Journal of Econometrics 146 (2008) 304–317

Contents lists available at ScienceDirect

Journal of Econometrics

journal homepage: www.elsevier.com/locate/jeconom

Forecasting economic time series using targeted predictors

Jushan Bai [a,b], Serena Ng [c,*]

[a] Department of Economics, NYU, 19 W 4th Street, New York, NY 10012, United States
[b] School of Economics and Management, Tsinghua University, Beijing, China
[c] Department of Economics, Columbia University, 420 W. 118 St., New York 10027, United States

ARTICLE INFO

Article history:
Available online 28 August 2008

Keywords:
Diffusion index
Factor models
LASSO
LARS
Hard thresholding

ABSTRACT

This paper studies two refinements to the method of factor forecasting. First, we consider the method of quadratic principal components that allows the link function between the predictors and the factors to be non-linear. Second, the factors used in the forecasting equation are estimated in a way to take into account that the goal is to forecast a specific series. This is accomplished by applying the method of principal components to 'targeted predictors' selected using hard and soft thresholding rules. Our three main findings can be summarized as follows. First, we find improvements at all forecast horizons over the current diffusion index forecasts by estimating the factors using fewer but informative predictors. Allowing for non-linearity often leads to additional gains. Second, forecasting the volatile one month ahead inflation warrants a high degree of targeting to screen out the noisy predictors. A handful of variables, notably relating to housing starts and interest rates, are found to have systematic predictive power for inflation at all horizons. Third, the targeted predictors selected by both soft and hard thresholding changes with the forecast horizon and the sample period. Holding the set of predictors fixed as is the current practice of factor forecasting is unnecessarily restrictive.

1. Introduction

In recent years, the method of "diffusion index forecasts", also known as factor augmented forecasts, has received the attention of both econometricians and practitioners. In the diffusion index forecasting methodology, the factors are first estimated from a large number of predictors, $(X_{1t}, \ldots, X_{Nt})$, by the method of principal components, and then augmented to a linear forecasting equation for y_{t+h} that consists of lags of y and other predictors. What makes the diffusion index (DI) methodology appealing is its capacity to incorporate information in a large number of predictors into the forecast in a simple and parsimonious way. However, this does not preclude refinements to the DI methodology. In particular, the methodology as it stands does not take into account the predictive ability of X_{it} for y_{t+h} when the factors are estimated. Furthermore, the framework is now confined to a linear relation between the predictors and the series to be forecasted.

Our goal is first to go beyond the linear principal components framework to permit a more flexible factor structure, and more importantly, to use only those predictors informative for y in estimating the factors. To this end, we consider two possible improvements to the DI framework. First, we allow the factors to be non-linearly related to the predictors by expanding the set of predictors to include non-linear functions of the observed variables. Second, and independently of whether non-linearity is being considered, we take explicit account that the object of interest is ultimately the forecast of y. Accordingly, we form principal components using a subset of those predictors that are tested to have predictive power for y. As this set of predictors change with y, we refer to these as 'targeted predictors'. Our approach therefore entertains more predictors than the current implementation of DI, but we will, in general, use fewer predictors to estimate the factors than the existing implementation of factor forecasting.

The primary focus of our analysis is how to reduce the influence of uninformative predictors for y within the confines of DI framework. We use 'hard' and 'soft' thresholding to determine which variables the factors are to be extracted from. The factors are the diffusion indices of the forecasting equation. Under hard thresholding, subset variable selection based on some pretest procedure is used to decide whether a predictor is 'in' or 'out'. Under soft thresholding, the top ranked predictors are kept, where the ordering of the predictors depends on the particular soft-thresholding rule used.

We consider the LASSO and the Elastic Net soft-thresholding rules, which are special cases of the 'Least Angle Regression' (LARS) algorithm developed in Efron et al. (2004). These soft thresholding methods have been used in biostatistics to study whether groups

* Corresponding author.
E-mail addresses: Jushan.Bai@nyu.edu (J. Bai), serena.ng@columbia.edu (S. Ng).

0304-4076/$ – see front matter © 2008 Elsevier B.V. All rights reserved.
doi:10.1016/j.jeconom.2008.08.010

of genes in a DNA microarray can be used to predict if a certain outcome (such as prostate cancer) occurs. Donoho and Johnstone (1994) provided many optimality results for soft-thresholding and showed that LASSO asymptotically comes close to being an ideal subset selector in terms of its function as an oracle. However, most of the theoretical and empirical analysis we are aware of assume iid data design. We are interested in the usefulness of soft-thresholding from the point of view of factor forecasting, which raises two specific issues. First, economic data are generally weakly dependent data and not iid, and it is not known how these methods perform. Furthermore, subset variable selection is only our intermediate object of interest, as ultimately, it is how the ordered predictors affect the factor estimates that determine forecast efficiency. To our knowledge, this is a new use of the soft-thresholding methodology. As we will see below, the results are encouraging.

Our primary evaluation of the different methods will be based on forecasts of inflation at different horizons and over different samples. The decision to focus on inflation is due in part to the fact that inflation forecasts are important to decision making for both private agents and government agencies. Inflation is chosen for this study also because inflation is well documented to be a challenging series to forecast, see, for example, Stock and Watson (1999). In particular, the reduction in mean-squared inflation forecast error from using the DI methodology has been found to be much smaller than forecasting real variables such as industrial production, see Boivin and Ng (2005). Testing the methodologies on inflation thus puts these methods considered to a non-trivial challenge. Additional results for other series support the main finding from analyzing inflation data that we can push the efficiency of DI forecasting one step further simply by forming the diffusion indices from targeted predictors.

2. Preliminaries

Suppose we are given data on a large number of predictors $X_t = (X_{1t}, \ldots, X_{Nt})'$, $i = 1, \ldots, N$ and $t = 1, \ldots, T$. We are interested in forecasting y^h_{T+h}, the annualized value of the variable y_t in period $T+h$. The precise definition of y^h_{t+h} depends on whether y_t is stationary or not. We will consider both possibilities and make precise the definition of y^h_{T+h} in the empirical section. Note that y can even be one of the predictors. If $N < T$, a forecast that makes use of all available predictors is $\hat{y}^h_{T+h|T} = \hat{\alpha}'W_T + \hat{\Gamma}'X_T$, where W_t is a vector of predetermined variables such as a constant and lags of y_{t+h}, $\hat{\alpha}$ and $\hat{\Gamma}$ are obtained from least squares estimation of

$$y^h_{t+h} = \alpha'W_t + \Gamma'X_t + \epsilon_{t+h}. \quad (1)$$

Although $\hat{\alpha}$ and $\hat{\Gamma}$ are $\sqrt{T}$ consistent, the mean-squared forecast error is increasing in N. Let $\hat{\sigma}^2_n = \hat{\epsilon}^{n\prime}\hat{\epsilon}^n/T$ be the sum of squared residuals from estimating a model with n predictors, divided by T. In principle, we can use information criteria of the form

$$N^* = \min_n \left[\log(\hat{\sigma}^2_n) + n\frac{C_T}{T}\right],$$

to select the optimal number of predictors. The FPE with $C_T = 2$ is designed specifically for forecasting, but the BIC with $C_T = \log T$ is also widely used. However, when the predictors have no natural ordering, in theory, there are 2^N sets of predictors to consider, rendering the procedure impractical.

The factor approach to an h period-ahead forecast is to estimate the forecasting equation using data for $t = 1, \ldots, T-h$:

$$y^h_{t+h} = \alpha'W_t + \beta(L)'\hat{f}_t + \epsilon_{t+h} \quad (2)$$

where $\hat{f}_t \subset \hat{F}_t$, $\beta(L)$ are coefficients pertaining to f_t and p of its lags, $\hat{F}_t$ are the principal component estimates of the $r \times 1$ vector F_t in the factor model

$$X_{it} = \lambda_i'F_t + e_{it} \quad (3)$$

or in matrix form

$$X_t = \Lambda F_t + e_t.$$

Eqs. (2) and (3) constitute the 'diffusion index' (DI) forecasting framework of Stock and Watson (2002). The DI forecast is $\hat{y}^h_{T+h|T} = \hat{\alpha}'W_T + \hat{\beta}'(L)\hat{f}_T$.

It is now understood that consistent estimation of the space spanned by F_t makes it possible to obtain $\sqrt{T}$ consistent estimates of α and β and $\min[\sqrt{N}, \sqrt{T}]$ consistent forecasts of the conditional mean, $y^h_{T+h|T}$, if $\sqrt{T}/N \to 0$ as $N, T \to \infty$. As shown in Bai and Ng (2006), we can treat $\hat{f}_t$ in the forecasting equation as though it is a vector of observed regressors. The forecasts generated by this methodology seem promising. Evaluations based on key macroeconomic variables find that the DI forecasts tend to do at least as well and often beat alternative methods such as forecast combination, empirical Bayes procedures, etc. See, for example, Stock and Watson (2006) and the references therein. An alternative method of factor forecasting, developed by Forni et al. (2005), also yield promising results. See Forni et al. (2001) and Boivin and Ng (2005).

One way of thinking about the DI methodology is that the factors provide a natural ranking for N mutually orthogonal linear combinations of X_t. If the bulk of the variation in X_t can be explained by a small number of these combinations, say, $N_{\max} << N$, the BIC or the FPE need only be evaluated $O(N_{\max})$ times (much smaller than 2^N) to arrive at f_t, the subset of F_t that best predicts y. As the principal component estimates of F_t are just linear combinations of X_t, the DI forecast can be written as

$$y^h_{t+h} = \alpha'W_t + \bar{\Gamma}'X_t + \epsilon_{t+h}$$

where $\bar{\Gamma}$ is a restricted version of Γ in (1). Viewed this way, the DI forecasts uses all N of the predictors in forecasting to the extent that $\bar{\Gamma}$ has no element that equals zero exactly.

In this paper, we consider refinements to the DI methodology using what we refer to as 'targeted diffusion index forecasts'. The thrust of the refinement is to target the factor estimates to the objective of forecasting y_t. More precisely, we seek a model

$$y^h_{t+h} = \alpha'W_t + \gamma'x_t + \epsilon_{t+h}$$

where the $k^* \times 1$ vector x_t is a subvector of X_t. Written differently,

$$y^h_{t+h} = \alpha'W_t + \Gamma^*X_t + \epsilon_{t+h} \quad (4)$$

where the vector Γ^* effectively puts a zero weight on those predictors that are not useful in forecasting y. We will propose two ways of defining Γ^*. Before turning to such an analysis, we introduce a generalization of the method of principal components which can be used whether or not the predictors are targeted.

2.1. Quadratic principal components

By the method of principal component, the estimates of F_t are linear combinations of X_{it} that minimize the sum of squared residuals of the linear model, $X_{it} = \lambda_i'F_t + e_{it}$. This presupposes a linear link function between the data and the latent factors. A more flexible approach is to consider a non-linear link function, $g(\cdot)$, such that

$$g(X_{it}) = \phi_i'J_t + e_{it},$$

where J_t are the common factors, and ϕ_i is the vector of factor loadings. Define X^*_t to be X_t augmented by some or all of the unique

cross-products of the elements of X_{it}, and let $X^* = (X_1^*, \ldots, X_T^*)$ be an $N^* \times T$ matrix. The second-order factor model is

$$X_t^* = \Phi J_t + e_t$$

where X_t^* is an $N^* \times 1$ vector. Estimation of J_t then proceeds by the usual method of principal components. If $X_t^* = \{X_{it}, X_{it}^2\}$, then $N^* = 2N$; we will henceforth refer to the procedure as SPC (squared principal components). In a previous version of this paper, the cross-product terms $X_{it}X_{jt}$, $i \neq j$ were also included, a method we referred to as QPC (quadratic principal components). The QPC is computationally demanding and was not noticeably better than the SPC. Results are therefore not included.

