2014

Sparse Factor Auto-Regression for Forecasting Macroeconomic Time Series with Very Many Predictors

Oliver Kurt Galvis

FLORIDA STATE UNIVERSITY

COLLEGE OF ARTS AND SCIENCES

SPARSE FACTOR AUTO-REGRESSION FOR FORECASTING MACROECONOMIC TIME SERIES WITH VERY MANY PREDICTORS

By

OLIVER KURT GALVIS BALBÁS

A Dissertation submitted to the
Department of Statistics
in partial fulfillment of the
requirements for the degree of
Doctor of Philosophy

Degree Awarded:
Summer Semester, 2014

Oliver Kurt Galvis Balbás defended this dissertation on July 07, 2014.
The members of the supervisory committee were:

Yiyuan She
Professor Directing Dissertation

Giray Ökten
University Representative

Paul Beaumont
Committee Member

Fred Huffer
Committee Member

Minjing Tao
Committee Member

The Graduate School has verified and approved the above-named committee members, and certifies that the dissertation has been approved in accordance with university requirements.

To God, La Virgen Del Valle, Yoleida, and Vittoria.

ACKNOWLEDGMENTS

First and foremost, I want to thank my advisor and mentor Dr. Yiyuan She for his guidance and shrewdness through difficult and stressful times throughout the years. His patience led me to complete my dissertation work. I am indebted for his dedication to develop my English writing skills, and obliged for the wonderful discussions about variable selection, forecasting, and other topics in statistical learning. I owe him most of my statistical abilities. He is a great teacher, amazing researcher, and superb family man.

Many thanks also to Dr. Giray Ökten for having opened the doors of FSU to me. I learned many things from our fruitful conversations. To Dr. Fred Huffer for his help during my early stages in the program. Working as his grader has been one of the most joyful times I've had at FSU. To Dr. Paul Beaumont for his willingness to discuss and explain key economic topics in a marvelous way. To Dr. Minjing Tao for her kindness and encouragement during my research work.

I would also like to thank Dr. Dan McGee for having opened the doors of the Department of Statistics to me. To Ken Baldauf for the opportunity and support given at PIC. To Dr. Steve Ramsier and Dr. Xu-Feng Niu for accepting my request to be part of the Statistical Consulting group. To Pamela McGhee, Alex Cohn, James Stricherz, Chauncey Richburg, and Marylou Tatis for their unconditional support.

Finally, I'm greatly thankful for my family. Tito, thank you for every conversation and correction. Kenneth and Flor Elena, thank you for the inspiration. René, Lele, and Faby, thank you! Orlando and Flor María, thank you for the love. Francisco and Silvina, thank you for your support. My deepest gratitude goes to Yoleida and Vittoria. Without you there would be no dissertation to defend and no dream to fight for. You revitalize every day of my life. Thanks for believing in me. I love you.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF SYMBOLS

Let $\boldsymbol{M}$ be a generic matrix.

m_t	an observation at time t
$\boldsymbol{m}_t$	a vector of observations with its last observation at time t
$\boldsymbol{M}_t$	a matrix with last vector of observations at time t
$\boldsymbol{M}^\mathsf{T}$	the transpose of $\boldsymbol{M}$
$\boldsymbol{M}^+$	the Moore-Penrose pseudo inverse of $\boldsymbol{M}$
$\mathrm{r}(\boldsymbol{M})$ or $\mathrm{rank}(\boldsymbol{M})$	the rank of $\boldsymbol{M}$
$\boldsymbol{M}_{ij}$	the entry of $\boldsymbol{M}$ in the ith row and jth column
$\|\boldsymbol{M}\|_F^2 = \sum_{i,j} c_{i,j}^2$	the Frobenius norm of $\boldsymbol{M}$
$\boldsymbol{I}$	the identity matrix
$\boldsymbol{\Gamma}$	the weight matrix
$\|\boldsymbol{M}\|_{2,0} = \sum_{j=1}^{p} 1_{\{c_j \neq 0\}}$	the hard-ridge penalty on $\boldsymbol{M}$
$\hat{\boldsymbol{\Sigma}}_{\boldsymbol{MM}}$	the sample covariance matrix of $\boldsymbol{M}$
$\boldsymbol{D}_{(\hat{\boldsymbol{\Sigma}}_{\boldsymbol{MM}})}$	the diagonal of the sample covariance matrix of $\boldsymbol{M}$
$\delta_{(\hat{\boldsymbol{\Sigma}}_{\boldsymbol{MM}})}$	the regularization parameter of $\hat{\boldsymbol{\Sigma}}_{\boldsymbol{MM}}$
λ	a regularization parameter
η	the ridge parameter
$\mathfrak{q}$	the cardinality constraint
r	the rank constraint

LIST OF ABBREVIATIONS

i.i.d.	independently and identically distributed
s.t.	subject to
AIC	Akaike's Information Criterion
AR(4)	Autoregressive of order 4
BIC	Bayesian Information Criterion
C-level factors	Category level factors
DFM	Dynamic Factor Model
e-Net	Elastic net
LARS	Least Angle Regression
LASSO	Least Absolute Shrinkage and Selection Operator
MLE	Maximum Likelihood Estimate
MSE	Mean Squared Error
NNP	Nuclear Norm Penalized
OLS	Ordinary Least Squares
PCA	Principal Component Analysis
PRCCR	Progressive Rank-Cardinality Constrained Algorithm
RRR	Reduced Rank Regression
RSC	Rank Selection Criterion
SCAD	Smoothly Clipped Absolute Deviation
SEL-RRR	Selectable Reduced-Rank Regression
SFAR	Sparse Factor Auto-Regression
SVD	Singular Value Decomposition
TCR	Target Category Response
TSR	Target Series Response
VAR(l)	Vector Autoregressive of order l

ABSTRACT

Forecasting a univariate target time series in high dimensions with very many predictors poses challenges in statistical learning and modeling. First, many nuisance time series exist and need to be removed. Second, from economic theories, a macroeconomic target series is typically driven by few latent factors constructed from some macroeconomic indices. Consequently, a high dimensional problem arises where deleting junk time series and constructing predictive factors simultaneously, are meaningful and advantageous for accuracy of the forecasting task. In macroeconomics, multiple categories are available with the target series belonging to one of them. With all series available we advocate constructing category level factors to enhance the performance of the forecasting task.

We introduce a novel methodology, the Sparse Factor Auto-Regression (SFAR) methodology, to construct predictive factors from a reduced set of relevant time series. SFAR attains dimension reduction via joint variable selection and rank reduction in high dimensional time series data. A multivariate setting is used to achieve simultaneous low rank and cardinality control on the matrix of coefficients where ℓ_0-contraint regulates the number of useful series and the rank constrain elucidates the upper bound for constructed factors. The doubly-constrained matrix is a nonconvex mathematical problem optimized via an efficient iterative algorithm with a theoretical guarantee of convergence. SFAR fits factors using a sparse low rank matrix in response to a target category series. Forecasting is then performed using lagged observations and shrinkage methods. We generate a finite sample data to verify our theoretical findings via a comparative study of the SFAR. We also analyze real-world macroeconomic time series data to demonstrate the usage of the SFAR in practice.

CHAPTER 1

MOTIVATION AND OVERVIEW

1.1 Economic and financial data motivation

In economic and finance, forecasting is an imperative task accomplished to identify the future trends of a variable of interest such as a stock's price movement or the demand for good and services, and reduce the uncertainty of its future outcomes such that an investment or a budget allocation can be executed for upcoming periods of time. Forecasting then can be described as the process of analyzing past and current observations of a variable to determine the variable's future outcomes. For example, a stock analyst can sustain an investment recommendation on a specific stock by forecasting its future behavior using the stock's past and current observations.

A set of historical observations indexed by time about the same phenomenon is known as time series. A time series is a sequence of observations in successive order collected through time at regular intervals. Examples of time series include the daily closing value of the Dow Jones Industrial Average, the monthly US Interest Rate, and trimestral Consumer Credit in the US to mention just a few. Given a target series of interest, also a *target series response* (TSR), a conventional model that utilizes some (or all) of its past information to provide accuracy in forecasting is the Auto-Regression (AR) model. In particular, the AR(4) which uses the 4 most recent observations of the target series for forecasting, is widely used in the macroeconomic literature and is considered a benchmark in the macroeconomic forecasting literature despite the existence of other conventional methods such as Auto-Regression Moving Average (ARMA) model.

However, nowadays, given the gigantic amount of information gathered around the world, data sets contain millions, even billions of entries, and consequently very large number of time series are reachable. These data sets typically contain a number of time series that even exceed the number of time point observations in the series, and consequently high dimensional time series data sets are ordinary. Specifically, high dimensional time series data sets can be found in finance and economics. With banks and government entities recording every monetary transaction during the last decade or more, it is hardly surprising these data sets are gigantic. High dimensional time series analysis

is commonplace in many fields including, among others, finance, economics, environmental and medical studies. For example, understanding the dynamics of the returns of large number of assets is the key for asset pricing, portfolio allocation, and risk management.

Moreover, modern financial and economic time series data often contain multiple categories, each containing a large amount of series with each series belonging to one of them. As a result, a target series response (TSR) corresponds to one of the categories. Usually, time series are classified based on the industry they belong, or the portion of the economy they represent, therefore it is expected that series be grouped based on a specific criterion. For instance, series containing information about the exchange rates in the US may be grouped in one category, those series related to the gross domestic product (GDP) should be grouped in another category, and so forth. The high dimensional macroeconomic time series data analyzed in Stock and Watson [62] grouped the series based on their association with several aspects of the US economy.

Clearly, one may consider using all the information at hand to forecast the target series. For example, in forecasting the stock price of a company, not only the past observations of the TSR, but also the past observations in other series in the same category and other categories can be used for forecasting. Thus, two sources of information are available. On one hand, one wants to take advantage from the past observations of the target series. To that end, one can make use of the Auto-Regression model, more specifically the AR(4). On the other hand, benefiting from all the other series represents a difficult task since a high dimensional data involves two difficulties for forecasting.

1.2 Challenges for forecasting in high dimensions

Forecasting with high dimensional multi-category time series data poses two difficulties. The **first difficulty** arises given the many time series nuisance that exist within the series available. From a philosophical viewpoint (i.e. following the Occam's Razor principle), in statistical modeling parsimonious models are preferred over complex model. Therefore, it is well understood that just a handful number of time series are useful to explain the future variability of the target series of interest. It is, however, challenging to identify the relevant subset of time series predictors and their appropriate lags that best explain the TSR.

The **second difficulty** appears from economic theories. Based on economic theories a few factors could explain a large fraction of the variance of many macroeconomic series. That is, a macroeconomic TSR of interest is typically driven by few latent factors constructed from some macroeconomic indices, see, e.g. Anderson [5], Sims [50], and Stock and Watson [65] which suggest the natural means of dimension reduction via factor construction. How to extract predictive factors (from a possibly subset of predictors) is, however, extremely difficult.

To give an illustration, consider the S&P 500 data (http : //finance.yahoo.com) which, contain raw stock price time series (monthly processed in this example) of the 500 most important public companies in terms of market capitalization for the last 30 years. The data is naturally grouped into 10 categories giving an example of high dimensional multi-category time series data. Now, suppose one wants to forecast the stock price of *Apple* which falls into the *information technology* (IT) category that contains 65 time series. Given the TSR, the conventional AR(4) can be applied. However, it seems reasonable to incorporate many other series as well. For example, **(i)** the past stock price observations of *Google* and *Intel* in the same category may be informative to forecast *Apple*, **(ii)** *Best Buy* and *Verizon* in the *Consumer Discretionary* and *Telecommunication Services* category, respectively may also seem to be relevant because both companies sell a considerable quantity of *Apple* products. This simply results in a high dimensional problem where constructing predictive factors and deleting nuisance time series simultaneously, are both meaningful and advantageous for accuracy to the forecasting task.

We are in an era of massive automated data collection, systematically obtaining many measurements, not knowing which ones will be relevant to the phenomenon of interest. Our task is to find needles (relevant variables) in a haystack (high dimensional data), teasing the relevant information (variables) out of a vast pile of glut (nuisance time series), see, Donoho in [12]. Modern statistical approaches such as shrinkage methods and dimension reduction techniques are critical to analyze high dimensional multi-category time series data. In this work, we study a dimension reduction via joint variable selection and factor extraction for forecasting a TSR of interest in high dimensional multi-category time series data. Variable selection may require some sparse-enforcing regularization techniques. Although the notion of sparsity gives rise to biased estimation in general, it has been proved to be very effective in many applications. It has been shown, see, e.g. Hastie, Tibshirani, and Friedman [23] that regularization techniques can often substantially reduce the variance

at the cost of a negligible increase in the bias. This can lead to substantial improvements in the accuracy, and is preferred when forecasting is a crucial goal. It is achieved by effectively identifying the important predictors and improving the model interpretability. Therefore, sparsity should be understood in a wider sense as a complexity reducer.

1.3 What do we propose to overcome both difficulties?

To overcome the second difficulty we propose to extract factors from the multi-category data, referred to as C-level factors. Concretely, given a *target category response*, denoted by TCR in this work, which contains the TSR our factors are extracted to best explain the TCR (and the series therein). However, the huge number of nuisance dimensions possesses a serious challenge since C-level factors shouldn't be constructed using all dimensions available.

There is a huge body of works for uncovering underlying factors or identifying relevant variables *alone*. Factors can be obtained by applying rank reduction techniques, for example, principal component analysis (PCA) [26], reduced-rank regression (RRR) [5, 27, 46], and dynamic factor models (DFM) [21, 59, 61]. All input variables are involved in constructing such factors and it is well understood that in our high dimensional setting, not every time series contribute to the TSR or TCR forecasting. Recently penalized variable selection techniques have been developed, see, for instance, Lasso [69], eNet [78], SCAD [15], among many others. However, such approaches ignore the low-rank concerns in financial and economic data. Our problem cannot be solved by implementing a rank reduction or variable selection technique on their own. Even more, as shown in Bunea, She and Wegkamp [10], it is prohibitive by implementing plain variable selection, regardless of the rank of the system, since it may overkill factor construction and consequently, ignore the rank constraint in building the forecasting model. The low rankness offers another type of parsimony and must be taken into account. Therefore, it seems plausible to construct factors using only a small group of relevant predictors, or to perform simultaneous rank constrained and variable selection such that meaningful and interpretable variables factors can be obtained.

In this document we propose a novel sparse factor auto-regression (SFAR) to tackle the aforementioned challenges in high dimensional multi-category time series data and enhance the forecasting accuracy of the TSR. Dimension reduction in the multivariate setting is achieved by imposing a rank constraint and a cardinality constraint on the regression coefficient matrix which regulate

the number of C-level factors to be constructed and the number of time series to be selected, respectively. The doubly-constrained problem is optimized via an efficient iterative algorithm with a theoretical guarantee of convergence. Orthogonal C-level factors are then added to a model to assist the TSR in modeling and forecasting.

The supervised methodology proposed should outperform unsupervised procedures. Unlike unsupervised procedures, the SFAR methodology performs simultaneous variable selection and C-level factor construction in response to a specific group of series in the TCR. Therefore, predictive C-level factors are constructed to assist the forecasting of all the series in the TCR.

1.4 Outline

Chapter 2 provides a literature review of the relevant methods and techniques to our proposed methodology. Section 2.1 details some modern shrinkage methods for variable selection with a brief introduction of the classical methods. Some of the existing regularization methods are described for single, and group of parameters. Section 2.2 introduces factor analysis and provides a brief explanation of how latent factors are uncovered. Section 3.3 elucidates the supervised reduced rank-regression (RRR). Section 3.4 describes dynamic factor models (DFM) and compares it to our proposed methodology. Similitudes and differences between methodologies are given.

Chapter 3 details the proposed Sparse Factor Auto-Regression (SFAR) and its methodology. This chapter elucidates the logic behind the construction of the SFAR together with the restrictions and conditions that must be satisfied in order to use it. A short explanation on why the MLE estimator cannot be used in our problem to obtain a more accurate prediction is also given. The joint constraints imposed on the coefficient matrix of the regression model are explained in details to understand their uses and consequences. Model notation is given and the forecasting methodology provided.

Chapter 4 gives details about the computational component of the model and the implemented algorithm. A new version of the selectable reduced rank-regression (SEL-RRR) estimator with both constraints is given and supported. The multivariate quantile thresholding rule is proposed to provide selectability power while a trick to decompose the matrix coefficient is also given. To provide evidence of how both constraints are simultaneously satisfied in computation the Progressive Rank-Cardinality Constrained Regression (PRCCR) algorithm is exhibited and explained.

Chapter 5 provides data analysis on synthetic and real world data. The forecasting methodologies implemented in this study are described. A synthetic data is generated and analyzed to validate theoretical findings. Then, a macroeconomic data set is analyzed to understand the behavior of SFAR in practice. Results are shown for $\mathbf{\Gamma} = \boldsymbol{I}$ and different setup of $\mathbf{\Gamma}$.

Chapter 6 conveys the conclusion and future work.

CHAPTER 2

LITERATURE REVIEW

In this chapter some of the old model selection selection criteria are quickly revised as an introductory subject to describe a few of the most representative shrinkage methods in the literature. Classic model selection criteria are still useful to pick models that describe an adequate and interpretable description of the relationship between variables. However, when interpretability and prediction accuracy are the objectives for analyzing high dimensional data these criteria may not provide a good model estimation.

Shrinkage methods, also known as regularization methods, were developed in an effort to select models that provide interpretability and accuracy in prediction by either retaining or discarding variables. These days, shrinkage methods can be implemented for univariate or multivariate models. Dimension reduction techniques are also considered as useful approach to reduce the model complexity and achieve both objectives. One may think that shrinkage methods and dimension reduction techniques work toward the same goal but it is not necessarily true. For instance, the former achieves variable selection upon the shrinkage method used, while the latter never accomplishes selectability.

2.1 Modern methods for model selection

Before describing shrinkage methods some classical model selection methods are introduced. Model selection is intended to find a subset of variables that are sufficient to best describe the relationship between inputs and output in the model. It is selected by comparing its performance with other models in order to choose the best one. Each model possesses either the same or different number of variables, and typically variables are selected given their explanatory power. It is expected that predictors with low explanatory power are removed. In other words, only the subset of relevant predictors to the variable of interest should remain. Model selection in statistical modeling argues in favor of the smallest model that best fits the data. Therefore, it is essential to an adequate inference.

The problem in classical model selection methods is to examine certain subsets and select the 'best subset' which either maximizes or minimizes an appropriate criterion. Two subsets are obvious: the best single variable and the complete set of variables. The problem lies in selecting an intermediate subset that is better than both of these extremes. Some popular methods for choosing the 'best subset' regression are Best Subset selection, Stepwise selection, Forward selection, and Backward selection. These methods result in the selection of the best model from a set of models as indicated by a pre-specified selection criteria. For example, the first method may use either the adjusted R-squared or Mallows' Cp [38] criteria. For the last three methods, the test-statistics uses either the F-statistics for a continuous dependent variable, or the G-statistics for a binary dependent variable. Consequently, all these methods can be seen as model comparison criteria rather than variable selection criteria since they will provide the model that best fits the data based on its criteria comparison. However, these methods do not need to optimize any reasonable criterion.

Other model selection criteria in linear regression analysis are available in the literature. For instance, the Akaike's AIC [3, 4], which is a large-sample approximation result, provides an asymptotically unbiased estimator of the expected Kullback-Leibler discrepancy between the generating and the fitted approximating model. The AIC is used to compare the candidate models which are already given, and select the best model in the candidate set which may *not* include the true model. Another example is the Schwarz's BIC [51], built on the large-sample (asymptotic) approximation of the Bayesian method provides a large-sample estimator of a transformation of the Bayesian posterior probability associated with the approximating model. The BIC also compares the given models and penalizes the number of parameters more strongly than the AIC. In addition, considerable efforts have been made by many researchers to obtain better criteria, by either modifying the existing ones or creating innovative new measures. Examples include Knight and Tibshiriani, [70], Pauler [42], Zheng and Loh [77], and Shen and Ye [55].

Traditional model selection criteria, such as Mallowss Cp, AIC, and BIC involve a combinatorial optimization problem, which is NP-hard, with computational time increasing exponentially with the dimensionality of the problem, see, Fan [16]. Another major drawback arises because parameter estimation and model selection are two different processes which can result in instability, see, Breiman [7] and complicated stochastic properties, see, Fan and Li [15].

In these times, dimension reduction methods become a necessity in the extreme case in which the number of observations T is less than the number of candidate variables p, also known as high dimensional problems. Due to recent advances in data collection and computing capability, modern statistical applications often involve a large number of candidate variables, making variable selection even more important for the purpose of building an effective regression model. In economics and finance, variable selection procedures can be used to choose a small subset of predictors from a large set of potentially useful variables, however, the performance of these methods ultimately rest on the few variables that are chosen, where "few" is usually 10 or less, see, Stock and Watson [64].

Variable selection is key in modeling time series. To address variable selection we may require some regularization technique, or a penalty function. One of the most active statistical areas that tries to give response to the challenge of variable selection is related to the penalized regression models family. From the seminal paper, presenting the ℓ_1 penalized regression model, better known as the lasso model, see, Tibshirani [69], subsequent or competitive approaches have grown in importance determining a very active area in variable and model selection such as LARS model, Efron *et al* [14], and Elastic Net, Zou and Hastie [78], among others.

Modeling high dimensional data is challenging. For a continuous response variable $y \in \mathbb{R}$, a simple yet useful approach is given by the linear model

$$y_t = \sum_{j=1}^{p} \boldsymbol{\beta}_j x_t^{(j)} + \epsilon_t \quad \text{for} \quad t = 1, \ldots, T, \tag{2.1.1}$$

where $\epsilon_1, \ldots, \epsilon_T$ are *i.i.d.*, independent of x_t, and $E[\epsilon_t] = 0$.

For simplicity and without loss of generality, we usually assume that the intercept is zero and that all covariates are centered and measured on the same scale. Both of these assumptions can be approximately achieved by empirical mean-centering and scaling with the standard deviation. The only unusual aspect of the linear model (2.1.1) is the fact that $p \gg n$. The vector notation of model (2.1.1) is

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon} \tag{2.1.2}$$

with response vector $\boldsymbol{y}$ of size $T \times 1$, design matrix $\boldsymbol{X}$ of dimensions $T \times p$, parameter vector $\boldsymbol{\beta}$ of magnitude $p \times 1$, and the $T \times 1$ vectors of errors $\boldsymbol{\epsilon}$. Assuming an univariate model, we minimize

$$PLS(\boldsymbol{\beta}, \lambda) = (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})^{\mathsf{T}}(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}) + \lambda \sum_{j=1}^{p} P(\beta_j). \tag{2.1.3}$$

where $\lambda \sum_{j=1}^{p} P(\beta_j)$ provides a very general sense of a penalized method since P can take any penalty form from those available in the literature.

2.1.1 Ridge regression

Ridge regression was introduced by Hoerl and Kennard [25]. It minimizes over $\boldsymbol{\beta}$, a criterion of the form

$$\hat{\boldsymbol{\beta}}^{(ridge)} = \underset{\boldsymbol{\beta}}{\operatorname{argmin}} \frac{1}{2} ||\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}||_2^2 + \lambda \sum_{j=1}^{p} \beta_j^2 \tag{2.1.4}$$

for a ridge parameter $\lambda \in [0, \infty]$. As λ ranges through $[0, \infty]$ the solution $\boldsymbol{\beta}(\lambda)$ traces out a path in $\mathbb{R}^p$. With an appropriate parameter tuning strategy, ridge can achieve good estimation and prediction accuracy. Adding the penalty reduces the variance of the estimate $\boldsymbol{\beta}(\lambda)$ while introducing bias. The second term in (2.1.4) is also called ℓ_2-penalty and shrunk coefficients toward zero but never make them exactly zero. That is, ridge regression does not promote sparsity (i.e. it does not perform variable selection) and may not produce an interpretable model. The analytic solution of ridge regression is

$$\hat{\boldsymbol{\beta}}^{(ridge)} = (\boldsymbol{X}^\top \boldsymbol{X} + \lambda \boldsymbol{I})^{-1} \boldsymbol{X}^\top \boldsymbol{y}. \tag{2.1.5}$$

Adding the term $\lambda \boldsymbol{I}$ results in a regular and invertible matrix even in the presence of multicolinearity. Ridge regression cannot promote sparsity because the range of solutions from (2.1.5) does not include zero even in the case where the term $\lambda \boldsymbol{I} = 0$. However, ridge regression can improve OLS, in the sense of MSE, by choosing a proper ridge parameter. This is, as $\lambda \downarrow 0$ we obtain the least squares solution, and for the case $\lambda \uparrow \infty$ we have $\hat{\boldsymbol{\beta}}_{\lambda=\infty}^{(ridge)} = 0$ (i.e. the intercept-only model). Typically, we just need to get a suitable scale of λ to guarantee good performance in regression and for that a very fine searching grid can be used.

2.1.2 Bridge regression

Two of the most important penalty functions ridge and lasso are special cases of the bridge regression penalty proposed by Frank and Friedman [20]. The penalized bridge regression is

$$\hat{\boldsymbol{\beta}}^{(bridge)} = \frac{1}{2} ||\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}||_2^2 \ + \lambda \sum_{j=1}^{p} |\beta_j|^\gamma. \tag{2.1.6}$$

The bridge regression estimate of $\boldsymbol{\beta}$ is defined as the $\hat{\boldsymbol{\beta}}$ that minimizes the penalized residual of square. Ridge regression, is equivalent to the bridge regression with $\gamma = 2$. For models derived

from the bridge regression we have that the solution of the convex ℓq-penalty does not satisfy the sparsity condition for $q > 1$. Therefore, the hard thresholding penalty satisfies the mathematical conditions to enforce sparsity and promote variable selection.

2.1.3 Lasso

The least absolute shrinkage and selection operator (lasso) proposed by Tibshirani [69] is the bridge regression with $\gamma = 1$. Lasso is a shrinkage method like ridge, with subtle but important differences. Lasso does both continuous shrinkage and automatic variable selection simultaneously. The lasso estimates are defined by

$$\hat{\boldsymbol{\beta}}^{(lasso)} = \underset{\boldsymbol{\beta}}{\operatorname{argmin}} \frac{1}{2}||\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}||_2^2 + \lambda \sum_{j=1}^{p} |\beta_j| \tag{2.1.7}$$

where λ is a nonnegative regularization parameter and the second term is known as the ℓ_1-penalty. When (2.1.4) and (2.1.7) are compared, the difference exists in the type of penalty implemented: the ℓ_2 ridge penalty $\sum_1^p \beta_j^2$ is replaced by the ℓ_1 lasso penalty $\sum_1^p |\beta_j|$. This latter constraint makes the solutions nonlinear in $\boldsymbol{y}$, and there is no close form expression as in ridge regression. The main benefit of lasso is that it can find sparse solutions, ones in which some or even most of the β_j are zero. Sparsity is desirable for interpretation.

When the predictor matrix is orthonormal, under equation (2.1.7), the lasso estimator can be written as $\hat{\beta}^{(lasso)} = (\boldsymbol{X}^\mathsf{T}\boldsymbol{X})^{-1}(\boldsymbol{X}^\mathsf{T}\boldsymbol{y} - \frac{\lambda}{2})_+$ where $(\cdot)_+$ denotes the nonnegative part of the entry. The regularization parameter λ controls the weight of the ℓ_1-penalty and the number of nonzero coefficients. As in Ridge, setting $\lambda = 0$ changes lasso to the OLS regression setup. On the other hand, as $\lambda \to \infty$, all the coefficients of $\hat{\boldsymbol{\beta}}$ shrink to zero, resulting in the null model.

Lasso has many desirable features that have made it a popular regression algorithm. It is at the same time a shrinkage estimator of $\boldsymbol{\beta}^{(OLS)}$ (OLS coefficients are shrunk towards the origin) and a variable selection technique [28], performing a kind of continuous subset selection, see, e.g. Hastie, Tibshirani, and Firedman [23]. Lasso continuously shrinks the coefficients toward zero as the regularization parameter increases, and some coefficients are shrunk to exactly zero reducing the model complexity. Lasso is computationally efficient for high dimensional data, see, Efron [14]. The entire lasso sequence of path can be generated by a slight modification of the LAR algorithm (also LARS), which is a procedure that efficiently combines lasso, forward-stagewise and LAR. The fact

that solution paths of LARS and lasso are piecewise linear gives them tremendous computational advantages when compared with other methods.

Many researchers have studied the properties of lasso. Donoho [13] proved the near minimax optimality of soft thresholding, which is a lasso shrinkage estimate with orthonormal predictors matrix. In addition, variable selection consistency has been studied. Zhao and Yu [76] proved that lasso is variable selection consistent when the irrepresentable condition holds. The irrepresentable condition, which depends mainly on the covariance of the predictor variables, states that lasso selects the true model consistently if and (almost) only if the predictors that are not in the true model are irrepresentable by predictors that are in the true model. They proved that a variable selection procedure is consistent if the probability of selecting exactly the set of variables with nonzero coefficients, that is identifying the subset $\{j : \beta_j \neq 0, j = 1, \cdots, p\}$, converges $\to 1$, and if the probability of point mass at $\beta_j = 0$ is equal to 1 where $j = 1, \cdots, p$. Knight and Fu [32] showed that with a fixed p, under certain conditions, lasso has the model selection consistency. Zhang and Huang [75] showed that it was not variable selection consistent without proper assumptions. Leng at al. [33] showed that when the prediction accuracy was used as a criterion to choose the regularization parameter λ, lasso was not variable selection consistent in general.

As variable selection becomes increasingly important in modern data analysis, the lasso is much more appealing due to its sparse representation. Although lasso has shown success in many situations, is has some limitations. Consider the following three scenarios:

1. In the $p > T$ case, lasso selects at most T variables before it saturates, because of the nature of the convex optimization problem. This seems to be a limiting feature for a variable selection method. Moreover, lasso is not well-defined unless the bound on the ℓ_1-norm of the coefficients is smaller than a certain value.

2. If there is a group of variables among which the pairwise correlations are very high, then lasso tends to select an arbitrary variable from the group and does not care which one is selected.

3. Usually, for $T > p$ situations, if there exist medium or lower correlations among predictors, it has been empirically observed that the prediction performance of lasso is dominated by the ridge regression [69].

Scenarios (1) and (2) declared by Zou and Hastie in [78] make lasso an inappropriate variable selection method in some situations.

2.1.4 Elastic net

Elastic net, proposed by Zou and Hastie in [78], proposes another regularization and variable selection method that is a convex combination of lasso, ℓ_1, and ridge, ℓ_2. Elastic net is particularly useful when the number of predictors p is much larger than the number of observations T. Especially in high dimensions elastic net is able to handle collinearity among predictors and select groups of correlated variables, also known as the grouping effect. The penalized regression model using elastic net is

$$\hat{\boldsymbol{\beta}}^{(elastic\ net)} = \underset{\boldsymbol{\beta}}{\operatorname{argmin}} \frac{1}{2}||\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}||_2^2 + \lambda_1 \sum_{j=1}^{p} |\beta_j| + \frac{\lambda_2}{2} \sum_{j=1}^{p} \beta_j^2 \tag{2.1.8}$$

where λ_1 and λ_2 are two nonnegative regularization parameters. The ℓ_1 part of the model generates a sparse model, and the ℓ_2 part removes the limitation on the number of selected variables, encourages grouping effect, and stabilizes the ℓ_1 regularization path. Combining ℓ_1 and ℓ_2 penalties tends to give a result in between, with fewer regression coefficients set to zero than in a pure ℓ_1 setting, and more shrinkage of the other coefficients. The amount of shrinkage is determined by tuning parameters λ_1 and λ_2. A value of zero always means no shrinkage (= maximum likelihood estimation) and a value of infinity means infinite shrinkage (= setting all regression coefficients to zero). Empirical evidence shows the naive elastic net does not perform satisfactorily since two shrinkage procedures (ridge and lasso) are included in it. Double shrinkage introduces unnecessary bias. Zou and Zhang [79] proposed the adaptive elastic net that combined the strengths of the quadratic penalty and the adaptive lasso shrinkage.

2.1.5 ℓ_0-penalty & $\ell_0 + \ell_2$-penalty

In some cases ℓ_0-penalty may be utilized to enforce sparsity. Intuitively the ℓ_0 penalty $||\boldsymbol{\beta}||_0 = \sum_{j=1}^{p} \mathbb{I}(|\beta_j| > 0)$, which is the limit of the $\ell\gamma$-penalty as $\gamma \to 0$, penalizes the number of nonzero coefficients in the model directly and thus is desirable for variable selection. The ℓ_0-penalty minimization problem is then

$$\min_{\boldsymbol{\beta}} \frac{1}{2}||\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}||_2^2 + \frac{\lambda^2}{2}||\boldsymbol{\beta}||_0 \tag{2.1.9}$$

where $\boldsymbol{\beta}$ is a vector. However, due to its nonconvexity and discontinuity at the origin, it is very hard to solve the corresponding optimization problem. Recently, She [52] solved the exact ℓ_0-penalized constrained problem via the Thresholding-based Iterative Selection Procedures **TISP**.

A perhaps better regularization approach is to fuse the ℓ_0 and ℓ_2 penalties in one equation. This approach given by Shè [53] assumes that two objectives, aligned with Occam's razor principle, are involved in the task of statistical learning and modeling: **1**) accurate prediction, and **2**) parsimonious model representation (or interpretability). Seen from **1**), a ridge, ℓ_2-penalty, is desired to account for noise and collinearity in the data. However, it never encourages sparsity. In the elastic net which uses a linear combination of the ℓ_1-penalty and the ℓ_2-penalty, the ridge part may counteract the parsimony requirements **2**) in the estimate, see, Zou and Has [78]. Yet the ℓ_1-norm already provides the tightest convex relaxation of the ℓ_0-norm. Therefore, to maintain accuracy and promote sparsity, one must take into account nonconvex penalties such as those of type $\ell_0+\ell_2$. The hard-ridge penalty or $\ell_0 + \ell_2$-penalty is represented in the following regression model

$$\hat{\boldsymbol{\beta}}^{(HR)} = \underset{\boldsymbol{\beta}}{\operatorname{argmin}} \ \frac{1}{2}||\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}||_2^2 + \frac{\lambda_2}{2}\sum_{j=1}^{p}|\beta_j| + \frac{\lambda_0^2}{2}||\boldsymbol{\beta}||_0 \tag{2.1.10}$$

where λ_2 is the ridge parameter promoting accuracy in prediction, and λ_0 represents the hard regularization parameter in charge of promoting sparsity in the vector of coefficients.

2.1.6 Group lasso

In real world applications predictors are often naturally grouped. For example, predictors may be grouped by multi-level categorical variables, or using a linear, quadratic or other nonlinear effect of the variables. Pursuing a between-groups sparsity is desired in those situations. That is, if one wants to remove a variable one will have to kill its predictor group as a whole rather than a single coefficient. In the analysis of multivariate time series, for instance, large amount of series are used for forecasting and thus, it seems adequate to discard groups of irrelevant predictors. Multicollinearity is also a common problem in the analysis of financial or economics large data sets. To overcome this problem, a group-level variable selection and regularization function seems convenient for selecting significant predictors and reducing the variance of the model. To address this grouping concern, Yuan and Lin [73] proposed a group lasso penalty which is intermediate between the ℓ_1-penalty used in the lasso and the ℓ_2-penalty used in ridge regression.

Consider a vector response model

$$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{C} + \boldsymbol{E} \tag{2.1.11}$$

where $\boldsymbol{Y}$ is a $T \times n$ matrix of response variables, $\boldsymbol{X}$ is the $T \times p$ matrix of predictors, $\boldsymbol{C}$ is a $p \times n$ matrix of unknowns, and $\boldsymbol{E}$ represents the $T \times p$ matrix of errors with independent entries. Let $\boldsymbol{c}_j \in \boldsymbol{C}$ represent the row-vector coefficient corresponding to a predictor $\boldsymbol{x}_i \in \boldsymbol{X}$ for $i = j$. At every time point t we have a response vector. Assume predictors are naturally grouped such that all p predictors fall into K groups. i.e. the design matrix is grouped into K blocks: $\boldsymbol{X} = (\boldsymbol{X}_1, \ldots, \boldsymbol{X}_K) \in \mathbb{R}^{T \times p}$, so that in model selection one wants to keep or kill a group of predictors as a whole. The predictor groups do not overlap but group sizes can be different. When there are p groups, each being a singleton, the model reduces to the common ungrouped case. The criterion of the group P_k-penalized is defined by

$$\frac{1}{2}||\boldsymbol{Y} - \boldsymbol{X}_k\boldsymbol{C}_k||_2^2 + \sum_{k=1}^{K} P_k(||\boldsymbol{C}_k||_2; \lambda_k) \tag{2.1.12}$$

where $\boldsymbol{C}_k$ are the coefficients associated with $\boldsymbol{X}_k$, and P_k is the penalty functions that can be discrete, nonconvex, and non differentiable at zero. Since the dimension p is greater than the sample size n there may exist a large number of nuisance features. Conveniently, groups can be represented by a singleton. Then, criterion (2.1.12) can be described as follows

$$\frac{1}{2}||\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{C}||_2^2 + \sum_{j=1}^{p} P(||\boldsymbol{c}_j||_2; \lambda) \tag{2.1.13}$$

where $\boldsymbol{X} = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_p)$, and $\boldsymbol{C} = (\boldsymbol{c}_1^\mathsf{T}, \ldots, \boldsymbol{c}_p^\mathsf{T})$ are defined in (2.1.11). Penalty P affects each vector coefficient in $\boldsymbol{C}$ rather than a group of coefficients in a particular block. In general, $P(\boldsymbol{C}; \lambda)$ represents the penalty with λ as the regularization parameter and $\boldsymbol{C}$ is assumed to be sparse. Usually, P is assumed to be an additive penalty in the sense that $P(\boldsymbol{C}; \lambda)$ is obtained by a vector P: $P(\boldsymbol{C}; \lambda) = \sum P(\boldsymbol{c}_j; \lambda)$.

Directly optimizing (2.1.13) can be tricky for a given penalty function. For example, the ℓ_0-penalty $\frac{\lambda^2}{2}||\boldsymbol{C}||_0 = \frac{\lambda^2}{2}|\{i : \boldsymbol{c}_i \neq 0\}|$ (where $|\cdot|$ is the set cardinality) is used for building a parsimonious model that is discrete and nonconvex. A class of estimators for grouped predictors is defined via an arbitrarily given thresholding rule to solve (2.1.13) for essentially any P, see, Shè [53]. Shè also provided the computational component to address nonconvex group of penalties for any GLM.

Throughout this study, a vector response or predictor is grouped as a special case of the univariate case where the specific grouping manner is the indexed time point observations from $t = 1, \ldots, T$

for the response variable. Instead, a predictor is conformed by a vector of the same length as the response but embracing lagged observations.

