Forecasting with Many Predictors: Allowing for Non-linearity

Eran Raviv[1,2] and Dick van Dijk[1,2,3]

[1] *Econometric Institute, Erasmus University Rotterdam*
[2] *Tinbergen Institute*
[3] *Erasmus Research Institute of Management*

November 28, 2014

Abstract

While there is an extensive literature concerning forecasting with many predictors, there are but few attempts to allow for non-linearity in such a 'data-rich environment'. Using macroeconomic data, we show that substantial gains in forecast accuracy can be achieved by including both squares and first level interactions of the original variables in a predictive regression model. In case the number of original variables is reasonably large this requires specific econometric considerations though, as the number of parameters to be estimated may greatly exceed the number of available observations. We propose a two-stage "screen and clean" procedure that enables estimation and forecasting in this 'ultrahigh-dimensional' setting. In the first stage, we perform univariate regressions to screen for truly interesting effects, controlling the False Discovery Rate. In the second step, we perform a standard bridge regression.

Keywords: Variable selection; Multiple testing; False discovery rate; Data-rich environment.
JEL-code: C53; C55; E37

[1] *Corresponding author:* Dick van Dijk, Econometric Institute, Erasmus University Rotterdam, PO Box 1738, 3000 DR Rotterdam, The Netherlands. Email: djvandijk@ese.eur.nl.

1 Introduction

Over the past decade, we have witnessed an ongoing steady increase in computing power, data collection and data storage facilities. As a result, a great number of potential explanatory variables can nowadays be utilized in economic modeling and forecasting. Practitioners and policy makers need no longer form their decisions according to measurements of only a few key variables, but they may consider a broad set of indicators. In this context, the familiar trade-off between model complexity and forecast accuracy[1] becomes of crucial importance. In order to obtain accurate parameter estimates, we ideally should have many observations for each parameter in the model, i.e. the ratio N/T should be close to zero, where N denotes the number of parameters (which in the context of a linear regression model corresponds with the number of explanatory variables) and T denotes the number of observations available for estimation. When we have only a few observations per parameter, i.e. when the ratio N/T is not close to zero (but not very large yet) things are more complicated, but several tools for modeling and forecasting in this case have been developed in recent years, see Stock and Watson (2006) for a survey.

Much less is known when facing a situation where the ratio N/T is large, say, over ten or even over a hundred. A motivating example from economic forecasting is the need to generate forecasts based on an extremely short estimation period (such that T is small) due to a perceived structural break in the data generating process. Another relevant case, which we consider in our empirical application, occurs when we wish to allow for non-linear relations between the dependent variable (or target variable) and the explanatory variables (or predictor variables). For clarity, assume that we are already in a data-rich (or "high-dimensional") environment with the ratio $N/T \approx 1$. A natural first step to allow for non-linearity is to augment the original explanatory variables with their squares. This doubles the number of parameters in the model, resulting in a ratio $2N/T \approx 2$. This situation is still econometrically manageable. However, when we also consider interaction terms between different variables the ratio inflates to $(N+N(N+1)/2)/T \approx (N+3)/2$. In this paper we use the term "ultrahigh-dimensional" to describe such a situation.[2] Progress in this area is found

[1] More complex models typically involve more unknown parameters. The resulting estimation uncertainty may worsen the model's forecast performance.

[2] In the statistics literature this is sometimes referred to as high-dimensional, but we differentiate it from

primarily in the statistics literature, mainly directed towards applications such as genomics, tumour classifications, signal processing and image analysis. For example, to classify a tumour, thousands of genes are monitored, each potentially helpful, while the number of patients is far smaller.

The literature on statistical modeling in ultra-high-dimensional settings is growing rapidly, steered to provide guidance both in terms of inference (see Meinshausen *et al.*, 2009; Wasserman and Roeder, 2009) and prediction (see Fan *et al.*, 2009; Fan and Lv, 2010). In this paper we focus on the latter issue. In this branch of literature, without exception, the dimensionality issue is handled by splitting the forecasting problem into two steps. First, a screening procedure is applied where the number of variables included in the predictive regression model is substantially reduced. The goal of screening is to go back to a manageable situation with a reasonable value for the ratio N/T, such that existing tools and methods can be used to construct a forecast in the second step. In this paper we follow the methodology introduced by Fan *et al.* (2009) and Fan and Lv (2010), where the initial screening is done by means of componentwise regression.[3] The predictive regression model in the second step then uses only those variables which have highest marginal utility, i.e. the highest marginal correlation with the target variable. A cutoff point defines how large the subset of chosen variables is. It is typical to regard a variable as being relevant if the t-statistic associated with its coefficient in the componentwise regression is larger in absolute value than a fixed threshold $t(\alpha)$ corresponding with a pre-determined significance level α, as in Bai and Ng (2008), for example. This procedure is commonly known as "hard-thresholding". In empirical applications, α is typically set at a conventional value such as 0.05 or 0.10. This is likely not adequate in an ultrahigh-dimensional environment, however, as it may not render a sufficient reduction in the number of explanatory variables that 'survive' the screening procedure.

The first contribution of this paper is to suggest an important modification of this screening step in order to make it useful for ultrahigh-dimensional settings. Our proposal is derived from the well-developed literature of multiple testing. Note that the hard-thresholding procedure as described above can be interpreted in this way, in the sense that a large number of t-tests is

the term "high-dimensional" in economics which describes the more manageable case where $N \approx T$.

