) = \frac{1}{2}(y_t - \Phi(z_t))^2$, the optimal solution is $\Phi(z) = E(y_t|z_t = z)$. The generic boosting algorithm for estimating $\Phi(z)$ based on observed data can be described as follows:

1. Initialize: $\widehat{\Phi}_0(z_t) = \overline{y}$ for each t.
2. For $m = 1, \ldots, M$:

(a) for $t = 1, \ldots T$, compute the negative gradient vector $u_t = \frac{-\partial C(y_t, \Phi)}{\partial \Phi}|_{\Phi=\widehat{\Phi}_{m-1}(z_t)}$. Under the quadratic loss function, $u_t = y_t - \widehat{\Phi}_{m-1}(z_t)$;

(b) fit a base learner (such as a regularized regression or a spline) to the gradient vector to yield $\widehat{\phi}_m$. For example, with regularized regression, $\widehat{\phi}_m(z_t) = z_t'\widehat{\beta}$, where $\widehat{\beta} = \text{argmin}_\beta \Sigma_{t=1}^T (u_t - z_t'\beta)^2 + \lambda\|\beta\|^2$ for some $\lambda > 0$.

3. Update $\widehat{\Phi}_m(\cdot) = \widehat{\Phi}_{m-1}(\cdot) + \nu\widehat{\phi}_m(\cdot)$, where $0 < \nu \leq 1$ is the step length.

The algorithm estimates $\Phi(z)$ as the sum of M estimated fitting procedures (or base learners), $\widehat{\phi}_m(z)$, to give $\widehat{\Phi}_M(z) = \widehat{\Phi}_0(z) + \nu\Sigma_{m=1}^M \widehat{\phi}_m(z)$. As indicated in Step 2, boosting is a 'stage-wise forward regression' as it shares the property that variables are included sequentially in a step-wise regression. Every boosting procedure entails the choice of a learner in step 2(b). A learner is 'weak' if it is simple, involves few parameters, and has a large bias relative to variance. Some popular learners are smoothing splines, least squares regression, and kernel regressions. Step 3 shows that boosting is an 'ensemble scheme' that aggregates many function estimates of the reweighted data (or the so-called psuedo residuals). Note that at the mth iteration $\nu \cdot \widehat{\phi}_m$ is added to the overall fit,

DOI: 10.1002/jae

and not the entire $\widehat{\phi}_m$. Thus boosting not only performs subset variable selection but also performs coefficient shrinkage.

Freund (1995) and Schapire (1990) introduced Adaboost as a classification device for $y \in \{-1, 1\}$, and with $C(y, \Phi(z)) = \exp(y\Phi(z))$. The remarkable resistance of Adaboost to overfitting has generated a lot of research. In an important paper, Friedman (2001) presented boosting as a gradient descent technique in function space that iteratively looks for the steepest descent of $\widehat{\Phi}(z)$ to reach the minimum of the loss function. This provides a formal link between boosting as a tool in machine-learning analysis and as a formal statistical procedure (see Friedman *et al*., 2000). Different loss functions will yield different boosting algorithms. Adaboost is now understood to be similar to maximizing the negative of the binomial likelihood. The usage of boosting has expanded from classification of univariate variables to (generalized) regressions, survival analysis, and to systems analysis. It has been used in biostatistical analyses to link health outcomes to gene expressions, and in classification analysis such as spam data that are often analyzed in the machine-learning literature.

Boosting has several practical and theoretical advantages when used to analyze data that are independent and identically distributed. It can handle high-dimensional data well with low computational cost, and when the data truly have a sparse structure it can produce models that do not tend to overfit. The bias–variance trade-off that underlies boosting is in sharp contrast to that of nonparametric estimation, such as a smoothing spline. In spline regressions, one often chooses the smoothing parameter, say λ, to control the bias–variance trade-off. In boosting with spline learners, one fixes a λ such that the base procedure has a low variance but possibly a high bias. This bias is then reduced by boosting iterations. Buhlmann and Yu (2003) showed that under quadratic loss (L_2) boosting with smoothing spline learners achieves minimax optimal mean-squared error (MSE) rates. The authors showed that at each iteration m the bias decreases while the variance increases at an exponential rate. It is this exponential trade-off that puts the MSE at the optimal rate achievable by smoothing splines.[2] Buhlmann and Hothorn (2007) provided an excellent introduction to boosting from a statistical perspective. Our interest is in the application of boosting to macroeconomic data, and we will now focus on a loss function and base learner that can most easily accommodate such data.

For the remainder of the paper, $\Phi(z)$ stands for the conditional mean, a scalar function of n-dimensional variable z, and $\widehat{\Phi}_m(z)$ is the boosting estimator of $\Phi(z)$ at the m-stage of boosting. In the L_2 boosting to be discussed below, these functions are evaluated at the data points $z_1, z_2, \ldots, z_T$, where z_t is $n \times 1$. We define

$$\Phi = (\Phi(z_1), \ldots, \Phi(z_T))', \text{ and } \widehat{\Phi}_m = (\widehat{\Phi}_m(z_1), \ldots, \widehat{\Phi}_m(z_T))'$$

each being a $T \times 1$ vector. Therefore, $\Phi(z)$ is a function and Φ is a vector. Their meaning can be discerned from the context. Since they represent the same object, clarity should not be affected even without making a distinction between the two. The same can be said for $\widehat{\Phi}_m(z)$ and $\widehat{\Phi}_m$. Also, we use $\widehat{\Phi}_{t,m}$ to denote the tth component of $\widehat{\Phi}_m$, as in Algorithms 1 and 2 below.

2.1. L_2 Boost of i.i.d. Data

As mentioned earlier, minimizing the expected quadratic loss $C(y, \Phi(z)) = |y - \Phi(z)|^2/2$ leads to the well-known result that $E(y|Z = z)$ is the population minimizer. Buhlmann and Yu (2003)

[2] A different ensemble method is bagging. While bagging is primarily a variance reduction technique, boosting also reduces bias via its flexibility in combining models (see, for example, Rosset, 2005).

DOI: 10.1002/jae

termed the boosting algorithm that minimizes quadratic loss as L_2-boost. If the base learner is a linear regression, step 2(b) under L_2-boost can be rewritten as $\widehat{\phi}_m = Su_{m-1}$, where u_{m-1} are the least-squares residuals at the end of step $m - 1$, and S is a boosting operator that maps $y_1, \ldots y_T$ to its fitted values. At the end of the m-step, the fitted conditional mean is $\widehat{\Phi}_m = B_m Y$ if $v = 1$, where $B_m = I_T - (I_T - S)^{m+1}$ and $Y = (y_1, \ldots, y_T)^{'}$. Note that if the eigenvalues of $I_T - S$ are all less than 1, $B_m \to I_T$ as $m \to \infty$, and this saturated model will give a perfect fit. We therefore want to terminate step 2 at some $m = M$. The choice of regularization parameter M will be discussed below.

If we had performed ordinary least squares (OLS) on all n potential predictors, we would get $\widehat{\Phi} = PY$, where P is the projection matrix formed from the n regressors. However, under boosting with OLS as the learner, $\widehat{\Phi}_m = B_m Y$ and $B_m \neq P$, so $\widehat{\Phi}_m \neq \widehat{\Phi}$. Thus, in contrast to OLS, which takes one greedy step towards the final model, boosting makes many small adjustments, where 'greedy' means that at each step the function that leads to the largest reduction of the error is added to the estimator. The resistance by boosting to overfitting has much to do with this stage-wise approach to model fitting.

We are interested in using boosting for estimation and prediction when (i) the number of predictors is large, (ii) the data are dependent, and (iii) some and possibly all of the predictors are 'generated'. The next subsection deals with the large number of dynamic predictors. Section 3 discusses how generated regressors will be accommodated.

2.2. L_2 Boost of Dynamic Models

Suppose y_t has zero mean. Let z_t be the vector of predictors considered, and let n be the dimension of z_t. In the absence of dynamics, the predictors are $z_t = (w_t^{'}, X_t^{'})^{'}$, where we recall that w_t are 'usual suspects' predictors that researchers always choose to include. In a dynamic context, the set of potential predictors are the current and lagged values of y_t and X_t as well as functions of these variables. Thus a reasonable base case set of predictors is $z_t = (Z_t, Z_{t-1}, \ldots Z_{t-\text{pmax}})$ where $Z_t = (y_{t-1}, w_t^{'}, X_{t1}, \ldots X_{tN}, X_{t1}^2, \ldots X_{tN}^2)^{'}$ for some prespecified pmax. We will denote the dimension of Z_t by $\overline{N}$.

Given the set of potential predictors z_t, we want to use boosting to fit a model for y_t. A question immediately arises: how to deal with the lags. We present two possibilities. The first one, which will be referred to as component-wise boosting, treats each lag of each variable as a separate predictor. The second considers each of the $\overline{N}$ variables and their lags jointly. We refer to this method as block-wise L_2 boost. Details are as follows.

Algorithm 1 Component-wise L_2 boost When the number of potential predictors is large, a convenient method, considered in Buhlmann and Yu (2003), is to fit learners using one predictor at a time. Let $z_{\cdot,i}$ denote the vector of T time series observations for the ith variable in the potential set. The predictor being selected at the mth round, denoted $z_{\cdot,i_m^*}$, is such that

$$i_m^* = \underset{i=1,\ldots n}{\text{argmin}} \sum_{t=1}^{T} (u_t - \widehat{\phi}_m(z_{t,i}))^2$$

DOI: 10.1002/jae

That is, variable i_m^* has the smallest sum-of-squared residuals amongst all predictors considered in the mth step. The component-wise L_2 boost with linear least squares as the base learner can be implemented as follows:

```
1. let $\widehat{\Phi}_{t,0} = \overline{y}$ for each t.
2. For $m = 1, \ldots M$:
  (a) for $t = 1, \ldots T$, let $u_t = y_t - \widehat{\Phi}_{t,m-1}$ be the ‘current residuals’;
  (b) for each $i = 1, \ldots n$, regress the current residual vector $u$ on $z_{.,i}$ (the $i$th regressor) to obtain
      $\widehat{b}_i$. Compute the $\widehat{e}_{.,i} = u - z_{.,i}\widehat{b}_i$, as well as $\mathrm{SSR}_i = \widehat{e}'_{.,i}\widehat{e}_{.,i}$;
  (c) let $i_m^*$ be such that $\mathrm{SSR}_{i_m^*} = \min_{i \in [1,\ldots n]} \mathrm{SSR}_i$;
  (d) let $\widehat{\phi}_m = z_{.,i_m^*}\widehat{b}_{i_m^*}$.
3. For $t = 1, \ldots T$, update $\widehat{\Phi}_{t,m} = \widehat{\Phi}_{t,m-1} + \upsilon\widehat{\phi}_{t,m},$ where $0 < \upsilon \leq 1$ is the step length.
```

If Z_t is $\overline{N} \times 1$ pmax lags are entertained, there will be $n = \overline{N} \times$ pmax elements in z_t. Component-wise boosting treats each of these n variables as distinct. As far as we are aware, treating each lag of a variable as a component has not been considered in the context of time series data.

