Sufficient Instruments Filter

Rajveer Jat*

University of California Riverside

December 19, 2025

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 regressor. 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 with weak and even 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.

1 Introduction

The instrumental variable (IV) approach is a cornerstone in addressing endogeneity issues in econometrics. When faced with a large set of instruments, standard two-staged least square (2SLS) estimator becomes inconsistent as noted in Bekker (1994). Another key challenge is the potential presence of weak instruments, where many instruments exhibit only a weak correlation

*Email: rjat001@ucr.edu. Corresponding address: Sproul Hall, W Campus Dr, Riverside, CA 92507.

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 (1997)).

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), 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.[1] While methods such as Newey (1990) have been developed for asymptotically efficient instrumental variables estimation 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 instrument(s), 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 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 functional structure, which may lead to model misspecification, thereby under-utilizing the information available. Misspecification is a major issue in economics discussed in seminal

[1] 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. Carrasco (2012)'s regularized 2SLS can potentially address both non-linearities and high-dimensionality of instruments, but it requires all instruments to be valid and not weak which is stronger condition than Bai & Ng (2010).

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 irrepresentable 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, Bai & Ng (2010) and Kapetanios & Marcellino (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 use them to estimate the "first-stage" as a non-parametric function, 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. Inclusion of 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 aforementioned challenges in a comprehensive and unified framework.

When we face many instruments, presence of correlation among them is likely to be more plausible than sparsity especially so in macro-finance contexts. In areas like industrial organizations, particularly when considering different functional forms of instruments or when lagged versions of a variable also serve as instruments, the possibility of instruments being correlated with each other is non-trivial. 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 irrepresentable 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 correlated (see Table-33 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.

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, our approach is supervised, therefore, it avoids the inclusion of irrelevant factors, thereby enhancing the efficiency of the estimation procedure. On top of these features, our estimator sufficiently reduces the dimensionality to further gain efficiency in the estimation procdure.

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. 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. 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 $E(x \mid \mathbf{f})$ is the first moment of the conditional distribution of x given $\mathbf{f}$, but the broader goal is to understand how $F(x \mid \mathbf{f})$ behaves as $\mathbf{f}$ varies. To simplify this, $\mathbf{f}$ can be replaced by $L \leq r$ linear combinations of its components, $\boldsymbol{\theta}_1'\mathbf{f}, \ldots, \boldsymbol{\theta}_L'\mathbf{f}$, without

losing information about $F(x \mid \mathbf{f})$. 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. These combinations of factors $(\boldsymbol{\theta}_1'\mathbf{f}, \ldots, \boldsymbol{\theta}_L'\mathbf{f})$ are called sufficient dimension reduction (SDR) indices and they serve as true instruments in our procedure, meeting the necessary relevance and exclusion restrictions. 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 number of SDR indices identified in the third layer, yielding $\widehat{x}$, equivalent to the first stage in 2SLS. In the final layer, we perform OLS regression using the exogenous variation $\widehat{x}$, 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 regressing 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. Scalers 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 $t = 1, 2, \ldots, T$ 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. $\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_L$ 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 $(\boldsymbol{\theta}_1'\mathbf{f}_t, \ldots, \boldsymbol{\theta}_L'\mathbf{f}_t)$ 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 $i = 1, 2, \ldots, N$. For the clarity of the presentation, we refer to "noisy instruments" $\{z_{it}\}$ as instruments, $\mathbf{f}_t$ as factors, and to true instruments $(\boldsymbol{\theta}_1'\mathbf{f}_t, \ldots, \boldsymbol{\theta}_L'\mathbf{f}_t)$ 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.

$$y_t = \beta_0 + \beta_1 x_t + \varepsilon_t \tag{2.1}$$

$$x_t = m\left(\boldsymbol{\theta}_1'\mathbf{f}_t, \ldots, \boldsymbol{\theta}_L'\mathbf{f}_t\right) + e_t \tag{2.2}$$

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

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

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

$$\mathbf{Z} = \mathbf{F}\mathbf{B}' + \mathbf{U}$$

Our parameter of interest is $\boldsymbol{\beta} = (\beta_0 \quad \beta_1)'$. In particular, we care about β_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.

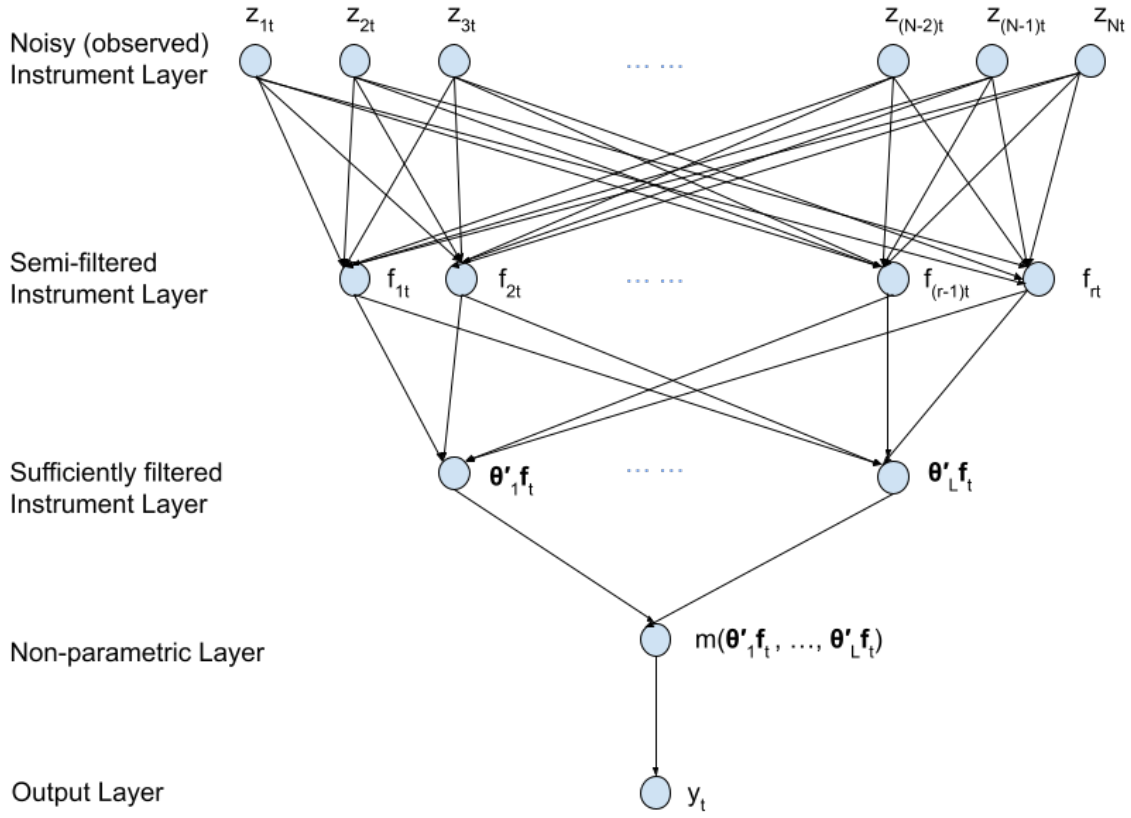

Figure 1: Architecture of the Sufficient Instrument Filter (SIF) estimation procedure

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 $\boldsymbol{\theta}_1'\mathbf{f}_t, \ldots, \boldsymbol{\theta}_L'\mathbf{f}_t$, 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 $\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_L$ are not identifiable without imposing structural conditions on $m(\cdot)$, however, we just need the subspace $S_{x|\mathbf{f}_t}$ spanned by $\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_L$, which can be identified. Therefore, throughout this paper, we refer to any orthonormal basis $\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_L$ of the central subspace $S_{x|\mathbf{f}_t}$ as sufficient dimension reduction directions, and their corresponding indices $\boldsymbol{\theta}_1'\mathbf{f}_t, \ldots, \boldsymbol{\theta}_L'\mathbf{f}_t$ 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. We already discussed that the 2SLS method becomes inconsistent when the number of instruments is high, therefore, we need to use a high-dimensional method to obtain $\widehat{x}_t$. 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[2]. 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 $\widehat{x}_t$, 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.

2.2 Identification

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] Jat & Padha (2024) is a recent non-linear and supervised forecasting method.

2.2.1 Relevancy Condition

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

$$E\Big[m(\boldsymbol{\theta_1}'\mathbf{f}_t, \ldots, \boldsymbol{\theta_L}'\mathbf{f}_t)x_t\Big] \neq 0 \tag{2.5}$$

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 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), $E(f_{jt}x_t) \neq 0$ *for all* $j = 1, 2, \ldots, r$, is not necessary for but is sufficient for our method. For example, if we take $\boldsymbol{\theta}_1$ as a vector of ones, $\boldsymbol{\theta}_2$ to $\boldsymbol{\theta}_L$ as a vector of zeros, and the function $m(\cdot)$ as a linear function, this gives us $E\Big[m(\boldsymbol{\theta_1}'\mathbf{f}_t, \ldots, \boldsymbol{\theta_L}'\mathbf{f}_t)x_t\Big] = E\Big[\boldsymbol{\theta_1}'\mathbf{f}_t x_t\Big] = E\Big[\mathbf{f}_t x_t\Big]$.

2.2.2 Exclusion Restriction

The exclusion restriction means that the true instruments $(\boldsymbol{\theta_1}'\mathbf{f}_t, \ldots, \boldsymbol{\theta_L}'\mathbf{f}_t)$ 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:

$$E\Big[m(\boldsymbol{\theta_1}'\mathbf{f}_t, \ldots, \boldsymbol{\theta_L}'\mathbf{f}_t)\varepsilon_t\Big] = 0$$

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 $\mathbf{f}_t$ from the large pool of available instruments $\{z_{it}\}$, $i = 1, 2, ..., N$. 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 indices: $\widehat{\boldsymbol{\theta}}_1\widehat{\mathbf{f}}_t, \ldots, \widehat{\boldsymbol{\theta}}_L\widehat{\mathbf{f}}_t$. 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 $\widehat{x}_t = \widehat{m}(\widehat{\boldsymbol{\theta}}_1\widehat{\mathbf{f}}_t, \ldots, \widehat{\boldsymbol{\theta}}_L\widehat{\mathbf{f}}_t)$. 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 $\widehat{x}_t$ which is very close to the true x_t.

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

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

$$\text{subject to } T^{-1}\mathbf{F}'\mathbf{F} = \mathbf{I}_{\boldsymbol{r}}, \quad \mathbf{B}'\mathbf{B} \text{ is diagonal} \tag{2.6}$$

Where $\mathbf{Z} = (\mathbf{z}_1, \ldots, \mathbf{z}_T), \mathbf{F}' = (\mathbf{f}_1, \ldots, \mathbf{f}_T)$, 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 $\widehat{\mathrm{F}}_r/\sqrt{T}$ 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 $\left\{\widehat{\mathbf{f}}_1, \ldots, \widehat{\mathbf{f}}_T\right\}$ 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, $E(\mathbf{z}_t \mid x_t)$. This is a key difference from Bai & Ng (2010), which uses unsupervised and linear counterpart $Cov(\boldsymbol{z}_t, x_t)$. By conditioning on the target x_t in the model-2.3, we derive

$$\operatorname{cov}\left(E\left(\mathbf{z}_t \mid x_t\right)\right) = \mathbf{B}\operatorname{cov}\left(E\left(\mathbf{f}_t \mid x_t\right)\right)\mathbf{B}'$$

Where we used the assumption that $E(\mathbf{u}_t \mid x_t) = 0$. Assuming that $E\left(\mathbf{b}'\mathbf{f}_t \mid \boldsymbol{\theta}_1'\mathbf{f}_t, \ldots, \boldsymbol{\theta}_L'\mathbf{f}_t\right)$ is a linear function of $\boldsymbol{\theta}_1'\mathbf{f}_t, \ldots, \boldsymbol{\theta}_L'\mathbf{f}_t$ for any $\mathbf{b} \in \mathbb{R}^r$, Li (1991) showed that $E(\mathbf{f}_t \mid x_t)$ is contained in the central subspace $S_{x|\mathbf{f}_t}$ spanned by $\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_L$. This important result implies that $S_{x|\mathbf{f}_t}$ contains the linear span of $\operatorname{cov}(E(\mathbf{f}_t \mid x_t))$. Therefore, we can get the SDR direction by investigating the top L eigenvectors of $\operatorname{cov}(E(\mathbf{f}_t \mid x_t))$. To see this, let $\boldsymbol{\Theta} = (\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_L)$. Then, for a $L \times 1$ coefficient function $\mathbf{a}(x_t)$, we can write

$$E\left(\mathbf{f}_t \mid x_t\right) = \Theta\mathbf{a}\left(x_t\right)$$

Therefore,

$$\operatorname{cov}\left(E\left(\mathbf{f}_t \mid x_t\right)\right) = \Theta E\left[\mathbf{a}\left(x_t\right)\mathbf{a}\left(x_t\right)^T\right]\Theta'$$

The quadratic form matrix $\operatorname{cov}\left(E\left(\mathbf{f}_t \mid x_t\right)\right)$ will have L eigenvalues bounded away from if $E\left[\mathbf{a}\left(x_t\right)\mathbf{a}\left(x_t\right)^T\right]$ is non-degenerate.

