Parameter Estimation on Dynamic Factor Augmented Regression Model

Wenxin (Ellie) Zou

Advisor: Professor Danna Zhang

Department of Mathematics

University of California San Diego

In partial fulfillment of the requirements

for the Mathematics Honors Program

August 2023

Acknowledgement

I would like to express my deepest gratitude to Professor Danna Zhang for her mentorship and guidance throughout my entire thesis journey. She has a tremendous capacity for developing students' skills, and the paper would not have been possible without having countless meetings and discussions with her regarding the research direction and other details of the paper. Additionally, I would like to give my special thanks to Dehao Dai, Professor Zhang's Ph.D. student, for sharing his insights and guiding me through every stage of the process. It is my great honor to have the opportunity to work with them, and it has been a remarkable learning experience that has greatly enriched my academic journey. Lastly, I want to thank my family and friends for supporting me and encouraging me all the time.

Contents

Abstract

In this paper, we propose a factor-augmented regression model with dependent noise, which combines dimension reduction and regression analysis. Existing research mainly focuses on independent noise but overlook the natural dependence structure in real applications. To this end, our model bridges the gap by relaxing the condition and introducing more practical dependence structures and moment assumptions. In particular, we use a regularization technique to address the high-dimensional regression issues and establish the estimation consistency. Furthermore, we conduct a simulation study concerning different dependencies in the noise to validate the convergence rate of our estimators and apply our proposed approach to the real U.S. macroeconomic dataset for its practical efficacy in capturing complex dynamics.

Chapter 1

Introduction

1.1 Background

In the contemporary landscape of data-driven research and analysis, the proliferation of high-dimensional data has emerged in various industries such as finance, medical imaging, and astronomy. However, traditional analysis has limitations in terms of dealing with high-dimensional data due to the complexity of matrix decomposition. In response, factor modeling has gained prominence as a viable alternative for addressing the challenges posed by high-dimensional data. Stock and Watson [1998] and Stock and Watson [2002b] as groundbreaking and pioneer work of factor model first introduce a method to extract and analyze information from a large number of economic time series data to estimate the state of the economy and predict business cycle fluctuations. Specifically, factor model is in a form $\boldsymbol{X} = \boldsymbol{F}\boldsymbol{B}^\top + \boldsymbol{U}$ where $\boldsymbol{F}$ is factor, $\boldsymbol{B}$ is loading matrix, and $\boldsymbol{U}$ is noise. By capturing common features, namely factors, the factor model lets $\boldsymbol{X}$ be decomposed and reveals latent factors that make complex high-dimensional data more interpretable. Alternative methods such as Principal Component Regression (PCR) and Ridge Regression (RIDGE) have a similar goal to achieve dimensionality reduction for high-dimensional data.

The factor model finds wide applications across diverse domains, notably emerging as a crucial component within the financial arena. In this context, historical endeavors have often aimed to identify an exhaustive set of features capable of comprehensively measuring overall economic activity. Nevertheless, these features are correlated, which leads us to the Fama-French three-factor model Fama and French [2004] that has been widely used in asset pricing analysis. Building upon the traditional Capital Asset Pricing Model (CAPM) Jagannathan et al. [1995], the Fama-French model extends the explanatory power of asset returns by incorporating three additional factors that capture additional sources of risk and return: (1) SMB represents the outperformance of small versus big companies, which accounts for the size of firms, (2) HML stands for the outperformance of high book/market versus low book/market companies, and (3) the third factor $r - r_f$ is the difference between the expected return of the market and the risk-free rate, which measures the excess return on the market. Thus, the Fama-French three-factor model, more generally, the factor model, effectively reduces the computation cost and makes high-dimensional data more approachable.

While the factor model succeeds in seizing common factors, it falls short in explaining how they act in the response variable, or the regression model. As a result, the Factor Augmented Regression Model (FARM) is an extension of the traditional factor model [Bai and Ng, 2002, Fan et al., 2011]. As is introduced in Fan et al. [2023], FARM incorporates both the latent factor and the idiosyncratic component into the covariates, and it is in a form

$$\boldsymbol{Y} = \boldsymbol{F}\boldsymbol{\gamma} + \boldsymbol{U}\boldsymbol{\beta} + \boldsymbol{e},$$

where $\boldsymbol{F}$ is the latent factor, $\boldsymbol{U}$ is the idiosyncratic component and $\varepsilon \in \mathbb{R}$ is the random noise that is independent of $\boldsymbol{F}$ and $\boldsymbol{U}$. Nevertheless, high-dimensional data is often sparse and only has a few active elements, so it presents a challenge to FARM since regression can lead to overfitting by incorporating inactive elements, consequently yielding inaccurate results. Thus, regularization techniques such as RIDGE and Elastic Net are needed to employ, and in our paper, we primarily focus on LASSO.

1.2 Contribution

LASSO stands for Least Absolute Shrinkage and Selection Operator, which extends the linear regression model by introducing an additional l_1 penalty term based on the absolute values of the coefficients, and it is in a form

$$Q(\boldsymbol{\beta}, \lambda) = Q(\boldsymbol{\beta}) + \lambda|\boldsymbol{\beta}|_1 = \sum_{i=1}^{n}(y_i - x_i^\top \boldsymbol{\beta})^2 + \lambda|\boldsymbol{\beta}|_1,$$

where β is the coefficient and λ is the regularization parameter that controls the level of regularization applied. The effectiveness of FARM with regularization is confirmed by Stock and Watson's work, in which they used a U.S. macroeconomic dataset and demonstrated how a massive amount of variables could be reduced to just a few. Therefore, compared to the regression model, LASSO achieves a balance between model simplicity and accuracy while also promoting sparse models with fewer parameters Huang et al. [2008]. In addition, we want to emphasize our assumptions. Many scholars have often employed stringent assumptions in factor models or FARM, assuming that the noise is mutually independent. However, inspired by Breitung and Tenhofen [2011] which expands the factor model and introduces the correlation in the idiosyncratic component, we introduce a milder condition for the noise in FARM.

Specifically, we consider the noise to be dependent and follow an autoregressive (AR) process. Furthermore, we assume the error to be i.i.d and independent to $\boldsymbol{F}$ and $\boldsymbol{U}$.

1.3 Structure

The paper is organized as follows. Chapter 2 illustrates the regular assumptions and corresponding properties in the Factor Model, introduces our dependence measure on the noise, and establishes the LASSO estimators' statistical properties of the Dynamic Factor Augmented Regression model. Chapter 3 presents the simulation results using the Dynamic Factor regression model with different dependencies $\phi = 0.1, 0.9$. In addition, we work on a real U.S. macroeconomic dataset to evaluate our model in Chapter 4, and we also discuss the data background and potential reasons for the ups and downs in the graphs. Lastly, the conclusion and discussions are in Chapter 5.

1.4 Notation

In this section, we want to introduce notations that will be consistently employed throughout this paper. For any vector $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_p)^\top \in \mathbb{R}^p$, $|\boldsymbol{\mu}|_2 = (\sum_{i=1}^p \mu_i^2)^{\frac{1}{2}}$, $|\boldsymbol{\mu}|_\infty = \max_i |\mu_i|$. Denote $\lambda_j(\boldsymbol{A})$ as the j-th largest eigenvalue of a nonnegative definitive matrix $\boldsymbol{A}$, $|\boldsymbol{A}|_2$ be the spectral norm of a matrix $\boldsymbol{A}$, and $|\boldsymbol{A}|_\mathbb{F}$ be the Frobenius norm of $\boldsymbol{A}$. For a random variable X, denote $\|X\|_p = (\mathbb{E}|X|^p)^{1/p}$. In addition, we let $[m] = \{1, \ldots, m\}$ for $m \in \mathbb{Z}$. Let $\|Z\|_{\psi_2} = \inf\{t > 0 : \mathbb{E}\exp(Z^2/t^2) \leqslant 2\}$ be the sub-Gaussian norm of a scalar random variable Z and $\|\boldsymbol{Z}\|_{\psi_2} = \sup_{|\boldsymbol{x}|_2=1} \|\boldsymbol{Z}\boldsymbol{x}\|_{\psi_2}$ be the sub-Gaussian norm of a random vector $\boldsymbol{Z}$. Denote $\mathbb{I}\{\cdot\}$ and $\boldsymbol{I}_K$ as the indicator function and the identity matrix in $\mathbb{R}^{K\times K}$, respectively. For a matrix $\boldsymbol{A} = [A_{jk}]$, we define $\|\boldsymbol{A}\|_\mathbb{F} = \sqrt{\sum_{jk} A_{jk}^2}$ as its Frobenius norm, and $\|\boldsymbol{A}\|_{\max} = \max_{jk} |A_{jk}|$ and $\|\boldsymbol{A}\|_\infty = \max_j \sum_k |A_{jk}|$ are its element-wise max-norm and matrix ℓ_∞-norm, respectively. In addition, denote $\lambda_{\min}(\boldsymbol{A})$ and $\lambda_{\max}(\boldsymbol{A})$ to be the minimal and maximal eigenvalues of $\boldsymbol{A}$, respectively. $|\mathcal{A}|$ is the cardinality of set $\mathcal{A}$. For $\{a_n\}_{n\geqslant 1}, \{b_n\}_{n\geqslant 1}$ to be two positive sequences, we denote $a_n = O(b_n)$ if there exists a positive constant C such that $a_n \leqslant C \cdot b_n$ and we write $a_n = o(b_n)$ if $a_n/b_n \to 0$. Similarly, the notations $a_n = O_\mathbb{P}(b_n)$ and $a_n = o_\mathbb{P}(b_n)$ remain the same as previously mentioned, besides the relationship of a_n/b_n holds with high probability.

Chapter 2

Dynamic Factor Augmented Regression Model

This section introduces a regularized estimation method for the factor-augmented sparse linear model and delivers the statistical properties. In general, suppose that we observe n independent and identically distributed (i.i.d.) random samples $\{(\boldsymbol{x}_t, Y_t)\}_{t=1}^n$ from $(\boldsymbol{x}, Y)$, which satisfy that

$$\boldsymbol{x}_t = \boldsymbol{B}\boldsymbol{f}_t + \boldsymbol{u}_t \quad \text{and} \quad Y_t = \boldsymbol{f}_t^\top \boldsymbol{\gamma}^\star + \boldsymbol{u}_t^\top \boldsymbol{\beta}^\star + e_t, \quad t = 1, \ldots, n, \tag{2.1}$$

where $\boldsymbol{f}_1, \ldots, \boldsymbol{f}_n \in \mathbb{R}^K$, $\boldsymbol{u}_1, \ldots, \boldsymbol{u}_n \in \mathbb{R}^d$ are i.i.d. realizations of $\boldsymbol{f}$, $\boldsymbol{u}$, respectively. In our framework, we can extend the original i.i.d. condition for $\boldsymbol{e}$ to follow a wide class of dependent structure. In addition, we can rewrite (2.1) in a more compact matrix form as follows,

$$\begin{aligned} \boldsymbol{X} &= \boldsymbol{F}\boldsymbol{B}^\top + \boldsymbol{U}, \\ \boldsymbol{Y} &= \boldsymbol{F}\boldsymbol{\gamma}^\star + \boldsymbol{U}\boldsymbol{\beta}^\star + \boldsymbol{e}, \end{aligned} \tag{2.2}$$

where $\boldsymbol{X} = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n)^\top$, $\boldsymbol{F} = (\boldsymbol{f}_1, \ldots, \boldsymbol{f}_n)^\top$, $\boldsymbol{U} = (\boldsymbol{u}_1, \ldots, \boldsymbol{u}_n)^\top$, $\boldsymbol{Y} = (Y_1, \ldots, Y_n)^\top$ and $\boldsymbol{e} = (e_1, \ldots, e_n)^\top$. Throughout the whole paper, we assume we only get access to observations $\{(\boldsymbol{x}_t, Y_t)\}_{t=1}^n$. Both the latent factors $\boldsymbol{F}$ and the idiosyncratic components $\boldsymbol{U}$ are unobserved and need to be estimated from the observed predictors $\boldsymbol{X}$. Thus, we shall first introduce the method of estimating $\boldsymbol{F}$ and $\boldsymbol{U}$, then establish the theoretical properties.

