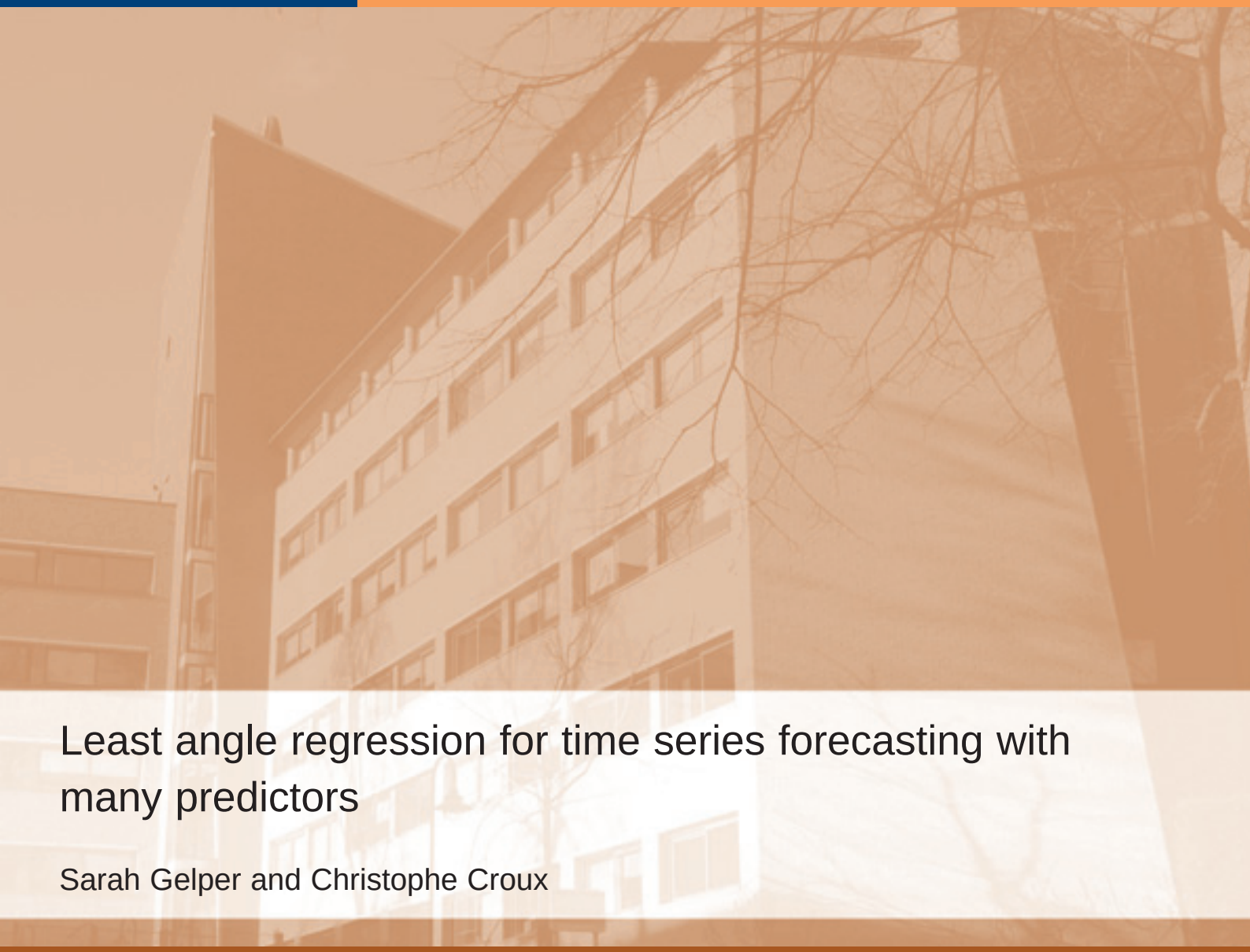
KATHOLIEKE UNIVERSITEIT
LEUVEN

Faculty of Business and Economics



Least angle regression for time series forecasting with many predictors

Sarah Gelper and Christophe Croux

DEPARTMENT OF DECISION SCIENCES AND INFORMATION MANAGEMENT (KBI)

KBI 0801

Least Angle Regression for Time Series Forecasting with Many Predictors

Sarah Gelper and Christophe Croux

Faculty of Business and Economics, Katholieke Universiteit Leuven, Naamsestraat 69, 3000 Leuven, Belgium.

Abstract: Least Angle Regression (LARS) is a variable selection method with proven performance for cross-sectional data. In this paper, it is extended to time series forecasting with many predictors. The new method builds parsimonious forecast models, taking the time series dynamics into account. It is a flexible method that allows for ranking the different predictors according to their predictive content. The time series LARS shows good forecast performance, as illustrated in a simulation study and two real data applications, where it is compared with the standard LARS algorithm and forecasting using diffusion indices.

Keywords: Macro-econometrics, Model selection, Penalized regression, Variable ranking.

1 INTRODUCTION

This article introduces a new method for forecasting univariate time series using many predictors. In various fields of application, large data sets are available and the problem of forecasting using many possible predictors is of interest. In this article, we focus on macro-economic forecasting, where the availability of time series data

is rapidly growing. Each of these time series could be informative for predicting any single economic indicator. Including all possible time series in one prediction model generally leads to poor forecasts, due to the increased variance of the estimates in a model that is too complex. Therefore, it is essential to identify the most informative predictors and separate them from the noise variables. Including only the former in the prediction model leads to parsimonious models with better forecast performance. This article presents a method which automatically identifies these most informative predictors. It is an extension of the Least Angle Regression (LARS) method proposed by Efron et al. (2004), and we call it Times Series LARS, or TS–LARS. The TS–LARS procedure takes the time series dynamics into account, since the predictive power of a time series is not only contained in the present values, but also in their lagged values. This requires a substantive extension of the standard LARS algorithm, which constitutes the main methodological contribution of this paper. TS–LARS is a flexible variable selection procedure as it allows us to select the predictors according to the macro-economic variable we want to forecast and for which horizon the forecast is made. This makes it possible to identify the short-term and long-term predictors for various macro-economic variables. The new TS–LARS method shows very good forecast performance, as will be illustrated in a simulation study and in two real data applications.

In cross-sectional analysis, extensive research has been done in the area of variable selection. The objective is to predict a univariate response making use of many covariates. Breiman (1995) introduced the non-negative garotte. This method fits a model that regresses the response on all covariates and shrinks the regression parameter estimates to zero. By deliberately setting many parameter estimates to zero, the non-negative garotte obtains variable selection. This is also achieved by the LASSO method, as proposed in Tibshirani (1996), where the LASSO puts a constraint on the parameters from an OLS fit.

More recently, Efron et al. (2004) proposed the LARS method, which can be considered as a computationally efficient version of the LASSO. The LARS method

first ranks the candidate predictors according to their predictive content. Parsimonious prediction models are then obtained by retaining only the highest ranked variables for model estimation. As such, a more compact model is selected, which can be fitted by standard estimation procedures, like OLS. The LARS algorithm is fast and shows good forecast performance in various applied fields. This is illustrated, for instance, in the biomedical application developed in Efron et al. (2004), in the field of bioinformatics in Bovelstad et al. (2007) and Saigo et al. (2007), and in Sulkava et al. (2006), who apply LARS in the context of environmental monitoring. Because of its good performance, several extensions of the LARS method have been proposed: for logistic regression (Keerthi and Shevade (2007)), multivariate regression (Simila and Tikka (2007)), analysis of variance (Yuan and Lin (2006) and Meier et al. (2008)), robust regression (Khan et al. (2007) and McCann and Welsch (2007)) and experimental design (Yuan et al. (2007)). This article extends the LARS algorithm to a time series context. To the best of our knowledge, no such modification of the LARS algorithm exists in the literature.

It is commonly recognized in time series analysis that lagged values of both the response variable, which we want to predict, and the predictors might contain predictive information. To account for these dynamic relationships, predictors are selected as blocks of present and lagged values of a time series. Selecting a time series as an important predictor corresponds to selecting the block of present and lagged values of the series. This excludes the possibility to select, for instance, the second lagged series of a certain predictor but not the first. A similar problem arises in the static case when working with categorical variables. Selecting one single categorical variable with more than two levels implies selecting several dummy variables. Blockwise variable selection for categorical variables has been studied by Yuan and Lin (2006) and Meier et al. (2008).

The need for automated variable selection in a time series context was the motivation for the development of the *Gets* method by Hoover and Krolzig (1999). The *Gets* procedure is a general-to-specific method which starts by estimating a

general unrestricted model. This model is then reduced by sequential testing and a post-search evaluation. It is an extensive set of steps and rules that result in a final parsimonious forecast model. An Ox package, called *PcGets*, was developed by Hendry and Krozlig (2001). The approach followed in this article is different, however, as the TS–LARS method starts from an auto-regressive model and then adds the predictors one by one following the LARS-computation scheme. There are at least two advantages of using TS–LARS as an alternative. First, the number of variables can be larger than the number of observations. This is not the case in the *Gets* procedure, which starts by fitting an unrestricted model. And secondly, the TS–LARS algorithm involves no testing.