It is noteworthy that $\hat{J}_t$ estimated from X_t^* is different from $\hat{K}_t$, where $\hat{K}_t$ are estimates of the factors in the model $X_{it}^2 = \psi_i' K_t + \eta_{it}$. Whereas $\hat{J}_t$ is a linear combination of the linear AND the quadratic terms of X_{it}, $\hat{K}_t$ is a linear combination of X_{it}^2. The latter is of interest when one is concerned with factors in the second moments of X_{it}. As our application concerns forecasting inflation, not its volatility, estimation of K_t will not be further considered.

Once the estimates of J_t are obtained using SPC, they are augmented to the forecasting equation in the same way as the standard DI. That is to say, $\hat{J}$ is the matrix of eigenvectors corresponding to the r largest eigenvalues of the $T \times T$ matrix $X^{*\prime}X^*$.[1] Note also that the forecasting equation remains linear whatever is g.

An alternative way of capturing non-linearity is to augment the forecasting equation to include functions of the factors. The simplest case, and one which we will refer to as PC^2 (squared factors), uses the following forecasting equation

$$y_{t+h}^h = \alpha' W_t + \beta_1' \hat{F}_t + \beta_2' \hat{F}_t^2 + \epsilon_{t+h}.$$

The PC^2 is conceptually distinct from SPC. While the PC^2 forecasting model allows the volatility of factors estimated by linear principal components to have predictive power for y, the SPC model allows the factors to be possibly non-linear functions of the predictors while maintaining a linear relation between the factors and y. Ludvigson and Ng (2007) found that the square of the first factor estimated from a set of financial factors (i.e., volatility of the first factor) is significant in the regression model for the mean excess returns. In contrast, factors estimated from the second moment of data (i.e., volatility factors) are much weaker predictors of excess returns.

Clearly, the expanded matrix X^* can be much higher dimension than X because of the quadratic terms. But inclusion of noisy predictors can potentially lead to inferior factor estimates. Consideration of quadratic principal components leads naturally to the issue of what predictors should be used to form the diffusion indices.

3. Targeted predictors

A practical question in DI forecasting is how much data are really needed? In Boivin and Ng (2006), it was found that expanding the sample size simply by adding data that bear little information about the factor components does not necessarily improve forecasts. We need to take into account the properties of the idiosyncratic errors when constructing principal components. When the data are too noisy, we can be better off throwing away some data even though they are available. Results in Stock and Watson (2004a) suggest that the weighted principal components, in the spirit of GLS, indeed provide better forecasts than the OLS-based principal component estimates.

As currently implemented, the factors are extracted from the same large data set, regardless of the series to be forecasted. It is conceivable that the series to be forecasted, y, is highly predictable by a subset of the N series, and this subset is different for different y. We now discuss two classes of procedures to isolate this subset of variables, which we call 'targeted predictors'.

3.1. Hard thresholding

The method of hard thresholding simply uses a statistical test to determine if the i-th predictor is significant without regard for the other predictors being considered. Boivin and Ng (2006) used the correlation coefficients of the errors to pick out the variables to be dropped. It thus exploits a particular hard-thresholding rule to decide which variables are to be used in factor analysis. However, the series to be forecasted was not taken into account.

Our implementation of hard thresholding is closest to Bair et al. (2006), who, like Boivin and Ng (2006), also suggested that the principal components estimated from a large group of variables (which in their analysis are genes) can be dominated by principal components estimated from a smaller set of predictors. They used the bivariate relation between y_{t+h} and X_{it} to screen the variables and referred to the resulting procedure as 'supervised principal components'.[2] 'Supervised learning' has been used to isolate out subsets of genes associated with certain disease when often, the number of genes (our N) is much larger than the cases (our T) under investigation. However, given the dependent nature of our data, our targeting (or supervising) cannot be based just on the bivariate relation between y_{t+h} and X_{it}. Instead, we need to consider this relation after controlling for other predictors W_t (such as lags of y_t) since a simple autoregressive forecast is always available as an alternative forecasting procedure. The details are:

a. For each $i = 1, \ldots, N$, perform a regression of y_t^h on W_{t-h} and X_{it-h}. In application, W_{t-h} includes a constant and four lags of y_t. Let t_i denote the t statistic associated with X_{it-h}.
b. Obtain a ranking of the marginal predictive power of X_{it} by sorting $|t_1|, |t_2|, \ldots, |t_N|$ in descending order.
c. Let k_α^* be the number of series whose $|t_i|$ exceeds a threshold significance level, α;
d. Let $x_t(\alpha) = (x_{[1]t}, \ldots, x_{[k_\alpha^*]t}')$ be the corresponding set of k_α^* targeted predictors. Estimate F_t from $x_t(\alpha)$ by the method of principal components to yield $\hat{F}_t$.
e. Estimate (2) using the BIC to select p and $\hat{f}_t \subset \hat{F}_t$.
f. The h period ahead forecast is $\hat{y}_{T+h|T}^h = \hat{\alpha}' W_T + \hat{\beta}'(L)\hat{f}_T$.

Instead of including W_{t-h} as regressors in step 1, an equivalent method is to perform regressions on $M_w y$ and $M_w X_i$, where M_w is the matrix that projects onto the space orthogonal to W, making $M_w y$ and $M_w X_i$ the residuals associated with these projections. An alternative to step (c) is to let k_α^* be N minus the smallest j such that $p_{[j]} \geq \frac{\alpha}{N-j+1}$, where $p_{[j]}$ is the j-th ordered p-value of the test. This Bonferroni-type procedure, due to Holm (1979), is more powerful and generally selects fewer variables than step (c) discussed above, but the top variables selected are quite similar and results will not be reported.

The above algorithm essentially uses only those variables whose marginal predictive power for y is significant at some prescribed level, α, in the factor analysis. After the targeted principal components are estimated, steps (d)–(f) are standard in the DI framework.

[1] In practice, the data are always demeaned and standardized before forming principal components.

[2] Another use of hard thresholding is 'bagging'. Inoue and Kilian (2008) orthogonalized the data on about 30 variables and used hard thresholding at each bootstrap sample to reduce forecasting variance.

4. Soft thresholding

Hard thresholding can be sensitive to small changes in the data because of the discreteness of the decision rule. Another drawback of selecting predictors one at a time is that it does not take into account the information in other predictors. We may end up selecting variables that are too 'similar'. It is well known that model averaging is effective only if we pool over variables that bear distinct information from each other.

We now consider 'soft thresholding' methods that perform subset selection and shrinkage simultaneously. In the context of (4), it estimates Γ^* and sets those elements corresponding to weak predictors to zero. It is in this sense that shrinkage and model selection are performed simultaneously. We now describe three procedures in this class.

4.1. LASSO

One way of dropping uninformative regressors is to use penalized regressions. Let RSS (α, β) be sum of squared residuals from a regression of y_{t+h}^h on *all* available regressors, W_t and X_{it}, $i = 1, \ldots, N$. The solution to

$$\min_{\beta,\alpha} \quad RSS + \lambda \sum_{j=1}^{N} \beta_j^2$$

for $0 \leq \lambda < \infty$ is the well-known ridge estimator that shrinks the least squares estimates of β_j towards zero. Note that $\sum_{j=1}^{n} \beta_j^2 = \|\beta\|_2^2$, the length of β given by the L_2 norm. By the nature of the L_2 penalty, the ridge estimates will almost never be zero exactly. In consequence, uninformative predictors can still inflate forecast error variance.

Consider replacing the L_2 penalty by an L_1 penalty $\|\beta\|_1 = \sum_{j=1}^{N} |\beta_j|$. The solution to

$$\min_{\beta,\alpha} \quad RSS + \lambda \sum_{j=1}^{N} |\beta_j|$$

is the LASSO estimator ('least absolute shrinkage selection operator') of Tibshirani (1996). The dual to this problem is

$$\min_{\beta,\alpha} \quad RSS \text{ subject to } \sum_{j=1}^{N} |\beta_j| \leq c$$

where the parameter $c \geq 0$ controls the amount of shrinkage. An important feature of the L_1 penalty is that some coefficient estimates can be exactly zero. The shrinkage under LASSO depends only on λ (or c) and the value of the unrestricted estimates, but not on the correlation of the predictors as is the case under ridge estimation. As shown in Fan and Li (2001), LASSO enjoys a 'sparsity property'; it estimates zero components of the true parameter vector exactly as zero with probability approaching one as the sample size increases. The asymptotic distribution of the estimator is the same whether or not the zero restrictions are imposed. LASSO thus possesses the 'oracle property' in the sense of Fan and Li (2001). That is, the asymptotic distribution of the estimator based on the overall model and the one based on the more parsimonious model coincide.

The LASSO estimator and the ridge estimator are special cases of bridge estimators which are solutions to

$$\min_{\beta} \quad RSS + \lambda \sum_{j} |\beta_j|^{\gamma}.$$

As discussed in Fu (1998), 'bridge' estimators have a Bayesian interpretation. The bridge penalty function $\sum_j |\beta_j|^{\gamma}$ can be thought of as the log prior distribution of the parameter vector, β. The prior distribution with $\gamma = 2$ is Gaussian, and the prior distribution with $\gamma = 1$ is a Laplace (or double exponential). A small γ favors models either with many parameters set to zero, or parameters with large absolute values from long tailed density. Large values of γ favor models with regression parameters of small but non-zero values from a normal like, or short tailed density. Mol et al. (2006) considered penalized regression models as an alternative to DI forecast and analyzed the problem from a Bayesian perspective. We stay within the DI framework and are interested in which regressors to use in the estimation of the factors.

If the regressors are orthogonal, the LASSO estimates, denoted $\tilde{\beta}$, are

$$\tilde{\beta}_i = \text{sign}\{\hat{\beta}_i\}(|\hat{\beta}_i| - \lambda/2)_+$$

where $\hat{\beta}_i$ is the unrestricted OLS estimate of β_i, $z_+ = z$ if $z > 0$ and 0 otherwise. Therefore, when the least squares coefficients are too small in absolute value, LASSO sets them to zero. Clearly, the LASSO estimate is a non-linear and non-differentiable function of the data. Fu (1998) proposed a shooting algorithm that iteratively solves for the LASSO estimates without using quadratic programming, but the method is unstable when $N > T$. Using convex theory, Osborne et al. (2000) showed that the solution path for $\tilde{\beta}$ is piecewise linear in c. More efficient algorithms are available by exploiting this feature. Our implementation of LASSO will be discussed below.

4.2. The elastic net

The LASSO estimator is an improvement over the ridge estimator when there are many zero coefficients in the true model, since the ridge estimator will never set the coefficients to zero exactly. However, LASSO is not without its drawback. Empirically, it seems that when there is high correlation in the predictors, LASSO is dominated by the ridge. Conceptually there are two problems as highlighted by Zou and Hastie (2005). First, if $N > T$, LASSO can select at most T variables. Second, if there is a group of variables with high pairwise coefficients, LASSO tends to select only one variable from the group and does not care which one. These concerns suggest that a convex combination of ridge and LASSO estimation might be desirable. The result is the 'elastic net' (EN) estimator of Zou and Hastie (2005).