Algorithm 2 Block-wise L_2 boost

```
1. Let $\widehat{\Phi}_{t,0} = \overline{y}$ for each $t$.
2. For $m = 1, \ldots M$:
  (a) for $t = 1, \ldots T$, let $u_t = y_t - \widehat{\Phi}_{t,m-1}$ be the ‘current residuals’,
  (b) for each $i = 1, \ldots \overline{N}$:
     (i) for $p = 1, \ldots$ pmax, estimate the model $u_t = a_1 Z_{t,i} + a_2 Z_{t-1,i} + \ldots a_p Z_{t-p,i} + \upsilon_{tp}$. For given
         $A_T$ that depends on the desire for a parsimonious model, compute

         $$\mathrm{IC}(p) = \log(\widehat{\sigma}^2_{\upsilon p}) + A_T \frac{p}{T} \qquad (3)$$

         where $\widehat{\sigma}^2_{\upsilon p} = T^{-1}\sum_{t=1}^{T} \widehat{\upsilon}^2_{tp}$;
     (ii) let $p_i^* = \mathrm{argmin}_p \mathrm{IC}(p)$;
     (iii) let $\widehat{b}_i$ be the least-squares estimates obtained by regressing $u_t$ on $z_{ti}$, where $z_{ti} = (Z_{t,i},$
           $\ldots Z_{t-p_i^*,i})'$. Compute the $\widehat{e}_{.,i} = u - z_{.,i}\widehat{b}_i$, as well as $\mathrm{SSR}_i = \widehat{e}'_{.,i}\widehat{e}_{.,i}$;
  (c) let $i_m^*$ be such that $\mathrm{SSR}_{i_m^*} = \min_{i \in [1,\ldots \overline{N}]} \mathrm{SSR}_i$;
  (d) let $\widehat{\phi}_m = z_{.,i_m^*}\widehat{b}_{i_m^*}$.
3. For $t = 1, \ldots T$, update $\widehat{\Phi}_{t,m} = \widehat{\Phi}_{t,m-1} + \nu\widehat{\phi}_{t,m}$, where $0 < \nu \leq 1$ is the step length.
```

Our block boosting treats lags of the same variable as a group. It has some similarity to the partial boosting device of Tutz and Binder (2005), which separates the ‘must have’ predictors from the rest by treating the former set of variables as a block. In our set-up, there is no ‘must have’ variable(s).

Note that component-wise L_2 boost selects one predictor at each iteration, but the same predictor can be selected more than once during the M iterations. Similarly, block-wise L_2 boost selects the predictor and its lags jointly at each iteration, and the same set of predictors can be selected more than once during M iterations. Algorithms 1 and 2 can both be desirable, depending on whether one is interested in a consistent estimate of the true but sparse model, or in an efficient estimate of $\Phi(z)$ in a mean-squared error sense. An example makes this precise. Suppose the data-generating process is $y_t = \beta_{11}y_{t-1} + \beta_{21}X_{t-2,1} + \beta_{32}X_{t-3,2} + e_{t+h}$, but β_{11} and β_{21} are small, though non-zero. We are given $N = 4$ predictors and want to entertain up to 4 lags of each variable as predictors of y. Given data on $(y_t, Z_t), t = 1, \ldots T - h$, Algorithm 1 is more likely to pick $X_{t-3,2}$ as the only predictor, and restrict β_{11}, β_{21}, as well as the coefficient on all other lags to zero. On the other hand, Algorithm 2 will likely pick lagged y along with X_1 and X_2 as predictors with lag lengths of 1, and 3 respectively, leaving β_{11}, β_{12} and β_{22} unconstrained. Algorithm 1 is more parsimonious, but may have a larger mean-squared error if the reduction in variance induced by the two restrictions is not enough to offset the corresponding increase in bias. Also, a procedure that produces a small in-sample forecast error need not perform well out-of-sample. Which algorithm is more desirable will depend on the application on hand.

Whether we use Algorithm 1 or 2, the objective of L_2 boost is still to estimate the conditional mean, $\Phi(z)$. If we let β be the $n \times 1$ parameter vector associated with the predictors $z_t = (z_{t1}, \ldots, z_{tn})'$, then the 'in-sample' fit is

$$\widehat{\Phi}_m(z) = \overline{y} + z'\widehat{\beta}_m$$

and the out-of-sample estimate, given predictor z_{T+h}, is

$$\widehat{\Phi}_m(z_{T+h}) = \overline{y} + z'_{T+h}\widehat{\beta}_m$$

The estimator $\widehat{\beta}_m$ can be shown to follow the recursion (with $\widehat{\beta}_0 = 0$);

$$\widehat{\beta}_m = \widehat{\beta}_{m-1} + \nu\widehat{b}^{\dagger}_m,$$

where $\widehat{b}^{\dagger}_m$ is non-zero only in the i^*_mth position in the case of component-wise boosting, and also non-zero in positions corresponding to lags of the i^*_mth variable in the case of block boosting. The non-zero element of $\widehat{b}^{\dagger}_m$ is equal to $\widehat{b}_{i^*_m}$. Thus, $\widehat{\beta}_m$ and $\widehat{\beta}_{m-1}$ differ only at positions relating to the i^*_mth variable. At the final step, $\widehat{\beta}_M$ will likely have many zero elements, which is a direct consequence of subset variable selection.

The degrees of freedom associated with boosting that stops at iteration m is

$$\text{d.f.}_m = \text{trace}(B_m)$$

where

$$B_m = B_{m-1} + \nu P^{(m)}(I_T - B_{m-1}) = I_T - \prod_{j=0}^{m}(I_T - \nu P^{(j)})$$

where $P^{(m)} = z_{.i^*_m}(z'_{.i^*_m}z_{.i^*_m})^{-1}z'_{.i^*_m}$ is the projection matrix formed by the i^*_mth regressor $z_{.i^*_m}$ in component-wise boosting, or a block of regressors in block boosting $(m = 1, 2, \ldots, M)$. The staring value $B_0 = \frac{1}{\nu}P^{(0)} = \iota_T\iota_T/T$, where ι_T is a $T \times 1$ vector of 1's. Also note that the vector of fitted values at stage m is $\widehat{\Phi}_m = B_mY$.

DOI: 10.1002/jae

The stopping parameter M can be defined by an information criterion.[3] Let $\widehat{\sigma}_m^2 = \sum_{t=1}^{T}(y_t - \widehat{\Phi}_{t,m})^2$. Then $M = \text{argmin}_m\, IC(m)$ where

$$IC(m) = \log(\widehat{\sigma}_m^2) + \frac{A_T \cdot \text{d.f.}_m}{T} \tag{4}$$

The modified AIC and BIC can be obtained with $A_T = 2$ and log (T), respectively. This is a modified IC because the complexity of the model is evaluated at d.f.$_m$. In contrast, the standard IC penalty uses the number of estimated β coefficients instead of d.f.$_m$. Because a regressor can be used in more than one of the M iterations, counting the number of non-zero coefficient estimates is not the best measure of model complexity. Criterion (4) permits the number of predictors to be larger than the number of observations (see Buhlmann, 2006).

The bias and in-sample variance of $\widehat{\Phi}_M(z)$ are, respectively,

$$\begin{aligned} \text{bias}(\widehat{y}|z) &= E[\widehat{\Phi}_M(z) - \Phi(z)] \\ \text{var}(\widehat{y}|z) &= \text{var}(\widehat{\Phi}_m(z)) = E[\widehat{\Phi}_M(z) - E\widehat{\Phi}_M(z)]^2 \end{aligned}$$

Because $\widehat{\Phi}_M(z)$ is a non-linear and non-differentiable function of y, the evaluation of sampling uncertainty is not a trivial exercise. Knight and Fu (2000) suggested using the bootstrap to obtain standard errors for both $\widehat{y}$ and $\widehat{\beta}_M$.

3. L_2 BOOSTING WITH ESTIMATED FACTORS

In Section 2, we merely noted $\tilde{F}_t$ as factors estimated from a large panel of data by the method of principal components. We now make precise how F_t is estimated. Let $X_t = (x_{1t}, x_{2t}, \ldots, x_{Nt})'$ and $\Lambda = (\lambda_1, \ldots \lambda_N)'$ so that the factor model (2) can be written in vector form:

$$X_t = \Lambda F_t + e_t$$

Let $X = (X_1, X_2, \ldots, X_N)$, a $T \times N$ data matrix, and let $F = (F_1, F_2, \ldots, F_T)'$, a $T \times r$ matrix. The factor model can be further rewritten in matrix form $X = F\Lambda' + e$. The principal components estimator for the factors, F, is $\tilde{F} = (\tilde{F}_1, \ldots, \tilde{F}_T)'$. This is a $T \times r$ matrix consisting of r eigenvectors (multiplied by $\sqrt{T}$) associated with the r largest eigenvalues of the matrix $XX'/(TN)$ in decreasing order. Thus $T^{-1}\sum_{t=1}^{T}\tilde{F}_t\tilde{F}_t' = I_r$. Then $\tilde{\Lambda} = (\tilde{\lambda}_1, \ldots, \tilde{\lambda}_N)' = X'\tilde{F}/T$. The number of factors r is chosen by the information criterion approach developed in Bai and Ng (2002). Also let $\tilde{V}$ be the $r \times r$ diagonal matrix consisting of the r largest eigenvalues of $XX'/(TN)$, and $H = \tilde{V}^{-1}(\tilde{F}'F/T)(\Lambda'\Lambda/N)$. The matrix H depends on N and T, but this dependence will be suppressed for notational simplicity.