[3] If, for example, we have N explanatory variables $\{x_i\}_{i=1}^{N}$ and a single target variable y, applying componentwise regression means running N individual regressions of the form $y_t = \alpha_{0i} + \alpha_{1i} x_{it} + \varepsilon_{it}, \quad i = 1, \ldots N$.

performed in the componentwise regressions. Due to the multiplicity problem, the probability of a type I error obviously increases, and the set of variables passing the screening step likely contains a substantial number of 'false positives'. A natural way to deal with this issue is to control the family wise error rate (FWER). We can commit ourselves to keep the type I error under $100\alpha\%$ in the spirit of Bonferroni by using the threshold $t(\alpha/K)$, where K denotes the number of tests, instead of the threshold $t(\alpha)$. The main drawback of this procedure is its low power, in the sense that it generally fails to detect a substantial number of truly relevant explanatory variables.[4] In our context of forecasting, we can afford to be less strict as the consequences of a type I error are less harmful than, say, falsely approving a drug. Our proposal therefore is to deal with the multiplicity problem by controlling the false discovery rate (FDR). This idea dates back to the seminal paper of Benjamini and Hochberg (1995). The FDR is the expected proportion of rejections that are actually true in population, where in our context a 'rejection' means that a variable passes the initial screening procedure and thus is considered to be helpful for forecasting. The procedure to control the FDR is less strict than the Bonferroni procedure by applying a dynamic threshold to judge the significance of individual t-statistics. Specifically, the threshold changes monotonically from $t(\alpha/K)$ for the largest t-statistic to $t(\alpha)$ for the smallest. As a result the FDR procedure is more powerful and will generally discover more variables that are truly important than the Bonferroni-based FWER procedure. The increased power of the FDR procedure is due to the fact that it controls the number of false discoveries as a proportion of all discoveries, not as a proportion of all tests. So in the case of many true discoveries, there is a wider margin for allowing false discoveries. The cost of this increase in power is by design including more variables that are unimportant compared to the more conservative FWER approach.

A second contribution of this paper lies in our empirical application, where we allow for non-linear relations between a high-dimensional set of explanatory variables and important macroeconomic target variables. Specifically, we aim to predict four key measures of the US real economy, namely industrial production, employment, income and sales, using a set of 126 macroeconomic and financial variables. We extend this set of predictors with their squares and first order interactions. We deal with the resulting ultrahigh-dimensional environment

[4] Holm (1979) offers a refinement of this procedure that is somewhat more powerful.

by first screening for important variables, controlling the FDR. In this way we end up with a manageable, albeit still large number of relevant variables. We then use ridge regression in the second step to estimate the coefficients in the predictive regression model and obtain the out-of-sample forecasts. We find that allowing for non-linearity in this way can be beneficial in terms of forecast accuracy. We note that we are not the first to consider the possibility of nonlinearities in the context of macroeconomic forecasting in a data-rich environment. Earlier attempts include Bai and Ng (2008) including squares of the explanatory variables and by introducing squared factors within a principal component (factor modelling) framework; Giovannetti (2013) using principal component analysis combined with spline regressions; and Exterkate *et al.* (2013) using kernel ridge regression. All three studies find that accounting for non-linearities can lead to a non-trivial improvement in forecast accuracy. Our empirical findings reinforce and further strengthen these results.

Efficiently extracting relevant information from a large number of explanatory variables, while at the same time upholding good forecast performance, is the focus of two main strands of literature. The first strand, referred to as "Diffusion Index" or "Principal Component Regression" modeling, considers summarizing the information from a large panel of predictor variables using a small number of factors, typically taken to be the first few principal components. Under the weak assumption that these factors are a good summary of the information available in the large panel, the factors may be used for prediction instead of the many individual original variables. Prominent contributions in this area are Stock and Watson (1999, 2002a,b, 2006), who drew considerable attention to the success of such methods taking a forecasting perspective, more recently exemplified in Stock and Watson (2012). A generalized version using spectral analysis for factor estimation is developed in Forni and Lippi (2001) and Forni *et al.* (2005). For inference in these class of models see Bai (2003) and Bai and Ng (2006). A survey of the extensive use of these models is found in the meta-analysis undertaken by Eickmeier and Ziegler (2008).

The second strand of literature offers an alternative in the form of shrinkage. In this approach, all individual variables are included in the (predictive) regression model. Obviously, in a data-rich environment this breeds vast estimation noise and leads to overfitting. These effects may be countered by shrinking the parameter estimates towards some target, which

typically is taken to be zero. In most approaches, shrinkage is achieved by penalizing the magnitude of the coefficients. For example, Ridge Regression (Hoerl and Kennard, 1970) minimizes the residual sum of squares plus a penalty in terms of the L_2-norm of the coefficients, while the Least Absolute Sum of Squares Operator (LASSO) uses a penalty in terms of the L_1-norm, see Tibshirani (1996) and Hesterberg *et al.* (2008) for reviews. Both Ridge Regression and LASSO are special cases of the so-called Bridge Regression (Fu, 1998). In a linear regression setting, both also have a Bayesian flavor and can be cast into a Bayesian framework with a specific choice of prior distribution. Over the years, next to the accumulating evidence favoring shrinkage in terms of gains in forecast accuracy, many variants have been suggested. The Elastic Net (Zou and Hastie, 2005), Adaptive LASSO (Zou, 2006) and the Random LASSO (Wang *et al.*, 2011) are really just a few examples. An interesting horserace between many of the methods mentioned above is conducted by Kim and Swanson (2014), who apply a large collection of models to a large-scale dataset of macroeconomic variables. They empirically demonstrate that a combination of the two approaches, shrinkage and dimension reduction, is highly effective for forecasting purposes. In our application, after the initial screening step we opt for Ridge Regression in the second step to obtain the forecast. We find non-trivial forecasting gains from extending the linear relation further, using squares and first order interactions of the original variables.