The idea of the elastic net is to stretch the fishing net that retains all the 'big fish'. Like LASSO, the EN simultaneously shrinks the estimates and performs model selection. The LASSO penalty is convex, but not strictly convex. Strict convexity enforces the grouping effect so that predictors with similar properties will have similar coefficients. The EN objective function is

$$\min_{\beta} \quad RSS + \lambda_1 \sum_{j=1}^{N} |\beta_j| + \lambda_2 \sum_{j=1}^{N} \beta_j^2.$$

The EN penalty is thus a convex combination of the LASSO and the ridge penalty and is strictly convex when $\frac{\lambda_2}{\lambda_1+\lambda_2} > 0$. A computationally appealing property of the EN is that it can be reformulated as a LASSO problem and hence solved using algorithms for LASSO. To see this, define new variables (when the W variables are absent) as follows:

$$X^+ = (1+\lambda_2)^{-1/2} \begin{pmatrix} X \\ \sqrt{\lambda_2} I_N \end{pmatrix} \qquad y^+ = \begin{pmatrix} y \\ 0_N \end{pmatrix}.$$

Let $\gamma = \frac{\lambda_1}{\sqrt{1+\lambda_2}}$. Then the EN estimator can be reformulated as

$$\beta^{++} = \underset{\beta}{\text{argmin}}\, RSS^+ + \gamma \sum_{j=1}^{N} |\beta_j|$$

with RSS^+ is the sum of squared residuals from a regression of y^+ on X^+. To remove a double shrinkage effect (which is in both LASSO and ridge), the EN estimator that proposed by Zou and Hastie (2005) is $\beta^+ = (1+\lambda_2)\beta^{++}$. As will be discussed below, our main interest is not so much in the point estimates, but the ordering of variables provided by the EN. With this in mind, we now turn to the implementation of LASSO and EN.

4.3. Least angle regressions

A widely-used variable selection method is the forward selection regression whereby the $(k+1)$-th predictor is added to the 'in' set if it has the maximum correlation with the residual vector from the k-step. The residual vector is then projected on the remaining predictors and a new predictor is found. Forward selection regressions tend to be too aggressive in the sense of eliminating too many predictors correlated with the ones included. Another popular method is forward stagewise regression, which is more cautious than forward selection regressions as it takes smaller steps towards the final model. Briefly, if $\hat{\mu}_k$ is the current estimate of y with k predictors and $\hat{c} = X'(y - \hat{\mu}_k)$ is the 'current correlation' (assuming each column of X is standardized), there exists a j such that $|\hat{c}_j|$ is maximized. Consider the updating rule $\hat{\mu}_{k+1} = \hat{\mu}_k + \hat{\gamma}\,\text{sign}(\hat{c}_j)X_j$. Forward selection sets $\hat{\gamma} = |\hat{c}_j|$ whereas forward stagewise regression sets $\hat{\gamma}$ to a small constant. As we will see below LASSO uses yet another $\hat{\gamma}$ and replaces X_j by some other quantity.

Efron et al. (2004) showed that LASSO and forward stagewise regressions are in fact special cases of what is known as LARS, or least angle regressions. At each step, the $\hat{\gamma}$ in LARS is endogenously chosen so that the algorithm proceeds equiangularly between the variables in the most correlated set (hence the 'least angle direction') until the next variable is found. After k steps, there are k variables in the active set. If we end after k steps, we will have an active set of k predictors, or in other words, the coefficients corresponding to the remaining $N-k$ predictors will be set to zero. If we continue until $k = N$, we will have a set K of indices of predictors ordered according to when they join the active set. How many coefficients to set to zero is thus re-cast into a stopping rule for k.

Formally, the LARS algorithm begins at $\hat{\mu}_0 = 0$. Suppose $\hat{\mu}$ is the current estimate and let $\hat{c} = X'(y - \hat{\mu})$. Define K as the set of indices corresponding to variables with the largest absolute correlations,

$$\hat{C} = \max_j |\hat{c}_j| \qquad K = \{j : |\hat{c}_j| = |\hat{C}|\}.$$

Let $s_j = \text{sign}(\hat{c}_j)$ and define the active matrix corresponding to K as

$$X_K = (s_j x_j)_{j\in K}.$$

Let $G_K = X_K' X_K$ and $A_K = (1_K' G_K^{-1} 1_K)^{-1/2}$, where 1_K is a vector of ones equaling the size of K. A unit equiangular vector with columns of the active set matrix X_K can be defined as

$$u_K = X_K w_K, \qquad w_K = A_K G_K^{-1} 1_K, \qquad a_K = X' u_K,$$

so that $X_K' u_K = A_K 1_K$. LARS then updates $\hat{\mu}$ as

$$\hat{\mu}^{\text{new}} = \hat{\mu} + \hat{\gamma} u_K$$

where

$$\hat{\gamma} = \min_{j\in K^c}{}^{+} \left(\frac{\hat{C} - \hat{c}_j}{A_K - a_j}, \frac{\hat{C} + \hat{c}_j}{A_K + a_j} \right)$$

where the minimum is taken over only the positive components.

LARS has several advantages. First, it gives us a ranking of the predictors when the presence of other predictors are taken into account, which is unlike the case of hard thresholding. Second, the algorithm implicitly avoids strongly correlated predictors, since if one of the correlated predictors is already included, the new residual will have a low correlation with variables strongly correlated with the variable just included. Third, LARS is not as 'greedy' as forward regressions which, when a good direction is found, it exploits the direction to a maximum. Fourth, LARS is fast; the computation cost is of the same order as the usual OLS. Indeed, starting from zero, the LARS solution paths grow piecewise linearly in a predictable way.

Superficially, LASSO and LARS seem quite different. However, as also shown in Efron et al. (2004), LASSO is in fact a special case of LARS that imposes the sign restriction that, if $\hat{\beta}^k$ is the vector of estimates at the k-th step, the sign of $\hat{\beta}_j^k$ must agree with the sign $\hat{c}_j$ for those j in the active set. Variables in the active set that fail the sign restriction can be 'kicked out' of the active set under LASSO. Therefore, unlike LARS, the size of the active set under LASSO need not be monotonically increasing. Under LASSO, the tuning parameter, λ, determines the severity of the penalty and thus how many parameters are set to zero. The LARS implementation of LASSO turns the choice of this tuning parameter into the choice of k, or in other words, the size of the active set. To determine k^*, one can use an information criterion such as the BIC. That is, have

$$k^* = \underset{k}{\text{argmin}}\, \text{BIC}(k) = \log(\hat{\sigma}_k^2) + k\frac{\log T}{T}.$$

By choosing k^*, the BIC also sets coefficients on the k^*+1 to N predictors as ordered by LARS/LASSO/EN to zero. The BIC can be replaced by the AIC, which is very similar to generalized cross-validation (GCV), and which Efron et al. (2004) considered.[3]

Our interests in soft thresholding was initiated by the sparsity property arising from the L_1 penalty of LASSO. If the number of non-zero coefficients is indeed very small, then one can simply use this sparse set of predictors for forecasting. However, when the model structure is not sufficiently sparse, i.e., when k^* is not sufficiently small, one is again faced with the problem that including too many predictors will inject excess sampling variability to the forecasts. In this situation, it seems reasonable to resort to dimension reduction by constructing diffusion indices from a selected subset of significant predictors. This being the case, the particular aspect of soft thresholding that turns out to be more useful for us is the ordering of the predictors provided by the LARS, and/or LARS implementation of LASSO. Specifically, the ordered set of variables will be used to construct diffusion indices. Of course, if k^* is too small, the principal component estimates will be imprecise. As we cannot know a priori how big is k^*, we always estimate the factors using a fixed number of series but these series will be ordered according to LARS/LASSO. On the other hand, if k^* is in fact small, we also entertain a forecasting model that uses k^* non-standardized predictors directly without forming factors. But it should be kept in mind that this is not a DI forecast.

In the empirical analysis to follow, we consider three methods of using LARS-ordered variables: (i) estimate principal components $\hat{F}_t$ from the first 30 series that LARS/LASSO/EN select; (ii) enter the first five ($k = 5$) predictors to the forecasting equation directly; (iii) enter $k = k^*$ predictors to the forecasting equation directly. Put differently, (ii) and (iii) use a small number of selected variables as predictors (i.e., as our $\hat{F}_t$ without principal components analysis), while the $\hat{F}_t$ in (i) are principal components estimated

[3] Comments on the LARS article reflect concerns by many that the GCV will overfit. The BIC evaluates models using in-sample errors, but an out-of-sample variant can also be considered. As discussed below, k^* is not of primary importance given that we have to estimate the factors. For this reason, alternative methods for determining k^* was not pursued.

from the first 30 series selected by LARS. Then the BIC is used to determine p and the corresponding $\hat{f}_t$ (a subset of $\hat{F}_t$) that enters the forecasting equation (2). For (i), we use 30 series because our experience has been that for well-behaved data, the principal component estimates have reasonably good properties when the number of cross-section units exceed 30. As well, even the tightest hard-thresholding criterion tends to pick out over 30 series. We therefore want to see if the DI methodology works well with as few as 30 predictors. It is important to emphasize that (i) is a DI forecast and as such, neither the LARS/LASSO parameter estimates, nor k^* per se, will be used directly because the variables selected by LARS/LASSO enter the forecast only via the factors.

As we will see, some of our subsamples considered has $N > T$ and some has N in the same order as T, even though for the whole sample $T > N$. In practice, we always use the X^+ and Y^+ data matrices so that we effectively implement what amounts to LARS-EN and LASSO-EN, with the difference between LARS-EN and LASSO-EN being whether or not to impose a sign restriction. As a matter of notation, we simply refer to the three methods as LA(5), LA(PC), and LA(k^*) and the use of EN is implicit. The LASSO results can be similarly defined. We also use the LARS algorithm to produce an ordering of the variables when x_{it} and x_{it}^2 are included. A complete list of the methods considered is given below:

PC	f_t estimated from all 132 of the X_{it} available;
SPC	f_t estimated from all 132 of X_{it} and X_{it}^2 available;
TPC	f_t are the targeted principal components estimated from a subset of available X_{it} where the subset is selected by hard thresholding;
TSPC	f_t are the targeted principal components estimated from X_{it} and X_{it}^2;
TSTPC	f_t are targeted principal components estimated from a subset of X_{it} and X_{it}^2;
PC^2	f_t estimated from all 132 of the X_{it} available, $[f_t, f_t^2]$ used in the forecasting equation;
TPC^2	f_t estimated from a subset of X_{it} and X_{it}^2, $[f_t, f_t^2]$ used in the forecasting equation;
LA(5)	f_t is the vector of 5 best predictors selected by LARS;
LA(PC)	f_t are the factors estimated from the 30 best predictors in $\{x_{it}\}$ as selected by LARS;
LA(k^*)	f_t is the vector of k^* best predictors selected by LARS;
LA(SPC)	f_t are the factors estimated from the 30 best predictors in $\{x_{it}, x_{it}^2\}$ as selected by LARS.