Since H is invertible, HF_t and F_t will provide the same prediction when used as predictors. However, both HF_t and F_t are unobservable. The next lemma, proved in Bai and Ng (2002) and Bai (2003), shows that $\tilde{F}_t$ is close to HF_t for all t, justifying the use of $\tilde{F}_t$ as predictor variables.

[3] For small samples, the corrected AIC, $\text{AIC}_c(m) = \log(\widehat{\sigma}_m^2) + \dfrac{1 + \text{d.f.}_m/T}{1 - (\text{d.f.}_m + 2)/T}$ is also used in the boosting literature.

DOI: 10.1002/jae

Lemma 1 *Let $H = \tilde{V}^{-1}(\tilde{F}'F/T)(\Lambda'\Lambda/N)$. Suppose Assumption (A) as stated in Bai and Ng (2006) holds. Let $\delta_{NT} = \min[N^{1/2}, T^{1/2}]$. Then*

(i) $\frac{1}{T}\sum_{t=1}^{T}\left\|\tilde{F}_t - HF_t\right\| = O_p(\delta_{NT}^{-2})$;

(ii) $\sqrt{N}(\tilde{F}_t - HF_t)\xrightarrow{d}N(0, V^{-1}Q\Gamma_t Q'V^{-1}) \equiv N(0, \text{Avar}(\tilde{F}_t))$ *where* $Q = \text{plim}\ \tilde{F}'F/T$, $V = \text{plim}\tilde{V}$, *and* $\Gamma_t = \lim_{N\to\infty}\frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{N}E(\lambda_i\lambda_j' e_{it}e_{jt})$.

Assumption A provides the conditions for an approximate factor model that allows some heteroskedasticity and correlation in the errors, both in the time series and cross-section dimensions. The results stated in Lemma 1 were used in Bai and Ng (2006) to show that if $\sqrt{T}/N \to 0$ as $N, T \to \infty$, the estimates obtained from the factor-augmented regressions are $\sqrt{T}$ consistent and asymptotically normal. In other words, the factors can be treated as though they are known. Recall that variables with a 'tilde' means that they are principal components estimates. Our set of potential predictors is thus $\tilde{z}_t^F = (\tilde{Z}_t^F, \tilde{Z}_{t-1}^F, \ldots \tilde{Z}_{t-\text{pmax}}^F)$, where $\tilde{Z}_t^F = (y_t, \tilde{F}_{t1}, \ldots, \tilde{F}_{tr}, \tilde{F}_{t1}^2, \ldots \tilde{F}_{tr}^2)$.

An issue that remains to be considered is whether a generated (estimated) predictor should be treated the same as an observed one. To better understand the problem, it is useful to first consider the mean-squared forecast error when all predictors are observed. For a generic regression model $y_{t+1} = z_t'\beta + e_{t+1}$, where z_t is $n \times 1$, the forecast is $\widehat{y}_{T+1|T} = z_T'\widehat{\beta}$. The MSE is

$$\begin{aligned} E[(\widehat{y}_{T+1|T} - y_{T+1})^2] &= E[(z_T'(\widehat{\beta} - \beta) - e_{T+1})^2] \\ &= \sigma^2 + E[z_T'(\widehat{\beta} - \beta)(\widehat{\beta} - \beta)'z_T] \end{aligned}$$

Assuming $\sqrt{T}(\widehat{\beta} - \beta) \sim N(0, \sigma^2 Q_z^{-1})$, where $Q_z = E(z_t z_t')$, T times the second term in brackets is a χ^2 random variable with n degrees of freedom whose expectation is n. This gives MSE $\approx \sigma^2(1 + n/T)$. Replacing σ^2 by the estimator $\frac{T}{T-n}\widehat{\sigma}_n^2$, where $\widehat{\sigma}_n^2$ is the MLE of σ^2, we obtained the estimated MSE, $\widehat{\text{MSE}} \approx \widehat{\sigma}_n^2(T + n)/(T - n)$. The log of the estimated MSE is approximated by

$$\log(\widehat{\text{MSE}}) \approx \log\widehat{\sigma}_n^2 + \frac{2n}{T-n}$$

As discussed in Brockwell and Davies (1991), this leads to the FPE (final predictor error) criterion for choosing n as the minimizer of log(MSE). The AIC given in (4) can similarly be motivated, except that the AIC replaces $T - n$ by T.

Now suppose $n_1 \leq n$ of the predictors are generated. Let $\tilde{z}_t = (z_{t1}, \ldots z_{t,n-n_1}, \widehat{z}_{t,n-n_1+1}, \ldots \widehat{z}_{t,n})$, where $\widehat{z}_{t,n-n_1+1} \ldots \widehat{z}_{t,n}$ are themselves estimates. The regression is

$$y_{t+1} = \tilde{z}_t'\beta + (z_t - \tilde{z}_t)'\beta + e_{t+1} = \tilde{z}_t'\beta + \upsilon_t$$

with $\upsilon_t = (z_t - \tilde{z}_t)'\beta + e_{t+1}$. The one-step-ahead prediction is

$$\begin{aligned} \widehat{y}_{T+1|T} - y_{T+1} &= \tilde{z}_T'\widehat{\beta} - z_T'\beta - e_{T+1} \\ &= \tilde{z}_T'(\widehat{\beta} - \beta) + (\tilde{z}_T - z_T)'\beta + (\tilde{z}_T - z_T)'(\widehat{\beta} - \beta) - e_{T+1} \end{aligned}$$

DOI: 10.1002/jae

Assuming the third term on the right-hand side is dominated by the first two, which will be the case if $\tilde{z}_T - z_T = o_p(1)$, we have

$$\text{MSE} \approx \sigma^2 + E[\tilde{z}_T'(\widehat{\beta}-\beta)(\widehat{\beta}-\beta)'\tilde{z}_T] + \beta' E[(\tilde{z}_T - z_T)(\tilde{z}_T - z_T)']\beta$$

The consequence of generated predictors is to add the last term to the MSE. If $\sqrt{T}(\widehat{\beta}-\beta) \sim N(0, \sigma^2 Q_z^{-1})$, the second term is again approximately $\sigma^2 n/T$. Next consider the contribution of the last term:

$$\beta' E[(\tilde{z}_T - z_T)(\tilde{z}_T - z_T)']\beta \tag{5}$$

In the existing literature on generated regressors, e.g., Pagan (1984), $\text{var}(\tilde{z}_t - z_t) = O_p(T^{-1})$ for all t. More specifically, under the assumption that $\sqrt{T}(\tilde{z}_T - z_T) \xrightarrow{d} N(0, \Sigma_z)$, the contribution of (5) is approximated by $\beta'\Sigma_z\beta/T$. Thus the MSE can be written as

$$\text{MSE} \approx \sigma^2\left(1 + \frac{n}{T} + \frac{c_n}{T}\right)$$

where $c_n = \beta'\Sigma_z\beta/\sigma^2$. Again, replacing σ^2 by $\frac{T}{T-n}\widehat{\sigma}_n^2$, and replacing c_n by $\widehat{c}_n = \widehat{\beta}'\widehat{\Sigma}_z\widehat{\beta}/\widehat{\sigma}_n^2$, we obtain the estimated MSE, $\widehat{\text{MSE}}$. Its log form is approximately

$$\log(\widehat{\text{MSE}}) \approx \log(\widehat{\sigma}_n^2) + \frac{2n}{T-n} + \frac{\widehat{c}_n}{T-n} \tag{6}$$

The denominator $T-n$ may be replaced by T akin to the AIC criterion. The above is a modified FPE for choosing n in the presence of generated predictors; a generated predictor is penalized more heavily than an observed one. Note the overall penalty still decreases at rate T. However, the additional penalty can be important in finite samples, a consideration that appears to have been overlooked.

Our current problem is similar but non-standard because $\tilde{z}_t$ are estimated by the method of principal components. More precisely, our $\tilde{z}_t$ is $\tilde{F}_t$, which is $\sqrt{N}$ consistent for $z_t = HF_t$, in contrast to standard generated regressors which are $\sqrt{T}$ consistent, as given in Lemma 1(ii). Thus the additional penalty is of $O(N^{-1})$. Based on the above analysis, we propose the following procedure. Suppose $\tilde{F}_t$ are the r factors estimated by the method of principal components from a T by N matrix of data:

(a) (lag length selection) For an arbitrary factor $j \in [1, r]$, let $\widehat{\sigma}_{jp}^2$ be the variance from estimating the distributed lag model $y_{t+1} = (\tilde{F}_{tj}\tilde{F}_{t-1,j}\ldots\tilde{F}_{t-p,j})'\psi_j + \varepsilon_{t+1,j,p}$ by OLS. Let

$$p_j^* = \underset{p=1,\ldots\text{pmax}}{\text{argmin}} \log(\widehat{\sigma}_{jp}^2) + A_T\frac{p}{T} + A_N\frac{p}{N} \tag{7}$$

(b) (boosting stopping rule) Consider a set of n potentially estimated predictors, $\tilde{z}_t = (\tilde{Z}_t, \ldots\tilde{Z}_{t-\text{pmax}})$. Given some pre-specified $\overline{M}$, boosting stops at step M where

$$M = \underset{m=1,\ldots\overline{M}}{\text{argmin}} \text{IC}(m), \quad \text{IC}(m) = \log(\widehat{\sigma}_m^2) + A_T\frac{\text{d.f.}_m}{T} + A_N\frac{c_m}{N} \tag{8}$$

DOI: 10.1002/jae

where $\text{d.f.}_m = \text{trace}(\tilde{B}_m)$, $\tilde{B}_m = I_T - \Pi_{j=0}^{m}(I_T - \mu\tilde{P}^{(j)})$, $\tilde{P}^{(m)}$ is similar to $P^{(m)}$ defined in the previous section with $z_{.i_m^*}$ replaced by $\tilde{z}_{.i_m^*}$, and c_m is an approximation to $\beta^{'}E[N(\tilde{z}_T - z_T)(\tilde{z}_T - z_T)^{'}]\beta/\sigma^2 = \beta^{'}\Sigma_z\beta/\sigma^2$, which can be consistently estimated. In simulations, the results are similar to using $c_m = m_1$, where m_1 is the number of estimated predictors. Indeed, $A_N(m_1/N)$ captures the main idea that there are m_1 predictors whose sampling errors vanish at rate N, and is what we use in applications. The AIC obtains when $A_T = A_N = 2$, and the BIC obtains when $A_T = \log(T), A_N = \log(N)$.