To obtain $cov(E\left(\mathbf{f}_t \mid x_t\right))$, we use sliced inverse regression methodology proposed by Li (1991). The estimator is given by:

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

We slice or divide the range of x_t into M number of intervals: $I_1, \ldots, I_M$ such that $P\left(x_t \in I_s\right) = 1/M$. Substituting $E\left(\mathbf{f}_t \mid x_t\right) = \Theta\mathbf{a}\left(x_t\right)$, we get:

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

If the matrix within the brackets in equation-2.6 is non-degenerate and M is at least $\max\{L, 2\}$, we are guaranteed to have L eigenvalues bounded away from zero. Therefore, the linear span generated by the eigenvectors corresponding to the L largest eigenvalues of $\Sigma_{\mathbf{f}|x}$ matches that generated by $\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_L$. To estimate $\Sigma_{\mathbf{f}|x}$ when factors are unobserved, we consistently derive the factors $\mathbf{f}_t$ from the factor model in equation 2.3, then utilize the estimated factors $\widehat{\mathbf{f}}_c$ along with 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 $\widehat{\Sigma}_{\widehat{\mathbf{f}}|x}$ as $N, T \to \infty$. Furthermore, the eigenvectors of $\widehat{\Sigma}_{\widehat{\mathbf{f}}|x}$ corresponding to the L largest eigenvalues, denoted as $\widehat{\boldsymbol{\theta}}_j (j \equiv 1, \ldots, L)$, 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 $\widehat{\boldsymbol{\theta}}_1'\widehat{\mathbf{f}}_t, \ldots, \widehat{\boldsymbol{\theta}}_t'\widehat{\mathbf{f}}_t$, the true instruments.

To accurately estimate $\Sigma_{\mathbf{f}|x}$, we will substitute the conditional expectations $E\left(\mathbf{f}_t \mid x_t \in I_s\right)$ with their sample equivalents. Let us denote the ordered statistics of $\left\{\left(x_t, \widehat{\mathbf{f}}_t\right)\right\}_{t=1,\ldots,T-1}$ by $\left\{\left(x_{(t)}, \widehat{\mathbf{f}}_{(t)}\right)\right\}_{t=1,\ldots,T}$ based on the values of x, arranged such that $x_{(2)} \leq \cdots \leq x_{(T)}$. We partition the range of x into M slices, with M typically being a fixed number. The first $M-1$ slices contain an equal number of observations, denoted as $c > 0$, while the last slice may contain fewer than c observations, which has minimal asymptotic impact. For clarity, we introduce a double

subscript notation $_{(s,j)}$, where $s = 1, \ldots, M$ indicates the slice number and $j = 1, \ldots, c$ denotes the index of an observation within a specific slice. Therefore, we can express $\left\{\left(x_t, \widehat{\mathbf{f}}_t\right)\right\}_{t=1,\ldots,T}$ as follows:

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

Using the estimated factors $\widehat{\mathbf{f}}$, we can represent the estimate $\widehat{\mathbf{\Sigma}}_{f|x}$ in the following form:

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

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 $\widehat{x}_t$. For simplicity, we use the local linear least square regression (Fan & Gijbels (1992)) to estimate $m(\cdot)$. We postpone the discussion on this step to the section-3.2.3 where we also discuss its asymptotic properties.

2.3.4 Estimation of Main Parameter

So far we have discussed how to estimate the $\widehat{x}_t$, the exogenous variation counterpart of x_t, popularly known as the "first-stage" in the 2SLS method. Now replace x_t by $\widehat{x}_t$ in equation-2.1 to estimate the parameter of interest $\boldsymbol{\beta} = (\beta_0 \quad \beta_1)'$ 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.

Algorithm 1	Sufficient Instrument Filter (SIF) Procedure
Step 1	Obtain the estimated factors $\left\{\widehat{\mathbf{f}}_t\right\}_{t=1,\ldots,T}$ from 2.5 and 2.6.
Step 2	Construct $\widehat{\Sigma}_{\mathbf{f}\|x}$ described in the equation-2.7.
Step 3	Obtain $\widehat{\boldsymbol{\theta}}_1, \cdots, \widehat{\boldsymbol{\theta}}_L$ from the L largest eigenvectors of $\widehat{\Sigma}_{\mathbf{f}\|x}$.
Step 4	Construct the predictive indices $\widehat{\boldsymbol{\theta}}_1'\widehat{\mathbf{f}}_t, \cdots, \widehat{\boldsymbol{\theta}}_L'\widehat{\mathbf{f}}_t$.
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 $\widehat{x}_t$.
Step 6	Use the $\widehat{x}_t$ obtained in step-5 in place of x_t in equation-2.1 and do OLS to get $\widehat{\boldsymbol{\beta}} = (\hat{\beta}_0 \quad \hat{\beta}_1)'$

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, $\widehat{\boldsymbol{\beta}}$. 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 ℓ_1 norm, respectively. The spectral norm is defined as the largest singular value of $\mathbf{B}$, while the ℓ_1 norm is the maximum absolute column sum. Additionally, the matrix ℓ_∞ norm, denoted $\|\mathbf{B}\|_\infty$, is the maximum absolute row sum. For symmetric matrices, the ℓ_1 norm is equivalent to the ℓ_∞ 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 $\boldsymbol{\beta}$. 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. $E\Big[m(\boldsymbol{\theta_1}'\mathbf{f}_t, \ldots, \boldsymbol{\theta_L}'\mathbf{f}_t)x_t\Big] \neq 0$
2. $E\Big[m(\boldsymbol{\theta_1}'\mathbf{f}_t, \ldots, \boldsymbol{\theta_L}'\mathbf{f}_t)\varepsilon_t\Big] = 0$
3. $E\Big[m(\boldsymbol{\theta_1}'\mathbf{f}_t, \ldots, \boldsymbol{\theta_L}'\mathbf{f}_t)e_t\Big] = 0$
4. *For all* i *and all* t, $E(u_{it}e_t) = 0$ *and* $E(u_{it}\varepsilon_t) = 0$
5. $\mathbb{E}[\varepsilon_t] = 0$, *and* $\sqrt{T}\left(\frac{1}{T}\sum_{t=1}^{T}\varepsilon_t - E(\varepsilon_t)\right) = O_p(1)$ *for all* t.
6. $\mathbb{E}[e_t] = 0$, *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. The Assumption-1.4 says that the relationship between the instruments and the endogenous regressor is through the factors and not through error $\boldsymbol{u}_t$. The Assumption-1.5 and 1.6 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 ε_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 endogenity 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.

We borrow our Assumption-2 from Fan *et al.* (2017), to make it easy to follow, we tried to keep the language the as theirs.

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

1. ***Pervasive Condition:*** *The loadings* $\mathbf{b}_i$ *satisfy* $\|\mathbf{b}_i\| \leq \mathcal{M}$ *for* $i = 1, \ldots, N$. *As* $N \to \infty$, *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* $E\left(\mathbf{b}'\mathbf{f}_t \mid \boldsymbol{\phi}_1'\mathbf{f}_t, \ldots, \boldsymbol{\phi}_L'\mathbf{f}_t\right)$ *is a linear function of* $\boldsymbol{\phi}_1'\mathbf{f}_t, \ldots, \boldsymbol{\phi}_L'\mathbf{f}_t$ *for any* $\mathbf{b} \in \mathbb{R}^N$, *where the vectors* $\boldsymbol{\phi}_i'$ *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 $E\left(\mathbf{f}_t \mid x_t\right)$ lies within the central subspace. Specifically, following Li (1991), Fan *et al.* (2017) states the following lemma:

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

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. For proof, one can see Fan *et al.* (2017).

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 σ-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, mean-zero, and mutually independent. Additionally,* $E\|\mathbf{f}_t\|^4 < \infty$ *and* $E\left(\|\mathbf{f}_t\|^2 \mid x_t\right) < \infty$. *For some positice constant* c*, the mixing coefficient* $\alpha(T) < c\rho^T$ *for all* $T \in \mathbb{Z}^+$ *and some* $\rho \in (0, 1)$. *Note that the* $\{\mathbf{f}_t\}_{t\geq 1}$, $\{\mathbf{u}_t\}_{t\geq 1}$, *and* $\{e_t\}_{t\geq 1}$ *are allowed to be serially correlated.*

The Assumption-3 is the same as Assumtion-3.2 of Fan *et al.* (2017) and 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 $E(\mathbf{u}_t \mid x_t) = 0$ for $t \geq 1$ suffices for the theory to hold. However, we retain this simplified assumption for clarity.

To consistently estimate factor and factor loading, we impose the following conditions on the residuals and dependencies in the factor model-2.3, similar to those in Bai (2003).

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), $E\left|N^{-1/2}(\mathbf{u}_s'\mathbf{u}_t - E(\mathbf{u}_s\mathbf{u}_t))\right|^4 \leq \mathcal{M}$.

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 $\boldsymbol{\beta} = (\beta_0 \quad \beta_1)$, 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*

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

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[3], 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.

Remark 1. *Under the Assumption-2.1,2.2, Assumption-3, and Assumption-4, the rotation matrix H asymptotically converges to identity matrix of order r. Therefore, we can obtain the actual factors, not just the rotation. For more details, please refer to Lemma-A.3 of Fan et al. (2017).*

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., $\widehat{\boldsymbol{\Sigma}}_{\mathbf{f}|x}$ 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.

[3]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.

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*

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

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

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

for $j = 1, \ldots, L$, *where* $\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_L$ *form an orthonormal basis for the central subspace* $S_{x|\mathbf{f}}$.

Proof. This result is proved in Theorem 3.1 of Fan *et al.* (2017). One need to replace their x_{it} with our z_{it} and their y_{t+1} with our x_t. □

Corollary 1 Under the same conditions of Lemma-3, for any $j = 1, 2, \ldots L$, 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 $k = 1, 2, \ldots, L$. For a given L dimensional point $\boldsymbol{w} = \{w_1, w_2, \ldots, w_L\}$ and a vector of bandwidths $\boldsymbol{h} = \{h_1, h_2, \ldots, h_L\}$, $\boldsymbol{\psi}_t = \left(\frac{\boldsymbol{w}_t - \boldsymbol{w}}{\boldsymbol{h}}\right) = \left(\frac{w_{t1}-w_1}{h_1}, \cdots, \frac{w_{tL}-w_L}{h_L}\right) = \left(\psi_{t1}, \ldots \psi_{tL}\right)$ is local deviation from it. We define the $L-$dimesnional kernel function to be the product of kernels with individual arguments. It weights an observation inversely based on its distance from our point $\boldsymbol{w}$:

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

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. $h_1 = h_2 = \ldots = h_L = h$. the joint density $g(x, w_1, \ldots, w_L)$ of our x_t and SDR indices can be given by the

following expression:

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

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

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

One can obtain the expression for conditional density of x given SDR indices $g(x \mid w_1, \ldots, w_L)$ by dividing the two expressions above. Using the definition of conditional mean which is $m(\boldsymbol{w}) = \int x_t g(x_t \mid \boldsymbol{w}) dx_t$. We use the local linear least square (LLLS) method for estimating $m(\boldsymbol{w})$. 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** $m(\cdot)$**:** $m(\boldsymbol{w})$ is twice continuously differentiable, and the second derivatives are bounded:*

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

2. ***Stationarity:** The process $\{(\boldsymbol{w}_t, e_t)\}$ is strictly stationary and ergodic.*

3. ***Mixing Condition:** The sequence $\{(\boldsymbol{w}_t, e_t)\}$ satisfies an α-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 $\boldsymbol{w}_t$ have bounded moments of order $2+\delta'$ for some $\delta' > 0$:*

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

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:*

$$h \to 0, \quad Th^L \to \infty \quad as \quad T \to \infty.$$

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

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 control 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. 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 asymptotic 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 $\widehat{m}(\boldsymbol{w})$ of the non-parametric function $m(\boldsymbol{w})$, has the following asymptotic properties:*

1. ***Bias**:*

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

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

2. ***Variance**:*

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

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

3. ***Asymptotic Normality**:*

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

where $\xrightarrow{d}$ denotes convergence in distribution, $tr(\cdot)$ 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 □

The expression $\widehat{m}(\widehat{\boldsymbol{w}}) - m(\boldsymbol{w})$ appears in the expression of $\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}$ which gives us asymptotic normality. The expression $\widehat{m}(\widehat{\boldsymbol{w}}) - m(\boldsymbol{w})$ has $\widehat{\boldsymbol{w}}$ as an argument instead of $\boldsymbol{w}$ which should add one more term in the variance but $\widehat{\boldsymbol{w}}$ goes to $\boldsymbol{w}$ at a faster rate relative to non-parametric rate, therefore that term does not survive in the computation of asymptotic variance.

