Teresa Buchen und Klaus Wohlrabe:

Forecasting with many predictors - Is boosting a viable alternative?

Munich Discussion Paper No. 2010-31

Department of Economics
University of Munich

Volkswirtschaftliche Fakultät
Ludwig-Maximilians-Universität München

Forecasting with many predictors - Is boosting a viable alternative?*

Teresa Buchen†
Klaus Wohlrabe‡
Ifo Institute for Economic Research, Poschingerstr. 5, 81679 Munich, Germany

September 6, 2010

Abstract

This paper evaluates the forecast performance of boosting, a variable selection device, and compares it with the forecast combination schemes and dynamic factor models presented in Stock and Watson (2006). Using the same data set and comparison methodology, we find that boosting is a serious competitor for forecasting US industrial production growth in the short run and that it performs best in the longer run.

JEL classification: C53, E27.
Keywords: Forecasting, Boosting, Cross-validation.

*We thank Mark W. Watson for making available his data set and thus enabling a direct comparison of the results. We also thank Kai Carstensen and Nikolay Robinzonov for helpful comments.

†Tel.: +49(0)89/9224-1222; fax:+49(0)89/9224-1463. *E-mail address:* buchen@ifo.de

‡Corresponding author. Tel.: +49(0)89/9224-1229; fax:+49(0)89/9224-1463. *E-mail address:* wohlrabe@ifo.de

1 Introduction

In recent years, a large body of research has developed that utilises many predictors for forecasting since both the availability of data and the computational power to handle them have increased tremendously. The crucial question that arises is which pieces of information are relevant for forecasting macroeconomic aggregates. There are two ways of exploiting a large number of time series without overfitting the forecasting model, information codensation and variable selection. The most common approaches, factor models and forecast combination schemes, perform information condensation (for a recent overview of factor models, see Stock and Watson (2010) and of forecast combination, see Timmermann (2006)).

This paper compares the forecast accuracy of boosting, a variable selection algorithm, with both forecast combination methods and factor models. An alternative approach would be to evaluate all possible combinations of variables according to some in-sample or out-of-sample criterion. However, since the number of combinations rises exponentially with the number of predictors, this method becomes infeasible when the number of variables is large. In this case, boosting provides an efficient solution to the variable selection problem. We present componentwise boosting which iteratively estimates an unknown function and in each iteration adds the variable with the largest contribution to the fit.

Boosting has been proposed in the machine learning community as a scheme for classification (Freund and Schapire, 1996) and further developed for regression problems by Friedman (2001). Until now, there are only very few applications to forecasting. Bai and Ng (2009) estimate the common factors of a large number of predictors and then select the most relevant factors by boosting. For the only considered forecast horizon of 12 months, they find that some form of boosting can improve the forecast as compared to standard factor models. However, in their empirical application to US data, boosting of factors is only advantageous for two out of five target variables, while for the others it is better to boost the predictors directly. Carriero et al. (2010) forecast a large number of variables using vector autoregressive

(VAR) models and compare the forecasting accuracy of several reduced-rank models with factor models, Bayesian VARs and multivariate boosting. The latter performs best when forecasting CPI inflation one month ahead.

The forecasting performance of boosting depends crucially on the number of iterations. While a small number of iterations leads to a large bias, a large number increases the variance since more and more predictors are added. In order to determine the stopping criterion, Bai and Ng (2009) use an Akaike criterion and Carriero et al. (2010) apply grid search. In this paper, we show that cross-validation leads to better results than the AIC. Moreover, it is computationally more efficient than grid search. Furthermore, we do not only compare the forecast accuracy of boosting with factor models, but also with forecast combination as another commonly used approach to incorporate many predictors.

As a basis of comparison, we build on Stock and Watson (2006) who examine the performance of different forecast combination schemes, factor models, Bayesian model averaging and empirical Bayes methods over several forecast horizons. Thereby, they are one of the few that compare the forecast accuracy of pooling of information versus pooling of forecasts.

We go one step further and include boosting into the horse race with these most prominent approaches to deal with large data sets. In our empirical application to US industrial production we use the same methods for forecast comparison and the same data set consisting of 131 economic time series from 1959 to 2003.

The remainder is organised as follows: While Section 2 outlines the boosting procedure, Section 3 describes the empirical application. Finally, Section 4 concludes.