The proposed lag length selection and the boosting stopping rule penalizes an additional regressor more heavily if that regressor is an estimated factor than if the regressor is observed. The additional penalty vanishes at rate N, which is the cross-section dimension of the panel of data from which the factors are estimated. The overall penalty of an additional predictor vanishes at rate min [N, T]. Factors that do not belong to the prediction equation (i.e., $\beta_j = 0$) cannot contribute to the sampling variability of y and thus will not be penalized beyond the usual bias–variance trade-off.

Note that if $\tilde{F}_{tj}$ is being estimated, so are its lags. Thus, when using block-wise boosting, (7) should be used in step (2) for lag length selection. Of course, under Algorithm 1 when each lag is treated as a separate regressor, (7) is irrelevant. On the other hand, (8) should be used as a boosting stopping rule when some and possibly all of the predictors are factor estimates.

The large sample properties of boosting when all the predictors are observed have been an active area of research in statistics. Zhang and Yu (2005) showed convergence of the boosting solution to the infinmum of the loss function over the linear span of the predictors. In an i.i.d. setting, Buhlmann (2006) showed that L_2 boost yields consistent estimates in high-dimensional linear models when the number of predictors is allowed to grow as fast as the sample size, and assuming that the true underlying model is sparse in terms of the L_1 norm of regression coefficients. More precisely, the author showed (in our notation) that $\int|\widehat{\Phi}_{m_T}(z) - \Phi(z)|^2 \mathrm{d}F_n(z) = o_p(1)$ as $T \to \infty$ with $m_T = o_p(T^{\eta})$ for some $\eta > 0$, where $F_n(z)$ is the cdf of z_t under i.i.d. assumption for z_t. Lutz and Buhlmann (2006) extend Buhlmann's result to a multivariate, time series setting and shows that L_2 boosting recovers the true sparse regression function even if the dimension of the predictors increase with the sample size.

Proposition 1 *Let $\widehat{\Phi}_M(z)$ be the boosting estimate for the conditional mean when all the predictors are observable, and let $\tilde{\Phi}_M(z)$ be the boosting estimate when some or all of the predictors are estimated by the method of principal components. If $\sqrt{T}/N \to 0$ and boosting terminates at step M with $\dfrac{M}{\min[\sqrt{N}, \sqrt{T}]} \to 0$, as $M, NandT \to \infty$, then $\frac{1}{T}\sum_{t=1}^{T}|\tilde{\Phi}_M(z_t) - \widehat{\Phi}_M(z_t)|^2 = o_p(1)$, and for each given z, $|\tilde{\Phi}_M(z) - \widehat{\Phi}_M(z)| = o_p(1)$.*

Because boosting repeatedly fits a model using the estimated predictors, the error from estimating the predictors accumulates. The proposition, puts an upper bound on the boosting stopping rule. If this condition is satisfied, then together with the result that boosting can consistently estimate the structure of the sparse model when the factors are observed, we have the result that boosting will also consistently estimate the sparse structure if the latent factors are replaced by the principal components estimates.

DOI: 10.1002/jae

4. SIMULATIONS AND APPLICATIONS

In this section, we simulate data from two data-generating processes (DGPs) to assess the effectiveness of boosting in a FAR framework. We will use C_j to denote component-wise boosting, and B_j to denote block-wise boosting, where j will be defined below.

DGP 1 For $j = 1, \dots, rmax$, $i = 1, \dots N$, and $t = 1, \dots T$:

$$x_{it} = \lambda_i' F_t + \sqrt{rmax} e_{it}$$
$$F_{jt} = \alpha_j F_{jt-1} + u_{jt}$$
$$e_{it} = \rho_i e_{it-1} + \sigma_\varepsilon \varepsilon_{it}$$

The parameter σ_ε controls the strength of the factors in the predictors. The larger is σ_ε, the smaller is the common relative to the idiosyncratic component. The objective is to forecast y_{t+h}, where

$$y_{t+h} = \beta_1 F_{1t} + \beta_2 F_{3t-4} + \beta_4 F_{6t}^2 + \sigma_y v_{t+h}$$

We let $\beta = (0.8, 0.5, 0.3)'$, $\alpha_j \sim U[0.2, 0.8]$, and $\rho_i \sim U[0.3, 8]$. These are drawn once and held fixed during simulations. The factor loadings are $\lambda_i \sim 0.5N(0, rmax)$, while the shocks are $(u_{jt}, \varepsilon_{it}, v_{t+h}) \sim N(0, I_3)$. We consider seven configurations of $(\sigma_\varepsilon, \sigma_y)$.

DGP 2 Generates x_{it} as in DGP 1, but y is specified as

$$y_{t+h} = \beta_1 F_{2t} + \beta_2 F_{2t-1} + \beta_3 F_{2t-2} + \beta_4 F_{4t-1} + \beta_5 F_{4t-2} + v_{t+h}$$

with $\beta = (0.8, 0.5, 0.2, 0.4, 0.4)$.

We use boosting to select predictors from the forecasting equation:

$$y_{t+h} = \kappa_0 + \alpha(L) y_t + \beta(L) \tilde{f}_t + \varepsilon_{t+h}$$

where the choice of $\tilde{f}_t$ is not limited to a subset $\tilde{F}_t$. More specifically, $\tilde{f}_t$ is formed from one of six potential predictor sets:

(i) $\tilde{f}_t = (x_{1t}, \dots x_{Nt})' = X_t$;
(ii) $\tilde{f}_t = (x_{1t}, \dots x_{Nt}, x_{1t}^2, \dots x_{Nt}^2)'$;
(iii) $\tilde{f}_t = (\tilde{F}_{1t}, \dots \tilde{F}_{rmax,t})'$, the *rmax* principal components of the data matrix $T \times N$ matrix $X = (X_1, \dots, X_N)$.
(iv) $\tilde{f}_t = (\tilde{F}_{1t}, \dots \tilde{F}_{rmax,t}, \tilde{F}_{1t}^2, \dots \tilde{F}_{rmax,t}^2)'$;
(v) $\tilde{f}_t = (\tilde{G}_{1t}, \dots \tilde{G}_{rmax,t})'$, the *rmax* principal components of the $T \times 6N$ matrix X^*, where $x_{it}^* = (x_{it}, x_{it}^2, x_{it-1}, x_{it-2}, x_{it-3}, x_{it-4})'$;
(vi) $\tilde{f}_t = (\tilde{G}_{1t}, \dots \tilde{G}_{rmax,t}, \tilde{G}_{1t}^2, \dots \tilde{G}_{rmax,t}^2)'$.

The first set of predictors consists of just the observed variables, $X = (X_1, X_2, \dots, X_N)$, and the second consists of the observed variables and their squared values. The third set is the *rmax* principal components of XX'. Predictor set (iv) adds the squares of the factors. Predictor set

DOI: 10.1002/jae

(v) forms principal components from an expanded dataset as suggested by a referee, and set (vi) further adds the squares of the principal components. We could have considered an even larger set of predictors by considering all the cross-products of x_{it}. Given that the DGPs do not include such terms, this is not considered.

For each predictor set, we use either block-wise or component-wise boosting to select which lags of y_t, and which elements of $\tilde{f}_t$ and their lags to enter the forecasting equation. Thus, for each DGP, there are 12 sets of results. We use the BIC both for lag length selection and for choosing the stopping rule, M. Thus $A_T = log(T)$ and $A_N = log(N)$. The results are slightly better when the BIC is used, and to conserve space the results for the AIC will not be reported.

We simulate 250 observations of the variables and discard the first 50 observations, leaving 200 observations for evaluation. The first estimation uses data from $t = 1, \ldots 101$ to perform a three-period-ahead forecast, i.e., $T + h = 104$. Then T is incremented by 1, the estimation is repeated, and a forecast for $T + h = 105$ is performed. The last forecast of $T + h = 200$ is based on estimation using data up to $T = 197$. Each forecast $\widehat{y}_{T+h|T}$ is then compared to y_{T+h}. We compare the different forecasting methods using three criteria. Based on the 98 forecasts, we compute (i) median bias, (ii) the root-mean-squared forecast error (RMSE) relative to the variance of y, and (iii) the mean-squared forecast error relative to an AR(4) forecast. The median bias is reported as the mean bias tends to be distorted by unusually bad forecasts. The AR(4) has often been used as a benchmark in the literature on diffusion index forecasting. A relative RMSE bigger than one means that the AR(4) outperforms the method considered. Normalizing by the variance of the series to be forecast provides a comparison that does not depend on the choice of benchmark.

4.1. Results

Before turning to the specific DGPs, some general observations are of note. First, the method that yields that lowest bias tends not to be the method that has the lowest RMSE. This is true whether we use the AIC or BIC to select the stopping rule. The modified information criterion that takes into account generated regressors tends to have smaller bias, but the best method tends not to depend on the correction factor. We report these results in Tables I–IV but they will not be separately discussed.