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

This section talks about the consistent estimation and the asymptotic normality of our estimator estimator $\boldsymbol{\beta}$, the main parameter of interest for inference.

Theorem 1. *(Consistency of $\widehat{\boldsymbol{\beta}}$) Under the assumptions 1-5, $\widehat{\boldsymbol{\beta}}$ is a consistent estimator of $\boldsymbol{\beta}$. Formally,*

$$\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta} = o_p(1)$$

Proof. We prove this result in appendix-A.1.2 □

There are in total seven terms that appear when we expand the closed-form expression of $\widehat{\boldsymbol{\beta}}$. Out of these seven terms, two terms dominate the other five. These two terms come from non-parametric estimation which determines the rate of convergence of our estimator $\widehat{\boldsymbol{\beta}}$. Now we state the asymptotic normality results:

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

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

$$\delta_{NT}\left(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta} - \mathscr{B}\right) \xrightarrow{d} \mathcal{N}\left(0, \left(\frac{1}{T}Q'Q\right)^{-1} \mathbf{V} \left(\frac{1}{T}Q'Q\right)^{-1}\right),$$

where, $\mathscr{B} = \left(\frac{Q'Q}{T}\right)^{-1} \begin{bmatrix} -\frac{1}{T}\sum_{t=1}^{T}\left(\frac{1}{2}\delta_{NT}^{-1}tr(\boldsymbol{G})\right)\beta_1 \\ \frac{1}{T}\sum_{t=1}^{T}\left(\frac{1}{2}\delta_{NT}^{-1}tr(\boldsymbol{G})\right)\left[m(\boldsymbol{w}_t) - \frac{1}{2}\delta_{NT}^{-1}tr(\boldsymbol{G})\right]\beta_1 \end{bmatrix}$ *and*

$$\boldsymbol{V} = \begin{bmatrix} V_{11} & V_{12} \\ V_{21} & V_{22} \end{bmatrix}.$$

Elements of $\mathbf{V}$ *are defined in the Appendix-A.1.3.*

Proof. The proof of this theorem is given in the Appendix-A.1.3 □

The elements of matrix $\mathbf{V}$ contain β_0 and β_1 and $m(\boldsymbol{w}_t)$ 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 $\delta_{NT} = \min\{N^{\frac{1}{2}}, T^{\frac{2}{L+4}}\}$, therefore we need both N and T to approach ∞ for our asymptotic results. One may ask why one should use our method when it has both bias and a slower rate, we answer that our method is flexible and this is the cost of the same similar to non-parametric estimators. However, we can kill bias and improve rates by putting more restrictions, which we leave for future research.

Remark 2. *(Asymptotic Bias) The asymptotic bias term $\mathscr{B}$ goes to zero but persists with δ_{NT} rate, which is standard in non-parametric estimation. To kill this bias term, 'undersmoothing' is employed i.e. choose a bandwidth h smaller than the optimal $h^* \propto T^{\frac{-1}{L+4}}$.*[4] *However, under-*

[4] Pagan & Ullah (1999) notes in their Chapter-3, Assumption A10 that $(\sqrt{Th})h^2 \to 0$ as $T \to \infty$ is needed to kill the bias term. At optimal bandwidth i.e. h^*, $(\sqrt{Th})h^2 \propto \sqrt{(TT^{\frac{-1}{L+4}})}\left(T^{\frac{-2}{L+4}}\right) \propto$ constant, does not go to zero as $T \to \infty$. Therefore, a smaller h i.e. undersmoothing is required.

smoothing means lesser data points availability around a point, therefore may inflate variance. Therefore, we stick to optimal bandwidth selection with bias term. Simulation evidences shows that bias decays as sample size grow.

Remark 3. *(On the convergence rate of our estimator): We take a conservative approach similar to the conservative rates of LASSO. Bühlmann & Van De Geer (2011) discuss in their theory for the LASSO section that under general conditions, the ℓ_2-norm of the LASSO estimator converges slower than $O_p\left(\sqrt{\frac{s\log p}{n}}\right)$. However, if further restrictions in the form of "compatibility conditions," which include "restricted eigenvalue conditions," are imposed, the estimator follows a faster rate of $O_p\left(\sqrt{\frac{s\log p}{n}}\right)$, where s is the number of nonzero parameters, p is the number of variables (dimensionality), and n is the sample size. Similar to this, we suggest that our estimator is slower than the $\sqrt{T}$ rate. After imposing additional conditions, we conjecture that we can improve the rate to $\sqrt{T}$. We leave the task of rate improvement for future research.*

3.2.5 On Semi-parametric Efficiency

There is a large literature on semi-parametric estimation which states that two-step estimators, where the first step is a function rather than a finite-dimensional parameter, can be $\sqrt{T}$-consistent, even though the convergence rate for the first-step functions is slower than $\sqrt{T}$. Newey (1994) and Newey & McFadden (1994) list a set of regularity conditions for $\sqrt{T}$-consistency of the second-stage estimator when the first-stage estimator could be a slower non-parametric function.

While it is not difficult to show that two second-order regularity conditions—*Linearization* and *Stochastic Equicontinuity*—required by Newey (1994) hold in our framework, we need to impose further restrictions on our model to verify the remaining conditions. This is because Newey (1994)'s results are developed for i.i.d. data and when the arguments of the first-stage function are known. In our framework, two major challenges arise: we have dependent data, and our arguments in the non-parametric function are generated regressors instead of regular variables. Therefore, verifying the remaining conditions is difficult and beyond the scope of this paper, so we defer this work to future research.

3.2.6 Bootstrap for Confidence Intervals

Since the expression for the asymptotic variance of our estimator is complicated, we suggest an alternative method for constructing confidence intervals. Given that our data exhibits serial correlation, bootstrap methods that preserve the serial dependence structure should be employed. Carlstein (1986) proposed the Block-bootstrap to handle time-dependent data. The assumptions stated in our paper for dependent data satisfy the conditions required (stationarity and α-mixing) for the Block-bootstrap introduced by Carlstein (1986), allowing us to apply the Block-bootstrap to construct a confidence interval for our estimator.

The key difference between the regular bootstrap and the Block-bootstrap is that, in Block-bootstrap, we divide the data into several blocks and sample the blocks rather than individual observations. One key consideration is choosing the block length, i.e., the number of observations in a block. If the block size is small, we will have more blocks and therefore better asymptotic properties because we resample blocks, not individual observations. However, the blocks must be long enough to preserve the dependence structure of the data. The weaker the serial correlation, the smaller will be the optimal length of the blocks. Carlstein (1986) and Hall *et al.* (1995) suggest that the block length should be approximately $T^{1/3}$, where T is the sample size.

One can draw B bootstrap samples to obtain B estimates and then use quantiles for constructing confidence intervals. In our empirical applications, we demonstrate the use of bootstrap intervals.

4 Simulation

We evaluate the performance of our method across various data-generating processes (DGPs). Four designs are considered: the first three highlight the core strengths of our approach—supervised learning, dimension reduction, and the ability to capture non-linearities. The last design gives a level playing field to two strands of literature: sparsity-based and factor-based methods.

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 (SIF) against three competitors: Ordinary Least Squares (OLS), Two-Stage Least Squares (2SLS), and the Factor Instrumental Variable (FIV) estimator from Bai & Ng (2010) for the first three designs. The Post-Lasso IV (PLIV) estimator from Belloni

et al. (2012) is not a suitable competitor in the first three designs because it is designed for settings with sparse instruments. As shown in Table-33, Belloni *et al.* (2012)'s method performs worse than simple OLS across various designs, and therefore, we do not include this method in comparisons. However, in Design-IV, we report PLIV along with FIV and our method's performance.

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, creating a level playing field for both our method and competing methods. We will set the true number of factors that drive the instrument set to five, but only three of them will be relevant for the endogenous regressor x_t. As a result, our supervised method should outperform unsupervised methods such as Bai & Ng (2010) due to its ability to filter out the factors that are 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 forThe 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:

$$y_t = \beta_0 + \beta_1 x_t + \varepsilon_t \tag{4.1}$$

$$x_t = \boldsymbol{\phi}' \mathbf{f}_t + e_t \tag{4.2}$$

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

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

We set $\beta_0 = 0$ and the main parameter of interest $\beta_1 = 2$. We let $r = 5$ and $\boldsymbol{\phi} = (0.8, 0.5, 0.3, 0, 0)'$, which means that five factors drive the instrument set but only three are related to the endogenous regressor x_t. 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 (ε_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 ρ, which is the correlation between η_t and ζ_t ending up relating the errors ε_t and e_t. The disturbances (η_t, ζ_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 $\gamma_j = \alpha_1 = \alpha_2 = 0.5$ for all $j = 1, 2, \ldots, r$. To save some space, in the rest of the text, we write γ_j, $j = 1, 2, \ldots, r$ in vector form. $\boldsymbol{\gamma} = 0.5$ means that $\gamma_j = 0.5$ for all $j = 1, 2, \ldots, r$. The parameter σ_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 $\boldsymbol{\gamma}$ randomly while we keep them fixed. We keep $\boldsymbol{\gamma}$ 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 $\boldsymbol{\gamma}$ 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 β_1 is 2. We run 500 replications and report the expectation of the estimate, i.e., $E(\widehat{\beta}_1)$, and its root mean squared error (RMSE). We consider three values of N for a given T: $N < T$, $N \sim T$, and $N > T$ to demonstrate the efficacy of our method in the high-dimensional setting. We also consider three values of ρ. The case $\rho = 0$ represents when the data (y_t, x_t) is i.i.d.; $\rho = 0.5$ indicates a small to moderate endogeneity case, as η_t accounts for about half of the variation in ε_t. Similarly, $\rho = 0.9$ represents a moderate to strong endogeneity case. This results in a total of nine (ρ, 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

				$E(\widehat{\beta}_1)$				RMSE$(\widehat{\beta}_1)$			
ρ	r	T	N	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 ($\boldsymbol{\gamma} = \alpha_1 = \alpha_2 = 0.5$, 500 Reps)
[Notes: True value of β_1 is 2. We cannot estimate 2SLS when $N > T$.]

and RMSE whenever there is an endogeneity ($\rho \neq 0$). There are five major observations from the results in Table-1. First, our method is a reliable approach, exhibiting the least root mean squared error and lower bias in the majority of cases. The bias and RMSE approach zero as the sample size T increases, confirming the asymptotic theory. Second, as expected, in the case of no endogeneity ($\rho = 0$), OLS is the best method in terms of the lowest RMSE; however, it is important to note that the OLS estimate is not unbiased due to the presence of dependent data. Third, the 2SLS method performs better than OLS when endogeneity is introduced, but only when the number of instruments is small. As the number of instruments increases, the variance of the 2SLS estimator rises, leading to a deterioration in its performance, a finding that is corroborated by existing literature Bekker (1994), Berry *et al.* (1995). Fourth, an increase in N relative to T has little effect on our estimates relative to the impact observed on other competing methods. This highlights our method's capability to produce stable estimates while effectively handling both low and high-dimensional cases. Finally, an increase in sample size T reduces the bias and RMSE across all methods.

