

# Sufficient Instruments Filter For Causal Discovery

Rajveer Jat\*

University of California Riverside

September 29, 2024

## Abstract

We introduce a novel five-layered deep learning-based tractable procedure to filter out sufficient information from many instruments for estimating parameters in regression models with endogenous regressors. Our method draws its merit from three key properties: the ability to incorporate supervision, the flexibility to accommodate non-linearity, and the capability for sufficient dimension reduction. We show that our method is consistent and asymptotically normal when many instruments are correlated. Simulation exercises show that this method consistently achieves lower bias and root mean squared error compared to competing benchmarks, across many specifications. We further validate our approach with two real-world applications in industrial organization and finance, yielding meaningful insights into causal relationships. Our method remains robust when the number of instruments exceeds the sample size, and performs well even with invalid observed instruments, as long as there exists at least one linear combination of common factors among the observed instruments that serves as a valid instrument.

**keywords:** Causal Inference, High-dimensions, Instrumental Variables, Dimension Reduction, Non-parametric, Supervised Learning.

---

\*Email: rjat001@ucr.edu.

# 1 Introduction

The instrumental variable (IV) approach is a cornerstone in addressing endogeneity issues in econometrics. When faced with a large set of instruments, two prominent strategies have emerged in the literature to harness them for causal inference. The first assumes sparsity, positing that only a subset of instruments is valid, as discussed in [Belloni \*et al.\* \(2012\)](#). The second approach, as in [Kapetanios & Marcellino \(2010\)](#) and [Bai & Ng \(2010\)](#), leverages the idea that instruments share common components. Both approaches, however, are restricted to a linear relationship between the endogenous regressor and the instruments.<sup>1</sup> While methods such as [Newey \(1990\)](#) have been developed to construct optimal instruments for nonlinear models, these non-parametric techniques can become computationally prohibitive when dealing with a large number of instruments. In this paper, we introduce a novel method that filters relevant information from numerous instruments (possibly more than the sample size) to achieve efficient estimation, accommodating both linear and nonlinear relationships between the endogenous regressor and instruments. Moreover, our approach permits the presence of weak and invalid instruments, provided that some linear combination(s) of their common components can serve as valid instruments, a weaker condition than [Bai & Ng \(2010\)](#).

A natural question is: Do we encounter many instruments in economics? The answer is yes, all the time. If  $z$  is a valid instrument, why not also consider its functional transformations, such as  $z^2$ ,  $\log z$ ,  $\sqrt{z}$ , among others? In the case of an  $\text{AR}(p)$  process, the  $(p+1)^{\text{th}}$  lag can serve as an instrument, so if  $z_{t-k}$  is an instrument, why not consider its lags, polynomial expansions, and interaction terms as well? Even without transformations, many linear instruments are not uncommon, as in industrial organization applications in [Berry \*et al.\* \(1995\)](#). [Olmstead \*et al.\* \(2007\)](#) is a similar work estimating price elasticities of water demand. Given that a wide array of instruments arises naturally, should we not make use of them? The answer depends on how we model the first stage. By restricting the first stage to a linear form, we impose a specific

---

<sup>1</sup>[Belloni \*et al.\* \(2012\)](#) incorporates non-linearities through a sieve or polynomial transformations but then instruments have to be low dimensional. Essentially their method is LASSO based which is linear.

functional structure, which may lead to model misspecification, thereby invalidating the inferences as discussed in [Newey \(1985\)](#). Misspecification is a major issue in economics discussed in seminal works such as [Lucas Jr \(1976\)](#). Hence, sufficiently leveraging the information from a large set of instruments is crucial for achieving efficient estimation of the structural parameters of interest.

Efficiently utilizing information from many instruments presents several key challenges. First, the relationship between endogenous regressors and instruments may be nonlinear and unknown, complicating model specification. Second, with many instruments, correlations among them can lead to the failure of the irrepressible condition (as noted by [Fan \*et al.\* \(2020\)](#)), resulting in sparsity-based LASSO-IV methods like [Belloni \*et al.\* \(2012\)](#) potentially selecting incorrect instruments. Non-parametric IV methods, such as [Newey \(1990\)](#), can address the first challenge. For the second, [Kapetanios & Marcellino \(2010\)](#) and [Bai & Ng \(2010\)](#) suggest using common components or factors as instruments. However, neither approach alone suffices in a more general setting when both the nonlinearity and instrument correlation are present.

A possible solution is to extract factors and estimate the endogenous regressor as a non-parametric function of these factors, a strategy that, to our knowledge, has yet to be explored. However, this introduces a third challenge: not all factors are necessarily relevant to the endogenous regressor. Including irrelevant factors can reduce efficiency, as observed in the forecasting literature (e.g., [Bai & Ng \(2008\)](#), [Kelly & Pruitt \(2015\)](#), [Fan \*et al.\* \(2017\)](#)), and also increases the dimensionality of the non-parametric function, thereby slowing convergence ([Pagan & Ullah \(1999\)](#)). The fourth challenge arises even if all factors are relevant—when their number is sizeable, as in [Fama & French \(1993\)](#) and [Fama & French \(2015\)](#), non-parametric estimation becomes slower due to convergence issues. Our proposed method seeks to address these challenges in a comprehensive and unified framework.

The issue of utilizing many instruments has been extensively studied, with a central concern

being that the 2SLS method may become inconsistent when faced with a large number of instruments (Bekker (1994)). Another key challenge is the potential presence of weak instruments, where many instruments exhibit only a weak correlation with the endogenous regressor. This phenomenon, known as the weak instrument problem, leads to invalid inference since the 2SLS estimator’s asymptotic distribution begins to resemble a Cauchy distribution, which is non-normal and has undefined moments (Phillips (1989), Staiger & Stock (1994)). When many instruments are sparse, Belloni *et al.* (2012)-like approach can address this problem through a selection of only the most relevant instruments.

However, in many macro-finance contexts, the assumption of sparsity may be less plausible, particularly when considering different functional forms of a variable or when lagged versions of a variable also serve as instruments. In such cases, instruments are likely to often share common components because they stem from the same sources of information, therefore exhibiting a factor structure, we discuss one such case in section-5.1. This factor structure introduces challenges for sparsity-based selection methods like LASSO, as the violation of the irrerepresentable condition can lead to incorrect instrument selection (Fan *et al.* (2020)). In a simulation setting with correlated instruments, we show that Belloni *et al.* (2012)’s sparsity-based approach cannot beat simple OLS, highlighting its limitations in settings where instruments are interrelated (see Table-32 in Appendix-B.4 for detailed results).

In economics and finance, when a large number of variables are correlated, factor models are commonly employed to capture the underlying structure (Chamberlain & Rothschild (1983)). For the case of correlated instruments, Bai & Ng (2010) developed a method to identify causal parameters of interest, under the assumption that a large number of instruments can be explained by a few unobservable factors. They demonstrated that these factors could serve as effective instruments. While their approach addresses the high dimensionality of the instrument set, it is limited to linear models and is unsupervised therefore may select factors that are irrelevant to the endogenous regressor introducing a source of inefficiency in the estimation

procedure.

This paper proposes a novel method that extends the applicability of factor-based approaches by accommodating both linear and non-linear relationships between endogenous regressors and instruments, even in high-dimensional settings. The non-linear aspect of our method mitigates the issue of functional form misspecification, while its capability to manage high-dimensional data allows for the optimal utilization of a large set of correlated instruments. Furthermore, since our approach is supervised, it avoids the inclusion of irrelevant factors, thereby enhancing the efficiency of the estimation procedure.

Our approach can be conceptualized as a five-layer deep learning architecture designed for tractability. In the first layer, we take  $N$  instruments as inputs, allowing  $N$  to exceed the sample size. The instruments can be noisy, weak, or even invalid, provided the underlying common components are identifiable. In the second layer, principal component analysis is employed to extract  $r$  common factors from the instruments, which serve as true instruments, meeting the necessary relevance and exclusion restrictions. The third layer employs sufficient dimension reduction to estimate the *Central Mean Subspace* (CMS) of the factors for the endogenous regressor, using the methodology introduced by Li (1991). This step further reduces the dimensionality of the factors. To elaborate, let  $x$  represent the endogenous regressor and  $\mathbf{f}$  the vector of factors serving as instruments. Suppose  $x \in \mathbb{R}$  and  $\mathbf{f} \in \mathbb{R}^r$  with joint cumulative distribution function  $F(x, \mathbf{f})$ . The conditional mean regression  $E(x | \mathbf{f})$  is the first moment of the conditional distribution of  $x$  given  $\mathbf{f}$ , but the broader goal is to understand how  $F(x | \mathbf{f})$  behaves as  $\mathbf{f}$  varies. To simplify this,  $\mathbf{f}$  can be replaced by  $L \leq r$  linear combinations of its components,  $\boldsymbol{\theta}'_1 \mathbf{f}, \dots, \boldsymbol{\theta}'_L \mathbf{f}$ , without losing information about  $F(x | \mathbf{f})$ . Thus, we have:

$$x \perp \mathbf{f} \mid \boldsymbol{\theta}' \mathbf{f},$$

where  $\boldsymbol{\theta}$  is an  $r \times L$  matrix. This formulation implies that the conditional distribution of

$x \mid \mathbf{f}$  depends on  $\mathbf{f}$  only through  $\boldsymbol{\theta}'\mathbf{f}$ , effectively reducing the dimensionality of the regression problem. If  $L < r$ , this significantly simplifies the regression. For example, when  $r = 5$  and the relationship between  $x$  and  $\mathbf{f}$  is linear,  $L = 1$  is sufficient to explain  $x$ , allowing a non-parametric model to be estimated efficiently. Even in non-linear cases, using  $L$  (with  $1 < L \leq r$ ) linear combinations of factors is more efficient than directly working with  $r$  factors. This supervised procedure is also advantageous even for linear models, as it filters the variations in the instruments relevant to  $x$ . For instance, with 100 instruments generated from five factors, but only three factors affecting  $x$ , our method identifies the three relevant factors, while methods like [Bai & Ng \(2010\)](#) would use all five, reducing efficiency. This advantage is highlighted in [Fan \*et al.\* \(2017\)](#) in the forecasting problem. Unlike unsupervised methods such as [Bai & Ng \(2010\)](#), which focus on the within-variance of the instrument set, our method leverages the covariance of the inverse regression function,  $E(\mathbf{z}_t \mid x_t)$ , ensuring that only factors relevant to  $x$  are considered. The fourth layer then estimates  $x$  as a non-parametric function of the  $L$  sufficient indices identified in the third layer, yielding  $\hat{x}$ , equivalent to the first stage in 2SLS. In the final layer, we perform linear regression using the exogenous variation  $\hat{x}$ , derived from the factors, in place of the endogenous regressor  $x$ , to estimate the causal parameter of interest.

To evaluate the performance of our method, we design several simulation experiments and compare our method against competitors. In the main simulation results, we report outcomes for three major designs, each dedicated to demonstrating performance gains due to three key properties: dimension reduction, supervision, and the ability to handle non-linearities. We keep both factors and errors serially correlated to better reflect real-world data. We show that the method outperforms competitors in the majority of cases where endogeneity is present. We also consider two empirical applications: one in industrial organization and the other in finance. The method yields meaningful insights, proving the applicability of our concepts to real-world problems.

For an observation, we use  $t$  subscript to indicate that our paper allows time series data. The method can equally be applied to the cross-section data. Further, we do not explicitly consider exogenous control variables. But if controls are necessary, one can always first get residuals of response and endogenous regressors by filtering out control variables, then use the respective residuals in our method, we demonstrate it through our empirical application in section-5.1. Vectors and matrices are represented by small and capital boldfaced letters respectively. Scalars are not boldfaced.

The remainder of the paper is organized as follows: Section 2 outlines the proposed procedure. Section 3 discusses the associated asymptotic theory. In Section 4, we evaluate the performance of the method through simulations. As a proof of concept, Section 5 applies our approach to real-world problems. Finally, Section 6 concludes the paper.

## 2 The Sufficient Instruments Filter

In this section, we introduce the structural framework, an overview of sufficient dimension reduction (SDR), and its relevance to the problem at hand. The identification assumptions are then outlined, followed by a detailed explanation of the estimation process across the various layers of the procedure. The section concludes with an algorithm summarizing the estimation process, alongside a discussion of the tuning parameters and their implications for the overall methodology.

### 2.1 Structural Framework

For  $t = 1, 2, \dots, T$  the dependent variable  $y_t$  and independent variable of interest  $x_t$  are endogenously related through a linear equation of the form 2.1. The endogeneity comes from the fact that  $E(x_t \varepsilon_t) \neq 0$ , which makes OLS estimates biased.  $\mathbf{f}_t$  is a  $r \times 1$  vector of fundamental variables which we call factors.  $\boldsymbol{\theta}_1, \dots, \boldsymbol{\theta}_L$  are  $r$  dimensional orthonormal vectors

called sufficient dimension reduction (SDR) directions ( $L \leq r$ ). Sufficient dimensions reduction (SDR) directions span the central subspace ( $S_{x|\mathbf{f}_t}$ ) required for conditional mean estimation of  $x_t$  (Cook (2009)) i.e. factors relate with  $x_t$  only through these SDR directions. Product of SDR directions with factors ( $\boldsymbol{\theta}'_1 \mathbf{f}_t, \dots, \boldsymbol{\theta}'_L \mathbf{f}_t$ ) are called SDR indices (more on this later) which are our instruments for  $x_t$  and which are allowed to be non-linearly related with  $x_t$  through the equation-2.2. However, the problem is that we neither observe the true instruments (SDR indices) nor factors, instead, we observe a large number of noisy versions of them,  $\{z_{it}\}$ , where  $i = 1, 2, \dots, N$ . For the clarity of the presentation, we refer to “noisy instruments”  $\{z_{it}\}$  as instruments,  $\mathbf{f}_t$  as factors, and to true instruments ( $\boldsymbol{\theta}'_1 \mathbf{f}_t, \dots, \boldsymbol{\theta}'_L \mathbf{f}_t$ ) as SDR indices for the rest of the paper to avoid any confusion. The number of such noisy instruments is  $N$  which can be large potentially more than the sample size.

$$y_t = \beta_0 + \beta_1 x_t + \varepsilon_t \quad (2.1)$$

$$x_t = m(\boldsymbol{\theta}'_1 \mathbf{f}_t, \dots, \boldsymbol{\theta}'_L \mathbf{f}_t) + e_t \quad (2.2)$$

$$z_{it} = \mathbf{b}'_i \mathbf{f}_t + u_{it}, \quad 1 \leq i \leq N, \quad 1 \leq t \leq T \quad (2.3)$$

$$f_{jt} = \gamma_j f_{jt-1} + v_{jt}, \quad 1 \leq j \leq r \quad (2.4)$$

For a  $j$ th factor,  $f_{jt}$  is the value at time  $t$  which is not observable. For  $i$ th instrument,  $z_{it}$  is the observed value at time  $t$ , we define  $\mathbf{z}_t = (z_{1t}, \dots, z_{Nt})$ , and  $T \times N$  matrix of instruments  $\mathbf{Z} = (\mathbf{z}_1, \mathbf{z}_2, \dots, \mathbf{z}_T)'$ .  $\mathbf{b}_i$  is an  $r \times 1$  vector of factor loadings for the instrument  $i$ , which in matrix form can be written as  $\mathbf{B} = (\mathbf{b}_1, \dots, \mathbf{b}_N)'$ .  $u_{it}$  is the error term or idiosyncratic noise in instrument  $i$  at time  $t$ , it can be represented in vector form as  $\mathbf{u}_t = (u_{1t}, \dots, u_{Nt})'$  and in matrix form  $\mathbf{U} = (\mathbf{u}_1, \mathbf{u}_2, \dots, \mathbf{u}_T)'$ .  $\mathbf{F} = (\mathbf{f}_1, \dots, \mathbf{f}_T)'$  is a  $T \times r$  matrix of factors. We can rewrite equation-2.3 in matrix form as:

$$\mathbf{Z} = \mathbf{FB}' + \mathbf{U}$$



Our parameter of interest is  $\beta = (\beta_0 \ \beta_1)'$ . In particular, we care about  $\beta_1$ , the causal effect of  $x_t$  on  $y_t$ . In equation (2.2),  $m(\cdot)$  is an unknown non-parametric function, and  $e_t$  is a stochastic error term that is independent of  $\mathbf{f}_t$  and  $u_{it}$ . One can see this model as a deep learning framework (Bengio *et al.* (2009)) which involves five layers of linear/nonlinear processes for dimension reduction and estimation with the added advantage of offering a scalable and explicit computational algorithm. Figure-1 presents the architecture of this procedure.

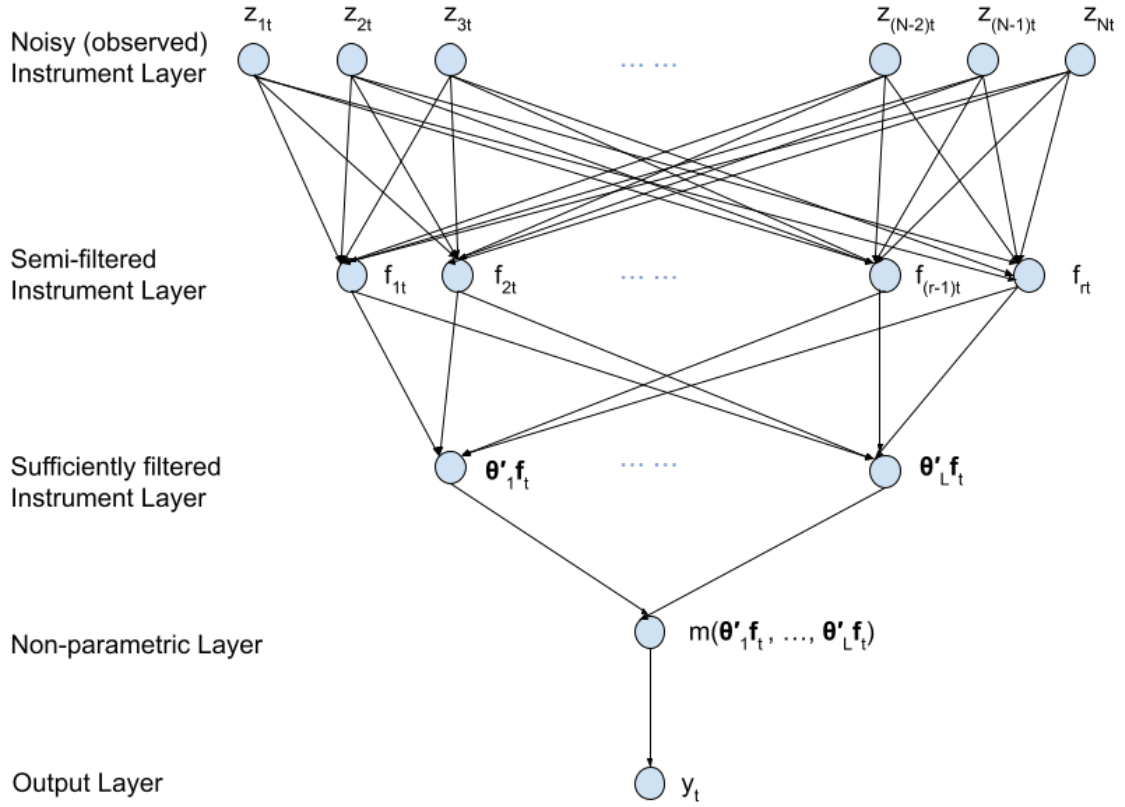


Figure 1: Architecture of the Sufficient Instrument Filter (SIF) estimation procedure

### 2.1.1 Sufficient Dimension Reduction

The model (2.2) says that the endogenous regressor  $x_t$  depends on the factors  $\mathbf{f}_t$  only through  $L$ -many indices  $\boldsymbol{\theta}'_1 \mathbf{f}_t, \dots, \boldsymbol{\theta}'_L \mathbf{f}_t$ , where  $L \leq r$ . For example, if factors are linearly related with  $x_t$ , then  $L = 1$  because a linear combination of factors can capture the underlying relationship. The main goal of the sufficient dimension reduction (SDR) procedure is to extract the directions of factors such that the directions are sufficient to find the best fit for  $x_t$ . In other words, when we pin down the span of the central subspace, there is nothing left in factors that can explain  $x_t$ , that is why this process is called sufficient dimension reduction.

While the individual directions  $\boldsymbol{\theta}_1, \dots, \boldsymbol{\theta}_L$  are not identifiable without imposing structural conditions on  $m(\cdot)$ , however, we just need the subspace  $S_{x|\mathbf{f}_t}$  spanned by  $\boldsymbol{\theta}_1, \dots, \boldsymbol{\theta}_L$ , which can be identified. Therefore, throughout this paper, we refer to any orthonormal basis  $\boldsymbol{\theta}_1, \dots, \boldsymbol{\theta}_L$  of the central subspace  $S_{x|\mathbf{f}_t}$  as sufficient dimension reduction directions, and their corresponding indices  $\boldsymbol{\theta}'_1 \mathbf{f}_t, \dots, \boldsymbol{\theta}'_L \mathbf{f}_t$  as sufficient indices.

This exercise effectively reduces the dimension of instruments from the diverging  $N$  to a fixed  $L$  to estimate the non-parametric function  $m(\cdot)$ , and thus greatly alleviates the curse of dimensionality making it possible for us to estimate the non-parametric function  $m(\cdot)$ . Further, we need  $L > 1$  only if there is a non-linear relationship between  $x_t$  and factors  $\mathbf{f}_t$ .

### 2.1.2 What is So Special in SDRs?

To uncover the causal effect of  $x_t$  on  $y_t$ , we should ideally be using a variable that can best explain  $x_t$  but is not endogenously related to  $y_t$  at the same time. The basic two-stages least square (2SLS) method is restricted to a linear regression of  $z_t$  on  $x_t$ , which is found to be not only inefficient but also inconsistent when the number of instruments is high (Bekker (1994)). When many instruments are available and potentially correlated, we have a high-dimensional method to obtain  $\hat{x}_t$ . In particular, we can use linear unsupervised PCA-based methods by Stock & Watson (2002), linear supervised methods Bai & Ng (2008), and Kelly & Pruitt (2015). These methods are based on Principal Component Regression (PCR) which is limited to using

the linear form of the factors<sup>2</sup>. What if the variable  $x_t$  is made up of non-linear combinations of the factors? One straightforward solution is to use non-parametric regression to obtain  $\hat{x}_t$ , however, it gets cursed by dimensionality if the true number of factors  $r$  increases. SDRs fuse the factors into  $L \leq r$  indices which uncovers the central subspace  $S_{x|\mathbf{f}_t}$ . The merit of the SDR-based method comes from three major advantages: first, it fuses factors into a smaller number of directions hence making it possible for us to estimate a non-parametric function  $m(\cdot)$  with a relatively much faster convergence rate. Second, it is a supervised method unlike PCA-based factor estimations of Bai & Ng (2010) therefore, it picks the directions in the instrument set  $Z$  which are relevant for the  $x_t$  more accurately. Third, it can capture the non-linear relationships between the  $x_t$  and  $\mathbf{f}_t$  through multiple SDR indices and  $m(\cdot)$ .

## 2.2 Identification of $\beta$

Instrument variable-based estimation requires an instrument to be a valid source of variation for the endogenous regressors. We call an instrument valid if it satisfies two conditions: relevancy and exclusion restriction. The relevancy means that our instruments are sufficiently related to the endogenous regressor. The exclusion restriction requires the instrument to affect the target only through the endogenous regressor. In this section, we pin down the conditions required for our procedure.

### 2.2.1 Relevancy Condition

Since the true factors are sufficient indices, the required relevancy condition is:

$$E\left[m(\boldsymbol{\theta}_1'\mathbf{f}_t, \dots, \boldsymbol{\theta}_L'\mathbf{f}_t)x_t\right] \neq 0 \quad (2.5)$$

This means that if any functional form of sufficient indices can explain the endogenous regressor  $x_t$ , the relevancy condition of our instruments will be satisfied. This is a much weaker condition

---

<sup>2</sup>Jat & Padha (2024) is a recent non-linear and supervised forecasting method.

than the previously required ones in the literature in two ways. The first is that unlike [Bai & Ng \(2010\)](#), we do not need all factors to be a valid instrument, as long as some linear combination(s) of them is(are) valid instrument(s), our method works. The second is that our relevancy condition does not require the linear form of the indices/factors to be related to the endogenous regressor, as long as they are related in any functional space, our relevancy condition is satisfied. The required condition of [Bai & Ng \(2010\)](#),  $E(f_{jt}x_t) = 0$  for all  $j = 1, 2, \dots, r$ , is not necessary for but is sufficient for our method. For example, take  $\theta_1$  as a vector of ones,  $\theta_2$  to  $\theta_L$  as a vector of zeros, and the function  $m(\cdot)$  as a linear function. This gives us  $E\left[m(\theta_1'f_t, \dots, \theta_L'f_t)x_t\right] = E\left[\theta_1'f_tx_t\right] = E\left[f_tx_t\right]$ .

### 2.2.2 Exclusion Restriction

The exclusion restriction means that the true instruments  $(\theta_1'f_t, \dots, \theta_L'f_t)$  should be able to affect the target  $y_t$  only through the endogenous regressor  $x_t$ . Translating it into an equation, the exclusion restriction we need is:

$$E\left[m(\theta_1'f_t, \dots, \theta_L'f_t)\varepsilon_t\right] = 0$$

In other words, the fundamental source of variation i.e. the factors should be independent of the error term in the equation-[2.1](#).

## 2.3 The Estimator

There are four steps involved in our estimation procedure. The first is to estimate factors  $f_t$  from the large pool of available instruments  $\{z_{it}\}$ ,  $i = 1, 2, \dots, N$ . There is a large literature on the estimation of factors using principal components such as [Stock & Watson \(2002\)](#) and [Bai \(2003\)](#). We follow the existing literature to consistently estimate factors using principal component analysis.

The second step is to estimate the SDR directions and indices:  $\hat{\theta}_1 \hat{\mathbf{f}}_t, \dots, \hat{\theta}_L \hat{\mathbf{f}}_t$ . We follow [Fan \*et al.\* \(2017\)](#) for estimating SDR directions using sliced inverse regression (SIR) which was originally developed by [Li \(1991\)](#). Most of our theory in SDR direction estimation is borrowed from [Fan \*et al.\* \(2017\)](#) and [Li \(1991\)](#). There exist other methods of SDR direction estimations such as parametric inverse regression (PIR) developed by [Bura & Cook \(2001\)](#). In this paper, we use SIR but for the sanity check, we verify the performance by replacing SIR with PIR.

The third step in our procedure is to estimate the non-parametric function  $m(\cdot)$  using SDR indices as arguments and  $x_t$  as the target variable (Eq-2.2). We use the local linear least square approach to estimate the non-parametric function  $m(\cdot)$ . The asymptotic properties of this method are extensively discussed in [Masry \(1996\)](#) and [Fan \(2018\)](#), we broadly use their results with some minor modifications to fit in our setting. As a result of the third step, we obtain  $\hat{x}_t = \hat{m}(\hat{\theta}_1 \hat{\mathbf{f}}_t, \dots, \hat{\theta}_L \hat{\mathbf{f}}_t)$ . This estimate is exogenous because it uses the fundamentally independent variation coming out of factors. This can be seen as the first stage of the two-stage least square (2SLS) estimator of instrument variable regression. The merit of our method comes from the fact that we use a procedure that can give us an exogenous  $\hat{x}_t$  which is very close to the true  $x_t$ .

The fourth and last step is to obtain the  $\hat{\beta}$  by using  $\hat{x}_t$  in place of  $x_t$  in equation 2.1. In this section, we describe our estimation procedure, the asymptotic theory of  $\hat{\beta}$  is developed in the section-3.

