Forecasting with Sufficient Dimension Reductions

Alessandro Barbarino* and Efstathia Bura†

September 14, 2015

Abstract

Factor models have been successfully employed in summarizing large datasets with few underlying latent factors and in building time series forecasting models for economic variables. When the objective is to forecast a target variable y with a large set of predictors $\mathbf{x}$, the construction of the summary of the $\mathbf{x}$s should be driven by how informative on y it is. Most existing methods first reduce the predictors and then forecast y in independent phases of the modeling process. In this paper we present an alternative and potentially more attractive alternative: summarizing $\mathbf{x}$ as it relates to y, so that all the information in the conditional distribution of $y|\mathbf{x}$ is preserved. These y-targeted reductions of the predictors are obtained using Sufficient Dimension Reduction techniques. We show in simulations and real data analysis that forecasting models based on sufficient reductions have the potential of significantly improved performance.

JEL Classification Number: C32, C53, C55, E17

Keywords: Forecasting, Factor Models, Principal Components, Partial Least Squares, Dimension Reduction, Diffusion Index.

1 Introduction

Methods able to identify and estimate a condensed latent structure that summarizes a large set of variables with few "factors" attracted attention early on in Statistics (Hotelling, 1933 [40]). The Economics and Econometrics literature caught on with investigations ranging from the estimation of the underlying factors that drive the economy, as in Geweke (1977) [34] and Sargent and Sims (1977) [51], to the description of asset prices as in Chamberlain and Rothschild (1983) [16], to applications in labor markets as in Engle and Watson (1981) [31]. In this body of work, the common thread is the usage of **factor models** and the focus is on identifying the common latent sources of correlation of a set of given variables without explicit reference to a target variable.

In part due to the availability of richer datasets and also to the seminal work of Stock and Watson (2002a) [54], factor models under the form of Dynamic Factor Models (DFM) have resurfaced in the Econometric literature in the past 15 years and are now a standard tool for both measuring comovement and forecasting time series. In contrast to the early applications of factor models and their use to measure comovement, a distinctive feature of applying DFM in forecasting is the inherent targeted nature of the process. Departing from being simply a tool to identify and estimate a latent structure in multivariate data, DFMs aim to 1) reduce the dimension of a large

*Alessandro Barbarino, Research and Statistics, Federal Reserve Board. *Address*: 20th St. & C St. NW Washington, DC 20551 USA. *Email*: alessandro.barbarino@frb.gov

†Efstathia Bura, Department of Statistics, George Washington University. *Address*: 801 22nd St. NW Washington, DC 20052 USA. *Email*: ebura@gwu.edu

panel of data to a sufficiently "lean" factor structure and 2) exploit such structure to forecast a target variable y.

Targeting comes into the picture only after a condensed latent structure is estimated and is resolved by postulating a linear relationship between the target variable y and the factors. The reduction step has so far been largely disconnected from the targeting step, likely a legacy of the origin of factor models.

Sufficient Dimension Reduction (SDR), a parallel, yet so far unrelated, collection of novel tools for reducing the dimension of multivariate data in regression problems without losing inferential information on the distribution of a target variable y, has emerged over the last twenty five years in Statistics. SDR focuses on finding sufficient (in a statistical sense) reductions of a potentially large set of explanatory variables with the aim of modeling a given target response y. In SDR, the reduction and the targeting are obtained simultaneously by exploiting the concept of statistical sufficiency. SDR aims to identify and estimate a sufficient function of the regressors, $\mathbf{R}(\mathbf{x})$, that preserves the information in the conditional distribution of $y|\mathbf{x}$. SDR also offers a powerful toolbox to analyze the link between the target y and the panel of regressors $\mathbf{x}$ in contrast to the potentially unjustified assumption that y depends linearly on some factors as in the DFM literature.

Since SDR methods preserve the information of the conditional distribution of $y|\mathbf{x}$ it should prove superior to current practice in producing density forecasts although we do not explore this aspect in this paper.

SDR is not the only approach that allows for simultaneous reduction and targeting, and the potential gains that can be obtained from linking the two modeling steps have been already acknowledged in the Econometrics literature. Bai and Ng (2008) [4] and De Mol, Giannone and Reichlin (2008) [25] explore the effectiveness of RIDGE regression, a penalty based regression, and other shrinkage estimators including the LASSO and execute their papers on the canvas laid out by Stock and Watson (2002b and 2002b) [54] [55]. In more recent contributions, Kelly and Pruit (2014)[43] and Groen and Kapetanios (2014) [35] explore variants of partial least squares (PLS) in order to obtain simultaneous reduction and targeting. In contrast to the innovative (in the Econometrics literature) statistical learning tools explored by these authors, SDR methods ensure the preservation of all the statistical information contained in the data as encapsulated in the conditional distribution of $y|\mathbf{x}$. SDR methods also allow to selectively preserve targeted statistical information regarding the conditional distribution of $y|\mathbf{x}$, such as the conditional mean, the conditional variance, or both, and clearly specify the assumptions required for the proper extraction of such information. In contrast, RIDGE regression is constrained by the forecasting functional form of the model, whereas PLS builds predictors that are little understood and their reliability ultimately depends on the particular application at hand using, quite arbitrarily, the covariance of the response and the original predictors.

Although the SDR methodology offers a potentially powerful source of new methods for the econometrician, several hurdles need be overcome. SDR has been developed and tested with statistical modeling in mind and its effectiveness has not been tested and proven in macro-forecasting. Most importantly, SDR has been developed for cross-sectional applications and adaptations are necessary for a successful application in forecasting and econometrics, analogously to the contributions of Stock and Watson (2002a and 2012b)[54] [55], Bai and Ng (2003) [2] and Forni, Hallin, Lippi and Reichlin (2005) [32] that enabled the application of principal components to a time series setting.

This paper takes a first stab at introducing SDR techniques to macro-forecasting by establishing a connection between the SDR and DFM bodies of literature, by extending some key SDR results to a time series setting and by offering a first assessment of the effectiveness of SDR methods in a real-world forecasting application. In order to draw analogies and highlight differences with results

in the DFM literature we conduct our real-world forecasting experiment with a large panel of macro variables as in Stock and Watson (2002a) [54] and Stock and Watson (2002b) [55] although our data source is the novel repository FRED-MD maintained by the St. Louis Fed and documented by McCracken and Ng (2015) [49]. The task of conducting extensive Monte-Carlo simulations drawing from Stock and Watson (2002a) [54] and Doz et al. (2012) [28] in order to compare the performance of competing methods is deferred to a companion paper (see Barbarino and Bura (2015) [7]).

In Section 2 we list the challenges of macro-forecasting in a data rich environment and describing the specific solution adopted in the DFM framework. We next propose an alternative forecasting framework based on SDR methods and contrast it with the DFM framework providing a connection between the two. In Section 3 we review shrinkage estimators that have been proposed within the DFM literature and that are tested in the empirical application. Section 4 contains the conceptual motivation for targeted reductions. Section 5 is a detailed exposition of the principles of SDR and our proposal for an SDR-based forecasting framework, including extensions to a time-series setting of sliced inverse regression (SIR), the SDR method we choose to present and apply in the empirical Section. Finally, as in Stock and Watson (2002a and 2002b) [54] [55] Section 6 contains the description and results of a horserace between the estimators reviewed in the paper on a large panel of macro variables in which the focus is on forecasting accuracy in predicting inflation and industrial production in an out-of-sample forecasting experiment. We find that SIR has similar or superior performance to PCR and always superior to PLS. The last Section concludes. Coverage of likelihood-based SDR methods is outside the scope of this exploratory paper and deferred to future work.

2 Forecasting with a Large Set of Explanatory Variables

A large set of p explanatory variables $\underset{(p\times 1)}{\mathbf{x}_t}$ is available to forecast a single variable $\underset{(1\times 1)}{y_t}$ using a sample of size T. The most immediate approach to the problem, as described, for instance, in the survey of Stock and Watson (2006) [57], is to consider all regressors to be potentially useful in modeling y_t and use OLS to estimate the model

$$y_{t+h} = \boldsymbol{\beta}'\mathbf{x}_t + \boldsymbol{\gamma}'\mathbf{w}_t + \varepsilon_{t+h} \tag{2.1}$$

where $\mathbf{w}_t$ may contain additional regressors such as lags of y_t[1]. Notice that although in this set-up all regressors are potentially useful in forecasting, they enter the model through just one linear combination namely $\boldsymbol{\beta}'\mathbf{x}_t+\boldsymbol{\gamma}'\mathbf{w}_t$. More than one linear combinations or projections of the regressors may instead be necessary to model y_t in order to preserve all the information that the covariates $\mathbf{x}_t$ and $\mathbf{w}_t$ carry about y_{t+h}. We will revisit this point later on in the paper.

Estimation in (2.1) via OLS can be problematic when p is large relative to T, or variables in $\mathbf{x}_t$ are nearly collinear, as is the case in the macro forecasting literature (see, e.g., Stock and Watson (2006) [57]). The variance of the prediction is of order p/T, so when p is large other estimation methods may dominate OLS, even under assumptions that guarantee that the OLS estimator is unbiased. Moreover, when $p > 3$, the OLS estimator is not even admissible under the mean square error (MSE) criterion,[2] a striking result by James and Stein (1961) that inaugurated research in shrinkage estimation.

[1] The assumed linearity of $E(y_{t+h})$ in the linear combinations $\beta'\mathbf{x}_t$ and $\gamma'\mathbf{w}_t$ and the presumed lack of dependence of the latter from the error ε_{t+h} are important assumptions that lead to ordinary least squares (OLS) as the estimator of choice for the parameters $\boldsymbol{\beta}$ and γ. The presence of the lags of y_t is meant to capture the dynamics in y_t and to avoid that such dynamics might impart non-orthogonality between the error and the regressors.

[2] Although among unbiased estimators the OLS estimator $\widehat{\theta}_{OLS}$ has minimum mean squared error, other estimators

When the number of predictors exceeds the number of observations ($p > T$), the OLS parameter estimates are not identifiable and multiple estimators of the parameter vector are solutions to the OLS problem. In this case, different shrinkage estimators, such as RIDGE and LASSO, can be viewed as specific criteria-based choices in the space of OLS-consistent solutions.

Furthermore, for typical datasets encountered in macro and finance forecasting, even if $p << T$, collinearity or ill-conditioning of the $\underset{(T\times p)}{\mathbf{X}_t}$ matrix, which collects all the observations on the vector $\mathbf{x}_t$, makes OLS predictions very unstable. This is likely to occur if the set of explanatory variables contains an index and several sub-indexes, e.g. industrial production (IP) along with its sub-indexes such as manufacturing IP or mining IP, or when variables are linked by identities or tight relationships, such as the inclusion of assets linked by arbitrage conditions.

2.1 The DFM Forecasting Framework

Pioneering results in Stock and Watson in a series of papers (1999, 2002a, 2002b) [53][54][55], relaunched the idea of working around the impossibility and/or lack of desirability of using OLS for estimation when p is large by:

(i) Positing that the set of explanatory variables $\mathbf{x}_t$ is, up to idiosyncratic noise, driven by a small $r < p$ set of latent factors $\underset{(r\times 1)}{\mathbf{f}_t}$ with

$$\mathbf{x}_t = \mathbf{\Lambda}\mathbf{f}_t + \mathbf{u}_t \tag{2.2}$$

where $\mathbf{f}_t$ and $\mathbf{u}_t$ are independent although $\mathbf{u}_t$ can be serially and cross-sectionally correlated. This setup generalizes the case of a classic factor structure.

(ii) Assuming that the forecasting model is

$$y_{t+h} = \boldsymbol{\lambda}'\mathbf{f}_t + \boldsymbol{\gamma}'\mathbf{w}_t + \varepsilon_{t+h} \tag{2.3}$$

The primary goal of considering the factors $\mathbf{f}_t$ is to introduce structure that reduces the high-dimensionality of the problem. The linear factor structure implies that although p is potentially large, the information content of the regressors is drastically reduced to r. However the factors are latent variables hence they need to be estimated. The complexity induced by the high dimensionality of the problem is traded off with the need of estimating additional fictional unobserved variables $\mathbf{f}_t$. Stock and Watson [54] provide conditions under which the factors are estimated via Principal Component Analysis (PCA) of the variance-covariance matrix of the observed predictors.

In fact, most estimators of the factors proposed in the literature turn out to be linear functions of the explanatory variables, $\mathbf{v}'\mathbf{x}_t$, where the matrix $\mathbf{v}$ of coefficients or weights corresponds to the particular method, e.g. PCR, RIDGE or PLS. The rank of the column space of $\mathbf{v}$ is the dimension of the problem detected by that particular estimation method. In this sense there is no unique dimension of the information contained in the explanatory variables. Rather the dimension of the problem is estimator-dependent.

When the factors are estimated via PCA, as in Stock and Watson (2002a) [54], $\mathbf{v}$ is a matrix whose columns are the leading principal components of $\mathbf{\Sigma} = (\mathbf{X}_T - \bar{\mathbf{X}}_T)'(\mathbf{X}_T - \bar{\mathbf{X}}_T)/T$. If the

$\widehat{\theta}$, some also linear, have uniformly smaller mean square error, by trading off bias for variance :

$$MSE\left(\widehat{\theta}, \theta\right) \leq MSE\left(\widehat{\theta}_{OLS}, \theta\right)$$

for all θ with strict inequality for some θ. Hence the OLS estimator is not admissible under the MSE criterion.

dynamic principal components of Forni, Lippi Hallin and Reichlin (2005) [32] are used, $\mathbf{v}$ also captures a summary of the dynamics in $\mathbf{x}_t$ as it contains the eigenvectors of the autocovariances $\mathbf{\Sigma}^{(k)} = (\mathbf{X}_{T-k} - \bar{\mathbf{X}}_{T-k})'(\mathbf{X}_T - \bar{\mathbf{X}}_T)/(T-k)$. In (Quasi-) Maximum Likelihoods methods as in Doz et al. (2012) [28] or Banbura and Modugno (2014) [6], the algorithm used to compute the likelihood heavily exploits the linearity of the underlying system and $\mathbf{v}$ is derived from the Kalman Filter. A main theoretical objective in the DFM literature has been to show that asymptotically, for both T and $p \to \infty$, the chosen $\mathbf{v}$, is such that $\mathbf{v}'\mathbf{x}_t$ consistently estimates the factors.

The forecasting equation (2.3) is secondary since attention is shifted to reducing the dimension of the set of explanatory variables, a process assumed to unveil the data generating process driving $\mathbf{x}_t$. The practical goal of forecasting the target y_t via **link equation** (2.3) relies on the assumption that the same factors that determine the marginal distribution of $\mathbf{x}_t$ also determine the conditional distribution of y_t and in the same functional form; that is, linearly.

The assumption of a factor structure and the ancillary role of the forecasting equation in Stock and Watson (2002a) [54] was adopted in the ensuing literature. For instance, Doz et al. (2012) [28] focus on the performance of different estimators in identifying the factors and show no interest in the forecasting accuracy of their estimators.

The assumption that the underlying DGP has a linear factor structure, while convenient, imposes restrictions on the conditional distribution of y given $\mathbf{x}$, which are difficult to pin down. In a follow-up to this paper (Barbarino and Bura (2015) [7]), we show by means of Monte-Carlo simulations and formally by exploiting recent results in Leeb (2013) [44] and, more importantly, the extension in Steinberger and Leeb (2015) [52] that a linear factor structure underlying the generation of both y and $\mathbf{x}$ coupled with the assumption of joint normality of the factors implies that a linear model is the correct specification for the conditional mean of y given $\mathbf{x}$, which is a rather restrictive model.

The formal result is stated in the following proposition.

Proposition 1 *Let y denote a response random variable and $\mathbf{x}$ a random p-vector of explanatory variables. Suppose that the response can be written as $y = \boldsymbol{\alpha}'\mathbf{f}$ where $\boldsymbol{\alpha} \in \mathbb{R}^r$ is unknown and that the set of explanatory variables can be written as $\mathbf{x} = \underset{(p\times r)}{\boldsymbol{\beta}}\mathbf{f}$ where $\boldsymbol{\beta}$ is such that:*

(i) for each $\boldsymbol{\alpha}$, the conditional mean of $\boldsymbol{\alpha}'\mathbf{f}$ given $\boldsymbol{\beta}'\mathbf{f} = \mathbf{x}$ is linear in $\mathbf{x} \in \mathbb{R}^p$;

(ii) for each $\boldsymbol{\alpha}$ the conditional variance of $\boldsymbol{\alpha}'\mathbf{f}$ given $\boldsymbol{\beta}'\mathbf{f} = \mathbf{x}$ is constant in $\mathbf{x} \in \mathbb{R}^p$.

Then, y can be decomposed into the sum of a linear function of $\mathbf{x}$ and a remainder or error term, as follows,

$$y = \mathbf{c}'\mathbf{x} + u \tag{2.4}$$

where $\mathbf{c} = \boldsymbol{\beta}\boldsymbol{\alpha} \in \mathbb{R}^p$ is an unknown parameter, $\mathrm{E}(u|\mathbf{x}) = 0$ and $\mathrm{var}(u|\mathbf{x})$ is constant.

Proof. *See Barbarino and Bura (2015) [7].* ■

A key point to note is that conditions (i) and (ii) in Proposition 1 are satisfied for any $\boldsymbol{\beta}$, if $\mathbf{f}$ is normally distributed, the assumption made in practice in the DFM literature especially likelihood-based but also non-parametric. Moreover, the implied model (2.4) is a standard linear model with the error term uncorrelated with the predictors so that the solution to the normal equations in the population is the OLS, $\mathbf{c}_{OLS} = (\mathbf{x}'\mathbf{x})^{-1}\mathbf{x}'y$. For a random sample, regressing $\underset{(T\times 1)}{\mathbf{y}}$ on $\underset{(T\times p)}{\mathbf{X}}$ using OLS yields $\hat{\mathbf{c}}_{OLS} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}$, which is optimal, in terms of statistical

efficiency, provided the sample size is larger than the number of predictors, i.e. $p < T$. Therefore the DFM model assumptions (2.2) and (2.3) imply that the forecasting model is approximately linear in the explanatory variables with errors uncorrelated with the predictors. As a consequence it should not come as a surprise that shrinkage methods such as RIDGE perform well in our simulations. Although ([52]) is valid when $(y_t, \mathbf{x}_t)$ is a random sample for $t = 1, \ldots, T$, on the basis of our simulations, we conjecture that the results in Proposition 1 are approximately true in a factor model, under the set of assumptions commonly made in the DFM literature where both the response and the predictors are potentially autocorrelated time series.

2.2 The SDR Forecasting Framework

In contrast to DFM, sufficient dimension reduction (SDR) methods depart from the pervasive linear factor assumption. As we show in greater detail in a companion paper (Barbarino and Bura (2015) [7]) and summarize in the following paragraph, SDR methods thrive when the relationship between the response and the predictors contains non-linearities whereas they do not have comparative advantage when a linear factor structure is the true DGP.

SDR works directly with observables and sidestep the assumption of a factor structure. Thus, instead of imposing an artificial latent factor structure on the panel $\mathbf{x}_t$, SDR methods **directly** seek to identify how many and which functions of the explanatory variables are needed to fully describe the conditional distribution function $F(y_{t+h}|\mathbf{x}_t)$, or its features, such as the conditional mean. SDR aims to identify and estimate functions of the predictors, $\mathbf{R}(\mathbf{x}_t)$, that are called reductions because they preserve all the information that $\mathbf{x}_t$ carries about y_{t+h} in the sense that $F(y_{t+h}|\mathbf{x}_t) = F(y_{t+h}|\mathbf{R}(\mathbf{x}_t))$. Obviously only if such functions are fewer than p do they represent proper reductions.

The reductions can be either linear or nonlinear functions. In this paper we focus on **moment-based** and **linear** SDR methods that obtain linear reductions in order to draw a more pertinent comparison with the DFM literature. Linear SDR methods lay down conditions under which it is possible to identify the number of and the *linear* combinations of the explanatory variables needed to "adequately" model y_t. They also provide estimation algorithms. More formally, linear SDR methods provide the means to estimate a matrix $\mathbf{v} : p \times d$, $0 \leq d \leq p$, such that $\mathbf{R}(\mathbf{x}_t) = \mathbf{v}'\mathbf{x}_t$.

Our proposed SDR-based forecasting framework is based on the following two conditions:

(i) The linearity condition

$$E\left[\mathbf{x}_t|\mathbf{v}'\mathbf{x}_t\right] = \mathbf{A}\mathbf{v}'\mathbf{x}_t \tag{2.5}$$

for any invertible matrix $\mathbf{A}$ and a $p \times d$ full rank matrix $\mathbf{v}$ with $0 \leq d \leq p$[3]

(ii) The **forward** model

$$y_{t+h} = g(\mathbf{v}'\mathbf{x}_t, \varepsilon_{t+h}) \tag{2.6}$$

Moment-based SDR methods place conditions on the marginal distribution of $\mathbf{x}_t$, such as the linearity assumption (2.5), which is a first moment condition and analogous to the linear factor structure (2.2) of DFM.[4] However, in contrast with the DFM setup, no dependence on underlying factors is postulated.

[3] The rank of $\mathbf{v}$ is the *structural dimension* of the regression and the case $d = 0$ signifies that y_{t+h} is independent of $\mathbf{x}_t$.

[4] Within the SDR literature the term "moment-based" catalogues estimators conceptually distinct from SDR "likelihood-based" methods that require assumptions on the distribution of $\mathbf{x}_t|y_{t+h}$. Likelihood-based SDR methods can be compared to likelihood based estimation methods for DFM however we do not pursue such comparison in this paper for clarity purposes.

The second equation (2.6) specifies the **forward** model and is analogous to the link equation in DFM, although the SDR framework allows more flexibility admitting a general $g(\cdot)$ instead of a linear function. Linear SDR methods are powerful tools that can determine the number of linear combinations of the explanatory variables $\mathbf{x}_t$ needed to model the response y_t and provide consistent estimators without the need to specify the functional form of the forward model; that is, without specifying the exact relationship between y_t and $\mathbf{v}'\mathbf{x}_t$. Linear SDR replaces a large number of explanatory variables by a few of their linear combinations without loss of inferential information; their number $d(= \text{rank}(\mathbf{v}))$ indicates the dimension of the regression problem. In our experience, fewer than the number of PCs are needed in order to generate a comparable MSFE, which is expected as the SDR estimator is targeted to y. As a result, the forecaster can concentrate on the estimation of $g(\cdot)$ with the option of also using non-parametric regression since the number of predictors is significantly reduced.