In the context of variable selection, lasso is often thought of as a convex surrogate for best-subset selection, see, Mazumder, Friedman and Hastie [40]. For grouped predictors Shè [53] provided a solution to the grouped ℓ_0-penalty $\sum_{j=1}^{p} \frac{\lambda^2}{2} 1_{||\boldsymbol{c}_j||\neq 0}$ which can attain more group sparsity than the group lasso. Bringing the idea of the ℓ_0 grouped predictors the following problem is minimized

$$\min_{\boldsymbol{C}} ||\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{C}||_2^2 + \frac{\lambda^2}{2}||\boldsymbol{C}||_0. \tag{2.1.14}$$

where $||\boldsymbol{C}||_0 = \sum_{j=1}^{p} ||\boldsymbol{c}_j||_0$.

A further rank-constraint group lasso is read in Bunea, She, and Wegkamp [10] where an additional rank-constraint is added in (2.1.13). Similarly, in this group setup the ℓ_0 regularization criterion is superior to that of ℓ_1 regularization since ℓ_0 is the ideal way to promote (group) sparsity. The ℓ_0 group constraint is defined by: $\frac{\lambda^2}{2}\sum 1_{\boldsymbol{c}_j\neq 0} = \frac{\lambda^2}{2}\#\text{ nz}(\boldsymbol{C}) = \frac{\lambda^2}{2}||\boldsymbol{c}_j||_0 := \frac{\lambda^2}{2}\times$ the set cardinality of $\boldsymbol{C}$.

In similar fashion, the hard-ridge group penalty, or $\ell_2 + \ell_0$-penalty defined by Shè [53] does simultaneous selection and shrinkage with a thresholding parameter λ and a ridge parameter η,

$$\min_{\boldsymbol{C}} ||\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{C}||_F^2 + \frac{\lambda^2}{2(1+\eta)}||\boldsymbol{C}||_{2,0} + \frac{\eta}{2}||\boldsymbol{C}||_F^2 \tag{2.1.15}$$

where $\|\boldsymbol{C}\|_{2,0} := \sum_j 1_{\boldsymbol{c}_j\neq 0}$. This hard-ridge penalty offers both selection and shrinkage into regularization, interplaying with each other during the iteration of nonorthogonal designs. As explained above the cardinality of $\boldsymbol{C}$ is essential regardless of the size of p. Both regularizations simultaneously provide variable selection (interpretability) and decorrelated groups (accuracy).

2.2 Factor analysis

In factor analysis the variables $y_1, y_2, \ldots, y_n$ are represented as linear combinations of a few random variables $f_1, f_2, \ldots, f_r$ $(r < n)$ called factors. Factors are underlying constructs or latent variables that explain an unobserved relationship between variables. Factor analysis is a dimension reduction tool which goal is to reduce the redundancy among the variables by using a smaller number of factors. When a particular subset of variables are highly correlated among themselves, then there may be an underlying variable (latent factor) that represents the variables in the subset.

If the other variables can be similarly grouped into subsets with a like pattern of correlations, then a few factors can represent these groups of variables. In economics, for example, the variables income, dollars in saving, and home value might be grouped to represent the concept of the economic status of research subjects. There has been a rapidly growing literature of applications of factor analysis techniques to economic time series. Examples of applications include forecasting [64], stock market returns [36], and interest rates [35].

The factor analysis model expresses each variable as a linear combination of underlying *common factors* $f_1, f_2, \ldots, f_r$ with an accompanying error term to account for that part of the variable that is unique. For $y_1, y_2, \ldots, y_T$ in any observation vector $\boldsymbol{y}$ with neglected mean, the model is as follows

$$\begin{aligned} y_1 &= \phi_{11} f_1 + \phi_{12} f_2 + \ldots + \phi_{1r} f_r + \epsilon_1 \\ &\vdots \\ y_T &= \phi_{T1} f_1 + \phi_{T2} f_2 + \ldots + \phi_{Tr} f_r + \epsilon_T. \end{aligned} \tag{2.2.1}$$

Ideally, r should be substantially smaller than T; otherwise we have to achieve a parsimonious description of the variables as functions of a few underlying factors. The coefficients ϕ_{ij} are called *loadings* and serve as weights, showing how each y_i individually depends on the f's. With appropriate assumptions, ϕ_{ij} indicates the importance of the jth factor f_j to the ith variable y_i and can be used in interpretation of f_j. For example, we describe or interpret f_2 by examining its coefficients, $\phi_{12}, \phi_{22}, \ldots, \phi_{n2}$. The larger loadings relate f_2 to the corresponding y's. From these y's, we infer a meaning of description of f_2. In a sample version it is convenient to have f_j orthogonal to each other. Encouraging orthogonality among factors guarantees that the information contained in one factor cannot be covered by any other factor in such a way that redundancy is avoided.

Most linear factor models used in economics and finance are macroeconomic, fundamental, and statistical factor models [31]. Statistical factor models, are obtained through a logical and mathematical analysis of the given variables. Factors are typically determined using reduced rank analysis via reduced-rank regression (RRR) or principal component analysis (PCA) which is a special case of RRR.

2.3 Reduced-Rank Regression - RRR

The reduced-rank regression model [5, 27, 46] achieves dimension reduction through restricting the rank of the coefficient matrix without performing variable selection. Anderson [5] was the first

to propose and study the RRR problem, for the case where $\boldsymbol{X}$, the set of predictor variables, is fixed. Izenman [27], introduced the term *reduce-rank regression* for this class of models and provide further study on the estimates. The reduced-rank regression model and its statistical properties were examined further by Robinson [48, 49], Rao [43], Tso [71], and Davies and Tso [11]. The monograph on reduced rank regression by Reinsel and Velu [46] has an excellent, comprehensive account of more recent development and extensions of the model including time series approach.

Consider a multivariate vector time series regression model of the form

$$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{C} + \boldsymbol{E} \tag{2.3.1}$$

as defined in (2.1.11). The usual description in (2.3.1) assumes that $\boldsymbol{C}$ is of full rank and can be estimated using least square or maximum likelihood estimation methods where each of the n individual response variables is regressed separately on the predictor variables, see, for instance, Izenman [28] for a discussion of this phenomenon. Estimators restricted to have rank less than or equal to a fixed number $r \leq p \wedge n$ were introduced to remedy this drawback. To undertake this problem $\boldsymbol{C}$ can be well approximated by a low rank matrix. Thus, in many practical situations, as in statistical applications, it is necessary to reduce the number of unknowns and find a matrix of smaller rank that comprises a large portion of the information contained in a matrix of larger rank. From an econometric approach, $\boldsymbol{Y}$ contains the multidimensional response time series of interest, and $\boldsymbol{X}$ includes a multidimensional time series of potential explanatory predictor variables. $\boldsymbol{C}$ is now the matrix containing the vector of coefficients of the series in $\boldsymbol{X}$ and $\boldsymbol{E}$ represents the noise.

Reduced-rank regression admits the $rank(\boldsymbol{C})$ is deficient and imposes a rank constraint in $\boldsymbol{C}$ assuming

$$rank(\boldsymbol{C}) = r \leq min(p, n). \tag{2.3.2}$$

Let $r(\boldsymbol{C})$ (or simply r) denote the $rank(\boldsymbol{C})$, nz($\boldsymbol{C}$) (or simply $J(\boldsymbol{C})$) denotes the index set of the nonzero rows of $\boldsymbol{C}$ and $|J(\boldsymbol{C})|$ its cardinality. The imposed rank constraint reduces the number of parameters in $\boldsymbol{C}$ to $r(T + |J(\boldsymbol{C})| - r)$ a substantially lower amount than the sample size T. Furthermore, as we can always reduce $\boldsymbol{X}$ of rank r to an $T \times r$ matrix with r orthogonal columns in $\mathbb{R}^T$ that span the same space as the columns of $\boldsymbol{X}$, the corresponding coefficient matrix will have r rows, so we can always assume that $|J(\boldsymbol{C})| \leq r$. If $\boldsymbol{C}$ is of full rank with no zero rows, the total number of parameters to be estimated reverts back to nr.

Another result from (2.3.2) implies that $\boldsymbol{C}$ can be written as a product of two lower dimensional matrices that are of full rank [27, 46]. $\boldsymbol{C}$ can be expressed as

$$\boldsymbol{C} = \boldsymbol{C}_1\boldsymbol{C}_2 \tag{2.3.3}$$

where $\boldsymbol{C}_1$ is a $p \times r$ tall matrix and $\boldsymbol{C}_2$ is a $r \times n$ fat matrix. Matrices $\boldsymbol{C}_1$ and $\boldsymbol{C}_2$ are essential in applying this methodology to estimate factors. Note that r columns in $\boldsymbol{C}_1$ can be viewed as a basis for the columns space of $\boldsymbol{C}$, while the r rows of $\boldsymbol{C}_2$ form a basis for the row space. Then model (2.3.1) can be written as follows

$$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{C}_1\boldsymbol{C}_2 + \boldsymbol{E} \tag{2.3.4}$$

where $\boldsymbol{X}\boldsymbol{C}_1$ is of reduced dimensions with only r components. Estimators of the coefficients matrix $\boldsymbol{C}$ minimize the weighed sum of squares with given rank, say r, with a positive-definite matrix of weights $\boldsymbol{\Gamma}$. For clarity, these reduced-rank estimators of given rank r are denoted here as $\hat{\boldsymbol{C}}_r$. For a positive semi-definite weight matrix $\boldsymbol{\Gamma}$, the estimator of $\boldsymbol{C}$ of given rank r may be found minimizing the weighted contained problem

$$\min_{\boldsymbol{C}} ||(\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{C})\boldsymbol{\Gamma}^{1/2}||_F^2 \; s.t. \; r(\boldsymbol{C}) \leq r. \tag{2.3.5}$$

To compute $\boldsymbol{C}_r$ for any $\boldsymbol{\Gamma}$, a computational efficient procedure suggested by Anderson [5] may be used with the ML sample covariance $\hat{\boldsymbol{\Sigma}}_{\boldsymbol{XX}}, \hat{\boldsymbol{\Sigma}}_{\boldsymbol{XY}} = \hat{\boldsymbol{\Sigma}}_{\boldsymbol{YX}}$, and $\hat{\boldsymbol{\Sigma}}_{\boldsymbol{YY}}$ where the components of the covariance matrices are calculated as: $\hat{\boldsymbol{\Sigma}}_{\boldsymbol{XX}} = T^{-1}\boldsymbol{X}^\mathsf{T}\boldsymbol{X}$; $\hat{\boldsymbol{\Sigma}}_{\boldsymbol{XY}} = T^{-1}\boldsymbol{X}^\mathsf{T}\boldsymbol{Y} = \hat{\boldsymbol{\Sigma}}_{\boldsymbol{YX}}^\mathsf{T}$; and $\hat{\boldsymbol{\Sigma}}_{\boldsymbol{YY}} = T^{-1}\boldsymbol{Y}^\mathsf{T}\boldsymbol{Y}$. No assumptions are made on the invertibility of the sample covariance matrix $\hat{\boldsymbol{\Sigma}}_{\boldsymbol{XX}}$ and $\hat{\boldsymbol{\Sigma}}_{\boldsymbol{YY}}$ unless otherwise noted. Assume $\hat{\boldsymbol{\Sigma}}_{\boldsymbol{XX}}$ is nonsingular. The step for computing $\boldsymbol{C}_r$ of given rank r are then:

1. Find the (normalized) eigenvectors $\hat{\boldsymbol{V}}_r = (\hat{\boldsymbol{v}}_1, \hat{\boldsymbol{v}}_2, \cdots, \hat{\boldsymbol{v}}_r)$, where $\hat{\boldsymbol{v}}_j$ is the eigenvalue corresponding to the j-th largest eigenvalue of the symmetric matrix

$$\hat{\boldsymbol{R}} = \boldsymbol{\Gamma}^{1/2}\hat{\boldsymbol{\Sigma}}_{\boldsymbol{YX}}\hat{\boldsymbol{\Sigma}}_{\boldsymbol{XX}}^{-1}\hat{\boldsymbol{\Sigma}}_{\boldsymbol{XY}}\boldsymbol{\Gamma}^{1/2} \tag{2.3.6}$$

2. Calculate the (full-rank) least-square estimator $\hat{\boldsymbol{C}} = \hat{\boldsymbol{\Sigma}}_{\boldsymbol{XX}}^{-1}\hat{\boldsymbol{\Sigma}}_{\boldsymbol{XY}}$. Then form

$$\hat{\boldsymbol{C}}_1 = \hat{\boldsymbol{C}}\boldsymbol{\Gamma}^{1/2}\hat{\boldsymbol{V}}_r \tag{2.3.7}$$

and

$$\hat{\boldsymbol{C}}_2 = \boldsymbol{V}_r^\mathsf{T}\boldsymbol{\Gamma}^{-1/2}. \tag{2.3.8}$$

3. Compute the estimator

$$\hat{\boldsymbol{C}}_r = \hat{\boldsymbol{C}}_1 \hat{\boldsymbol{C}}_2 \tag{2.3.9}$$

where $\hat{\boldsymbol{C}}_1 = \hat{\boldsymbol{C}}[, 1 : r]$ denotes the matrix $\boldsymbol{C}_1$ retaining only its first r columns and $\hat{\boldsymbol{C}}_2 = \hat{\boldsymbol{C}}[1 : r,]$ denotes the matrix $\boldsymbol{C}_2$ retaining only the first r rows.

Note that, even if the product in (2.3.1) is identified, matrices $\hat{\boldsymbol{C}}_1$ and $\hat{\boldsymbol{C}}_2$ are not unique unless we impose further normalizing conditions, since for any nonsingular $r \times r$-dimensional matrix $\boldsymbol{M}$ it holds that $\hat{\boldsymbol{C}}_r = \hat{\boldsymbol{C}}_1 \hat{\boldsymbol{C}}_2 = \hat{\boldsymbol{C}}_1 \boldsymbol{M}^{-1} \boldsymbol{M} \hat{\boldsymbol{C}}_2 = \tilde{\boldsymbol{C}}_1 \tilde{\boldsymbol{C}}_2$ where $\tilde{\boldsymbol{C}}_1 = \hat{\boldsymbol{C}}_1 \boldsymbol{M}^{-1}$ and $\tilde{\boldsymbol{C}}_2 = \boldsymbol{M} \hat{\boldsymbol{C}}_2$. Hence, $\hat{\boldsymbol{C}}_1$ and $\hat{\boldsymbol{C}}_2$ yield a unique decomposition of $\hat{\boldsymbol{C}}_r$ when conditions $\hat{\boldsymbol{C}}_1 (\boldsymbol{X}^\mathsf{T} \boldsymbol{X}) \hat{\boldsymbol{C}}_1$ and $\hat{\boldsymbol{C}}_2 \boldsymbol{\Sigma} \hat{\boldsymbol{C}}_2^\mathsf{T} = \boldsymbol{I}_r$ are considered.

Model (2.3.4) can also be given a *parametric* or factor analysis representation as $\boldsymbol{Y} = \boldsymbol{F} \hat{\boldsymbol{C}}_2 + \boldsymbol{E}$ with $\boldsymbol{F} = \boldsymbol{X} \hat{\boldsymbol{C}}_1$, where $\boldsymbol{X}$ is $T \times p$ is the matrix of vector predictors of full rank, $\hat{\boldsymbol{C}}_1$ is a $p \times r$ matrix of factor loadings, and $\hat{\boldsymbol{C}}_2$ is a $r \times n$ orthogonal matrix. The factor analysis approach provides the number of factors in the model as a consequence of a rank restriction imposed on the coefficient matrix and estimates parameters by reducing the metric between the estimated and real matrix of coefficients.

Applications can be found in genomic data analysis [9] and financial econometrics [46], among others. The *rank selection criteria* (RSC) introduced by Bunea, She and Wegkamp [9] selects the optimal reduced rank estimator of the coefficient matrix in multivariate response regression models under the assumption the variance σ^2 is unknown. RSC is a procedure with very low computational complexity and when compared with the *nuclear norm penalized* (NNP), [74, 41], shows to provide a more parsimonious model. Giraud [22] proposed and analyzed a criterion to handle the case where σ^2 is unknown. His theory requires no assumption on the design matrix $\boldsymbol{X}$ and applies mainly in high dimensions.

The RRR model can be used to generalize the classical multivariate regression model by relaxing the implicit constraint on the rank of the coefficient matrix. More importantly, by carefully choosing the input vector $\boldsymbol{X}$, the output vector $\boldsymbol{Y}$, and the matrix of weights $\boldsymbol{\Gamma}$, RRR can be used to play an important role as a unifying treatment of several classical multivariate procedures that were developed separately from each other. For example, if we set $\boldsymbol{X} = \boldsymbol{Y}$ $(n = p)$ by making the output variables identical to the input variables, and set $\boldsymbol{\Gamma} = \boldsymbol{I}$ then this problem has the Hotellings PCA [26] form, see, Izenman [28].

In the context of modeling macroeconomic time series, Sims [50] discusses the notion of *index* models. The indexes, $\boldsymbol{X}^* = \boldsymbol{X}\boldsymbol{C}$, which are smaller in number, are constructed from a large set of time series predictors in $\boldsymbol{X}$ and are found to drive the vector of observed response variables $\boldsymbol{Y}$. More recent developments of reduced-rank models considered the vector autoregressive time series case where predictors in $\boldsymbol{X}$ represented lagged values of $\boldsymbol{Y}$. Among others, see, Reinsel [47], Velu *et al* [72], Ahn and Reinsel [1, 2].

Rao [44] has shown a stronger result that the solution for $\boldsymbol{C}_1$ and $\boldsymbol{C}_2$ in equation (2.3.4) also simultaneously minimizes the eigenvalues of the weights matrix provided that $\boldsymbol{\Gamma} = \boldsymbol{\Sigma}_{\boldsymbol{YY}}^{-1}$. This was proved using the Poincare separation theorem. Robinson [49] has proved a similar result, which indicates that the solutions of $\boldsymbol{C}_1$ and $\boldsymbol{C}_2$ when the corresponding sample quantities are substituted and $\boldsymbol{\Gamma} = \tilde{\boldsymbol{\Sigma}}_{\boldsymbol{YY}}^{-1}$, are gaussian estimates under the reduced-rank model (2.3.1), that is, the maximum likelihood estimates under the assumption of normality of $\boldsymbol{E}$.

The classical multivariate regression methods are based on the assumption that (i) the regression coefficient matrix is of full rank and (ii) the error term in the model are independent. When the data are in the form of time series the assumption of serial independence of errors is often not appropriate. In these circumstances, it is important to account for a serial correlation in the errors. Robinson [48] considered the reduced-rank model for time series data in the frequency domain with a general structure for errors. In particular, assuming the errors followed a general stationary process, Robinson studied the asymptotic properties of both the least squares and other more efficient estimators of the regression coefficient matrix. However, often the serially correlated disturbances will satisfy more restrictive assumptions and, in these circumstances, it would seem possible to exploit this structure to obtain estimates which may have better small sample properties than the estimators proposed by Robinson. In his work, Robinson proved the canonical correlation analysis (CCA) for time series is equivalent to the problem of estimating a linear regression matrix with $\boldsymbol{C}$ of less than full rank. After reparameterizing $\boldsymbol{C}$ some estimates of the new parameters, obtained by solving an eigenvalue problem closely related to canonical correlations, are found to be consistent and efficient when the residuals are mutually independent. When the residuals are generated by an autocorrelated model, stationary time series estimates are still consistent and obey a central limit theorem, but are no longer efficient. Alternatively, more general estimates were suggested which are efficient in the presence of serial correlation, see, Robinson [48]. In the

analysis of the time series vector autoregressive as predictors in model (2.3.1) via reduced-rank regression we consider Robinson's approach [48] such that stationary autoregressive vectors can be considered under the least square estimator approach.

The classical time series book by Brillinger [8] provides a detailed proof of the use of OLS regression method for time series under normality assumption or not. Brillinger also highlighted the particular importance of the case $\Gamma = \boldsymbol{I}$. Thus, the scenario where weights of the predictors are equal, the OLS regression holds. Brillinger brilliantly described the three cases for the OLS regression models. He first considered the case $r = rank(\boldsymbol{C}) = min(p, n)$ where none real reduction in dimension is required. Later, he studied the case where $n = p$ so if we have the same number of series in $\boldsymbol{Y}$ to match those in $\boldsymbol{X}$ then we would have the problem whose solution lead to a PCA of the spectral design matrix. Finally, his attention was directed to the case where a real dimension reduction is achieved. This case analyzed the setup $r = rank(\boldsymbol{C}) \leq min(p, n)$ and $n \leq p$ leading to a reduced-rank regression framework. Theorem 10.2.3 in [8] justifies the use OLS regression structure to analyze time series when $n \leq p$ and variates are asymptotically normal.

2.4 Dynamic Factor Model - DFM

DFM is based on dynamic factor analysis and its premise is that a small number of unobserved common dynamic factors produce the observed comovements of economic time series, see, e.g. Sargent and Sims [50], Stock and Watson [58, 57]. Common dynamic factors are driven by the most relevant series for the purposes of conducting prediction. In macroeconomic forecasting, for example, p can be very large, often larger than T available for model fitting. This large dimensional problem is simplified by modeling the covariability of the series in terms of a relatively few number of unobserved latent factors. Forecasting then can be carried out in a two-step process. First, a time series of the factors is estimated from the predictors; second, the relationship between the variables to be forecasted and the factors is estimated by a linear regression.

The dynamic factor model expresses $\boldsymbol{x}_t$ as a distributed lag of a small number of unobserved common factors, plus a possibly correlated idiosyncratic disturbance

$$x_{it} = \tilde{\boldsymbol{\lambda}}_i(\mathfrak{L})\mathfrak{g}_t + u_{it}, \;\; i = 1, \ldots, N \tag{2.4.1}$$

$$u_{it} = \delta_i(\mathfrak{L})u_{it-1} + \xi_{it}, \tag{2.4.2}$$

where $\mathfrak{g}_t$ is the q-vector of unobserved dynamic factors, $\tilde{\boldsymbol{\lambda}}_i(\mathfrak{L})$ is a q-vector of dynamic loadings, u_{it} is the disturbance, and ξ_{it} is i.i.d $\sim \mathrm{N}(0, \sigma_\xi^2)$. Terms $\mathfrak{g}_t$ and u_{it} are assumed to be uncorrelated. In addition, the disturbance terms are taken to be mutually uncorrelated at all leads and lags. In particular, no serial correlation in individual idiosyncratic terms. Factor models arise naturally in economics. For example, x_{it} is the GDP growth rate for country i in period t, $\boldsymbol{f}_t$ is a vector common shocks, $\boldsymbol{\lambda}_i$ is the heterogeneous impact of the shocks, and u_{it} is the country-specific growth rate.

DFM is convenient when the idiosyncratic errors are serially uncorrelated. Then, with $|\delta_i| < 1$ and using the AR property of the AR model it follows

$$\begin{aligned} u_{it} &= \delta_i(1)u_{i(t-1)} + \xi_{it} \\ &= \delta_i^2(2)u_{i(t-2)} + \xi_{it} + \delta_i(1)\xi_{i(t-1)} \\ &\dots \\ &= [\xi_{it} + \delta_i(1)\xi_{i(t-1)} + \delta_i^2(2)\xi_{i(t-2)} + \delta_i^3(3)\xi_{i(t-3)} + \dots] = \sum_{j=0}^{\infty} \delta_i^j(j)\xi_{i(t-j)}. \end{aligned}$$

Then, since only a large finite number of lags is considered, we have

$$\begin{aligned} &= \frac{1}{1 - \boldsymbol{\delta}_i(\mathfrak{L})\mathfrak{L}}\xi_{it} \\ (1 - \boldsymbol{\delta}_i(\mathfrak{L})\mathfrak{L})u_{it} &= \xi_{it} \\ \boldsymbol{D}(\mathfrak{L})u_{it} &= \xi_{it} \end{aligned} \tag{2.4.3}$$

where $\boldsymbol{D}(\mathfrak{L}) = 1 - \boldsymbol{\delta}_i(\mathfrak{L})\mathfrak{L}$. Equation (2.4.3) shows that when u_{it} is multiplied by $\boldsymbol{D}(\mathfrak{L})$, it will follow an i.i.d. process. Thus, to achieve serially uncorrelated errors we need to multiply both sides of the equation (2.4.1) by $\boldsymbol{D}(\mathfrak{L})$. This multiplication yields

$$x_{it} = \boldsymbol{\lambda}_i(\mathfrak{L})\mathfrak{g}_t + \boldsymbol{\delta}_i(\mathfrak{L})x_{it} + \xi_{it}, \tag{2.4.4}$$

where $\boldsymbol{\lambda}_i(\mathfrak{L}) = (1 - \boldsymbol{\delta}_i(\mathfrak{L})\mathfrak{L})\tilde{\boldsymbol{\lambda}}_i(\mathfrak{L})$. Then, equation (2.4.4) can be rewritten as

$$x_{it} = \boldsymbol{\lambda}_i(\mathfrak{L})\mathfrak{g}_t + \boldsymbol{\delta}_i(\mathfrak{L})x_{it} + \xi_{it} \tag{2.4.5}$$

$$\mathfrak{g}_t = \boldsymbol{\Gamma}(\mathfrak{L})\mathfrak{g}_{t-1} + \boldsymbol{e}_t \tag{2.4.6}$$

where the evolution of the dynamic factors is modeled as following a VAR process, $\boldsymbol{\Gamma}(\mathfrak{L})$ is a matrix lag polynomial, and $\boldsymbol{e}_t$ is a disturbance vector. A DFM-q can be written in a different form with r

finite factors. Suppose that $\boldsymbol{\lambda}(\mathfrak{L})$ has a finite degree $s-1$, and let $\boldsymbol{f}_t = (\mathfrak{g}_t', \ldots, \mathfrak{g}_{t-s+1}')'$ or a subset of these lags. Let $\boldsymbol{f}_t \in \mathbb{R}^r$, with $s \leq r \leq qs$. Then the DFM in (2.4.5) and (2.4.6) can be written as

$$\boldsymbol{x}_t = \boldsymbol{\Lambda}\boldsymbol{f}_t + \boldsymbol{D}(\mathfrak{L})\boldsymbol{x}_{t-1} + \boldsymbol{\xi}_t \tag{2.4.7}$$

$$\boldsymbol{f}_t = \boldsymbol{\Phi}(\mathfrak{L})\boldsymbol{f}_{t-1} + \boldsymbol{G}\boldsymbol{e}_t \tag{2.4.8}$$

where $\boldsymbol{\Lambda}$ is the loading matrix, $\boldsymbol{f}_t$ consists of current and (possibly) lagged values of the q dynamic factors, $\boldsymbol{\Phi}(\mathfrak{L})$ is a sparse matrix with some entries from $\boldsymbol{\Gamma}(\mathfrak{L})$, and $\boldsymbol{G}$ is a $r \times q$ matrix.

2.4.1 Factor estimation via Principal Components

Dynamic factors can be estimated nonparametrically using the method of PC, see, Stock and Watson [59]. Factors estimated via (2.4.7) and (2.4.8) find the minimization problem as

$$\min_{\boldsymbol{f}_1,\ldots,\boldsymbol{f}_T,\boldsymbol{\Lambda},\boldsymbol{D}(\mathfrak{L})} T^{-1} \sum_{t=1}^{T} [(\boldsymbol{I} - \boldsymbol{D}(\mathfrak{L})\mathfrak{L})\boldsymbol{x}_t - \boldsymbol{\Lambda}\boldsymbol{f}_t]'[(\boldsymbol{I} - \boldsymbol{D}(\mathfrak{L})\mathfrak{L})\boldsymbol{x}_t - \boldsymbol{\Lambda}\boldsymbol{f}_t] \tag{2.4.9}$$

where $\boldsymbol{f}_1, \ldots, \boldsymbol{f}_t, \boldsymbol{\Lambda}$, and $\boldsymbol{D}(\mathfrak{L})\mathfrak{L}$ need to be estimated jointly. Finding a solution for the minimization problem in (2.4.9) is somewhat challenging. To attenuate this scenario Stock and Watson [61] assumed $\boldsymbol{D}(\mathfrak{L})\mathfrak{L}$ is known. As a result, the minimization problem notably reduces its complexity as it now requires to estimate only $\boldsymbol{f}_1, \ldots, \boldsymbol{f}_t$ and $\boldsymbol{\Lambda}$ jointly, which is a more solvable problem.

With $\boldsymbol{D}(\mathfrak{L})\mathfrak{L}$ given, the problem becomes equivalent to solve

$$\boldsymbol{X} = \boldsymbol{F}\boldsymbol{\Lambda}' + \boldsymbol{U} \tag{2.4.10}$$

where now $\boldsymbol{X} = (\tilde{\boldsymbol{x}}_1, \ldots, \tilde{\boldsymbol{x}}_T)'$ with $\tilde{\boldsymbol{x}}_t \triangleq (\boldsymbol{D}(\mathfrak{L})\mathfrak{L})\boldsymbol{x}_t$, $\boldsymbol{\Lambda} = (\boldsymbol{\lambda}_1, \ldots, \boldsymbol{\lambda}_N)'$ is a $N \times r$ matrix of factor loadings, $\boldsymbol{F} = (\boldsymbol{f}_1, \ldots, \boldsymbol{f}_T)'$ is a $T \times r$ matrix of factors, and $\boldsymbol{U} = (\boldsymbol{u}_1, \ldots, \boldsymbol{u}_N)'$ is the $T \times N$ matrix of disturbances. In real world applications the unknown matrix of lagged values of $\boldsymbol{x}_t$ is assumed to be known so the minimization problem can be easily accomplished. Reinsel and Velu [46], supports the use of principal components in estimating the factors.

Theorem 1. *Let $\boldsymbol{S}$ be a matrix of order $m \times n$ with rank m. The Euclidean norm, $tr[\boldsymbol{S}-\boldsymbol{P})(\boldsymbol{S}-\boldsymbol{P})']$, is minimum among matrices $\boldsymbol{P}$ of the same order as $\boldsymbol{S}$ but of rank $r \leq m$, when $\boldsymbol{P} = \boldsymbol{M}\boldsymbol{M}'\boldsymbol{S}$, where $\boldsymbol{M}$ is $m \times r$ and the columns of $\boldsymbol{M}$ are the first r (normalized) eigenvectors of $\boldsymbol{S}\boldsymbol{S}'$, that is, the normalized eigenvectors corresponding to the largest eigenvalues of $\boldsymbol{S}\boldsymbol{S}'$.*

By analogy to regression, $\boldsymbol{\Lambda}$ and $\boldsymbol{F}$ can be estimated using the Frobenius norm as follows

$$\min_{\boldsymbol{F},\boldsymbol{\Lambda}} \ (T)^{-1}||\boldsymbol{X} - \boldsymbol{F}\boldsymbol{\Lambda}'||_F^2 \text{ subject to } \boldsymbol{\Lambda}'\boldsymbol{\Lambda} = \boldsymbol{I}_r. \tag{2.4.11}$$

Then, isolating $\boldsymbol{F}$ in (2.4.10) (i.e. $\boldsymbol{F} = \boldsymbol{X}\boldsymbol{\Lambda}$), we estimate $\boldsymbol{\Lambda}$ as follows,

$$\begin{aligned}
&\min_{\boldsymbol{F},\boldsymbol{\Lambda}} \ (T)^{-1}\text{tr}[(\boldsymbol{X} - \boldsymbol{F}\boldsymbol{\Lambda}')(\boldsymbol{X} - \boldsymbol{F}\boldsymbol{\Lambda}')'] \text{ subject to } \boldsymbol{\Lambda}'\boldsymbol{\Lambda} = \boldsymbol{I}_r \\
&\min_{\boldsymbol{F},\boldsymbol{\Lambda}} \ (T)^{-1}\text{tr}[(\boldsymbol{X}\boldsymbol{X}' - \boldsymbol{X}\boldsymbol{\Lambda}\boldsymbol{F}' - \boldsymbol{F}\boldsymbol{\Lambda}'\boldsymbol{X}' + \boldsymbol{F}\boldsymbol{\Lambda}'\boldsymbol{\Lambda}\boldsymbol{F}')] \\
&\min_{\boldsymbol{\Lambda}} \ (T)^{-1}\text{tr}[(\boldsymbol{X}\boldsymbol{X}' - \boldsymbol{X}\boldsymbol{\Lambda}\boldsymbol{\Lambda}'\boldsymbol{X}' - \boldsymbol{X}\boldsymbol{\Lambda}\boldsymbol{\Lambda}'\boldsymbol{X}' + \boldsymbol{X}\boldsymbol{\Lambda}\boldsymbol{\Lambda}'\boldsymbol{\Lambda}\boldsymbol{\Lambda}'\boldsymbol{X}')] \\
&\min_{\boldsymbol{\Lambda}} \ (T)^{-1}\text{tr}[\boldsymbol{X}\boldsymbol{X}'(\boldsymbol{I} - \boldsymbol{\Lambda}\boldsymbol{\Lambda}')] \text{ subject to } \boldsymbol{\Lambda}'\boldsymbol{\Lambda} = \boldsymbol{I}_r
\end{aligned}$$

Minimizing the above quantity with respect to $\boldsymbol{\Lambda}$ is equivalent to maximizing

$$\begin{aligned}
&\max_{\boldsymbol{\Lambda}} \ (T)^{-1}\text{tr}[\boldsymbol{\Lambda}\boldsymbol{X}\boldsymbol{X}'\boldsymbol{\Lambda}')] \text{ subject to } \boldsymbol{\Lambda}'\boldsymbol{\Lambda} = \boldsymbol{I}_r \\
&\max_{\boldsymbol{\Lambda}} \ \text{tr}[\boldsymbol{\Lambda}'\boldsymbol{\Sigma}_{\boldsymbol{X}\boldsymbol{X}}\boldsymbol{\Lambda}] \text{ subject to } \boldsymbol{\Lambda}'\boldsymbol{\Lambda} = \boldsymbol{I}_r.
\end{aligned} \tag{2.4.12}$$

Result in Theorem 1 follows that the goal of maximizing $\boldsymbol{\Lambda}$ in problem (2.4.12) is achieved by choosing the columns of $\boldsymbol{\Lambda}$ to be the orthonormal eigenvectors of $\boldsymbol{\Sigma}_{\boldsymbol{X}\boldsymbol{X}}$ that correspond to the first r largest eigenvalues. The positive square roots of the eigenvalues of $\boldsymbol{\Sigma}_{\boldsymbol{X}\boldsymbol{X}}$ are referred to as the singular values of the matrix $\boldsymbol{X}$. In general, a matrix $\boldsymbol{X}$ of dimensions $T \times N$ and rank k can be expressed in the *singular value decomposition* (SVD) as $\boldsymbol{X} = \boldsymbol{V}\boldsymbol{D}\boldsymbol{W}$, where $\boldsymbol{D} = \text{diag}(\boldsymbol{d}_1, \ldots, \boldsymbol{d}_k)$ with $\boldsymbol{d}_1^2 > \ldots > \boldsymbol{d}_k^2$ being the nonzero eigenvalues of $\boldsymbol{\Sigma}_{\boldsymbol{X}\boldsymbol{X}}$, $\boldsymbol{V} = (\boldsymbol{v}_1, \ldots, \boldsymbol{v}_k)$ is a $T \times k$ matrix such that $\boldsymbol{V}'\boldsymbol{V} = \boldsymbol{I}$. The columns of $\boldsymbol{V}$ are the normalized eigenvectors of $\boldsymbol{\Sigma}_{\boldsymbol{X}\boldsymbol{X}}$. Thus, $\boldsymbol{\Lambda}$ equals the first r eigenvectors in $\boldsymbol{V}$ ($r \leq k$), and factors are estimated using $\boldsymbol{F} = \boldsymbol{\Lambda}\boldsymbol{X}$ subject to $\boldsymbol{\Lambda}'\boldsymbol{\Lambda} = \boldsymbol{I}_r$.

The solution to the minimization problem (2.4.11) leads to the estimation of the factors via PCA. Nonetheless, in high dimensions estimated factors produced via conventional procedures may lead to inconsistency issues. For example, Johnstone and Lu [29] proved that the eigenvectors estimated via PCA are consistent if and only if $\frac{N}{T} \to 0$. An immediate consequence is that in high dimensions ($N \gg T$) factors may not be constructed through conventional PCA. Another drawback of using PCA to construct factors in high dimensions is that all predictors are considered relevant. To avoid both drawbacks, our proposed methodology performs variable selection while constructing the C-level factors, and reduces the amount of assumptions in the DFM methodology.

2.4.2 Dynamic affinity between DFM and SFAR methodologies

A recursive iteration can be used to show a similarity between the DFM and SFAR in factor construction. Consider the multivariate vector time series response

$$\boldsymbol{X}_t = \boldsymbol{X}_{t-1}\boldsymbol{C} + \boldsymbol{E}_t, \tag{2.4.13}$$

where $\boldsymbol{X}_{t-1}$ contains the most recent observations of the series in $\boldsymbol{X}_t$ at time t, $\boldsymbol{C}$ is a coefficient matrix. Matrix $\boldsymbol{E}_t$ is the error component of the series at time t. The relationship between lags of observations is used as the starting point to elucidate the connection between latent factors $\boldsymbol{F}_{t-1}$ and $\boldsymbol{X}_t$. We use the fact that $\boldsymbol{C} = \boldsymbol{C}_1\boldsymbol{C}_2$ where $\boldsymbol{C}_1$ is a $p \times r$ matrix so factors are constructed as

$$\boldsymbol{F}_{t-1} = \boldsymbol{X}_{t-1}\boldsymbol{C}_1. \tag{2.4.14}$$

and equation (2.4.13) is now

$$\boldsymbol{X}_t = \boldsymbol{X}_{t-1}\boldsymbol{C}_1\boldsymbol{C}_2 = \boldsymbol{F}_{t-1}\boldsymbol{C}_2 \tag{2.4.15}$$

where $\boldsymbol{C}_2$ is the coefficient matrix. Therefore, there exists a linear association between observations at time t and factors constructed using observations at time $t-1$. Following the same procedure it can be shown that for two lags of observations the same relationship exist. Given $\boldsymbol{X}_{t+1} = [\boldsymbol{X}_t \;\; \boldsymbol{X}_{t-1}]$ and by replacing it in (2.4.15), we have

$$\boldsymbol{X}_{t+1} = [\boldsymbol{X}_t \;\; \boldsymbol{X}_{t-1}]\boldsymbol{C}_1\boldsymbol{C}_2 = \boldsymbol{F}_t\boldsymbol{C}_2 \tag{2.4.16}$$

where

$$\begin{aligned}
\boldsymbol{F}_t &= [\boldsymbol{X}_t \;\; \boldsymbol{X}_{t-1}]\boldsymbol{C}_1 && \text{by (2.4.16)} \\
&= [\boldsymbol{X}_{t-1}\boldsymbol{C}_1\boldsymbol{C}_2 \;\; \boldsymbol{X}_{t-2}\boldsymbol{C}_1\boldsymbol{C}_2]\boldsymbol{C}_1 && \text{by (2.4.15)} \\
&= [\boldsymbol{F}_{t-1}\boldsymbol{C}_2 \;\; \boldsymbol{F}_{t-2}\boldsymbol{C}_2]\boldsymbol{C}_1 && \text{by (2.4.14)} \\
&= \begin{bmatrix}\boldsymbol{F}_{t-1} & \boldsymbol{F}_{t-2}\end{bmatrix} \begin{bmatrix}\boldsymbol{C}_2 & 0 \\ 0 & \boldsymbol{C}_2\end{bmatrix} \boldsymbol{C}_1. && (2.4.17)
\end{aligned}$$

Equation (2.4.17) shows the straight association between lagged factors and observations at time $t+1$. In a more general framework, it can be shown that the association holds for any lag order. Let $\boldsymbol{X}_{t-m}$ denote the observations of the time series in $\boldsymbol{X}$ from time $t-m+1$ to t. Then,

$$\boldsymbol{X}_{t+1} = \boldsymbol{X}_{t-m}\boldsymbol{C}_1\boldsymbol{C}_2 = \boldsymbol{F}_t\boldsymbol{C}_2$$

where

$$\begin{aligned}
\boldsymbol{F}_t &= [\boldsymbol{X}_t \;\; \boldsymbol{X}_{t-1} \;\; \dots \;\; \boldsymbol{X}_{t-m+1}]\boldsymbol{C}_1 \\
&= [\boldsymbol{X}_{t-1}\boldsymbol{C}_1\boldsymbol{C}_2 \;\; \boldsymbol{X}_{t-2}\boldsymbol{C}_1\boldsymbol{C}_2 \;\; \dots \;\; \boldsymbol{X}_{t-m}\boldsymbol{C}_1\boldsymbol{C}_2]\boldsymbol{C}_1 \\
&= [\boldsymbol{F}_{t-1}\boldsymbol{C}_2 \;\; \boldsymbol{F}_{t-2}\boldsymbol{C}_2 \;\; \dots \;\; \boldsymbol{F}_{t-m}\boldsymbol{C}_2]\boldsymbol{C}_1 \\
&= \begin{bmatrix}\boldsymbol{F}_{t-1} & \boldsymbol{F}_{t-2} & \dots & \boldsymbol{F}_{t-m}\end{bmatrix} \begin{bmatrix} \boldsymbol{C}_2 & 0 & \cdots & 0 \\ 0 & \boldsymbol{C}_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \boldsymbol{C}_2 \end{bmatrix} \boldsymbol{C}_1. \qquad (2.4.18)
\end{aligned}$$

As shown in equation (2.4.18), for any lag order the SFAR methodology takes advantage of lagged factors to estimate current factors. Clearly, the SFAR methodology intrinsically imitates the recursive iteration of DFM to estimate factors although in a more informative way. Factors from SFAR are constructed using information in $\boldsymbol{X}_t$. Instead, DFM uses a VAR representation of the lagged factors.