### 2.3.1 Estimation of Factors

Estimation of factors using principal components is an established literature. We temporarily assume that the number of underlying factors  $r$  is known to us. Consider the following constrained least squares problem:

$$\left(\widehat{\mathbf{B}}_r, \widehat{\mathbf{F}}_r\right) = \arg \min_{(\mathbf{B}, \mathbf{F})} \|\mathbf{Z} - \mathbf{B}\mathbf{F}'\|_F^2 \quad (2.5)$$

$$\text{subject to } T^{-1}\mathbf{F}'\mathbf{F} = \mathbf{I}_r, \quad \mathbf{B}'\mathbf{B} \text{ is diagonal} \quad (2.6)$$

Where  $\mathbf{Z} = (\mathbf{z}_1, \dots, \mathbf{z}_T)$ ,  $\mathbf{F}' = (\mathbf{f}_1, \dots, \mathbf{f}_T)$ , and  $\|\cdot\|_F$  denotes the Frobenius norm. This is a classical principal components problem, and it has been widely used to extract underlying common factors (Stock & Watson (2002), Bai & Ng (2002); Bai & Ng (2013)). The constraints in 2.6 correspond to the normalization. The minimizers  $\widehat{\mathbf{F}}_r$  and  $\widehat{\mathbf{B}}_r$  are such that the columns of  $\widehat{\mathbf{F}}_r/\sqrt{T}$  are the eigenvectors corresponding to the  $r$  largest eigenvalues of the  $T \times T$  matrix  $\mathbf{Z}'\mathbf{Z}$ , and  $\widehat{\mathbf{B}}_r = T^{-1}\mathbf{X}\widehat{\mathbf{F}}_r$ . To simplify notation, we use  $\widehat{\mathbf{B}} = \widehat{\mathbf{B}}_r$ ,  $\widehat{\mathbf{F}} = \widehat{\mathbf{F}}_r$ , and  $\{\widehat{\mathbf{f}}_1, \dots, \widehat{\mathbf{f}}_T\}$  throughout this paper. To choose  $r$ , we use Ahn & Horenstein (2013)'s method.

### 2.3.2 Estimation of SDR Directions

Our SDR direction estimation procedure fully leverages the information from the instrument, through the covariance matrix of the inverse regression function,  $E(\mathbf{z}_t | x_t)$ . This is a key difference from Bai & Ng (2010), which uses unsupervised and linear counterpart  $Cov(\mathbf{z}_t, x_t)$ . By conditioning on the target  $x_t$  in the model-2.3, we derive

$$\text{cov}(E(\mathbf{z}_t | x_t)) = \mathbf{B} \text{cov}(E(\mathbf{f}_t | x_t)) \mathbf{B}'$$

Where we used the assumption that  $E(\mathbf{u}_t | x_t) = 0$ . This salient feature does not impose any structure on the covariance matrix of  $\mathbf{u}_t$ . Li (1991) showed that  $E(\mathbf{f}_t | x_t)$  is contained in the central subspace  $S_{x|\mathbf{f}_t}$  spanned by  $\boldsymbol{\theta}_1, \dots, \boldsymbol{\theta}_L$  provided that  $E(\mathbf{b}'\mathbf{f}_t | \boldsymbol{\theta}'_1\mathbf{f}_t, \dots, \boldsymbol{\theta}'_L\mathbf{f}_t)$  is a linear function of  $\boldsymbol{\theta}'_1\mathbf{f}_t, \dots, \boldsymbol{\theta}'_L\mathbf{f}_t$  for any  $\mathbf{b} \in \mathbb{R}^r$ . This important result implies that  $S_{x|\mathbf{f}_t}$  contains the linear span of  $\text{cov}(E(\mathbf{f}_t | x_t))$ . Thus, it is promising to estimate sufficient directions by investigating the top  $L$  eigenvectors of  $\text{cov}(E(\mathbf{f}_t | x_t))$ . To see this, let  $\boldsymbol{\Theta} = (\boldsymbol{\theta}_1, \dots, \boldsymbol{\theta}_L)$ .

Then, we can write

$$E(\mathbf{f}_t | x_t) = \Theta \mathbf{a}(x_t)$$

for some  $L \times 1$  coefficient function  $a(\cdot)$ , according to the aforementioned result by Li (1991).

As a result,

$$\text{cov}(E(\mathbf{f}_t | x_t)) = \Theta E[\mathbf{a}(x_t) \mathbf{a}(x_t)^T] \Theta'$$

This matrix has  $L$  nonvanishing eigenvalues if  $E[\mathbf{a}(x_t) \mathbf{a}(x_t)^T]$  is non-degenerate. Their corresponding eigenvectors have the same linear span as  $\boldsymbol{\theta}_1, \dots, \boldsymbol{\theta}_L$ .

It is difficult to directly estimate the covariance of the inverse regression curve  $E(\mathbf{f}_t | x_t)$  and obtain sufficient predictive indices. Fortunately, it is possible if we slice the data into  $M$  small parts or slices and then estimate. Li (1991) developed a method to estimate the sliced covariance estimate. We use their method to estimate the inverse covariance of the underlying factors. The estimator is:

$$\Sigma_{\mathbf{f}|x} = \frac{1}{M} \sum_{s=1}^M E(\mathbf{f}_t | x_t \in I_s) E(\mathbf{f}_t' | x_t \in I_s)$$

where the range of  $x_t$  is divided into  $M$  slices  $I_1, \dots, I_M$  such that  $P(x_t \in I_s) = 1/M$ . By substituting  $E(\mathbf{f}_t | x_t) = \Theta \mathbf{a}(x_t)$ , we have

$$\Sigma_{\mathbf{f}|x} = \Theta \left[ \frac{1}{M} \sum_{s=1}^M E(\mathbf{a}(x_t) | x_t \in I_s) E(\mathbf{a}(x_t) | x_t \in I_s)^T \right] \Theta' \quad (2.6)$$

This matrix has  $L$  nonvanishing eigenvalues as long as the matrix in the bracket is non-degenerate. In this case, the linear span created by the eigenvectors of the  $L$  largest eigenvalues of  $\Sigma_{\mathbf{f}|x}$  is the same as that spanned by  $\boldsymbol{\theta}_1, \dots, \boldsymbol{\theta}_L$ . Note that  $M \geq \max\{L, 2\}$  is required for the matrix inside the bracket in equation-2.6 to be non-degenerate. To estimate  $\Sigma_{\mathbf{f}|x}$

with unobserved factors, we consistently estimate factors  $\mathbf{f}_t$  from the factor model 2.3 and then use the estimated factors  $\widehat{\mathbf{f}}_c$  and the observed target  $x_t$  to approximate the sliced estimate  $\Sigma_{\mathbf{f}|x}$ .

In Section-3, we shall show that under mild conditions,  $\Sigma_{\mathbf{f}|x}$  is consistently estimated by  $\widehat{\Sigma}_{\mathbf{f}|x}$  as  $N, T \rightarrow \infty$ . Furthermore, the eigenvectors of  $\widehat{\Sigma}_{\mathbf{f}|x}$  corresponding to the  $L$  largest eigenvalues, denoted as  $\widehat{\boldsymbol{\theta}}_j (j \equiv 1, \dots, L)$ , will converge to the corresponding eigenvectors of  $\Sigma_{\mathbf{f}|x}$ , which actually span the aforementioned central subspace  $S_{x|\mathbf{f}_t}$ . This will yield consistent estimates of sufficient indices  $\widehat{\boldsymbol{\theta}}_1' \widehat{\mathbf{f}}_t, \dots, \widehat{\boldsymbol{\theta}}_L' \widehat{\mathbf{f}}_t$ , the true instruments.

To effectively estimate  $\Sigma_{\mathbf{f}|x}$ , we shall replace the conditional expectations  $E(\mathbf{f}_t | x_t \in I_s)$  by its sample counterpart. Denote the ordered statistics of  $\left\{ \left( x_t, \widehat{\mathbf{f}}_t \right) \right\}_{t=1, \dots, T-1}$  by  $\left\{ \left( x_{(t)}, \widehat{\mathbf{f}}_{(t)} \right) \right\}_{t=1, \dots, T}$  according to the values of  $x$ , where  $x_{(2)} \leq \dots \leq x_{(T)}$ . We divide the range of  $x$  into  $M$  slices, where  $M$  is typically fixed. Each of the first  $M - 1$  slices contains the same number of observations  $c > 0$ , and the last slice may have fewer than  $c$  observations, which exerts little influence asymptotically. For ease of presentation, we introduce a double script  $_{(s,j)}$  in which  $s = 1, \dots, M$  refers to the slice number and  $j = 1, \dots, c$  is the index of an observation in a given slice. Thus, we can write  $\left\{ \left( x_t, \widehat{\mathbf{f}}_t \right) \right\}_{t=1, \dots, T}$  as

$$\left\{ \left( x_{(s,j)}, \widehat{\mathbf{f}}_{(s,j)} \right) : x_{(s,j)} = x_{(c(s-1)+j+1)}, \widehat{\mathbf{f}}_{(s,j)} = \widehat{\mathbf{f}}_{(c(s-1)+j)} \right\}_{s=1, \dots, M; j=1, \dots, c}.$$

Based on the estimated factors  $\widehat{\mathbf{f}}$ , we have the estimate  $\widehat{\Sigma}_{\mathbf{f}|x}$  in the form of

$$\widehat{\Sigma}_{\mathbf{f}|x} = \frac{1}{M} \sum_{s=1}^M \left[ \frac{1}{c} \sum_{j=1}^c \widehat{\mathbf{f}}_{(s,j)} \right] \left[ \frac{1}{c} \sum_{j=1}^c \widehat{\mathbf{f}}_{(s,j)} \right]'. \quad (2.7)$$



### 2.3.3 Estimation of Non-parametric Function $m(\cdot)$

Given the estimated low-dimensional SDR indices in the previous section, we can employ one of the well-developed nonparametric regression techniques to estimate  $m(\cdot)$  to obtain  $\hat{x}_t$ . For simplicity, we use the local linear least square regression (Fan & Gijbels (1992)) to estimate  $m(\cdot)$ . We postpone the discussion on this procedure to the section-3.2.3 where we also discuss its asymptotic properties.

### 2.3.4 Estimation of $\beta$

So far we have discussed how to estimate the exogenous variation counterpart of  $x_t$ , popularly known as the “first-stage” in the 2SLS method. Now we use the  $\hat{x}_t$  in equation-2.1 to estimate the parameter of interest  $\beta = (\beta_0 \ \beta_1)'$  using least squares. This is like the second stage in the two-staged least squares (2SLS) estimation procedure. We summarize the proposed estimation procedure in Algorithm-1.

<b>Algorithm 1</b> Sufficient Instrument Filter (SIF) Procedure	
Step 1	Obtain the estimated factors $\{\hat{\mathbf{f}}_t\}_{t=1,\dots,T}$ from 2.5 and 2.6.
Step 2	Construct $\hat{\Sigma}_{\mathbf{f} x}$ described in the equation-2.7.
Step 3	Obtain $\hat{\boldsymbol{\theta}}_1, \dots, \hat{\boldsymbol{\theta}}_L$ from the $L$ largest eigenvectors of $\hat{\Sigma}_{\mathbf{f} x}$ .
Step 4	Construct the predictive indices $\hat{\boldsymbol{\theta}}_1' \hat{\mathbf{f}}_t, \dots, \hat{\boldsymbol{\theta}}_L' \hat{\mathbf{f}}_t$ .
Step 5	Use the local linear least squared regression (Fan & Gijbels (1992)) to estimate $m(\cdot)$ with indices from Step 4, and hence to get $\hat{x}_t$ .
Step 6	Use the $\hat{x}_t$ obtained in step-5 in place of $x_t$ in equation-2.1 and do OLS to get $\hat{\beta} = (\hat{\beta}_0 \ \hat{\beta}_1)'$

### 2.3.5 Tuning Parameters

We need to tune three hyperparameters in our procedure, the number of factors  $r$ , the number of indices  $L$ , and the number of slices  $M$  to be considered for sliced inverse regression. The number of factors is estimated by the eigenvalue ratio test proposed by [Ahn & Horenstein \(2013\)](#). There are several tests available to choose the number of sufficient directions  $L$ ; these tests are discussed in [Li \(2018\)](#). [Fan \*et al.\* \(2017\)](#) observed that the number of slices  $M$  does not seem to matter much as long as it is greater than  $\max\{L, 2\}$ ; therefore, we set  $M = 10$ , the number used in [Fan \*et al.\* \(2017\)](#). We also verify that the performance of the method does not crucially depend on the choice of  $M$  in simulation exercises by putting  $M = \{8, 15\}$ .

Due to the presence of two tuning parameters in the sliced inverse regression (SIR) method of SDR direction estimation, a section of the literature raises questions about its robustness. We therefore cross-check its performance by replacing SIR with another SDR method called Parametric Inverse Regression (PIR), developed by [Bura & Cook \(2001\)](#), in our procedure. In simulation exercises (see [Appendix-B.3](#)), we verify that SIR's performance is similar to PIR. Therefore, the tuning parameters of SIR are not crucially affecting the method's performance.

## 3 Asymptotic Theory

In this section, we present the asymptotic results for the estimation steps and derive the asymptotic distribution of the estimator of interest,  $\hat{\beta}$ . We establish the convergence rates associated with each estimation step. Let's introduce the necessary notation for clarity, for a vector  $\mathbf{b}$ , let  $\|\mathbf{b}\|$  denote its Euclidean norm. For a matrix  $\mathbf{B}$ ,  $\|\mathbf{B}\|$  and  $\|\mathbf{B}\|_1$  represent the spectral norm and the  $\ell_1$  norm, respectively. The spectral norm is defined as the largest singular value of  $\mathbf{B}$ , while the  $\ell_1$  norm is the maximum absolute column sum. Additionally, the matrix  $\ell_\infty$  norm, denoted  $\|\mathbf{B}\|_\infty$ , is the maximum absolute row sum. For symmetric matrices, the  $\ell_1$  norm is equivalent to the  $\ell_\infty$  norm. We also define the smallest and largest eigenvalues

of a matrix as  $\lambda_{\min}(\cdot)$  and  $\lambda_{\max}(\cdot)$ , respectively.

### 3.1 Assumptions

In this section, we introduce four sets of assumptions. The first set covers the assumptions required for the identification of our causal parameter of interest  $\beta$ . Assumptions-2 is for the identification of factors, loadings, and SDR directions. Assumption-3 and 4 put structure on the data generating processes and error structures. Assumption-5 talks about the conditions required for consistently estimating a non-parametric relationship between the endogenous regressor and sufficient indices, the true instruments.

**Assumption 1.** (*Identification Assumptions*)

1.  $E\left[m(\boldsymbol{\theta}_1' \mathbf{f}_t, \dots, \boldsymbol{\theta}_L' \mathbf{f}_t)x_t\right] \neq 0$
2.  $E\left[m(\boldsymbol{\theta}_1' \mathbf{f}_t, \dots, \boldsymbol{\theta}_L' \mathbf{f}_t)\varepsilon_t\right] = 0$
3.  $E\left[m(\boldsymbol{\theta}_1' \mathbf{f}_t, \dots, \boldsymbol{\theta}_L' \mathbf{f}_t)e_t\right] = 0$
4.  $E(\mathbf{f}_t \mathbf{u}_t) = 0$
5. For all  $i$  and all  $t$ ,  $E(u_{it}e_t) = 0$  and  $E(u_{it}\varepsilon_t) = 0$
6.  $\mathbb{E}[\varepsilon_t] = 0$ , and  $\sqrt{T} \left( \frac{1}{T} \sum_{t=1}^T \varepsilon_t - E(\varepsilon_t) \right) = O_p(1)$  for all  $t$ .
7.  $\mathbb{E}[e_t] = 0$ , and  $\sqrt{T} \left( \frac{1}{T} \sum_{t=1}^T e_t - E(e_t) \right) = O_p(1)$  for all  $t$ .

The first and second assumptions of Assumption-1 are relevancy and exclusion restrictions respectively, required for the validity of the instrument. Relative to the nearest literature (Bai & Ng (2010)), our relevancy condition is weaker. The exclusion restriction is a sufficient condition for all functional forms of  $m(\cdot)$ , which may turn out to be a stronger one if the true underlying model is linear. However, when the true model is non-linear, unlike ours the exclusion condition of Bai & Ng (2010) is no longer sufficient.

The Assumption-1.3 is required by the construction of the non-parametric model. Similarly, the Assumption-1.4 is required by construction in the equation-2.3. The Assumption-1.5 says that the relationship between the instruments and the endogenous regressor is through the factors and not through error  $\mathbf{u}_t$ . The Assumption-1.6 and 1.7 ensure that the partial sum of errors grows at a rate proportional to  $\sqrt{T}$ . This assumption satisfies as long as the serial correlation decays sufficiently fast, the assumptions trivially hold if errors  $\varepsilon_t$  and  $e_t$  are i.i.d.  $E(e_t \varepsilon_t)$  are allowed to be non-zero hence introducing the  $x_t$  and  $y_t$  endogeneity in the model. All error terms are allowed to be serially correlated. The  $u_{it}$  are allowed to be both serially and cross-sectionally correlated, more structure on the errors is in the Assumption-3 and 4.

**Assumption 2. (*Factors, Loadings, SDR's Basic Assumption*)**

1. **Pervasive Condition:** The loadings  $\mathbf{b}_i$  satisfy  $\|\mathbf{b}_i\| \leq \mathcal{M}$  for  $i = 1, \dots, N$ . As  $N \rightarrow \infty$ , there exist two positive constants  $c_1$  and  $c_2$  such that:

$$c_1 < \lambda_{\min} \left( \frac{1}{N} \mathbf{B}' \mathbf{B} \right) < \lambda_{\max} \left( \frac{1}{N} \mathbf{B}' \mathbf{B} \right) < c_2$$

2. **Identification:**  $T^{-1} \mathbf{F}' \mathbf{F} = \mathbf{I}_K$ , and  $\mathbf{B}' \mathbf{B}$  is a diagonal matrix with distinct entries.
3. **Linearity:** The expectation  $E(\mathbf{b}' \mathbf{f}_t \mid \phi'_1 \mathbf{f}_t, \dots, \phi'_L \mathbf{f}_t)$  is a linear function of  $\phi'_1 \mathbf{f}_t, \dots, \phi'_L \mathbf{f}_t$  for any  $\mathbf{b} \in \mathbb{R}^N$ , where the vectors  $\phi'_i$  are derived from model 2.2.

Assumption 2.1 is commonly referred to as the pervasive condition, which ensures that the factors influence a substantial portion of the noisy instruments (Bai & Ng (2002)). Assumption 2.2 relates to the PC1 condition in Bai & Ng (2013), which removes rotational indeterminacy in the individual columns of  $\mathbf{F}$  and  $\mathbf{B}$ . Assumption 2.3, known as the linearity condition, is standard in dimension reduction literature. It holds when the distribution of  $\mathbf{f}_t$  is elliptically symmetric and is asymptotically justified when the dimension of  $\mathbf{f}_t$  is large, see Fan et al.

(2017) for references. Assumption 2.3 guarantees that the (centered) inverse regression curve  $E(\mathbf{f}_t | x_t)$  lies within the central subspace. Specifically, following Li (1991), we can derive the following lemma:

**Lemma 1.** *Under model 2.2 and Assumption 2.3, the centered inverse regression curve  $E(\mathbf{f}_t | x_t) - E(\mathbf{f}_t)$  is contained within the linear subspace spanned by  $\phi'_k \text{cov}(\mathbf{f}_t)$ , where  $k = 1, \dots, L$ .*

Lemma-1 forms the basis for sliced inverse regression. With this lemma, estimating the SDR directions without the knowledge of functional form  $m(\cdot)$  is possible.

We assume the data generation process is strongly mixing to ensure that the influence of past information gradually diminishes. Let  $\mathcal{F}_\infty^0$  and  $\mathcal{F}_T^\infty$  represent the  $\sigma$ -algebras generated by  $\{(\mathbf{f}_t, \mathbf{u}_t, e_t) : t \leq 0\}$  and  $\{(\mathbf{f}_t, \mathbf{u}_t, e_t) : t \geq T\}$  respectively. Define the mixing coefficient as:

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

**Assumption 3.** *(Data Generating Process)  $\{\mathbf{f}_t\}_{t \geq 1}$ ,  $\{\mathbf{u}_t\}_{t \geq 1}$ , and  $\{e_t\}_{t \geq 1}$  are strictly stationary processes and mutually independent. Additionally,  $E\|\mathbf{f}_t\|^4 < \infty$  and  $E(\|\mathbf{f}_t\|^2 | x_t) < \infty$ . For some positive constant  $c$ , the mixing coefficient  $\alpha(T) < c\rho^T$  for all  $T \in \mathbb{Z}^+$  and some  $\rho \in (0, 1)$ . Structural errors  $e_t$  and  $\varepsilon_t$  are allowed to be correlated, which is a source of endogeneity in the model.*

The independence between  $\{\mathbf{f}_t\}_{t \geq 1}$  and  $\{\mathbf{u}_t\}_{t \geq 1}$  (or  $\{e_t\}_{t \geq 1}$ ) aligns with Assumption A(d) in Bai & Ng (2013). Independence between  $\{\mathbf{u}_t\}_{t \geq 1}$  and  $\{e_t\}_{t \geq 1}$  can be relaxed to reflect a more realistic data generation process. For example, assuming  $E(\mathbf{u}_t | x_t) = 0$  for  $t \geq 1$  suffices for the theory to hold. However, we retain this simplified assumption for clarity. To ensure boundedness

Moreover, we impose the following conditions on the residuals and dependencies in the factor model-2.3, similar to those in Bai (2003), ensuring consistent estimation of common factors and factor loadings.

**Assumption 4.** (*Residuals and Dependence*) *There exists a positive constant  $\mathcal{M} < \infty$ , independent of  $N$  and  $T$ , such that:*

1.  $E(\mathbf{u}_t) = \mathbf{0}$ , and  $E|u_{it}|^8 \leq \mathcal{M}$ .
2.  $\|\Sigma_u\|_1 \leq \mathcal{M}$ , and for every  $i, j, t, s > 0$ ,  $(NT)^{-1} \sum_{i,j,t,s} |E(u_{it}u_{js})| \leq \mathcal{M}$
3. For every  $(t, s)$ ,  $E|N^{-1/2}(\mathbf{u}'_s \mathbf{u}_t - E(\mathbf{u}_s \mathbf{u}_t))|^4 \leq \mathcal{M}$ .

The Assumption-4 allows idiosyncratic errors and factors to be serially correlated but not too strongly. These assumptions on errors and their dependence are standard in the factor estimation literature, one can refer to Bai & Ng (2010) or Fan *et al.* (2017) for further detail.

## 3.2 Results

We prove the consistencies of the intermediate steps involved in our procedure. Since we do inferences on the parameter  $\beta = (\beta_0 \ \beta_1)$ , we characterize its asymptotic distribution.

### 3.2.1 Consistency of Factor Estimation

Define  $\omega_{N,T} = N^{-1/2} + T^{-1/2}$ .

**Lemma 2.** *under Assumptions 2.1, 2.2, 3 and 4, we have the following*

$$\frac{1}{T} \sum_{t=1}^T \|\hat{\mathbf{f}}_t - \mathbf{H}\mathbf{f}_t\|^2 = O_p(\omega_{N,T}^2)$$

*Proof.* This result is proved in Theorem 1 of Bai & Ng (2002). □

This lemma establishes the estimated factor(s) convergence to the true factors up to a rotation. It is well known in the literature on factor models<sup>3</sup>, that true underlying factor(s) are not identifiable; we instead estimate a rotated version of the true factors, which preserves their span. One important point to emphasize is that we use the indices of these factors. Therefore, the indices will automatically rotate back the factors to span the central mean subspace of the  $x_t$ .

### 3.2.2 Consistency of SDR Directions Estimation

The subsequent result details the rate at which the sliced covariance estimate of the inverse regression curve (i.e.,  $\widehat{\Sigma}_{\mathbf{f}|x}$  as defined in 2.7) converges under the spectral norm. It also suggests a similar convergence rate for the estimated SDR directions corresponding to sufficient indices. For simplicity, the number of factors  $r$  and the number of slices  $M$  are assumed to be constant, this assumption facilitates a faster convergence rate but is not crucial.

**Lemma 3.** *Assuming Assumptions 2.1-2.3 are satisfied and letting  $\omega_{N,T} = N^{-1/2} + T^{-1/2}$ , then under the model-2.2 and its corresponding factor model 2.3, it holds that*

$$\left\| \widehat{\Sigma}_{\mathbf{f}|x} - \Sigma_{\mathbf{f}|x} \right\| = O_p(\omega_{N,T})$$

*If the  $L$  largest eigenvalues of  $\Sigma_{\mathbf{f}|x}$  are positive and distinct, the eigenvectors  $\widehat{\theta}_1, \dots, \widehat{\theta}_L$  corresponding to these  $L$  largest eigenvalues of  $\widehat{\Sigma}_{\mathbf{f}|x}$  provide a consistent estimate of the directions  $\theta_1, \dots, \theta_L$ , with rates*

$$\left\| \widehat{\theta}_j - \theta_j \right\| = O_p(\omega_{N,T})$$

*for  $j = 1, \dots, L$ , where  $\theta_1, \dots, \theta_L$  form an orthonormal basis for the central subspace  $S_{x|\mathbf{f}}$  spanned by  $\phi_1, \dots, \phi_L$ .*

*Proof.* This result is proved in Theorem 3.1 of Fan *et al.* (2017). One need to replace their  $x_{it}$

---

<sup>3</sup>This feature of inherent unidentifiability has been emphasized in Bai (2003) among other papers. The normalization imposed in assumption 2.2 is done to handle this issue.

with our  $z_{it}$  and their  $y_{t+1}$  with our  $x_t$ . □

**Corollary 1** Under the same conditions of Lemma-3, for any  $j = 1, 2, \dots, L$ , we have

$$\widehat{\boldsymbol{\theta}}_j' \widehat{\mathbf{f}}_t \xrightarrow{p} \boldsymbol{\theta}_j' \mathbf{f}_t$$

This corollary states that the sufficient indices can be consistently estimated as a consequence of the Lemma-3. For the proof, one can refer to [Fan \*et al.\* \(2017\)](#).

### 3.2.3 Consistency of Non-parametric Function Estimation

For notational simplicity, let's denote SDR indices  $\boldsymbol{\theta}_k \mathbf{f}_t$  by  $w_k$ , where  $k = 1, 2, \dots, L$ . For a given  $L$  dimensional point  $\mathbf{w} = \{w_1, w_2, \dots, w_L\}$  and a vector of bandwidths  $\mathbf{h} = \{h_1, h_2, \dots, h_L\}$ ,  $\boldsymbol{\psi}_t = \left( \frac{\mathbf{w}_t - \mathbf{w}}{\mathbf{h}} \right) = \left( \frac{w_{t1} - w_1}{h_1}, \dots, \frac{w_{tL} - w_L}{h_L} \right) = (\psi_{t1}, \dots, \psi_{tL})$  is local deviation from it. We define the  $L$ -dimensional kernel function to be the product of kernels with individual arguments. It weights an observation inversely based on its distance from our point  $\mathbf{w}$ :

$$\mathcal{K}(\boldsymbol{\psi}) = \mathcal{K}(\psi_1) \times \dots \times \mathcal{K}(\psi_L)$$

While different bandwidth for different variables makes sense, in our setting variables are scaled, therefore for simplicity, we use the same bandwidth for all SDR indices i.e.  $h_1 = h_2 = \dots = h_L = h$ . the joint density  $g(x, w_1, \dots, w_L)$  of our  $x_t$  and SDR indices can be given by the following expression:

$$g(x, w_1, \dots, w_L) = \frac{1}{Th^{L+1}} \sum_{t=1}^T \mathcal{K}\left(\frac{x_t - x}{h}\right) \mathcal{K}\left(\frac{w_{t1} - w}{h}\right) \times \dots \times \mathcal{K}\left(\frac{w_{tL} - w}{h}\right)$$



Similarly, the joint density of the SDR indices is given by:

$$\begin{aligned}
g(\mathbf{w}) = g(w_1, \dots, w_L) &= \frac{1}{Th^L} \sum_{t=1}^T \mathcal{K}\left(\frac{w_{t1} - w}{h}\right) \times \dots \times \mathcal{K}\left(\frac{w_{tL} - w}{h}\right) \\
&= \frac{1}{Th^L} \sum_{t=1}^T \mathcal{K}(\psi_{t1}) \times \dots \times \mathcal{K}(\psi_{tL}) \\
&= \frac{1}{Th^L} \sum_{t=1}^T \mathcal{K}(\psi_t)
\end{aligned}$$

One can obtain the expression for conditional density of  $x$  given SDR indices  $g(x \mid w_1, \dots, w_L)$  by dividing the two expressions above. Using the definition of conditional mean which is  $m(\mathbf{w}) = \int x_t g(x_t \mid \mathbf{w}) dx_t$ . We use the local linear least square (LLLS) method for estimating  $m(\mathbf{w})$ . Masry (1996) and Fan (2018) discuss asymptotic properties of multivariate non-parametric estimation in time series data. Assumption-5 lists the assumptions required for consistent estimation of the non-parametric function  $m(\cdot)$ .