The remainder of this article is organized as follows. Section 2 presents the TS–LARS algorithm in detail. It is described how the TS–LARS procedure results in a ranking of the predictors according to their incremental predictive content for the response variable. It is also discussed how the number of time series to be included in the final prediction model and the lag lengths should be selected. Section 3 extends the method of forward selection to a time series context and gives a short overview of forecasting using diffusion indices; both are alternatives to the TS–LARS for forecasting using many predictors. The excellent performance of the TS–LARS method is shown in Section 4 by means of a simulation study, and in Section 5 on two large macro-economic data sets. The first application selects predictors for forecasting aggregate US industrial production growth, from a total of 131 potential predictors. The second application concerns the prediction of Belgian industrial production growth using 75 time series measuring European-wide economic sentiment. Section 6 presents a short summary of the results and some concluding remarks.

2 LARS FOR TIME SERIES

Suppose we observe a time series, denoted by y_t, for which we want to predict future values. We observe a large number m of candidate predictors $x_{j,t}$. Index t is the time index and j the index of the predictor time series, which ranges from 1 to m. These can be used for obtaining h-step-ahead forecast of the response. We consider the linear time series model

$$\begin{aligned} y_{t+h} &= \beta_{0,0}y_t + \ldots + \beta_{0,p_0}y_{t-p_0} + \beta_{1,0}x_{1,t} + \ldots + \beta_{1,p_1}x_{1,t-p_1} + \ldots \\ &\quad + \beta_{m,0}x_{m,t} + \ldots + \beta_{m,p_m}x_{m,t-p_m} + \varepsilon_{t+h}\,, \end{aligned} \tag{1}$$

with $h \geq 1$ the forecast horizon. Model (1) explains y_{t+h} in terms of current and past values of the response itself and all the predictors. The past of the response variable is included up to lag p_0 and the past of predictor j is included up to lag p_j for $j = 1, \ldots, m$. The above can be written in matrix-notation as

$$y = \underline{y}\beta_0 + \sum_{j=1}^{m} \underline{x}_j\beta_j + \varepsilon\,, \tag{2}$$

where y is the response vector of length T. On the right hand side of equation (2), $\underline{y}$ is the $T \times (1 + p_0)$ matrix of lagged values of y from lag h to $p_0 + h$, and β_0 is the associated autoregressive parameter vector of size $1 + p_0$. Each predictor variable x_j $(j = 1, \ldots, m)$ enters model (2) as $\underline{x}_j$, a $T \times (1 + p_j)$ matrix of lagged values of x_j, and β_j is the accompanying parameter vector of size $(1 + p_j)$. The error term, a vector of length T, is denoted by ε, and it is assumed that each component has zero mean. We assume that all time series, i.e. response and predictors, are covariance stationary. Furthermore, to simplify the calculations and similar as in Efron et al. (2004), all variables are standardized so that they have mean zero and unit variance. Hence, no intercept is included in model (2).

Not all predictors in model (2) are relevant, i.e. many of the β_js are zero. The aim of the TS–LARS method is to identify which of these β_js are non-zero vectors and to obtain accurate estimates of them. Of course, in reality, many β_j's will

not be exactly zero, but will be very small. Adding them to the regression model increases the variability of the parameter estimates, while not improving the forecast performance. The TS–LARS algorithm aims at obtaining a more parsimonious model than (2), with better predictive power. To that end, the procedure first ranks the predictors, as will be discussed in Section 2.1. Once the ranking is obtained, only the highest ranked predictors will be included in the final prediction model. The exact number of predictors to include in the selected model is obtained using information criteria. A detailed discussion can be found in Section 2.2.

2.1 Predictor ranking with TS–LARS

This section outlines how the predictors are ranked according to the TS–LARS procedure. We start by fitting an autoregressive model to the response variable, excluding the predictors, using OLS. The corresponding residual series is retained and its standardized version is denoted by z_0. This scaling of z_0 is without loss of generality, and simplifies the algebra. The autoregressive model can be improved upon by including predictors, at least if the latter have incremental predictive power. The TS–LARS procedure ranks the predictors according to how much they contribute to improving upon the autoregressive fit.

In the first step, the residual series z_0 serves as the response. The first ranked predictor is that x_j, which has the highest R^2 measure

$$R^2(z_0 \sim \underline{x}_j)\,,$$

for $j = 1, \ldots, m$. Here, $R^2(y \sim x)$ denotes the R^2 measure of an OLS regression of the vector y on the variables contained in the columns of the matrix x. Recall that $\underline{x}_j$ is a matrix of lagged x_j values. The predictor with largest R^2, which we denote by $x_{(1)}$, is the first ranked predictor, hence the subscript (1). It is the first predictor included in the *active set* A. At every stage of the procedure, the active set contains all predictors ranked so far. The complement of the active set A^c contains

all predictors which have not been ranked yet. As the procedure continues, all predictors are added one by one to the active set.

Denote the hat-matrix corresponding to the first active predictor by $H_{(1)}$, which is the projection matrix on the space spanned by the columns of $\underline{x}_1$,

$$H_{(1)} = \underline{x}_{(1)}(\underline{x}'_{(1)}\underline{x}_{(1)})^{-1}\underline{x}'_{(1)}\,.$$

Furthermore, let $\hat{z}_0 = H_{(1)}z_0$ be the vector of fitted values. The current response z_0 is updated by removing the effect of $x_{(1)}$:

$$z_1 = z_0 - \gamma_1\hat{z}_0\,, \tag{3}$$

where γ_1 still needs to be determined. The scalar γ_1 is called the *shrinking factor* and takes values between zero and one. If $\gamma_1 = 1$, then z_1 simply contains the OLS residuals of regressing z_0 on $\underline{x}_{(1)}$. But the good performance of LARS results from shrinking the OLS parameters towards zero. This is achieved by multiplying them with the shrinking factor γ_1, which is chosen as the smallest positive value, such that for a predictor x_j, with $j \in A^c$, it holds that

$$R^2(z_0 - \gamma_1\hat{z}_0 \sim \underline{x}_{(1)}) = R^2(z_0 - \gamma_1\hat{z}_0 \sim \underline{x}_j)\,. \tag{4}$$

Condition (4) is an extension of the *equi-correlation* condition of the LARS procedure developed by Efron et al. (2004). In standard LARS, one adds single variables one by one, whereas in our case we add blocks of lagged values of a time series. For standard LARS, the R^2 in equation (4) reduces to a squared correlation, and equation (4) is trivial to solve. For TS–LARS, as is shown in the Appendix, condition (4) is equivalent to solving the following quadratic equation in γ:

$$z'_0(H_{(1)}-H_j)z_0+z'_0(H_{(1)}H_j+H_jH_{(1)}-2H_{(1)})z_0\gamma+z'_0(H_{(1)}-H_{(1)}H_jH_{(1)})z_0\gamma^2 = 0\,, \tag{5}$$

with H_j the projection matrix on the space spanned by $\underline{x}_j$, so $H_j = \underline{x}_j(\underline{x}'_j\underline{x}_j)^{-1}\underline{x}'_j$. We show in the Appendix that for every j in A^c, we find two solutions for condition (5), of which at least one between zero and one. The shrinking factor γ_1 is then

chosen as the smallest positive solution to condition (5), taken over all indices j in the non-active set A^c.

Equation (5) can be written in a more compact form, avoiding the use of multiple matrix multiplications. Denote the standardized version of $\hat{z}_0$ by $\tilde{x}_{(1)}$, so

$$\tilde{x}_{(1)} = \frac{\hat{z}_0}{s_1} \qquad \text{for} \qquad s_1^2 = \frac{\hat{z}_0' \hat{z}_0}{T-1}\,.$$

In this paper, all variances and covariances are computed with denominator $(T-1)$, with T the number of observations. In the Appendix, it is shown that equation (5) is equivalent with

$$(T-1)s_1^2 - z_0' H_j z_0 + 2(z_0' H_j \tilde{x}_{(1)} - (T-1)s_1)\,(s_1\gamma) + ((T-1) - \tilde{x}_{(1)}' H_j \tilde{x}_{(1)})(s_1\gamma)^2\,, \quad (6)$$

which is computationally faster to solve than equation (5).

The TS–LARS algorithm chooses the shrinking parameter γ_1 in equation (3) simultaneously with the next predictor entering the active set. In particular, the second time series in the active set is the one with index j yielding the smallest positive value of γ_1. Denote this predictor by $x_{(2)}$, where the subscript (2) indicates it is the second ranked predictor in the active set. The active set now contains two predictors and the response z_1 is obtained according to equation (3). Then we scale the response z_1 to unit variance for numerical convenience.