2.4.3 Forecasting

Factors extracted via DFM are commonly used to forecast a target series of interest. Let $\boldsymbol{y}_t$ denote the time series to forecast at time t. Also, assume h stands for the number of periods ahead to be forecasted. For example, since the time point observations in $\boldsymbol{y}_t$ are collected at regular periods of time, y_{t+h} represents the value of the target series h periods ahead. Assume $\boldsymbol{X}_t$ admits a dynamic factor model representation with r factors in $\boldsymbol{F}_t$. Also, let $\boldsymbol{X}_{t-m}$ to have finite lag order m. Those assumptions permit to estimate the TSR h periods ahead as follows

$$\hat{y}_{t+h} = \boldsymbol{y}_{t-m+1}\hat{\boldsymbol{c}} + \boldsymbol{f}_t\hat{\boldsymbol{c}}' + e_{t+h}, \qquad (2.4.19)$$

where $\boldsymbol{y}_{t-m+1}$ is a vector containing m lags of observations of the TSR, $\boldsymbol{f}_t$ is a vector of dynamic factors, $\boldsymbol{c}$ and $\boldsymbol{c}'$ are estimated coefficients in response to the target series, and e_{t+h} is the error component. The h periods ahead forecast of $\boldsymbol{y}_t$ can be computed by regressing y_{t+h} on $\boldsymbol{f}_t$, and y_t and its lags. Model (2.4.19) is equivalent to the one described in Stock and Watson [59].

Forecasting multiple periods ahead can be computed in two ways. The iterated one-period ahead model uses the AR forecast with model parameters estimated recursively via OLS, and forecasts are constructed recursively using the estimated coefficients. In turn, the direct multi-period ahead

forecast performs a direct estimate of the parameters as the recursive minimizers of the mean squared error (MSE) of the h periods ahead criterion function. Accordingly, the parameters are estimated by the OLS regression in which the regressors are constant.

Choosing between iterated and direct forecasts involves a tradeoff between bias and estimation variance: the iterated method produces more efficient parameter estimates than the direct method, but it is prone to bias if the one-period ahead model is misspecified. Ignoring estimation uncertainty, if both the iterated model and the direct model have l lags of the dependent variable but the true autoregressive order exceeds it, then the asymptotic mean squared forecast error (MSFE) of the direct forecast typically is less than (and cannot exceed) the MSFE of the iterated forecast, see, e.g. Findley [18]. Because the relative efficiency of iterated vs. direct forecasts is theoretically ambiguous and depends on the unknown population best linear projection, the question of which method to choose is an empirical one, see, Marcellino, Stock, and Watson [39] and the references therein.

What forecasting method should prevail has been heavily studied in the literature. Contribution to the theory of iterated vs. direct forecast include Tiao and Tsay [68], and Kang [30], among others. Empirically studies for the performance of iterated vs. direct forecast belong to Findley [18, 19] and Liu [34]. In their study, Marcellino, Stock, and Watson [39] concluded in favor of the iterated forecast as it tends to have a smaller mean squared forecast error (MSFE).

In this work we preferred the direct multi-period forecast for two reasons. First, the direct method is required for comparison purposes. The second is related to the idea that the forecasting error of the direct forecast is not cumulative and therefore should be smaller than its similar in the iterated approach.

CHAPTER 3

SPARSE FACTOR AUTO-REGRESSION

In this chapter we provide details of the Sparse Factor Auto-Regression (SFAR). The SFAR is convenient for forecasting macroeconomic time series with very many predictors. Its framework attempts to provide interpretability and accuracy in forecasting. Interpretability by identifying the most relevant variables that best explain the target category of interest and series therein, and accuracy by utilizing the relevant predictors in constructing the C-level factors. C-level factors are extracted in compliance with some conditions that works in favor of the predictability power of the model.

Throughout this chapter we define the notation used in the SFAR, and provide the SFAR framework that takes advantage of the high-dimensions, multi-category, and large-p attributes of the data. Then, we provide a strategy to fit the model and the computational component of the SFAR.

3.1 Notation

The SFAR uses notation as follows: given a time series response, we use y_t to denote its observation at time t, and $\boldsymbol{y}_t = (y_{t-T+1}, \ldots, y_t)^\mathsf{T}$ to be its vector representation containing T observations from time $t-T+1$ to t. (Be aware that when the subscript t is used with vector $\boldsymbol{y}$ it represents the ending time point). A matrix response is denoted by $\boldsymbol{Y}_t = [\boldsymbol{y}_{t,1}, \ldots, \boldsymbol{y}_{t,n}] \in \mathbb{R}^{T\times n}$ where $\boldsymbol{y}_{t,i}$ represents the ith vector time series response with T observations from time $t-T+1$ to t, for $i = 1, \ldots, n$. We call $\boldsymbol{y}_t$ the *target series response* (TSR), and $\boldsymbol{Y}_t$ the *target category response* (TCR). Throughout this dissertation, $\boldsymbol{y}_t$ is assumed to be contained in $\boldsymbol{Y}_t$.

Without loss of generality, suppose one uses the information available up to time $t-h$ for $h \geq 1$, to train a model in response to $\boldsymbol{y}_t$. Let x_t stand for the observation at time t of a (raw) time series $\boldsymbol{x}_t = (x_{t-T+1}, \ldots, x_t)^\mathsf{T}$ with T observations from time $t-T+1$ to t. We use $\tilde{p}$ to represent the number of raw time series available. Therefore, for training purposes, $\boldsymbol{x}_{t-h} = (x_{t-h-T+1}, \ldots, x_{t-h})^\mathsf{T}$ is a predictor for $\boldsymbol{y}_t$ from time $t-h-T+1$ to $t-h$. Let m be the lag order for a raw time series

to be considered. Then, the predictor matrix can be defined as follows

$$\boldsymbol{X}_{t-h-m+1:t-h} = \Big[[\boldsymbol{x}_{t-h,1}, \ldots, \boldsymbol{x}_{t-h,\tilde{p}}] \ldots [\boldsymbol{x}_{t-h-m+1,1} \ldots, \boldsymbol{x}_{t-h-m+1,\tilde{p}}]\Big] \in \mathbb{R}^{T\times \tilde{p}m},$$

where the subscript $t-h-m+1:t-h$ indicates the time interval of predictors to model $\boldsymbol{y}_t$. Given any j, $\boldsymbol{x}_{t-h,j} \in \mathbb{R}^T$ denotes the first lag of the jth time series predictor from $t-h-T+1$ to $t-h$, while $\boldsymbol{x}_{t-h-m+1,j} \in \mathbb{R}^T$ is the mth lag of the same time series from $t-h-m-T+2$ to $t-h-m+1$. Let $p = \tilde{p}m$ denote the total number of predictors. For notational simplicity, we also write $\boldsymbol{X}_{t-h-m+1:t-h} \in \mathbb{R}^{T\times \tilde{p}m}$ as $\boldsymbol{X}_{t-h} \in \mathbb{R}^{T\times p}$. Notice that $\boldsymbol{y}_t$, $\boldsymbol{Y}_t$, and $\boldsymbol{X}_{t-h}$ are used for training purposes, while, y_s with $s > t$, is the goal in forecasting.

In addition, we represent by $\boldsymbol{Z}_{t-h-l+1:t-h} \in \mathbb{R}^{T\times w}$ a general design matrix containing potentially useful time series predictors and their lags, in response to the TSR. Typically, the lagged observations of the TSR, say, $(y_{t-1}, y_{t-2}, y_{t-3}, y_{t-4})$ are included, as in AR(4). For notational simplicity, we also write $\boldsymbol{Z}_{t-h-l+1:t-h}$ as $\boldsymbol{Z}_{t-h}$.

Finally, for any matrix $\boldsymbol{C} \in \mathbb{R}^{p\times n}$ we use the notation $\|\boldsymbol{C}\|_{2,0}$ to stand for the number of nonzero row vectors in $\boldsymbol{C}$, and rank($\boldsymbol{C}$) or r($\boldsymbol{C}$) to denote the rank of matrix $\boldsymbol{C}$.

3.2 The SFAR framework

In this section, we describe the framework of the Sparse Factor Auto-Regression (SFAR) that takes advantage of the available multi-category data for constructing parsimonious category-level factors (C-level factors) in TSR to forecast. The complete SFAR is given as follows

M1. *Target Series Prediction (TSP):* $\boldsymbol{y}_t = \boldsymbol{Z}_{t-h}\boldsymbol{B} + \boldsymbol{F}_{t-h}\boldsymbol{B}' + \boldsymbol{\varepsilon}_t$

M2. *Category Factor Extraction (CFE):* $\boldsymbol{F}_{t-h} = \boldsymbol{X}_{t-h}\boldsymbol{L}$

M3. *Target Category Correspondence (TCC):* $\boldsymbol{Y}_t = \boldsymbol{F}_{t-h}\boldsymbol{B}'' + \boldsymbol{E}_t$

subject to the following constraints

C1. *Orthogonality Constraint:* $\boldsymbol{F}^\top\boldsymbol{F}$ *is diagonal*

C2. *Cardinality-Rank Constraints:* $\boldsymbol{L} \in \mathbb{R}^{p\times r}$ *and* $\|\boldsymbol{L}\|_{2,0} \le \mathfrak{q}$

where $\|\boldsymbol{L}\|_{2,0} \le \mathfrak{q}$ indicates that $\|\boldsymbol{L}\|_{2,0}$ contains at most $\mathfrak{q}$ nonzero rows. To emphasize the dependence of $\boldsymbol{L}$ on $\mathfrak{q}$ and r, we also write $\boldsymbol{L}(\mathfrak{q}, r)$.

In the SFAR framework there are three different predictor matrices: $\boldsymbol{Z}_{t-h} \in \mathbb{R}^{T\times w}$ is formed using raw time series that serve in (M1). On the other hand, the predictor matrix $\boldsymbol{X}_{t-h} \in \mathbb{R}^{T\times p}$, also formed based on (potentially more) raw time series, assists in (M3), and $\boldsymbol{F}_{t-h} \in \mathbb{R}^{T\times r}$ which contains the newly constructed factors to explain $\boldsymbol{y}_t$ (or y_t), also collaborates in (M3) . The unknowns in SFAR include the coefficient matrices $\boldsymbol{B} \in \mathbb{R}^{w\times w}$, $\boldsymbol{B}' \in \mathbb{R}^{r\times r}$, $\boldsymbol{B}'' \in \mathbb{R}^{r\times n}$, and the factor loading matrix $\boldsymbol{L} \in \mathbb{R}^{p\times r}$ subject to the nontrivial constraints (C1) and (C2). Finally, $\boldsymbol{E}_t \in \mathbb{R}^{T\times n}$ and $\boldsymbol{\varepsilon}_{t+h}$ are the error components.

The SFAR possesses a hybrid target-category framework which encompasses: **a)** the multivariate-linear level information in $\boldsymbol{Z}_{t-h}$ and **b)** the category-level data in $\boldsymbol{X}_{t-h}$ to describe a univariate TSR. First, (M3) is used to train the C-level factors contained in $\boldsymbol{F}_t$. As a result, matrix $\boldsymbol{L}$ is of reduced columns. Matrix $\boldsymbol{L}$ is estimated from the category-level in $\boldsymbol{X}_{t-h}$, and used to construct C-level factors, as shown in (M2). The factor loading matrix $\boldsymbol{L}$, is critical in factors construction because it: **a)** has a row sparse design with at most $\mathfrak{q}$ nonzero rows aligned with a subset of relevant predictors in $\boldsymbol{X}_{t-h}$, **b)** regulates the number of factors to be constructed based on dimensionality r (giving a rank-driven attribute to the SFAR) and **c)** provides the dynamic attribute to the factors constructed. At last, (M1) combines the multivariate-linear information in $\boldsymbol{Z}_{t-h}$ and C-level factors in $\boldsymbol{F}_{t-h}$ to build a forecasting model in response to TSR. The SFAR solves the problem of encompassing two levels of information to forecast a TSR of interest.

To illustrate the meaning of the constraints, we give the following example. Suppose in the S&P 500 data example, one sets the parameters as $m = 2$, $\mathfrak{q} = 20$, and $r = 3$. Then, a prediction model for the TSR, $\boldsymbol{y}_t$, is constructed using only a subset of at most 20 relevant predictors (based on 3 factors) from the pool of $p = 1000$ candidate predictors in $\boldsymbol{X}_{t-h}$. As discussed previously, the smallest value of r is consistent with economic theories. In addition, the reason for using a value of $\mathfrak{q} \ll 1000$ is due to the presence of very many nuisance time series predictors in response to the TCR. Therefore, by removing them significant predictors are revealed, and one expects to improve the predictive performance. The SFAR provides a convenient and effective strategy for C-level factors construction using the multi-category data. At last, predictors in $\boldsymbol{Z}_{t-h}$ are combined with these C-level factors in $\boldsymbol{F}_{t-h}$ in the TSR regression framework to build a forecasting model.

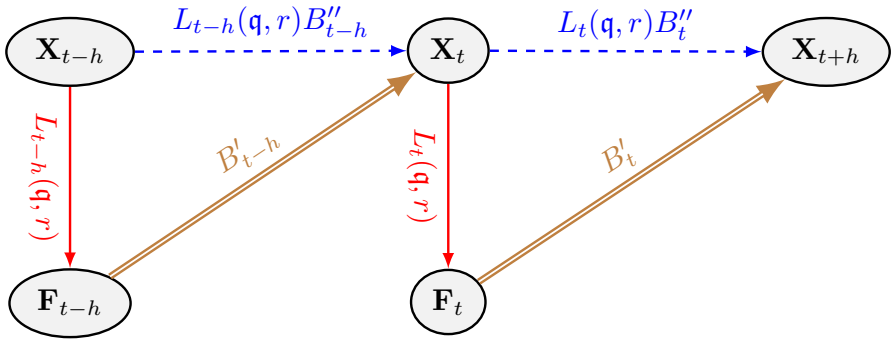
Our proposed SFAR relies on the cardinality-rank constraint, interplaying with a sparsity design and rank restriction to achieve model interpretability and prediction accuracy. In addition, the

orthogonality constraint, ensures that the C-level factors are uncorrelated to each other. The strength of the SFAR is attributed to its ability to precisely explain the TCR and series therein by adding the *target specific component* to the C-level factors. The target specific component, responsible for the predictive power of the factors, is added by constraint (C2) and the supervised learning (M3) guaranteeing that extracted factors best describe the TCR (regardless of the size of p).

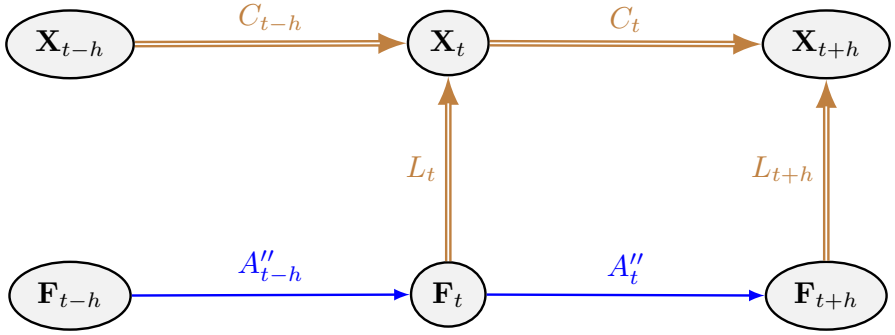
Distinct to the size of p, the lag order m is essential to the SFAR because it: **i)** benefits variable selection by making the candidate pool even larger, and **ii)** makes the significant (actual) lag order to extract factors identifiable. Then, given m the SFAR estimates it via (M3). Therefore, initial m is only restricted to the past observations of the (raw) time series available. With m given the SFAR extracts factors in a supervised manner (in particular, RRR), as (implicitly) shown in (M3). However, other methodologies may require the actual m to be chosen *a priori*. For example, dynamic factor model (DFM) assumes the actual lag order is known *a priori*. DFM utilizes all the predictors and their (*known*) lags in an unsupervised manner (specifically, PCA) to construct factors given the extremely difficulty involved in fitting a model to jointly identify the lag order, extract factors, and estimate coefficients. Another relevant difference lies in approximating the TCR. The SFAR in (M3), approximates the TCR via extracted factors, while DFM uses constructed factors and past observations of the predictors.

Figure (3.1) shows the mechanisms of the SFAR and DFM to build factors (and approximate the TCR), and their dynamic nature. For clarity, we write $\boldsymbol{L}$ as $\boldsymbol{L}_{t-h}$ and $\boldsymbol{L}_t$, and $\boldsymbol{B}''$ as $\boldsymbol{B}''_{t-h}$ and $\boldsymbol{B}''_t$ in the following figures to make clear they refer to different lag orders.

To approximate $\boldsymbol{X}_t$ (or TCR), the SFAR mechanism in subfigure (3.1a) starts training the coefficient matrices $\boldsymbol{L}_{t-h}(\mathfrak{q}, r)$ and $\boldsymbol{B}''_{t-h}$, as shown by the dashed line in subfigure (3.1a). Then, factors $\boldsymbol{F}_{t-h}$ are extracted using only the $\mathfrak{q}$ relevant predictors in response to $\boldsymbol{X}_t$, as seen from the solid line to $\boldsymbol{F}_{t-h}$. At last, the double line ending at $\boldsymbol{X}_{t-h}$ displays its approximation via $\boldsymbol{F}_{t-h}\boldsymbol{B}''_{t-h}$. In contrast DFM, uses the mechanism in subfigure (3.1b) to train $\boldsymbol{F}_t$ (from constructed factors) through $\boldsymbol{F}_{t-h}\boldsymbol{B}''_{t-h}$ as shown by the solid line concluding at $\boldsymbol{F}_t$. Next, $\boldsymbol{X}_t$ is approximated via $\boldsymbol{F}_{t-h}\boldsymbol{B}''_{t-h} + \boldsymbol{X}_{t-h}\boldsymbol{C}''_{t-h}$ as illustrated by both double lines finishing at $\boldsymbol{X}_t$. The SFAR is a three-step mechanism for approximating $\boldsymbol{X}_t$ while the DFM is a two-step mechanism.



(a) Mechanism of the SFAR and its dynamic nature



(b) Mechanism of the DFM and its dynamic nature

Figure 3.1: Mechanisms to approximate the TCR via factor construction

Despite the differences, factors constructed via the proposed sparse factor auto-regression are close in spirit to those from DFM. As shown in Figure (3.1), DFM extracts factors as an approximation of those in previous periods. Therefore, factor extraction to approximate $\boldsymbol{X}_{t+h}$ via mechanism in subfigure (3.1b) occurs as follows:

$$\begin{aligned}\boldsymbol{F}_{t+h} &\approx \boldsymbol{F}_t \boldsymbol{B}''_t \\ &\approx \boldsymbol{F}_{t-h} \boldsymbol{B}''_{t-h} \boldsymbol{B}''_t.\end{aligned}$$

Likewise, factors extracted via SFAR are an approximation of factors in previous periods as can be easily seen through a recursive iteration. Thus, approximating $\boldsymbol{X}_{t+h}$ through the mechanism in

subfigure (3.1a) takes place as follows:

$$\begin{aligned}
\boldsymbol{X}_{t+h} &\approx [\boldsymbol{X}_t,\ \boldsymbol{X}_{t-h}]\ \boldsymbol{L}_{t+h}(\mathfrak{q}, r)\boldsymbol{B}''_{t+h} \\
&\approx [\boldsymbol{X}_t\boldsymbol{L}_t(\mathfrak{q}, r)\boldsymbol{B}''_t,\ \boldsymbol{X}_{t-h}\boldsymbol{L}_{t-h}(\mathfrak{q}, r)\boldsymbol{B}''_{t-h}]\ \boldsymbol{L}_{t+h}(\mathfrak{q}, r)\boldsymbol{B}''_{t+h} \\
&\approx [\boldsymbol{F}_t\boldsymbol{B}''_t,\ \boldsymbol{F}_{t-h}\boldsymbol{B}''_{t-h}]\ \boldsymbol{L}_{t+h}(\mathfrak{q}, r)\boldsymbol{B}''_{t+h} \\
&\approx \begin{bmatrix}\boldsymbol{F}_t & \boldsymbol{F}_{t-h}\end{bmatrix} \begin{bmatrix}\boldsymbol{B}''_t & 0 \\ 0 & \boldsymbol{B}''_{t-h}\end{bmatrix} \boldsymbol{L}_{t+h}(\mathfrak{q}, r)\boldsymbol{B}''_{t+h} \approx \boldsymbol{F}_{t+h}\boldsymbol{B}''_{t+h}
\end{aligned}$$

where $\boldsymbol{F}_{t+h}$ encompasses $\boldsymbol{F}_t$ and $\boldsymbol{F}_{t-h}$ as in DFM. Clearly the relationship holds for any lag order m. The recursive iteration above corroborates that C-level factors extracted via SFAR include the dynamic component of the DFM shown in subfigure (3.1b).

3.3 SFAR via penalized maximum likelihood estimate (MLE)

Consider a multivariate vector response model

$$\boldsymbol{Y}_t = \boldsymbol{X}_{t-h}\boldsymbol{C} + \boldsymbol{E}_t \tag{3.3.1}$$

where $\boldsymbol{C}$ is a $p \times n$ matrix of unknowns, and $\boldsymbol{E}_t$ is the $T \times p$ matrix of errors with independent entries.

One may be tempted to estimate $\boldsymbol{B}$ by performing an OLS regression. For the OLS regression to apply in (3.3.1) the vector predictors in $\boldsymbol{X}$ must be treated as being uncorrelated with $\boldsymbol{E}_t$. However, if $\boldsymbol{X}$ is simultaneously determined with $\boldsymbol{Y}$ and is correlated with $\boldsymbol{E}_t$, the OLS regression would be inconsistent. Therefore, as long as the error matrix $\boldsymbol{E}_t$ is a white noise process, or more generally, is stationary and independent of $\boldsymbol{X}$, model (3.3.1) can be estimated by the OLS.

Model (3.3.1) is a model in which a group of variables is in turn explained by its own lagged values, plus current and past values of other variables. Such models are known as Vector Auto-Regression (VAR). VAR is used to summarize the dynamics of macroeconomic data. The VAR framework provides a systematic way to capture rich dynamics in multiple time series, and its statistical component is easy to interpret, see, Stock and Watson [63]. As Sims [56] and others argued in a series of influential papers, VARs held out the promise of providing a coherent and credible approach to data description and forecasting macroeconomic issues.

Consider model (3.3.1) with $l = 1$, which is a VAR(1) model, stated as

$$\boldsymbol{y}_t = \boldsymbol{y}_{t-1}\boldsymbol{A} + \boldsymbol{u}_t \tag{3.3.2}$$

where $\boldsymbol{A}$ is a matrix of coefficients and $\boldsymbol{u}_t \sim N(0, \boldsymbol{\Omega_u})$. Evidently, (3.3.2) may have a larger lag order, however, without loss of generality, the VAR(1) model can be assumed since any VAR(l) process can be written in VAR(1) form, see, Helmut [37].

The primary principle on which LS is based is maximum likelihood (ML). Suppose we have observed p-vectors of size $T; y_1, \ldots, y_T$. The approach will be to calculate the probability density

$$f(\boldsymbol{y}_1, \boldsymbol{y}_2, \ldots, \boldsymbol{y}_p; \boldsymbol{A}), \tag{3.3.3}$$

which may be loosely viewed as the probability of having observed those particular vectors predictors. The maximum likelihood estimate (MLE) of $\boldsymbol{A}$ is the value for which those vectors are most likely to have been observed, that is, it is the value of $\boldsymbol{A}$ that maximizes (3.3.3).

Given $\boldsymbol{u}_t$ is a white noise, the joint density of the first t observations is then

$$f(\boldsymbol{y}_1, \boldsymbol{y}_2, \ldots, \boldsymbol{y}_p|\boldsymbol{A}) = f(\boldsymbol{y}_{t-1}, \boldsymbol{y}_{t-2}, \ldots, \boldsymbol{y}_1|\boldsymbol{A}) f(\boldsymbol{y}_t|\boldsymbol{y}_{t-1}; \boldsymbol{A}).$$

where the likelihood of the complete sample can thus be calculated as

$$L(\boldsymbol{A}|\boldsymbol{y}_1, \boldsymbol{y}_2, \ldots, \boldsymbol{y}_p) = f(\boldsymbol{y}_1|\boldsymbol{A}) \prod_{t=2}^{p} f(\boldsymbol{y}_t|\boldsymbol{y}_{t-1}; \boldsymbol{A}) \tag{3.3.4}$$

The conditional ML estimates are trivial to compute. Moreover, if the sample size T is sufficiently large, the first observation makes a negligible contribution to the total likelihood. As a result we can write equation (3.3.4) as

$$f(\boldsymbol{A}|\boldsymbol{y}_2, \ldots, \boldsymbol{y}_p) = \prod_{t=2}^{p} (2\pi)^{-T/2} \boldsymbol{\Omega_u}^{-1/2} \exp\left\{ -\frac{1}{2} (\boldsymbol{y} - \boldsymbol{y}_{t-1}\boldsymbol{A})' \boldsymbol{\Omega_u} (\boldsymbol{y}_t - \boldsymbol{y}_{t-1}\boldsymbol{A}) \right\}$$

where the conditional ML estimate of $\boldsymbol{A}$ can be obtained by solving

$$\min_{\boldsymbol{A}} \frac{1}{2} \sum_{t=2}^{p} ||\boldsymbol{y}_t - \boldsymbol{y}_{t-1}\boldsymbol{A}||_2^2. \tag{3.3.5}$$

Given the matrix form of equation (3.3.2) we conveniently set $\boldsymbol{A} = \boldsymbol{C}$, the matrix of responses $\boldsymbol{Y}_t = \boldsymbol{y}_2, \ldots, \boldsymbol{y}_p$; and the matrix of predictors $\boldsymbol{X}_{t-h} = \boldsymbol{y}_1, \ldots, \boldsymbol{y}_{p-1}$. A VAR model results with Auto-Regression predictors in $\boldsymbol{X}$ and predictors in $\boldsymbol{Y}_t$. Then the problem is restated as follows

$$\boldsymbol{C}_{ML} = \underset{\boldsymbol{C}}{\operatorname{argmin}}\ \ell(\boldsymbol{A}) = \frac{1}{2} ||\boldsymbol{Y}_t - \boldsymbol{X}_{t-h}\boldsymbol{C}||_F^2. \tag{3.3.6}$$

In most applications the parameters of an auto-regression model are estimated via conditional maximum likelihood since the estimate of the parameters can be obtained from an OLS regression on y_t on a finite number of lags of its own values. However, variables in macroeconomic time series data are linearly correlated. Therefore, the ML estimate is not an ideal choice for modeling macroeconomic time series data. In practice, there is usually high collinearity in $\boldsymbol{X}_{t-h}$, especially when the data contains dynamic series, and the number of observations is limited. Moreover, the ML estimate does not enforce sparsity and consequently $\boldsymbol{C}_{ML}$ is difficult to interpret. To improve prediction accuracy and obtain an interpretable model, shrinkage estimation is necessary. It can be done by adding a penalty or constraint to the coefficient matrix.

We propose to regularize the model and study it as a whole. Regularization achieves interpretability and forecasting accuracy by penalizing predictors in $\boldsymbol{X}_{t-h}$ through a constraint imposed on their corresponding row-coefficients in $\boldsymbol{C}$ such that irrelevant predictors are removed entirely. For example, we can use the penalized maximum likelihood (PML) estimation

$$\boldsymbol{C}_{PML} = \underset{\boldsymbol{C}}{\operatorname{argmin}}\ \ell(\boldsymbol{A}) + P(\boldsymbol{A};\lambda) \tag{3.3.7}$$

where λ is the regularization parameter, P is the penalty function, and denote $P(\boldsymbol{A};\lambda) = \sum_j P(\boldsymbol{c}_j;\lambda)$. Many additive penalties have been heavily studied in the literature and can be applied here. For example, Lasso, Tibshirani [69] solves the ℓ_1 penalization problem. It is mathematically elegant and easy to solve but suffers from drawbacks such as selection inconsistency and estimation bias, as well as an incapability of dealing with collinearity. A nonconvex option is the $\ell_0+\ell_2$ hard-ridge, see, She [53] which advocates row sparsity. However, due to nonconvexity, the solutions only converge to local optimum and depend on the choices of the initial points. As a result selection is not stable. Rather than additive penalties, either convex or nonconvex, constraints can also be used to achieve selectability within collinear predictors.

As explained in the following section using a constraint rather than an additive penalty on $\boldsymbol{C}$ provides an ideal setup to, for example, select a determined set of variables without the bothersome work of tuning a regularization parameter and specifying parameters that are meaningless and lack interpretability. Model (3.3.1) provides the starting point in developing a strategy to overcome the difficulties involved in forecasting time series with very many predictors. In the next section we provide a strategy for training the SFAR and performing the forecasting of the target series of interest.

3.4 A three-stage SFAR fitting strategy

Directly fitting (M1) - (M3) subject to (C1) and (C2) is computationally prohibitive. Instead we propose a three-stage fitting strategy of the SFAR which encompasses the three stages: **1)** *selectable reduced-rank regression*, **2)** *orthogonal factor construction*, and **3)** *forecasting model training*.

Stage 1: Selectable Reduced-Rank Regression. This stage integrates (M2) into (M3), and trains a multivariate model in response to the TCR, subject to the rank and cardinality constraints in (C2). Given a multivariate model with coefficient matrix $\boldsymbol{C}$ (which corresponds to $\boldsymbol{LB''}$), we impose a row sparsity pattern on $\boldsymbol{C}$ (as well as low rankness). Let P be a sparsity promoting penalty (such as the convex group ℓ_1, see, Yuan [73]), thus, the problem in stage 1 can be formulated as follows

$$\hat{\boldsymbol{C}}(\mathfrak{q}, r) = \min_{\boldsymbol{C}} ||\boldsymbol{Y}_t - \boldsymbol{X}_{t-h}\boldsymbol{C}||_F^2 + \frac{\eta}{2}||\boldsymbol{C}||_F^2 + \sum_{j=1}^{p} P(\|\boldsymbol{c}_j\|, \lambda) \;\; s.t. \;\; \text{rank}(\boldsymbol{C}) \leq r. \tag{3.4.1}$$

referred to as the selectable reduced rank regression estimator. Using the convex group ℓ_1 as the penalty choice seems to be convenient given its computational efficiency. Nevertheless, we advocate nonconvex penalties such as the group ℓ_0, see She [53], since they enforce further parsimony in $\boldsymbol{C}$ (more details in section (4.3)). The ridge penalty in (3.4.1) is necessary because of the collinearity issue in high dimensions among time series data (often seen in finance and economic data).

The parameter tuning issuance cannot be ignored. One usually specifies a grid of λ values and obtains solution paths to choose the best. However, this is computationally expensive and the ideal mode comparison criterion is unknown in our setup. Therefore, rather than using a penalty, we propose to impose an ℓ_0 constraint to have a cardinality control of the number of relevant predictors in $\boldsymbol{X}_{t-h}$. Consequently, the doubly constrained regression problem is

$$\hat{\boldsymbol{C}}(\mathfrak{q}, r) = \min_{\boldsymbol{C}} ||\boldsymbol{Y}_t - \boldsymbol{X}_{t-h}\boldsymbol{C}||_F^2 + \frac{\eta}{2}||\boldsymbol{C}||_F^2 \;\; s.t. \;\; \text{rank}(\boldsymbol{C}) \leq r, \|\boldsymbol{C}\|_{2,0} \leq \mathfrak{q}. \tag{3.4.2}$$

where $\|\boldsymbol{C}\|_{2,0} = \sum_{j=1}^{p} 1_{\{c_j \neq 0\}}$, $\text{rank}(\boldsymbol{C}) \leq r$ is the rank restriction on $\boldsymbol{C}$, and $\|\boldsymbol{C}\|_F^2 = \sum_{i,j} c_{i,j}^2$ is the Frobenius norm of matrix $\boldsymbol{C}$. Equivalently, this fully incorporates the constraint (C2) on $\boldsymbol{L}$. The selectable reduced-rank regression estimator imposes a simultaneous cardinality and rank constraints on $\boldsymbol{C}$ such that $\hat{\boldsymbol{C}}(\mathfrak{q}, r)$ has only $\mathfrak{q}$ nonzero rows, and reduced rank (r). A direct control over the number of predictors to be selected provides some advantages. For example, the user can conveniently specify a desired number of predictors based on prior knowledge instead of a somewhat

meaningless regularization parameter λ. This poses a great challenge in computation which will be treated in section (4.3). The guaranteed dimension reduction in $\boldsymbol{C}$ helps extracting meaningful factors in forecasting.

Stage 2: Orthogonal factor construction. This stage uses the estimate $\hat{\boldsymbol{C}}(\mathfrak{q}, r)$ from Stage 1 to extract predictive C-level factors $\boldsymbol{F}_{t-h}$ for forecasting the TSR. Because of (C2), $\hat{\boldsymbol{C}}$ has an inherent low rankness and sparse rows. Factors are constructed using linear combinations of the predictors in $\boldsymbol{X}_{t-h}$, and are preferred to be decorrelated, as seen in (C1). We propose a simple means to solve the problem. With $\hat{\boldsymbol{C}}(\mathfrak{q}, r)$ and $\boldsymbol{X}_{t-h}$ available, perform spectral decomposition on $\boldsymbol{X}_{t-h}\hat{\boldsymbol{C}}(\mathfrak{q}, r)$ as follows

$$\hat{\boldsymbol{C}}^{\mathsf{T}}(\mathfrak{q}, r)\boldsymbol{X}_{t-h}^{\mathsf{T}}\boldsymbol{X}_{t-h}\hat{\boldsymbol{C}}(\mathfrak{q}, r) = \boldsymbol{U}\boldsymbol{D}\boldsymbol{U}^{\mathsf{T}}$$

where $\boldsymbol{U} \in \mathbb{R}^{p\times r}$ is an orthogonal matrix satisfying $\boldsymbol{U}^{\mathsf{T}}\boldsymbol{U} = \boldsymbol{I}$, and $\boldsymbol{D}$ is an $r \times r$ diagonal matrix. Then,

$$\boldsymbol{X}_{t-h}\hat{\boldsymbol{C}}(\mathfrak{q}, r) = \boldsymbol{X}_{t-h}\hat{\boldsymbol{C}}(\mathfrak{q}, r)\boldsymbol{U}\boldsymbol{U}^{\mathsf{T}},$$

and the C-level factors are given by

$$\boldsymbol{F}_{t-h} \triangleq \boldsymbol{X}_{t-h}(\hat{\boldsymbol{C}}(\mathfrak{q}, r)\boldsymbol{U})$$

Additionally, this provides a matrix of factors $\boldsymbol{F}_{t-h} \in \mathbb{R}^{T\times r}$ with only r columns. The dimension reduction can be dramatic when r is much smaller than p.

Next, we verify the orthogonality (C1).

$$\boldsymbol{F}_{t-h}^{\mathsf{T}}\boldsymbol{F}_{t-h} = \boldsymbol{U}^{\mathsf{T}}(\boldsymbol{X}_{t-h}\hat{\boldsymbol{C}}(\mathfrak{q}, r))^{\mathsf{T}}\boldsymbol{X}_{t-h}(\hat{\boldsymbol{C}}(\mathfrak{q}, r)\boldsymbol{U}) = \boldsymbol{U}^{\mathsf{T}}\boldsymbol{U}\boldsymbol{D}\boldsymbol{U}^{\mathsf{T}}\boldsymbol{U} = \boldsymbol{D} \tag{3.4.3}$$

Equation in (3.4.3) shows that constructed factors are orthogonal to each other, which is preferable in model fitting and in improving the forecasting accuracy. After Stage 1 and 2 are completed, (M3) and (M2) are have been implemented, with constraints (C1) and (C2) satisfied.

Stage 3: Forecasting model training. The final stage uses (M1) to build a forecasting model of the TSR by combining potentially interesting (raw) predictors in $\boldsymbol{Z}_{t-h}$ and factor predictors in $\boldsymbol{F}_{t-h}$, in a regression framework. The fitting only involves a univariate response regression.

However, in some situations, it is possible that $\boldsymbol{Z}_{t-h}$ still contains a large number of predictors. Fortunately, this is a recently heavily studied problem and many modern variable selection approaches such as the Lasso, (Tibshirani [69]), can be applied.

