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*

December 12, 2013

Abstract

While there is an extensive literature concerning forecasting in a data-rich environment, there are but few attempts to allow for non-linearity in such cases. A non-linear extension in a data-rich environment quickly induces instabilities (due to the *curse of dimensionality*) which require specific econometric considerations. When augmenting an already large number of potentially important explanatory variables with their squares and first order interactions, the result is a situation with very high ratio of number of parameters to number of data points available for estimation. We apply a two step procedure. In the first step we screen for truly interesting effects by way of simple t-tests, accounting for the large number by controlling False Discovery Rate; this is the main thrust of the paper. In the second step we use Ridge-Regression in order to mitigate potential overfitting. Using macroeconomic data, we show that accuracy gains are achieved by allowing for both squares and first level interactions of the original explanatory variables.

[1] *Corresponding author:* Eran Raviv, Econometric Institute, Erasmus University Rotterdam, PO Box 1738, 3000 DR Rotterdam, The Netherlands. Email: `eran.raviv@apg-am.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 now be utilized in the practice of economic modelling and forecasting. Nowadays, practitioners and policy makers need not form their decisions according to measurements of only a few key variables but they may consider a broad set of indicators (Mönch, 2008, and references therein). 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 fairly close to zero, where N denotes the number of parameters (or variables included in the model) and T denotes the number of observations available for estimation. When we have only a few or even a single observation per parameter, i.e. when the ratio N/T is not close to zero (but not very large yet), the situation is more complicated, but at this point in time we have several excellent ways to handle it. We provide a brief review of these techniques later in this section.

Much less is known when facing a case 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 example, which we consider in our empirical application, is when we wish to allow for non-linear relations between the dependent variable (or target variable) and the explanatory variables. For clarity, assume that we are already in a data-rich (or "high-dimensional") environment and 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 primarily in the statistical 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,

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

[2] In the statistical literature this is sometimes referred to as high-dimensional, but we differentiate it from the term "high-dimensional" in economics which describes the more manageable case.

steered to provide guidance both in terms of inference (see Meinshausen et al., 2009; Wasserman and Roeder, 2009) and prediction (see Fan and Lv, 2010; Fan et al., 2009). In this paper we focus on the latter issue. In this branch of literature, with no 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 prediction 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. A good example of such a two-step procedure is the supervised principal component approach advocated by Bair et al. (2006). 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 applying componentwise regressions.[3] The prediction 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 decide that a variable is important if the t-statistic associated with its coefficient is larger in absolute value than a fixed threshold $t(\alpha)$ corresponding with a predetermined significance level α, as in Bai and Ng (2008), for example. This is commonly known as hard-thresholding. In empirical applications of this procedure, 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 a modification to this screening step in order to make it useful for ultrahigh-dimensional settings. The modification we suggest is pulled from the well-developed literature of multiple testing. Note that hard-thresholding procedure can be interpreted in this way, in the sense that a large number of t-tests is performed. 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 Boole's inequality, that is, instead of using the threshold $t(\alpha)$ we use the threshold $t(\alpha/K)$, where K denotes the number of tests. The main drawback of this procedure is its low power, in the sense that it generally fails to detect truly relevant explanatory variables.[4] In our context of forecasting, we can afford to be considerably less stringent, meaning that the severity of a type I error in our context is far

[3] If, for example, we have N explanatory variables $\{x_i\}_1^N$ and one target y, applying componentwise regression means running N individual regressions of the form $y_t = \alpha_0 + \alpha_i x_{ti} + \varepsilon_{ti}, \quad i = 1, \ldots N$.

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

less 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 by 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 stringent that 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 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 is in our empirical application, where we allow for non-linearities in the relations between a high-dimensional set of explanatory variables and the 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 show that allowing for non-linearity in the prediction model can be very beneficial in terms of forecast accuracy. We are not the first to consider this possibility; earlier attempts were made by Bai and Ng (2008) including squares of the explanatory variables and by introducing squared factors within a principal component (factor modelling) framework; by Giovannelli (2012) using kernel principal component analysis and artificial neural networks; and by 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. We extend the set of predictors with the squares and first order interactions of the original variables. We deal with the resulting ultrahigh-dimensional environment 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 where a wide array of existing methods, outlined shortly, are available.