The results for DGP 1 are reported in Table I. The column labeled $X(C)$ refers to component-wise boosting with $\tilde{f}_t = (x_{1t}, \ldots, x_{Nt})$, i.e., predictor set (i) of Section 4. The $X(B)$ column refers to block boosting with the same predictor set as $X(C)$. Similarly, the $X^2(C)$ column refers to component-wise boosting with $\tilde{f}_t$ being predictor set (ii) of Section 4, and $F(C)$ refers to predictor set (iii), and so on. Whether one considers bias or RMSE as criteria, boosting the factors tends to be better, in general, than boosting the observed variables. This should not be surprising given the DGP. For this DGP, component-wise boosting tends to yield smaller errors than block boosting. The performance of the methods considered deteriorates when σ_y and/or σ_e are large. For example, when $\sigma_e = 2$, adding x_{it}^2 as potential predictors or using them to form principal components seems to increase forecast errors. In some cases, the simple AR(4) is better than many of the methods considered. In cases when boosting X is best, boosting the factors usually does not fare significantly worse. However, when boosting the factors is best, it tends to outperform boosting the observed variables by at least 10% of the RMSE.

Under DGP 2, y is predicted by F_2 and F_4, and their lags. Unlike DGP 1, some of the predictors are dynamically related. It is clear from Table II that boosting the factors gives smaller MSE and

DOI: 10.1002/jae

Table I(A). DGP 1 (BIC)

$$y_{t+h} = \beta_1 F_{1t} + \beta_2 F_{3t-4} + \beta_4 F_{6t}^2 + \sigma_y \varepsilon_{t+h}$$
$$x_{it} = \lambda_i' F_t + e_{it}$$

(σ_e, σ_y)	1	2	3	4	5	6	7	8	9	10	11	12	13	Best
	DI	$X(C)$	$X(B)$	$X^2(C)$	$X^2(B)$	$F(C)$	$F^2(C)$	$F(B)$	F^2(B)	$G(C)$	$G^2(C)$	$G(B)$	$G^2(B)$	
(1.0, 1.0)	0.033	0.076	−0.094	0.161	0.191	0.066	0.083	0.087	−0.006	0.139	0.018	0.146	0.104	9
	1.033	1.014	0.936	1.025	0.955	0.915	0.983	0.890	0.916	1.021	1.081	0.999	1.006	8
	0.918	0.884	0.753	0.902	0.784	0.719	0.830	0.680	0.721	0.897	1.005	0.858	0.871	8
(2.0, 1.0)	0.239	0.237	0.246	0.335	0.385	0.341	0.336	0.407	0.355	0.220	0.256	0.379	0.290	10
	1.027	1.042	1.047	1.068	1.062	1.050	1.128	1.016	1.028	1.087	1.118	1.042	1.067	8
	0.901	0.928	0.936	0.975	0.963	0.942	1.087	0.881	0.903	1.010	1.068	0.927	0.973	8
(0.5, 1.0)	0.035	−0.112	−0.103	−0.037	−0.231	0.035	0.008	0.021	0.021	0.045	0.031	0.008	0.012	12
	0.846	0.860	0.825	0.867	0.850	0.941	0.941	0.917	0.917	0.954	0.976	0.978	0.982	3
	0.631	0.652	0.600	0.663	0.637	0.782	0.781	0.741	0.741	0.802	0.840	0.844	0.850	3
(1.0, 2.0)	0.427	0.320	0.248	0.443	0.418	0.183	0.138	0.394	0.351	0.616	0.468	0.445	0.414	7
	1.004	1.047	0.984	1.053	1.021	0.977	0.973	1.006	1.007	1.007	1.015	0.995	1.001	7
	0.962	1.044	0.923	1.057	0.994	0.909	0.903	0.965	0.967	0.966	0.982	0.944	0.956	7
(1.0, 0.5)	0.204	−0.022	−0.014	0.075	0.106	0.091	0.070	0.037	0.069	0.028	0.048	0.144	0.181	3
	0.914	0.801	0.823	0.847	0.870	0.885	0.903	0.869	0.899	0.922	0.912	0.922	0.916	2
	1.013	0.779	0.822	0.869	0.917	0.950	0.987	0.916	0.979	1.031	1.007	1.031	1.018	2
(2.0, 2.0)	0.451	0.583	0.572	0.694	0.504	0.492	0.435	0.546	0.560	0.365	0.201	0.524	0.436	11
	1.028	0.998	1.012	1.053	1.034	1.019	1.003	1.017	1.013	1.021	1.006	1.037	1.039	2
	0.976	0.920	0.945	1.023	0.987	0.957	0.928	0.954	0.948	0.961	0.933	0.992	0.997	2
(2.0, 2.0)	0.451	0.583	0.572	0.694	0.504	0.492	0.435	0.546	0.560	0.365	0.201	0.524	0.436	11
	1.028	0.998	1.012	1.053	1.034	1.019	1.003	1.017	1.013	1.021	1.006	1.037	1.039	2
	0.976	0.920	0.945	1.023	0.987	0.957	0.928	0.954	0.948	0.961	0.933	0.992	0.997	2
(0.5, 0.5)	−0.040	−0.162	−0.044	−0.175	−0.026	−0.076	−0.107	−0.052	−0.132	−0.210	−0.140	−0.293	−0.270	5
	0.681	0.759	0.717	0.786	0.764	0.718	0.708	0.720	0.693	0.938	0.912	0.903	0.898	1
	0.460	0.570	0.509	0.612	0.578	0.511	0.496	0.513	0.475	0.872	0.823	0.807	0.798	1

For each parameterization, row 1 is median bias, row 2 is MSE/var(y), and row 3 is MSE relative to that of an AR(4). C, component-wise boosting; B, block-wise boosting; X, predictor set (i) of Section 4; X^2, predictor set (ii); F, predictor set (iii), etc.

DOI: 10.1002/jae

Table I(B). DGP 1 (adjusted BIC)

(σ_e, σ_y)	1	2	3	4	5	6	7	8	9	10	11	12	13	Best
	DI	$X(C)$	$X(B)$	$X^2(C)$	$X^2(B)$	$F(C)$	$F^2(C)$	$F(B)$	F^2(B)	$G(C)$	$G^2(C)$	$G(B)$	$G^2(B)$	
(1.0, 1.0)	0.033	0.076	−0.094	0.161	0.191	0.182	0.078	0.033	0.033	−0.050	0.053	0.044	0.034	8
	1.033	1.014	0.936	1.025	0.955	0.918	1.002	0.879	0.890	1.035	1.122	1.012	1.037	8
	0.918	0.884	0.753	0.902	0.784	0.724	0.863	0.664	0.681	0.921	1.081	0.880	0.924	8
(2.0 1.0)	0.239	0.237	0.246	0.335	0.385	0.121	0.309	0.382	0.332	0.123	0.183	0.391	0.304	6
	1.027	1.042	1.047	1.068	1.062	1.071	1.182	1.022	1.050	1.109	1.151	1.040	1.067	8
	0.901	0.928	0.936	0.975	0.963	0.979	1.193	0.892	0.942	1.050	1.131	0.924	0.973	8
(0.5, 1.0)	0.035	−0.112	−0.103	−0.037	−0.231	−0.012	−0.045	−0.029	−0.020	−0.044	−0.057	−0.128	−0.083	6
	0.846	0.860	0.825	0.867	0.850	0.922	0.939	0.885	0.878	0.954	1.004	0.973	0.971	3
	0.631	0.652	0.600	0.663	0.637	0.749	0.777	0.690	0.680	0.803	0.890	0.834	0.831	3
(1.0, 2.0)	0.427	0.320	0.248	0.443	0.418	0.082	0.018	0.337	0.305	0.529	0.195	0.319	0.285	7
	1.004	1.047	0.984	1.053	1.021	0.964	0.980	0.992	0.994	1.016	1.055	0.983	0.991	6
	0.962	1.044	0.923	1.057	0.994	0.886	0.915	0.937	0.942	0.983	1.061	0.921	0.936	6
(1.0, 0.5)	0.204	−0.022	−0.014	0.075	0.106	0.108	0.102	0.104	0.158	−0.130	0.052	0.059	0.093	3
	0.914	0.801	0.823	0.847	0.870	0.898	0.909	0.867	0.905	0.926	0.914	0.906	0.909	2
	1.013	0.779	0.822	0.869	0.917	0.978	1.001	0.911	0.992	1.040	1.012	0.994	1.002	2
(2.0, 2.0)	0.451	0.583	0.572	0.694	0.504	0.409	0.591	0.607	0.607	0.303	0.214	0.430	0.500	11
	1.028	0.998	1.012	1.053	1.034	1.028	1.027	1.019	1.011	1.028	1.012	1.050	1.053	2
	0.976	0.920	0.945	1.023	0.987	0.975	0.973	0.959	0.944	0.976	0.946	1.018	1.023	2
(0.5, 0.5)	−0.040	−0.162	−0.044	−0.175	−0.026	−0.104	−0.107	−0.014	−0.056	−0.042	−0.103	−0.219	−0.137	8
	0.681	0.759	0.717	0.786	0.764	0.704	0.702	0.690	0.663	0.955	0.910	0.913	0.922	9
	0.460	0.570	0.509	0.612	0.578	0.490	0.488	0.472	0.436	0.903	0.821	0.825	0.842	9

For each parameterization, row 1 is median bias, row 2 is MSE/var(y), and row 3 is MSE relative to that of an AR(4).