The rest of this paper is organized as follows. Section 2 outlines the proposed two-step procedure with the aim of forecasting in ultrahigh-dimensional situations. The focus is on our suggestion for the initial screening procedure based on controlling the FDR. Section 3 introduces the empirical application by discussing the data set and several implementation issues. Section 4 describes the empirical results. Section 5 concludes.

2 Forecasting in an ultrahigh-dimensional environment

We follow the convention in the literature on modeling and forecasting in an ultrahigh-dimensional environment and split the problem into two parts. The first part consists of an initial screening procedure, which aims to reduce the set of explanatory variables to a manageable size. The second part then uses the selected subset of explanatory variables to

estimate a predictive regression model for the target variable and to construct a forecast. We frame the discussion in this section in terms of the subsequent empirical application, where the ultrahigh-dimensional environment arises because of the desire to allow for non-linearities in the relations between the explanatory variables and the target. The same principles apply to different settings, including the situation where the number of available predictor variables is ultrahigh to start with.

2.1 Step 1: Screening based on controlling the FDR

Our aim is to construct an effective forecasting procedure that allows for non-linearities between the explanatory variables and the target series, in an ultrahigh-dimensional setting. For this purpose we augment the original predictor variables with their squares and first-order interactions. The resulting variables are collected in the $T \times N$ matrix $\boldsymbol{X}$, where N denotes the total number of variables (i.e. the sum of the number of original variables, their squares and first-order interactions) and T denotes the number of available observations. All variables in $\boldsymbol{X}$ are standardized to have mean zero and variance one. In the following we use 'explanatory variable', 'variable' or 'predictor' to describe a column in $\boldsymbol{X}$, be it an original variable, its square, or an interaction term between two original variables. In case the number of original variables is already fairly large (relative to the time dimension T), including their squares and interactions will lead to the situation that $N \gg T$. Hence, conventional predictive regression models cannot be applied. Furthermore, even while in theory techniques such as principal component regression (PCR) or ridge regression may be able to handle this situation, in practice they are likely to suffer from problems if they are applied directly, using the full matrix $\boldsymbol{X}$. Specifically, it is reasonable to assume that most variables in $\boldsymbol{X}$ are not related to the target, especially since an interaction term may be deemed important only if it provides information in addition to the original variables. A technique such as PCR is known to be negatively affected by the inclusion of (many) irrelevant variables, see Boivin and Ng (2006) and Bai and Ng (2008), among others. For this reason it is useful to reduce the set of variables before applying such techniques.

Let y_{t+h} denote the target variable at time $t+h$, where h is the forecast horizon. In order to select a subset of the available predictor variables, we conduct a univariate predictive

regression for each variable x_{it}, $i = 1, \ldots, N$:

$$y_{t+h} = \beta_{0i} + \beta_{1i} x_{it} + \beta_{2i} x_{kt} \mathbb{1}_{\{x_{it}=x_{kt}x_{lt}\}} + \beta_{3i} x_{lt} \mathbb{1}_{\{x_{it}=x_{kt}x_{lt}\}} + \varepsilon_{i,t+h}, \qquad t = 1, \ldots, T-h, \tag{1}$$

where $k, l = 1, \ldots, N$ and $k \neq l$, and $\mathbb{1}_{\{C\}}$ is the indicator function which takes the value 1 if the condition C is true and 0 otherwise. In (1), the condition C is whether the variable x_i under consideration is an interaction term between (original) variables x_k and x_l. If so, these original variables are also included in the regression, implicating that selection of interaction term means that it has predictive ability in addition to the original variables.[5] The relevance of variable i is judged by the t-statistic associated with the least squares estimate of the coefficient β_{1i} in (1). Typically, we select those variables for which the corresponding (two-sided) p-values are below a predetermined significance level α. This componentwise design, maybe due to its simplicity and ease of implementation, is increasing in popularity, with support from the forecast combination literature (Elliott *et al.*, 2013; Samuels and Sekkel, 2013; Rossi and Sekhposyan, 2014) and, as in this case, alternative screening procedures (Bair *et al.*, 2006; Bai and Ng, 2008; Fan *et al.*, 2009; Fan and Lv, 2010).

The screening procedure based on the componentwise regression in (1) can be viewed as a multiple hypothesis testing problem, since it involves judging the significance of N different test statistics. In the ultrahigh-dimensional setting where N is extremely large, using a fixed significance level α implies that many variables are likely to be mistaken as relevant only because of the large number of tests conducted (unless α is set to an extremely low value, of course, but this may reduce the discriminating power of the procedure to identify relevant predictors, as discussed in the introduction). We propose a refinement of this procedure to account for this feature. In order to do so we invoke the FDR of Benjamini and Hochberg (1995). Controlling the FDR means controlling the (unknown) quantity:

$$\mathsf{E}\left(\frac{V}{V+S}\right), \tag{2}$$

[5] In case x_i is the square of an original variable we do not include the original variable in (1). This is motivated by the fact that the original variables are assumed to be standardized to have mean zero (and unit variance), so that the correlation between the variable and its square will be (close to) zero by construction.

where V is the number of false rejections and S is the number of correct rejections. Hence, instead of controlling the proportion of false rejections relative to the total number of tests (as in the traditional procedure described above), we aim to control the proportion of false rejections relative to the total number of rejections. The motivation for controlling this quantity is twofold. First, while making a type I error twice when we select four variables is 'unacceptable', it is much less problematic when we select a few hundred variables. Controlling the FDR allows just that, with more false rejections being allowed as long as we also discover more truly important variables. This is in sharp contrast with the FWER procedure, which is ignorant to the number of true discoveries. Second, although undesirable since we unnecessary inflate estimation noise, the implications of false positives in our forecasting context are far less severe than, say, in case of approving an ineffective drug for production. In that sense we can be less strict and in return, gain a higher number of true discoveries.