4.1.3 Discussion

In this simulation setting, the true $m(\cdot)$ is linear, providing an ideal environment for the Factor Index Variable (FIV) approach of Bai & Ng (2010) to function effectively. The key distinction here is that not all factors in the instrument set are relevant to the endogenous regressor. Our supervised method selectively identifies only the relevant factors for x_t during the sufficient index step, in contrast to Bai & Ng (2010), which incorporates all available factors. This supervisory capability allows our method to estimate $\widehat{x}_t$ more accurately than competing methods, thereby yielding a more efficient estimator. The advantages of our approach become increasingly evident as the sample size T and the strength of endogeneity (ρ) grow, as the gap in RMSE between our method and competitors widens.

One might wonder why our method exhibits less bias than OLS, even in the absence of endogeneity. This is attributed to the serial correlation present in the error terms ε_t and e_t. When these error terms are made serially uncorrelated, OLS outperforms our method under no endogeneity, as shown in Table 6 in the appendix B.2. Under conditions of endogeneity, the 2SLS estimator is expected to outperform OLS, a trend observable in the rows where $N = 25$. However, as the number of instruments N increases, the performance of 2SLS declines due to the increasing variance in the first-stage estimation, as noted in Berry *et al.* (1995) and discussed

in Belloni *et al.* (2012).

It is important to recognize that our method not only benefits from supervision but also involves the dimensionality reduction of r factors to L Sufficient Dimension Reductions (SDRs). Thus, the observed advantages may stem from both sources. The SDR step is critical to our method, making it challenging to disentangle the contributions of each source to the overall performance gains.

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. $\boldsymbol{\gamma} = 0.5$ but $\alpha_1 = \alpha_2 = 0$ in the Table-6 in the appendix-B.2. Also, we report results for $\gamma_j = \alpha_1 = \alpha_2 = 0$ in Table- 8 and for $(\gamma_j = 0, \alpha_1 = \alpha_2 = 0.5)$ in Table-9. Note that errors (ε_t, e_t) are serially uncorrelated but are still endogenous because η_t, and ζ_t are still correlated. The summary of these additional results is qualitatively the same as the findings observed so far in Table-1. Therefore, this verifies and robustifies the conclusion that our method can supervise the process to filter out the relevant factors.

4.2 Design-II: Gains from Dimension Reduction

In addition to the supervision and the ability to capture the non-linearities, another strength of our method is to achieve the sufficient dimension reduction required to obtain the $E(x_t \mid \mathbf{f}_t)$. While Bai & Ng (2010) achieves a dimensional reduction in the sense that N number of instruments are summarized into r number of factors where $r < N$ is a small number. However, in addition to this, our method further combines these r 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 $E(x_t \mid \mathbf{f}_t)$, 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 $\boldsymbol{\phi} = \{0.8, 0.5, 0.3\}$ in equations 4.1 to 4.4, ensuring that supervision will not

provide an advantage. Furthermore, by keeping $m(\cdot)$ as a linear function, we make sure that capturing non-linearities will not yield benefits. Therefore, if our method performs better than others (FIV of Bai & Ng (2010) in particular), the improvement should stem from the additional dimension reductions that reduce variance in the process. We maintain the rest of the settings as described in section 4.1. The results of this simulation design are presented in Table 2.

4.2.2 Results and Discussions

We observe that our method consistently outperforms the nearest competitors in terms of both bias and RMSE, demonstrating that further dimension reduction by combining factors into indices enhances estimation accuracy. Given that the remaining observations are similar to those outlined in the previous simulation design section (Section 4.1), we omit them here to conserve space.

It may seem surprising that a more complex, non-parametric method outperforms the simpler FIV approach of Bai & Ng (2010), especially when the true data-generating process aligns more closely with their method. To address this, we conducted several experiments on simplified DGPs to compare the performance of SDR-based and OLS-based two-step procedures. Our findings indicate that SDR combined with a non-parametric approach outperforms OLS-based FIV, particularly in the presence of endogeneity and serial correlation. The relative performance of the SDR-based method improves further as the strength of endogeneity, serial correlation, or both increases. Additionally, the SDR-based method demonstrates superior performance when the dimensionality N increases. Theoretically, while the SDR-based method may be marginally less effective than the OLS-based method when the DGP is IID, it consistently outperforms in more complex data environments.

To maintain transparency, we provide our code and a detailed report of the SDR vs. OLS experiments, accessible here. We encourage readers to explore the code, verify our results, and share any insights that might help refine our findings—we would greatly appreciate your feedback.

In conclusion, while it may be challenging to outperform OLS (or OLS-based methods) when the true DGP is linear, our method shows clear advantages when the DGP deviates from IID assumptions.

				$E(\widehat{\beta}_1)$				RMSE$(\widehat{\beta}_1)$			
ρ	r	T	N	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 ($\boldsymbol{\gamma} = 0.5$, 500 Reps)
[Notes: True value of β_1 is 2. We cannot estimate 2SLS when $N > T$.]

4.2.3 More Results for Robustness

Under $r = 3$ and linear $m(\cdot)$ case, we consider various combinations of $\boldsymbol{\gamma}$ and (α_1, α_2). The results for autocorrelated factors with serially uncorrelated errors i.e. $\boldsymbol{\gamma} = 0.5$ and $\alpha_1 = \alpha_2 = 0$ given in Table-10 in the appendix-B.2. Also, the result for $\boldsymbol{\gamma} = \alpha_1 = \alpha_2 = 0$ specification is in Table-12 and for $\boldsymbol{\gamma} = 0$, $\alpha_1 = \alpha_2 = 0.5$ in Table-13. The qualitative results are the same as discussed for the $\boldsymbol{\gamma} = \alpha_1 = \alpha_2 = 0.5$ 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 $m(\cdot)$ in equation 4.2 is a linear function. This can lead to a misspecified model; therefore, we estimate $m(\cdot)$ as a non-parametric function that captures any present non-linearities. In our design for this section, we keep all factors relevant for x_t but related in a non-linear fashion. This ensures that the improved performance of our method arises not from the supervision but from capturing the non-linearities.

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:

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

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 $S_{x|\mathbf{f}}$ 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 β_1 and the root mean squared error (RMSE) of the estimate calculated using 500 replications in Table-3.

				$E(\widehat{\beta}_1)$				RMSE($\widehat{\beta}_1$)			
ρ	r	T	N	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 ($\boldsymbol{\gamma} = 0.5$, 500 Reps)
[Notes: True value of β_1 is 2. We cannot estimate 2SLS when $N > T$.]

Most of the observations from Table 3 reiterate the findings of Table 1 discussed in section 4.1.

One distinct observation in Table 3 is that the gains from capturing non-linearity are not as apparent for smaller sample sizes. However, as the sample size begins to grow, our method's performance improves, surpassing that of the competitors.

4.3.3 More Results for Robustness

Under $r = 3$ and non-linear $m(\cdot)$ case, we consider various combinations of γ and (α_1, α_2). The results for auto-correlated factors with serially uncorrelated errors i.e. $\boldsymbol{\gamma} = 0.5$ and $\alpha_1 = \alpha_2 = 0$ given in Table-14 in the appendix-B.2. Also, the result for $\boldsymbol{\gamma} = \alpha_1 = \alpha_2 = 0$ specification is in Table-16 and for $\boldsymbol{\gamma} = 0$, $\alpha_1 = \alpha_2 = 0.5$ in Table-17. The qualitative results are the same as discussed for the $\boldsymbol{\gamma} = \alpha_1 = \alpha_2 = 0.5$ specification in the main analysis of this design.

4.4 Design-4: Both Sparsity and Factor Structure Present in Instruments

When an applied researcher uses many instruments, she may not know whether there is a sparse or factor structure in the instrument set. Therefore, it is important to demonstrate how our method compares with competitors in a situation where instruments are neither purely sparse nor purely from a factor structure. In this design, we allow half of the instruments to share a factor structure, while the remaining half is sparse. This design can be seen as a middle ground between sparsity-based methods (e.g., Belloni *et al.* (2012)) and factor structure-based methods such as Bai & Ng (2010) and our method.

4.4.1 Design

We allow half of the instruments (from $i = 1$ to $N/2$) to share the factor structure given in equation-4.8. The remaining half of the instruments are sparse, as defined in equations 4.9 and 4.10. The sparse structure implies that some instruments are related to x_t, while many may just represent noise. An exogenous term $e_{2t} \sim \mathcal{N}(0, 1)$ is introduced to make certain sparse instruments related to x_t. Equation 4.9 ensures that $N/8$ instruments are related to x_t through e_t, while the remaining $3N/8$ instruments are noise.

The data-generating process is:

$$y_t = \beta_0 + \beta_1 x_t + \varepsilon_t \tag{4.6}$$

$$x_t = \boldsymbol{\phi}' \mathbf{f}_t + e_{2t} + e_t \tag{4.7}$$

$$z_{it} = \mathbf{b}_i' \mathbf{f}_t + \sigma_z u_{it}, \quad 1 \leq i \leq N/2, \quad 1 \leq t \leq T \tag{4.8}$$

$$z_{it} = \sigma_z (e_{2t} + u_{it}), \quad N/2 + 1 \leq i \leq 5N/8, \quad 1 \leq t \leq T \tag{4.9}$$

$$z_{it} = \sigma_z u_{it}, \quad 5N/8 + 1 \leq i \leq N, \quad 1 \leq t \leq T \tag{4.10}$$

$$f_{jt} = \gamma_j f_{jt-1} + v_{jt}, \quad 1 \leq j \leq r \tag{4.11}$$

We set $\beta_0 = 0$ and the main parameter of interest $\beta_1 = 2$. We let $r = 3$ and $\boldsymbol{\phi} = (0.8, 0.5, 0.3)'$, which means that all factors are relevant for x_t and influence the instruments that share factor structure. We keep $m(\cdot)$ to be linear to keep the design a level-playing field for three competing methods: Belloni *et al.* (2012), Bai & Ng (2010), and ours. The rest of the details are the same as the Design-I discussed in section-4.1.

4.4.2 Results

We denote Belloni *et al.* (2012)'s method as PLIV i.e. Post-Lasso IV in the tables. The results are presented in Table-4

The secular observation from the table-4 is that the performance of three competing methods is comparable. Therefore, when the researcher suspects that both factor structure and sparse structure may be present in the instrument set, employing both Belloni *et al.* (2012) or our approach is likely to yield the same outcome.

4.5 More Designs for Robustness

We have demonstrated so far that our method performs better due to three major strengths: being a supervised method, achieving considerable dimension reduction, and its ability to handle non-linearities. One may be curious to know how our method performs when all these challenges are present simultaneously. 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 section 4.1.

The results are reported in Table 19 in Appendix B.2. The results are qualitatively the same as those in Table 3. We observe a slightly higher RMSE for the competing method FIV,

			$E(\widehat{\beta}_1)$			RMSE$(\widehat{\beta}_1)$		
ρ	T	N	SIF	FIV	PLIV	SIF	FIV	PLIV
0	200	50	2.10	2.06	2.08	0.12	0.11	0.11
0	200	150	2.10	2.06	2.09	0.13	0.11	0.12
0	200	250	2.11	2.06	2.09	0.13	0.11	0.12
0.5	200	50	2.52	2.46	2.50	0.54	0.48	0.51
0.5	200	150	2.52	2.37	2.46	0.53	0.38	0.47
0.5	200	250	2.52	2.33	2.45	0.53	0.35	0.46
0.9	200	50	2.91	2.83	2.89	0.91	0.84	0.90
0.9	200	150	2.92	2.67	2.84	0.93	0.68	0.84
0.9	200	250	2.92	2.61	2.83	0.93	0.62	0.83

Table 4: Simulations using Design-4 and AR Errors
[Notes: True value of β_1 is 2. ρ represents extent of endogeneity, $\rho = 0$ means no-endogeneity. Errors and factors are AR(1) with AR(1) coefficient being 0.5. Other Parameters: $r = 3$, number of replications = 500]

which may be due to selecting irrelevant factors. We also consider different combinations of uncorrelated errors and factors, and the results are reported in Appendix B.2. The qualitative results remain consistent.

