Finance and Economics Discussion Series
Divisions of Research & Statistics and Monetary Affairs
Federal Reserve Board, Washington, D.C.

A Unified Framework for Dimension Reduction in Forecasting

Alessandro Barbarino and Efstathia Bura

2017-004

Please cite this paper as:
Barbarino, Alessandro, and Efstathia Bura (2017). "A Unified Framework for Dimension Reduction in Forecasting," Finance and Economics Discussion Series 2017-004. Washington: Board of Governors of the Federal Reserve System, https://doi.org/10.17016/FEDS.2017.004.

A Unified Framework for Dimension Reduction in Forecasting*

Alessandro Barbarino† and Efstathia Bura‡

January 12, 2017

Abstract

Factor models are widely used in summarizing large datasets with few underlying latent factors and in building time series forecasting models for economic variables. In these models, the reduction of the predictors and the modeling and forecasting of the response y are carried out in two separate and independent phases. We introduce a potentially more attractive alternative, Sufficient Dimension Reduction (SDR), that summarizes $\mathbf{x}$ as it relates to y, so that all the information in the conditional distribution of $y|\mathbf{x}$ is preserved. We study the relationship between SDR and popular estimation methods, such as ordinary least squares (OLS), dynamic factor models (DFM), partial least squares (PLS) and RIDGE regression, and establish the connection and fundamental differences between the DFM and SDR frameworks. We show that SDR significantly reduces the dimension of widely used macroeconomic series data with one or two sufficient reductions delivering similar forecasting performance to that of competing methods in macro-forecasting.

JEL Classification Number: C32, C53, C55, E17

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

*The results and opinions expressed in this paper reflect the views of the authors and should not be attributed to the Federal Reserve Board.

†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 and Institute of Statistics and Mathematical Methods in Economics, Vienna University of Technology. *Address*: 801 22nd St. NW Washington, DC 20052 USA. *Email*: ebura@gwu.edu

1 Introduction

This paper introduces Sufficient Dimension Reduction (SDR) to macro-forecasting and studies the relationship between SDR and popular estimation methods in forecasting, such as ordinary least squares (OLS), dynamic factor models (DFM), partial least squares (PLS) and RIDGE regression. In particular, we establish the connection and fundamental differences between the DFM and SDR frameworks. We also extend some key SDR results to a time series setting and offer a first assessment of the effectiveness of SDR methods in a real-world forecasting application.[1]

The availability of richer datasets, in conjunction with the seminal work of Stock and Watson (1998) [69], has attracted widespread interest in Dynamic Factor Models (DFM) that resulted in a large body of research.[2] The core contributions in DFM have largely focused on the assumptions needed in order to identify and estimate asymptotically a latent factor structure in multivariate data, as outlined in the early work of Stock and Watson (1998) [69] and in subsequent studies by Bai and Ng (2002) [4], Bai (2003) [3], and the more recent contributions by Onatski (2009 and 2010) [63][64], Alessi, Barigozzi and Capasso (2010) [2] and Ahn and Horenstein (2013) [1], among others.[3].

A considerable number of studies, following Stock and Watson (1998, 2002a and 2002b) [69][70][71], have focused on the use of DFMs in forecasting. In attempts to boost the forecasting performance of DFMs, selected tools from the plethora of reduction methods in statistical and machine learning were explored in the econometric literature. Notable early examples are Bai and Ng (2008) [5], DeMol, Giannone and Reichlin (2008) [33] and, more recently, Stock and Watson (2012) [76], Kelly and Pruitt (2015) [55] and Groen and Kapetanios (2014) [44]. The survey by Ng (2013) [62] provides additional references, as well as a discussion on targeting, a central issue in this paper.

[1]The study by Barbarino and Bura (2015) [8] also provides an introduction to SDR techniques in macro-forecasting, however it is more applied in nature and it provides fewer details on the connection between SDR and DFM frameworks.

[2]Surveys on the DFM literature include Stock and Watson (2006 and 2011) [73][75] and Bai and Ng (2008) [6]. In this paper we exclusively discuss moment-based estimation methods (non-parametric, in some parlance) sidestepping the treatment of model-based estimation (maximum likelihood) and the use of the Kalman filter. In part as a consequence of this choice the empirical application will abstract from issues specific to nowcasting such as data with mixed frequencies and "ragged edges" recently surveyed by Bambura, Giannone and Reichlin (2011) [7].

[3]In this paper we concentrate on static factors in order to keep matters simple and draw an uncluttered comparison between SDR and DFMs and our references reflect this choice. Fundamental contributions on the factor structure by Forni, Hallin, Lippi and Reichlin (1998, 2000, 2001, 2004, 2005) and further developments are neatly organized and referenced in the surveys by Stock and Watson (2006 and 2011) [73][75]

Dimension reduction methods in regression fall into two categories: *variable* or *model selection*, where a subset of the original predictors is selected for modeling the response, and *feature extraction*, where linear combinations of the regressors, frequently referred to as "derived," replace the original regressors. The underlying assumption in variable selection is that the individual predictors have independent effects on the response, which is typically violated in econometric time series that have varying degrees of correlation. Thus, the focus in the econometrics literature, including this paper, is on derived predictor methods.

Studies that exploit a factor structure in forecasting follow a recurrent theme: 1) First the dimension of a large panel of data is reduced to a sufficiently "lean" factor structure and 2) then the factors are used to forecast a target variable y. The reduction step has so far been largely disconnected from the targeting step, likely a legacy of the origin of factor models. 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.[4]

In contrast, Sufficient Dimension Reduction (SDR) is a 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.[5] SDR focuses on finding sufficient, in the statistical sense, reductions of a large set of explanatory variables in order to model a target response y. The reduction and targeting are carried out simultaneously as SDR identifies a sufficient function of the regressors, $\mathbf{R}(\mathbf{x})$, that preserves the information in the conditional distribution of $y|\mathbf{x}$.

SDR methods do not resort to a latent factor structure. It is assumed that the data generating process (DGP) directly generates the set of regressors $\mathbf{x}$ without the mediation of latent factors, and conditions that ensure identification and estimation are placed directly on the marginal distribution of $\mathbf{x}$. Finally, specifying the link between the target y and the panel of regressors $\mathbf{x}$ is not required in SDR, further differentiating it from OLS, DFM, RIDGE or PLS, which all conjecture that y depends linearly on common latent factors.

[4]Stock and Watson (2006) [73] devote a paragraph on the two possible ways of accommodating a target variable y on p. 526, however all assume that y is generated by the same set of unobserved factors as the other variables in the panel.

[5]Li (1991) [57] introduced the concept of inverse regression as a dimension reduction tool in Sliced Inverse Regression (SIR) and SIR will be the SDR method of choice in this paper. Cook and his collaborators formalized the field in several papers (e.g. Cook and Weisberg (1991) [31]; Cook (1994), (1998b), (2007) [22][24][26]; Cook and Lee (1999) [29]; Bura and Cook (2001a and 2001b) [13][14]; Cook and Yin (2001) [32]; Chiaromonte, Cook and Li (2002) [20]; Cook and Ni (2005) [30]; Cook and Forzani (2008 and 2009) [27][28]) and a book (Cook 1998a [23]), where much of the SDR terminology was introduced.

SDR was developed for cross-sectional applications and the associated inference results, such as consistency of the estimates of the sufficient reductions and tests for dimension, were derived under the assumption of random samples. We show that our sufficient reduction estimators are consistent for covariance stationary predictor time series. Under additional assumptions, we can also obtain their asymptotic normality and tests of dimension.

In Section 2 we discuss the challenges of macro-forecasting in a data rich environment and describe the DFM framework approach. We also propose an alternative forecasting framework based on SDR methods and contrast it with the DFM framework, providing a connection between the two. Section 3 is a general exposition 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 5. Section 3 also contains the conceptual motivation for targeted reductions.

In Section 4 we review other linear dimension reduction methods that have been proposed within the DFM framework and are used in the empirical application. In order to draw analogies and highlight differences with results in the macro-forecasting literature, a real-world forecasting experiment with a large panel of macro variables, as in Stock and Watson (2002a and 2002b) [70][71], is conducted, although our data source is the novel repository FRED-MD maintained by the St. Louis Fed and documented by McCracken and Ng (2015) [60]. Section 5 contains the description and results of a horserace between the estimators we consider in the paper focusing on forecasting accuracy in predicting various targets in an out-of-sample forecasting experiment. We find that SIR achieves similar forecasting performance as the other methods with the important difference that in most cases it does so with just one or two linear combinations of the predictors. Methods that are constrained to use just one linear combination, such as OLS and RIDGE do not perform well in general, and PCR and PLS usually need more than two components to achieve their minimum mean square forecast error. We conclude in Section 6.

2 A Unified Framework for Forecasting with a Large Set of Explanatory Variables

A large set of p explanatory variables $\mathbf{x}_t$ is available to forecast a single variable y_t using a sample of size T. In statistical and machine learning this problem is approached by considering all regressors

as potentially useful in modeling y_t via the model

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

where $\mathbf{w}_t$ may contain additional regressors such as lags of y_t, in a time-series context. If all available information up to time t and time-series dynamics are well captured by (2.1), it is natural to assume that $\mathrm{E}(\varepsilon_{t+h}) = 0$, and that both $\boldsymbol{\alpha}'\mathbf{x}_t$ and $\boldsymbol{\gamma}'\mathbf{w}_t$ are uncorrelated with ε_{t+h}. To simplify notation, we incorporate all predictors into $\mathbf{x}_t$ and drop the term $\boldsymbol{\gamma}'\mathbf{w}_t$ to obtain the forecast model

$$y_{t+h} = \boldsymbol{\alpha}'\mathbf{x}_t + \varepsilon_{t+h} \tag{2.2}$$

The optimal forecast from (2.2)

$$y_{t+h|t} = \boldsymbol{\alpha}'\mathbf{x}_t \tag{2.3}$$

is unfeasible since $\boldsymbol{\alpha}$ is unknown. However, the assumptions underlying model (2.2) are conducive to ordinary least squares (OLS) as the estimator of choice for the parameters $\boldsymbol{\alpha}$, leading to the feasible forecast

$$y_{t+h|t} = \widehat{\boldsymbol{\alpha}}'_{OLS}\mathbf{x}_t \tag{2.4}$$

In this set-up all regressors are potentially useful in forecasting. Yet, estimation in (2.2) via OLS can be problematic when p is large relative to T, or variables in $\mathbf{x}_t$ are nearly collinear, as is often the case in the macro forecasting literature [see, e.g., Stock and Watson (2006) [73]].[6] The variance of the prediction (2.4) is of order p/T. Thus, when p is large, estimation methods that trade bias with variance may dominate OLS under the mean squared error (MSE) criterion, even under assumptions that guarantee that the OLS estimator is unbiased. In particular, when $p \geq 3$, the OLS estimator is not admissible under the MSE criterion.[7]

In their survey, Stock and Watson (2006) [73] start their discussion with model (2.2). However, after pointing out difficulties with such model and OLS and in line with their previous work [see

[6] 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.

[7] The MSE of any estimator $\hat{\theta}$ of a parameter θ is given by $MSE(\hat{\theta}) = \mathrm{Bias}^2(\hat{\theta}) + \mathrm{var}(\hat{\theta})$. Although among unbiased estimators the OLS estimator $\widehat{\theta}_{OLS}$ has minimum mean squared error, other estimators $\widehat{\theta}$, some also linear, may have uniformly smaller MSE by trading off bias for variance, so that

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

for all θ with strict inequality for some θ. James and Stein (1961) showed that, for $p \geq 3$, the OLS estimator is not admissible under the MSE; a striking result that inaugurated research in shrinkage estimation.

Stock and Watson (1998 and 2002a) [69][70]], they resolve to use their factorial structure apparatus assuming *from the outset* that the forecast model is

$$y_{t+h} = \boldsymbol{\gamma}_y' \mathbf{f}_t + \varepsilon_{t+h} \tag{2.5}$$

with corresponding unfeasible forecast,

$$y_{t+h|t} = \boldsymbol{\gamma}_y' \mathbf{f}_t \tag{2.6}$$

The added difficulty in this setup is that not only the parameters $\boldsymbol{\gamma}_y$, but also $\mathbf{f}_t$ are unknown and need to be estimated. Stock and Watson showed that their factorial structure together with additional assumptions ensure the factors are identifiable and estimable by *principal components*, say $\widehat{\mathbf{f}}_t$, which in turn can be plugged in (2.5)

$$y_{t+h|t} = \boldsymbol{\gamma}_y' \widehat{\mathbf{f}}_t + \varepsilon_{t+h} \tag{2.7}$$

They also obtained that the OLS estimation of (2.7) closely approximates asymptotically the unfeasible forecast (2.6).

The DFM approach to the forecasting equation raises two issues: How applicable the *reduced* population model (2.5) relative to the population model (2.2) is for forecasting, and whether the dimension of the exogenous variables $\mathbf{x}$ in the population model (2.2) can be reduced so that one can replace it with population model (2.5) and still obtain accurate prediction.

Regarding the first question, Groen and Kapetanios (2014) [44] showed that, *under a factor structure*, the reduced forecasting population model (2.5) may be too restrictive a shortcut, and concluded that the population model (2.2) entails fewer misspecification errors. We go one step further in reducing misspecification risk and in subsequent sections we show how our SDR approach allows to further relax (2.2) to

$$y_{t+h} = g(\boldsymbol{\alpha}' \mathbf{x}_t, \varepsilon_{t+h}) = g(\boldsymbol{\alpha}_1' \mathbf{x}_t, \ldots, \boldsymbol{\alpha}_d' \mathbf{x}_t, \varepsilon_{t+h}) \tag{2.8}$$

where $\boldsymbol{\alpha}$ denotes a $p \times d$ matrix of rank $d < p$. SDR allows the target to be a non linear function $g(\cdot)$ of the predictors $\mathbf{x}_t$. Non-linear g's allow for more than one linear combinations or projections of the regressors to model $\mathbf{y}_t$ in order to preserve all the information that the covariates $\mathbf{x}_t$ carry about y_{t+h}.[8]

[8] By contrast, when $g(\cdot)$ comprises of multiple linear combinations, as for instance in PCR, they can all be combined in one.

We focus on the second question in Proposition 1, which states the conditions under which it is possible to reduce the dimension of $\mathbf{x}$ *without the need of assuming an underlying factor structure.* These conditions have not been considered in the DFM forecasting literature or by Groen and Kapetanios (2014) [44] . Moreover, they pivot the attention to the probability distributions that satisfy them. Proposition 1 assumes that population model (2.2) is correct, replaces the assumption of an underlying factor structure with a set of "leaner" conditions and concludes that reduced model (2.5) is a good approximation to population model (2.2). The absence of an underlying factor structure and companion identification assumptions implies that model (2.5) is not necessarily tied to PCR and offers a framework that subsumes PCR, PLS and RIDGE, and other estimators provided that they are linear projections of the predictors.

Proposition 1 *Suppose* $\mathbf{x}$ *is a random p-vector with finite first two moments. Let* $y = \boldsymbol{\alpha}'\mathbf{x}$ *where* $\boldsymbol{\alpha} \in \mathbb{R}^p$ *is unknown, and* $\mathbf{f} = \boldsymbol{\beta}'\mathbf{x}$*, where the* $p \times r$ *matrix* $\boldsymbol{\beta}$ *is such that*

(i) for each $\boldsymbol{\alpha}$, $\mathrm{E}(\boldsymbol{\alpha}'\mathbf{x}|\boldsymbol{\beta}'\mathbf{x} = \mathbf{f})$ *is linear in* $\mathbf{f} \in \mathbb{R}^r$ *(LC);*

(ii) for each $\boldsymbol{\alpha}$, $\mathrm{Var}(\boldsymbol{\alpha}'\mathbf{x}|\boldsymbol{\beta}'\mathbf{x} = \mathbf{f})$ *is constant in* $\mathbf{f} \in \mathbb{R}^r$ *(CVC).*

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

$$y = \mu_y + \mathbf{c}'(\mathbf{f} - \mathrm{E}(\mathbf{f})) + \epsilon \tag{2.9}$$

where $\mathbf{c} = (\boldsymbol{\beta}'\boldsymbol{\Sigma}_{\mathbf{x}}\boldsymbol{\beta})^{-}\boldsymbol{\beta}'\boldsymbol{\Sigma}_x\boldsymbol{\alpha} \in \mathbb{R}^r$, $\mu_y = \mathrm{E}(y)$, $\mathrm{E}(\epsilon|\mathbf{f}) = 0$ *and* $\mathrm{Var}(\epsilon|\mathbf{f})$ *is constant.* [9]

Proof. Let $\mathrm{E}(\mathbf{x}) = \boldsymbol{\mu}_x$, $\mathrm{Var}(\mathbf{x}) = \boldsymbol{\Sigma}_x$, and for simplicity, without loss of generality assume $\mathrm{E}(\mathbf{f}) = 0$. Condition (i) implies that $\mathrm{E}(y|\mathbf{f}) = \mathrm{E}(\boldsymbol{\alpha}'\mathbf{x}|\boldsymbol{\beta}'\mathbf{x}) = \mu_y + \mathbf{c}'(\boldsymbol{\beta}'\mathbf{x})$ for some constants μ_y and $\mathbf{c} \in \mathbb{R}^r$. Therefore, $\mathrm{E}(y) = \mu_y$ and $\mathbf{c} = \mathrm{Var}(\boldsymbol{\beta}'\mathbf{x})^{-}\,\mathrm{Cov}(\boldsymbol{\beta}'\mathbf{x}, \boldsymbol{\alpha}'\mathbf{x}) = (\boldsymbol{\beta}'\boldsymbol{\Sigma}_x\boldsymbol{\beta})^{-}\boldsymbol{\beta}'\boldsymbol{\Sigma}_x\boldsymbol{\alpha}$, the population OLS slope (see Goldberger 1991, p. 54 [43]). Next, $\mathrm{Var}(y) = \mathrm{Var}(\mathrm{E}(y|\mathbf{f})) + \mathrm{E}(\mathrm{Var}(y|\mathbf{f})) = \mathrm{Var}\left(\mu_y + \mathbf{c}'(\boldsymbol{\beta}'\mathbf{x})\right) + \sigma^2 = \mathbf{c}'\boldsymbol{\beta}'\boldsymbol{\Sigma}_x\boldsymbol{\beta}\mathbf{c} + \sigma^2$. That is, $\sigma^2 = \mathrm{var}(y|\mathbf{f}) = \mathrm{var}(y) - \mathbf{c}'\boldsymbol{\beta}'\boldsymbol{\Sigma}_x\boldsymbol{\beta}\mathbf{c}$, which is constant by condition (ii). Combining the above obtains (2.9) with $\epsilon = y - \mathrm{E}(y|\mathbf{f})$. ■

We draw attention to the similarity between (2.2) and the requirement in Proposition 1 that $y = \boldsymbol{\alpha}'\mathbf{x}$, which can be relaxed to $y = \boldsymbol{\alpha}'\mathbf{x} + \varepsilon$, with $\mathrm{E}(\varepsilon) = 0$ and $\mathrm{cov}(\varepsilon, \mathbf{x}) = 0$. Proposition 1 ascertains that one can replace $\mathbf{x}$ by a lower dimensional projection $\mathbf{f}$ in the forecasting model if the distribution of the predictors $\mathbf{x}$ satisfies the *linearity condition* (i) (henceforth LC) and *constant*

[9] $\mathbf{A}^{-}$ denotes a generalized inverse of a matrix $\mathbf{A}$.

variance condition (ii) (henceforth CVC), and then use the reduced model (2.9) to forecast y using OLS.

An important aspect of Proposition 1 is that any $\boldsymbol{\beta}$ that satisfies conditions (LC) and (CVC) produces a good approximation to (2.2). In this regard, Proposition 1 suggests a common umbrella under which one can organize and understand population objects in the form of feature extraction and the estimators that are their sample counterparts.

Specifically, focusing on the conditional expectation in (2.9),

$$\begin{aligned} \mathrm{E}(y-\mu_y|\boldsymbol{\beta}'\mathbf{x}) &= \boldsymbol{\alpha}'\boldsymbol{\Sigma}_x\boldsymbol{\beta}(\boldsymbol{\beta}'\boldsymbol{\Sigma}_x\boldsymbol{\beta})^{-}\boldsymbol{\beta}'(\mathbf{x}-\mathrm{E}(\mathbf{x})) \\ &= \boldsymbol{\alpha}'\mathbf{P}'_{\boldsymbol{\beta}(\boldsymbol{\Sigma}_x)}(\mathbf{x}-\mathrm{E}(\mathbf{x})) \end{aligned} \tag{2.10}$$

where $\mathbf{P}_{\boldsymbol{\beta}(\boldsymbol{\Sigma}_x)} = \boldsymbol{\beta}(\boldsymbol{\beta}'\boldsymbol{\Sigma}_x\boldsymbol{\beta})^{-}\boldsymbol{\beta}'\boldsymbol{\Sigma}_x$ is the projection operator onto $\mathcal{R}(\boldsymbol{\beta})$ relative to the inner product $(\mathbf{a},\mathbf{b}) = \mathbf{a}'\boldsymbol{\Sigma}_x\mathbf{b}$. If $\boldsymbol{\beta}$ is of full rank r, then $\mathbf{P}_{\boldsymbol{\beta}(\boldsymbol{\Sigma}_x)} = \boldsymbol{\beta}(\boldsymbol{\beta}'\boldsymbol{\Sigma}_x\boldsymbol{\beta})^{-1}\boldsymbol{\beta}'\boldsymbol{\Sigma}_x$ and $\mathbf{c} = (\boldsymbol{\beta}'\boldsymbol{\Sigma}_{\mathbf{x}}\boldsymbol{\beta})^{-1}\boldsymbol{\beta}'\boldsymbol{\Sigma}_x\boldsymbol{\alpha}$. Importantly, the calculation of $\mathbf{c}$ and $\mathbf{P}_{\boldsymbol{\beta}(\boldsymbol{\Sigma})}$ does not require $\boldsymbol{\Sigma}_x = \mathrm{var}(\mathbf{x})$ be invertible.[10]

For a given $\mathbf{x}$ value, how close the predicted y value, $\mathrm{E}(y|\mathbf{x})$ will be to the truth is reflected by the length or norm of $\mathbf{I}-\mathbf{P}_{\boldsymbol{\beta}(\boldsymbol{\Sigma}_x)}$ in (2.10), which is controlled solely by $\boldsymbol{\beta}$. In consequence, ordering estimators $\boldsymbol{\beta}'\mathbf{x}$ with respect to their forecasting accuracy is tantamount to identifying $\boldsymbol{\beta}$'s satisfying (LC) and (CVC) with smaller norm. As a result, principal components need not be the best aggregation method.