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" modelling, 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, 2002, 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. Following this approach, we consider all of the individual variables, which in our data-rich environment breeds vast estimation noise and leads to overfitting the model. We counter these effects by shrinking the parameter estimates. Actually the shrinking is done by penalizing the magnitude of the coefficients. Ridge Regression (Hoerl and Kennard, 1970) minimizes the residual sum of squares plus a penalty in terms of the $L2$-norm of the coefficients, while the Least Absolute Sum of Squares Operator (LASSO) uses a penalty in terms of the $L1$-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 flavour and can be cast into a Bayesian framework with a specific choice of prior distribution. Over the years, next to the accumulating evidence favouring shrinkage in terms of accuracy gains, many variants have been suggested. The Elastic Net (Zou and Hastie, 2005), Adaptive LASSO (Zou, 2006) Bootstrapped LASSO, (Bach, 2008) and the Random LASSO (Wang et al., 2011) are really just a few examples. The Bayesian regression (or shrinkage regression) is yet another variant. In a Bayesian regression we place a prior distribution on both the regression coefficients and their variances. Given the prior distribution, the posterior distribution is then used to to obtain point estimates and confidence intervals (see Geweke and Whiteman, 2006; Ghosh, 2009, for more details). It has been shown to be a powerful addition in terms of forecast accuracy, examples related to economic forecasting are given in Doan and Sims (1984), Carriero et al. (2011) De Mol et al. (2008) and Koop (2013). Many of the methods just mentioned can be found in Kim and Swanson (2013), 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 the paper is organized as follows. Section 2 outlines the proposed two-step procedure with the aim of forecasting in ultrahigh-dimensional situations. The main 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 modelling 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 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 a high-dimensional setting. We start by augmenting the original predictors 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. 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 explanatory 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 is 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 series 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_{ti}, $i = 1, \ldots, N$:

$$y_{t+h} = \beta_{0i} + \beta_{1i}x_{ti} + \beta_{3i}x_{tk}\mathbb{1}_{\{x_{ti}=x_{tk}x_{tl}\}} + \beta_{4i}x_{tl}\mathbb{1}_{\{x_{ti}=x_{tk}x_{tl}\}} + \varepsilon_{t+h,i}, \qquad t = 1, \ldots, T-h, \quad (1)$$

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 i under consideration is an interaction term between variables k and l. If so, the original variables are also included in the regression, implicating that selection of interaction term means that it has predictive ability in excess of the original variables. 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 (Rossi and Sekhposyan, 2013; Elliott et al., 2013; Samuels and Sekkel, 2013) and, as in this case, alternative screening procedures (Fan and Lv, 2010; Fan et al., 2009; Bai and Ng, 2008; Bair et al., 2006).

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:

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

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, since for instance type I error which is made twice when we test four variables is 'unacceptable' however, it is 'acceptable' when we test few hundreds. Second, the implications of a false positive. Controlling the FDR allows just that, with more false rejections made possible as long as we discover more truly important variables. In contrast with the FWER procedure which is ignorant to the number of true discoveries. Second, although undesirable since we unnecessary inflate estimation noise, implications of false positives are far less severe than, say, approving an ineffective drug for production out of the very many drugs which are tested. In that sense we can be less stringent 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, $i = 1, \ldots, N$, 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

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

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 k and 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

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

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. Indeed, similar concerns voiced by Bickel in Fan et al. (2009). 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 in the first screening step 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 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 as 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 which 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). 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 L2-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 is given by:

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

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 Empirical application: Data, implementation issues 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 U.S. macroeconomic and financial variables at the monthly frequency for the period April 1959 - September 2009. 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 our target variable 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: $y_{t+h,h} = \frac{1200}{h} \ln\left(\frac{\nu_{t+h}}{\nu_t}\right)$, where ν_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 is denoted as y_{t+1}.