We further replicate the simulation exercise for linear and non-linear $m(\cdot)$ with eight true factors ($r = 8$), where only three are relevant for the endogenous regressor x_t. All results are tabulated in Appendix B.2.

After analyzing more than 30 tables of various simulation specifications, we conclude that our method decisively filters out the relevant factors and combines them in a manner that is relevant to the endogenous regressor. 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) (hereafter, BLP) 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 BLP's empirical application on automobile demand using Chernozhukov *et al.* (2015)'s data.

We conducted a 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. Moreover, the first two PCs are strongly related to the endogenous regressor, *price*. This suggests that there could be a factor structure in the instrument set that may drive the endogenous regressor (price), implying that our method can be employed. Philosophically, the decision-making process of an automobile buyer might arise from a low-rank latent factor structure, which is reflected in many instruments considered in the BLP problem.

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:

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

$$p_{it} = z_{it}'\delta_0 + x_{it}'\gamma_0 + u_{it} \tag{5.2}$$

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 this 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[5]. 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 $\widehat{\tilde{p}}_j$ in the equation-5.1 to estimate the parameter of interest α_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. 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.[6] 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]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

[6]We need Assumption-1.1 i.e. $E\Big[m(\boldsymbol{\theta_1}'\mathbf{f}_j, \ldots, \boldsymbol{\theta_L}'\mathbf{f}_j)x_j\Big] \neq 0$ for relevancy. However, if factors are correlated with x_t, it means that for $m(\cdot)$ linear, $\boldsymbol{\theta}_1$ being vector of ones and other $\boldsymbol{\theta}_j$ being vectors of zeros, our relevancy holds. It may also satisfy through more functional forms but we at least guarantee one.

5.1.3 Results

We estimate the own-price elasticity of automobile demand, given by the parameter α_0 of equation-5.1. We draw random samples with replacement multiple times (bootstrap, 500 times) and estimate α_0. We report the bootstrapped confidence intervals for our method 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 for $\widehat{\alpha}_0$ estimated by our method are $[-0.161, -0.125]$ with a mean of -0.152. Further, if the bias survives, it implies that our estimator is closer to Chernozhukov *et al.* (2015)'s estimate of -0.185, because the bias is positive. The coefficient estimated in the original paper Berry *et al.* (1995) was -0.142. Our stable results over many sample-splitting procedures give confidence in the estimates. Therefore, we can argue that, similar to Berry *et al.* (1995) and Chernozhukov *et al.* (2015), our method also captures the important features of the data. This exercise serves as 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® 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 tradable security. We therefore, use its tradable counterpart, the iShares S&P Small-Cap 600 Value ETF[7]. 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® combines three leading indices, the S&P 500®, the S&P MidCap 400®, and the S&P SmallCap 600®, which covers approximately 90% of U.S. market capitalization. The S&P SmallCap 600® is a float-adjusted market cap weighted index of the smallest 600 companies

[7]This ETF, launched in July 2000, is traded under the symbol IJS on major trading platforms.

in the S&P 1500®. 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.

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:

$$y_t = \alpha + \beta R_t^* + \eta_t \tag{5.3}$$

where R_t^* is the return on the market portfolio, which is unobservable. We are interested in the parameter β, 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 Dow Jones Industrial Average (DJIA) index. We can write the observed proxy R_t as:

$$R_t = R_t^* + e_t \tag{5.4}$$

where e_t is the error term that satisfies the assumptions outlined in the theoretical framework. 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 $\varepsilon_t = e_t - \beta\eta_t$ is correlated with R_t; this makes the 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 for instruments that have been a part of the DJIA at any time during our analysis period to avoid any trivial analysis. Therefore, there is no overlap among the 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&P 500 firm (instrument variable) follow the approximate factor model (Chamberlain & Rothschild (1983)) structure, where the factors are strong in the sense that they follow assumptions 1-3. This can be written as the following equation:

$$z_{it} = \Lambda_i \mathbf{f}_t + u_{it} \tag{5.5}$$

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 represents 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 for a more general setting, we treat the factors as unobservable.

In this setting, R_t^* can be viewed as a sufficient index of the unobserved factor variables ($\mathbf{f}_t$) for R_t; these unobserved factors drive both the instruments $\mathbf{z}_i \quad i = 1, 2, .., N$, 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 their findings.

5.2.2 Data

For daily returns on the S&P 500 companies' data, we use the Center for Research in Security Prices (CRSP) 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, the IJS ETF. We then calculate the returns based on the closing price of the trading day. Table 5 contains the list of the companies that have been a part of the Dow Jones Industrial Average (DJIA) at any point between 2001 and 2023. We remove these companies from the set of instruments to ensure that the R_t measured by the DJIA is not a trivial combination of companies in the instrument. Our data runs from 2001 to 2023. We use the method from Hamilton & Xi (2024) to transform the data into a stationary series.

5.2.3 Validity of Instruments

Relevancy The required relevancy condition is $E\Big[m(\boldsymbol{\theta}_1'\mathbf{f}_t, \ldots, \boldsymbol{\theta}_L'\mathbf{f}_t)R_t\Big] \neq 0$. Since factors are not observable, therefore, we use their consistent estimates, this translates the relevancy condition to $E\Big[\widehat{m}(\widehat{\boldsymbol{\theta}}_1'\widehat{\mathbf{f}}_t, \ldots, \widehat{\boldsymbol{\theta}}_L'\widehat{\mathbf{f}}_t)R_t\Big] \neq 0$. This condition is weaker than otherwise required $E(z_{it}R_t) \neq 0$ for all $i = 1, 2, \ldots, N$ in 2SLS method and $E(f_{jt}R_t) \neq 0$ for all $j = 1, 2, \ldots, r$ 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. Further, y_t i.e. IJS is unlikely to drive our instruments because the aggregate market capital of small firms in IJS is less than 4% of the stock market capital.

5.2.4 Results

We found the value of CAPM *beta* of IJS 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. The OLS estimate is about 1.05, much lower than expected.

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. Developing a semi-parametric efficiency theory for this five-layered procedure is left for future research.

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*, 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.

Bühlmann, Peter, & Van De Geer, Sara. 2011. *Statistics for high-dimensional data: methods, theory and applications*. Springer Science & Business Media.

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.

Carrasco, Marine. 2012. A regularization approach to the many instruments problem. *Journal of Econometrics*, **170**(2), 383–398.

Chamberlain, Gary, & Rothschild, Michael. 1983. Arbitrage, Factor Structure, and Mean-Variance Analysis on Large Asset Markets. *Econometrica*, 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.

Hall, Peter, Horowitz, Joel L., & Jing, Bing-Yi. 1995. On blocking rules for the bootstrap with dependent data. *Biometrika*, **82**(3), 561–574.

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.

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*, 809–837.

Newey, Whitney K. 1994. The asymptotic variance of semiparametric estimators. *Econometrica: Journal of the Econometric Society*, 1349–1382.

Newey, Whitney K, & McFadden, Daniel. 1994. Large sample estimation and hypothesis testing. *Handbook of econometrics*, **4**, 2111–2245.

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, & Stock, James H. 1997. Instrumental Variables Regression with Weak Instruments. *Econometrica*, **65**(3), 557–586.

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 $\hat{m}(\boldsymbol{w})$. 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 $\boldsymbol{P}_{tT} = e_t \mathcal{K}_h(\boldsymbol{w}_t - \boldsymbol{w})$ that need to be taken care of. The variance of $\hat{m}(\boldsymbol{w})$ is given by:

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

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

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

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

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

Since e_t is assumed to have finite variance and the kernel $\mathcal{K}$ is bounded and integrates to 1, $e_t^2 \mathcal{K}_h^2(\boldsymbol{w}_t - \boldsymbol{w})$ 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 $\hat{m}(\boldsymbol{w})$ is defined as the solution to the weighted least squares problem:

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

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

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

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

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

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

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

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

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

Variance:

The variance of the estimator is given by:

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

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

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

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

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

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

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

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

This completes the proof. □

A.1.2 Proof of Theorem-1

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

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

Where $\widehat{X}$ is the predicted value from instruments which is a function of factors. We used hats on both m and $\boldsymbol{w}$ 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 $\widehat{\boldsymbol{\beta}}$.

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

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

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

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

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

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

that $\frac{\partial m(w_{t1},\cdots,w_{tL})}{\partial w_j}$ and $\frac{\partial^2 m(w_{t1},\cdots,w_{tL})}{\partial w_j \partial w_k}$ are $O(1)$. Therefore, we can write $\widehat{m}(w_{t1},\cdots,w_{tL})$ as:

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

Using Corollary-1, we know that $\widehat{w}_j = \widehat{\boldsymbol{\theta}}_j'\widehat{\mathbf{f}}_t \to \boldsymbol{\theta}_j'\mathbf{f}_t = w_j$ for all $j = 1,\ldots,L$ 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 $j = 1,\ldots,L$. Therefore,

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

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:

$$\widehat{m}(\widehat{w}_{t1},\cdots,\widehat{w}_{tL}) = m(w_{t1},\cdots,w_{tL}) + O_p(h^2) + O_p\Big(\frac{1}{Th^L}\Big) + O_p(\omega_{NT})$$

Lemma-4 says that the Bias of $\widehat{m}(\cdot)$ 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, $O_p\Big(\frac{1}{Th^L}\Big) = O_p\Big(\frac{1}{T(T^{\frac{-1}{L+4}})^L}\Big) = O_p\Big(T^{\frac{-4}{L+4}}\Big)$ 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 $O_p(\delta_{NT}^{-1})$. Hence,

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

□

Let's get back to the expression of $\widehat{\boldsymbol{\beta}}$, we have $\widehat{\boldsymbol{\beta}} = (\widehat{X}'\widehat{X})^{-1}\widehat{X}'y$ where $\widehat{X} = [1 \quad \widehat{m}(\widehat{\boldsymbol{w}})]$. We want to expand $\widehat{X}'\widehat{X}$ and $\widehat{X}'y$ in terms of Q. Matrices with true values are:

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

and

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

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

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

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

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

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

$$\begin{aligned} \widehat{m}(\widehat{\boldsymbol{w}}_t)^2 - m(\boldsymbol{w}_t)^2 &= \left(O_p(h^2) + O_p(\omega_{NT})\right)\left(\widehat{m}(\widehat{\boldsymbol{w}}_t) + m(\boldsymbol{w}_t)\right) \\ &= \Big(O_p(h^2) + O_p(\omega_{NT})\Big)\Big(O(1)\Big) \\ &= O_p(\delta_{NT}^{-1}) \end{aligned}$$

Therefore,

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

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

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

where:

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

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:

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

Note that $y_t = \beta_0 + \beta_1 x_t + \varepsilon = [1 \quad x_t]\boldsymbol{\beta} + \varepsilon_t = [1 \quad m(\boldsymbol{w}_t) + e_t]\boldsymbol{\beta} + \varepsilon_t = [1 \quad m(\boldsymbol{w}_t)]\boldsymbol{\beta} + e_t\beta_1 + \varepsilon_t$. Therefore, stacking t subscript terms, we can write $y = Q\boldsymbol{\beta} + e\beta_1 + \varepsilon$:

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

Therefore,

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

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 :

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

The term $\sum_{t=1}^{T} m(\boldsymbol{w}_t)$ is $O(T)$ because $m(\boldsymbol{w}_t)$ is $O(1)$ (Assumption-5.1 and Assumption-5.5). Therefore, the $Q'Q$ matrix is of order $O(T)$. Similarly,

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

Note that β_1 is non-stochastic and is constant, therefore, can be replaced by $O(1)$. From Assumption-1.6, 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 1.6, $\sum_{t=1}^{T} m(\boldsymbol{w}_t)e_t = O_p(\sqrt{T})$. 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,

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

From Assumption-1.5, ε_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, $\frac{1}{T}\sum_{t=1}^{T} m(\boldsymbol{w}_t)\varepsilon_t = O_p(\sqrt{T})$. Therefore,

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

□

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

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

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

We get the second line by substituting $x_t = m(\boldsymbol{w}_t) + e_t$ (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,

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

□

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

Proof. We have already shown,

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

The terms in Δ are $\sum_{t=1}^{T} O_p(\delta_{NT}^{-1}) = O_p(T\delta_{NT}^{-1}) = O_p(T(N^{-1/2} + T^{-2/(L+4)}))$
$= O_p(N^{-1/2}T + T^{(L+2)/(L+4)})$. Since $N > 0$, therefore the $N^{-1/2}T$-order term will be smaller than the terms of order T. Hence the terms in Δ 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 $(Q'Q)^{-1}\,\Delta\,(Q'Q+\Delta)^{-1}\,Q'Q\boldsymbol{\beta}$ to $(Q'Q)^{-1}\,\Delta O_p(1)\boldsymbol{\beta}$. Since $\boldsymbol{\beta}$ is

a constant with respect to time, this term's order is decided by $(Q'Q)^{-1}\Delta$. Therefore,

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

□

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

Proof. In the previous claims, we have shown $Q'Q$ is of order $O(T)$, Δ is of order $O_p(T\delta_{NT}^{-1})$, $(Q'Q+\Delta)^{-1}$ is of order $O_p(T^{-1})$, and $Q'e$ is of order $O_p(\sqrt{T})$.