2.1 Factor Estimation

Suppose we observe n independent and identically distributed (i.i.d.) random samples $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n \in \mathbb{R}^d$ from the factor model

$$\boldsymbol{x}_t = \boldsymbol{B}\boldsymbol{f}_t + \boldsymbol{u}_t, \tag{2.3}$$

where $\boldsymbol{f}_1, \ldots, \boldsymbol{f}_n$ and $\boldsymbol{u}_1, \ldots, \boldsymbol{u}_n$ are i.i.d. realizations of $\boldsymbol{f}$ and $\boldsymbol{u}$, respectively. Recall that the latent variables $(\boldsymbol{f}_t, \boldsymbol{u}_t)$ are not observed under the factor model (2.3) and only the predictor variable $\boldsymbol{x}$ is observable. More specifically, for any non-singular matrix $\boldsymbol{S} \in \mathbb{R}^{K\times K}$, we have $\boldsymbol{x} = \boldsymbol{B}\boldsymbol{f} + \boldsymbol{u} = (\boldsymbol{B}\boldsymbol{S})(\boldsymbol{S}^{-1}\boldsymbol{f}) + \boldsymbol{u}$. To resolve this issue, we impose the following conditions [Bai, 2003, Fan et al., 2013]:

$$\mathrm{Cov}(\boldsymbol{f}) = \boldsymbol{I}_K \quad \text{and} \quad \boldsymbol{B}^\top \boldsymbol{B} \quad \text{is diagonal.}$$

Consequently, the constrained least squares estimator of $(\boldsymbol{F}, \boldsymbol{B})$ based on $\boldsymbol{X}$ is given by

$$\begin{aligned} (\hat{\boldsymbol{F}}, \hat{\boldsymbol{B}}) = &\underset{\boldsymbol{F}\in\mathbb{R}^{n\times K}, \boldsymbol{B}\in\mathbb{R}^{d\times K}}{\arg\min} \sum_{i=1}^{d}\sum_{t=1}^{n}(x_{it} - \boldsymbol{b}_i^\top \boldsymbol{f}_t)^2 \\ &\text{subject to} \quad n^{-1}\boldsymbol{F}^\top\boldsymbol{F} = \boldsymbol{I}_K \quad \text{and} \quad \boldsymbol{B}^\top\boldsymbol{B} \quad \text{is diagonal.} \end{aligned}$$

The columns of $\hat{\boldsymbol{F}}/\sqrt{n}$ are the eigenvectors corresponding to the largest K eigenvalues of the matrix $\boldsymbol{X}\boldsymbol{X}^\top$ and $\hat{\boldsymbol{B}}^\top = (\hat{\boldsymbol{F}}^\top\hat{\boldsymbol{F}})^{-1}\hat{\boldsymbol{F}}^\top\boldsymbol{X} = n^{-1}\hat{\boldsymbol{F}}^\top\boldsymbol{X}$. And the least squares estimator for $\boldsymbol{U}$ is given by $\hat{\boldsymbol{U}} = \boldsymbol{X} - \hat{\boldsymbol{F}}\hat{\boldsymbol{B}}^\top = (\boldsymbol{I}_n - n^{-1}\hat{\boldsymbol{F}}\hat{\boldsymbol{F}}^\top)\boldsymbol{X}$.

Now we first introduce some regularity conditions following from Fan et al. [2023].

Assumption 2.1 (Factors)**.** *There exists a positive constant* $c_0 < \infty$ *such that* $\|\boldsymbol{f}\|_{\psi_2} \le c_0$.

Assumption 2.2 (Factor Loadings)**.** *There exists a constant* $c_0 > 1$ *such that* $d/c_0 \le \lambda_{\min}(\boldsymbol{B}^\top\boldsymbol{B}) \le \lambda_{\max}(\boldsymbol{B}^\top\boldsymbol{B}) \le dc_0$ *and* $|\boldsymbol{B}|_{\max} \le c_0$.

Assumption 2.3 (Idiosyncratic Error)**.**

1. *There exists a positive constant* $c_1 < \infty$ *such that* $\|\boldsymbol{u}\|_{\psi_2} \le c_1$. *If let* $\boldsymbol{\Sigma} = \mathrm{Cov}(\boldsymbol{u})$, *then* $\mathbb{E}|\boldsymbol{u}^\top\boldsymbol{u} - \mathrm{tr}(\boldsymbol{\Sigma})|^4 \le c_1 d^2$.
2. *There exist a positive constant* $c_2 < 1$ *such that* $c_2 \le \lambda_{\min}(\boldsymbol{\Sigma})$, $|\boldsymbol{\Sigma}|_1 \le 1/c_2$ *and* $\min_{1\le k,\ell\le d}\mathrm{Var}(u_k u_\ell) \ge c_2$.

Remark 2.1. Assumptions 2.1–2.3 are standard assumptions in the studies of large dimensional factor models. We refer to Bai [2003], Fan et al. [2013] and Fan et al. [2023] for more details.

Next, we provide the theoretical results related to consistent factor estimation in the following proposition which directly follows from Proposition 2.1 in Fan et al. [2023].

Theorem 2.1 (Proposition 2.1 in [Fan et al., 2023])**.** *Assume that* $\log n = o(d)$. *Let* $\boldsymbol{H} = n^{-1}\boldsymbol{V}^{-1}\hat{\boldsymbol{F}}^\top\boldsymbol{F}\boldsymbol{B}^\top\boldsymbol{B}$, *where* $\boldsymbol{V} \in \mathbb{R}^{K\times K}$ *is a diagonal matrix consisting of the first* K *largest eigenvalues of the matrix* $n^{-1}\boldsymbol{X}\boldsymbol{X}^\top$. *Then, under Assumptions 2.1-2.3, we have*

1. $|\hat{\boldsymbol{F}} - \boldsymbol{F}\boldsymbol{H}^\top|_{\mathbb{F}}^2 = O_{\mathbb{P}}(n/d + 1/n)$.

2. *For any* $\mathcal{I} \subset \{1, 2, \dots, d\}$, *we have*

$$\max_{\ell \in \mathcal{I}} \sum_{t=1}^{n} |\hat{u}_{t\ell} - u_{t\ell}|^2 = O_{\mathbb{P}}(\log |\mathcal{I}| + n/d).$$

3. $|\boldsymbol{H}^\top \boldsymbol{H} - \boldsymbol{I}_K|_{\mathbb{F}}^2 = O_{\mathbb{P}}(1/n + 1/d)$.

4. $\max_{\ell \in [d]} |\hat{\boldsymbol{b}}_\ell - \boldsymbol{H}\boldsymbol{b}_\ell|_2^2 = O_{\mathbb{P}}\{(\log d)/n\}$.

Remark 2.2 (Consistency of K)**.** In practice, the number of latent factors K is typically unknown and it is an important issue to determine K. There have been various methods proposed in the literature to estimate the number K [Ahn and Horenstein, 2013, Bai and Ng, 2002, Fan et al., 2022, Lam and Yao, 2012]. Our theories always work as long as we replace K by any consistent estimator $\hat{K}$, i.e. we only require

$$\mathbb{P}(\hat{K} = K) \to 1, \text{ as } n \to \infty.$$

2.2 Estimation of Regression Parameters

The high dimension time series where the dimension d can be much larger than the sample size n implies that only a few predictors could be contributed and the true parameter vector can be assumed as a sparse vector. Then the estimator for the unknown parameter vectors $\boldsymbol{\beta}^\star$ and $\boldsymbol{\gamma}^\star$ of our factor augmented linear model is defined as follows:

$$(\hat{\boldsymbol{\beta}}_\lambda, \hat{\boldsymbol{\gamma}}) = \underset{\boldsymbol{\beta} \in \mathbb{R}^d, \boldsymbol{\gamma} \in \mathbb{R}^K}{\arg\min} \left\{ \frac{1}{2n} |\boldsymbol{Y} - \hat{\boldsymbol{U}}\boldsymbol{\beta} - \hat{\boldsymbol{F}}\boldsymbol{\gamma}|_2^2 + \lambda |\boldsymbol{\beta}|_1 \right\}, \tag{2.4}$$

where $\lambda > 0$ is a tuning parameter. It is hard to directly get the solution of $\boldsymbol{\gamma}$ first. Therefore, we need to find an equivalent formula of the loss function (2.4). Projecting onto the column space of $\hat{\boldsymbol{F}}$, we can get the residuals of the response vector $\boldsymbol{Y}$ given by

$$\tilde{\boldsymbol{Y}} = (\boldsymbol{I}_n - \hat{\boldsymbol{P}})\boldsymbol{Y},$$

where $\hat{\boldsymbol{P}} = n^{-1}\hat{\boldsymbol{F}}\hat{\boldsymbol{F}}^\top$ is the corresponding projection matrix. Recall that $\hat{\boldsymbol{U}} = (\boldsymbol{I}_n - \hat{\boldsymbol{P}})\boldsymbol{X}$ implies $\hat{\boldsymbol{F}}$ are perpendicular to $\hat{\boldsymbol{U}}$, i.e. $\hat{\boldsymbol{F}}^\top \hat{\boldsymbol{U}} = 0$. Hence, the solution of (2.4) is equivalent to

$$\hat{\boldsymbol{\beta}}_\lambda = \underset{\boldsymbol{\beta} \in \mathbb{R}^d}{\arg\min} \left\{ \frac{1}{2n} |\tilde{\boldsymbol{Y}} - \hat{\boldsymbol{U}}\boldsymbol{\beta}|_2^2 + \lambda |\boldsymbol{\beta}|_1 \right\},$$
$$\hat{\boldsymbol{\gamma}} = (\hat{\boldsymbol{F}}^\top \hat{\boldsymbol{F}})^{-1} \hat{\boldsymbol{F}}^\top \boldsymbol{Y} = n^{-1} \hat{\boldsymbol{F}}^\top \boldsymbol{Y}.$$

In the next subsection, we will discuss the LASSO estimator.

2.2.1 LASSO estimator

LASSO, namely Least Absolute Shrinkage, and Selection Operator, serves as a regularization technique within linear regression in the high dimensional scenario. Its primary objective is to introduce a regular assumption of parameter sparsity to the model by adding a penalty term to the loss function of the Ordinary Least Squares (OLS) method. Suppose that we have n covariates with d-dimension x_{ij} and n corresponding responses y_i. If we consider an intercept β_0 in the linear model [Huang et al., 2008] given by

$$y_i = \beta_0 + \sum_{j=1}^{d} x_{ij}\beta_j + \varepsilon_i, \quad i = 1, 2, \ldots, n.$$

The estimator $\hat{\beta}$lasso can be represented as follows:

$$\hat{\beta}^{\text{lasso}} = \underset{\beta}{\text{argmin}} \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{d} x_{ij}\beta_j \right)^2$$
$$\text{subject to } \sum_{j=1}^{d} |\beta_j| \leq t.$$

When we consider the concept of constrained optimization, the LASSO estimate is also equivalent to **Lagrangian form** [Zou, 2006]

$$\hat{\beta}^{\text{lasso}} = \underset{\beta}{\text{argmin}} \left\{ \frac{1}{2} \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{d} x_{ij}\beta_j \right)^2 + \lambda \sum_{j=1}^{d} |\beta_j| \right\},$$

Here, we can denote $|\boldsymbol{\beta}|_1 := \sum_{j=1}^{d} |\beta_j|$ and $|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}|_2 = \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{d} x_{ij}\beta_j \right)^2$. Therefore, it can be regarded as a constrained optimization problem to find the optimal solution. Furthermore, LASSO aims to control the absolute size of the coefficients β_j. Specifically, λ measures the connection between LASSO and the Lagrangian form, and when λ is small, the constraint is loose, enabling more coefficients to maintain non-zero terms, and vice versa.

Note that it is vital to choose appropriate t since if t is sufficiently small, it will lead to some coefficients being exactly 0 and thereby achieve covariates selection. The standard tuning parameter $s = t / \sum_{j=1}^{p} |\hat{\beta}_j|$. In general, we can use k-fold cross-validation to choose a suitable λ. When applying 10-fold cross-validation, a value of $\hat{s} \approx 0.36$, for example, results in the four coefficients approach 0 [Hastie et al., 2009]. Hence, LASSO is a powerful regularization technique that balances predictive accuracy and variable selection.

Then, we discuss and compare the approaches for the regression model: Ridge regression and LASSO. Suppose the input matrix $\boldsymbol{X}$ is an orthonormal matrix, the two procedures have explicit solutions. Ridge regression is also a regularization regression model with l_2 penalty, i.e.

$$\hat{\beta}^{\text{ridge}} = \underset{\beta}{\operatorname{argmin}} \left\{ \frac{1}{2} \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{d} x_{ij}\beta_j \right)^2 + \lambda \sum_{j=1}^{d} |\beta_j|^2 \right\},$$

Each method applies a simple transformation to the least squares estimate β_j, as detailed in Table 2.1 .

Estimator	Formula
LASSO	$\text{sgn}(\hat{\beta}_j)(\lvert\hat{\beta}_j\rvert - \lambda)_+$
Ridge	$\hat{\beta}_j/(1+\lambda)$

Table 2.1: Estimators of β_j in the case of orthonormal columns of $\boldsymbol{X}$. λ is the constant chosen by the corresponding techniques; sgn denotes the sign of its argument (± 1), and x_+ denotes "positive part" of x.

Next, we use a figure to illustrate their relationship. Figure 2.1 depicts the LASSO (left) and Ridge regression (right) when there are only two parameters (i.e. two dimensions in the figure). The residual sum of squares can be regarded as a series of elliptical contours, centered at the least squares estimator. The constraint region for Ridge regression is the disk $|\boldsymbol{\beta}|_2^2 = \beta_1^2 + \beta_2^2 \le t^2$, while that for lasso is the diamond $|\boldsymbol{\beta}|_1 = |\beta_1| + |\beta_2| \le t$. Both methods find the first point where the elliptical contours reach the constraint region. The disk can have usual solutions, while the diamond has corners; if the solution occurs at a corner, then it has one parameter β_j equal to zero. When $d > 2$, there are many more opportunities for the estimated parameters to be zero (see Tibshirani [1996]).