Table II(A). DGP 2 (BIC)

$$y_{t+h} = \beta_1 F_{2t} + \beta_2 F_{2t-1} + \beta_3 F_{2t-2} + \beta_4 F_{4t-1} + \beta_5 F_{4t-2} + \varepsilon_{t+h}$$

$$x_{it} = \lambda_i' F_t + e_{it}$$

(σ_e, σ_y)	1	2	3	4	5	6	7	8	9	10	11	12	13	Best
	DI	$X(C)$	$X(B)$	$X^2(C)$	$X^2(B)$	$F(C)$	$F^2(C)$	$F(B)$	$F^2(B)$	$G(C)$	$G^2(C)$	$G(B)$	$G^2(B)$	
(1.0, 1.0)	0.166	0.228	0.681	1.228	1.082	0.343	0.409	0.247	0.213	1.374	0.882	1.149	0.905	1
	0.798	0.858	0.872	1.010	1.016	0.733	0.762	0.703	0.740	1.064	1.154	0.971	0.997	8
	0.779	0.899	0.929	1.247	1.262	0.656	0.709	0.603	0.670	1.383	1.627	1.152	1.214	8
(2.0, 1.0)	0.779	0.377	0.438	0.268	0.982	0.141	0.359	0.601	0.696	0.469	0.878	0.543	0.421	6
	1.069	1.002	1.006	1.088	1.120	1.007	1.044	1.044	1.048	1.113	1.224	1.090	1.083	2
	0.966	0.849	0.856	1.001	1.059	0.857	0.922	0.922	0.929	1.048	1.266	1.005	0.991	2
(0.5, 1.0)	0.165	0.140	0.200	−1.051	−1.161	0.017	0.098	0.024	0.014	−0.010	−0.303	−0.561	−1.020	10
	0.595	0.547	0.520	1.024	1.009	0.541	0.567	0.549	0.555	0.978	1.110	0.950	0.960	3
	0.413	0.348	0.315	1.222	1.187	0.342	0.375	0.351	0.360	1.115	1.437	1.051	1.073	3
(1.0, 2.0)	−0.290	−0.162	−0.616	0.146	−0.172	−0.302	−0.343	−0.213	−0.213	−0.001	−0.382	−0.164	−0.418	10
	0.981	0.923	0.885	1.026	1.012	0.804	0.796	0.853	0.841	1.120	1.184	1.002	1.082	7
	1.058	0.935	0.860	1.156	1.125	0.710	0.695	0.799	0.776	1.377	1.538	1.102	1.284	7
(1.0, 0.5)	−0.172	−0.283	−0.129	−0.068	−0.246	−0.356	−0.113	−0.215	−0.014	−0.498	−0.233	−0.324	−0.258	9
	0.927	0.846	0.794	1.082	1.053	0.793	0.791	0.783	0.803	0.960	1.084	0.946	0.970	8
	0.962	0.801	0.707	1.311	1.242	0.705	0.700	0.686	0.721	1.032	1.315	1.003	1.054	8
(2.0, 2.0)	0.946	1.291	1.268	0.881	1.282	0.373	0.396	0.702	0.668	0.770	0.673	1.088	0.972	6
	1.033	0.929	0.952	1.134	1.149	0.794	0.813	0.837	0.857	1.094	1.061	1.078	1.091	6
	0.993	0.803	0.842	1.195	1.228	0.587	0.614	0.651	0.683	1.112	1.046	1.080	1.107	6
(0.5, 0.5)	0.059	−0.003	−0.024	−0.485	−0.531	−0.227	−0.056	−0.265	−0.232	−1.005	−0.825	−0.775	−0.642	2
	0.619	0.446	0.438	1.133	1.045	0.481	0.485	0.509	0.506	1.115	1.143	1.017	1.055	3
	0.429	0.223	0.214	1.437	1.221	0.259	0.263	0.290	0.286	1.390	1.460	1.157	1.246	3

For each parameterization, row 1 is median bias, row 2 is MSE/var(y), and row 3 is MSE relative to that of an AR(4).

Table II(B). DGP 2 (adjusted BIC)

(σ_e, σ_y)	1	2	3	4	5	6	7	8	9	10	11	12	13	Best
	DI	$X(C)$	$X(B)$	$X^2(C)$	$X^2(B)$	$F(C)$	$F^2(C)$	$F(B)$	F^2(B)	$G(C)$	$G^2(C)$	$G(B)$	$G^2(B)$	
(1.0, 1.0)	0.166	0.228	0.681	1.228	1.082	0.267	0.379	0.390	0.312	1.285	0.881	1.102	0.940	1
	0.798	0.858	0.872	1.010	1.016	0.746	0.769	0.691	0.718	1.146	1.167	1.006	1.039	8
	0.779	0.899	0.929	1.247	1.262	0.680	0.723	0.584	0.630	1.604	1.666	1.237	1.319	8
(2.0, 1.0)	0.779	0.377	0.438	0.268	0.982	0.105	0.160	0.510	0.677	0.060	0.949	0.479	0.199	10
	1.069	1.002	1.006	1.088	1.120	1.016	1.053	1.032	1.052	1.129	1.261	1.076	1.088	2
	0.966	0.849	0.856	1.001	1.059	0.872	0.937	0.900	0.935	1.076	1.343	0.978	1.001	2
(0.5, 1.0)	0.165	0.140	0.200	−1.051	−1.161	0.010	0.141	0.175	0.163	−0.235	−0.489	−0.764	−0.941	6
	0.595	0.547	0.520	1.024	1.009	0.529	0.566	0.548	0.557	0.997	1.135	0.957	1.013	3
	0.413	0.348	0.315	1.222	1.187	0.326	0.373	0.350	0.362	1.157	1.501	1.068	1.196	3
(1.0, 2.0)	−0.290	−0.162	−0.616	−0.146	−0.172	−0.365	−0.290	−0.254	−0.257	−0.104	−0.415	−0.339	−0.351	10
	0.981	0.923	0.885	1.026	1.012	0.792	0.790	0.834	0.832	1.154	1.216	0.986	1.089	7
	1.058	0.935	0.860	1.156	1.125	0.689	0.686	0.763	0.759	1.462	1.624	1.068	1.302	7
(1.0, 0.5)	−0.172	−0.283	−0.129	−0.068	−0.246	−0.275	−0.088	−0.204	−0.235	−0.494	−0.237	−0.382	−0.221	4
	0.927	0.846	0.794	1.082	1.053	0.788	0.791	0.772	0.795	0.973	1.097	0.955	0.997	8
	0.962	0.801	0.707	1.311	1.242	0.696	0.700	0.667	0.707	1.059	1.347	1.021	1.112	8
(2.0, 2.0)	0.946	1.291	1.268	0.881	1.282	0.489	0.311	0.449	0.466	1.046	0.589	0.965	0.537	7
	1.033	0.929	0.952	1.134	1.149	0.775	0.807	0.828	0.854	1.131	1.071	1.108	1.106	6
	0.993	0.803	0.842	1.195	1.228	0.558	0.605	0.637	0.678	1.190	1.067	1.141	1.138	6
(0.5, 0.5)	0.059	−0.003	−0.024	−0.485	−0.531	−0.227	−0.041	−0.122	−0.239	−0.857	−0.684	−0.777	−0.705	2
	0.619	0.446	0.438	1.133	1.045	0.481	0.486	0.552	0.558	1.166	1.176	1.035	1.079	3
	0.429	0.223	0.214	1.437	1.221	0.259	0.265	0.341	0.349	1.520	1.547	1.198	1.301	3

For each parameterization, row 1 is median bias, row 2 is MSE/var(y), and row 3 is MSE relative to that of an AR(4).

often smaller bias than boosting the observed predictors. Forming the factors from the expanded dataset (which includes the lags and quadratic terms in x) tends to yield forecasts strongly inferior to even an AR(4).

Overall, the results show that boosting the observed variables is inferior to boosting the factors when the data have a strong factor structure. In results not reported, we also considered a case in which the dependence of y on X is concentrated on a small number of variables. It is not surprising that in such cases boosting the observed variables can perform quite well. As for component versus block boosting, this too is DGP specific. What is promising, however, is that the boosting methods considered have the potential to reduce forecast errors. Practitioners need to carefully evaluate which method is best for their application on hand as the difference between the best method and worst method can be non-trivial.

4.2. Applications

In this subsection we applied the same procedures to five series: inflation; the change in Federal Funds rate; growth rate of industrial production; growth rate of employment; and the unemployment rate. In each case we are interested in forecasting the series 12 months ahead. The predictors consist

Table III. MSE for forecasting monthly inflation: $h = 12$

Start	End	Adjusted BIC													
		1	2	3	4	5	6	7	8	9	10	11	12	13	Best
		DI	$X(C)$	$X(B)$	$X^2(C)$	$X^2(B)$	$F(C)$	$F^2(C)$	$F(B)$	F^2(B)	$G(C)$	G^2(C)	$G(B)$	$G^2(B)$	
73.1	80.1	0.607	0.733	0.570	0.981	0.666	0.737	0.777	0.609	0.700	0.800	0.906	0.614	0.815	3
73.1	90.1	0.678	0.730	0.566	0.854	0.636	0.697	0.721	0.597	0.626	0.761	0.819	0.568	0.650	3
73.1	00.1	0.695	0.765	0.577	0.886	0.648	0.722	0.744	0.605	0.630	0.773	0.829	0.573	0.647	12
84.1	90.1	0.816	0.871	0.703	0.941	0.758	0.822	0.841	0.707	0.706	0.858	0.881	0.645	0.649	12
84.1	00.1	0.949	1.097	0.863	1.211	0.948	1.013	1.039	0.846	0.841	1.009	1.023	0.767	0.768	12
84.1	02.1	1.011	1.259	1.014	1.401	1.112	1.103	1.129	0.935	0.933	1.084	1.090	0.863	0.865	12
90.1	00.1	0.942	1.234	0.964	1.471	1.123	1.038	1.132	0.863	0.854	1.113	1.596	0.843	1.089	12
73.1	02.1	0.713	0.820	0.630	0.976	0.725	0.774	0.816	0.640	0.666	0.844	1.040	0.617	0.760	12

Table IV. MSE for forecasting Federal Funds rate: $h = 12$

Start	End	Adjusted BIC													
		1	2	3	4	5	6	7	8	9	10	11	12	13	Best
		DI	$X(C)$	$X(B)$	$X^2(C)$	$X^2(B)$	$F(C)$	$F^2(C)$	$F(B)$	F^2(B)	$G(C)$	G^2(C)	$G(B)$	$G^2(B)$	
73.0	80.0	0.684	0.828	0.768	0.845	0.873	0.701	0.732	0.730	0.758	0.798	0.781	0.726	0.693	1
73.0	90.0	0.658	0.657	0.627	0.819	0.693	0.656	0.678	0.670	0.699	0.616	0.620	0.606	0.599	13
73.0	100.0	0.623	0.627	0.618	0.787	0.736	0.612	0.618	0.633	0.648	0.582	0.576	0.583	0.570	13
84.0	90.0	0.621	0.652	0.660	0.836	0.757	0.630	0.641	0.634	0.649	0.622	0.623	0.614	0.616	12
84.0	100.0	0.504	0.464	0.485	0.534	0.642	0.425	0.413	0.514	0.485	0.429	0.403	0.421	0.412	11
84.0	102.0	0.602	0.553	0.536	0.551	0.612	0.540	0.528	0.604	0.568	0.465	0.443	0.470	0.465	11
90.0	100.0	0.646	0.483	0.500	0.483	0.552	0.535	0.520	0.618	0.574	0.382	0.366	0.429	0.421	11
73.0	102.0	0.701	0.668	0.670	0.788	0.752	0.692	0.693	0.705	0.707	0.613	0.614	0.625	0.621	10

See footnote to Table I(a).

of a panel of 132 series.[4] These are monthly time series available from 1960 : 1 to 2003 : 12 for a total of $T = 528$ observations. From this large panel of data, eight factors are estimated and denoted $\tilde{F}_t$.