Controlling the FDR is achieved with the following procedure. We order the p-values p_i, $i = 1, \ldots, N$, associated with the least squares estimate of the coefficient β_{1i} in (1) in increasing order and denote them by $p_{(i)}$, such that

$$p_{(1)} < p_{(2)} < \cdots < p_{(N)}. \tag{3}$$

We then select the variables associated with the m smallest p-values, where

$$m = \max\{j : p_{(j)} \leq \frac{j}{N}\alpha,\ j = 1, \ldots, N\}, \qquad \text{for given } 0 < \alpha < 1. \tag{4}$$

In words, m is such that all ordered p-values up to and including the m-th one are smaller than the increasing sequence $\frac{j}{N}\alpha$, but the $(m+1)$-st one is not. Note that the number of variables in the resulting subset is determined by the strength of the marginal correlation *and* by the number of variables we test.

Naturally, when variables x_k and x_l are correlated, the t-statistics for β_{1k} and for β_{1l} in (1) are correlated. While the FDR procedure as described above is designed for independent tests, Benjamini and Yekutieli (2001) suggest a correction for correlated tests. The correction is general in the sense that it does not depend on the specific form of the correlation structure

between the different tests. Instead of using (4), we set m as

$$m = \max\{j : p_{(j)} \leq \frac{j}{N(\frac{1}{2} + \log(N))}\alpha,\ j = 1, \ldots, N\}. \tag{5}$$

This procedure provides us with a subset of variables that are considered to be most important for prediction. We do not pursue a theoretical justification so admittedly, we may be left with a subset that does not contain all truly relevant variables. We leave further theoretical developments for future research and at the moment, merely use this screening procedure as a quantitative selection device.

2.2 Step 2: Forecasting based on ridge regression

We collect the variables selected by means of the FDR-based screening procedure and denote the new reduced matrix of explanatory variables as $\tilde{\boldsymbol{X}}$. These are used in the predictive regression model

$$y_{t+h} = \tilde{\mathbf{x}}_t \boldsymbol{\beta} + \varepsilon_{t+h}, \qquad t = 1, \ldots, T-h, \tag{6}$$

where $\tilde{\mathbf{x}}_t$ denotes the t-th row in $\tilde{\boldsymbol{X}}$. Given the ultrahigh dimension of the initial problem, we are likely to be left with a large number of explanatory variables in $\tilde{\mathbf{x}}_t$ still, compared with the number of observations available for estimation. This is a situation particularly prone to the danger of overfitting. In order to mitigate this effect, and given the fact that the variable selection in the first step has been done according to marginal importance, we use ridge regression to estimate the coefficients $\boldsymbol{\beta}$ in the predictive regression model (6). Shrinkage, by means of ridge regression or related techniques, has long been proven to be a powerful tool to prevent overfitting. This also fits the focus of this paper which is more on improving out-of-sample performance by allowing for non-linearities, and less on the inference side. When one is more interested in inference, instead of using ridge regression, shrinkage can be applied via the LASSO or Adaptive LASSO. These methods have the advantage of shrinking coefficients exactly to zero and thus effectively achieving a further reduction in the subset of predictor variables that are considered relevant.[6]

[6] In the more common case where $T > \tilde{N}$, it is observed that prediction performance of ridge regression is better than that of the LASSO, when cross-correlation in the explanatory matrix is high (Tibshirani, 1996), but in general there is no evidence for universal dominance of one method over the other, see Fu (1998).

Formally, we minimize the residual sum of squares plus a penalty in term of the L_2-norm of the coefficients:

$$RSS(\lambda) = (\boldsymbol{y} - \tilde{\boldsymbol{X}}\boldsymbol{\beta})'(\boldsymbol{y} - \tilde{\boldsymbol{X}}\boldsymbol{\beta}) + \lambda\boldsymbol{\beta}'\boldsymbol{\beta}, \tag{7}$$

where $\boldsymbol{y}$ is the vector of observations on the target variable and the shrinkage coefficient $\lambda > 0$. The solution to this minimization problem is given by:

$$\hat{\boldsymbol{\beta}}^{RR} = (\tilde{\boldsymbol{X}}'\tilde{\boldsymbol{X}} + \lambda I)^{-1}\tilde{\boldsymbol{X}}'\boldsymbol{y}. \tag{8}$$

The h-step ahead point forecast for y_{T+h} is then obtained as

$$\hat{y}_{T+h|T} = \tilde{\mathbf{x}}_T\hat{\boldsymbol{\beta}}^{RR}.$$

3 Data, implementation and benchmark forecasts

In this section we introduce the application that we use to assess the empirical usefulness of our proposed forecasting procedure in an ultrahigh-dimensional environment. We describe the data, several relevant implementation issues, and competing methods that are used as benchmarks for comparison.

3.1 Data

Our data set comprises a large number of US macroeconomic and financial variables at the monthly frequency for the period April 1959 - September 2009. Following Stock and Watson (2002b), Bai and Ng (2008) and related studies, we consider a total of 126 variables including various measurements of production, consumption, income, sales, employment, monetary aggregates, prices, interest rates, and exchange rates. All variables are transformed to stationarity by taking logarithms and/or first differences as described in Stock and Watson (2005). We use four key indicators of real economic activity as target variables to be predicted: Industrial production, personal income, manufacturing & trade sales and employment. The target series are transformed to represent annualized h-month percentage growth rates: $y_{t+h,h} = \frac{1200}{h}\ln\left(\frac{w_{t+h}}{w_t}\right)$, where w_t is the original series. We consider four short- and medium-

term forecast horizons, namely $h = 1$, 3, 6 and 12 months ahead. To simplify the notation, the one-month growth rate ($h = 1$) is denoted as y_{t+1}.