2.2.2 Dependence Measure

Traditionally, Fan et al. [2023] considered the i.i.d realization of noise $\boldsymbol{e}$ which has finite sub-Gaussian norm. Therefore, we introduce a mild condition for the random noise. In general, some studies adopted the mixing conditions such as the α-mixing in the literature like Fan et al. [2013]. They consider $\boldsymbol{f}_t$ and $\boldsymbol{u}_t$ as a stationary time series with zero mean. Let $\mathcal{F}_{-\infty}^{0}$ and $\mathcal{F}_{T}^{\infty}$ denote the σ-algebras generated by $\{(\boldsymbol{f}_t, \boldsymbol{u}_t) : t \le 0\}$ and $\{(\boldsymbol{f}_t, \boldsymbol{u}_t) : t \ge T\}$ respectively. They define the mixing coefficient

$$\alpha(T) = \sup_{A \in \mathcal{F}_{-\infty}^{0}, B \in \mathcal{F}_{T}^{\infty}} |\mathbb{P}(A)\mathbb{P}(B) - \mathbb{P}(A \cap B)|$$

In our paper, to make the dependence measure easier to implement, we will introduce the framework in Wu and Wu [2016]. Let ε_t, $t \in \mathbb{Z}$ be i.i.d random variables and

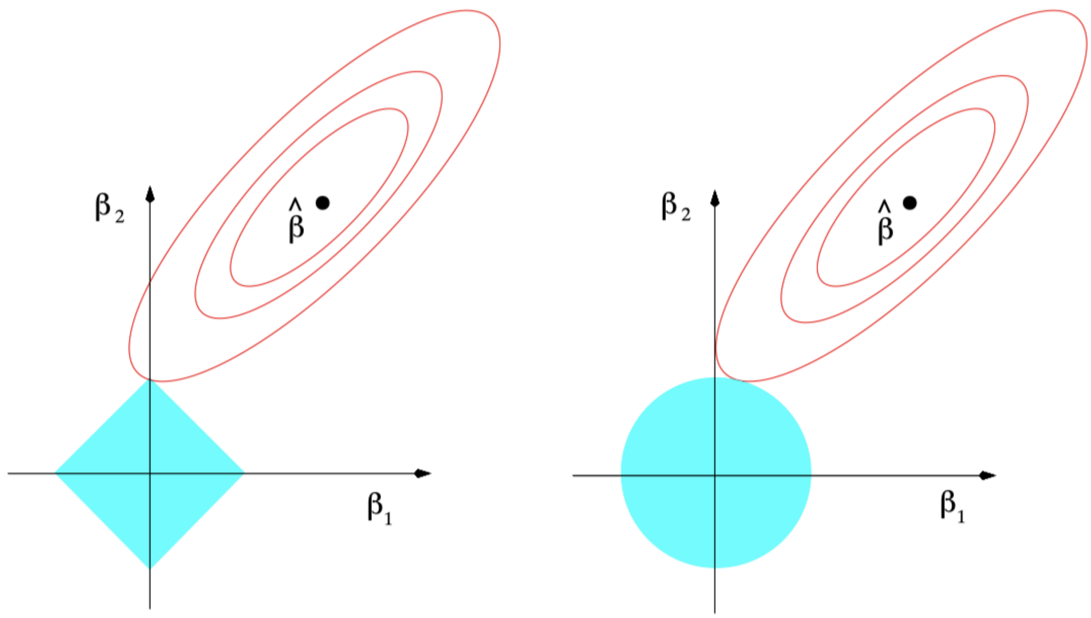

Figure 2.1: Estimation picture for the LASSO (left) and Ridge regression (right). Shown are the contours of the error and constraint functions. The solid blue areas are the constraint regions $|\boldsymbol{\beta}|_1 = |\beta_1| + |\beta_2| \leq t$ and $|\boldsymbol{\beta}|_2^2 = \beta_1^2 + \beta_2^2 \leq t^2$, respectively, while the red ellipses are the contours of the least squares error function.

corresponding σ-field $\mathcal{F}_s^t = (\varepsilon_s, \varepsilon_{s+1}, \ldots, \varepsilon_t)$ generated by the innovations $\varepsilon_s, \ldots, \varepsilon_t$. Set $\mathcal{F}^t := \mathcal{F}_{-\infty}^t$. Assume that the stationary time series $\{e_t\}_t$ has causal form

$$e_t := g(\mathcal{F}^t) = g(\ldots, \varepsilon_s, \varepsilon_{s+1}, \ldots, \varepsilon_t) \tag{2.5}$$

where $g(\cdot)$ are real-valued corresponding measurable functions such that e_t can be well-defined. Following Wu [2005] and Wu and Wu [2016], given the causal form (2.5), if $\|e_t\|_q < \infty$ for some $q \geq 1$, we can define the functional dependence measure as

$$\delta_{t,q} = \|e_t - e_t^*\|_q = \|e_t - g(\mathcal{F}^{t,\{0\}})\|_q = \|g(\mathcal{F}^t) - g(\mathcal{F}^{t,\{0\}})\|_q,$$

where the coupled processes are $e_t^* = g(\mathcal{F}^{t,\{0\}})$ with

$$\mathcal{F}^{t,\{0\}} = (\ldots, \varepsilon_{-1}, \varepsilon_0', \varepsilon_1, \ldots, \varepsilon_{t-1}, \varepsilon_t)$$

where ε_0' is i.i.d copy of ε_0. We can assume this kind of short-dependence given by

$$\Delta_{0,q} := \sum_{t=0}^{\infty} \delta_{t,q} < \infty$$

For fixed m, $\Delta_{m,q}$ measures the cumulative effect of ε_0 on $\{e_t\}_{t \geq m}$. Assume that the geometric moment contracting (GMC) condition is satisfied for each component process: There exists a constant $\rho \in (0,1)$ such that

$$\|e_.\|_q := \sup_{m\geq 0} \rho^{-m} \sum_{t=m}^{\infty} \delta_{t,q} < \infty \tag{2.6}$$

Now we introduce the extension of dependence-adjusted norm following from Wu and Wu [2016].

Definition 2.1. *A (weakly) one-dimensional stationary time series $\{X_t\}_{t\geq 1} \in \mathcal{L}^q$ holds for all $q \geq 2$ and, for some $\nu \geq 0$, define the following dependence-adjusted Orlicz norm as*

$$\|X_.\|_{\psi_\nu} := \sup_{q\geq 2} q^{-\nu} \|X_.\|_q = \sup_{q\geq 2} q^{-\nu} \sum_{t=0}^{\infty} \|X_t - X_t^*\|_q \tag{2.7}$$

We can provide a formal assumption for e_t below.

Assumption 2.4 (Dependent Noise)**.**

1. $(e_t)_{t\geq 1}$ *is weakly stationary with mean 0, i.e.* $\mathbb{E}e_t = 0$ *for any* $t \leq T$.
2. $\mathbb{E}e_t u_{it} = \mathbb{E}e_t f_{jt} = 0$ *for all* $t \leq T$, $i \leq p$ *and* $j \leq K$.
3. *If* $(e_t)_{t\geq 1} \in \mathcal{L}^q$ *holds for all* $q > 2$ *and, for some* $\nu \geq 0$,

$$\|e_.\|_{\psi_\nu} := \sup_{q\geq 2} q^{-\nu} \|e_.\|_q < \infty.$$

Moreover, there exists some $0 < \rho < 1$ *and* $\nu \geq 0$ *such that* $\|e_.\|_{\psi_\nu} < \infty$.

Before discussing the theoretical properties of $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$, we introduce some important and useful lemmas based on the dependence structure.

Lemma 2.1 (Theorem 3 from Wu and Wu [2016])**.** *Under Assumption 2.4, let* $S_n = \sum_{t=1}^{n} e_t$ *and* $\alpha = 2/(1+2\nu)$. *Then for* $x > 0$, *there exists a positive constant* C_α *only depending on* α *such that*

$$\mathbb{P}\left(S_n/\sqrt{n} \geq x\right) \leq C_\alpha \exp\left(-\frac{x^\alpha}{2e\alpha \|e_.\|_{\psi_\nu}^\alpha}\right).$$

Lemma 2.2. *Let* $S_n = \sum_{i=1}^{n} e_i$. *Denote* $\delta_{i,q} := \|e_i - e_{i,\{0\}}\|_q$. *For any* $m \geq 0$, *define* $e_{i,m} = \mathbb{E}(e_i|\varepsilon_{i-m+1}, \ldots, \varepsilon_i)$, $S_{n,m} = \sum_{i=1}^{n} e_{i,m}$ *and* $\Delta_{m,q} = \sum_{j=m}^{\infty} \delta_{j,q}$. *Under the same condition as Lemma 2.1, for any* $x > 0$,

$$\mathbb{P}\left((S_n - S_{n,m})/\sqrt{n} \geq x\right) \leq C_\alpha \exp\left(-\frac{x^\alpha}{C_{\rho,\alpha} \rho^{\alpha m/2} \|e_.\|_{\psi_\nu}^\alpha}\right),$$

where $\alpha = 2/(1+2\nu)$ *and* $C_{\rho,\alpha}$ *only depends on* α *and* ρ.

2.2.3 Estimation Error

In general, when we discuss the sparsity of parameters, we usually introduce a cone set to study its properties. Therefore, for any subset $\mathcal{S} \subset \{1, 2, \ldots, d\}$, define a convex cone $\mathcal{C}(\mathcal{S}, 3) := \{\boldsymbol{\delta} \in \mathbb{R}^d : |\boldsymbol{\delta}_{\mathcal{S}^c}|_1 \leq |\boldsymbol{\delta}_{\mathcal{S}}|_1\}$. Given the parameter in assumptions, we also write

$$\mathcal{V}_{n,d} = \frac{n}{d} + \sqrt{\frac{\log d}{n}} + \sqrt{\frac{n \log d}{d}} \tag{2.8}$$

Now, we provide some technical lemmas helpful to establish the estimation error of $\hat{\boldsymbol{\beta}}$ and $\hat{\boldsymbol{\gamma}}$.

Lemma 2.3. *Assume that* $\lambda \geq \frac{2}{n}|\hat{\boldsymbol{U}}^\top(\tilde{\boldsymbol{Y}} - \hat{\boldsymbol{U}}\boldsymbol{\beta}^*)|_\infty$ *and for some positive constant* $\kappa(\mathcal{S}_*, 3)$,

$$\kappa(\mathcal{S}_*, 3) := \min_{\mathcal{S}_* \subset \{1,\ldots,p\}, |S_*| \leq s} \min_{0 \neq \boldsymbol{v} \in \mathcal{C}(\mathcal{S}_*, 3)} \frac{\boldsymbol{v}^\top \tilde{\boldsymbol{\Sigma}} \boldsymbol{v}}{|\boldsymbol{v}|_2^2} > 0.$$

Then we have $\hat{\boldsymbol{\beta}}_\lambda - \boldsymbol{\beta}^* \in \mathcal{C}(\mathcal{S}_*, 3)$,

$$|\hat{\boldsymbol{\beta}}_\lambda - \boldsymbol{\beta}^*|_2 \leq \frac{3\lambda\sqrt{|\mathcal{S}_*|}}{\kappa(\mathcal{S}_*, 3)} \text{ and } |\hat{\boldsymbol{U}}(\hat{\boldsymbol{\beta}}_\lambda - \boldsymbol{\beta}^*)|_2^2 \leq \frac{9n\lambda^2|\mathcal{S}_*|}{\kappa(\mathcal{S}_*, 3)}.$$

Lemma 2.4. *Under Assumption 2.1-2.3, for any vector* $\boldsymbol{\varphi} \in \mathbb{R}^K$ *with* $|\boldsymbol{\varphi}|_2 = O(1)$, *we have*

$$|\hat{\boldsymbol{U}}^\top \boldsymbol{F} \boldsymbol{\varphi}|_\infty = O_{\mathbb{P}}(\mathcal{V}_{n,d}). \tag{2.9}$$

Lemma 2.5. *Under Assumptions 2.1-2.3, we have*

$$|\hat{\boldsymbol{U}}^\top \hat{\boldsymbol{U}} - \boldsymbol{U}^\top \boldsymbol{U}|_{\max} = O_{\mathbb{P}}\left(\frac{n}{d} + \log d\right).$$

Proof of Lemma 2.3, 2.4, 2.5. See the proof in Fan et al. [2023]. □

Lemma 2.6. *Under the Assumptions 2.1-2.4, there exists a positive constant* $C > 0$ *such that*

(i) $|\boldsymbol{F}^\top \boldsymbol{e}|_2 = O_{\mathbb{P}}(\sqrt{n})$,

(ii) $|\boldsymbol{U}^\top \boldsymbol{e}|_\infty = O_{\mathbb{P}}(\sqrt{n(\log d)^{1+2\nu}})$.