Since the second and all following steps have the same structure, we generalize from here on to step k. At the beginning of step k, the active set A contains k active or ranked predictors $x_{(1)}, x_{(2)}, \ldots, x_{(k)}$, with $k \geq 2$. The current response is denoted by z_{k-1}. Let $\tilde{x}_{(i)}$ be the standardized vector of fitted values $H_{(i)} z_{i-1}$ for $i = 1, \ldots, k$. First, we look for the *equiangular vector* u_k, which is defined as the vector having equal correlation with all vectors $\tilde{x}_{(1)}, \tilde{x}_{(2)}, \ldots, \tilde{x}_{(k)}$. This correlation is denoted by a_k

$$a_k = \mathrm{Cor}(u_k, \tilde{x}_{(1)}) = \mathrm{Cor}(u_k, \tilde{x}_{(2)}) = \ldots = \mathrm{Cor}(u_k, \tilde{x}_{(k)})\,. \quad (7)$$

The equiangular vector u_k is easy to obtain (e.g. Khan et al. (2007) and Efron et al. (2004)). Let R_k be the $(k \times k)$ correlation matrix computed from $\tilde{x}_{(1)}, \tilde{x}_{(2)}, \ldots, \tilde{x}_{(k)}$

and $\mathbf{1_k}$ a vector of ones of length k. The equiangular vector u_k is then a weighted sum of $\tilde{x}_{(1)}, \ldots, \tilde{x}_{(k)}$:

$$u_k = (\tilde{x}_{(1)}, \tilde{x}_{(2)}, \ldots, \tilde{x}_{(k)})w_k \qquad \text{with} \qquad w_k = \frac{R_k^{-1}\mathbf{1_k}}{\sqrt{\mathbf{1_k}' R_k^{-1}\mathbf{1_k}}}\,.$$

Note that the equiangular vector has unit variance.

Afterwards, the response is updated by moving along the direction of the equiangular vector

$$z_k = z_{k-1} - \gamma_k u_k\,. \tag{8}$$

The shrinking factor γ_k is chosen as the smallest positive solution such that for a predictor x_j not in the active set it holds that

$$R^2(z_{k-1} - \gamma u_k \sim \tilde{x}_{(k)}) = R^2(z_{k-1} - \gamma u_k \sim \underline{x}_j)\,. \tag{9}$$

The associated predictor, denoted by $x_{(k+1)}$, is then added to the active set. Once γ_k is obtained, we can update the response as in equation (8) and the new response is than standardized and again denoted by z_k. In the Appendix, we prove the following lemma.

Lemma 1 *For every step $k \geq 1$ in the TS–LARS algorithm, it holds that*

(a) The current response z_{k-1} has equal and positive correlation with all $\tilde{x}_{(1)}, \ldots, \tilde{x}_{(k)}$ in the active set:

$$r_k = Cor(z_{k-1}, \tilde{x}_{(1)}) = \ldots = Cor(z_{k-1}, \tilde{x}_{(k)}) \geq 0\,. \tag{10}$$

(b) For every j not in the active set, it holds that

$$R^2(z_{k-1} \sim \underline{x}_j) \leq r_k^2\,.$$

(c) For the solution γ_k to (9) it holds that $0 \leq \gamma_k \leq r_k/a_k$.

From the above lemma, it follows that, in equation (9), the index k can be replaced by any other number from 1 to k. Using the coefficient of correlation r_k, defined be (10), it is shown in the Appendix that condition (9) is equivalent to solving the following quadratic equation in γ:

$$(T-1)r_k^2 - z'_{k-1}H_j z_{k-1} + 2(z'_{k-1}H_j u_k - (T-1)a_k r_k)\gamma + ((T-1)a_k^2 - u'_k H_j u_k)\gamma^2 = 0\,. \tag{11}$$

The TS–LARS algorithm solves equation (11) for every j in A^c and retains the smallest positive solution over all j in A^c. The associated variable is added to the active set and denoted by $x_{(k+1)}$. This procedure is continued either until all predictors have been ranked, or until more predictors have been ranked than there are observations in each time series.

2.2 Variable and lag length selection

After ranking the predictors, only the highest ranked ones will be included in the final prediction model. The variable selection problem reduces to choosing the number of highest ranked predictors to be included in the prediction model. This number can be chosen according to different information criteria. Efron et al. (2004) propose using the C_p information criterion, but we prefer to use the Bayesian Information Criterion (BIC). The BIC is well known to be a good information criterion in time series analysis, as discussed, for instance, in Qian and Zhao (2007). At every stage of the LARS procedure, we perform an OLS fit to model (1), where only the predictors in the active set are included. We store these BIC values and choose the model with optimal BIC, selecting k^* predictors. We allow k^* to be equal to zero. If this is the case, no predictors are included and a pure autoregressive model is selected.

An additional difficulty in a time series context is that the vector of lag lengths $(p_0, p_1, \ldots, p_m)$ is unknown. The lag length of the autoregressive component, p_0, can be chosen at the beginning of the procedure by optimizing the BIC for the autoregressive model. Concerning the selection of p_1 to p_m, we assume fixed lag

lengths for the predictors, i.e. $p = p_1 = p_2 = \ldots = p_m$. For different values of the lag length p, we run the TS–LARS algorithm, and obtain a selected model with a number of predictors minimizing BIC. The final prediction model, then, is the model obtained by minimizing BIC further over all the considered values of p.

In principle, it is possible to include an automated mechanism into the TS–LARS method, allowing for different lag lengths for different predictors. In every TS–LARS step, one finds the optimal lag length for every time series in the non-active set with respect to the current response. Given the large number of predictors, such an approach will be computationally too demanding.

3 OTHER METHODS

In this section, we review two other forecasting methods applicable when many predictors are available. The first one is the straightforward time series extension of forward variable selection. The second one is forecasting with diffusion indices, following the dynamic factor model approach of Stock and Watson (2002a). In Sections 4 and 5, the performance of TS–LARS method will be compared with these two approaches.

Forward Selection of Predictors

Forward selection is a well known variable ranking and selection method considered in many textbooks on applied regression analysis, such as Dielman (2001). It is a competitor of the LARS method, and its extension to the time series context will be called the TS–FS. The first step of the TS–FS method consists of regressing the response y_t on its own past only and retaining the residuals z_0 of this auto-regressive model. As in the TS–LARS procedure, the first ranked predictor $x_{(1)}$ is the one with the largest R^2 value

$$R^2(z_0 \sim \underline{x}_j)\,,$$

for $j = 1, \ldots, m$. Retain the residuals z_1 of regressing z_0 on $\underline{x}_{(1)}$ by OLS. No shrinkage is involved here. The second ranked predictor, then, is the one with largest value of

$$R^2(z_1 \sim \underline{x}_j)\,,$$

for all predictors x_j which have not yet been selected. This procedure is continued until all predictors have been ranked.

The selection of the number of predictors to be included in the prediction model and the lag length selection are carried out by minimizing the BIC, in the same way as explained in Section 2.2. Since every step uses only one block of variables as regressors, the forward selection method allows to rank all predictors, regardless of the number of observations. This is in contrast with backward variable selections, which requires the number of observations in each time series to be larger than the number of variables.

Forecasting Using Diffusion Indices

Time series forecasting using many predictors is receiving increasing attention in the literature. For a recent overview, we refer to Stock and Watson (2006), who discuss three approaches in depth: forecast combination, Bayesian model averaging and dynamic factor models. In forecast combination, many forecasts from different models are combined to obtain one final forecast. In the Bayesian framework, model averaging is achieved by assigning probabilities to each model. Recent developments in forecast combination and Bayesian model averaging can be found in Ekkund and Karlsson (2007). The methods presented in this article are different in the sense that we obtain forecasts from one parsimonious model in which the included predictors are automatically selected from the many candidate predictors.

The method we compare with is the basic Dynamic Factor Model (DFM), which uses diffusion indices or latent factors. These diffusion indices are extracted from the predictors and are then used in a final prediction model. The underlying idea

is that there are a few unobserved forces driving the economy or influencing both the predictors and the response (see e.g. Stock and Watson (2002a), Forni et al. (2005), and Bai and Ng (2007)). The dynamic factor model is based on this idea and tries to extract these latent factors from the predictors. The latent factors are denoted by $f_1, f_2, \ldots$ and extracted from the predictors by the method of principal components. The forecast model associated with the DFM regresses the response y_{t+h} on current and lagged values of the response itself and on current and lagged values of the factors

$$y_{t+h} = \beta_{0,0} y_t + \ldots + \beta_{0,p_0} y_{t-p_0} + \beta_{1,0} f_{1,t} + \ldots + \beta_{1,p_1} f_{1,t-p_1} + \ldots + \beta_{k,p_k} f_{k,t-p_k} + \varepsilon_{t+h}\,, \quad (12)$$

with k the number of latent factors. Note that the autoregressive model is a special case of the DFM, when the optimal number of factors to include is zero. In this paper, we choose the number of factors and the lag lengths according to the BIC. The simulation results in Stock and Watson (2002a) indicate that the BIC is a good information criterion for selecting the number of factors. More advanced methods for estimating the number of factors can be found in Bai and Ng (2002) and Dante and Watson (2007).