How restrictive is the linearity condition? – The linearity condition is satisfied for **any v** by any elliptically contoured vector of predictors. Importantly, Hall and Li (1993) [36] showed that, as the cross-section gets large and $p \to \infty$, such condition is ever more likely to be satisfied provided that the dimension of the problem d remains small relative to p. More recently, Steinberger and Leeb (2015) [52] showed that as $d/\log p$ becomes smaller, the discrepancy between $E[\mathbf{x}_t|\mathbf{v}'\mathbf{x}_t]$ and $\mathbf{A}\mathbf{v}'\mathbf{x}_t$ in (2.5) goes to zero and the linearity condition holds approximately.

Comparison of SDR and DFM Assumptions – As outlined in Proposition 1, the assumption of a linear factorial structure along with normality imply that a linear model is the correctly specified model. By contrast moment-based SDR requires the linearity condition (2.5) hold for the linear projections $\mathbf{v}'\mathbf{x}$ that satisfy the general forward regression model $F(y|\mathbf{x}) = F(y|\mathbf{v}'\mathbf{x})$. Therefore, DFM assumes more restrictive conditions than SDR on the marginal distribution of $\mathbf{x}$.

Dimension of the Regression – The SDR framework allows for more general models to describe the relationship between y and $\mathbf{x}$. As a consequence, for example, the finding that on the same dataset four PCA-estimated factors produce the same MSFE as two SDR predictors is not contradictory. The PC-based DFM framework ignores non-linearities in the DGP so that a larger number of factors is required to approximate non-linearities in an incorrectly specified forward regression. This is analogous to the effect of dynamic misspecification as shown in Bai and Ng (2007) [3] where one needs a larger set of static factors in order to approximate a given set of (true) dynamic factors. In their setting a larger set of static factors approximate a polynomial in the lag operator. In our setting, a larger number of static factors is needed to approximate non-linearities which are captured by SDR methods.

Remark 1 *When SDR finds that two or more linear combinations are needed ($d \geq 2$) in the forward model (2.6) it means that the forward regression function contains non-linear functions of the linear combinations $\mathbf{v}'\mathbf{x}_t$. Thus, a correct specification of the forward regression entails non-linearities that the DFM framework misses.*

Robustness to Model Misspecification – SDR techniques frequently yield very few derived predictors, which allows non-parametric estimation of the relationship between the response y_t and the (linear combinations of) the explanatory variables $\mathbf{x}_t$. As a result, the prohibitive curse of dimensionality problem in non-parametric regression is circumvented .

Issues in Large p Problems – SDR techniques are data intensive relative to shrinkage estimators such as Principal Components (PC), RIDGE or Partial Least Squares (PLS). As we will see, when the number of predictors p is larger than the sample size, direct application of SIR is not possible

since $rank(\mathbf{\Sigma}) \leq \min(T-1, p)$ and $\mathbf{\Sigma}$ is not invertible. Moreover, when T and p are of the same order, or when the columns of $\mathbf{X}$ are highly correlated, $\mathbf{\Sigma}$ is ill-conditioned and its inverse is numerically unstable resulting in non-robust estimation. In this sense SDR methods suffer from the same limitations as OLS. Therefore we also evaluate a regularized version of our estimators, following Chiaromonte and Martinelli (2002) [22] and Li and Li (2004) [46] who used principal components as an intermediate step in order to eliminate PC directions in which the random vector $\mathbf{x}$ is degenerate. Bernard-Michel et al. (2011) show that preprocessing the data with PC in order to eliminate degenerate projections of $\mathbf{x}$ and then applying SIR is a special case of their Gaussian Regularized SIR, where a Gaussian prior is introduced on the unknown parameters of the inverse regression. In a companion paper (Barbarino and Bura (2015) [7]), we show that even though regularization can be a useful approach in large p settings, a more appealing work-around to the ill-conditioning or non-invertibility of $\mathbf{\Sigma}$ is the extension of SDR methods using Krylov subspaces.

Difficulties in Estimating the Forward Model – SDR approaches obtain optimal results in an interactive modeling setting. That is, the sequence of modeling steps in SDR is to (1) reduce and estimate the $d(< p)$ SDR-derived predictors, and (2) obtain visual assessments via scatterplots of the response versus the SDR-predictors to form a forward regression model $g(\cdot)$ that best describes the data at hand. This process cannot be carried out in an automatic fashion so that $g(\cdot)$ be recursively estimated prior to the computation of the out-of-sample forecast. Instead, we simply use the linear SDR predictors, which are linear combinations of $\mathbf{x}_t$, with no further transformations as input to a linear forward model for the response. As a result, when the estimated dimension of the problem turns out to be greater than one, the forecasting horserace will penalize the SDR estimator by means of misspecification of $g(\cdot)$. Such difficulty is not faced by the applied forecaster as they only need SDR predictors for one-shot forecast (even when repeated over time) in which case they can easily evaluate the presence of non-linearities and decide on the appropriate modeling of $g(\cdot)$.

3 Dimension Reduction via Linear Combinations

Several estimators in the literature form linear combinations of the explanatory variables $\mathbf{v}'\mathbf{x}_t$ as a data reduction step before fitting the model used for prediction. In this Section, we provide a brief review of some such widely used estimators that are also used in the empirical Section. We start with OLS, move on to principal component regression (PCR), which is the prevalent method in dynamic factor analysis, and RIDGE regression. The last leg of our tour provides the fundamentals of the partial least squares (PLS) algorithm. We cast these methods in a shared framework of minimization of an objective function, which is what distinguishes individual methods, and discuss how the resulting estimators exploit the eigen-structure of the data matrix $\mathbf{X}$. A more in-depth treatment under a common unifying framework is discussed in Barbarino and Bura (2015) [7]. To motivate the discussion about the features and relative drawbacks and advantages of different methods, we start off from a simple data generating model, the multivariate normal distribution.

The Normal Data Generating Process – The simplest DGP that implies 1-dimensional linear reduction is a Normal DGP where the predictors and the response are jointly normal:

$$\begin{pmatrix} \mathbf{x} \\ y \end{pmatrix} \sim N\left(\begin{pmatrix} \boldsymbol{\mu}_{\mathbf{x}} \\ \mu_y \end{pmatrix}, \begin{pmatrix} \mathbf{\Sigma} & \boldsymbol{\sigma}_{xy} \\ \boldsymbol{\sigma}'_{xy} & \sigma^2_y \end{pmatrix} \right)$$

Under this assumption the best predictor under quadratic loss is the linear regression function,

$$E\left(y|\mathbf{x}\right) = \mu_y + \boldsymbol{\beta}'\left(\mathbf{x} - \mu_{\mathbf{x}}\right)$$

with $\boldsymbol{\beta} = \boldsymbol{\Sigma}^{-1}\boldsymbol{\sigma}_{xy}$. Therefore, in a Normal DGP the relationship between $\mathbf{x}$ and y is entirely and exhaustively encapsulated in one linear combination of the predictors. We note that even a small departure from this model may result in linear reductions of the predictors no longer being exhaustive (see Bura and Forzani (2015) [14]). Importantly, as shown in Proposition 1, the same DGP is implied by a normal and linear factor structure.

We consider various competing ways of estimating the population parameter $\boldsymbol{\beta}$ next.

Ordinary Least Squares – The OLS coefficient is the solution to the following maximization problem:

$$\max_{\{\beta\}} Corr^2\left(y, \mathbf{x}'\boldsymbol{\beta}\right) \tag{3.1}$$

OLS selects *one and only one* linear combination with the property that it maximizes the correlation between the response and $\mathbf{x}'\boldsymbol{\beta}$. Assuming that $\boldsymbol{\Sigma}$ is full rank, the unique solution to (3.1) is

$$\boldsymbol{\beta}_{OLS} = \boldsymbol{\Sigma}^{-1}\boldsymbol{\sigma}_{xy} \tag{3.2}$$

and the OLS prediction of the response y_{t_0} at an observed $\mathbf{x}_{t_0}$ is

$$y_{OLS} = \mathbf{x}'_{t_0}\boldsymbol{\beta}_{OLS}$$

Principal Component Regression (PCR) – PCR operates in two steps. First, the linear combinations that maximize the variance of $\mathbf{x}$ and that are mutually orthogonal are the solution to the following maximization problem

$$\begin{array}{c} \max \\ \{\mathbf{c}_k\} \\ \boldsymbol{\beta}'_k\boldsymbol{\beta}_k=1 \\ \left\{\boldsymbol{\beta}'_k\boldsymbol{\Sigma}\boldsymbol{\beta}_i=0\right\}_{i=1}^{k-1} \end{array} \quad Var(\mathbf{x}'\boldsymbol{\beta}_k) \tag{3.3}$$

A maximum of p such linear combinations, called principal components, can be extracted. Secondly, y is regressed on the first $M \leq p$ PCs. The number of PCs, M, is a meta parameter chosen by the user. The solution to (3.3) is

$$\boldsymbol{\beta}_{PCR}(M) = \boldsymbol{\Sigma}^{-}_{PCR}(M)\boldsymbol{\sigma}_{xy} \tag{3.4}$$

If $M = p$, then $\boldsymbol{\beta}_{PCR}(M) = \boldsymbol{\beta}_{OLS}$. The pseudo-inverse $\boldsymbol{\Sigma}^{-}_{PCR}(M)$ used to compute the solution, which can be shown to be a Moore-Penrose inverse, will depend on M. Notice that $\boldsymbol{\Sigma}^{-}_{PCR}(M)$ is an estimate of a truncated $\boldsymbol{\Sigma}^{-1}$ by retaining only the first M components that explain a specified amount of variance in the predictors. Therefore, PCA reduces the dimension of the input predictor vector $\mathbf{x}$ by finding orthogonal linear combinations that maximize var $(\mathbf{a}'\mathbf{x})$. Such operation entirely disregards the response y, so that although the $\mathbf{x}$ principal components contain as much information as the entire predictor vector, they are ordered in relevance to $\mathbf{x}$ and not to y. Targeting enters only in the second stage through the term $\boldsymbol{\sigma}_{xy}$ in (3.4). The PCR prediction is

$$y_{PCR} = \mathbf{x}'_{t_0}\boldsymbol{\beta}_{PCR}$$

RIDGE Regression – RIDGE regression has been reviewed in the macro-forecasting literature by De Mol et al. (2008) [25] in connection to Bayesian regression. The RIDGE estimator picks one and only one linear combination of the data as it can be shown that RIDGE regression minimizes the least squares criterion on the sphere with radius a:

$$\max_{\{\beta\}} Corr^2\left(y, \mathbf{x}'\boldsymbol{\beta}\right) \text{ subject to } \sum_{j=1}^{p} \beta_j^2 \leq a \tag{3.5}$$

The solution to (3.6) is

$$\boldsymbol{\beta}_{RR}(\lambda) = (\boldsymbol{\Sigma} + \lambda \mathbf{I})^{-1} \boldsymbol{\sigma}_{xy} \tag{3.6}$$

where λ denotes the Lagrange multiplier in the constrained maximization (3.5) and is a function of a. It is also a meta parameter playing a role similar to M in PCR. When $\lambda = 0$ the maximization is the same as in OLS. As $\lambda > 0$, the solution deviates from the OLS solution and it can be shown that the penalization works to shrink more those directions with smallest variance although RIDGE does not truncate directions in such draconian way as PCR. The relationship with the OLS coefficient (3.2) is

$$\boldsymbol{\beta}_{RR}(\lambda) = \left(\mathbf{I} + \lambda \boldsymbol{\Sigma}^{-1}\right)^{-1} \boldsymbol{\Sigma}^{-1} \boldsymbol{\sigma}_{xy} = \left(\mathbf{I} + \lambda \boldsymbol{\Sigma}^{-1}\right)^{-1} \boldsymbol{\beta}_{OLS}$$

The RIDGE prediction at $\mathbf{x}_{t_0}$ is

$$y_{RR} = \mathbf{x}'_{t_0} \boldsymbol{\beta}_{RR}(\lambda)$$

Partial Least Squares – PLS, an increasingly popular method of dimension reduction, followed a peculiar trajectory in Econometrics. Originally developed by H. Wold [63] in the mid 70s, it did not gain much traction in Econometrics and swiftly fell into oblivion. By contrast, it garnered a lot of attention in Chemometrics, a field that produced a large volume of PLS studies in the late 80s and early 90s (see, for example, the instructive work of Helland (1988) [38]). Only recently has the method resurfaced in Econometrics with the work of Kelly and Pruitt (2014) [43] and Groen and Kapetanios [35] within macro-forecasting applications as PLS handles well cases in which $p > T$.

The PLS algorithm induces a simultaneous bi-linear decomposition of both the target variable and the panel of regressors[5]. That is, factors $\underset{(T\times 1)}{\mathbf{f}_i}$ and loadings $\underset{(p\times 1)}{q_i}$ and $\underset{(1\times 1)}{p_i}$ are generated at each step so that the factors are orthogonal and the following decompositions are carried out concurrently

$$\mathbf{x} = q'_1 \mathbf{f}_1 + q'_2 \mathbf{f}_2 + \ldots + q'_u \mathbf{f}_u + \mathbf{E}_u$$
$$y = p'_1 \mathbf{f}_1 + p'_2 \mathbf{f}_2 + ... + p'_u \mathbf{f}_u + e_u$$

The suffix u denotes the last step of the procedure. The algorithm always converges in the sense that after p steps the factors will be identically zero. Using the recursive formulas for the factors and the loadings, one can show that PLS prediction admits a linear form similar to the predictions of the other estimators at $\mathbf{x}_{t_0}$:

$$y_{PLS} = \mathbf{x}'_{t_0} \boldsymbol{\beta}_{PLS}(u)$$

where

$$\boldsymbol{\beta}_{PLS}(u) = \mathbf{W}_u \left(\mathbf{W}'_u \boldsymbol{\Sigma} \mathbf{W}_u\right)^{-1} \mathbf{W}'_u \boldsymbol{\sigma}_{xy} \tag{3.7}$$

The matrix $\mathbf{W}_u = (\mathbf{w}_1, \ldots, \mathbf{w}_u)$ is obtained after u recursions of the algorithm by stacking the weights generated at each step. Such weights are initialized with $\mathbf{w}_1 = \boldsymbol{\sigma}_{xy}$, and, for $u > 1$,

$$\mathbf{w}_u = \boldsymbol{\sigma}_{xy} - \boldsymbol{\Sigma} \mathbf{W}_{u-1} \left(\mathbf{W}'_{u-1} \boldsymbol{\Sigma} \mathbf{W}_{u-1}\right)^{-1} \mathbf{W}'_u \boldsymbol{\sigma}_{xy}$$

generates the subsequent weights. Notice that the weights are "weighted" covariances of the predictors and the response.

[5] In the appendix we show how the decomposition naturally leads to the factorial structure.

4 Reductions as Eigen-Decompositions

In order to set the stage for the sequel, we now focus on the different targets of the eigen-decompositions the methods in Section 3 entail.

Eigen-Decompositions Underpinning PCR and PLS – PCR targets the extraction of derived inputs that maximize the variance of the explanatory variables. The PCs are left eigenvectors from the eigen-decomposition of $\text{var}(\mathbf{x})$. PLS has (initial) target $\text{cov}(\mathbf{x}, y)$, which reveals its targeted nature and that, by focusing on the principal directions of $\text{cov}(\mathbf{x}, y)$, it is hard-wired to extract **linear signals**.[6]

Eigen-Decomposition Underpinning SDR – SDR methods that will be introduced in the next Section carry out an eigen-decomposition of a target, called the **seed** or **kernel**, in order to isolate directions of principal variation in relevance to y.

The SDR method used in the empirical applications, sliced inverse regression (SIR), is based on the eigen-decomposition of $\text{var}[E(\mathbf{x}|y)]$. To gain intuition of why such target works, we resort to a classical probabilistic identity satisfied by any random vector $\mathbf{x}$ with finite second moment and conditioning random variable or vector y,

$$\underbrace{\text{var}(\mathbf{x})}_{\text{identified by PCA}} = \underbrace{\text{var}[\text{E}(\mathbf{x}|y)]}_{\text{identified by linear SDR}} + \underbrace{\text{E}[\text{var}(\mathbf{x}|y)]}_{\text{noise}} \tag{4.1}$$

Suppose the range of y is sliced in non-overlapping bins and $\mathbf{x}|y$ is the restriction of $\mathbf{x}$ in the bins defined by the slices of y. The right hand side of (4.1) obtains that the variance of $\mathbf{x}$ can be split into two parts:

- $\text{var}[\text{E}(\mathbf{x}|y)]$ or *between* slice variation in $\mathbf{x}$, and
- $\text{E}[\text{var}(\mathbf{x}|y)]$ or *within* slice variation.

In ANOVA, the first summand is the **signal** that y carries about $\mathbf{x}$ since it represents variation of the average value of $\mathbf{x}$ associated with different values of y from the overall $\mathbf{x}$ mean. The second element represents **noise**, i.e. deviations of $\mathbf{x}$ from its overall average across bins, hence unrelated to y.

From this perspective, since PCA performs an eigenanalysis of $\text{var}(\mathbf{x})$, noise in $\text{E}[\text{var}(\mathbf{x}|y)]$ may attenuate or suppress the signal in $\text{var}[\text{E}(\mathbf{x}|y)]$ and result in PCs that are little related to y. PLS targets $\text{cov}(\mathbf{x}, y)$ and can potentially suppress non-linear signal. By contrast, a method that focuses on the eigen-analysis of $\text{var}[E(\mathbf{x}|y)]$ produces derived inputs ordered according to their importance with respect to y and has the capacity to preserve non-linear signals. As we will see next, centering on the signal and ignoring the noise is what sufficient dimension reduction in general and, in particular, sliced inverse regression is designed to do.

4.1 Sufficiency and Inverse Reductions

Definition 2 *A reduction $\mathbf{R} : \mathbb{R}^p \to \mathbb{R}^q$, where $q \leq p$, is sufficient if it satisfies $y|\mathbf{x} \sim y|\mathbf{R}(\mathbf{x})$ or equivalently*

$$F(y|\mathbf{x}) = F(y|\mathbf{R}(\mathbf{x})) \tag{4.2}$$

[6] To be precise, when y is a scalar, $\text{cov}(\mathbf{x}, y)$ is a vector and its eigen-decomposition returns the vector itself. However it is useful to think of the PLS algorithm as a sequence of eigen-decompositions when comparing the methodology with PCA. When y is multivariate the PLS algorithm entails eigen-decomposition of the matrix $\text{cov}(\mathbf{x}, y)$.

A consequence of the definition of sufficiency is that, since (4.2) can be written as $F(y|\mathbf{x},\mathbf{R}(\mathbf{x})) = F(y|\mathbf{R}(\mathbf{x}))$ we have

$$y \perp\!\!\!\perp \mathbf{x}|\mathbf{R}$$

Consequently, $\mathbf{R}(\mathbf{x})$ is called a **forward reduction.** Although the term "sufficient" was originally coined to highlight the information preserving role of $\mathbf{R}(\mathbf{x})$, it turns out that there is a specific link with the Fisherian concept of statistical sufficiency (see Cook (2007)) [19] as we will see shortly. Before doing so, we introduce the concept of **inverse reduction**.

Definition 3 *A function $\mathbf{R}: \mathbb{R}^p \to \mathbb{R}^q$, where $q \leq p$, is an inverse reduction if*

$$\mathbf{x}|(\mathbf{R}(\mathbf{x}), y) \overset{d}{=} \mathbf{x}|\mathbf{R}(\mathbf{x}) \tag{4.3}$$

If one treats y as a parameter, (4.3) states that $\mathbf{R}(\mathbf{x})$ is a sufficient statistic for y and it contains all information $\mathbf{x}$ contains about y. Thus, it is a *sufficient reduction* for the *forward* regression of y on $\mathbf{x}$. Proposition 2 provides the formal statement and proof of this fact.

Proposition 2 *Assume that the random vector $\left(y, \mathbf{x}'\right)'$ has a joint distribution and let $\mathbf{R}(\mathbf{x})$ be a measurable function of the predictor vector. Then,*

$$F(y|\mathbf{x}) = F(y|\mathbf{R}(\mathbf{x})) \quad \textit{iff} \quad \mathbf{x}|(\mathbf{R}(\mathbf{x}), y) \overset{d}{=} \mathbf{x}|\mathbf{R}(\mathbf{x})$$

Proof. *Denote $\mathbf{R}(\mathbf{x})$ with $\mathbf{R}$. Assume $F(y|\mathbf{x}) = F(y|\mathbf{R}(\mathbf{x}))$ so that $y \perp\!\!\!\perp \mathbf{x}|\mathbf{R}$ and $F(y,\mathbf{x}|\mathbf{R}) = F(\mathbf{x}|\mathbf{R})F(y|\mathbf{R})$. Therefore,*

$$F(\mathbf{x}|\mathbf{R}) = \frac{F(y,\mathbf{x}|\mathbf{R})}{F(y|\mathbf{R})} = \frac{F(y,\mathbf{x},\mathbf{R})}{F(y|\mathbf{R})F(\mathbf{R})} = \frac{F(y,\mathbf{x},\mathbf{R})}{F(y,\mathbf{R})} = F(\mathbf{x}|y,\mathbf{R})$$

To prove the reverse statement we start with the definition of conditional distribution of $y|(\mathbf{x},\mathbf{R})$

$$F(y|\mathbf{x},\mathbf{R}) = \frac{F(y,\mathbf{x},\mathbf{R})}{F(\mathbf{x},\mathbf{R})} = \frac{F(\mathbf{x}|y,\mathbf{R})F(y,\mathbf{R})}{F(\mathbf{x}|\mathbf{R})F(\mathbf{R})}$$

Using the condition $\mathbf{x}|(\mathbf{R}(\mathbf{x}), y) \overset{d}{=} \mathbf{x}|\mathbf{R}(\mathbf{x})$, which is equivalent to $F(\mathbf{x}|y,\mathbf{R}) = F(\mathbf{x}|\mathbf{R})$ and simplifying, one obtains $F(y|\mathbf{x},\mathbf{R}) = F(y|\mathbf{R})$. ■

Proposition 2 sheds light on why inverse regression is a powerful tool for the identification of sufficient reductions of the predictors: if a function $\mathbf{R}(\mathbf{x})$ is a sufficient *statistic* for the inverse regression, it is also a sufficient *reduction* for the forward regression. This implies that the econometrician is free to choose the most convenient way to determine a sufficient reduction, either from the forward or inverse regression. An immediate advantage of inverse regression is that it treats each predictor separately instead of treating the panel as a block. That is, a large p-dimensional forward regression (potentially non-linear) problem is split in p univariate regression problems, which are easily modeled if y is univariate (or has a small dimension) even if p is large. Furthermore, inverse regression allows a plethora of estimation methods, also non-parametric, where the curse of dimensionality would make modeling of the forward regression practically impossible. Therefore, the method can result in significantly more accurate estimation than a linear forward regression model.