Proof of Lemma 2.6. (i) By the relationship between l_2 norm and l_∞ norm, we have

$$|\boldsymbol{F}^\top \boldsymbol{e}|_2 = \left|\sum_{t=1}^n \boldsymbol{f}_t e_t\right|_2 \leq \sqrt{K} \left|\sum_{t=1}^n \boldsymbol{f}_t e_t\right|_\infty.$$

Since Assumption 2.1 implies $\|f_{jt}\|_2 \leq \sqrt{2}Kc_0$ for some $K > 0$, we can get

$$\sum_{t=m}^{\infty} \|f_{jt}e_t - f_{jt}e_t^*\|_\tau = \sum_{t=1}^{\infty} \|f_{jt}(e_t - e_t^*)\|_\tau$$
$$\leq \sum_{t=1}^{\infty} \|f_{jt}\|_2 \|e_t - e_t^*\|_q \leq \sqrt{2}Kc_0\Delta_{0,q},$$

Thus, by Lemma 2.1, for $x > 0$, we have

$$\max_i \mathbb{P}\left(\left|\frac{1}{n}\sum_{t=1}^{n} f_{it}e_t\right| \geq x\right) \leq C\exp\{-C'(\sqrt{n}x/\|e_\cdot\|_{\psi_\nu})^{2/(1+2\nu)}\},$$

where constants C, C' only depend on ν. Using the Bonferroni inequality,

$$\mathbb{P}(\left|\boldsymbol{F}^\top \boldsymbol{e}\right|_2 \geq x) \leq \mathbb{P}\left(\sqrt{K}\max_{i\leq K}\left|\sum_{t=1}^{n} f_{it}e_t\right| \geq x\right)$$
$$\leq K\max_i \mathbb{P}\left(\left|\frac{1}{n}\sum_{t=1}^{n} f_{it}e_t\right| \geq x/(\sqrt{K}n)\right).$$

Now we choose a suitable $x = C\sqrt{n}$. For all large enough $C > 0$,

$$K\exp\{-C(x/\sqrt{Kn}\|e_\cdot\|_{\psi_\nu})^{2/(1+2\nu)}\} \to 0.$$

(ii)Similarly, we can get

$$\sum_{t=m}^{\infty} \|u_{jt}e_t - u_{jt}^*e_t^*\|_\tau \leq C\Delta_{0,q},$$

where $\tau = 2q/(2+q)$. Thus, by Lemma 2.1 and Bonferroni inequality, for $x > 0$, we have

$$\mathbb{P}\left(\max_{i\leq d}\left|\sum_{t=1}^{n} u_{it}e_t\right| \geq x\right) \leq d\max_i \mathbb{P}\left(\left|\sum_{t=1}^{n} u_{it}e_t\right| \geq x\right)$$
$$\leq dC\exp\{-C'(x/\sqrt{n}\|e_\cdot\|_{\psi_\nu})^{2/(1+2\nu)}\},$$

where constants C, C' only depend on ν. Now, we can choose a large enough $x = C(\log d)^{1/2+\nu}\sqrt{n}$. For all large enough $C > 0$,

$$d\exp\{-C'(x/\sqrt{n}\|e_\cdot\|_{\psi_\nu})^{2/(1+2\nu)}\} = O\left(\frac{1}{d}\right).$$

□

Lemma 2.7. *Under Assumptions 2.1-2.4, for any set $\mathcal{S} \subset \{1, 2, \ldots, p\}$ with*

$$|\mathcal{S}_*| \left(\frac{1}{d} + \frac{\log d}{n} \right) \to 0, \tag{2.10}$$

there exists a constant $\kappa(S, 3) > 0$ such that

$$\frac{\boldsymbol{v}^\top \tilde{\boldsymbol{\Sigma}} \boldsymbol{v}}{|\boldsymbol{v}|_2^2} \geq \kappa(\mathcal{S}, 3) = \frac{\lambda_{\min}(\boldsymbol{\Sigma})}{4},$$

with a high probability.

Lemma 2.8. *Under Assumptions 2.1-2.4, we have*

$$|(\hat{\boldsymbol{U}} - \boldsymbol{U})^\top \boldsymbol{e}|_\infty = O_{\mathbb{P}} \left(\sqrt{\frac{n}{d}} + \sqrt{\log d} \right).$$

Proof of Lemma 2.7, 2.8. See the proof in Fan et al. [2023]. □

Lemma 2.9. *Under Assumptions 2.1-2.4, we have*

$$|\hat{\boldsymbol{U}}^\top (\tilde{\boldsymbol{Y}} - \hat{\boldsymbol{U}} \boldsymbol{\beta}^*)|_\infty = O_{\mathbb{P}} \left(\sqrt{n (\log d)^{1+2\nu}} + \mathcal{V}_{n,d} |\boldsymbol{\varphi}^*|_2 \right).$$

Proof of Lemma 2.9. By Lemma 2.4,2.6 and 2.7, we have $|\hat{\boldsymbol{U}}^\top \boldsymbol{F} \boldsymbol{\varphi}^*|_\infty = O_{\mathbb{P}}(\mathcal{V}_{n,d} |\boldsymbol{\varphi}^*|_2)$ and

$$|\hat{\boldsymbol{U}}^\top \boldsymbol{e}|_\infty \leq |(\hat{\boldsymbol{U}} - \boldsymbol{U})^\top \boldsymbol{e}|_\infty + |\boldsymbol{U}^\top \boldsymbol{e}|_\infty = O_{\mathbb{P}} \left(\sqrt{n (\log d)^{1+2\nu}} \right).$$

Thus, combining the two inequalities implies

$$\begin{aligned} |\hat{\boldsymbol{U}}^\top (\tilde{\boldsymbol{Y}} - \hat{\boldsymbol{U}} \boldsymbol{\beta}^*) &= |\hat{\boldsymbol{U}}^\top \boldsymbol{e} + \hat{\boldsymbol{U}}^\top \boldsymbol{F} \boldsymbol{\varphi}^*|_\infty \leq |\hat{\boldsymbol{U}}^\top \boldsymbol{e}|_\infty + |\hat{\boldsymbol{U}}^\top \boldsymbol{F} \boldsymbol{\varphi}^*|_\infty \\ &= O_{\mathbb{P}} \left(\sqrt{n (\log d)^{1+2\nu}} + \mathcal{V}_{n,d} |\boldsymbol{\varphi}^*|_2 \right). \end{aligned}$$

□

Now we state the main result concerning estimation consistency under the new Assumptions 2.1-2.4.

Theorem 2.2 (Theorem 2.2 in Fan et al. [2023])**.** *If $\boldsymbol{\varphi}^* = \boldsymbol{\gamma}^* - \boldsymbol{B}^\top \boldsymbol{\beta}^*$, then under Assumptions 2.1-2.4, we have*

$$|\hat{\boldsymbol{\gamma}} - \boldsymbol{H} \boldsymbol{\gamma}^*|_2 = O_{\mathbb{P}} \left\{ \frac{1}{\sqrt{n}} + \left(\frac{1}{\sqrt{n}} + \frac{1}{\sqrt{d}} \right) |\boldsymbol{\varphi}^*|_2 + \left(\sqrt{\frac{\log |\mathcal{S}_\star|}{n}} + \frac{1}{\sqrt{d}} \right) |\boldsymbol{\beta}^*|_1 \right\}$$

where $\mathcal{S}_* = \{j : \beta_j^* \neq 0, 1 \leq j \leq p\}$ and $|\mathcal{S}_*|$ is cardinality of set $\mathcal{S}_*$.

Proof of Theorem 2.2. Apply similar proof in Fan et al. [2023]. □

Theorem 2.3. *If*

$$|\mathcal{S}_*|\left(\frac{1}{d}+\frac{\log d}{n}\right)\to 0,$$

then under Assumptions 2.1-2.4, choosing appropriate $\lambda \geq \frac{2}{n}|\hat{\boldsymbol{U}}^\top(\tilde{\boldsymbol{Y}}-\hat{\boldsymbol{U}}\boldsymbol{\beta}^*)|_\infty$*, we have* $\hat{\boldsymbol{\beta}}_\lambda-\boldsymbol{\beta}^*\in\mathcal{C}(\mathcal{S}_*,3)$ *and*

$$|\hat{\boldsymbol{\beta}}_\lambda-\boldsymbol{\beta}^*|_2 = O_{\mathbb{P}}\left(\sqrt{\frac{|\mathcal{S}_*|(\log d)^{1+2\nu}}{n}}+\frac{\mathcal{V}_{n,d}|\boldsymbol{\varphi}^*|_2\sqrt{|\mathcal{S}_*|}}{n}\right) \tag{2.11}$$

where $\mathcal{V}_{n,d}$ *is defined in (2.8).*

Proof of Theorem 2.3. Applying Lemma 2.3 with Lemma 2.7 and 2.9 and using the fact that $|\boldsymbol{v}|_2\leq\sqrt{|\boldsymbol{v}|}|\boldsymbol{v}|_1$ can obtain (2.11). □

Chapter 3

Simulation

For data generation, we set the number of factors $K = 2$, dimension of covariate $d = 100$, sparsity $s = 3$. We select the corresponding number of observations n according to the ratio $s\sqrt{(\log d)^{1+2\nu}/T}$ that takes uniform grids in [0.30, 0.60]. We replicate the experiment 500 times. In addition, we assume the noise follows the linear Autoregressive (AR) model with MA(∞) representation, i.e.

$$e_t = \phi e_{t-1} + \varepsilon_t = \sum_{k=0}^{\infty} \phi^k \varepsilon_{t-k}, t = 1, \ldots, n$$

where ϕ satisfies $|\phi| < 1$ and the innovation ε_t follows the Gaussian distribution $\mathcal{N}(0, 0.5^2)$. We generate every entry in factors $\boldsymbol{F}$ and idiosyncratic error $\boldsymbol{U}$ following from the standard Gaussian distribution, every entry in factor loadings $\boldsymbol{B}$ following from the uniform distribution Unif$(-1, 1)$. Moreover, according to the linearity of noise, we can choose the dependency ϕ to be 0.1 and 0.9. Specifically, since the variance of e_t is

$$\text{Var}(e_t) = \frac{\sigma_\varepsilon^2}{1 - \phi^2} = \frac{0.5^2}{1 - \phi^2},$$

$\phi = 0.1$ and 0.9 imply the dependence within the noise and their standard deviations are far away from $\boldsymbol{F}$ and $\boldsymbol{U}$. We compared the results under $\phi = 0.1$ and $\phi = 0.9$. Then, we rescale the e_t by standardizing their corresponding standard deviation to 0.5.

In Figure 3.1 and 3.2, the red lines represent the estimation results using our model while the blue lines denote the results using the traditional LASSO method. The means of the distance between our estimators and true parameters approaches 0.2, while the others are above 0.6. Also, the standard deviation for our estimators is less than that for traditional LASSO estimators. Therefore, we find that our estimators outperform the original LASSO estimators even though the dependency is stronger.

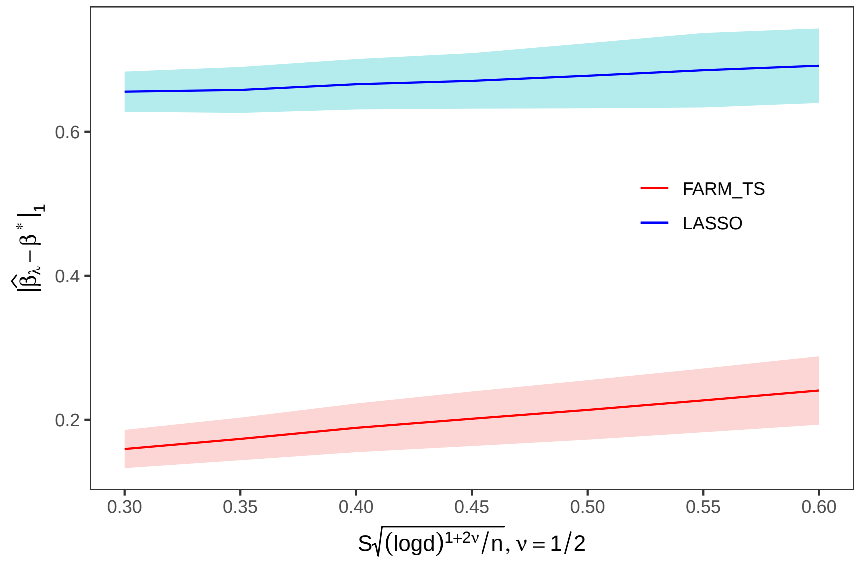

Figure 3.1: Accuracy for $\hat{\beta}_\lambda$ with $\text{dist}(\hat{\beta}_\lambda, \beta^\star) := |\hat{\beta}_\lambda - \beta^\star|_1$ based on 500 replication under dependency of noise $\phi = 0.1$.

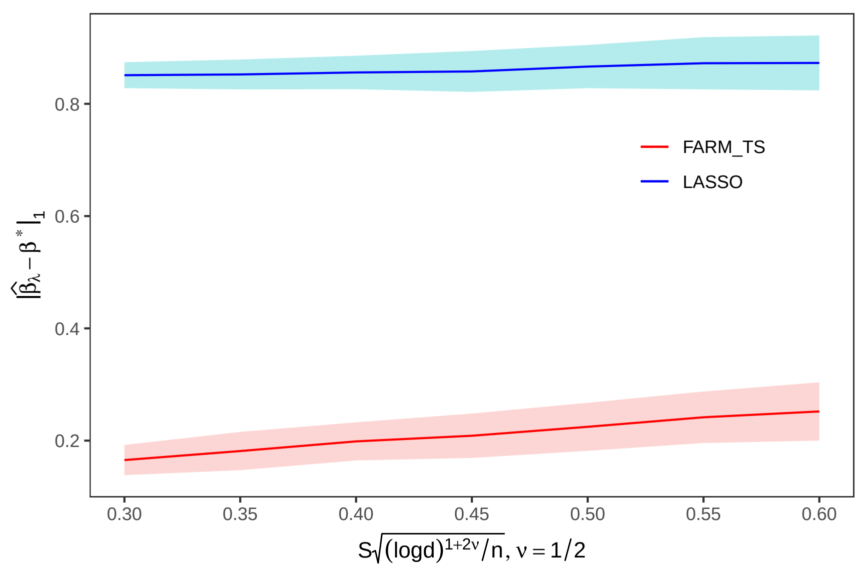

Figure 3.2: Accuracy for $\hat{\beta}_\lambda$ with $\text{dist}(\hat{\beta}_\lambda, \beta^\star) := |\hat{\beta}_\lambda - \beta^\star|_1$ based on 500 replication under dependency of noise $\phi = 0.9$.

Chapter 4

Real Data Analysis

4.1 Background and Motivation

In this section, we introduce a macroeconomic dataset named FRED-MD [McCracken and Ng, 2020] to evaluate the performance of our model. There are 210 quarterly U.S. macroeconomic variables in this dataset. They exhibit correlation since they relate to various aspects of the economy thus driven by similar factors. In our study, we pick out 2 variables named All Employees: Retail Trade (Thousands of Persons) and GOV:FED as our response variables while letting the rest be the covariates. Specifically, USTRADE stands for All Employees: Retail Trade (Thousands of Persons); GOV:FED represents Real Government Consumption Expenditures and Gross Investment: Federal (Percent Change from Preceding Period).