Forecasting y_s with $s > t$ using the estimates $\hat{\boldsymbol{L}}$, $\hat{\boldsymbol{B}}$, and $\hat{\boldsymbol{B}}'$ is performed as follows. Suppose the observations of the TSR are available up to time t, and one wants to forecast y_{t+h} for $h \geq 1$. To that end, one uses the newly constructed category level factors in $\hat{\boldsymbol{F}}_{t-h}$, which incorporate the dynamic feature provided by the loading factor matrix $\hat{\boldsymbol{L}}$, and its coefficient matrix $\hat{\boldsymbol{B}}'$. In addition, predictors in $\boldsymbol{Z}_{t-h}$ and its coefficient matrix $\hat{\boldsymbol{B}}$ are also used.

Therefore, to compute $\hat{y}_s$ one period ahead (M1) is used as follows

$$\hat{\boldsymbol{y}}_{t+h} = \boldsymbol{z}_t^{\mathsf{T}} \hat{\boldsymbol{B}} + \boldsymbol{f}_t^{\mathsf{T}} \hat{\boldsymbol{B}}' \tag{3.4.4}$$

where, for example, $\boldsymbol{z}_t^{\mathsf{T}}$ corresponds to the (transpose) vector of the past observations of y_{t+h} contained in $\boldsymbol{Z}$.

In case one is interested in long term forecasting $y_{t+\kappa h}$ for $\kappa \geq 2$, that is, forecasting y_{t+h} more than one period ahead, one may use the iterated one period ahead forecast. This method forecasts estimates $\hat{y}_{t+h}$ using information up to time t, as shown in (3.4.4). Then, $\hat{y}_{t+2h}$ is estimated via (3.4.4) with $\hat{y}_{t+h}$ replacing the furthest observation in $\boldsymbol{z}_t^{\mathsf{T}}$. Next, $\hat{y}_{t+3h}$ is estimated with $\hat{y}_{t+2h}$ and $\hat{y}_{t+h}$ replacing the furthest observations in $\boldsymbol{z}_t^{\mathsf{T}}$. The procedure is repeated until the desired κ is forecasted. However, in this document the direct multi-period ahead forecast is implemented through the *rolling window scheme* technique, see, e.g. Stock and Watson [62], which is described in section 4.

CHAPTER 4

JOINTLY RANK-CARDINALITY CONSTRAINS IN FACTOR REGRESSION

Two practical concerns are always present when it comes to forecasting time series in high dimensions with very many predictors: accuracy and interpretability. In this chapter, we propose a novel means for rank-constrained variable screening which can remarkably reduce the computational cost for applications in ultrahigh dimensions. The key is to change the group selection penalty to a group ℓ_0 constraint and add in progressive squeezing operations. The ℓ_0-constraint enforces a sparsity design in the coefficient matrix as a manner to remove nuisance time series. Thus, enforcing a row-sparsity pattern in the coefficient matrix is critical to identify a handful relevant set of predictors in response to the TSR, while the rank constraint uses selected predictors to construct factors.

We propose the computational component of the SFAR. First, the selectable reduced-rank regression estimator which is a doubly-constrained estimator of the coefficient matrix $\boldsymbol{C}$ is defined. Moreover, the multivariate quantile thresholding rule, essential to perform variable selection by eliminating a vector predictor entirely, is introduced. The computational part to solve the three-stage SFAR fitting strategy is developed. At last, the PRCCR algorithm is also given.

4.1 Selectable Reduced-Rank Regression

Consider the multivariate vector response model (3.3.1) which is conveniently restated below

$$\boldsymbol{Y}_t = \boldsymbol{X}_{t-h}\boldsymbol{C} + \boldsymbol{E}_t \tag{4.1.1}$$

where $\boldsymbol{Y}_t$ is a $T \times n$ matrix of vector responses, $\boldsymbol{X}_{t-h}$ is the $T \times p$ matrix of predictors, $\boldsymbol{C}$ is a $p \times n$ matrix of unknowns, and $\boldsymbol{E}_t$ represents the $T \times p$ matrix of errors.

To enforce sparsity in $\boldsymbol{C}$ one can use additive penalties like those mentioned in chapter 2. For example, one may apply a group penalty of the form

$$\hat{\boldsymbol{C}} = \min_{\boldsymbol{C}} \frac{1}{2}||\boldsymbol{Y}_t - \boldsymbol{X}_{t-h}\boldsymbol{C}||_F^2 + \frac{\eta}{2}\sum_{j=1}^{p} P(\boldsymbol{c}_j; \lambda) \tag{4.1.2}$$

where the penalty choice in (4.1.2) can be as flexible as the group ℓ_1 by Yuan and Lin [73]: $\sum_{j=1}^{p}\lambda_j\|\boldsymbol{c}_j\|_2$, or $\lambda\|\boldsymbol{C}\|_{2,1}$ when $\lambda_j=\lambda$, and the group ℓ_0: $\sum_{j=1}^{p}(\lambda_j^2/2)1_{\boldsymbol{c}_j\neq 0}$, or $(\lambda^2/2)\|\boldsymbol{C}\|_{2,0}$ when $\lambda_j=\lambda$. Other choices include group SCAD, group l_p, and group *hard-ridge*, see, She [53].

Sparsity can be promoted by using the adequate penalty. The chosen penalty must achieve row-sparsity in $\boldsymbol{C}$ such that irrelevant predictors are removed in their totality. Additionally, it is desirable for the penalty to handle collinearity among vector time series. From those available the *hard-ridge* penalty is desirable since it simultaneously enforces sparsity and removes collinearity among predictors. Thus, one may apply the group $\ell_2{+}\ell_0$ *hard-ridge* penalty as follows

$$\hat{\boldsymbol{C}}=\min_{\boldsymbol{C}}||\boldsymbol{Y}_t-\boldsymbol{X}_{t-h}\boldsymbol{C}||_F^2+\frac{\lambda^2}{2(1+\eta)}\|\boldsymbol{C}\|_{2,0}+\frac{\eta}{2}\|\boldsymbol{C}\|_F^2 \tag{4.1.3}$$

where $\|\boldsymbol{C}\|_{2,0}:=\sum_j 1_{\boldsymbol{c}_j\neq 0}$, and $\frac{\eta}{2}\|\boldsymbol{C}\|_F^2$ penalizes $\sqrt{\sum_{i=1}^{r}d_i}$ where d_i are the singular values of $\boldsymbol{C}$. Equation (4.1.3), which is equivalent to the minimization problem in (2.1.15), achieves parsimony in estimation with only relevant predictors in $\boldsymbol{C}$. Suppose one wants to select a specific number of relevant predictors. To that end, parameters λ and η need to be tuned and thus the computational effort raises. Instead, imposing a constraint on $\boldsymbol{C}$ rather than tuning λ seems appropriate. Accordingly the ℓ_0-constraint can be used as defined by: $\frac{\lambda^2}{2}\sum 1_{\boldsymbol{c}_j\neq 0}=\frac{\lambda^2}{2}\#\text{ nz}(\boldsymbol{C})=\frac{\lambda^2}{2}||\boldsymbol{c}_j||_0:=\frac{\lambda^2}{2}\times$ the set cardinality of $\boldsymbol{C}$. Consequently, imposing a constraint over the number of row vector coefficients in (4.1.3) conveniently replaces the term $\frac{\lambda^2}{2(1+\eta)}||\boldsymbol{C}||_{2,0}$ by the preferred ℓ_0-constraint: $\|\boldsymbol{C}\|_{2,0}=\sum_j 1_{\boldsymbol{c}_j\neq 0}=\mathfrak{q}$.

We propose to use the ℓ_0-constraint over $\boldsymbol{C}$ to facilitate the variable selection process. With the ℓ_0-constraint the minimization problem is:

$$\hat{\boldsymbol{C}}=\min_{\boldsymbol{C}}\|\boldsymbol{Y}_t-\boldsymbol{X}_{t-h}\boldsymbol{C}\|_F^2+\frac{\eta}{2}\|\boldsymbol{C}\|_F^2\ \ s.t.\ \ \|\boldsymbol{C}\|_{2,0}\leq\mathfrak{q}. \tag{4.1.4}$$

where $\|\boldsymbol{C}\|_{2,0}\leq\mathfrak{q}$ represents the ℓ_0-constraint. The ℓ_0-constraint reduces the computationally effort in (4.1.4) to only one parameter. This constraint represents an advantage for users of the SFAR since the number of relevant variables can be selected *a priori*.

The minimization problem in (4.1.4) efficiently solves the variable selection issue. Nonetheless, it may not be able to provide an interpretable model yet. To achieve interpretability we may reduce the selected $\mathfrak{q}$ relevant predictors by constructing r linear combinations of those predictors. In that way, only the selected $\mathfrak{q}$ predictors appropriately contribute in factor construction. This reduction

can be accomplished by imposing a rank constraint on $\boldsymbol{C}$ in (4.1.4). Adding the rank constraint results in

$$\hat{\boldsymbol{C}} = \min_{\boldsymbol{C}} \|\boldsymbol{Y}_t - \boldsymbol{X}_{t-h}\boldsymbol{C}\|_F^2 + \frac{\eta}{2}\|\boldsymbol{C}\|_F^2 \;\; s.t. \;\; \text{rank}(\boldsymbol{C}) \leq r, \;\; \|\boldsymbol{C}\|_{2,0} \leq \mathfrak{q} \tag{4.1.5}$$

where $\text{rank}(\boldsymbol{C}) \leq r$ indicates that at most r factors are constructed. The rank constraint provides a factor-driven model that is interpretable. In particular, our factor construction procedure constructs C-level orthogonal factors to improve the predictability power of the model. When both constraints are simultaneously implemented we have the **joint rank-sparsity constrained estimator** providing interpretability and accuracy for time series forecasting in high dimensions with very many predictors.

The desired number of C-level factors constructed is a linear function of the rank constraint imposed on $\boldsymbol{C}$. As a result, the SFAR is a *rank-driven* C-level factors methodology. The rank-driven approach of the SFAR provides another advantage for users as they may conveniently set the value of r *a priori* based on speculations, experience, or a particular goal, among other reasons.

The estimator in (4.1.5) is the Selectable Reduced-Rank Regression (SEL-RRR) used in Stage 1. The rank constraint brings low rankness for the new feature matrix, while the group ℓ_0-constraint enforces sparsity on $\boldsymbol{C}$ for feature selection. This procedure acts like the lasso at the group level: depending on a regularization parameter λ, an entire group of predictors may drop out of the model. The group lasso yields sparsity within a group. That is, if a group of parameters is zero, they all will be zero. The *hard-ridge* penalty supports the idea of thresholding function. For $P(\boldsymbol{c}_j; \lambda)$ a predictor is considered if $P(\tau; \lambda) = \frac{\lambda^2}{2} 1_{\tau \neq 0}$. Then, the implicitly penalty within $\|\boldsymbol{C}\|_{2,0} \leq \mathfrak{q}$ in (4.1.5) is $P = \sum_j \frac{\lambda^2}{2} 1_{\boldsymbol{c}_j \neq 0}$ where $\frac{\lambda^2}{2}$ represents the number of nonzero rows in $\boldsymbol{C}$. The SEL-RRR can handle very large values of p and is of particular interest to our computational procedure. A doubly-constrained estimator ensures that both practical concerns are met.

4.2 Multivariate Quantile Thresholding Rule

We encourage using thresholding rules rather than different forms of the penalty function to tackle the computational challenge for nonconvex P even though there is a universal connection between thresholding rules and penalty functions, see, She [52]. One direct reason is that different P's may result in the same estimator and the same thresholding. Moreover, starting with thresholding functions facilitates the computation.

Definition 4.2.1 (Threshold function)**.** *A threshold function is a real valued function* $\Theta(\tau;\lambda)$ *defined for* $-\infty < \tau < \infty$ *with* λ *as the parameter (*$0 \leq \lambda < \infty$*) such that*

1) $\Theta(\tau;\lambda) = -\Theta(\tau;\lambda)$,

2) $\Theta(\tau;\lambda) \leq \Theta(\tau^{\mathsf{T}};\lambda)$ *for* $\tau \leq \tau^{\mathsf{T}}$,

3) $\lim_{\tau\to\infty} \Theta(\tau;\lambda) = \infty$, *and*

4) $0 \leq \Theta(\tau;\lambda) \leq \tau$ *for* $0 \leq \tau < \infty$.

In words, $\Theta(\cdot;\lambda)$ is an odd monotone unbounded shrinkage rule for τ, at any λ. A multivariate version of the thresholding operator Θ is defined componentwise if either τ or λ is replaced by a vector. For any vector $\boldsymbol{c} \in \mathbb{R}^n$, $\vec{\Theta}(\boldsymbol{c};\lambda) := \boldsymbol{c}\Theta(\|\boldsymbol{c}\|_2;\lambda)/\|\boldsymbol{c}\|_2$ for $\boldsymbol{c} \neq 0$ and 0 otherwise. For any matrix $\boldsymbol{C} \in \mathbb{R}^{p\times n}$ with $\boldsymbol{C} = [\boldsymbol{c}_1, \cdots, \boldsymbol{c}_p]^{\mathsf{T}}$, $\vec{\Theta}(\boldsymbol{C};\lambda) := [\vec{\Theta}(\boldsymbol{c}_1;\lambda), \cdots, \vec{\Theta}(\boldsymbol{c}_p;\lambda)]^{\mathsf{T}}$.

Our trick to tackle the rank constraint in (4.1.5) is to write $\boldsymbol{C} = \boldsymbol{S}\boldsymbol{V}^{\mathsf{T}}$ with $\boldsymbol{V}$ being orthogonal and $\boldsymbol{S}$ the orthogonal projection of $\hat{\boldsymbol{C}}$ onto a lower dimensional space r. Using the decomposition $\boldsymbol{C} = \boldsymbol{S}\boldsymbol{V}^{\mathsf{T}}$ we have that when $\boldsymbol{V}$ is given $\boldsymbol{S}$ is easily estimated, and in turn when $\boldsymbol{V}$ is given $\boldsymbol{S}$ can be solved by the thresholding-based iterative selection procedures (TISP) for model selection and shrinkage, see, She [52]. A variant of particular interest is to estimate $\boldsymbol{C} = \boldsymbol{S}\boldsymbol{V}^{\mathsf{T}}$ via the threshold function in the SEL-RRR estimator.

Throughout our research we used a multivariate version of the threshold function which can be applied to vectors rather than elementwise. It can be seen as a net with big holes useful to retain only those $\mathfrak{q}$ vector predictors whose 2-norm is $\geq$ a threshold λ. The shrinkage rule $\Theta(\|\boldsymbol{c}_j\|_2;\lambda)$ serves as an indicator function that states whether the vector predictor under evaluation is incorporated to the final model. The ratio $\boldsymbol{c}_j/\|\boldsymbol{c}_j\|_2$ computes the magnitude of the selected predictors onto the projections. Notice that the threshold function has no concern about the direction of the new dimensions but on their magnitudes. As a result, the selected predictors will be those with the most contribution to describe the variability of the category data response. It is worth to mention that the threshold multivariate function works only in supervised settings since predictors are selected in response to some variable(s) of interest.

Employing various thresholding rules, one can reach all commonly used P, including group versions of ℓ_1, ℓ_0, SCAD, ℓ_p, elastic net. In particular, She [53] showed that the *hard-ridge thresholding* rule that fuses the hard-thresholding and the ridge-thresholding is

$$\Theta_{HR}(\tau;\lambda,\eta) = \begin{cases} 0, & \text{if } |\tau| < \lambda \\ \frac{\tau}{1+\eta}, & \text{if } |\tau| \geq \lambda. \end{cases} \tag{4.2.1}$$

Setting

$$q(\theta;\lambda,\eta)=\begin{cases}\frac{(1+\eta)(\lambda-|\theta|)^2}{2}, & \text{if } 0<|\theta|<\lambda\\ 0, & \text{if } \theta=0 \text{ or } |\theta|>\lambda\end{cases} \tag{4.2.2}$$

he obtained the $\ell_0+\ell_2$ penalty: $P(\theta)=\frac{1}{2}\eta\theta^2+\frac{1}{2}\frac{\lambda^2}{1+\eta}1_{\theta\neq 0}$, which corresponds to

$$\frac{1}{2}\eta\|\boldsymbol{C}\|_F^2+\frac{1}{2}\frac{\lambda^2}{1+\eta}\|\boldsymbol{C}\|_{2,0} \tag{4.2.3}$$

in (4.1.3). The *hard-ridge* penalty may be of interest in statistical learning tasks that have jointly concerns of accuracy and parsimony: the group ℓ_0 portion enforces parsimonious selection, while the ridge portion shrinks $\boldsymbol{C}$ to compensate for large noise and decorrelates the input variables in large-p applications tackling the collinearity among predictors.

The *quantile thresholding rule* $\Theta^{\#}(\cdot;\mathfrak{q},\eta)$ as a variant of the *hard-ridge* thresholding (4.2.1). Given $1\leq\mathfrak{q}\leq p$ and $\eta\geq 0$, $\Theta^{\#}(\boldsymbol{c};\mathfrak{q},\eta):\mathbb{R}^p\to\mathbb{R}^p$ is defined for any $\boldsymbol{c}\in\mathbb{R}^p$ such that the $\mathfrak{q}$ largest components of $\boldsymbol{c}$ (in absolute value) are shrunk by a factor of $(1+\eta)$ and the remaining components are all set to be zero. In case of ties, a random tie breaking rule is used. A matrix version $\vec{\Theta}^{\#}$ to be used in our problem is defined as

$$\vec{\Theta}^{\#}(\boldsymbol{C};\mathfrak{q},\eta)=\mathrm{diag}\{\Theta^{\#}(\boldsymbol{g}(\boldsymbol{C});\mathfrak{q},\eta)\}\boldsymbol{C}^{\circ}$$

for any $\boldsymbol{C}=[\boldsymbol{c}_1,\cdots,\boldsymbol{c}_p]^T\in\mathbb{R}^{p\times n}$, where $\boldsymbol{g}(\boldsymbol{C}):=[\|\boldsymbol{c}_j\|_2]_{p\times 1}$ and $\boldsymbol{C}^{\circ}=(\mathrm{diag}\{|\boldsymbol{g}(\boldsymbol{C})|\})^{+}\boldsymbol{C}$ with $^{+}$ standing for the Moore-Penrose pseudoinverse.

4.3 Selectable Category Factor Regression

One of the key steps of the three-stage fitting strategy of the SFAR model is to fit a penalized multivariate regression model for the target category response matrix subject to both rank and cardinality constraint. This, combined with Stage 2, guarantees extracting a small number of predictive factors from a *subset* of predictors identified in their row coefficients. With such joint rank reduction and nuisance dimension removement, the high dimensional challenge in multivariate problems can be addressed. Our particular interest lies in the selectable reduced-rank regression which offers direct cardinality and rank control in addition to the ridge shrinkage to deal with large noise and collinearity.

To deal with the computational challenge, WLOG, we write $\boldsymbol{C} = \boldsymbol{S}\boldsymbol{V}^{\mathsf{T}}$ with $\boldsymbol{V} \in \mathbb{O}^{n\times r}$ being orthogonal to get rid of the rank constraint. It is easy to show that (3.4.2) is equivalent to solving

$$(\hat{\boldsymbol{S}}, \hat{\boldsymbol{V}}) = \min_{(\boldsymbol{S},\boldsymbol{V})} ||\boldsymbol{Y}_t\boldsymbol{V} - \boldsymbol{X}_{t-h}\boldsymbol{S}||_F^2 + \frac{\eta}{2}||\boldsymbol{S}||_F^2 \;\; s.t. \;\; \|\boldsymbol{S}\|_{2,0} \leq \mathfrak{q}, \boldsymbol{V}^{\mathsf{T}}\boldsymbol{V} = \boldsymbol{I}. \tag{4.3.1}$$

We then use an alternative strategy in optimization. Fixing $\boldsymbol{V}$, estimator (4.3.1) minimizes only on the rank reduced $\boldsymbol{S}$ with orthogonal $\boldsymbol{V}$ free of regularization. This estimator is very flexible in the sense that it allows estimation to be over $\boldsymbol{S}$ with $\boldsymbol{V}$ given, and vice versa. Therefore, with $\boldsymbol{V}$ fixed estimator (4.3.1) becomes

$$\hat{\boldsymbol{S}} = \min_{\boldsymbol{S}} ||\boldsymbol{Y}_t\boldsymbol{V} - \boldsymbol{X}_{t-h}\boldsymbol{S}||_F^2 + \frac{\eta}{2}||\boldsymbol{S}||_F^2 \;\; s.t. \;\; \|\boldsymbol{S}\|_{2,0} \leq \mathfrak{q}. \tag{4.3.2}$$

The $\vec{\Theta}^{\#}$ to be used in our problem is defined as $\vec{\Theta}^{\#}(\boldsymbol{S}; \mathfrak{q}, \eta) = \text{diag}\{\Theta^{\#}(\boldsymbol{g}(\boldsymbol{S}); \mathfrak{q}, \eta)\}\boldsymbol{S}^{\circ}$ for any $\boldsymbol{S} = [\boldsymbol{s}_1, \cdots, \boldsymbol{s}_p]^{\mathsf{T}} \in \mathbb{R}^{p\times r}$, where $\boldsymbol{g}(\boldsymbol{S}) : [\|\boldsymbol{s}_j\|_2]_{p\times 1}$ and $\boldsymbol{S}^{\circ} = (\text{diag}\{|\boldsymbol{g}(\boldsymbol{S})\})^{+}\boldsymbol{S}$ with $^{+}$ standing for the Moore-Penrose pseudoinverse.

To iteratively optimize $\boldsymbol{S}$ we use:

$$\hat{\boldsymbol{S}} \leftarrow \vec{\Theta}^{\#}\left(\frac{1}{K}\boldsymbol{X}_{t-h}^{\mathsf{T}}\boldsymbol{Y}_t\boldsymbol{V} + (\boldsymbol{I} - \frac{1}{K}\boldsymbol{X}_{t-h}^{\mathsf{T}}\boldsymbol{X}_{t-h})\boldsymbol{S}^{(0)}; \mathfrak{q}, \eta\right). \tag{4.3.3}$$

The initial point of $\boldsymbol{S}$ affects the final solution due to nonconvexity. In general, we construct $\boldsymbol{S}^{(0)}$ as $(\boldsymbol{X}_{t-h}^{\mathsf{T}}\boldsymbol{X}_{t-h})^{+}\boldsymbol{X}_{t-h}^{\mathsf{T}}\boldsymbol{Y}_t\boldsymbol{V}_r$, based on the reduced-rank regression estimate $\hat{\boldsymbol{C}} = (\boldsymbol{X}_{t-h}^{\mathsf{T}}\boldsymbol{X}_{t-h})^{+}\boldsymbol{X}_{t-h}^{\mathsf{T}}\boldsymbol{Y}_t\boldsymbol{V}_r\boldsymbol{V}_r^{\mathsf{T}}$, where $\boldsymbol{V}_r$ is formed by the top r eigenvectors obtained from

$$\hat{\boldsymbol{R}} = \boldsymbol{\Gamma}^{1/2}\hat{\boldsymbol{\Sigma}}_{\boldsymbol{YX}}\hat{\boldsymbol{\Sigma}}_{\boldsymbol{XX}}^{-1}\hat{\boldsymbol{\Sigma}}_{\boldsymbol{XY}}\boldsymbol{\Gamma}^{1/2}. \tag{4.3.4}$$

where $\boldsymbol{\Gamma}$ is a positive-definite matrix of weights, and $\hat{\boldsymbol{\Sigma}}_{\boldsymbol{XX}}, \hat{\boldsymbol{\Sigma}}_{\boldsymbol{XY}} = \hat{\boldsymbol{\Sigma}}_{\boldsymbol{YX}}$, and $\hat{\boldsymbol{\Sigma}}_{\boldsymbol{YY}}$ are the ML sample covariance matrices calculated as $\hat{\boldsymbol{\Sigma}}_{\boldsymbol{XX}} = T^{-1}\boldsymbol{X}_{t-h}^{\mathsf{T}}\boldsymbol{X}_{t-h}$; $\hat{\boldsymbol{\Sigma}}_{\boldsymbol{XY}} = T^{-1}\boldsymbol{X}_{t-h}^{\mathsf{T}}\boldsymbol{Y}_t = \hat{\boldsymbol{\Sigma}}_{\boldsymbol{YX}}^{\mathsf{T}}$; and $\hat{\boldsymbol{\Sigma}}_{\boldsymbol{YY}} = T^{-1}\boldsymbol{Y}_t^{\mathsf{T}}\boldsymbol{Y}_t$, respectively. Optimization on $\boldsymbol{S}$ is then executed via $\Theta^{\#}(\cdot; \mathfrak{q}, \eta)$ to tackle the computational challenge for nonconvex P. Ridge parameter $\eta > 0$ provides control over the bias-variance tradeoff and guarantees selection stability by reaching a good balance in selection in the presence of collinearity.

On the other hand, when $\boldsymbol{S}$ is given $\boldsymbol{V}$ is estimated via

$$\hat{\boldsymbol{V}} = \min_{\boldsymbol{V}} ||\boldsymbol{Y}_t\boldsymbol{V} - \boldsymbol{X}_{t-h}\boldsymbol{S}||_F^2 \;\; s.t. \;\; \boldsymbol{V}^{\mathsf{T}}\boldsymbol{V} = \boldsymbol{I} \tag{4.3.5}$$

and then the optimal $\boldsymbol{V}$ can be obtained by procrustes rotation: first perform SVD on $\boldsymbol{Y}_t^{\mathsf{T}}\boldsymbol{X}_{t-h}\boldsymbol{S}$ getting $\boldsymbol{U}_w\boldsymbol{D}_w\boldsymbol{V}_w^{\mathsf{T}}$, where $\boldsymbol{D}_w \in \mathbb{R}^{r\times r}$, then $\hat{\boldsymbol{V}} = \boldsymbol{U}_w\boldsymbol{V}_w^{\mathsf{T}}$.

The computation is described in Algorithm 2. The algorithm is simple to implement and has low computational complexity. In addition the matrix multiplication and thresholding operations, Step (3) performs an SVD, but $\boldsymbol{W}$ has only r columns, and the rank values of practical interest are usually small.

Algorithm 1 Rank-Cardinality Constrained Regression - RCCR

given $1 \le r \le p$, $1 \le \mathfrak{q} \le p$, $\eta \ge 0$, $\boldsymbol{S}^{(0)} \in \mathbb{R}^{p\times r}$, M_{inner}: maximum number of inner iterations, M_{outer}: maximum number of outer iterations

(1) $j \leftarrow 0$, $K \leftarrow \|\boldsymbol{X}_{t-h}\|_2^2$

while $\|\boldsymbol{C}^{(t)} - \boldsymbol{C}^{(t-1)}\|$ (if existing) is not small enough and $j < M_{outer}$ **do**

(2) $j \leftarrow j+1$

(3) Let $\boldsymbol{W} \leftarrow \boldsymbol{Y}_t^{\mathsf{T}}\boldsymbol{X}_{t-h}\boldsymbol{S}^{(j)}$ and perform SVD: $\boldsymbol{W} = \boldsymbol{U}_w\boldsymbol{D}_w\boldsymbol{V}_w^{\mathsf{T}}$, where $\boldsymbol{D}_w \in \mathbb{R}^{r\times r}$

(4) $\boldsymbol{V}^{(j)} \leftarrow \boldsymbol{U}_w\boldsymbol{V}_w^{\mathsf{T}}$

(5) Perform the inner iterations:

(5.1) $j' \leftarrow 0$, $\tilde{\boldsymbol{S}}^{(0)} \leftarrow \boldsymbol{S}^{(j-1)}$.

while $\|\tilde{\boldsymbol{S}}^{(j')} - \tilde{\boldsymbol{S}}^{(j'-1)}\|$ (if existing) is not small enough and $j' > M_{inner}$ **do**

(5.2) $j' \leftarrow j'+1$

(5.3) $\tilde{\boldsymbol{S}}^{(j')} \leftarrow \vec{\Theta}^{\#}\left(\frac{1}{K}\boldsymbol{X}_{t-h}^{\mathsf{T}}\boldsymbol{Y}_t\boldsymbol{V}^{(j-1)} + (\boldsymbol{I} - \frac{1}{K}\boldsymbol{X}_{t-h}^{\mathsf{T}}\boldsymbol{X}_{t-h})\tilde{\boldsymbol{S}}^{(j'-1)}; \mathfrak{q}, \eta\right)$

end while

(5.4) $\boldsymbol{S}^{(j)} \leftarrow \tilde{\boldsymbol{S}}^{(j')}$.

(6) $\boldsymbol{C}^{(j)} \leftarrow \boldsymbol{S}^{(j)}(\boldsymbol{V}^{(j)})^{\mathsf{T}}$

end while

deliver $\widehat{\boldsymbol{C}} = \boldsymbol{C}^{(j)}$.

For generality, we prove a result for any $\boldsymbol{Y}_t \in \mathbb{R}^{T\times n}$ and any $K \ge \|\boldsymbol{X}_{t-h}\|_2^2$. Given $1 \le r \le p$, $1 \le \mathfrak{q} \le p$, $\eta \ge 0$, and an arbitrary starting point $\boldsymbol{S}^{(0)} \in \mathbb{R}^{p\times r}$, let $(\boldsymbol{S}^{(j)}, \boldsymbol{V}^{(j)}, \boldsymbol{C}^{(j)})$ $(j = 1, 2, \cdots)$ denote the sequence of iterates generated by Algorithm 2. Then, given $K \ge \|\boldsymbol{X}_{t-h}\|_2^2$ the original problem is as follows

$$(\hat{\boldsymbol{S}}, \hat{\boldsymbol{V}}) = \min_{\boldsymbol{S},\boldsymbol{V}} \frac{1}{2}||\boldsymbol{Y}_t\boldsymbol{V} - \boldsymbol{X}_{t-h}\boldsymbol{S}||_F^2 + \frac{1}{2}\|\boldsymbol{Y}(\boldsymbol{I} - \boldsymbol{V}\boldsymbol{V}^{\mathsf{T}})\|_F^2 + \frac{\eta}{2}||\boldsymbol{S}||_F^2 \;\; s.t. \;\; \|\boldsymbol{S}\|_{2,0} \le \mathfrak{q}, \boldsymbol{V}^{\mathsf{T}}\boldsymbol{V} = \boldsymbol{I} \tag{4.3.6}$$

where $\eta' = \eta K$. The objective function (4.3.6) is the Alternating-Optimization function which solves for $\boldsymbol{S}$ when $\boldsymbol{V}$ is given and vice versa. Therefore, we directly solve the scaling objective

function defined as follows

$$F(\boldsymbol{C}) = \frac{\|\boldsymbol{Y}_t - \boldsymbol{X}_{t-h}\boldsymbol{C}\|_F^2}{2K} + \frac{\eta\|\boldsymbol{C}\|_F^2}{2} = \frac{\|\boldsymbol{Y}_t - \boldsymbol{X}_{t-h}\boldsymbol{S}\boldsymbol{V}^\top\|_F^2}{2K} + \frac{\eta\|\boldsymbol{C}\|_F^2}{2} = F(\boldsymbol{S}, \boldsymbol{V}).$$

where parameter η can be used as an adaptive parameter. However, tuning η would increase the computational cost. Conveniently, the magnitude of η has been chosen to be fixed and focus the attention on main parameters $\mathfrak{q}$ and r. Choosing an appropriate value for $\mathfrak{q}$ allow the quantile thresholding rule to select the most meaningful predictors in response to the TCR. The size of $\mathfrak{q}$ can be defined based on some preference or as a function of T. For instance, $\mathfrak{q} = \zeta T$ where $0 < \zeta < 1$.

Theorem 2. *Suppose $K \geq \|\boldsymbol{X}_{t-h}\|_2^2$. Then, $F(\boldsymbol{C}^{(j)})$ is decreasing: $F(\boldsymbol{C}^{(j)}) - F(\boldsymbol{C}^{(j+1)}) \geq (1 - \frac{\|\boldsymbol{X}_{t-h}\|_2^2}{K})\|\boldsymbol{S}^{(j)} - \boldsymbol{S}^{(j+1)}\|_F^2$, and $\boldsymbol{C}^{(j)}$ obeys $rank(\boldsymbol{C}^{(j)}) \leq r$ and $\|\boldsymbol{C}^{(j)}\|_{2,0} \leq \mathfrak{q}$ for any $j \geq 1$.*

Theorem 2 guarantees the function-value decreasing property under rank-cardinality constraints. Additionally, theorem 2 ensures local convergence to the optimal minimum of F. Therefore, Algorithm 2 conducts rank constrained variable screening by solving a constrained optimization problem:

$$\min_{\boldsymbol{C}} \frac{1}{2K}\|\boldsymbol{Y}_t - \boldsymbol{X}_{t-h}\boldsymbol{C}\|_F^2 + \frac{\eta}{2}\|\boldsymbol{C}\|_F^2 \text{ s.t. } \|\boldsymbol{C}\|_{2,0} \leq \mathfrak{q}, rank(\boldsymbol{C}) \leq r. \tag{4.3.7}$$

A finite-sample oracle inequality shows the globally optimal solutions of (4.3.7). In the theorem below we assume

$$\boldsymbol{Y}_t = \boldsymbol{X}_{t-h}\boldsymbol{C}^* + \boldsymbol{E}_t, \tag{4.3.8}$$

where $\boldsymbol{Y}_t \in \mathbb{R}^{T\times n}, \boldsymbol{X}_{t-h} \in \mathbb{R}^{T\times p}$, and $\text{vec}\,(\boldsymbol{E}_t)$ is *sub-Gaussian* with mean zero and scale bounded by σ; see Definition D.0.1 in Section D. Sub-Gaussian examples include Gaussian random variables and bounded random variables such as Bernoulli. Note that the entries of $\boldsymbol{E}$ may *not* be iid. We use $\lesssim$ to denote an inequality that holds up to a multiplicative numerical constant. Let $q := \text{rank}(\boldsymbol{X}_{t-h})$.

Theorem 3. *Suppose the model (4.3.8) with the sub-Gaussian noise contamination holds. Let $\hat{\boldsymbol{C}} \in \arg\min_{\boldsymbol{C}:rank(\boldsymbol{C})\leq r, \|\boldsymbol{C}\|_{2,0}\leq\mathfrak{q}} \frac{1}{2}\|\boldsymbol{Y}_t - \boldsymbol{X}_{t-h}\boldsymbol{C}\|_F^2 + \frac{\eta}{2}\|\boldsymbol{C}\|_F^2$. Then, the following oracle inequality holds for any $\boldsymbol{C} \in \mathbb{R}^{p\times n}$ with $rank(\boldsymbol{C}) \leq r$ and $\|\boldsymbol{C}\|_{2,0} \leq \mathfrak{q}$:*

$$\begin{aligned}&\mathbb{E}\left[\|\boldsymbol{X}_{t-h}\hat{\boldsymbol{C}} - \boldsymbol{X}_{t-h}\boldsymbol{C}^*\|_F^2\right] \\ \lesssim &\|\boldsymbol{X}_{t-h}\boldsymbol{C} - \boldsymbol{X}_{t-h}\boldsymbol{C}^*\|_F^2 + \eta\|\boldsymbol{C}\|_F^2 + \sigma^2\{(q \wedge \mathfrak{q})r + nr + \mathfrak{q}\log(ep/\mathfrak{q})\} + \sigma^2.\end{aligned} \tag{4.3.9}$$

Nonasymptotic oracle inequalities an be proved for the global solutions of the selectable reduced-rank regression. This nonasymptotic result does not require any regularity condition (such as low coherence) on the design matrix. Prediction error rate can be derived from this oracle inequality. For example, when $\eta = 0$, $\mathfrak{q} = \mathfrak{q}^*$ and $r = r^*$, where r is the true r, the risk is bounded by σ^2 times $\{(q \wedge \mathfrak{q}^*)r^* + nr^* + \mathfrak{q}^* \log(ep/\mathfrak{q}^*)\}$ which is low when r^* and $\mathfrak{q}^*$ are small. This is very attractive in high-dimensional multivariate statistical applications. The oracle inequality provides finer error control: the existence of the bias term $\|\boldsymbol{X}_{t-h}\boldsymbol{C} - \boldsymbol{X}_{t-h}\boldsymbol{C}^*\|_F^2$ can handle approximately sparse and/or approximately low rank signals; the ridge term offers further bias-variance trade-off in choosing a reference signal $\boldsymbol{C}$.

In the asymptotic arena results require some assumptions. The classical assumption setup assumes finite parameters p, $\mathfrak{q}$, and r are fixed, usually to a small number, while $T \to \infty$. Under these conditions the average of the risk function $\|\boldsymbol{X}_{t-h}\boldsymbol{C} - \boldsymbol{X}_{t-h}\boldsymbol{C}^*\|_F^2 \to 0$ due to factor $\frac{1}{pT}$. Likewise, the average of $\eta\|\boldsymbol{C}\|_F^2$ also goes to zero for typical values of η. Consequently, term $\sigma^2\{(q \wedge \mathfrak{q})r + nr + \mathfrak{q}\log(ep/\mathfrak{q})\}$ will bound the prediction error. In particular, for small values of $q(\ll T)$, $\mathfrak{q}$ (typically a few) and r (in order of 1 or 2) the oracle inequality is bounded to a very small error.

These types of problems are of great interest in (ultra)high-dimensional applications because under the sparsity assumption it is not difficult for one to specify an upper bound $\mathfrak{q}$ for the number of relevant features $\|\boldsymbol{C}^*\|_{2,0}$. Usually, $\mathfrak{q}$ satisfies $\mathfrak{q} < n$, and so after the screening process, one faces a low-dimensional problem. Running Algorithm 2 on the screened dataset significantly reduces the computation burden.

Algorithm 2 is closely connected with the SIS, see, Fan and Lv [17]. Given a univariate response (denoted by $\boldsymbol{y}_t$) model, with $\boldsymbol{S}^{(0)} = \boldsymbol{0}$, the first iteration ranks all features based on $\boldsymbol{X}_{t-h}^{\mathsf{T}}\boldsymbol{y}_t$, which amounts to the SIS. Of course, this screening is purely based on marginal statistics, and may be improper in the presence of many (correlated) predictors. Algorithm 2 iterates to lessen such greediness. The iterative quantile screening technique applies to the sole variable selection, group variable selection, or rank reduction.

We adopt an '*annealing + squeezing*' idea to further reduce computational load for big data. Let $\mathfrak{q}$ be an upper bound of the number of relevant x variables and r be an upper bound of the number of necessary underlying factors. Define a cooling schedule $Q(\mathfrak{t})$ $(1 \leq \mathfrak{t} \leq \mathfrak{T})$ with $Q(1) = p$ and

$Q(\mathfrak{T}) = \mathfrak{q}$. We propose a **progressive** quantile thresholding-based iterative screening procedure as follows.

1. Initialization: $\boldsymbol{d} \leftarrow [1\ 2\ \cdots\ p]$.
2. Given each $\mathfrak{t}$, run Steps (1)-(6) of Algorithm 2 with $\mathfrak{q}$ replaced by $Q(\mathfrak{t})$, and add 'squeezing' operations afterwards:

$$(7)\ \tilde{\boldsymbol{d}} \leftarrow \{j : \boldsymbol{g}(\boldsymbol{S}^{(t)})[j] \neq 0\}, \boldsymbol{d} \leftarrow \boldsymbol{d}[\tilde{\boldsymbol{d}}], \boldsymbol{S}^{(j)} \leftarrow \boldsymbol{S}^{(j)}[\tilde{\boldsymbol{d}},], \boldsymbol{X} \leftarrow \boldsymbol{X}[, \tilde{\boldsymbol{d}}].$$

3. Repeat this for $\mathfrak{t} = 1, \cdots, \mathfrak{T}$, and deliver the final remaining dimensions indexed by $\boldsymbol{d}$.