The current DI methodology has two limitations: $\tilde{f}_t$ is ordered according to $\tilde{F}_t$; and the dynamic structure is rather restrictive. As discussed earlier, both problems arise because there is no easy way to select a subset of predictors when the predictors have no natural ordering. Here, the choice of $\tilde{f}_t$ is not limited to a subset or functions of $\tilde{F}_t$. As in the simulation subsection, we consider six predictor sets (i)–(vi) described earlier, except that X_t is a 132×1 vector of macroeconomic series, instead of simulated series. The results for inflation are reported in Table III. Once again, C denotes component-wise boosting, and B block-boosting.

For brevity, we only report the average RMSE relative to an AR(4) forecast, using the adjusted BIC to select the boosting stopping rule. An entry below one means it beats the AR(4) forecast. We evaluate the methods over eight subsamples. The first column of Table III is the performance of the DI forecast. Notably, except for the subsamples that start in 1984, the DI produces smaller errors than an autoregressive forecast. We note in passing that there are generally some reductions in RMSE when the estimated predictors are more heavily penalized. When many estimated factors are being considered as predictors, the additional penalty properly takes into account sampling variability to avoid choosing too many estimated predictors. The question is whether boosting can do better. The answer to this is, yes, by block boosting. Block boosting clearly outperforms component-wise boosting and always produces smaller RMSE than the DI. The results show that there is no gain by considering squared values of estimated factors (columns 7, 9, 11, 13). When directly boosting observable variables, including quadratic variables adds no benefit either (columns 4 and 5). However, boosting the principal components formed from adding quadratic variables and lags performs quite well, as shown by the result in column 12. This suggests that some predictable variation in inflation is contained in the higher-order terms.

Columns 2–5 present results from boosting the 132 predictors, plus their squared terms. Again, block boosting is better than component-wise boosting. Boosting the observable variables performs well except for the subperiods of 1984.1–2002.1 and 1990.1–2000.1. The performance can be

Table V. MSE for forecasting industrial production: $h = 12$

Start	End	Adjusted BIC													
		1	2	3	4	5	6	7	8	9	10	11	12	13	Best
		DI	$X(C)$	$X(B)$	$X^2(C)$	$X^2(B)$	$F(C)$	$F^2(C)$	$F(B)$	F^2(B)	$G(C)$	G^2(C)	$G(B)$	$G^2(B)$	
73.0	80.0	0.911	0.665	0.651	0.668	0.670	0.763	0.676	0.771	0.722	0.604	0.693	0.545	0.664	12
73.0	90.0	0.843	0.679	0.684	0.665	0.625	0.721	0.701	0.749	0.727	0.557	0.587	0.584	0.626	10
73.0	100.0	0.983	0.426	0.431	0.406	0.359	0.891	0.883	0.905	0.887	0.567	0.568	0.643	0.636	5
84.0	90.0	0.978	0.402	0.409	0.373	0.320	0.896	0.885	0.911	0.889	0.560	0.543	0.653	0.625	5
84.0	100.0	0.999	0.269	0.283	0.256	0.224	0.910	0.861	0.934	0.877	0.556	0.545	0.654	0.634	5
84.0	102.0	0.970	0.265	0.267	0.208	0.191	0.883	0.823	0.913	0.848	0.489	0.479	0.590	0.571	5
90.0	100.0	0.966	0.251	0.258	0.206	0.196	0.850	0.808	0.887	0.831	0.482	0.478	0.580	0.567	5
73.0	102.0	0.953	0.394	0.422	0.358	0.323	0.885	0.875	0.894	0.878	0.511	0.514	0.589	0.589	5

See footnote to Table I(a).

[4] The data are taken from Mark Watson's web site: http://www.princeton.edu/ mwatson. The four series are PUNEW, FYFF, IP and CES002.

DOI: 10.1002/jae

period specific. As noted earlier, adding squared terms does not improve the performance. In comparison, the results in column 12 are more stable. Stock and Watson (2002) suggest that the DI approach may be less susceptible to parameter instability than the classical approach to prediction. Our results support this conjecture.

Results for the Federal Funds rate, industrial production, and employment are presented in Tables IV, V and VI. For all but one subsample and only for the Federal Funds rate, one of the methods considered yields a lower RMSE than the DI forecast. For all three series, some form of quadratic term appears to help forecast the series of interest. While component-wise boosting of the predictors works systematically well for industrial production and employment, boosting factors formed from functions of the observed data works better for the Federal Funds rate. Table VII presents results for the unemployment rate. For this series, the standard diffusion index forecast tends to outperform the boosting alternatives. However, unemployment rate is possibly non-stationary. Accordingly, Table VIII reports results for forecasting the change in unemployment rate. Now boosting again outperforms the DI. These results make clear that a method that forecasts a series well may not forecast another series just as well. The need to search for the best methodology as the environment changes remains the reality of forecasting economic time series.

5. CONCLUSION

Boosting is a tool for analyzing high-dimension data in the machine-learning literature and in biostatistics. This paper considers the usefulness of boosting in economic analysis. In particular, boosting is used to select estimated factors to be augmented to a standard forecasting equation. It has the advantage that it does not require a priori ordering of the predictors or their lags as conventional model selection procedures do. We also discuss how to account for the fact that our predictors are estimated by the method of principal components in forecasting applications.

As the zero restrictions are imposed by boosting on the parameters in a stochastic manner, one might want to take into account the sampling variability due to model selection on the parameter estimates. One possibility is to cast boosting in terms of linear estimation subject to stochastic constraints. In this regard, boosting can be thought of as a Theil–Goldberger mixed estimator

Table VI. MSE for forecasting employment: $h = 12$

Start	End	Adjusted BIC													
		1	2	3	4	5	6	7	8	9	10	11	12	13	Best
		DI	$X(C)$	$X(B)$	$X^2(C)$	$X^2(B)$	$F(C)$	$F^2(C)$	$F(B)$	F^2(B)	$G(C)$	G^2(C)	$G(B)$	$G^2(B)$	
73.0	80.0	0.658	0.581	0.577	0.634	0.626	0.727	0.649	0.740	0.690	0.531	0.589	0.547	0.597	10
73.0	90.0	0.788	0.606	0.614	0.601	0.556	0.721	0.716	0.747	0.739	0.523	0.538	0.590	0.604	10
73.0	100.0	0.883	0.431	0.432	0.399	0.359	0.894	0.896	0.905	0.883	0.525	0.520	0.623	0.608	5
84.0	90.0	0.908	0.427	0.427	0.382	0.337	0.901	0.896	0.915	0.886	0.533	0.513	0.637	0.611	5
84.0	100.0	0.927	0.279	0.284	0.240	0.212	0.916	0.862	0.939	0.864	0.513	0.495	0.626	0.600	5
84.0	102.0	0.942	0.247	0.241	0.200	0.197	0.895	0.818	0.929	0.840	0.449	0.432	0.564	0.537	5
90.0	100.0	0.926	0.207	0.206	0.170	0.176	0.865	0.805	0.902	0.821	0.425	0.414	0.540	0.518	4
73.0	102.0	0.913	0.400	0.423	0.362	0.332	0.885	0.881	0.896	0.867	0.481	0.477	0.578	0.567	5

See footnote to Table I(a).

DOI: 10.1002/jae

Table VII. MSE for forecasting unemployment rate: $h = 12$

Start	End	Adjusted BIC													
		1	2	3	4	5	6	7	8	9	10	11	12	13	Best
		DI	$X(C)$	$X(B)$	$X^2(C)$	$X^2(B)$	$F(C)$	$F^2(C)$	$F(B)$	F^2(B)	$G(C)$	G^2(C)	$G(B)$	$G^2(B)$	
73.0	80.0	0.331	0.523	0.478	0.609	0.542	0.404	0.553	0.439	0.515	0.337	0.486	0.476	0.510	1
73.0	90.0	0.453	0.662	0.605	0.738	0.705	0.484	0.569	0.522	0.567	0.476	0.568	0.578	0.601	1
73.0	100.0	0.550	0.878	0.791	0.962	0.877	0.671	0.730	0.674	0.703	0.637	0.711	0.696	0.722	1
84.0	90.0	0.656	1.083	0.994	1.181	1.071	0.763	0.778	0.738	0.751	0.776	0.801	0.773	0.799	1
84.0	100.0	0.785	1.325	1.193	1.516	1.333	0.958	0.995	0.894	0.915	0.917	0.951	0.894	0.938	1
84.0	102.0	0.834	1.137	1.046	1.270	1.143	0.919	0.933	0.865	0.878	0.885	0.903	0.853	0.886	1
90.0	100.0	0.938	1.153	1.154	1.309	1.297	1.000	1.005	0.914	0.916	0.956	0.975	0.913	0.930	12
73.0	102.0	0.702	0.936	0.891	1.040	1.003	0.790	0.828	0.745	0.769	0.757	0.790	0.772	0.772	1

See footnote to Table I(a).