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

□

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

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

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

□

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

Proof. In the previous claims, we have shown that Δ and Γ are of order $O_p(T\delta_{NT}^{-1})$. Therefore, we can write,

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

□

After learning the behavior of all terms involved in the expression of $\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}$ (equation-A.1), using Claim-2 to Claim-8, we can write,

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

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 $O_p(\delta_{NT}^{-1}) = O_p(N^{\frac{-1}{2}} + T^{\frac{-2}{5}})$ which goes to zero slower than the first and second terms which are $O_p(T^{\frac{-1}{2}})$. Therefore, asymptotically ($N \to \infty$ and $T \to \infty$), 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, with L fixed, $N \to \infty$ and $T \to \infty$, we can write:

$$\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta} = O_p(\delta_{NT}^{-1})$$

Since $\delta_{NT}^{-1} = N^{-1/2} + T^{\frac{-2}{(L+4)}}$, therefore as $N \to \infty$ and $T \to \infty$, δ_{NT}^{-1} will go to zero. Therefore $\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}$ will go to zero, hence $\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta} = o_p(1)$. Therefore, $\widehat{\boldsymbol{\beta}}$ is a consistent estimator of $\boldsymbol{\beta}$. □

A.1.3 Proof of Theorem-2

Proof. We have discussed in the proof of theorem-1, that the the expression of $(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta})$ 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,

$$\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta} = \left(Q'Q\right)^{-1}\Gamma - \left(Q'Q\right)^{-1}\Delta\left(Q'Q + \Delta\right)^{-1}Q'Q\boldsymbol{\beta} + o_p(\delta_{NT}^{-1})$$

Let's analyze $\left(Q'Q + \Delta\right)^{-1}Q'Q$ term:

$$\begin{aligned}\left(Q'Q + \Delta\right)^{-1}Q'Q =& \Big[(Q'Q)^{-1} - (Q'Q)^{-1}\Delta\left(Q'Q + \Delta\right)^{-1}\Big]Q'Q \\ =& I - \Big[(Q'Q)^{-1}\Delta\left(Q'Q + \Delta\right)^{-1}Q'Q\Big]\end{aligned}$$

The second term in the expression above is of smaller order, therefore, we can approximate $(Q'Q + \Delta)^{-1} Q'Q$ by an identity matrix, therefore, we can write:

$$\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta} = \left(Q'Q\right)^{-1} \Big(\Gamma - \Delta\boldsymbol{\beta}\Big) + o_p(\delta_{NT}^{-1})$$

Let's solve for $\Gamma - \Delta\boldsymbol{\beta}$. The matrix Γ is:

$$\begin{aligned}
\Gamma &= \begin{bmatrix} 0 \\ \sum_{t=1}^{T} (\widehat{m}(\widehat{\boldsymbol{w}}_t) - m(\boldsymbol{w}_t))(\beta_0 + x_t\beta_1 + \varepsilon_t) \end{bmatrix} \\
&= \begin{bmatrix} 0 \\ \sum_{t=1}^{T} (\widehat{m}(\widehat{\boldsymbol{w}}_t) - m(\boldsymbol{w}_t))(\beta_0 + (m(\boldsymbol{w}_t) + e_t)\beta_1 + \varepsilon_t) \end{bmatrix}
\end{aligned}$$

and

$$\begin{aligned}
\Delta\boldsymbol{\beta} &= \begin{bmatrix} 0 & \sum_{t=1}^{T}(\widehat{m}(\widehat{\boldsymbol{w}}_t) - m(\boldsymbol{w}_t)) \\ \sum_{t=1}^{T}(\widehat{m}(\widehat{\boldsymbol{w}}_t) - m(\boldsymbol{w}_t)) & \sum_{t=1}^{T}(\widehat{m}(\widehat{\boldsymbol{w}}_t)^2 - m(\boldsymbol{w}_t)^2) \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix} \\
&= \begin{bmatrix} \sum_{t=1}^{T}(\widehat{m}(\widehat{\boldsymbol{w}}_t) - m(\boldsymbol{w}_t))\beta_1 \\ \sum_{t=1}^{T}(\widehat{m}(\widehat{\boldsymbol{w}}_t) - m(\boldsymbol{w}_t))\beta_0 + \sum_{t=1}^{T}(\widehat{m}(\widehat{\boldsymbol{w}}_t)^2 - m(\boldsymbol{w}_t)^2)\beta_1 \end{bmatrix}
\end{aligned}$$

Therefore,

$$\begin{aligned}
\Gamma - \Delta\boldsymbol{\beta} &= \begin{bmatrix} -\sum_{t=1}^{T}(\widehat{m}(\widehat{\boldsymbol{w}}_t) - m(\boldsymbol{w}_t))\beta_1 \\ \sum_{t=1}^{T}(\widehat{m}(\widehat{\boldsymbol{w}}_t) - m(\boldsymbol{w}_t))\big((m(\boldsymbol{w}_t) + e_t)\beta_1 + \varepsilon_t\big) - \sum_{t=1}^{T}(\widehat{m}(\widehat{\boldsymbol{w}}_t)^2 - m(\boldsymbol{w}_t)^2)\beta_1 \end{bmatrix} \\
&= \begin{bmatrix} -\sum_{t=1}^{T}(\widehat{m}(\widehat{\boldsymbol{w}}_t) - m(\boldsymbol{w}_t))\beta_1 \\ \sum_{t=1}^{T} (\widehat{m}(\widehat{\boldsymbol{w}}_t) - m(\boldsymbol{w}_t)) [2m(\boldsymbol{w}_t) - \widehat{m}(\widehat{\boldsymbol{w}}_t)] \beta_1 + \sum_{t=1}^{T} (\widehat{m}(\widehat{\boldsymbol{w}}_t) - m(\boldsymbol{w}_t)) \left(e_t\beta_1 + \varepsilon_t\right) \end{bmatrix}
\end{aligned}$$

Usually, the rate of the non-parametric estimator is bias-square and variance is $\sqrt{Th^L}$, but we have generated regressors instead of regular variables, therefore the rate $N^{1/2}$ will also appear, as shown in Claim-1. Hence, the rate of $(\widehat{m}(\widehat{\boldsymbol{w}}_t) - m(\boldsymbol{w}_t))$ is δ_{NT}^{-1}.

We can divide the expression $(\widehat{m}(\widehat{\boldsymbol{w}}_t) - m(\boldsymbol{w}_t))$ into bias and variance parts using Lemma-4:

$$\widehat{m}(\widehat{\boldsymbol{w}}) - m(\boldsymbol{w}) = \frac{1}{2}\delta_{NT}^{-1}\mathrm{tr}(\boldsymbol{G}) + \frac{1}{\delta_{NT}}\sum_{t=1}^{T} e_t\mathcal{K}_h(\boldsymbol{w}_t - \boldsymbol{w})$$

Let's divide matrix $\Gamma - \Delta\boldsymbol{\beta}$ into two parts containing bias and variance. Rewrite the matrix by substituting $2m(\boldsymbol{w}_t) - \widehat{m}(\widehat{\boldsymbol{w}}_t)$:

$$2m(\boldsymbol{w}_t) - \widehat{m}(\widehat{\boldsymbol{w}}_t) = m(\boldsymbol{w}_t) - \frac{1}{2}\delta_{NT}^{-1}\mathrm{tr}(\boldsymbol{G}) - \frac{1}{\delta_{NT}}\sum_{t=1}^{T} e_t \mathcal{K}_h(\boldsymbol{w}_t - \boldsymbol{w})$$

Bias Part (without e_t and ε_t):

$$\begin{bmatrix} -\sum_{t=1}^{T}\left(\frac{1}{2}\delta_{NT}^{-1}\mathrm{tr}(\boldsymbol{G})\right)\beta_1 \\ \sum_{t=1}^{T}\left(\frac{1}{2}\delta_{NT}^{-1}\mathrm{tr}(\boldsymbol{G})\right)\left[m(\boldsymbol{w}_t) - \frac{1}{2}\delta_{NT}^{-1}\mathrm{tr}(\boldsymbol{G})\right]\beta_1 \end{bmatrix}$$

Variance Part (with e_t and ε_t):

$$\begin{bmatrix} -\sum_{t=1}^{T}\left(\frac{1}{\delta_{NT}}\sum_{t=1}^{T} e_t\,\mathcal{K}_h(\boldsymbol{w}_t - \boldsymbol{w})\right)\beta_1 \\ \sum_{t=1}^{T}\left(\frac{1}{\delta_{NT}}\sum_{t=1}^{T} e_t\,\mathcal{K}_h(\boldsymbol{w}_t - \boldsymbol{w})\right)\left[m(\boldsymbol{w}_t) - \frac{1}{2\delta_{NT}}\mathrm{tr}(\boldsymbol{G})\right]\beta_1 \\ +\sum_{t=1}^{T}\left(\frac{1}{\delta_{NT}}\sum_{t=1}^{T} e_t\,\mathcal{K}_h(\boldsymbol{w}_t - \boldsymbol{w})\right)(e_t\beta_1 + \varepsilon_t) \end{bmatrix}_{2\times 1}$$

$$\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta} = \left(Q'Q\right)^{-1}\begin{bmatrix} -\sum_{t=1}^{T}\left(\frac{1}{2}\delta_{NT}^{-1}\mathrm{tr}(\boldsymbol{G})\right)\beta_1 \\ \sum_{t=1}^{T}\left(\frac{1}{2}\delta_{NT}^{-1}\mathrm{tr}(\boldsymbol{G})\right)\left[m(\boldsymbol{w}_t) - \frac{1}{2}\delta_{NT}^{-1}\mathrm{tr}(\boldsymbol{G})\right]\beta_1 \end{bmatrix}$$

$$+(Q'Q)^{-1}\begin{bmatrix} -\sum_{t=1}^{T}\left(\frac{1}{\delta_{NT}}\sum_{t=1}^{T} e_t\,\mathcal{K}_h(\boldsymbol{w}_t - \boldsymbol{w})\right)\beta_1 \\ \sum_{t=1}^{T}\left(\frac{1}{\delta_{NT}}\sum_{t=1}^{T} e_t\,\mathcal{K}_h(\boldsymbol{w}_t - \boldsymbol{w})\right)\left[m(\boldsymbol{w}_t) - \frac{1}{2\delta_{NT}}\mathrm{tr}(\boldsymbol{G})\right]\beta_1 \\ +\sum_{t=1}^{T}\left(\frac{1}{\delta_{NT}}\sum_{t=1}^{T} e_t\,\mathcal{K}_h(\boldsymbol{w}_t - \boldsymbol{w})\right)(e_t\beta_1 + \varepsilon_t) \end{bmatrix}_{2\times 1}$$