Algorithm 2 Progressive Rank-Cardinality Constrained Regression - PRCCR

given $1 \leq r \leq p$, $1 \leq \mathfrak{q} \leq p$, $\eta \geq 0$, $\boldsymbol{S}^{(0)} \in \mathbb{R}^{p \times r}$, $Q(\mathfrak{t})$ for $1 \leq \mathfrak{t} \leq \mathfrak{T}$, M_{inner}: maximum number of inner iterations, M_{outer}: maximum number of outer iterations

(1) $j \leftarrow 0$, $K \leftarrow \|\boldsymbol{X}_{t-h}\|_2^2$

while $\|\boldsymbol{C}^{(t)} - \boldsymbol{C}^{(t-1)}\|$ (if existing) is not small enough and $j < M_{outer}$ **do**

(2) $j \leftarrow j + 1$

(3) Let $\boldsymbol{W} \leftarrow \boldsymbol{Y}_t^{\mathsf{T}} \boldsymbol{X}_{t-h} \boldsymbol{S}^{(j)}$ and perform SVD: $\boldsymbol{W} = \boldsymbol{U}_w \boldsymbol{D}_w \boldsymbol{V}_w^{\mathsf{T}}$, where $\boldsymbol{D}_w \in \mathbb{R}^{r \times r}$

(4) $\boldsymbol{V}^{(j)} \leftarrow \boldsymbol{U}_w \boldsymbol{V}_w^{\mathsf{T}}$

(5) Perform the inner iterations:

(5.1) $j' \leftarrow 0$, $\tilde{\boldsymbol{S}}^{(0)} \leftarrow \boldsymbol{S}^{(j-1)}$.

while $\|\tilde{\boldsymbol{S}}^{(j')} - \tilde{\boldsymbol{S}}^{(j'-1)}\|$ (if existing) is not small enough and $j' > M_{inner}$ **do**

(5.2) $j' \leftarrow j' + 1$

(5.3) $\tilde{\boldsymbol{S}}^{(j')} \leftarrow \vec{\Theta}^{\#}\left(\frac{1}{K}\boldsymbol{X}_{t-h}^{\mathsf{T}}\boldsymbol{Y}_t\boldsymbol{V}^{(j-1)} + (\boldsymbol{I} - \frac{1}{K}\boldsymbol{X}_{t-h}^{\mathsf{T}}\boldsymbol{X}_{t-h})\tilde{\boldsymbol{S}}^{(j'-1)}; \mathfrak{q}, \eta\right)$

end while

(5.4) $\boldsymbol{S}^{(j)} \leftarrow \tilde{\boldsymbol{S}}^{(j')}$.

(6) $\boldsymbol{C}^{(j)} \leftarrow \boldsymbol{S}^{(j)}(\boldsymbol{V}^{(j)})^{\mathsf{T}}$

(7) $\tilde{\boldsymbol{d}} \leftarrow \{j : \boldsymbol{g}(\boldsymbol{S}^{(j)})[j] \neq 0, \boldsymbol{d} \leftarrow \boldsymbol{d}[\tilde{\boldsymbol{d}}], \boldsymbol{S}^{(j)} \leftarrow \boldsymbol{S}^{(j)}[\tilde{\boldsymbol{d}},], \boldsymbol{X}_{t-h} \leftarrow \boldsymbol{X}_{t-h}[, \tilde{\boldsymbol{d}}]$

end while

deliver $\widehat{C} = \boldsymbol{C}^{(j)}$.

Because of the squeezing operations, the sizes of $\boldsymbol{W}$ and $\boldsymbol{S}$ keep dropping as j increases. This is particularly helpful for very high dimensional data. Empirically, we set $M_{outer} = M_{inner} = 1$, and find the sigmoidal decay cooling schedule $Q(\mathfrak{t}) = [2p/(1 + \exp(\alpha\mathfrak{t}))]$ with $\alpha = 0.01$ achieve good balance between selection and efficiency. The Sigmoid function possesses the ability to slow values around p to rapidly decrease to q and finally keep values around $\mathfrak{q}$ to optimize $\boldsymbol{C}$.

Algorithm (2) is very flexible in the sense of estimation. For any given $\boldsymbol{S}$ the algorithm estimates $\boldsymbol{V}$ and vice versa. It works in large-p applications where the number of features grows faster than the number of observations. Indeed, since the number of observations grows only through time, the number of variables is expected to increase to a higher rate. The algorithm tackles large values of p via its *progressive variable screening* attribute.

Algorithm (2) is designed to minimize over all $p \times n$ matrices $\boldsymbol{C}^{(j)}$ with rank r. This is the same as taking the minimum over all $p \times r$ matrices $\boldsymbol{S}^{(j)}$ and all $n \times r$ matrices $\boldsymbol{V}^{\mathsf{T}(j)}$. So the optimization problem in $\boldsymbol{C}^{(j)}$ becomes an optimization problem in $\boldsymbol{S}^{(j)}$ and $\boldsymbol{V}^{\mathsf{T}(j)}$. Since the main goal is to reduce the Euclidean norm between matrices we must look over $tr[(\boldsymbol{Y}_t - \boldsymbol{X}_{t-h}\boldsymbol{C})(\boldsymbol{Y}_t - \boldsymbol{X}_{t-h}\boldsymbol{C})]$ where tr stands for trace. Solving the distance between matrices we obtain $tr[(\boldsymbol{Y}_t - \boldsymbol{X}_{t-h}\boldsymbol{C})(\boldsymbol{Y}_t - \boldsymbol{X}_{t-h}\boldsymbol{C})] = tr(\boldsymbol{Y}_t\boldsymbol{Y}_t^{\mathsf{T}}) - tr(\boldsymbol{Y}_t^{\mathsf{T}}\boldsymbol{X}_{t-h}\boldsymbol{S}\boldsymbol{V}^{\mathsf{T}})$ with $\boldsymbol{C} = \boldsymbol{S}\boldsymbol{V}^{\mathsf{T}}$. Minimizing this solution is equivalent to maximizing the second term $tr(\boldsymbol{Y}_t^{\mathsf{T}}\boldsymbol{X}_{t-h}\boldsymbol{S}\boldsymbol{V}^{\mathsf{T}})$. Then, for $\boldsymbol{W} \leftarrow \boldsymbol{Y}_t^{\mathsf{T}}\boldsymbol{X}_{t-h}\boldsymbol{S}$ we need to find $\mathbf{max}[tr(\boldsymbol{W}^{\mathsf{T}}\boldsymbol{V})] \in \mathbb{R}^{n \times r}$. To obtain a global maximum for $tr(\boldsymbol{W}^{\mathsf{T}}\boldsymbol{V})$ it suffices to find an upper bound in $\boldsymbol{W}$ and $\boldsymbol{S}$. By von Neumann's trace inequality we know that $tr(\boldsymbol{W}^{\mathsf{T}}\boldsymbol{V}) \leq \sum_i d_i(\boldsymbol{W})$ where d_i represents the singular values from $\boldsymbol{W}$. Then, we are required to estimate the SVD of $\boldsymbol{W}$ by setting $\boldsymbol{W} = \boldsymbol{U}_w\boldsymbol{D}_w\boldsymbol{V}_{\boldsymbol{w}}^{\mathsf{T}}$ and $\boldsymbol{D}_w \in \mathbb{R}^{r \times r}$. It follows that $\boldsymbol{V}^{(j)} = \boldsymbol{U}_w\boldsymbol{V}_w^{\mathsf{T}}$ achieves the upper bound $\Sigma_i d_i(\boldsymbol{W})$.

The PRCCR inherently reduces the computational effort associated in estimating $\hat{\boldsymbol{C}}$. Optimization of $\hat{\boldsymbol{C}}$ is performed within the outer iteration while the optimization of $\tilde{\boldsymbol{S}}^{(j')}$ is performed in the inner iteration. Therefore, one run may suffice for the algorithm to achieve convergence in both matrices since when $\tilde{\boldsymbol{S}}^{(j')}$ is optimized $\hat{\boldsymbol{C}}$ is optimized as well.

4.3.1 Weight Matrix - $\boldsymbol{\Gamma}$

Macroeconomic time series grouped in the same category share common information regarding a particular component of the economy. Commonality among series means that the response variables are not independent. Otherwise, each response variable will account for a particular problem. Using this common information can improve the forecasting accuracy of a target series. The most generalized form incorporates this knowledge through a weighted matrix. Reduced-rank regression (RRR) uses $\boldsymbol{\Gamma}$ as the weight matrix, see, (2.3.5). For a positive definite weight matrix $\boldsymbol{\Gamma}$, the estimator of the coefficient matrix $\boldsymbol{C}$ of given rank r may be found minimizing the weighted

contained problem

$$\min_{\boldsymbol{C}} ||(\boldsymbol{Y}_t - \boldsymbol{X}_{t-h}\boldsymbol{C})\boldsymbol{\Gamma}^{1/2}||_F^2 \; s.t. \; r(\boldsymbol{C}) \leq r.$$

where $\boldsymbol{C}$ can be expressed as the product of two matrices of smaller rank. Then, coefficient matrix $\boldsymbol{C}$ can be written as $\boldsymbol{C} = \boldsymbol{C}_1\boldsymbol{C}_2$.

The choice of the weight matrix is very important in RRR. The SFAR uses the choice of weight $\boldsymbol{\Gamma} = \boldsymbol{I}$ but other alternatives can be implemented. Two very popular choices of the weight matrix have been studied in the literature, see, e.g. Reinsel and Velu [46]. These choices set: $\boldsymbol{\Gamma} = \boldsymbol{\Sigma}_E^{-1}$ and $\boldsymbol{\Gamma} = \hat{\boldsymbol{\Sigma}}_{\boldsymbol{YY}}^{-1}$. However, the former has two drawbacks. If $\boldsymbol{\Sigma}_E^{-1}$ is not *i.i.d.*, then a new criterion needs to be developed with a weighting matrix that decorrelates the correlated model. On the other hand, $\boldsymbol{\Sigma}_E^{-1}$ can not be estimated directly making it an inconvenient option. Consequently, $\boldsymbol{\Gamma} = \hat{\boldsymbol{\Sigma}}_{\boldsymbol{YY}}^{-1}$ rises as the convenient option. Even more, as stated in section 2.3, the solution for $\boldsymbol{C}_1$ and $\boldsymbol{C}_2$ simultaneously minimizes the eigenvalues of the weight matrix providing that $\boldsymbol{\Gamma} = \hat{\boldsymbol{\Sigma}}_{\boldsymbol{YY}}^{-1}$. This particular setup matrix assigns the ML estimates for $\boldsymbol{C}$ in the large sample setup. Through the population version established by Reinsel and Velu [46], the relationship to $\boldsymbol{\Gamma} = \hat{\boldsymbol{\Sigma}}_{\boldsymbol{YY}}^{-1}$ is direct, provided that $\hat{\boldsymbol{\Sigma}}_{\boldsymbol{YY}}$ is positive definite. However, in high dimensions details are absent. It is noted that when $p > T$, $\boldsymbol{\Sigma}_{\boldsymbol{YY}}$ is singular and not invertible. An alternative is to use the Moore-Penrose generalized inverse. However, estimates from this will show to be unstable as finding them is reduced to taking the inverse of the singular values. Therefore, if the singular values are small, they will approach ∞. More recently, estimators with regularization parameters have been utilized to bypass this issue. That is, the sample residual covariance could be regularized so that $\boldsymbol{\Gamma} = (\hat{\boldsymbol{\Sigma}}_{\boldsymbol{YY}} + \delta\boldsymbol{I})^{-1}$ where δ is the regularization parameter.

Consequently two variations of $\boldsymbol{\Gamma}$ are tested in this study:

1. $\boldsymbol{\Gamma} = (\hat{\boldsymbol{\Sigma}}_{\boldsymbol{YY}} + \delta)^{-1}$
2. $\boldsymbol{\Gamma} = (\boldsymbol{D}_{(\hat{\boldsymbol{\Sigma}}_{\boldsymbol{YY}})} + \delta)^{-1}$

where $\boldsymbol{D}_{(\hat{\boldsymbol{\Sigma}}_{\boldsymbol{YY}})}$ stands for the diagonal of the sample covariance matrix of $\boldsymbol{Y}_t$. Notice that values of delta can differ from one category and/or combination of parameters $(r, \mathfrak{q})$ to another.

CHAPTER 5

COMPARATIVE STUDY OF THE SFAR

Several methodologies exist to forecast a univariate time series. In this chapter, forecasting results for synthetic and empirical data are exhibited in a comparative study of the SFAR versus some up-to-date forecasting methodologies. Simulated data are generated as indicated in subsection (5.2), while empirical data are the same as in Stock and Watson in [62]. To forecast series multiple steps ahead the direct forecast is favored and implemented through the rolling window forecasting scheme.

To undertake a comparison between different forecasting methods and contrast their performance on both types of data, we use Ridge regression (Ridge), AutoRegression model of order 4 (AR(4)), Dynamic Factor Models (DFM), and Sparse Factor Auto-Regression (SFAR) methodology. As encountered in the macroeconomic literature, the AR(4) is used as the benchmark to measure the performance of all other methods. Two groups of forecasting methods can be identified. One group includes Ridge and AR(4) using lagged observations as predictors. The second group encompasses DFM5 and SFAR methods which use a combination of lagged observations and orthogonal factors. Stock and Watson [60] provide extensive use of many forecasting methods in a comparative study for empirical data. This chapter starts describing the implemented forecasting methodologies including univariate and multivariate forms.

5.1 Forecasting methodologies description

- Ridge. The OLS regression methodology uses the least square method to estimate regression coefficients. However, a non-singular solution is found when this method is applied in high dimensions. To overcome non-singularity different shrinkage methods can be applied. In this work we regularize the coefficient matrix via ridge regression. Thus, to perform the forecasting of a target series of interest $\boldsymbol{y}_t$, it follows:

  1. Estimate ridge coefficients via $\hat{\boldsymbol{c}}_t^{(ridge)} = (\boldsymbol{X}_{t-h}^{\mathsf{T}}\boldsymbol{X}_{t-h} + \lambda\boldsymbol{I})^{-1}\boldsymbol{X}_{t-h}^{\mathsf{T}}\boldsymbol{y}_t$ where λ is the ridge parameter, and $\boldsymbol{I}$ is an identity matrix.

2. Compute the forecasting as $\hat{y}_{t+h} = \boldsymbol{x}_t \hat{\boldsymbol{c}}_t^{(ridge)}$ where $\boldsymbol{x}_t$ contains the observation of the predictors at time t.

- AR(4). The autoregression model for forecasting, in particular the AR(4), uses the four most recent observations as predictors. To estimate the regression coefficients we favored the equally weighted OLS regression with neglected intercept. Given a target series $\boldsymbol{y}_t$, the information up to time t is used to train coefficients, and estimate $\hat{y}_{t+h}$ as follows.

  1. Estimate the 4-dimensional coefficient vector $\boldsymbol{\phi}_t = [\phi_1, \phi_2, \phi_3, \phi_4]$ by regressing $\boldsymbol{y}_t \sim \boldsymbol{X}_{t-h}$, where $\boldsymbol{X}_{t-h} = [\boldsymbol{y}_{t-h-1} \;\; \boldsymbol{y}_{t-h-2} \;\; \boldsymbol{y}_{t-h-3} \;\; \boldsymbol{y}_{t-h-4}]$.
  2. Perform the forecasting via $\hat{y}_{t+h} = \sum_{j=0}^{3} y_{t-j}\phi_{j+1}$

- DFM. Dynamic Factor Models is used as stated in Stock and Watson [62]. DFM are built via OLS estimation since it offers the best forecasting performance versus GLM and WLS see, Stock and Watson [60]. OLS is also computationally convenient. In particular, the DFM-5 uses the first five principal components (PCs) ordered according to the magnitude of the largest eigenvalues with which they are associated to construct factors; the remaining PCs are omitted. Then, the predictor matrix is constructed and the forecasting coefficients estimated via OLS without shrinkage with neglected intercept. For estimation, the DFM-5 methodology works as follows.

  1. Estimate the factors via $\hat{\boldsymbol{F}}_{t-h}^{(DFM5)} = \boldsymbol{X}_{t-h}\hat{\boldsymbol{P}}_{t-h}$ where $\hat{\boldsymbol{P}}_{t-h}$ is the $T \times 5$ matrix containing the PCs. Since matrix $\hat{\boldsymbol{P}}_{t-h}$ has low rank approximation the spectral decomposition is preferred. Given a square matrix, we can write

  $$\begin{aligned} \boldsymbol{X}_{t-h}^{\mathsf{T}}\boldsymbol{X}_{t-h} &= \boldsymbol{P}\boldsymbol{D}\boldsymbol{P}^{\mathsf{T}} \\ \boldsymbol{X}_{t-h} &= \boldsymbol{X}_{t-h}\boldsymbol{P}\boldsymbol{P}^{\mathsf{T}} \\ \boldsymbol{X}_{t-h} &= \boldsymbol{F}_{t-h}^{(DFM5)}\boldsymbol{P}^{\mathsf{T}} \end{aligned}$$

  where $\boldsymbol{P}\boldsymbol{P}^{\mathsf{T}} = \boldsymbol{I}$ with $\boldsymbol{P} \in \mathbb{R}^{T\times 5}$, $\boldsymbol{D}$ is diagonal containing the eigenvalues, and $\boldsymbol{F}_{t-h} = \boldsymbol{X}_{t-h}\boldsymbol{P}$. Be aware that for factor construction $\boldsymbol{X}_{t-h}$ contains lags of information of the series.

  2. Estimate the coefficient vector $\hat{\boldsymbol{c}}_t^{(DFM)} = (\hat{\boldsymbol{A}}_t^{\mathsf{T}}\hat{\boldsymbol{A}}_t)^{-1}\hat{\boldsymbol{A}}_t^{\mathsf{T}}\boldsymbol{y}_t$, where $\boldsymbol{y}_t$ is the target series and the predictor matrix is

  $$\hat{\boldsymbol{A}} = [\boldsymbol{y}_{t-h-1} \;\; \boldsymbol{y}_{t-h-2} \;\; \boldsymbol{y}_{t-h-3} \;\; \boldsymbol{y}_{t-h-4} \;\; \hat{\boldsymbol{F}}_{t-h}^{(DFM5)}].$$

  3. The forecasting is estimated as $\hat{y}_{t+h} = \boldsymbol{a}_t\hat{\boldsymbol{c}}_t^{(DFM5)}$, where $\boldsymbol{a}_t$ is a 9-dimensional vector containing factors at time t and the most recent four observations of $\boldsymbol{y}_{t+h}$.

- SFAR. The sparse factor autoregression model is implemented as described in section (3.2). SFAR performs all the steps below for each target category response as follows. Given a matrix of coefficients $\boldsymbol{C} \in \mathbb{R}^{p\times n}$

1. Decompose $\boldsymbol{C}$ as the product of two matrices of lower rank via $\boldsymbol{C} = \boldsymbol{S}\boldsymbol{V}_r^{\mathsf{T}}$ with $\boldsymbol{V}_r \in \mathbb{O}^{n\times r}$ being orthogonal, and construct $\hat{\boldsymbol{S}}^{(0)} = (\boldsymbol{X}_{t-h}^{\mathsf{T}}\boldsymbol{X}_{t-h})^{+}\boldsymbol{X}_{t-h}^{\mathsf{T}}\boldsymbol{Y}_t\boldsymbol{V}_r$ where $\hat{\boldsymbol{S}}^{(0)} \in \mathbb{R}^{p\times r}$.
2. Perform the PRCCR algorithm $\mathcal{T}$ times as indicated by the cooling schedule plan to estimate the doubly-constrained coefficient matrix $\hat{\boldsymbol{C}}(\mathfrak{q}, r)$.
3. Extract C-level factors via $\boldsymbol{F}_{t-h} = \boldsymbol{X}_{t-h}(\hat{\boldsymbol{C}}(\mathfrak{q}, r)\boldsymbol{U}_r)$ where $\boldsymbol{U}_r$ is an orthogonal matrix satisfying $\boldsymbol{U}^{\mathsf{T}}\boldsymbol{U} = \boldsymbol{I}$ extracted from the spectral decomposition
$$\hat{\boldsymbol{C}}^{\mathsf{T}}(\mathfrak{q}, r)\boldsymbol{X}_{t-h}^{\mathsf{T}}\boldsymbol{X}_{t-h}\hat{\boldsymbol{C}}(\mathfrak{q}, r) = \boldsymbol{U}\boldsymbol{D}\boldsymbol{U}^{\mathsf{T}}.$$
4. Construct the general design matrix $\boldsymbol{Z} = [\boldsymbol{y}_{t-h-1}\ \boldsymbol{y}_{t-h-2}\ \boldsymbol{y}_{t-h-3}\ \boldsymbol{y}_{t-h-4}]$, and create a new predictor matrix $\boldsymbol{G}_{t-h} = [\boldsymbol{Z}_{t-h}\ \ \boldsymbol{X}_{t-h}]$
5. Estimate the regression coefficient matrix using ridge regression as follows: $\hat{\boldsymbol{C}}_t^{(ridge)} = (\boldsymbol{G}_{t-h}^{\mathsf{T}}\boldsymbol{G}_{t-h} + \lambda\boldsymbol{I})^{-1}\boldsymbol{G}_{t-h}^{\mathsf{T}}\boldsymbol{Y}_t$ where λ is selected from a grid of values. By shrinking the estimates the bias-variance tradeoff is introduced, and consequently an improvement in forecasting accuracy of the target series is expected.
6. Forecast the target series via $\hat{y}_{it+h} = \boldsymbol{g}_t\hat{\boldsymbol{c}}_{it}^{(ridge)}$ where $\boldsymbol{g}_t$ is a $(m+r)$-dimensional vector in $\boldsymbol{G}_t$ containing the most recent m observations of the target series in $\boldsymbol{z}_t$ and estimated C-level factors $\boldsymbol{f}_{t-h}$, and $\hat{\boldsymbol{c}}_{it}^{(ridge)}$ for $i = 1, \ldots, n$ is a T-dimensional vector in $\hat{\boldsymbol{C}}_t^{(ridge)}$.

Conveniently, parameters in the SFAR can be set or tuned. With numerous parameters involved there are many possible setting combinations and therefore it seems suitable to fix them. Parameters of SFAR are set as follows: the ridge regularization parameter is fixed at $\eta = 0.1$ to introduce shrinkage, and gain efficiency. Lag order parameters of Z_{t-h} and X_{t-h} are respectively set to $l = 4$ and $m = (1, 2, 4)$. Using different values of m allows to contrast the predictive power of the C-level factors when constructed using different lag orders. The cooling schedule selected to achieve the desired cardinality $\mathfrak{q}$ follows a Sigmoid function with decay parameter 0.1. Main estimation parameters r and $\mathfrak{q}$ are also set. For the C-level factors constructed r, we prefer to keep it within the range of $r = (1, 2, 3)$. The cardinality control is also bounded to $\mathfrak{q} = (2, 3, 4, 5)$ since parsimony is critical to improve the prediction performance of the fitted model. All parameters are used to forecast target series at $h = (1, 2, 4)$.

5.1.1 Rolling window scheme

Evaluation of the forecasting methods described above is accomplished through the so-called rolling MSE, a conventional measure in econometrics, see e.g. Stock and Watson [62] and He, She

and Wu [24]. Suppose $\tilde{T}$ observations $x_1, \ldots, x_{\tilde{T}}$ are available. Let the rolling window size be T ($T < \tilde{T}$) and h the horizon, where T stands for the sample size for training, and $\tilde{T}$ represents the full sample. Standing at time t, we use the most recent ($T = 100$) observations to estimate $\boldsymbol{F}_{t-h}$. Then, $\boldsymbol{Z}_{t-h}$ and $\boldsymbol{F}_{t-h}$ are used to forecast y_{t+h}, denoting the forecast as $\hat{y}_{t+h}$, and the forecasting error as $e_t^h = \|y_{t+h} - \hat{y}_{t+h}\|_2^2$. This process is repeated for $t = T, \ldots, T + N - 1$ as we shift the window. N is the number of window shifting that satisfies $1 < N < \tilde{T} - T - h - 1$. Then the rolling MSE for horizon h is defined as $MSE^h_{Rolling} = \frac{1}{N}\sum_{t=T}^{T+N-1} e_t^h$. When using $\hat{\boldsymbol{Z}}_{t-h}$ and $\hat{\boldsymbol{F}}_{t-h}$ to forecast $\hat{y}_{t+h}$, we do pseudo out-of-sample forecasting, that is, we assume observations after t are not available and consequently we need to do h-step ahead forecast.

5.2 Synthetic data analysis

A finite sample data is generated to corroborate the theoretical findings. In this section emphasis is made on the C-level factor extraction and cardinality control. This section starts with a detailed description of the simulated data process generation followed by the results of the comparative study.

5.2.1 Data generation

Synthetic data are generated as follows,

$$\boldsymbol{Y}_t = \boldsymbol{X}_{t-h}\tilde{\boldsymbol{C}} + \boldsymbol{E}_t \tag{5.2.1}$$

where $\boldsymbol{X}_{t-h}$ is a $T \times p$ matrix of predictors that mimics a vector autoregressive (VAR) model, $\tilde{\boldsymbol{C}}$ is a $p \times p$ matrix of coefficients, and $\boldsymbol{E}_t$ is the error component. Each component in equation (5.2.1) is generated as follows.

1. The coefficient matrix $\tilde{\boldsymbol{C}} \in \mathbb{R}^{p\times p}$ has the form

$$\tilde{\boldsymbol{C}} = \begin{bmatrix} \boldsymbol{C} \\ 0 \end{bmatrix} = \begin{bmatrix} \phi \boldsymbol{C}_1 \boldsymbol{C}_2 \\ 0 \end{bmatrix} \tag{5.2.2}$$

   where $|\phi| < 1$ is the stationary parameter, $\boldsymbol{C}_1$ is a $\mathfrak{q} \times r$ matrix and $\boldsymbol{C}_2$ is a $r \times p$ matrix. All entries in $\boldsymbol{C}_1$ and $\boldsymbol{C}_2$ are i.i.d. $N(0, 1)$. In this way, $\tilde{\boldsymbol{C}}$ contains r C-level factors and $\mathfrak{q}$ nonzero rows.

2. The noise component is represented by matrix $\tilde{\boldsymbol{E}}_t = [\tilde{\boldsymbol{e}}_{\tilde{t},1} \ldots \boldsymbol{e}_{\tilde{t},\tilde{p}}] \in \mathbb{R}^{\tilde{T}-1\times\tilde{p}}$, and generated from a multivariate normal distribution $MVN(0, \Sigma)$ with $\Sigma = \sigma^2\boldsymbol{I}$.

3. The predictor matrix $\boldsymbol{X}_{t-h}$ obeys an autoregression (AR) model structure with finite lag order m of the form $[\boldsymbol{X}_{t-h-1}\,\boldsymbol{X}_{t-h-2}\,\ldots\,\boldsymbol{X}_{t-h-m}]$. The AR process in the design matrix is essential to make sure that each observation of the series is generated from its predecessor plus some random shock, and are indexed by time t. It is also critical because indexed observations in $\boldsymbol{Y}_t$ must be h periods apart from those in $\boldsymbol{X}_{t-h}$ as needed for training purposes. To that end, matrix $\boldsymbol{X}_t = [\boldsymbol{x}_{\tilde{t},1}\,\ldots\,\boldsymbol{x}_{\tilde{t},\tilde{p}}]$ for $\tilde{t} = 1,\ldots,\tilde{T}$ which contains the burn-in time period in $\boldsymbol{X}_t$, is iteratively generated as follows

$$\begin{aligned}
\boldsymbol{x}_{\tilde{t}-\tilde{T}+2,.} &= \boldsymbol{x}_{\tilde{t}-\tilde{T}+1,.}\tilde{\boldsymbol{C}} + \boldsymbol{e}_{\tilde{t}-\tilde{T}+2,.} \\
\boldsymbol{x}_{\tilde{t}-\tilde{T}+3,.} &= \boldsymbol{x}_{\tilde{t}-\tilde{T}+2,.}\tilde{\boldsymbol{C}} + \boldsymbol{e}_{\tilde{t}-\tilde{T}+3,.} \\
&\vdots \\
\boldsymbol{x}_{\tilde{t},.} &= \boldsymbol{x}_{\tilde{t}-1,.}\tilde{\boldsymbol{C}} + \boldsymbol{e}_{\tilde{t},.}
\end{aligned} \tag{5.2.3}$$

where $\boldsymbol{x}_{\tilde{t}-\tilde{T}+1,.}$ is a p-dimensional random vector $\sim N(\mu,\sigma^2)$. Notation $\boldsymbol{x}_{\tilde{t}-\tilde{T}+1,.}$ denotes the observations of $\tilde{p}$ raw time series predictors at time $\tilde{t}-\tilde{T}+1$. Then, the burn-in period is removed from $\boldsymbol{X}_t$, and matrix $\boldsymbol{X}_t \in \mathbb{R}^{\tilde{T}\times\tilde{p}}$ is created. Matrix $\boldsymbol{X}_t$ is a high dimensional time series data. Then, for a value of h, $T \leq \tilde{T}-h-1$, and lag order m the predictor matrix $\boldsymbol{X}_{t-h} \in \mathbb{R}^{T\times p}$ is constructed with $p = \tilde{p}m$.

4. Response matrix $\boldsymbol{Y}_t \in \mathbb{R}^{T\times n}$ is extracted from $\boldsymbol{X}_t$ using a subset J for $J < \tilde{p}$ of indexed predictors where $\sum_{i=1}^{k} J_i = \tilde{p}$ and k stands for the maximum number of response categories.

In general, the data generating system described above is governed by the set of relevant predictors aligned with $\mathfrak{q}$ nonzero rows in $\tilde{\boldsymbol{C}}$ to extract r C-level factors. Then for each category response, C-level factors are constructed using a subset $\mathfrak{q}$ of relevant predictors from those in the predictor pool. Consequently, at most r factors and the lag observations of the target series are relevant for forecasting. As explained above, different settings are reported for simulation. In addition, the correlation among series is set $\rho = 0.5$. Despite the fact that in some cases time series may be highly correlated this value provides a good estimate of the collinearity among series for the purpose of simulation. A total of six settings are to be reported in the simulated data analysis.

5.2.2 Comparative analysis - $\boldsymbol{\Gamma} = \boldsymbol{I}$

Synthetic data are generated for a coefficient matrix with $r(\boldsymbol{C}) = 1$. Tables A.1, A.2, A.3 and A.4 show results for $r(\boldsymbol{C}) = 1$. The matrix of (raw) predictors is created with $p = 120$ and $\tilde{T} = 230$. However, after removing the burn-in period a total of $\tilde{T} = 200$ indexed observations remain in the predictor matrix. The total 120 time series predictors are split in 10 categories as follows:

Table 5.1: Distribution of 10 simulated categories

Category	1	2	3	4	5	6	7	8	9	10
# of Series	9	13	13	10	10	10	13	14	11	17

For training purposes $T = 100$ observations are used. The number of predictors is determined by the lag order m so that when $m = 1$ then $\boldsymbol{X}_{t-h} \in \mathbb{R}^{100\times 120}$, and for $m = 4$ the predictor matrix is $\boldsymbol{X}_{t-h} \in \mathbb{R}^{100\times 480}$. For notation simplicity and better analysis of the tables, let SFAR$(r, \mathfrak{q})$ denote the SFAR with r C-level factors constructed using a number $\mathfrak{q}$ of relevant predictors.

Table A.1 exhibits the percentile distributions of the MSE for all forecasting methods and series relative to the AR(4) for $r(\boldsymbol{C}) = 1$. Information is presented for one-, two-, and four-quarters ahead using the pseudo out-of-sample with $m = 1$. For $h = 1$, results suggest the SFAR$(1, 2)$ is the most convenient method for forecasting with improvements over 75% of the series. For $h = 2$, the SFAR provided a good performance with the SFAR$(1, 2)$ as the best alternative for forecasting. Similar results can be observed for $h = 4$. For variable selection, SFAR identifies and extracts a small group of relevant predictors. For instance, according to the SFAR$(1, 2)$ a group of 65 variables (from the pool of 120) suffices to forecast all categories for $h = 1$ and $m = 1$, and outperforms other methods. Therefore, a 45.8% reduction of junk time series is achieved. Figure 5.1 shows the top 27 selected of the group of 65 predictors with each of them selected at least 10 times. Clearly, the SFAR overcomes both difficulties stated in the introduction to provide parsimony and accuracy in forecasting. Evidently, Ridge performed poorly with values around 4 times those obtained from the SFAR$(1, 2)$. In turn, the DFM5 performed very competitively for all percentiles confirming its predictive power in forecasting.

Forecasting when 4 lags of observations are used to construct C-level factors deteriorates the performance of the DFM5 and SFAR. The AR(4) led most percentiles for all horizons with some exemptions. For instance, the upper 75% percentile of the series when $h = 4$ is dominated by the AR(4). The SFAR$(1, 2)$ used a total of 61 out of the 480 predictors to extract a C-level factor, forecast series and improve percentiles. Figure 5.2 shows the top 21 predictors for all categories with variables selected at least 10 times. For factor construction, notice that variable 398 was selected more than 300 times to construct one C-level factor. Variable 398 stands for row variable 38 with

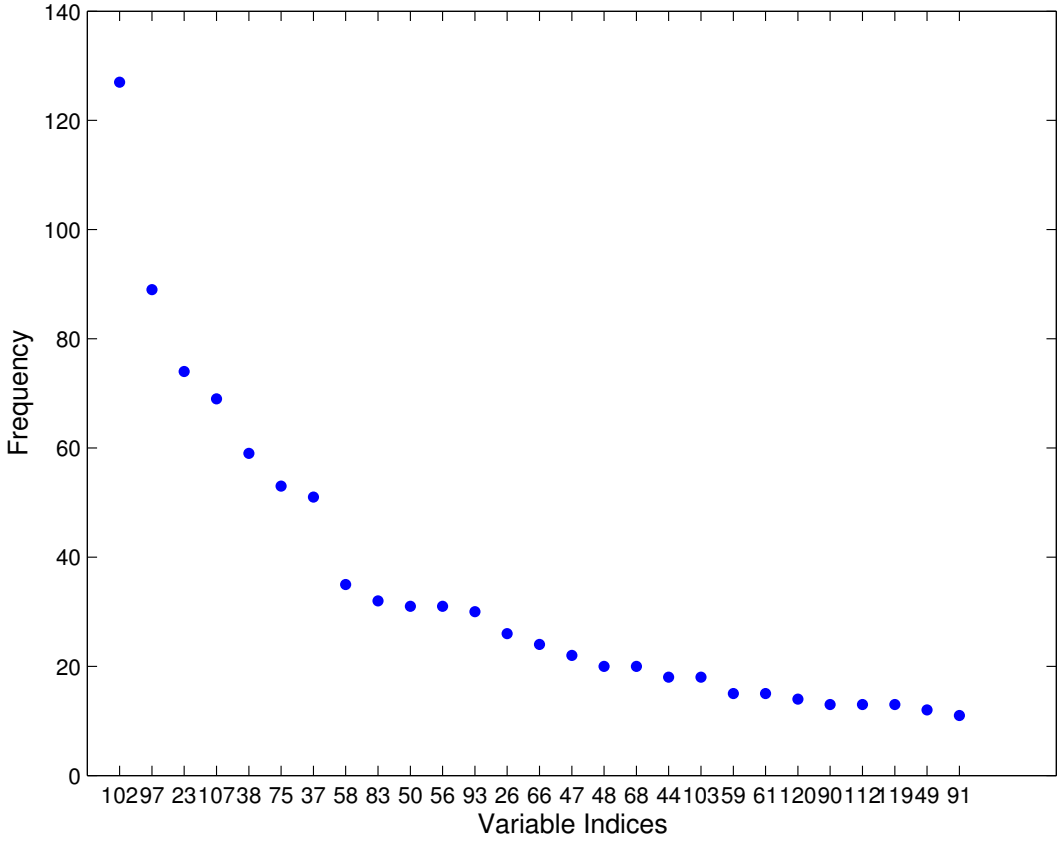


Figure 5.1: Frequency of the selected variables to extract C-level factors for all simulated category of series via SFAR$(1, 2)$ for $h = 1$ and $m = 1$.

$m = 4$ and belongs to category 6. Conveniently, SFAR facilitates not only variable extraction but also provides detailed information about the selected variable.

The dimension reduction achieved by the cardinality control via quantile thresholding rule is dramatic. The number of free parameters in the doubly-constrained $\boldsymbol{C}$ is only $r(\mathfrak{q} + n - r)$ extraordinarily less than the full $pn = 120n$ in the original $\boldsymbol{C}$. Thus, implementation of the selectable reduced-rank regression in Stage 1 is critical to achieve dimension reduction, and control the cardinality set and factor construction.

Comparison of the median of the MSE by all categories of series and methods can be observed in tables A.2 and A.4. Ridge performed poorly through all categories of series and horizons. Table

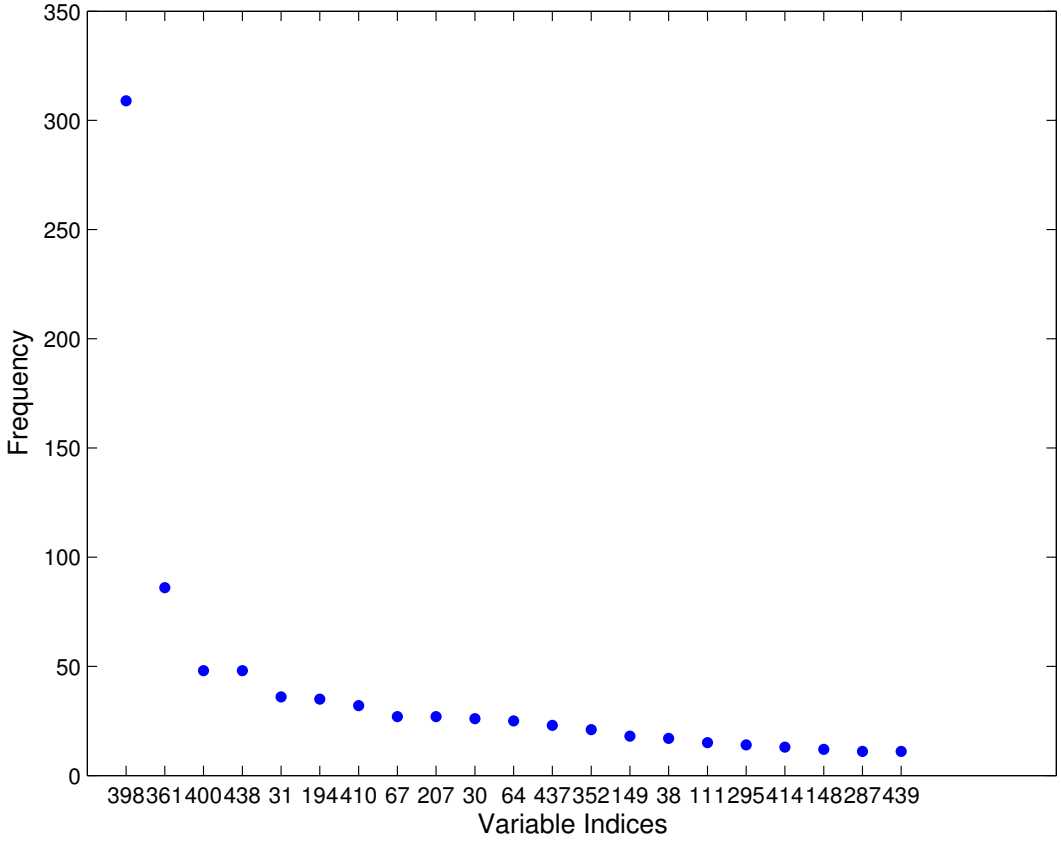


Figure 5.2: Frequency of the selected variables to extract C-level factors for all simulated category of series via SFAR$(1, 2)$ for $h = 4$ and $m = 4$.