2 Boosting

Boosting is a forward stagewise modelling algorithm that iteratively estimates an unknown function, which can be linear or nonlinear. We estimate the following autoregressive distributed lag (ADL) model:

$$\mathrm{E}\left(y_{t+h}^{h}|\boldsymbol{z}_t,\boldsymbol{\delta}\right)=\boldsymbol{\delta'Z}=\mu+\sum_{i=1}^{p}\alpha_i y_{t+1-i}+\sum_{j=1}^{N}\sum_{i=1}^{p}\beta_i^{(j)}x_{t+1-i}^{(j)}=:F(\boldsymbol{z}_t,\boldsymbol{\delta}),\quad(1)$$

where $\boldsymbol{x}=(x^1,...,x^N)$ contains all exogenous predictors, and p and N denote the number of lags and variables, respectively. Those variables, however, that are not chosen, obtain a zero restriction. The main ingredients of the boosting algorithm are the base learner and the loss function. While the base learner $f(.)$ is a simple fitting procedure, such as OLS, the loss function $L(.)$ is needed for the variable selection. The loss function that is most often used for regression problems is squared error (L_2) loss:

$$L(y_t,F(\boldsymbol{z}_t,\boldsymbol{\delta}))=\frac{1}{2}(y_t-F(\boldsymbol{z}_t,\boldsymbol{\delta}))^2.\quad(2)$$

For multi-dimensional datasets, Bühlmann and Yu (2003) suggest to use componentwise boosting where the base learner is applied to one variable at a time. Note that with componentwise boosting, the lags of one predictor are treated as separate variables such that the algorithm simultaneously selects variables and lags. So from all $p+N\times p$ potential predictors $z_{t,k}$, the variable z_{t,k_m^*} minimising the loss function is selected in each iteration m.

The algorithm for componentwise L_2 boosting can be summarised as follows:

1. Initialise $\hat{f}_{t,0}(.)=\bar{y}$ for each t. Set $m=0$.
2. Increase m by 1. For $t=1,...,T$, compute the negative gradient $-\frac{\partial L(y_t,F)}{\partial F}$ and evaluate at $\hat{f}_{t,m-1}(\boldsymbol{z}_t,\hat{\boldsymbol{\delta}}^{[m-1]})$: $u_t=y_t-\hat{f}_{t,m-1}(\boldsymbol{z}_t,\hat{\boldsymbol{\delta}}^{[m-1]})$.
3. For $k=1,...,p+N\times p$, regress the negative gradient vector u_t on $z_{t,k}$ and compute $SSR_k=\sum_{t=1}^{T}(u_t-z_{t,k}\hat{\theta}_k)^2$.
4. Choose z_{t,k_m^*} such that $SSR_{k^*}=\arg\min_{k\in N}SSR_k$.
5. Let $\hat{f}_{t,m}=z_{t,k_m^*}\hat{\theta}_{k_m^*}$.
6. For $t=1,...,T$, update $\hat{f}_{t,m}(.)=\hat{f}_{t,m-1}(.)+\nu\hat{f}_{t,m}(.)$, where $0<\nu<1$.

7. Iterate steps 2 to 6 until $m = M$

The final function estimate results as the sum of the M base learner estimates multiplied by the shrinkage parameter ν:

$$\hat{F}(\boldsymbol{z}_t, \hat{\boldsymbol{\delta}}^{[M]}) = \sum_{m=0}^{M} \nu \hat{f}_m(\boldsymbol{z}_t, \hat{\boldsymbol{\theta}}^{[m]}). \tag{3}$$

In order to reduce the variance, Friedman (2001) proposed to combine variable selection with shrinkage and introduced the step size ν into the boosting algorithm. Overfitting is also prevented by stopping the procedure at iteration M. The stopping criterion can be obtained by cross-validation or a modification of the Akaike criterion. For further details on boosting, see Bühlmann and Hothorn (2007).

3 Application to US Data

3.1 Data

The data set is the same used in Stock and Watson (2006). Covering the period from 1959 to 2003 it contains US industrial production as target series and 130 monthly time series from three broad categories: real economy, money and prices, and financial markets. The series were transformed to stationarity and standardised according to Stock and Watson (2004).

3.2 Methods

Following Stock and Watson (2006), we forecast the h-month growth of industrial production at an annual rate, where $h = 1, 3, 6$ and 12. The forecasts are computed directly and pseudo-out-of-sample using a recursive scheme with a forecast period from 1974:7 to 2003:12-h. When evaluating the forecast accuracy, we use the relative mean squared forecast errors (MSFEs), where the benchmark is an AR(AIC) model:

$$\mathrm{E}\left(y_{t+h}^{h}|\boldsymbol{y}_t\right) = \alpha + \sum_{i=1}^{p} \beta_i y_{t+1-i}, \text{ where} \tag{4}$$
$$y_{t+h}^{h} = (1200/h)\ln(IP_{t+h}/IP_t).$$

For the boosting procedure, we estimate the ADL model in Equation (1) using a linear weak learner (OLS) and an L_2-loss function. Since the boosting algorithm is relatively insensitive to the value of the shrinkage parameter ν – as long as it is sufficiently small – we set it to the commonly used value of 0.1 (Lutz and Bühlmann, 2006). The crucial parameter is the stopping criterion M, which we determine both by the AIC and bootstrapped cross-validation.