This implies:

$$\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta} = \left(\frac{Q'Q}{T}\right)^{-1}\begin{bmatrix} -\frac{1}{T}\sum_{t=1}^{T}\left(\frac{1}{2}\delta_{NT}^{-1}\mathrm{tr}(\boldsymbol{G})\right)\beta_1 \\ \frac{1}{T}\sum_{t=1}^{T}\left(\frac{1}{2}\delta_{NT}^{-1}\mathrm{tr}(\boldsymbol{G})\right)\left[m(\boldsymbol{w}_t) - \frac{1}{2}\delta_{NT}^{-1}\mathrm{tr}(\boldsymbol{G})\right]\beta_1 \end{bmatrix}$$

$$+\left(\frac{Q'Q}{T}\right)^{-1}\begin{bmatrix} -\frac{1}{T}\sum_{t=1}^{T}\left(\frac{1}{\delta_{NT}}\sum_{t=1}^{T} e_t\,\mathcal{K}_h(\boldsymbol{w}_t - \boldsymbol{w})\right)\beta_1 \\ \sum_{t=1}^{T}\left(\frac{1}{T\delta_{NT}}\sum_{t=1}^{T} e_t\,\mathcal{K}_h(\boldsymbol{w}_t - \boldsymbol{w})\right)\left[m(\boldsymbol{w}_t) - \frac{1}{2\delta_{NT}}\mathrm{tr}(\boldsymbol{G})\right]\beta_1 \\ +\sum_{t=1}^{T}\left(\frac{1}{T\delta_{NT}}\sum_{t=1}^{T} e_t\,\mathcal{K}_h(\boldsymbol{w}_t - \boldsymbol{w})\right)(e_t\beta_1 + \varepsilon_t) \end{bmatrix}_{2\times 1}$$

Let's call $\mathscr{B} = \left(\frac{Q'Q}{T}\right)^{-1} \begin{bmatrix} -\frac{1}{T}\sum_{t=1}^{T}\left(\frac{1}{2}\delta_{NT}^{-1}\text{tr}(\boldsymbol{G})\right)\beta_1 \\ \frac{1}{T}\sum_{t=1}^{T}\left(\frac{1}{2}\delta_{NT}^{-1}\text{tr}(\boldsymbol{G})\right)\left[m(\boldsymbol{w}_t) - \frac{1}{2}\delta_{NT}^{-1}\text{tr}(\boldsymbol{G})\right]\beta_1 \end{bmatrix}$, the bias term. Then, We can write:

$$\delta_{NT}\left(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta} - \mathscr{B}\right) = \left(\frac{Q'Q}{T}\right)^{-1} \begin{bmatrix} -\frac{1}{T}\sum_{t=1}^{T}\left(\sum_{t=1}^{T} e_t\,\mathcal{K}_h(\boldsymbol{w}_t - \boldsymbol{w})\right)\beta_1 \\ \sum_{t=1}^{T}\left(\frac{1}{T}\sum_{t=1}^{T} e_t\,\mathcal{K}_h(\boldsymbol{w}_t - \boldsymbol{w})\right)\left[m(\boldsymbol{w}_t) - \frac{1}{2\delta_{NT}}\text{tr}(\boldsymbol{G})\right]\beta_1 \\ +\sum_{t=1}^{T}\left(\frac{1}{T}\sum_{t=1}^{T} e_t\,\mathcal{K}_h(\boldsymbol{w}_t - \boldsymbol{w})\right)(e_t\beta_1 + \varepsilon_t) \end{bmatrix}_{2\times 1}$$

As $\delta_{NT} \to \infty$ i.e. both $N \to \infty$ and $T \to \infty$, we can ignore $\frac{1}{2}\delta_{NT}^{-1}\text{tr}(\boldsymbol{G})$ term in the second row because it goes to zero compared to $\boldsymbol{w}_t$. Therefore, we can write:

$$\delta_{NT}\left(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta} - \mathscr{B}\right) = \left(\frac{Q'Q}{T}\right)^{-1} \begin{bmatrix} -\frac{1}{T}\sum_{t=1}^{T}\left(\sum_{t=1}^{T} e_t\mathcal{K}_h(\boldsymbol{w}_t - \boldsymbol{w})\right)\beta_1 \\ \sum_{t=1}^{T}\left(\frac{\sum_{t=1}^{T} e_t\mathcal{K}_h(\boldsymbol{w}_t - \boldsymbol{w})}{T}\right) m(\boldsymbol{w}_t)\beta_1 + \sum_{t=1}^{T}\left(\frac{\sum_{t=1}^{T} e_t\mathcal{K}_h(\boldsymbol{w}_t - \boldsymbol{w})}{T}\right)(e_t\beta_1 + \varepsilon_t) \end{bmatrix}$$

Let's derive the expression of variance of $\delta_{NT}\left(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta} - \mathscr{B}\right)$.

$$Var\left(\delta_{NT}(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta} - \mathscr{B}))\right) = \left(\frac{Q'Q}{T}\right)^{-1} \boldsymbol{V} \left(\frac{Q'Q}{T}\right)^{-1}$$

Where $\boldsymbol{V}$ is a 2x2 matrix. The processes for ε_t and e_t are defined as follows:

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

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

where η_t and ζ_t are white noise terms with variances σ_η^2 and σ_ζ^2, respectively. The correlation between these noise terms is given by: $\rho = \text{cor}(\eta_t, \zeta_t)$ The variance of ε_t can be expressed as:

$$\text{Var}(\varepsilon_t) = \frac{\sigma_\eta^2}{1-\alpha_1^2} \qquad \text{Var}(e_t) = \frac{\sigma_\zeta^2}{1-\alpha_2^2}.$$

The covariance can be derived using the properties of AR processes and the correlation ρ:

$$\text{Cov}(\varepsilon_t, e_t) = \rho\sqrt{\text{Var}(\varepsilon_t)\text{Var}(e_t)} = \rho\sqrt{\frac{\sigma_\eta^2}{1-\alpha_1^2} \cdot \frac{\sigma_\zeta^2}{1-\alpha_2^2}}.$$

Let's represent the elements of the 2×2 matrix $\boldsymbol{V}$ as:

$$\boldsymbol{V} = \begin{bmatrix} V_{11} & V_{12} \\ V_{21} & V_{22} \end{bmatrix}.$$

$\underline{\mathbf{V_{11}}}$:

$$V_{11} = \left(\frac{1}{T}\sum_{s=1}^{T} \mathcal{K}_h(\boldsymbol{w}_s - \boldsymbol{w})\right)^2 \cdot \text{Var}(e_t) \cdot \beta_1^2 = \left(\frac{1}{T}\sum_{s=1}^{T} \mathcal{K}_h(\boldsymbol{w}_s - \boldsymbol{w})\right)^2 \sigma_e^2 \beta_1^2$$

$\underline{\mathbf{V_{22}}}$:

V_{22} is the variance of the following term:

$$\sum_{t=1}^{T}\left(\frac{1}{T}\sum_{s=1}^{T} e_s \mathcal{K}_h(\boldsymbol{w}_s - \boldsymbol{w})\right) m(\boldsymbol{w}_t)\beta_1 + \sum_{t=1}^{T}\left(\frac{1}{T}\sum_{s=1}^{T} e_s \mathcal{K}_h(\boldsymbol{w}_s - \boldsymbol{w})\right)(e_t\beta_1 + \varepsilon_t)$$

Let's name the first term V_{221} and second term V_{222}. Then

$$V_{22} = E(V_{221}^2) + 2E(V_{221}V_{222}) + E(V_{222}^2)$$

Denote $Cov(e_t, e_s) = \kappa_e(t-s)$ and $Cov(\varepsilon_t, \varepsilon_s) = \kappa_\varepsilon(t-s)$, and $Cov(\varepsilon_t, e_s) = \kappa_{\varepsilon e}(t-s)$

$$\begin{aligned}
E(V_{221}^2) &= E\left[\left(\sum_{t=1}^{T}\left(\frac{1}{T}\sum_{s=1}^{T} e_s \mathcal{K}_h(\boldsymbol{w}_s - \boldsymbol{w})\right) m(\boldsymbol{w}_t)\beta_1\right)^2\right] \\
&= E\Bigg[\left(\sum_{t=1}^{T}\left(\frac{1}{T}\sum_{s=1}^{T} e_s \mathcal{K}_h(\boldsymbol{w}_s - \boldsymbol{w})\right) m(\boldsymbol{w}_t)\beta_1\right) \times \\
&\qquad\qquad \left(\sum_{t'=1}^{T}\left(\frac{1}{T}\sum_{s'=1}^{T} e_{s'} \mathcal{K}_h(\boldsymbol{w}_{s'} - \boldsymbol{w})\right) m(\boldsymbol{w}_{t'})\beta_1\right)\Bigg] \\
&= \beta_1^2 E\left[\sum_{t=1}^{T}\sum_{t'=1}^{T}\left(\frac{1}{T}\sum_{s=1}^{T} e_s \mathcal{K}_h(\boldsymbol{w}_s - \boldsymbol{w})\right) m(\boldsymbol{w}_t) \cdot \left(\frac{1}{T}\sum_{s'=1}^{T} e_{s'} \mathcal{K}_h(\boldsymbol{w}_{s'} - \boldsymbol{w})\right) m(\boldsymbol{w}_{t'})\right] \\
&= \frac{\beta_1^2}{T^2} E\left[\sum_{t=1}^{T}\sum_{t'=1}^{T}\sum_{s=1}^{T}\sum_{s'=1}^{T} e_s e_{s'} \mathcal{K}_h(\boldsymbol{w}_s - \boldsymbol{w})\mathcal{K}_h(\boldsymbol{w}_{s'} - \boldsymbol{w}) m(\boldsymbol{w}_t) m(\boldsymbol{w}_{t'})\right] \\
&= \frac{\beta_1^2}{T^2} \sum_{t=1}^{T}\sum_{t'=1}^{T}\sum_{s=1}^{T}\sum_{s'=1}^{T} \kappa_e(s-s') \mathcal{K}_h(\boldsymbol{w}_s - \boldsymbol{w})\mathcal{K}_h(\boldsymbol{w}_{s'} - \boldsymbol{w}) m(\boldsymbol{w}_t) m(\boldsymbol{w}_{t'})
\end{aligned}$$

$$
\begin{aligned}
E(V_{222}^2) &= E\left[\left(\sum_{t=1}^{T}\left(\frac{1}{T}\sum_{s=1}^{T} e_s \mathcal{K}_h(\boldsymbol{w}_s - \boldsymbol{w})\right)(e_t\beta_1 + \varepsilon_t)\right)^2\right] \\
&= \frac{1}{T^2} E\Bigg[\sum_{t=1}^{T}\sum_{t'=1}^{T}\sum_{s=1}^{T}\sum_{s'=1}^{T} e_s e_{s'} \mathcal{K}_h(\boldsymbol{w}_s - \boldsymbol{w})\mathcal{K}_h(\boldsymbol{w}_{s'} - \boldsymbol{w}) \\
&\qquad\qquad \left(e_t e_{t'}\beta_1^2 + e_t\varepsilon_{t'}\beta_1 + \varepsilon_t e_{t'}\beta_1 + \varepsilon_t\varepsilon_{t'}\right)\Bigg] \\
&= \frac{1}{T^2}\sum_{t=1}^{T}\sum_{t'=1}^{T}\sum_{s=1}^{T}\sum_{s'=1}^{T} \kappa_e(s-s')\mathcal{K}_h(\boldsymbol{w}_s - \boldsymbol{w})\mathcal{K}_h(\boldsymbol{w}_{s'} - \boldsymbol{w}) \\
&\quad \cdot \left(\kappa_e(t-t')\beta_1^2 + \kappa_{\varepsilon e}(t-t')\beta_1\sigma_\varepsilon + \kappa_\varepsilon(t-t')\right)
\end{aligned}
$$

$$
\begin{aligned}
E(V_{221}V_{222}) &= E\Bigg[\left(\sum_{t=1}^{T}\left(\frac{1}{T}\sum_{s=1}^{T} e_s \mathcal{K}_h(\boldsymbol{w}_s - \boldsymbol{w})\right) m(\boldsymbol{w}_t)\beta_1\right) \times \\
&\qquad\qquad \left(\sum_{t=1}^{T}\left(\frac{1}{T}\sum_{s'=1}^{T} e_{s'} \mathcal{K}_h(\boldsymbol{w}_{s'} - \boldsymbol{w})\right)(e_t\beta_1 + \varepsilon_t)\right)\Bigg] \\
&= \frac{\beta_1^2}{T^2} E\left[\sum_{t=1}^{T}\sum_{t'=1}^{T}\sum_{s=1}^{T}\sum_{s'=1}^{T} e_s e_{s'} \mathcal{K}_h(\boldsymbol{w}_s - \boldsymbol{w})\mathcal{K}_h(\boldsymbol{w}_{s'} - \boldsymbol{w}) m(\boldsymbol{w}_t) e_{t'}\right] \\
&\quad + \frac{\beta_1}{T^2} E\left[\sum_{t=1}^{T}\sum_{t'=1}^{T}\sum_{s=1}^{T}\sum_{s'=1}^{T} e_s e_{s'} \mathcal{K}_h(\boldsymbol{w}_s - \boldsymbol{w})\mathcal{K}_h(\boldsymbol{w}_{s'} - \boldsymbol{w}) m(\boldsymbol{w}_t)\varepsilon_{t'}\right] \\
&= \frac{\beta_1^2}{T^2}\sum_{t=1}^{T}\sum_{t'=1}^{T}\sum_{s=1}^{T}\sum_{s'=1}^{T} \kappa_e(s-s')\mathcal{K}_h(\boldsymbol{w}_s - \boldsymbol{w})\mathcal{K}_h(\boldsymbol{w}_{s'} - \boldsymbol{w}) m(\boldsymbol{w}_t)\kappa_e(t-t') \\
&\quad + \frac{\beta_1}{T^2}\sum_{t=1}^{T}\sum_{t'=1}^{T}\sum_{s=1}^{T}\sum_{s'=1}^{T} \kappa_e(s-s')\mathcal{K}_h(\boldsymbol{w}_s - \boldsymbol{w})\mathcal{K}_h(\boldsymbol{w}_{s'} - \boldsymbol{w}) m(\boldsymbol{w}_t)\kappa_{e\varepsilon}(t-t')
\end{aligned}
$$