If only the linearity condition (LC) holds; that is, the first conditional moment of y given $\mathbf{f}$ is linear in $\mathbf{f}$ but its conditional variance is no longer constant, the following corollary states that the forward model can be heteroskedastic.

Corollary 1 *Suppose* $\mathbf{x}$ *is a random p-vector with finite first moment. Let* $y = \boldsymbol{\alpha}'\mathbf{x}$ *where* $\boldsymbol{\alpha} \in \mathbb{R}^p$ *is unknown, and* $\mathbf{f} = \boldsymbol{\beta}'\mathbf{x}$*, where the* $p \times r$ *matrix* $\boldsymbol{\beta}$ *is such that*

(i) for each $\boldsymbol{\alpha}$, $\mathrm{E}(\boldsymbol{\alpha}'\mathbf{x}|\boldsymbol{\beta}'\mathbf{x}=\mathbf{f})$ *is linear in* $\mathbf{f} \in \mathbb{R}^r$.

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

$$y = \mu_y + \mathbf{c}'(\mathbf{f}-\mathrm{E}(\mathbf{f})) + \epsilon$$

with $\mathbf{c} = (\boldsymbol{\beta}'\boldsymbol{\Sigma}_{\mathbf{x}}\boldsymbol{\beta})^{-}\boldsymbol{\beta}'\boldsymbol{\Sigma}_x\boldsymbol{\alpha} \in \mathbb{R}^r$, $\mu_y = \mathrm{E}(y)$, $\mathrm{E}(\epsilon|\mathbf{f}) = 0$ *and* $\mathrm{Var}(\epsilon|\mathbf{f}) = \sigma^2(\mathbf{f})$. □

[10] Therefore, this set-up allows the possibility that OLS might not be well defined.

The proof of Corollary 1 follows the proof of Proposition 1 by writing $y = \mathrm{E}(y|\mathbf{f}) + y - \mathrm{E}(y|\mathbf{f}) = E(y|\mathbf{f}) + \epsilon$, with the difference that $\mathrm{var}(\epsilon) = \mathrm{var}(y|\mathbf{f}) = \sigma^2(\mathbf{f})$ may depend on $\mathbf{f} = \boldsymbol{\beta}'\mathbf{x}$. If $\boldsymbol{\beta}$ is of full rank r, then $\mathbf{c} = (\boldsymbol{\beta}'\boldsymbol{\Sigma}_{\mathbf{x}}\boldsymbol{\beta})^{-1}\boldsymbol{\beta}'\boldsymbol{\Sigma}_x\boldsymbol{\alpha} \in \mathbb{R}^r$.

Comparing DGPs – In the DFM literature it is assumed that the source of randomness of the data generating process (DGP) stems from the latent factors $\mathbf{f}$ that generate the observable data $\mathbf{x}$. As a consequence, since the factors are unobserved, mostly unverifiable assumptions that guarantee their identifiability are required. If the source of randomness is instead in the observables $\mathbf{x}$, as in Proposition 1, then the "factors" $\mathbf{f}$ become simple linear reductions of the information in the observables. Such linear reductions might be chosen and formed by the econometrician or a statistical agency, and not inherently by the DGP. Different reductions will in general possess different properties, depending on the $\boldsymbol{\beta}$ used to construct them and the econometrician in principle can study such properties. Moreover, Proposition 1 pivots the focus back on the observables $\mathbf{x}$ and on the distributional assumptions on the observables under which conditions (LC) and (CVC) are satisfied.[11]

LC and CVC under Normality – Conditions (LC) and (CVC) of Proposition (1) are difficult to verify in practice given that the set of $\boldsymbol{\beta} \in \mathbb{R}^{p\times r}$ for which they hold is unknown. However, they are satisfied for any $\boldsymbol{\beta}$ if $\mathbf{x}$ is multivariate normal. In Section 5.2, we show that joint normality in our panel of macro variables is rejected.

LC under Ellipticity – Condition (LC) on the marginal distribution of the predictors, which is fundamental for SDR and reappears in (3.3), is satisfied for all $\boldsymbol{\beta} \in \mathbb{R}^{p\times r}$, if the predictors have an elliptically contoured distribution [See Eaton (1986)[37]]. The elliptically contoured family of distributions includes the multivariate normal and Student's t. The important intuition of De Mol et al. (2008) [33] is that strong cross-correlation of the variables included in macro panels explains the similar performance of PCR, RIDGE and LASSO. Herein, we provide a different interpretation of the comparable forecasting performance of these feature selection methods: as shown in Section 5.2 ellipticity appears to be an empirical characteristic of the large panel of macro variables, so that (LC) is satisfied and Corollary 1 applies. In this case, Corollary 1 implies that all estimators that can be written as a linear transformation of $\mathbf{x}$ will be roughly equivalent in summarizing the large $\mathbf{x}$ vector and linear regression forecast models will likely result in approximately equal MSFEs.

[11] Although conditions (i) and (ii) are difficult to check in practice they are based on observables, progress advocated by Bai (2003) [3], who acknowledges the problems arising from DFM assumptions placed on unobservable quatities.

LC and CVC under More General DGPs – Leeb (2013) [56] and Steinberger and Leeb (2015) [68] studied when conditions (LC) and (CVC) in Proposition 1 are satisfied. Building on a line of work initiated by Diaconis and Freedman (1984) [34] and Hall and Li (1993) [45], Steinberger and Leeb (2015) [68] showed that under comparatively mild conditions on the distribution of $\mathbf{x}$, both conditions (LC) and (CVC) are approximately satisfied for most matrices $\boldsymbol{\beta}$. In other words, they showed that most conditional means are approximately linear and most conditional variances are approximately constant for a large class of non-Gaussian multivariate distributions, when the conditioning is on lower-dimensional projections provided that p is sufficiently large relative to r. Conditions (LC) and (CVC) are also shown to hold as $p \to \infty$, where r may increase with p at the order $r = o(\log(p))$. This is an important result as the requirement of linear conditional means and constant conditional variances, which initially seems as quite restrictive, is actually satisfied, in an approximate sense, by a large class of distributions. Regarding (LC) in particular, Steinberger and Leeb (2015) [68] 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)[68]], then the conditional mean of $\mathbf{x}$ given $\boldsymbol{\beta}'\mathbf{x}$ is linear in $\boldsymbol{\beta}'\mathbf{x}$ for a $p \times r$ matrix $\boldsymbol{\beta}$, as $p \to \infty$ and r is either fixed or grows very slowly at the rate $r/\log p \to 0$. An appealing feature of these results is that they rely on bounds that can be estimated from data. Steinberger and Leeb's result therefore ascertains that condition (LC) is satisfied by a large class of predictor distributions. Thus, first-moment SDR estimators, such as Sliced Inverse Regression (SIR) in the ensuing Section 4.5, can be widely used to estimate basis elements of the column space of $\mathbf{v}$ in the reduction $\mathbf{R}(\mathbf{x}_t) = \mathbf{v}'\mathbf{x}_t$.

Summary – Proposition 1 takes a step back relative to the typical DFM shortcut of assuming that population model (2.5) is true and provides conditions under which it is possible to reduce the information of a large set of observables $\mathbf{x}$ so that (2.5) is the true model for forecasting y. Additionally, it replaces the many heterogeneous assumptions found in the DFM literature with a simple set of conditions and offers a common paradigm for OLS, PCR, RIDGE and PLS and other feature extraction estimators. As a consequence it provides an insight for the success of simple linear forward regression models in modeling and forecasting.

2.1 The DFM Forecasting Framework

The basic building blocks of the DFM forecasting framework generalize the case of a classic factor structure assuming that

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

$$\underset{(p\times 1)}{\mathbf{x}_t} = \underset{(p\times r)}{\mathbf{\Gamma}} \underset{(r\times 1)}{\mathbf{f}_t} + \underset{(p\times 1)}{\mathbf{u}_t} \tag{2.11}$$

where $\mathbf{f}_t$ and $\mathbf{u}_t$ are independent or weakly dependent, and $\mathbf{u}_t$ can be both serially and cross-sectionally correlated.

(ii) The forecasting model is (2.5) allowing for response lags as predictors,

$$y_{t+h} = \boldsymbol{\gamma}_y' \mathbf{f}_t + \boldsymbol{\gamma}' \mathbf{w}_t + \varepsilon_{t+h} \tag{2.12}$$

An important feature of this framework is the assumption that the same factors that determine the marginal distribution of $\mathbf{x}_t$ also determine the conditional distribution of y_t. Forecasting equation (2.12), the centerpiece of Section 2, receives only marginal attention in the DFM literature since all effort is shifted to reducing the dimension of the set of explanatory variables.

The linear factor structure ascertains that the information content of the regressors has "dimension" r, in the sense that the $r < p$ factors $\mathbf{f}_t$ in (2.11) introduce structure that reduces the high-dimensionality of the problem. The complexity in the high dimensionality of the problem is traded off with the need of identifying and estimating additional fictional unobserved variables, the latent $\mathbf{f}_t$.

In addition to the linear factor structure, Stock and Watson (2002) [70] provided a set of assumptions under which the factors can be identified and estimated via *principal components*. Such assumptions have been subsequently modified and updated according to various needs, also to justify alternative estimators.

2.2 The SDR Forecasting Framework

In contrast to DFM, SDR methods depart from the linear factor assumption. In fact, SDR methods thrive when the relationship between the target variable and the predictors contains non-linearities. The SDR apparatus, which we introduce in detail in Section 3, can be viewed as a generalization of the setup of Proposition 1 that allows the target variable to be a non-linear function of the predictors. Instead of imposing an artificial latent factor structure on the panel $\mathbf{x}_t$, SDR works **directly** with observables and seeks to identify how many and which functions of the explanatory variables are needed to fully describe the conditional cumulative distribution function $F(y_{t+h}|\mathbf{x}_t)$. Specifically, SDR aims to identify and estimate functions of the predictors $\mathbf{R}(\mathbf{x}_t)$, so

that $F\left(y_{t+h}|\mathbf{x}_t\right) = F\left(y_{t+h}|\mathbf{R}(\mathbf{x}_t)\right)$. These functions are called reductions because they preserve all the information that $\mathbf{x}_t$ carries about y_{t+h}. Obviously, only if such functions are fewer than p do they represent proper reductions.

Reductions can be either linear or nonlinear functions of the panel data. In order to draw a more pertinent comparison with the DFM literature, we focus on **linear moment-based** SDR methods, which place conditions on the marginal distribution of $\mathbf{x}_t$, such as the linearity assumption (2.13), a first moment condition that coincides with condition (LC) in Proposition 1 and Corollary 1 and is analogous to the linear factor structure (2.11) of DFM.[12]

The key ingredients of the proposed SDR-based forecasting framework are:

(i) The *linearity condition* (LC) on the marginal distribution of $\mathbf{x}_t$

$$\mathrm{E}\left[\mathbf{x}_t|\boldsymbol{\beta}'\mathbf{x}_t\right] = \mathbf{A}\boldsymbol{\beta}'\mathbf{x}_t \tag{2.13}$$

for a matrix $\mathbf{A}$ and a $p \times d$ full rank matrix $\boldsymbol{\beta}$ with $0 \leq d \leq p$.[13]

(ii) The *forecasting model*

$$y_{t+h} = g(\boldsymbol{\beta}'\mathbf{x}_t, \varepsilon_{t+h}) \tag{2.14}$$

Linear SDR aims to identify a $p \times d$ matrix $\boldsymbol{\beta}$, of rank $d < p$, so that $\mathbf{R}\left(\mathbf{x}_t\right) = \boldsymbol{\beta}'\mathbf{x}_t$. However, in contrast with the DFM setup, no dependence on underlying factors is postulated. The forecast, or forward in SDR, model is specified in (2.14) and is analogous to (2.2), although $g(\cdot)$ is a general 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 forecasting model; that is, without specifying the exact relationship between y_t and $\boldsymbol{\beta}'\mathbf{x}_t$. They replace a large number of explanatory variables by a few linear combinations without loss of inferential information; their number $d = \mathrm{rank}(\boldsymbol{\beta})$ is the dimension of the regression problem. Because SDR targets y, typically fewer sufficient reductions than PCs are needed in order to generate a comparable mean squared forecast error (MSFE). As

[12]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.

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

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.

2.3 Comparison of Assumptions in the Two Approaches

Proposition 1, Corollary 1, the SDR and DFM approaches all impose conditions on the marginal distribution of $\mathbf{x}$. However, Proposition 1, Corollary 1 and SDR place restrictions on the observables, $\mathbf{x}_t$. Even though conditions (i) and (ii) in Proposition 1 are difficult to verify in practice, they are based on the observed $\mathbf{x}_t$, a definite progress also advocated by Bai (2003) [3] relative to assumptions made on unobservable quantities such as latent factors and idiosyncratic errors in the DFMs. Moreover, the absence of latent variables implies that there is no need to consider double asymptotics in order to study the statistical properties of SDR estimators for macroeconomic data, which will be derived with standard large–T asymptotics for the specific SDR method, Sliced Inverse Regression (SIR), we implement. The latter requires only the linearity condition (2.13), or equivalently, condition (LC) of Corollary 1 to hold for the linear projections $\boldsymbol{\beta}'\mathbf{x}$ that satisfy the general forecasting regression model $F(y|\mathbf{x}) = F(y|\boldsymbol{\beta}'\mathbf{x})$.

The SDR approach indicates useful directions in which to explore the assumptions made in the DFM literature by relating them to conditions (LC) and (CVC).

Factors as Linear Projections of the Regressors – If $\boldsymbol{\Gamma}'\boldsymbol{\Gamma}$ is invertible, right multiplying by $(\boldsymbol{\Gamma}'\boldsymbol{\Gamma})^{-1}\boldsymbol{\Gamma}'$ both sides of (2.11) obtains

$$\mathbf{f}_t = (\boldsymbol{\Gamma}'\boldsymbol{\Gamma})^{-1}\boldsymbol{\Gamma}'\mathbf{x}_t + (\boldsymbol{\Gamma}'\boldsymbol{\Gamma})^{-1}\boldsymbol{\Gamma}'\mathbf{u}_t \tag{2.15}$$

Moreover, if $\boldsymbol{\Gamma}'\mathbf{u}_t = 0$ then $\mathbf{f}_t$ is a linear transformation of $\mathbf{x}_t$ with $(\boldsymbol{\Gamma}'\boldsymbol{\Gamma})^{-1}\boldsymbol{\Gamma}'$ corresponding to $\boldsymbol{\beta}$ in Proposition 1. Asymptotic invertibility of $\boldsymbol{\Gamma}'\boldsymbol{\Gamma}$ is required in all core DFM papers, for instance in Assumption F1.a in Stock and Watson (2002a) [70], or Assumption B in Bai and Ng (2002) [4]. Assumption B(ii) in De Mol et al. (2008) [33] requires invertibility also in finite samples. These conditions imply that each factor has a nontrivial contribution to the variance of $\mathbf{x}_t$, that is the factors are "evenly spread" across the predictors. Condition $\boldsymbol{\Gamma}'\mathbf{u}_t = 0$ is automatically satisfied in the population and on average in the sample when non-random factor loadings are considered. In the case of random factor loadings, it is required they be independent of the factors and idiosyncratic errors (see Bai (2003) [3]) or assumption F2.1 in Stock and Watson (2002a) [70]).

LC and Factorial Structure – When $\mathbf{x}_t$ is generated by a factorial structure (2.11), then

$$\mathrm{E}(\boldsymbol{\alpha}'\mathbf{x}_t|\mathbf{f}_t) = \boldsymbol{\alpha}'\,\mathrm{E}(\boldsymbol{\Gamma}\mathbf{f}_t + \mathbf{u}|\mathbf{f}_t) = \boldsymbol{\alpha}'\boldsymbol{\Gamma}\mathbf{f}_t + \boldsymbol{\alpha}'\,\mathrm{E}(\mathbf{u}|\mathbf{f}_t) \tag{2.16}$$

If $\mathrm{E}(\mathbf{u}|\mathbf{f}_t) = 0$, then $\mathrm{E}(\boldsymbol{\alpha}'\mathbf{x}_t|\mathbf{f}_t)$ is linear in $\mathbf{f}_t$ for any α. Condition $\mathrm{E}(\mathbf{u}_t|\mathbf{f}_t) = 0$ is a standard assumption in classical factor models and is also assumed in De Mol et al. (2008) [33]. It is relaxed in core DFM papers (see assumption D in Bai and Ng (2002) [4]) by requiring that the covariance between $\mathbf{u}_t$ and $\mathbf{f}_t$ is asymptotically zero (e.g., see assumptions F1 and M1 in Stock and Watson (2002a) [70] or assumptions A through D in Bai and Ng (2002) [4]). The condition is instead implied by the stronger independence assumption between $\mathbf{u}_t$ and $\mathbf{f}_t$ in order to obtain inferential results beyond the estimate of the space spanned by the factors (see, for instance, Assumption FTE in Bai and Ng (2008) [6]).

CVC and Factorial Structure – Under a factorial structure for $\mathbf{x}_t$,

$$\mathrm{Var}(\boldsymbol{\alpha}'\mathbf{x}_t|\mathbf{f}_t) = \boldsymbol{\alpha}'[\underbrace{\mathrm{Var}(\boldsymbol{\Gamma}\mathbf{f}_t|\mathbf{f}_t)}_{=0} + \mathrm{Var}(\mathbf{u}_t|\mathbf{f}_t) + 2\underbrace{\mathrm{Cov}(\boldsymbol{\Gamma}\mathbf{f}_t, \mathbf{u}_t|\mathbf{f}_t)}_{=0}]\boldsymbol{\alpha} \tag{2.17}$$

$$= \boldsymbol{\alpha}' \mathrm{Var}(\mathbf{u}_t|\mathbf{f}_t)\boldsymbol{\alpha} \tag{2.18}$$

Therefore condition (CVC) in Proposition (1) holds in a factor model whenever $\mathrm{Var}(\mathbf{u}_t|\mathbf{f}_t)$ is constant. Independence of $\mathbf{u}_t$ and $\mathbf{f}_t$ suffices for ensuring constancy, as is the case in classical factor models. In the DFM framework, assumptions that place bounds on both the moments of $\mathbf{u}_t$ and the dependence of $\mathbf{u}_t$ on $\mathbf{f}_t$ go in the direction yet fall short of ensuring that the constant variance assumption holds.

Factorial Structure Underpinning Alternative Estimators – Although the most common estimator in DFMs is principal components, most of the papers that experiment with alternative estimators also assume that the data are generated by a factorial structure. De Mol et al. (2008) [33] assume the approximate factorial structure (2.11) and DFM forecast model (2.5). Within that framework they show that the population OLS regression forecast (2.3) converges to the DFM population forecast (2.6). They also show that the sample RIDGE forecast, for a wisely tuned choice of the RIDGE meta parameter, converges to the population OLS regression forecast (2.3) hence to the DFM population forecast (2.6). The authors have shown a result which is approximately the converse of Proposition 1: under a factor structure, the (assumed true) reduced population model (2.5) can be replaced with population model (2.2). The theory in the paper is silent on which estimator is better, however our results in the empirical section suggest that PCR and RIDGE perform very similarly.[14] The theoretical results in De Mol et al. (2008) [33] can be interpreted through the lenses

[14]It is also shown that LASSO achieves MSFE similar to PCR and RIDGE.

of Proposition 1 and be taken to suggest that the factorial structure plus the extra assumptions that they impose induce near-elipticity in the population.

Groen and Kapetanios (2014) [44] also adopt a factor structure although they depart from forecast model (2.5) after pointing out its various shortcomings and revert to forecast model (2.2). They show that under forecast model (2.2), PCR can be dominated by PLS and RIDGE.

Kelly and Pruitt (2015) [55] essentially study the PLS estimator under a factor structure and forecast model (2.5). Their factorial structure allows for both relevant and irrelevant factors and the latter can distort the performance of PCR. They show in simulations that PLS, a targeted method, has the advantage of not being affected by irrelevant factors. In practice, in our out-of-sample forecasting experiment with a large panel of macro variables PLS forecasting performace does not appear to dominate PCR.

In order to understand why PCR can dominate PLS in practical situations it is necessary to delve deeper into the intricacies of PLS geometry as done by Carrasco and Rossi (2016) [18]. Among others, they study PLS, PCR, and RIDGE both under a factor structure and in a "ill-posed" problem in which the eigenvalues of $\mathbf{\Sigma}_x$ are bounded and the smallest eigenvalue declines to zero fast as p increases, as would be the case, for example, when the additional regressors are strongly correlated with those already included in the panel. They establish that, under an ill-posed problem, the regularization bias of PLS is smaller than that of PCR, whereas the estimation error may be larger with an uncertain effect on their relative MSE.