5. Results

The variable we are interested in forecasting, y_{t+h}, is the logarithm of PUNEW, or CPI all items, using factors estimated from some or all of the 132 predictors and or their cross-products.[4] These are monthly time series available from 1960:1 to 2003:12 for a total of $T = 528$ observations. As argued by Nelson and Plosser (1982) and Beveridge and Nelson (1981), many of those series are I(1) non-stationary or contain an I(1) components, the data are therefore transformed by taking logs, first or second differences when necessary, as in Stock and Watson (2006). In particular, the logarithm of CPI is assumed to be integrated of order two. Following Stock and Watson (2002), define

$$y_{t+h}^h = \frac{1200}{h} \cdot (y_{t+h} - y_t) - 1200 \cdot (y_t - y_{t-1})$$

and let

$$z_t = 1200 \cdot (y_t - y_{t-1}) - 1200 \cdot (y_{t-1} - y_{t-2}).$$

For $h = 1, 6, 12$, and 24, the factor augmented forecast given information in time t is

$$\hat{y}_{t+h|t}^h = \hat{\alpha}_0 + \hat{\alpha}_1'(L) z_t + \hat{\beta}_1'(L) \hat{f}_t$$

where the number of lags of z_t and $\hat{f}_t$ are determined by the BIC with the maximum number of lags set to six when the sample size permits, and is reduced to four otherwise. In the notation of the preceding discussion, W_t consists of a constant and lags of z_t. To simplify notation, $\hat{f}_t$ generically denotes estimated factors used for forecasting, where $\hat{f}_t \subset \hat{F}_t$. In all cases, $\hat{F}_t$ is a 10×1 vector. That is, we select the factors used for forecasting from the first ten estimated factors. It is understood that $\hat{F}_t$ are estimated using different number of series. Although we are forecasting the change in inflation, we will continue to refer to the forecasts as inflation forecasts.

Our main interest is in figuring out n, the number of variables used to estimate F_t. Given that W_t are lags of inflation, the question more precisely phrased is what variables have predictive power for inflation after controlling for lags of inflation themselves. We do not restrict the optimal predictors to be the same for every time period. Instead, the predictors are selected at each t, and the forecasting equation is re-estimated after new factors are estimated. Our first estimation consists of ten years of data (120 data points) starting in 1960:3; the sample is extended one month at a time. There is one forecasting equation for each h. The last observation used in estimation when $h = 12$ is 2002:12. For each h, we have about 400 out-of-sample forecasts. We use the average of the forecast errors to evaluate the different procedures. We will refer to the ratio of the MSE for a given method to the MSE of an AR(4) as RMSE (relative mean-squared error). That is,

$$\text{RMSE(method)} = \frac{\text{MSE(method)}}{\text{MSE(AR(4))}}.$$

An entry less than one indicates that the specified method is superior to the simple AR(4) forecast.

Because parameter instability is prevalent in economic time series, a method that forecasts well in one sample is not guaranteed to forecast well in another sample period. Therefore, in addition to the full sample analysis, we also consider seven forecast subsamples: 70:3–80:12, 80:3–90:12, 90:3–00:12, 70:3–90:12, 70:3–00:12, 80:3–00:12, and 70:3–03:12. For example, in the case of 70:3–00:12, the first forecast of 70:3 is based on estimation up to 60:3–70:3-h. The last forecast is for 00:12, and it uses parameters estimated for the sample 60:3–00:12-h. Table A.1 provides summary statistics for both y^h and the level of inflation over these samples. Notably, the mean of inflation over the estimation sample can be higher or lower than the forecast sample. Table A.2 then shows that five of the seven forecast samples considered had decelerating inflation. Note that inflation is the most volatile when $h = 1$. This feature will help understand the results to follow.

5.1. Number of variables chosen

We use the t statistic and three cutoff points for hard thresholding:

1. hard thresholding, $|t| > 1.28$;
2. hard thresholding, $|t| > 1.65$;
3. hard thresholding, $|t| > 2.58$;

[4] The data are taken from Mark Watson's web site http://www.princeton.edu/~mwatson.

Table 1
Fraction of variables selected with frequency *freq*

	h	1	6	12	24	1	6	12	24
$t = 1.28$	*freq*	x_{it}				x_{it}^2			
	[0, .2]	0.311	0.311	0.205	0.258	0.515	0.295	0.265	0.242
	[.2, .4]	0.083	0.045	0.061	0.030	0.038	0.068	0.076	0.053
	[.4, .6]	0.114	0.023	0.053	0.038	0.098	0.061	0.068	0.091
	[.6, .8]	0.318	0.182	0.174	0.197	0.235	0.303	0.174	0.280
	[.8, 1.0]	0.174	0.439	0.508	0.477	0.114	0.273	0.417	0.333
$t = 1.65$	*freq*	x_{it}				x_{it}^2			
	[0, .2]	0.455	0.379	0.318	0.303	0.652	0.402	0.348	0.348
	[.2, .4]	0.091	0.008	0.023	0.030	0.045	0.038	0.045	0.023
	[.4, .6]	0.114	0.053	0.068	0.030	0.076	0.083	0.053	0.068
	[.6, .8]	0.242	0.167	0.182	0.242	0.189	0.258	0.250	0.326
	[.8, 1.0]	0.098	0.394	0.409	0.394	0.038	0.220	0.303	0.235
$t = 2.58$	*freq*	x_{it}				x_{it}^2			
	[0, .2]	0.742	0.447	0.439	0.386	0.886	0.568	0.530	0.477
	[.2, .4]	0.038	0.015	0.015	0.023	0.015	0.045	0.030	0.053
	[.4, .6]	0.114	0.053	0.068	0.023	0.068	0.098	0.068	0.061
	[.6, .8]	0.083	0.280	0.311	0.364	0.030	0.174	0.288	0.288
	[.8, 1.0]	0.023	0.205	0.167	0.205	0.000	0.114	0.083	0.121
$\lambda_2 = .5$	*freq*	x_{it}				x_{it}^2			
	[0, .2]	0.992	0.947	0.932	0.879	0.992	0.958	0.943	0.928
	[.2, .4]	0.000	0.023	0.023	0.045	0.004	0.027	0.011	0.030
	[.4, .6]	0.000	0.015	0.023	0.045	0.000	0.004	0.015	0.023
	[.6, .8]	0.008	0.015	0.023	0.023	0.004	0.008	0.023	0.011
	[.8, 1.0]	0.000	0.000	0.000	0.008	0.000	0.004	0.008	0.008

The three threshold values are critical values of the t test at the two-tailed 10%, 5%, and 1% levels. For soft thresholding, Efron et al. (2004) found that LARS and LASSO tend to give extremely similar results, and this is also our experience with LARS-EN and LASSO-EN. The EN entails a choice λ_2 but the results are not very sensitive to λ_2. This is because we always form principal components from the first 30 series ordered by LARS. Thus, the DI forecasts do not strongly depend on k^*. To conserve space, we only report results for $\lambda_2 = (1.5, .5, .25)$.

For each t in 1970:1–2003:12-h, we keep track of whether each of the X_{it} is being selected as predictor. Under hard thresholding, variable i is 'in' if the t statistic associated with X_i in a regression of y_t^h on W_{t-h} and X_{it-h} exceeds a threshold. For LARS, we report the frequency that a variable is one of the first k^* variables in the ordered set. Naturally, the chosen set of predictors depends on the forecast horizon. Averaging over t gives us the average frequency that a predictor is 'in'. Table 1 reports this average frequency evaluated at five equally spaced bins. In other words, Table 1 tabulates how many of the 132 potential predictors is selected with frequency between 0 and .2, .2 and .4, and so forth.

Several features in Table 1 are noteworthy. First, the number of variables that exceeds the threshold increases with the forecast horizon, h. With the most liberal threshold of 1.28, .508 of the variables exceed the threshold in over 80% of the sample periods considered when $h = 12$. But this frequency falls to only .174 when $h = 1$. Inevitably, the higher the threshold, the less often the variables fall into the .8–1 frequency range. When $h = 1$ and under the tightest threshold of 2.58, only .023 of the variables are significant with a frequency above .8. As h increases, more variables are significant more often. The results thus imply that there are fewer variables with good predictive power for short than for long horizon forecasts. The LARS is expected to select sparse models, which is evident from Table 1. In fact, LARS selects more parsimonious models than when the tightest of our hard threshold is used. The right panel of Table 1 shows the frequency that X_{it}^2 passes the threshold. Not surprisingly, fewer X_{it}^2 are systematically significant than X_{it}.

Table 2 lists the ten most frequently selected variables in all samples considered. The left panel are the top predictors used in linear principal components. The top squared-predictors are listed in the right panel. Since LARS orders the variables taking other predictors into account, the right panel reports the top predictors in $\{x_{it}, x_{it}^2\}$. Squared terms have a prefix '2'.

Variables that systematically predict short horizon forecasts are different from those for long horizon forecasts. For hard thresholding, the interest rate variables (FYGT*) dominate the top ten list, while real variables such as purchasing manager's index (PMI), vendor's deliveries (PMDEL), new orders (PMNO), employment (LHU*) and housing (HS*) data follow. For $h = 12$, real M2 (FM2DQ) is the dominant variable irrespective of the threshold. However, the remaining variables in the top ten list are real variables like housing starts (HS*), employment/help wanted (CES* and LH*). Interest rate variables, which have high predictive power when $h = 1$, now ranked below 50 in terms of marginal predictive power.

The usual suspects also show up in list of systematic predictors selected by LARS, and these variables also differ by forecast horizon. The Fed Funds rate (FYFF), bond rates (FYGT*) the NAPM employment index (PMEMP), the purchasing manger index (PMI) along with the housing market variables have individual predictive power for inflation. The list is not markedly different from the one given by hard thresholding. Perhaps one difference is that LARS select variables from more diverse categories. As well, LARS favors related measures of inflation (which are in the top 20 and hence used in the factor analysis even though they are not the top 10 list), such as GMDC* (consumption deflators) and PU* (components of the CPI), while hard thresholding does not.

The right panel of Table 2 gives the top 10 series amongst the X_{it}^2 in terms of predictive power for inflation. The square of many variables found to be the top ten predictors by hard thresholding continues to be important. Of note is that financial variables, such as FSPXE (price earnings ratio), FSPCOM (composite stock market index) are also important. Under LARS, all variables are considered jointly. Thus, the squared variables compete with the first order variables for a good ranking. Surprisingly, the square of some interest rate variables are very highly ranked and even dominate variables that were highly ranked when the first order