3.2 Implementation issues

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 (small), M (medium) and L (large). Note that the lag of the dependent variable is an explanatory variable which is counted in the original 126 variables. This variable is regarded as important and is included in the final subset regardless of it's p-value, for all three specifications, S, M and L.

To determine an appropriate value for the ridge parameter λ in (7) we use a data-driven strategy based on cross-validation (CV), for a recent survey see Arlot and Celisse (2010). 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 such as BIC. A full-blown (leave-one-out) CV procedure would computationally be very costly, mostly 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. We modify the procedure in a way that honours 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). Specifically, we obtain h-month ahead forecasts from the ridge regression (using the pre-selected subset of the predictors) according to a fine grid of different λ values. At time t we select the value of λ that delivers the smallest root mean squared prediction error (RMSE) for the most recent 12 forecasts. In situations where the number of variables chosen in the initial screening step is smaller than six (corresponding with 5% of the window length) no shrinkage is applied, i.e. the penalty parameter λ is set automatically to zero. Note that this procedure implies that the ridge parameter may vary over time as well as across forecast horizons. For computing we use the *glmnet* in R software (Friedman et al., 2010).

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 fix the lag length p at 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 (2002), 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

$$y_{t+h,h} = \boldsymbol{\alpha}'\boldsymbol{z}_t + \boldsymbol{\gamma}'\boldsymbol{f}_t + \varepsilon_{t+h,h}, \tag{10}$$

where $\boldsymbol{z}_t$ is a vector of 'preferred' predictors such as a constant (as in our case) or lagged dependent variable, and $\boldsymbol{f}_t$ is a vector of r factors presumed to properly span the variance in the set of predictors X. Typically the first few principal components of the covariance matrix of X are used for this purpose. In our empirical application we use two different specifications for the set of predictors X: The first consists of the original variables only (abbreviated here as PC), the second also includes their squares (abbreviated here as SPC).[5] The decision regarding the number of principal components to include in (10) is made in a similar fashion to that described above. We create 12 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. [6]

4 Results

In this section we report the forecasting results obtained for the four 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 Root Mean Squared Error (RMSE), by far the most common evaluation metric in the literature (Gneiting, 2011). We also test whether differences in forecast accuracy are statistically significant using a Diebold-Mariano type test statistic (**?**). **?** (GW) show that the test-statistic has the same (standard normal) asymptotic distribution under the null of equal predictive ability even under non-vanishing estimation uncertainty, as is in our rolling window estimation procedure. We do not compare all possible pairs of forecasts, but only examine the best FDR-based forecast versus the

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

$$y_{t+h,h} = \boldsymbol{\alpha}'\boldsymbol{z}_t + \boldsymbol{\gamma}_1'\boldsymbol{f}_t + \boldsymbol{\gamma}_2'\boldsymbol{f}_t^2 + \varepsilon_{t+h,h}.$$

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.

[6] 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. Since computation time is substantially lower, using BIC information criterion is preferred and we use CV only for coherence with the other models.

best forecast in the benchmark category.

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 asterisks indicate a significant improvement of the best FDR-based forecasts over the most accurate benchmark forecasts at the 10% (*) and 5% (**) significance levels.

The table suggests three main conclusions. First, using the FDR-based screening procedure offers, sometimes substantial improvements in forecast accuracy compared to the benchmark methods. The FDR-based forecasts achieve the lowest RMSE in 14 out of the 16 cases considered. The only exceptions are for sales at the $h = 3$ month horizon and for industrial production at the $h = 1$ month horizon, with PC performing better in both cases. While forecasting gains are observed across all four target series and all forecast horizons, it also appears that the improvements are largest at longer horizons. Differences in RMSE between the best FDR-based forecasts and the best benchmark method are small and typically insignificant for $h = 1$ and 3. However, for forecast horizons of $h = 6$ and 12 months the FDR approach leads to more substantial and significant improvements, with gains in RMSE up to 10% 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.