Most importantly, inverse regression accomplishes another goal in connecting sufficient reductions with the classical concept of a sufficient statistic: the "parameter" to be estimated and predicted is the whole time-series y_t.

4.2 Linear Reductions and Moment-Based SDR

Even though sufficient reductions need not be linear, **moment-based SDR** was developed under the requirement that $\mathbf{R}(\mathbf{x}_t)$ be linear. In linear SDR, $\mathbf{R}(\mathbf{x}_t)$ is a projection $\mathbf{P}_{\mathcal{S}}\mathbf{x}_t$ onto a lower-dimensional subspace $\mathcal{S}$ of $\mathbb{R}^p$ that incurs no loss of information about the conditional distribution $F(y_{t+h}|\mathbf{x}_t)$ or selected features thereof. If the mean squared error loss is used to evaluate the accuracy of the forecast, the conditional mean $\mathrm{E}(y_{t+h}|\mathbf{x}_t)$ is of interest and the goal is to find a reduction $\mathbf{R}(\mathbf{x}_t)$ such that $\mathrm{E}(y_{t+h}|\mathbf{x}_t) = \mathrm{E}(y_{t+h}|\mathbf{R}(\mathbf{x}_t))$. In density forecasting, the whole conditional distribution is the target and a reduction $\mathbf{R}(\mathbf{x}_t)$ such that $F(y_{t+h}|\mathbf{x}_t) = F(y_{t+h}|R(\mathbf{x}_t))$ is sought. In the rest of this Section we suppress subscripts keeping in mind that y is used in place of y_{t+h} and $\mathbf{x}$ in place of $\mathbf{x}_t$.

In this Section we focus the discussion on the identification (and peripherally to existence and uniqueness) of linear sufficient reductions and show how to exploit inverse regression to identify them. We require at the very outset the reduction be linear:

Condition 1 *Suppose the reduction* $\mathbf{R}(\mathbf{x})$ *is sufficient and a linear function of* $\mathbf{x}$*; that is, it satisfies*

$$F(y|\mathbf{x}) = F(y|\mathbf{R}(\mathbf{x})) \tag{4.4}$$

and

$$\mathbf{R}(\mathbf{x}) = \mathbf{a}'\mathbf{x}$$

for some $p \times d$ *matrix* $\mathbf{a}$.

Notice that the definition of sufficiency implies that we can only identify the subspace spanned by a linear reduction, span($\mathbf{a}$), rather than $\mathbf{a}$ per se, since $F(y|\mathbf{a}'\mathbf{x}) = F(y|\mathbf{b}'\mathbf{x})$ for all matrices $\mathbf{a}$ and $\mathbf{b}$ such that span$(\mathbf{a})$ = span$(\mathbf{b})$. A subspace spanned by the columns of a matrix $\mathbf{a}$ with $F(y|\mathbf{x}) = F(y|\mathbf{a}'\mathbf{x})$ is called a **dimension reduction subspace** (DRS).

Existence and Uniqueness – A linear reduction, although a trivial one, always exists, since one can always set $\mathbf{R}(\mathbf{x}) = \mathbf{x} = \mathbf{I}_p\mathbf{x}$. For the same reason a DRS is not generally unique. SDR's objective is to identify a **minimal reduction**, that is a DRS with minimum dimension as well as conditions that insure existence and uniqueness. Uniqueness and minimality are jointly guaranteed by focusing on the intersection of all DRS; such intersection, if it is itself a DRS, is called the **central subspace**. The latter exists under reasonably mild conditions on the marginal distribution of $\mathbf{x}$, such as convexity of its support. We refer to Cook (1998)[17] for more details and henceforth restrict attention to those regressions for which a central subspace exists.

The identification of a minimal sufficient reduction or, equivalently, the identification of a basis for the central subspace requires moment conditions on the marginal distribution of the predictor vector $\mathbf{x}$.

Condition 2 (Linear Design Condition) *There exists a full rank* $p \times d$ *matrix* $\mathbf{v}$ *such that*

$$\mathrm{E}\left[\mathbf{x}|\mathbf{v}'\mathbf{x}\right] = \mathbf{A}\mathbf{v}'\mathbf{x} \tag{4.5}$$

for an invertible matrix $\mathbf{A}$.

In general, the linearity condition (4.5) on the marginal distribution of the predictors is difficult to verify as it requires knowledge of $\mathbf{v}$. Nevertheless, it is satisfied for all $\mathbf{v} \in \mathbb{R}^{p\times d}$ if the predictors have an elliptically contoured distribution [See Eaton (1986)[30]]. The elliptically contoured family

of distributions includes the multivariate normal and Student's t distributions. Moreover, Steinberger and Leeb (2015) [52] showed that under comparatively mild conditions on the distribution of $\mathbf{x}$ the condition (4.5) is likely to be satisfied as p grows and the cross-section becomes large. Specifically, they showed that if a random p-vector $\mathbf{x}$ has a Lebesgue density, the mean of certain functions of $\mathbf{x}$ is bounded and that certain moments of $\mathbf{x}$ are close to what they would be in the Gaussian case [see the bounds (b1) and (b2) in Th. 2.1, Steinberger and Leeb (2015)[52]], then the conditional mean of $\mathbf{x}$ given $\mathbf{v}'\mathbf{x}$ is linear in $\mathbf{v}'\mathbf{x}$ for a $p \times d$ matrix $\mathbf{v}$, as $p \to \infty$ and d is either fixed or grows very slowly at the rate $d/\log p \to 0$. An appealing feature of these results is that they rely on bounds that can be estimated from data.

Steinberger and Leeb's result is of fundamental importance in SDR since it ascertains that the linearity condition (4.5) is satisfied by a large class of predictor distributions. Thus, first-moment SDR estimators, such as SIR in the ensuing Section 4.3, can be widely used to estimate basis elements of the column space of $\mathbf{v}$ in the reduction $\mathbf{R}(\mathbf{x}_t) = \mathbf{v}'\mathbf{x}_t$.

The following lemma links the linearity condition with inverse regression and points to a means to find the reduction.

Lemma 1 *Assume $\mathbf{R}(\mathbf{x}) = \mathbf{v}'\mathbf{x}$ satisfies (4.4), that is, it is a sufficient reduction, and the linearity condition (4.5) is satisfied for $\mathbf{v}$. Then*

$$\boldsymbol{\Sigma}^{-1}\left[\mathrm{E}\left(\mathbf{x}|y\right) - \mathrm{E}\left(\mathbf{x}\right)\right] \in \mathrm{span}\left(\mathbf{v}\right)$$

where $\boldsymbol{\Sigma} = \mathrm{cov}\left(\mathbf{x}\right)$. Equivalently,

$$\mathrm{span}\left(\boldsymbol{\Sigma}^{-1}\left[\mathrm{E}\left(\mathbf{x}|y\right) - \mathrm{E}\left(\mathbf{x}\right)\right]\right) \subseteq \mathrm{span}\left(\mathbf{v}\right)$$

Proof. See Corollary 10.1 in Cook (1998)[17] and Theorem 3.1 in Li (1991)[45]. ■

Lemma 1 obtains that the centered and scaled inverse regression function "lives" in a subspace, the inverse regression subspace, spanned by the columns of $\mathbf{v}$. That is, as y varies in $\mathbb{R}$, the random vector $\boldsymbol{\Sigma}^{-1}\left[\mathrm{E}\left(\mathbf{x}|y\right) - \mathrm{E}\left(\mathbf{x}\right)\right]$ is contained in a subspace that is spanned by the columns of $\mathbf{v}$. Therefore, in order to identify the sufficient reduction we need to identify a basis that generates the subspace that contains $\boldsymbol{\Sigma}^{-1}\left[\mathrm{E}\left(\mathbf{x}|y\right) - \mathrm{E}\left(\mathbf{x}\right)\right]$ as y varies. The following proposition provides the answer.

Proposition 3 *The column space of the matrix $\boldsymbol{\Sigma}^{-1}\mathrm{var}\left(\mathrm{E}\left(\mathbf{x}|y\right)\right)$ spans the same subspace as the subspace spanned by $\boldsymbol{\Sigma}^{-1}\left[\mathrm{E}\left(\mathbf{x}|y\right) - \mathrm{E}\left(\mathbf{x}\right)\right]$. That is,*

$$\mathrm{span}\left(\boldsymbol{\Sigma}^{-1}\mathrm{Var}(\mathrm{E}(\mathbf{x}|y))\right) = \mathrm{span}\left(\boldsymbol{\Sigma}^{-1}\left[\mathrm{E}(\mathbf{x}|y) - E(\mathbf{x})\right]\right) \subseteq \mathrm{span}\left(\mathbf{v}\right)$$

Proof. See Proposition 11.1 in Cook (1998)[17], an extension of Proposition 2.7 in Eaton (1983)[29], and Lemma 1. ■

Lemma 1 and Proposition 3 draw a link between the distribution of the data and the subspace that we wish to identify. Notice that in general the column space of $\boldsymbol{\Sigma}^{-1}\mathrm{var}\left(\mathrm{E}(\mathbf{x}|y)\right)$ provides only partial coverage of the central subspace since the inverse regression subspace can be a proper subset of the central subspace.

Under additional conditions one can show that more exhaustive capturing of the central subspace is possible. Other inverse regression moments, such as $\mathrm{E}\left(\boldsymbol{\Sigma} - \mathrm{Var}(\mathrm{E}(\mathbf{x}|y))\right)^2$, also live in the central subspace under additional conditions on the marginal distribution of the predictors (Cook and Weisberg (1991) [23]). In order not to clutter the present exposition, we focus only on the first inverse regression moment $\mathrm{E}(\mathbf{X}|Y)$ in order to introduce SDR methodology to the econometrics

literature via the simple and widely used Sliced Inverse Ression (SIR, Li (1991)[45]).[7] In general, linear moment-based SDR methods provide a way of identifying the number and coefficients (up to rotations, as in the DFM literature) of the linear combinations of the predictors in the forward forecasting equation. A feature of SDR methods, which can be viewed both as an advantage and a downside, is that they are silent regarding the functional form of the forward regression. They obtain the linear combinations of $\mathbf{x}_t$ that are needed in the forecasting equation in order to adequately reduce $\mathbf{x}_t$. When the number of SDR directions is 1 or 2, a plot of the response versus the reduction(s) can visually inform forward regression modeling. Dimension 2 or larger indicates that the forward model involves non-linear functions of the reductions. It is important to note that SDR methods do not remove the need to model the response but rather reduce significantly the complexity of modeling and uncover the structural dimension of the forward regression problem, i.e. how many derived linear combinations of the original predictors suffice to completely explain y.

4.3 Sliced Inverse Regression

Several estimators have been proposed in order to estimate the central subspace. We focus on the first and most widely used: Sliced Inverse Regression (SIR) proposed by Li (1991)[45]. SIR is a semiparametric method for finding dimension reduction subspaces in regression. It is based on the results of Section 4.3 and uses a sample counterpart[8] to $\mathbf{\Sigma}^{-1}\text{Var}\left(\text{E}\left(\mathbf{x}|y\right)\right)$, the population object that lives in the subspace generated by the coefficient matrix of the reduction. The name derives from using the inverse regression of $\mathbf{x}$ on the sliced response y to estimate the reduction. For a univariate y, the method is particularly easy to implement, SIR's step functions being a very simple nonparametric approximation to $\text{E}(\mathbf{x}|y)$.

Implementation of SIR – In order to estimate $\mathbf{M} = \text{var}(\text{E}(\mathbf{x}|y))$, the range of the observed responses $\mathbf{Y} = (y_1, \ldots, y_T)'$ is divided in J disjoint slices $S_1, \ldots, S_J$ whose union is the range of $\mathbf{Y}$. We denote the overall sample mean of the sample predictor matrix $\mathbf{X}$ by $\bar{\mathbf{X}} = (\sum_{t=1}^{T} x_{t1}/T, \ldots, \sum_{t=1}^{T} x_{tp}/T)'$, and for $j = 1, \ldots, J$, we let $\bar{\mathbf{X}}_j = \sum_{y_t \in S_j} \mathbf{X}_t/n_j$, where n_j is the number of y_t's in slice S_j. The covariance matrix of $\mathbf{x}$ is estimated by the sample covariance matrix $\widehat{\mathbf{\Sigma}} = \sum_{t=1}^{T}(\mathbf{X}_t - \bar{\mathbf{X}})(\mathbf{X}_t - \bar{\mathbf{X}})'/T$, and the SIR seed $\mathbf{M}$ with

$$\widehat{\mathbf{M}} = \sum_{j=1}^{J} \frac{n_j}{T}(\bar{\mathbf{X}}_j - \bar{\mathbf{X}})(\bar{\mathbf{X}}_j - \bar{\mathbf{X}})'$$

The spectral value decomposition of $\widehat{\mathbf{M}}$ yields its d left eigenvectors $\widehat{\mathbf{u}}_1, \ldots, \widehat{\mathbf{u}}_d$ that correspond to its d largest eigenvalues, $\hat{\lambda}_1 > \hat{\lambda}_2 > \ldots > \hat{\lambda}_d$. The matrix $\widehat{\mathbf{B}} = \widehat{\mathbf{\Sigma}}^{-1}(\widehat{\mathbf{u}}_1, \ldots, \widehat{\mathbf{u}}_d) = (\boldsymbol{b}_1, ..., \boldsymbol{b}_d)$ estimates $\mathbf{v}$ in $\mathbf{R}(\mathbf{x}) = \mathbf{v}'\mathbf{x}$ of Lemma 1. The SIR predictors to replace $\mathbf{X}$ in the forward regression are the columns of the $T \times d$ matrix $\mathbf{X}\widehat{\mathbf{B}} = (\mathbf{X}\boldsymbol{b}_1, ..., \mathbf{X}\boldsymbol{b}_d)$. The number of SIR directions, d, is typically estimated using asymptotic weighted chi-square tests (Bura and Cook (2001a)[12], Bura and Yang (2011)[15]), information criteria such as AIC and BIC, or permutation tests (Yin and Cook (2001)[24]. We note that these tests are valid under the assumption of iid draws from the joint distribution of $(y, \mathbf{x})$, which is typically not the case for econometric data.

How SIR works: SIR finds the directions of maximum variance between slices, with T data points collapsed in J slice means clustered according to y labels (slices). In the extreme case of

[7] For instance, although SIR may not exhaustively identify the central subspace, it can be shown that SIR is exhaustive when $\mathbf{x}|y$ is multivariate normal with constant variance-covariance matrix (See Cook (2007)[19]).

[8] Stricly speaking, SIR, through slicing, forces a discretization of y rather than the actual realizations of y, however it is possible to show that the space spanned by the slices is a subset of the central subspace.

$J = T$, i.e. when each slice corresponds to a single y observation, $\mathbf{M}$ becomes $\mathbf{\Sigma}$, the sample covariance of $\mathbf{x}$, and SIR is identical to PCA. However, for $J < T$, the variance (noise) of the components within the same slice is suppressed in favor of their signal, which makes SIR much more efficient in identifying $\mathbf{x}$ components (projections) targeted to y.

4.3.1 Statistical Properties

In Proposition 4,we show that the SIR directions are consistent estimators of directions in the central subspace for all $\mathbf{x}_t$ satisfying the linear design condition (4.5) and conditional distributions of $\mathbf{x}_t|y_{t+h}$, $h = 1, 2, ...$ with finite second moments.

Proposition 4 *Assume that the time series $\mathbf{x}_t$ and $\mathbf{x}_t|(y_{t+h} = s)$, $s = 1, \ldots, J$, $t = 1, \ldots$, $h = 0, 1, \ldots$, are both covariance-statonary with absolutely summable autocovariances, i.e. $\sum_{l=-\infty}^{\infty} |\sigma_{jj}(l)| < \infty, \sum_{l=-\infty}^{\infty} |\sigma_{jj|y_{t+h}}(l)| < \infty, j = 1, ..., p$. Then, the SIR directions are consistent estimators of directions in the central subspace for all $\mathbf{x}_t$ satisfying the linear design condition (4.5).*

Proof. SIR is based on the covariance matrix $\mathbf{M}_h = \text{cov}(\text{E}(\mathbf{x}_t|y_{t+h}))$, $t = 1, \ldots, T$. If y_t is discrete and finite, we can assume $y_t \in \{1, 2, \ldots, J\}$ without loss of generality. Let $p_s = \text{Prob}(y_{t+h} = s)$ and $\mathbf{m}_s = \text{E}(\mathbf{x}_t|y_{t+h} = s)$, $s = 1, \ldots, J$. Then,

$$\text{cov}(\text{E}(\mathbf{x}_t|y_{t+h})) = \sum_{s=1}^{J} p_s(\mathbf{m}_s - \boldsymbol{\mu})(\mathbf{m}_s - \boldsymbol{\mu})'$$

As a result of the second order stationarity with absolutely summable autocovariances of $\mathbf{x}_t$ and $\mathbf{x}_t|(y_{t+h} = s)$, $s = 1, \ldots, J$, $t = 1, \ldots$, $h = 0, 1, \ldots$, the sample moments $\bar{\mathbf{X}}$ and $\widehat{\mathbf{m}}_s = \bar{\mathbf{X}}_s = \sum_{y_{t+h}=s} \mathbf{X}_t/n_s$, where n_s is the number of y_t's equal to s, are both consistent as $T, n_s \to \infty$. Also, $\hat{p}_s = n_s/T \to p_s$. Therefore,

$$\widehat{\mathbf{M}}_h = \sum_{s=1}^{J} \hat{p}_s(\widehat{\mathbf{m}}_s - \bar{\mathbf{X}})(\widehat{\mathbf{m}}_s - \bar{\mathbf{X}})' \xrightarrow{p} \mathbf{M}_h$$

as it is a continuous function of consistent estimators. Consequently, the eigenvectors of $\widehat{\mathbf{M}}_h$, $\widehat{\mathbf{u}}_k$, $k = 1, \ldots, p$, converge to the corresponding eigenvectors of $\mathbf{M}_h$. Moreover, since the sample covariance matrix $\widehat{\mathbf{\Sigma}}$ is consistent for $\mathbf{\Sigma}$, the SIR predictors $\widehat{\mathbf{\Sigma}}^{-1}\widehat{\mathbf{u}}_k$, $k = 1, \ldots, d$ are consistent for the d columns of $\mathbf{v}$ in the sufficient reduction $\mathbf{R}(\mathbf{x}_t) = \mathbf{v}'\mathbf{x}_t$. Notation and results for stationary and ergodic time series that we use are provided in Appendix D.

When y is continuous, it is replaced with a discrete version $\tilde{y}$ based on partitioning the observed range of Y into J fixed, non-overlapping slices. Since $y \perp\!\!\!\perp \mathbf{x}|\mathbf{v}'\mathbf{x}$ yields that $\tilde{y} \perp\!\!\!\perp \mathbf{x}|\mathbf{v}'\mathbf{x}$, we have $S_{\tilde{Y}|\mathbf{x}} \subseteq S_{Y|\mathbf{x}}$. In particular, provided that J is sufficiently large, $S_{\tilde{y}|\mathbf{x}} \approx S_{y|\mathbf{x}}$, and there is no loss of information when y is replaced by $\tilde{y}$. ■

Under more restrictive assumptions on the processes $\mathbf{x}_t$ and $\mathbf{x}_t|(y_{t+h} = s)$, $s = 1, \ldots, J$, $t = 1, \ldots$, $h = 0, 1, \ldots$, it can also be shown that their sample means are approximately normally distributed for large T (see Appendix D). Under the same assumptions we can then obtain that $\widehat{\mathbf{M}}_h$ is asymptotically normal following similar arguments as Bura and Yang (2011)[15] who obtained the asymptotic distribution of $\widehat{\mathbf{M}}$ when the data are iid draws from the joint distribution of $(y, \mathbf{x})$.

4.3.2 Inverse Regression as Extraction of Targeted Factors

Most SDR methodology is based on inverse regression. In general, inverse regression focuses attention on the set of p inverse regressions

$$\mathbf{x} = \mathbf{a} + \mathbf{B}\mathbf{f}(y) + \mathbf{e} \tag{4.6}$$

where y is substituted with $\mathbf{f}(y)$ that contains functions of y whose choice reflects different inverse regression based SDR methods. Such functions play the role of "observed factors" and in practice, in addition to contemporaneous and lagged values of y, may contain various functions of y such as polynomials. For example, SIR effectively approximates $\mathbf{f}(y)$ with step functions; parametric inverse regression (PIR) (Bura and Cook (2001b) [13]) and principal fitted components (PFC, Cook and Forzani (2008)[20]) approximate $\mathbf{f}(y)$ with continuous functions of the response. These three SDR methods essentially analyze and extract the first few PCs of the space of the fitted values in (4.6). The term $\mathbf{f}(y)$ plays the role of a factor structure, but, in contrast with the DFM, it is observable. Intuitively the inverse regression approach replaces $\mathbf{x}$ with its projection on $\mathbf{f}(y)$ and in so doing it extracts its "targeted" factor structure.