Conceptually, the TS–LARS method has two advantages over the dynamic factor model. First, the TS–LARS method takes the response variable and the forecast horizon into account while building the forecast model, as was also considered by Heij et al. (2007). This allows us to use different types of information from the many predictors depending on the value to forecast. In the dynamic factor model, on the other hand, the response is not at all involved in extracting the latent factors from the data, while in reality, different predictors are important for different responses and different forecast horizons. Secondly, the model selected by TS–LARS is directly interpretable in terms of the original predictors. This contrasts with the dynamic factor model which can be hard to interpret since it is not always clear what the extracted factors measure.

4 SIMULATION STUDY

In this section, we compare the TS–LARS procedure with four other methods for univariate time series forecasting using many predictors. First, we compare with a static approach, where we only use current values of the response and the predictors for predicting y_{t+h}, taking all lag lengths in model (1) equal to zero. Variables will be ranked using standard LARS, as described in Efron et al. (2004). No blockwise selection of predictors is required, hence we simply refer to this method as LARS. It might be that TS–LARS and LARS select the same model, but in most applications lagged values of the predictors will be included by the TS–LARS algorithm. The simulation study will show that taking the dynamics of the series into account leads to a much better performance. Both LARS and TS–LARS are compared with time series forward selection (TS–FS), and forecasting with diffusion indices in the dynamic factor model (DFM), as discussed in Section 3.

Different aspects of the procedures will be studied. First, we will compare the predictor ranking and selection performance of TS–LARS, LARS and TS–FS. Are the highest ranked predictors indeed the relevant ones? Note that the DFM method is not performing a ranking, and neither a selection of the predictors, and hence will not be included in this comparison. We also study whether the number of relevant predictors and the lag length are appropriately selected using the BIC criterion. Secondly, the forecast performance of the variable selection methods TS–LARS, LARS and TS–FS is compared with the DFM approach. For studying the forecast performance, two different simulation models are considered: a linear time series model as in equation (1) and a latent factor model as in equation (12). We add the latent factor setting as a second simulation scheme for forecast comparison, as one would expect that the DFM method gives the better results in that setting. It turns out, however, that the TS-LARS procedure outperforms the other methods considered for both simulation schemes.

Simulation Schemes

The first simulation setting generates time series according to a linear time series model. The data generating process is given by

$$y_{t+1} = \beta_{0,0} y_t + \beta_{0,1} y_{t-1} + \sum_{j=1}^{20} \beta_{j,0} x_{j,t} + \sum_{j=1}^{20} \beta_{j,1} x_{j,t-1} + \varepsilon_{t+1}\,, \tag{13}$$

where ε_{t+1} is i.i.d. $N(0,2)$. Only five of the $m = 20$ candidate predictors are relevant, while the remaining 15 predictors are redundant. The lagged values of the dependent variable enter the model with two lags. Denote $\beta_j = (\beta_{j,0}, \beta_{j,1})'$, then the regression parameters are

$$\beta_0 = (0.4\,,\, 0.1)',\ \beta_1 = (4\,,\, 2)',\ \beta_2 = (3\,,\, 1.5)',$$

$$\beta_3 = (2\,,\, 1)',\ \beta_4 = (1\,,\, 0.5)' \text{ and } \beta_5 = (0.5\,,\, 0.25)'\,.$$

For the redundant predictors x_6 to x_{20} we have $\beta_j = \mathbf{0}$, for $j = 6, \ldots, 20$. Furthermore, the predictors are auto- and cross-correlated. The first two relevant predictors are simulated from the following VAR(1) model:

$$\begin{pmatrix} x_{1,t} \\ x_{2,t} \end{pmatrix} = \begin{pmatrix} 0.5 & 0.3 \\ 0.3 & 0.5 \end{pmatrix} \begin{pmatrix} x_{1,t-1} \\ x_{2,t-1} \end{pmatrix} + \begin{pmatrix} e_1 \\ e_2 \end{pmatrix}.$$

Additionally, four of the redundant predictors are simulated according to the VAR(1) model

$$\begin{pmatrix} x_{6,t} \\ x_{7,t} \\ x_{8,t} \\ x_{9,t} \end{pmatrix} = \begin{pmatrix} 0.5 & 0.3 & 0.1 & 0 \\ 0.3 & 0.5 & 0 & 0.1 \\ 0.1 & 0 & 0.5 & 0.3 \\ 0 & 0.1 & 0.3 & 0.5 \end{pmatrix} \begin{pmatrix} x_{6,t-1} \\ x_{7,t-1} \\ x_{8,t-1} \\ x_{9,t-1} \end{pmatrix} + \begin{pmatrix} e_6 \\ e_7 \\ e_8 \\ e_9 \end{pmatrix}.$$

All the other predictors are generated from the following AR(1) model:

$$x_{i,t} = a_i x_{i,t-1} + e_i \text{ for } i = 3, 4, 5, 10, 11, \ldots, 20\,,$$

where the auto-regressive parameter a_i is chosen randomly between 0 and 0.8, according to a uniform distribution. The error components e_i for $i = 1, \ldots, 20$ are independent of each other, and all follow an i.i.d. $N(0,1)$ process.

The second simulation setting uses a latent factor model and relates to the dynamic factor model, as discussed in Section 3. In this setting, we simulate two latent factors, denoted by L_1 and L_2, that follow the VAR model

$$\begin{pmatrix} L_{1,t} \\ L_{2,t} \end{pmatrix} = \begin{pmatrix} 0.5 & 0.3 \\ 0.3 & 0.5 \end{pmatrix} \begin{pmatrix} L_{1,t-1} \\ L_{2,t-1} \end{pmatrix} + \begin{pmatrix} l_{1,t} \\ l_{2,t} \end{pmatrix},$$

where l_1 and l_2 are i.i.d. $N(0,1)$. The relevant predictors are generated as

$$\begin{aligned} x_{1,t} &= 3L_{1,t} + e_1\,, \\ x_{2,t} &= 0.5L_{1,t} + e_2\,, \\ x_{3,t} &= 3L_{2,t} + e_3\,, \\ x_{4,t} &= 0.5L_{2,t} + e_4\,, \\ x_{5,t} &= 0.5L_{1,t} + 0.3L_{2,t} + e_5\,, \end{aligned}$$

where e_1 to e_5 are i.i.d. $N(0,1)$ noise components. The response depends on lagged values of the latent factors

$$y_{t+1} = \beta_{0,0}y_t + \beta_{0,1}y_{t-1} + 2L_{1,t} + 2L_{1,t-1} + L_{2,t} + L_{2,t-1} + \varepsilon_{t+1}$$

where $\beta_{0,0}$, $\beta_{0,1}$ and ε_{t+1} are as before in the previous setting. The redundant predictors are simulated in the same way as in the first simulation setting. For both simulation schemes, we generate $M = 2000$ data sets, each consisting of 20 candidate predictor series and one series to predict, where all series are of length $T = 150$.

Ranking of predictors and model selection

We generate time series according to the model (13), yielding a total number of relevant predictors equal to five in every simulation run. We compare the predictor ranking obtained by the TS–LARS, LARS and TS–FS procedures using recall-curves. A recall-curve plots the number of relevant predictors among the first k

ranked predictors, where k ranges from 1 to the total number of predictors. The steeper the recall-curve goes towards its maximum value of five, the better.

Figure 1 shows the recall-curves of the TS–LARS, LARS and TS–FS methods applied for forecast horizon $h = 1$, and averaged over all simulation runs. The TS–LARS method has the steepest recall-curve and thus shows the best performance in terms of predictor ranking. On average, the relevant predictors are ranked higher by the TS–LARS procedure than by the LARS and TS–FS procedures. The latter two perform equally well. For example, when retaining the highest $k = 5$ ranked predictors, we see from Figure 1 that, on average, more than 4 relevant predictors will be in this set of 5. In contrast, the other two methods have an average recall below 4 out of 5. For all methods, it is observed that the recall curve increases quickly, and then tends rather slowly to the maximum value 5. The reason is that the explanatory power of the fifth predicting time series is very low, having a small value of β_5 relative to the error variance. Hence, it is difficult to distinguish this fifth predictor from the redundant ones.

In Section 2.2, we proposed using the BIC as a criterion for both the selection of predictors and the lag length. We can select an incorrect model both by selecting an incorrect set of predictors or by selecting an incorrect lag length.