Table 2
Most frequently selected predictors

	Rank	$h=1$	$h=6$	$h=12$	$h=24$	$h=1$	$h=6$	$h=12$	$h=24$
		x_{it}				x_{it}^2			
$t=1.28$	1	FYGT10	FYGT10	FM2DQ	FSPIN	HSMW	FSPXE	FSPXE	FSPXE
	2	FYGT5	FYGT5	HSBNE	FSPCOM	HSBNE	FSPCOM	FSDXP	FSDXP
	3	FYGT1	FM2DQ	HSMW	FM2DQ	PMEMP	FM2DQ	FSPIN	FSPIN
	4	FYGM6	A0M008	CES088	HSBMW	PMP	PMNO	FSPCOM	FSPCOM
	5	FMFBA	PMNO	CES053	HSBNE	PUXM	HSBNE	HSBNE	FM2DQ
	6	HSMW	HSBNE	CES048	HSMW	PMI	HSMW	HSMW	HSBMW
	7	LHU27	HSMW	CES046	a0m001	PMNO	LHU15	CES088	HSBNE
	8	HSBNE	CES155	CES002	CES048	HSBMW	LHUR	CES048	HSMW
	9	PMEMP	A0M048	LHU15	CES046	HSBR	FSDXP	CES003	a0m001
	10	PMP	CES053	LHUR	CES002	LHU26	FSPIN	LHUR	CES053
$t=1.96$	1	FYGT10	FYGT5	FM2DQ	FM2DQ	PMEMP	PMNO	HSMW	FM2DQ
	2	FYGT5	HSBNE	HSMW	HSMW	PMP	HSBNE	FSPCOM	HSMW
	3	FYGT1	HSMW	CES048	a0m001	HSMW	HSMW	LHUR	CES053
	4	FYGM6	A0M005	LHELX	CES046	HSBNE	LHUR	FSDXP	a0m001
	5	LHU27	LHU27	LHEL	A0M051	PMI	FSPXE	HSBNE	FSPCOM
	6	FMFBA	LHU15	A0M051	FSPIN	PMDEL	FSDXP	FSPXE	LHUR
	7	HSMW	LHELX	HSBNE	LHU15	PMNV	FSPCOM	HSBMW	FSDXP
	8	PMEMP	LHEL	A0M224R	HSBWST	HSBMW	LHU15	CES088	HSBWST
	9	FYGM3	A0M051	CES046	HSBMW	HSBR	HSBMW	FSPIN	HSBMW
	10	HSBNE	FYGT10	FYGT5	HSWST	PMNO	PMP	PMNO	HSWST
$t=2.32$	1	FYGT5	HSMW	FM2DQ	FM2DQ	PMEMP	HSMW	HSMW	HSWST
	2	FYGT10	LHELX	HSMW	HSWST	PMI	PMNO	HSBNE	HSBWST
	3	FYGT1	LHEL	HSBNE	HSBNE	CES015	HSBNE	PMNO	HSBNE
	4	FYGM6	LHU15	LHELX	LHELX	PMP	PMP	PMP	sfygm3
	5	FYFF	PMNO	LHEL	LHEL	PMNO	PMEMP	HSBMW	HSMW
	6	FYGM3	LHU27	LHU27	HSMW	CES017	LHU15	PMEMP	sFYGM6
	7	PMEMP	HSBNE	LHU15	LHU15	FSDXP	PMI	PMI	PMNO
	8	PMI	A0M051	LHUR	A0M051	FYAAAC	HSBMW	a0m001	HSBMW
	9	PMDEL	FYGT5	A0M051	A1M092	HSBMW	HSBR	FYAAAC	PMP
	10	LHU27	CES002	PMNO	PMNO	HSBR	HSBSOU	FYBAAC	FYBAAC
$\lambda_2=.5$	1	FYFF	LHEL	LHEL	FYBAAC	FYFF	LHEL	LHEL	LHEL
	2	FSPXE	FYFF	FYBAAC	LHEL	FSPXE	FYFF	FYBAAC	FYBAAC
	3	A0M027	FYBAAC	PMNO	HSBMW	2FYFF	FYBAAC	FYFF	2sFYGM6
	4	FSDXP	HSBR	HSBMW	HSBR	FSDXP	2HSBR	FSDXP	FYFF
	5	FYGT5	PMNO	HSBR	FYFF	2FYBAAC	PMNO	PMNO	2HSBMW
	6	IPS307	HSBMW	HSFR	HSFR	LHEL	2HSFR	2sFYGM6	2HSFR
	7	HHSNTN	HSFR	FSDXP	sFYGM6	FYGT5	2FYBAAC	2sfygm3	2HSBR
	8	PMP	FSDXP	FYFF	HSBNE	A0M027	HSBR	2PMI	2PMNO
	9	FYGT10	CP90	PMI	LHU27	LHU27	2sfygm3	2HSBR	2sFYGT5
	10	LHUR	LHU27	PMEMP	PMNO	IPS307	2A0m082	2HSBMW	2HSBNE

term was considered only, an example being the square of FYBAAC (2FYBAAC). Taken together, the results suggest that there is some additional information in quadratic variables that has not been exploited by the linear PC.

Table 3 gives information similar to Table 1, but highlight the dependence of the number of 'in' variables to the sample period. Table 3 reports the average n (for the intermediate hard thresholding of 1.65) and k^* (for EN-LARS with $\lambda_2 = .5$) noting that for PC, SPC, LA(PC), LA(SPC), n is fixed. Notably, n and k^* increases with h. Furthermore, k^* is always much smaller than n. For a given estimation sample, k^* ranges from 1 to 22 depending on the sample and the forecast horizon. At $h = 12$, k^* is below 10 on average. Only in rare occasions do k^* exceeds 15. In contrast, even under the tightest hard thresholding (not reported), it is not uncommon for n to be above 45 on average.

5.2. Forecast errors

The RMSE for three thresholds and three values of λ_2 are reported in Tables 4–7 for $h = 1, 6, 12$ and 24, respectively. A quick glance at the results reveal that most of the entries are below 1, showing that there are generally efficiency gains in doing factor forecasts. The RMSE falls with h irrespective of how targeting is done. This indicates that there is more to be gained from factor forecasting when the objective is annual rather than short horizon forecasts of inflation. Indeed, when $h = 1$, the RMSEs rarely fall below .9. Forecasting monthly inflation during 1970 and 1980 has been especially difficult for *all* methods. In fact, most of the methods do worse than the simple AR(4) forecast. Notably, supply shocks in this period are thought to be responsible for the high level and volatility of inflation and there were important changes in monetary policy. These uncertainties could be responsible for the failure of any state variable to have systematic predictive power for short-run inflation.

A tight (large value) threshold is desirable to eliminate the influence of noisy predictors. The gain from TPC over PC is expected to be smallest when the data are noisy and the threshold is loose. When $h = 1$, the RMSE is indeed largest with 1.28 as threshold. For $h = 6, 12, 24$, more predictors have genuine predictive power for inflation, and a threshold of 1.28 is adequate. Setting too high a threshold for annual inflation bears the cost of throwing away informative predictors. Setting the threshold to 1.65 seems to strike a pretty good balance.

For each h and each forecast sub-sample, the method that produces the smallest RMSE is highlighted. Notably, regardless of h and forecast horizon, the best forecast always involves some non-linear component, and most often obtained by using some form of SPC. The PC is dominated by some other method in every single case. For $h = 6$, 12, and 24, the LA(SPC) is systematically the best procedure, especially when $\lambda_2 = .25$ is used. The typical reduction in RMSE is around 10 basis points. Even though the best method

Table 3
Average number of variables selected, threshold = 1.65, $\lambda_2 = .5$

Sample	PC	SPC	TPC	TSPC	TSTPC	PC^2	TPC^2	LA(PC)	LA(10)	LA(k^*)	LA(SPC)
$h = 1$											
70.1–80.1	132	264	32.174	60.264	47.736	132	32.174	30	5	2.719	30
80.1–90.1	132	264	62.421	108.124	95.066	132	62.421	30	5	2.074	30
90.1–00.1	132	264	73.884	114.215	107.975	132	73.884	30	5	2.000	30
70.1–90.1	132	264	47.299	84.195	71.407	132	47.299	30	5	2.390	30
70.1–00.1	132	264	56.152	94.177	83.576	132	56.152	30	5	2.260	30
80.1–00.1	132	264	68.154	111.162	101.515	132	68.154	30	5	2.037	30
70.1–03.9	132	264	57.757	96.641	86.354	132	57.757	30	5	2.231	30
$h = 6$											
70.1–80.1	132	264	43.562	74.959	69.058	132	43.562	30	5	7.694	30
80.1–90.1	132	264	70.868	124.802	115.248	132	70.868	30	5	5.661	30
90.1–00.1	132	264	72.132	126.446	118.149	132	72.132	30	5	7.273	30
70.1–90.1	132	264	57.195	99.842	92.120	132	57.195	30	5	6.689	30
70.1–00.1	132	264	62.158	108.690	100.776	132	62.158	30	5	6.886	30
80.1–00.1	132	264	71.494	125.627	116.697	132	71.494	30	5	6.469	30
70.1–03.9	132	264	63.003	110.355	102.433	132	63.003	30	5	7.038	30
$h = 12$											
70.1–80.1	132	264	64.223	122.752	108.645	132	64.223	30	5	14.992	30
80.1–90.1	132	264	87.264	169.711	154.190	132	87.264	30	5	10.727	30
90.1–00.1	132	264	89.000	171.405	157.818	132	89.000	30	5	11.769	30
70.1–90.1	132	264	75.714	146.158	131.344	132	75.714	30	5	12.867	30
70.1–00.1	132	264	80.127	154.548	140.141	132	80.127	30	5	12.504	30
80.1–00.1	132	264	88.124	170.552	155.992	132	88.124	30	5	11.249	30
70.1–03.9	132	264	80.634	155.553	141.130	132	80.634	30	5	12.769	30
$h = 24$											
70.1–80.1	132	264	73.132	144.347	127.008	132	73.132	30	5	15.355	30
80.1–90.1	132	264	92.769	185.901	165.091	132	92.769	30	5	14.917	30
90.1–00.1	132	264	96.430	188.463	169.388	132	96.430	30	5	15.025	30
70.1–90.1	132	264	82.917	165.058	145.963	132	82.917	30	5	15.158	30
70.1–00.1	132	264	87.418	172.853	153.767	132	87.418	30	5	15.119	30
80.1–00.1	132	264	94.606	187.199	167.257	132	94.606	30	5	14.979	30
70.1–03.9	132	264	87.418	172.853	153.767	132	87.418	30	5	15.119	30

Table 4
RMSE, $h = 1$

Sample	PC	SPC	TPC	TSPC	TSTPC	PC^2	TPC^2	LA(PC)	LA(10)	LA(k^*)	LA(SPC)
	Threshold = 1.28							$\lambda_2 = .25$			
70.1–80.1	1.015	1.007	0.968	**0.927**	1.023	1.016	1.002	1.009	1.014	1.015	1.114
80.1–90.1	0.982	0.925	0.960	0.955	0.914	0.988	0.969	**0.877**	0.971	0.958	0.931
90.1–00.1	0.963	0.959	0.938	0.968	0.953	0.945	**0.923**	0.990	0.959	0.961	1.052
70.1–90.1	0.998	0.964	0.964	0.942	0.965	1.001	0.985	**0.938**	0.992	0.985	1.013
70.1–00.1	0.990	0.963	0.960	**0.947**	0.963	0.993	0.976	0.947	0.988	0.982	1.019
80.1–00.1	0.972	0.934	0.955	0.959	0.924	0.978	0.958	**0.906**	0.969	0.960	0.960
70.1–03.9	0.979	0.961	0.951	**0.937**	0.952	0.982	0.961	0.937	0.977	0.971	1.006
	Threshold = 1.65							$\lambda_2 = .5$			
70.1–80.1	1.015	1.007	**0.944**	0.974	1.001	1.016	0.970	1.036	1.014	1.015	1.001
80.1–90.1	0.982	0.925	0.960	0.934	0.930	0.988	0.891	0.875	0.971	0.958	**0.874**
90.1–00.1	0.963	0.959	0.990	0.983	0.984	**0.945**	0.961	1.003	0.959	0.959	0.981
70.1–90.1	0.998	0.964	0.953	0.948	0.963	1.001	**0.927**	0.949	0.992	0.985	0.932
70.1–00.1	0.990	0.963	0.960	0.951	0.964	0.993	**0.929**	0.959	0.988	0.982	0.938
80.1–00.1	0.972	0.934	0.969	0.943	0.939	0.978	0.902	0.908	0.969	0.959	**0.897**
70.1–03.9	0.979	0.961	0.960	0.938	0.950	0.982	**0.928**	0.950	0.977	0.970	0.936
	Threshold = 2.58							$\lambda_2 = 1.5$			
70.1–80.1	1.015	1.007	1.043	1.072	1.065	1.016	1.050	1.035	0.977	1.015	**0.951**
80.1–90.1	0.982	0.925	0.874	0.922	0.900	0.988	0.874	**0.873**	0.968	0.963	0.900
90.1–00.1	0.963	0.959	0.991	0.989	0.988	**0.945**	0.991	1.000	0.959	0.974	1.068
70.1–90.1	0.998	0.964	0.952	0.988	0.977	1.001	0.956	0.948	0.973	0.988	**0.924**
70.1–00.1	0.990	0.963	0.954	0.985	0.975	0.993	0.956	0.957	0.971	0.984	**0.945**
80.1–00.1	0.972	0.934	**0.895**	0.933	0.916	0.978	0.895	0.906	0.967	0.963	0.941
70.1–03.9	0.979	0.961	0.945	0.969	0.961	0.982	0.948	0.945	0.963	0.972	**0.941**

is sample dependent for $h = 1$, the LA(PC)/LA(SPC) is never far behind the best procedure.