In summary, our SDR approach is complementary to the DFM framework. Proposition 1 is likely a better organizing framework for interpreting empirical results *if the final objective is prediction or forecasting*, that is *if there is a natural candidate target variable* y. However, if the purpose is to identify the basic forces driving a panel of variables, the DFM framework remains a very effective device.

3 Sufficient Dimension Reductions

In this section we define general sufficient reductions and introduce the tools we need from moment-based SDR focusing on linear sufficient reductions.[15]

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*

[15]Recently, Bura and Forzani (2015) [15] and Bura, Duarte and Forzani (2016) [16] derived non-linear sufficient reductions for elliptically contoured and exponential family inverse predictors, respectively.

equivalently

$$F(y|\mathbf{x}) = F(y|\mathbf{R}(\mathbf{x})) \tag{3.1}$$

A consequence of the definition of sufficiency is that, since (3.1) can be written as $F(y|\mathbf{x}, \mathbf{R}(\mathbf{x})) = F(y|\mathbf{R}(\mathbf{x}))$ we have $y \perp\!\!\!\perp \mathbf{x}|\mathbf{R}$, where $\perp\!\!\!\perp$ denotes statistical independence. The function $\mathbf{R}(\mathbf{x})$ is called a **forward reduction**. Although the term "sufficient" was originally coined to capture the information preserving role of $\mathbf{R}(\mathbf{x})$, there is a specific link with the Fisherian concept of statistical sufficiency (see Cook (2007) [26]). First, 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*

$$\mathbf{x}|(\mathbf{R}(\mathbf{x}), y) \stackrel{d}{=} \mathbf{x}|\mathbf{R}(\mathbf{x}) \tag{3.2}$$

If one views y as a parameter, (3.2) 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 $\mathbf{x}$. Then,*

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

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 $\mathbf{x}|(\mathbf{R}(\mathbf{x}), y) \stackrel{d}{=} \mathbf{x}|\mathbf{R}(\mathbf{x})$, 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 one is free to choose the most convenient way to determine a sufficient reduction, either from the forward or

inverse regression. An 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, including non-parametric, where the curse of dimensionality would make modeling of the forward regression practically impossible. Most importantly, inverse regression identifies a function of the predictor data that encapsulates all the information they contain about the "parameter" to be estimated and predicted, which is the whole time-series y_t.

3.1 Linear Reductions and Moment-Based SDR

Moment-based SDR was developed under the requirement the reduction be linear. In *linear* SDR, $\mathbf{R}(\mathbf{x}_t)$ is a projection of $\mathbf{x}_t$ onto a lower-dimensional subspace 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. 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 $\mathbf{x}_t$.

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.

Condition 1 *Suppose the reduction* $\mathbf{R}(\mathbf{x})$ *is sufficient and a linear function of* $\mathbf{x}$; *that is, it satisfies (3.1) and* $\mathbf{R}(\mathbf{x}) = \boldsymbol{\alpha}'\mathbf{x}$ *for some* $p \times d$ *matrix* $\boldsymbol{\alpha}$.

Let $\mathcal{R}(\mathbf{A})$ denote the column space of a matrix $\mathbf{A}$. The definition of sufficiency implies that we can only identify the subspace spanned by a linear reduction, $\mathcal{R}(\boldsymbol{\alpha})$, rather than $\boldsymbol{\alpha}$ per se, since $F(y|\boldsymbol{\alpha}'\mathbf{x}) = F(y|\mathbf{b}'\mathbf{x})$ for all matrices $\boldsymbol{\alpha}$ and $\mathbf{b}$ such that $\mathcal{R}(\boldsymbol{\alpha}) = \mathcal{R}(\mathbf{b})$. A subspace spanned by the columns of a matrix $\boldsymbol{\alpha}$ with $F(y|\mathbf{x}) = F(y|\boldsymbol{\alpha}'\mathbf{x})$ 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 dimension reduction subspace 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 ensure 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)[23] for

more details and, henceforth, restrict attention to those regressions for which a central subspace exists.

The identification of a 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*

$$\mathrm{E}\left[\mathbf{x}|\mathbf{v}'\mathbf{x}\right] = \mathbf{A}\mathbf{v}'\mathbf{x} \tag{3.3}$$

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

The following lemma links the linearity condition with inverse regression and points to where the reduction can be found.

Lemma 1 *Assume $\mathbf{R}(\mathbf{x}) = \mathbf{v}'\mathbf{x}$ satisfies (3.1), that is, it is a sufficient reduction, and the linearity condition (3.3) is satisfied for $\mathbf{v}$. Then*

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

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

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

Proof. See Corollary 10.1 in Cook (1998)[23] and Theorem 3.1 in Li (1991) [57]. ■

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 $\boldsymbol{\Sigma_{\mathbf{x}}}^{-1}\left[\mathrm{E}\left(\mathbf{x}|y\right) - \mathrm{E}\left(\mathbf{x}\right)\right]$ is contained in a subspace that is spanned by the columns of $\mathbf{v}$. The following proposition provides a means to identify such a space.

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

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

Proof. See Proposition 11.1 in Cook (1998)[23], an extension of Proposition 2.7 in Eaton (1983)[36], and Lemma 1. ■

Lemma 1 and Proposition 3 establish a link between the distribution of the data and the subspace we wish to identify. Notice that in general the column space of $\mathbf{\Sigma_x}^{-1} \operatorname{Var}(\mathrm{E}(\mathbf{x}|y))$ 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 $\mathrm{E}\left(\mathbf{\Sigma_x} - \operatorname{Var}(\mathrm{E}(\mathbf{x}|y)\right)^2$, also live in the central subspace under the additional constant variance condition on the marginal distribution of the predictors (Cook and Weisberg (1991) [31]). As our goal is to introduce SDR methodology to the econometrics literature, we avoid cluttering the present exposition and focus on the first inverse regression moment $\mathrm{E}(\mathbf{x}|y)$ via the simple and widely used Sliced Inverse Regression (SIR, Li (1991)[57]).[16]

Linear moment-based SDR methods reduce significantly the complexity of modeling and uncover the structural dimension of the forward regression problem, i.e. how many derived linear combinations (*directions*) of the original predictors suffice to explain y. They estimate the number and coefficients (up to rotations, as in the DFM literature) of the linear combinations of the predictors in the forward forecasting equation. SDR does not specify the functional form of the forward regression. 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.

3.2 Sliced Inverse Regression

Sliced Inverse Regression (SIR), the first and most widely used linear SDR method, was proposed by Li (1991)[57]. SIR is a semiparametric method for finding a dimension reduction subspace in regression. It is based on the results of Section 3.1 and uses a sample counterpart to $\mathbf{\Sigma}^{-1} \operatorname{Var}(\mathrm{E}(\mathbf{x}|y))$.[17] 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 simple nonparametric approximation to $\mathrm{E}(\mathbf{x}|y)$.

Implementation of SIR – In order to estimate $\mathbf{M} = \operatorname{cov}(\mathrm{E}(\mathbf{x}|y))$, the range of the observed re-

[16] Although SIR generally identifies a subset of 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) [26] and Bura and Forzani (2015) [15]).

[17] SIR discretizes y through slicing when y is continuous. It can be shown that the space spanned by the slice predictor means is a subset of the central subspace (e.g., Cook (1998) [23]).

sponses $\mathbf{Y} = (y_1, \ldots, y_T)'$ 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 by $\bar{\mathbf{x}} = (\sum_{t=1}^{T} x_{t1}/T, \ldots, \sum_{t=1}^{T} x_{tp}/T)'$, and 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, for $j = 1, \ldots, J$. The covariance matrix of $\mathbf{x}_t$ is estimated by the sample covariance matrix $\widehat{\mathbf{\Sigma}}_x = \sum_{t=1}^{T}(\mathbf{x}_t - \bar{\mathbf{x}})(\mathbf{x}_t - \bar{\mathbf{x}})'/T$, and the SIR seed matrix $\mathbf{M}$ with

$$\widehat{\mathbf{M}} = \sum_{j=1}^{J} \frac{n_j}{T}(\bar{\mathbf{x}}_j - \bar{\mathbf{x}})(\bar{\mathbf{x}}_j - \bar{\mathbf{x}})' = \widehat{\text{var}}\left(\hat{\text{E}}(\mathbf{x}|y)\right)$$

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 $\widehat{\mathbf{B}} = \widehat{\mathbf{\Sigma}}_x^{-1}(\widehat{\mathbf{u}}_1, \ldots, \widehat{\mathbf{u}}_d) = (\boldsymbol{b}_1, ..., \boldsymbol{b}_d)$ 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)[13], Bura and Yang (2011)[17]), information criteria such as AIC and BIC, or permutation tests (Yin and Cook (2001)[32]. These tests, though, are valid under the assumption of random 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 $J = T$; i.e., when each slice corresponds to a single y observation, $\widehat{\mathbf{M}}$ becomes $\widehat{\mathbf{\Sigma}}_x$, 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 $\mathbf{x}$ projections targeted to y.

3.3 Consistency of SIR Estimators

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 (3.3) and have conditional distributions $\mathbf{x}_t|y_{t+h}$, $h = 1, 2, ...$ 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.* $\sum_{l=-\infty}^{\infty} |\sigma_{jj}(l)| < \infty, \sum_{l=-\infty}^{\infty} |\sigma_{jj|y_{t+h}}(l)| < \infty, j = 1, \ldots, p$. *Then, the SIR directions are consistent estimators of directions in the central subspace for all* $\mathbf{x}_t$ *satisfying the linear design condition (3.3).*

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 $p_s = \Pr(y_{t+h} = s)$ 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 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 $\widehat{\boldsymbol{\Sigma}}_x$ is consistent for $\boldsymbol{\Sigma}_x$, the SIR predictors $\widehat{\boldsymbol{\Sigma}}_x^{-1}\widehat{\mathbf{u}}_k$, $k = 1, \ldots, d$ 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 we use are provided in Appendix B.

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 B). Under the same assumptions we can then obtain that $\widehat{\mathbf{M}}_h$ is asymptotically normal following similar arguments as Bura and Yang (2011)[17] who derived the asymptotic distribution of $\widehat{\mathbf{M}}$ when the data are i.i.d. draws from the joint distribution of $(y, \mathbf{x})$.

3.4 Inverse Regression as Extraction of Targeted Factors

In general, inverse regression focuses on the set of p inverse regressions

$$\mathbf{x} = \mathbf{a} + \mathbf{B}\mathbf{f}(y) + \mathbf{e} \tag{3.4}$$

where y is substituted by the vector of functions of y, $\mathbf{f}(y)$, 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.

SIR is a simple case of (3.4), where $\mathbf{f}(y) = (f_1(y), \ldots, f_{J-1}(y))'$ is a vector of step functions with

$$f_s(y) = I(y \in S_s) - \frac{n_s}{T}, \quad s = 1, \ldots, J-1,$$

where J is the number of the disjoint slices $S_1, \ldots, S_J$ whose union is the range of the y-values, $I(\cdot)$ is the indicator function, and n_s is the number of observations in S_s. Parametric inverse regression (PIR) (Bura and Cook (2001b) [14]) and principal fitted components (PFC) (Cook and Forzani (2008)[27]) 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 (3.4). The term $\mathbf{f}(y)$ is analogous to a factor structure, but it is observable. Intuitively, inverse regression replaces $\mathbf{x}$ with its projection on $\mathbf{f}(y)$ and, in so doing, it extracts its "targeted" factor structure.

4 Dimension Reduction Methods via Linear Combinations

In this section we review some widely used estimators, which form linear combinations of the explanatory variables $\boldsymbol{\beta}'\mathbf{x}_t$ as a data reduction step prior to fitting the model used for prediction. We cast OLS, PCR, a method often used to extract factors in dynamic factor analysis, RIDGE and PLS regression in a shared framework of maximization of an objective function that distinguishes them from each other. We also present SIR as the solution to a maximization problem and conclude with a comparison of the estimators.

To motivate the discussion about the relative drawbacks and advantages of the different methods, we start from a simple data generating model, where the predictors and the response are jointly normal, which is the simplest DGP that implies 1-dimensional linear reduction:

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

We assume $\boldsymbol{\Sigma}_x$ is invertible throughout this section. Under this setting, the best predictor under quadratic loss is the linear regression function,

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

with $\boldsymbol{\beta} = \boldsymbol{\Sigma}_x^{-1}\boldsymbol{\sigma}_{xy} = \boldsymbol{\beta}_{OLS}$. Thus, in a normal DGP the relationship between $\mathbf{x}$ and y is entirely and exhaustively encapsulated in one linear combination of the predictors and OLS is the optimal population estimator and spans the central subspace. When joint normality does not hold, under the assumptions of Section 3, $\boldsymbol{\beta}_{OLS}$ remains one of the basis elements of the central subspace, even though *more* linear reductions may be required to exhaustively capture the information that $\mathbf{x}$ carries on y. If the assumptions of Section 3 do not hold, even small departures from normality can result in linear reductions no longer being exhaustive (see Bura and Forzani (2015) [15]).

4.1 Ordinary Least Squares (OLS)

The OLS coefficient is the solution to the following maximization problem:

$$\max_{\{\boldsymbol{\beta}\}} \text{Corr}^2(y, \mathbf{x}'\boldsymbol{\beta}) \tag{4.1}$$

The first order conditions of the problem lead to the the normal equations $\boldsymbol{\Sigma}_x\boldsymbol{\beta} = \boldsymbol{\sigma}_{xy}$ and assuming that $\boldsymbol{\Sigma}_x$ is full rank, the unique solution to (4.1) is $\boldsymbol{\beta}_{OLS} = \boldsymbol{\Sigma}_x^{-1}\boldsymbol{\sigma}_{xy}$. Therefore OLS selects the *one and only one* linear combination $\mathbf{x}'\boldsymbol{\beta}$ with maximum correlation with the target y. The OLS prediction of the response y at an observed $\mathbf{x}_0$ is

$$y_{OLS} = \mathbf{x}_0'\boldsymbol{\beta}_{OLS}$$

The spectral value decomposition of $\boldsymbol{\Sigma}_x = \mathbf{U}\Lambda\mathbf{U}'$, where $\boldsymbol{\Lambda}$ is the diagonal matrix with the eigenvalues $\lambda_1 \geq \ldots \geq \lambda_p$ of $\boldsymbol{\Sigma}_x$ on its main diagonal, and $\mathbf{U}$ is the $p \times p$ orthogonal matrix of the corresponding eigenvectors, yields $\boldsymbol{\Sigma}_x^{-1} = \mathbf{U}\Lambda^{-1}\mathbf{U}'$. As a result the OLS solution can be written as $\boldsymbol{\beta}_{OLS} = \mathbf{U}\Lambda\mathbf{U}'\boldsymbol{\sigma}_{xy} = \sum_i^p \frac{1}{\lambda_i}\mathbf{u}_i\mathbf{u}_i'\boldsymbol{\sigma}_{\mathbf{x}y}$. Notice that the population model we assume implies $\mathcal{R}(\boldsymbol{\Sigma}_x) = \mathcal{R}(\boldsymbol{\Sigma}_x^{-1}) = \mathbb{R}^p$. In practice, in typical samples encountered in macro and finance forecasting, even if $p << T$, collinearity or ill-conditioning imply that the estimator of $\boldsymbol{\Sigma}_x$ is non-invertible or numerically unstable resulting in grossly inaccurate OLS predictions.

4.2 RIDGE Regression (RR)

RIDGE regression has been reviewed in the macro-forecasting literature by De Mol et al. (2008) [33] in connection with Bayesian regression. RIDGE regression minimizes the least squares criterion on the sphere with radius a:

$$\max_{\substack{\{\boldsymbol{\beta}\} \\ \{\sum_{j=1}^p \beta_j^2 \leq a\}}} \text{Corr}^2(y, \mathbf{x}'\boldsymbol{\beta}) \tag{4.2}$$

The first order conditions leads to modified normal equations and the solution to (4.2) is

$$\boldsymbol{\beta}_{RR}(\kappa) = (\boldsymbol{\Sigma}_x + \kappa\mathbf{I})^{-1}\boldsymbol{\sigma}_{xy} = (\mathbf{I} + \kappa\boldsymbol{\Sigma}_x^{-1})^{-1}\boldsymbol{\beta}_{OLS} \tag{4.3}$$

where κ denotes the Lagrange multiplier in the constrained maximization (4.2) and is a function of a. It is also a meta parameter indexing the RIDGE family. When $\kappa = 0$, the maximization is the same as in OLS. As $\kappa > 0$ increases, the solution deviates from the OLS solution and the penalization shrinks the coefficients of the variables with smaller variance more. Nevertheless, RIDGE does not remove predictors from the regression. The RIDGE prediction at $\mathbf{x}_0$ is

$$y_{RR} = \mathbf{x}_0'\boldsymbol{\beta}_{RR}(\kappa).$$

As OLS, RIDGE estimates the response mean with one linear combination of the predictors. In practice the advantage of RIDGE over OLS comes from the fact that RIDGE does not suffer from the $p < T$ limitation of OLS, it can handle a large number of predictors regardless of the sample size, and delivers typically more stable estimation and prediction when $\boldsymbol{\Sigma}_x$ is ill-conditioned.

4.3 Principal Component Regression (PCR)

PCR operates in two stages. First, the linear combinations that maximize the variance of $\mathbf{x}$ and are mutually orthogonal are extracted as the solution to the following maximization problem

$$\max_{\substack{\{\mathbf{u}_i\} \\ \{\mathbf{u}_i'\mathbf{u}_i=1\} \\ \{\mathbf{u}_i'\boldsymbol{\Sigma}_x\mathbf{u}_j=0\}_{j=1}^{i-1}}} \mathrm{Var}(\mathbf{x}'\mathbf{u}_i) \tag{4.4}$$

A maximum of p such linear combinations whose coefficients are the eigen-vectors corresponding to the largest m eigenvalues of $\boldsymbol{\Sigma}_x$, called principal components, can be extracted. Secondly, y is regressed on the first $m(\leq p)$ principal components (PCs), $\widehat{\mathbf{x}} = (\mathbf{x}'\mathbf{u}_1, \ldots, \mathbf{x}'\mathbf{u}_m)$, with $\boldsymbol{\beta}$ solving

$$\max_{\{\boldsymbol{\beta}\}} \mathrm{Corr}^2(y, \widehat{\mathbf{x}}'\boldsymbol{\beta}). \tag{4.5}$$

The number of PCs, m, is a meta parameter chosen by the user. The solution to (4.5) is

$$\boldsymbol{\beta}_{PCR}(m) = \boldsymbol{\Sigma}_x^{-}(m)\boldsymbol{\sigma}_{xy} \tag{4.6}$$

and can be shown to be the minimum norm solution to the normal equations in the subspace spanned by the linearly independent columns of $\boldsymbol{\Sigma}_x$. The pseudo-inverse $\boldsymbol{\Sigma}_x^{-}(m)$ in (4.6) depends on m. It is obtained by truncating $\boldsymbol{\Sigma}_x^{-1}$ to retain only the first m components that explain a

specified amount of variance in the predictors. Such operation entirely disregards the response y and the p principal components 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 (4.6). The PCR-based prediction is

$$y_{PCR} = \mathbf{x}_0' \boldsymbol{\beta}_{PCR}.$$

To further appreciate how PCR operates consider that since $\boldsymbol{\Sigma}_x$ is full rank and $\boldsymbol{\sigma}_{\mathbf{x}y} \in \mathbb{R}^p$, then $\boldsymbol{\sigma}_{\mathbf{x}y} = \sum_{j=1}^{p} c_j \mathbf{u}_j$ for $c_j \in \mathbb{R}$, and $\mathbf{u}_j$, $j = 1, \ldots, p$ are the eigenvectors of $\boldsymbol{\Sigma}_x$. Therefore, $\boldsymbol{\Sigma}_x \boldsymbol{\sigma}_{\mathbf{x}y} = \boldsymbol{\sigma}_{\mathbf{x}y}$; that is, the covariance of $\mathbf{x}$ and y is contained in the span of $\boldsymbol{\Sigma}_x$. When $m = p$ in (4.6), $\boldsymbol{\beta}_{PCR} = \boldsymbol{\Sigma}_x^{-1} \boldsymbol{\sigma}_{\mathbf{x}y} = \boldsymbol{\beta}_{OLS}$. When $m < p$ PCs are used, since $\mathbf{u}_j$ are orthonormal, (4.6) yields

$$\boldsymbol{\beta}_{PCR}(m) = (\mathbf{u}_1, \ldots, \mathbf{u}_m) \operatorname{diag}(\lambda_1^{-1}, \ldots, \lambda_m^{-1})(\mathbf{u}_1, \ldots, \mathbf{u}_m)' \sum_{j=1}^{p} c_j \mathbf{u}_j = \sum_{j=1}^{m} \frac{c_j}{\lambda_j} \mathbf{u}_j$$

That is, only the part of $\boldsymbol{\sigma}_{\mathbf{x}y}$ in the span of the *first* m PCs is captured in $\boldsymbol{\beta}_{PCR}(m)$ and contributes to the PCR-based prediction. Therefore, if the m eigenvectors that correspond to the m largest eigenvalues of $\boldsymbol{\Sigma}_x$ happen to miss $\boldsymbol{\sigma}_{\mathbf{x}y}$ or part of it, then PCR's performance will be sub-optimal in capturing the linear signal about the target. In applications, ill-conditioning of $\boldsymbol{\Sigma}_x$ implies that PCR with an appropriate choice of m dominates OLS in samples typically encountered in the macro-finance literature.