5 Empirical Application

5.1 Data and Out-of-Sample Forecasting Exercise

Next we want to put our estimators to work comparing them in a classical pseudo out-of-sample macro-forecasting horserace against the parsimonious AR(4). We pick inflation (CPIAUCSL) and industrial production as our targets (INDPRO). Then we look for a large set of regressors in order to feed into our models as much information on the macroeconomy as possible. We would like to adopt a "standard" dataset containing a large number of US macro variables however the various forecasting studies that use large panels of US macro variables, aside from a core set of agreed upon macro variables, is quite inhomogenous regarding the specific set of non-core variables to be used in forming the **initial** dataset from which to choose or combine the variables. Given that one of the tasks of the forecasting exercise is choosing the most useful variables to forecast a given target variable, the choice of the initial dataset seems indeed crucial. The next paragraphs highlights the rugged landscape of data sources used in some of the most important studies in the DFM macro-forecasting literature and our final choice.

Settling on a Shared Data Source – A problem faced by any researcher in the macro-forecasting field is the lack of comparability across studies due to the multitude of data sources and data vintages used in the literature, a phenomenon that has been pervasive until recently. Luckily an initiative spearheaded by McCracken and Ng and documented in McCracken and Ng (2015) [49] has set out on the project to impose some discipline in the current and future production of macro-forecasting studies. One outcome of the project has been the creation of a dataset of 132 macro variables called FRED-MD and updated in real time by the same staff that maintains the popular FRED database.[9] We embraced their initiative and adopted FRED-MD as our dataset of choice although the dataset comes with some limitations that we discuss to some extent below and more in detail in the appendix.

Alternative Data Sources in Other DFM Studies – FRED-MD has fewer variables than the quarterly dataset of 144 variables used by Stock and Watson in [58], apparently the most exploited

[9] The data can be dowloaded from Michael McCracken's website at the St. Louis Fed at the address: https://research.stlouisfed.org/econ/mccracken/fred-databases/

dataset in quarterly studies. It has also fewer variables than the dataset of 143 variables used by Stock and Watson in [60]. The latter is a quarterly study but the dataset posted by Mark Watson contains monthly variables. Finally our dataset contains fewer variables than the 149 regressors used by Stock and Watson in [54] or the 215 series used by Stock and Watson in [55].[10] In turn, both studies draw upon the work in the seminal NBER working paper [53] by Stock and Watson. The recently published quarterly study [60] by Stock and Watson on shrinkage methods uses 143 series but in the dataset posted online by Mark Watson one can find only 108 monthly variables and 79 quarterly variables. Although we have chosen the most up-to-date dataset available online at the cost of omitting some variables given the data intensity of our statistical techniques we also tried to run our pseudo-out-of-sample forecasts using some of the richer datasets mentioned above, and we have noticed a slight deterioration of forecasting performance when using our dataset of choice signalling that either some of the variables that we do not include might be marginally helpful in improving the forecast or that data revisions played a role.[11] Definitely the inclusion of the great recession and the subsequent recovery, a period not covered by the mentioned studies, appears to impart a substantive deterioration in the forecasting performance of the estimators that we review. The following table summarizes several relevant studies and the salient characteristics of their dataset and statistical methods used and McCracken and Ng (2015) [49] have an informative chronology of the evolution of the large panel of macro datasets.

Table 1: Summary of Datasets Characteristics of Several Relevant Studies in the DFM literature

Study	**# Var.**	**Freq.**	**Online**	**Data Span**	**Meth.**
Stock and Watson (1998) [53]	224	m	no	1959-1997	PCR, VAR
Stock and Watson (2002) [54]	149	m	no	1959-1999	PCR, VAR
Stock and Watson (2002) [55]	215	m	no	1959-1999	PCR, VAR
Bai and Ng (2008) [4]	132	m	no	1959-2003	SPC, EN
Stock and Watson (2005) [56]	132	m	yes	1959-2003	VAR, SVAR, PCR
Boivin and Ng (2006) [10]	147	m	no	1960-1998	WPCR
Stock and Watson (2008) [58]	144	q	yes	1959-2006	Split-Sample PCR
Stock and Watson (2012) [60]	108	m, q	yes	1960-2008	PCR, IC, PT BMA, EB, B
Jurado et al. (2015) [42]	135	m	yes	1959-2010	PCR

FRED-MD – The dataset is called FRED-MD (where MD stands for monthly dataset) and it contains a balanced panel with monthly data from 1960m1 to 2014m12, covering 54 years totaling 648 monthly observations.. We choose to work with monthly data since the companion quarterly dataset FRED-QD is not available yet. Moreover, our SDR forecasting procedure is quite data intensive and monthly data seem to be a better workbench to test our estimators at this juncture. The dataset is described in McCracken and Ng (2015) [49] along with a discussion of some data adjustments needed to construct the panel. We note that a major shortcoming of the dataset is that core CPI and non-farm payroll employment, two of the most watched series by forecasters and FED officials have not been included so far. FRED-MD contain very limited real-time vintages making real-time forecasting unfeasible at the moment.

[10] The series in this study came from the **DRI/McGraw-Hill Basic Economics** database, formerly named **Citibase**.

[11] We are currently working on retrieving an updated dataset with real-time vintages with about 150 regressors as used by Stock and Watson in some of their work.

Manipulation of FRED-MD – Some variables in the dataset have a large number of missing data. Rather than running an EM algorithm to fill in the missing data and achieve a balanced panel as done by McCraken and Ng (2015) [49], who in turn follow Stock and Watson (2002b) [55], we exclude them. The five excluded variables are: ACOGNO="New Orders for Consumer Goods", ANDENOx="New Orders for Nondefense Capital Goods", OILPRICE="Crude Oil, spliced WTI and Cushing", TWEXMMTH="Trade Weighted U.S. Dollar Index: Major Currencies" and UMCSENT="Consumer Sentiment Index". We do not apply any cleaning of outliers.

Data Transformations and Forecast Targets – We adopt the transformations suggested by McCracken and Ng (2015) [49] and coded in the second row of the original downloaded dataset. We opt to present our results for a "nominal" forecast target, CPI inflation with mnemonics CPIAUCSL, and a "real" target, total industrial production with mnemonics INDPRO. We follow the literature and instead of forecasting the chosen target variables h months ahead we forecast the average realization of the variable in the h months ahead period. Hence the transformation of the target variable dictates the forecast target. For instance in the case of inflation, a variable marked as I(2) and transformed as

$$y_t = \Delta^2 \log\left(CPI_t\right)$$

we generate the target

$$y_{h+t}^h = \frac{1200}{h} ln\left(\frac{CIP_{t+h}}{CIP_t}\right) - 1200 \ln\left(\frac{CPI_t}{CPI_{t-1}}\right)$$

Industrial production is a variable marked as I(1) and tranformed by

$$y_t = \Delta \log\left(IP_t\right)$$

and the resulting target will be

$$y_{h+t}^h = \frac{1200}{h} ln\left(\frac{IP_{t+h}}{IP_t}\right)$$

The Pseudo Out-of-Sample Forecasting Scheme – We align the data as shown in the following Figure by placing the target on the same line of the regressors in an ideal h-step ahead OLS scheme. The transformed and aligned data are available on request. We conduct our forecasting exercise at horizons $h = 1, 3, 6, 12, 24$. Both these are relevant horizons in practice and they are enough to allow exploration of possible variation across horizons within each forecasting method. For practical reasons, as common in the literature, we adopt h-step ahead regression rather than iterated in order to avoid the simulation and feeding of exogenous regressors. As indicated in Figure 1, an advantage of PCR, a non-supervised method, is that the principal components can be re-computed as soon as a new line of obervations becomes available in the recursive scheme. The superscript of the PC component in the table highlights this point: for instance when estimating PCR to forecast 3-steps ahead, a 3-step ahead regression is run in $t = 1984m01$ regressing y_{t+3}^3 on the PCs for $1984m01$ but computed using data through $t + 1 = 1984m02$. Then new data through $1982m02$ is used to forecast y_{t+3+1}^3 and prediction is formed

$$\widehat{y}_{t+3+1}^3 = \alpha_3 + \gamma_3\left(L\right) y_t + \beta_3\left(L\right) \widehat{PC}^{t+3+1}\left(t\right)$$

and compared with the realized data (the comparison involving the two yellow cells in Figure 1).

Figure 1: Data Alignement and Out-of-Sample Forecasting Scheme

Row	Date t	$y^{h=3}(t+h)$	$y^{h=1}(t+h)$	y(t)	y(t-1)	X(t)	PC(t)	PLS(t)	yh3hat	yh1hat
1	1959m01	NA	NA	y(1959m01)	NA	X(1959m01)			NA	NA
2	1959m02	$y^{h=3}$(1959m05)	$y^{h=1}$(1959m03)	y(1959m02)	y(1959m01)	X(1959m02)			NA	NA
3	1959m03	$y^{h=3}$(1959m06)							NA	NA
...	...	...	...	...	...	...			NA	NA
...	...	...	...	...	...	...			NA	NA
13	1960m01	y^{h}(1960m04)	$y^{h=1}$(1960m02)	y(1960m01)	y(1959m12)	X(1960m01)			NA	NA
...	...	...	...	...	...	...			NA	NA
...	...	...	...	...	...	...			NA	NA
...	...	...	...	...	...	...			NA	NA
...	...	...	...	...	...	...			NA	NA
...	...	...	...	...	...	...			NA	NA
301	1984m01	$y^{h=3}$(1984m04)	$y^{h=1}$(1984m02)	y(1984m01)	y(1983m12)	X(1984m01)	$PC^{1984m02}$(1984m01)	$PLS^{1984m01}$(1984m01)	NA	NA
302	1984m02	$y^{h=3}$(1984m05)	$y^{h=1}$(1984m03)	y(1984m02)	y(1984m01)	X(1984m02)			$yhat^{h=3}$(1984m05)	$yhat^{h=1}$(1984m03)
303	1984m03	$y^{h=3}$(1984m06)	$y^{h=1}$(1984m04)	y(1984m03)	y(1984m02)	X(1984m03)			$yhat^{h=3}$(1984m06)	$yhat^{h=1}$(1984m04)
				y(1984m04)	y(1984m03)	X(1984m04)				
...	...	...	...	...	...	...			...	...
...	...	...	...	...	...	...			...	...
...	...	$y^{h=3}$(2015m05)	$y^{h=1}$(2015m03)	y(2015m02)	y(2015m01)	X(2015m02)			...	...
675	2015m03	$y^{h=3}$(2015m06)	$y^{h=1}$(2015m04)	y(2015m03)	...	...			$yhat^{h=3}$(2015m06)	$yhat^{h=1}$(2015m04)
676	2015m04	NA	$y^{h=1}$(2015m05)	y(2015m04)	...	...			NA	$yhat^{h=1}$(2015m05)
677	2015m05	NA	$y^{h=1}$(2015m06)	y(2015m05)	...	...			NA	$yhat^{h=1}$(2015m06)
678	2015m06	NA	NA	y(2015m06)	...	...			NA	NA

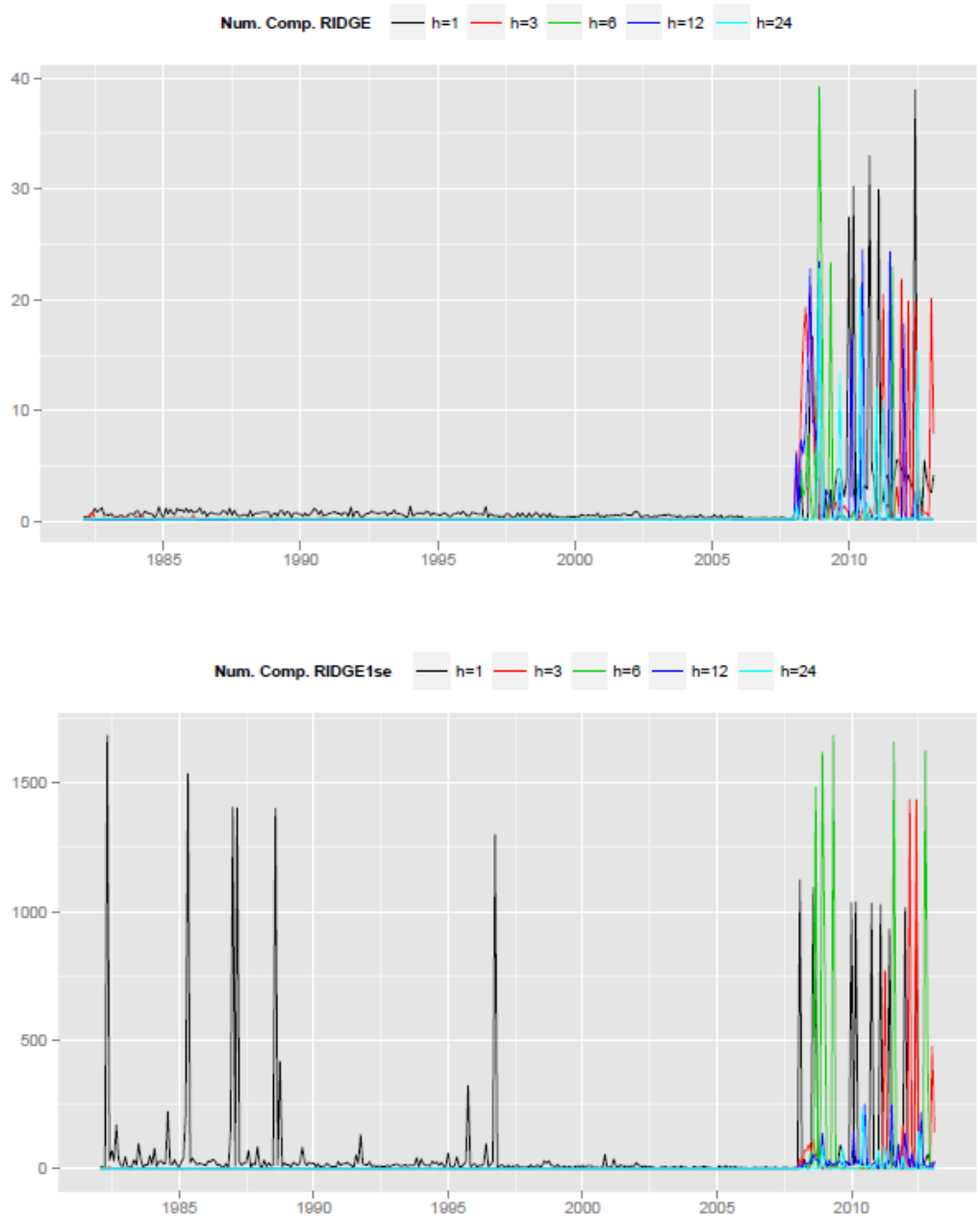
We report results for different sub-samples however our out-of-sample forecasting exercise uses a recursive window rather then a moving one.

5.2 Estimation Details and Results on the Number of Components

The practical implementation of the estimators summarized in an earlier Section necessitates the choice of several details. In this Section we present these choices by estimation procedure and some sensitivity analysis of the results.

RIDGE – In order to implement RIDGE we used the R package `glmnet` by Friedman et al. (2015) [33]. The shrinkage parameter λ needs to be chosen prior to the estimation. DeMol et al. (2008) [25] resorted to fit a grid of values and report MSFE for all. We tried some of the values of their grid and we discuss their forecast performance below. However we also opted to fully exploit the full functionalities of `glmnet` and let the data suggest a value for lambda using *nfold* cross-validation (where $n = 8$). As a result of the cross-validation we obtain sequences for λs at each forecasting step that we plot in Figure 2. In the top plot are the values of $\lambda_{\min}$ that minimize the cross-validation error. The bottom plot contains the values of λ_{1se} corresponding to a 1 standard deviation of the cross-validation sequence (the default). A striking pattern is revealed in the plot: it appears that forecasting at any horizon becomes increasingly difficult after the great recession with a sudden surge both in the volatility and average level of λ, a sign of RIDGE attempting to shrink more as a reaction in the increase difficulty in prediction. The discouraging results in terms of MSFE that we report in the next Section appear to be the flip side of this feature.

Figure 2: CPIAUCSL: λ_{min} and λ_{1se} Selected by Cross-Validated RIDGE over the Recursive Forecasting Window.

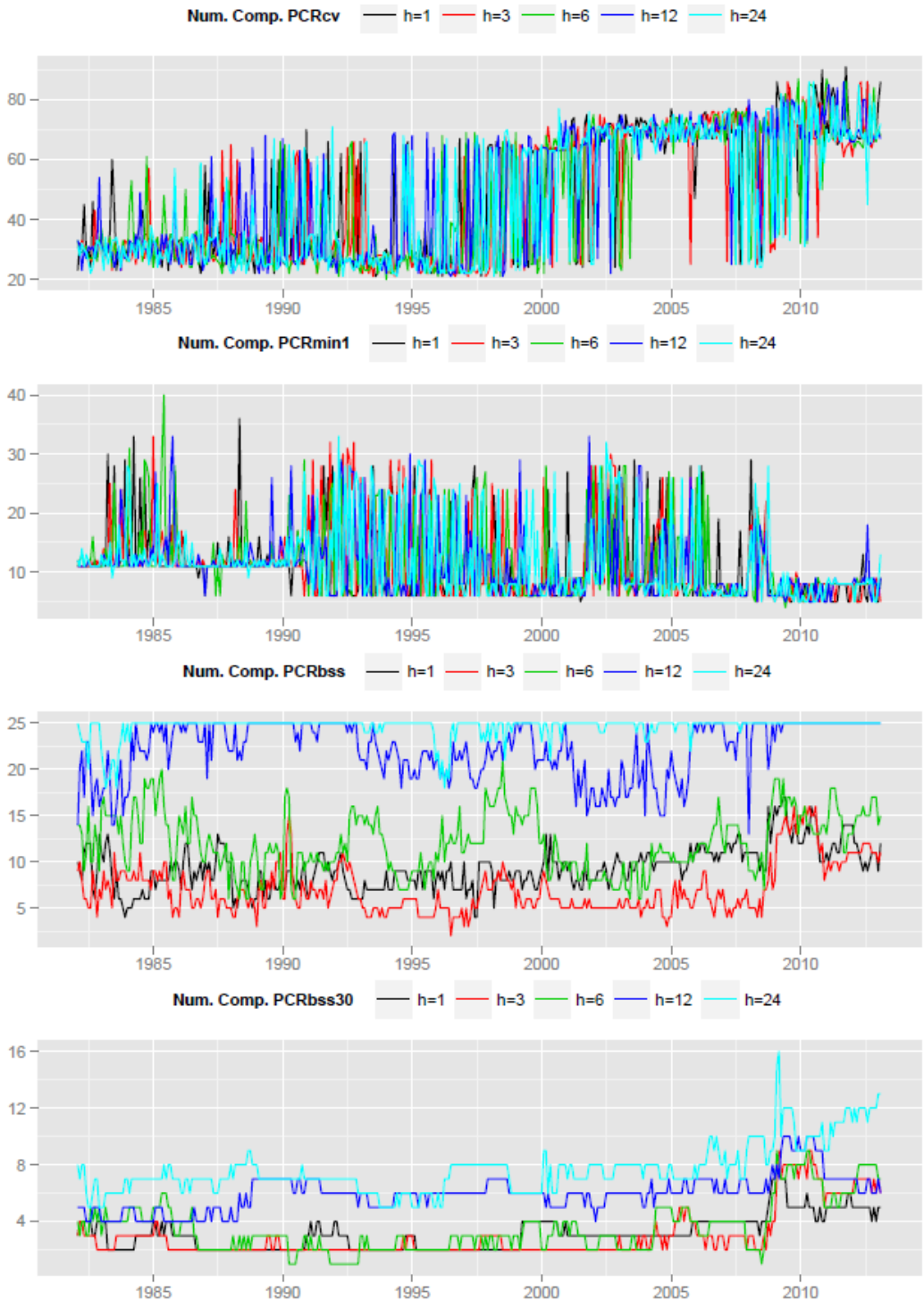


PCR – We estimate many different versions of PCR in order to cover a wide menu of choices of the truncation meta-parameter M. First of all we follow Stock and Watson (2002) [55] and estimate a series of PCRs (in this context known as diffusion indexes with acronym DIAR) with a constant M throughout the forecasting experiment. We looked at values for $M = 1, 2, 3, 4, 5, 6, 8, 10, 20, 30$ generating models PCRn1 through PCRn30. When forecasting inflation we find that 8 components explain about 45% of the variance of the panel matching the findings in McCracken and Ng (2015)

[49]. We note that simulations from an approximate factor model in Barbarino and Bura (2015) [7] in which the true number of factors is 8 result in at least 75% of explained variance on average across simulations.

We use the R-package `pls` by Mevik et al. (2013) [50] which also implements PCR where the number of components M is selected via cross validation.. The cross-validation in-sample MSFE has an interesting contour over the number of components with one or two local minima and a global minimum all of them depending on the forecast horizon. The global minimum on average is achieved only with a rather large number of components especially toward the end of the sample as shown in the first panel in Figure 3. That is a striking feature considering that in simulations reported in our companion paper Barbarino and Bura (2015) [7] cross-validation tends to gravitate around a number of components close to the number of true factors. The results in our empirical investigations in this paper suggest that either the true number of factors generating FRED-MD is indeed very large or that some non-linearities (such as change in regime) or other features of the data disrupt somewhat the effectiveness of cross-validation after the early 2000s. Cross-validation appears to generate a quite volatile choice for the number of components over the experiment at all forecast horizons. The evolution of the first local minimum is shown in the second panel of Figure 3 (model PCRmin1), which although may be less volatile it is still quite unstable. Another possibility is implementing best subset selection (BSS) in the choice of the components: model PCRbss is allowed to pick any component and PCRbss30 only in the first 30.To implement BSS, we use the `leaps` R-package by Thomas Lumley (2009) [47]. BSS does not impose a hierarchical ordering of the components (in which if component #3 is used also component #2 is used). Rather it uses a backward search running very many regressions from larger models to smaller ones eliminating regressors using the BIC criterion. The last two panels of Figure 3 highlight that forecasting at longer horizons seems more difficult and requires more components. Also for these models it appears that forecasting in the last part of the window requires more components on balance.