We choose quarterly data from March 1967 to December 2019 and apply recommended transformations to the data. When discussing prediction and inference of real data analysis, some scholars tend to select more compact, stationary, or even normally distributed data; however, such data options are invariably limited. Therefore, unlike Fan et al. [2023], which excludes the period from November 2007 to July 2010 due to the global financial crisis causing significant financial breaks and rendering the data non-stationary, we believe that after performing transformations, such as (code: 1 = no transformation, 2 = first difference Δx_t, 3 = second difference $\Delta^2 x_t$, 4 = $\log(x_t)$, 5 = first difference of logged variables $\Delta \log(x_t)$, 6 = second difference of logged variables $\Delta^2 \log(x_t)$), the data will be stationary. Specifically, we apply code 5 to "USTRADE" and code 1 to "GOV:FED". Moreover, we can check QQ plots of these response variables before the inference and prediction to determine whether they approximately follow a Gaussian distribution. Thus, we incorporate all periods in our analysis.

We employ models PCR, LASSO, and RIDGE to illustrate the performance of our proposed model. Using the moving window approach with a window size w of different months, we analyze and compare the prediction results obtained by employing these models. For example, given the window size $w = 120$, for each period and model, we utilize the panel data indexing from 1 for each of the one time periods, for all 120, we use the 120 previous observation pairs $\{(\boldsymbol{x}_{t-120}, Y_{t-120}), \ldots, (\boldsymbol{x}_{t-1}, Y_{t-1})\}$ to train a model and output a prediction $\widehat{Y}_t$ and in-sample mean $\bar{Y}_t = \frac{1}{120}\sum_{i=t-120}^{t-1} Y_i$. We evaluate the model prediction by introducing out-of-sample R^2 given by

$$R^2 := 1 - \frac{\sum_{t=121}^{T}(Y_t - \widehat{Y}_t)^2}{\sum_{t=121}^{T}(Y_t - \bar{Y}_t)^2},$$

where T denotes the number of total data points in a given period.

4.2 Discussion

We pick up two target response variables and check their normal QQplot in the whole period: from March 1967 to December 2019. Obviously, from Figure 4.1 and 4.2, we can conclude that the two variables we select do follow a Gaussian distribution. It is because we use some tricky methods to make them more stationary than the original one. It also implies the big economic crisis does not affect this kind of economy index roughly.

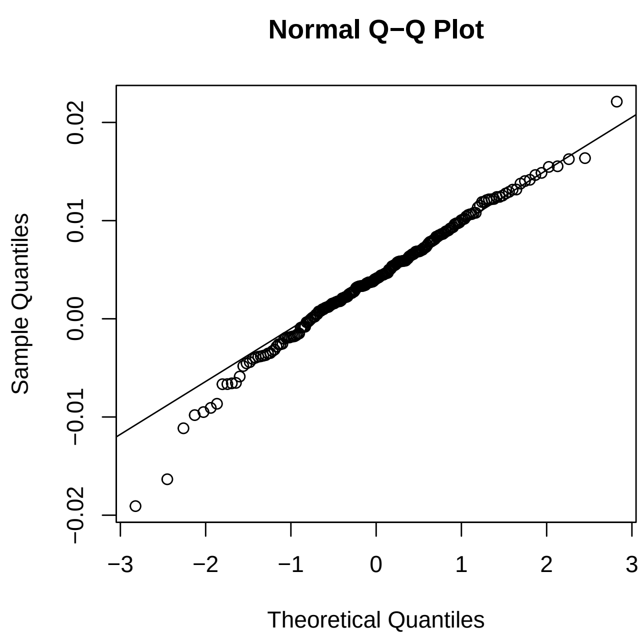

Figure 4.1: Normal QQPlot of "USTRADE" data in periods: from March 1967 to December 2019

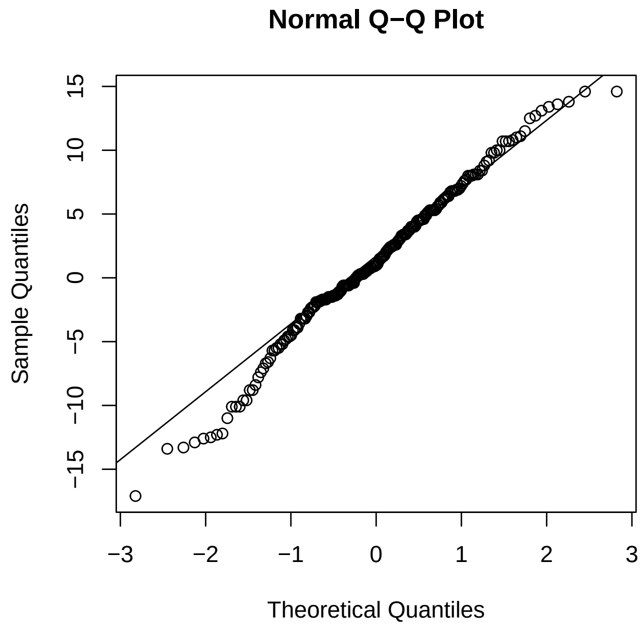

Figure 4.2: Normal QQPlot of "GOV:FED" or "Gov:Fed" data in time periods: From March 1967 to December 2019

Then, we select the time window as $w = 90$ to run the code. Table 4.1 implies that the national data as a response variable is much more stationary and follows a normal distribution approximately. Also, our model displays higher R^2 compared to other methods. For instance, GOV:FED has $R^2 = 0.950$ in our model while R^2 is 0.945, 0.100, 0.048 under other models, respectively; USTRADE has R^2 0.950 in our model while R^2 is 0.926, 0.624, 0.586 under other models, respectively.

Data	DFARM	LASSO	RIDGE	PCR
GOV:FED	0.950	0.945	0.100	0.048
USTRADE	0.950	0.926	0.624	0.586

Table 4.1: Out-of-sample R^2 for predicting GOV:FED and USTRADE data using different models in different time windows quarterly from March 1967 to December 2019.

Finally, we present the out-of-sample prediction results for the 'PCDGx' dataset using the optimal window size. Our model (red dashed line) demonstrates the closest alignment with the true observed values, particularly in capturing peaks, and then followed by LASSO, exhibiting commendable performance. The moving average performs the worst, with almost a flat line around the value 0.

What's more, considering the graph is for the result of 'GOV:FED' data, which is an indicator of the government's real consumption expenditures and gross investment,

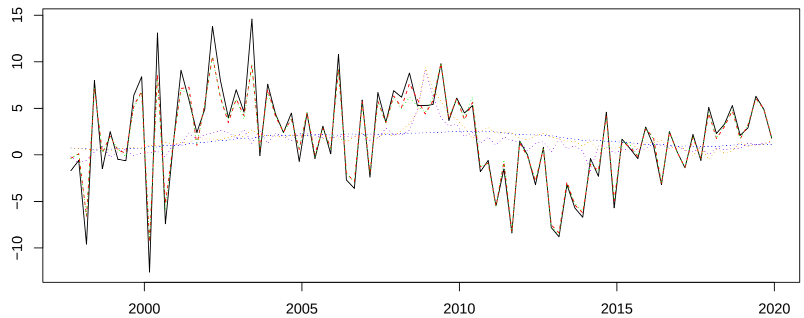

Figure 4.3: Out-of-sample prediction results for 'GOV:FED' data in periods: from September 1997 to December 2019. The black line represents the true observed values, the red dashed line stands for the predictions made by our model, and the green, purple, blue, and orange dot lines represent the predictions made by using LASSO, RIDGE, in-sample mean (moving average with the corresponding window size), and PCR, respectively.

we can discern a strong correlation between the depicted pattern and the economic conditions during that period. Notably, the year 2008 marked one of the most severe financial crises since the Great Depression. Given this historical context, it might come as a surprise that the graph shows only minor fluctuations around the value 5, with the presence of two peaks around 2008. However, 'GOV:FED' is influenced by various factors besides the economic cycle, such as government policies, social demands, and infrastructure needs. On the one hand, governments often adopt counter-cyclical fiscal policies during crises to increase spending and investment to offset the negative impact of reduced private sector spending. For instance, intending to inject capital into the banking system and prevent a collapse of the financial sector, the U.S. enacted the Emergency Economic Stabilization Act (EESA) in October 2008. On the other hand, there can be lag effects, causing the impact of a financial crisis to manifest in the graph with a delay. It explains the line's declining trend from 2010 onward, followed by a gradual resurgence after 2013.

Similarly, we output the out-of-sample prediction outcomes for 'USTRADE' data using the optimal window size. While the moving average still yields the weakest performance, the predictions from our model, LASSO, and PCR demonstrate comparable effectiveness.

In addition, as previously discussed, Figure 4.3 displays a time-delayed reflection of the economic situation, but Figure 4.4 illustrates a more current state of the economy since we can see the two most distinct valleys in 2002 and 2008. The year 2002 marks the worst tumble of the stock market since 1987, while 2008 is the notable financial crisis as mentioned before. The timely ups and downs could be attributed to the fact

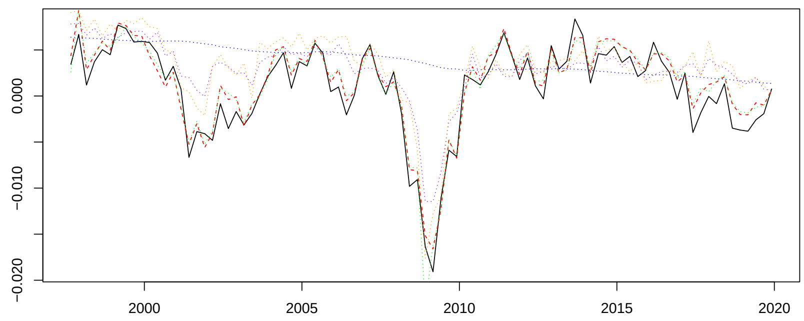

Figure 4.4: Out-of-sample prediction results for 'USTRADE' data in periods: from September 1997 to December 2019. The captions are the same as those in the Previous Figure.

that 'USTRADE' is an indicator of the retail trade of all employees, so it is a key component of consumer spending that indicates shifts in consumer confidence and disposable income, and thereby it provides a more timely snapshot of the economic situation.

Chapter 5

Conclusion and Discussion

In the paper, we introduce the Dynamic Factor Augmented Regression Model as an approach to address the challenges posed by high-dimensional time series data. While early research often overlooked the natural dependence structure in the model, opting for independent noise, we extend the model's conditions and assume the noise follows an autoregressive (AR) process. Moreover, regularization techniques are incorporated to address issues arising from regression with high-dimensional data. In the simulation analysis, in contrast to LASSO, our model consistently demonstrates superior performance in minimizing the L_1 estimation error ($|\hat{\boldsymbol{\beta}}_\lambda - \boldsymbol{\beta}^*|_1$), regardless of the strength of the dependency parameter ($\phi = 0.1$ or 0.9). Moreover, our model maintains an error rate of less than 0.2 even when subjected to increased convergence rates $S\sqrt{(\log d)^{1+2v}/n}$. When our model is tested on authentic US economic data, it closely follows the actual values, particularly at catching peaks and valleys. The selection of 'GOV:FED' and 'USTRADE' as response variables encompasses both macroeconomic and microeconomic dimensions. While 'GOV:FED' reflects economic conditions with a discernible lag effect, 'USTRADE' provides more immediate insights, enhancing our model's ability to predict broader economic trends beyond isolated macro or micro perspectives. In the future, robustness can also be discussed in our model which requires more conditions on our data.

Appendix A

Real Data Background

FRED-QD is a quarterly frequency companion to FRED-MD. It is designed to emulate the dataset used in "Disentangling the Channels of the 2007-2009 Recession" by Stock and Watson (2012, NBER WP No. 18094) but also contains several additional series. The columns denote the following: (i) ID denotes the series number, (ii) SW ID denotes the series number in SW (2012), (iii) TCODE denotes one of the following data transformations for a series x : (1) no transformation; (2) Δx_t; (3) $\Delta^2 x_t$; (4) $\log(x_t)$; (5) $\Delta \log(x_t)$; (6) $\Delta^2 \log(x_t)$. (7) $\Delta(x_t/x_{t-1} - 1.0)$, (iv) SW FACTORS denotes whether a series was used in SW (2012) when constructing factors (i.e. 1 is yes and 0 is no), (v) FRED MNEMONIC denotes the mnemonic we use for the dataset, (vi) SW MNEMONIC denotes the mnemonic used in SW (2012), and (vii) DESCRIPTION gives a brief definition of the series. The series is loosely grouped based on SW (2012).

Details on the construction of the data will be forthcoming, but a few general comments are in order. First, if the FRED mnemonic does not end in " x " then the series comes directly from the FRED database (e.g. PCECC96; real PCE). Otherwise, the series is a modified variant of a series from FRED (e.g. PCDGx; nominal PCE durables are manually deflated using the PCE price index). The exception to this rule is the S&P data, which is taken from public sources. Lastly, monthly frequency series are aggregated to a quarterly frequency using averages.