As concerns the selection of predictors, four scenarios can occur. The best scenario would be to select exactly the set of 5 relevant predictors. Another possibility is under-selection, i.e. all selected predictors are relevant, but not all relevant predictors are selected. Over-selection, on the other hand, means that all relevant predictors are selected as well as some redundant predictors. The remaining scenario is the selection of a model including some relevant and some redundant predictors. The results for forecast horizon $h = 1$ are summarized in Table 1. In 26% of the simulation runs, the TS–LARS selects exactly the set of relevant predictors, as compared to only 1% for the LARS and 8% for the TS–FS procedures. In 42% of the simulation runs, the TS–LARS method did not succeed in selecting all relevant predictors. But under-selection is more a problem of the LARS and TS–FS methods,

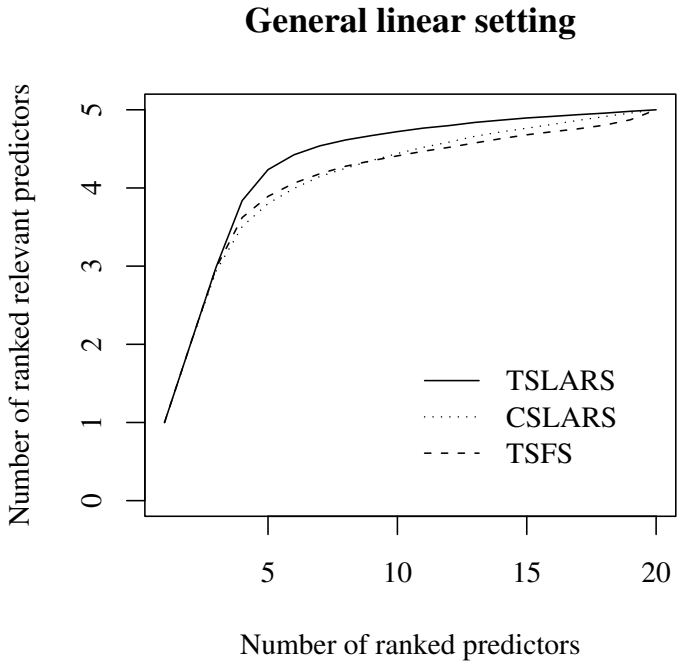


Figure 1: Recall curves averaged over the 2000 simulation runs. The full line represents the TS–LARS method, the dashed line the TS–FS method and the dotted line the LARS method.

Table 1: Percentage of simulation runs where number of relevant predictors was correctly identified and where it was under- or over-selected.

	Correct	Under	Over
TS–LARS	0.26	0.42	0.01
LARS	0.01	0.50	0.02
TS–FS	0.08	0.61	0.00

Table 2: Percentage of correctly selected, under- and over-selected lag lengths.

	Correct	Under	Over
TS–LARS	0.87	0.11	0.02
TS–FS	0.91	0.04	0.05

selecting a model where some of the relevant predictors are missing in 50% and 61% of all runs, respectively. In this simulation study, over-selection is not an issue. This means that it almost never occurs that redundant predictors are selected, when all relevant ones are already in the model.

We further look at the performance of the TS–LARS and TS–FS procedures with respect to lag length selection. The LARS method is not considered here because it does not perform any lag length selection. There are three possible scenarios: under-, over- or correct selection of the lag length. The results, again for $h = 1$, are summarized in Table 2. In a large majority of the simulation runs, both the TS–LARS and the TS–FS algorithms select the correct lag length. In the other cases, the lag length is taken too short rather than too long.

To conclude, we may say that the BIC criterion succeeds reasonably well in specifying the correct model, both in terms of the number of relevant predictors and in terms of the lag length. The fully correctly specified model with $p = 1$ and $k = 5$ relevant predictors, is retrieved in 24% of the simulation runs using the TS–LARS

method. This is substantially higher than for the TS–FS method, which has a hit rate of only 8%. When interpreting these rather low numbers, one should take into account that the number of candidate models is very large, and the time series of moderate length $T = 150$. Moreover, even when the number of relevant predictors is not correctly specified, the resulting model may still have very good forecasting performance, as will be discussed in the remainder of this section.

Forecast performance

We study the out-of-sample forecast performance of the TS–LARS, LARS and TS–FS procedures and make a comparison with the DFM method discussed in Section 3. In every simulation run, we fit the models selected by the different procedures to the simulated data-set of length 150. We then obtain one to five step ahead forecasts for observations 151 to 155. These forecasts are compared to the realized values of the time series as simulated according to the data generating process. The performance of the forecast methods is then compared using the Mean Squared Forecast Error at horizon h:

$$\text{MSFE}_h = \frac{1}{M}\sum_{i=1}^{M}(y_{i,150+h} - \hat{y}_{i,150+h})^2\,,$$

where the index i indicates the simulation runs and $M = 2000$ is the number of simulation runs.

The results for the first simulation setting are presented in Table 3, where the simulated out-of-sample MSFE_h is reported for $h = 1, \ldots, 5$. First of all, we observe that for all methods the MSFE_h steeply increases in h, as is to be expected. Most striking is that for almost all forecast horizons the TS–LARS method yields the smallest MSFE_h. In 14 out of 15 cases, as shown by applying a paired t-test (p-value reported in the table), the MSFE_h of the TS-LARS method is significantly smaller than for the other methods. At horizon 5 the LARS method performs best, but the difference with TS–LARS is not significant.

The better performance of TS–LARS with respect to the static LARS shows

Table 3: Simulated MSFE_h at several forecast horizons using the TS–LARS, LARS, TS–FS and DFM method. Series are generated from a linear time series model. The smallest MSFEs are indicated in italics; p-values of a paired t-test for equal MSFE_h with respect to TS–LARS are given between parentheses.

	$h=1$	$h=2$	$h=3$	$h=4$	$h=5$
TS–LARS	*55.49*	*127.36*	*220.74*	*316.30*	403.26
LARS	76.55	145.70	234.65	322.20	*403.13*
	(< 0.01)	(< 0.01)	(< 0.01)	(0.03)	(0.67)
TS–FS	59.08	132.32	227.96	333.03	423.87
	(< 0.01)	(< 0.01)	(0.01)	(< 0.01)	(< 0.01)
DFM	73.64	149.25	247.55	337.19	420.34
	(< 0.01)	(< 0.01)	(0.01)	(< 0.01)	(0.01)

that it is worth taking lagged values of the series into account. More interesting is that TS-LARS outperforms both the more simple TS–FS method, and the popular DFM approach. The worse performance of the DFM method could be explained by the fact that all predictors enter in the construction of the factors, even the redundant ones. Of course, the latter will receive a small weight in the construction of the factor, but they still increase variability of the estimates, leading to poorer forecasting performance.

The second simulation scheme is a latent factor model, where one expects the DFM approach to yield better results. However, as can be seen from Table 4, the results are similar to those obtained before. For almost all forecast horizons, the TS–LARS method yields the smallest MSFE_h, and the differences with the other methods are significant in most cases. An exception is $h = 5$, where the DFM method is better, but the difference with TS–LARS is not significant. It is, however, worth mentioning that for $h = 5$ the DFM approach mostly selects $k = 0$ predictors, meaning that a pure autoregressive model is taken. The other methods based on

Table 4: Simulated MSFE_h at several forecast horizons using the TS–LARS, LARS, TS–FS and DFM method. Series are generated from a latent factor model. The smallest MSFEs are indicated in italics; p-values of a paired t-test for equal MSFE_h with respect to TS–LARS are given between parentheses.

	$h = 1$	$h = 2$	$h = 3$	$h = 4$	$h = 5$
TS–LARS	*20.60*	*38.52*	*66.58*	*99.55*	124.74
LARS	23.92	43.78	71.83	102.10	124.67
	(< 0.01)	(< 0.01)	(< 0.01)	(0.01)	(0.64)
TS–FS	24.75	42.58	72.48	105.17	138.18
	(< 0.01)	(< 0.01)	(< 0.01)	(< 0.01)	(< 0.01)
DFM	29.60	47.84	73.28	102.35	*122.76*
	(< 0.01)	(< 0.01)	(< 0.01)	(0.12)	(0.40)

selection of predictors, i.e. LARS and TS–FS, also generally yield better results than forecasting using diffusion indices.

We conclude that for the simulation schemes under consideration, the TS–LARS method outperforms its competitors, in particular for short term forecasting.

5 APPLICATIONS

5.1 Predicting Aggregate Industrial Production Growth in the US Using Many Economic Indicators.

To illustrate the performance of the TS–LARS method using real data, we repeat the example elaborated by Stock and Watson (2002a) and Stock and Watson (2002b) for forecasting US industrial production growth. The data used are available on Mark W. Watson's homepage[1] and were previously used in, for instance, Stock and Watson

[1] http://www.princeton.edu/∼mwatson/

(2005) and de Poorter et al. (2007). The data set includes 132 time series covering many different aspects of the US economy, such as price indices, unemployment rates, interest rates and money supply numbers. There is no missingness in the data and we have 528 monthly observations at our disposal, ranging from January 1960 to December 2003. The aim is to obtain accurate predictions of the industrial production growth rate, while the remaining 131 time series are candidates to be included in the forecast model. The industrial production growth rate is obtained as the log-differences of real US industrial production. A precise description of all 132 series can be found in Stock and Watson (2005). They also describe how the original series can be transformed to be stationary. For some series, this requires differencing, for others taking a log-transformation or log-differencing. We apply the same transformations.