Second, in the context of the FDR-based procedure, allowing for nonlinear relations between the predictors and the target variable improves forecast accuracy, especially when allowing for interaction effects among different original variables. 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 always dominated either by those that include the squares (M) or 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, we find that allowing for interactions improved forecast accuracy for 12 out of the 16 cases considered. Also here we find that especially at longer forecast horizons allowing for more complex nonlinear relations is beneficial.

Third, allowing for nonlinearity 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: Results - out-of-sample accuracy

		Personal Income				Manufacturing & Trade Sales			
	h:	1	3	6	12	1	3	6	12
Benchmarks									
AR(4)		0.692	0.725	0.858	0.869	0.675	0.825	0.860	0.848
PC		0.695	0.708	0.837	0.886	0.656	**0.786**	0.850	0.894
SPC		0.680	0.729	0.870	0.909	0.665	0.797	0.849	0.872
FDR-RR									
S		0.665	0.706	0.819	0.833	0.660	0.793	0.831	0.823
M		**0.664**	**0.700**	0.809	0.824	**0.657**	0.789	0.824	0.817
L		0.683	0.716	**0.779***	**0.794****	0.672	0.817	**0.770***	**0.746****
		Industrial Production				Employment			
	h:	1	3	6	12	1	3	6	12
Benchmarks									
AR(4)		0.856	0.944	0.928	0.836	0.885	1.097	0.991	0.856
PC		**0.784**	0.867	0.854	0.821	0.908	1.132	1.024	0.887
SPC		0.779	0.898	0.911	0.857	0.954	1.257	1.117	0.927
FDR-RR									
S		0.823	0.893	0.87	0.781	1.035	1.296	1.102	0.884
M		0.808	0.877	0.857	0.770	0.984	1.238	1.063	0.865
L		0.807	**0.834**	**0.802**	**0.749***	**0.869**	**1.070**	**0.922**	**0.815**

Note: The table reports root mean squared prediction errors (RMSE) for forecasts of the h-month growth rate over the period May 1970 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. Asterisks indicate a significant difference between the best performing FDR-RR method and the best benchmark method according to a one-sided Diebold-Mariano test at the 10% (*) and 5% (**) significance levels.

We next turn to examine whether the benefits from allowing for non-linear relations between the predictors and the target variable are stable over time. In order to do so, Figure 1 presents 10-year rolling RMSEs of the different FDR-based forecasts relative to the RMSE of PC-based forecasts, for the 12 months horizon. We observe that for all series, including squares of the original predictors (M) delivers similar or smaller RMSE quite consistently. Once interactions are introduced performance is much improved in all four series roughly until 1995. Evidently, this is the main driver for the compelling results presented in Table 1.

Taken together, Figure 1 induces further confidence in our FDR-based forecast method, with dramatic 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.