Group 1: NIPA

	ID	SW ID	TCODE	SW FACTORS	FRED MNEMONIC	SW MNEMONIC	DESCRIPTION
1	1	1	5	0	GDPC1	GDP	Real Gross Domestic Product, 3 Decimal (Billions of Chained 2012 Dollars)
2	2	2	5	0	PCECC96	Consumption	Real Personal Consumption Expenditures (Billions of Chained 2012 Dollars)
3	3	3	5	1	PCDGx	Cons:Dur	Real personal consumption expenditures: Durable goods (Billions of Chained 2012 Dollars), deflated using PCE
4	4	4	5	1	PCESVx	Cons:Svc	Real Personal Consumption Expenditures: Services (Billions of 2012 Dollars), deflated using PCE
5	5	5	5	1	PCNDx	Cons:NonDur	Real Personal Consumption Expenditures: Nondurable Goods (Billions of 2012 Dollars), deflated using PCE
6	6	6	5	0	GPDIC1	Investment	Real Gross Private Domestic Investment, 3 decimal (Billions of Chained 2012 Dollars)
7	7	7	5	0	FPIx	FixedInv	Real private fixed investment (Billions of Chained 2012 Dollars), deflated using PCE
8	8	8	5	1	Y033RC1Q027SBEAx	Inv:Equip&Software	Real Gross Private Domestic Investment: Fixed Investment: Nonresidential: Equipment (Billions of Chained 2012 Dollars), deflated using PCE
9	9	9	5	1	PNFIx	FixInv:NonRes	Real private fixed investment: Nonresidential (Billions of Chained 2012 Dollars), deflated using PCE
10	10	10	5	1	PRFIx	FixedInv:Res	Real private fixed investment: Residential (Billions of Chained 2012 Dollars), deflated using PCE
11	11	11	1	1	A014RE1Q156NBEA	Inv:Inventories	Shares of gross domestic product: Gross private domestic investment: Change in private inventories (Percent)
12	12	12	5	0	GCEC1	Gov.Spending	Real Government Consumption Expenditures & Gross Investment (Billions of Chained 2012 Dollars)
13	13	13	1	1	A823RL1Q225SBEA	Gov:Fed	Real Government Consumption Expenditures and Gross Investment: Federal (Percent Change from Preceding Period)
14	14	14	5	1	FGRECPTx	Real Gov Receipts	Real Federal Government Current Receipts (Billions of Chained 2012 Dollars), deflated using PCE
15	15	15	5	1	SLCEx	Gov:State&Local	Real government state and local consumption expenditures (Billions of Chained 2012 Dollars), deflated using PCE
16	16	16	5	1	EXPGSC1	Exports	Real Exports of Goods & Services, 3 Decimal (Billions of Chained 2012 Dollars)
17	17	17	5	1	IMPGSC1	Imports	Real Imports of Goods & Services, 3 Decimal (Billions of Chained 2012 Dollars)
18	18	18	5	0	DPIC96	Disp-Income	Real Disposable Personal Income (Billions of Chained 2012 Dollars)
19	19	19	5	0	OUTNFB	Ouput:NFB	Nonfarm Business Sector: Real Output (Index 2012=100)
20	20	20	5	0	OUTBS	Output:Bus	Business Sector: Real Output (Index 2012=100)
21	21	21	5	0	OUTMS	Output:Manuf	Manufacturing Sector: Real Output (Index 2012=100)
22	190	n.a.	2	0	B020RE1Q156NBEA		Shares of gross domestic product: Exports of goods and services (Percent)
23	191	n.a.	2	0	B021RE1Q156NBEA		Shares of gross domestic product: Imports of goods and services (Percent)

Group 2: Industrial Production

	ID	SW ID	TCODE	SW FACTORS	FRED MNEMONIC	SW MNEMONIC	DESCRIPTION
1	22	22	5	0	INDPRO	IP:Total index	Industrial Production Index (Index 2012=100)
2	23	23	5	0	IPFINAL	IP:Final products	Industrial Production: Final Products (Market Group) (Index 2012=100)
3	24	24	5	0	IPCONGD	IP:Consumer goods	Industrial Production: Consumer Goods (Index 2012=100)
4	25	25	5	0	IPMAT	IP:Materials	Industrial Production: Materials (Index 2012=100)
5	26	26	5	1	IPDMAT	IP:Dur gds materials	Industrial Production: Durable Materials (Index 2012=100)
6	27	27	5	1	IPNMAT	IP:Nondur gds materials	Industrial Production: Nondurable Materials (Index 2012=100)
7	28	28	5	1	IPDCONGD	IP:Dur Cons. Goods	Industrial Production: Durable Consumer Goods (Index 2012=100)
8	29	29	5	1	IPB51110SQ	IP:Auto	Industrial Production: Durable Goods: Automotive products (Index 2012=100)
9	30	30	5	1	IPNCONGD	IP:NonDur Cons God	Industrial Production: Nondurable Consumer Goods (Index 2012=100)
10	31	31	5	1	IPBUSEQ	IP:Bus Equip	Industrial Production: Business Equipment (Index 2012=100)
11	32	32	5	1	IPB51220SQ	IP:Energy Prds	Industrial Production: Consumer energy products (Index 2012=100)
12	33	33	1	1	TCU	Capu Tot	Capacity Utilization: Total Industry (Percent of Capacity)
13	34	34	1	1	CUMFNS	Capu Man.	Capacity Utilization: Manufacturing (SIC) (Percent of Capacity)
14	194	n.a.	5	0	IPMANSICS		Industrial Production: Manufacturing (SIC) (Index 2012=100)
15	195	n.a.	5	0	IPB51222S		Industrial Production: Residential Utilities (Index 2012=100)
16	196	n.a.	5	0	IPFUELS		Industrial Production: Fuels (Index 2012=100)

Group 3: Employment and Unemployment

	ID	SW ID	TCODE	SW FACTORS	FRED MNEMONIC	SW MNEMONIC	DESCRIPTION
1	35	35	5	0	PAYEMS	Emp:Nonfarm	All Employees: Total nonfarm (Thousands of Persons)
2	36	36	5	0	USPRIV	Emp:Private	All Employees: Total Private Industries (Thousands of Persons)
3	37	37	5	0	MANEMP	Emp:mfg	All Employees: Manufacturing (Thousands of Persons)
4	38	38	5	0	SRVPRD	Emp:Services	All Employees: Service-Providing Industries (Thousands of Persons)
5	39	39	5	0	USGOOD	Emp:Goods	All Employees: Goods-Producing Industries (Thousands of Persons)
6	40	40	5	1	DMANEMP	Emp:DurGoods	All Employees: Durable goods (Thousands of Persons)
7	41	41	5	0	NDMANEMP	Emp:Nondur Goods	All Employees: Nondurable goods (Thousands of Persons)
8	42	42	5	1	USCONS	Emp:Const	All Employees: Construction (Thousands of Persons)
9	43	43	5	1	USEHS	Emp:Edu&Health	All Employees: Education & Health Services (Thousands of Persons)
10	44	44	5	1	USFIRE	Emp:Finance	All Employees: Financial Activities (Thousands of Persons)
11	45	45	5	1	USINFO	Emp:Infor	All Employees: Information Services (Thousands of Persons)
12	46	46	5	1	USPBS	Emp:Bus Serv	All Employees: Professional & Business Services (Thousands of Persons)
13	47	47	5	1	USLAH	Emp:Leisure	All Employees: Leisure & Hospitality (Thousands of Persons)
14	48	48	5	1	USSERV	Emp:OtherSvcs	All Employees: Other Services (Thousands of Persons)
15	49	49	5	1	USMINE	Emp:Mining/NatRes	All Employees: Mining and logging (Thousands of Persons)
16	50	50	5	1	USTPU	Emp:Trade&Trans	All Employees: Trade, Transportation & Utilities (Thousands of Persons)
17	51	51	5	0	USGOVT	Emp:Gov	All Employees: Government (Thousands of Persons)
18	52	52	5	1	USTRADE	Emp:Retail	All Employees: Retail Trade (Thousands of Persons)
19	53	53	5	1	USWTRADE	Emp:Wholesal	All Employees: Wholesale Trade (Thousands of Persons)
20	54	54	5	1	CES9091000001	Emp:Gov(Fed)	All Employees: Government: Federal (Thousands of Persons)
21	55	55	5	1	CES9092000001	Emp:Gov (State)	All Employees: Government: State Government (Thousands of Persons)
22	56	56	5	1	CES9093000001	Emp:Gov (Local)	All Employees: Government: Local Government (Thousands of Persons)
23	57	57	5	0	CE16OV	Emp:Total (HHSurve)	Civilian Employment (Thousands of Persons)
24	58	58	2	0	CIVPART	LF Part Rate	Civilian Labor Force Participation Rate (Percent)
25	59	59	2	0	UNRATE	Unemp Rate	Civilian Unemployment Rate (Percent)
26	60	60	2	0	UNRATESTx	Urate_ST	Unemployment Rate less than 27 weeks (Percent)
27	61	61	2	0	UNRATELTx	Urate_LT	Unemployment Rate for more than 27 weeks (Percent)
28	62	62	2	1	LNS14000012	Urate:Age16-19	Unemployment Rate - 16 to 19 years (Percent)
29	63	63	2	1	LNS14000025	Urate:Age>20 Men	Unemployment Rate - 20 years and over, Men (Percent)
30	64	64	2	1	LNS14000026	Urate:Age>20 Women	Unemployment Rate - 20 years and over, Women (Percent)
31	65	65	5	1	UEMPLT5	U:Dur<5wks	Number of Civilians Unemployed - Less Than 5 Weeks (Thousands of Persons)
32	66	66	5	1	UEMP5TO14	U:Dur5-14wks	Number of Civilians Unemployed for 5 to 14 Weeks (Thousands of Persons)
33	67	67	5	1	UEMP15T26	U:dur>15-26wks	Number of Civilians Unemployed for 15 to 26 Weeks (Thousands of Persons)
34	68	68	5	1	UEMP27OV	U:Dur>27wks	Number of Civilians Unemployed for 27 Weeks and Over (Thousands of Persons)
35	69	69	5	1	LNS13023621	U:Job losers	Unemployment Level - Job Losers (Thousands of Persons)
36	70	70	5	1	LNS13023557	U:LF Reenty	Unemployment Level - Reentrants to Labor Force (Thousands of Persons)
37	71	71	5	1	LNS13023705	U:Job Leavers	Unemployment Level - Job Leavers (Thousands of Persons)
38	72	72	5	1	LNS13023569	U:New Entrants	Unemployment Level - New Entrants (Thousands of Persons)

Group 3: Employment and Unemployment, continued

	ID	SW ID	TCODE	SW FACTORS	FRED MNEMONIC	SW MNEMONIC	DESCRIPTION
39	73	73	5	1	LNS12032194	Emp:SlackWk	Employment Level - Part-Time for Economic Reasons, All Industries (Thousands of Persons)
40	74	74	5	0	HOABS	EmpHrs:Bus Sec	Business Sector: Hours of All Persons (Index 2012=100)
41	75	75	5	0	HOAMS	EmpHrs:mfg	Manufacturing Sector: Hours of All Persons (Index 2012=100)
42	76	76	5	0	HOANBS	EmpHrs:nfb	Nonfarm Business Sector: Hours of All Persons (Index 2012=100)
43	77	77	1	1	AWHMAN	AWH Man	Average Weekly Hours of Production and Nonsupervisory Employees: Manufacturing (Hours)
44	78	78	2	1	AWHNONAG	AWH Privat	Average Weekly Hours Of Production And Nonsupervisory Employees: Total private (Hours)
45	79	79	2	1	AWOTMAN	AWH Overtime	Average Weekly Overtime Hours of Production and Nonsupervisory Employees: Manufacturing (Hours)
46	80	80	1	0	HWIx	HelpWnted	Help-Wanted Index
47	197	n.a.	2	0	UEMPMEAN		Average (Mean) Duration of Unemployment (Weeks)
48	198	n.a.	2	0	CES0600000007		Average Weekly Hours of Production and Nonsupervisory Employees: Goods-Producing
49	220	n.a.	2	0	HWIURATIOx		Ratio of Help Wanted/No. Unemployed
50	221	n.a.	5	0	CLAIMSx		Initial Claims

Group 4: Housing

	ID	SW ID	TCODE	SW FACTORS	FRED MNEMONIC	SW MNEMONIC	DESCRIPTION
1	81	81	5	0	HOUST	Hstarts	Housing Starts: Total: New Privately Owned Housing Units Started (Thousands of Units)
2	82	82	5	0	HOUST5F	Hstarts >5units	Privately Owned Housing Starts: 5-Unit Structures or More (Thousands of Units)
3	83	83	5	1	PERMIT	Hpermits	New Private Housing Units Authorized by Building Permits (Thousands of Units)
4	84	84	5	1	HOUSTMW	Hstarts:MW	Housing Starts in Midwest Census Region (Thousands of Units)
5	85	85	5	1	HOUSTNE	Hstarts:NE	Housing Starts in Northeast Census Region (Thousands of Units)
6	86	86	5	1	HOUSTS	Hstarts:S	Housing Starts in South Census Region (Thousands of Units)
7	87	87	5	1	HOUSTW	Hstarts:W	Housing Starts in West Census Region (Thousands of Units)
8	179	190	5	1	USSTHPI	Real Hprice:OFHEO	All-Transactions House Price Index for the United States (Index 1980 Q1=100)
9	180	191	5	1	SPCS10RSA	Real CS_10	S&P/Case-Shiller 10-City Composite Home Price Index (Index January 2000 = 100)
10	181	192	5	1	SPCS20RSA	Real CS_20	S&P/Case-Shiller 20-City Composite Home Price Index (Index January 2000 = 100)
11	227	n.a.	5	0	PERMITNE		New Private Housing Units Authorized by Building Permits in the Northeast Census Region (Thousands, SAAR)
12	228	n.a.	5	0	PERMITMW		New Private Housing Units Authorized by Building Permits in the Midwest Census Region (Thousands, SAAR)
13	229	n.a.	5	0	PERMITS		New Private Housing Units Authorized by Building Permits in the South Census Region (Thousands, SAAR)
14	230	n.a.	5	0	PERMITW		New Private Housing Units Authorized by Building Permits in the West Census Region (Thousands, SAAR)