Using the TS–LARS algorithm, we obtain a ranking of the predictors according to their additional predictive power for one-month-ahead US industrial production growth. The top-5 ranking, as obtained by the TS–LARS method, is presented in Table 5. The highest ranked predictor, the NAPM new orders index, is a survey-based indicator constructed by the National Association of Purchasing Management. More than 300 purchasing and supply executives from across the US indicate recent evolutions in their business with respect to orders, prices and unemployment among other things. The TS–LARS algorithm suggests that the number of new orders is the most important predictor for future growth in industrial production. Other important predictors quantify the number of job openings in newspapers, interest rates, wages and consumer expectations. The smallest BIC value is obtained by including the 4 highest ranked predictors up to one lag in the forecasting model.

To compare the forecast performance of the different methods, we divide the data in an in-sample and an out-of-sample part. The in-sample part of the data ranges from January 1960 to December 1981, with a length of $R = T/2$. The selected model is recursively estimated for $t = R, \ldots, T - h$, to obtain a series of out-of-sample h-step-ahead forecasts. The number of included predictors and the

Table 5: Top-5 ranking of predictor series for predicting one step ahead US industrial production growth, as obtained by the TS–LARS algorithm.

Ranking	Predictor
1	NAPM new orders index
2	Index of help-wanted advertising in newspapers
3	3-month FF spread
4	Average hourly earnings: construction
5	Michigan index of consumer expectations

Table 6: Out-of-sample MSFE_h ($\times 10000$) obtained by the TS–LARS, LARS, TS–FS and DFM methods, for different forecast horizons. The smallest MSFE_h value is in italics; p-values of the Diebold-Mariano test for equal forecast accuracy compared to TS–LARS are between parentheses.

	$h=1$	$h=2$	$h=3$	$h=6$	$h=12$
TS–LARS	*60.81*	*58.68*	*64.82*	*71.13*	*81.79*
LARS	62.41	59.60	73.71	75.50	*81.79*
	(0.57)	(0.79)	(0.02)	(0.02)	
TS–FS	65.31	74.21	72.19	78.04	87.35
	(0.42)	(0.01)	(0.04)	(0.04)	(0.08)
DFM	69.28	66.72	72.36	79.80	84.57
	(0.07)	(0.03)	(0.11)	(0.11)	(0.64)

lag length are obtained as discussed in Section 2.2. The mean squared forecast error at horizon h is computed as

$$\text{MSFE}_h = \frac{1}{T-h-R+1} \sum_{t=R}^{T-h} (y_{t+h} - \hat{y}_{t+h})^2 .$$

The MSFE_h is presented in Table 6 for the procedures TS–LARS, LARS, TS–FS and DFM, at several forecast horizons h. First of all, it is clear that the prediction model selected by the TS–LARS method results in the smallest MSFE_h at every horizon. To check whether the observed differences in MSFE_h are significant, p-values are reported for a Diebold-Mariano test for equal out-of-sample predictive accuracy in comparison with the TS–LARS (Diebold and Mariano (1995)). The significance results vary according to the forecast horizon. For short-term forecasts ($h = 1$ and $h = 2$), there is no significant difference between the three methods that rely on variable selection method, LARS, TS–LARS and TS–FS. However, TS–LARS does significantly better than DFM. For 6-month forecasts, the TS–LARS approach gives significantly lower MSFE_h than both other variable selection procedures. In particular, the difference between TS–LARS and LARS is pronounced, showing that accounting for dynamic relationships, as done by the TS–LARS, is worthwhile. At long forecast horizons ($h = 12$), all methods perform comparably. Note that the TS–LARS selects a model here with only current values of the selected predictors, and therefore yields an identical MSFE_h as standard LARS.

The good performance of the TS–LARS method on this well-known data set is striking and promising. The number of retained predicting time series for TS–LARS ranges from $k = 4$, for forecast horizon one, to $k = 2$, for forecast horizon 12, yielding very parsimonious models. Finally, including a total of 12 factors in the forecast model, as was done in Stock and Watson (2002b), does not improve the MSFE_h for the DFM approach (results available upon request).

5.2 Predicting Country Specific Industrial Production Using Sentiment Indicators.

As a second application, we study the predictive content of European consumer and production sentiment surveys. We are interested in forecasting Belgian industrial production growth and use sentiment surveys from all over Europe. For every EU15 country, we use 5 sentiment indicators: consumer, industrial, retail, services and construction sentiment indicators, resulting in a total of 75 sentiment indicators. All these indicators are potential predictors for Belgian industrial production growth. The data range from April 1995 to October 2007, resulting in a total of 151 observations. All the data are publicly available on the Eurostat website[2].

Running the TS–LARS algorithm with forecast horizon one leads to the top-5 ranking as presented in Table 7. Using the BIC, one retains the first two predictors, each with one lag. The industrial confidence indicators in France and Belgium are identified as the most powerful predictors for short term forecasting of Belgian industrial production growth. This suggests that, in the short run, industrial production growth might be strongly driven by confidence of the industrial sector itself. Other high ranked predictors, not included in the selected model, are consumer confidence in the Netherlands and retail confidence in Germany and France. It is interesting to note that the five highest ranked predictors are from Belgium itself or neighboring countries. For one-year-ahead forecasting ($h = 12$) of Belgian industrial production growth, the TS–LARS method selects the Belgian and German consumer confidence indicators to be included in the prediction model. So, in the long run, industrial production growth is more accurately predicted by consumer confidence than by confidence in the industrial sector.

To compare the forecast performance of the TS–LARS method to the LARS, TS–FS and DFM, we compare out-of-sample forecasts. The in-sample range includes the first 75 observations and we proceed in the same way as in Section 5.1. The results of

[2] ec.europa.eu/eurostat

Table 7: Top-5 ranking of predictor series for predicting one step ahead Belgian industrial production growth, as obtained by the TS–LARS algorithm.

Ranking	Predictor
1	Industrial Confidence, France
2	Industrial Confidence, Belgium
3	Consumer Confidence, the Netherlands
4	Retail Confidence, Germany
5	Retail Confidence, France

the out-of-sample forecast comparisons are presented in Table 8, where the MSFE_h is reported for the TS–LARS, LARS, TS–FS and DFM methods and for several forecast horizons. For forecasting one-month-ahead industrial production growth, the TS–LARS gives the best results. Even more, the TS–LARS is significantly better than all three other methods. No significant differences are obtained for two-months-ahead forecasts, but for three-months-ahead we again see that the TS–LARS performs significantly better than the other three considered methods. For longer forecast horizons, all methods perform comparably.

6 CONCLUSION

The LARS procedure, as presented in Efron et al. (2004), is a fast and well-performing method for model selection with cross-sectional data. We propose a new version of the LARS especially designed for time series data and call it the TS–LARS. The new procedure takes the dynamics of the time series into account by selecting blocks of variables. Each block consists of a number of lagged series of a predictor. The good performance of the TS–LARS method is demonstrated by means of a simulation study. It performs better than the standard LARS method, which does not take the dynamics into account. Moreover, in terms of forecast

Table 8: Out-of-sample MSFE_h ($\times 1000$) obtained by the TS–LARS, LARS, TS–FS and DFM methods, for different forecast horizons. The smallest MSFE_h value is in italics; p-values of the Diebold-Mariano test for equal forecast accuracy compared to TS–LARS are between parentheses.

	$h=1$	$h=2$	$h=3$	$h=6$	$h=12$
TS–LARS	*22.63*	16.98	*17.30*	17.17	19.68
LARS	25.95	*14.24*	20.73	17.18	*19.48*
	(0.01)	(0.09)	(0.01)	(0.89)	(0.75)
TS–FS	26.60	15.76	21.28	*17.10*	*19.48*
	(0.02)	(0.13)	(<0.01)	(0.56)	(0.75)
DFM	26.36	15.72	26.09	17.18	20.74
	(<0.01)	(0.45)	(<0.01)	(0.89)	(0.25)

performance, the TS–LARS method shows better results than the dynamic factor model.

In the first empirical application, we apply the TS–LARS method to a large data set to identify which predictors are most informative for future US monthly industrial production growth rates. The highest ranked short-term predictor is the "NAPM new orders index" indicating that evolutions in the number of placed orders is a good predictor for industrial production growth. The second application studies the predictive content of EU business and consumer surveys for future Belgian monthly industrial production growth rates. The TS–LARS procedure indicates that the industrial confidence indicators of both France and Belgium are the most informative short-term predictors. The models selected by the TS–LARS procedure shows good forecast performance as compared to a dynamic factor model, especially for short term forecasting.