Figure 3: CPIAUCSL: Number of Components Selected by Cross-Validated and Best Subset PCR over the Recursive Forecasting Window.



We also run PCR with selection of the components according to Bai and Ng (2002) [2] and pick their favorite PCp2 and the alternative ICp2 criteria. Also in this exercise we find instabilities

worth of note. We can more or less reproduce the results in McCracken and Ng (2015) [49] in which it is shown that such criteria select about 8 to 10 components (as remarked above this number of components explains less than half of the variance in the panel). However especially PCp2 is very sensitive to the maximum number of allowed components that has to be entered in their computation (a meta-parameter that we denote with $k_{\max}$). Setting $k_{\max} = 30$, a natural choice since 30 components explain about 90% of the variance of the dataset (model PCRpcp2, top plot in the Figure below), we obtain that PCp2 selects many more components than with $k_{\max} = 15$ (model PCRpcp2b, 3rd panel below). For $k_{\max} = 50$ PCp2 selects way more components, more than cross validation and very frequently all components, about 120 on average. By contrast ICp2 appears to be more stable when $k_{\max}$ moves from 15 to 30. ICp2 becomes very unstable when $k_{\max}$ is above 70. The selection of the number of components for these criteria is shown in Figure 4. Notice that these criteria, being untargeted are not affected by the forecast horizon at hand.

PLS – We implement PLS using the `pls` package in [50]. We choose the truncation meta-parameter u with cross-validation. We implement two models. In PLSRcv we include y_t and its lags directly in the $\mathbf{X}$ matrix whereas in model PLSRcvd we do not include them and in a second step we run OLS of the target on y_t and its lags and the PLS components. In contrast with PCR, at short horizons u is not very large. At longer horizons, $h = 12$ and $h = 24$ PLS needs a large number of components, similar to PCR. The great recession appears to require more components also at shorter horizons.

Figure 4: CPIAUCSL: Number of Components Selected by PCp2 and ICP2 Criteria over Recursive Forecasting Window ($k_{max} = 30$ in top 2 panels, $k_{max} = 15$ in bottom 2 panels).

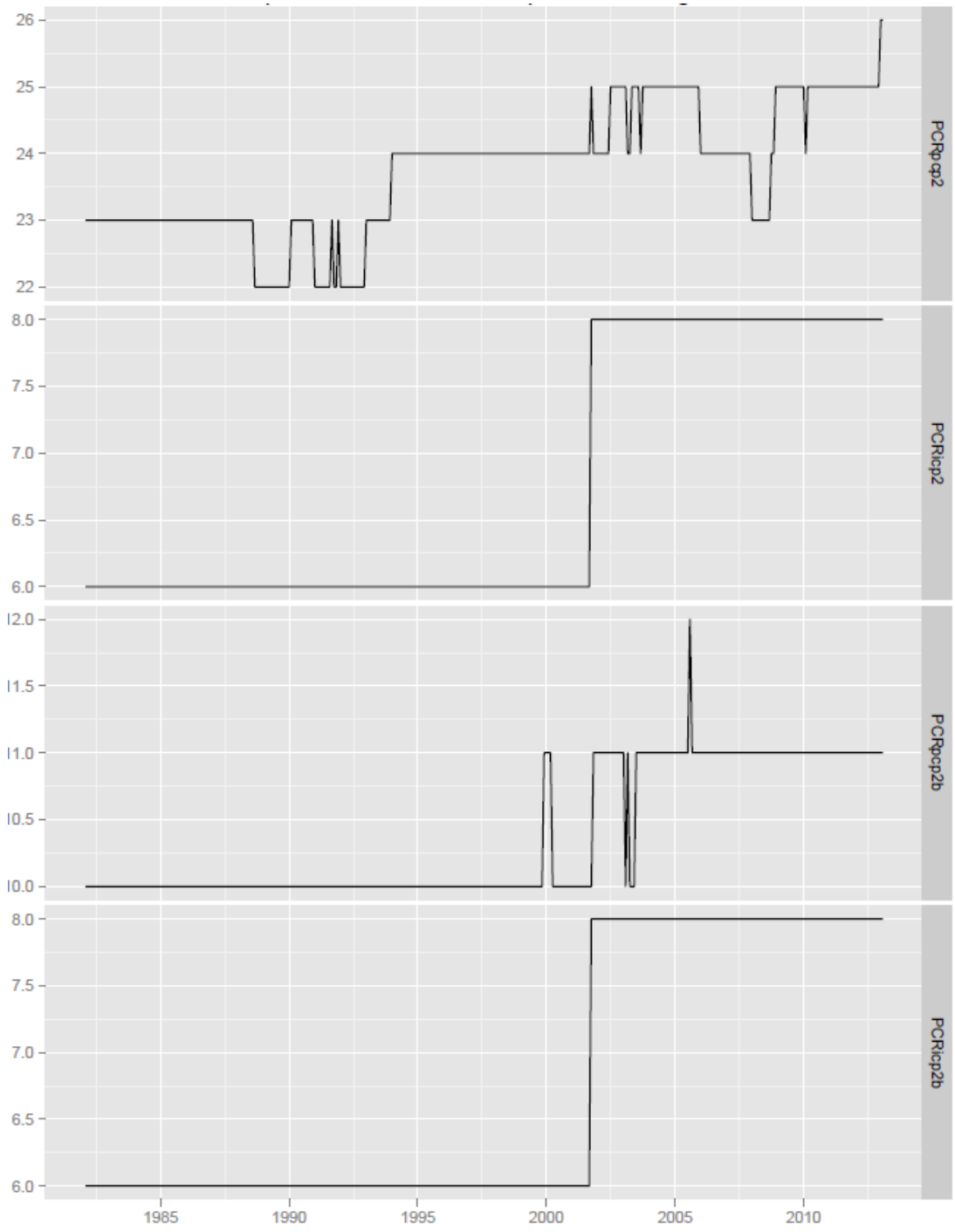


Figure 5: CPIAUCSL: Number of Components Selected by Cross-Validation of PLS over Recursive Forecasting Window.

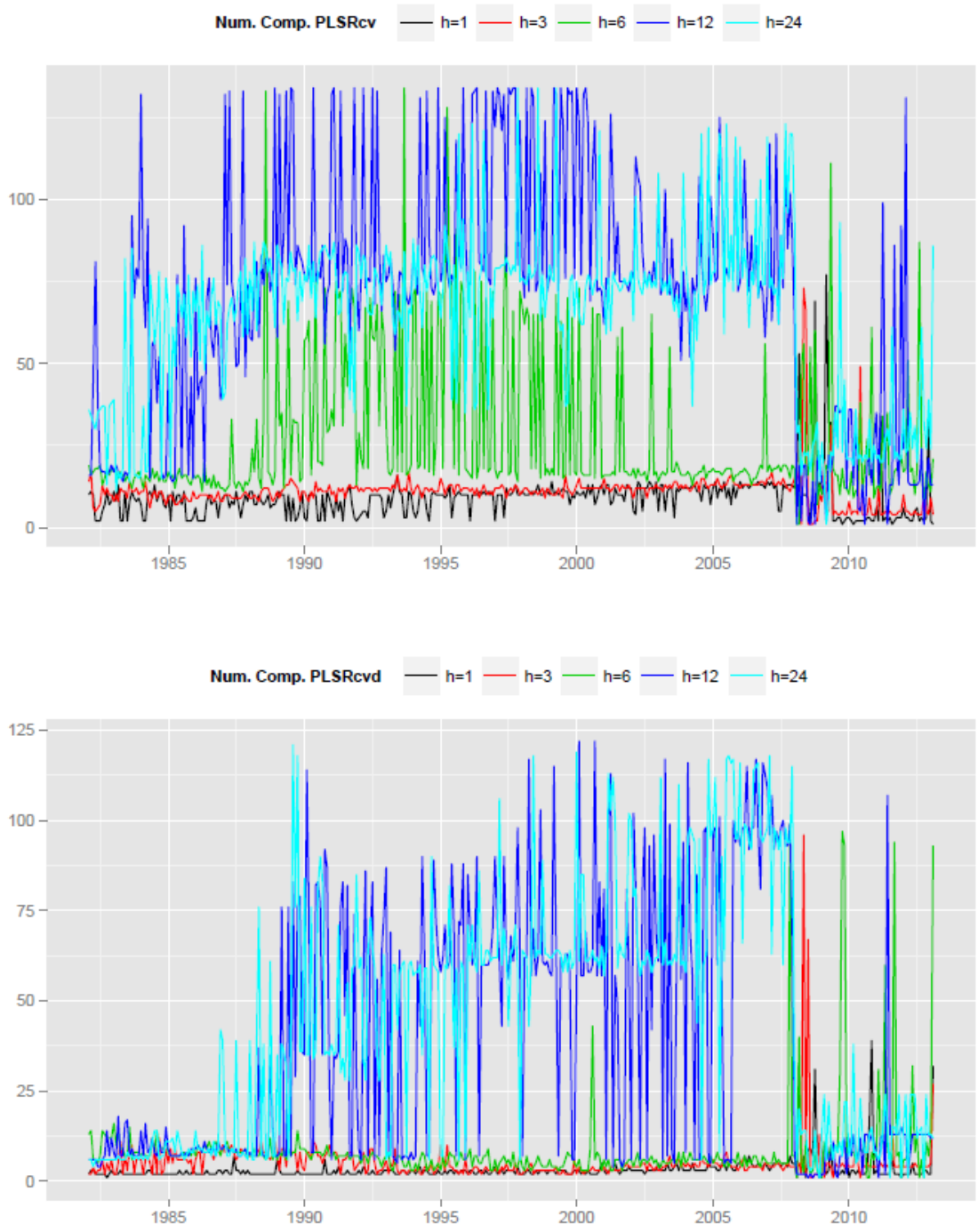


Table 2 compares PCR and PLS on the basis of explained variance with 10 components. Despite the much smaller number of components the superior targeting nature of PLS relative to PCR is evident as 10 components explain about 70% of the variance CPIAUCSL and 60% of the variance

in INDPRO. Not only does PLS explains a larger fraction of the variance in the target but also it explains a fraction of the variance of the panel similar to the fraction explained by PCR.

Table 2: Comparing PCR and PLS on the Basis of Explained Variances by 10 Comp.

		CPIAUCSL	**INDPRO**
PCR10	%var(X)	52%	52%
	%var(y_{t+h})	33%	40%
PLS10	%var(X)	50%	46%
	%var(y_{t+h})	70%	60%

Tables 3 and 4 wrap up the results on the average number of components and their volatility across our estimators.

Table 3: Average of Num. Components Selected (Estimation Sample: 1960m01-1982m01-2013m01)

	CPIAUCSL					**INDPRO**				
horizon	h=1	h=3	h=6	h=12	h=24	h=1	h=3	h=6	h=12	h=24
RIDGE	1.43	0.91	0.54	0.73	0.47	14.78	9.23	5.43	7.41	3.18
RIDGE1se	69.94	15.19	26.05	6.31	2.70	192.24	70.51	35.10	25.39	12.89
PCRcv	47.90	46.93	46.18	48.39	46.57	117.38	117.06	117.00	117.13	117.11
PCRmin1	11.38	11.57	11.56	11.35	11.21	34.30	35.75	36.21	35.79	36.37
PCRmin2	23.37	23.43	23.68	23.02	23.19	57.72	59.60	60.95	60.26	60.16
PCRpcp2	23.73	23.73	23.73	23.73	23.73	23.53	23.53	23.53	23.53	23.53
PCRpcp2b	10.37	10.37	10.37	10.37	10.37	10.37	10.37	10.37	10.37	10.37
PCRicp2	6.74	6.74	6.74	6.74	6.74	6.74	6.74	6.74	6.74	6.74
PCRicp2b	6.74	6.74	6.74	6.74	6.74	6.74	6.74	6.74	6.74	6.74
PCRbss	9.30	7.06	12.10	22.23	24.61	10.53	12.41	18.49	17.02	17.94
PCRbss30	3.15	2.98	3.53	5.97	7.67	3.88	4.12	4.40	7.28	8.73
PLSRcv	9.01	10.97	27.85	70.86	63.99	1.39	2.37	13.18	2.06	4.17
PLSRcvd	3.23	5.24	8.37	39.31	44.79	1.26	2.47	4.58	2.18	3.94

Table 4: Standard Deviation of Num. Components Selected (Estimation Sample: 1960m01-1982m01-2013m01)

	CPIAUCSL					INDPRO				
horizon	h=1	h=3	h=6	h=12	h=24	h=1	h=3	h=6	h=12	h=24
RIDGE	3.88	3.15	2.95	2.98	2.07	9.86	9.23	5.43	7.41	3.18
RIDGE1se	241.51	115.36	192.18	26.19	16.66	91.86	70.51	35.10	25.39	12.89
PCRcv	20.91	21.15	21.06	20.88	20.80	2.55	117.06	117.00	117.13	117.11
PCRmin1	6.50	6.67	6.36	5.83	6.16	25.76	35.75	36.21	35.79	36.37
PCRmin2	8.18	7.77	7.46	7.41	7.35	27.51	59.60	60.95	60.26	60.16
PCRpcp2	0.95	0.95	0.95	0.95	0.95	0.91	23.53	23.53	23.53	23.53
PCRpcp2b	0.49	0.49	0.49	0.49	0.49	0.48	10.37	10.37	10.37	10.37
PCRicp2	0.96	0.96	0.96	0.96	0.96	0.96	6.74	6.74	6.74	6.74
PCRicp2b	0.96	0.96	0.96	0.96	0.96	0.96	6.74	6.74	6.74	6.74
PCRbss	2.34	2.74	3.27	3.02	1.18	3.96	12.41	18.49	17.02	17.94
PCRbss30	1.18	1.71	1.79	1.26	1.77	1.14	4.12	4.40	7.28	8.73
PLSRcv	6.98	5.78	23.09	37.27	26.39	0.68	2.37	13.18	2.06	4.17
PLSRcvd	3.49	6.27	11.29	36.88	34.18	0.44	2.47	4.58	2.18	3.94

SIR – As mentioned SIR requires a large sample to yield reliable estimates. In our companion paper Barbarino and Bura (2015) [7] we develop a Krylov subspaces version of SIR that can handle cases where $p > T$. However in line with the exploratory nature of this paper we wanted to test the performance of a standard SIR which necessitated pre-processing of the data over samples where $p > T$ or p is of order close to T. When we pre-condition the data we use 20 or 30 PCs, a number sufficient to preserve between 75% and 90% of the total variance in the panel. The idea is that by retaining a sufficient number of components not too much information on the conditional predictive density of $y_{t+h}|x_t$ is lost and the application of SIR on the reduced data can still identify and estimate a SDR subspace. We describe in detail the algorithm use to compute our pre-processed SIR in the Appendix. Although we are experimenting with non-parametric regression techniques in order to model the forward regression at this point we report only results obtained using a linear forward regression that models the dynamics of y_t and includes SIR components. This solution is suboptimal and likely negatively affects the forecasting accuracy of our estimator as it omits including non-linear terms which are implied by the higher than one dimension of the SIR predictors. Despite this fact our results are encouraging. Regarding the choice of the dimension of the SDR subspace, we tried to apply both the asymptotic weighted-chi square in Bura and Cook (2001b) [13] and the permutation test in Cook and Yin (2001) [24], however both proved to be very unstable and unreliable in our time-series settings. While we are working on the development of a test appropriate for our time-series environment, in this paper we estimate the dimension to be the number of SIR predictors resulting in the most accurate forecasts under several scenarios of constructing the SIR predictors and forward forecasting models. We verify that almost never a dimension larger than two is beneficial in the forecasting exercise. Notice that in the factor literature the information criteria used to select the number of components are somewhat cumbersome, involving some "model-mining" (for instance see the preceding discussion on $k_{\max}$). Our two-step procedure can be viewed as an alternative way of selecting the PC components whereby optimal selection is achieved by SDR techniques achieving as we will see shortly extreme parsimony.[12] We will also show that for large samples standard SIR applied to the raw variables

[12] This might have the flavor of a optimal weighting scheme in the extraction of the factors as suggested by Boivin

has competitive performance using only one or two SIR predictors.

5.3 Results: Forecasting Performance

We now turn to the analysis of the forecasting performance of the estimators that we have lined up in this study. We concentrate on the mean square forecast error (MSFE) as a measure of performance although broadly similar results are obtained using the mean absolute error criterion. The forecasting performance can depend on the range of the sample it is based on. This is particularly true as the "great recession" is covered in our data. To study such effect and also to be able to draw inference unencumbered by such effect, we consider several sample ranges.

Estimators that Use One Component – AR(4), OLS and RIDGE use only one linear combination to form their forecast. Also RIDGE does so. In Table 5 we report MSFE for these three methods at five different horizons. OLS, as expected, is greatly affected by the relatively small sample size. This said, in simulations conducted in a companion paper [7] we found that OLS are much more competitive when the data are generated by an exact factor model and only large deviation from such DGP or extreme paucity of observations disrupt the forecast efficiency of OLS. We do show in that paper that indeed an exact factor model implies that OLS is the correct model to use. We were surprised by the sub-par performance of RIDGE as it performs well very few times. This is in contrast with results in DeMol et al. (2008) [25]. We did feed into our RIDGE estimator (models RIDGE141 through RIDGE3532) also the parameters suggested in DeMol et al. (2008) [25] with little success. Data revisions, sample and estimation procedures may explain the difference.

Table 5: OLS, AR4 and RIDGE: MSFE Relative to AR4 (Estimation Sample: 1960m01-1982m01-2013m01)

	CPIAUCSL					**INDPRO**				
horizon	h=1	h=3	h=6	h=12	h=24	h=1	h=3	h=6	h=12	h=24
OLS	2.11	6.07	8.92	7.77	1.07	1.73	1.33	4.61	32.47	7.33
AR4	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
RIDGE	0.92	2.05	5.57	7.66	1.51	0.97	0.87	1.11	6.04	6.99
RIDGE1se	1.00	1.68	5.33	6.66	1.47	0.94	0.95	0.99	1.70	2.55
RIDGE141	1.10	1.56	2.04	2.36	2.73	0.92	0.98	1.00	1.05	1.02
RIDGE288	1.11	1.59	2.09	2.43	2.84	0.96	1.08	1.08	1.05	1.01
RIDGE292	1.11	1.59	2.09	2.43	2.84	0.96	1.08	1.08	1.05	1.01
RIDGE528	1.12	1.60	2.11	2.47	2.89	1.01	1.17	1.15	1.06	1.01
RIDGE582	1.12	1.60	2.12	2.47	2.89	1.01	1.18	1.16	1.07	1.01
RIDGE949	1.12	1.61	2.13	2.49	2.92	1.05	1.24	1.19	1.07	1.01
RIDGE3532	1.12	1.62	2.15	2.52	2.96	1.11	1.34	1.27	1.09	1.01

Principal Components Regression – We now turn to the performance of the diffusion index models in Table 6. PCR appears to be effective when forecasting one month ahead for both targets however forecasting inflation with PCR hits a wall at longer horizons. Including enough components appears to be key in general. Industrial production appears to be an easier to forecast and methods that select a relatively stable number of components are the most successful, such as fixing 8 or 10 components or using PCp2 or ICp2. Also BSS restricted to 30 components seems to be working

and Ng (2006) [10].

consistently well across horizons. PCR with 10 PCs has consistently good performance for CPI but not for industrial production.

Table 6: Principal Component Regression: MSFE Relative to AR4 (Estimation Sample: 1960m01-1982m01-2013m01)

	CPIAUCSL					**INDPRO**				
horizon	h=1	h=3	h=6	h=12	h=24	h=1	h=3	h=6	h=12	h=24
PCRcv	0.97	1.02	0.96	1.31	1.83	1.17	7.67	15.88	7.66	3.77
PCRmin1	0.95	1.07	1.11	1.18	1.37	1.00	1.06	1.04	1.00	0.90
PCRmin2	0.99	1.05	1.06	1.22	1.29	1.02	1.12	1.12	1.03	0.95
PCRpcp2	0.96	1.01	1.02	1.10	1.25	1.04	1.01	0.99	0.93	0.89
PCRicp2	0.95	1.06	1.11	1.19	1.38	0.92	0.89	0.94	0.96	1.06
PCRpcp2b	0.93	1.07	1.12	1.20	1.40	0.94	0.87	0.93	0.92	0.93
PCRicp2b	0.95	1.06	1.11	1.19	1.38	0.92	0.89	0.94	0.96	1.06
PCRbss	0.97	1.09	1.04	1.03	1.10	1.05	1.03	1.05	1.01	0.89
PCRbss30	0.97	1.06	1.06	1.16	1.30	0.96	0.93	0.95	0.91	0.90
PCRn1	1.02	1.05	1.08	1.13	1.34	0.96	0.96	0.98	0.99	1.01
PCRn2	1.03	1.07	1.08	1.11	1.22	0.95	0.94	1.00	1.01	1.08
PCRn3	1.03	1.07	1.09	1.14	1.27	0.94	0.93	0.98	0.99	1.07
PCRn4	0.98	1.07	1.09	1.16	1.33	0.93	0.93	0.98	0.99	1.08
PCRn5	0.97	1.07	1.10	1.18	1.39	0.93	0.94	0.99	1.00	1.07
PCRn6	0.95	1.05	1.10	1.19	1.40	0.93	0.93	1.00	1.01	1.08
PCRn8	0.95	1.06	1.11	1.19	1.38	0.92	0.85	0.92	0.94	1.05
PCRn10	0.93	1.07	1.12	1.21	1.41	0.94	0.87	0.93	0.93	0.94

Partial Least Squares Regression – The general impression from Table 7 is that cross-validation is very effective when forecasting industrial production at all horizons whereas fixing the number of components causes a deterioration of the MSFEs. The opposite seems to be true when forecasting inflation, in which case fixing the number of components to 6 or 8 seems the most appropriate choice. On balance, PLS comes out of the horserace as one of the best performers especially for inflation.