4.4 Partial Least Squares (PLS)

PLS is an increasingly popular method of dimension reduction that has recently resurfaced in econometrics, within macro-forecasting applications, with Kelly and Pruitt (2015) [55] and Groen and Kapetanios (2014) [44], mostly because PLS handles regressions where $p > T$.[18]

PLS solves the maximization problem [see Stone and Brooks (1990) [77]]:

$$\max_{\substack{\{\boldsymbol{\beta}_i\} \\ \{\boldsymbol{\beta}_i'\boldsymbol{\beta}_i = 1\} \\ \{\boldsymbol{\beta}_i'\boldsymbol{\Sigma}_x\boldsymbol{\beta}_j = 0\}_{j=1}^{j-1}}} \operatorname{Corr}^2(y, \mathbf{x}'\boldsymbol{\beta}_i) \operatorname{Var}(\mathbf{x}'\boldsymbol{\beta}_i) \tag{4.7}$$

and combines in one step the two maximizations carried out separately by PCR, thus "coming closer" to OLS. Analogously to the extraction of PCs, the maximization in (4.7) can in principle

[18] PLS followed an uneven trajectory in Econometrics. Originally developed by H. Wold [81] in the mid 70's, it did not gain much traction in Econometrics and swiftly fell practically 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 80's and early 90's (see, for example, Helland (1988) [48]).

be computed using the eigen-decomposition of $\text{Cov}(\mathbf{x}, y) = \sigma_{xy}$, so that it extracts it signal by focusing on the principal directions of $\text{Cov}(\mathbf{x}, y)$.[19] However, the most efficient way to find the PLS solution is by applying a algorithm that avoids the computation of $\mathbf{\Sigma}_x^{-1}$ and it has been shown to be the *conjugate gradient* method applied to the normal equations [see Wold et al. (1984) [82]].[20] One can show that prediction from PLS admits a linear form at $\mathbf{x}_0$,

$$y_{PLS} = \mathbf{x}_0' \boldsymbol{\beta}_{PLS}(s)$$

where $\boldsymbol{\beta}_{PLS}(s) = \mathbf{W}_s(\mathbf{W}_s'\mathbf{\Sigma}_x\mathbf{W}_s)^{-1}\mathbf{W}_s'\boldsymbol{\sigma}_{xy}$.[21] The algorithm always converges in the sense that after p steps it stops with the solution $\boldsymbol{\beta}_{PLS}(p) = \mathbf{\Sigma}_x^{-1}\boldsymbol{\sigma}_{xy} = \boldsymbol{\beta}_{OLS}$. However, it is crucial to understand that contrary to PCR even after $s < p$ steps the population algorithm may reach the OLS solution, depending on the eigen-structure of $\mathbf{\Sigma}_x^{-1}$, as we explain next.

The eigen-representation of $\boldsymbol{\beta}_{OLS} = \mathbf{\Sigma}_x^{-1}\boldsymbol{\sigma}_{xy} = \sum_{j=1}^{p} \frac{1}{\lambda_j}\mathbf{u}_j\mathbf{u}_j'\boldsymbol{\sigma}_{xy}$ can be simplified in two ways. First, remove all eigenvectors orthogonal to the signal; i.e., remove $\mathbf{u}_j$ with $\mathbf{u}_j'\boldsymbol{\sigma}_{xy} = 0$. Secondly, eigenvectors corresponding to eigenvalues λ_j with multiplicity $g \geq 2$ are condensed in a linear combination $\mathbf{u}_j^*$ with weights that preserve their aggregate contribution to $\boldsymbol{\beta}_{OLS}$, so that $\mathbf{u}_j^*\mathbf{u}_j^{*\prime}\boldsymbol{\sigma}_{xy} = \sum_{j=1}^{g}\mathbf{u}_j\mathbf{u}_j'\boldsymbol{\sigma}_{xy}$. It can be shown that $\mathbf{u}_j^*$ is itself an eigenvector, orthonormal to the eigenvectors associated with the remaining eigenvalues $\lambda_{i\neq j}$. These two operations can reduce the number of eigenvectors from p to $s < p$, and $\boldsymbol{\beta}_{OLS} = \boldsymbol{\beta}_{PLS} = \sum_{j=1}^{s}\mathbf{u}_j^*\mathbf{u}_j^{*\prime}\boldsymbol{\sigma}_{xy}$. The formula of $\boldsymbol{\beta}_{PLS}(s)$ reveals that PLS exhaustively captures the linear signal $\boldsymbol{\sigma}_{xy}$ also when $s < p$, in contrast to PCR with $m < p$. In practice, s is unknown and setting $s << p$ implies some "truncation" of $\boldsymbol{\beta}_{OLS}$ so that the realized performance of PLS relative to PCR will depend both on the sample and on the unknown eigen-structure of $\mathbf{\Sigma}_x$.

[19] When y is a scalar, $\text{Cov}(\mathbf{x}, y)$ is a vector and its eigen-decomposition is degenerate returning the vector itself, however when the response is multivariate, PLS entails a proper eigen-decomposition of the matrix $\text{Cov}(\mathbf{x}, \mathbf{y})$.

[20] In this section we treat population objects therefore the PLS algorithm we discuss is the *population PLS algorithm* studied by Helland (1990) [49].

[21] The matrix $\mathbf{W}_s = (\mathbf{w}_1, \ldots, \mathbf{w}_s)$ is obtained after s 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 $s > 1$,

$$\mathbf{w}_s = \boldsymbol{\sigma}_{xy} - \mathbf{\Sigma}_x\mathbf{W}_{s-1}(\mathbf{W}_{s-1}'\mathbf{\Sigma}_x\mathbf{W}_{s-1})^{-1}\mathbf{W}_s'\boldsymbol{\sigma}_{xy}$$

generates the subsequent weights. The latter are "weighted" covariances of the predictors and the response. Helland (1988) [48] showed that the matrix of weights $\mathbf{W}_s$ spans the Krylov subspace $\mathcal{K}_s(\mathbf{\Sigma}_x, \boldsymbol{\sigma}_{xy}) = \text{span}\{\boldsymbol{\sigma}_{\mathbf{x}\mathbf{y}}, \mathbf{\Sigma}_x\boldsymbol{\sigma}_{\mathbf{x}\mathbf{y}}, \ldots, \mathbf{\Sigma}_x^{s-1}\boldsymbol{\sigma}_{\mathbf{x}\mathbf{y}}\}$ and subsequent work showed that the PLS solution to the normal equations is the minimum norm solution in such subspace.

4.5 Sliced Inverse Regression (SIR and RSIR)

Interestingly also for the SIR estimator one can show that the estimates of the coefficients of the linear combinations of the predictors are solutions to a maximization problem in the same vein as OLS and the other dimension reduction methods reviewed so far. The SIR components $\boldsymbol{\beta}_i'\mathbf{x}$, are the solution to the maximization problem

$$\max_{\substack{\{\boldsymbol{\beta}_i\} \\ \{\boldsymbol{\beta}_i'\boldsymbol{\Sigma}_x\boldsymbol{\beta}_j=0\}_{j=1}^{i-1}}} \text{Corr}^2(\text{E}(\mathbf{x}'\boldsymbol{\beta}_i|y), \mathbf{x}'\boldsymbol{\beta}_i), \tag{4.8}$$

which obtains a set of directions $\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_p$ satisfying $\text{cov}(\boldsymbol{\beta}_i'\mathbf{x}, \boldsymbol{\beta}_j'\mathbf{x}) = 0$, $i \neq j$, and $\text{cov}(\boldsymbol{\beta}_i'\mathbf{x}, \boldsymbol{\beta}_i'\mathbf{x}) = 1$ (see Theorem 6.1 on p. 62 in Li (2000) [58]). Under the normal DGP, $\text{E}(\mathbf{x}'\boldsymbol{\beta}|y) = \boldsymbol{\mu}_x'\boldsymbol{\beta} + \boldsymbol{\sigma}_{xy}'\boldsymbol{\beta}(y - \mu_y)/\sigma_y^2$ is a linear function of y. Therefore, $\text{Corr}^2(\text{E}(\mathbf{x}'\boldsymbol{\beta}_i|y), \mathbf{x}'\boldsymbol{\beta}_i) = \text{Corr}^2(y, \mathbf{x}'\boldsymbol{\beta}_i)$ and the maximization problem of population SIR returns the OLS estimator capturing the linear signal $\boldsymbol{\sigma}_{xy}$ entirely, just as PLS does. The solution to (4.8) is obtained by solving the generalized eigenvalue problem $\text{Var}(\text{E}(\mathbf{x}|y))\mathbf{v}_j = \lambda_j\boldsymbol{\Sigma}_x\mathbf{v}_j$. In practice, when $\boldsymbol{\Sigma}_x$ is ill-conditioned, the SIR generalized eigenvalue problem is unstable resulting highly variable SIR components (although not as much as OLS in our empirical findings). In the empirical application, we use regularized SIR (RSIR) that substitutes $\boldsymbol{\Sigma}_x$ with an approximation based on many principal components.

4.6 Comparison of Estimator Families

Table 1 summarizes the estimators in this paper. It reports the meta parameter indexing each family of estimators, the target based "signal", that is the statistic encapsulating the relationship of the target y with $\mathbf{x}$, and the scaling matrix of the signal for each estimator. It also reports the corresponding parameter each estimator uses in the prediction along with its eigen-representation.

Table 1: Summary of Estimators and Their Solutions.

	Meta Parameter	*Inverse*	*Signal*	$\boldsymbol{\beta}$	*Eigen-decomposition*
OLS	–	$\Sigma_{\mathbf{x}}^{-1}$	$\boldsymbol{\sigma}_{\mathbf{x}y}$	$\Sigma_{\mathbf{x}}^{-1}\boldsymbol{\sigma}_{\mathbf{x}y}$	$\sum_i^p \frac{1}{\lambda_i}\mathbf{u}_i\mathbf{u}_i'\boldsymbol{\sigma}_{\mathbf{x}y}$
RIDGE	κ	$(\Sigma_{\mathbf{x}} + \kappa\mathbf{I})^{-1}$	$\boldsymbol{\sigma}_{\mathbf{x}y}$	$(\Sigma_{\mathbf{x}} + \kappa\mathbf{I})^{-1}\boldsymbol{\sigma}_{\mathbf{x}y}$	$\sum_i^p \frac{1}{\lambda_i+\kappa}\mathbf{u}_i\mathbf{u}_i'\boldsymbol{\sigma}_{\mathbf{x}y}$
PCR	m	$\Sigma_{\mathbf{x}}^{-}(m)$	$\boldsymbol{\sigma}_{\mathbf{x}y}$	$\Sigma_{\mathbf{x}}^{-}(m)\boldsymbol{\sigma}_{\mathbf{x}y}$	$\sum_i^m \frac{1}{\lambda_i}\mathbf{u}_i\mathbf{u}_i'\boldsymbol{\sigma}_{\mathbf{x}y}$
PLS	s	$\Sigma_{\mathbf{x}}^{-}(s, y)$	$\boldsymbol{\sigma}_{\mathbf{x}y}$	$\Sigma_{\mathbf{x}}^{-}(s, y)\boldsymbol{\sigma}_{\mathbf{x}y}$	$\sum_i^s \frac{1}{\lambda_i}\mathbf{u}_i^*\mathbf{u}_i^{*\prime}\boldsymbol{\sigma}_{\mathbf{x}y}$
SIR	d	$\Sigma_{\mathbf{x}}^{-1}$	$\Sigma_{\text{E}(\mathbf{x}\|y)}(d)$	$\Sigma_{\mathbf{x}}^{-1}\Sigma_{\text{E}(\mathbf{x}\|y)}(d)$	$\Sigma_{\mathbf{x}}^{-1}\sum_i^d \widetilde{\lambda}_i\widetilde{\mathbf{u}}_i\widetilde{\mathbf{u}}_i'$
RSIR	d	$\Sigma_{\mathbf{x}}^{-}(m)$	$\Sigma_{\text{E}(\widehat{\mathbf{x}}\|y)}(d)$	$\Sigma_{\mathbf{x}}^{-}(m)\Sigma_{\text{E}(\widehat{\mathbf{x}}\|y)}(d)$	$\sum_i^m \frac{1}{\lambda_i}\mathbf{u}_i\mathbf{u}_i'\sum_i^d \widetilde{\lambda}_i\widetilde{\mathbf{u}}_i\widetilde{\mathbf{u}}_i'$

The covariance $\boldsymbol{\sigma}_{xy}$ is the signal relating target and explanatory variables for OLS, RIDGE, PCR and PLS. What differentiates these methods is the choice of the inverse in the scaling matrix. Since in practice problems arise from the multicollinearity and ill-conditioning of the design matrix $\boldsymbol{\Sigma}_{\mathbf{x}}$ in OLS, the alternative estimators offer remedies to the "big data" problem by adopting alternative scaling matrices. The difference among estimators boils down to the choice of the generalized inverse that approximates $\Sigma_{\mathbf{x}}^{-1}$ and of the meta-parameter that indexes them.

Specifically, RIDGE approximates $\Sigma_{\mathbf{x}}^{-1}$ "in excess" by inflating its main diagonal with $(\Sigma_{\mathbf{x}} + \kappa\mathbf{I})^{-1}$. The eigen-decomposition of the RIDGE parameters in Table 1 shows all eigenvalues are retained (the summation is up to p) but the weight of eigenvectors associated to smallest eigenvalues is shrunk to zero. By contrast, both PCR and PLS use truncation. PCR deletes eigenvalues close to zero, similar to PLS. However, PLS also makes use of the target and deletes all eigenvalues that are orthogonal to the signal as quantified by $\boldsymbol{\sigma}_{\mathbf{x}y}$.

SIR distinguishes itself from the rest by adopting an altogether different and potentially richer signal, the covariance of the expected value of $\mathbf{x}|y$ and truncates excessive noise with the choice of d. In addition, by replacing the covariance by the expected inverse mean, $\mathrm{E}(\mathbf{x}|y)$, SIR can identify non-linear signal if present.

RSIR is a modified SIR estimator in the absence of abundant data, as in our empirical application. RSIR replaces the raw explanatory variables $\mathbf{x}$ with a subset of principal components, denoted by $\widehat{\mathbf{x}}$ in the table.

4.7 Target Signal in Eigen-Decompositions

To gain further intuition on why SIR has the potential of being more accurate, we resort to the classical variance decomposition satisfied by any random vector $\mathbf{x}$ with finite second moments and conditioning random variable, or vector, y,

$$\underbrace{\mathrm{Var}(\mathbf{x})}_{\text{targeted in PCA}} = \underbrace{\mathrm{Var}[\mathrm{E}(\mathbf{x}|y)]}_{\text{targeted in linear SDR}} + \underbrace{\mathrm{E}[\mathrm{var}(\mathbf{x}|y)]}_{\text{noise}} \tag{4.9}$$

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 variance identity (4.9) reveals that the variance of $\mathbf{x}$ can be split into two parts: 1) $\mathrm{Var}[\mathrm{E}(\mathbf{x}|y)]$ or *between* slice variation in $\mathbf{x}$, and 2) $\mathrm{E}[\mathrm{Var}(\mathbf{x}|y)]$ or *within* slice variation. In analysis of variance, $\mathrm{Var}[\mathrm{E}(\mathbf{x}|y)]$ is the *signal* that $\mathbf{x}$ carries about y since it represents variation of the average value of $\mathbf{x}$ associated with different values of y from the overall $\mathbf{x}$ mean,

whereas E[Var($\mathbf{x}|y$)] represents *noise*, i.e. deviations of $\mathbf{x}$ from its overall average across bins, hence unrelated to y.

Since PCA performs an eigenanalysis of Var($\mathbf{x}$), the noise in E[Var($\mathbf{x}|y$)] may attenuate or suppress the signal in Var[E($\mathbf{x}|y$)] and result in PCs that are weakly related to y. PLS targets Cov($\mathbf{x}, y$) and can potentially suppress non-linear signal. By contrast, a method that focuses on the eigen-analysis of 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. Centering on the signal and ignoring the noise is what sufficient dimension reduction is designed to do.

5 Empirical Application

After reviewing data and methodology, this section contains a classical pseudo out-of-sample (OOS) horserace among the estimators defined in Sections 4.5 and 4. We examine their accuracy and robustness against AR(4), a simple and parsimonious process.

5.1 Data and Out-of-Sample Forecasting Exercise

The first step in our empirical implementation is choosing the macro-panel on which to run our experiment from the rather rugged landscape of data sources used in some of the most important studies in the DFM macro-forecasting literature.

FRED-MD – The multitude of data sources and data vintages used in the literature has hindered comparability across studies in macro-forecasting. Many forecast studies have a core set of variables in common, thought to capture the bulk of ongoing macro activity across varying periods. However, there is no unanimity regarding other details such as the specific set of non-core variables and most importantly the data vintage. An initiative spearheaded by McCracken and Ng, and documented in McCracken and Ng (2015) [60], has set about to impose some discipline on the current and future production of macro-forecasting studies. An outcome of their project has been the creation of the macro-panel called FRED-MD.[22]. We embrace their initiative and adopt FRED-MD as our dataset of choice. FRED-MD contains a balanced panel of 132 variables with monthly data from January 1960 (1960m1) resulting in a macro-panel covering more then 50 years.[23] The dataset

[22]FRED-MD is updated in real time by the same staff that maintains the popular FRED database. The data can be dowloaded from Michael McCracken's website at the St. Louis Fed

[23]FRED-MD has fewer variables than the quarterly dataset of 144 variables used by Stock and Watson in [74], the most widely used dataset in quarterly studies. It has also fewer variables than the dataset of 143 variables used by

is described in McCracken and Ng (2015) [60] along with a discussion of some data adjustments needed to construct the panel and a useful chronology and summary of the main alternative macro-panels that have been used in the DFM literature. We choose to work with monthly data since the companion quarterly dataset FRED-QD is not available yet and our SDR forecasting procedure is data intensive. A shortcoming of the dataset is that core PCE inflation and non-farm payroll employment, two of the most watched series by forecasters and Federal Reserve staff have not been included yet. FRED-MD contains very limited real-time vintages making real-time forecasting unfeasible at the moment. Since mid-2016, the ISM requested to discontinue the publication of their data, widely used as leading indicators of productions, therefore we use the 2016-05 vintage of FRED-MD, the last vintage to contain the ISM data. Moreover given the intent of creating a balanced dataset starting in 1960, the authors had to exclude important variables routinely used in small scale or "judgemental" OLS regressions to forecast target variables of interest to policy makers and market practitioneers alike.

Handling of missing data in FRED-MD – Five variables in the dataset have a large number of missing data. Rather than running an EM algorithm to fill in the missing values and achieve a balanced panel, as done by McCraken and Ng (2015) [60] and Stock and Watson (2002b) [71], we prefer to exclude them avoiding to bias the sample in favor of PCR. The five excluded variables are: new orders for consumer goods (`ACOGNO`), new orders for nondefense capital goods (`ANDENOx`), crude oil, spliced WTI and cushing (`OILPRICE`), trade weighted U.S. dollar index: major currencies (`TWEXMMTH`) and consumer sentiment index (`UMCSENT`). We do not apply any filter for outliers.

Forecast Targets – In principle, any variable in FRED-MD can be used as target. We focus on a small set of variables that are closely watched and forecasted by monetary authorities and professional forecasters. The FED dual mandate implies that inflation and labor market measures are closely monitored. In the absence of Core PCE inflation in FRED-MD, CPI inflation (`CPIAUCSL`) is the most natural candidate. Key labor market variables in FRED-MD are total non-farm payrolls (`PAYEMS`), the unemployment rate (`UNRATE`) and civilian labor force participation (`CLF16OV`); together these three variables capture the three margins of labor utilization and slack that have been featured the most in discussing the timing of the next monetary policy tighten-

Stock and Watson in [76]. The latter is a quarterly study but the dataset posted by Mark Watson contains variables observed monthly. The FRED-MD dataset also contains fewer variables than the 149 regressors used by Stock and Watson in [70], or the 215 series used by Stock and Watson in [71] from the **DRI/McGraw-Hill Basic Economics** database, formerly named **Citibase**.

ing cycle. Wage growth is another important indicator of slack; we choose to forecast average hourly earnings in goods-producing industries (`CES0600000008`), the most top-level aggregate for wages in FRED-MD. We also forecast industrial production (`INDPRO`), a classical measure of real activity available at monthly frequency. Finally, we include real personal consumption expenditures (`DPCERA3M086SBEA`), a key component in tracking models of GDP and one of its main determinants, real personal income (`RPI`).

Data Transformations – We adopt the transformations suggested by McCracken and Ng (2015) [60] and coded in the second row of the original downloaded dataset. 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. The transformation of the target variable dictates the forecast target. For example, in the case of inflation, a variable marked as I(2) and transformed as

$$y_t = \Delta^2 \log(CPI_t),$$

we generate the target

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

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

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

and the resulting target is

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

The Pseudo OOS Forecasting Scheme – We conduct a standard out-of-sample (OOS) forecasting experiment with a recursive window at horizons $h = 1, 3, 6, 12, 24$.[24]. These are relevant horizons in practice and allow exploration of possible variation across horizons within each forecasting method. As is common in the literature, we adopt h-step ahead regression rather than iterated.

5.2 Normality and Ellipticity Tests

Proposition 1 and the accompanying discussion brought attention to the importance of the distribution of the predictors.

[24]The transformed and aligned data are available on request.