3.2 Implementation

We use a moving window with a fixed length of 10 years to specify and estimate all forecasting models. For our proposed two-step procedure as described in Section 2, this means that each month both the FDR-based screening and the ridge regression are implemented using the most recent $T = 120$ observations. Using a fixed length moving window is a simple and popular way to counter, at least partially, the effects of possible structural breaks in the data generating process (Pesaran *et al.*, 2006).

In order to obtain a good insight with regards to the possible gains in forecast accuracy due to allowing for non-linear relations between the predictors and the target series, we apply our proposed procedure using three different sets of predictor variables: (1) only the 126 original variables, (2) the original variables together with their squares, and (3) the original variables together with their squares and first-order interactions. In the remainder, these three cases are labeled S (for 'Small'), M ('Medium') and L ('Large'). In addition, following Stock and Watson (2002b) we account for autocorrelation in the target variables by including p lags of the one-month growth rate, y_{t-j}, $j = 0, 1, \ldots, p-1$, in both the univariate regression in (1) in the initial screening phase as well as in the final predictive regression model in (6). We fix the lag length p at 4 throughout the analysis.

To determine an appropriate value for the ridge parameter λ in (7) we use a data-driven strategy based on cross-validation (CV), see Arlot and Celisse (2010) for a recent survey. The main advantage of using a CV-type procedure here is that we can tailor it to our needs, focusing on prediction, unlike in-sample oriented methods based on information criteria, for example. A full-blown (leave-one-out) CV procedure is prohibitively costly though, due to the use of the 10-year moving window for specification and estimation. In addition, CV is problematic due to the time series nature of our data and the fact that we are also interested in more than one period ahead forecasts. We modify the procedure in a way that honors the temporal dependence structure in the data, while at the same time making it out-of-sample oriented and computationally feasible (even though still costly). Using the pre-selected subset

of the predictors (recall this subset is allowed to change from one period to the next), we obtain h-month ahead forecasts from the ridge regression according to a fine grid of different λ values. We use the subsequent 36 months as a validation period, and select the value of λ that delivers the smallest root mean squared prediction error (RMSE) for the validation period. The first forecast which enters the evaluation is that after the 13 years used for both training (10 years) and validation (3 years). As a result, the evaluation period runs from May 1974 until September 2009. Also note that this procedure implies that the ridge parameter may vary over time as well as across forecast horizons.

3.3 Benchmark forecasts

We consider three competing methods to deal with the (ultra)high-dimensional environment as benchmarks for comparison. Following Bai and Ng (2008) and Stock and Watson (2012), the first benchmark model we use is a univariate autoregressive (AR) model of the form

$$y_{t+h,h} = \alpha + \phi_1 y_t + \cdots + \phi_p y_{t-p+1} + \varepsilon_{t+h,h}, \tag{9}$$

where we again set $p = 4$. Note that the 'predictors' in this case are lagged one-month growth rates, irrespective of the forecast horizon h.

The second benchmark is the diffusion index (DI) model of Stock and Watson (2002a), also called (dynamic) principal component regression (PCR). It is widely used in macroeconomic forecasting and has been recently shown by (Stock and Watson, 2012) to be a tough benchmark to beat. In this approach, forecasts are obtained from the predictive regression model

$$y_{t+h,h} = \alpha + \phi_1 y_t + \cdots + \phi_p y_{t-p+1} + \boldsymbol{\gamma}' \boldsymbol{f}_t + \varepsilon_{t+h,h}, \tag{10}$$

where $\boldsymbol{f}_t$ is a vector of r factors presumed to properly span the variance in the set of predictors $\boldsymbol{X}$. Typically the first few principal components of the covariance matrix of $\boldsymbol{X}$ are used for this purpose. In our empirical application we use two different specifications for the set of predictors $\boldsymbol{X}$: The first consists of the original variables only (abbreviated here as PC), the second also includes their squares (abbreviated here as SPC).[7] The decision regarding

[7] We also considered a third possibility by including first-order interactions of the original variables (QPC).

the number of principal components to include in (10) is made in a similar fashion to that described above. We create 36 out-of-sample forecasts, using different numbers of factors ranging from 1 to 10. We pick the number of factors that delivers the lowest RMSE based on those forecasts.[8]

4 Results

In this section we report the forecasting results obtained for the four real macroeconomic time series for the four forecast horizons considered. The FDR-based procedure using the S, M and L sets of predictors is contrasted against the three benchmark forecasts AR, PC and SPC. We evaluate the out-of-sample performance using the Root Mean Squared Error (RMSE), by far the most common evaluation metric in the literature (Gneiting, 2011).

Table 1 presents the RMSEs for each of the forecasts relative to the RMSE of a random-walk or 'no-change' forecast (i.e. $y_{t,h}$ is used as a forecast of $y_{t+h,h}$. For each target series and forecast horizon, the method that achieves the lowest RMSE is highlighted in bold.

The table suggests three main conclusions. First, using the FDR-based screening procedure generally offers improvements in forecast accuracy compared to the benchmark methods. The FDR-based forecasts achieve the lowest RMSE in 12 out of the 16 cases considered. While forecasting gains are observed across all forecast horizons, it also appears that the improvements are largest at longer horizons. For $h = 6$ and 12 months the FDR approach leads to quite substantial improvements, with gains in RMSE up to 8% relative to the best benchmark. This empirical finding is in line with results reported in Bai and Ng (2008), among others, where the added value of non-linearity also is found to be much more pronounced for longer horizons than for shorter horizons. The benchmark forecasts, in particular PC, are most difficult to improve upon for employment. This is in agreement with Exterkate *et al.* (2013),

This approach resulted in dramatically worse forecasting performance, in line with the negative results obtained by Bai and Ng (2008). Results are therefore not reported here, but available upon request. Apart from SPC, another way to allow for non-linearity is to add the squared factors to equation (10), i.e.,

$$y_{t+h,h} = \alpha + \phi_1 y_t + \cdots + \phi_p y_{t-p+1} + \boldsymbol{\gamma}_1' \boldsymbol{f}_t + \boldsymbol{\gamma}_2' \boldsymbol{f}_t^2 + \varepsilon_{t+h,h}.$$

Both Bai and Ng (2008) and Exterkate *et al.* (2013) find this specification to be dominated by SPC and hence we do not apply it here.