A.2 contains information for one-, two-, and four-quarters ahead with $r(\boldsymbol{C}) = 1$ and $m = 1$. For all categories and forecasted horizons the SFAR$(1, 2)$ rised as the best forecasting method. This result favors the idea that only a small set of predictors suffices to forecast all categories. Even more, according to the results, C-level factors show a more predictive power than global factors.

Table A.4 shows different results with the AR(4) rising as the option with the best overall performance for all horizons. As claimed by the results, incorporating additional training information seems to harm methods with orthogonal predictors. For instance, for $h = 1$ and $m = 4$, none of the SFAR models improved the AR(4) although performance of the proposed methodology is very competitive. A reduction in the performance of the DFM5 relative to the AR(4) can also be seen

against results from $m = 1$. This finding argues in favor of using the most recent information for model fitting to strengthen the predictive power of the C-level factors and thus, the forecasting accuracy.

Results reported by all tables advocate implementing the SFAR to achieve competitive performance in forecasting and interpretability for afterwards analysis. Differently to other forecasting methods studied, the SFAR identifies the adequate lag order and relevant variables in response to the target category and series therein. Methods with orthogonal factors showed sensitivity to the amount of information used for model fitting. On the other hand, Ridge benefited from using larger amount of information. It is clear that SFAR provides an advantage over other methods in terms of parsimony and predictive factors. SFAR can be tested under more favorables conditions. For instance, tuning some of the parameters such as m and $\mathfrak{q}$ to maximize the SFAR performance, may be convenient. In particular, the idea of tuning m is intriguing given the impact it has on the predictive power of the C-level factors.

5.3 Macroeconomic data

As described in Stock and Watson [62], the empirical data set consists of quarterly observations on 143 U.S. macroeconomic time series from 1960:II through 2008:IV, for a total of 195 quarterly observations, with earlier observations used for lagged values of regressors as necessary. They grouped the series into 13 categories as listed in table (5.2). From those, only 12 categories are considered (Category 13, Consumer Expectations, was removed) in our analysis since SFAR requires the response matrix to be multivariate. Series are transformed by taking logarithms and/or differencing. [1] Of the 143 series in the data set, 34 are of high-level aggregates. Consequently, only 108 lower-level disaggregated series are considered along the target category responses.

5.3.1 Comparative analysis - $\boldsymbol{\Gamma} = \boldsymbol{I}$

Table B.1 shows the percentile distributions of the MSE for all forecasting methodologies and series relative to the AR(4) for $m = 1$. Information is presented for one-, two-, and four-quarters ahead using the pseudo out-of-sample scheme and $\boldsymbol{\Gamma} = \boldsymbol{I}$. DFM5 rised as the best option for $h = 1$ with improvement over the lower 75% of the series. The SFAR$(1, 2)$ became the best alternative

[1] A complete description of the data set including transformations can be found at http : //www.princeton.edu/ ∼ mwatson/papers/stock_watson_generalized_shrinkage_February_2011.pdf.

Table 5.2: Categories of series in the empirical data set

Group	Brief Description	Examples of Series	# of Series
1	GDP components	GDP, consumption, investment	16
2	IP	IP, capacity utilization	14
3	Employment	Sectoral & total employment and hours	20
4	Unemp. rate	Unemp rates, total and by duration	7
5	Housing	Housing starts, total and by region	6
6	Inventories	NAPM inventories, new orders	6
7	Prices	Price indexes, commodity prices	37
8	Wages	Average hourly earnings, unit labor cost	6
9	Interest rates	Treasuries, Corporate, term spreads	13
10	Money	M1, M2, business loans, consumer credit	7
11	Exchange rates	Average & selected trading partners	5
12	Stock prices	Various stock price indexes	5

for the upper 25%. Forecasting 6-months ahead the SFAR$(2,3)$ stands as the best forecasting method with refinements at percentiles $(25, 50, 95)$, while the AR(4) achieved a good performance at $(5, 75)$ percentiles. The DFM5 deteriorated its performance in comparison with $h = 1$. One-year ahead shows competitive performance between DFM5 and SFAR$(2,3)$ for percentiles $(25, 75)$ with the latter improving also for 95 percentile. In turn, the AR(4) enhanced percentiles $(5, 50)$ with small improvements over the SFAR$(2,3)$. Ridge performed poorly regardless of the horizon forecasted. Notice how Table B.1 supports the implementation of the SFAR$(2,3)$ for forecasting macroeconomic time series when 1-lag of observations is used to construct C-level factors.

Table B.2 reports the median of the MSE by forecasting methods and categories of series, relative to the AR(4) for $m = 1$. Results for one-quarter, two- and fourth-quarters ahead are presented. Ridge regression performed poorly for all categories of series and horizons. Even though Ridge solves the singularity problem, it doesn't provide dimension reduction and orthogonal factors weakening the predictive power of the fitted model. Overall, mixed results can be seen in table B.2. As anticipated in table B.1 the DFM5 provides the smallest MSE for all categories of series for $h = 1$. Its overall performance came from improvements in categories *GDP, Unemployment, Interest rates* and *Stock prices* categories over the AR(4). However, for categories *Housing, Money, Exchange rates* and *Stock prices* the SFAR$(3,5)$ provided a better performance than the DFM5.

Similarly, other SFAR (combinations of $\mathfrak{q}$ and r) provided enhancements over AR(4) for several categories. For instance, category 11, *Exchange rates*, is enhanced by more than 6% via SFAR$(1,3)$ suggesting that C-level factors extracted using the three most relevant predictors in response to *Exchange rates* is adequate to forecast such category. Because of the rolling window scheme, C-level factors are constructed a total of 94 times. Consequently, a slight variation in the selected variables used to construct factors can be noticed. SFAR easily identified these variables. For example, a C-level factor constructed via SFAR$(1,3)$ to forecast series in *Exchange rates* at time $t+1$ (the first rolling window) is constructed using *Sfygm6*, (fygm6-fygm3) in Cat: 9, *Sfygt1*, (fygt1-fygm3) in Cat: 9, and *PMNO*, (NAPM new ordrs) in Cat: 6, as follows:

$$\boldsymbol{z}_1 = -0.0429 \cdot \mathit{Sfygm6} - 0.0442 \cdot \mathit{Sfygt1} + 0.0194 \cdot \mathit{PMNO} \tag{5.3.1}$$

where the intercept is neglected due to normalization. SFAR selects the most relevant predictors to construct specific predictive factors in response to the target category for each window.

Figure 5.3 shows this selected group of variables. It shows 34 out of 42 variables used to construct a C-level factor across 94 forecasting windows. This indicates that 38.89% of the initial predictors are enough to improve the forecasting accuracy of the Exchange rate category. The top 12 variables spread across 6 categories and encompass 57% of the total number (282) of selected variables. These variables are described in table 5.3. According to table 5.3 the SFAR$(1,3)$ utilizes 12 variables, from a pool of 108 candidates, to explain more than 50% of the Exchange rates category. Differently, DFM5 utilizes all 108 predictors.

To forecast 6-months ahead the SFAR$(1,3)$ provided the best overall performance. In that line, the SFAR$(2,3)$ and SFAR$(2,4)$ also improved the overall performance of the benchmark. The SFAR$(1,3)$ enhanced the forecasting of 9 out of 12 categories studied including difficult to forecast categories such as *Prices, Interest rates, Exchange rates*, and *Stock prices*. In general, the SFAR, for any combination of r and $\mathfrak{q}$ provided the smallest MSE of all other methods. The DFM5 did a good job forecasting the *Unemployment* category although its performance for $h=1$ was better. Competitive results provided by the SFAR argue in favor of achieving simultaneous variable selection and factor construction to forecast macroeconomic time series. Remarkable is the 11.54% of reduction in terms of MSE for category *Housing*, given by the SFAR$(1,3)$ over the AR(4). Variations in the results for different lag orders m can be associated to the sensitiveness of

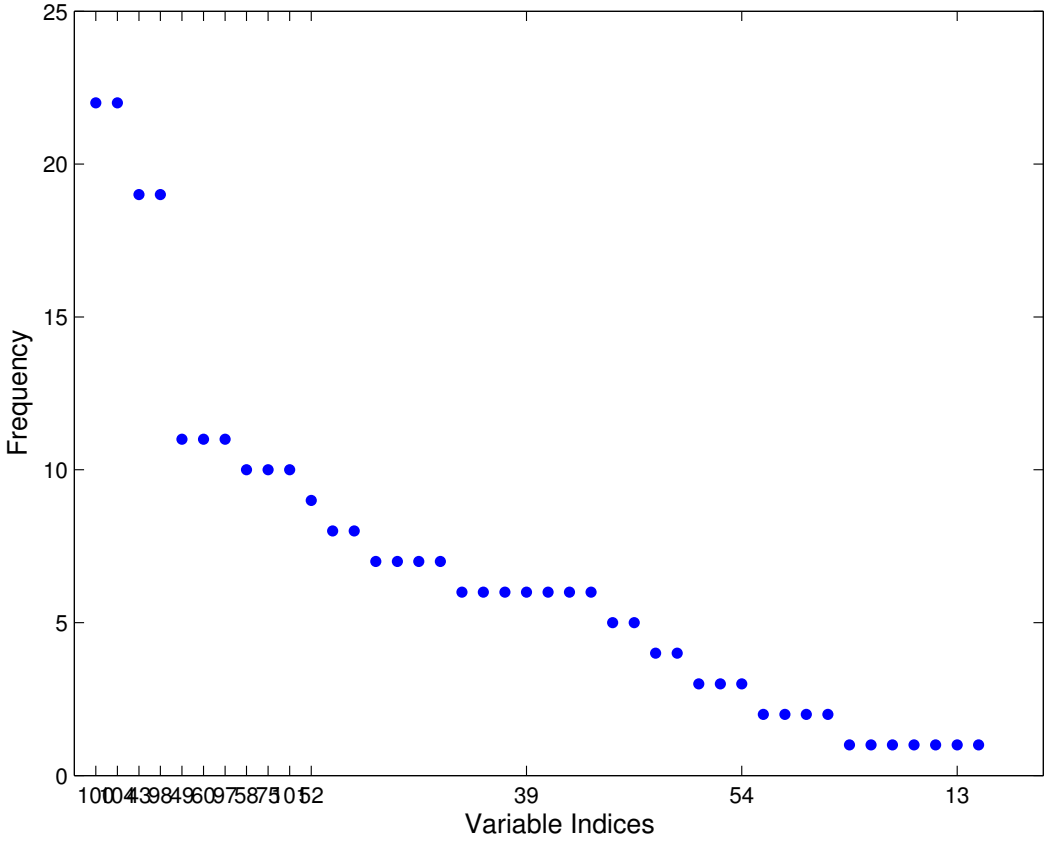


Figure 5.3: Frequency of the selected variables to extract C-level factors for Exchange rates via SFAR$(1,3)$ for $h=1$ and $m=1$.

the initial point $\boldsymbol{S}^{(0)}$ of algorithm 2 to initial conditions of the model. Due to nonconvexity the ℓ_0-constraint can provide a slightly different selection for different parameters even though categories of series are the same.

Even though the SFAR outperformed other methods and provided parsimony and accuracy in forecasting this could not be case for $h=4$. Results for $h=4$ with $m=1$ can be observed in sub-table at the bottom of table B.2. For $h=4$ method utilizing orthogonal factors diminished its predictive power against other horizons. In particular, the overall performance of the SFAR$(2,3)$ suffered a reduction of less than 1% versus results at $h=2$. Only the SFAR$(2,3)$ outperformed every other method in forecasting *Interest rates*. Moreover, it also enhanced *Stock prices* and *Prices*

Table 5.3: Top 12 selected variables to extract a C-level factor for Exchange rates

	ID	Frequency	Name	Short Description	Category
1	100	22	FSPXE	S&P PE ratio	Stock prices
2	104	22	PMNO	NAPM new ordrs	Inventories
3	43	19	PMP	NAPM prodn	IP
4	98	19	FSPIN	S&P: indust	Stock prices
5	49	11	CES017	Emp: dble gds	Employment
6	60	11	LHUR	U: all	Unemployment rate
7	97	11	FSPCOM	S&P 500	Stock prices
8	58	10	LHELX	Help wanted/emp	IP
9	75	10	FYGT1	1 yr T-bond	Interest rates
10	101	10	FSDJ	DJIA	Stock prices
11	52	9	CES048	Emp: TTU	Employment
12	39	8	IPS34	IP: dble mats	IP

among other categories to account for less that 1% away from the top forecaster AR(4). Results suggest the AR(4) is the top performer due to the poor performance showed by the SFAR$(2,3)$ in categories *Housing* and *Money*. Identifying and extracting a handful set of three useful predictors to construct two C-level factors evidence the powerfulness of algorithm 2 to execute the selectable reduced-rank regression while the Alternating-Optimization objective function in (4.3.6) guarantees an optimal solution for $\boldsymbol{S}$ and/or $\boldsymbol{V}$.

Using $m = 2$ to construct C-level factors somewhat changed the picture given at $m = 1$ in table B.1. Table B.3 shows the distribution by forecasting method for one-, two-, and fourth-quarter ahead for $m = 2$ for all series relative to the AR(4). For $h = 1$, The SFAR$(1,2)$ outperformed other methods at percentiles $(25, 50, 95)$. The DFM5 reduced its performance against its results for $m = 1$. Ridge, instead, notably improved its performance at $m = 1$ and $h = 1$. For $h = 2$ the AR(4) improved at percentiles $(5, 50, 75)$ while the SFAR$(1,3)$ did so for remaining percentiles. For $h = 4$ the AR(4) rised as a consistent alternative for forecasting. Percentile of distributions in table B.3 offered mix results with the SFAR$(2,3)$ as the preferred option among SFAR models.

Table B.2 displays the results for one-quarter, half- and one-year ahead for $m = 2$ by forecasting method for all categories and series, relative to the AR(4). Results of forecasting one-period ahead show the DFM5 as the best option overall providing the best forecasting performance for

categories *GDP* and *Unemployment*. On the other hand, the SFAR$(2,3)$ outperformed all methods for forecasting categories *Inventories* and *Interest rates*. However, it performed poorly for other categories such as *Employment*. Notice how category *Money* responds better as the number of factors and selected variable increase. For $h = 2$ the performance of the SFAR models deteriorated with the SFAR$(2,3)$ supplying the smallest MSE for all the macroeconomic time series. The AR(4) benchmark supplied the best forecast overall although this is not true for all categories. For instance, categories *GDP, Employment, Inventories* and Prices were better forecasted via SFAR$(2,4)$, SFAR$(2,4)$, DFM5, and SFAR$3,5$ respectively. To forecast one-year ahead, again, the AR(4) rised as the best overall option. The SFAR, in general, deteriorated even more its performance as well as the DFM5. As observed in table B.2 forecasting four-periods ahead finds an opportunity for improvement in the SFAR.

Forecasting series using $m = 4$ provided mixed results. Table B.5 shows the distribution by forecasting method for one-, two-, and fourth-quarter ahead for $m = 4$ for all series relative to the AR(4). For $h = 1$, distributions of the SFAR$(1,2)$ and DFM5 assorted dominance over displayed percentiles. Distribution for $h = 2$ found the AR(4) as the best alternative at $(5, 50, 75)$ percentiles with the remaining percentiles improved by the SFAR$(1,2)$. Results for $h = 2$ and $m = 4$ are similar to those at $h = 2$ and $m = 2$. For $h = 4$ the SFAR$(2,3)$ offered a convenient option by improving the 1st and 3rd quantiles, and percentile 95%. On the other hand, the AR(4) showed good performance for $(5, 50)$ percentiles. Broadly speaking, different methods provided similar outcomes in terms of forecasting performance for $h = 2$ and consequently, a preferred method was not identified.

Table B.6 provides evidence of a competitive performance among forecasting methods. Table B.6 displays the results for one-quarter, half- and one-year ahead for $m = 4$ by forecasting method for all categories and series, relative to the AR(4). Of interest is the analysis of the SFAR for four specific categories: *Price, Interest rates, Exchange rates* and *Stock prices* since they are considered to be the most difficult categories in terms of forecasting. For $h = 1$, the AR(4) rised as the appropriate option to forecast the *Price* category. *Interest rates* and *Stock prices* found the SFAR$(2,3)$ as the most suitable forecasting method. Finally, *Exchange rates* was enhanced via the SFAR$(1,2)$. Using the SFAR$(2,3)$ seems to be a good option for those categories when $h = 1$. For $h = 2$, most of these categories encountered the SFAR$(2,3)$ as a convenient forecasting option. An exception can be seen

for *Stock prices* where the AR(4) was the best forecaster. However, the SFAR$(2,3)$ positioned itself less than 1% away. For $h = 4$, the SFAR$(2,3)$ rised as one of the two alternatives to simultaneously enhance categories *Prices* and *Stock prices*. The other two categories found the AR(4) as the best forecasting alternatives. As claimed by the results, the SFAR$(2,3)$ provides a convenient alternative to forecast these categories. However, including additional lag of observations to construct C-level factors appears to lessen its performance.

5.3.2 Comparative analysis for different $\mathbf{\Gamma}$

Adding different combinations of $\mathbf{\Gamma}$ through variations of $(\hat{\mathbf{\Sigma}}_{\boldsymbol{YY}}, \boldsymbol{D}_{(\hat{\mathbf{\Sigma}}_{\boldsymbol{YY}})}, \boldsymbol{I})$ and δ offered mixed results for different forecasted horizons. From our experiments there is not a universal choice of $\mathbf{\Gamma}$ and δ. Results claim that $\mathbf{\Gamma}$ is sensitive to the initial parameters of the model and even within the same category different $\mathbf{\Gamma}$s can be observed for different models. Nonetheless, a pair of $\mathbf{\Gamma}$ and δ were selected from testing $\mathbf{\Gamma} = (\hat{\mathbf{\Sigma}}_{\boldsymbol{YY}}, \boldsymbol{D}_{(\hat{\mathbf{\Sigma}}_{\boldsymbol{YY}})}, \boldsymbol{I})$ and $\delta = (0, 0.001, 0.1, 0.5, 1)$ for each category.

Table C.2 displays the results for one-quarter, half- and one-year ahead for $m = 1$ by forecasting method for all categories and series, relative to the AR(4). This table shows a reduction in the forecasting accuracy per SFAR model. For instance, the accuracy of the SFAR for forecasting categories *GDP, IP, Employment* and *Inventories* slumped for implemented combinations of $\mathbf{\Gamma}$ and δ. However, a different picture is drawn for $h = 2$ where the performance of the SFAR models improved in comparison with results in table B.2. In fact, SFAR$(1,3)$ and SFAR$(3,4)$ achieved enhancements over 9 out of 12 categories against the AR(4), with the former providing the best overall performance. Overall results argue in favor of adding a weight matrix $\mathbf{\Gamma}$ and parameter δ to improve the forecasting accuracy. Remarkable is the fact that SFAR$(1,2)$ is now the best forecasting alternative for $h = 4$.

The improvement offered by adding $\mathbf{\Gamma}$ and δ is very small when $m = 2$. Table C.3 displays the results for one-quarter, half- and one-year ahead for $m = 2$ by forecasting method for all categories and series, relative to the AR(4). For $h = 1$ the overall performance of the SFAR$(2,3)$ improved less than 1% with enhancements in GDP, Housing, Inventories and Unemployment categories among others. However, its accuracy for categories *IP* and *Employment* notably diminished. For $h = 2$ and $h = 4$, the improvement can be noticed for specific categories. For example, for $h = 2$ the accuracy performance of the SFAR for *Unemployment* improved versus the same results in table B.4. The same happened for category *GDP* when $h = 4$.

Results showed in table C.4 favored adding $\mathbf{\Gamma}$ and δ only for $h = 4$. Other horizons can not take advantage of the relation among series to improve the forecasting accuracy. In general, using $m = 4$ suggests using two C-level factors constructed using the three most relevant variables to enhance the forecasting accuracy of the categories. However, there are still many experiments to perform in order to identify, in case it exists, a better combination of $\mathbf{\Gamma}$ and δ suitable to the initial parameters of the model and forecasting horizon. Therefore, selecting the weight matrix $\mathbf{\Gamma}$ and values of δ can be exhausting since for different initial conditions these values may vary. Tuning parameter δ is difficult and may increase the computational burden of the methodology although in some cases refinement may compensate for the effort.

CHAPTER 6

DISCUSSION AND FUTURE WORK

In this work, a new methodology, Sparse Factor Auto-Regression, to forecast univariate time series in high dimensions with very many predictors is proposed and studied. SFAR is a three-step methodology that enforces group sparsity and low rank modeling to extract predictive category-level factors in response to a target category, useful to forecast a target series of interest. In the first step, a doubly-constrained multivariate selectable reduced-rank regression model proved to be sensible in satisfying practical concerns. Step 2 extracts uncorrelated orthogonal predictors (C-level factors) to reduce model dimensionality even more and comply with linear regression assumptions. The final step uses a hybrid forecasting model using lagged observations and uncorrelated orthogonal predictors in a regression framework to exploit all sources of information available and measure the predictive power of the factors.

A reliable SFAR feature lies in its capacity to identify and construct an adequate number of factors restricted to an upper bound defined in the rank constraint. It is also versatile enough to handle a very large number of predictors without decreasing the forecasting accuracy of the target series. The proposed ℓ_0 constraint and the variable screening of the PRCCR algorithm offer enough flexibility in the interplay between $\mathfrak{q}$ and r while optimizing a doubly-constrained coefficient matrix in response to a target category response. Variable screening performance was found to be sensitive to the decay rate function. However, different experiments revealed Sigmoid function as the best alternative to optimize the coefficient matrix.

The proposed SFAR was compared to Ridge, AR(4), and DFM5 with the AR(4) as the benchmark. This benchmark agrees with other analysis in the macroeconomic literature, see, e.g. [65]. SFAR proved competitive against well established forecasting methods in the literature in simulations. Ridge performed poorly in general. DFM5 offered comparative results to those of the SFAR and AR(4) although for selected values of $\mathfrak{q}$ and r the SFAR outperformed other methods at some percentiles. The SFAR was found to lose much of its predictive power for $m = 4$.

Empirical results showed a different picture. SFAR provided good performance overall although its performance diminished for larger horizons and values of m. This phenomenon raised the question if there exists a difference between grouping patterns among categories. It is still hypothesized that more homogenous categories can construct C-level factors with more predictive power. To ascertain this issue a variation in the weight matrix Γ was introduced. Results showed an improvement, important in some cases, for all SFAR across categories.

The analysis of results in the comparative study from simulated and empirical data showed SFAR as a consistent methodology to provide parsimony and forecasting accuracy. Moreover, our analysis identified the lag order m as a critical parameter, beside $\mathfrak{q}$ and r, in extracting predictive C-level factors. For instance, SFAR performed better for $m = 1$ than for $m = 4$. This occurrence consents to the assumption that the more recent lags should provide more reliable information than the more distant ones, see, Bańbura [6]. Consequenlty, a tuning strategy for m may serve to identify the right amount of information needed for model fitting.

Results argue in favor of using the SFAR$(2, 3)$ for $m = 2$ to forecast all categories of series. However, more combinations of r and $\mathfrak{q}$ are to be tested in the future. Moreover, different associations of $\mathbf{\Gamma}$ and δ are to be tested by model as well, to identify a point of agreement, for all categories and per category to provide a reliable structure of the macroeconomic time series categories.

Direct multi-period ahead forecast was favored in this study to compare SFAR with existent methodologies through the MSE via rolling window scheme. However, a different alternative such as the iterated one-period ahead forecast technique could also be implemented to compare and measure the predictive power of the factors through time. Evidently, the multi-period approach is convenient as it estimates factors at each rolling window. Nonetheless, it is not realistic. The iterated one-period ahead manner can provide a different picture of the factors performance. For instance, questions such as: how many periods ahead can one go in forecasting using the same set of factors, and what is the impact of parameters $\mathfrak{q}$, r, and m for different horizons, can be raised in real world applications, and answered through the implementation of the iterated one-period ahead forecasting technique.

APPENDIX A

TABLES: SIMULATED RESULTS FOR $\Gamma = I$

Table A.1: Distribution of the MSE, for $m = 1$ in simulated data with $r(\boldsymbol{C}) = 1$, by forecasting method, relative to the AR(4) for $\boldsymbol{\Gamma} = \boldsymbol{I}$.

	Percentiles				
Method	0.050	0.250	0.500	0.750	0.950
			$h = 1$		
Ridge	4.4931	4.5883	4.8713	5.0565	5.3164
AR(4)	**1.0000**	**1.0000**	**1.0000**	**1.0000**	**1.0000**
DFM5	1.0225	1.0174	1.0421	1.0200	1.0271
SFAR $(1, 2)$	1.0052	0.9990	0.9995	0.9818	0.9890
SFAR $(1, 3)$	1.0227	1.0060	0.9993	0.9783	0.9952
SFAR $(2, 3)$	0.9958	1.0044	1.0032	0.9817	0.9901
			$h = 2$		
Ridge	4.1958	4.3402	4.7273	4.8539	5.5462
AR(4)	**1.0000**	**1.0000**	**1.0000**	**1.0000**	**1.0000**
DFM5	1.0260	1.0199	1.0170	1.0177	1.0188
SFAR $(1, 2)$	0.9976	0.9816	0.9943	0.9843	0.9826
SFAR $(1, 3)$	1.0213	0.9949	0.9955	0.9959	0.9847
SFAR $(2, 3)$	1.0204	0.9974	0.9852	0.9957	0.9828
			$h = 4$		
Ridge	4.4939	4.5658	4.8660	4.9969	5.3353
AR(4)	**1.0000**	**1.0000**	**1.0000**	**1.0000**	**1.0000**
DFM5	1.0295	1.0326	1.0254	1.0271	1.0319
SFAR $(1, 2)$	0.9785	0.9754	0.9914	0.9974	0.9907
SFAR $(1, 3)$	0.9865	0.9727	0.9859	0.9971	1.0038
SFAR $(2, 3)$	1.0019	0.9783	0.9971	1.0023	0.9949

Table A.2: Median results, for $m = 1$ with $r(\boldsymbol{C}) = 1$, of the MSE by forecasting method and category of series, relative to AR(4) for $\boldsymbol{\Gamma} = \boldsymbol{I}$.

Category	Ridge	**AR(4)**	DFM5	SFAR $(1,2)$	SFAR $(1,3)$	SFAR $(2,3)$
			$h = 1$			
1	4.7584	**1.0000**	1.0206	0.9919	0.9871	0.9794
2	4.5256	**1.0000**	1.0262	0.9952	0.9951	0.9954
3	4.4884	**1.0000**	1.0311	0.9964	0.9939	1.0007
4	4.8562	**1.0000**	1.0218	0.9858	0.9885	0.9853
5	4.8900	**1.0000**	1.0217	0.9753	0.9860	0.9754
6	4.7800	**1.0000**	1.0189	0.9866	0.9921	1.0004
7	4.5932	**1.0000**	1.0216	0.9811	0.9982	1.0074
8	4.7221	**1.0000**	1.0184	0.9994	1.0058	1.0030
9	4.8602	**1.0000**	1.0254	0.9889	0.9917	0.9786
10	4.8437	**1.0000**	1.0201	0.9996	1.0033	1.0033
Overall	4.7318	**1.0000**	1.0226	0.9900	0.9942	0.9929
			$h = 2$			
1	4.5015	**1.0000**	1.0197	0.9690	0.9872	0.9839
2	4.8179	**1.0000**	1.0195	0.9961	1.0033	1.0086
3	4.5683	**1.0000**	1.0207	0.9901	1.0104	0.9972
4	4.8814	**1.0000**	1.0223	0.9799	0.9930	0.9852
5	4.7267	**1.0000**	1.0222	0.9785	0.9846	0.9931
6	4.5972	**1.0000**	1.0174	0.9812	0.9920	0.9879
7	4.8334	**1.0000**	1.0351	1.0005	1.0098	1.0133
8	4.6124	**1.0000**	1.0285	0.9685	0.9863	0.9841
9	4.5651	**1.0000**	1.0325	0.9965	0.9993	0.9887
10	4.4999	**1.0000**	1.0176	0.9719	0.9803	0.9857
Overall	4.6604	**1.0000**	1.0235	0.9832	0.9946	0.9928
			$h = 4$			
1	4.7067	**1.0000**	1.0160	0.9782	0.9949	0.9897
2	4.9334	**1.0000**	1.0265	0.9944	1.0044	1.0057
3	4.5404	**1.0000**	1.0223	0.9844	0.9831	0.9829
4	4.7169	**1.0000**	1.0176	0.9920	1.0006	0.9905
5	4.6291	**1.0000**	1.0225	0.9828	0.9856	0.9832
6	4.8642	**1.0000**	1.0173	0.9892	0.9934	0.9937
7	5.0359	**1.0000**	1.0233	0.9892	1.0071	1.0105
8	4.7539	**1.0000**	1.0212	0.9863	0.9983	0.9910
9	4.8015	**1.0000**	1.0253	0.9968	0.9997	1.0069
10	4.8169	**1.0000**	1.0219	0.9672	0.9811	0.9804
Overall	4.7799	**1.0000**	1.0214	0.9860	0.9948	0.9934

Table A.3: Distribution of the MSE, for $m = 4$ in simulated data with $r(\boldsymbol{C}) = 1$, by forecasting method, relative to the AR(4) for $\boldsymbol{\Gamma} = \boldsymbol{I}$.

	Percentiles				
Method	0.050	0.250	0.500	0.750	0.950
			$h = 1$		
Ridge	1.2006	1.2174	1.2340	1.2417	1.2596
AR(4)	**1.0000**	**1.0000**	**1.0000**	**1.0000**	**1.0000**
DFM5	1.0541	1.0446	1.0440	1.0479	1.0398
SFAR $(1, 2)$	1.0182	0.9888	0.9952	1.0072	1.0028
SFAR $(1, 3)$	1.0162	1.0143	1.0380	1.0239	1.0360
SFAR $(2, 3)$	1.0464	1.0178	1.0096	1.0101	1.0251
			$h = 2$		
Ridge	1.2138	1.2025	1.2269	1.2171	1.2141
AR(4)	**1.0000**	**1.0000**	**1.0000**	**1.0000**	**1.0000**
DFM5	1.0682	1.0268	1.0417	1.0165	1.0338
SFAR $(1, 2)$	1.0168	0.9979	1.0140	0.9991	1.0075
SFAR $(1, 3)$	1.0591	1.0409	1.0415	1.0339	1.0167
SFAR $(2, 3)$	1.0575	1.0281	1.0155	1.0182	1.0181
			$h = 4$		
Ridge	1.2154	1.2421	1.2459	1.2563	1.2664
AR(4)	**1.0000**	**1.0000**	**1.0000**	**1.0000**	**1.0000**
DFM5	1.0436	1.0299	1.0493	1.0356	1.0406
SFAR $(1, 2)$	0.9998	0.9940	1.0026	1.0086	1.0338
SFAR $(1, 3)$	1.0323	1.0196	1.0358	1.0331	1.0537
SFAR $(2, 3)$	0.9997	1.0169	1.0291	1.0221	1.0445

Table A.4: Median results, for $m = 4$ with $r(\boldsymbol{C}) = 1$, of the MSE by forecasting method and category of series, relative to AR(4) for $\boldsymbol{\Gamma} = \boldsymbol{I}$.

Category	Ridge	**AR(4)**	DFM5	SFAR $(1,2)$	SFAR $(1,3)$	SFAR $(2,3)$
			$h = 1$			
1	1.2083	**1.0000**	1.0462	1.0057	1.0308	1.0355
2	1.2353	**1.0000**	1.0414	0.9969	1.0319	1.0270
3	1.1933	**1.0000**	1.0463	0.9927	1.0455	1.0273
4	1.1942	**1.0000**	1.0388	0.9858	1.0165	1.0113
5	1.2021	**1.0000**	1.0379	1.0172	1.0533	1.0117
6	1.1918	**1.0000**	1.0328	0.9986	1.0295	1.0491
7	1.1884	**1.0000**	1.0371	0.9930	1.0218	0.9972
8	1.2024	**1.0000**	1.0434	0.9990	1.0302	1.0273
9	1.2090	**1.0000**	1.0271	1.0052	1.0319	1.0262
10	1.2282	**1.0000**	1.0409	1.0084	1.0189	1.0286
Overall	1.2053	**1.0000**	1.0392	1.0003	1.0310	1.0241
			$h = 2$			
1	1.1836	**1.0000**	1.0622	1.0249	1.0455	1.0431
2	1.2147	**1.0000**	1.0338	1.0216	1.0652	1.0529
3	1.1847	**1.0000**	1.0360	1.0161	1.0422	1.0447
4	1.2038	**1.0000**	1.0430	1.0049	1.0446	1.0349
5	1.1895	**1.0000**	1.0405	1.0125	1.0405	1.0394
6	1.1865	**1.0000**	1.0329	1.0051	1.0394	1.0113
7	1.1988	**1.0000**	1.0398	1.0118	1.0162	1.0178
8	1.2084	**1.0000**	1.0439	1.0134	1.0268	1.0323
9	1.2298	**1.0000**	1.0430	1.0109	1.0353	1.0299
10	1.2179	**1.0000**	1.0331	1.0047	1.0323	1.0238
Overall	1.2018	**1.0000**	1.0408	1.0126	1.0388	1.0330
			$h = 4$			
1	1.1940	**1.0000**	1.0508	1.0204	1.0468	1.0378
2	1.2131	**1.0000**	1.0331	0.9838	1.0216	1.0081
3	1.1792	**1.0000**	1.0291	0.9900	0.9993	1.0046
4	1.2159	**1.0000**	1.0369	1.0200	1.0421	1.0504
5	1.2109	**1.0000**	1.0368	1.0011	1.0238	1.0146
6	1.2553	**1.0000**	1.0512	1.0133	1.0463	1.0401
7	1.2172	**1.0000**	1.0444	1.0004	1.0155	1.0077
8	1.2125	**1.0000**	1.0313	0.9979	1.0173	1.0253
9	1.2333	**1.0000**	1.0500	1.0107	1.0375	1.0489
10	1.2553	**1.0000**	1.0246	0.9993	1.0080	1.0205
Overall	1.2187	**1.0000**	1.0388	1.0037	1.0258	1.0258

APPENDIX B

TABLES: MACRO RESULTS FOR $\mathbf{\Gamma} = \boldsymbol{I}$

Table B.1: Distribution of the MSE, for $m = 1$, by forecasting methods, relative to the AR(4) for $\mathbf{\Gamma} = \boldsymbol{I}$.

	Percentiles				
Method	0.050	0.250	0.500	0.750	0.950
			$h = 1$		
Ridge	24.8862	27.7574	22.8585	15.2527	16.0663
AR(4)	**1.0000**	**1.0000**	**1.0000**	**1.0000**	**1.0000**
DFM5	0.9803	0.9991	0.9665	0.9617	1.0413
SFAR $(1,2)$	1.1289	1.0742	0.9939	1.0164	0.8902
SFAR $(1,3)$	1.1615	1.1095	0.9813	1.0623	0.9396
SFAR $(2,3)$	1.2017	1.0971	0.9753	1.0034	0.9987
SFAR $(2,4)$	1.1440	1.0857	1.0359	1.0154	1.0199
SFAR $(3,4)$	1.1764	1.0946	1.0443	0.9639	0.9751
SFAR $(3,5)$	1.1302	1.0432	1.0089	1.0252	1.0431
			$h = 2$		
Ridge	15.8045	20.9175	19.3059	10.1909	12.2770
AR(4)	**1.0000**	**1.0000**	**1.0000**	**1.0000**	**1.0000**
DFM5	1.1586	1.0303	1.0800	1.0800	1.0033
SFAR $(1,2)$	1.1576	0.9963	1.0344	1.0721	0.9055
SFAR $(1,3)$	1.1113	0.9778	1.0354	1.0871	0.8709
SFAR $(2,3)$	1.1654	0.9754	0.9873	1.0715	0.9037
SFAR $(2,4)$	1.1213	0.9838	1.0203	1.0823	0.8782
SFAR $(3,4)$	1.0546	0.9869	1.0417	1.0883	0.9262
SFAR $(3,5)$	1.1057	0.9626	1.0191	1.0824	0.9346
			$h = 4$		
Ridge	15.5000	18.6983	17.9581	10.2552	10.2864
AR(4)	**1.0000**	**1.0000**	**1.0000**	**1.0000**	**1.0000**
DFM5	1.1659	0.9797	1.0839	0.9068	1.1187
SFAR $(1,2)$	1.1164	0.9909	1.0069	0.9884	0.9609
SFAR $(1,3)$	1.0941	0.9899	1.0466	0.9822	0.9759
SFAR $(2,3)$	1.0773	0.9840	1.0070	0.8980	0.9548
SFAR $(2,4)$	1.0832	0.9827	1.0441	0.9089	0.9328
SFAR $(3,4)$	1.0666	1.0021	1.0061	0.9076	0.9418
SFAR $(3,5)$	1.0783	0.9788	1.0384	0.9566	0.9494

Table B.2: Median results, for $m = 1$, of the MSE by forecasting method and category of series, relative to AR(4) for $\boldsymbol{\Gamma} = \boldsymbol{I}$.