Tests of Joint Normality – Joint normality of the explanatory variables $\mathbf{x}$ emerged as sufficient for both (LC) and (CVC) conditions in Proposition 1 to be satisfied for any $\boldsymbol{\beta}$. The joint distribution of FRED-MD variables, *after transformations*, appears to be far from normal, failing standard tests as reported in Table 2. Moreover, the lack of normality cannot be attributed to the presence of outliers. The result is not surprising considering that our panel also includes financial variables that are known to have "fat tails."

Table 2: Tests for Joint Multivariate Normality and Ellipticity in FRED-MD.

(a) Normality Tests.

Mardia mSkewness = 5497.98	p-value = 0.000
Mardia mKurtosis = 18381.1	p-value = 0.000
Doornik-Hansen	p-value = 0.000

(b) Ellipticity Test.

Q Statistic = 118.63
chi2(122) = 148.78
p-value = .569

Ellipticity Test – Predictor ellipticity ensures the linearity condition (LC) or, equivalently, the linear design condition (3.3) is satisfied. The elliptically contoured family contains fat tailed distributions. We performed the multivariate elliptical symmetry test using the semiparametric rank-based procedure as proposed by Cassart (2007) [19] and implemented by Verardi and Croux (2009) [78]. We carried out the test across all the time periods and variables in FRED-MD that we consider in our forecasting experiment. The test has p-value of 0.569 and does not reject the null hypothesis of ellipticity. In consequence, since Corollary 1 applies, we expect the estimators, aside from sample variation, to perform similarly with respect to predictive accuracy.

5.3 Estimation Details

The implementation of the estimation methods discussed in this paper necessitates practical choices that we present herein along with some sensitivity analyses of the results.

OLS and Judgmental OLS – OLS denotes the linear regression model with inputs all the non-target variables in FRED-MD, the contemporaneous value of the target plus its first four lags. We did not consider predictors available outside of FRED-MD that have routinely been employed by practitioners, for instance measures of flows in and out of the labor market, since usually they are not available starting in 1960. Forecast accuracy of OLS is expected to be subpar even though in some rare instances OLS performs better than AR(4). *JOLS* denotes linear regression on the

contemporaneous value of the target and its first four lags plus a very small number of "indicators" judgmentally chosen from FRED-MD in accordance to conventional wisdom followed by forecast practitioners.[25] *JOLS* does not perform well in our experiments, however this could well be due to the restriction of picking the indicators within FRED-MD.

RIDGE – The RIDGE regularization parameter κ needs to be chosen prior to estimation.[26] DeMol et al. (2008) [33] fit a grid of κ values and report MSFE for all. We have included those values in our implementation of RIDGE and we denote such RIDGE estimators with *RIDGE.*κ, where κ is the fixed scalar used. *RIDGE.min* denotes the RIDGE estimator in which κ has been chosen to minimize the cross-validation (CV) mean square error. *RIDGE.1se* denotes the RIDGE estimator in which κ is chosen one standard deviation away from the minimum of the CV errror, an option that works well in some applications (see [42] for more details).[27]

PCR – The PCR forecasting models are linear regressions on a varying number of PCs and the AR(4) component (contemporaneous plus four lags of the target). They differ by the selection of the PCs at each OOS forecasting step. First, we follow Stock and Watson (2002) [71] and estimate a series of PCRs (also known as diffusion indexes) with a constant number of PC's (m) throughout the forecasting experiment. We considered $m = 1, ..., 30$ generating models denoted *PC.1* through *PC.30*. We also considered a host of PC selection criteria, collectively referred to panel information criteria (ICp), proposed in the DFM literature, including PCp1, PCp2, ICp1 and BIC3 proposed by Bai and Ng [4], the criterion proposed by Onatski [64], and the ER and GR criteria proposed by Anh and Horenstein [1] (these are models *PC.PCp1* through *PC.GR*).[28]

PCR with Best Subset Selection – The above criteria for PC selection have one characteristic

[25]The chosen indicators for `INDPRO` are ISM new orders (`NAPMNOI`), a traditional leading indicator of manufacturing production, as well as an index of inventories (`ISRATIOx`), durable goods orders (`AMDMNOx`), unemployment claims (`CLAIMSx`) and the BAA corporate bonds spread (`BAAFFM`). Unemployment claims (`CLAIMSx`) are used to form *JOLS* for total private payrolls (`PAYEMS`) and retail sales (`RETAILx`), real personal income (`RPI`) and the S&P500 index (`S.P.500`) are used in the univariate regression for consumption (`DPCERA3M086SBEA`). Explanatory variables for the unemployment rate (`UNRATE`) were unemployment claims and mean unemployment duration (`UEMPMEAN`).

[26]To implement RIDGE we used the `R` package `glmnet` by Friedman et al. (2015) [42].

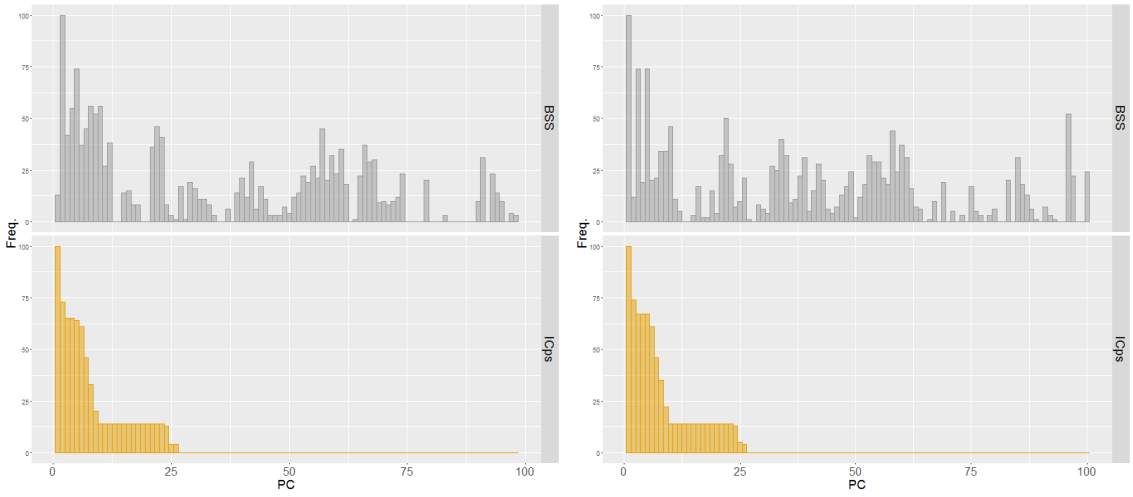
[27]In some plots *RIDGE.min* and *RIDGE.1se* are collectively denoted as *RIDGE.CV* to enhance clarity.

[28]Although we can reproduce the results in McCracken and Ng (2015) [60], where the Bai-Ng criteria are found to select about 8 to 10 components, especially PCp2 is very sensitive to the maximum number of allowed components and it can very quickly call for the inclusion of 50 or more PC. Other criteria are more stable. However they all seem to require more components as the great recession enters into the forecasting window. Other PCR specifications always deliver better forecasting results.

in common: if the k^{th} PC is included also all prior PCs, from the 1^{st} to the $(k-1)^{th}$, are included. SDR and PLS produce a re-ordering of the PCs according to the particular target at hand, with the potential of selecting information in the panel encapsulated in PCs further down the tail. For this reason, we also applied best subset selection (BSS) on all PCs in the panel. We discovered that BSS often selects PCs in the tail of the distribution, as far as the 94^{th} component, and that, in general, variants of BSS perform better than the DFM information-based selection criteria discussed above.[29] BSS requires the specification of a criterion to evaluate models: we experimented with the default Mallows *Cp* and *BIC* but also considered *BIC3* and *PCp1*. A typical pattern across horizons, targets and sub-samples is portrayed in Figure 1, where for target `INDPRO` at horizon $h = 12$ the average frequency of selection of each PC over the OOS experiment by BSS using *Cp*, *BIC*, *BIC3* and *PCp1* (top portion of panel 1a) is contrasted with the corresponding average frequency across all the ICp criteria we tested (bottom portion of panel 1a). Both portions of panel 1a experience a sharp drop at about 10 components. However, BSS is free to select PCs far out in the tail and it does so frequently indicating that signal may be better extracted by targeted methods.[30] Regarding `CPIAUCSL` in panel 1b BSS is very selective also for the first 10 leading PCs, selecting PC2 rarely whereas it selects PC out in the tail, such as PC99 very often.

Figure 1: Av. Num. of Comp. Selected by BSS versus ICps Criteria Over 1992-2010.

(a) Target `INDPRO` at h=12. (b) Target `CPIAUCSL` at h=12.



[29] We implemented BSS using the `leaps` R-package by Thomas Lumley (2009) [59] with the backward search option starting from large models.

[30] We also used CV to select PCs in unreported results as its performance was largely dominated by the other approaches.

PLS – The forecasting equation for PLS consists of a linear regression on a baseline AR(4) augmented with a fixed number s of PLS components, for $s = 1, \ldots, 30$.[31] We also selected s with cross-validation, however do not report the results as the forecasting performance was not competitive.[32] The superior targeting nature of PLS relative to PCR in sample is manifested by 10 PLS components explaining about 70% of the variance in `CPIAUCSL`, and 60% of the variance in `INDPRO`, as compared to 33% and 40%, respectively, explained by 10 PCs. Moreover, PLS explains a larger fraction of the variance in the target with no adverse consequences on the variance of the whole panel it is able to capture. However, we see in Section 5.6 that the out-of-sample forecasting performance of PCR often dominates PLS.

SIR – We have implemented two versions of SIR [33] in Section 4.5: *SIRa* refers to SIR components derived from the inverse regression of $\mathbf{x}_t$ on y_{t+h}, and *SIRb* from the inverse regression of $\mathbf{x}_t$ on y_t. Once the SIR components are estimated, the target is regressed via OLS on the AR(4) component and a fixed number of SIR components (from 1 to 8). The number of SIR components in cross-sectional i.i.d. data can be estimated with the asymptotic weighted-chi squared test (Bura and Cook (2001b) [14]), or the permutation test in Cook and Yin (2001) [32]. However both tests proved to be very unstable. We empirically verified that a dimension larger than two is rarely beneficial in this forecasting exercise.

Regularized SIR – For SIR to be effective a substantial amount of data is required, at least $T > 5p$, or 5 observations per predictor.[34] Our estimation sample is small so we supplemented the standard SIR estimator with *RSIR,* a regularized form of SIR (see Appendix 6 for more details). The SIR predictors were computed from the inverse regression of the first 30 or 70 principal components. For most samples, 70 PCs explain about 80-90% of the variability in the predictors, so that not much information on the conditional predictive density of $y_{t+h}|x_t$ is lost and SIR on the reduced data can still identify and estimate part of the SDR subspace. RSIR is a two-step procedure that can be interpreted as a form of weighted PCR, with the weights being approximately informed by the conditional distribution of $y_{t+h}|x_t$.[35]. Fan, Xue and Yao (2015) [39] also apply SIR on PCs although their approach is conceptually and practically distinct. First they estimate the number

[31] We implemented PLS using the `pls` `R`-package in [61].

[32] We did not experiment with the cross-validation criteria recently reviewed by Carrasco and Rossi [18], neither did we select the PLS components with criteria suggested by Krämer and Sugiyama (2011) [54].

[33] We have implemented SIR using the `R`-package `dr` by Weisberg [80].

[34] Ideally, the sample size satisfies $T > 10p$.

[35] Alternative weighting schemes were proposed by Boivin and Ng (2006) [10]

of PCs using standard ICps criteria, then they apply SIR to condense them in fewer components. Their approach is rooted in the DFM assumption that the predictors and the response are driven by the same unobserved common factors, which are identified by the first few ordered PCs and shared by both the response and the observed predictors. Such assumption, as we argued earlier on, not only can it not be formally checked but also ignores potentially useful modeling information in $\mathbf{x}$ about y. In contrast, we use PCA as a data pre-processing tool, necessitated by the practical limitation that SIR either cannot be applied ($p > T$), or is unstable ($p \approx T$), and retain as many PCs as practicable so that as little as possible information in $\mathbf{x}$ is sacrificed. RSIR is the only SDR option in samples where the OOS experiment begins early and it often performs well in larger samples as well. The notation we use for the different SIR–based models is summarized in Table 3.

Table 3: SIR–based Forecasting Models

Type of SIR	*SIR predictors*	*Forward Model*
SIRa	X on y_{t+h}	y_{t+h} on y_t, 4 lags of y_t and SIR predictor(s)
SIRb	X on y_t	y_{t+h} on y_t, 4 lags of y_t and SIR predictor(s)
R8SIRa	First 8 PCs on y_{t+h}	y_{t+h} on y_t, 4 lags of y_t and SIR predictor(s)
R30SIRa	First 30 PCs on y_{t+h}	y_{t+h} on y_t, 4 lags of y_t and SIR predictor(s)
R70SIRa	First 70 PCs on y_{t+h}	y_{t+h} on y_t, 4 lags of y_t and SIR predictor(s)

SIR and Non-Linearities – By drastically reducing the number of predictor components required to model the target, SDR allows accommodating forecasting models that are nonlinear in the components. We fitted forecasting models linear in the AR(4) component and nonparametric in the SIR component(s).[36] The forecasting accuracy results do not support non-linearities in the **conditional mean** of the target variable. Plots of the residuals of the fitted forecasting model versus the second SIR component indicate that the variance of the forecasting models varies for most targets, so that the signal in the second SIR predictor is contained in the conditional variance of the targets (see Bura and Forzani (2015)). We did not investigate the issue further as it is unclear how to model the non-constant variance in an automated fashion over the OOS forecasting experiment. Nevertheless, the finding that the conditional mean of the target does not contain non-linearities in the predictors is important as it explains why SIR linear and non-linear regressions

[36]Specifically we relied on the model by Robinson (1988) [67] with cross-validation as in Racine and Li (2004) [66]. The `R`-package `np` developed by Racine and Hayfield (2014) [65] was used.

do not have a clear forecasting advantage over the other estimators discussed in this paper, except that of greatly reducing the number of required components.

5.4 Overview of Forecasting Performance

The mean squared forecast error (MSFE) relative to the baseline AR(4) is used as a measure of forecasting performance throughout.[37]

No Estimator Family Has an Edge – Overall we found that no single family of estimators has a universal edge in forecasting our targets. This is in agreement with the finding that the predictors are jointly elliptically contoured in Section 5.2 and the discussion at the end of Section 5.3. Table 4 reports the best estimators among all estimators discussed in this paper for all targets over 5 horizons and 3 of the subsamples that we considered[38] and offers a bird's eye view of our many forecast experiments. Table 4 reveals there is no clear "winner" estimator or estimator family. In some instances, an estimator family is superior to the competitors at forecasting a specific target over a specific subsample. For example, RIDGE is a particularly effective method to forecast `PAYEMS` in the recent recovery as evident from the bottom portion of Table 4. However, this is an exception.

Dependence of Results on Sample – The forecasting performance greatly depends on the window of the sample it is based on as well as the target. For instance, the inclusion of the "great recession" imparts degradation in accuracy for all estimators. In order to study the robustness of our estimators we also report RMSFE over rolling windows in Section 5.7.

Overall Performance of SIR and RSIR – From Table 4 we can see that regularized SIR frequently is the best performing estimator for a given target-horizon pair. Regularization even on the first 8 leading PCs appears to be working quite well, although most often more PCs are needed. In fact, we found that a larger number of PCs is a more robust choice when forecasting at longer horizons ($h = 12$ and $h = 24$), suggesting that important information resides in the last PCs that explain smaller fractions of predictor variance. Most frequently, RSIR summarizes in one or two components the information encapsulated in 8, 30 or 70 PCs. SIR on the raw data, without preconditioning appears in the table just a couple of times. This is expected since SIR is data intensive and the size of the sample does not allow to exploit the full potential of standard SIR.

[37] We obtained broadly similar results, which we do not report, using the mean absolute forecast error.

[38] We considered subsample (a) as it excludes the two last most severe recessions. Subsample (b) includes the "great recession" and subsample (c) only includes the recent recovery.

Targeted vs Untargeted Estimators – Overall, we found that targeted estimators require fewer components than untargeted ones for a given level of accuracy. In particular, most SIR–type estimators in Table 4 make use of only one component to deliver the most efficient form of targeting. In a few cases, the second SIR predictor appears to capture some signal, which may be a byproduct of unmodeled heteroskedasticity in the forecasting regression model.

Forecasting Performance by Target – The forecasting performance of the estimators considered in this paper varies with the target. In general, beating AR(4) is a difficult task. Prices and wages appear to be very difficult to forecast, a well-known fact in the forecasting literature. The same is true for real consumption (`DPCERA`), further justifying recent efforts to use "big data" sources, such as payments and credit card data, in trying to improve the forecast of such an important variable (70% of GDP). A notable exemption is RIDGE that is particularly effective in forecasting `PAYEMS` over the recovery, whereas PCR appears to have an edge in forecasting wages at longer horizons. For labor force participation, real personal income, and, to a somewhat lesser extent, the unemployment rate are far easier to improve upon their forecast relative to AR(4).

Table 4: Best Estimators by Target and by Forecast Horizon (MSFE Relative to AR4).

(a) Recursive Out-of-Sample Window from 1992 to 2007.

	Forecast Horizon									
	h=1		*h=3*		*h=6*		*h=12*		*h=24*	
Target	*Estimator*	*MSFE*	*Estimator*	*MSFE*	*Estimator*	*MSFE*	*Estimator*	*MSFE*	*Estimator*	*MSFE*
INDPRO	RIDGE[b]-119	0.87	RIDGE[b]-141	0.95	R70SIRb[c,d]-2	0.95	R8SIRb[c,d]-2	0.9	R8SIRb[c,d]-2	0.9
PAYEMS	R30SIRb[c,d]-1	0.96	R30SIRb[c,d]-1	0.94	R30SIRb[c,d]-2	0.87	R30SIRb[c,d]-2	0.82	R30SIRb[c,d]-2	0.75
UNRATE	PC-1	0.93	PLS-1	0.76	RIDGE[b]-0.4	0.76	RIDGE[b]-0.3	0.78	PC-4	0.88
CLF16O	RIDGE[b]-141	0.94	PC-16	0.87	PC-23	0.81	PC-16	0.74	PC-16	0.68
CPIAUC	PC.BS[h]-8.7	0.89	SIRb[d]-2	0.94	PC.ON[f]-1.2	0.97	R8SIRb[c,d]-2	0.95	R70SIRb[c,d]-2	0.99
CES060	PLS-1	0.99	PLS-1	1	R8SIRb[c,d]-1	0.97	AR4	1	AR4	1
DPCER	AR4	1	AR4	1	AR4	1	AR4	1	PCF.BS[i]-9	0.92
RPI	R8SIRb[c,d]-1	0.93	R8SIRb[c,d]-1	0.88	R8SIRb[c,d]-1	0.88	RFSIRb[c,d]-1	0.83	SIRb[d]-2	0.88

(b) Recursive Out-of-Sample Window from 1992 to 2016.

	Forecast Horizon									
	h=1		*h=3*		*h=6*		*h=12*		*h=24*	
Target	*Estimator*	*MSFE*	*Estimator*	*MSFE*	*Estimator*	*MSFE*	*Estimator*	*MSFE*	*Estimator*	*MSFE*
INDPRO	PC-15	0.95	PC-15	0.94	SIRb[d]-1.4	0.96	R70SIRb[c,d]-3	0.94	PC-16	0.89
PAYEMS	R8SIRb[c,d]-1	0.88	R8SIRb[c,d]-1	0.82	R8SIRb[c,d]-1	0.89	R30SIRb[c,d]-2	0.9	PC-26	0.81
UNRATE	PC-15	0.9	PC-3	0.74	PC-3	0.73	PC-3	0.76	PC-3	0.8
CLF16O	PC-11	0.88	PC-17	0.69	PC-20	0.56	PC-17	0.41	PC-16	0.34
CPIAUC	PLS-3	0.9	RFSIRb[c,d]-5.1	0.91	SIRb[d]-4	0.97	R8SIRb[c,d]-1	0.99	R8SIRb[c,d]-1	0.99
CES060	AR4	1	SIRb[d]-1	0.99	R8SIRb[c,d]-1	0.94	PC-1	0.92	PC-3	0.82
DPCER	PC.ON[f]-1.5	0.93	PC-1	0.92	RFSIRb[c,d]-1	0.94	R8SIRb[c,d]-1	0.94	R8SIRb[c,d]-2	0.95
RPI	PLS-1	0.91	PLS-1	0.85	PLS-1	0.82	PLS-1	0.8	R8SIRb[c,d]-2	0.87

(c) Recursive Out-of-Sample Window from 2010 to 2016.