Recall that LA(PC) is a diffusion index forecast with factors estimated from the 30 best predictors as determined by LARS. In contrast, LA(5) and LA(k^*) use the selected variables directly into (3), the forecasting equation. The finding that LA(PC) works better underscores the point that selecting the 'best' variables for forecasting is not enough. Information beyond the 'best' variables can improve prediction, and in our analysis anywhere between 5 to 29 additional variables can be useful (i.e., $30 - k^*$). But use of 30 predictors will introduce excess sampling variability, and the DI deals with this by using the factors to perform dimension

Table 5
RMSE, $h = 6$

Sample	PC	SPC	TPC	TSPC	TSTPC	PC^2	TPC^2	LA(PC)	LA(10)	LA(k^*)	LA(SPC)
	Threshold = 1.28							$\lambda_2 = .25$			
70.1–80.1	0.712	0.661	0.715	0.705	0.707	0.765	0.718	0.665	0.763	0.812	**0.653**
80.1–90.1	0.654	0.601	0.647	0.588	0.586	0.673	0.678	0.571	0.582	0.686	**0.543**
90.1–00.1	0.660	0.632	0.750	0.645	0.648	0.660	0.750	0.651	0.787	0.796	**0.609**
70.1–90.1	0.675	0.623	0.672	0.635	0.633	0.709	0.692	0.608	0.651	0.734	**0.585**
70.1–00.1	0.671	0.622	0.680	0.634	0.633	0.701	0.697	0.610	0.667	0.741	**0.587**
80.1–00.1	0.652	0.604	0.663	0.595	0.594	0.667	0.688	0.582	0.618	0.706	**0.554**
70.1–03.9	0.670	0.623	0.680	0.634	0.631	0.697	0.696	0.609	0.660	0.712	**0.587**
	Threshold = 1.65							$\lambda_2 = .5$			
70.1–80.1	0.712	**0.661**	0.721	0.744	0.707	0.765	0.711	0.688	0.739	0.797	0.676
80.1–90.1	0.654	0.601	0.660	0.579	0.590	0.673	0.675	0.590	0.604	0.699	**0.547**
90.1–00.1	0.660	0.632	0.656	0.647	0.652	0.660	0.656	0.666	0.757	0.872	**0.606**
70.1–90.1	0.675	0.623	0.682	0.644	0.637	0.709	0.689	0.629	0.655	0.736	**0.597**
70.1–00.1	0.671	0.622	0.677	0.642	0.636	0.701	0.683	0.631	0.667	0.752	**0.597**
80.1–00.1	0.652	0.604	0.656	0.588	0.598	0.667	0.668	0.600	0.631	0.731	**0.556**
70.1–03.9	0.670	0.623	0.678	0.641	0.635	0.697	0.681	0.626	0.664	0.722	**0.597**
	Threshold = 2.58							$\lambda_2 = 1.5$			
70.1–80.1	0.712	0.661	0.713	0.769	0.784	0.765	0.699	0.684	0.730	0.748	**0.659**
80.1–90.1	0.654	0.601	0.682	0.580	0.591	0.673	0.682	0.604	0.606	0.683	**0.552**
90.1–00.1	0.660	0.632	0.639	0.635	0.639	0.660	0.639	0.667	0.757	0.872	**0.626**
70.1–90.1	0.675	0.623	0.693	0.653	0.666	0.709	0.688	0.632	0.653	0.707	**0.594**
70.1–00.1	0.671	0.622	0.685	0.649	0.661	0.701	0.680	0.634	0.664	0.726	**0.597**
80.1–00.1	0.652	0.604	0.671	0.587	0.596	0.667	0.671	0.611	0.632	0.717	**0.564**
70.1–03.9	0.670	0.623	0.682	0.646	0.656	0.697	0.677	0.629	0.664	0.710	**0.598**

Table 6
RMSE, $h = 12$

Sample	PC	SPC	TPC	TSPC	TSTPC	PC^2	TPC^2	LA(PC)	LA(10)	LA(k^*)	LA(SPC)
	Threshold = 1.28							$\lambda_2 = .25$			
70.1–80.1	0.631	0.595	0.652	0.608	0.622	0.644	0.645	0.580	0.614	0.683	**0.523**
80.1–90.1	0.575	0.582	0.565	0.585	0.586	0.633	0.518	0.566	0.560	0.680	**0.482**
90.1–00.1	0.723	0.699	0.686	0.699	0.703	0.703	0.693	**0.623**	0.658	1.194	0.679
70.1–90.1	0.603	0.589	0.608	0.597	0.604	0.639	0.580	0.573	0.587	0.682	**0.502**
70.1–00.1	0.611	0.597	0.613	0.604	0.611	0.642	0.588	0.573	0.590	0.730	**0.516**
80.1–00.1	0.594	0.597	0.581	0.600	0.601	0.639	0.542	0.568	0.570	0.766	**0.510**
70.1–03.9	0.609	0.597	0.616	0.603	0.609	0.639	0.593	0.570	0.592	0.754	**0.517**
	Threshold = 1.65							$\lambda_2 = .5$			
70.1–80.1	0.631	0.595	0.659	0.654	0.636	0.644	0.612	0.599	0.623	0.691	**0.562**
80.1–90.1	0.575	0.582	0.689	0.573	0.591	0.633	0.661	0.569	0.566	0.702	**0.477**
90.1–00.1	0.723	0.699	0.616	0.698	0.703	0.703	**0.613**	0.681	0.665	1.088	0.675
70.1–90.1	0.603	0.589	0.675	0.613	0.613	0.639	0.638	0.584	0.594	0.698	**0.519**
70.1–00.1	0.611	0.597	0.665	0.618	0.619	0.642	0.631	0.590	0.597	0.733	**0.531**
80.1–00.1	0.594	0.597	0.669	0.589	0.605	0.639	0.644	0.583	0.576	0.764	**0.506**
70.1–03.9	0.609	0.597	0.665	0.615	0.616	0.639	0.632	0.587	0.597	0.751	**0.531**
	Threshold = 2.58							$\lambda_2 = 1.5$			
70.1–80.1	0.631	0.595	0.686	0.651	0.647	0.644	0.663	0.633	0.627	0.685	**0.591**
80.1–90.1	0.575	0.582	0.663	0.576	0.594	0.633	0.663	0.573	0.564	0.664	**0.514**
90.1–00.1	0.723	0.699	0.707	0.686	0.690	0.703	0.707	0.810	**0.643**	1.033	0.702
70.1–90.1	0.603	0.589	0.675	0.612	0.620	0.639	0.664	0.603	0.595	0.675	**0.552**
70.1–00.1	0.611	0.597	0.675	0.617	0.624	0.642	0.665	0.620	0.596	0.707	**0.564**
80.1–00.1	0.594	0.597	0.665	0.589	0.605	0.639	0.665	0.608	0.571	0.723	**0.542**
70.1–03.9	0.609	0.597	0.669	0.614	0.620	0.639	0.660	0.615	0.596	0.722	**0.564**

reduction. The results here show that forming diffusion indices from targeted predictors is a refinement to the DI methodology that is worthy of undertaking.

Our results thus far have been based on inflation. One may wonder if the approach of targeted principal components is always useful for forecasting series other than inflation, and if there might be cases when using targeted predictors without forming principal components might do just as well. To this end, we present the RMSE for the growth rate of four other series: personal income, retail sales, industrial production, and total employment. The log level of all four series are assumed to be differenced stationary. Accordingly, $y^h_{t+h} = \frac{1200}{h}(y_{t+h} - y_t)$ and $z_t = 1200(y_t - y_{t-1})$. The hard threshold is set at 1.65, while λ_2 is set to .25. These parameters have not been tuned to the data under investigation. Our main purpose is just to illustrate that the effectiveness of targeted predictors extends beyond forecasting inflation.

Table 8 report results for $h = 12$. Notably, for each of the series and for each of the samples, there is always a targeted DI forecast that outperforms the standard DI forecast (labeled PC). For personal income and retail sales, LA(5) and LA(k^*) often perform well, and in these two cases, targeted predictors directly is effective; there is no need to estimate targeted principal components because k^* is small enough (often between 5 and 10). However, for industrial production and employment, using targeted predictors directly would be undesirable. Targeted DI forecasts, especially hard thresholding, tend to perform better than PC. The results for these two series support the main finding

Table 7
RMSE, $h = 24$

Sample	PC	SPC	TPC	TSPC	TSTPC	PC^2	TPC^2	LA(PC)	LA(10)	LA(k^*)	LA(SPC)
	Threshold = 1.28							$\lambda_2 = .25$			
70.1–80.1	0.532	0.554	0.606	0.558	**0.474**	0.620	0.573	0.486	0.588	0.586	0.542
80.1–90.1	0.506	0.520	0.564	0.580	0.555	0.601	0.560	**0.442**	0.676	0.762	0.555
90.1–00.1	0.546	0.550	0.626	0.637	0.600	0.551	0.630	**0.447**	0.811	0.836	0.706
70.1–90.1	0.522	0.540	0.588	0.571	0.515	0.613	0.569	**0.467**	0.633	0.670	0.550
70.1–00.1	0.523	0.540	0.590	0.575	0.520	0.608	0.572	**0.464**	0.646	0.683	0.562
80.1–00.1	0.512	0.523	0.571	0.586	0.560	0.592	0.568	**0.441**	0.695	0.773	0.577
70.1–03.9	0.523	0.540	0.590	0.575	0.520	0.608	0.572	**0.464**	0.646	0.683	0.562
	Threshold = 1.65							$\lambda_2 = .5$			
70.1–80.1	0.532	0.554	0.501	0.542	0.545	0.620	0.553	**0.497**	0.607	0.593	0.609
80.1–90.1	0.506	0.520	0.561	0.528	0.497	0.601	0.561	**0.431**	0.696	0.715	0.564
90.1–00.1	0.546	0.550	0.630	0.529	0.552	0.551	0.630	**0.500**	0.832	0.954	0.756
70.1–90.1	0.522	0.540	0.531	0.538	0.525	0.613	0.559	**0.468**	0.652	0.652	0.590
70.1–00.1	0.523	0.540	0.538	0.537	0.525	0.608	0.563	**0.470**	0.666	0.675	0.602
80.1–00.1	0.512	0.523	0.569	0.527	0.502	0.592	0.569	**0.440**	0.716	0.751	0.592
70.1–03.9	0.523	0.540	0.538	0.537	0.525	0.608	0.563	**0.470**	0.666	0.675	0.602
	Threshold = 2.58							$\lambda_2 = 1.5$			
70.1–80.1	0.532	0.554	0.519	0.555	0.550	0.620	0.544	**0.544**	0.613	0.563	0.600
80.1–90.1	0.506	0.520	0.545	0.494	0.422	0.601	0.533	**0.460**	0.744	0.678	0.573
90.1–00.1	0.546	0.550	0.651	0.525	0.528	0.551	0.651	**0.521**	0.922	1.153	0.780
70.1–90.1	0.522	0.540	0.534	0.528	0.491	0.613	0.541	**0.507**	0.678	0.618	0.589
70.1–00.1	0.523	0.540	0.541	0.527	0.493	0.608	0.548	**0.507**	0.696	0.660	0.604
80.1–00.1	0.512	0.523	0.558	0.497	0.436	0.592	0.548	**0.468**	0.770	0.751	0.604
70.1–03.9	0.523	0.540	0.541	0.527	0.493	0.608	0.548	**0.507**	0.696	0.660	0.604