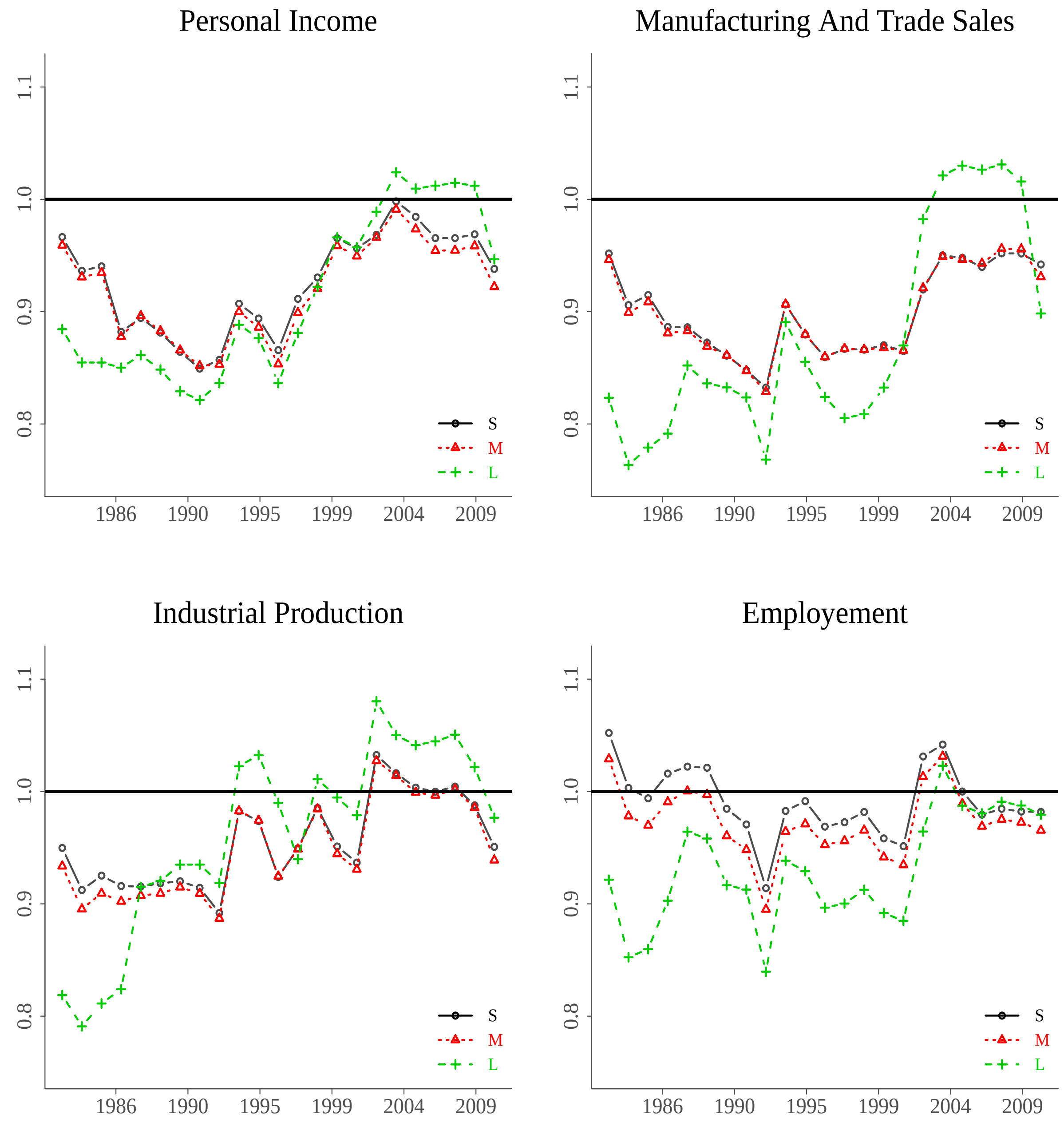

Figure 1: Ten years rolling RMSE. The graph shows the ratio of the RMSE of the $FDR - RR$ method to the RMSE of the PC benchmark model, for the 12 months horizon. The horizontal line at 1 represents equal RMSE between the PC method and the $FDR - RR$ method.

5 Discussion and Conclusion

Our two step $FDR - RR$ procedure can also be viewed as *diversified* shrinkage. There are two groups of variables, the group which was excluded in the first screening step, and the group which is eventually used for forecasting. Each group of corresponding coefficients have its own penalty parameter, coefficients of variables which are excluded in the first step have very high value of λ (which shrinks them to zero), while the λ of the remaining coefficients is determined as described in section 3.1. In that sense, our proposal fits into the framework recently outlined in Stock and Watson (2012) with a specific shrinkage function.

Results presented encourage further research towards other possible ways to allow a beyond-linear relation. The research into this area is not yet abundant, but empirical evidence in this- and other papers suggest a new source of information can be exploited using modern statistical methodology. One such direction may be the exploration of the fast growing literature which combines dimension reduction and sparsity. Included in this are 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).

In sum, we proposed a refinement to the line of research suggested by Fan et al. (2009), and Fan and Lv (2010) in the form of controlling the FDR under general correlation structure. By that, accounting for (1) the large number of variables considered, and (2) the cross-correlation. We demonstrated how to apply it in a *general* ultra-high-dimensional setting using an intuitive and practical two step procedure. Results from our empirical application, targeted towards application in financial and/or macro-economic forecasting, add to the evidence that allowing for a non-linear relation leads to substantial accuracy gains.