[8] We also experimented with choosing the number of factors using BIC, which is more common in this case. Results in terms of forecast accuracy are qualitatively similar.

where it is also documented that PC delivers relatively accurate forecasts for this series.

Second, in the context of the FDR-based procedure, allowing for non-linear relations between the predictors and the target variable improves forecast accuracy. In some instances (in particular for industrial production) allowing for interaction effects among different original variables offers material gains. This finding emerges from comparing the relative RMSE values for the FDR-based forecasts using the S, M, and L sets of predictors. We observe that forecasts based on the set of original variables only (S) are (on average) dominated either by those that only include the squares (M) or also the first-order-interactions (L). In fact, in most cases both the M- and L-based forecasts achieve a lower RMSE than the S-based forecasts. Comparing the results for the M- and L-based forecasts directly is more subtle. For horizons 6 and 12 months horizons, we find that allowing for interactions improves forecast accuracy for 3 out of the 4 series considered. Also here we find that especially at longer forecast horizons allowing for more complex non-linear relations is beneficial.

Third, allowing for non-linearity by including squared principal components, as in the SPC approach, does not lead to forecast improvements. In fact, in the large majority of cases, PC performs better than SPC, a result again in line with Bai and Ng (2008). A plausible explanation for this finding is the fact that in the SPC approach *all* variables load on the factors, i.e. the factor loadings are not sparse. This may prompt inaccurate factor estimates, which eventually harms forecast accuracy. Strong support for this argument can be found in Bai and Ng (2008) where it is found that restricting the number of variables which enter the factor construction achieves substantially better results.

[Table 1 about here.]

We next examine whether the benefits from allowing for non-linear relations between the predictors and the target variable are stable over time. Table 1 presents results for the complete evaluation period May 1974-September 2009, which covers more than 35 years. Hence, a relevant question is whether the superior forecasting performance of the FDR-RR method arises because of a 'steady stream' of more accurate forecasts throughout the entire period, or whether this is due to specific sub-periods. Figure 1 helps in answering this question, by presenting 10-year rolling RMSEs of the different FDR-based forecasts relative to the RMSE

of PC-based forecasts, for the 12 months horizon. Several features are noteworthy. First, and most importantly, we do observe variation in the (relative) performance of the FDR-RR method over time. For all four target series, performance is relatively strong compared with the PC method until the late 1990s, but around the turn of the millenium it worsens such that the RMSE of the FDR-based forecasts actually exceeds that of the PC-based forecasts. Such behavior of a well performing method over some period which looses it's edge over some other periods is also encountered in the inflation forecasting literature Stock and Watson (2009). Second, Figure 1 generalizes the finding in Table 1 that incorporating interactions improves performance, which is observed almost uniformly across series and over time. Compared to just including the original predictors (S), including their squares (M) delivers similar or smaller RMSE quite consistently. Once interactions are introduced (L) performance is further improved.

Taken together, Figure 1 induces further confidence in our FDR-based forecast method, with forecasting gains observed in some periods for all four series. That said, there are periods in which adding interactions negatively impacts accuracy compared with just using the original variables and their squares. In some short periods the L-based approach is even worse than the PC method. Since it is hard to foresee in advance which specification is best for a given period, in order to stabilize performance one might consider averaging forecasts from the three specifications, S, M and L together with other models or external forecasts.

[Figure 1 about here.]

5 Conclusion

In this paper, we propose a novel procedure for forecasting in an ultrahigh-dimensional environment, where the number of predictor variables greatly exceeds the number of observations available for estimating the predictive regression model. The procedure consists of two steps. In the first step, the set of predictor variables is reduced by selecting those variables that appear most informative (from a forecasting perspective) for the target series. The novelty introduced in this step is that the selection is performed while controlling the false discovery rate (FDR), instead of applying a "hard thresholding"-type approach that controls the family

wise error rate (FWER). In the second step we use ridge regression to estimate the predictive regression model. Shrinkage is applied here, as the number of predictor variables selected in the first step may still be large and so may harm forecasting performance by way of overfitting.

Our proposed two-step procedure can also be viewed as *diversified* shrinkage. Each group of coefficients in the predictive regression model has its own penalty parameter λ, with coefficients of variables excluded in the first step having an extremely high value of λ (which shrinks them to zero), while the λ of the remaining coefficients is determined as usual using CV. In that sense, our procedure fits into the general framework recently given in Stock and Watson (2012) with a specific shrinkage function.

We apply the proposed two-step procedure in an empirical forecasting exercise, where the target series are four monthly measures of the US real economy. The ultrahigh-dimensional environment arises here because of the desire to accommodate possible non-linear relations between the (126) predictors and the target. This is achieved by allowing for squares and cross-products of the original variables in the predictive regression model. We document that substantial improvements in forecast accuracy can be achieved by (i) allowing for such non-linearities, and (ii) by applying the FDR-based variable selection procedure. The improved predictive ability is most substantial for longer forecast horizons.

Our empirical results encourage further research towards other possible ways to allow for a non-linear relations between predictor variables and target series. The research into this area is not yet abundant, but empirical evidence in this- and other papers suggest this can be exploited using modern statistical methods. One such direction may be the exploration of the fast growing literature which combines dimension reduction and sparsity. This includes papers discussing sparse partial least squares (Boulesteix and Strimmer, 2007, and references therein), and the promising method of sparse principal component analysis (Zou *et al.*, 2006).