In this article, the TS–LARS method acted as a variable selection technique, after which a standard OLS-fit was applied to the selected model. Our experience is

that the selected model typically contains only few predictors, avoiding overfitting by OLS of the selected model. In the examples considered in this paper, parsimonious prediction models were obtained by making use of an OLS fit. In cases where the TS–LARS method would yield a fairly large prediction model, estimation by OLS can be improved upon, since it is well known that OLS estimators may have high variance in models with too many predictors. A solution would be to use a penalized version of OLS, for example the LASSO estimators (Tibshirani (1996)), instead. Alternatively, one could use TS–LARS both as a variable selection and as a model fitting procedure, using the shrinked OLS regression estimates computed in the TS–LARS algorithm outlined in Section 2.1.

A distinct feature of TS–LARS is that it allows for a ranking of the different predictors. This ranking differs according to the series one wants to predict and the forecast horizon. The highest ranked time series are the ones that need to be followed up closely by the forecaster. If data monitoring is expensive or time-consuming, then one might continue to track only the highly ranked series, and not the less informative ones. To conclude, we think that the use of TS–LARS algorithm for predicting time series using many predictors yields parsimonious and easy-to-interpret models, with good forecasting performance. While the good properties of the LARS algorithm have been well documented for cross-sectional data, this paper present its extension to the time series setting. The obtained results are very promising and might generate a new stream of research on forecasting with high dimensional time series.

Acknowledgements: This research was supported by the K.U.Leuven Research Fund and the Fonds voor Wetenschappelijk Onderzoek (Contract number G.0594.05).

Appendix

Proof of equation (5). The left hand side of equation (4) can be rewritten as

$$R^2(z_0 - \gamma\hat{z}_0 \sim \underline{x}_{(1)}) = 1 - \frac{(z_0 - \gamma\hat{z}_0)'(I - H_{(1)})(z_0 - \gamma\hat{z}_0)}{(z_0 - \gamma\hat{z}_0)'(z_0 - \gamma\hat{z}_0)}\,.$$

with

$$\begin{aligned}(z_0 - \gamma\hat{z}_0)'(I - H_{(1)})(z_0 - \gamma\hat{z}_0) &= (z_0 - \gamma H_{(1)}z_0)'(I - H_{(1)})(z_0 - \gamma H_{(1)}z_0) \\ &= z_0'(I - \gamma H_{(1)})(I - H_{(1)})(I - \gamma H_{(1)})z_0 \\ &= z_0'(I - H_{(1)})z_0, \qquad \text{(A.14)}\end{aligned}$$

where we used the well known properties $H_{(1)}' = H_{(1)}$ and $H_{(1)}H_{(1)} = H_{(1)}$ of the projection matrix. Note that the value of (A.14) does not depend on γ anymore. The right hand side of equation (4) equals

$$R^2(z_0 - \gamma\hat{z}_0 \sim \underline{x}_j) = 1 - \frac{(z_0 - \gamma\hat{z}_0)'(I - H_j)(z_0 - \gamma\hat{z}_0)}{(z_0 - \gamma\hat{z}_0)'(z_0 - \gamma\hat{z}_0)},$$

with

$$\begin{aligned}(z_0 - \gamma\hat{z}_0)'(I - H_j)(z_0 - \gamma\hat{z}_0) &= (z_0 - \gamma H_{(1)}z_0)'(I - H_j)(z_0 - \gamma H_{(1)}z_0) \quad \text{(A.15)} \\ &= z_0'[(I - \gamma H_{(1)})(I - H_j)](I - \gamma H_{(1)})z_0 \\ &= z_0'\big(I - H_j + (H_{(1)}H_j + H_jH_{(1)} - 2H_{(1)})\gamma \\ &\quad + (H_{(1)} - H_{(1)}H_jH_{(1)})\gamma^2\big)z_0\,. \qquad \text{(A.16)}\end{aligned}$$

So equation (4) holds if and only if (A.14) and (A.16) are equal to each other. This results in the following quadratic equation (5) for γ.

Finally, note that (5) will always have a root between zero and one, and this for every j not in the active set. Indeed, for $\gamma = 0$ we have $R^2(z_0 \sim \underline{x}_{(1)}) \geq R^2(z_0 \sim \underline{x}_j)$, by definition of the first index in the active set. On the other hand, for $\gamma = 1$ we have $0 = R^2(z_0 - \hat{z}_0 \sim \underline{x}_{(1)}) \leq R^2(z_0 - \hat{z}_0 \sim \underline{x}_j)$. Hence, there must exist a γ between zero and one for which condition (4), and then also condition (5) holds.

Proof of equation (6). Recall that $\hat{z}_0 = H_{(1)}z_0$ and $\tilde{x}_{(1)} = \hat{z}_0/s_1$. The following relations hold:

$$z_0'H_{(1)}z_0 = z_0'H_{(1)}H_{(1)}z_0 = \hat{z}_0'\hat{z}_0 = (T-1)s_1^2 \tag{A.17}$$

$$\begin{aligned} z_0'(H_{(1)}H_j + H_jH_{(1)} - 2H_{(1)})z_0 &= \hat{z}_0'H_jz_0 + z_0'H_j\hat{z}_0 - 2\hat{z}_0'H_{(1)}\hat{z}_0' \\ &= 2(s_1\tilde{x}_{(1)}'H_jz_0 - \hat{z}_0'\hat{z}_0') \\ &= 2s_1(\tilde{x}_{(1)}'H_jz_0 - (T-1)s_1) \end{aligned} \tag{A.18}$$

$$\begin{aligned} z_0'(H_{(1)} - H_{(1)}H_jH_{(1)})z_0 &= \hat{z}_0'\hat{z}_0 - \hat{z}_0'H_j\hat{z}_0 \\ &= s_1^2((T-1) - \tilde{x}_{(1)}'H_j\,\tilde{x}_{(1)}) \end{aligned} \tag{A.19}$$

Inserting relations (A.17), (A.18) and (A.19) in (5) results in (6).

Proof of Lemma 1. The three statements of the lemma will be proven by induction. For $k = 1$, we know that

$$r_1 = \mathrm{Cor}(z_0, \tilde{x}_{(1)}) = \mathrm{Cor}(z_0, \hat{z}_0) \geq 0\,,$$

so (a) holds for $k = 1$. Since the index (1) by construction yields the largest $R^2(z_0 \sim \underline{x}_j)$, (b) holds for $k = 1$. For proving (c) for $k = 1$, note that we already showed that (5) always has a solution between zero and one. Every solution of (5) multiplied by s_1 then solves (6) for $k = 1$, since

$$R^2(z_0 - \gamma\hat{z}_0 \sim \underline{x}_{(1)}) = R^2(z_0 - \gamma s_1\tilde{x}_{(1)} \sim \underline{x}_{(1)}) = R^2(z_0 - \gamma s_1\tilde{x}_{(1)} \sim \tilde{x}_{(1)})\,.$$

From $a_1 = 1$, since $\tilde{x}_{(1)}$ is the equiangular vector in the first step, and from $r_1 = s_1$, since z_0 has unit variance, it follows that (c) holds for $k = 1$.

Suppose now that the three statements of Lemma 1 hold for $k-1$. Now we will prove that they also hold for step k. First of all, we have for every i in $1, \ldots, k-1$

$$\begin{aligned} \mathrm{Cor}(z_{k-1}, \tilde{x}_{(i)}) &= \mathrm{Cov}(z_{k-1}, \tilde{x}_{(i)}) \\ &= \mathrm{Cov}(z_{k-2} - \gamma_{k-1}u_{k-1}, \tilde{x}_{(i)}) \\ &= r_{k-1} - \gamma_{k-1}a_{k-1} \geq 0\,, \end{aligned} \tag{A.20}$$

which does not depend on i. Here, we used (a) for $k-1$, and (7). The inequality holds since (c) is assumed to hold for $k-1$. Furthermore, since condition (9) holds in step $k-1$, it follows that

$$R^2(z_{k-1} \sim \tilde{x}_{(k-1)}) = R^2(z_{k-1} \sim \underline{x}_{(k)}) \quad \text{or} \quad \mathrm{Cor}^2(z_{k-1} \sim \tilde{x}_{(k-1)}) = \mathrm{Cor}^2(z_{k-1} \sim \tilde{x}_{(k)})\,.$$

Since $\tilde{x}_{(k)}$ is proportional to the fitted values in the OLS regression of z_{k-1} on $\underline{x}_{(k)}$, one also has $\mathrm{Cor}(z_{k-1}, \tilde{x}_{(k-1)}) = \mathrm{Cor}(z_{k-1}, \tilde{x}_{(k)}) \geq 0$. We conclude that (a) holds for step k.