References

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

Ashley R. 1998. A new technique for postsample model selection and validation. *Journal of Economic Dynamics and Control* **22**: 647–665.

Bach FR. 2008. Bolasso: model consistent lasso estimation through the bootstrap. In *Proceedings of the 25th international conference on Machine learning.* ACM, 33–40.

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

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

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

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

Benjamini Y, Hochberg Y. 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.

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

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

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

Carriero A, Kapetanios G, Marcellino M. 2011. Forecasting large datasets with bayesian reduced rank multivariate models. *Journal of Applied Econometrics* **26**: 735Ű761.

De Mol C, Giannone D, Reichlin L. 2008. Forecasting using a large number of predictors: Is Bayesian shrinkage a valid alternative to principal components? *Journal of Econometrics* **146**: 318–328.

Doan LR T, Sims C. 1984. Forecasting and conditional projection using realistic prior distributions. *Econometric Reviews* **3**: 1–144.

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

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

Exterkate P, Groenen PJ, Heij C, van Dijk D. 2013. Nonlinear forecasting with many predictors using kernel ridge regression. CREATES Research Papers 2013-16, School of Economics and Management, University of Aarhus.

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

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

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

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

Friedman J, Hastie T, Tibshirani R. 2010. Regularization paths for generalized linear models via coordinate descent. *Journal of statistical software* **33**: 1.

Fu WJ. 1998. Penalized regressions: The bridge versus the lasso. *Journal of Computational and Graphical Statistics* **7**: 397–416.

Geweke J, Whiteman C. 2006. Bayesian forecasting. *Handbook of economic forecasting* **1**: 3–80.

Ghosh JK. 2009. Bayesian methods: A social and behavioral sciences approach, second edition by jeff gill. *International Statistical Review* **77**: 301–302.

Giovannelli A. 2012. Nonlinear forecasting using large datasets: Evidences on us and euro area economies. *CEIS Working Paper* .

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

Hesterberg T, Choi NH, Meier L, Fraley C. 2008. Least angle and l1 penalized regression: A review. *Statistics Surveys* **2**: 61–93.

Hoerl AE, Kennard RW. 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* : 65–70.

Kim HH, Swanson NR. 2013. Forecasting financial and macroeconomic variables using data reduction methods: New empirical evidence. *Journal of Econometrics* : to appear.

Koop GM. 2013. Forecasting with medium and large bayesian vars. *Journal of Applied Econometrics* **28**: 177–203.

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

Mönch E. 2008. Forecasting the yield curve in a data-rich environment: A no-arbitrage factor-augmented var approach. *Journal of Econometrics* **146**: 26–43.

Pesaran MH, Pettenuzzo D, Timmermann A. 2006. Forecasting time series subject to multiple structural breaks. *Review of Economic Studies* **73**: 1057–1084.

Rossi B, Sekhposyan T. 2013. Evaluating predictive densities of us output growth and inflation in a large macroeconomic data set. *International Journal of Forecasting* : to appear.

Samuels JD, Sekkel R. 2013. Forecasting with many models: Model confidence sets and forecast combination. Working Papers 13-11, Bank of Canada.

Stock JH, Watson MW. 1999. Forecasting inflation. *Journal of Monetary Economics* **44**: 293–335.

Stock JH, Watson MW. 2002. Macroeconomic forecasting using diffusion indexes. *Journal of Business and Economic Statistics* **20**: 147–162.

Stock JH, Watson MW. 2005. Implications of dynamic factor models for var analysis. Technical report, National Bureau of Economic Research.

Stock JH, Watson MW. 2006. *Forecasting with Many Predictors*, volume 1 of *Handbook of Economic Forecasting*, chapter 10. Elsevier, 515–554.

Stock JH, Watson MW. 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. B* **58**: 267–288.

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

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

Yu WC, Salyards DM. 2009. Parsimonious modeling and forecasting of corporate yield curve. *Journal of Forecasting* **28**: 73–88.

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

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

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