Table 8
RMSE, $h = 12$, other variables

Sample	PC	SPC	TPC	TSPC	TSTPC	PC^2	TPC^2	LA(PC)	LA(10)	LA(k^*)	LA(SPC)
Personal income a0m051											
70.1–80.1	0.545	0.422	0.465	0.461	0.439	0.563	0.447	0.461	**0.343**	0.401	0.372
80.1–90.1	0.902	0.986	0.944	0.973	1.103	0.846	0.933	**0.753**	0.966	1.054	1.073
90.1–00.1	1.106	0.984	1.096	1.242	1.226	1.122	1.172	1.171	1.019	**0.951**	1.175
70.1–90.1	0.673	0.623	0.635	0.645	0.676	0.666	0.620	0.566	**0.565**	0.632	0.621
70.1–00.1	0.796	0.722	0.766	0.813	0.831	0.796	0.777	0.738	**0.695**	0.724	0.777
80.1–00.1	1.012	0.982	1.027	1.114	1.167	0.994	1.062	**0.975**	0.999	1.003	1.127
70.1–03.9	0.817	**0.716**	0.782	0.791	0.839	0.803	0.783	0.743	0.719	0.749	0.776
Retail sales: a0m05											
70.1–80.1	0.620	0.676	0.633	0.687	0.682	0.615	0.637	0.674	**0.485**	0.502	0.637
80.1–90.1	0.559	0.582	**0.514**	0.564	0.569	0.590	0.520	0.594	0.713	0.675	0.586
90.1–00.1	1.158	1.032	1.142	**0.994**	1.197	1.164	1.144	1.154	1.163	1.043	1.332
70.1–90.1	0.601	0.650	0.601	0.654	0.651	0.606	0.606	0.654	**0.555**	0.556	0.623
70.1–00.1	0.716	0.725	0.713	0.722	0.762	0.722	0.717	0.757	0.678	**0.655**	0.770
80.1–00.1	0.840	0.785	0.808	**0.762**	0.860	0.860	0.812	0.857	0.916	0.843	0.936
70.1–03.9	0.726	0.729	0.723	0.725	0.771	0.732	0.727	0.765	0.685	**0.668**	0.780
Industrial production: ips10											
70.1–80.1	0.247	0.275	0.197	0.219	0.190	0.223	0.192	0.201	0.300	0.224	**0.144**
80.1–90.1	0.846	0.862	0.820	0.946	1.001	0.846	0.813	**0.692**	0.788	0.834	0.821
90.1–00.1	1.055	**0.974**	1.327	1.040	1.707	1.168	1.277	1.423	1.802	1.451	1.340
70.1–90.1	0.442	0.462	0.399	0.455	0.452	0.426	0.393	**0.359**	0.459	0.420	0.363
70.1–00.1	0.497	0.508	0.483	0.508	0.566	0.493	0.474	0.456	0.581	0.513	**0.452**
80.1–00.1	0.898	0.890	0.944	0.972	1.171	0.925	0.928	**0.866**	1.029	0.980	0.946
70.1–03.9	0.551	0.536	0.526	0.548	0.589	0.530	0.519	0.513	0.648	0.604	**0.498**
Total employment: ces002											
70.1–80.1	0.524	0.603	0.487	0.514	0.470	0.508	0.484	**0.383**	0.394	0.428	0.427
80.1–90.1	0.644	0.738	0.656	0.650	0.654	0.644	0.619	**0.591**	0.791	0.772	0.760
90.1–00.1	0.947	1.075	0.965	1.010	1.048	**0.930**	0.965	1.149	1.207	1.063	1.273
70.1–90.1	0.569	0.655	0.549	0.566	0.539	0.558	0.534	**0.459**	0.539	0.552	0.548
70.1–00.1	0.616	0.705	0.601	0.619	0.604	0.605	0.588	**0.545**	0.624	0.617	0.639
80.1–00.1	0.730	0.829	0.744	0.748	0.769	0.725	**0.717**	0.749	0.914	0.859	0.906
70.1–03.9	0.696	0.734	0.689	0.652	0.670	0.667	0.672	**0.609**	0.759	0.680	0.677

Threshold = 1.65, $\lambda_2 = .25$.

observed for inflation that estimating the factors from targeted predictors can yield more precise DI forecasts.

Why does targeting predictors produce better forecasts? As is evident from Tables 1–3, n and/or the composition of the n series vary with both the estimation sample and forecasting horizon h. In contrast, n is always 132 in the standard DI methodology regardless of h and the sample period. Allowing the number of series to change with the sample thus provides the targeted DI with additional flexibility to adapt to parameter instability in the data. As well, the role of non-linearity in DI forecasts is also sample

Table A.1
Summary statistics

Est. sample	$h=1$	$h=6$	$h=12$	$h=24$	Fcst sample	$h=1$	$h=6$	$h=12$	$h=24$
1: 60:3–80:12-h	4.867	4.808	4.748	4.690	70:3–80:12	7.218	7.393	7.515	7.803
2: 60:3–90:12-h	4.899	4.923	4.959	5.036	80:3–90:12	4.914	4.708	4.615	4.328
3: 60:3–00:12-h	4.396	4.416	4.444	4.520	90:3–00:12	2.836	2.831	2.761	2.694
4: 60:3–90:12-h	4.899	4.923	4.959	5.036	70:3–90:12	6.066	6.050	6.065	6.065
5: 60:3–00:12-h	4.396	4.416	4.444	4.520	70:3–00:12	4.989	4.977	4.964	4.941
6: 60:3–00:12-h	4.396	4.416	4.444	4.520	80:3–00:12	3.875	3.769	3.688	3.511
7: 60:3–03:12-h	4.220	4.281	4.376	4.520	70:3–03:12	4.680	4.739	4.790	4.941

Mean $\pi^h = \frac{1200}{h}(y_{t+h} - y_t)$ over estimation and forecast samples.

Table A.2
Summary statistics: y^h over forecast samples

	$h=1$		$h=6$		$h=12$		$h=24$	
Forecast sample	Mean	Std	Mean	Std	Mean	Std	Mean	Std
1: 70:3–80:12	0.089	3.683	0.264	2.916	0.387	3.020	0.674	3.555
2: 80:3–90:12	−0.047	3.063	−0.254	3.265	−0.346	3.086	−0.634	3.307
3: 90:3–00:12	−0.065	2.092	−0.070	1.846	−0.139	1.749	−0.207	1.834
4: 70:3–90:12	0.021	3.381	0.005	3.100	0.020	3.069	0.020	3.488
5: 70:3–00:12	−0.008	3.010	−0.020	2.744	−0.033	2.700	−0.055	3.037
6: 80:3–00:12	−0.056	2.618	−0.162	2.648	−0.243	2.505	−0.420	2.677
7: 70:3–03:12	−0.022	3.069	−0.025	2.747	−0.048	2.727	−0.055	3.037

Table A.3
Data sources, transformations, and definitions

Short name	Mnemonic	Fast or slow?	Tran	Description
PI	a0m052	S	Δln	Personal income (AR, Bil. Chain 2000 $) (TCB)
PI less transfers	a0m051	S	Δln	Personal income less transfer payments (AR, Bil. Chain 2000 $) (TCB)
Consumption	a0m224_r	S	Δln	Real consumption (AC) a0m224/gmdc (a0m224 is from TCB)
M&T sales	a0m057	S	Δln	Manufacturing and trade sales (Mil. Chain 1996 $) (TCB)
Retail sales	a0m059	S	Δln	Sales of retail stores (Mil. Chain 2000 $) (TCB)
IP: total	ips10	S	Δln	Industrial production index — Total index
IP: products	ips11	S	Δln	Industrial production index — Products, total
IP: final prod	ips299	S	Δln	Industrial production index — Final products
IP: cons gds	ips12	S	Δln	Industrial production index — Consumer goods
IP: cons dble	ips13	S	Δln	Industrial production index — Durable consumer goods
IP: cons nondble	ips18	S	Δln	Industrial production index — Non-durable consumer goods
IP: bus eqpt	ips25	S	Δln	Industrial production index — Business equipment
IP: matls	ips32	S	Δln	Industrial production index — Materials
IP: dble matls	ips34	S	Δln	Industrial production index — Durable goods materials
IP: non-dble matls	ips38	S	Δln	Industrial production index — Non-durable goods materials
IP: mfg	ips43	S	Δln	Industrial production index — Manufacturing (Sic)
IP: res util	ips307	S	Δln	Industrial production index — Residential utilities
IP: fuels	ips306	S	Δln	Industrial production index — Fuels
NAPM prodn	pmp	S	lv	Napm production index (%)
Cap util	a0m082	S	Δlv	Capacity utilization (Mfg) (TCB)
Help wanted indx	lhel	S	Δlv	Index of help-wanted advertising in newspapers (1967 = 100; Sa)
Help wanted/emp	lhelx	S	Δlv	Employment: Ratio; help-wanted Ads: No. unemployed Clf
Emp CPS total	lhem	S	Δln	Civilian labor force: Employed, total (Thous., Sa)
Emp CPS non-ag	lhnag	S	Δln	Civilian labor force: Employed, non-agric. industries (Thous., Sa)
U: all	lhur	S	Δlv	Unemployment rate: All workers, 16 years & over (%, Sa)
U: mean duration	lhu680	S	Δlv	Unemploy. By duration: Average (mean) duration in weeks (Sa)
U < 5 wks	lhu5	S	Δln	Unemploy. By duration: Persons unempl. less than 5 wks (Thous., Sa)
U 5–14 wks	lhu14	S	Δln	Unemploy. By duration: Persons unempl. 5–14 wks (Thous., Sa)
U 15+ wks	lhu15	S	Δln	Unemploy. By duration: Persons unempl. 15 wks + (Thous., Sa)
U 15–26 wks	lhu26	S	Δln	Unemploy. By duration: Persons unempl. 15–26 wks (Thous., Sa)
U 27+ wks	lhu27	S	Δln	Unemploy. By duration: Persons unempl. 27 wks + (Thous., Sa)
UI claims	a0m005	S	Δln	Average weekly initial claims, unemploy. Insurance (Thous.) (TCB)
Emp: total	ces002	S	Δln	Employees on non-farm payrolls: Total private
Emp: gds prod	ces003	S	Δln	Employees on non-farm payrolls — Goods-producing
Emp: mining	ces006	S	Δln	Employees on non-farm payrolls — Mining
Emp: const	ces011	S	Δln	Employees on non-farm payrolls — Construction
Emp: mfg	ces015	S	Δln	Employees on non-farm payrolls — Manufacturing
Emp: dble gds	ces017	S	Δln	Employees on non-farm payrolls — Durable goods
Emp: non-dbles	ces033	S	Δln	Employees on non-farm payrolls — Non-durable goods
Emp: services	ces046	S	Δln	Employees on non-farm payrolls — Service-providing
Emp: TTU	ces048	S	Δln	Employees on non-farm payrolls — Trade, transportation, and utilities
Emp: wholesale	ces049	S	Δln	Employees on non-farm payrolls — Wholesale trade
Emp: retail	ces053	S	Δln	Employees on non-farm payrolls — Retail trade
Emp: FIRE	ces088	S	Δln	Employees on non-farm payrolls — Financial Activities
Emp: Govt	ces140	S	Δln	Employees on non-farm payrolls — Government
Emp-hrs non-ag	a0m048	S	Δln	Employee hours in non-ag. establishments (AR, Bil. hours) (TCB)