	Forecast Horizon									
	h=1		*h=3*		*h=6*		*h=12*		*h=24*	
Target	*Estimator*	*MSFE*	*Estimator*	*MSFE*	*Estimator*	*MSFE*	*Estimator*	*MSFE*	*Estimator*	*MSFE*
INDPRO	RIDGE[b]-949	0.96	RIDGE[b]-949	0.93	RIDGE[b]-3532	0.94	R30SIRb[c,d]-3	0.74	R70SIRb[c,d]-6	0.39
PAYEMS	RIDGE[b]-141	0.74	RIDGE[b]-141	0.56	RIDGE[b]-288	0.5	RIDGE[b]-288	0.44	R8SIRa[c,d]-2	0.43
UNRATE	PCF.BS[i]-11	0.85	PLS-4	0.64	SIRa[d]-7	0.55	PC.BS[h]-1	0.38	PC-3	0.31
CLF16O	PC-7	0.78	PCF.BS[i]-11	0.52	RIDGE[b]-0.6	0.39	PLS-5	0.33	PC-16	0.05
CPIAUC	PC.BS[h]-1	0.96	SIRb[d]-1.3	0.9	R70SIRb[c,d]-3	0.94	PC-5	0.87	PLS-12	0.84
CES060	SIRa[d]-7	0.94	SIRb[d]-1	0.99	R8SIRb[c,d]-3.5	0.94	R8SIRa[c,d]-2	0.87	SIRa[d]-7	0.8
DPCER	PC-5	0.87	PC-5	0.7	PC-5	0.48	PC-17	0.44	R8SIRb[c,d]-3	0.54
RPI	R8SIRa[c,d]-1	0.91	RFSIRa[c,d]-1.6	0.79	PLS-8	0.78	PLS-20	0.64	PLS-21	0.48

[a]After each estimator the number of components used.

[b]The number after RIDGE Comp. is value of regularization param.

[c]R#SIR is regularized SIR. # refers to number of leading PCs used for regularization.

[d]In type–a SIR target is y_{t+h}. In type–b SIR target is y_t.

[e]RFSIR is regularized SIR on most frequently selected PCs by out-of-sample best subset selection.

[f]PC.ON is PCR in which number of components is chosen using Onatski criterion.

[g]PC.ICP1 is PCR in which number of components is chosen using Bai-Ng criterion.

[h]PC.BS is PCR in which number of components is chosen by best subset selection.

[i]PCF.BS is PCR in which number of components is chosen by best subset selection applied on most frequently chosen PCs in out-of-sample.

5.5 Forecasting Performance of Regularized SIR

Figures 2 and 3 plot relative MSFE with respect to AR(4) of each estimator versus the number of components for the eight target variables. Continuous lines denote type–a regularized SIR and dotted lines denote type–b. We observe that, in general, MSFE is increasing as the number of components increases indicating that one or two components deliver the best performance.

Performance of Regularized SIR by Number of Components – In Figures 2 and 3 we see that rarely using 70 PCs in the regularization is beneficial. 30 PCs appear in general to strike a good balance between the difficulty of SIR with large data and small sample size and capturing information in the tail of the PC distribution. The violet lines corresponding to regularized SIR on just a few PCs in the tail of the distribution show it to fare well across all variables, especially at longer forecast horizons. By contrast, in general regularization on just 8 PCs exhibits deteriorating performance at longer horizons. Information in the tail of the PC distribution appears to be important for forecasting at longer horizons.

Performance of Regularized SIR by Type – We can also infer from Figures 2 and 3 that for prices, wages, consumption, income and payrolls, type–b regularized SIR performs best (lowest point of dotted lines above lowest point of continuous lines), whereas for industrial production, the unemployment rate and labor force participation, type–a regularized SIR is best.

Performance of Regularized SIR by Target – Prices (`CPIAUC`), wages (`CES060`) and consumption (`DPCERA`) are traditionally challenging to forecast. SIR also cannot improve much upon the baseline AR(4), as can be seen in Figure 2. Forecasting industrial production (`INDPRO`) is difficult and regularized SIR appears to have a consistent edge on the AR(4) only in the long-run. The variables for which regularized SIR delivers the best results are labor force participation (`CLF160`) and real personal income (`RPI`). In the long-run, the forecast of payrolls (`PAYEMS`) from AR(4) can be improved upon by regularized SIR, whereas most of the forecasting gains appear to happen in the short to medium-run for the unemployment rate (`UNRATE`).

Figure 2: Relative MSFE vs Number of Components for SIR Estimators for Prices, Wages, Real Consumption and Real Income (Out-of-Sample from 1992 to 2016).

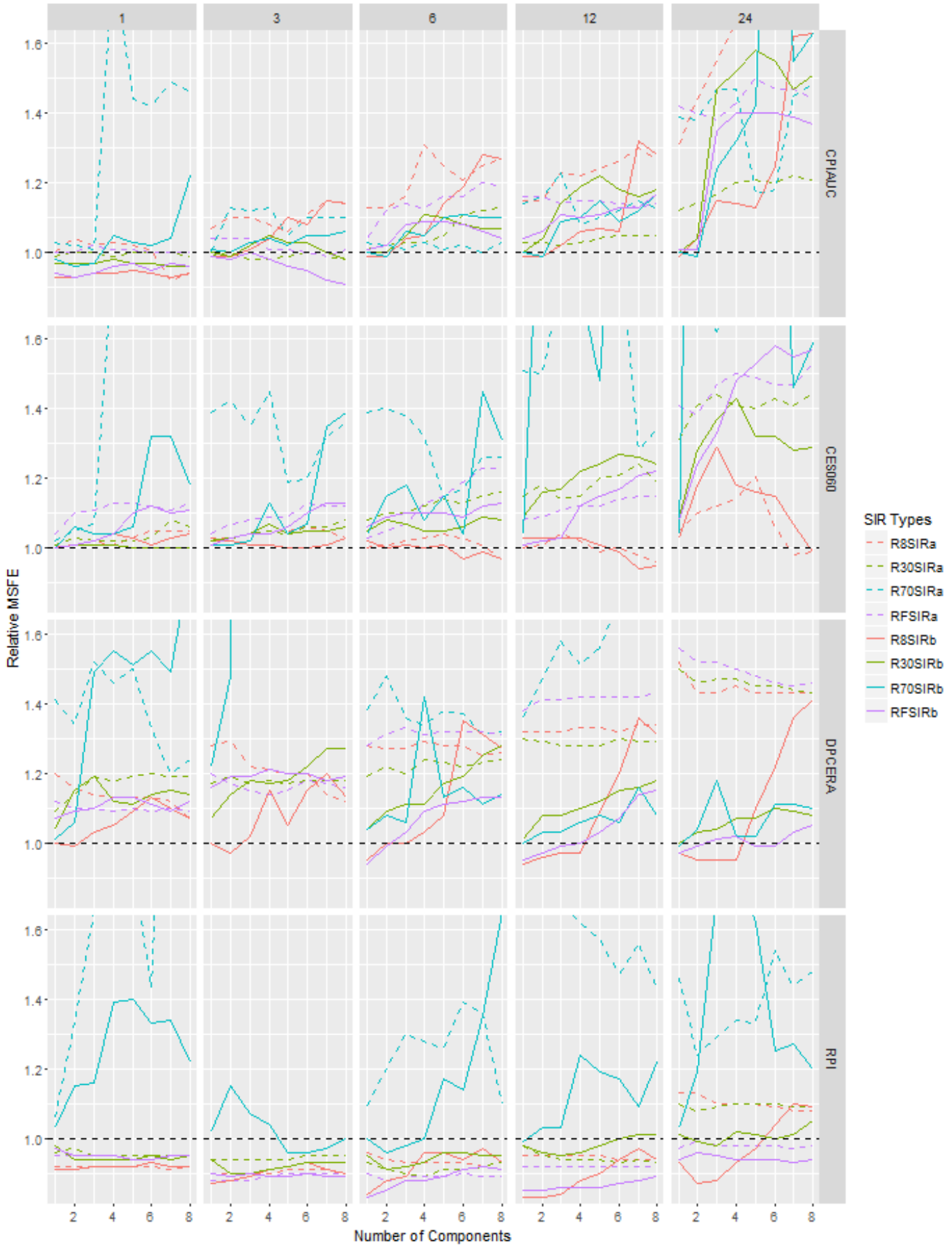
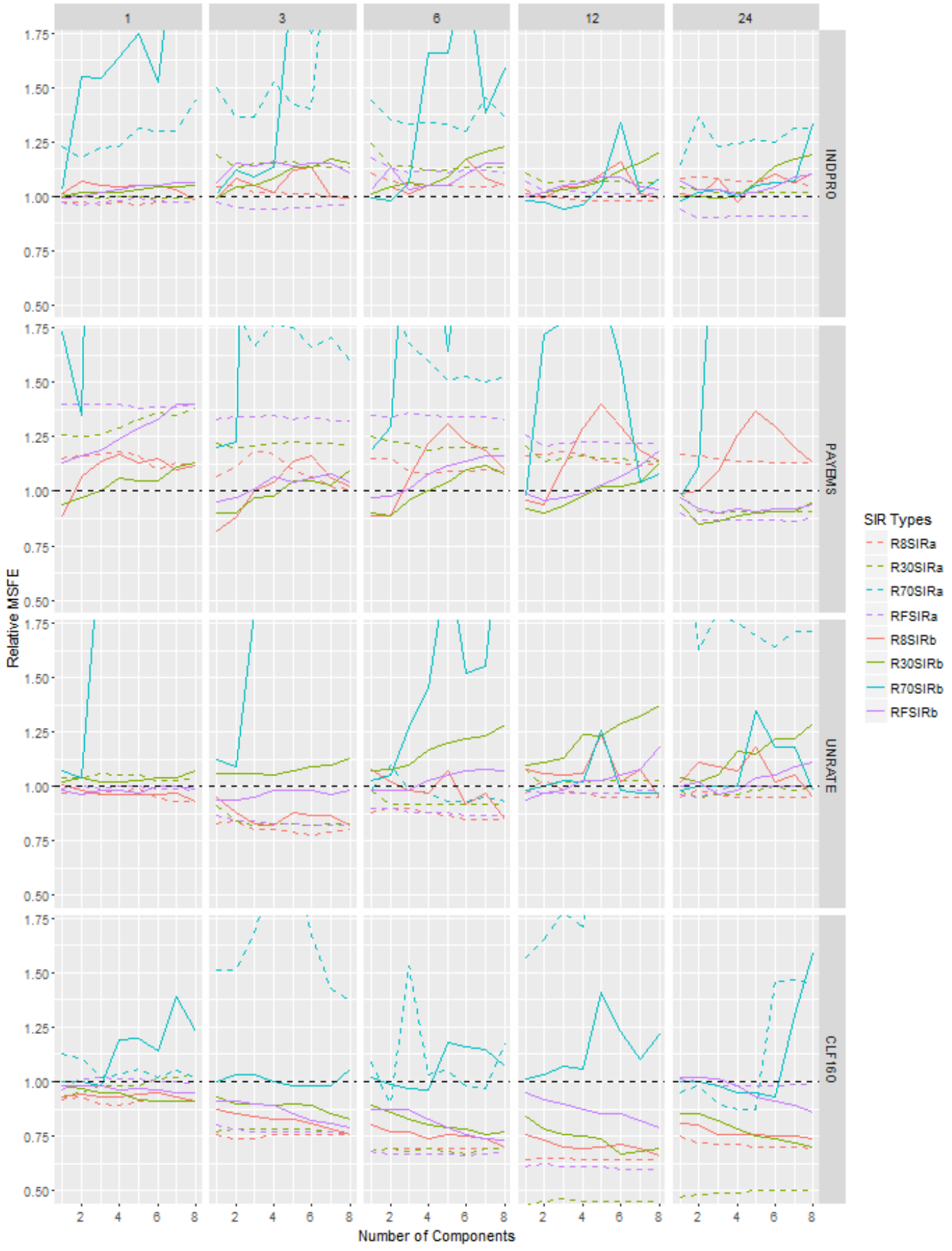


Figure 3: Relative MSFE vs Number of Components for SIR Estimators for Labor Market Variables and Industrial Production (Out-of-Sample from 1992 to 2016).



5.6 Comparing Forecasting Performance Across Estimator Families

Figures 4 and 5 plot the relative MSFE to AR(4) against the number of components for PCR, PLS and the best between type–a and type–b regularized SIR with 8, 30 and 70 components estimator "families." Regularized SIR on PCs selected by best subset among all the PCs of the predictors is also reported (RFSIR). The dashed horizontal line at 1 is the AR(4) reference line.

PLS and PCR – In Figure 4 we see that PLS outperforms PCR in forecasting `CPIAUC`, but PCR is better at forecasting wages (`CES060`), especially at longer horizons and at $h = 24$ PCR is essentially the only estimator capable of beating AR(4). Also consumption (`DPCERA`) is better predicted using PCR, whereas PLS is better than PCR in forecasting real personal income (`RPI`). Turning to Figure 5, overall PCR appears to perform better when it uses many components. With one or two components, PLS usually dominates PCR.

RSIR – From both Figures 4 and 5 we conclude that regularized SIR summarizes the information encapsulated in many PCs with very few components (one or two) and performs better or on a par with PCR and PLS.

The Effect of Targeting on Number of Components – MSFEs of targeted methods, PLS and SIR, are in general increasing in the number of components. This suggests that targeting methods do not require many components. In this respect, the extreme targeted nature of RSIR is reflected in the steepness of the gradient of its MSFE in the number of components, almost always steeper than that of PLS. Frank and Friedman (1993) [41] pointed out that PLS achieves its best performance with a smaller number of components than PCR. We show that SIR–type methods offer more extreme data compression than PLS. By contrast, the MSFE of PCR varies in a non-monotone fashion with the number of components: In Figure 4 PCR delivers its best forecasting performance with just a few components, whereas in Figure 5 it needs a large number of components to achieve competitive advantage over other estimators.

RFSIR and PCF – The magenta dots in Figure 4 and Figure 5 report the MSFE of RFSIR showing it to often outperform or be on a par with other types of regularized SIR. We do not report PCF, that is PCR computed on exactly the same PCs, since such PCR specification was not consistently outperforming other PCR specifications (it made into the list of best of estimators in Table 4 only twice). Both specifications are based on the same PCs, however SIR is able to utilize information in the tail of the PC distribution much more effectively, as reflected by its lower MSFE.

Figure 4: Relative MSFE vs Number of Components for Prices, Wages, Real Consumption and Real Income (Out-of-Sample from 1992 to 2016).

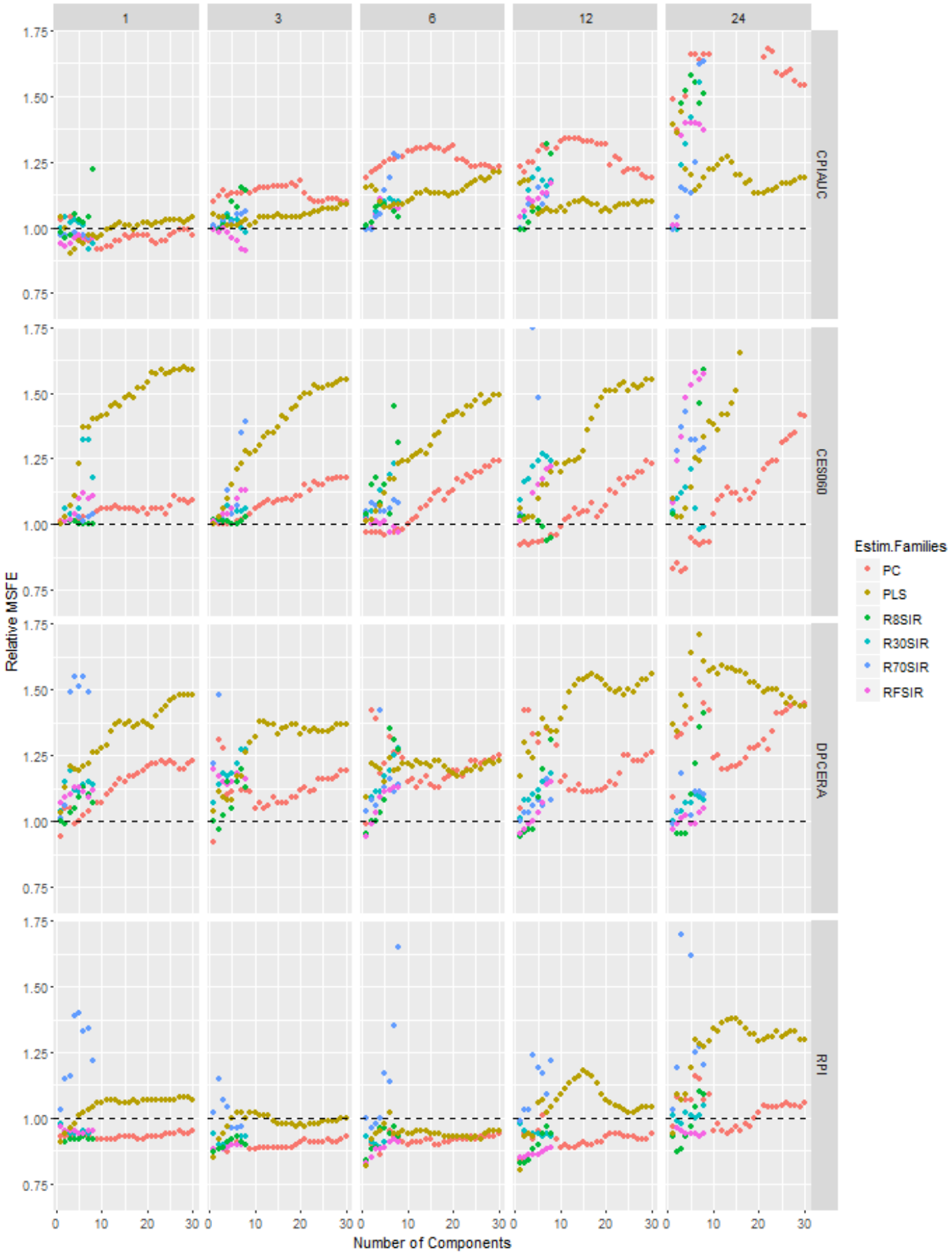
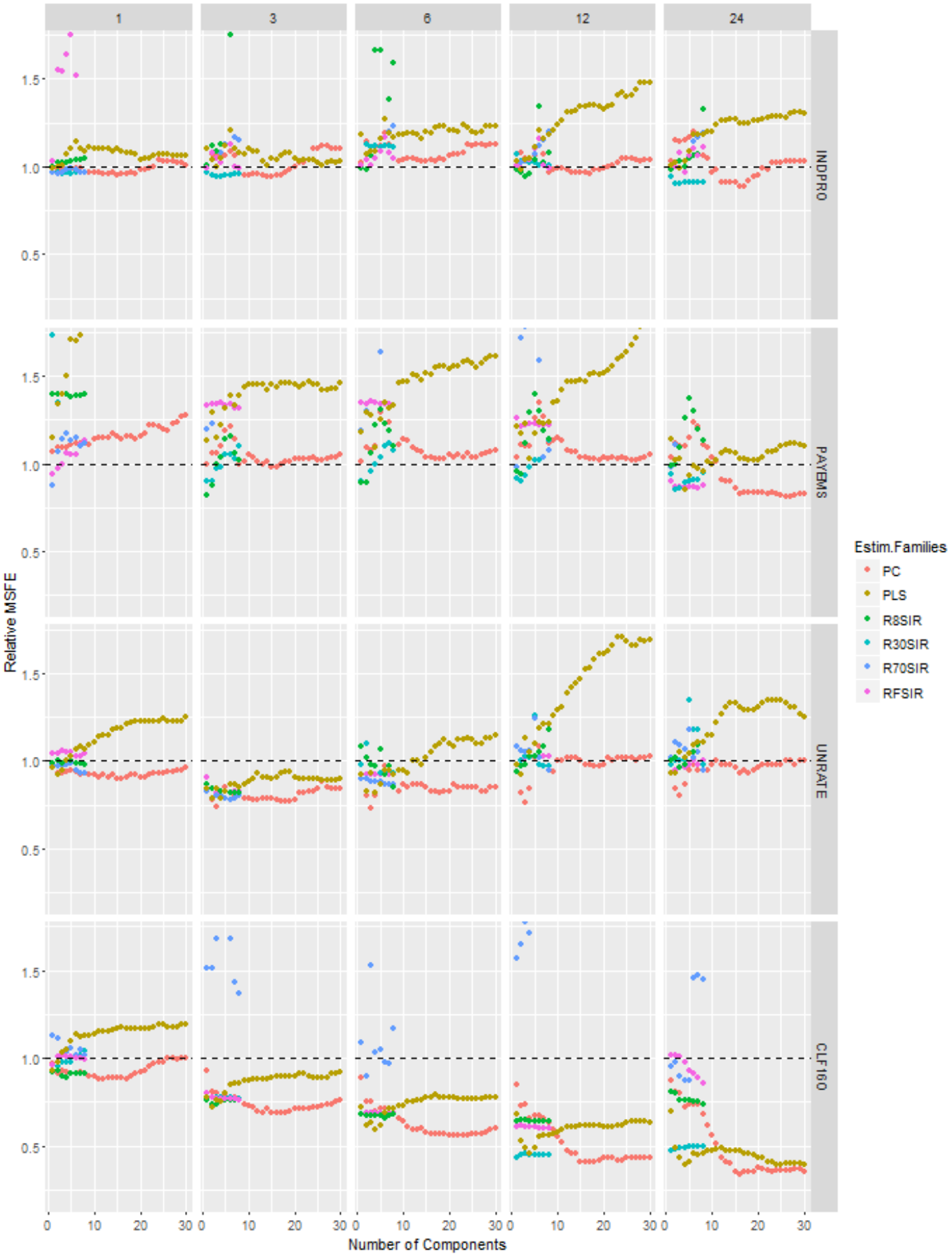


Figure 5: Relative MSFE vs Number of Components for Labor Market Variables and Industrial Production (Out-of-Sample from 1992 to 2016).