**Assumption 5.** (*Kernel, Smoothness of  $m(\cdot)$ , Moments, Bandwidth*)

1. **Smoothness of  $m(\cdot)$ :**  $m(\mathbf{w})$  is twice continuously differentiable, and the second derivatives are bounded:

$$\sup_{\mathbf{w}} \left| \frac{\partial^2 m(\mathbf{w})}{\partial w_i \partial w_j} \right| < \infty, \quad \text{for all } i, j \in \{1, \dots, L\}.$$

2. **Stationarity:** The process  $\{(\mathbf{w}_t, e_t)\}$  is strictly stationary and ergodic.
3. **Mixing Condition:** The sequence  $\{(\mathbf{w}_t, e_t)\}$  satisfies an  $\alpha$ -mixing condition with mixing coefficients  $\alpha(k)$  that decay sufficiently fast, i.e. for some  $\delta > 0$ :

$$\sum_{k=1}^{\infty} \alpha(k)^{\delta/(2+\delta)} < \infty$$

4. **Moment Conditions:** The error term  $e_t$  has finite second moment  $\mathbb{E}[e_t^2] = \sigma^2$  and may follow an autoregressive process. The covariates  $\mathbf{w}_t$  have bounded moments of order  $2 + \delta'$  for some  $\delta' > 0$ :

$$\mathbb{E}[\|\mathbf{w}_t\|^{2+\delta'}] < \infty.$$

5. **Kernel Function:** The kernel function  $K(\cdot)$  is a symmetric, bounded, and integrable function with compact support, satisfying:

$$\int K(\boldsymbol{\psi}) d\boldsymbol{\psi} = 1, \quad \int \boldsymbol{\psi} K(\boldsymbol{\psi}) d\boldsymbol{\psi} = 0, \quad 0 < \int \boldsymbol{\psi} \boldsymbol{\psi}^\top K(\boldsymbol{\psi}) d\boldsymbol{\psi} = \boldsymbol{\kappa}_2 < \infty.$$

6. **Bandwidth:** The bandwidth  $h$  depends on the sample size  $T$  and satisfies:

$$h \rightarrow 0, \quad Th^L \rightarrow \infty \quad \text{as } T \rightarrow \infty.$$

Specifically,  $h$  is chosen such that  $Th^{L+4} \rightarrow 0$  as  $T \rightarrow \infty$ .

The Assumption-5.1 means that the non-parametric function  $m(\cdot)$  is twice continuously differentiable. Bounding the second derivative ensures that there is no abrupt change in the function. The Assumption-5.2 and 5.3 controls the serial dependence of errors. They are explicitly mentioned but can be inferred from the Assumption-3. For the central limit theorem to hold in non-parametric estimation, we need  $(2 + \delta)^{th}$  moment for some  $\delta > 0$  to hold, which is what Assumption-5.4 states. Assumption-5.5 puts structure on kernel function. In particular, it should integrate to one which is analogous to the sum of weights adding up to one in a discrete case. A symmetric kernel function ensures that the first moment is zero i.e.  $\int \boldsymbol{\psi} K(\boldsymbol{\psi}) d\boldsymbol{\psi} = 0$ , this helps us in simplifying the Taylor series expansion to prove the asymp-

otic theory.  $0 < \int \boldsymbol{\psi} \boldsymbol{\psi}^\top K(\boldsymbol{\psi}) d\boldsymbol{\psi} = \boldsymbol{\kappa}_2 < \infty$  ensures that the second moment is bounded which appears in the asymptotic distribution of the estimator of  $m(\cdot)$ . The last assumption Assumption-5.6 ensures that the bandwidth selection should be done in a way that there is enough sample size available for consistent estimation. We now formally state the asymptotic results of the non-parametric estimation step.

**Lemma 4.** *Under the Assumptions 5.1-5.6, the local linear estimator  $\hat{m}(\mathbf{w})$  of the non-parametric function  $m(\mathbf{w})$ , has the following asymptotic properties:*

1. **Bias:**

$$\mathbb{E}[\hat{m}(\mathbf{w})] - m(\mathbf{w}) = \frac{1}{2} h^2 \text{tr}(\mathbf{G}) + o(h^2),$$

where  $\mathbf{G} = \nabla^2 m(\mathbf{w}) \cdot \int \boldsymbol{\psi} \boldsymbol{\psi}^\top K(\boldsymbol{\psi}) d\boldsymbol{\psi}$ ,  $\nabla^2 m(\mathbf{w})$  is the Hessian matrix of second derivatives of  $m(\mathbf{w})$ .

2. **Variance:**

$$\text{Var}(\hat{m}(\mathbf{w})) = \frac{\sigma^2}{Th^L g(\mathbf{w})} \int K^2(\boldsymbol{\psi}) d\boldsymbol{\psi} + o\left(\frac{1}{Th^L}\right),$$

where  $g(\mathbf{w})$  is the joint density of the covariates  $\mathbf{w}_t$  at point  $\mathbf{w}$ , and  $\sigma^2$  is the variance of the error term  $e_t$ .

3. **Asymptotic Normality:**

$$\sqrt{Th^L} \left( \hat{m}(\mathbf{w}) - m(\mathbf{w}) - \frac{1}{2} h^2 \text{tr}(\mathbf{G}) \right) \xrightarrow{d} \mathcal{N} \left( 0, \frac{\sigma^2}{g(\mathbf{w})} \int K^2(\boldsymbol{\psi}) d\boldsymbol{\psi} \right),$$

where  $\xrightarrow{d}$  denotes convergence in distribution,  $\text{tr}(\cdot)$  denotes the trace of a matrix, and the integrals are taken over the multivariate space of  $\boldsymbol{\psi}$ .

*Proof.* The proof is given in the appendix -A.1.1 □

### 3.2.4 Main Result: Consistency and Asymptotic Normality of $\hat{\beta}$

We make inferences on the estimator  $\beta$ , therefore, we derive the asymptotic normality result of the same in this section.

**Theorem 1.** (*Consistency of  $\hat{\beta}$* ) Define  $\delta_{NT} = N^{-1/2} + T^{-2/L+4}$ . Under the assumptions [1-5](#),

$$\left(\hat{\beta} - \beta\right) = O_p(\delta_{NT})$$

where  $L$  is the number of sufficient dimension reduction directions.

*Proof.* We prove this result in appendix-[A.1.2](#) □

We note that the convergence rate of our estimator is slower than the parametric rate of  $\sqrt{T}$  given that  $L \geq 1$ . If the underlying true model is linear then we have  $L = 1$  and the convergence rate is  $T^{0.4}$ . This is the cost we need to pay to accommodate the features of our method. Having more instruments is not a problem in our case it helps us get a faster rate if  $T$  is large. Further, having a rate of  $T^{-2/L+4}$  does not necessarily mean that our method will be slow, it's a theoretical guarantee that the rate of convergence won't be slower than this. Our method in practice may turn out to be faster than the rate stated in this theorem.

**Theorem 2.** (*Asymptotic Normality of  $\hat{\beta}$* )

Define  $\phi_{NT} = \min\{N^{\frac{1}{2}}, T^{\frac{2}{L+4}}\}$  and  $Q_t = [1 \quad m(\mathbf{w}_t)]$ . Then, under Assumption-[1-5](#), we have,

$$\phi_{NT}(\hat{\beta} - \beta) \xrightarrow{d} \mathcal{N}\left(0, \left(\frac{1}{T}Q'Q\right)^{-1} \frac{\mathbf{V}}{\delta_{NT}^2} \left(\frac{1}{T}Q'Q\right)^{-1}\right),$$

where,  $\mathbf{V}$  is a  $2 \times 2$  matrix, the two columns of  $\mathbf{V}$  are given as:

$$\mathbf{V}_1 = \begin{bmatrix} \beta_1^2 \left(\sum_{t=1}^T (\hat{m}(\hat{\mathbf{w}}_t) - m(\mathbf{w}_t))\right)^2 \\ \beta_1 \left(\sum_{t=1}^T (\hat{m}(\hat{\mathbf{w}}_t) - m(\mathbf{w}_t))\right) \left(\beta_0 \sum_{t=1}^T (\hat{m}(\hat{\mathbf{w}}_t) - m(\mathbf{w}_t)) + \beta_1 \sum_{t=1}^T (\hat{m}(\hat{\mathbf{w}}_t)^2 - m(\mathbf{w}_t)^2)\right) \end{bmatrix}$$

and

$$\mathbf{V}_2 = \begin{bmatrix} \beta_1 \left( \sum_{t=1}^T (\hat{m}(\hat{\mathbf{w}}_t) - m(\mathbf{w}_t)) \right) \left( \beta_0 \sum_{t=1}^T (\hat{m}(\hat{\mathbf{w}}_t) - m(\mathbf{w}_t)) + \beta_1 \sum_{t=1}^T (\hat{m}(\hat{\mathbf{w}}_t)^2 - m(\mathbf{w}_t)^2) \right) \\ \left( \beta_0 \sum_{t=1}^T (\hat{m}(\hat{\mathbf{w}}_t) - m(\mathbf{w}_t)) + \beta_1 \sum_{t=1}^T (\hat{m}(\hat{\mathbf{w}}_t)^2 - m(\mathbf{w}_t)^2) \right)^2 + \left( \sum_{t=1}^T (\hat{m}(\hat{\mathbf{w}}_t) - m(\mathbf{w}_t)) y_t \right)^2 \end{bmatrix}$$

*Proof.* The proof of this theorem is given in the Appendix-[A.1.3](#)  $\square$

The elements of matrix  $\mathbf{V}$  contain  $\beta_0$  and  $\beta_1$  and  $m(\mathbf{w}_t)$  which are not observable. However, we have their consistent estimators as proved in Theorem-[1](#) and Lemma-[4](#) respectively. Therefore we can get a consistent estimator of the variance expression stated here.

Since  $\phi_{NT} = \min\{N^{\frac{1}{2}}, T^{\frac{2}{L+4}}\}$ , therefore we need both  $N$  and  $T$  to approach  $\infty$  for our asymptotic results. If  $N$  is large then the rate is dominated by  $T$  and is slower than the parametric  $\sqrt{T}$  rate. This is the price we need to pay for the appealing features of our method.

### 3.2.5 Bootstrap for Confidence Intervals

Since the expression of the asymptotic variance of our estimator is complicated, we suggest an alternate method for the construction of the confidence intervals. Since our data allows serial correlation, bootstrap methods that preserve the serial dependence structure e.g. Block-bootstrap ([Carlstein \(1986\)](#)) should be used. The assumptions stated in our paper for dependent data satisfy the assumptions required for the block bootstrap introduced by [Carlstein \(1986\)](#). Many other bootstrap procedures for dependent data is discussed in [Lahiri \(2013\)](#).

## 4 Simulation

We evaluate the performance of our method across various data-generating processes (DGPs). Three distinct designs are considered to highlight the core strengths of our approach: supervised learning, dimension reduction, and the ability to capture non-linearities. For each scenario, we present both the expected value and root mean squared error (RMSE) of our

estimator in comparison to alternative methods. The bias can be derived from the expected value by subtracting the true value, which is set to 2 in all simulations.

We compare our method against three competitors: Ordinary Least Squares (OLS), Two-Stage Least Squares (2SLS), and the Factor Instrumental Variable (FIV) estimator from [Bai & Ng \(2010\)](#). The Post-Lasso IV (PLIV) estimator from [Belloni \*et al.\* \(2012\)](#) is not a suitable competitor because it is designed for settings with sparse instruments. As shown in Table-32, the [Belloni \*et al.\* \(2012\)](#)'s method does worse than simple OLS across various designs, and therefore, we do not include this method in comparisons.

#### 4.1 Design-I : Demonstrating Supervision Capability

Our main goal in this subsection is to show the supervising ability of our method. Therefore, we will keep  $m(\cdot)$  linear, a ground playing field for both our and competing methods. We will set the true number of factors that drive the instrument set equal to five but only three of them will be relevant for the endogenous regressor  $x_t$ . Therefore our supervised method should be able to achieve better performance relative to the unsupervised methods such as [Bai & Ng \(2010\)](#) because of the ability to filter out the factors relevant for  $x_t$ .

##### 4.1.1 The Data Generating Process

The broader structure of this design is an approximate factor model in the sense that endogenous regressor  $x_t$  is linearly related to the latent factors  $\mathbf{f}_t$ . This is the setting that [Bai & Ng \(2010\)](#) considers in their paper except the fact that not all factors are relevant for  $x_t$ . Capital asset price model (CAPM) is one such example in the literature ([Campbell & Shiller](#)

(1988); Polk *et al.* (2006)). The data-generating process is:

$$y_t = \beta_0 + \beta_1 x_t + \varepsilon_t \quad (4.1)$$

$$x_t = \phi' \mathbf{f}_t + e_t \quad (4.2)$$

$$z_{it} = \mathbf{b}_i' \mathbf{f}_t + \sigma_z u_{it}, \quad 1 \leq i \leq N, 1 \leq t \leq T \quad (4.3)$$

$$f_{jt} = \gamma_j f_{jt-1} + v_{jt}, \quad 1 \leq j \leq r \quad (4.4)$$

We set  $\beta_0 = 0$  and the main parameter of interest  $\beta_1 = 2$ . We let  $r = 5$  and  $\phi = (0.8, 0.5, 0.3, 0, 0)'$ , which means that five factors drive the instrument set but only three relate to the endogenous regressor. A linear combination of factors is useful for the endogenous regressor  $x_t$ , meaning that one direction or linear combination is enough to describe  $x_t$ , therefore  $L = 1$ . Factor loadings  $\mathbf{b}_i$  are drawn from the uniform distribution  $U[1, 2]$ . We didn't choose zero loading to ensure the factors drive the instruments. We allow factors  $\mathbf{f}_t$  and errors  $(\varepsilon_t, e_t)$  to be serially correlated following AR(1) process. The errors  $(\varepsilon_t, e_t)$  are generated from the following processes:

$$\varepsilon_t = \alpha_1 \varepsilon_{t-1} + \eta_t$$

$$e_t = \alpha_2 e_{t-1} + \zeta_t$$

We control the endogeneity by a parameter  $\rho$ , which is the correlation between  $\eta_t$  and  $\zeta_t$  ending up relating the errors  $\varepsilon_t$  and  $e_t$ . The disturbances  $(\eta_t, \zeta_t)$  are drawn from a joint-normal distribution with mean zero and variance-covariance matrix  $\Sigma = \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}$ . The error terms  $u_{it}$  and  $v_{jt}$  are each generated from standard normal distributions. For simplicity, we set all AR(1) coefficients involved equal to  $\gamma = \alpha_1 = \alpha_2 = 0.5$ . The parameter  $\sigma_z$  controls the influence of factors on the instrument set. To ensure that the instruments are noisy, we set  $\sigma_z = 0.25$ . In addition to the presence of irrelevant factors in  $x_t$ , the only difference between

Bai & Ng (2010) and our design is how the coefficients are drawn. For example, they generate  $\gamma_j$  from random while we keep it fixed. We keep  $\gamma_j$  fixed to be parsimonious to show that the variation in the performance stems from the feature we are targeting. We also check that if we allow  $\gamma_j$  to be random as in Bai & Ng (2010), nothing changes in the qualitative results.

#### 4.1.2 Results

The true value of the parameter of interest  $\beta_1$  is 2. We run 500 replications and report the expectation of estimate ( $E(\hat{\beta}_1)$ ) and its root mean squared error (RMSE). We consider three values of  $N$  for given  $T$ :  $N < T$ ,  $N \sim T$ , and  $N > T$  to demonstrate the efficacy of our method in the high-dimensional setting. We consider three values of  $\rho$ . The  $\rho = 0$  is the case when data  $(y_t, x_t)$  is i.i.d.,  $\rho = 0.5$  is a small to moderate endogeneity case because  $\eta_t$  makes about half of the variation in  $\varepsilon_t$ . Similarly,  $\rho = 0.9$  is a moderate to strong endogeneity case. This in total makes nine  $(\rho, N)$  combinations for a given  $T$ . We report results for  $T = \{100, 200, 400\}$ . Table-1 presents the results.

One secular observation from the results is that our method is better in terms of both bias and RMSE whenever there is an endogeneity ( $\rho \neq 0$ ). Results can be summarized into the following points:

1. Our method is a reliable/dependable method with the least root mean squared error and less bias in the majority of the cases. The bias and RMSE approaches zero as the sample size  $T$  increases which confirms the asymptotic theory.
2. As expected, under no endogeneity ( $\rho = 0$ ) case, OLS is the best method in the sense of the lowest RMSE.
3. The 2SLS does better than OLS when endogeneity is introduced but only when the number of instruments is small. When the number of instruments increases, the variance of the 2SLS estimator increases, due to which its performance deteriorates. This



$\rho$	r	T	N	$E(\hat{\beta}_1)$				RMSE( $\hat{\beta}_1$ )			
				SIF	OLS	2SLS	FIV	SIF	OLS	2SLS	FIV
0	5	100	25	2.06	2.06	2.01	2.01	0.38	0.15	0.27	0.28
0	5	100	75	2.07	2.06	2.06	2.06	0.36	0.15	0.15	0.15
0	5	100	125	2.07	2.06	-	1.88	0.37	0.15	-	0.59
0	5	200	50	2.05	2.06	2.05	2.04	0.27	0.11	0.11	0.11
0	5	200	150	2.05	2.06	2.06	2.06	0.27	0.11	0.11	0.11
0	5	200	250	2.05	2.06	-	1.96	0.27	0.11	-	0.38
0	5	400	100	2.03	2.05	2.04	2.05	0.18	0.09	0.09	0.09
0	5	400	300	2.02	2.05	2.05	2.05	0.18	0.09	0.09	0.10
0	5	400	500	2.02	2.05	-	1.98	0.18	0.09	-	0.28
0.5	5	100	25	2.13	2.37	2.23	2.26	0.39	0.41	0.38	0.41
0.5	5	100	75	2.13	2.37	2.50	2.52	0.39	0.41	0.53	0.55
0.5	5	100	125	2.12	2.37	-	1.80	0.38	0.41	-	0.65
0.5	5	200	50	2.09	2.33	2.86	2.90	0.29	0.35	0.87	0.92
0.5	5	200	150	2.08	2.33	2.50	2.51	0.28	0.35	0.51	0.52
0.5	5	200	250	2.07	2.33	-	1.93	0.28	0.35	-	0.37
0.5	5	400	100	2.05	2.31	2.88	2.91	0.20	0.32	0.89	0.93
0.5	5	400	300	2.04	2.31	2.45	2.46	0.20	0.32	0.46	0.47
0.5	5	400	500	2.04	2.31	-	1.97	0.20	0.32	-	0.28
0.9	5	100	25	2.20	2.64	2.37	2.42	0.44	0.66	0.47	0.52
0.9	5	100	75	2.22	2.64	2.74	2.76	0.43	0.66	0.76	0.78
0.9	5	100	125	2.22	2.64	-	1.77	0.43	0.66	-	0.74
0.9	5	200	50	2.15	2.61	3.01	3.01	0.31	0.62	1.02	1.06
0.9	5	200	150	2.12	2.61	2.74	2.75	0.30	0.62	0.75	0.76
0.9	5	200	250	2.12	2.61	-	1.92	0.29	0.62	-	0.38
0.9	5	400	100	2.07	2.59	3.02	3.04	0.21	0.59	1.03	1.06
0.9	5	400	300	2.07	2.59	2.71	2.71	0.20	0.59	0.71	0.72
0.9	5	400	500	2.06	2.59	-	1.96	0.20	0.59	-	0.28

Table 1: Simulations Using Design-I with AR Errors ( $\gamma = \alpha_1 = \alpha_2 = 0.5$ , 500 Reps)  
[Notes: True value of  $\beta_1$  is 2. We cannot estimate 2SLS when  $N > T$ .]

observation is known in the literature (Bekker (1994), Berry *et al.* (1995)), we confirm it.

4. The non-parametric estimation step in our method introduces a bias in our estimator which is visible at a lower sample size but goes to zero as sample size increases.
5. Increase in  $N$  relative to  $T$  hardly affects our estimate, while it appears to do so to other competing methods. This underscores our method's ability to produce stable estimates by handling both low and high-dimensional cases well.
6. Increase in sample-size  $T$  reduces the bias of all the methods.

#### 4.1.3 Discussion

The true  $m(\cdot)$  in this simulation setting is linear which is the ideal environment for the FIV of Bai & Ng (2010) to work, the only difference here is that not all factors present in the instrument set are relevant to the endogenous regressor. Since our method is supervised, it selects only the relevant factors for  $x_t$  in the sufficient index step unlike Bai & Ng (2010) which picks all the factors. Because of this supervising ability, our method can get  $\hat{x}_t$  better than the competition methods, hence producing a more efficient estimator. The merit of our method becomes clearer when the sample size  $T$  and strength of endogeneity ( $\rho$ ) increases as the difference between the RMSE of our methods relative to the competitor becomes wider.

One can ask why our method is less biased than OLS even when there is no endogeneity, it is because the error terms  $\varepsilon_t$  and  $e_t$  are serially correlated. When we make these error terms serially uncorrelated, the OLS performs better than ours, see table-?? in the appendix-B.2. Under endogeneity, the 2SLS estimator should perform better than the OLS which is visible from the rows with  $N = 25$ , however when the number of instruments  $N$  increases the performance of 2SLS deteriorates because the variance of the first stage estimation increases as noted in Berry *et al.* (1995) and discussed in Belloni *et al.* (2012).

#### 4.1.4 More Results for Robustness

We replicate the exercise done above with various combinations of the presence of serial correlation in factors and errors. In particular, we report the results for serially correlated factors with serially uncorrelated errors i.e.  $\gamma = 0.5$  but  $\alpha_1 = \alpha_2 = 0$  in the Table-5 in the appendix-B.2. Also, we report results for  $\gamma = \alpha_1 = \alpha_2 = 0$  in Table- 7 and for  $(\gamma = 0, \alpha_1 = \alpha_2 = 0.5)$  in Table-8. Note that errors  $(\varepsilon_t, e_t)$  are serially uncorrelated but are still endogenous because  $\eta_t$ , and  $\zeta_t$  are still correlated. The summary of these additional results is qualitatively the same as the findings observed so far in Table-1. Therefore, we verify and robustify the conclusion that our method is able to supervise the process to filter out the relevant factors.

## 4.2 Design-II: Gains from Dimension Reduction

In addition to the supervision and ability to capture the non-linearities, one more strength of our method is to achieve the sufficient dimension reduction required to obtain the  $E(x_t | \mathbf{f}_t)$ . While Bai & Ng (2010) achieves a dimensional reduction in the sense that  $N$  number of instruments are summarized in  $r$  number of factors where  $r < N$  is a small number. However, in addition to this, our method further combines the  $r$  number of factors into a smaller  $L$  number of indices required for the estimation of the conditional mean of the endogenous regressor (equation-4.2). This is an additional dimension reduction step that also performs the supervision, but we only focus on gains from the dimension reductions in this section. The gains stem from the fact that we have  $L \leq r$  number of variables to estimate  $E(x_t | \mathbf{f}_t)$ , therefore lesser variance in the estimation procedure.

### 4.2.1 The Design

To demonstrate this strength, we make all factors relevant by setting the true number of factors  $r = 3$  and true  $\phi = \{0.8, 0.5, 0.3\}$  in equation-4.1 to 4.4, therefore supervision will not give us an advantage. Further, keeping  $m(\cdot)$  to be a linear function will ensure that capturing the non-linearities will not benefit us. Therefore, if our method does better than others, it should

be coming from the additional dimension reductions which reduces the variance in the process. We keep the rest of the setting the same as section-4.1. The results of this simulation design are tabulated in the table-2.

#### 4.2.2 Results and Discussions

We see that our method outperforms the nearest competitors on both the bias and RMSE grounds. Therefore, achieving further dimension reduction by combining factors into indices improves the estimation. The rest of the observations are similar to the one noted in the previous simulation design section-4.1, therefore to save some space, we omit them here.

#### 4.2.3 More Results for Robustness

Under  $r = 3$  and linear  $m(\cdot)$  case, we consider various combinations of  $\gamma$  and  $(\alpha_1, \alpha_2)$ . The results for autocorrelated factors with serially uncorrelated errors i.e.  $\gamma = 0.5$  and  $\alpha_1 = \alpha_2 = 0$  given in Table-9 in the appendix-B.2. Also, the result for  $\gamma = \alpha_1 = \alpha_2 = 0$  specification is in Table-11 and for  $\gamma = 0, \alpha_1 = \alpha_2 = 0.5$  in Table-12. The qualitative results are the same as discussed for the  $\gamma = \alpha_1 = \alpha_2 = 0.5$  specification in the main analysis of this design. This cements the findings that the dimension reduction property of our method leads to gains in terms of lower RMSE and lower bias of the final estimator.

### 4.3 Design-III: Gains of Handling Non-Linearities

One major assumption in Bai & Ng (2010) is that  $m(\cdot)$  of equation-4.2 is a linear function. This can be a misspecified model, therefore we estimate  $m(\cdot)$  as a non-parametric function that captures the non-linearities if present. In our design in this section, we keep all factors relevant for  $x_t$  but are related in a non-linear fashion. This will ensure that better performance of our method will not come from the supervision but from capturing the non-linearities.

$\rho$	r	T	N	$E(\hat{\beta}_1)$				RMSE( $\hat{\beta}_1$ )			
				SIF	OLS	2SLS	FIV	SIF	OLS	2SLS	FIV
0	3	100	25	2.06	2.06	2.00	2.01	0.37	0.15	0.28	0.29
0	3	100	75	2.07	2.06	2.06	2.06	0.37	0.15	0.15	0.15
0	3	100	125	2.07	2.06	-	1.93	0.37	0.15	-	0.42
0	3	200	50	2.04	2.06	2.05	2.04	0.27	0.11	0.11	0.11
0	3	200	150	2.04	2.06	2.06	2.06	0.27	0.11	0.11	0.11
0	3	200	250	2.04	2.06	-	1.97	0.27	0.11	-	0.29
0	3	400	100	2.02	2.05	2.04	2.05	0.18	0.09	0.09	0.09
0	3	400	300	2.02	2.05	2.05	2.05	0.18	0.09	0.09	0.10
0	3	400	500	2.02	2.05	-	1.99	0.18	0.09	-	0.20
0.5	3	100	25	2.10	2.37	2.22	2.28	0.38	0.41	0.37	0.43
0.5	3	100	75	2.10	2.37	2.51	2.52	0.38	0.41	0.54	0.56
0.5	3	100	125	2.10	2.37	-	1.89	0.37	0.41	-	0.41
0.5	3	200	50	2.06	2.33	2.88	2.90	0.27	0.35	0.89	0.95
0.5	3	200	150	2.06	2.33	2.51	2.51	0.27	0.35	0.52	0.53
0.5	3	200	250	2.06	2.33	-	1.96	0.27	0.35	-	0.28
0.5	3	400	100	2.03	2.31	2.89	2.92	0.19	0.32	0.90	0.94
0.5	3	400	300	2.03	2.31	2.45	2.46	0.19	0.32	0.46	0.47
0.5	3	400	500	2.03	2.31	-	1.98	0.19	0.32	-	0.21
0.9	3	100	25	2.15	2.64	2.36	2.42	0.41	0.66	0.46	0.54
0.9	3	100	75	2.16	2.64	2.75	2.77	0.41	0.66	0.77	0.79
0.9	3	100	125	2.17	2.64	-	1.88	0.41	0.66	-	0.42
0.9	3	200	50	2.09	2.61	3.02	3.01	0.28	0.62	1.03	1.08
0.9	3	200	150	2.09	2.61	2.75	2.75	0.28	0.62	0.76	0.76
0.9	3	200	250	2.09	2.61	-	1.95	0.28	0.62	-	0.28
0.9	3	400	100	2.05	2.59	3.03	3.02	0.20	0.59	1.04	1.07
0.9	3	400	300	2.05	2.59	2.71	2.71	0.21	0.59	0.71	0.72
0.9	3	400	500	2.05	2.59	-	1.97	0.20	0.59	-	0.21

Table 2: Simulations using Design-II & AR Errors ( $\gamma = 0.5$ , 500 Reps)  
[Notes: True value of  $\beta_1$  is 2. We cannot estimate 2SLS when  $N > T$ .]