(*continued on next page*)

Table A.3 (*continued*)

Short name	Mnemonic	Fast or slow?	Tran	Description
Avg hrs	ces151	S	lv	Avg weekly hrs of prod or non-sup workers on private non-farm payrolls — Goods-producing
Overtime: mfg	ces155	S	Δlv	Avg weekly hrs of prod or non-sup workers on private non-farm payrolls — Mfg overtime hours
Avg hrs: mfg	a0m001	S	lv	Average weekly hours, Mfg. (h) (TCB)
NAPM empl	pmemp	S	lv	Napm employment index (%)
Starts: non-farm	hsfr	S	ln	Housing starts: Non-farm (1947–58); Total farm & Non-farm (1959-) (Thous., Saar)
Starts: NE	hsne	F	ln	Housing starts: Northeast (Thous. U.) S.A.
Starts: MW	hsmw	F	ln	Housing starts: Midwest (Thous. U.) S.A.
Starts: South	hossou	F	ln	Housing starts: South (Thous. U.) S.A.
Starts: West	hswst	F	ln	Housing starts: West (Thous. U.) S.A.
BP: total	hsbr	F	ln	Housing authorized: Total new priv housing units (Thous. Saar)
BP: NE	hsbne*	F	ln	Houses authorized by build. Permits: Northeast (Thou. U.) S.A.
BP: MW	hsbmw*	F	ln	Houses authorized by build. Permits: Midwest (Thou. U.) S.A.
BP: South	hsbsou*	F	ln	Houses authorized by build. Permits: South (Thou. U.) S.A.
BP: West	hsbwst*	F	ln	Houses authorized by build. Permits: West (Thou. U.) S.A.
PMI	pmi	F	lv	Purchasing managers' index (Sa)
NAPM new ordrs	pmno	F	lv	Napm new orders index (%)
NAPM vendor del	pmdel	F	lv	Napm vendor deliveries index (%)
NAPM invent	pmnv	F	lv	Napm inventories index (%)
Orders: cons gds	a0m008	F	Δln	Mfrs' new orders, consumer goods and materials (Bil. Chain 1982 $) (TCB)
Orders: dble gds	a0m007	F	Δln	Mfrs' new orders, Durable goods industries (Bil. Chain 2000 $) (TCB)
Orders: cap gds	a0m027	F	Δln	Mfrs' new orders, non-defense capital goods (Mil. Chain 1982 $) (TCB)
Unf orders: dble	a1m092	F	Δln	Mfrs' unfilled orders, durable goods indus. (Bil. Chain 2000 $) (TCB)
M&T invent	a0m070	F	Δln	Manufacturing and trade inventories (Bil. Chain 2000 $) (TCB)
M&T invent/sales	a0m077	F	Δlv	Ratio, Mfg. and trade inventories to sales (Based on chain 2000 $) (TCB)
M1	fm1	F	Δ^2ln	Money stock: M1 (Curr, Trav. Cks, Dem Dep, other Ck'able Dep) (Bil$, Sa)
M2	fm2	F	Δ^2ln	Money stock: M2 (M1+ O'nite Rps, Euro$, G/P&B/D Mmmfs&Sav&Sm time Dep (Bil$, Sa)
M3	fm3	F	Δ^2ln	Money stock: M3 (M2+ Lg time Dep, Term RP's&Inst only Mmmfs) (Bil$, Sa)
M2 (real)	fm2dq	F	Δln	Money supply — M2 in 1996 dollars (Bci)
MB	fmfba	F	Δ^2ln	Monetary base, Adj for reserve requirement changes (Mil$, Sa)
Reserves tot	fmrra	F	Δ^2ln	Depository Inst reserves: Total, Adj for reserve Req Chgs (Mil$, Sa)
Reserves non-bor	fmrnba	F	Δ^2ln	Depository Inst reserves: Non-borrowed, Adj Res Req Chgs (Mil$, Sa)
C&I loans	fclnq	F	Δ^2ln	Commercial & industrial loans outstanding in 1996 dollars (Bci)
ΔC&I loans	fclbmc	F	lv	Wkly Rp Lg Com'l Banks: Net change Com'l & Indus Loans (Bil$, Saar)
Cons credit	ccinrv	F	Δ^2ln	Consumer credit outstanding — Non-revolving (G19)
Inst cred/PI	a0m095	F	Δlv	Ratio, consumer installment credit to personal income (Pct.) (TCB)
S&P 500	fspcom	F	Δln	S&P's common stock price index: Composite (1941–43 = 10)
S&P: indust	fspin	F	Δln	S&P's common stock price index: Industrials (1941–43 = 10)
S&P div yield	fsdxp	F	Δlv	S&P's composite common stock: dividend yield (% per annum)
S&P PE ratio	fspxe	F	Δln	S&P's composite common stock: Price-earnings ratio (%, Nsa)
Fed Funds	fyff	F	Δlv	Interest rate: Federal funds (Effective) (% per annum, Nsa)
Comm paper	cp90	F	Δlv	Commercial paper rate (AC)
3 mo T-bill	fygm3	F	Δlv	Interest rate: US Treasury Bills, Sec Mkt, 3-Mo. (% per ann, Nsa)
6mo T-bill	fygm6	F	Δlv	Interest rate: US Treasury Bills, Sec Mkt, 6-Mo. (% per ann, Nsa)
1 yr T-bond	fygt1	F	Δlv	Interest rate: US Treasury Const Maturities, 1-Yr. (% per ann, Nsa)
5 yr T-bond	fygt5	F	Δlv	Interest rate: US Treasury Const Maturities, 5-Yr. (% per ann, Nsa)
10 yr T-bond	fygt10	F	Δlv	Interest rate: US Treasury Const Maturities, 10-Yr. (% per ann, Nsa)
Aaa bond	fyaaac	F	Δlv	Bond yield: Moody's Aaa Corporate (% per annum)
Baa bond	fybaac	F	Δlv	Bond yield: Moody's Baa Corporate (% per annum)
CP-FF spread	scp90	F	lv	cp90-fyff (AC)
3 mo-FF spread	sfygm3	F	lv	fygm3-fyff (AC)
6 mo-FF spread	sfygm6	F	lv	fygm6-fyff (AC)
1 yr-FF spread	sfygt1	F	lv	fygt1-fyff (AC)
5 yr-FF spread	sfygt5	F	lv	fygt5-fyff (AC)
10 yr-FF spread	sfygt10	F	lv	fygt10-fyff (AC)
Aaa-FF spread	sfyaaac	F	lv	fyaaac-fyff (AC)

and series specific. The refinements to the DI forecasts we consider have greater flexibility to adapt to features in the data than the DI.

6. Concluding comments

The present analysis suggests that a useful way to improve forecasts is to use targeted predictors. The targeted-DI forecasts show non-trivial reduction in inflation forecast errors, while using targeted predictors alone is adequate in reducing the RMSE of some macroeconomic series. Allowing for non-linearity can also yield additional gains. How to use the targeted predictors is specific to the series to be forecasted, and it is this tailoring of the series to be forecasted that generates the reduction in forecast errors. While both hard and soft thresholding tend to perform better than no targeting at all, the improvements are larger and more systematic with soft thresholding. Ordering the variables using the LARS-LASSO EN algorithm seems effective in selecting horizon and sample dependent predictors for forecasting economic time series.

Acknowledgements

We would like to thank an anonymous referee, an associate editor, Jeremy Piger (discussant) and participants at the Beveridge–Nelson decomposition anniversary conference for helpful comments. We also acknowledge financial support from the NSF (grants SES-0551275, SES-0549978).

Appendix. Data

Table A.3 lists the short name of each series, its mnemonic (the series label used in the source database), the transformation

applied to the series, and a brief data description. All series are from the Global Insights Basic Economics Database, unless the source is listed (in parentheses) as TCB (The Conference Board's Indicators Database) or column, ln denotes logarithm, Δln and Δ^2ln denote the first and second difference of the logarithm, lv denotes the level of the series, and Δlv denotes the first difference of the series.

References

Bai, J., Ng, S., 2006. Confidence intervals for diffusion index forecasts and inference with factor-augmented regressions. Econometrica 74 (4), 1133–1150.

Bair, E., Hastie, T., Paul, D., Tibshirani, R., 2006. Prediction by supervised principal components. Journal of the American Statistical Association 101 (473), 119–137.

Beveridge, S., Nelson, C.R., 1981. A new approach to decomposition of economic time series into permanent and transitory components with particular attention to measurement of the business cycle. Journal of Monetary Economics.

Boivin, J., Ng, S., 2005. Undertanding and comparing factor based forecasts. International Journal of Central Banking 1 (3), 117–152.

Boivin, J., Ng, S., 2006. Are more data always better for factor analysis. Journal of Econometrics 132, 169–194.

Donoho, D., Johnstone, I., 1994. Ideal spatial adaptiation by wavelet shrinkage. Biometrika 81, 425–455.

Efron, B., Hastie, T., Johnstone, I., Tibshirani, R., 2004. Least angle regression. Annals of Statistics 32 (2), 407–499.

Fan, J., Li, R., 2001. Variable selection via nonconcave penalized likelihood and its oracle properties. Journal of the American Statistical Association 96, 1348–1360.

Forni, M., Hallin, M., Lippi, M., Reichlin, L., 2001. Do financial variables help in forecasting inflation and real activity in the Euro area. Manuscript. www.dynfactor.org.

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.

Fu, W., 1998. Penalized regressions: The bridge vs. the Lasso. Journal of Computational and Graphical Statistics 7 (3), 397–416.

Holm, S., 1979. A simple sequentially rejective multiple test procedure. Scandinavian Journal of Statistics 6, 65–70.

Inoue, A., Kilian, L., 2008. How useful is bagging in forecasting economic time series? A case study of US CPI inflation. Journal of the American Statistical Association 103 (482), 511–522.

Ludvigson, S., Ng, S., 2007. The empirical risk return relation: A factor analysis approach. Journal of Financial Economics 83, 171–222.

Mol, C., Giannone, D., Reichlin, L., 2006. Forecasting using a large number of predictors: Is Bayesian regression a valid alternative to principal components. Unpublished.

Nelson, R., Plosser, C., 1982. Trends and random walks in macroeconomic time series. Journal of Monetary Economics 10, 139–162.

Osborne, M.A., Presnell, B., Turlach, B., 2000. A new approach to variable selection in least squares problem. IMA Journal of Numerical Analysis 20 (3), 389–403.

Stock, J.H., Watson, M.W., 1999. Forecasting inflation. Journal of Monetary Economics 44 (2), 293–335.

Stock, J.H., Watson, M.W., 2002. Forecasting using principal components from a large number of predictors. Journal of the American Statistical Association 97, 1167–1179.

Stock, J.H., Watson, M.W., 2004a. An empirical comparison of methods for forecasting using many predictors. Mimeo. Princeton University.

Stock, J.H., Watson, M.W., 2006. Forecasting with many predictors. In: Elliott, Graham, Granger, Clive, Timmermann, Allan (Eds.), Handbook of Forecasting.

Tibshirani, R., 1996. Regression shrinkage and selection via the lasso. Journal of Royal Statistical Society Series B 58 (1), 267–288.

Zou, H., Hastie, T., 2005. Regularization and variable selection via the elastic net. Journal of Royal Statistical Society, Series B 67 (2), 301–320.