References

Arlot, S. and A. Celisse (2010), A survey of cross-validation procedures for model selection, *Statistics Surveys* **4**, 40–79.

Bai, J. (2003), Inferential theory for factor models of large dimensions, *Econometrica* **71**, 135–171.

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

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

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

Benjamini, Y. and D. Yekutieli (2001), The control of the false discovery rate in multiple testing under dependency, *Annals of Statistics* **29**, 1165–1188.

Benjamini, Y. and Y. Hochberg (1995), Controlling the false discovery rate: A practical and powerful approach to multiple testing, *Journal of the Royal Statistical Society. Series B* **57**, 289–300.

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

Boulesteix, A.-L. and K. Strimmer (2007), Partial least squares: A versatile tool for the analysis of high-dimensional genomic data, *Briefings in Bioinformatics* **8**, 32–44.

Eickmeier, S. and C. Ziegler (2008), How successful are dynamic factor models at forecasting output and inflation? A meta-analytic approach, *Journal of Forecasting* **27**, 237–265.

Elliott, G., A. Gargano and A. Timmermann (2013), Complete subset regressions, *Journal of Econometrics* **177**, 357–373.

Exterkate, P., P.J.F. Groenen, C. Heij and D. van Dijk (2013), Nonlinear forecasting with many predictors using kernel ridge regression, CREATES Research Papers 2013-16.

Fan, J. and J. Lv (2010), A selective overview of variable selection in high dimensional feature space, *Statistica Sinica* **20**, 101–148.

Fan, J., R. Samworth and Y. Wu (2009), Ultrahigh dimensional feature selection: Beyond the linear model, *The Journal of Machine Learning Research* **10**, 2013–2038.

Forni, M., M. Hallin, M. Lippi and L. Reichlin (2005), The generalized dynamic factor model, *Journal of the American Statistical Association* **100**, 830–840.

Forni, M. and M. Lippi (2001), The generalized dynamic factor model: Representation theory, *Econometric Theory* **17**, 1113–1141.

Fu, W.J. (1998), Penalized regressions: The bridge versus the lasso, *Journal of Computational and Graphical Statistics* **7**, 397–416.

Giovannetti, B.C. (2013), Nonlinear forecasting using factor-augmented models, *Journal of Forecasting* **32**, 32–40.

Gneiting, T. (2011), Making and evaluating point forecasts, *Journal of the American Statistical Association* **106**, 746–762.

Hesterberg, T., N.H. Choi, L. Meier and C. Fraley (2008), Least angle and L1 penalized regression: A review, *Statistics Surveys* **2**, 61–93.

Hoerl, A.E. and R.W. Kennard (1970), Ridge regression: applications to nonorthogonal problems, *Technometrics* **12**, 69–82.

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

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

Meinshausen, N., L. Meier and P. Bühlmann (2009), P-values for high-dimensional regression, *Journal of the American Statistical Association* **104**, 1671–1681.

Pesaran, M.H., D. Pettenuzzo and A. Timmermann (2006), Forecasting time series subject to multiple structural breaks, *Review of Economic Studies* **73**, 1057–1084.

Rossi, B. and T. Sekhposyan (2014), Evaluating predictive densities of US output growth and inflation in a large macroeconomic data set, *International Journal of Forecasting* **30**, 662–682.

Samuels, J.D. and R. Sekkel (2013), Forecasting with many models: Model confidence sets and forecast combination, Bank of Canada Working Paper 2013-11.

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

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

Stock, J.H. and M.W. Watson (2002b), Macroeconomic forecasting using diffusion indexes, *Journal of Business and Economic Statistics* **20**, 147–162.

Stock, J.H. and M.W. Watson (2005), Implications of dynamic factor models for VAR analysis, NBER Working Paper No. 11467.

Stock, J.H. and M.W. Watson (2006), Forecasting with Many Predictors, in G. Elliott, C.W.J. Granger and A. Timmermann (eds.), *Handbook of Economic Forecasting Vol. 1*, Elsevier, pp. 515–554.

Stock, J.H. and M.W. Watson (2009), Phillips curve inflation forecasts, in J. Fuhrer, Y.K. Kodrzycki, J. Sneddon Little and G.P. Olivei (eds.), *Understanding Inflation and the Implications for Monetary Policy, a Phillips Curve Retrospective*, MIT Press.

Stock, J.H. and M.W. Watson (2012), Generalized shrinkage methods for forecasting using many predictors, *Journal of Business and Economic Statistics* **30**, 481–493.

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

Wang, S., B. Nan, S. Rosset and J. Zhu (2011), Random lasso, *The Annals of Applied Statistics* **5**, 468–485.

Wasserman, L. and K. Roeder (2009), High dimensional variable selection, *Annals of Statistics* **37**, 2178–2201.

Zou, H. (2006), The adaptive lasso and its oracle properties, *Journal of the American Statistical Association* **101**, 1418–1429.

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

Zou, H., T. Hastie and R. Tibshirani (2006), Sparse principal component analysis, *Journal of Computational and Graphical Statistics* **15**, 265–286.