### 4.3.1 Design

One simple but widely used example of non-linearity in economics is allowing interactions between variables. Fan *et al.* (2017) gives an example of the interaction between financial dependence and economic growth, we borrow their example. We keep the equation 4.1, 4.3, and 4.4 as it is. The equation-4.2 is now:

$$x_t = f_{1t}(f_{2t} + f_{3t} + 1) + e_t \quad (4.5)$$

Similar to the section-4.2, we set  $r = 3$  with all factors being relevant for the endogenous regressor. This will ensure that if our method does better, it will not come from the ability to supervise. The true sufficient directions are the vectors in the plane  $S_{f|x}$  generated by  $\phi_1 = (1, 0, 0)'$  and  $\phi_2 = (0, 1, 1)'/\sqrt{2}$ . In other words, we need  $L = 2$  to sufficiently capture the non-linearity considered in this case. Had we used the linear model of Bai & Ng (2010), only one sufficient direction would have been captured, thereby missing the interaction structure. The rest of the elements of the data-generating process are the same as the section-4.1.

### 4.3.2 Results

We report the estimate of the  $\beta_1$  and the root mean squared error (RMSE) of the estimate calculated using 500 replications in Table-3.

Most of the observations from Table-3 reiterates the findings of Table-1 discussed in the section-4.1. One different observation in Table-3 is that the gains of capturing non-linearity are not that visible for smaller sample sizes, however, as the sample size starts growing, our method's performance becomes better beating the competitors. This is because the non-parametric step is involved in our method whose convergence rate is slower than the parametric rate.

$\rho$	r	T	N	$E(\hat{\beta}_1)$				$\text{RMSE}(\hat{\beta}_1)$			
				SIF	OLS	2SLS	FIV	SIF	OLS	2SLS	FIV
0	3	100	25	2.24	2.05	2.01	2.02	0.46	0.16	0.30	0.32
0	3	100	75	2.23	2.05	2.05	2.05	0.43	0.16	0.18	0.18
0	3	100	125	2.23	2.05	-	1.67	0.42	0.16	-	2.02
0	3	200	50	2.20	2.05	2.06	2.05	0.31	0.12	0.13	0.14
0	3	200	150	2.19	2.05	2.05	2.05	0.30	0.12	0.13	0.13
0	3	200	250	2.19	2.05	-	1.87	0.30	0.12	-	0.91
0	3	400	100	2.14	2.04	2.05	2.05	0.20	0.09	0.11	0.11
0	3	400	300	2.14	2.04	2.04	2.04	0.20	0.09	0.10	0.10
0	3	400	500	2.14	2.04	-	1.93	0.20	0.09	-	0.53
0.5	3	100	25	2.28	2.28	2.23	2.26	0.46	0.33	0.41	0.44
0.5	3	100	75	2.28	2.28	2.41	2.42	0.46	0.33	0.46	0.48
0.5	3	100	125	2.28	2.28	-	1.57	0.46	0.33	-	1.83
0.5	3	200	50	2.22	2.25	2.93	2.96	0.33	0.28	0.95	1.00
0.5	3	200	150	2.22	2.25	2.41	2.41	0.32	0.28	0.43	0.43
0.5	3	200	250	2.21	2.25	-	1.87	0.32	0.28	-	1.05
0.5	3	400	100	2.15	2.23	2.94	2.97	0.22	0.24	0.95	0.99
0.5	3	400	300	2.15	2.23	2.36	2.37	0.22	0.24	0.38	0.38
0.5	3	400	500	2.15	2.23	-	1.92	0.21	0.24	-	0.55
0.9	3	100	25	2.35	2.47	2.36	2.41	0.54	0.51	0.49	0.53
0.9	3	100	75	2.35	2.47	2.60	2.62	0.50	0.51	0.64	0.66
0.9	3	100	125	2.36	2.47	-	1.59	0.51	0.51	-	1.98
0.9	3	200	50	2.25	2.45	3.10	3.10	0.34	0.46	1.12	1.16
0.9	3	200	150	2.25	2.45	2.60	2.61	0.34	0.46	0.62	0.62
0.9	3	200	250	2.25	2.45	-	1.84	0.34	0.46	-	0.92
0.9	3	400	100	2.17	2.43	3.11	3.12	0.23	0.44	1.12	1.16
0.9	3	400	300	2.17	2.43	2.57	2.57	0.23	0.44	0.57	0.58
0.9	3	400	500	2.17	2.43	-	1.90	0.22	0.44	-	0.56

Table 3: Simulations using Design-III & AR Errors ( $\gamma = 0.5$ , 500 Reps)  
[Notes: True value of  $\beta_1$  is 2. We cannot estimate 2SLS when  $N > T$ . ]

### 4.3.3 More Results for Robustness

Under  $r = 3$  and non-linear  $m(\cdot)$  case, we consider various combinations of  $\gamma$  and  $(\alpha_1, \alpha_2)$ . The results for auto-correlated factors with serially uncorrelated errors i.e.  $\gamma = 0.5$  and  $\alpha_1 = \alpha_2 = 0$  given in Table-13 in the appendix-B.2. Also, the result for  $\gamma = \alpha_1 = \alpha_2 = 0$  specification is in Table-15 and for  $\gamma = 0, \alpha_1 = \alpha_2 = 0.5$  in Table-16. The qualitative results are the same as discussed for the  $\gamma = \alpha_1 = \alpha_2 = 0.5$  specification in the main analysis of this design.

### 4.4 More Designs for Robustness

We have demonstrated so far that our method performs better because of three major strengths: being a supervised method, achieving considerable dimension reduction, and the ability to handle non-linearities. One may be curious to know how our method does when all these challenges are present together. We consider five factors ( $r = 5$ ) driving the instrument set but only three are relevant for  $x_t$ , and  $m(\cdot)$  is a non-linear function. The rest of the specifications are the same as discussed in the section-4.1. The results are reported in the Table-18 in the appendix-B.2. The results are qualitatively the same as of Table-3. We see a slightly higher RMSE for the competing method FIV which may be because it is selecting the irrelevant factors. We consider different combinations of uncorrelated errors and factors and report the results in the appendix-B.2. We find that the qualitative results remain the same. We then replicate the simulation exercise of linear and non-linear  $m(\cdot)$  for eight true factors ( $r = 8$ ) where only three are relevant for the endogenous regressor  $x_t$ . All tables are reported in the appendix-B.2.

After analyzing more than 30 tables of different simulation specifications, we conclude that our method is decisively able to filter out the factors and then combine them in a manner that is relevant to the endogenous regressor. This powerful procedure then culminates into an efficient estimate of the causal parameter of interest  $\beta_1$  of equation-4.1. We now move to apply our method to real world problems.



## 5 Empirical Applications

We consider two applications, one is in the automobile industry of empirical industrial organization (IO) on cross-section data and the other one is in finance on time series data. We discuss the results and insights.

### 5.1 Price Elasticity of Demand for Automobiles

Berry *et al.* (1995) points out that the price elasticity of automobile demand estimated by the 2SLS method is too small to make sense. Expanding on this, Chernozhukov *et al.* (2015) demonstrates that the inconsistency in 2SLS estimation can be addressed by incorporating higher-order polynomials and interaction terms of the instrumental variable (IV) with the control variables. In this exercise, we replicate empirical application on automobile demand by Chernozhukov *et al.* (2015).

We do the principal component analysis of the set of instruments used in Chernozhukov *et al.* (2015) and found that about 85% of the information in their instrument set can be explained by the first two principal components. Also, the first two PCs are strongly related to the endogenous regressor *price*. This indicates that there could be factor structure in the instrument sets which may be driving the endogenous regressor (price), therefore, our method can be used.

#### 5.1.1 Model

Let the market demand of an automobile product  $i$  in period or market  $t$  is given by  $y_{it} = \log(s_{it}) - \log(s_{0t})$ , where  $s_{it}$  is the market share of the product  $i$  and the subscript 0 is the outside option in the market or period  $t$ .  $p_{it}$  is the price that is endogenously related to the automobile demand.  $x_{it}$  is the set of control variables that are observed including product characteristics, and  $z_{it}$  is the set of instruments. Chernozhukov *et al.* (2015) uses the following

basic specification of [Berry et al. \(1995\)](#)'s model:

$$y_{it} = \log(s_{it}) - \log(s_{0t}) = \alpha_0 p_{it} + x'_{it} \beta_0 + \varepsilon_{it} \quad (5.1)$$

$$p_{it} = z'_{it} \delta_0 + x'_{it} \gamma_0 + u_{it} \quad (5.2)$$

We assume that the high-dimensional set of instrument  $z_{it}$  has a factor structure that influences all the instruments and the endogenous regressor  $p_{it}$ , i.e.

$$z_{it} = \lambda_i f_t + \nu_{it}$$

In our application the period is one year apart therefore we treat a car in period  $t$  differently from period  $t - 1$ , in other words, we treat  $it$  as one datapoint  $j$ <sup>4</sup>. We first obtain  $\tilde{y}_j$  and  $\tilde{p}_j$  as the residuals of the regression of control variables  $x_j$  (used in [Berry et al. \(1995\)](#)) on  $y_j$  and  $p_j$  respectively. Then estimate-5.2 as  $\tilde{p}_j = m(\theta'_i f_t) + e_j$  where  $f_t$  are the factors that drives instruments  $z_j$  and endogenous regressor  $\tilde{p}_j$ . Use the  $\hat{\tilde{p}}_j$  in the equation-5.1 to estimate the parameter of interest  $\alpha_0$ .

### 5.1.2 Data

We use the final data made available in the replication package of [Chernozhukov et al. \(2015\)](#). The original paper [Berry et al. \(1995\)](#) uses five controls and ten instruments. [Chernozhukov et al. \(2015\)](#) builds on it and uses 24 control variables and 48 instruments to fit the application into their sparse setting. Unlike them, we need not use many controls so we keep on using the same control variables of the original paper [Berry et al. \(1995\)](#), but, since our paper allows correlated instruments, therefore, we use all the instruments suggested by both papers.

For the validity of the instruments, one can refer to [Berry et al. \(1995\)](#) and [Chernozhukov et al. \(2015\)](#). The only difference is that we are using the common components of their instruments.

---

<sup>4</sup>Our theory is valid for both IID and serially correlated data, but for one-year apart datapoints, the serial correlation is not that prevalent, so for simplicity, we treat it a cross-section

If a set of instruments satisfies the exclusion restriction, their common component and their linear combination should also satisfy the same which is our instrument in this application. We check the relevancy condition also and find that the factors are significantly correlated with the endogenous regressor with a p-value less than 0.01.<sup>5</sup> The instrument variable literature suggests that a weak instrument is not likely to occur if F-statistics is more than 10, we found our F-stat to be more than 100. Further, the possibility of weak factors is less likely to occur because our method selects the combinations of factors that best relate with  $p_j$  through supervision. Overall, the instruments used by [Berry \*et al.\* \(1995\)](#) and [Chernozhukov \*et al.\* \(2015\)](#) can serve as noisy instruments in our framework. It is natural to see because if the  $z_{it}$  are the instruments, their common component will also be the instruments. Therefore, the technical conditions for the validity of the instruments are satisfied in our data.

### 5.1.3 Results

We estimate the own-price elasticity of automobile demand given by the parameter  $\alpha_0$  of equation-5.1. We draw random samples with replacement multiple times (bootstrap, 500 times) and estimate  $\alpha_0$ , we report the bootstrapped confidence intervals for ours and [Chernozhukov \*et al.\* \(2015\)](#)'s method. The basic inversely sloped demand curve suggests that this estimate should be negative (significantly less than zero).

We find that the bootstrapped 95% confidence intervals of  $\hat{\alpha}_0$  estimated by our method is  $[-0.161, -0.125]$  with mean  $-0.152$ . For [Chernozhukov \*et al.\* \(2015\)](#)'s method, the interval is  $[-0.207, -0.154]$  with mean  $-0.192$  which is very close to the  $-0.185$ , the mean estimate given in their paper. The coefficient estimated in the original paper [Berry \*et al.\* \(1995\)](#) was  $-0.142$ . Our stable results over many sample-splitting pose confidence in the estimates. Further, our results reproduce the results of the [Chernozhukov \*et al.\* \(2015\)](#) and our estimates are close to theirs. Therefore, we can argue that similar to [Chernozhukov \*et al.\* \(2015\)](#), our method also

---

<sup>5</sup>We need Assumption-1.1 i.e.  $E[m(\theta_1' \mathbf{f}_j, \dots, \theta_L' \mathbf{f}_j)x_j] \neq 0$  for relevancy. However, if factors are correlated with  $x_t$ , it means that for  $m(\cdot)$  linear,  $\theta_1$  being vector of ones and other  $\theta_j$  being vectors of zeros, our relevancy holds. It may also satisfy through more functional forms but we at least guarantee one.

captures the important features of the data. This exercise confirms as a proof of concept that our method works.

## 5.2 CAPM Beta of BlackRock's SmallCap ETF

Firms with small market capital are attractive choices for investment because of their high (potentially multi-bagger) growth potential but they also bring along a higher level of risk relative to the market. Quantifying this risk is a question of interest for many investors. The simplest measure used for risk assessment by an investor is *Beta* of the stock estimated by the Capital Asset Pricing Model (CAPM).

The S&P SmallCap 600<sup>®</sup> is an index that provides investors with a benchmark for small-sized companies in the U.S. equities market that meet investability and financial viability criteria. However, the index is not a tradeable security. We, therefore, use its tradeable counterpart, the iShares S&P Small-Cap 600 Value ETF<sup>6</sup>. It is a passive exchange-traded fund (ETF) designed by BlackRock to mimic the return on S&P Small-Cap 600. We aim to estimate the CAPM *Beta* of this tradeable security. This ETF is known by its symbol IJS in the market, we'll use IJS to refer to our target variable in this application.

The S&P 1500<sup>®</sup> combines three leading indices, the S&P 500<sup>®</sup>, the S&P MidCap 400<sup>®</sup>, and the S&P SmallCap 600<sup>®</sup>, which covers approximately 90% of U.S. market capitalization. The S&P SmallCap 600<sup>®</sup> is a float-adjusted market cap weighted index of the smallest 600 companies in the S&P 1500<sup>®</sup>. The market capital of all companies in this index makes around 4% of the US market capitalization, therefore, one can say that this index is small and is unlikely to cause factors that drive the US economy or market.

---

<sup>6</sup>This ETF, launched in July 2000, is traded by symbol **IJS** on major trading platforms.

### 5.2.1 Model

Let  $y_t$  be the return on our IJS ETF. Using the capital asset pricing model (CAPM) theory we can write,

$$y_t = \alpha + \beta R_t^* + \eta_t \quad (5.3)$$

where  $R_t^*$  is the return on the market portfolio which is not observed. We are interested in the parameter  $\beta$  which is popularly known as *CAPM beta*. The true market return  $R_t^*$  is not observed, instead, what we observe is a proxy for the market portfolio  $R_t$ , such as the return on the DJIA index. We can write the observed proxy  $R_t$  as:

$$R_t = R_t^* + e_t \quad (5.4)$$

Where  $e_t$  is the error term satisfying the assumptions outlined in the theory part. Substituting  $R_t$  in place of  $R_t^*$  in the equation for  $y_t$ , we get:

$$y_t = \alpha + \beta R_t + \varepsilon_t$$

where  $\varepsilon_t = e_t - \beta e_t$  is correlated with  $R_t$ , this makes CAPM beta estimated using OLS biased. To solve this problem, we need instruments.

We use the return on the Dow Jones Industrial Average (DJIA) as a proxy for market return and firms in the S&P 500 for the instruments. We do not consider the firms that have been a part of the DJIA at any point in time during our analysis period to avoid any trivial analysis. Therefore there is no overlap of firms in our target variable ( $y_t$ ), instrument variable set  $\mathbf{z}_t$ , and the endogenous regressor  $R_t$ . Suppose that the returns on an S&P500 firm (instrument variable) follow the approximate factor model ([Chamberlain & Rothschild \(1983\)](#)) structure where factors are strong in the sense that they follow assumption-1-3. This can be written as the following equation:

$$z_{it} = \Lambda_i \mathbf{f}_t + u_{it} \quad (5.5)$$

Here  $\mathbf{f}_t$  is a vector of possibly unobservable factors. If  $\mathbf{f}_t$  is equal to  $R^* - r^*$  where  $r^*$  is the risk-free return, the equation-5.5 is a CAPM equation. Similarly, if  $\mathbf{f}_t$  contains Fama-French factors, the equation-5.5 can be seen as the Fama-French factor model. To allow a more general setting, we treat the factors as unobservables.

In this setting, the  $R_t^*$  can be seen as a sufficient index of the unobserved factor variables ( $\mathbf{f}_t$ ) for  $R_t$ , these unobserved factors are driving both the instruments  $\mathbf{z}_i$   $i = 1, 2, \dots, N$ , and the endogenous regressor  $R_t$ . In the remaining part of this section, we discuss the technical conditions related to the validity of the instruments and discuss the findings.

### 5.2.2 Data

For daily returns on the S&P500 companies' data, we use the Center for Research in Security Prices (CRSP)'s data on security prices traded in the market. We use the Wall Street Journal's website for the historical daily price data of the Dow Jones Industrial Average (DJIA) and the target variable, IJS ETF. We then calculate the returns based on the closed price of the trading day. Table-4 contains the list of the companies that have been a part of the Dow Jones Industrial Average (DJIA) at any point between 2001 to 2023. We remove these companies from the set of instruments to make sure that the  $R_t$  measured by DJIA is not a trivial combination of companies in the instrument.

We supplement our main analysis with a similar exercise on monthly data where we replace the instruments with Federal Reserve Economic Data monthly data (FRED-MD) and the endogenous regressor by the S&P500 Index. We use FRED-MD data compiled by [McCracken & Ng \(2016\)](#) which contain data on both new instruments and new endogenous regressor. For our target variable IJS ETF ( $y_t$ ), we use the Wall Street Journal to download the monthly index level, from which we compute the monthly returns. Our data runs from 2001 to 2023. We use [Hamilton & Xi \(2024\)](#)'s method to transform the data into a stationary series.

### 5.2.3 Validity of Instruments

**Relevancy** The required relevancy condition is  $E[m(\boldsymbol{\theta}_1' \mathbf{f}_t, \dots, \boldsymbol{\theta}_L' \mathbf{f}_t) R_t] \neq 0$ . Since factors are not observable, therefore, we use their consistent estimates, this translates the relevancy condition to  $E[\widehat{m}(\widehat{\boldsymbol{\theta}}_1' \widehat{\mathbf{f}}_t, \dots, \widehat{\boldsymbol{\theta}}_L' \widehat{\mathbf{f}}_t) R_t] \neq 0$ . This condition is weaker than otherwise required  $E(z_{it} R_t) \neq 0$  for all  $i = 1, 2, \dots, N$  in 2SLS method and  $E(f_{jt} R_t) \neq 0$  for all  $j = 1, 2, \dots, r$  of Bai & Ng (2010). In other words, many of the (noisy) instruments  $z_i$  are allowed to be invalid i.e. not satisfying the  $E(z_{it} R_t) \neq 0$ . We test the estimated factors from the instrument set to well correlate with our endogenous regressor DJIA with a p-value less than 0.01. The F-statistics is more than the 10, typically used in the IV literature. Therefore the relevancy condition is satisfied.

**Exclusion Condition** The exclusion restriction requires that our instruments affect  $y_t$  solely through the observed proxy for market return,  $R_t$  (the return on the DJIA). Since the true instruments are SDR indices of factors, we need these SDR indices (linear combinations of factors) to influence  $y_t$  only through  $R_t$ , implying there should be no direct effect of the factors on  $y_t$ . We employ S&P500 firms as ‘noisy’ instruments, where the true instrument is the common factor affecting all of these firms. As shown in Gabaix (2011), the largest firms in the U.S. explain a significant portion of aggregate market fluctuations. In our context, we argue that the common factors among S&P500 firms influence general market sentiment, as reflected in the DJIA. Therefore, these common factors are unlikely to directly impact all the smaller firms which constitute the IJS ETF, without first affecting the DJIA. It is more plausible that the common factors influence small firms through general market sentiment, and thus through  $R_t$ , the DJIA.

### 5.2.4 Results

We found the value of  $Beta$  to be equal to 1.51, and we reject the null hypothesis that the CAPM Beta of IJS ETF is less than or equal to one with 99% confidence. It means that the

S&P SmallCap 600 Index measured by IJS ETF is statistically more volatile than the market. This result makes sense because small-cap companies are usually more risky than the market.

In a supplementary exercise to this, we use macroeconomic variables in FRED-MD data ( $z$ ) as instruments and define the monthly returns on the S&P500 Index to be our  $R_t$ . The common factors driving the FRED-MD data are our instruments that need to satisfy the exclusion restrictions. Note that our target variable the S&P SmallCap 600 Index constitutes only about 4% of the total US market capitalization, which is a small part, therefore, is unlikely to drive the fundamental factors in the US economy. Whenever news about macroeconomic variables becomes public, it affects the market return  $R_t$  which then affects the other variables such as our  $y_t$ . Markets react and settle quickly, the other channel of macroeconomic variables affecting  $y_t$  is too slow compared to the main channel of market returns. Therefore, the exclusion restriction holds in our application. We again reject the null that Beta is one for IJS ETF.

## 6 Conclusion

In this paper, we introduce a novel method for causal inference using instrumental variables when the number of instruments is large. Our approach offers three key advantages: the ability to incorporate supervision, the flexibility to manage non-linearity, and the capability for sufficient dimension reduction. These attributes contribute to a more efficient estimation of the causal parameter of interest. Through extensive simulation exercises, we demonstrate the effectiveness of our method, showing that it consistently achieves lower bias and root mean squared error compared to alternative approaches across a variety of specifications. Additionally, we apply the method to two real-world case studies, yielding meaningful insights.



## References

- Ahn, Seung C, & Horenstein, Alex R. 2013. Eigenvalue ratio test for the number of factors. *Econometrica*, **81**(3), 1203–1227.
- Bai, Jushan. 2003. Inferential theory for factor models of large dimensions. *Econometrica*, **71**(1), 135–171.
- Bai, Jushan, & Ng, Serena. 2002. Determining the number of factors in approximate factor models. *Econometrica*, **70**(1), 191–221.
- Bai, Jushan, & Ng, Serena. 2008. Forecasting economic time series using targeted predictors. *Journal of Econometrics*, **146**(2), 304–317.
- Bai, Jushan, & Ng, Serena. 2010. Instrumental variable estimation in a data-rich environment. *Econometric Theory*, **26**(6), 1577–1606.
- Bai, Jushan, & Ng, Serena. 2013. Principal components estimation and identification of static factors. *Journal of econometrics*, **176**(1), 18–29.
- Bekker, Paul A. 1994. Alternative approximations to the distributions of instrumental variable estimators. *Econometrica: Journal of the Econometric Society*, 657–681.
- Belloni, Alexandre, Chen, Daniel, Chernozhukov, Victor, & Hansen, Christian. 2012. Sparse models and methods for optimal instruments with an application to eminent domain. *Econometrica*, **80**(6), 2369–2429.
- Bengio, Yoshua, *et al.* 2009. Learning deep architectures for AI. *Foundations and trends® in Machine Learning*, **2**(1), 1–127.
- Berry, Steven, Levinsohn, James, & Pakes, Ariel. 1995. Automobile Prices in Market Equilibrium. *Econometrica*, **63**(4), 841–890.

- Bura, Efstathia, & Cook, R Dennis. 2001. Estimating the structural dimension of regressions via parametric inverse regression. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, **63**(2), 393–410.
- Campbell, John Y, & Shiller, Robert J. 1988. Stock prices, earnings, and expected dividends. *the Journal of Finance*, **43**(3), 661–676.
- Carlstein, Edward. 1986. The use of subseries values for estimating the variance of a general statistic from a stationary sequence. *The annals of statistics*, 1171–1179.
- Chamberlain, Gary, & Rothschild, Michael. 1983. Arbitrage, Factor Structure, and Mean-Variance Analysis on Large Asset Markets. *Econometrica: Journal of the Econometric Society*, 1281–1304.
- Chernozhukov, Victor, Hansen, Christian, & Spindler, Martin. 2015. Post-selection and post-regularization inference in linear models with many controls and instruments. *American Economic Review*, **105**(5), 486–490.
- Cook, R Dennis. 2009. *Regression graphics: Ideas for studying regressions through graphics*. John Wiley & Sons.
- Fama, Eugene F, & French, Kenneth R. 1993. Common risk factors in the returns on stocks and bonds. *Journal of financial economics*, **33**(1), 3–56.
- Fama, Eugene F, & French, Kenneth R. 2015. A five-factor asset pricing model. *Journal of financial economics*, **116**(1), 1–22.
- Fan, Jianqing. 2018. *Local polynomial modelling and its applications: monographs on statistics and applied probability 66*. Routledge.
- Fan, Jianqing, & Gijbels, Irene. 1992. Variable bandwidth and local linear regression smoothers. *The Annals of Statistics*, 2008–2036.

- Fan, Jianqing, Xue, Lingzhou, & Yao, Jiawei. 2017. Sufficient forecasting using factor models. *Journal of econometrics*, **201**(2), 292–306.
- Fan, Jianqing, Ke, Yuan, & Wang, Kaizheng. 2020. Factor-adjusted regularized model selection. *Journal of Econometrics*, **216**(1), 71–85.
- Gabaix, Xavier. 2011. The granular origins of aggregate fluctuations. *Econometrica*, **79**(3), 733–772.
- Hamilton, James D, & Xi, Jin. 2024. *Principal Component Analysis for Nonstationary Series*. Tech. rept. National Bureau of Economic Research.
- Jat, Rajveer, & Padha, Daanish. 2024. Kernel Three-pass Regression Filter. *The 2024 California Econometrics Conference*.
- Kapetanios, George, & Marcellino, Massimiliano. 2010. Factor-GMM estimation with large sets of possibly weak instruments. *Computational Statistics & Data Analysis*, **54**(11), 2655–2675.
- Kelly, Bryan, & Pruitt, Seth. 2015. The three-pass regression filter: A new approach to forecasting using many predictors. *Journal of Econometrics*, **186**(2), 294–316.
- Lahiri, Soumendra Nath. 2013. *Resampling methods for dependent data*. Springer Science & Business Media.
- Li, Bing. 2018. *Sufficient dimension reduction: Methods and applications with R*. Chapman and Hall/CRC.
- Li, Ker-Chau. 1991. Sliced inverse regression for dimension reduction. *Journal of the American Statistical Association*, **86**(414), 316–327.
- Lucas Jr, Robert E. 1976. Econometric policy evaluation: A critique. *Pages 19–46 of: Carnegie-Rochester conference series on public policy*, vol. 1. North-Holland.

- Masry, Elias. 1996. Multivariate regression estimation local polynomial fitting for time series. *Stochastic Processes and their Applications*, **65**(1), 81–101.
- McCracken, Michael W, & Ng, Serena. 2016. FRED-MD: A monthly database for macroeconomic research. *Journal of Business & Economic Statistics*, **34**(4), 574–589.
- Newey, Whitney K. 1990. Efficient instrumental variables estimation of nonlinear models. *Econometrica: Journal of the Econometric Society*, 809–837.
- Newey, William K. 1985. Maximum Likelihood Specification Testing and Conditional Moment Tests. *Econometrica*, **53**(5), 1047–1070.
- Olmstead, Sheila M, Hanemann, W Michael, & Stavins, Robert N. 2007. Water demand under alternative price structures. *Journal of Environmental economics and management*, **54**(2), 181–198.
- Pagan, Adrian, & Ullah, Aman. 1999. *Nonparametric econometrics*. Cambridge University Press, Cambridge.
- Phillips, Peter CB. 1989. Partially identified econometric models. *Econometric Theory*, **5**(2), 181–240.
- Polk, Christopher, Thompson, Samuel, & Vuolteenaho, Tuomo. 2006. Cross-sectional forecasts of the equity premium. *Journal of Financial Economics*, **81**(1), 101–141.
- Staiger, Douglas O, & Stock, James H. 1994. *Instrumental variables regression with weak instruments*.
- Stock, James H, & Watson, Mark W. 2002. Forecasting using principal components from a large number of predictors. *Journal of the American statistical association*, **97**(460), 1167–1179.