#	Category	Ridge	**AR(4)**	DFM5	SFAR (1, 2)	SFAR (1, 3)	SFAR (2, 3)	SFAR (2, 4)	SFAR (3, 4)	SFAR (3, 5)
					$h = 1$					
1	GDP	18.1340	**1.0000**	0.8575	1.0084	1.0221	0.9749	0.9745	1.0138	0.9997
2	IP	21.5913	**1.0000**	1.0633	1.0261	1.0596	1.0476	1.1031	1.1241	1.1057
3	Emp	34.3293	**1.0000**	1.0587	1.2299	1.2059	1.2692	1.2079	1.2900	1.2647
4	Unemp	17.7063	**1.0000**	0.6783	0.9885	1.0304	0.8513	0.8615	0.8178	0.8809
5	Housing	19.6026	**1.0000**	1.0307	0.9917	1.1079	1.0462	1.0654	0.9996	0.9611
6	Invent	53.6031	**1.0000**	1.0319	0.8893	0.9344	0.9984	1.0170	0.9802	1.0470
7	Prices	25.1174	**1.0000**	1.1068	1.0799	1.0511	1.0754	1.0483	1.0319	1.0911
8	Wages	20.4803	**1.0000**	1.0097	1.0185	1.0316	1.0174	1.0153	1.0132	1.0208
9	I. rates	20.9838	**1.0000**	0.9885	0.9733	0.9818	0.9829	1.0144	1.0011	0.9977
10	Money	18.3896	**1.0000**	1.0110	1.0558	1.0216	0.9189	0.9219	0.8856	0.8772
11	E. rates	18.5152	**1.0000**	1.0274	0.9685	0.9373	0.9556	0.9513	0.9609	0.9746
12	Stock	16.7258	**1.0000**	0.9753	0.9843	0.9895	0.9842	0.9913	0.9842	0.9592
	Overall	23.7649	**1.0000**	0.9866	1.0179	1.0311	1.0102	1.0143	1.0085	1.0150
					$h = 2$					
1	GDP	19.7166	**1.0000**	0.9336	0.9599	0.9776	0.9340	0.9564	0.9929	0.9883
2	IP	18.4553	**1.0000**	1.0034	1.0157	1.0856	1.0218	1.0373	1.0567	1.0171
3	Emp	23.6078	**1.0000**	0.9831	1.0588	1.0472	1.0399	1.0464	1.0501	1.0408
4	Unemp	18.6960	**1.0000**	0.8869	1.0565	0.9962	0.9914	0.9480	0.9519	1.0035
5	Housing	12.4436	**1.0000**	1.0436	1.0214	1.0642	1.0583	1.1010	1.0963	1.0659
6	Invent	31.8403	**1.0000**	0.9584	0.9023	0.8846	0.9801	0.9457	0.9830	0.9926
7	Prices	15.4604	**1.0000**	1.0248	0.9866	0.9824	1.0140	0.9750	0.9902	0.9869
8	Wages	20.7056	**1.0000**	1.0244	0.9767	0.9900	0.9721	0.9531	0.9479	0.9493
9	I. rates	20.2604	**1.0000**	1.1361	0.9704	0.9484	0.9475	0.9797	0.9745	0.9858
10	Money	15.0565	**1.0000**	1.1375	1.0727	0.9963	1.0215	1.0203	1.0014	1.0145
11	E. rates	17.7842	**1.0000**	1.0587	1.0076	0.9855	0.9992	1.0192	1.0155	1.0100
12	Stock	17.3238	**1.0000**	1.0273	1.0020	0.9990	0.9930	0.9945	0.9868	0.9808
	Overall	19.2792	**1.0000**	1.0181	1.0025	0.9964	0.9977	0.9981	1.0039	1.0030
					$h = 4$					
1	GDP	16.3232	**1.0000**	1.0043	0.9638	0.9490	0.9550	0.9539	0.9902	0.9658
2	IP	14.9523	**1.0000**	0.9963	1.0091	1.0166	1.0220	1.0235	1.0407	0.9953
3	Emp	17.1151	**1.0000**	0.9520	1.0506	1.0377	1.0426	1.0521	1.0625	1.0393
4	Unemp	15.1694	**1.0000**	0.7952	0.9423	0.9538	0.8985	0.9103	0.8684	0.9458
5	Housing	7.4447	**1.0000**	1.0146	1.0748	1.0791	1.0771	1.1279	1.1426	1.1014
6	Invent	18.0097	**1.0000**	0.9406	0.9816	0.9840	1.0058	0.9540	0.9860	1.0128
7	Prices	14.8364	**1.0000**	1.0419	0.9909	0.9899	0.9840	0.9827	1.0021	0.9782
8	Wages	21.5682	**1.0000**	1.0991	1.0256	1.0349	1.0219	1.0111	1.0517	1.0187
9	I. rates	21.6965	**1.0000**	1.1125	1.0162	1.0250	0.9831	1.0343	1.0297	1.0847
10	Money	18.1060	**1.0000**	1.2713	1.0893	1.0261	1.0632	1.0649	1.1045	1.1082
11	E. rates	16.8720	**1.0000**	1.0541	0.9908	0.9986	1.0063	1.0086	1.0024	1.0087
12	Stock	16.2929	**1.0000**	1.0565	0.9967	0.9990	0.9914	0.9820	0.9836	0.9833
	Overall	16.5322	**1.0000**	1.0282	1.0110	1.0078	1.0042	1.0088	1.0220	1.0202

Table B.3: Distribution of the MSE, for $m = 2$, by forecasting methods, relative to the AR(4) for $\boldsymbol{\Gamma} = \boldsymbol{I}$.

	Percentiles				
Method	0.050	0.250	0.500	0.750	0.950
			$h = 1$		
Ridge	3.1617	3.1111	2.7716	2.3539	1.9387
AR(4)	**1.0000**	**1.0000**	**1.0000**	**1.0000**	**1.0000**
DFM5	1.0088	1.0160	0.9901	0.9459	0.9498
SFAR $(1, 2)$	1.1716	0.9943	0.9461	1.0344	0.9225
SFAR $(1, 3)$	1.1873	1.0661	1.0021	1.0254	0.9464
SFAR $(2, 3)$	1.1530	0.9978	1.0107	1.0003	0.9156
SFAR $(2, 4)$	1.0585	1.0129	0.9993	0.9286	0.9768
SFAR $(3, 4)$	1.0961	1.0075	1.0234	0.9675	0.9627
SFAR $(3, 5)$	1.0882	1.0051	1.0519	1.0733	1.0183
			$h = 2$		
Ridge	2.6967	2.6629	2.7120	2.0451	1.7361
AR(4)	**1.0000**	**1.0000**	**1.0000**	**1.0000**	**1.0000**
DFM5	1.0709	1.0664	1.0685	1.0665	1.0175
SFAR $(1, 2)$	1.1269	0.9881	1.1045	1.0781	1.0380
SFAR $(1, 3)$	1.0864	0.9952	1.1277	1.0775	0.9728
SFAR $(2, 3)$	1.0894	0.9826	1.0927	1.0423	1.0191
SFAR $(2, 4)$	1.0240	0.9820	1.0831	1.0674	1.0482
SFAR $(3, 4)$	1.0363	0.9845	1.1093	1.0318	1.0337
SFAR $(3, 5)$	1.0699	0.9914	1.1060	1.0434	1.0424
			$h = 4$		
Ridge	2.1292	2.5331	2.7048	2.0041	1.5671
AR(4)	**1.0000**	**1.0000**	**1.0000**	**1.0000**	**1.0000**
DFM5	1.1526	0.9255	1.0779	0.9076	1.1503
SFAR $(1, 2)$	1.0987	0.9659	1.0712	0.9736	1.1526
SFAR $(1, 3)$	1.0671	0.9727	1.1245	1.0136	1.0934
SFAR $(2, 3)$	1.0988	0.9414	1.1124	0.9554	1.1116
SFAR $(2, 4)$	1.0498	0.9539	1.0751	0.9319	1.1426
SFAR $(3, 4)$	1.0339	0.9246	1.0780	0.9322	1.1635
SFAR $(3, 5)$	1.0386	0.9407	1.1085	0.9307	1.1345

Table B.4: Median results, for $m = 2$, of the MSE by forecasting method and category of series, relative to AR(4) for $\boldsymbol{\Gamma} = \boldsymbol{I}$.

#	Category	Ridge	**AR(4)**	DFM5	SFAR (1, 2)	SFAR (1, 3)	SFAR (2, 3)	SFAR (2, 4)	SFAR (3, 4)	SFAR (3, 5)
					$h = 1$					
1	GDP	2.3108	**1.0000**	0.9108	0.9811	0.9817	0.9788	0.9409	0.9431	0.9465
2	IP	2.6451	**1.0000**	1.0340	1.1116	1.0935	1.1620	1.1791	1.2043	1.2240
3	Emp	3.6446	**1.0000**	1.1681	1.3591	1.3967	1.2608	1.1287	1.2388	1.1936
4	Unemp	1.8865	**1.0000**	0.6620	0.8974	0.8477	0.8251	0.7365	0.7774	0.7565
5	Housing	2.5280	**1.0000**	1.1199	1.0525	1.0632	1.0544	1.0162	1.0701	1.1223
6	Invent	3.9751	**1.0000**	0.9473	0.9418	0.9796	0.9294	0.9926	0.9695	1.0284
7	Prices	2.5697	**1.0000**	1.0636	1.1015	1.0888	1.0912	1.0712	1.0505	1.1069
8	Wages	2.4641	**1.0000**	0.9973	0.9880	0.9948	1.0407	1.0288	1.0677	1.0738
9	I. rates	4.0165	**1.0000**	0.9695	1.0081	1.0473	0.9293	1.0304	1.0405	1.1530
10	Money	3.0904	**1.0000**	0.9672	1.0167	1.0166	0.9296	0.9232	0.8899	0.8836
11	E. rates	1.9291	**1.0000**	1.0066	0.9712	1.0150	0.9972	1.0719	1.0275	1.0514
12	Stock	1.6840	**1.0000**	1.0162	0.9825	1.0285	0.9769	1.0117	0.9680	1.0009
	Overall	2.7286	**1.0000**	0.9885	1.0343	1.0461	1.0146	1.0109	1.0206	1.0451
					$h = 2$					
1	GDP	2.3626	**1.0000**	0.9599	0.9548	0.9748	0.9446	0.9362	0.9590	0.9561
2	IP	2.5293	**1.0000**	1.0995	1.1481	1.2095	1.1925	1.2340	1.2745	1.3271
3	Emp	2.5516	**1.0000**	0.9960	0.9858	0.9526	0.9582	0.9289	0.9599	0.9492
4	Unemp	2.0818	**1.0000**	0.8734	1.1190	1.1609	1.0681	1.0970	1.0656	1.0838
5	Housing	1.9473	**1.0000**	1.1252	1.0814	1.1105	1.0934	1.0572	1.0997	1.1136
6	Invent	2.6906	**1.0000**	0.9295	0.9587	0.9579	0.9354	0.9854	0.9644	0.9697
7	Prices	2.1724	**1.0000**	1.0431	0.9885	0.9874	0.9639	0.9771	0.9662	0.9475
8	Wages	2.6681	**1.0000**	0.9943	1.0455	1.0444	1.0416	1.0355	0.9926	1.0142
9	I. rates	3.6186	**1.0000**	0.9904	1.0399	1.1747	0.9519	1.1458	1.1911	1.2355
10	Money	2.2142	**1.0000**	1.1142	1.1208	1.1283	1.1646	1.1449	1.0788	1.0721
11	E. rates	2.1627	**1.0000**	1.0292	0.9756	0.9598	0.9456	0.9119	0.9444	0.9377
12	Stock	1.6140	**1.0000**	1.0721	1.0192	1.0724	1.0539	1.1073	1.1265	1.1521
	Overall	2.3844	**1.0000**	1.0189	1.0364	1.0611	1.0261	1.0468	1.0519	1.0632
					$h = 4$					
1	GDP	2.3558	**1.0000**	1.0308	0.9303	1.0087	1.0371	1.0700	1.0158	1.0303
2	IP	2.2719	**1.0000**	1.0243	1.1500	1.1733	1.1481	1.2090	1.2094	1.2473
3	Emp	2.1003	**1.0000**	0.9449	0.9852	0.9818	0.9075	0.9180	0.9285	0.9389
4	Unemp	1.7710	**1.0000**	0.8061	0.9708	1.0607	0.9418	0.9572	0.9316	0.9367
5	Housing	1.4569	**1.0000**	1.0849	1.1060	1.1423	1.1115	1.1000	1.1690	1.1683
6	Invent	1.8882	**1.0000**	0.9750	0.9631	0.9759	0.9955	1.0439	1.0455	0.9915
7	Prices	1.6303	**1.0000**	1.0133	0.9659	0.9598	0.9859	0.9539	1.0011	0.9512
8	Wages	2.9071	**1.0000**	1.0742	1.1160	1.1462	1.1129	1.1589	1.1495	1.1465
9	I. rates	3.6562	**1.0000**	0.9813	1.4182	1.4919	1.4688	1.4920	1.4191	1.4630
10	Money	2.1639	**1.0000**	1.2431	1.0195	1.0223	1.0242	1.0201	1.0570	1.1013
11	E. rates	2.1794	**1.0000**	1.0230	1.0103	1.0083	1.0308	1.0506	1.0470	1.0691
12	Stock	1.8585	**1.0000**	1.1432	1.0469	1.0238	1.0469	1.1221	1.0960	1.0658
	Overall	2.1866	**1.0000**	1.0287	1.0569	1.0829	1.0676	1.0913	1.0891	1.0925

Table B.5: Distribution of the MSE, for $m = 4$, by forecasting methods, relative to the AR(4) for $\boldsymbol{\Gamma} = \boldsymbol{I}$.

	Percentiles				
Method	0.050	0.250	0.500	0.750	0.950
			$h = 1$		
Ridge	1.7952	1.7037	2.0143	1.5287	1.2428
AR(4)	**1.0000**	**1.0000**	**1.0000**	**1.0000**	**1.0000**
DFM5	0.9945	1.0442	0.9810	0.9530	0.9123
SFAR $(1, 2)$	1.2280	1.0471	0.9625	1.0247	0.8974
SFAR $(1, 3)$	1.2024	1.0237	1.0005	1.0684	0.9380
SFAR $(2, 3)$	1.1369	1.0480	1.0109	1.0279	0.8652
SFAR $(2, 4)$	1.1079	1.0783	1.0028	1.0270	0.8972
SFAR $(3, 4)$	1.1762	1.0242	0.9859	1.0023	0.9165
SFAR $(3, 5)$	1.1725	1.0707	1.0568	1.0346	0.9569
			$h = 2$		
Ridge	1.6774	1.5710	1.6933	1.4099	1.1299
AR(4)	**1.0000**	**1.0000**	**1.0000**	**1.0000**	**1.0000**
DFM5	1.0349	1.0663	1.0197	1.0824	0.9789
SFAR $(1, 2)$	1.0990	0.9982	1.0971	1.0746	0.8628
SFAR $(1, 3)$	1.1657	0.9882	1.1358	1.0803	0.9188
SFAR $(2, 3)$	1.1471	1.0067	1.0975	1.0543	0.9284
SFAR $(2, 4)$	1.1225	0.9877	1.1003	1.0345	0.9263
SFAR $(3, 4)$	1.1210	0.9940	1.0738	1.0277	0.9266
SFAR $(3, 5)$	1.0832	0.9861	1.0937	1.0208	0.9789
			$h = 4$		
Ridge	1.4656	1.5572	1.8175	1.3469	1.0691
AR(4)	**1.0000**	**1.0000**	**1.0000**	**1.0000**	**1.0000**
DFM5	1.1204	0.9324	1.0434	0.9356	1.0999
SFAR $(1, 2)$	1.0380	0.9677	1.0638	1.0311	0.9581
SFAR $(1, 3)$	1.0497	0.9491	1.0964	1.1131	0.9901
SFAR $(2, 3)$	1.0974	0.9910	1.0876	0.9982	0.9765
SFAR $(2, 4)$	1.0648	0.9739	1.0818	0.9940	0.9858
SFAR $(3, 4)$	1.0810	0.9862	1.1227	1.0696	1.0188
SFAR $(3, 5)$	1.0124	0.9989	1.1410	1.0259	1.0451

Table B.6: Median results, for $m = 4$, of the MSE by forecasting method and category of series, relative to AR(4) for $\boldsymbol{\Gamma} = \boldsymbol{I}$.

#	Category	Ridge	**AR(4)**	DFM5	SFAR (1, 2)	SFAR (1, 3)	SFAR (2, 3)	SFAR (2, 4)	SFAR (3, 4)	SFAR (3, 5)
					$h = 1$					
1	GDP	1.7377	**1.0000**	0.9848	0.9342	0.9671	0.9498	0.9488	0.9357	0.9467
2	IP	1.7496	**1.0000**	1.2204	1.0571	1.1389	1.1491	1.2385	1.2487	1.4240
3	Emp	2.1357	**1.0000**	1.3872	1.2945	1.3713	1.2525	1.2412	1.3724	1.3319
4	Unemp	1.2283	**1.0000**	0.6795	0.9059	0.9967	0.8721	0.8846	0.8504	0.8474
5	Housing	2.2101	**1.0000**	1.1615	1.0337	1.0202	1.0737	1.0342	1.0343	1.0481
6	Invent	2.3175	**1.0000**	0.9109	0.9097	0.9491	0.8706	0.8923	0.9178	0.9552
7	Prices	1.6643	**1.0000**	1.0623	1.1007	1.0920	1.0855	1.0827	1.0542	1.0475
8	Wages	1.6617	**1.0000**	0.9705	1.0691	1.0918	1.0749	1.0873	1.0703	1.1209
9	I. rates	2.3055	**1.0000**	1.0774	1.0760	1.2159	0.9838	1.1669	1.1210	1.1131
10	Money	1.5589	**1.0000**	1.0259	1.0177	1.1275	0.9286	0.9420	0.9075	0.9230
11	E. rates	1.3597	**1.0000**	1.0084	0.9967	1.0204	1.0092	1.0412	1.0635	1.0972
12	Stock	1.3322	**1.0000**	1.0282	0.9795	0.9794	0.9708	0.9586	0.9851	1.0621
	Overall	1.7717	**1.0000**	1.0431	1.0312	1.0809	1.0184	1.0432	1.0467	1.0764
					$h = 2$					
1	GDP	1.5976	**1.0000**	0.9635	0.9495	0.9863	0.9487	0.9512	0.9282	0.9455
2	IP	1.6805	**1.0000**	1.1765	1.1031	1.1554	1.2069	1.2595	1.3019	1.3587
3	Emp	1.7021	**1.0000**	1.0356	0.9714	0.9871	0.9971	0.9969	1.0377	1.0012
4	Unemp	1.4023	**1.0000**	0.9269	1.1016	1.1925	1.0618	1.0462	1.0763	1.0757
5	Housing	1.8079	**1.0000**	1.0971	1.0732	1.0291	1.0950	1.0670	1.1048	1.0967
6	Invent	1.7152	**1.0000**	0.9101	0.9471	0.9522	0.9084	0.9178	0.9407	0.9859
7	Prices	1.4844	**1.0000**	1.0475	0.9903	0.9926	0.9697	0.9630	0.9459	0.9357
8	Wages	1.8153	**1.0000**	0.9826	1.0487	1.0802	1.0610	1.0532	1.0198	1.0181
9	I. rates	2.2005	**1.0000**	1.0625	0.9865	1.2381	0.9998	1.1539	1.1894	1.1194
10	Money	1.6316	**1.0000**	1.0333	1.1108	1.1383	1.1422	1.1738	1.0913	1.0844
11	E. rates	1.5701	**1.0000**	1.0340	0.9017	0.9680	0.9914	1.0169	1.0023	1.0658
12	Stock	1.3059	**1.0000**	1.0830	1.0413	1.0274	1.0046	1.0387	1.1047	1.1697
	Overall	1.6594	**1.0000**	1.0294	1.0188	1.0623	1.0322	1.0532	1.0619	1.0714
					$h = 4$					
1	GDP	1.6631	**1.0000**	1.0508	0.9216	0.9683	1.0284	1.0501	0.9889	1.0272
2	IP	1.4464	**1.0000**	1.0526	1.1116	1.1386	1.1495	1.1738	1.1655	1.1741
3	Emp	1.3910	**1.0000**	0.9435	0.8900	0.9217	0.9668	0.9975	0.9732	0.9894
4	Unemp	1.2456	**1.0000**	0.7598	0.9849	1.0064	0.8956	0.8811	0.9485	0.9027
5	Housing	1.3936	**1.0000**	1.0623	1.0983	1.1064	1.1063	1.0815	1.1465	1.1121
6	Invent	1.2581	**1.0000**	0.9556	1.0036	0.9688	0.9297	0.9403	0.9469	1.0019
7	Prices	1.1713	**1.0000**	0.9930	0.9677	0.9491	0.9910	0.9739	0.9602	0.9562
8	Wages	1.8579	**1.0000**	1.0403	1.0680	1.1219	1.1047	1.0985	1.0722	1.1174
9	I. rates	2.1362	**1.0000**	1.0297	1.3278	1.5569	1.4843	1.5553	1.5357	1.5598
10	Money	1.3471	**1.0000**	1.1533	1.0184	1.0201	1.0174	1.0163	1.0697	1.1194
11	E. rates	1.8203	**1.0000**	1.0148	1.0053	1.0404	1.0297	1.0684	1.0520	1.0793
12	Stock	1.4942	**1.0000**	1.1161	1.0198	1.0056	0.9901	0.9980	1.0336	1.0544
	Overall	1.5187	**1.0000**	1.0143	1.0348	1.0670	1.0578	1.0696	1.0744	1.0911

APPENDIX C

TABLES: MACRO RESULTS $\mathbf{\Gamma} \neq \boldsymbol{I}$

Table C.1: Implemented combinations of $\mathbf{\Gamma}$ and δ per category of series.

#	Category	$\mathbf{\Gamma}$	δ
1	GDP	$\hat{\boldsymbol{\Sigma}}_{\boldsymbol{YY}}$	0
2	IP	$\boldsymbol{D}_{(\hat{\boldsymbol{\Sigma}}_{\boldsymbol{YY}})}$	0
3	Emp	$\hat{\boldsymbol{\Sigma}}_{\boldsymbol{YY}}$	0
4	Unemp	$\boldsymbol{D}_{(\hat{\boldsymbol{\Sigma}}_{\boldsymbol{YY}})}$	0
5	Housing	$\boldsymbol{D}_{(\hat{\boldsymbol{\Sigma}}_{\boldsymbol{YY}})}$	0.001
6	Invent	$\boldsymbol{D}_{(\hat{\boldsymbol{\Sigma}}_{\boldsymbol{YY}})}$	0
7	Prices	$\hat{\boldsymbol{\Sigma}}_{\boldsymbol{YY}}$	0.001
8	Wages	$\hat{\boldsymbol{\Sigma}}_{\boldsymbol{YY}}$	0.001
9	I. rates	$\hat{\boldsymbol{\Sigma}}_{\boldsymbol{YY}}$	1
10	Money	$\boldsymbol{D}_{(\hat{\boldsymbol{\Sigma}}_{\boldsymbol{YY}})}$	0
11	E. rates	$\hat{\boldsymbol{\Sigma}}_{\boldsymbol{YY}}$	0
12	Stock	$\boldsymbol{D}_{(\hat{\boldsymbol{\Sigma}}_{\boldsymbol{YY}})}$	0

Table C.2: Median results, for $m = 1$, of the MSE by forecasting method and category of series, relative to AR(4) for variations of $\mathbf{\Gamma}$.

#	Category	Ridge	**AR(4)**	DFM5	SFAR (1, 2)	SFAR (1, 3)	SFAR (2, 3)	SFAR (2, 4)	SFAR (3, 4)	SFAR (3, 5)
					$h = 1$					
1	GDP	18.1340	**1.0000**	0.8575	1.0435	1.0315	0.9947	1.0004	1.0103	1.0218
2	IP	21.5913	**1.0000**	1.0633	1.0483	1.0185	1.0501	1.0964	0.9868	0.9622
3	Emp	34.3293	**1.0000**	1.0587	1.1454	1.1449	1.2929	1.2722	1.3105	1.2675
4	Unemp	17.7063	**1.0000**	0.6783	1.0078	1.0130	1.0121	1.0066	0.9006	0.8845
5	Housing	19.6026	**1.0000**	1.0307	0.9781	1.0145	1.0517	0.9810	1.0179	0.9847
6	Invent	53.6031	**1.0000**	1.0319	0.9535	1.0648	1.0903	1.0663	1.0185	1.0987
7	Prices	25.1174	**1.0000**	1.1068	1.0890	1.0651	1.0088	1.0155	1.0458	1.0791
8	Wages	20.4803	**1.0000**	1.0097	1.0219	1.0279	1.0132	1.0111	1.0212	1.0198
9	I. rates	20.9838	**1.0000**	0.9885	0.9839	0.9631	0.9835	1.0060	0.9671	0.9833
10	Money	18.3896	**1.0000**	1.0110	1.0530	0.9558	0.8986	0.9390	1.0322	0.9312
11	E. rates	18.5152	**1.0000**	1.0274	0.9626	0.9661	0.9579	0.9624	0.9666	0.9577
12	Stock	16.7258	**1.0000**	0.9753	0.9948	0.9826	0.9825	0.9729	0.9661	0.9724
	Overall	23.7649	**1.0000**	0.9866	1.0235	1.0206	1.0280	1.0275	1.0203	1.0136
					$h = 2$					
1	GDP	19.7166	**1.0000**	0.9336	1.0049	1.0025	0.9876	0.9919	0.9692	0.9794
2	IP	18.4553	**1.0000**	1.0034	1.0045	0.9480	1.0243	1.0167	0.9664	0.9537
3	Emp	23.6078	**1.0000**	0.9831	1.0034	0.9804	1.0126	1.0191	1.0229	1.0126
4	Unemp	18.6960	**1.0000**	0.8869	1.1155	1.0794	1.0842	1.1116	0.9489	1.0073
5	Housing	12.4436	**1.0000**	1.0436	1.0023	0.9970	1.0465	1.0083	1.0901	1.0529
6	Invent	31.8403	**1.0000**	0.9584	0.9416	0.9185	0.9302	0.9579	0.9821	0.9420
7	Prices	15.4604	**1.0000**	1.0248	0.9832	0.9832	0.9693	0.9542	0.9544	0.9400
8	Wages	20.7056	**1.0000**	1.0244	0.9759	0.9885	0.9740	0.9511	0.9580	0.9528
9	I. rates	20.2604	**1.0000**	1.1361	0.9714	0.9439	0.9822	0.9798	0.9650	0.9771
10	Money	15.0565	**1.0000**	1.1375	0.9746	0.9249	1.0507	1.0054	1.0279	1.0186
11	E. rates	17.7842	**1.0000**	1.0587	0.9673	0.9412	0.9676	0.9655	0.9940	0.9978
12	Stock	17.3238	**1.0000**	1.0273	0.9700	1.0059	0.9875	0.9981	0.9854	0.9947
	Overall	19.2792	**1.0000**	1.0181	0.9929	0.9761	1.0014	0.9967	0.9887	0.9857
					$h = 4$					
1	GDP	16.3232	**1.0000**	1.0043	1.0042	1.0425	0.9933	1.0210	0.9875	0.9885
2	IP	14.9523	**1.0000**	0.9963	0.9551	0.9414	0.9761	0.9400	0.9695	0.9538
3	Emp	17.1151	**1.0000**	0.9520	0.9925	0.9729	1.0119	1.0089	0.9984	0.9624
4	Unemp	15.1694	**1.0000**	0.7952	0.8984	0.9469	0.9267	0.9200	0.9729	0.9676
5	Housing	7.4447	**1.0000**	1.0146	1.0440	1.0756	1.0773	1.0858	1.1052	1.0863
6	Invent	18.0097	**1.0000**	0.9406	0.9763	0.9720	0.9884	0.9953	0.9895	1.0167
7	Prices	14.8364	**1.0000**	1.0419	0.9824	0.9923	0.9362	0.9297	0.9553	0.9406
8	Wages	21.5682	**1.0000**	1.0991	1.0253	1.0357	1.0198	1.0073	1.0386	1.0297
9	I. rates	21.6965	**1.0000**	1.1125	1.0233	0.9782	1.0012	1.0343	1.0134	0.9894
10	Money	18.1060	**1.0000**	1.2713	1.0301	1.0534	1.1322	1.1198	1.1061	1.0641
11	E. rates	16.8720	**1.0000**	1.0541	0.9918	1.0130	0.9834	0.9836	0.9691	0.9743
12	Stock	16.2929	**1.0000**	1.0565	0.9866	0.9904	0.9822	0.9771	1.0005	0.9993
	Overall	16.5322	**1.0000**	1.0282	0.9925	1.0012	1.0024	1.0019	1.0088	0.9977

Table C.3: Median results, for $m = 2$, of the MSE by forecasting method and category of series, relative to AR(4) for variations of $\mathbf{\Gamma}$.

#	Category	Ridge	**AR(4)**	DFM5	SFAR (1, 2)	SFAR (1, 3)	SFAR (2, 3)	SFAR (2, 4)	SFAR (3, 4)	SFAR (3, 5)
					$h = 1$					
1	GDP	2.3108	**1.0000**	0.9108	0.9897	0.9730	0.9696	0.9161	0.9984	1.0070
2	IP	2.6451	**1.0000**	1.0340	1.1432	1.2430	1.2961	1.3429	1.3849	1.4331
3	Emp	3.6446	**1.0000**	1.1681	1.1788	1.2238	1.4623	1.4356	1.4321	1.3705
4	Unemp	1.8865	**1.0000**	0.6620	0.7297	0.6919	0.7129	0.6701	0.7167	0.7144
5	Housing	2.5280	**1.0000**	1.1199	1.0464	1.0266	0.9932	1.0442	1.0585	1.0888
6	Invent	3.9751	**1.0000**	0.9473	0.9715	1.0526	0.9039	0.9828	0.9534	0.9086
7	Prices	2.5697	**1.0000**	1.0636	1.1092	1.0946	1.0992	1.1180	1.1636	1.1261
8	Wages	2.4641	**1.0000**	0.9973	1.0002	1.0165	1.0083	1.0014	1.0631	1.0912
9	I. rates	4.0165	**1.0000**	0.9695	1.0682	1.0673	0.8625	0.9721	0.9528	1.1827
10	Money	3.0904	**1.0000**	0.9672	1.1202	1.1244	0.9069	0.9160	0.9348	0.9330
11	E. rates	1.9288	**1.0000**	1.0066	0.9458	0.9753	0.9606	0.9674	0.9886	0.9879
12	Stock	1.6840	**1.0000**	1.0162	0.9446	0.9136	0.9379	0.9870	0.9522	0.9846
	Overall	2.7286	**1.0000**	0.9885	1.0206	1.0336	1.0094	1.0295	1.0499	1.0690
					$h = 2$					
1	GDP	2.3626	**1.0000**	0.9599	0.9735	0.9847	0.9669	0.9432	0.9466	0.9831
2	IP	2.5293	**1.0000**	1.0995	1.1314	1.2217	1.2813	1.3709	1.3564	1.4078
3	Emp	2.5516	**1.0000**	0.9960	1.0280	1.0446	1.0500	1.0484	1.0224	1.0152
4	Unemp	2.0818	**1.0000**	0.8734	0.9669	0.9327	1.0763	1.0379	1.0058	0.9912
5	Housing	1.9473	**1.0000**	1.1252	1.0675	1.0587	1.0807	1.0880	1.1038	1.1050
6	Invent	2.6906	**1.0000**	0.9295	0.9520	0.9936	0.9162	0.9993	0.9932	0.9934
7	Prices	2.1724	**1.0000**	1.0431	0.9884	0.9925	0.9982	1.0038	1.0972	1.0775
8	Wages	2.6681	**1.0000**	0.9943	1.0358	1.0506	1.0089	1.0091	0.9895	0.9959
9	I. rates	3.6186	**1.0000**	0.9904	1.0196	1.0645	0.9405	1.0539	0.9518	1.1141
10	Money	2.2142	**1.0000**	1.1142	1.0855	1.1184	1.1157	1.1797	1.1102	1.1641
11	E. rates	2.1627	**1.0000**	1.0292	0.9626	0.9995	1.0201	1.0313	0.9542	0.9868
12	Stock	1.6140	**1.0000**	1.0721	1.1073	1.1682	1.0695	1.1010	1.1111	1.1279
	Overall	2.3844	**1.0000**	1.0189	1.0265	1.0525	1.0437	1.0722	1.0535	1.0802
					$h = 4$					
1	GDP	2.3558	**1.0000**	1.0308	0.9652	0.9771	0.9246	0.9747	1.0044	1.0067
2	IP	2.2719	**1.0000**	1.0243	1.0801	1.0334	1.0952	1.1520	1.0835	1.1613
3	Emp	2.1003	**1.0000**	0.9449	1.0008	1.0174	0.9978	1.0117	1.0048	1.0029
4	Unemp	1.7710	**1.0000**	0.8061	0.8866	0.8722	0.9464	0.9227	0.9344	0.8824
5	Housing	1.4569	**1.0000**	1.0849	1.0728	1.1126	1.0909	1.0979	1.1608	1.1468
6	Invent	1.8882	**1.0000**	0.9750	1.0090	0.9459	0.9577	0.9775	0.9496	0.9494
7	Prices	1.6303	**1.0000**	1.0133	0.9695	0.9664	0.9968	0.9979	0.9870	0.9928
8	Wages	2.9071	**1.0000**	1.0742	1.1144	1.1437	1.1068	1.1407	1.1659	1.1877
9	I. rates	3.6562	**1.0000**	0.9813	1.1840	1.3330	1.1528	1.4474	1.3328	1.5371
10	Money	2.1639	**1.0000**	1.2431	1.0132	1.0171	1.0148	1.0158	1.0214	1.1165
11	E. rates	2.1794	**1.0000**	1.0230	1.0036	1.0177	1.0312	1.0317	1.0131	1.0522
12	Stock	1.8585	**1.0000**	1.1432	1.0560	1.0719	1.0112	1.0472	1.0446	1.0361
	Overall	2.1866	**1.0000**	1.0287	1.0296	1.0424	1.0272	1.0681	1.0585	1.0893

Table C.4: Median results, for $m = 4$, of the MSE by forecasting method and category of series, relative to AR(4) for variations of $\mathbf{\Gamma}$.

#	Category	Ridge	**AR(4)**	DFM5	SFAR (1, 2)	SFAR (1, 3)	SFAR (2, 3)	SFAR (2, 4)	SFAR (3, 4)	SFAR (3, 5)
					$h = 1$					
1	GDP	1.7377	**1.0000**	0.9848	1.0044	0.9981	0.9979	0.9895	1.0347	1.0094
2	IP	1.7496	**1.0000**	1.2204	1.1407	1.2442	1.2061	1.2539	1.2415	1.3912
3	Emp	2.1357	**1.0000**	1.3872	1.2460	1.2847	1.3352	1.3360	1.3597	1.4119
4	Unemp	1.2283	**1.0000**	0.6795	0.7491	0.7454	0.7390	0.7599	0.7992	0.7947
5	Housing	2.2101	**1.0000**	1.1615	1.0363	1.0248	1.0253	1.0203	1.0788	1.0527
6	Invent	2.3175	**1.0000**	0.9109	1.0006	1.0115	0.8925	0.8505	0.9878	0.9774
7	Prices	1.6643	**1.0000**	1.0623	1.1007	1.0949	1.0762	1.0878	1.0993	1.1302
8	Wages	1.6617	**1.0000**	0.9705	1.0346	1.0806	1.0756	1.0523	1.0640	1.0707
9	I. rates	2.3055	**1.0000**	1.0774	1.0138	1.0451	1.0342	1.0965	1.0440	1.1566
10	Money	1.5589	**1.0000**	1.0259	1.1206	1.1038	0.9345	0.9708	0.9283	0.9729
11	E. rates	1.3589	**1.0000**	1.0084	0.9727	0.9806	0.9603	0.9534	0.9904	0.9918
12	Stock	1.3322	**1.0000**	1.0282	0.9716	1.0250	1.0972	1.2007	1.1039	1.1016
	Overall	1.7717	**1.0000**	1.0431	1.0326	1.0532	1.0312	1.0476	1.0610	1.0884
					$h = 2$					
1	GDP	1.5976	**1.0000**	0.9635	1.0005	1.0003	1.0039	0.9938	0.9968	1.0022
2	IP	1.6805	**1.0000**	1.1765	1.1333	1.2127	1.2116	1.3071	1.2866	1.3521
3	Emp	1.7021	**1.0000**	1.0356	1.0216	0.9920	1.0106	0.9892	1.0672	1.0719
4	Unemp	1.4023	**1.0000**	0.9269	1.0807	1.1119	1.0918	1.1218	1.0512	1.0777
5	Housing	1.8079	**1.0000**	1.0971	1.0635	1.0458	1.0836	1.0767	1.0960	1.0579
6	Invent	1.7152	**1.0000**	0.9101	0.9986	0.9415	0.8511	0.8971	0.9098	0.9336
7	Prices	1.4844	**1.0000**	1.0475	0.9762	0.9845	1.0213	1.0039	1.0663	1.0852
8	Wages	1.8153	**1.0000**	0.9826	1.0508	1.0751	1.0481	1.0627	0.9900	1.0497
9	I. rates	2.2005	**1.0000**	1.0625	0.9865	0.9862	1.0158	1.1309	1.0565	1.1630
10	Money	1.6316	**1.0000**	1.0333	1.1855	1.1784	1.1336	1.1805	1.0396	1.1069
11	E. rates	1.5701	**1.0000**	1.0340	0.9985	1.0145	0.9842	0.9930	1.0075	1.0096
12	Stock	1.3059	**1.0000**	1.0830	1.0298	1.1082	1.1180	1.2274	1.0492	1.1483
	Overall	1.6594	**1.0000**	1.0294	1.0438	1.0543	1.0478	1.0820	1.0514	1.0882
					$h = 4$					
1	GDP	1.6631	**1.0000**	1.0508	1.0254	1.0326	0.9346	0.9445	0.9984	1.0555
2	IP	1.4464	**1.0000**	1.0526	0.9810	0.9719	1.0094	1.0125	1.0283	1.0424
3	Emp	1.3910	**1.0000**	0.9435	0.9677	0.9727	0.9701	0.9763	0.9591	0.9916
4	Unemp	1.2456	**1.0000**	0.7598	0.9366	0.9276	0.8952	0.8948	0.9327	0.9065
5	Housing	1.3936	**1.0000**	1.0623	1.0746	1.0843	1.1062	1.0737	1.1190	1.0667
6	Invent	1.2581	**1.0000**	0.9556	0.9665	0.9306	0.9057	0.9338	0.8984	0.9412
7	Prices	1.1713	**1.0000**	0.9930	0.9732	0.9495	0.9791	0.9683	0.9793	0.9821
8	Wages	1.8579	**1.0000**	1.0403	1.0673	1.0847	1.0604	1.0778	1.1035	1.1439
9	I. rates	2.1362	**1.0000**	1.0297	1.0989	1.3038	1.1232	1.4945	1.4109	1.5459
10	Money	1.3471	**1.0000**	1.1533	1.0212	1.0439	1.0175	1.0206	1.0530	1.1249
11	E. rates	1.8203	**1.0000**	1.0148	0.9981	0.9858	0.9949	1.0114	1.0224	1.0316
12	Stock	1.4942	**1.0000**	1.1161	1.0090	1.1232	1.0592	1.1521	1.0947	1.1455
	Overall	1.5187	**1.0000**	1.0143	1.0100	1.0342	1.0046	1.0467	1.0500	1.0815

APPENDIX D

PROOFS

Proof of Theorem 3. Given any matrix $\boldsymbol{A}$, we use $CS(\boldsymbol{A})$ and $RS(\boldsymbol{A})$ to denote its column space and row space, respectively. Denote by $\mathbf{P}_{\boldsymbol{A}}$ the orthogonal projection matrix onto $CS(\boldsymbol{A})$, i.e., $\mathbf{P}_{\boldsymbol{A}} = \boldsymbol{A}(\boldsymbol{A}^T\boldsymbol{A})^+\boldsymbol{A}^T$, where $^+$ stands for the Moore-Penrose pseudoinverse, and $\mathbf{P}_{\boldsymbol{A}}^{\perp}$ the projection onto its orthogonal complement. For notation simplicity assume $\boldsymbol{X}_{t-h} = \boldsymbol{X}$. Given any index set $\mathcal{J} \subset [p]$, $\mathbf{P}_{\mathcal{J}}$ is short for $\mathbf{P}_{\boldsymbol{X}_{\mathcal{J}}}$ when there is no ambiguity. In the theorem, the noise matrix has sub-Gaussian marginal tails.