To prove (b), suppose that there exists a j not belonging to the active set in step k such that

$$r_k^2 = R^2(z_{k-1} \sim \tilde{x}_{(k)}) < R^2(z_{k-1} \sim \underline{x}_j)\,. \tag{A.21}$$

We use the shortened notations $f_0(\gamma) = R^2(z_{k-2} - \gamma u_{k-1} \sim \tilde{x}_{(k-1)})$, for the left hand side of condition (9) at step $k-1$, and $f_j(\gamma) = R^2(z_{k-2} - \gamma u_{k-1} \sim \underline{x}_j)$, for the right hand side of condition (9) at step $k-1$. By definition, γ_{k-1} is the first crossing of the two curves $f_{(k)}(\gamma)$ and $f_0(\gamma)$, for $\gamma > 0$. Since (b) is assumed to be hold at step $k-1$, we have $f_0(0) \geq f_j(0)$. But definition (8) and (A.21) imply that

$$f_0(\gamma_{k-1}) = f_{(k-1)}(\gamma_k) < f_j(\gamma_{k-1})\,.$$

Hence, there must exist a crossing of $f_0(\gamma)$ and $f_j(\gamma)$ at a $\tilde{\gamma}$ strictly smaller than γ_{k-1}. This contradicts the definition of γ_{k-1}, so (A.21) can not hold, proving statement (b).

Finally, let us show that (c) holds at step k. It is sufficient to show that there always exists a solution of equation (9) in the interval $[0, r_k/a_k]$. Now note that for $\gamma = 0$, the left hand side of (9) is larger than the right hand side, since we have already proven (b) for step k. On the other hand, for $\gamma = r_k/a_k$, the left hand side of equation (9) equals 0 (as can be seen from equation (A.20)), which is of course smaller than the right hand side of equation (9). Hence, there exists a positive solution, smaller than r_k/a_k to condition (9).

Proof of equation (11). The left hand side of equation (9) can be rewritten as

$$\begin{aligned} R^2(z_{k-1} - \gamma u_k \sim \tilde{x}_{(k)}) &= \mathrm{Cor}^2(z_{k-1} - \gamma u_k, \tilde{x}_{(k)}) \\ &= \frac{(\mathrm{Cov}(z_{k-1}, \tilde{x}_{(k)}) - \gamma \mathrm{Cov}(u_k, \tilde{x}_k))^2}{\mathrm{Var}(z_{k-1} - \gamma u_k)} \\ &= \frac{(r_k - \gamma a_k)^2}{\mathrm{Var}(z_{k-1} - \gamma u_k)} \,. \end{aligned}$$

We used here equation (7) and the fact that $\tilde{x}_{(k)}$ has been standardized. The right hand side of condition (9) can be written as

$$\begin{aligned} R^2(z_{k-1} - \gamma u_k \sim \underline{x}_j) &= 1 - \frac{(z_{k-1} - \gamma u_k)'(I - H_j)(z_{k-1} - \gamma u_k)}{(z_{k-1} - \gamma u_k)'(z_{k-1} - \gamma u_k)} \\ &= \frac{(z_{k-1} - \gamma u_k)' H_j (z_{k-1} - \gamma u_k)}{(z_{k-1} - \gamma u_k)'(z_{k-1} - \gamma u_k)} \,, \end{aligned}$$

where j does not belong to the active set. So condition (9) is equivalent to

$$\begin{aligned} \frac{(r_k - \gamma a_k)^2}{\frac{1}{T-1}(z_{k-1} - \gamma u_k)'(z_{k-1} - \gamma u_k)} &= \frac{(z_{k-1} - \gamma u_k)' H_j (z_{k-1} - \gamma u_k)}{(z_{k-1} - \gamma u_k)'(z_{k-1} - \gamma u_k)} \\ \Leftrightarrow (T-1)(r_k - \gamma a_k)^2 &= (z_{k-1} - \gamma u_k)' H_j (z_{k-1} - \gamma u_k) \\ \Leftrightarrow (T-1)(r_k^2 - 2a_k r_k \gamma + a_k^2 \gamma^2) &= z'_{k-1} H_j z_{k-1} - 2 z'_{k-1} H_j u_k \gamma + u'_k H_j u_k \gamma^2 \,. \end{aligned}$$

After rearranging the terms, equation (11) follows.

References

Bai, J. and Ng, S. (2002), "Determining the Number of Factors in Approximate Factor Models," *Econometrica*, 70, 191–221.

— (2007), "Determining the Number of Primitive Shocks in Factor Models," *Journal of Business and Economic Statistics*, 25, 52–60.

Bovelstad, H.; Nygard, S.; Storvold, H.; Aldrin, M.; Borgan, O.; Frigessi, A. and Lingjaerde, O. (2007), "Predicting Survival from Microarray Data, a Comparative Study," *Bioinformatics*, 23, 2080–2087.

Breiman, L. (1995), "Better subset regression using the nonnegative garrote," *Technometrics*, 37, 373–384.

Dante, A. and Watson, M. (2007), "Consistent Estimation of the Number of Dynamic Factors in a Large N and T Panel," *Journal of Business and Economic Statistics*, 25, 91–96.

de Poorter, M.; Ravazzolo, F. and van Dijk, D. (2007), "Predicting the Term Sturcture of Interest Rates," *Working Paper*.

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

Dielman, T. (2001), *Applied Regression Analysis*, Duxbury Thomson Learning.

Efron, B.; Hastie, T.; Johnstone, I. and Tishirani, R. (2004), "Least Angle Regression," *The Annals of Statistics*, 32, 407–451.

Ekkund, J. and Karlsson, S. (2007), "Forecast combination and model averaging using predictive measures," *Econometric Review*, 26, 329–363.

Forni, M.; Lippi, M. and Reichlin, L. (2005), "The Generalized Dynamic Factor Model: One-Sided Estimation and Forecasting," *Journal of the American Statistical Association*, 100, 830–840.

Heij, C.; Groenen, P. and van Dijk, D. (2007), "Forecast comparison of principal component regression and principal covariate regression," *Computational Statistics and Data Analysis*, 51, 3612–3625.

Hendry, D. and Krozlig, H. (2001), *Automatic Econometric Model Selection*, London: Timberlake Consultants Press.

Hoover, K. and Krolzig, H. (1999), "Data Mining Reconsidered: Encompassing and the General-to-specific Approach to Specification Search," *Econometrics Journal*, 2, 167–191.

Keerthi, S. and Shevade, S. (2007), "A fast tracking algorithm for generalized LARS/LASSO," *IEEE Transactions on Neural Networks*, 18, 1826–1830.

Khan, J.; Van Aelst, S. and Zamar, R. (2007), "Robust Linear Model Selection Based on Least Angle Regression," *Journal of the American Statistical Association*, 102, 1289–1299.

McCann, L. and Welsch, R. (2007), "Robust variable selection using least angele regression and elemental set sampling," *Computational Statistics and Data Analysis*, 52, 1289–1299.

Meier, L.; van de Geer, S. and Buhlmann, P. (2008), "The Group LASSO for Logistic Regression," *Journal of the Royal Statistical Society, Series B*, 70, 53–71.

Qian, G. and Zhao, X. (2007), "On time series model selection involving many candidate ARMA models," *Computational Statistics and Data Analysis*, 51, 6180–6196.

Saigo, H.; Uno, T. and Tsuda, K. (2007), "Mining Complex Genotypic Features for Predicting HIV-1 Drug Resistance," *Bioinformatics*, 23, 2455–2462.

Simila, T. and Tikka, J. (2007), "Input selection and shrinkage in multiresponse linear regression," *Computational Statistics and Data Analysis*, 52, 406–422.

Stock, J. and Watson, M. (2002a), "Forecasting Using Principal Components From a Large Number of Predictors," *Journal of the American Statistical Association*, 97, 1167–1179.

— (2002b), "Macroeconomic Forecasting Using Diffusion Indexes," *Journal of Business and Economics Statistics*, 20, 147–162.

— (2006), "Forecasting with many predictors," *Handbook of Economic Forecasting, Elliott, G., Grange, C.W.J., Timmermann, A.*, 515–554.

Stock, J. and Watson, W. (2005), "Implications of Dynamic Factors for VAR Analysis," *Working Paper*.

Sulkava, M.; Tikka, J. and Hollmen, J. (2006), "Sparse regression for analyzing the development of foliar nutrient concentrations in coniferous trees," *Ecological Modelling*, 191, 118–130.

Tibshirani, R. (1996), "Regression shrinkage and selection via the LASSO," *Journal of the Royal Statistical Society Series B*, 58, 267–288.

Yuan, M.; Joseph, V. and Lin, Y. (2007), "An efficient variable selection approach for analyzing designed experiments," *Technometrics*, 49, 430–439.

Yuan, M. and Lin, Y. (2006), "Model selection and estimation in regression with grouped variables," *Journal of the Royal Statistical Society Series B*, 68, 49–67.