## A Technical Appendix

### A.1 Proofs of Theoretical Results

#### A.1.1 Proof of Lemma 4

*Proof.* We aim to establish the bias, variance, and asymptotic normality of the local linear estimator  $\hat{m}(\mathbf{w})$ . Let's verify the Lindeberg-Feller CLT Conditions hold in our estimation procedure. The first condition required by Lindeberg-Feller CLT is *Normalization Condition*: it requires that the covariance matrix of the sum of random vectors converges to a positive definite matrix. For the local linear estimator, we have the random vectors  $\mathbf{P}_{tT} = e_t \mathcal{K}_h(\mathbf{w}_t - \mathbf{w})$  that need to be taken care of. The variance of  $\hat{m}(\mathbf{w})$  is given by:

$$\text{Var}(\hat{m}(\mathbf{w})) = \frac{\sigma^2}{Th^L g(\mathbf{w})} \int \mathcal{K}^2(\boldsymbol{\psi}) d\boldsymbol{\psi} + o\left(\frac{1}{Th^L}\right).$$

As long as this variance converges to a positive definite matrix as  $T \rightarrow \infty$ , the normalization condition is satisfied. The second condition is called *Lindeberg Condition*. It requires:

$$\frac{1}{\|\boldsymbol{\Sigma}\|} \sum_{t=1}^T \mathbb{E} [\|\mathbf{P}_{tT}\|^2 \mathbf{1}\{\|\mathbf{P}_{tT}\| > \epsilon \|\boldsymbol{\Sigma}\|\}] \xrightarrow{T \rightarrow \infty} 0,$$

for any  $\epsilon > 0$ . Which in our context, becomes:

$$\frac{1}{\sigma^2} \sum_{t=1}^T \mathbb{E} [e_t^2 \mathcal{K}_h^2(\mathbf{w}_t - \mathbf{w}) \mathbf{1}\{e_t^2 \mathcal{K}_h^2(\mathbf{w}_t - \mathbf{w}) > \epsilon \sigma^2\}] \xrightarrow{T \rightarrow \infty} 0.$$

Since  $e_t$  is assumed to have finite variance and the kernel  $\mathcal{K}$  is bounded and integrates to 1,  $e_t^2 \mathcal{K}_h^2(\mathbf{w}_t - \mathbf{w})$  should be well-behaved. For large  $T$ , the contribution of large deviations of  $e_t$  is controlled by the indicator function, satisfying the Lindeberg condition if the tails of  $e_t$  are not too heavy. Thus, the local linear estimator satisfies the conditions required for the Lindeberg-Feller Central Limit Theorem. Now we can proceed with the proof.

**Bias:** The local linear estimator  $\hat{m}(\mathbf{w})$  is defined as the solution to the weighted least

squares problem:

$$\hat{m}(\mathbf{w}) = \arg \min_{a, \mathbf{b}} \sum_{t=1}^T \left( x_t - a - \mathbf{b}^\top (\mathbf{w}_t - \mathbf{w}) \right)^2 \mathcal{K}_h(\mathbf{w}_t - \mathbf{w}),$$

where  $\mathcal{K}_h(\boldsymbol{\psi}) = h^{-L} \mathcal{K}\left(\frac{\boldsymbol{\psi}}{h}\right)$  is the scaled multivariate kernel.

Expanding  $m(\mathbf{w}_t)$  around  $\mathbf{w}$  using a Taylor series, we have:

$$m(\mathbf{w}_t) \approx m(\mathbf{w}) + (\mathbf{w}_t - \mathbf{w})^\top \nabla m(\mathbf{w}) + \frac{1}{2} (\mathbf{w}_t - \mathbf{w})^\top \nabla^2 m(\mathbf{w}) (\mathbf{w}_t - \mathbf{w}) + \dots$$

Since  $\hat{m}(\mathbf{w})$  is estimated using a weighted least squares criterion, where weights are determined by a kernel function  $\mathcal{K}_h(\mathbf{w}_t - \mathbf{w})$  with bandwidth  $h$ . Substituting this into the weighted least squares criterion, we can approximate  $\hat{m}(\mathbf{w})$  as:

$$\hat{m}(\mathbf{w}) \approx m(\mathbf{w}) + \frac{1}{2} \frac{\sum_{t=1}^T (\mathbf{w}_t - \mathbf{w})^\top \nabla^2 m(\mathbf{w}) (\mathbf{w}_t - \mathbf{w}) \mathcal{K}_h(\mathbf{w}_t - \mathbf{w})}{\sum_{t=1}^T \mathcal{K}_h(\mathbf{w}_t - \mathbf{w})}$$

Taking expectations and using the properties of the kernel  $K(\boldsymbol{\psi})$ , we obtain:

$$\mathbb{E}[\hat{m}(\mathbf{w})] - m(\mathbf{w}) = \frac{1}{2} h^2 \text{tr}(\mathbf{G}) + o(h^2),$$

where  $\mathbf{G} = \nabla^2 m(\mathbf{w}) \cdot \int \boldsymbol{\psi} \boldsymbol{\psi}^\top \mathcal{K}(\boldsymbol{\psi}) d\boldsymbol{\psi}$ .

### Variance:

The variance of the estimator is given by:

$$\text{Var}(\hat{m}(\mathbf{w})) = \text{Var} \left( \frac{\sum_{t=1}^T (m(\mathbf{w}_t) + e_t) \mathcal{K}_h(\mathbf{w}_t - \mathbf{w})}{\sum_{t=1}^T \mathcal{K}_h(\mathbf{w}_t - \mathbf{w})} \right).$$

Since  $m(\mathbf{w}_t)$  is smooth, the dominant term in the variance comes from the errors  $e_t$ , leading to:

$$\text{Var}(\hat{m}(\mathbf{w})) = \frac{\sigma^2}{Th^L g(\mathbf{w})} \int \mathcal{K}^2(\boldsymbol{\psi}) d\boldsymbol{\psi} + o\left(\frac{1}{Th^L}\right),$$

where  $g(\mathbf{w})$  is the joint density of  $\mathbf{w}_t$  at  $\mathbf{w}$ .

**Asymptotic Normality:** To establish asymptotic normality, we use the Lindeberg-Feller central limit theorem for mixing processes. Our target estimator can be expressed as:

$$\sqrt{Th^L} \left( \hat{m}(\mathbf{w}) - m(\mathbf{w}) - \frac{1}{2} h^2 \text{tr}(\mathbf{G}) \right) = \frac{1}{\sqrt{Th^L}} \sum_{t=1}^T e_t \mathcal{K}_h(\mathbf{w}_t - \mathbf{w}) + o(1).$$

By the Lindeberg-Feller central limit theorem, the sum converges in distribution to a normal random variable:

$$\sqrt{Th^L} \left( \hat{m}(\mathbf{w}) - m(\mathbf{w}) - \frac{1}{2} h^2 \text{tr}(\mathbf{G}) \right) \xrightarrow{d} \mathcal{N} \left( 0, \frac{\sigma^2}{g(\mathbf{w})} \int \mathcal{K}^2(\psi) d\psi \right).$$

This completes the proof.  $\square$

### A.1.2 Proof of Theorem-1

*Proof.* Let  $Q_t = [1 \quad m(\mathbf{w}_t)]$  and  $\beta = [\beta_0 \quad \beta_1]'$ . Let's represent  $\hat{X}_t = [1 \quad \hat{x}_t] = [1 \quad \hat{m}(\hat{\mathbf{w}}_t)]$ . We get  $Q$  when we stack  $Q_t$  and similarly  $\hat{X}$  by stacking  $\hat{X}_t$ . The expression of our target estimator  $\beta$  is:

$$\hat{\beta} = (\hat{X}' \hat{X})^{-1} \hat{X}' y$$

Where  $\hat{X}$  is the predicted value from instruments which is a function of factors. We used hats on both  $m$  and  $\mathbf{w}$  because we need to estimate them, given that neither the functional form  $m(\cdot)$  nor the factors (or their SDR indices) are observed. We'll prove an intermediary result stated in Claim-1 and then will return to the  $\hat{\beta}$ .

*Claim-1:*  $\hat{m}(\hat{\mathbf{w}}_t) - m(\mathbf{w}_t) = O_p(\delta_{NT})$  where  $\delta_{NT} = N^{-1/2} + T^{-2/L+4}$ .

*Proof.* Let's write down the Taylor series expansion of  $\hat{m}(\hat{\mathbf{w}}_t) = \hat{m}(\hat{w}_{t1}, \dots, \hat{w}_{tL})$  around a point  $\mathbf{w}_t = (w_{t1}, \dots, w_{tL})$ :

$$\begin{aligned}\widehat{m}(\widehat{w}_{t1}, \dots, \widehat{w}_{tL}) &= \widehat{m}(w_{t1}, \dots, w_{tL}) + \sum_{j=1}^L (\widehat{w}_{tj} - w_{tj}) \frac{\partial \widehat{m}(w_{t1}, \dots, w_{tL})}{\partial w_j} \\ &\quad + \frac{1}{2!} \sum_{j=1}^L \sum_{k=1}^L (\widehat{w}_{tj} - w_{tj})(\widehat{w}_{tk} - w_{tk}) \frac{\partial^2 \widehat{m}(w_{t1}, \dots, w_{tL})}{\partial w_j \partial w_k} + \dots\end{aligned}$$

Using Lemma-4, we can write the first term in the expression above i.e.  $\widehat{m}(w_{t1}, \dots, w_{tL})$  as

$$\widehat{m}(w_{t1}, \dots, w_{tL}) = m(w_{t1}, \dots, w_{tL}) + \frac{1}{2} h^2 \text{tr}(\mathbf{G}) + O_p\left(\frac{1}{Th^L}\right) + o_p(h^2)$$

Where  $\mathbf{G} = \nabla^2 m(\mathbf{w}) \cdot \int \boldsymbol{\psi} \boldsymbol{\psi}^\top \mathcal{K}(\boldsymbol{\psi}) d\boldsymbol{\psi}$ , the matrix  $\mathbf{G}$  is finite by assumption-5.5. We can obtain the convergence rate of the partial derivative  $\frac{\partial \widehat{m}(w_{t1}, \dots, w_{tL})}{\partial w_j}$  using the chain rule.  $\widehat{m}(w_{t1}, \dots, w_{tL})$  can be replaced by  $m(w_{t1}, \dots, w_{tL})$  and small order terms. Therefore the derivative of the estimated  $\frac{\partial \widehat{m}(w_{t1}, \dots, w_{tL})}{\partial w_j}$  will be of the same order as derivative of  $m(w_{t1}, \dots, w_{tL})$  which is  $O(1)$  by Assumption-5.1. Similarly, the second-order derivative terms  $\frac{\partial^2 \widehat{m}(w_{t1}, \dots, w_{tL})}{\partial w_j \partial w_k}$  will be  $O(1)$ . Since we have assumed that the function  $m(\cdot)$  is twice continuously differentiable, we can say that  $\frac{\partial m(w_{t1}, \dots, w_{tL})}{\partial w_j}$  and  $\frac{\partial^2 m(w_{t1}, \dots, w_{tL})}{\partial w_j \partial w_k}$  are  $O(1)$ . Therefore, we can write  $\widehat{m}(w_{t1}, \dots, w_{tL})$  as:

$$\begin{aligned}\widehat{m}(\widehat{w}_{t1}, \dots, \widehat{w}_{tL}) &= m(w_{t1}, \dots, w_{tL}) + O_p(h^2) + O_p\left(\frac{1}{Th^L}\right) + \sum_{j=1}^L (\widehat{w}_{tj} - w_{tj}) O(1) \\ &\quad + \frac{1}{2!} \sum_{j=1}^L \sum_{k=1}^L (\widehat{w}_{tj} - w_{tj})(\widehat{w}_{tk} - w_{tk}) O(1) + \text{higher order terms}\end{aligned}$$

Using Corollary-1, we know that  $\widehat{w}_j = \widehat{\boldsymbol{\theta}}_j' \widehat{\mathbf{f}}_t \rightarrow \boldsymbol{\theta}_j' \mathbf{f}_t = w_j$  for all  $j = 1, \dots, L$  with a rate  $\omega_{NT} = N^{-1/2} + T^{-1/2}$ . Which means that  $(\widehat{w}_{tj} - w_{tj}) = O_p(\omega_{NT})$  for all  $j = 1, \dots, L$ .



Therefore,

$$\begin{aligned}
\hat{m}(\hat{w}_{t1}, \dots, \hat{w}_{tL}) &= m(w_{t1}, \dots, w_{tL}) + O_p(h^2) + O_p\left(\frac{1}{Th^L}\right) + \sum_{j=1}^L O_p(\omega_{NT})O(1) \\
&\quad + \frac{1}{2!} \sum_{j=1}^L \sum_{k=1}^L O_p(\omega_{NT})O_p(\omega_{NT})O_p(h^2) \\
&= m(w_{t1}, \dots, w_{tL}) + O_p(h^2) + O_p\left(\frac{1}{Th^L}\right) + LO_p(\omega_{NT})O(1) + L^2O_p(\omega_{NT})^2O(1)
\end{aligned}$$

Since  $L$  is fixed, we can ignore it for convergence rate purposes.  $O_p(\omega_{NT})^2$  is of smaller order relative to  $O_p(\omega_{NT})$ , therefore we drop it. Also  $O_p(\omega_{NT})O(1) = O_p(\omega_{NT})$ , therefore, we have:

$$\hat{m}(\hat{w}_{t1}, \dots, \hat{w}_{tL}) = m(w_{t1}, \dots, w_{tL}) + O_p(h^2) + O_p\left(\frac{1}{Th^L}\right) + O_p(\omega_{NT})$$

Lemma-4 says that the Bias of  $\hat{m}(\cdot)$  is of order  $O_p(h^2)$  which implies that bias-square is of order  $O_p(h^4)$ , on the other hand variance is  $O_p(\frac{1}{Th^L})$ . Therefore the optimal bandwidth minimizing mean squared error is  $h_{opt} \propto T^{-1/(L+4)}$ , which translates to  $O_p(h^2) = O_p(T^{-2/(L+4)})$ . Since  $L \geq 1$ , therefore the fastest this rate could be is  $O_p(T^{-2/5})$  when  $L = 1$ . Therefore, it will dominate  $T^{-1/2}$  part in  $O_p(\omega_{NT})$ . Further,  $O_p\left(\frac{1}{Th^L}\right) = O_p\left(\frac{1}{T(T^{\frac{1}{L+4}})^L}\right) = O_p\left(T^{\frac{-4}{L+4}}\right)$  which goes to zero faster than the  $O_p(T^{-2/(L+4)})$ , therefore, this term can be ignored relative to  $O_p(h^2)$ .

Therefore,  $O_p(h^2) + O_p(\omega_{NT}) = O_p(N^{-1/2} + T^{-2/L+4})$ , which we call  $O_p(\delta_{NT})$ . Hence,

$$\begin{aligned}
\hat{m}(\hat{w}_{t1}, \dots, \hat{w}_{tL}) - m(w_{t1}, \dots, w_{tL}) &= O_p(h^2) + O_p(\omega_{NT}) \\
&= O_p(N^{-1/2} + T^{-2/L+4}) \\
&= \delta_{NT}
\end{aligned}$$

□

Let's get back to the expression of  $\hat{\beta}$ , we have  $\hat{\beta} = (\hat{X}'\hat{X})^{-1}\hat{X}'y$  where  $\hat{X} = [1 \quad \hat{m}(\hat{\mathbf{w}})]$ . We

want to expand  $\widehat{X}'\widehat{X}$  and  $\widehat{X}'y$  in terms of  $Q$ . Matrices with true values are:

$$Q'Q = \begin{bmatrix} T & \sum_{t=1}^T m(\mathbf{w}_t) \\ \sum_{t=1}^T m(\mathbf{w}_t) & \sum_{t=1}^T m(\mathbf{w}_t)^2 \end{bmatrix},$$

and

$$Q'y = \begin{bmatrix} \sum_{t=1}^T y_t \\ \sum_{t=1}^T m(\mathbf{w}_t)y_t \end{bmatrix}.$$

For  $\widehat{X}$ , we can write  $\widehat{X}'\widehat{X} = Q'Q + \Delta$ , where  $\Delta$  captures the deviations caused by  $\widehat{m}(\widehat{\mathbf{w}}) - m(\mathbf{w})$ . Specifically:

$$\Delta = \begin{bmatrix} 0 & \sum_{t=1}^T (\widehat{m}(\widehat{\mathbf{w}}_t) - m(\mathbf{w}_t)) \\ \sum_{t=1}^T (\widehat{m}(\widehat{\mathbf{w}}_t) - m(\mathbf{w}_t)) & \sum_{t=1}^T (\widehat{m}(\widehat{\mathbf{w}}_t)^2 - m(\mathbf{w}_t)^2) \end{bmatrix}.$$

$\Delta_{1,2} = \Delta_{2,1}$ , the two of the elements of the  $\Delta$  are  $\widehat{m}(\widehat{\mathbf{w}}_t) - m(\mathbf{w}_t) = O_p(h^2) + O_p(\omega_{NT})$ , let's see the  $\Delta_{2,2}$ :

$$\widehat{m}(\widehat{\mathbf{w}}_t)^2 - m(\mathbf{w}_t)^2 = (\widehat{m}(\widehat{\mathbf{w}}_t) - m(\mathbf{w}_t)) (\widehat{m}(\widehat{\mathbf{w}}_t) + m(\mathbf{w}_t)).$$

The first term in multiplication is:  $\widehat{m}(\widehat{\mathbf{w}}_t) - m(\mathbf{w}_t) = O_p(h^2) + O_p(\omega_{NT})$ . We can write the second term as:  $\widehat{m}(\widehat{\mathbf{w}}_t) + m(\mathbf{w}_t) = (\widehat{m}(\widehat{\mathbf{w}}_t) - m(\mathbf{w}_t)) + 2m(\mathbf{w}_t) = O_p(h^2) + O_p(\omega_{NT}) + O(1)$ .  $m(\mathbf{w}_t)$  is  $O(1)$  by assumption-5. Therefore,

$$\widehat{m}(\widehat{\mathbf{w}}_t)^2 - m(\mathbf{w}_t)^2 = (O_p(h^2) + O_p(\omega_{NT})) (\widehat{m}(\widehat{\mathbf{w}}_t) + m(\mathbf{w}_t)) = (O_p(h^2) + O_p(\omega_{NT})) (O(1)) = O_p(\delta_{NT})$$

Therefore,

$$\Delta = \begin{bmatrix} 0 & \sum_{t=1}^T O_p(\delta_{NT}) \\ \sum_{t=1}^T O_p(\delta_{NT}) & \sum_{t=1}^T O_p(\delta_{NT}) \end{bmatrix}.$$

Similarly, expand  $\widehat{X}'y$ :

$$\widehat{X}'y = Q'y + \Gamma,$$

where:

$$\Gamma = \begin{bmatrix} 0 \\ \sum_{t=1}^T (\widehat{m}(\widehat{\mathbf{w}}_t) - m(\mathbf{w}_t))y_t \end{bmatrix} = \begin{bmatrix} 0 \\ \sum_{t=1}^T O_p(\delta_{NT})y_t \end{bmatrix}$$

For two matrices  $A$  and  $B$ , we have the following identity (one can verify the same by post-multiplying  $(A + B)$ ):

$$(A + B)^{-1} = A^{-1} - A^{-1}B(A + B)^{-1}$$

Set  $Q'Q = A$  and  $\Delta = B$ , we can write  $(\widehat{X}'\widehat{X})^{-1} = (Q'Q + \Delta)^{-1}$  as:

$$(Q'Q + \Delta)^{-1} = (Q'Q)^{-1} - (Q'Q)^{-1} \Delta (Q'Q + \Delta)^{-1}$$

Note that  $y_t = \beta_0 + \beta_1 x_t + \varepsilon = [1 \quad x_t]\beta + \varepsilon_t = [1 \quad m(\mathbf{w}_t) + e_t]\beta + \varepsilon_t = [1 \quad m(\mathbf{w}_t)]\beta + e_t\beta_1 + \varepsilon_t$ .

Therefore, stacking  $t$  subscript terms, we can write  $y = Q\beta + e\beta_1 + \varepsilon$ :

$$\begin{aligned} \widehat{\beta} &= (\widehat{X}'\widehat{X})^{-1} \widehat{X}'y = (Q'Q + \Delta)^{-1} (Q'y + \Gamma) \\ &= \left( (Q'Q)^{-1} - (Q'Q)^{-1} \Delta (Q'Q + \Delta)^{-1} \right) (Q'(Q\beta + e\beta_1 + \varepsilon) + \Gamma) \\ &= \beta + (Q'Q)^{-1} Q'e\beta_1 + (Q'Q)^{-1} Q'\varepsilon + (Q'Q)^{-1} \Gamma - (Q'Q)^{-1} \Delta (Q'Q + \Delta)^{-1} Q'Q\beta \\ &\quad - (Q'Q)^{-1} \Delta (Q'Q + \Delta)^{-1} Q'e\beta_1 - (Q'Q)^{-1} \Delta (Q'Q + \Delta)^{-1} Q'\varepsilon - (Q'Q)^{-1} \Delta (Q'Q + \Delta)^{-1} \Gamma \end{aligned}$$

Therefore,

$$\begin{aligned} \widehat{\beta} - \beta &= (Q'Q)^{-1} Q'e\beta_1 + (Q'Q)^{-1} Q'\varepsilon + (Q'Q)^{-1} \Gamma - (Q'Q)^{-1} \Delta (Q'Q + \Delta)^{-1} Q'Q\beta \\ &\quad - (Q'Q)^{-1} \Delta (Q'Q + \Delta)^{-1} Q'e\beta_1 - (Q'Q)^{-1} \Delta (Q'Q + \Delta)^{-1} Q'\varepsilon - (Q'Q)^{-1} \Delta (Q'Q + \Delta)^{-1} \Gamma \end{aligned} \tag{A.1}$$

There are seven terms on the right-hand side in equation-A.1. To organize the discussion, we state and prove several claims each one related to one term in the equation-A.1.

*Claim-2:*  $(Q'Q)^{-1} Q'e\beta_1 = O_p(T^{-1/2})$

*Proof.* The matrix  $Q'Q$  is :

$$\begin{aligned} Q'Q &= \begin{bmatrix} T & \sum_{t=1}^T m(\mathbf{w}_t) \\ \sum_{t=1}^T m(\mathbf{w}_t) & \sum_{t=1}^T m(\mathbf{w}_t)^2 \end{bmatrix} \\ &= \begin{bmatrix} O(T) & O(T) \\ O(T) & O(T) \end{bmatrix} \end{aligned}$$

The term  $\sum_{t=1}^T m(\mathbf{w}_t)$  is  $O(T)$  because  $m(\mathbf{w}_t)$  is  $O(1)$  (Assumption-5.1 and Assumption-5.5).

Therefore, the  $Q'Q$  matrix is of order  $O(T)$ . Similarly,

$$Q'e\beta_1 = \begin{bmatrix} \sum_{t=1}^T e\beta_1 \\ \sum_{t=1}^T m(\mathbf{w}_t)e\beta_1 \end{bmatrix} = \begin{bmatrix} O_p(\sqrt{T}) \\ O_p(\sqrt{T}) \end{bmatrix}$$

Note that  $\beta_1$  is non-stochastic and is constant, therefore, can be replaced by  $O(1)$ . From Assumption-1.7,  $e_t$  are allowed to be weakly serially correlated so that  $\sum_{t=1}^T e_t$  is  $O_p(\sqrt{T})$ . By Assumption-1.3 and 1.7,  $\sum_{t=1}^T m(\mathbf{w}_t)e_t = O_p(\sqrt{T})$ . This implies

$$(Q'Q)^{-1} Q'e\beta_1 = O(T^{-1}) O_p(\sqrt{T}) O(1) = O_p(T^{-1/2})$$

□

*Claim-3:*  $(Q'Q)^{-1} Q'\varepsilon = O_p(T^{-1/2})$

*Proof.* In Claim-2, we have already shown that  $Q'Q = O(T)$ . Similarly,

$$Q'\varepsilon = \begin{bmatrix} \sum_{t=1}^T \varepsilon_t \\ \sum_{t=1}^T m(\mathbf{w}_t)\varepsilon_t \end{bmatrix} = \begin{bmatrix} O_p(\sqrt{T}) \\ O_p(\sqrt{T}) \end{bmatrix}$$

From Assumption-1.6,  $\varepsilon_t$  are allowed to be weakly serially correlated so that  $\sum_{t=1}^T \varepsilon_t$  is  $O_p(\sqrt{T})$ .

By by exclusion restriction stated as Assumption-1.2,  $\frac{1}{T} \sum_{t=1}^T m(\mathbf{w}_t)\varepsilon_t = O_p(\sqrt{T})$ . Therefore,

$$(Q'Q)^{-1}Q'\varepsilon = O(T^{-1})O_p(\sqrt{T}) = O_p(T^{-1/2})$$

□

*Claim-4:*  $(Q'Q)^{-1}\Gamma = O_p(\delta_{NT})$

*Proof.* This term involves  $\Gamma$  in multiplication with the inverse of  $Q'Q$ . Let's have a look at the behavior of matrix  $\Gamma$ .

$$\begin{aligned} \Gamma &= \begin{bmatrix} 0 \\ \sum_{t=1}^T (\hat{m}(\hat{\mathbf{w}}_t) - m(\mathbf{w}_t))y_t \end{bmatrix} = \begin{bmatrix} 0 \\ \sum_{t=1}^T (\hat{m}(\hat{\mathbf{w}}_t) - m(\mathbf{w}_t))(\beta_0 + x_t\beta_1 + \varepsilon_t) \end{bmatrix} \\ &= \begin{bmatrix} 0 \\ \sum_{t=1}^T O_p(\delta_{NT})\beta_0 + \sum_{t=1}^T O_p(\delta_{NT})m(\mathbf{w}_t)\beta_1 + \sum_{t=1}^T O_p(\delta_{NT})e_t\beta_1 + \sum_{t=1}^T O_p(\delta_{NT})\varepsilon_t \end{bmatrix} \\ &= \begin{bmatrix} 0 \\ \sum_{t=1}^T O_p(\delta_{NT}) \end{bmatrix} = \begin{bmatrix} 0 \\ O_p(T\delta_{NT}) \end{bmatrix} \end{aligned}$$

We get the second line by substituting  $x_t = m(\mathbf{w}_t) + e_t$  (equation-2.2). Note that the latter two terms in  $\Gamma_{2,1}$  element are multiplied with  $O_p(T^{-1/2})$  random disturbances and therefore are going to be of smaller order than the first two. Therefore, the overall order of the  $\Gamma_{2,1}$  is

dominated by the first two terms. Hence,

$$(Q'Q)^{-1}\Gamma = O(T^{-1}) \sum_{t=1}^T O_p(T\delta_{NT}) = O(T^{-1})O_p(T\delta_{NT}) = O_p(\delta_{NT})$$

□

*Claim-5:*  $(Q'Q)^{-1} \Delta (Q'Q + \Delta)^{-1} Q'Q\beta = O_p(\delta_{NT})$

*Proof.* We have already shown,

$$\Delta = \begin{bmatrix} 0 & \sum_{t=1}^T O_p(\delta_{NT}) \\ \sum_{t=1}^T O_p(\delta_{NT}) & \sum_{t=1}^T O_p(\delta_{NT}) \end{bmatrix} \quad \text{and} \quad Q'Q = \begin{bmatrix} O(T) & O(T) \\ O(T) & O(T) \end{bmatrix}$$

The terms in  $\Delta$  are  $\sum_{t=1}^T O_p(\delta_{NT}) = O_p(T\delta_{NT}) = O_p(T(N^{-1/2} + T^{-2/L+4}))$   
 $= O_p(N^{-1/2}T + T^{L+2/L+4})$ . Since  $N > 0$ , therefore the  $N^{-1/2}T$ -order term will be smaller than the terms of order  $T$ . Hence the terms in  $\Delta$  will be of smaller order compared to  $T$ -order terms in  $Q'Q$ . Therefore, we can approximate  $(Q'Q + \Delta)^{-1}(Q'Q)$  as  $O_p(1)$ . Using continuous mapping theorem, this reduces  $(Q'Q)^{-1} \Delta (Q'Q + \Delta)^{-1} Q'Q\beta$  to  $(Q'Q)^{-1} \Delta O_p(1)\beta$ . Since  $\beta$  is a constant with respect to time, this term's order is decided by  $(Q'Q)^{-1}\Delta$ . Therefore,

$$(Q'Q)^{-1} \Delta (Q'Q + \Delta)^{-1} Q'Q\beta = O(T^{-1}) \sum_{t=1}^T O_p(\delta_{NT}) \times O_p(T^{-1})O(T^1) = O_p(\delta_{NT})$$

□

*Claim-6:*  $(Q'Q)^{-1} \Delta (Q'Q + \Delta)^{-1} Q'e\beta_1 = O_p(\delta_{NT})O_p(T^{-1/2})$

*Proof.* In the previous claims, we have shown  $Q'Q$  is of order  $O(T)$ ,  $\Delta$  is of order  $O_p(T\delta_{NT})$ ,

$(Q'Q + \Delta)^{-1}$  is of order  $O_p(T^{-1})$ , and  $Q'e$  is of order  $O_p(\sqrt{T})$ .

$$\begin{aligned} (Q'Q)^{-1} \Delta (Q'Q + \Delta)^{-1} Q'e\beta_1 &= O(T^{-1})O_p(T\delta_{NT})O_p(T^{-1})O_p(T^{1/2})O(1) \\ &= O_p(\delta_{NT})O_p(T^{-1/2}) \end{aligned}$$

□

*Claim-7:*  $(Q'Q)^{-1} \Delta (Q'Q + \Delta)^{-1} Q'\varepsilon = O_p(\delta_{NT})O_p(T^{-1/2})$

*Proof.* Following Claim-5, this is straightforward. We have already shown that  $Q'\varepsilon$  is of order  $O_p(\sqrt{T})$ . Therefore,

$$\begin{aligned} (Q'Q)^{-1} \Delta (Q'Q + \Delta)^{-1} Q'\varepsilon &= O(T^{-1}) \sum_{t=1}^T O_p(\delta_{NT}) \times O_p(T^{-1})O_p(T^{1/2}) \\ &= O_p(\delta_{NT})O_p(T^{-1/2}) \end{aligned}$$

□

*Claim-8:*  $(Q'Q)^{-1} \Delta (Q'Q + \Delta)^{-1} \Gamma = O_p(\delta_{NT}^2)$

*Proof.* In the previous claims, we have shown that  $\Delta$  and  $\Gamma$  are of order  $O_p(T\delta_{NT})$ . Therefore, we can write,

$$\begin{aligned} (Q'Q)^{-1} \Delta (Q'Q + \Delta)^{-1} \Gamma &= O(T^{-1})O_p(T\delta_{NT})O_p(T^{-1})O_p(T\delta_{NT}) \\ &= O_p(\delta_{NT}^2) \end{aligned}$$

□

After learning the behavior of all terms involved in the expression of  $\hat{\beta} - \beta$ , the equation-A.1.