Group 5: Inventories, Orders, and Sales

	ID	SW ID	TCODE	SW FACTORS	FRED MNEMONIC	SW MNEMONIC	DESCRIPTION
1	88	89	5	0	CMRMTSPLx	MT Sales	Real Manufacturing and Trade Industries Sales (Millions of Chained 2012 Dollars)
2	89	90	5	1	RSAFSx	Ret. Sale	Real Retail and Food Services Sales (Millions of Chained 2012 Dollars), deflated by Core PCE
3	90	91	5	1	AMDMNOx	Orders (DurMfg)	Real Manufacturers' New Orders: Durable Goods (Millions of 2012 Dollars), deflated by Core PCE
4	91	92	5	1	ACOGNOx	Orders(ConsGoods/Mat.)	Real Value of Manufacturers' New Orders for Consumer Goods Industries (Millions of 2012 Dollars), deflated by Core PCE
5	92	93	5	1	AMDMUOx	UnfOrders(DurGds)	Real Value of Manufacturers' Unfilled Orders for Durable Goods Industries (Millions of 2012 Dollars), deflated by Core PCE
6	93	94	5	1	ANDENOx	Orders(NonDefCap)	Real Value of Manufacturers' New Orders for Capital Goods: Nondefense Capital Goods Industries (Millions of 2012 Dollars), deflated by Core PCE
7	94	96	5	1	INVCQRMTSPL	MT Invent	Real Manufacturing and Trade Inventories (Millions of 2012 Dollars)
8	222	n.a.	5	0	BUSINVx		Total Business Inventories (Millions of Dollars)
9	223	n.a.	2	0	ISRATIOx		Total Business: Inventories to Sales Ratio

Group 6: Prices

	ID	SW ID	TCODE	SW FACTORS	FRED MNEMONIC	SW MNEMONIC	DESCRIPTION
1	95	97	6	0	PCECTPI	PCED	Personal Consumption Expenditures: Chain-type Price Index (Index 2012=100)
2	96	98	6	0	PCEPILFE	PCED_LFE	Personal Consumption Expenditures Excluding Food and Energy (Chain-Type Price Index) (Index 2012=100)
3	97	99	6	0	GDPCTPI	GDP Defl	Gross Domestic Product: Chain-type Price Index (Index 2012=100)
4	98	100	6	1	GPDICTPI	GPDI Defl	Gross Private Domestic Investment: Chain-type Price Index (Index 2012=100)
5	99	101	6	1	IPDBS	BusSec Defl	Business Sector: Implicit Price Deflator (Index 2012=100)
6	100	102	6	0	DGDSRG3Q086SBEA	PCED_Goods	Personal consumption expenditures: Goods (chain-type price index)
7	101	103	6	0	DDURRG3Q086SBEA	PCED_DurGoods	Personal consumption expenditures: Durable goods (chain-type price index)
8	102	104	6	0	DSERRG3Q086SBEA	PCED_Serv	Personal consumption expenditures: Services (chain-type price index)
9	103	105	6	0	DNDGRG3Q086SBEA	PCED_NDurGoods	Personal consumption expenditures: Nondurable goods (chain-type price index)
10	104	106	6	0	DHCERG3Q086SBEA	PCED_HouseholdServ.	Personal consumption expenditures: Services: Household consumption expenditures (chain-type price index)
11	105	107	6	1	DMOTRG3Q086SBEA	PCED_MotorVec	Personal consumption expenditures: Durable goods: Motor vehicles and parts (chain-type price index)
12	106	108	6	1	DFDHRG3Q086SBEA	PCED_DurHousehold	Personal consumption expenditures: Durable goods: Furnishings and durable household equipment (chain-type price index)
13	107	109	6	1	DREQRG3Q086SBEA	PCED_Recreation	Personal consumption expenditures: Durable goods: Recreational goods and vehicles (chain-type price index)
14	108	110	6	1	DODGRG3Q086SBEA	PCED_OthDurGds	Personal consumption expenditures: Durable goods: Other durable goods (chain-type price index)
15	109	111	6	1	DFXARG3Q086SBEA	PCED_Food_Bev	Personal consumption expenditures: Nondurable goods: Food and beverages purchased for off-premises consumption (chain-type price index)
16	110	112	6	1	DCLORG3Q086SBEA	PCED_Clothing	Personal consumption expenditures: Nondurable goods: Clothing and footwear (chain-type price index)
17	111	113	6	1	DGOERG3Q086SBEA	PCED_Gas_Enrgy	Personal consumption expenditures: Nondurable goods: Gasoline and other energy goods (chain-type price index)
18	112	114	6	1	DONGRG3Q086SBEA	PCED_OthNDurGds	Personal consumption expenditures: Nondurable goods: Other nondurable goods (chain-type price index)
19	113	115	6	1	DHUTRG3Q086SBEA	PCED_Housing-Utilities	Personal consumption expenditures: Services: Housing and utilities (chain-type price index)
20	114	116	6	1	DHLCRG3Q086SBEA	PCED_HealthCare	Personal consumption expenditures: Services: Health care (chain-type price index)
21	115	117	6	1	DTRSRG3Q086SBEA	PCED_TransSvg	Personal consumption expenditures: Transportation services (chain-type price index)

Group 6: Prices, continued

	ID	SW ID	TCODE	SW FACTORS	FRED MNEMONIC	SW MNEMONIC	DESCRIPTION
22	116	118	6	1	DRCARG3Q086SBEA	PCED_RecServices	Personal consumption expenditures: Recreation services (chain-type price index)
23	117	119	6	1	DFSARG3Q086SBEA	PCED_FoodServ_Acc.	Personal consumption expenditures: Services: Food services and accommodations (chain-type price index)
24	118	120	6	1	DIFSRG3Q086SBEA	PCED_FIRE	Personal consumption expenditures: Financial services and insurance (chain-type price index)
25	119	121	6	1	DOTSRG3Q086SBEA	PCED_OtherServices	Personal consumption expenditures: Other services (chain-type price index)
26	120	122	6	0	CPIAUCSL	CPI	Consumer Price Index for All Urban Consumers: All Items (Index 1982-84=100)
27	121	123	6	0	CPILFESL	CPI_LFE	Consumer Price Index for All Urban Consumers: All Items Less Food & Energy (Index 1982-84=100)
28	122	124	6	0	WPSFD49207	PPI:FinGds	Producer Price Index by Commodity for Finished Goods (Index 1982=100)
29	123	125	6	0	PPIACO	PPI	Producer Price Index for All Commodities (Index 1982=100)
30	124	126	6	1	WPSFD49502	PPI:FinConsGds	Producer Price Index by Commodity for Finished Consumer Goods (Index 1982=100)
31	125	127	6	1	WPSFD4111	PPI:FinConsGds(Food)	Producer Price Index by Commodity for Finished Consumer Foods (Index 1982=100)
32	126	128	6	1	PPIIDC	PPI:IndCom	Producer Price Index by Commodity Industrial Commodities (Index 1982=100)
33	127	129	6	1	WPSID61	PPI:IntMat	Producer Price Index by Commodity Intermediate Materials: Supplies & Components (Index 1982=100)
34	128	133	5	1	WPU0531	Real Price:NatGas	Producer Price Index by Commodity for Fuels and Related Products and Power: Natural Gas (Index 1982=100)
35	129	134	5	1	WPU0561	Real Price:Oil	Producer Price Index by Commodity for Fuels and Related Products and Power: Crude Petroleum (Domestic Production) (Index 1982=100)
36	130	135	5	0	OILPRICEx	Real Crudeoil Price	Real Crude Oil Prices: West Texas Intermediate (WTI) - Cushing, Oklahoma (2012 Dollars per Barrel), deflated by Core PCE
37	205	n.a.	6	0	WPSID62		Producer Price Index: Crude Materials for Further Processing (Index 1982=100)
38	206	n.a.	6	0	PPICMM		Producer Price Index: Commodities: Metals and metal products: Primary nonferrous metals (Index 1982=100)
39	207	n.a.	6	0	CPIAPPSL		Consumer Price Index for All Urban Consumers: Apparel (Index 1982-84=100)
40	208	n.a.	6	0	CPITRNSL		Consumer Price Index for All Urban Consumers: Transportation (Index 1982-84=100)
41	209	n.a.	6	0	CPIMEDSL		Consumer Price Index for All Urban Consumers: Medical Care (Index 1982-84=100)
42	210	n.a.	6	0	CUSR0000SAC		Consumer Price Index for All Urban Consumers: Commodities (Index 1982-84=100)

Group 6: Prices, continued

	ID	SW ID	TCODE	SW FACTORS	FRED MNEMONIC	SW MNEMONIC	DESCRIPTION
43	211	n.a.	6	0	CUSR0000SAD		Consumer Price Index for All Urban Consumers: Durables (Index 1982-84=100)
44	212	n.a.	6	0	CUSR0000SAS		Consumer Price Index for All Urban Consumers: Services (Index 1982-84=100)
45	213	n.a.	6	0	CPIULFSL		Consumer Price Index for All Urban Consumers: All Items Less Food (Index 1982-84=100)
46	214	n.a.	6	0	CUSR0000SA0L2		Consumer Price Index for All Urban Consumers: All items less shelter (Index 1982-84=100)
47	215	n.a.	6	0	CUSR0000SA0L5		Consumer Price Index for All Urban Consumers: All items less medical care (Index 1982-84=100)
48	233	n.a.	6	0	CUSR0000SEHC		CPI for All Urban Consumers: Owners' equivalent rent of residences (Index Dec 1982=100)

Group 7: Earnings and Productivity

	ID	SW ID	TCODE	SW FACTORS	FRED MNEMONIC	SW MNEMONIC	DESCRIPTION
1	131	136	5	0	AHETPIx	Real AHE:PrivInd	Real Average Hourly Earnings of Production and Nonsupervisory Employees: Total Private (2012 Dollars per Hour), deflated by Core PCE
2	132	137	5	0	CES2000000008x	Real AHE:Const	Real Average Hourly Earnings of Production and Nonsupervisory Employees: Construction (2012 Dollars per Hour), deflated by Core PCE
3	133	138	5	0	CES3000000008x	Real AHE:MFG	Real Average Hourly Earnings of Production and Nonsupervisory Employees: Manufacturing (2012 Dollars per Hour), deflated by Core PCE
4	134	139	5	1	COMPRMS	CPH:Mfg	Manufacturing Sector: Real Compensation Per Hour (Index 2012=100)
5	135	140	5	1	COMPRNFB	CPH:NFB	Nonfarm Business Sector: Real Compensation Per Hour (Index 2012=100)
6	136	141	5	1	RCPHBS	CPH:Bus	Business Sector: Real Compensation Per Hour (Index 2012=100)
7	137	142	5	1	OPHMFG	OPH:mfg	Manufacturing Sector: Real Output Per Hour of All Persons (Index 2012=100)
8	138	143	5	1	OPHNFB	OPH:nfb	Nonfarm Business Sector: Real Output Per Hour of All Persons (Index 2012=100)
9	139	144	5	0	OPHPBS	OPH:Bus	Business Sector: Real Output Per Hour of All Persons (Index 2012=100)
10	140	145	5	0	ULCBS	ULC:Bus	Business Sector: Unit Labor Cost (Index 2012=100)
11	141	146	5	1	ULCMFG	ULC:Mfg	Manufacturing Sector: Unit Labor Cost (Index 2012=100)
12	142	147	5	1	ULCNFB	ULC:NFB	Nonfarm Business Sector: Unit Labor Cost (Index 2012=100)
13	143	148	5	1	UNLPNBS	UNLPay:nfb	Nonfarm Business Sector: Unit Nonlabor Payments (Index 2012=100)
14	216	n.a.	6	0	CES0600000008		Average Hourly Earnings of Production and Nonsupervisory Employees: Goods-Producing (Dollars per Hour)

Group 8: Interest Rates

	ID	SW ID	TCODE	SW FACTORS	FRED MNEMONIC	SW MNEMONIC	DESCRIPTION
1	144	149	2	1	FEDFUNDS	FedFunds	Effective Federal Funds Rate (Percent)
2	145	150	2	1	TB3MS	TB-3Mth	3-Month Treasury Bill: Secondary Market Rate (Percent)
3	146	151	2	0	TB6MS	TM-6MTH	6-Month Treasury Bill: Secondary Market Rate (Percent)
4	147	153	2	0	GS1	TB-1YR	1-Year Treasury Constant Maturity Rate (Percent)
5	148	154	2	0	GS10	TB-10YR	10-Year Treasury Constant Maturity Rate (Percent)
6	149	155	2	0	MORTGAGE30US	Mort-30Yr	30-Year Conventional Mortgage Rate© (Percent)
7	150	156	2	0	AAA	AAA Bond	Moody's Seasoned Aaa Corporate Bond Yield© (Percent)
8	151	157	2	0	BAA	BAA Bond	Moody's Seasoned Baa Corporate Bond Yield© (Percent)
9	152	158	1	1	BAA10YM	BAA_GS10	Moody's Seasoned Baa Corporate Bond Yield Relative to Yield on 10-Year Treasury Constant Maturity (Percent)
10	153	159	1	1	MORTG10YRx	MRTG_GS10	30-Year Conventional Mortgage Rate Relative to 10-Year Treasury Constant Maturity (Percent)
11	154	160	1	1	TB6M3Mx	tb6m_tb3m	6-Month Treasury Bill Minus 3-Month Treasury Bill, secondary market (Percent)
12	155	161	1	1	GS1TB3Mx	GS1_tb3m	1-Year Treasury Constant Maturity Minus 3-Month Treasury Bill, secondary market (Percent)
13	156	162	1	1	GS10TB3Mx	GS10_tb3m	10-Year Treasury Constant Maturity Minus 3-Month Treasury Bill, secondary market (Percent)
14	157	163	1	1	CPF3MTB3Mx	CP_Tbill Spread	3-Month Commercial Paper Minus 3-Month Treasury Bill, secondary market (Percent)
15	201	n.a.	2	0	GS5		5-Year Treasury Constant Maturity Rate
16	202	n.a.	1	0	TB3SMFFM		3-Month Treasury Constant Maturity Minus Federal Funds Rate
17	203	n.a.	1	0	T5YFFM		5-Year Treasury Constant Maturity Minus Federal Funds Rate
18	204	n.a.	1	0	AAAFFM		Moody's Seasoned Aaa Corporate Bond Minus Federal Funds Rate
19	225	n.a.	2	0	CP3M		3-Month AA Financial Commercial Paper Rate
20	226	n.a.	1	0	COMPAPFF		3-Month Commercial Paper Minus Federal Funds Rate