Definition D.0.1. *ξ is called a sub-Gaussian random variable if there exist constants $C, c > 0$ such that $\mathbb{P}\{|\xi| \geq t\} \leq Ce^{-ct^2}, \forall t > 0$. The scale ($\psi_2$-norm) for ξ is defined as $\sigma(\xi) = \inf\{\sigma > 0 : \mathbb{E}\exp(\xi^2/\sigma^2) \leq 2\}$. $\boldsymbol{\xi} \in \mathbb{R}^p$ is called a sub-Gaussian random vector with scale bounded by σ if all one-dimensional marginals $\langle \boldsymbol{\xi}, \boldsymbol{a}\rangle$ are sub-Gaussian satisfying $\|\langle \boldsymbol{\xi}, \boldsymbol{a}\rangle\|_{\psi_2} \leq \sigma\|\boldsymbol{a}\|_2$, $\forall \boldsymbol{a} \in R^p$.*

Sub-Gaussian examples include Gaussian random variables and bounded random variables such as Bernoulli. In the proof, we assume vec $(\boldsymbol{E})$ is sub-Gaussian. Note that the entries of $\boldsymbol{E}$ may not be iid.

From the construction of $\hat{\boldsymbol{C}}$, we have

$$\frac{1}{2}\|\boldsymbol{X}\hat{\boldsymbol{C}} - \boldsymbol{X}\boldsymbol{C}^*\|_F^2 \leq \frac{1}{2}\|\boldsymbol{X}\boldsymbol{C} - \boldsymbol{X}\boldsymbol{C}^*\|_F^2 + P_2(\boldsymbol{C}) - P_2(\hat{\boldsymbol{C}}) + \langle \boldsymbol{E}, \boldsymbol{X}\hat{\boldsymbol{C}} - \boldsymbol{X}\boldsymbol{C}\rangle \tag{D.0.1}$$

where $P_2(\boldsymbol{C})$ stands for $\frac{\eta}{2}\|\boldsymbol{C}\|_F^2$. Define $\mathcal{J}(\boldsymbol{C}) = \{j : \boldsymbol{c}_j \neq \boldsymbol{0}\}$, $J(\boldsymbol{C}) = |\mathcal{J}(\boldsymbol{C})| = \|\boldsymbol{C}\|_{2,0}$. Let $\boldsymbol{\Delta} = \hat{\boldsymbol{C}} - \boldsymbol{C}$, $\hat{\mathcal{J}} = \mathcal{J}(\hat{\boldsymbol{C}})$, $\mathcal{J} = \mathcal{J}(\boldsymbol{C})$, $J = J(\boldsymbol{C})$, $\hat{J} = J(\hat{\boldsymbol{C}})$, $r = r(\boldsymbol{C})$, $\hat{r} = r(\hat{\boldsymbol{C}})$. Let $\mathbf{P}_{rs}$ and $\mathbf{P}_{\mathcal{J}}$ be the orthogonal projections onto the row space of $\boldsymbol{X}_{\mathcal{J}}\boldsymbol{C}_{\mathcal{J}}$ and the column space of $\boldsymbol{X}_{\mathcal{J}}$, respectively. Decompose $\boldsymbol{X}\boldsymbol{\Delta}$ as follows

$$\begin{aligned}
\boldsymbol{X}\boldsymbol{\Delta} =& \boldsymbol{X}\boldsymbol{\Delta}\mathbf{P}_{rs} + \boldsymbol{X}\boldsymbol{\Delta}\mathbf{P}_{rs}^{\perp} \\
=& \mathbf{P}_{\mathcal{J}}\boldsymbol{X}\boldsymbol{\Delta}\mathbf{P}_{rs} + \mathbf{P}_{\mathcal{J}}^{\perp}\boldsymbol{X}\boldsymbol{\Delta}\mathbf{P}_{rs} + \boldsymbol{X}_{\hat{\mathcal{J}}}\hat{\boldsymbol{C}}_{\hat{\mathcal{J}}}\mathbf{P}_{rs}^{\perp} \\
=& \mathbf{P}_{\mathcal{J}}\boldsymbol{X}\boldsymbol{\Delta}\mathbf{P}_{rs} + \mathbf{P}_{\mathcal{J}}^{\perp}\boldsymbol{X}_{\hat{\mathcal{J}}}\hat{\boldsymbol{C}}_{\hat{\mathcal{J}}}\mathbf{P}_{rs} + \boldsymbol{X}_{\hat{\mathcal{J}}}\hat{\boldsymbol{C}}_{\hat{\mathcal{J}}}\mathbf{P}_{rs}^{\perp}.
\end{aligned}$$

Clearly, $\|\boldsymbol{X}\boldsymbol{\Delta}\|_F^2 = \|\mathbf{P}_{\mathcal{J}}\boldsymbol{X}\boldsymbol{\Delta}\mathbf{P}_{rs}\|_F^2 + \|\mathbf{P}_{\mathcal{J}}^{\perp}\boldsymbol{X}_{\hat{\mathcal{J}}}\hat{\boldsymbol{C}}_{\hat{\mathcal{J}}}\mathbf{P}_{rs}\|_F^2 + \|\boldsymbol{X}_{\hat{\mathcal{J}}}\hat{\boldsymbol{C}}_{\hat{\mathcal{J}}}\mathbf{P}_{rs}^{\perp}\|_F^2$. The noise term in (D.0.1) is now

$$\begin{aligned}\langle \boldsymbol{E}, \boldsymbol{X}\boldsymbol{\Delta}\rangle =& \langle \boldsymbol{E}, \mathbf{P}_{\mathcal{J}}\boldsymbol{X}\boldsymbol{\Delta}\mathbf{P}_{rs}\rangle + \langle \boldsymbol{E}, \mathbf{P}_{\mathcal{J}}^{\perp}\boldsymbol{X}_{\hat{\mathcal{J}}}\hat{\boldsymbol{C}}_{\hat{\mathcal{J}}}\mathbf{P}_{rs}\rangle + \langle \boldsymbol{E}, \boldsymbol{X}_{\hat{\mathcal{J}}}\hat{\boldsymbol{C}}_{\hat{\mathcal{J}}}\mathbf{P}_{rs}^{\perp}\rangle \\ \equiv & I + II + III. \end{aligned} \tag{D.0.2}$$

To bound I and III, we introduce Lemma 1. Define $P_o(\boldsymbol{C}) = \sigma^2\{(q \wedge J(\boldsymbol{C}) + n - r(\boldsymbol{C}))r(\boldsymbol{C}) + J(\boldsymbol{C})\log(ep/J(\boldsymbol{C}))\}$ with $q = \mathrm{rank}(\boldsymbol{X})$; for convenience, we also denote it by $P_o(J(\boldsymbol{C}), r(\boldsymbol{C}))$.

Lemma 1. *Suppose vec$(\boldsymbol{E})$ is sub-Gaussian with mean zero and ψ_2-norm bounded by σ. Given $\boldsymbol{X} \in \mathbb{R}^{T\times p}$, $1 \le J \le p$, $1 \le r \le J \wedge n$, define $\Gamma_{J,r} = \{\boldsymbol{A} \in \mathbb{R}^{T\times n} : \|\boldsymbol{A}\|_F \le 1, rank(\boldsymbol{A}) \le r, CS(\boldsymbol{A}) \subset CS(\boldsymbol{X}_{\mathcal{J}})$ for some $\mathcal{J} : |\mathcal{J}| = J\}$. Let $P_o'(J,r) = \sigma^2\{(q \wedge J)r + (n-r)r + \log\binom{p}{J}\}$. Then for any $t \ge 0$,*

$$\mathbb{P}\left(\sup_{\boldsymbol{A}\in\Gamma_{J,r}} \langle \boldsymbol{E}, \boldsymbol{A}\rangle \ge t\sigma + \sqrt{L \cdot P_o'(J,r)}\right) \le C\exp(-ct^2), \tag{D.0.3}$$

where $L, C, c > 0$ are universal constants.

Proof. By Definition D.0.1, for any fixed $\boldsymbol{A}$, $\langle \boldsymbol{E}, \boldsymbol{A}\rangle$ is a mean-centered sub-Gaussian random variable with scale bounded by $\sigma\|\boldsymbol{A}\|_F$. Therefore, $\{\langle \boldsymbol{E}, \boldsymbol{A}\rangle : \boldsymbol{A} \in \Gamma_{J,r}\}$ is a stochastic process with sub-Gaussian increments; Dudley's entropy integral can be used to bound its supremum, see, e.g., [67]. The induced metric on $\Gamma_{J,r}$ is Euclidean: $d(\boldsymbol{A}_1, \boldsymbol{A}_2) = \sigma\|\boldsymbol{A}_1 - \boldsymbol{A}_2\|_F$.

To compute the metric entropy $\log\mathcal{N}(\varepsilon, \Gamma_{J,r}, d)$, where $\mathcal{N}(\varepsilon, \Gamma_{J,r}, d)$ is the smallest cardinality of an ε-net that covers $\Gamma_{J,r}$ under d, we characterize each matrix in $\Gamma_{J,r}$ using its row/column spaces, motivated by [45]. Given $\boldsymbol{A} \in \Gamma_{J,r}$, its column space must be contained in $CS(\boldsymbol{X}_{\mathcal{J}})$ for some $\mathcal{J}$ with $|\mathcal{J}| = J$; its row space must be contained in an r-dimensional subspace in $\mathbb{R}^n$. Hence

$$\boldsymbol{A} = \boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^T, \tag{D.0.4}$$

where $\boldsymbol{\Sigma} \in \mathbb{R}^{(J\wedge q)\times r}$ and $\mathbf{P}_{\boldsymbol{U}} = \mathbf{P}_{\boldsymbol{X}_{\mathcal{J}}}$. $\boldsymbol{\Sigma}$ is in a $(J \wedge q) \times r$-dimensional unit ball (denoted by $B_{(J\wedge q)\times r}$).

The number of candidate $\boldsymbol{X}_{\mathcal{J}}$ is $\binom{p}{J}$. By a standard volume argument,

$$\mathcal{N}(\varepsilon, B_{(J\wedge q)\times r}, d') \le (C_0/\varepsilon)^{(J\wedge q)\times r},$$

where d' is the Euclidean distance in $\mathbb{R}^{(J\wedge q)\times r}$ and C_0 is a universal constant. Note that $RS(\boldsymbol{A})$ is a point on the Grassmann manifold (denoted by $G_{n,r}$) of all r-dimensional subspaces of $\mathbb{R}^n$. Equipped with the metric d'' given by the operator norm $\|\boldsymbol{V}_1\boldsymbol{V}_1^T-\boldsymbol{V}_2\boldsymbol{V}_2^T\|_2$ for any $\boldsymbol{V}_1,\boldsymbol{V}_2\in\mathbb{O}^{n\times r}$, we have $\mathcal{N}(\varepsilon,G_{n,r},d'')\le(C_1/\varepsilon)^{r(n-r)}$, where C_1 is a universal constant [66]. We claim that

$$\mathcal{N}(\varepsilon,\Gamma_{J,r},d)\le\binom{p}{J}\left(\frac{C\sigma}{\varepsilon}\right)^{(J\wedge q)\times r+r(n-r)}. \tag{D.0.5}$$

In fact, given any $\boldsymbol{A}_1\in\Gamma_{J,r}$, with $\boldsymbol{A}_1=\boldsymbol{U}_1\boldsymbol{\Sigma}_1\boldsymbol{V}_1^T$ according to (D.0.4), find $\boldsymbol{V}_2$ and $\boldsymbol{\Sigma}_2$ such that $\|\boldsymbol{V}_1\boldsymbol{V}_1^T-\boldsymbol{V}_2\boldsymbol{V}_2^T\|_2\le\varepsilon$ and $\|\boldsymbol{\Sigma}_1\boldsymbol{V}_1^T\boldsymbol{V}_2-\boldsymbol{\Sigma}_2\|_F\le\varepsilon$, then, for $\boldsymbol{A}_2=\boldsymbol{U}_1\boldsymbol{\Sigma}_2\boldsymbol{V}_2^T$,

$$\begin{aligned}\|\boldsymbol{A}_1-\boldsymbol{A}_2\|_F&\le\|\boldsymbol{A}_1-\boldsymbol{A}_1\boldsymbol{V}_2\boldsymbol{V}_2^T\|_F+\|\boldsymbol{U}_1\boldsymbol{\Sigma}_1\boldsymbol{V}_1^T\boldsymbol{V}_2\boldsymbol{V}_2^T-\boldsymbol{U}_1\boldsymbol{\Sigma}_2\boldsymbol{V}_2^T\|_F\\&\le(Tr\{\boldsymbol{A}_1^T\boldsymbol{A}_1(\mathbf{P}_{\boldsymbol{V}_1}-\mathbf{P}_{\boldsymbol{V}_2})^2\})^{1/2}+\|\boldsymbol{\Sigma}_1\boldsymbol{V}_1^T\boldsymbol{V}_2-\boldsymbol{\Sigma}_2\|_F\\&\le(\|\boldsymbol{A}_1\|_F^2\|\mathbf{P}_{\boldsymbol{V}_1}-\mathbf{P}_{\boldsymbol{V}_2}\|_2^2)^{1/2}+\varepsilon\le 2\varepsilon.\end{aligned}$$

Using Dudley's integral bound, we obtain

$$\mathbb{P}\left(\sup_{\boldsymbol{A}\in\Gamma_{J,r}}\langle\boldsymbol{E},\boldsymbol{A}\rangle\ge t\sigma+L\int_0^\sigma\sqrt{\log\mathcal{N}(\varepsilon,\Gamma_{J,r},d)}\,\mathrm{d}\varepsilon\right)\le C\exp(-ct^2).$$

Simple computation yields

$$\begin{aligned}\int_0^\sigma\sqrt{\log\mathcal{N}(\varepsilon,\Gamma_{J,r},d)}\,\mathrm{d}\varepsilon)&\lesssim\sigma\sqrt{\log\binom{p}{J}}+\sigma\sqrt{(J\wedge q)\times r+r(n-r)}\\&\lesssim\sqrt{P_o'(J,r)}\end{aligned}$$

due to the Cauchy-Schwarz inequality. □

Terms I and III on the right hand side of (D.0.2) can handled in a similar manner. For instance,

$$\begin{aligned}&\langle\boldsymbol{E},\mathbf{P}_{\mathcal{J}}\boldsymbol{X}\boldsymbol{\Delta}\mathbf{P}_{rs}\rangle-\frac{1}{a}\|\mathbf{P}_{\mathcal{J}}\boldsymbol{X}\boldsymbol{\Delta}\mathbf{P}_{rs}\|_F^2-bLP_o(J,r)\\\le&\|\mathbf{P}_{\mathcal{J}}\boldsymbol{X}\boldsymbol{\Delta}\mathbf{P}_{rs}\|_F\langle\boldsymbol{E},\mathbf{P}_{\mathcal{J}}\boldsymbol{X}\boldsymbol{\Delta}\mathbf{P}_{rs}/\|\mathbf{P}_{\mathcal{J}}\boldsymbol{X}\boldsymbol{\Delta}\mathbf{P}_{rs}\|_F\rangle-2\sqrt{\frac{b}{a}}\|\mathbf{P}_{\mathcal{J}}\boldsymbol{X}\boldsymbol{\Delta}\mathbf{P}_{rs}\|_F\sqrt{LP_o(J,r)}\\\le&\frac{1}{a'}\|\mathbf{P}_{\mathcal{J}}\boldsymbol{X}\boldsymbol{\Delta}\mathbf{P}_{rs}\|_F^2+\frac{a'}{4}\sup_{1\le J\le p,1\le r\le n\wedge J}\left(\sup_{\boldsymbol{A}\in\Gamma_{J,r}}\langle\boldsymbol{E},\boldsymbol{A}\rangle-2\sqrt{\frac{b}{a}}\sqrt{LP_o(J,r)}\right)_+^2\\\equiv&\frac{1}{a'}\|\mathbf{P}_{\mathcal{J}}\boldsymbol{X}\boldsymbol{\Delta}\mathbf{P}_{rs}\|_F^2+\frac{a'}{4}\sup_{1\le J\le p,1\le r\le n\wedge J}R_{J,r}^2\\\equiv&\frac{1}{a'}\|\mathbf{P}_{\mathcal{J}}\boldsymbol{X}\boldsymbol{\Delta}\mathbf{P}_{rs}\|_F^2+\frac{a'}{4}R^2,\end{aligned}$$

for any $a, b, a' > 0$. Lemma 1 provides a control of R^2:

$$\begin{aligned}
&\mathbb{P}(R \geq t\sigma) \\
\leq &\sum_{J=1}^{p} \sum_{r=1}^{n \wedge J} \mathbb{P}(R_{J,r} \geq t\sigma) \\
\leq &\sum_{J=1}^{p} \sum_{r=1}^{n \wedge J} \mathbb{P}(\sup_{\boldsymbol{A} \in \Gamma_{J,r}} \langle \boldsymbol{E}, \boldsymbol{A} \rangle - \sqrt{LP_o(J,r)} \geq t\sigma + (2\sqrt{b/a} - 1)\sqrt{LP_o(J,r)}) \\
\leq &\sum_{J=1}^{p} \sum_{r=1}^{n \wedge J} C \exp(-ct^2) \exp\left\{-c\left((2\sqrt{b/a} - 1)^2 L \cdot P_o(J,r)/\sigma^2\right)\right\} \\
\leq &C' \exp(-ct^2),
\end{aligned}$$

from which it follows that $\mathbb{E}R^2 \leq C\sigma^2$.

Similarly, we introduce the following lemma to deal with *II*.

Lemma 2. *Suppose vec*$(\boldsymbol{E})$ *is sub-Gaussian with mean zero and* ψ_2*-norm bounded by* σ*. Given* $\boldsymbol{X} \in \mathbb{R}^{T \times p}$*,* $1 \leq J, J' \leq p$*,* $1 \leq r \leq J \wedge n$*, define* $\Gamma_{J',J,r} = \{\boldsymbol{A} \in \mathbb{R}^{T \times n} : \|\boldsymbol{A}\|_F \leq 1, rank(\boldsymbol{A}) \leq r, CS(\boldsymbol{A}) \subset CS(\mathbf{P}_{\mathcal{J}'}^{\perp} \mathbf{P}_{\mathcal{J}})$ *for some* $\mathcal{J}', \mathcal{J} \subset [p]$ *satisfying* $|\mathcal{J}'| = J', |\mathcal{J}| = J\}$*. Let* $P_o''(J', J, r) = \sigma^2\{(q \wedge J \wedge (p - J'))r + (n - r)r + \log\binom{p}{J} + \log\binom{p}{J'}\}$*. Then for any* $t \geq 0$*,*

$$\mathbb{P}\left(\sup_{\boldsymbol{A} \in \Gamma_{J',J,r}} \langle \boldsymbol{E}, \boldsymbol{A} \rangle \geq t\sigma + \sqrt{L \cdot P_o''(J', J, r)}\right) \leq C \exp(-ct^2), \tag{D.0.6}$$

where $L, C, c > 0$ *are universal constants.*

Proof. The proof is similar to that of Lemma 1 based on the entropy integral bound. The rate P_o'' can be obtained by calculating the degrees-of-freedom and inflation bounds. The details are omitted. □

Similar to the treatment of I, applying Lemma 2 gives (noticing that $P_o''(J, \hat{J}, \hat{r}) \leq P_o(J, r) + P_0(\hat{J}, \hat{r})$):

$$\begin{aligned}
&\langle \boldsymbol{E}, \mathbf{P}_{\hat{\mathcal{J}}}^{\perp} \boldsymbol{X}_{\hat{\mathcal{J}}} \hat{\boldsymbol{C}}_{\hat{\mathcal{J}}} \mathbf{P}_{rs} \rangle - \frac{1}{a} \|\mathbf{P}_{\hat{\mathcal{J}}}^{\perp} \boldsymbol{X}_{\hat{\mathcal{J}}} \hat{\boldsymbol{C}}_{\hat{\mathcal{J}}} \mathbf{P}_{rs}\|_F^2 - bL\{P_o(J,r) + P_0(\hat{J}, \hat{r})\} \\
\leq &\frac{1}{a'} \|\mathbf{P}_{\mathcal{J}} \boldsymbol{X} \boldsymbol{\Delta} \mathbf{P}_{rs}\|_F^2 + \frac{a'}{4} R'^2,
\end{aligned}$$

with $\mathbb{E}R'^2 \leq C\sigma^2$.

In summary, for any $a, b, a', b' > 0$, we have

$$\begin{aligned}
&\langle \boldsymbol{E}, \boldsymbol{X}\boldsymbol{\Delta}\rangle \\
\leq&(\frac{1}{a}+\frac{1}{a'})\|\boldsymbol{X}\boldsymbol{\Delta}\|_F^2 + bL\{P_o(J,r)+P_o(J,r)+P_0(\hat{J},\hat{r})+P_o(\hat{J},\hat{r})\} \\
&+\frac{a'}{4}(2R^2+R'^2) \\
\leq&(\frac{1}{a}+\frac{1}{a'})(1+b')\|\boldsymbol{X}\boldsymbol{C}-\boldsymbol{X}\boldsymbol{C}^*\|_F^2 + (\frac{1}{a}+\frac{1}{a'})(1+\frac{1}{b'})\|\boldsymbol{X}\hat{\boldsymbol{C}}-\boldsymbol{X}\boldsymbol{C}^*\|_F^2 \\
&+2bL\{P_o(J,r)+P_o(\hat{J},\hat{r})\}+\frac{a'}{4}(2R^2+R'^2).
\end{aligned}$$

Choosing $(\frac{1}{a}+\frac{1}{a'})(1+\frac{1}{b'}) < \frac{1}{2}$, $4b > a$, and noticing that $J \leq \mathfrak{q}, \hat{J} \leq \mathfrak{q}, \hat{r} \leq r$, we obtain the oracle inequality as desired.

Lemma 3. *Given any $\boldsymbol{Y}_t \in \mathbb{R}^{T\times n}$, $\hat{\boldsymbol{C}} = \vec{\Theta}^{\#}(\boldsymbol{Y}_t; \mathfrak{q}, \eta)$ is a globally optimal solution to*

$$\min_{\boldsymbol{C}} \frac{1}{2}\|\boldsymbol{Y}_t - \boldsymbol{C}\|_2^2 + \frac{\eta}{2}\|\boldsymbol{C}\|_2^2 =: f(\boldsymbol{C};\eta) \quad \text{s.t. } J(\boldsymbol{C}) \leq \mathfrak{q}. \tag{D.0.7}$$

Let $\mathcal{J} \subset [p]$ with $|\mathcal{J}| = \mathfrak{q}$. Assuming $\boldsymbol{C}[\mathcal{J}^c,] = \boldsymbol{0}$, we get the optimal solution $\hat{\boldsymbol{C}}$ with $\hat{\boldsymbol{C}}[\mathcal{J},] = \boldsymbol{Y}_t[\mathcal{J},]/(1+\eta)$. It follows that $f(\hat{\boldsymbol{C}};\eta) = \frac{1}{2}\|\boldsymbol{Y}_t\|_F^2 + \frac{1}{2(1+\eta)}\|\boldsymbol{Y}_t[\mathcal{J}^c,]\|_F^2$. Hence the group quantile thresholding $\vec{\Theta}^{\#}(\boldsymbol{Y}_t; \mathfrak{q}, \eta)$ yields a global minimizer.

It remains to study the minimization of $g(\boldsymbol{S}) = \|\boldsymbol{Y}_t' - \boldsymbol{X}\boldsymbol{S}\|_F^2/(2K) + \eta\|\boldsymbol{S}\|_F^2/2$ s.t. $\|\boldsymbol{S}\|_{2,0} \leq \mathfrak{q}$, where $\boldsymbol{Y}_t' = \boldsymbol{Y}_t\boldsymbol{V}$. Construct a surrogate function $G(\boldsymbol{S}, \tilde{\boldsymbol{S}}) = g(\tilde{\boldsymbol{S}}) + \langle(\boldsymbol{I} - \boldsymbol{X}^T\boldsymbol{X}/K)(\boldsymbol{S}-\tilde{\boldsymbol{S}}), \boldsymbol{S}-\tilde{\boldsymbol{S}}\rangle$. The rest of the proof follows the same lines as, She [54]. The details are omitted.

REFERENCES

[1] Sung K. Ahn and Gregory C. Reinsel. Nested reduced-rank autogressive models for multiple time series. *Journal of the American Statistical Association*, 83(403):pp. 849–856, 1988.

[2] Sung K. Ahn and Gregory C. Reinsel. Estimation for partially nonstationary multivariate autoregressive models. *Journal of the American Statistical Association*, 85(411):pp. 813–823, 1990.

[3] Htrotugu Akaike. Maximum likelihood identification of gaussian autoregressive moving average models. *Biometrika*, 60(2):255–265, 1973.

[4] Htrotugu Akaike. A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, 19:716–723, 1974.

[5] T. W. Anderson. Estimating linear restrictions on regression coefficients for multivariate normal distributions. *Annals of Mathematical Statistics*, 22:327–351, 1951.

[6] M. Bańbura, D. Giannone, and L. Reichlin. Large bayesian vector auto regressions. *Journal of Applied Econometrics*, 25(1):71–92, 2010.

[7] Leo Breiman. Heuristics of instability and stabilization in model selection. *The Annals of Statistics*, 24(6):2350–2383, December 1996.

[8] D.R. Brillinger. *Time Series: Data Analysis and Theory*. Classics in Applied Mathematics. Society for Industrial and Applied Mathematics, 2001.

[9] F. Bunea, Y. She, and M. Wegkamp. Optimal selection of reduced rank estimators of high-dimensional matrices. *ArXiv e-prints*, 2010.

[10] Florentina Bunea, Yiyuan She, and Marten Wegkamp. Joint variable and rank selection for parsimonious estimation of high dimensional matrices. *Annals of Statistics*, 2012. to appear.

[11] P. T. Davies and M. K-S. Tso. Procedures for reduced-rank regression. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, 31(3):pp. 244–255, 1982.

[12] David L. Donoho. Aide-memoire. high-dimensional data analysis: The curses and blessings of dimensionality, 2000.

[13] David L. Donoho, Iain M. Johnstone, Gerard Kerkyacharian, and Dominique Picard. Wavelet shrinkage: Asymptopia? *Journal of the Royal Statistical Society. Series B (Methodological)*, 57(2):pp. 301–369, 1995.

[14] Bradley Efron, Trevor Hastie, Iain Johnstone, and Robert Tibshirani. Least angle regression. *Annals of Statistics*, 32(2):407–499, 2004.

[15] J. Fan and R. Li. Variable selection via nonconcave penalized likelihood and its oracle properties. *J. Amer. Statist. Assoc.*, 96:1348–1360, 2001.

[16] Jianqing Fan and Runze Li. Statistical challenges with high dimensionality: feature selection in knowledge discovery. In *International Congress of Mathematicans*, Aug. 2006.

[17] Jianqing Fan and Jinchi Lv. Sure independence screening for ultrahigh dimensional feature space. *Journal Of The Royal Statistical Society Series B*, 70(5):849–911, 2008.

[18] D. F. Findley. On the Use of Multiple Models for Multi-Period Forecasting. *Proceedings of the Business and Statistics Section, American Statistical Association*, pages 528–531, 1983.

[19] D. F Findley. Model selection for multi-step-ahead forecasting. *Proceedings of the 7th Symposium on Identification and System Parameter*, pages 1039–1044, 1985.

[20] Ildiko E. Frank and Jerome H. Friedman. A statistical view of some chemometrics regression tools. *Technometrics*, 35(2):pp. 109–135, 1993.

[21] J. Geweke. The dynamic factor analysis of economic time series. *D.J. Aigner and A.S. Goldberger, eds., Latent Variables in Socio-Economic Models, (North-Holland, Amsterdam)*, 1977.

[22] C. Giraud. Low rank multivariate regression. *2010arXiv1009.5165G*, 2010.

[23] T.J. Hastie, R.J. Tibshirani, and J.J.H. Friedman. *The Elements of Statistical Learning*. Springer series in statistics. Springer-Verlag New York, 2009.

[24] Yuejia He, Yiyuan She, and Dapeng Wu. Stationary-sparse causality network learning. *Journal of Machine Learning Research*, 14:3073–3104, 2013.

[25] A.E. Hoerl and R. Kennard. Ridge regression: biased estimation for nonorthogonal problems. *Technometrics*, 12:55–67, 1970.

[26] H. Hotelling. Analysis of a complex of statistical variables into principal components. *journal of Educational Psychology*, 24:417–441, 498–520, 1933.

[27] A.J. Izenman. Reduced-rank regression for the multivariate linear model. *journal of Multivariate Analysis*, 5:248–262, 1975.

[28] A.J. Izenman. *Modern Multivariate. Statistical Techniques: Regression, Classification and Manifold Learning*. Springer, New York, 2008.

[29] Iain M Johnstone and Arthur Yu Lu. Sparse principal components analysis. *arXiv preprint arXiv:0901.4392*, 2004.

[30] In-Bong Kang. Multi-period forecasting using different models for different horizons: an application to us economic time series data. *International Journal of Forecasting*, 19(3):387–400, 2003.

[31] J. Knight and S. Satchell. *Linear Factor Models in Finance*. Quantitative Finance. Elsevier Science, 2004.

[32] Keith Knight and Wenjiang Fu. Asymptotics for lasso-type estimators. *Annals of Statistics*, pages 1356–1378, 2000.

[33] Chenlei Leng, Yi Lin, and Grace Wahba. A note on the Lasso and related procedures in model selection. *Statistica Sinica*, 16(4):1273–1284, 2006.

[34] Jin-Lung Lin and C. W. J. Granger. Forecasting from non-linear models in practice. *Journal of Forecasting*, 13(1):1–9, 1994.

[35] Marco Lippi and Daniel L. Thornton. A dynamic factor analysis of the response of u.s. interest rates to news. LEM Papers Series 2004/05, Laboratory of Economics and Management (LEM), Sant'Anna School of Advanced Studies, Pisa, Italy, March 2004.

[36] Sydney C. Ludvigson and Serena Ng. The empirical risk-return relation: A factor analysis approach. *Journal of Financial Economics*, 83(1):171–222, January 2007.

[37] Helmut Lütkepohl. *New introduction to multiple time series analysis*. Springer, 2nd edition, 2007.

[38] C. L. Mallows. Some comments on Cp. *Technometrics*, 15:661–675, 1973.

[39] Massimiliano Marcellino, James H Stock, and Mark W Watson. A comparison of direct and iterated multistep ar methods for forecasting macroeconomic time series. *Journal of Econometrics*, 135(1):499–526, 2006.

[40] Rahul Mazumder, Jerome H. Friedman, and Trevor Hastie. Sparsenet: Coordinate descent with nonconvex penalties. *Journal of the American Statistical Association*, 106(495):1125–1138, 2011.

[41] S. Negahban and M. J. Wainwright. Estimation of (near) low-rank matrices with noise and high-dimensional scaling. *2009arXiv0912.5100N*, 2009.

[42] Donna K. Pauler. The Schwarz criterion and related methods for normal linear models. *Biometrika*, 85(1):13–27, March 1998.

[43] C. R. Rao. Matrix approximations and reduction of dimensionality in multivariate statistical analysis. *Journal of Multivariate Analysis*, 5:3–22, 1980.

[44] C. Radhakrishna Rao. Separation theorems for singular values of matrices and their applications in multivariate analysis. *Journal of Multivariate Analysis*, 9(3):362–377, September 1979.

[45] Benjamin Recht, Maryam Fazel, and Pablo A. Parrilo. Guaranteed minimum-rank solutions of linear matrix equations via nuclear norm minimization. *SIAM Rev.*, 52(3):471–501, 2010.

[46] G.C. Reinsel and R.P. Velu. *Multivariate Reduced-Rank Regression: Theory and Applications*. Springer, New York, 1998.

[47] Gregory Reinsel. Some results on multivariate autoregressive index models. *Biometrika*, 70(1):pp. 145–156, 1983.

[48] P. M. Robinson. Generalized canonical analysis for time series. *Journal of Multivariate Analysis*, 3:141–160, 1973.

[49] P. M. Robinson. Identification, estimation and large-sample theory for regressions containing unobservable variables. *International Economic Review*, 15(3):pp. 680–692, 1974.

[50] T.J. Sargent and C.A. Sims. Business cycle modeling without pretending to have too much a-priori economic theory. *New Methods in Business Cycle Research, ed. by C. Sims et al., Minneapolis: Federal Reserve Bank of Minneapolis*, 1977.

[51] G. Schwarz. Estimating the dimension of a model. *The Annals of Statistics*, 6:461–464, 1978.

[52] Yiyuan She. Thresholding-based iterative selection procedures for model selection and shrinkage. *Electronic journal of Statistics*, 3:384–415, 2009.

[53] Yiyuan She. An iterative algorithm for fitting nonconvex penalized generalized linear models with grouped predictors. *Computational Statistics and Data Analysis*, 9:2976–2990, 2012.

[54] Yiyuan She. Selectable factor extraction in high dimensions. *journal = 2014arXiv:1403.6212 [stat.ME]*, 2014.

[55] Xiaotong Shen and Jianming Ye. Adaptive model selection. *Journal of the American Statistical Association*, 97(457):210–221, 2002.

[56] Christopher A. Sims. Macroeconomics and reality. *Econometrica*, 48(1):pp. 1–48, 1980.

[57] James Stock and Mark W. Watson. *A Probability Model of the Coincident Economic Indicators*, pages 63–90. Cambridge University Press, 1991.

[58] James H. Stock and Mark W. Watson. New indexes of coincident and leading economic indicators. In *NBER Macroeconomics Annual 1989, Volume 4*, NBER Chapters, pages 351–409. National Bureau of Economic Research, Inc, Jan-Jun 1989.

[59] James H. Stock and Mark W. Watson. Forecasting Using Principal Components From a Large Number of Predictors. *Journal of the American Statistical Association*, 97(460):1167–1179, December 2002.

[60] James H Stock and Mark W Watson. An empirical comparison of methods for forecasting using many predictors. *Manuscript, Princeton University*, 2005.

[61] James H. Stock and Mark W. Watson. Implications of dynamic factor models for var analysis. NBER Working Papers 11467, National Bureau of Economic Research, Inc, July 2005.

[62] James H. Stock and Mark W. Watson. Generalized shrinkage methods for forecasting using many predictors. *Journal of Business & Economic Statistics*, 30(4):481–493, October 2012.

[63] J.M. Stock and M.W. Watson. Vector autoregressions. *JEP*, 15(4):101–115, 2001.

[64] J.M. Stock and M.W. Watson. Macroeconomic forecasting using diffusion indexes. *JBES*, 2(2):147–162, 2002.

[65] J.M. Stock and M.W. Watson. Dynamic factor models. *Oxford Handbook of Forecasting, Michael P. Clements and David F. Hendry (eds), 2011, Oxford: Oxford University Press*, 2010.

[66] S Szarek. Nets of grassmann manifold and orthogonal groups. In *Proceedings of Banach Spaces Workshop*, pages 169–185. University of Iowa Press, 1982.

[67] M. Talagrand. *The Generic Chaining: Upper and Lower Bounds of Stochastic Processes*. Springer Monographs in Mathematics. Springer, 2005.

[68] George C Tiao and Ruey S Tsay. Some advances in non-linear and adaptive modelling in time-series. *Journal of forecasting*, 13(2):109–131, 1994.

[69] R. Tibshirani. Regression shrinkage and selection via the lasso. *JRSSB*, 58:267–288, 1996.

[70] Robert Tibshirani and Keith Knight. The covariance inflation criterion for adaptive model selection. *J. Roy. Statist. Soc. B*, 55:757–796, 1999.

[71] M. K.-S. Tso. Reduced-rank regression and canonical analysis. *Journal of the Royal Statistical Society. Series B (Methodological)*, 43(2):pp. 183–189, 1981.

[72] Raja P. Velu, Gregory C. Reinsel, and Dean W. Wichern. Reduced rank models for multiple time series. *Biometrika*, 73(1):105–118, 1986.

[73] M Yuan and Y. Lin. Model selection and estimation in regression with grouped variables. *JRSSB*, 68:49–67, 2006.

[74] Ming Yuan, Ali Ekici, Zhaosong Lu, and Renato Monteiro. Dimension reduction and coefficient estimation in multivariate linear regression. *JRBBS*, 69(3):329–346, 2007.

[75] Cun-Hui Zhang and Jian Huang. The sparsity and bias of the lasso selection in high-dimensional linear regression. *The Annals of Statistics*, 36(4):pp. 1567–1594, 2008.

[76] Peng Zhao and Bin Yu. On model selection consistency of lasso. *Journal of Machine Learning Research*, 7(2):2541, 2007.

[77] X. Zheng and W.Y. Loh. A consistent variable selection criterion for linear models with high-dimensional covariates. *Statistica Sinica*, 7:311–326, 1997.

[78] H. Zou and T. Hastie. Regularization and variable selection via the elastic net. *JRSSB*, 67(2):301–320, 2005.

[79] Hui Zou and Hao Helen Zhang. On the adaptive elastic-net with a diverging number of parameters. *The Annals of Statistics*, 37(4):pp. 1733–1751, 2009.

BIOGRAPHICAL SKETCH

Oliver Kurt Galvis Balbás was born in Puerto Ordaz, Venezuela to Wolfgang Kenneth Galvis and Flor Elena Balbás. He is one of 4 siblings: Wolfgang René, Florelena, and Fabiola. In October 2002, he received his Bachelor degree in Education with major in Physics and Math from UCAB in Caracas, Venezuela. Four years later, he earned his MBA in Marketing from IESA Business School in Caracas, Venezuela. After spending 3 years in the industry, he moved to Tallahassee, FL and enrolled in the Master of Science in Statistics program in the Department of Statistics at Florida State University, which was completed in 2012. Afterwards, he continued through the Doctorate program in Statistics. He defended his dissertation on July 7th, 2014. His areas of interest are high dimensional data analysis, dimension reduction techniques, segmentation, marketing analytics, classification methods, conjoint analysis, data mining, and time series and forecasting.