Using Claim-2 to Claim-8, we can write,

$$\begin{aligned}\widehat{\beta} - \beta = & O_p(T^{-\frac{1}{2}}) + O_p(T^{-\frac{1}{2}}) + O_p(\delta_{NT}) + O_p(\delta_{NT}) \\ & + O_p(\delta_{NT})O(T^{-1/2}) + O_p(\delta_{NT})O(T^{-1/2}) + O_p(\delta_{NT}^2)\end{aligned}$$

Note that the third and fourth terms will always dominate other terms. Since  $L \geq 1$ , the fastest rate possible for the second term is  $O_p(\delta_{NT}) = O_p(N^{-\frac{1}{2}} + T^{-\frac{2}{5}})$  which goes to zero slower than the first and second terms which are  $O_p(T^{-\frac{1}{2}})$ . Therefore, asymptotically ( $N \rightarrow \infty$  and  $T \rightarrow \infty$ ), the first and second terms will be dominated and, therefore can be ignored. Similarly, the fifth, sixth, and seventh terms will also be dominated by the third and fourth terms and therefore can be ignored in asymptotics. Therefore, the  $\widehat{\beta}$  is consistent estimator for  $\beta$  with  $O_p(\delta_{NT}) = O_p(N^{-\frac{1}{2}} + T^{-\frac{2}{L+4}})$ , where  $L$  is the number of SDR indices. In other words, with fixed  $L$ ,  $N \rightarrow \infty$  and  $T \rightarrow \infty$ , we can write:

$$\widehat{\beta} - \beta = O_p(\delta_{NT})$$

□

### A.1.3 Proof of Theorem-2

*Proof.* We have discussed in the proof of theorem-1, that the the expression of  $(\widehat{\beta} - \beta)$  is made up of several terms out of which only two terms survive. Therefore, for the asymptotics, we



ignore the other terms that go to zero at a faster rate. Therefore we can write,

$$\begin{aligned}
\widehat{\beta} - \beta &= (Q'Q)^{-1} \Gamma + (Q'Q)^{-1} \Delta (Q'Q + \Delta)^{-1} Q'Q\beta + o_p(\delta_{NT}) \\
&= (Q'Q)^{-1} \left[ \Gamma + \Delta (Q'Q + \Delta)^{-1} Q'Q\beta \right] + o_p(\delta_{NT}) \\
(\widehat{\beta} - \beta) &= \left( \frac{1}{T} Q'Q \right)^{-1} \frac{1}{T} \left[ \Gamma + \Delta (Q'Q + \Delta)^{-1} Q'Q\beta \right] + o_p(\delta_{NT}) \\
\frac{1}{\delta_{NT}} (\widehat{\beta} - \beta) &= \left( \frac{1}{T} Q'Q \right)^{-1} \frac{1}{T\delta_{NT}} \left[ \Gamma + \Delta (Q'Q + \Delta)^{-1} Q'Q\beta \right] + \frac{1}{\delta_{NT}} o_p(\delta_{NT})
\end{aligned}$$

We can write  $\frac{1}{\delta_{NT}} = \frac{1}{N^{\frac{-1}{2}} + T^{\frac{-2}{L+4}}}$  as  $\min\{N^{\frac{1}{2}}, T^{\frac{2}{L+4}}\}$ . Let's call  $\phi_{NT} = \min\{N^{\frac{1}{2}}, T^{\frac{2}{L+4}}\}$ . We can write the variance of the expression mentioned above as:<sup>7</sup>

$$\begin{aligned}
\text{Var} \left( \phi_{NT} (\widehat{\beta} - \beta) \right) &= \left( \frac{1}{T} Q'Q \right)^{-1} \frac{1}{T^2 \delta_{NT}^2} \text{Var} \left( \Gamma + \Delta (Q'Q + \Delta)^{-1} Q'Q\beta \right) \left( \frac{1}{T} Q'Q \right)^{-1} \\
&= \left( \frac{1}{T} Q'Q \right)^{-1} \frac{1}{T^2 \delta_{NT}^2} \left( \Gamma \Gamma' + \Delta (Q'Q + \Delta)^{-1} Q'Q\beta\beta' Q'Q (Q'Q + \Delta)^{-1} \Delta' \right) \left( \frac{1}{T} Q'Q \right)^{-1}
\end{aligned}$$

As  $N, T \rightarrow \infty$ ,  $(Q'Q + \Delta)^{-1} \rightarrow (Q'Q)^{-1}$ , therefore we can write  $Q'Q(Q'Q + \Delta)^{-1} = (Q'Q + \Delta)^{-1} Q'Q = I$ . It allows us to write:

$$\text{Var} \left( \phi_{NT} (\widehat{\beta} - \beta) \right) = \left( \frac{1}{T} Q'Q \right)^{-1} \frac{1}{T^2 \delta_{NT}^2} \left( \Gamma \Gamma' + \Delta \beta \beta' \Delta' \right) \left( \frac{1}{T} Q'Q \right)^{-1}$$

$$\Gamma \Gamma' = \begin{bmatrix} 0 & 0 \\ 0 & \left( \sum_{t=1}^T (\widehat{m}(\widehat{\mathbf{w}}_t) - m(\mathbf{w}_t)) y_t \right)^2 \end{bmatrix}$$

We note now find the expression for  $\Delta \beta \beta' \Delta' = \Delta \beta (\Delta \beta)'$ .

---

<sup>7</sup>Note that  $\frac{1}{\delta_{NT}} o_p(\delta_{NT}) = o_p(1)$  does not survive relative to  $O_p(1)$ , therefore is ignored for the variance.

$$\Delta = \begin{bmatrix} 0 & \sum_{t=1}^T (\hat{m}(\hat{\mathbf{w}}_t) - m(\mathbf{w}_t)) \\ \sum_{t=1}^T (\hat{m}(\hat{\mathbf{w}}_t) - m(\mathbf{w}_t)) & \sum_{t=1}^T (\hat{m}(\hat{\mathbf{w}}_t)^2 - m(\mathbf{w}_t)^2) \end{bmatrix}, \quad \beta = \begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix},$$

$$\Delta\beta = \begin{bmatrix} \sum_{t=1}^T (\hat{m}(\hat{\mathbf{w}}_t) - m(\mathbf{w}_t))\beta_1 \\ \sum_{t=1}^T (\hat{m}(\hat{\mathbf{w}}_t) - m(\mathbf{w}_t))\beta_0 + \sum_{t=1}^T (\hat{m}(\hat{\mathbf{w}}_t)^2 - m(\mathbf{w}_t)^2)\beta_1 \end{bmatrix}$$

$$\Delta\beta\beta'\Delta' = \begin{bmatrix} \sum_{t=1}^T (\hat{m}(\hat{\mathbf{w}}_t) - m(\mathbf{w}_t))\beta_1 \\ \sum_{t=1}^T (\hat{m}(\hat{\mathbf{w}}_t) - m(\mathbf{w}_t))\beta_0 + \sum_{t=1}^T (\hat{m}(\hat{\mathbf{w}}_t)^2 - m(\mathbf{w}_t)^2)\beta_1 \end{bmatrix} \times$$

$$\begin{bmatrix} \beta_1 \sum_{t=1}^T (\hat{m}(\hat{\mathbf{w}}_t) - m(\mathbf{w}_t)) & \beta_0 \sum_{t=1}^T (\hat{m}(\hat{\mathbf{w}}_t) - m(\mathbf{w}_t)) + \beta_1 \sum_{t=1}^T (\hat{m}(\hat{\mathbf{w}}_t)^2 - m(\mathbf{w}_t)^2) \end{bmatrix}$$

The individual elements of the matrix have long expressions, making it difficult to fit the full matrix within the page. So we look at each one column at a time.

$$\Delta\beta\beta'\Delta'_1 = \begin{bmatrix} \beta_1^2 \left( \sum_{t=1}^T (\hat{m}(\hat{\mathbf{w}}_t) - m(\mathbf{w}_t)) \right)^2 \\ \beta_1 \left( \sum_{t=1}^T (\hat{m}(\hat{\mathbf{w}}_t) - m(\mathbf{w}_t)) \right) \left( \beta_0 \sum_{t=1}^T (\hat{m}(\hat{\mathbf{w}}_t) - m(\mathbf{w}_t)) + \beta_1 \sum_{t=1}^T (\hat{m}(\hat{\mathbf{w}}_t)^2 - m(\mathbf{w}_t)^2) \right) \end{bmatrix}$$

$$\Delta\beta\beta'\Delta'_2 = \begin{bmatrix} \beta_1 \left( \sum_{t=1}^T (\hat{m}(\hat{\mathbf{w}}_t) - m(\mathbf{w}_t)) \right) \left( \beta_0 \sum_{t=1}^T (\hat{m}(\hat{\mathbf{w}}_t) - m(\mathbf{w}_t)) + \beta_1 \sum_{t=1}^T (\hat{m}(\hat{\mathbf{w}}_t)^2 - m(\mathbf{w}_t)^2) \right) \\ \left( \beta_0 \sum_{t=1}^T (\hat{m}(\hat{\mathbf{w}}_t) - m(\mathbf{w}_t)) + \beta_1 \sum_{t=1}^T (\hat{m}(\hat{\mathbf{w}}_t)^2 - m(\mathbf{w}_t)^2) \right)^2 \end{bmatrix}$$

Let's denote the matrix  $\mathbf{V} = \Gamma\Gamma' + \Delta\beta\beta'\Delta'$ . Then the first column of this matrix is given by:

$$\mathbf{V}_1 = \begin{bmatrix} \beta_1^2 \left( \sum_{t=1}^T (\hat{m}(\hat{\mathbf{w}}_t) - m(\mathbf{w}_t)) \right)^2 \\ \beta_1 \left( \sum_{t=1}^T (\hat{m}(\hat{\mathbf{w}}_t) - m(\mathbf{w}_t)) \right) \left( \beta_0 \sum_{t=1}^T (\hat{m}(\hat{\mathbf{w}}_t) - m(\mathbf{w}_t)) + \beta_1 \sum_{t=1}^T (\hat{m}(\hat{\mathbf{w}}_t)^2 - m(\mathbf{w}_t)^2) \right) \end{bmatrix}$$

The second column of this matrix  $\mathbf{V}$  is given by:

$$\mathbf{V}_2 = \begin{bmatrix} \beta_1 \left( \sum_{t=1}^T (\hat{m}(\hat{\mathbf{w}}_t) - m(\mathbf{w}_t)) \right) \left( \beta_0 \sum_{t=1}^T (\hat{m}(\hat{\mathbf{w}}_t) - m(\mathbf{w}_t)) + \beta_1 \sum_{t=1}^T (\hat{m}(\hat{\mathbf{w}}_t)^2 - m(\mathbf{w}_t)^2) \right) \\ \left( \beta_0 \sum_{t=1}^T (\hat{m}(\hat{\mathbf{w}}_t) - m(\mathbf{w}_t)) + \beta_1 \sum_{t=1}^T (\hat{m}(\hat{\mathbf{w}}_t)^2 - m(\mathbf{w}_t)^2) \right)^2 + \left( \sum_{t=1}^T (\hat{m}(\hat{\mathbf{w}}_t) - m(\mathbf{w}_t)) y_t \right)^2 \end{bmatrix}$$

We notice that the elements of matrix  $\mathbf{V}$  contain  $\beta_0$  and  $\beta_1$  and  $m(\mathbf{w}_t)$  which are not observable. However, we do not need  $m(\mathbf{w}_t)$ , we only need the  $\hat{m}(\hat{\mathbf{w}}) - m(\mathbf{w}_t)$  which is the bias of non-parametric estimation step and is known to us. Further, we have the consistent estimates of  $\beta_0$  and  $\beta_1$  as proved in theorem-1, therefore, we can obtain the consistent estimate of the matrix  $\mathbf{V}$ . We can finally write:

$$\text{Var} \left( \phi_{NT}(\hat{\beta} - \beta) \right) = \left( \frac{1}{T} Q'Q \right)^{-1} \frac{\mathbf{V}}{T^2 \delta_{NT}^2} \left( \frac{1}{T} Q'Q \right)^{-1}$$

This allows us to write

$$\phi_{NT}(\hat{\beta} - \beta) \xrightarrow{d} \mathcal{N} \left( 0, \left( \frac{1}{T} Q'Q \right)^{-1} \frac{\mathbf{V}}{T^2 \delta_{NT}^2} \left( \frac{1}{T} Q'Q \right)^{-1} \right),$$

Where  $\phi_{NT} = \min\{N^{\frac{1}{2}}, T^{\frac{2}{L+4}}\}$ , if  $N$  is higher then  $\phi_{NT}$  is dominated by  $T$  and is slower than the parametric  $\sqrt{T}$  rate. This completes the proof. □

## B Supplementary Appendix: Tables, Algorithms, Data

### B.1 Data

We use the Center for Research in Security Prices (CRSP)'s data on security prices traded in the market. We use the Wall Street Journal's website for the historical daily price data of the Dow Jones Industrial Average (DJIA) and IJS ETF. We then calculate the returns based on the closed price of the trading day. Table-4 contains the list of the companies that have been a part of the Dow Jones Industrial Average (DJIA) at any point between 2001 to 2023. FRED-MD is a monthly database for macroeconomic research available on this [website](#). The detailed description of this data is available in [McCracken & Ng \(2016\)](#).

Name	Ticker	Year A/D	Name	Ticker	Yead A/D
AT&T	T	2004 (D)	Verizon	VZ	2004 (A)
International Paper	IP	2004 (D)	Kodak	KODK	2004 (D)
American Intern'l Group	AIG	2004 (A)	Pfizer	PFE	2004 (A)
Bank of America	BAC	2008 (A)	Altria	MO	2008 (D)
Honeywell	HON	2008 (D)	Chevron	CVX	2008 (A)
American Intern'l Group	AIG	2008 (D)	Kraft	KHC	2008 (A)
Travelers Companies	TRV	2009(A)	Citigroup	C	2009 (D)
General Motors	GM	2009 (D)	Cisco Systems	CSCO	2009 (A)
UnitedHealth Group	UNH	2012 (A)	Kraft	KHC	2012 (D)
Alcoa	AA	2013 (D)	Nike	NKE	2013 (A)
Bank of America	BAC	2013 (D)	Goldman Sachs	GS	2013 (A)
Hewlett-Packard	HPQ	2013 (D)	Visa	V	2013 (A)
AT&T	T	2015 (D)	Apple	AAPL	2015 (A)
General Electric	GE	2018 (D)	Walgreens	WBA	2018 (A)
DowDuPont Inc.	DWDP	2019 (D)	Dow Inc	DOW	2019 (A)
ExxonMobil	XOM	2020 (D)	Salesforce	CRM	2020 (A)
Raytheon	RTX	2020 (D)	Honeywell	HON	2020 (A)
Pfizer	PFE	2020 (D)	Amgen	AMGN	2020 (A)
Microsoft Corporation	MSFT	NC	Walmart Inc.	WMT	NC
JPMorgan Chase & Co.	JPM	NC	Boeing	BA	NC
The Procter & Gamble Company	PG	NC	Intel	INTC	NC
Johnson & Johnson	JNJ	NC	Walt Disney	DIS	NC
Coca-Cola	KO	NC	Merck Inc.	MRK	NC
McDonald's Corporation	MCD	NC	Caterpillar	CAT	NC
The Home Depot	HD	NC	IBM	IBM	NC
American Express	AXP	NC	3M Company	MMM	NC

Table 4: Firms Added (A), Deleted (D), and No Change (NC) in Dow Jones Industrial Average during 2001-2023

## B.2 Simulations

This section provides the results of various simulation designs and their sub-cases. In addition to three designs considered in the main text, we also provide a section “more designs” to check the robustness of the results.

### B.2.1 Design-I

This subsection contains the simulation results of the design discussed in section-4.1. We provide a small table that tabulates the simulation tables in this sub-section.

Table-No	Serially Correlated Factors?	Serially Correlated Errors?
Table-5	Yes	No
Table-6	Yes	Yes
Table-7	No	No
Table-8	No	Yes

### B.2.2 Design-II

This subsection contains the simulation results of the design discussed in section-4.2. We provide a small table that tabulates the simulation tables in this sub-section.

Table-No	Serially Correlated Factors?	Serially Correlated Errors?
Table-9	Yes	No
Table-10	Yes	Yes
Table-11	No	No
Table-12	No	Yes

$\rho$	r	T	N	$E(\hat{\beta}_1)$				RMSE( $\hat{\beta}_1$ )			
				SIF	OLS	2SLS	FIV	SIF	OLS	2SLS	FIV
0	5	100	25	2.04	2.01	1.96	1.96	0.26	0.10	0.19	0.20
0	5	100	75	2.04	2.01	2.01	2.01	0.25	0.10	0.11	0.10
0	5	100	125	2.04	2.01	-	1.90	0.24	0.10	-	0.35
0	5	200	50	2.02	2.01	2.00	2.00	0.17	0.07	0.07	0.08
0	5	200	150	2.03	2.01	2.00	2.00	0.17	0.07	0.07	0.07
0	5	200	250	2.03	2.01	-	1.96	0.17	0.07	-	0.24
0.5	5	100	25	2.09	2.33	2.18	2.20	0.30	0.36	0.27	0.30
0.5	5	100	75	2.10	2.33	2.44	2.46	0.28	0.36	0.46	0.48
0.5	5	100	125	2.10	2.33	-	1.86	0.27	0.36	-	0.41
0.5	5	200	50	2.06	2.30	2.72	2.75	0.20	0.31	0.72	0.77
0.5	5	200	150	2.06	2.30	2.43	2.43	0.19	0.31	0.44	0.44
0.5	5	200	250	2.05	2.30	-	1.94	0.19	0.31	-	0.25
0.9	5	100	25	2.15	2.56	2.28	2.32	0.33	0.57	0.34	0.39
0.9	5	100	75	2.15	2.56	2.64	2.66	0.31	0.57	0.66	0.67
0.9	5	100	125	2.13	2.56	-	1.85	0.30	0.57	-	0.46
0.9	5	200	50	2.09	2.53	2.84	2.87	0.21	0.53	0.85	0.89
0.9	5	200	150	2.08	2.53	2.63	2.63	0.20	0.53	0.63	0.64
0.9	5	200	250	2.08	2.53	-	1.94	0.20	0.53	-	0.25

Table 5: Simulations using Design-I & Non-AR Errors ( $\gamma = 0.5$ , 500 Reps)  
[Notes: True value of  $\beta_1$  is 2. We cannot estimate 2SLS when  $N > T$ .]

$\rho$	r	T	N	$E(\hat{\beta}_1)$				$\text{RMSE}(\hat{\beta}_1)$			
				SIF	OLS	2SLS	FIV	SIF	OLS	2SLS	FIV
0	5	100	25	2.06	2.06	2.01	2.01	0.38	0.15	0.27	0.28
0	5	100	75	2.07	2.06	2.06	2.06	0.36	0.15	0.15	0.15
0	5	100	125	2.07	2.06	-	1.88	0.37	0.15	-	0.59
0	5	200	50	2.05	2.06	2.05	2.04	0.27	0.11	0.11	0.11
0	5	200	150	2.05	2.06	2.06	2.06	0.27	0.11	0.11	0.11
0	5	200	250	2.05	2.06	-	1.96	0.27	0.11	-	0.38
0	5	400	100	2.03	2.05	2.04	2.05	0.18	0.09	0.09	0.09
0	5	400	300	2.02	2.05	2.05	2.05	0.18	0.09	0.09	0.10
0	5	400	500	2.02	2.05	-	1.98	0.18	0.09	-	0.28
0.5	5	100	25	2.13	2.37	2.23	2.26	0.39	0.41	0.38	0.41
0.5	5	100	75	2.13	2.37	2.50	2.52	0.39	0.41	0.53	0.55
0.5	5	100	125	2.12	2.37	-	1.80	0.38	0.41	-	0.65
0.5	5	200	50	2.09	2.33	2.86	2.90	0.29	0.35	0.87	0.92
0.5	5	200	150	2.08	2.33	2.50	2.51	0.28	0.35	0.51	0.52
0.5	5	200	250	2.07	2.33	-	1.93	0.28	0.35	-	0.37
0.5	5	400	100	2.05	2.31	2.88	2.91	0.20	0.32	0.89	0.93
0.5	5	400	300	2.04	2.31	2.45	2.46	0.20	0.32	0.46	0.47
0.5	5	400	500	2.04	2.31	-	1.97	0.20	0.32	-	0.28
0.9	5	100	25	2.20	2.64	2.37	2.42	0.44	0.66	0.47	0.52
0.9	5	100	75	2.22	2.64	2.74	2.76	0.43	0.66	0.76	0.78
0.9	5	100	125	2.22	2.64	-	1.77	0.43	0.66	-	0.74
0.9	5	200	50	2.15	2.61	3.01	3.01	0.31	0.62	1.02	1.06
0.9	5	200	150	2.12	2.61	2.74	2.75	0.30	0.62	0.75	0.76
0.9	5	200	250	2.12	2.61	-	1.92	0.29	0.62	-	0.38
0.9	5	400	100	2.07	2.59	3.02	3.04	0.21	0.59	1.03	1.06
0.9	5	400	300	2.07	2.59	2.71	2.71	0.20	0.59	0.71	0.72
0.9	5	400	500	2.06	2.59	-	1.96	0.20	0.59	-	0.28

Table 6: Simulations using Design-I & AR Errors ( $\gamma = 0.5$ , 500 Reps)

[Notes: True value of  $\beta_1$  is 2. We cannot estimate 2SLS when  $N > T$ .]



$\rho$	r	T	N	$E(\hat{\beta}_1)$				RMSE( $\hat{\beta}_1$ )			
				SIF	OLS	2SLS	FIV	SIF	OLS	2SLS	FIV
0	5	100	25	2.05	2.01	1.95	1.96	0.27	0.10	0.20	0.21
0	5	100	75	2.04	2.01	2.01	2.01	0.26	0.10	0.10	0.10
0	5	100	125	2.05	2.01	-	1.95	0.25	0.10	-	0.39
0	5	200	50	2.02	2.01	2.00	2.00	0.17	0.07	0.07	0.08
0	5	200	150	2.03	2.01	2.00	2.00	0.17	0.07	0.07	0.07
0	5	200	250	2.03	2.01	-	1.96	0.17	0.07	-	0.24
0.5	5	100	25	2.12	2.35	2.21	2.24	0.32	0.37	0.30	0.32
0.5	5	100	75	2.12	2.35	2.47	2.48	0.41	0.37	0.49	0.50
0.5	5	100	125	2.13	2.35	-	1.91	0.30	0.37	-	0.44
0.5	5	200	50	2.06	2.30	2.72	2.74	0.20	0.31	0.72	0.77
0.5	5	200	150	2.06	2.30	2.43	2.43	0.19	0.31	0.44	0.44
0.5	5	200	250	2.05	2.30	-	1.94	0.19	0.31	-	0.25
0.9	5	100	25	2.19	2.59	2.33	2.36	0.37	0.60	0.39	0.42
0.9	5	100	75	2.21	2.59	2.68	2.69	0.35	0.60	0.69	0.70
0.9	5	100	125	2.20	2.59	-	1.89	0.36	0.60	-	0.49
0.9	5	200	50	2.09	2.53	2.84	2.88	0.21	0.53	0.85	0.89
0.9	5	200	150	2.08	2.53	2.63	2.63	0.20	0.53	0.63	0.64
0.9	5	200	250	2.08	2.53	-	1.94	0.20	0.53	-	0.25

Table 7: imulations using Design-I & Non-AR Errors ( $\gamma = 0$ , 500 Reps)  
[Notes: True value of  $\beta_1$  is 2. We cannot estimate 2SLS when  $N > T$ .]

$\rho$	r	T	N	$E(\hat{\beta}_1)$				$RMSE(\hat{\beta}_1)$			
				SIF	OLS	2SLS	FIV	SIF	OLS	2SLS	FIV
0	5	100	25	2.09	2.06	2.01	2.02	0.35	0.13	0.25	0.26
0	5	100	75	2.08	2.06	2.06	2.06	0.34	0.13	0.13	0.13
0	5	100	125	2.07	2.06	-	1.94	0.36	0.13	-	0.48
0	5	200	50	2.06	2.07	2.05	2.05	0.23	0.10	0.09	0.09
0	5	200	150	2.05	2.07	2.06	2.06	0.23	0.10	0.10	0.10
0	5	200	250	2.05	2.07	-	1.99	0.23	0.10	-	0.33
0	5	400	100	2.03	2.06	2.05	2.05	0.17	0.09	0.08	0.08
0	5	400	300	2.03	2.06	2.06	2.06	0.17	0.09	0.09	0.09
0	5	400	500	2.03	2.06	-	2.00	0.17	0.09	-	0.24
0.5	5	100	25	2.14	2.39	2.27	2.31	0.39	0.42	0.39	0.43
0.5	5	100	75	2.17	2.39	2.54	2.56	0.38	0.42	0.57	0.59
0.5	5	100	125	2.14	2.39	-	1.88	0.35	0.42	-	0.50
0.5	5	200	50	2.10	2.35	2.94	2.94	0.26	0.36	0.95	0.97
0.5	5	200	150	2.07	2.35	2.54	2.55	0.24	0.36	0.55	0.56
0.5	5	200	250	2.07	2.35	-	1.96	0.23	0.36	-	0.32
0.5	5	400	100	2.05	2.32	2.95	2.95	0.18	0.33	0.96	0.98
0.5	5	400	300	2.04	2.32	2.48	2.49	0.18	0.33	0.49	0.50
0.5	5	400	500	2.04	2.32	-	1.99	0.18	0.33	-	0.25
0.9	5	100	25	2.21	2.67	2.42	2.47	0.42	0.69	0.50	0.55
0.9	5	100	75	2.23	2.67	2.79	2.80	0.43	0.69	0.81	0.82
0.9	5	100	125	2.22	2.67	-	1.86	0.41	0.69	-	0.53
0.9	5	200	50	2.16	2.64	3.09	3.09	0.30	0.64	1.09	1.11
0.9	5	200	150	2.14	2.64	2.79	2.80	0.28	0.64	0.80	0.80
0.9	5	200	250	2.13	2.64	-	1.95	0.26	0.64	-	0.32
0.9	5	400	100	2.06	2.62	3.09	3.09	0.19	0.62	1.10	1.11
0.9	5	400	300	2.06	2.62	2.74	2.75	0.19	0.62	0.75	0.75
0.9	5	400	500	2.06	2.62	-	1.98	0.19	0.62	-	0.25

Table 8: Simulations using Design-I & AR Errors ( $\gamma = 0$ , 500 Reps)

[Notes: True value of  $\beta_1$  is 2. We cannot estimate 2SLS when  $N > T$ .]