Table 7: Partial Least Squares Regression: MSFE Relative to AR4 (Estimation Sample: 1960m01-1982m01-2013m01)

	CPIAUCSL					**INDPRO**				
horizon	h=1	h=3	h=6	h=12	h=24	h=1	h=3	h=6	h=12	h=24
PLSRcv	1.02	1.09	1.06	1.01	0.87	0.93	0.87	0.95	0.89	0.86
PLSRcvd	0.96	1.00	1.02	1.09	1.09	0.96	0.90	0.97	0.95	0.87
PLSRn2	0.99	0.98	1.00	1.05	1.19	0.94	0.92	0.96	0.92	0.91
PLSRn4	0.90	0.94	0.94	1.00	1.14	1.09	1.08	1.02	0.93	0.85
PLSRn6	0.95	0.96	0.94	0.99	1.11	1.11	1.18	1.12	1.02	0.92
PLSRn8	0.97	0.98	0.95	0.98	1.07	1.08	1.11	1.08	1.00	0.91
PLSRn10	0.98	0.96	0.93	0.95	1.02	1.11	1.14	1.09	1.04	0.93

Sliced Inverse Regression – Table 8 reports SIR MSFE's relative to AR(4). Pre-conditioned

SIR (the first 4 lines of the table) on 30 or 20 PCs turns out to deliver some good results. The last two lines of the table report the results of using principal component regression. SIR is capable of summarizing in just one or two components the information encapsulated in 20 or 30 PCs. SIR improves the dismal performance of PCR in forecasting inflation in the medium term. We view the gain in parsimony and modeling as the major advantage of using SIR in this instance. As mentioned earlier our results are certainly adversely affected by forward model misspecification given that it appears that two SIR components capture all relevant information on the conditional distribution of $y_{t+h}^h|x_t$. Existing methods in the SDR literature exploit regression graphic devices in this case which are not easily ported in a pseudo forecasting experiment with estimation repeated hundreds of times. We are working to efficiently implement non-parametric methods in order to solve this problem. Finally we also report estimates for SIR on the raw data, without pre-conditioning. The sample is already long-enough to deliver a performance that slightly beats the AR(4) at short horizons although the estimation has been carried out using an inverse regression of the regressors on y_t rather than y_{t+h}^h in order to preserve as much information as possible in the estimation algorithm, an additional source of mispecification.

Table 8: Sliced Inverse Regression: MSFE Relative to AR4 (Estimation Sample: 1960m01-1982m01-2013m01)

	CPIAUCSL					**INDPRO**				
horizon	h=1	h=3	h=6	h=12	h=24	h=1	h=3	h=6	h=12	h=24
PC30SIRdr1OLS	0.99	0.97	0.95	1.01	1.13	1.07	1.04	1.12	0.95	0.94
PC30SIRdr2OLS	1.00	0.97	0.95	1.01	1.13	1.03	1.00	1.05	0.93	0.91
PC20SIRdr1OLS	0.99	0.97	1.02	1.07	1.18	0.99	0.97	1.03	0.94	0.94
PC20SIRdr2OLS	0.97	0.99	1.04	1.07	1.20	0.98	0.93	0.97	0.91	0.90
SIRdr1OLS	0.98	1.00	1.01	1.01	1.03	0.97	1.01	1.02	1.02	1.04
SIRdr2OLS	0.97	0.99	1.01	1.00	1.05	0.96	1.00	1.01	1.03	1.04
PCRn20	0.96	1.07	1.11	1.20	1.41	0.99	0.97	0.97	0.92	0.88
PCRn30	0.99	1.03	1.04	1.12	1.26	1.05	1.05	1.02	0.94	0.88

A Sub-Sample Comparison of SIR and PCR – In Tables 10-13, we report relative MSFEs with respect to AR(4) focusing attention to PCR, as the most widely used dimension reduction method in macro-economic forecasting, and SIR based forecasting models, whose regularized version also uses the PCs of the $\mathbf{x}$ variables. Because the SIR predictors are driven by the inverse regression of the predictors on the response, in a time series context, where the contemporaneous target variable and its lags can be used as predictors, different choices of variables to consider as predictors and response lead to different SIR models. The four different SIR based models we use are defined in Table 9 as follows. The second column describes how the SIR predictors are formed. For example, in SIRa, the SIR predictors are obtained from using all $\mathbf{x}$-predictors and y_t and its 4 lags as the $\mathbf{X}$-predictor matrix and y_{t+h} as the response that is sliced in the SIR algorithm of Section 4.3. When the PCs are used, then the regularized SIR algorithm in Appendix B is applied. The third column defines the forward forecasting model. For SIRa, for example, y_{t+h} is regressed on a linear model with inputs the corresponding SIR predictors from the second column. In Tables 10 and 12, only the regularized version of SIR is used as the starting sample is small relative to the number of predictors.

Table 9: SIR based Forecasting Models

	SIR Predictors (inverse regression)	**Forward Model**
SIRa	X or PCs and $y_t + 4$ lags on y_{t+h}	y_{t+h} on SIR predictors
SIRb	X or PCs and $y_t + 4$ lags on y_{t+h}	y_{t+h} on $y_t + 4$ lags and SIR predictors
SIRc	X or PCs on y_{t+h}	y_{t+h} on $y_t + 4$ lags and SIR predictors
SIRd	X or PCs on y_t	y_{t+h} on $y_t + 4$ lags and SIR predictors

For both inflation and industrial production, the general pattern across forecasting windows and horizons is that SIR, either standard or regularized, has similar performance to PCR. For the longest horizon of 24 months, SIR with has better performance. The only exception is for industrial production over the period 2003:01-2014:12 for models SIRa, SIRb and SIRc (see relative MSFEs in Table 13) where SIR predictors based on all 129 $\mathbf{x}$ variables are used. In contrast, over 2010:01-2014:12, the performance is on par with the other methods. This finding confirms that the sample size has a dramatic impact in SIR performance. Notably, SIRd, where the SIR predictors are built using only y_t, does not appear to be affected by the size of the sample. In effect, for these econometric series, SIRd exhibits overall the best performance for both PC-based and standard SIR across periods and horizons. PCR typically needs 10 components to achieve its best performance across horizons and time windows. In sum, SIR is shown to achieve what it is designed to do; that is, significantly reduce the dimension of the forecasting problem.

Table 10: Relative MSFE with respect to AR(4) for predicting CPIAUCSL

CPIAUCSL	**1971:01-2014:12**					**1982:01-2014:12**				
horizon	h=1	h=3	h=6	h=12	h=24	h=1	h=3	h=6	h=12	h=24
PCRn4	0.96	0.974	0.95	0.894	0.767	0.97	1.07	1.092	1.153	1.33
PCRn5	0.96	0.98	0.959	0.909	0.791	0.97	1.07	1.102	1.178	1.4
PCRn10	0.94	0.982	0.976	0.913	0.763	0.93	1.07	1.119	1.202	1.41
PCRn20	1	0.996	0.989	0.934	0.779	0.96	1.07	1.117	1.202	1.41
PC20SIRa	0.99	1.008	1.022	1.01	0.805	0.935	1.07	0.935	1.3	1.45
PC20SIRb	0.954	0.983	0.999	0.938	0.763	0.947	1.05	0.947	1.194	1.35
PC20SIRc	0.972	1.001	1.015	0.933	0.768	0.972	1	0.972	1.067	1.2
PC20SIRd	0.984	1.004	0.982	0.953	0.939	0.977	1.01	0.977	0.995	1.05

Table 11: Relative MSFE with respect to AR(4) for predicting CPIAUCSL

CPIAUCSL	**2003:01-2014:12**					**2010:01-2014:12**				
horizon	h=1	h=3	h=6	h=12	h=24	h=1	h=3	h=6	h=12	h=24
PCRn4	0.967	1.08	1.144	1.229	1.372	0.962	1.091	1.121	1.136	1.124
PCRn5	0.961	1.083	1.161	1.267	1.469	0.963	1.078	1.105	1.188	1.204
PCRn10	0.901	1.079	1.183	1.311	1.516	1.001	1.134	1.185	1.293	1.3
PCRn20	0.918	1.065	1.174	1.311	1.493	1.065	1.134	1.161	1.264	1.221
PC20SIRa	0.921	1.063	1.217	1.414	1.537	1.1	1.136	1.21	1.508	1.406
PC20SIRb	0.935	1.05	1.185	1.291	1.408	1.073	1.102	1.157	1.24	1.192
PC20SIRc	0.963	0.998	1.079	1.068	1.162	1.117	1.023	1.053	1.057	1.314
PC20SIRd	0.968	1.005	0.993	0.99	1	1.045	1.04	0.986	0.973	1.208
SIRa	57.416	30.219	6.788	7.059	8.975	1.597	1.345	1.39	1.339	1.917
SIRb	25.035	18.929	4.659	5.897	8.183	1.209	1.122	1.192	1.238	1.78
SIRc	20.473	25.865	2.388	11.89	1.902	1.057	1.145	1.062	1.378	2.225
SIRd	0.948	0.981	0.991	0.989	1.012	0.982	0.997	1.023	0.999	1.012

Table 12: Relative MSFE with respect to AR(4) for predicting CPIAUCSL

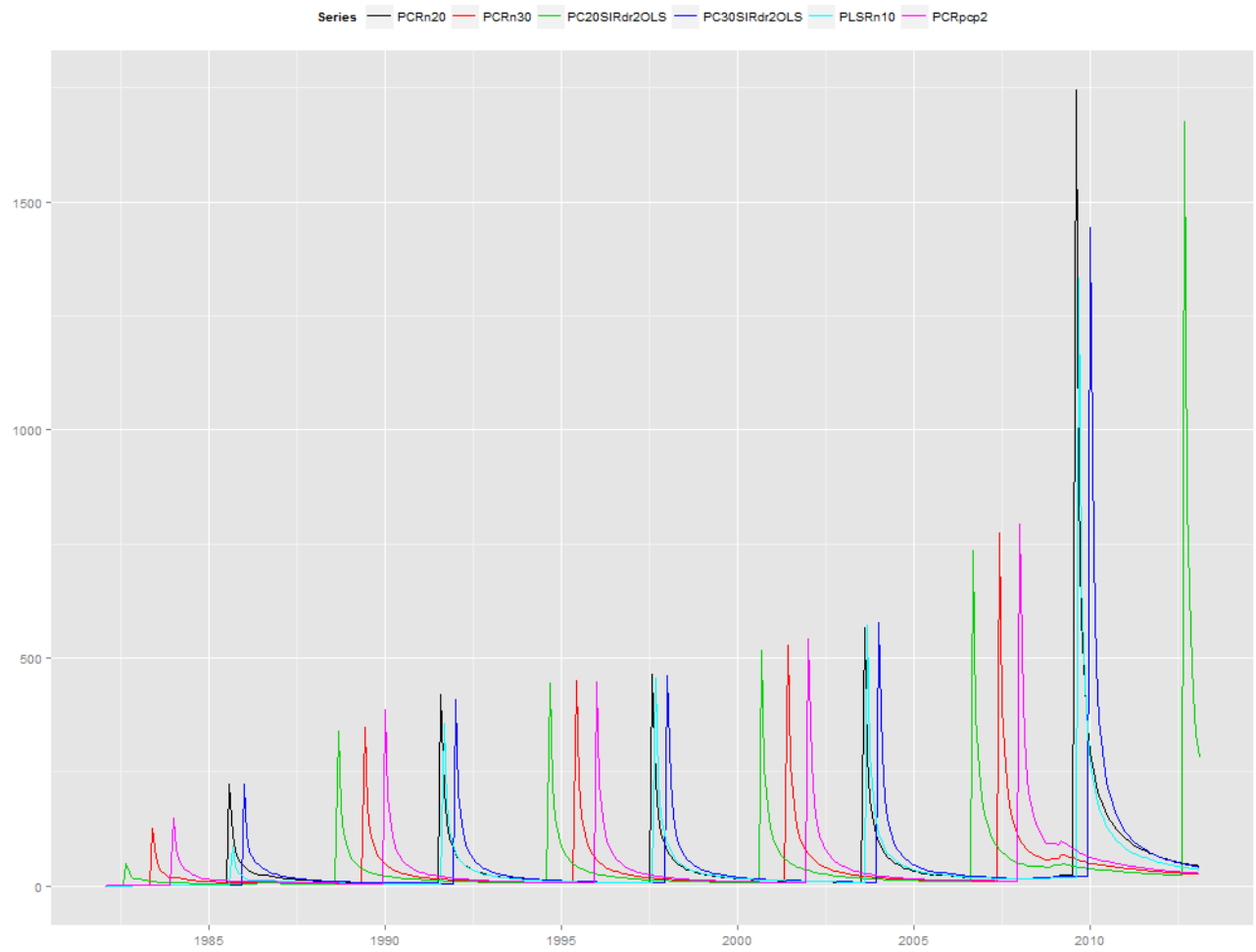
INDPRO	**1971:01-2014:12**					**1982:01-2014:12**				
horizon	h=1	h=3	h=6	h=12	h=24	h=1	h=3	h=6	h=12	h=24
PCRn4	0.904	0.8	0.761	0.717	0.75	0.943	0.947	0.996	0.989	1.078
PCRn5	0.897	0.8	0.769	0.718	0.74	0.937	0.946	0.999	0.996	1.069
PCRn10	0.914	0.83	0.807	0.713	0.65	0.932	0.884	0.94	0.927	0.946
PCRn20	0.954	0.9	0.868	0.741	0.64	0.987	0.986	0.983	0.922	0.882
PC20SIRa	0.946	0.86	0.867	0.742	0.64	0.977	0.93	0.977	0.921	0.91
PC20SIRb	0.94	0.88	0.872	0.741	0.64	0.996	0.95	0.996	0.917	0.901
PC20SIRc	0.914	0.89	0.819	0.693	0.62	0.975	0.95	0.975	0.912	0.907
PC20SIRd	1.008	1.02	1.002	0.973	1	1.035	1.036	1.035	1.002	0.976

Table 13: Relative MSFE with respect to AR(4) for predicting CPIAUCSL

INDPRO	**2003:01-2014:12**					**2010:01-2014:12**				
horizon	h=1	h=3	h=6	h=12	h=24	h=1	h=3	h=6	h=12	h=24
PCRn4	0.972	1.015	1.024	0.942	1.032	1.058	1.449	1.675	2.092	11.594
PCRn5	0.987	1.009	1.022	0.925	0.994	1.043	1.34	1.38	0.875	4.311
PCRn10	0.994	0.946	0.965	0.849	0.814	1.086	1.629	1.937	2.405	4.521
PCRn20	1.038	0.986	0.959	0.783	0.657	1.037	1.619	1.962	2.78	4.701
PC20SIRa	1.008	0.912	0.965	0.77	0.685	0.986	1.423	1.594	1.863	6.566
PC20SIRb	1.02	0.952	0.976	0.768	0.673	1.006	1.517	1.626	1.868	6.835
PC20SIRc	1.019	0.961	0.954	0.774	0.679	1.062	1.597	1.653	1.985	7.596
PC20SIRd	1.027	1.043	1.012	0.973	0.922	1.045	1.183	1.085	1.203	3.425
SIRa	18.339	2.83	9.592	9.422	10.488	1.966	2.729	2.769	3.659	12.07
SIRb	15.967	2.668	9.295	8.723	10.446	1.815	2.652	2.648	3.6	11.782
SIRc	10.115	1.312	7.347	5.717	6.459	1.697	2.37	2.672	4.013	11.232
SIRd	0.93	1.008	1.033	1.044	1.043	0.956	1.125	1.149	1.29	1.378

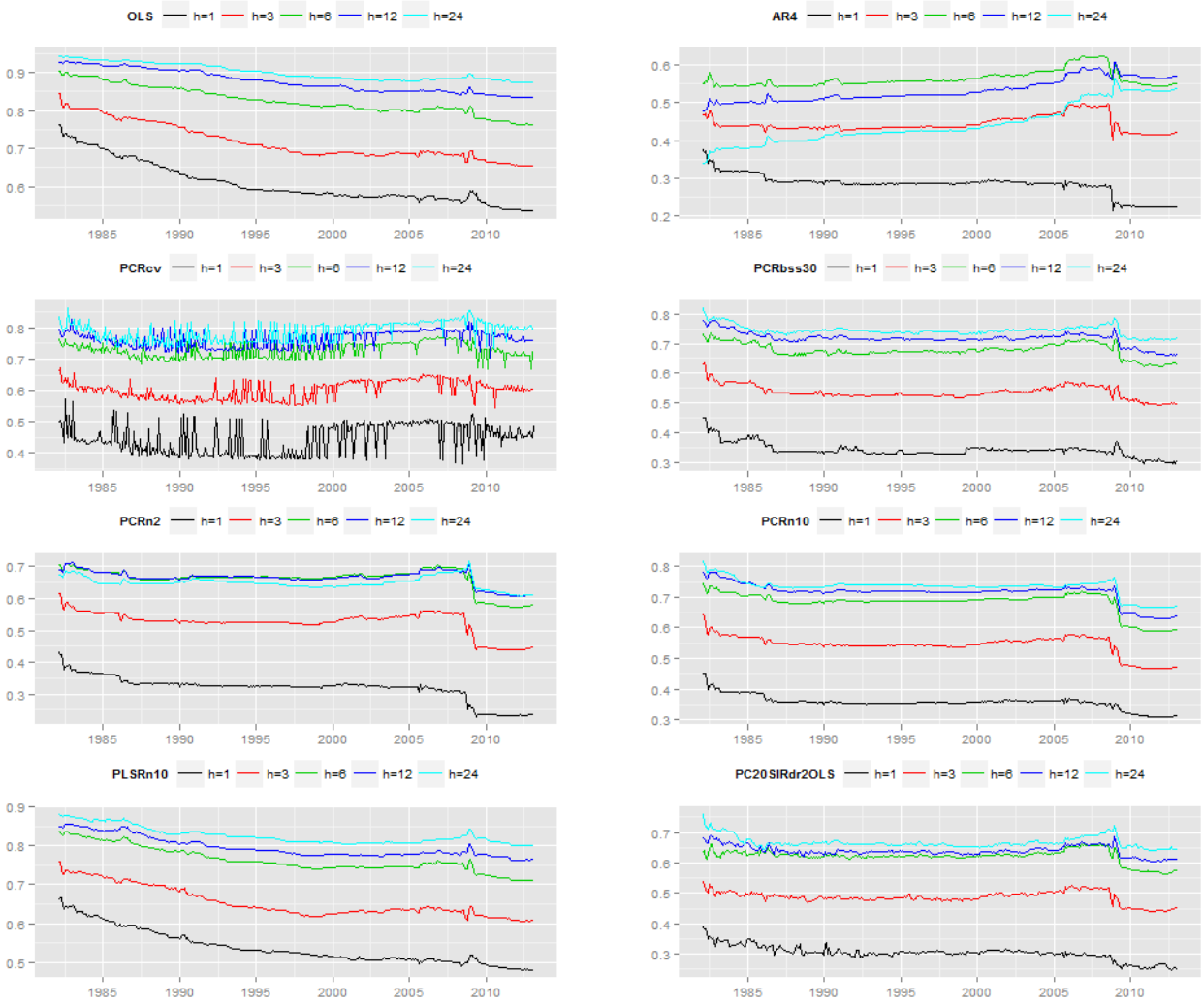
Evolution of the MSFE – Inflation appears definitely harder to forecast than industrial production, consistent with findings in the forecasting literature. Are there specific periods in which the forecast performance of our estimators deteriorates? Figure 6 portrays the evolution of the MSFE in forecasting inflation 12 months ahead for selected estimators (so each point represents the MSFE up to that point). There are definitely periods in which forecasting inflation is harder, however it seems that these periods vary by estimator with SIR computed out of 20 PCs being the first to react negatively to bad data entering the sample through the rolling window. Using asa a metric the height of the spikes it looks like PC20SIRdr performs at par if not better than other estimators except in the very last portion of the window.

Figure 6: CPIAUCSL: Rolling MSFE for Selected Estimators



In-Sample Fit – As is evident from Figure 7 there is no obvious relationship between in-sample fit and the out-of-sample forecasting performance commented above. For instance, OLS thanks to the very large number of variables, some subcomponents of the target variable itself, produces very high R-squared and bad forecasting results.

Figure 7: CPIAUCSL: R-Squared for Selected Estimators over Pseudo Forecasting Window



6 Conclusions

7 Summary and Conclusions

In this paper we (1) introduced sufficient dimension reduction methodology to econometric forecasting and focused on linear moment-based SDR, (2) derived properties of the SIR SDR estimator for covariance stationary series, (3) cast OLS, PCR, RIDGE and PLS in a common framework , and (4) studied the forecasting performance of these four methods, as well as SIR, using the FRED-MD data set put together by McCracken and Ng (2015) [49]. The empirical results indicate that PCR, PLS and SIR do not exhibit drastically different performance. The competitive edge of SIR is its parsimony: it attains practically the same forecasting accuracy using one or two linear combinations of the predictors. In contrast, both PLS and PCR require many components, in many cases more than ten. OLS and RIDGE are not found to be competitive for these data and the time periods we considered in our forecasting exercise.

There are several issues that impede the performance of SIR and which can be improved upon. Dimension two or higher in SIR indicates the presence of nonlinear relationships between the response and the SIR predictors. In such cases, plots of the response versus the SIR predictors would inform the construction of a more appropriate forward model. As the forecasting experiment was carried out in an automatic fashion, we could not visually assess the nature of nonlinearities and nonlinear functions of the SIR predictors were not included in the forecasting model. Gains in forecasting accuracy can potentially be realized by the inclusion of nonlinear SIR terms in the forward model.

SDR in general, and SIR in particular, require a large sample size to yield reliable results. The sample size of the FRED-MD data set is not large enough for SIR to be optimally used. For some periods in the forecasting exercise, SIR predictors were extremely unstable as the sample covariance matrix of the raw predictors was close to ill-conditioned. We develop a Krylov subspaces version of SIR to address this issue in a separate paper. Nevertheless, both these issues amount to limitations that need to be addressed for SIR, or SDR in general, to be properly applied to such data and deserve future empirical and theoretical research.