Table VIII. MSE for forecasting change in unemployment rate: $h = 12$

Start	End	Adjusted BIC													
		1	2	3	4	5	6	7	8	9	10	11	12	13	Best
		DI	$X(C)$	$X(B)$	$X^2(C)$	$X^2(B)$	$F(C)$	$F^2(C)$	$F(B)$	F^2(B)	$G(C)$	G^2(C)	$G(B)$	$G^2(B)$	
73.0	80.0	0.835	0.635	0.634	0.648	0.670	0.739	0.663	0.771	0.701	0.541	0.628	0.541	0.625	10
73.0	90.0	0.774	0.650	0.663	0.640	0.603	0.737	0.728	0.766	0.741	0.549	0.573	0.600	0.628	10
73.0	100.0	0.600	0.431	0.433	0.401	0.355	0.915	0.919	0.932	0.922	0.608	0.603	0.700	0.685	5
84.0	90.0	0.599	0.426	0.428	0.387	0.338	0.920	0.917	0.936	0.923	0.614	0.595	0.711	0.685	5
84.0	100.0	0.607	0.322	0.331	0.297	0.274	0.916	0.886	0.943	0.907	0.605	0.597	0.697	0.682	5
84.0	102.0	0.695	0.345	0.341	0.253	0.240	0.893	0.844	0.923	0.875	0.542	0.536	0.635	0.622	5
90.0	100.0	0.683	0.336	0.338	0.258	0.253	0.853	0.819	0.891	0.850	0.539	0.537	0.624	0.619	5
73.0	102.0	0.723	0.397	0.423	0.315	0.288	0.889	0.893	0.897	0.889	0.506	0.506	0.603	0.598	5

See footnote to Table I(a).

for which the variance of the estimator takes into account randomness of the constraints. The extension is not straightforward because of the non-linearity and non-differentiability of $\widehat{\Phi}_M(z)$ in y. We leave this for future research.

APPENDIX

Lemma A1 *Let $\tilde{P} = \tilde{F}(\tilde{F}'\tilde{F})^{-1}\tilde{F}'$, where $\tilde{F}$ is a $T \times r$ matrix of factors estimated from a $T \times N$ panel of data by the method of principal components. Also let $P = F(F'F)^{-1}F'$, where F is the factor matrix. Let $\delta_{NT} = \min[\sqrt{N}, \sqrt{T}]$:*

(i) $\|\tilde{P} - P\| = O_p(\delta_{NT}^{-1})$;
(ii) $\|(\tilde{F}'\tilde{F})^{-1}\tilde{F}' - (F'F)^{-1}F'\| = \frac{1}{\sqrt{T}}O_p(\delta_{NT}^{-1})$.

DOI: 10.1002/jae

This follows from the argument in the proof of Lemma 2 of Bai and Ng (2002), and Lemma 1(i) of this paper. Part (ii) assumes H is an identity matrix, which holds under the assumption that $F'F/T = I_r$ and $\Lambda'\Lambda/N$ is a diagonal matrix with distinct elements. With $H = I_r$, Lemma A1 holds with F_j and F_j (the jth column, $j = 1, 2, \ldots, r$). The details are omitted.

Proof of Proposition 1 The fitted value $\tilde\Phi_M$ equals $\tilde B_M Y$ and $\widehat\Phi_M = B_M Y$:

$$\frac{1}{T}\sum_{t=1}^{T} |\tilde\Phi_M(z_t) - \widehat\Phi_M(z_t)|^2 = \frac{1}{T}\left\|\tilde\Phi_M - \tilde\Phi_M\right\|^2 \le \left\|\tilde B_M - B_M\right\|^2(\left\|Y\right\|^2/T)$$

Note $\left\|Y\right\|^2/T = \frac{1}{T}\sum_{t=1}^T \left\|y_t\right\|^2 = O_p(1)$, and

$$\tilde B_M - B_M = \prod_{m=1}^{M} \tilde a_m - \prod_{m=1}^{M} a_m = (\tilde a_1 - a_1)A_1 + B_1(\tilde a_2 - a_2)A_2 + \ldots B_{M-1}(\tilde a_M - a_M)A_M$$

where $\tilde a_m = I_T - \tilde P^{(m)}, a_m = I_T - P^{(m)}$, and $A_m = \prod_{j=m+1}^M \tilde a_j$. But a_j and $\tilde a_j$ are projection matrices whose largest eigenvalue is one, and thus $\left\|a_j\right\| \le 1$ and $\left\|\tilde a_j\right\| \le 1$. It follows that $\left\|A_m\right\| \le 1$ and $\left\|B_m\right\| \le 1$ for all m. Furthermore, $(\tilde a_m - a_m) = P^{(m)} - \tilde P^{(m)}$ and $\left\|P^{(m)} - \tilde p^{(m)}\right\| = O_p(\delta_{NT}^{-1})$ by Lemma A1(i). It follows that

$$\|B_{j-1}(\tilde a_j - a_j)A_j\| \le \| B_{j-1}\|\|\tilde a_j - a_j\|\|A_j\| \le \|\tilde a_j - a_j\| = o_p(\delta_{NT}^{-1})$$

and $\left\|\tilde B_M - B_M\right\| = O_p(M/\delta_{NT})$. Thus if $M/\delta_{NT} \to 0$, then $\frac{1}{T}\sum_{t=1}^T |\tilde\Phi_M(z_t) - \widehat\Phi_M(z_t)|^2 \xrightarrow{p} 0$. Next, $\tilde\Phi_M(z) = \overline{y} + z'\tilde\beta_M$ and $\widehat\Phi_M(z) = \overline{y} + z'\widehat\beta_M$. Note that (for $M \ge 1$)

$$\widehat\beta_M = \nu(\widehat b_1^\dagger + \cdots + \widehat b_M^\dagger), \tilde\beta_M = \nu(\tilde b_1^\dagger + \cdots + \tilde b_M^\dagger)$$

where $\tilde b_m^\dagger = (\tilde z'_{i_m^*}\tilde z_{i_m^*})^{-1}\tilde z'_{i_m^*}Y$, and $\widehat b_m^\dagger$ is defined similarly with $z_{i_m^*}$ replacing $\tilde z_{i_m^*}, m = 1, 2, \ldots, M$. Thus

$$|\tilde\Phi(z) - \widehat\Phi(z)| \le |\nu|\|z\|[\|\tilde b_1^\dagger - \widehat b_1^\dagger\| + \cdots + \|\tilde b_M^\dagger - \widehat b_M^\dagger\|]$$

By Lemma A1(ii):

$$\|\tilde b_m^\dagger - \widehat b_m^\dagger\| \le \|\tilde z'_{i_m^*}\tilde z_{i_m^*})^{-1} z'_{i_m^*} - (z'_{i_m^*} z_{i_m^*})^{-1} z'_{i_m^*}\|\|Y\| \le O_p(\delta_{NT}^{-1})T^{-1/2}\|Y\| = O_p(\delta_{NT}^{-1})$$

It follows that

$$|\tilde\Phi(z) - \widehat\Phi(z)| \le |\nu|\|z\|M O_p(\delta_{NT}^{-1})$$

and the above converges to zero if $M/\delta_{NT} \to 0$.

ACKNOWLEDGEMENTS

We would like to thank the editor and three anonymous referees for many helpful comments. We also acknowledge financial support from the NSF through grants SES-0 551 275 and SES-0 549 978.

DOI: 10.1002/jae

REFERENCES

Bai J. 2003. Inferential theory for factor models of large dimensions. *Econometrica* **71**: 135–172.

Bai J, Ng S. 2002. Determining the number of factors in approximate factor models. *Econometrica* **70**: 191–221.

Bai J, Ng S. 2006. Confidence intervals for diffusion index forecasts and inference with factor-augmented regressions. *Econometrica* **74**: 1133–1150.

Bai J, Ng S. 2008. Forecasting economic time series using targeted predictors. *Journal of Econometrics* **146**: 304–317.

Brockwell PJ, Davies RA. 1991. *Time Series Theory and Methods* (2nd edn). Springer: New York.

Buhlmann P. 2006. Boosting for high-dimensional linear models. *Annals of Statistics* **54**: 559–583.

Buhlmann P, Hothorn T. 2007. Boosting algorithms: regularization, prediction and model fitting. *Statistical Science* **22**: 477–505.

Buhlmann P, Yu B. 2003. Boosting with the L_2 loss: regression and classification. *Journal of the American Statistical Assocation* **98**: 324–339.

DeMol C, Giannone D, Reichlin L. 2006. Forecasting using a large number of predictors: is Bayesian regression a valid alternative to principal components? *ECB Working Paper 700*.

Forni M, Hallin M, Lippi M, Reichlin L. 2005. The generalized dynamic factor model, one sided estimation and forecasting. *Journal of the American Statistical Association* **100**: 830–840.

Freund Y. 1995. Boosting a weak learning algorithm by majority. *Information and Computation* **121**: 256–285.

Friedman J. 2001. Greedy function approxmiation: a gradient boosting machine. *Annals of Statistics* **29**: 1189–1232.

Friedman J, Hastie T, Tibshirani R. 2000. Additive logistic regression: a statistical view. *Annals of Statistics* **28**: 337–374.

Knight K, Fu W. 2000. Asymptotics for lasso type estimators. *Annals of Statistics* **28**: 1356–1378.

Lutz RW, Buhlmann P. 2006. Boosting for high-multivariate responses in high-dimensional linear regression. *Statistica Sinica* **16**: 471–494.

Pagan A. 1984. Econometric issues in the analysis of regressions with generated regressors. *International Economic Review* **25**: 221–247.

Rosset S. 2005. Robust boosting and its relation to bagging. Mimeo, IBM.

Schapire RE. 1990. The strength of weak learnability. *Machine Learning* **5**: 197–227.

Stock JH, Watson MW. 2002. Forecasting using principal components from a large number of predictors. *Journal of the American Statistical Association* **97**: 1167–1179.

Tutz G, Binder H. 2005. Boosting ridge regression. Mimeo, Universitat Munchen.

Zhang T, Yu B. 2005. Boosting with early stopping: convergence and consistency. *Annals of Statistics* **33**: 1538–1579.

DOI: 10.1002/jae