5.7 Robustness over Subsamples

The panels in Figure 6 through Figure 8 show side-by-side the relative MSFE computed recursively over the 1992 to 2016 and over the 2010 to 2016 out-of-sample windows for some of the targets we considered.[39] Each point of each time series plotted is the MSFE up to that time point for the best estimators in each family at the end of the forecast window. For example, if SIR with one component (denoted as SIR.1) is the best among standard SIRs at the end of the window, its MSFE is selected for plotting. Only the best among the regularized SIRs is plotted. The estimator with lowest MSFE among *all* estimators at the end of the window for a particular target-horizon combination coincides with Table 4. The grey bands track the upper and lower bounds of MSFE and provide a comprehensive view of the performance of all estimators we considered. The evolution of MSFE provides an assessment of the robustness of the various estimators considered in Table 4. We can follow how close other estimators are to the best performers in Table 4 and how robust the best estimators are as the recursive out-of-sample window expands. Horizons $h = 3$, $h = 6$ and $h = 24$ are shown in order to explore the evolution of the MSFE in short, medium and long term forecasting. The MSFE plotting range is 0.6 to 1.4.[40]

Effect of the Great Recession – The plots in the left column of each figure show that although the great recession had an impact on some of the plotted estimators, especially RIDGE, most of the plotted estimators beat AR(4) over that period. That is, they appear to be better equipped to capture the non-linearities characterizing the beginning of the recession or benefit from the information contained in the panel as compared to the simplistic AR(4). This said, the grey bands spike up in the recession capturing the fact that many estimators break down over that period.

Performance over the Recent Expansion – The columns on the right show MSFEs over the recent expansionary period. A simple comparison with the column to the left shows that, in general, the various estimators perform better than AR(4) when recessions are excluded from the out-of-sample window. Yet, over longer out-of-sample windows, the accumulation of large forecast errors in certain periods may result in inferior forecasting performance than AR(4). This suggests that selection of the "meta parameters" in Table 1 indexing the various estimator families is best done on rolling windows rather than recursive windows. It is also suggestive of some structural instability to which

[39] MSFE is relative to the AR(4). We considered only some targets to economize on space; plots for all targets are available on request.

[40] Estimators for which components have been selected with *best subset selection* are denoted with ".BS". For instance, "PC.BS" denotes PCR where at each step the components are selected by best subset selection with cp, bic, bic3, or icp1. The specific criterion used can be found in Table 4.

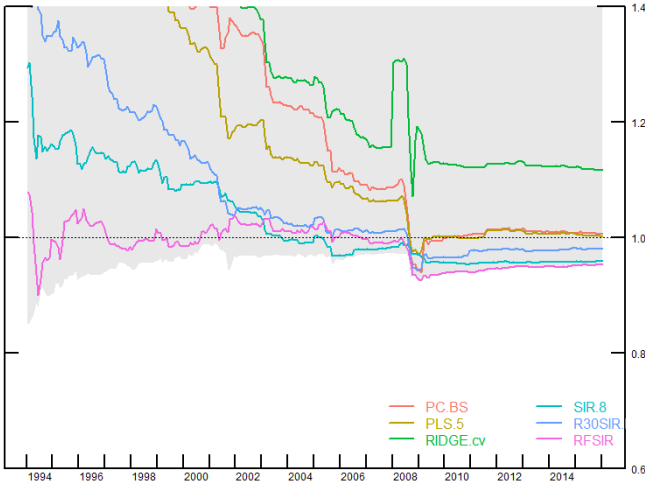
the AR(4) is somewhat less susceptible. Although RIDGE is often dominated by other estimators, it has a competitive edge in forecasting `PAYEMS` in the recent expansion.

Performance across Forecast Horizons – Comparing panels across rows and focusing on the grey bands we deduce that overall the estimators studied perform better relative to the AR(4) at longer horizons. Possibly the richer information set relative to AR(4) exploited by the various estimators reviewed is helpful in forming better long-run projections.

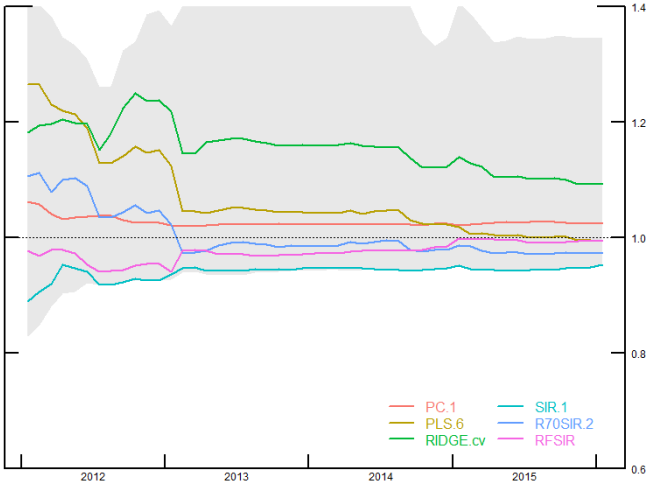
Robustness of SIR Estimators – SIR without regularization rarely performs well when the estimation sample size is smaller (left columns). It becomes more competitive as the estimation sample increases (right columns). By contrast, regularized SIR is consistent in its performance over different forecasting windows, targets and horizons and stands out as the most valuable application of SDR techniques in our forecast experiments. RSIR summarizes many PCs into one or two components with no deterioration of forecasting performance.

Figure 6: Evolution of Relative MSFE in in Out-of-Sample of Best Estimators for CPIAUCSL.

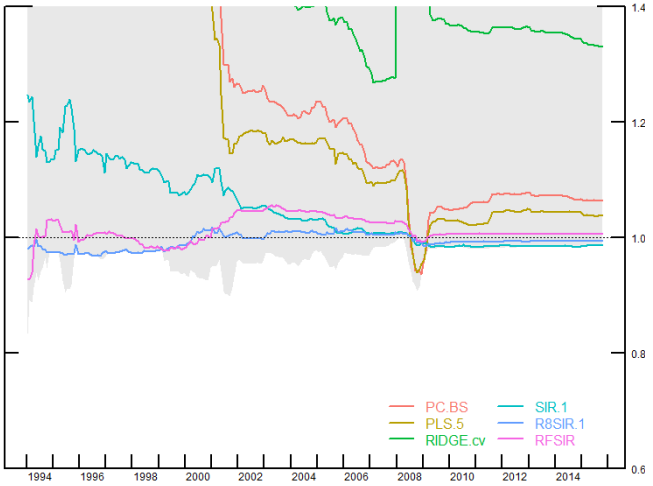
(a) MSFE from 1992 to 2016 at h=3



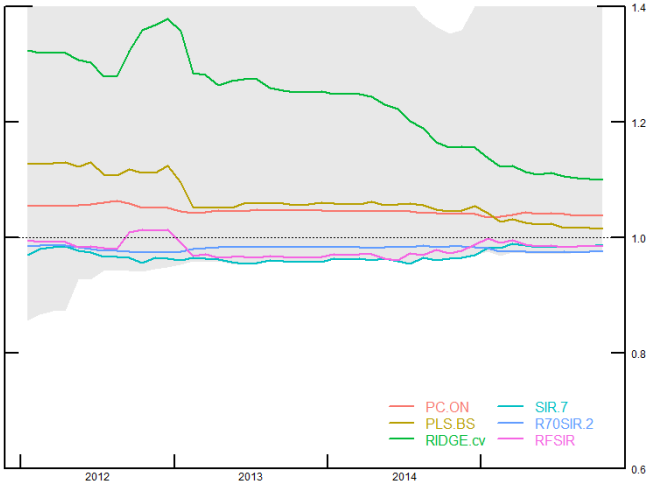
(b) MSFE from 2010 to 2016 at h=3



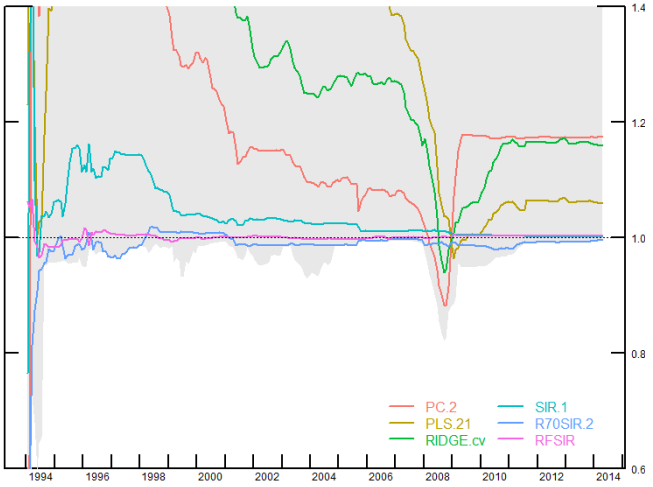
(c) MSFE from 1992 to 2016 at h=6



(d) MSFE from 2010 to 2016 at h=6



(e) MSFE from 1992 to 2016 at h=24



(f) MSFE from 2010 to 2016 at h=24

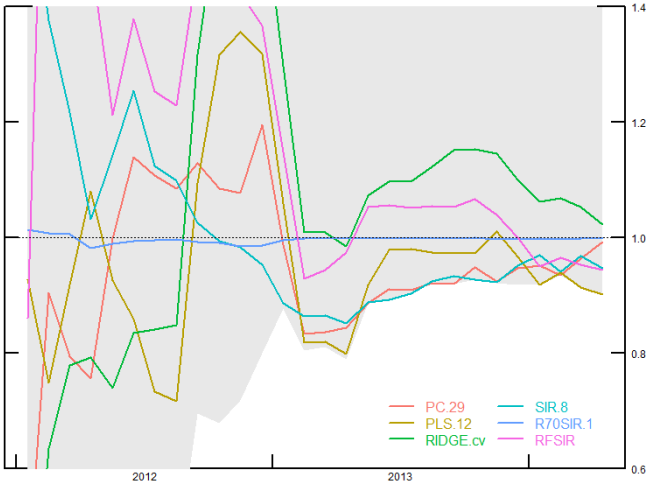
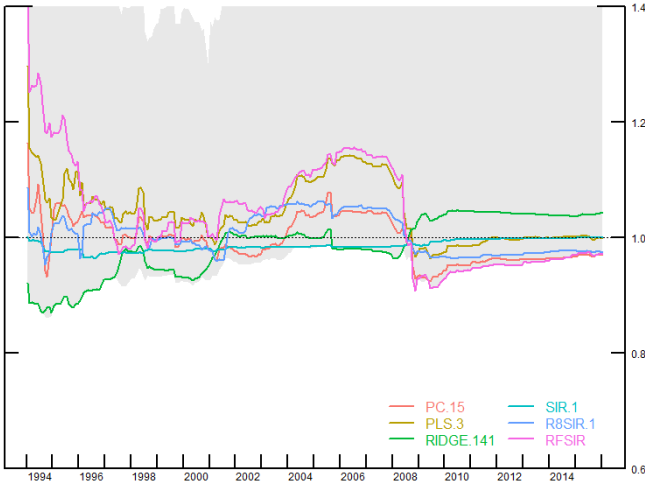
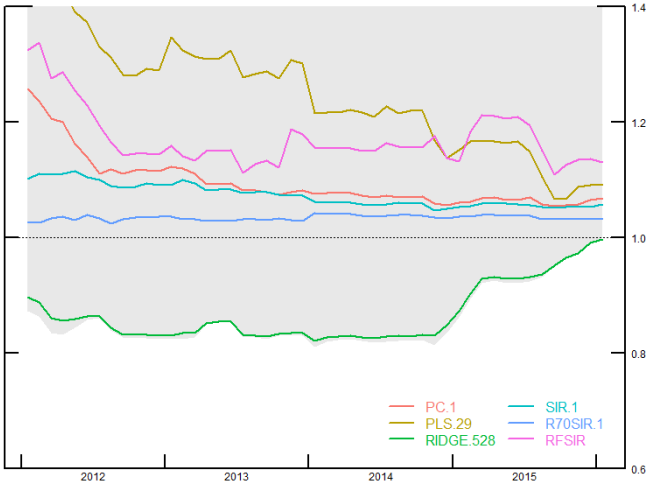


Figure 7: Evolution of Relative MSFE in Out-of-Sample of Best Estimators for INDPRO.

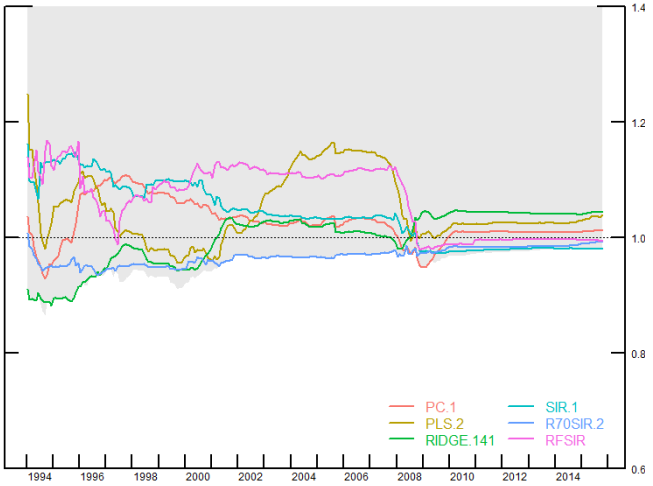
(a) MSFE from 1992 to 2016 at h=3



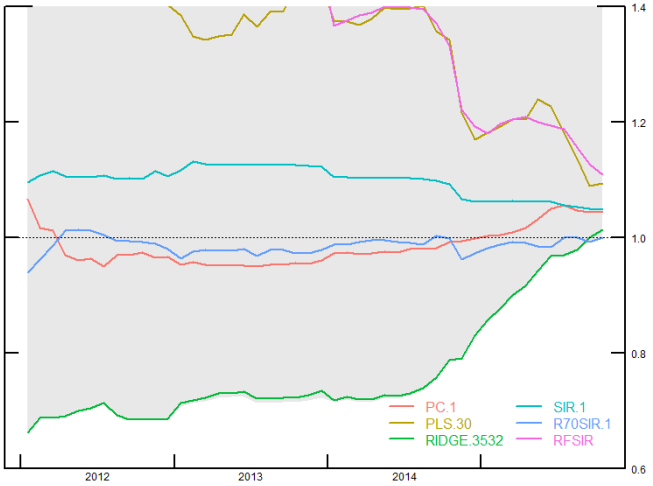
(b) MSFE from 2010 to 2016 at h=3



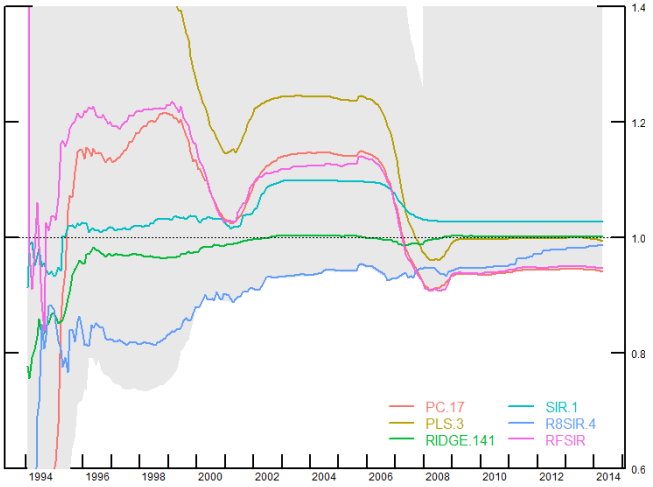
(c) MSFE from 1992 to 2016 at h=6



(d) MSFE from 2010 to 2016 at h=6



(e) MSFE from 1992 to 2016 at h=24



(f) MSFE from 2010 to 2016 at h=24

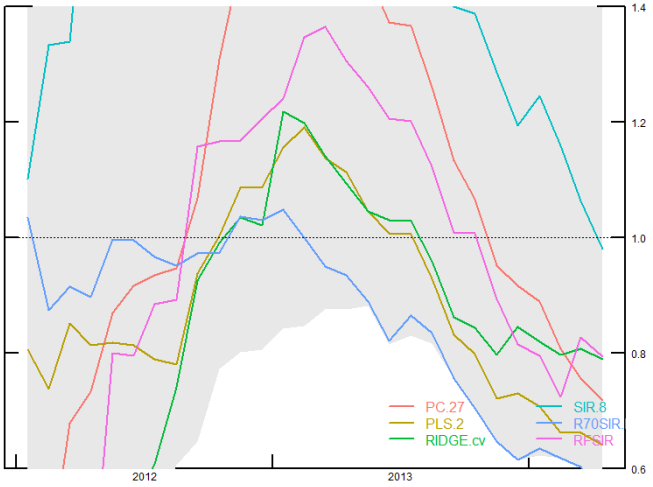
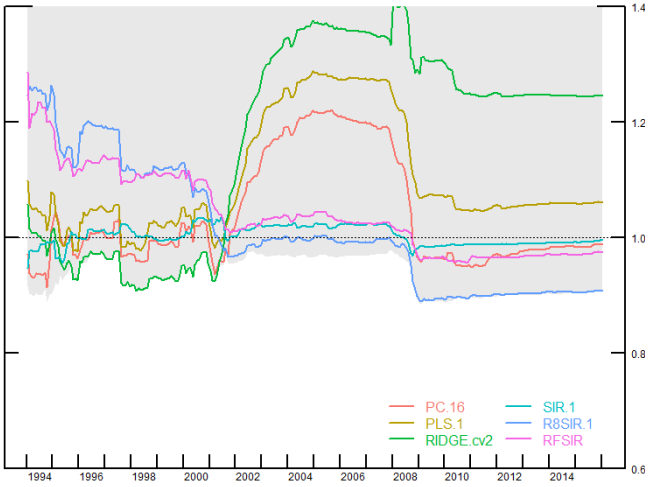
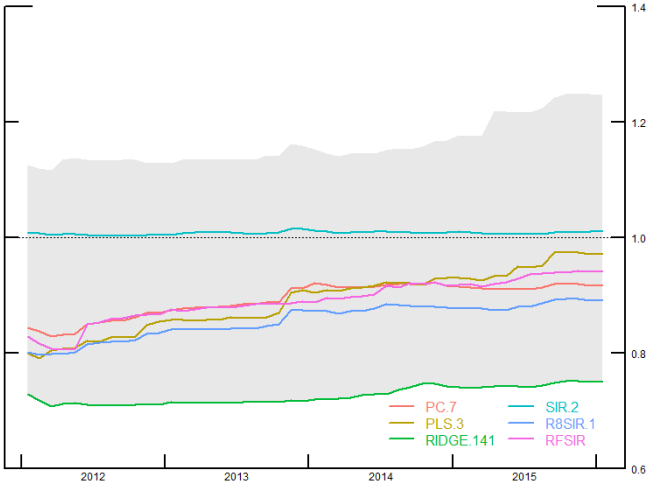


Figure 8: Evolution of Relative MSFE in Out-of-Sample of Best Estimators for PAYEMS.

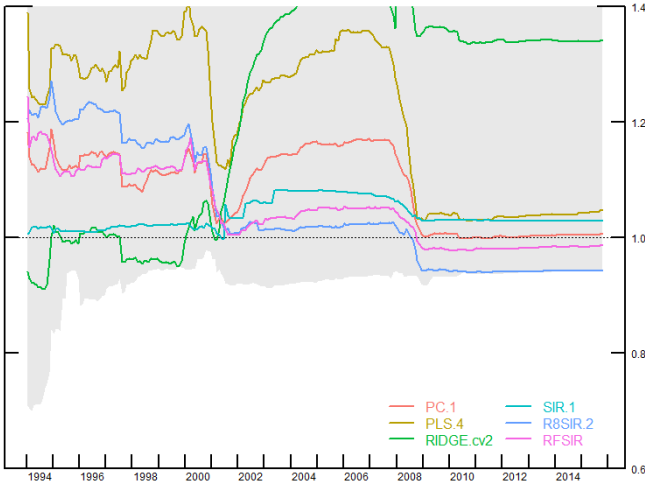
(a) MSFE from 1992 to 2016 at h=3



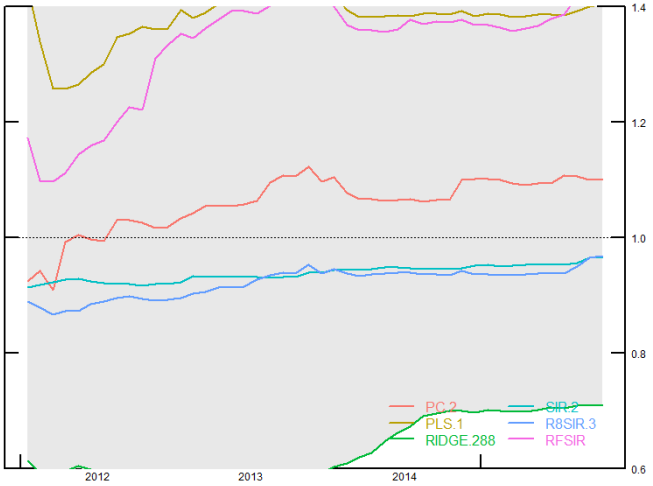
(b) MSFE from 2010 to 2016 at h=3



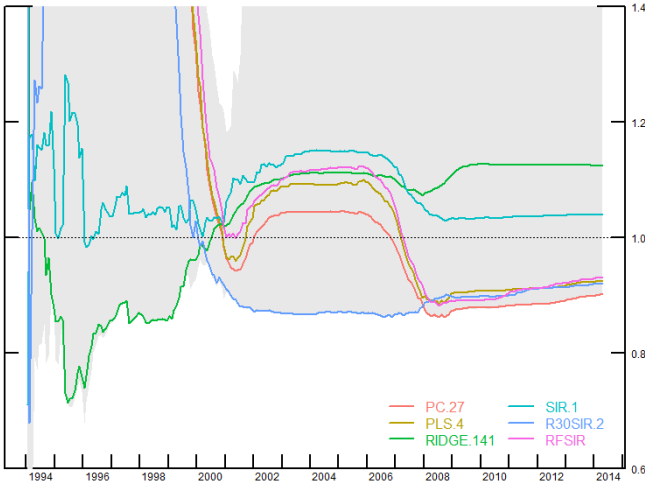
(c) MSFE from 1992 to 2016 at h=6



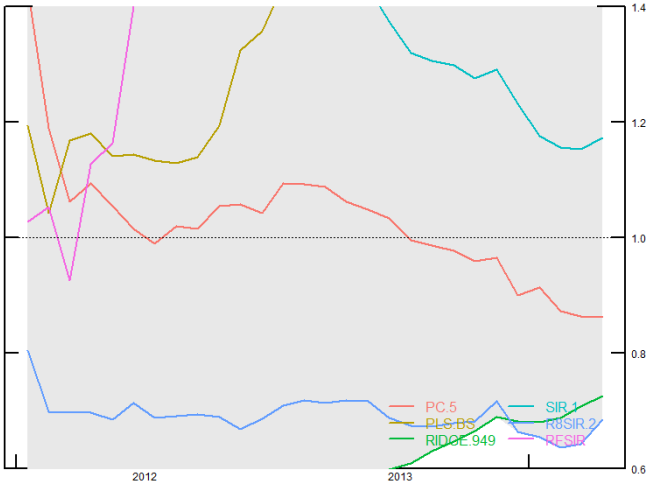
(d) MSFE from 2010 to 2016 at h=6



(e) MSFE from 1992 to 2016 at h=24



(f) MSFE from 2010 to 2016 at h=24



6 Summary and Conclusions

We (a) introduced sufficient dimension reduction methodology to econometric forecasting and focused on linear moment-based SDR; (b) derived properties of the SIR SDR estimator for covariance stationary series; (c) presented a common framework for OLS, PCR, RIDGE and PLS; and (d) studied the forecasting performance of these four methods and SIR, using the FRED-MD data set put together by McCracken and Ng (2015) [60].