Appendix A: List and Description of Variables

The following set of tables summarizes the variables in FRED-MD. Each table collects variables by statistical data release imparting an organization of the variables slightly rearranged relative to the tables in McCracken and Ng (2015) [49]. Grouping by statistical data release report is more useful both because the timing of the data release is different (although the timing is not exploited in the present study) and because some variables are aggregates of more detailed information in any one statistical data release and share the information and possible biases of that statistical release. Our reordering allows a better bird's eye view on the sources of information. We briefly describe each statistical data release below. In addition each table reports, under the column **T,** the transformation used[13]. The **G** column denotes the grouping chosen by McCracken and Ng (2015) [49] in turn not too dissimilar from groupings operated in other DFM studies. The **FRED-MD** column reports the variable mnemonics in the original FRED-MD datasets. The **Description** column permits to identify the series. The remaining two columns denote the Global Insight code and description; the GSI description allows to map the individual series with datasets in older papers. In some papers not all variables are used to compute principal components, a strategy

[13] The transformations closely follow McCracken and Ng (2015) [49] who in turn follow Stock and Watson.

followed by Stock and Watson (2005) [56], who add an additional column containing a dummy to denote whether the variable was used in the computation of the PCs. We include all variables when computing PCs hence we do not need such additional column. Asterisked series are adjusted by McCracken and Ng (2015) (see [49] for details).

Variables Directly Measuring Output – The most reliable and used data containing measures of output at a monthly frequency come from the IP system within the statistical release G.17 produced at the Federal Reserve Board and covering industrial production. The IP system contains information on about 200 sectors at NAICS 4-digits level and covers the manufacturing, mining and utilities sectors. The last variable is capacity utilization in manufacturing, one of the few observable measures of slack also from the G.17, computed as $\frac{\text{manufacturing IP}}{\text{manufacturing capacity}}$; manufacturing capacity is estimated by staff at the FRB using the quarterly survey of capacity (in turn run by the BLS) and included in the G.17 publication. The G.17 publication contains information on about 94 subaggregates at NAICS 4 digit level whereas the IP system used to produce it is based on 200+ atoms. Apart from the top aggregate INDPRO, the next seven rows represent the splitting and regrouping of the 200+ atoms in so called "market" groups. The last market group is split in two subaggregates, durable and non-durable materials. Manufacturing IP is a subaggregate of IP at the same level as Utilities. Fuels IP is an odd series to be included in this dataset given its idiosyncratic pattern and its higher level of detail. Notice that 25% of final industrial production data (that is after all revisions have taken place), are estimated from employment data (in the second table of this Section), implying that this set of variables and the set in the second tables might be strongly linked or have a factor in common.

Table 14: Output Variables from the IP System

id	T	G	FRED-MD	Description	GSI Description
6	5	1	INDPRO	IP Index	IP: total
7	5	1	IPFPNSS	IP: Final Products and Nonindustrial Supplies	IP: products
8	5	1	IPFINAL	IP: Final Products (Market Group)	IP: final prod
9	5	1	IPCONGD	IP: Consumer Goods	IP: cons gds
10	5	1	IPDCONGD	IP: Durable Consumer Goods	IP: cons dble
11	5	1	IPNCONGD	IP: Nondurable Consumer Goods	IP: cons nondble
12	5	1	IPBUSEQ	IP: Business Equipment	IP: bus eqpt
13	5	1	IPMAT	IP: Materials	IP: matls
14	5	1	IPDMAT	IP: Durable Materials	IP: dble matls
15	5	1	IPNMAT	IP: Nondurable Materials	IP: nondble matls
16	5	1	IPMANSICS	IP: Manufacturing (SIC)	IP: mfg
17	5	1	IPB51222s	IP: Residential Utilities	IP: res util
18	5	1	IPFUELS	IP: Fuels	IP: fuels
20	2	1	CUMFNS	Capacity Utilization: Manufacturing	Cap util

Variables Measuring Income and Consumption – Personal Income, personal consumption expenditures and PCE deflators are released monthly by the BEA. Retail sales are released by the Census Bureau.

Table 15: Variables Helpful in Estimating Consumption

id	T	G	FRED-MD	Description	GSI Description
1	5	1	RPI	Real Personal Income	PI
2	5	1	W875RX1	Real personal income ex transfer receipts	PI less transfers
3	5	4	DPCERA3M086SBEA	Real personal consumption expenditures	Real Consumption
4*	5	4	CMRMTSPLx	Real Manu. and Trade Industries Sales	MT sales
5*	5	4	RETAILx	Retail and Food Services Sales	Retail sales
123	6	7	PCEPI	Personal Cons. Expend.: Chain Price Index	PCE defl
124	6	7	DDURRG3M086SBEA	Personal Cons. Expend: Durable goods	PCE defl: dlbes
125	6	7	DNDGRG3M086SBEA	Personal Cons. Expend: Nondurable goods	PCE defl: nondble
126	6	7	DSERRG3M086SBEA	Personal Cons. Expend: Services	PCE defl: service

Variables Measuring Employment and Unemployment – The second table contains information on variables measuring employment, data produced by the Bureau of Labor Statistcs (BLS). The first two rows refer to data from the Current Population Survey (CPS). The rest of the table refers to variables from the Current Employment Statistics (CES) a program run each month that surveys approximately 143,000 businesses and government agencies, representing approximately 588,000 individual worksites. The last 3 variables contain miscellaneous information on the labor market. CLAIMS=unemployment claims, is a variable originally released at weekly frequency and comes from the states unemployment insurance system. HWI=Help-Wanted Index for United States is assembled by the Conference Board and recently it has been corrected by Barnichon (2010) [8]. Obvious candidates missing in the datasets are labor market indicators part of the FED labor market dashboard, such as data from the JOLTS survey.

Table 16: Employment Variables from Household CPS and Payroll CES Surveys

id	T	G	FRED-MD	Description	GSI Description
23	5	2	CLF16OV	Civilian Labor Force	Emp CPS total
24	5	2	CE16OV	Civilian Employment	Emp CPS nonag
25	2	2	UNRATE	Civilian Unemployment Rate	U: all
26	2	2	UEMPMEAN	Average Duration of Unemployment (Weeks)	U: mean duration
27	5	2	UEMPLT5	Civilians Unemployed - Less Than 5 Weeks	U < 5 wks
28	5	2	UEMP5TO14	Civilians Unemployed for 5-14 Weeks	U 5-14 wks
29	5	2	UEMP15OV	Civilians Unemployed - 15 Weeks Over	U 15+ wks
30	5	2	UEMP15T26	Civilians Unemployed for 15-26 Weeks	U 15-26 wks
31	5	2	UEMP27OV	Civilians Unemployed for 27 Weeks and Over	U 27+ wks
33	5	2	PAYEMS	All Employees: Total nonfarm	Emp: total
34	5	2	USGOOD	All Employees: Goods-Producing Industries	Emp: gds prod
35	5	2	CES1021000001	All Employees: Mining and Logging: Mining	Emp: mining
36	5	2	USCONS	All Employees: Construction	Emp: const
37	5	2	MANEMP	All Employees: Manufacturing	Emp: mfg
38	5	2	DMANEMP	All Employees: Durable goods	Emp: dble gds
39	5	2	NDMANEMP	All Employees: Nondurable goods	Emp: nondbles
40	5	2	SRVPRD	All Employees: Service-Providing Industries	Emp: services
41	5	2	USTPU	All Employees: Trade, Transportation Utilities	Emp: TTU
42	5	2	USWTRADE	All Employees: Wholesale Trade	Emp: wholesale
43	5	2	USTRADE	All Employees: Retail Trade	Emp: retail
44	5	2	USFIRE	All Employees: Financial Activities	Emp: FIRE
45	5	2	USGOVT	All Employees: Government	Emp: Govt
46	1	2	CES0600000007	Avg Weekly Hours : Goods-Producing	Avg hrs
47	2	2	AWOTMAN	Avg Weekly Overtime Hours : Manufacturing	Overtime: mfg
48	1	2	AWHMAN	Avg Weekly Hours : Manufacturing	Avg hrs: mfg
49	1	2	NAPMEI	ISM Manufacturing: Employment Index	NAPM empl
127	6	2	CES0600000008	Avg Hourly Earnings : Goods-Producing	AHE: goods
128	6	2	CES2000000008	Avg Hourly Earnings : Construction	AHE: const
129	6	2	CES3000000008	Avg Hourly Earnings : Manufacturing	AHE: mfg
32*	5	2	CLAIMSx	Initial Claims	UI claims
21*	2	2	HWI	Help-Wanted Index for United States	Help wanted indx
22*	2	2	HWIURATIO	Ratio of Help Wanted/No. Unemployed	Help wanted/unemp

Variables Measuring Construction Activity – The third table collects the variables that have leading properties in signaling changes in activity in the construction sector. Permits variables come from the Census' building permits monhtly survey of 9,000 selected permit-issuing places adjusted once a year with an annual census of an additional 11,000 permit places that are not in the monthly sample. Housing starts come from the Survey of Construction, a multi-stage stratified random sample that selects approximately 900 building permit-issuing offices, and a sample of more than 70 land areas not covered by building permits. Data from the national association of home buildiers such as existing home sales were not included oin the dataset.

Table 17: Leading Indicators of the Construction Sector

id	T	G	FRED-MD	Description	GSI Descr
50	4	3	HOUST	Housing Starts: Total New Privately Owned	Starts: nonfarm
51	4	3	HOUSTNE	Housing Starts, Northeast	Starts: NE
52	4	3	HOUSTMW	Housing Starts, Midwest	Starts: MW
53	4	3	HOUSTS	Housing Starts, South	Starts: South
54	4	3	HOUSTW	Housing Starts, West	Starts: West
55	4	3	PERMIT	New Private Housing Permits (SAAR)	BP: total
56	4	3	PERMITNE	New Private Housing Permits, Northeast (SAAR)	BP: NE
57	4	3	PERMITMW	New Private Housing Permits, Midwest (SAAR)	BP: MW
58	4	3	PERMITS	New Private Housing Permits, South (SAAR)	BP: South
59	4	3	PERMITW	New Private Housing Permits, West (SAAR)	BP: West

Variables Measuring Orders and Inventories – These variables are from the M3 survey run by the U.S. Census Bureau. The M3 is based upon data reported from manufacturing establishments with $500 million or more in annual shipments. Units may be divisions of diversified large companies, large homogenous companies, or single-unit manufacturers in 89 industry categories. The M3 provides statistics on manufacturers' value of shipments, new orders (net of cancellations), end-of-month order backlog (unfilled orders), end-of-month total inventory, materials and supplies, work-in-process, and finished goods inventories (at current cost or market value). Data are collected and tabulated predominantly by 6-digit NAICS (North American Industry Classification System). The most watched series from this survey is ANDENO="New Orders for Nondefense Capital Goods" since it excludes certain highly volatile goods (and not so informative on the business cycle) from new orders. Such series unfortunately has a short history and it is excluded in our estimation.

Table 18: Variables from the M3 Survey

id	T	G	FRED-MD	Description	GSI Description
3	5	4	DPCERA3M086SBEA	Real personal consumption expenditures	Real Consumption
4*	5	4	CMRMTSPLx	Real Manu. and Trade Industries Sales	MT sales
5*	5	4	RETAILx	Retail and Food Services Sales	Retail sales
64	5	4	ACOGNO	New Orders for Consumer Goods	Orders: cons gds
65*	5	4	AMDMNOx	New Orders for Durable Goods	Orders: dble gds
66*	5	4	ANDENOx	New Orders for Nondefense Capital Goods	Orders: cap gds
67*	5	4	AMDMUOx	Unfilled Orders for Durable Goods	Unf orders: dble
68*	5	4	BUSINVx	Total Business Inventories	MT invent
69*	2	4	ISRATIOx	Total Business: Inventories to Sales Ratio	MT invent/sales

Variables Measuring the Money Stock and Reserves – These data come mainly from the FRB H.6 statistical release.

Table 19: Variables Measuring the Money Stock and Bank Reserves

id	T	G	FRED-MD	Description	GSI Description
70	6	5	M1SL	M1 Money Stock	M1
71	6	5	M2SL	M2 Money Stock	M2
72	5	5	M2REAL	Real M2 Money Stock	M2 (reaal)
73	6	5	AMBSL	St. Louis Adjusted Monetary Base	MB
74	6	5	TOTRESNS	Total Reserves of Depository Institutions	Reserves tot
75	7	5	NONBORRES	Reserves Of Depository Institutions, Nonborrowed	Reserves nonbor

Variables Measuring Credit – These variables are mainly drawn from various FRB statistical releases such as G.19 and G.20.

Table 20: Variables Measuring Credit

id	T	G	FRED-MD	Description	GSI Descr
76	6	5	BUSLOANS	Commercial and Industrial Loans, All Commercial Banks	CI loan plus
77	6	5	REALLN	Real Estate Loans at All Commercial Banks	DCI loans
78	6	5	NONREVSL	Total Nonrevolving Credit Owned and Securitized Outstanding	Cons credit
79*	2	5	CONSPI	Nonrevolving consumer credit to Personal Income	Inst credit/PI
131	6	5	MZMSL	MZM Money Stock	N.A.
132	6	5	DTCOLNVHFNM	Consumer Motor Vehicle Loans Outstanding	N.A.
133	6	5	DTCTHFNM	Total Consumer Loans and Leases Outstanding	N.A.
134	6	5	INVEST	Securities in Bank Credit at All Commercial Banks	N.A.

Variables Measuring Interest Rates – The following table contains variables measuring interest rates, yields and spreads. Most variables are from statistical releases by the FRB such as the H.15.

Table 21: Interest Rates, Yields and Spreads

id	T	G	FRED-MD	Description	GSI Descr
84	2	6	FEDFUNDS	Effective Federal Funds Rate	Fed Funds
85*	2	6	CP3Mx	3-Month AA Financial Commercial Paper Rate	Comm paper
86	2	6	TB3MS	3-Month Treasury Bill:	3 mo T-bill
87	2	6	TB6MS	6-Month Treasury Bill:	6 mo T-bill
88	2	6	GS1	1-Year Treasury Rate	1 yr T-bond
89	2	6	GS5	5-Year Treasury Rate	5 yr T-bond
90	2	6	GS10	10-Year Treasury Rate	10 yr T-bond
91	2	6	AAA	Moody's Seasoned Aaa Corporate Bond Yield	Aaa bond
92	2	6	BAA	Moody's Seasoned Baa Corporate Bond Yield	Baa bond
93*	1	6	COMPAPFFx	3-Month Commercial Paper Minus FEDFUNDS	CP-FF spread
94	1	6	TB3SMFFM	3-Month Treasury C Minus FEDFUNDS	3 mo-FF spread
95	1	6	TB6SMFFM	6-Month Treasury C Minus FEDFUNDS	6 mo-FF spread
96	1	6	T1YFFM	1-Year Treasury C Minus FEDFUNDS	1 yr-FF spread
97	1	6	T5YFFM	5-Year Treasury C Minus FEDFUNDS	5 yr-FF spread
98	1	6	T10YFFM	10-Year Treasury C Minus FEDFUNDS	10 yr-FF spread
99	1	6	AAAFFM	Moody's Aaa Corporate Bond Minus FEDFUNDS	Aaa-FF spread
100	1	6	BAAFFM	Moody's Baa Corporate Bond Minus FEDFUNDS	Baa-FF spread
101	5	6	TWEXMMTH	Trade Weighted U.S. Dollar Index: Major Currencies	Ex rate: avg
102*	5	6	EXSZUSx	Switzerland / U.S. Foreign Exchange Rate	Ex rate: Switz
103*	5	6	EXJPUSx	Japan / U.S. Foreign Exchange Rate	Ex rate: Japan
104*	5	6	EXUSUKx	U.S. / U.K. Foreign Exchange Rate	Ex rate: UK
105*	5	6	EXCAUSx	Canada / U.S. Foreign Exchange Rate	EX rate: Canada

Variables Measuring Prices – The variables are from the BLS CPI and PPI statistical releases. For PPI more than 100,000 price quotations per month are organized into three sets of PPIs: (1) Final demand-Intermediate demand (FD-ID) indexes, (2) commodity indexes, and (3) indexes for the net output of industries and their products. The CPIs are based on prices of food, clothing, shelter, fuels, transportation fares, charges for doctors'

and dentists' services, drugs, and other goods and services that people buy for day-to-day living. Prices are collected each month in 87 urban areas across the country from about 6,000 housing units and approximately 24,000 retail establishments-department stores, supermarkets, hospitals, filling stations, and other types of stores and service establishments.

Table 22: Measures of Prices

id	T	G	FRED-MD	Description	GSI Descr
106	6	7	PPIFGS	PPI: Finished Goods	PPI: fin gds
107	6	7	PPIFCG	PPI: Finished Consumer Goods	PPI: cons gds
108	6	7	PPIITM	PPI: Intermediate Materials	PPI: int matls
109	6	7	PPICRM	PPI: Crude Materials	PPI: crude matls
110*	6	7	OILPRICEx	Crude Oil, spliced WTI and Cushing	Spot market price
111	6	7	PPICMM	PPI: Metals and metal products:	PPI: nonferrous
113	6	7	CPIAUCSL	CPI : All Items	CPI-U: all
114	6	7	CPIAPPSL	CPI : Apparel	CPI-U: apparel
115	6	7	CPITRNSL	CPI : Transportation	CPI-U: transp
116	6	7	CPIMEDSL	CPI : Medical Care	CPI-U: medical
117	6	7	CUSR0000SAC	CPI : Commodities	CPI-U: comm.
118	6	7	CUUR0000SAD	CPI : Durables	CPI-U: dbles
119	6	7	CUSR0000SAS	CPI : Services	CPI-U: services
120	6	7	CPIULFSL	CPI : All Items Less Food	CPI-U: ex food
121	6	7	CUUR0000SA0L2	CPI : All items less shelter	CPI-U: ex shelter
122	6	7	CUSR0000SA0L5	CPI : All items less medical care	CPI-U: ex med

Variables Measuring the Stock Market – These data are elaborated by Standard & Poor.

Table 23: Measures of the Stock Market from Standard and Poor

id	T	G	FRED-MD	Description	GSI Descr
80*	5	8	SP 500	SP's Common Stock Price Index: Composite	SP 500
81*	5	8	SP: indust	SP's Common Stock Price Index: Industrials	SP: indust
82*	2	8	SP div yield	SP's Composite Common Stock: Dividend Yield	SP div yield
83*	5	8	SP PE ratio	SP's Composite Common Stock: Price-Earnings Ratio	SP PE ratio

Diffusion Indexes from Manufacturing and Consumer Surveys – The last table mostly collects the diffusion indexes from the Institute for Supply Management (ISM)[14]. These variables are released the first day of month,following the reference month, hence they are quite timely and are used by several institutions in the produciton of their high frequency forecasts. However, since the literature on large panels of macro variables carries out monthly pseudo-forecast experiments and we follow that tradition, we do not exploit the full potential of these variables. Hence the only variable that likely has the most forecasting power is "new orders" a natural measure of future activity. Notice that these variables are diffusion indexes, that is they essentially capture the fraction of respondents that say that activity is up[15]. Given that they are diffusion indexes (fractions) they are stable and they are left in levels in the estimation. The dataset does not include data from other manufacturing surveys used in the construction of activity indexes and nowcasting such as the Philly Fed BOS survey or the Richmond Fed survey as well as information from the services

[14]Formerly known as the National Association of Purchasing Managers (NAPM).

[15]The ISM also reports other interesting diffusion indexes such as "new export orders", or "level of inventories", but these variables are available only starting from the 1990s, too short a time series to include them in pseudo-out-of-sample forecasting experiments. The same is true for the recently introduced diffusion indexes from the Markit survey.

surveys: most likely the choice of the authors was dictated by the span of available data. The last variable in the table is the closely watched Consumer Sentiment Index from the University of Michigan used in the forecast of comsuption expenditures. Other informative sub-indexes of the Michigan survey were not included in the dataset.

Table 24: Diffusion Indexes from the ISM Manufacturing Survey and the UM Consumer Survey

id	T	G	FRED-MD	Description	GSI Descr
19	1	1	NAPMPI	ISM Manufacturing: Production Index	NAPM prodn
29	1	2	NAPMEI	ISM Manufacturing: Employment Index	NAPM empl
60	1	4	NAPM	ISM : PMI Composite Index	PMI
61	1	4	NAPMNOI	ISM : New Orders Index	NAPM new ordrs
62	1	4	NAPMSDI	ISM : Supplier Deliveries Index	NAPM vendor del
63	1	4	NAPMII	ISM : Inventories Index	NAPM Invent
112	1	7	NAPMPRI	ISM Manufacturing: Prices Index	NAPM com price
130*	2	4	UMCSENTx	Consumer Sentiment Index	Consumer expect

8 Appendix B: Regularized SIR Algorithm

In relevance to the forecasting model (2.1), the response is y_{t+h}, $t = 1, \ldots, T, \ldots$, and the predictors consist of a group of p exogenous variables $\mathbf{x}_t = (x_{t1}, \ldots, x_{tp})'$ and the current response value y_t along with L of its lags, which is denoted by $\mathbf{W}_t = (y_{t-1}, \ldots, y_{t-L})'$.

1. Carry-out PCA on the sample predictor matrix $\mathbf{X}_T : T \times p$

   a. Compute the spectral decomposition of $\widehat{\Sigma} = \widehat{\mathbf{V}}\widehat{\mathbf{D}}\widehat{\mathbf{V}}'$, where $\widehat{\mathbf{V}} = (\widehat{\mathbf{v}}_1, \ldots, \widehat{\mathbf{v}}_p)$ are the $\widehat{\Sigma}$ eigenvectors, and $\widehat{\mathbf{D}} = diag(\hat{\lambda}_1, \ldots, \hat{\lambda}_p)$ is the diagonal matrix with the eigenvalues of $\widehat{\Sigma}$ arranged in decreasing order.

   b. Let M be the number of principal components that capture most of the variability in $\mathbf{X}$, either by formal tests such as Bai and Ng (2002) [2] or by simply surveying the scree plot, i.e. the plot of the ordered eigenvalues versus component number. A scree plot displays the proportion of the total variation in a dataset that is explained by each of the components in a principle component analysis. Using the scree plot, the number of components is estimated to be the number corresponding to the "elbow" of the plot.

   c. Let $F_1 = \widehat{\mathbf{v}}_1'\mathbf{X}, \ldots, F_M = \widehat{\mathbf{v}}_M'\mathbf{X}$ be the retained principal factors of $\mathbf{X}$.