Group 9: Money and Credit

	ID	SW ID	TCODE	SW FACTORS	FRED MNEMONIC	SW MNEMONIC	DESCRIPTION
1	158	167	5	0	BOGMBASEREALx	Real Mbase	Monetary Base (Millions of 1982-84 Dollars), deflated by CPI
2	159	168	5	0	IMFSLx	Real InsMMF	Real Institutional Money Funds (Billions of 2012 Dollars), deflated by Core PCE
3	160	169	5	0	M1REAL	Real m1	Real M1 Money Stock (Billions of 1982-84 Dollars), deflated by CPI
4	161	170	5	0	M2REAL	Real m2	Real M2 Money Stock (Billions of 1982-84 Dollars), deflated by CPI
5	162	171	5	0	MZMREAL	Real mzm	Real MZM Money Stock (Billions of 1982-84 Dollars), deflated by CPI
6	163	172	5	1	BUSLOANSx	Real C&Lloand	Real Commercial and Industrial Loans, All Commercial Banks (Billions of 2012 U.S. Dollars), deflated by Core PCE
7	164	173	5	1	CONSUMERx	Real ConsLoans	Real Consumer Loans at All Commercial Banks (Billions of 2012 U.S. Dollars), deflated by Core PCE
8	165	174	5	1	NONREVSLx	Real NonRevCredit	Total Real Nonrevolving Credit Owned and Securitized, Outstanding (Billions of 2012 Dollars), deflated by Core PCE
9	166	175	5	1	REALLNx	Real LoansRealEst	Real Real Estate Loans, All Commercial Banks (Billions of 2012 U.S. Dollars), deflated by Core PCE
10	167	176	5	1	REVOLSLx	Real RevolvCredit	Total Real Revolving Credit Owned and Securitized, Outstanding (Billions of 2012 Dollars), deflated by Core PCE
11	168	177	5	0	TOTALSLx	Real ConsuCred	Total Consumer Credit Outstanding (Billions of 2012 Dollars), deflated by Core PCE
12	169	178	1	1	DRIWCIL	FRBSLO_Consumers	FRB Senior Loans Officer Opions. Net Percentage of Domestic Respondents Reporting Increased Willingness to Make Consumer Installment Loans
13	199	n.a.	6	0	TOTRESNS		Total Reserves of Depository Institutions (Billions of Dollars)
14	200	n.a.	7	0	NONBORRES		Reserves Of Depository Institutions, Nonborrowed (Millions of Dollars)
15	217	n.a.	6	0	DTCOLNVHFNM		Consumer Motor Vehicle Loans Outstanding Owned by Finance Companies (Millions of Dollars)
16	218	n.a.	6	0	DTCTHFNM		Total Consumer Loans and Leases Outstanding Owned and Securitized by Finance Companies (Millions of Dollars)
17	219	n.a.	6	0	INVEST		Securities in Bank Credit at All Commercial Banks (Billions of Dollars)

Group 10: Household Balance Sheets

	ID	SW ID	TCODE	SW FACTORS	FRED MNEMONIC	SW MNEMONIC	DESCRIPTION
1	170	179	5	0	TABSHNOx	Real HHW:TASA	Real Total Assets of Households and Nonprofit Organizations (Billions of 2012 Dollars), deflated by Core PCE
2	171	181	5	1	TLBSHNOx	Real HHW:LiabSA	Real Total Liabilities of Households and Nonprofit Organizations (Billions of 2012 Dollars), deflated by Core PCE
3	172	182	5	0	LIABPIx	liab_PDISA	Liabilities of Households and Nonprofit Organizations Relative to Personal Disposable Income (Percent)
4	173	183	5	1	TNWBSHNOx	Real HHW:WSA	Real Net Worth of Households and Nonprofit Organizations (Billions of 2012 Dollars), deflated by Core PCE
5	174	184	1	0	NWPIx	W_PDISA	Net Worth of Households and Nonprofit Organizations Relative to Disposable Personal Income (Percent)
6	175	185	5	1	TARESAx	Real HHW:TA_RESA	Real Assets of Households and Nonprofit Organizations excluding Real Estate Assets (Billions of 2012 Dollars), deflated by Core PCE
7	176	186	5	1	HNOREMQ027Sx	Real HHW:RESA	Real Real Estate Assets of Households and Nonprofit Organizations (Billions of 2012 Dollars), deflated by Core PCE
8	177	188	5	1	TFAABSHNOx	Real HHW:FinSA	Real Total Financial Assets of Households and Nonprofit Organizations (Billions of 2012 Dollars), deflated by Core PCE
9	224	n.a.	2	0	CONSPIx		Nonrevolving consumer credit to Personal Income

Group 11: Exchange Rates

	ID	SW ID	TCODE	SW FACTORS	FRED MNEMONIC	SW MNEMONIC	DESCRIPTION
1	182	193	5	1	TWEXMMTH	Ex rate:major	Trade Weighted U.S. Dollar Index: Major Currencies (Index March 1973=100)
2	183	194	5	1	EXUSEU	Ex rate:Euro	U.S. / Euro Foreign Exchange Rate (U.S. Dollars to One Euro)
3	184	195	5	1	EXSZUSx	Ex rate:Switz	Switzerland / U.S. Foreign Exchange Rate
4	185	196	5	1	EXJPUSx	Ex rate:Japan	Japan / U.S. Foreign Exchange Rate
5	186	197	5	1	EXUSUKx	Ex rate:UK	U.S. / U.K. Foreign Exchange Rate
6	187	198	5	1	EXCAUSx	EX rate:Canada	Canada / U.S. Foreign Exchange Rate

Group 12: Other

	ID	SW ID	TCODE	SW FACTORS	FRED MNEMONIC	SW MNEMONIC	DESCRIPTION
1	188	199	1	1	UMCSENTx	Cons. Expectations	University of Michigan: Consumer Sentiment (Index 1st Quarter 1966=100)
2	189	200	2	1	USEPUINDXM	PolicyUncertainty	Economic Policy Uncertainty Index for United States

Group 13: Stock Markets

	ID	SW ID	TCODE	SW FACTORS	FRED MNEMONIC	SW MNEMONIC	DESCRIPTION
1	178	189	1	1	VXOCLSx	VXO	CBOE S&P 100 Volatility Index: VXO
2	231	n.a.	5	0	NIKKEI225		Nikkei Stock Average
3	232	n.a.	5	0	NASDAQCOM		NASDAQ Composite (Index Feb 5, 1971=100)
4	245	180	5	0	S&P 500		S&P's Common Stock Price Index: Composite
5	246	n.a.	5	0	S&P: indust		S&P's Common Stock Price Index: Industrials
6	247	n.a.	2	0	S&P: div yield		S&P's Composite Common Stock: Dividend Yield
7	248	n.a.	5	0	S&P PE ratio		S&P's Composite Common Stock: Price-Earnings Ratio

Group 14: Non-Household Balance Sheets

	ID	SW ID	TCODE	SW FACTORS	FRED MNEMONIC	SW MNEMONIC	DESCRIPTION
1	192	n.a.	2	0	GFDEGDQ188S		Federal Debt: Total Public Debt as Percent of GDP (Percent)
2	193	n.a.	2	0	GFDEBTNx		Real Federal Debt: Total Public Debt (Millions of 2012 Dollars), deflated by PCE
3	234	n.a.	5	0	TLBSNNCBx		Real Nonfinancial Corporate Business Sector Liabilities (Billions of 2012 Dollars), Deflated by Implicit Price Deflator for Business Sector IPDBS
4	235	n.a.	1	0	TLBSNNCBBDIx		Nonfinancial Corporate Business Sector Liabilities to Disposable Business Income (Percent)
5	236	n.a.	5	0	TTAABSNNCBx		Real Nonfinancial Corporate Business Sector Assets (Billions of 2012 Dollars), Deflated by Implicit Price Deflator for Business Sector IPDBS
6	237	n.a.	5	0	TNWMVBSNNCBx		Real Nonfinancial Corporate Business Sector Net Worth (Billions of 2012 Dollars), Deflated by Implicit Price Deflator for Business Sector IPDBS
7	238	n.a.	2	0	TNWMVBSNNCBBDIx		Nonfinancial Corporate Business Sector Net Worth to Disposable Business Income (Percent)
8	239	n.a.	5	0	TLBSNNBx		Real Nonfinancial Noncorporate Business Sector Liabilities (Billions of 2012 Dollars), Deflated by Implicit Price Deflator for Business Sector IPDBS
9	240	n.a.	1	0	TLBSNNBBDIx		Nonfinancial Noncorporate Business Sector Liabilities to Disposable Business Income (Percent)
10	241	n.a.	5	0	TABSNNBx		Real Nonfinancial Noncorporate Business Sector Assets (Billions of 2012 Dollars), Deflated by Implicit Price Deflator for Business Sector IPDBS
11	242	n.a.	5	0	TNWBSNNBx		Real Nonfinancial Noncorporate Business Sector Net Worth (Billions of 2012 Dollars), Deflated by Implicit Price Deflator for Business Sector IPDBS
12	243	n.a.	2	0	TNWBSNNBBDIx		Nonfinancial Noncorporate Business Sector Net Worth to Disposable Business Income (Percent)
13	244	n.a.	5	0	CNCFx		Real Disposable Business Income, Billions of 2012 Dollars (Corporate cash flow with IVA minus taxes on corporate income, deflated by Implicit Price Deflator for Business Sector IPDBS)

Bibliography

S. C. Ahn and A. R. Horenstein. Eigenvalue ratio test for the number of factors. *Econometrica*, 81(3):1203–1227, 2013.

J. Bai. Inferential theory for factor models of large dimensions. *Econometrica*, 71(1): 135–171, 2003.

J. Bai and S. Ng. Determining the number of factors in approximate factor models. *Econometrica*, 70(1):191–221, 2002.

J. Breitung and J. Tenhofen. Gls estimation of dynamic factor models. *Journal of the American Statistical Association*, 106(495):1150–1166, 2011.

E. F. Fama and K. R. French. The capital asset pricing model: Theory and evidence. *Journal of economic perspectives*, 18(3):25–46, 2004.

J. Fan, Y. Liao, and M. Mincheva. High dimensional covariance matrix estimation in approximate factor models. *Annals of statistics*, 39(6):3320, 2011.

J. Fan, Y. Liao, and M. Mincheva. Large covariance estimation by thresholding principal orthogonal complements. *Journal of the Royal Statistical Society. Series B, Statistical methodology*, 75(4), 2013.

J. Fan, J. Guo, and S. Zheng. Estimating number of factors by adjusted eigenvalues thresholding. *Journal of the American Statistical Association*, 117(538):852–861, 2022.

J. Fan, Z. Lou, and M. Yu. Are latent factor regression and sparse regression adequate? *Journal of the American Statistical Association*, 0(0):1–77, 2023.

T. Hastie, R. Tibshirani, J. H. Friedman, and J. H. Friedman. *The elements of statistical learning: data mining, inference, and prediction*, volume 2. Springer, 2009.

J. Huang, S. Ma, and C.-H. Zhang. Adaptive lasso for sparse high-dimensional regression models. *Statistica Sinica*, pages 1603–1618, 2008.

R. Jagannathan, E. R. McGrattan, et al. The capm debate. *Federal Reserve Bank of Minneapolis Quarterly Review*, 19(4):2–17, 1995.

C. Lam and Q. Yao. Factor modeling for high-dimensional time series: inference for the number of factors. *The Annals of Statistics*, pages 694–726, 2012.

M. McCracken and S. Ng. Fred-qd: A quarterly database for macroeconomic research. Technical report, National Bureau of Economic Research, 2020.

J. H. Stock and M. W. Watson. Diffusion indexes, 1998.

J. H. Stock and M. W. Watson. Forecasting using principal components from a large number of predictors. *Journal of the American statistical association*, 97(460): 1167–1179, 2002b.

R. Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, 58(1):267–288, 1996.

W. B. Wu. Nonlinear system theory: another look at dependence. *Proceedings of the National Academy of Sciences of the United States of America*, 102(40):pp. 14150–14154, 2005.

W. B. Wu and Y. N. Wu. Performance bounds for parameter estimates of high-dimensional linear models with correlated errors. *Electronic Journal of Statistics*, 10(1):352–379, 2016.

H. Zou. The adaptive lasso and its oracle properties. *Journal of the American statistical association*, 101(476):1418–1429, 2006.