The empirical results indicate that PCR and regularized SIR exhibit similar performance, which can be attributed to the ellipticity, strong cross-correlation and noise in the data. 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. PLS and RIDGE were not found to be competitive for these data and the time periods we considered in our forecasting exercise, except in few instances.

The performance of SIR, and SDR in general, is impeded by the size of the sample. 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 very unstable as the sample covariance matrix of the raw predictors was ill-conditioned. Our first stab at the problem consisted of preprocessing the data in regularized SIR. We are working on developing SDR methods that bypass the inversion of the sample covariance matrix of the predictors.

Furthermore, dimension two or higher in SIR indicates the presence of nonlinear relationships between the response and the SIR predictors. Gains in forecasting accuracy can potentially be realized by the inclusion of nonlinear SIR terms, either in the mean or variance components of the forecasting model. Plots of the response versus the SIR predictors would inform the construction of a more appropriate forward model. Even though this is very challenging to incorporate in an out-of-sample recursive automated forecasting experiment such as in our empirical application, it can be easily done in real time forecasting.

We conclude by noting that an important contribution of SDR methodology is conceptual as it shifts the focus from a hypothetical and practically untestable lower-dimensional latent structure to reductions of the observed data.

Appendix A: Regularized SIR Algorithm

In relevance to the forecasting model (2.2), 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 $\widehat{\mathbf{\Sigma}}_x = \widehat{\mathbf{V}}\widehat{\mathbf{D}}\widehat{\mathbf{V}}'$, where $\widehat{\mathbf{V}} = (\widehat{\mathbf{v}}_1, \ldots, \widehat{\mathbf{v}}_p)$ are the $\widehat{\mathbf{\Sigma}}_x$ eigenvectors, and $\widehat{\mathbf{D}} = diag(\hat{\lambda}_1, \ldots, \hat{\lambda}_p)$ is the diagonal matrix with the eigenvalues of $\widehat{\mathbf{\Sigma}}_x$ 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) [4] 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 principal 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 $\hat{f}_1 = \widehat{\mathbf{v}}_1'\mathbf{x}, \ldots, \hat{f}_m = \widehat{\mathbf{v}}_m'\mathbf{x}$ be the retained principal factors of $\mathbf{x}$.

2. Let $\widetilde{\mathbf{x}}_t = (\hat{f}_{t1}, \ldots, \hat{f}_{tM}, y_t, y_{t-1}, \ldots, y_{t-L})' = (\tilde{X}_1, \ldots, \tilde{X}_{m+L+1})'$ 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 slice S_j of the range of the y_t values.

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 $\widehat{\Lambda} = \text{diag}(\hat{l}_1, \ldots, \hat{l}_q)$, $\hat{l}_1 > \hat{l}_2 > \ldots > \hat{l}_q$ 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 $\hat{l}_1, \hat{l}_2, \ldots, \hat{l}_q$, respectively.

7. Using any dimension estimation method that applies, let $\hat{d}$ be the estimate of the dimension d of the regression.

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

Appendix B: Covariance Stationary Series

A sequence of random variables x_{jt} is covariance stationary or weakly stationary if and only if $\exists \mu_j \in \mathbb{R}\colon \text{E}(x_{jt}) = \mu_j, \forall t > 0$, and

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

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 $\text{E}(x_{jt}^2) < \infty\ \forall t$, is also weakly stationary. 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 $\text{E}(\mathbf{x}_t) = \boldsymbol{\mu} = (\text{E}(x_{1t}), \text{E}(x_{2t}), ..., \text{E}(x_{pt}))' = (\mu_1, \mu_2, \ldots, \mu_p)'$ and the variance/covariance matrix

$$\begin{aligned}
\mathbf{\Sigma}_x(0) &= \text{var}(\mathbf{x}_t) = \left((\mathbf{x}_t - \boldsymbol{\mu})(\mathbf{x}_t - \boldsymbol{\mu})'\right) = \text{E}\left(\mathbf{x}_t\mathbf{x}_t' - \boldsymbol{\mu}\boldsymbol{\mu}'\right) = \\
\mathbf{\Sigma}_x(h) &= \text{cov}(\mathbf{x}_{t+h}, \mathbf{x}_t) = \text{E}\left((\mathbf{x}_{t+h} - \boldsymbol{\mu})({}_{t+h} - \boldsymbol{\mu})'\right) = \text{E}\left(\mathbf{x}_{t+h}\mathbf{x}_t' - \boldsymbol{\mu}\boldsymbol{\mu}'\right)
\end{aligned}$$

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)). 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 $\boldsymbol{\mu}$, $\mathbf{\Sigma}_x(0)$, and $\mathbf{\Sigma}_x(h)$ are estimated from $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_t$ using the sample moments $\bar{\mathbf{x}}_T = \sum_{t=1}^{T} \mathbf{x}_t/T$, $\widehat{\mathbf{\Sigma}}_x(0) = \frac{1}{T}\sum_{t=1}^{T}(\mathbf{x}_t - \bar{\mathbf{x}})(\mathbf{x}_t - \bar{\mathbf{x}})'$, and

$$\widehat{\mathbf{\Sigma}}_x(h) = \begin{cases} \frac{1}{T}\sum_{t=1}^{T-h}(\mathbf{x}_{t+h} - \bar{\mathbf{x}})(\mathbf{x}_t - \bar{\mathbf{x}})' & \text{if } 0 \leq h \leq T-1 \\ \hat{\mathbf{\Gamma}}(h)' & \text{if } -T+1 \leq h < 0 \end{cases}$$

The ergodic theorem obtains that if $\mathbf{x}_t$ is a strictly stationary and ergodic time series, then, as $T \to \infty$, $\bar{\mathbf{x}}_T \xrightarrow{p} \boldsymbol{\mu}$, $\widehat{\boldsymbol{\Sigma}}_x(0) \xrightarrow{p} \boldsymbol{\Sigma}_x(0)$, $\widehat{\boldsymbol{\Sigma}}_x(h) \xrightarrow{p} \boldsymbol{\Sigma}_x(h)$.

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 $\sqrt{T}\ \bar{\mathbf{x}}_T$ converges to $\mathbf{V}$ when T increases indefinitely.

Appendix C: List and Description of Variables

The following set of tables summarizes the variables in FRED-MD. Each table collects variables by statistical data release allowing a bird's eye view on the sources of information in the dataset and imparting an organization of the variables somewhat different relative to the tables in McCracken and Ng (2015) [60].[41] We also briefly describe each statistical data release. Column "T" reports the transformation used.[42] Column "FRED-MD" reports variable mnemonics in FRED-MD dataset. Column "Description" permits to interpret the series.[43]. Contrary to Stock and Watson (2005) [72] we include all variables when computing PCs hence we do not need to flag variables not used in their computation. Asterisked series are adjusted by McCracken and Ng (2015) (see [60] for details).

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

[41] Column "G" reports the grouping chosen by McCracken and Ng (2015) [60], in turn not too dissimilar from groupings operated in other DFM studies. The groupings codes are: (1) *Output and Income*; (2) *Labor Market*; (3) *Consumption and Orders*; (4) *Orders and Inventories*; (5) *Money and Credit*; (6) *Interest rate and Exchange Rates*; (7) *Prices*; (8) *Stock Market*. We do not use groupings meta data in our empirical analysis.

[42] The transformations closely follow McCracken and Ng (2015) [60] who in turn follow Stock and Watson. For series x_t transformation codes are: (1) no transformation; (2) Δx_t; (3) $\Delta^2 x_t$; (4) $log(x_t)$; (5) $\Delta log(x_t)$; (6) $\Delta^2 log(x_t)$; (7) $\Delta(x_t/x_t 11.0)$.

[43] The remaining two columns denote the Global Insight code and description; the GSI description allows to map the individual series into datasets used in older papers.

by the Census Bureau. Both retail sales and `UMCSENTx`, the consumer sentiment index from the University of Michigan, are often used in forecasting PCE (other informative sub-indexes of the Michigan survey were not included in the dataset), however given its short history we excluded `UMCSENTx` from our dataset.

Table 5: Variables Measuring Income and 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
130*	2	4	UMCSENTx	Consumer Sentiment Index	Consumer expect

Variables Measuring Industrial Production – 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 94+ sectors at NAICS 4-digit level and covers the manufacturing, mining and utilities sectors. `INDPRO`, the first variable in Table 6 is the top aggregate of the IP system and the next seven rows in Table 6 represent the splitting and regrouping of the 200+ atoms in the IP system in so called "market" groups. It was an odd choice to include `IPFUELS` in FRED-MD given its idiosyncratic pattern and the fact that it is an atom at a 6-digits NAICS. `CUMFNS` is one of the few observable measures of slack and it is computed as $\frac{\text{manufacturing IP}}{\text{manufacturing capacity}}$.[44].

[44]Manufacturing capacity is estimated by staff at the Federal Reserve Board using the Quarterly Survey of Plant Capacity (in turn run by the Census Bureau)

Table 6: Variables from the Industrial Production 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

Diffusion Indexes from ISM Manufacturing Survey – Table 7 collects some of the diffusion indexes from the Institute for Supply Management (ISM).[45] These variables are released the first day of the month following the reference month, hence they are quite timely and mostly useful in a nowcasting experiment, although some variables might also contain signal for several months ahead such as "new orders" a natural measure of future activity. These variables are diffusion indexes, that is they are expressed as the fraction of respondents that say that activity is up, they are stable and therefore they are left in levels in the estimation.

Table 7: Diffusion Indexes from the ISM

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

[45] The ISM is formerly known as the National Association of Purchasing Managers (NAPM). The ISM also reports other interesting diffusion indexes such as "new export orders", or "level of inventories", however these variables are available only starting from the 1990s consequently they have not been included in FRED-MD. The same is true for the recently introduced diffusion indexes from the Markit survey and data from the services and manufacturing surveys released by the regional FEDs.

Variables Measuring Orders and Inventories – Table 8 reports data 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 in 89 industry categories. 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 8: 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

Labor Market Variables – Table 9 contains variables produced by the Bureau of Labor Statistics (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) [9]. Obvious candidates missing in the datasets are labor market indicators Federal Reserve staff watches, such as data from the JOLTS survey.

Table 9: Labor Market Variables

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 – Table 10 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 in the dataset.

Table 10: 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 the Money Stock and Reserves – These data come mainly from the Federal Reserve Board H.6 statistical release.

Table 11: 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 the Stock Market – These data are elaborated by Standard & Poor's.

Table 12: 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

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

Table 13: 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 – Table 14 contains variables measuring interest rates, yields, spreads and exchange rates that used to be contained in the statistical release H.15 by the Federal Reserve Board. The federal funds rate, interest rates on treasuries and commercial paper are still published in H.15 however publication of the other data has been discontinued as of October 2016 (however it does not affect our study since we use the May 2016 vintage of FRED-MD).

Table 14: 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 – Table 15 collects the variables 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 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 points of sale.

Table 15: 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

REFERENCES

[1] Ahn, S. C., and Horenstein, A. R. (2013), "Eigenvalue Ratio Test for the Numbers of Factors," *Econometrica* 81(3), 1203-1227.

[2] Alessi, L., Barigozzi, M., Capasso, M. (2010) "Improved Penalization for Determining the Number of Factors in Approximate Factor Models," *Statistics & Probability Letters* 80(23-24), 1-15.

[3] Bai, J. (2003). "Inferential Theory for Factor Models of Large Dimensions," *Econometrica* 71, 135-171.

[4] Bai, J. and Ng, S. (2002), "Determining the Number of Factors in Approximate Factor Models," *Econometrica* 70(1), 191-221.

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

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

[7] Banbura, M., Giannone, D., and Reichlin, L. (2011), "Nowcasting," in Clements, M.P., Hendry, D.F. (Eds.), *Handbook on Economic Forecasting*. Oxford University Press, pp. 63-90.

[8] Barbarino, A. and Bura, E. (2015). "Forecasting with Sufficient Dimension Reductions", Finance and Economics Discussion Series 2015-074. Washington: Board of Governors of the Federal Reserve System, `http://dx.doi.org/10.17016/FEDS.2015.074`.

[9] Barnichon R. (2010), "Building a Composite Help-Wanted Index," *Economics Letters,* 109(3), 175-178.

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

[11] Breitung, J. and S. Eickmeier (2005), "Dynamic Factor Models," Deutsche Bundesbank Discussion Paper 38/2005.

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

[13] 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.

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

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

[16] Bura, E., Duarte, S. and Forzani, L. (2016) "Sufficient reductions in regressions with exponential family inverse predictors," *Journal of the American Statistical Association* 111, 1-17.

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

[18] Carrasco, M. and Rossi, B. (2016), "In-Sample Inference and Forecasting in Misspecified Factor Models", *Journal of Business and Economic Statistics*, 34:3, 313-338, DOI: 10.1080/07350015.2016.1186029.

[19] Cassart, D. (2007) "Optimal Tests for Symmetry," PhD dissertation, Université Libre de Bruxelles.

[20] Chiaromonte, F., Cook, R.D. and Li, B. (2002), "Sufficient dimension reduction in regression with categorical predictors," *The Annals of Statistics* 30, 475-497.

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

[22] Cook, R. D. (1994), "Using dimension-reduction subspaces to identify important inputs in models of physical systems," in *Proc. Sect. Phys. Eng. Sc.*, p. 18-25. Alexandria, VA: American Statistical Association.

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

[24] Cook, R. D. (1998b), "Principal Hessian directions revisited (with discussion)," *Journal of the American Statistical Association* 93, 84-94.

[25] 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.)

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

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

[28] Cook, R. D. and Forzani, L. (2009), "Likelihood-Based Sufficient Dimension Reduction," *Journal of the American Statistical Association* 104, 197-208.

[29] Cook, R. D. and Lee, H. (1999), "Dimension reduction in binary response regression," *Journal of the American Statistical Association* 94, 1187-1200.

[30] Cook, R.D. and Ni, L. (2005), "Sufficient dimesion reduction via inverse regression: A minimum discrepancy approach," *Journal of the American Statistical Association* 100, 410-428.

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

[32] 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.

[33] 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`.

[34] Diaconis, P. and Freedman, D. (1984) "Asymptotics of graphical projection pursuit," *The Annals of Statistics* 12, 793-815.

[35] 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`.

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

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

[38] Eickmeier, S., and Ziegler, C. (2008), "How Successful are Dynamic Factor Models at Forecasting Output and Inflation? A Meta-Analytic Approach," *Journal of Forecasting* 27, 237-265.

[39] Fan, J., Xue, L. and Yao, J. (2015), "Sufficient Forecasting Using Factor Models," *ArXiv e-prints*, `http://arxiv.org/pdf/1505.07414.pdf`.

[40] 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.

[41] Frank, I., and Friedman, J. (1993). A Statistical View of Some Chemometrics Regression Tools. *Technometrics,* 35(2), 109-135. doi:10.2307/1269656

[42] Friedman, J., Hastie, T. and Tibshirani, R. (2015), "Regularization Paths for Generalized Linear Models via Coordinate Descent," *Journal of Statistical Software* 33(1), 1-22, `http://www.jstatsoft.org/v33/i01/`.

[43] Goldberger, A. (1991), *A Course in Econometrics,* Harvard University Press, Cambridge, Massachusetts.

[44] 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.

[45] 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.

[46] Hastie, T. J. and Tibshirani, R. J. (1990), *Generalized Additive Models,* Chapman & Hall/CRC.

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

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

[49] Helland, I. (1990), "Partial Least Squares Regression and Statistical Models," *Scandinavian Journal of Statistics* Vol. 17, No. 2 (1990), pp. 97-114.

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

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

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

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

[54] Krmer, N. and Sugiyama, M. (2011), "The Degrees of Freedom of Partial Least Squares Regression", *Journal of the American Statistical Association*, 106:494, 697-705, DOI: 10.1198/jasa.2011.tm10107.

[55] Kelly, B. T. and Pruitt, S. (2015), "The three-pass regression filter: A new approach to forecasting using many predictors," *Journal of Econometrics* 186, 294-316. `http://dx.doi.org/10.1016/j.jeconom.2015.02.011`.

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

[57] Li, K. C. (1991), "Sliced Inverse Regression for Dimension Reduction (with discussion)," *Journal of the American Statistical Association* 86, 316-342.

[58] Li, K. C. (2000), High dimensional data analysis via the SIR/PHD approach.

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

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

[61] Mevik B.H., Wehrens R., and Liland K. H. (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.

[62] Serena Ng (2013), "Variable Selection in Predictive Regressions". In *Handbook of Economic Forecasting,* Graham Elliott, Clive Granger, Allan Timmerman (eds.), North Holland.

[63] Onatski, A. (2009), "Testing Hypotheses about the Number of Factors in Large Factor Models," *Econometrica* 77(5), 1447-1479.

[64] Onatski, A. (2010), "Determining the Number of Factors from Empirical Distribution of Eigenvalues," *Review of Economics and Statistics* 92(4), 1004-1016.

[65] Racine, J.S. and Hayfield, T. (2009), "Nonparametric kernel smoothing methods for mixed data types," available on CRAN as `np`. See also "Nonparametric Econometrics: The np Package" by the same authors in *Journal of Statistical Software* 27(5), 2008.

[66] Racine JS, Li Q (2004), "Nonparametric Estimation of Regression Functions with both Categorical and Continuous Data," *Journal of Econometrics* 119(1), 99-130.

[67] Robinson PM (1988), "Root-N Consistent Semiparametric Regression," *Econometrica* 56, 931-954.

[68] Steinberger, L. and Leeb, H. (2015), "On conditional moments of high-dimensional random vectors given lower-dimensional projections," `http://arxiv.org/pdf/1405.2183.pdf`.

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

[70] Stock, J. H. and Watson, M. (2002), "Forecasting Using Principal Components from a Large Number of Predictors," *Journal of the American Statistical Association* 97(460), 1167-1179.

[71] Stock, J. H. and Watson, M. (2002), "Macroeconomic Forecasting Using Diffusion Indexes". *Journal of Business and Economic Statistics,* 20(2), 147-162.

[72] Stock, J. H. and Watson, M. (2005), "Implications of Dynamic Factor Models for VAR Analysis," NBER Working Paper No. 11467, `http://www.nber.org/papers/w11467`

[73] 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.

[74] 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.

[75] Stock, J. H. and Watson, M. (2011), "Dynamic Factor Models," in Clements, M.P., Hendry, D.F. (Eds.), *Handbook on Economic Forecasting,* Oxford University Press.

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

[77] Stone, M. and Brooks. R. J. (1990), "Continuum Regression: Cross-Validated Sequentially Constructed Prediction Embracing Ordinary Least Squares, Partial Least Squares and Principal Components Regression," *Journal of the Royal Statistical Society Series B (Methodological),* 52(2), 237-269.

[78] Verardi, V. and Croux, C. (2009), "Robust regression in Stata," *Stata Journal* 9(3), 439-453.

[79] Villegas, C. (1976), "On A Multivariate Central Limit Theorem for Stationary Bilinear Processes," *Stochastic Processes and their Applications* 4(2), 121-133.

[80] Weisberg, S. (2002), "Dimension Reduction Regression in R," *Journal of Statistical Software* 7(1), 1-22.

[81] Wold, S. (1978), "Cross Validatory Estimation of the Number of Components in Factor and Principal Component Models," *Technometrics* 20, 397-405.

[82] Wold, S, Ruhe, A., Wold, H., and Dunn II, W.J. (1984), "The Collinearity Problem in Linear Regression. the Partial Least Squares (PLS) Approach to Generalized Inverses", *Siam J. ScI. Stat. Comput.* Vol. 5, No. 3, September 1984