2. Let $\widetilde{\mathbf{X}}_t = (F_{t1}, \ldots, F_{tM}, Y_t, Y_{t-1}, \ldots, Y_{t-L})' = (\tilde{X}_1, \ldots, \tilde{X}_{M+L+1})'$ be the $(M + L + 1) \times 1$ vector of adjusted predictors, and let $q = M + L + 1$ where L denotes the lags of y_t.

3. Set $\bar{\widetilde{\mathbf{X}}} = (\bar{\tilde{X}}_1, \ldots, \bar{\tilde{X}}_q)^T$, where $\bar{\tilde{X}}_i = \sum_{t=1}^{T} \tilde{X}_{it}/T$, $i = 1, \ldots, q$.

4. For $j = 1, \ldots, J$, let $\bar{\widetilde{\mathbf{X}}}_j = \sum_{y_t \in S_J} \widetilde{\mathbf{X}}_t/n_j$, where n_j is the number of Y_t's in S_j.

5. Compute

$$\widehat{\mathbf{M}} = \sum_{j=1}^{J} \frac{n_j}{T} (\bar{\widetilde{\mathbf{X}}}_j - \bar{\widetilde{\mathbf{X}}})(\bar{\widetilde{\mathbf{X}}}_j - \bar{\widetilde{\mathbf{X}}})'$$

6. Compute the SVD of $\widehat{\mathbf{M}} = \widehat{\mathbf{U}}\widehat{\Lambda}\widehat{\mathbf{U}}^T$, where $\widehat{\Lambda} = diag(\hat{l}_1, \ldots, \hat{l}_q)$, $\hat{l}_1 > \hat{l}_2 > \ldots > \hat{l}_q$ are the eigenvalues of $\widehat{\mathbf{M}}$ and $\widehat{\mathbf{U}} = (\widehat{\mathbf{u}}_1, \ldots, \widehat{\mathbf{u}}_q)$ is the $q \times q$ orthonormal matrix of its eigenvectors that correspond to $\hat{l}_1, \hat{l}_2, \ldots, \hat{l}_q$.

7. Estimate the dimension d of the regression as $\hat{d}$ using any dimension estimation method that applies.

8. The SIR predictors are $\text{SIR}_1 = \widetilde{\Sigma}^{-1}\widehat{\mathbf{u}}_1\mathbf{X}, \ldots, \text{SIR}_{\hat{d}} = \widetilde{\Sigma}^{-1}\widehat{\mathbf{u}}_d\mathbf{X}$, where $\widetilde{\Sigma} = \sum_{t=1}^{T}(\widetilde{\mathbf{X}}_t - \bar{\widetilde{\mathbf{X}}})(\widetilde{\mathbf{X}}_t - \bar{\widetilde{\mathbf{X}}})'/T$ is the sample covariance matrix of the adjusted predictors $\tilde{\mathbf{X}}$.

Appendix C: Covariance-Stationary Time Series Properties

A sequence of random variables x_{jt} is covariance stationary or weakly stationary if and only if

$$\exists \mu_j \in \mathbb{R} \colon \mathrm{E}(x_{jt}) = \mu_j, \forall t > 0$$

and

$$\forall t' \geq 0, \exists \gamma_{jt'} \in \mathbb{R} \colon \mathrm{cov}(x_{jt}, x_{j,t-t'}) = \mathrm{E}[(x_{jt} - \mu_j)(x_{j,t-t'} - \mu_j)] = \gamma_{j,t-t'} = \gamma_j(t - t') = \gamma_j(h), \forall t > t'$$

In other words, all the terms of the sequence have mean μ, and the hth lag autocovariance, $\mathrm{cov}(x_{jt}, x_{j,t-t'})$, depends only on t' and not on t, so that x_{jt} has time invariant first and second moments. Thus, if x_{jt} is a weakly stationary time series, then the vector $\mathbf{x}_t = (x_{1t}, x_{2t}, ..., x_{pt})$ and the time-shifted vector $\mathbf{x}_{t+h} = (x_{1,t+h}, x_{2,t+h}, ..., x_{p,t+h})$ have the same mean vectors and covariance matrices for every integer h and positive integer t. A strictly stationary sequence is one in which the joint distributions of these two vectors are the same. Weak stationarity does not imply strict stationarity but a strictly stationary time series with $\mathrm{E}(x_{jt}^2) < \infty \; \forall t$ is also weakly stationary. A useful result is that any function of a weakly (strictly) stationary time series is also a weakly (strictly) stationary time series. A stationary time series x_{jt} is ergodic if sample moments converge in probability to population moments.

A multivariate time series $\mathbf{x}_t = (x_{1t}, x_{2t}, ..., x_{pt})$ is covariance stationary and ergodic if all of its component time series are stationary and ergodic. The mean of $\mathbf{x}_t$ is defined as the $(T \times 1)$ vector $\mathrm{E}(\mathbf{x}_t) = \boldsymbol{\mu} = (\mathrm{E}(x_{1t}), \mathrm{E}(x_{2t}), ..., \mathrm{E}(x_{pt}))' = \left(\mu_1, \mu_2, ..., \mu_p\right)'$ and the variance/covariance matrix

$$\begin{aligned}
\boldsymbol{\Sigma}(0) &= \mathrm{var}(\mathbf{x}_t) = \left((\mathbf{x}_t - \boldsymbol{\mu})(\mathbf{x}_t - \boldsymbol{\mu})'\right) = \mathrm{E}\left(\mathbf{x}_t\mathbf{x}_t' - \boldsymbol{\mu}\boldsymbol{\mu}'\right) = \\
&= \begin{bmatrix} \mathrm{var}(x_{1t}) & \mathrm{cov}(x_{1t}, x_{2t}) & \cdots & \mathrm{cov}(x_{1t}, x_{pt}) \\ \mathrm{cov}(x_{2t}, x_{1t}) & \mathrm{var}(x_{2t}) & \cdots & \mathrm{cov}(x_{2t}, x_{pt}) \\ \vdots & \vdots & \ddots & \vdots \\ \mathrm{cov}(x_{pt}, x_{1t}) & \cdots & \cdots & \mathrm{var}(x_{pt}) \end{bmatrix} \\
\boldsymbol{\Sigma}(h) &= \mathrm{cov}(\mathbf{x}_{t+h}, \mathbf{x}_t) = \mathrm{E}\left((\mathbf{x}_{t+h} - \boldsymbol{\mu})(\mathbf{x}_{t+h} - \boldsymbol{\mu})'\right) = \mathrm{E}\left(\mathbf{x}_{t+h}\mathbf{x}_t' - \boldsymbol{\mu}\boldsymbol{\mu}'\right)
\end{aligned}$$

If x_{jt} is a stationary time series with mean μ_j and autocovariance function $\gamma_j(h)$, $\bar{X}_j = \sum_{t=1}^{T} X_{jt}/T$ converges in mean square to μ_j if $\gamma(T) \to 0$ as $T \to \infty$ (see Prop. 2.4.1, p. 58 in Brockwell and Davis (2002), prop. 10.5, p. 279 in Hamilton (1994)). The consistency of the estimator $\bar{\mathbf{X}}$ is established by applying the proposition to each of the component time series x_{jt},

$j = 1, \ldots, p$ (Prop. 7.3.1, p. 234, Brockwell and Davis (2002)). A sufficient condition to ensure ergodicity (consistency) for second moments is $\sum_{h=-\infty}^{\infty} |\gamma_{jj}(h)| < \infty$ (Prop. 7.3.1, p. 234, Brockwell and Davis (2002)).

The parameters $\boldsymbol{\mu}$, $\boldsymbol{\Sigma}(0)$, and $\boldsymbol{\Sigma}(h)$ are estimated from $\mathbf{X}_1, \mathbf{X}_2, ..., \mathbf{X}_T$ using the sample moments:

$$\begin{aligned}
\bar{\mathbf{X}} &= \frac{1}{T}\sum_{t=1}^{T}\mathbf{X} \\
\hat{\boldsymbol{\Sigma}}(0) &= \frac{1}{T}\sum_{t=1}^{T}\left(\mathbf{X}_t - \bar{\mathbf{X}}\right)\left(\mathbf{X}_t - \bar{\mathbf{X}}\right)' \\
\hat{\boldsymbol{\Sigma}}(h) &= \begin{cases} \frac{1}{T}\sum_{t=1}^{T-h}\left(\mathbf{X}_{t+h} - \bar{\mathbf{X}}\right)\left(\mathbf{X}_t - \bar{\mathbf{X}}\right)' & \text{if } 0 \le h \le T-1 \\ \hat{\boldsymbol{\Gamma}}(h)' & \text{if } -T+1 \le h < 0 \end{cases}
\end{aligned}$$

The ergodic theorem obtains that if $\mathbf{x}_t$ is a strictly stationary and ergodic time series then as $T \to \infty$

$$\bar{\mathbf{X}} \xrightarrow{p} \boldsymbol{\mu} \tag{8.1}$$

$$\hat{\boldsymbol{\Sigma}}(0) \xrightarrow{p} \boldsymbol{\Sigma}(0) \tag{8.2}$$

$$\hat{\boldsymbol{\Sigma}}(h) \xrightarrow{p} \boldsymbol{\Sigma}(h) \tag{8.3}$$

Under more restrictive assumptions on the process $\mathbf{x}_t$ it can also be shown that $\bar{\mathbf{X}}_T$ is approximately normally distributed for large T. Determination of the covariance matrix of this distribution is quite complicated. For example, the following is a CLT for a covariance stationary m-dependent vector process (Villegas (1976), Thm. 5.1). A stochastic vector process $\mathbf{x}_1, \mathbf{x}_2, \ldots$ is m-dependent if the two sets of random vectors $\mathbf{x}_1, \ldots, \mathbf{x}_r$ and $\mathbf{x}_s, \ldots, \mathbf{x}_n$ are independent whenever $s - r > m$.

Theorem 4 *If $\mathbf{x}_1, \mathbf{x}_2, \ldots$ is a stationary m-dependent second-order vector process, then:*

(i) the distribution of $\sqrt{T}(\bar{\mathbf{X}}_T - \boldsymbol{\mu})$ converges to a (possibly degenerate) normal distribution with zero mean vector and covariance matrix

$$\mathbf{V} = \sum_{h=-m}^{m} \boldsymbol{\Sigma}(h)$$

where $\boldsymbol{\Sigma}(h)$ is the covariance matrix of $\mathbf{x}_t$ and $\mathbf{x}_{t+h}$;

(ii) the covariance matrix of $\sum_{t=1}^{T}(\mathbf{X}_1 + \mathbf{X}_2 + \ldots + \mathbf{X}_T)/\sqrt{T}$ converges to $\mathbf{V}$ when T increases indefinitely.

REFERENCES

[1] Adragni, K.P. and Cook R.D. (2009). "Sufficient dimension reduction and prediction in regression". *Phil. Trans. R. Soc. A*, **367**, 4385-4405.

[2] Bai, J. and Ng, S. (2002). "Determining the Number of Factors in Approximate Factor Models". *Econometrica*. Vol. 70, No. 1 (Jan., 2002) , pp. 191-221.

[3] Bai, J. and Ng, S. (2007). "Determining the Number of Primitive Shocks in Factor Models", *Journal of Business and Economic Statistics*, 2007, 25:1, p.52-60.

[4] Bai, J. and Ng, S. (2008a). "Forecasting Economic Time Series Using Targeted Predictors", *Journal of Econometrics* 146, 304-317.

[5] Bai, J. and Ng, S. (2008b). "Large Dimensional Factor Analysis", *Foundations and Trends in Econometrics*, 2008, 3:2, 89-163.

[6] Banbura M. and Modugno, M. (2014). "Maximum Likelihood Estimation of Factor Models on Datasets with Arbitrary Patterns of Missing Data". *J. Appl. Econ.* 29: 133–160 (2014).

[7] Barbarino, A. and Bura, E. (2015). "A Unifying Framework for Big Data". Forthcoming Working Paper 2015.

[8] Barnichon R. (2010). "Building a composite Help-Wanted Index". Economics Letters, Dec 2010.

[9] Bernard-Michel, C., Gardes, L. and Girad, S. (2009). Gaussian Regularized Sliced Inverse Regression, *Statistics and Computing*, 19(1), 85-98.

[10] Boivin, J. and Ng, S. (2006). "Are more data always better for factor analysis?". *Journal of Econometrics* 132, p. 169-194.

[11] Brockwell, P. J. and Davis, R.A. (2002). *Introduction to Time Series and Forecasting*, 2nd edition, Springer-Verlag, New York.

[12] Bura, E. and Cook, R.D. (2001a), "Estimating the structural dimension of regressions via parametric inverse regression," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 63, 393–410.

[13] Bura, E. and Cook, R.D. (2001b), "Extending SIR: The Weighted Chi-Square Test," *Journal of the American Statistical Association*, 96, 996-1003.

[14] Bura, E. and Forzani, L. (2015), "Sufficient reductions in regressions with elliptically contoured inverse predictors," *Journal of the American Statistical Association*, 110, 420-434.

[15] Bura, E. and Yang, J. (2011), "Dimension Estimation in Sufficient Dimension Reduction: A Unifying Approach," *Journal of Multivariate Analysis*, 102, 130-142.

[16] Chamberlain, G. and Rothchild, M. (1983). "Arbitrage, Factor Structure, and Mean-Variance Analysis on Large Asset Markets". *Econometrica*, 51(5), 1281-1304.

[17] Cook R.D. (1998a) *Regression Graphics: Ideas for studying regressions through graphics*. New York: Wiley.

[18] Cook, R. D. (2000). "SAVE: A method for dimension reduction and graphics in regression". *Communications in Statistics: Theory Methods*, **29**, 2109-2121. (Invited paper for a special millennium issue on regression.)

[19] Cook R.D. (2007). "Fisher lecture: Dimension reduction in regression". *Statistical Science*, **22**, 1-26.

[20] Cook, R.D., and Forzani, L. (2008), "Principal Fitted Components for Dimension Reduction in Regression", *Statistical Science*, 23, 485-501.

[21] Cook, R. D., Bing, L. and Francesca Chiaromonte. "Dimension Reduction in Regression without Matrix Inversion". *Biometrika* (2007), 94, 3, pp. 569–584.

[22] Chiaromonte, F. and Martinelli, J. (2002). Dimension reduction strategies for analyzing global gene expression data with a response. *Mathematical Biosciences*, 176, 123–144.

[23] Cook, R.D., and Weisberg, S. (1991), "Discussion of *Sliced inverse regression for dimension reduction", Journal of the American Statistical Association*, 86, 328-332.

[24] Cook, R. D. and Yin, X. (2001). Dimension reduction and visualization in discriminant analysis (with discussion), *Australian & New Zealand Journal of Statistics*, 43(2), 147-199.

[25] De Mol, C., Giannone, D. and Reichlin, L. (2008). "Forecasting using a large number of predictors: Is Bayesian shrinkage a valid alternative to principal components?". *Journal of Econometrics*, 146(2), 318-328. ISSN 0304-4076, http://dx.doi.org/10.1016/j.jeconom.2008.08.011.

[26] Diaconis, P. and Freedman, D. (1984). "Asymptotics of graphical projection pursuit". *Ann. Statist.* 12 793-815.

[27] Doz, C., Giannone, D. and Reichlin, L. (2011). "A two-step estimator for large approximate dynamic factor models based on Kalman filtering". *Journal of Econometrics*, 164(1), 188-205, ISSN 0304-4076, http://dx.doi.org/10.1016/j.jeconom.2011.02.012.

[28] Doz, C., Giannone, D. and Reichlin, L. (2012). "A Quasi—Maximum Likelihood Approach for Large, Approximate Dynamic Factor Models". *The Review of Economics and Statistics*. 94(4), 1014-1024.

[29] Eaton, M.L. (1983). *Multivariate Statistics. A Vector Space Approach*. New York: John Wiley & Sons, Inc.

[30] Eaton, M. L. (1986). "A characterization of spherical distributions". *Journal of Multivariate Analysis*, 20, 272-276.

[31] Engle, R. and Watson, M. (1981)."A One-Factor Multivariate Time Series Model of Metropolitan Wage Rates". *Journal of the American Statistical Association*, 76(376), 774-781.

[32] Forni, M., Hallin, M., Lippi, M. and Reichlin, L. (2005). "The Generalized Dynamic Factor Model: One-Sided Estimation and Forecasting". *Journal of the American Statistical Association*. 100(471), 830-840.

[33] Jerome Friedman, Trevor Hastie, Noah Simon, Rob Tibshirani (2015). "Lasso and Elastic-Net Regularized Generalized Linear Models". Available on CRAN as `glmnet`.

[34] Geweke, J. (1977). "The Dynamic Factor Analysis of Economic Time Series Models in Latent variables in socio-economic models", ed. by D. J. Aigner and A. S. Goldberger. North Holland.

[35] Groen, J. and Kapetanios, G. (2014). "Revisiting Useful Approaches to Data-Rich Macroeconomic Forecasting". Federal Reserve Bank of New York Staff Reports No.237, May 2008. Revised 2014.

[36] Hall, P. and Li, K. C. (1993). "On almost linearity of low dimensional projections from high dimensional data." *The Annals of Statistics*, 21, 867–889.

[37] Hamilton, J. D. (1994). *Time Series Analysis*. Princeton University Press, Princeton, New Jersey.

[38] Helland, I. (1988). "On the Structure of Partial Least Squares Regression". *Commun. Statist. Simula.* 17(2), 581-607 (1988).

[39] Hoerl, A. E. and Kennard, R. W. (1970). "Ridge Regression: Applications to Nonorthogonal Problems". *Technometrics*, 12(1), 69-82.

[40] Hotelling, H. (1933). "Analysis of a complex of statistical variables into principal components". *J. Educ. Psychol.*, 24, 417–441, 498–520.

[41] Ipsen, I. C. F. and Meyer, C. D. (1998). The Idea behind Krylov Methods. *The American Mathematical Monthly*, **105**(10), 889-899.

[42] Jurado, K., Ludvigson, S. and Serena Ng (2015). "Measuring Uncertainty". American Economic Review, 105(3): 1177-1216.

[43] Bryan T. Kelly and Seth Pruitt (2014). "The three-pass regression filter: A new approach to forecasting using many predictors". 2014/5 Fama-Miller Working Paper.

[44] Leeb, H. (2013)."On the conditional distributions of low-dimensional projections from high-dimensional data". *The Annals of Statistics*, 41, 464-483.

[45] Li, K. C. (1991). "Sliced inverse regression for dimension reduction (with discussion)". *Journal of the American Statistical Association*, **86**, 316-342.

[46] Li, L. and Li, H. (2004). "Dimension reduction methods for microarrays with application to censored survival data". *Bioinformatics*, 20(18), 3406–3412.

[47] Lumley, T. (2009), "Regression Subset Selection". Available on CRAN as `leaps`.

[48] Magnus, J. R. (1988). *Linear Structures*. Oxford University Press, New York .

[49] McCracken, M.W. and Ng, S. (2015). "FRED-MD: A Monthly Database for Macroeconomic Research". Working Paper 2015-012A, June 2015.

[50] Bjørn-Helge Mevik, Ron Wehrens and Kristian Hovde Liland (2013). "Partial Least Squares and Principal Component regression". Available on CRAN as `pls`. See also "The PLS Package: Principal Components and Partial Least Squares Regression in R" by the same authors in *Journal of Statistical Software* (2007), vol.18 Issue 2.

[51] Thomas Sargent and Christopher Sims (1977). "Business Cycle Modeling Without Pretending to Have too Much A Priori Economic Theory" in New Methods in business cycle research: Proceedings from a conference, ed. by Christopher Sims. Federal Reserve Bank of Minneapolis.

[52] Steinberger, L. and Leeb, H. (2015). "On conditional moments of high-dimensional random vectors given lower-dimensional projections". WP 2015.

[53] Stock, J. H. and Watson, M. (1998). "Diffusion Indexes". NBER Working Paper Series #6702, August 1998.

[54] Stock, J. H. and Watson, M. (2002). "Forecasting Using Principal Components from a Large Number of Predictors". *Journal of the American Statistical Association*, 2002.

[55] Stock, J. H. and Watson, M. (2002). "Macroeconomic Forecasting Using Diffusion Indexes". *Journal of Business and Economic Statistics*, April 2002, Vol. 20 No. 2, 147-162.

[56] Stock, J. H. and Watson, M. (2005). "Implications of Dynamic Factor Models for VAR Analysis". Working Paper, Revised June 2005

[57] Stock, J. H. and Watson, M. (2006). "Macroeconomic Forecasting Using Many Predictors". *Handbook of Economic Forecasting*, Graham Elliott, Clive Granger, Allan Timmerman (eds.), North Holland, 2006.

[58] Stock, J. H. and Watson, M. (2008). "Forecasting in Dynamic Factor Models Subject to Structural Instability" in *The Methodology and Practice of Econometrics, A Festschrift in Honour of Professor David F. Hendry*, Jennifer Castle and Neil Shephard (eds), 2008, Oxford: Oxford University Press.

[59] Stock, J. H. and Watson, M. (2010). "Dynamic Factor Models". in *Oxford Handbook of Forecasting*, Michael P. Clements and David F. Hendry (eds), 2011, Oxford: Oxford University Press.

[60] Stock, J. H. and Watson, M. (2012). "Generalized Shrinkage Methods for Forecasting Using Many Predictors". *Journal of Business and Economic Statistics*, 30:4 (2012), 481-493.

[61] Strang, G. (1978). *Linear Algebra and its Applications*, 3rd ed. Philadelphia, PA: Saunders.

[62] Villegas, C. (1976). "On a multivariate central limit theorem for stationary bilinear processes". *Stochastic Processes and their Applications*, 4(2), 121-133.

[63] Wold, S. (1978). "Cross validatory estimation of the number of components in factor and principal component models". *Technometrics* 20, 397-405.