$\rho$	r	T	N	$E(\hat{\beta}_1)$				RMSE( $\hat{\beta}_1$ )			
				SIF	OLS	2SLS	FIV	SIF	OLS	2SLS	FIV
0	3	100	25	2.04	2.01	1.95	1.96	0.24	0.10	0.19	0.20
0	3	100	75	2.04	2.01	2.01	2.01	0.24	0.10	0.10	0.10
0	3	100	125	2.04	2.01	-	1.94	0.23	0.10	-	0.26
0	3	200	50	2.02	2.01	2.00	2.00	0.16	0.07	0.07	0.07
0	3	200	150	2.02	2.01	2.00	2.00	0.16	0.07	0.07	0.07
0	3	200	250	2.02	2.01	-	1.97	0.16	0.07	-	0.18
0.5	3	100	25	2.08	2.33	2.17	2.22	0.27	0.36	0.26	0.32
0.5	3	100	75	2.07	2.33	2.45	2.46	0.26	0.36	0.47	0.48
0.5	3	100	125	2.08	2.33	-	1.92	0.26	0.36	-	0.27
0.5	3	200	50	2.04	2.30	2.72	2.76	0.18	0.31	0.73	0.79
0.5	3	200	150	2.04	2.30	2.43	2.43	0.18	0.31	0.44	0.44
0.5	3	200	250	2.04	2.30	-	1.96	0.18	0.31	-	0.19
0.9	3	100	25	2.11	2.56	2.28	2.34	0.28	0.57	0.33	0.41
0.9	3	100	75	2.11	2.56	2.65	2.66	0.29	0.57	0.66	0.67
0.9	3	100	125	2.10	2.56	-	1.92	0.29	0.57	-	0.27
0.9	3	200	50	2.05	2.53	2.85	2.84	0.19	0.53	0.85	0.91
0.9	3	200	150	2.06	2.53	2.63	2.64	0.19	0.53	0.64	0.64
0.9	3	200	250	2.06	2.53	-	1.96	0.20	0.53	-	0.19

Table 9: Simulations using Design-II & Non-AR Errors ( $\gamma = 0.5$ , 500 Reps)  
[Notes: True value of  $\beta_1$  is 2. We cannot estimate 2SLS when  $N > T$ .]

$\rho$	r	T	N	$E(\hat{\beta}_1)$				RMSE( $\hat{\beta}_1$ )			
				SIF	OLS	2SLS	FIV	SIF	OLS	2SLS	FIV
0	3	100	25	2.06	2.06	2.00	2.01	0.37	0.15	0.28	0.29
0	3	100	75	2.07	2.06	2.06	2.06	0.37	0.15	0.15	0.15
0	3	100	125	2.07	2.06	-	1.93	0.37	0.15	-	0.42
0	3	200	50	2.04	2.06	2.05	2.04	0.27	0.11	0.11	0.11
0	3	200	150	2.04	2.06	2.06	2.06	0.27	0.11	0.11	0.11
0	3	200	250	2.04	2.06	-	1.97	0.27	0.11	-	0.29
0	3	400	100	2.02	2.05	2.04	2.05	0.18	0.09	0.09	0.09
0	3	400	300	2.02	2.05	2.05	2.05	0.18	0.09	0.09	0.10
0	3	400	500	2.02	2.05	-	1.99	0.18	0.09	-	0.20
0.5	3	100	25	2.10	2.37	2.22	2.28	0.38	0.41	0.37	0.43
0.5	3	100	75	2.10	2.37	2.51	2.52	0.38	0.41	0.54	0.56
0.5	3	100	125	2.10	2.37	-	1.89	0.37	0.41	-	0.41
0.5	3	200	50	2.06	2.33	2.88	2.90	0.27	0.35	0.89	0.95
0.5	3	200	150	2.06	2.33	2.51	2.51	0.27	0.35	0.52	0.53
0.5	3	200	250	2.06	2.33	-	1.96	0.27	0.35	-	0.28
0.5	3	400	100	2.03	2.31	2.89	2.92	0.19	0.32	0.90	0.94
0.5	3	400	300	2.03	2.31	2.45	2.46	0.19	0.32	0.46	0.47
0.5	3	400	500	2.03	2.31	-	1.98	0.19	0.32	-	0.21
0.9	3	100	25	2.15	2.64	2.36	2.42	0.41	0.66	0.46	0.54
0.9	3	100	75	2.16	2.64	2.75	2.77	0.41	0.66	0.77	0.79
0.9	3	100	125	2.17	2.64	-	1.88	0.41	0.66	-	0.42
0.9	3	200	50	2.09	2.61	3.02	3.01	0.28	0.62	1.03	1.08
0.9	3	200	150	2.09	2.61	2.75	2.75	0.28	0.62	0.76	0.76
0.9	3	200	250	2.09	2.61	-	1.95	0.28	0.62	-	0.28
0.9	3	400	100	2.05	2.59	3.03	3.02	0.20	0.59	1.04	1.07
0.9	3	400	300	2.05	2.59	2.71	2.71	0.21	0.59	0.71	0.72
0.9	3	400	500	2.05	2.59	-	1.97	0.20	0.59	-	0.21

Table 10: Simulations using Design-II & AR Errors ( $\gamma = 0.5$ , 500 Reps)  
[Notes: True value of  $\beta_1$  is 2. We cannot estimate 2SLS when  $N > T$ .]

$\rho$	r	T	N	$E(\hat{\beta}_1)$				RMSE( $\hat{\beta}_1$ )			
				SIF	OLS	2SLS	FIV	SIF	OLS	2SLS	FIV
0	3	100	25	2.05	2.01	1.95	1.95	0.26	0.10	0.20	0.20
0	3	100	75	2.04	2.01	2.01	2.01	0.26	0.10	0.10	0.10
0	3	100	125	2.05	2.01	-	1.96	0.26	0.10	-	0.27
0	3	200	50	2.03	2.01	2.00	2.00	0.18	0.07	0.07	0.07
0	3	200	150	2.03	2.01	2.00	2.00	0.18	0.07	0.07	0.07
0	3	200	250	2.03	2.01	-	1.98	0.18	0.07	-	0.19
0.5	3	100	25	2.09	2.35	2.20	2.23	0.29	0.37	0.29	0.32
0.5	3	100	75	2.09	2.35	2.47	2.48	0.29	0.37	0.49	0.50
0.5	3	100	125	2.09	2.35	-	1.95	0.28	0.37	-	0.29
0.5	3	200	50	2.07	2.32	2.78	2.79	0.19	0.33	0.78	0.81
0.5	3	200	150	2.05	2.32	2.46	2.46	0.19	0.33	0.46	0.47
0.5	3	200	250	2.05	2.32	-	1.98	0.19	0.33	-	0.19
0.9	3	100	25	2.12	2.59	2.32	2.36	0.32	0.60	0.37	0.42
0.9	3	100	75	2.13	2.59	2.68	2.69	0.31	0.60	0.69	0.70
0.9	3	100	125	2.13	2.59	-	1.94	0.32	0.60	-	0.30
0.9	3	200	50	2.08	2.56	2.91	2.88	0.20	0.56	0.91	0.94
0.9	3	200	150	2.07	2.56	2.67	2.67	0.21	0.56	0.67	0.67
0.9	3	200	250	2.07	2.56	-	1.98	0.21	0.56	-	0.20

Table 11: Simulations using Design-II & Non-AR Errors ( $\gamma = 0$ , 500 Reps)  
[Notes: True value of  $\beta_1$  is 2. We cannot estimate 2SLS when  $N > T$ .]

$\rho$	r	T	N	$E(\hat{\beta}_1)$				RMSE( $\hat{\beta}_1$ )			
				SIF	OLS	2SLS	FIV	SIF	OLS	2SLS	FIV
0	3	100	25	2.06	2.06	2.00	2.01	0.33	0.13	0.25	0.27
0	3	100	75	2.06	2.06	2.06	2.06	0.33	0.13	0.13	0.13
0	3	100	125	2.06	2.06	-	1.96	0.34	0.13	-	0.35
0	3	200	50	2.05	2.07	2.05	2.05	0.23	0.10	0.09	0.09
0	3	200	150	2.04	2.07	2.07	2.06	0.23	0.10	0.10	0.09
0	3	200	250	2.04	2.07	-	1.99	0.24	0.10	-	0.25
0.5	3	100	25	2.10	2.39	2.26	2.30	0.35	0.42	0.38	0.43
0.5	3	100	75	2.11	2.39	2.55	2.56	0.35	0.42	0.57	0.59
0.5	3	100	125	2.10	2.39	-	1.93	0.35	0.42	-	0.36
0.5	3	200	50	2.09	2.35	2.95	2.92	0.25	0.36	0.96	0.97
0.5	3	200	150	2.06	2.35	2.55	2.55	0.23	0.36	0.55	0.56
0.5	3	200	250	2.07	2.35	-	1.98	0.24	0.36	-	0.24
0.9	3	100	25	2.16	2.67	2.41	2.44	0.39	0.69	0.49	0.54
0.9	3	100	75	2.17	2.67	2.79	2.81	0.40	0.69	0.81	0.83
0.9	3	100	125	2.16	2.67	-	1.92	0.38	0.69	-	0.37
0.9	3	200	50	2.12	2.64	3.09	3.03	0.27	0.64	1.10	1.10
0.9	3	200	150	2.09	2.64	2.80	2.80	0.26	0.64	0.80	0.80
0.9	3	200	250	2.09	2.64	-	1.97	0.26	0.64	-	0.24

Table 12: Simulations using Design-II & AR Errors ( $\gamma = 0$ , 500 Reps)  
[Notes: True value of  $\beta_1$  is 2. We cannot estimate 2SLS when  $N > T$ .]

### B.2.3 Design-III

This subsection contains the simulation results of the design discussed in section-4.3. We provide a small table that tabulates the simulation tables in this sub-section.

Table-No	Serially Correlated Factors?	Serially Correlated Errors?
Table-13	Yes	No
Table-14	Yes	Yes
Table-15	No	No
Table-16	No	Yes

$\rho$	r	T	N	$E(\hat{\beta}_1)$				RMSE( $\hat{\beta}_1$ )			
				SIF	OLS	2SLS	FIV	SIF	OLS	2SLS	FIV
0	3	100	25	2.21	2.01	1.96	1.96	0.33	0.12	0.22	0.23
0	3	100	75	2.20	2.01	2.01	2.01	0.31	0.12	0.13	0.13
0	3	100	125	2.20	2.01	-	1.79	0.31	0.12	-	0.98
0	3	200	50	2.16	2.00	2.00	2.00	0.23	0.09	0.10	0.10
0	3	200	150	2.15	2.00	2.00	2.00	0.23	0.09	0.09	0.09
0	3	200	250	2.15	2.00	-	1.86	0.23	0.09	-	0.60
0.5	3	100	25	2.25	2.23	2.17	2.20	0.36	0.27	0.29	0.32
0.5	3	100	75	2.24	2.23	2.35	2.36	0.34	0.27	0.38	0.39
0.5	3	100	125	2.24	2.23	-	1.69	0.35	0.27	-	1.30
0.5	3	200	50	2.19	2.21	2.77	2.80	0.26	0.23	0.79	0.84
0.5	3	200	150	2.18	2.21	2.33	2.34	0.25	0.23	0.35	0.35
0.5	3	200	250	2.18	2.21	-	1.86	0.25	0.23	-	0.62
0.9	3	100	25	2.29	2.39	2.27	2.30	0.40	0.42	0.36	0.40
0.9	3	100	75	2.29	2.39	2.51	2.52	0.40	0.42	0.53	0.54
0.9	3	100	125	2.29	2.39	-	1.67	0.39	0.42	-	1.35
0.9	3	200	50	2.22	2.37	2.92	2.96	0.28	0.38	0.93	0.99
0.9	3	200	150	2.21	2.37	2.49	2.50	0.26	0.38	0.51	0.51
0.9	3	200	250	2.20	2.37	-	1.85	0.26	0.38	-	0.63

Table 13: Simulations using Design-III & Non-AR Errors ( $\gamma = 0.5$ , 500 Reps)  
[Notes: True value of  $\beta_1$  is 2. We cannot estimate 2SLS when  $N > T$ .]



$\rho$	r	T	N	$E(\hat{\beta}_1)$				$\text{RMSE}(\hat{\beta}_1)$			
				SIF	OLS	2SLS	FIV	SIF	OLS	2SLS	FIV
0	3	100	25	2.24	2.05	2.01	2.02	0.46	0.16	0.30	0.32
0	3	100	75	2.23	2.05	2.05	2.05	0.43	0.16	0.18	0.18
0	3	100	125	2.23	2.05	-	1.67	0.42	0.16	-	2.02
0	3	200	50	2.20	2.05	2.06	2.05	0.31	0.12	0.13	0.14
0	3	200	150	2.19	2.05	2.05	2.05	0.30	0.12	0.13	0.13
0	3	200	250	2.19	2.05	-	1.87	0.30	0.12	-	0.91
0	3	400	100	2.14	2.04	2.05	2.05	0.20	0.09	0.11	0.11
0	3	400	300	2.14	2.04	2.04	2.04	0.20	0.09	0.10	0.10
0	3	400	500	2.14	2.04	-	1.93	0.20	0.09	-	0.53
0.5	3	100	25	2.28	2.28	2.23	2.26	0.46	0.33	0.41	0.44
0.5	3	100	75	2.28	2.28	2.41	2.42	0.46	0.33	0.46	0.48
0.5	3	100	125	2.28	2.28	-	1.57	0.46	0.33	-	1.83
0.5	3	200	50	2.22	2.25	2.93	2.96	0.33	0.28	0.95	1.00
0.5	3	200	150	2.22	2.25	2.41	2.41	0.32	0.28	0.43	0.43
0.5	3	200	250	2.21	2.25	-	1.87	0.32	0.28	-	1.05
0.5	3	400	100	2.15	2.23	2.94	2.97	0.22	0.24	0.95	0.99
0.5	3	400	300	2.15	2.23	2.36	2.37	0.22	0.24	0.38	0.38
0.5	3	400	500	2.15	2.23	-	1.92	0.21	0.24	-	0.55
0.9	3	100	25	2.35	2.47	2.36	2.41	0.54	0.51	0.49	0.53
0.9	3	100	75	2.35	2.47	2.60	2.62	0.50	0.51	0.64	0.66
0.9	3	100	125	2.36	2.47	-	1.59	0.51	0.51	-	1.98
0.9	3	200	50	2.25	2.45	3.10	3.10	0.34	0.46	1.12	1.16
0.9	3	200	150	2.25	2.45	2.60	2.61	0.34	0.46	0.62	0.62
0.9	3	200	250	2.25	2.45	-	1.84	0.34	0.46	-	0.92
0.9	3	400	100	2.17	2.43	3.11	3.12	0.23	0.44	1.12	1.16
0.9	3	400	300	2.17	2.43	2.57	2.57	0.23	0.44	0.57	0.58
0.9	3	400	500	2.17	2.43	-	1.90	0.22	0.44	-	0.56

Table 14: Simulations using Design-III & AR Errors ( $\gamma = 0.5$ , 500 Reps)  
[Notes: True value of  $\beta_1$  is 2. We cannot estimate 2SLS when  $N > T$ . ]

$\rho$	r	T	N	$E(\hat{\beta}_1)$				$\text{RMSE}(\hat{\beta}_1)$			
				SIF	OLS	2SLS	FIV	SIF	OLS	2SLS	FIV
0	3	100	25	2.21	2.01	1.95	1.96	0.35	0.11	0.22	0.23
0	3	100	75	2.19	2.01	2.01	2.01	0.32	0.11	0.12	0.12
0	3	100	125	2.19	2.01	-	1.84	0.32	0.11	-	1.34
0	3	200	50	2.17	2.01	2.00	2.00	0.24	0.08	0.09	0.09
0	3	200	150	2.17	2.01	2.00	2.00	0.24	0.08	0.08	0.08
0	3	200	250	2.17	2.01	-	1.94	0.24	0.08	-	0.63
0.5	3	100	25	2.29	2.28	2.21	2.23	0.43	0.31	0.31	0.33
0.5	3	100	75	2.28	2.28	2.40	2.41	0.40	0.31	0.43	0.44
0.5	3	100	125	2.28	2.28	-	1.77	0.39	0.31	-	1.19
0.5	3	200	50	2.22	2.25	2.83	2.85	0.28	0.26	0.84	0.86
0.5	3	200	150	2.21	2.25	2.39	2.39	0.27	0.26	0.40	0.40
0.5	3	200	250	2.21	2.25	-	1.93	0.26	0.26	-	0.52
0.9	3	100	25	2.34	2.46	2.32	2.35	0.44	0.48	0.39	0.42
0.9	3	100	75	2.34	2.46	2.58	2.59	0.47	0.48	0.60	0.61
0.9	3	100	125	2.34	2.46	-	1.72	0.45	0.48	-	1.32
0.9	3	200	50	2.25	2.44	2.97	2.96	0.30	0.44	0.98	1.00
0.9	3	200	150	2.24	2.44	2.57	2.57	0.29	0.44	0.57	0.58
0.9	3	200	250	2.24	2.44	-	1.91	0.28	0.44	-	0.55

Table 15: Simulations using Design-III & Non-AR Errors ( $\gamma = 0$ , 500 Reps)  
[Notes: True value of  $\beta_1$  is 2. We cannot estimate 2SLS when  $N > T$ .]

$\rho$	r	T	N	$E(\hat{\beta}_1)$				RMSE( $\hat{\beta}_1$ )			
				SIF	OLS	2SLS	FIV	SIF	OLS	2SLS	FIV
0	3	100	25	2.24	2.05	2.01	2.01	0.39	0.15	0.27	0.28
0	3	100	75	2.23	2.05	2.06	2.06	0.39	0.15	0.15	0.15
0	3	100	125	2.22	2.05	-	1.87	0.37	0.15	-	1.32
0	3	200	50	2.21	2.06	2.05	2.05	0.29	0.11	0.11	0.11
0	3	200	150	2.21	2.06	2.06	2.06	0.29	0.11	0.11	0.11
0	3	200	250	2.20	2.06	-	1.97	0.29	0.11	-	0.62
0.5	3	100	25	2.33	2.32	2.27	2.30	0.46	0.36	0.41	0.43
0.5	3	100	75	2.31	2.32	2.48	2.49	0.44	0.36	0.51	0.52
0.5	3	100	125	2.30	2.32	-	1.64	0.43	0.36	-	2.53
0.5	3	200	50	2.25	2.28	3.00	2.99	0.32	0.30	1.01	1.02
0.5	3	200	150	2.23	2.28	2.47	2.47	0.30	0.30	0.48	0.48
0.5	3	200	250	2.23	2.28	-	1.94	0.30	0.30	-	0.60
0.9	3	100	25	2.40	2.55	2.41	2.44	0.52	0.57	0.51	0.54
0.9	3	100	75	2.38	2.55	2.69	2.70	0.50	0.57	0.72	0.73
0.9	3	100	125	2.37	2.55	-	1.78	0.51	0.57	-	2.46
0.9	3	200	50	2.29	2.52	3.16	3.14	0.35	0.52	1.17	1.19
0.9	3	200	150	2.27	2.52	2.69	2.69	0.34	0.52	0.69	0.70
0.9	3	200	250	2.27	2.52	-	1.93	0.33	0.52	-	0.66

Table 16: Simulations using Design-III & AR Errors ( $\gamma = 0$ , 500 Reps)  
[Notes: True value of  $\beta_1$  is 2. We cannot estimate 2SLS when  $N > T$ .]

#### B.2.4 More Designs

This subsection contains the simulation results of designs with features of the already discussed ones. We provide a small table that tabulates the simulation tables in this sub-section. To save space, we denote “Serially Correlated Factors” by “AR Factors?”. Supervision and dimension reduction are required in all the designs in this section. Non-linearity is sometimes required, but sometimes may not be. We add a new column to indicate whether non-linearity is to be taken care of.

Table-No	True Number of Factors( $r$ )	AR Factors	AR Errors	$m(\cdot)$ is Non-linear
Table-17	5	Yes	No	Yes
Table-18	5	Yes	Yes	Yes
Table-19	5	No	No	Yes
Table-20	5	No	Yes	Yes
Table-21	8	Yes	No	No
Table-22	8	Yes	No	Yes
Table-23	8	Yes	Yes	No
Table-24	8	Yes	Yes	Yes
Table-25	8	No	No	No
Table-26	8	No	No	Yes
Table-27	8	No	Yes	No
Table-28	8	No	Yes	Yes

$\rho$	r	T	N	$E(\hat{\beta}_1)$				RMSE( $\hat{\beta}_1$ )			
				SIF	OLS	2SLS	FIV	SIF	OLS	2SLS	FIV
0	5	100	25	2.23	2.01	1.96	1.97	0.42	0.12	0.22	0.23
0	5	100	75	2.23	2.01	1.96	1.97	0.42	0.12	0.22	0.23
0	5	100	125	2.22	2.01	2.01	1.49	0.39	0.12	0.12	2.51
0	5	200	50	2.18	2.00	2.00	2.00	0.26	0.09	0.10	0.10
0	5	200	150	2.17	2.00	2.00	2.00	0.28	0.09	0.09	0.09
0	5	200	250	2.17	2.00	-	1.82	0.27	0.09	-	0.94
0.5	5	100	25	2.32	2.23	2.17	2.19	0.49	0.27	0.30	0.31
0.5	5	100	75	2.28	2.23	2.34	2.35	0.41	0.27	0.38	0.39
0.5	5	100	125	2.27	2.23	2.23	1.53	0.41	0.27	0.27	1.27
0.5	5	200	50	2.23	2.21	2.76	2.79	0.32	0.23	0.78	0.82
0.5	5	200	150	2.21	2.21	2.33	2.34	0.29	0.23	0.35	0.35
0.5	5	200	250	2.21	2.21	-	1.81	0.29	0.23	-	0.78
0.9	5	100	25	2.27	2.23	2.23	1.53	0.41	0.27	0.27	1.27
0.9	5	100	75	2.34	2.39	2.50	2.51	0.47	0.42	0.53	0.54
0.9	5	100	125	2.32	2.39	2.39	1.53	0.45	0.42	0.42	1.26
0.9	5	200	50	2.25	2.37	2.90	2.93	0.33	0.38	0.92	0.96
0.9	5	200	150	2.23	2.37	2.83	2.53	0.31	0.38	0.45	0.48
0.9	5	200	250	2.23	2.37	-	1.80	0.31	0.38	-	0.79

Table 17: Simulations in Non-Linear Design & Non-AR Errors ( $\gamma = 0.5$ , 500 Reps)  
[Notes: True value of  $\beta_1$  is 2. We cannot estimate 2SLS when  $N > T$ .]

$\rho$	r	T	N	$E(\hat{\beta}_1)$				$RMSE(\hat{\beta}_1)$			
				SIF	OLS	2SLS	FIV	SIF	OLS	2SLS	FIV
0	5	100	25	2.29	2.05	2.01	2.03	0.59	0.16	0.30	0.31
0	5	100	75	2.26	2.05	2.05	2.05	0.49	0.16	0.17	0.17
0	5	100	125	2.24	2.05	-	1.61	0.47	0.16	-	2.30
0	5	200	50	2.22	2.05	2.05	2.05	0.37	0.12	0.14	0.14
0	5	200	150	2.21	2.05	2.05	2.05	0.35	0.12	0.13	0.13
0	5	200	250	2.20	2.05	-	1.87	0.35	0.12	-	1.14
0	5	400	100	2.16	2.04	2.05	2.05	0.23	0.09	0.11	0.11
0	5	400	300	2.16	2.04	2.04	2.04	0.23	0.09	0.10	0.10
0	5	400	500	2.16	2.04	-	1.91	0.23	0.09	-	0.86
0.5	5	100	25	2.35	2.28	2.23	2.26	0.59	0.33	0.41	0.43
0.5	5	100	75	2.31	2.28	2.40	2.41	0.60	0.33	0.45	0.47
0.5	5	100	125	2.33	2.28	-	1.33	0.54	0.33	-	2.19
0.5	5	200	50	2.26	2.25	2.90	2.93	0.40	0.28	0.93	0.97
0.5	5	200	150	2.25	2.25	2.40	2.40	0.37	0.28	0.42	0.43
0.5	5	200	250	2.25	2.25	-	1.81	0.37	0.28	-	1.21
0.5	5	400	100	2.18	2.23	2.93	2.95	0.24	0.24	0.94	0.98
0.5	5	400	300	2.17	2.23	2.36	2.37	0.24	0.24	0.38	0.38
0.5	5	400	500	2.17	2.23	-	1.91	0.24	0.24	-	0.78
0.9	5	100	25	2.45	2.47	2.37	2.40	0.70	0.51	0.50	0.53
0.9	5	100	75	2.42	2.47	2.60	2.61	0.77	0.51	0.64	0.65
0.9	5	100	125	2.41	2.47	-	1.36	0.60	0.51	-	2.27
0.9	5	200	50	2.30	2.45	3.07	3.10	0.41	0.46	1.09	1.13
0.9	5	200	150	2.29	2.45	2.60	2.60	0.40	0.46	0.61	0.62
0.9	5	200	250	2.28	2.45	-	1.79	0.39	0.46	-	1.25
0.9	5	400	100	2.20	2.43	3.09	3.11	0.27	0.44	1.11	1.14
0.9	5	400	300	2.20	2.43	2.66	2.67	0.26	0.44	0.67	0.68
0.9	5	400	500	2.20	2.43	-	1.90	0.26	0.44	-	0.81

Table 18: Simulations in Non-Linear Design & AR Errors ( $\gamma = 0.5$ , 500 Reps)  
[Notes: True value of  $\beta_1$  is 2. We cannot estimate 2SLS when  $N > T$ .]

$\rho$	r	T	N	$E(\hat{\beta}_1)$				RMSE( $\hat{\beta}_1$ )			
				SIF	OLS	2SLS	FIV	SIF	OLS	2SLS	FIV
0	5	100	25	2.23	2.01	1.96	1.97	0.43	0.11	0.23	0.23
0	5	100	75	2.20	2.01	2.01	2.01	0.37	0.11	0.12	0.12
0	5	100	125	2.20	2.01	-	1.74	0.37	0.11	-	1.20
0	5	200	50	2.18	2.00	2.00	2.00	0.26	0.09	0.10	0.10
0	5	200	150	2.17	2.00	2.00	2.00	0.28	0.09	0.09	0.09
0	5	200	250	2.17	2.00	-	1.82	0.27	0.09	-	0.94
0.5	5	100	25	2.34	2.28	2.21	2.23	0.49	0.31	0.31	0.33
0.5	5	100	75	2.32	2.28	2.40	2.41	0.45	0.31	0.42	0.43
0.5	5	100	125	2.31	2.28	2.28	1.68	0.43	0.31	0.31	1.73
0.5	5	200	50	2.23	2.21	2.76	2.80	0.32	0.23	0.78	0.82
0.5	5	200	150	2.21	2.21	2.33	2.34	0.29	0.23	0.35	0.35
0.5	5	200	250	2.21	2.21	-	1.81	0.29	0.23	-	0.78
0.9	5	100	25	2.42	2.46	2.33	2.35	0.56	0.48	0.40	0.42
0.9	5	100	75	2.40	2.46	2.58	2.59	0.57	0.48	0.60	0.60
0.9	5	100	125	2.39	2.46	2.46	1.65	0.54	0.48	0.48	1.95
0.9	5	200	50	2.25	2.37	2.90	2.92	0.33	0.38	0.92	0.96
0.9	5	200	150	2.24	2.37	2.49	2.50	0.32	0.38	0.50	0.51
0.9	5	200	250	2.23	2.37	-	1.80	0.31	0.38	-	0.79

Table 19: Simulations in Non-Linear Design & Non-AR Errors ( $\gamma = 0$ , 500 Reps)  
[Notes: True value of  $\beta_1$  is 2. We cannot estimate 2SLS when  $N > T$ .]