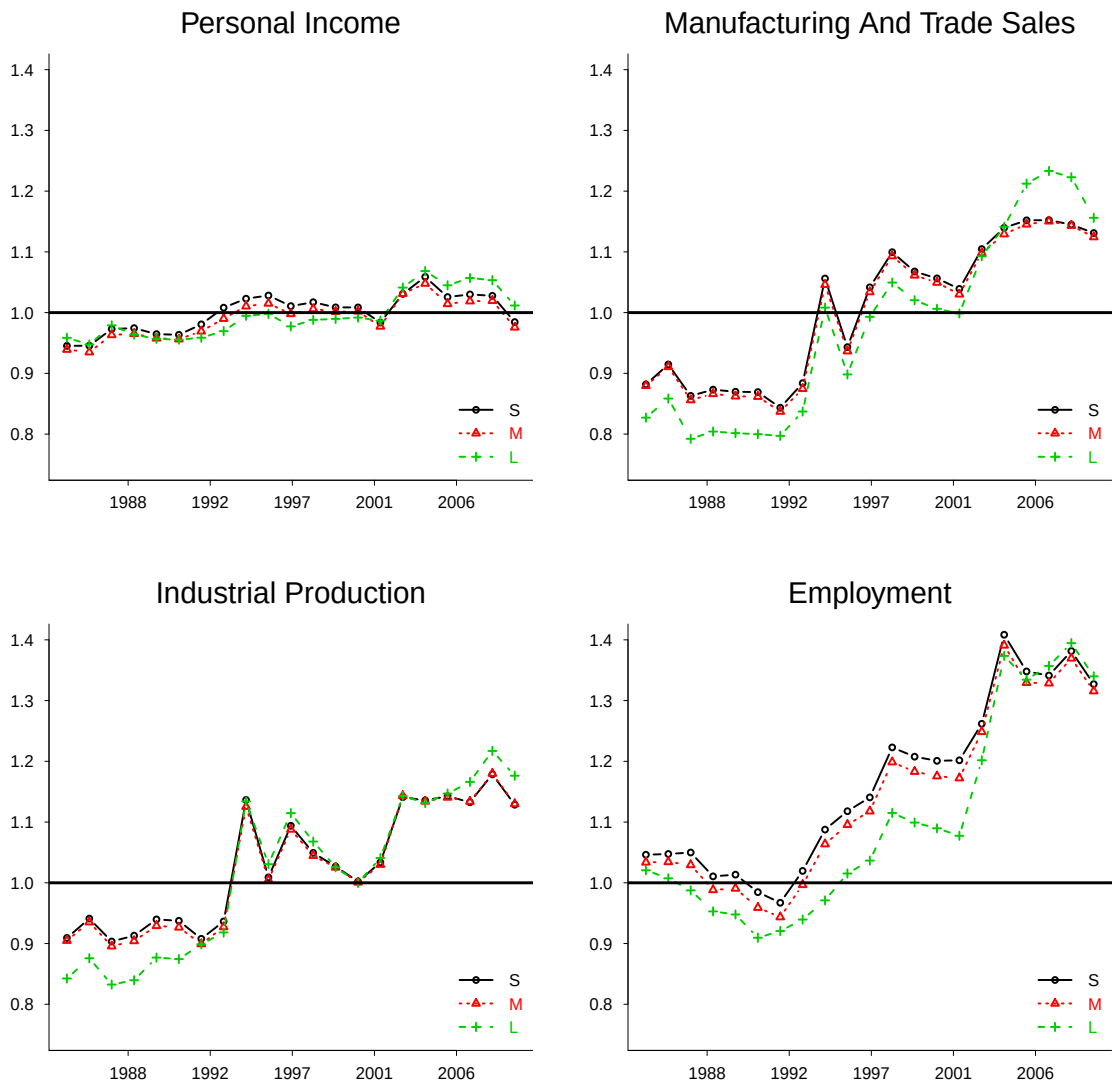

Figure 1: Ten years rolling RMSE. The graph shows the ratio of the RMSE of the $FDR-RR$ method to the RMSE of the PC benchmark model, for the 12 months horizon. The horizontal line at 1 represents equal RMSE between the PC method and the $FDR-RR$ method.

		Personal Income				Manufacturing & Trade Sales			
	h:	1	3	6	12	1	3	6	12
Benchmarks									
AR(4)		0.687	0.715	0.837	0.839	0.666	0.824	0.837	0.820
PC		0.858	0.714	0.879	0.864	0.672	0.830	0.855	0.856
SPC		**0.686**	0.743	0.870	0.903	0.659	0.838	0.834	0.875
FDR-RR									
S		0.673	0.706	0.833	0.841	0.659	0.798	0.819	0.816
M		0.672	**0.701**	**0.826**	**0.834**	**0.656**	**0.792**	0.811	0.810
L		0.687	0.727	0.844	0.843	0.658	0.809	**0.785**	**0.757**
		Industrial Production				Employment			
	h:	1	3	6	12	1	3	6	12
Benchmarks									
AR(4)		0.842	0.937	0.884	0.797	0.918	1.113	0.945	**0.795**
PC		0.955	0.885	0.834	0.798	0.915	**0.972**	**0.929**	0.830
SPC		0.822	0.917	0.830	0.888	0.942	1.023	0.936	0.901
FDR-RR									
S		0.833	0.915	0.867	0.795	1.119	1.337	1.113	0.919
M		0.818	0.898	0.853	0.790	1.058	1.276	1.075	0.906
L		**0.783**	**0.857**	**0.813**	**0.774**	**0.904**	1.079	0.960	0.886

Table 1: Results - out-of-sample accuracy

The table reports root mean squared prediction errors (RMSE) for forecasts of the h-month growth rate over the period May 1974 to September 2009, relative to the RMSE of a no-change forecast. For each series and each forecast horizon, the lowest RMSE achieved across all forecast methods is printed in boldface. PC denotes the principal component regression based on 10. SPC denotes the case where the principal components are allowed to load also on the squares of the original variables. FDR-RR indicates the forecast method with initial screening based on controlling the FDR followed by a ridge regression for the forecasting model. Three sets of predictors are considered in the screening phase: S - only the 126 original variables; M - the original variables together with their squares, and L - the original variables together with their squares and first-order interactions.