3.3 Results

The results are summarised in Table 1. As the entries are MSFEs relative to the AR benchmark, numbers less than 1 indicate an MSFE improvement over the benchmark forecast. It can be seen that the relative forecast performance of boosting improves with increasing forecast horizon. Moreover, the forecast errors are always smaller when cross-validation is used to determine the stopping criterion instead of the Akaike criterion. This is due to the fact that cross-validation tends to result in a smaller number of iterations and thus generates smaller models. Apart from the one-month forecast based on the AIC, boosting is always able to beat the benchmark. Furthermore, the boosting forecasts are competitive over all horizons and in most cases better than the combination forecasts. While the dynamic factor models perform best in the short and medium run, boosting based on cross-validation produces the best forecast 12 months ahead.

4 Conclusion

This paper introduces the variable selection method boosting into a horse race between factor models and forecast combination, two prominent ap-

proaches to deal with large numbers of predictors in forecasting. In an application to US industrial production, we show that boosting is a serious competitor, especially when cross-validation is used to determine the number of iterations. Based on a single data set and target variable, it is not possible to draw any general conclusions about the forecasting performance of boosting. However, it has been shown that boosting is a viable and computationally efficient alternative to other methods using many predictors.

References

Bai, J. and Ng, S. (2009). Boosting diffusion indices. *Journal of Applied Econometrics*, 24:607–629.

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 the American Statistical Association*, 98:324–339.

Carriero, A., Kapetanios, G., and Marcellino, M. (2010). Forecasting large datasets with bayesian reduced rank multivariate models. *Journal of Applied Econometrics*, forthcoming.

Freund, Y. and Schapire, R. (1996). Experiments with a new boosting algorithm. In *Proceedings of the Thirteenth International Conference on Machine Learning*, pages 148–156. Citeseer.

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

Lutz, R. and Bühlmann, P. (2006). Boosting for high-multivariate responses in high-dimensional linear regression. *Statistica Sinica*, 16:471.

Stock, J. and Watson, M. (2004). An empirical comparison of methods for forecasting using many predictors. Technical report.

Stock, J. and Watson, M. (2006). Forecasting with many predictors. In Graham, E., Granger, C., and Timmermann, A., editors, *Handbook of Economic Forecasting*, volume 1, pages 516–550. Amsterdam: North Holland.

Stock, J. and Watson, M. (2010). Dynamic factor models. In Clements, M. and Hendry, D., editors, *Oxford Handbook of Economic Forecasting*. Oxford: Oxford University Press.

Timmermann, A. (2006). Forecast combinations. In Granger, C., Elliot, G., and Timmermann, A., editors, *Handbook of Economic Forecasting*, pages 135–196. Amsterdam: North Holland.

Table 1: Forecasting with many predictors: Accuracy comparison

Method	1	3	6	12
Stock and Watson (2006)				
Univariate benchmark				
AR(AIC)	1.00	1.00	1.00	1.00
AR(4)	0.99	1.00	0.99	0.99
Multivariate Forecasts				
(1) OLS	1.78	1.45	2.27	2.39
(2) Combination forecasts				
Mean	0.95	0.93	0.87	0.87
SSR-weighted average	0.85	0.95	0.96	1.16
(3) DFM				
PCA(3,4)	0.83	**0.70**	0.74	0.87
Diagonal weighted PC(3,4)	0.83	0.73	0.83	0.96
Weighted PC(3,4)	**0.82**	**0.70**	**0.66**	0.76
(4) BMA				
X's, $g = 1/T$	0.83	0.79	1.18	1.50
Principal components, $g = 1$	0.85	0.75	0.83	0.92
Principal components, $g = 1/T$	0.85	0.78	1.04	1.50
(5) Empirical Bayes				
Parametric/g-prior	1.00	1.04	1.56	1.92
Parametric/mixed normal prior	0.93	0.75	0.81	0.89
(6) Boosting				
Linear lerner (AIC)	1.02	0.91	0.86	0.82
Linear learner (cross-validation)	0.86	0.81	0.79	**0.63**

Notes: Entries are relative MSFEs, relative to the AR(AIC) benchmark. The smallest MSFE ratio is in bold. All forecasts are recursive, and the MSFEs were computed over the period 1974:7-(2003:12-h). For details on (1) to (5) see Stock and Watson (2006).