$\rho$	r	T	N	$E(\hat{\beta}_1)$				RMSE( $\hat{\beta}_1$ )			
				SIF	OLS	2SLS	FIV	SIF	OLS	2SLS	FIV
0	5	100	25	2.27	2.05	2.02	2.02	0.53	0.15	0.27	0.28
0	5	100	75	2.23	2.05	2.06	2.06	0.43	0.15	0.15	0.15
0	5	100	125	2.24	2.05	-	1.74	0.45	0.15	-	1.56
0	5	200	50	2.22	2.06	2.05	2.05	0.34	0.11	0.11	0.11
0	5	200	150	2.21	2.06	2.06	2.06	0.32	0.11	0.11	0.11
0	5	200	250	2.21	2.06	-	1.94	0.32	0.11	-	0.88
0	5	400	100	2.17	2.05	2.05	2.05	0.23	0.09	0.09	0.10
0	5	400	300	2.17	2.05	2.05	2.05	0.23	0.09	0.09	0.09
0	5	400	500	2.17	2.05	-	1.97	0.22	0.09	-	0.59
0.5	5	100	25	2.38	2.32	2.28	2.30	0.56	0.36	0.41	0.43
0.5	5	100	75	2.37	2.32	2.47	2.49	0.54	0.36	0.51	0.52
0.5	5	100	125	2.36	2.32	-	1.69	0.53	0.36	-	1.58
0.5	5	200	50	2.28	2.28	2.98	2.98	0.37	0.30	0.99	1.01
0.5	5	200	150	2.26	2.28	2.47	2.47	0.34	0.30	0.48	0.48
0.5	5	200	250	2.25	2.28	-	1.90	0.33	0.30	-	0.92
0.5	5	400	100	2.20	2.26	2.99	3.01	0.26	0.27	1.00	1.02
0.5	5	400	300	2.19	2.26	2.41	2.42	0.24	0.27	0.42	0.43
0.5	5	400	500	2.19	2.26	-	1.96	0.24	0.27	-	0.59
0.9	5	100	25	2.48	2.55	2.43	2.45	0.64	0.57	0.52	0.54
0.9	5	100	75	2.46	2.55	2.69	2.70	0.64	0.57	0.71	0.73
0.9	5	100	125	2.45	2.55	-	1.64	0.60	0.57	-	1.68
0.9	5	200	50	2.34	2.52	3.15	3.12	0.41	0.52	1.15	1.17
0.9	5	200	150	2.31	2.52	2.68	2.69	0.37	0.52	0.69	0.69
0.9	5	200	250	2.30	2.52	-	1.86	0.37	0.52	-	0.92
0.9	5	400	100	2.23	2.50	3.16	3.16	0.28	0.50	1.16	1.18
0.9	5	400	300	2.23	2.50	2.80	3.16	0.27	0.50	0.70	0.78
0.9	5	400	500	2.22	2.50	-	1.93	0.26	0.50	-	0.59

Table 20: Simulations in Non-Linear Design & AR Errors ( $\gamma = 0$ , 500 Reps)  
[Notes: True value of  $\beta_1$  is 2. We cannot estimate 2SLS when  $N > T$ .]



$\rho$	r	T	N	$E(\hat{\beta}_1)$				RMSE( $\hat{\beta}_1$ )			
				SIF	OLS	2SLS	FIV	SIF	OLS	2SLS	FIV
0	8	100	25	2.07	2.01	1.96	1.97	0.33	0.10	0.19	0.20
0	8	100	75	2.05	2.01	2.01	2.01	0.24	0.10	0.11	0.11
0	8	100	125	2.04	2.01	-	1.84	0.23	0.10	-	0.48
0	8	200	50	2.02	2.01	2.00	2.00	0.17	0.07	0.08	0.08
0	8	200	150	2.02	2.01	2.00	2.00	0.16	0.07	0.07	0.07
0	8	200	250	2.03	2.01	-	1.93	0.16	0.07	-	0.33
0.5	8	100	25	2.14	2.33	2.18	2.20	0.39	0.36	0.27	0.30
0.5	8	100	75	2.15	2.33	2.44	2.45	0.31	0.36	0.46	0.47
0.5	8	100	125	2.14	2.33	-	1.78	0.29	0.36	-	0.50
0.5	8	200	50	2.11	2.30	2.71	2.74	0.22	0.31	0.71	0.74
0.5	8	200	150	2.06	2.30	2.43	2.43	0.19	0.31	0.43	0.44
0.5	8	200	250	2.06	2.30	-	1.92	0.18	0.31	-	0.32
0.9	8	100	25	2.21	2.56	2.30	2.32	0.44	0.57	0.36	0.38
0.9	8	100	75	2.21	2.56	2.64	2.65	0.34	0.57	0.65	0.66
0.9	8	100	125	2.21	2.56	-	1.77	0.34	0.57	-	0.53
0.9	8	200	50	2.14	2.53	2.83	2.86	0.24	0.53	0.84	0.87
0.9	8	200	150	2.10	2.53	2.63	2.63	0.21	0.53	0.63	0.63
0.9	8	200	250	2.09	2.53	-	1.91	0.20	0.53	-	0.33

Table 21: Simulations in Linear Design & Non-AR Errors ( $\gamma = 0.5$ , 500 Reps)  
[Notes: True value of  $\beta_1$  is 2. We cannot estimate 2SLS when  $N > T$ .]

$\rho$	r	T	N	$E(\hat{\beta}_1)$				RMSE( $\hat{\beta}_1$ )			
				SIF	OLS	2SLS	FIV	SIF	OLS	2SLS	FIV
0	8	100	25	2.28	2.01	1.96	1.96	0.70	0.12	0.23	0.23
0	8	100	75	2.21	2.01	2.01	2.01	0.39	0.12	0.13	0.13
0	8	100	125	2.20	2.01	-	1.51	0.42	0.12	-	1.41
0	8	200	50	2.18	2.00	2.00	2.00	0.30	0.09	0.10	0.11
0	8	200	150	2.17	2.00	2.00	2.00	0.28	0.09	0.09	0.09
0	8	200	250	2.16	2.00	-	1.79	0.26	0.09	-	1.86
0.5	8	100	25	2.44	2.23	2.18	2.19	0.77	0.27	0.30	0.31
0.5	8	100	75	2.33	2.23	2.34	2.35	0.49	0.27	0.38	0.38
0.5	8	100	125	2.30	2.23	-	1.44	0.48	0.27	-	1.53
0.5	8	200	50	2.27	2.21	2.74	2.77	0.37	0.23	0.76	0.79
0.5	8	200	150	2.23	2.21	2.33	2.33	0.31	0.23	0.35	0.35
0.5	8	200	250	2.22	2.21	-	1.69	0.31	0.23	-	1.42
0.9	8	100	25	2.51	2.39	2.28	2.30	0.77	0.42	0.37	0.39
0.9	8	100	75	2.39	2.39	2.50	2.51	0.54	0.42	0.52	0.53
0.9	8	100	125	2.32	2.39	-	1.53	0.45	0.42	-	1.26
0.9	8	200	50	2.32	2.37	2.88	2.91	0.42	0.38	0.90	0.93
0.9	8	200	150	2.27	2.37	2.49	2.49	0.35	0.38	0.50	0.50
0.9	8	200	250	2.23	2.37	-	1.80	0.31	0.38	-	0.79

Table 22: Simulations in Non-Linear Design & Non-AR Errors ( $\gamma = 0.5$ , 500 Reps)  
[Notes: True value of  $\beta_1$  is 2. We cannot estimate 2SLS when  $N > T$ .]

$\rho$	r	T	N	$E(\hat{\beta}_1)$				$\text{RMSE}(\hat{\beta}_1)$			
				SIF	OLS	2SLS	FIV	SIF	OLS	2SLS	FIV
0	8	200	50	2.05	2.06	2.05	2.04	0.26	0.11	0.11	0.11
0	8	200	150	2.06	2.06	2.06	2.06	0.26	0.11	0.11	0.11
0	8	200	250	2.06	2.06	-	1.94	0.26	0.11	-	0.53
0	8	400	50	2.03	2.05	2.03	2.03	0.18	0.09	0.09	0.09
0	8	400	150	2.03	2.05	2.04	2.05	0.18	0.09	0.09	0.09
0	8	400	250	2.03	2.05	2.05	2.05	0.18	0.09	0.09	0.09
0	8	400	100	2.03	2.05	2.04	2.04	0.18	0.09	0.09	0.09
0	8	400	300	2.03	2.05	2.05	2.05	0.18	0.09	0.09	0.10
0	8	400	500	2.03	2.05	-	1.96	0.18	0.09	-	0.36
0.5	8	200	50	2.14	2.33	2.84	2.87	0.31	0.35	0.85	0.89
0.5	8	200	150	2.11	2.33	2.49	2.50	0.28	0.35	0.51	0.51
0.5	8	200	250	2.10	2.33	-	1.90	0.27	0.35	-	0.51
0.5	8	400	50	2.09	2.31	2.95	2.96	0.23	0.32	0.96	1.00
0.5	8	400	150	2.05	2.31	2.75	2.76	0.20	0.32	0.76	0.77
0.5	8	400	250	2.05	2.31	2.49	2.57	0.19	0.32	0.52	0.49
0.5	8	400	100	2.06	2.31	2.87	2.89	0.20	0.32	0.87	0.90
0.5	8	400	300	2.05	2.31	2.45	2.46	0.19	0.32	0.46	0.47
0.5	8	400	500	2.05	2.31	-	1.97	0.19	0.32	-	0.36
0.9	8	200	50	2.21	2.61	2.99	3.03	0.35	0.62	1.00	1.04
0.9	8	200	150	2.17	2.61	2.74	2.74	0.32	0.62	0.75	0.75
0.9	8	200	250	2.16	2.61	-	1.88	0.31	0.62	-	0.53
0.9	8	400	50	2.13	2.59	3.06	3.05	0.24	0.59	1.07	1.10
0.9	8	400	150	2.08	2.59	2.78	2.79	0.21	0.59	0.78	0.79
0.9	8	400	250	2.08	2.59	-	2.79	0.21	0.59	-	0.57
0.9	8	400	100	2.10	2.59	3.01	3.02	0.22	0.59	1.02	1.04
0.9	8	400	300	2.08	2.59	2.70	2.71	0.21	0.59	0.71	0.71
0.9	8	400	500	2.08	2.59	-	1.96	0.21	0.59	-	0.36

Table 23: Simulations in Linear Design & AR Errors ( $\gamma = 0.5$ , 500 Reps)  
[Notes: True value of  $\beta_1$  is 2. We cannot estimate 2SLS when  $N > T$ .]

$\rho$	r	T	N	$E(\hat{\beta}_1)$				RMSE( $\hat{\beta}_1$ )			
				SIF	OLS	2SLS	FIV	SIF	OLS	2SLS	FIV
0	8	200	50	2.21	2.05	2.05	2.05	0.35	0.12	0.14	0.14
0	8	200	150	2.21	2.05	2.05	2.05	0.35	0.12	0.13	0.13
0	8	200	250	2.21	2.05	-	1.72	0.37	0.12	-	2.11
0	8	400	50	2.18	2.04	2.04	2.04	0.25	0.09	0.11	0.11
0	8	400	150	2.17	2.04	2.04	2.04	0.24	0.09	0.11	0.11
0	8	400	250	2.17	2.04	2.04	2.04	0.24	0.09	0.10	0.10
0	8	400	100	2.18	2.04	2.05	2.05	0.24	0.09	0.11	0.11
0	8	400	300	2.17	2.04	2.04	2.04	0.24	0.09	0.10	0.10
0	8	400	500	2.17	2.04	-	1.83	0.24	0.09	-	1.20
0.5	8	200	50	2.32	2.25	2.87	2.90	0.50	0.28	0.90	0.93
0.5	8	200	150	2.26	2.25	2.39	2.40	0.41	0.28	0.42	0.42
0.5	8	200	250	2.25	2.25	-	1.76	0.40	0.28	-	1.69
0.5	8	400	50	2.24	2.23	3.11	3.14	0.32	0.24	1.13	1.18
0.5	8	400	150	2.20	2.23	2.72	2.73	0.27	0.24	0.74	0.75
0.5	8	400	250	2.19	2.23	2.30	2.47	0.27	0.24	0.28	0.48
0.5	8	400	100	2.21	2.23	2.91	2.93	0.28	0.24	0.92	0.95
0.5	8	400	300	2.19	2.23	2.36	2.36	0.27	0.24	0.37	0.38
0.5	8	400	500	2.19	2.23	-	1.90	0.26	0.24	-	1.10
0.9	8	200	50	2.38	2.45	3.05	3.08	0.52	0.46	1.07	1.10
0.9	8	200	150	2.32	2.45	2.59	2.59	0.44	0.46	0.61	0.61
0.9	8	200	250	2.31	2.45	2.45	1.71	0.43	0.46	0.46	1.62
0.9	8	400	50	2.28	2.43	3.26	3.27	0.34	0.44	1.27	1.32
0.9	8	400	150	2.24	2.43	3.01	2.92	0.30	0.44	1.01	0.97
0.9	8	400	250	2.23	2.43	2.66	2.39	0.29	0.44	-	0.67
0.9	8	400	100	2.24	2.43	3.08	3.10	0.31	0.44	1.09	1.12
0.9	8	400	300	2.23	2.43	2.66	2.38	0.29	0.44	0.78	0.67
0.9	8	400	500	2.22	2.43	-	1.87	0.28	0.44	-	1.14

Table 24: Simulations in Non-Linear Design & AR Errors ( $\gamma = 0.5$ , 500 Reps)  
[Notes: True value of  $\beta_1$  is 2. We cannot estimate 2SLS when  $N > T$ .]

$\rho$	r	T	N	$E(\hat{\beta}_1)$				$RMSE(\hat{\beta}_1)$			
				SIF	OLS	2SLS	FIV	SIF	OLS	2SLS	FIV
0	8	100	25	2.10	2.01	1.95	1.96	0.36	0.10	0.20	0.21
0	8	100	75	2.05	2.01	2.01	2.01	0.27	0.10	0.10	0.10
0	8	100	125	2.04	2.01	-	1.91	0.26	0.10	-	0.51
0.5	8	100	25	2.17	2.35	2.22	2.24	0.41	0.37	0.31	0.32
0.5	8	100	75	2.21	2.35	2.47	2.47	0.34	0.37	0.48	0.49
0.5	8	100	125	2.19	2.35	-	1.87	0.33	0.37	-	0.52
0.9	8	100	25	2.26	2.59	2.35	2.37	0.48	0.60	0.40	0.43
0.9	8	100	75	2.29	2.59	2.68	2.68	0.42	0.60	0.69	0.70
0.9	8	100	125	2.28	2.59	-	1.85	0.39	0.60	-	0.55

Table 25: Simulations in Linear Design & Non-AR Errors ( $\gamma = 0$ , 500 Reps)  
[Notes: True value of  $\beta_1$  is 2. We cannot estimate 2SLS when  $N > T$ .]

$\rho$	r	T	N	$E(\hat{\beta}_1)$				$RMSE(\hat{\beta}_1)$			
				SIF	OLS	2SLS	FIV	SIF	OLS	2SLS	FIV
0	8	100	25	2.28	2.01	1.96	1.96	0.57	0.11	0.22	0.22
0	8	100	75	2.19	2.01	2.01	2.01	0.41	0.11	0.12	0.12
0	8	100	125	2.17	2.01	-	1.84	0.36	0.11	-	3.12
0.5	8	100	25	2.46	2.28	2.23	2.23	0.67	0.31	0.32	0.33
0.5	8	100	75	2.38	2.28	2.39	2.40	0.53	0.31	0.42	0.43
0.5	8	100	125	2.36	2.28	-	1.63	0.50	0.31	-	2.92
0.9	8	100	25	2.54	2.46	2.35	2.36	0.77	0.48	0.42	0.43
0.9	8	100	75	2.46	2.46	2.57	2.58	0.57	0.48	0.59	0.60
0.9	8	100	125	2.40	2.46	-	1.63	0.52	0.48	-	1.73

Table 26: Simulations in Non-Linear Design & Non-AR Errors ( $\gamma = 0$ , 500 Reps)  
[Notes: True value of  $\beta_1$  is 2. We cannot estimate 2SLS when  $N > T$ .]

$\rho$	r	T	N	$E(\hat{\beta}_1)$				$\text{RMSE}(\hat{\beta}_1)$			
				SIF	OLS	2SLS	FIV	SIF	OLS	2SLS	FIV
0	8	400	100	2.03	2.06	2.04	2.05	0.17	0.09	0.08	0.08
0	8	400	300	2.03	2.06	2.06	2.06	0.17	0.09	0.09	0.09
0	8	400	500	2.03	2.06	-	1.99	0.17	0.09	-	0.30
0.5	8	400	100	2.07	2.32	2.94	2.95	0.20	0.33	0.95	0.97
0.5	8	400	300	2.05	2.32	2.48	2.49	0.18	0.33	0.49	0.50
0.5	8	400	500	2.05	2.32	-	1.99	0.18	0.33	-	0.31
0.9	8	400	100	2.10	2.62	3.09	3.10	0.21	0.62	1.09	1.11
0.9	8	400	300	2.08	2.62	2.74	2.75	0.19	0.62	0.75	0.75
0.9	8	400	500	2.08	2.62	-	1.98	0.19	0.62	-	0.31

Table 27: Simulations in Linear Design & AR Errors ( $\gamma = 0$ , 500 Reps)  
[Notes: True value of  $\beta_1$  is 2. We cannot estimate 2SLS when  $N > T$ .]

$\rho$	r	T	N	$E(\hat{\beta}_1)$				$\text{RMSE}(\hat{\beta}_1)$			
				SIF	OLS	2SLS	FIV	SIF	OLS	2SLS	FIV
0	8	400	100	2.18	2.05	2.05	2.05	0.24	0.09	0.09	0.09
0	8	400	300	2.17	2.05	2.05	2.05	0.23	0.09	0.09	0.09
0	8	400	500	2.17	2.05	-	1.93	0.23	0.09	-	0.81
0.5	8	400	100	2.23	2.26	2.98	2.99	0.29	0.27	0.99	1.01
0.5	8	400	300	2.21	2.26	2.41	2.42	0.27	0.27	0.42	0.43
0.5	8	400	500	2.20	2.26	-	1.94	0.26	0.27	-	0.76
0.9	8	400	100	2.27	2.50	3.15	3.15	0.32	0.50	1.15	1.17
0.9	8	400	300	2.25	2.50	2.61	2.52	0.30	0.50	0.71	0.89
0.9	8	400	500	2.24	2.50	-	1.91	0.29	0.50	-	0.78

Table 28: Simulations in Non-Linear Design & AR Errors ( $\gamma = 0$ , 500 Reps)  
[Notes: True value of  $\beta_1$  is 2. We cannot estimate 2SLS when  $N > T$ .]

### B.3 PIR Simulations

Parametric Inverse Regression (PIR) is a supervised method of finding sufficient directions needed for the conditional mean. This method was developed by [Bura & Cook \(2001\)](#). In this section, we show the performance of our SIF method if we replace the sliced inverse regression (SIR) in the SDR direction estimation step by PIR.

$\rho$	r	T	N	$E(\hat{\beta}_1)$				RMSE( $\hat{\beta}_1$ )			
				SIF	OLS	2SLS	FIV	SIF	OLS	2SLS	FIV
0	5	100	25	2.06	2.01	1.95	1.96	0.28	0.10	0.20	0.21
0	5	100	75	2.05	2.01	2.01	2.01	0.28	0.10	0.10	0.10
0	5	100	125	2.05	2.01	-	1.95	0.27	0.10	-	0.39
0	5	200	50	2.04	2.01	2.00	2.00	0.18	0.07	0.07	0.07
0	5	200	150	2.04	2.01	2.00	2.00	0.18	0.07	0.07	0.07
0	5	200	250	2.04	2.01	-	1.98	0.18	0.07	-	0.25
0.5	5	100	25	2.13	2.35	2.21	2.24	0.33	0.37	0.30	0.32
0.5	5	100	75	2.15	2.35	2.47	2.48	0.31	0.37	0.49	0.50
0.5	5	100	125	2.14	2.35	-	1.91	0.29	0.37	-	0.44
0.5	5	200	50	2.10	2.32	2.77	2.79	0.22	0.33	0.78	0.80
0.5	5	200	150	2.08	2.32	2.46	2.46	0.20	0.33	0.46	0.46
0.5	5	200	250	2.08	2.32	-	1.97	0.20	0.33	-	0.26
0.9	5	100	25	2.19	2.59	2.33	2.36	0.36	0.60	0.39	0.42
0.9	5	100	75	2.21	2.59	2.68	2.69	0.35	0.60	0.69	0.70
0.9	5	100	125	2.20	2.59	-	1.89	0.33	0.60	-	0.49
0.9	5	200	50	2.14	2.56	2.90	2.92	0.25	0.56	0.91	0.93
0.9	5	200	150	2.11	2.56	2.67	2.67	0.22	0.56	0.67	0.67
0.9	5	200	250	2.10	2.56	-	1.96	0.21	0.56	-	0.27

Table 29: PIR Simulations in Design-I (4.1) & Non-AR Errors ( $\gamma = 0$ , 500 Reps)  
[Notes: True value of  $\beta_1$  is 2. We cannot estimate 2SLS when  $N > T$ .]

We observe similar patterns of SIF's performance as of the main text. This verifies that the use of the SIR method as the SDR direction estimator is a good choice.

$\rho$	r	T	N	$E(\hat{\beta}_1)$				RMSE( $\hat{\beta}_1$ )			
				SIF	OLS	2SLS	FIV	SIF	OLS	2SLS	FIV
0	3	100	25	2.27	2.39	2.27	2.30	0.37	0.42	0.36	0.40
0	3	100	75	2.26	2.39	2.51	2.52	0.36	0.42	0.53	0.54
0	3	100	125	2.26	2.39	-	1.67	0.35	0.42	-	1.35
0	3	200	50	2.02	2.01	2.00	2.00	0.16	0.07	0.07	0.07
0	3	200	150	2.02	2.01	2.00	2.00	0.16	0.07	0.07	0.07
0	3	200	250	2.02	2.01	-	1.97	0.16	0.07	-	0.18
0.5	3	100	25	2.07	2.33	2.17	2.22	0.25	0.36	0.26	0.31
0.5	3	100	75	2.08	2.33	2.45	2.46	0.26	0.36	0.47	0.48
0.5	3	100	125	2.07	2.33	-	1.92	0.25	0.36	-	0.27
0.5	3	200	50	2.04	2.30	2.72	2.77	0.18	0.31	0.73	0.79
0.5	3	200	150	2.04	2.30	2.43	2.43	0.19	0.31	0.44	0.44
0.5	3	200	250	2.04	2.30	-	1.96	0.19	0.31	-	0.19
0.9	3	100	25	2.10	2.56	2.28	2.34	0.28	0.57	0.33	0.41
0.9	3	100	75	2.11	2.56	2.65	2.66	0.28	0.57	0.66	0.67
0.9	3	100	125	2.10	2.56	-	1.92	0.28	0.57	-	0.27
0.9	3	200	50	2.06	2.53	2.85	2.84	0.19	0.53	0.85	0.91
0.9	3	200	150	2.06	2.53	2.63	2.64	0.20	0.53	0.64	0.64
0.9	3	200	250	2.06	2.53	-	1.96	0.20	0.53	-	0.19

Table 30: PIR Simulations in Design-II (4.2) with Non-AR Errors ( $\gamma = 0.5$ , 500 Reps)  
[Notes: True value of  $\beta_1$  is 2. We cannot estimate 2SLS when  $N > T$ .]



$\rho$	r	T	N	$E(\hat{\beta}_1)$				RMSE( $\hat{\beta}_1$ )			
				SIF	OLS	2SLS	FIV	SIF	OLS	2SLS	FIV
0	3	100	25	2.19	2.01	1.96	1.96	0.30	0.12	0.22	0.23
0	3	100	75	2.18	2.01	2.01	2.01	0.30	0.12	0.13	0.13
0	3	100	125	2.18	2.01	-	1.79	0.29	0.12	-	0.98
0	3	200	50	2.14	2.00	2.00	2.00	0.21	0.09	0.10	0.10
0	3	200	150	2.13	2.00	2.00	2.00	0.20	0.09	0.09	0.09
0	3	200	250	2.13	2.00	-	1.86	0.20	0.09	-	0.60
0.5	3	100	25	2.23	2.23	2.17	2.20	0.33	0.27	0.29	0.32
0.5	3	100	75	2.22	2.23	2.35	2.36	0.32	0.27	0.38	0.39
0.5	3	100	125	2.21	2.23	-	1.69	0.31	0.27	-	1.30
0.5	3	200	50	2.18	2.21	2.77	2.80	0.24	0.23	0.79	0.84
0.5	3	200	150	2.17	2.21	2.33	2.34	0.23	0.23	0.35	0.35
0.5	3	200	250	2.17	2.21	-	1.86	0.24	0.23	-	0.62
0.9	3	100	25	2.27	2.39	2.27	2.30	0.37	0.42	0.36	0.40
0.9	3	100	75	2.26	2.39	2.51	2.52	0.36	0.42	0.53	0.54
0.9	3	100	125	2.26	2.39	-	1.67	0.35	0.42	-	1.35
0.9	3	200	50	2.20	2.37	2.92	2.95	0.26	0.38	0.93	0.99
0.9	3	200	150	2.19	2.37	2.49	2.50	0.25	0.38	0.51	0.51
0.9	3	200	250	2.19	2.37	-	1.85	0.25	0.38	-	0.63

Table 31: PIR Simulations in Design-III (4.3) & Non-AR Errors ( $\gamma = 0.5$ , 500 Reps)  
[Notes: True value of  $\beta_1$  is 2. We cannot estimate 2SLS when  $N > T$ .]

#### B.4 Performance of Belloni *et al.* (2012) When Many Instruments are Correlated

Belloni *et al.* (2012) is a sparsity-based procedure that efficiently estimates the first stage of instrumental variable regression by selecting the important instruments from a pool of available ones. However, when the true instrument set is not sparse, that is if the instrument set is correlated, this method should not work well. We verify this idea in this section. Table-32 shows the performance of the OLS method and Post-Lasso-based IV method of Belloni *et al.* (2012). The true value of the parameter of interest  $\beta_1$  is 2. In the table-32, we report the actual estimate and the root mean squared error using 500 repetitions of simulation using Design-II (4.2). We choose Design-II so that one can't say that Belloni *et al.* (2012)'s method may not be doing well because of non-linearities or the need for supervision. We also show that in Design-III as well, Belloni *et al.* (2012)'s method does not do well. The aim of this section is that when the underlying structure of many instruments is not sparse, we need another method, therefore, this paper is an important contribution.

Simulation Design	$\rho$	T	N	$E(\hat{\beta}_1)$		RMSE( $\hat{\beta}_1$ )	
				OLS	Post-LASSO	OLS	Post-LASSO
Design-II (4.2)	0	200	50	2.06	2.05	0.11	0.11
Design-II (4.2)	0	200	150	2.06	2.05	0.11	0.11
Design-II (4.2)	0	200	250	2.06	2.05	0.11	0.11
Design-II (4.2)	0.5	200	50	2.33	2.93	0.35	0.94
Design-II (4.2)	0.5	200	150	2.33	2.82	0.35	0.83
Design-II (4.2)	0.5	200	250	2.33	2.73	0.35	0.75
Design-II (4.2)	0.9	200	50	2.61	3.05	0.62	1.06
Design-II (4.2)	0.9	200	150	2.61	2.96	0.62	0.97
Design-II (4.2)	0.9	200	250	2.61	2.92	0.62	0.93
Design-III (4.3)	0	200	50	2.05	2.06	0.12	0.14
Design-III (4.3)	0	200	150	2.05	2.07	0.12	0.14
Design-III (4.3)	0	200	250	2.05	2.06	0.12	0.13
Design-III (4.3)	0.5	200	50	2.25	3.07	0.28	1.10
Design-III (4.3)	0.5	200	150	2.25	2.92	0.28	0.96
Design-III (4.3)	0.5	200	250	2.25	2.82	0.28	0.87
Design-III (4.3)	0.9	200	50	2.45	3.21	0.46	1.24
Design-III (4.3)	0.9	200	150	2.45	3.07	0.46	1.10
Design-III (4.3)	0.9	200	250	2.45	2.98	0.46	1.02

Table 32: Expected value and RMSE of estimates of OLS and Post-LASSO method of [Belloni et al. \(2012\)](#)

[Notes: True value of  $\beta_1$  is 2. Other Parameters:  $\gamma = 0.5$ ,  $\rho = 0.5$ ,  $r = 3$ , number of replication =500]