Collecting terms, the expression of V_{22} is:

$$
\begin{aligned}
V_{22} &= E(V_{221}^2) + 2E(V_{221}V_{222}) + E(V_{222}^2) \\
&= \frac{1}{T^2}\sum_{t=1}^{T}\sum_{t'=1}^{T}\sum_{s=1}^{T}\sum_{s'=1}^{T} \kappa_e(s-s')\mathcal{K}_h(\boldsymbol{w}_s - \boldsymbol{w})\mathcal{K}_h(\boldsymbol{w}_{s'} - \boldsymbol{w}) \\
&\quad \cdot \Bigg[\beta_1^2 m(\boldsymbol{w}_t)m(\boldsymbol{w}_{t'}) + 2\beta_1^2 m(\boldsymbol{w}_t)\kappa_e(t-t') + 2\beta_1 m(\boldsymbol{w}_t)\kappa_{e\varepsilon}(t-t') \\
&\quad + \beta_1^2\kappa_e(t-t') + 2\beta_1\kappa_{e\varepsilon}(t-t') + \kappa_\varepsilon(t-t')\Bigg]
\end{aligned}
$$

$\underline{\mathbf{V_{12} = V_{21}}}$:

$$
\begin{aligned}
V_{12} = V_{21} &= V_{121} + V_{122} \\
&= E\left[\left(-\frac{1}{T}\sum_{t=1}^{T}\left(\sum_{t=1}^{T} e_t \mathcal{K}_h(\boldsymbol{w}_t - \boldsymbol{w})\right)\beta_1\right) \cdot \left(\sum_{t'=1}^{T}\left(\frac{1}{T}\sum_{s=1}^{T} e_s \mathcal{K}_h(\boldsymbol{w}_s - \boldsymbol{w})\right) m(\boldsymbol{w}_{t'})\beta_1\right)\right] \\
&\quad + E\left[\left(-\frac{1}{T}\sum_{t=1}^{T}\left(\sum_{t=1}^{T} e_t \mathcal{K}_h(\boldsymbol{w}_t - \boldsymbol{w})\right)\beta_1\right) \cdot \left(\sum_{t'=1}^{T}\left(\frac{1}{T}\sum_{s=1}^{T} e_s \mathcal{K}_h(\boldsymbol{w}_s - \boldsymbol{w})\right) (e_{t'}\beta_1 + \varepsilon_{t'})\right)\right] \\
&= -\frac{\beta_1^2}{T^2}\sum_{t=1}^{T}\sum_{t'=1}^{T}\sum_{s=1}^{T}\sum_{s'=1}^{T} \kappa_e(t-s)\mathcal{K}_h(\boldsymbol{w}_t - \boldsymbol{w})\mathcal{K}_h(\boldsymbol{w}_s - \boldsymbol{w})m(\boldsymbol{w}_{t'}) \\
&\quad - \frac{\beta_1^2}{T^2}\sum_{t=1}^{T}\sum_{t'=1}^{T}\sum_{s=1}^{T}\sum_{s'=1}^{T} \kappa_e(t-s)\kappa_e(t'-s)\mathcal{K}_h(\boldsymbol{w}_t - \boldsymbol{w})\mathcal{K}_h(\boldsymbol{w}_s - \boldsymbol{w}) \\
&\quad - \frac{\beta_1}{T^2}\sum_{t=1}^{T}\sum_{t'=1}^{T}\sum_{s=1}^{T}\sum_{s'=1}^{T} \kappa_e(t-s)\kappa_{e\varepsilon}(t'-s)\mathcal{K}_h(\boldsymbol{w}_t - \boldsymbol{w})\mathcal{K}_h(\boldsymbol{w}_s - \boldsymbol{w})
\end{aligned}
$$

$$
\begin{aligned}
V_{12} = V_{21} &= V_{121} + V_{122} \\
&= E\left[\left(-\frac{1}{T}\sum_{t=1}^{T}\left(\sum_{t=1}^{T} e_t \mathcal{K}_h(\boldsymbol{w}_t - \boldsymbol{w})\right)\beta_1\right) \cdot \left(\sum_{t'=1}^{T}\left(\frac{1}{T}\sum_{s=1}^{T} e_s \mathcal{K}_h(\boldsymbol{w}_s - \boldsymbol{w})\right) m(\boldsymbol{w}_{t'})\beta_1\right)\right] \\
&\quad + E\left[\left(-\frac{1}{T}\sum_{t=1}^{T}\left(\sum_{t=1}^{T} e_t \mathcal{K}_h(\boldsymbol{w}_t - \boldsymbol{w})\right)\beta_1\right) \cdot \left(\sum_{t'=1}^{T}\left(\frac{1}{T}\sum_{s=1}^{T} e_s \mathcal{K}_h(\boldsymbol{w}_s - \boldsymbol{w})\right) (e_{t'}\beta_1 + \varepsilon_{t'})\right)\right] \\
&= -\frac{\beta_1^2}{T^2}\sum_{t=1}^{T}\sum_{t'=1}^{T}\sum_{s=1}^{T}\sum_{s'=1}^{T} \kappa_e(t-s)\mathcal{K}_h(\boldsymbol{w}_t - \boldsymbol{w})\mathcal{K}_h(\boldsymbol{w}_s - \boldsymbol{w})m(\boldsymbol{w}_{t'}) \\
&\quad - \frac{\beta_1^2}{T^2}\sum_{t=1}^{T}\sum_{t'=1}^{T}\sum_{s=1}^{T}\sum_{s'=1}^{T} \kappa_e(t-s)\kappa_e(t'-s)\mathcal{K}_h(\boldsymbol{w}_t - \boldsymbol{w})\mathcal{K}_h(\boldsymbol{w}_s - \boldsymbol{w}) \\
&\quad - \frac{\beta_1}{T^2}\sum_{t=1}^{T}\sum_{t'=1}^{T}\sum_{s=1}^{T}\sum_{s'=1}^{T} \kappa_e(t-s)\kappa_{e\varepsilon}(t'-s)\mathcal{K}_h(\boldsymbol{w}_t - \boldsymbol{w})\mathcal{K}_h(\boldsymbol{w}_s - \boldsymbol{w}).
\end{aligned}
$$

With this, we can say that:

$$
\delta_{NT}(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta} - \mathscr{B}) \xrightarrow{d} \mathcal{N}\left(0, \left(\frac{1}{T}Q'Q\right)^{-1} \boldsymbol{V} \left(\frac{1}{T}Q'Q\right)^{-1}\right),
$$

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-5 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

Name	Ticker	Year A/D	Name	Ticker	Year 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 5: Firms Added (A), Deleted (D), and No Change (NC) in Dow Jones Industrial Average during 2001–2023

detailed description of this data is available in McCracken & Ng (2016).

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-6	Yes	No
Table-7	Yes	Yes
Table-8	No	No
Table-9	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-10	Yes	No
Table-11	Yes	Yes
Table-12	No	No
Table-13	No	Yes

				$E(\widehat{\beta}_1)$				RMSE$(\widehat{\beta}_1)$			
ρ	r	T	N	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 6: Simulations using Design-I & Non-AR Errors ($\boldsymbol{\gamma} = 0.5$, 500 Reps)
[Notes: True value of β_1 is 2. We cannot estimate 2SLS when $N > T$.]

				$E(\widehat{\beta}_1)$				RMSE($\widehat{\beta}_1$)			
ρ	r	T	N	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 7: Simulations using Design-I & AR Errors ($\boldsymbol{\gamma} = 0.5$, 500 Reps)
[Notes: True value of β_1 is 2. We cannot estimate 2SLS when $N > T$.]

				$E(\widehat{\beta}_1)$				RMSE$(\widehat{\beta}_1)$			
ρ	r	T	N	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 8: imulations using Design-I & Non-AR Errors ($\boldsymbol{\gamma} = 0$, 500 Reps)
[Notes: True value of β_1 is 2. We cannot estimate 2SLS when $N > T$.]

				$E(\widehat{\beta}_1)$				RMSE($\widehat{\beta}_1$)			
ρ	r	T	N	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 9: Simulations using Design-I & AR Errors ($\boldsymbol{\gamma} = 0$, 500 Reps)
[Notes: True value of β_1 is 2. We cannot estimate 2SLS when $N > T$.]

				$E(\widehat{\beta}_1)$				RMSE$(\widehat{\beta}_1)$			
ρ	r	T	N	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 10: Simulations using Design-II & Non-AR Errors ($\boldsymbol{\gamma} = 0.5$, 500 Reps)
[Notes: True value of β_1 is 2. We cannot estimate 2SLS when $N > T$.]

				$E(\widehat{\beta}_1)$				RMSE($\widehat{\beta}_1$)			
ρ	r	T	N	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 11: Simulations using Design-II & AR Errors ($\boldsymbol{\gamma} = 0.5$, 500 Reps)
[Notes: True value of β_1 is 2. We cannot estimate 2SLS when $N > T$.]

				$E(\widehat{\beta}_1)$				RMSE$(\widehat{\beta}_1)$			
ρ	r	T	N	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 12: Simulations using Design-II & Non-AR Errors ($\boldsymbol{\gamma} = 0$, 500 Reps)
[Notes: True value of β_1 is 2. We cannot estimate 2SLS when $N > T$.]

				$E(\widehat{\beta}_1)$				RMSE$(\widehat{\beta}_1)$			
ρ	r	T	N	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 13: Simulations using Design-II & AR Errors ($\boldsymbol{\gamma} = 0$, 500 Reps)
[Notes: True value of β_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-14	Yes	No
Table-15	Yes	Yes
Table-16	No	No
Table-17	No	Yes

				$E(\widehat{\beta}_1)$				RMSE$(\widehat{\beta}_1)$			
ρ	r	T	N	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 14: Simulations using Design-III & Non-AR Errors ($\boldsymbol{\gamma} = 0.5$, 500 Reps)
[Notes: True value of β_1 is 2. We cannot estimate 2SLS when $N > T$.]

				$E(\widehat{\beta}_1)$				RMSE($\widehat{\beta}_1$)			
ρ	r	T	N	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 15: Simulations using Design-III & AR Errors ($\boldsymbol{\gamma} = 0.5$, 500 Reps)
[Notes: True value of β_1 is 2. We cannot estimate 2SLS when $N > T$.]

				$E(\widehat{\beta}_1)$				RMSE$(\widehat{\beta}_1)$			
ρ	r	T	N	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 16: Simulations using Design-III & Non-AR Errors ($\boldsymbol{\gamma} = 0$, 500 Reps)
[Notes: True value of β_1 is 2. We cannot estimate 2SLS when $N > T$.]

				$E(\widehat{\beta}_1)$				RMSE$(\widehat{\beta}_1)$			
ρ	r	T	N	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 17: Simulations using Design-III & AR Errors ($\boldsymbol{\gamma} = 0$, 500 Reps)
[Notes: True value of β_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-18	5	Yes	No	Yes
Table-19	5	Yes	Yes	Yes
Table-20	5	No	No	Yes
Table-21	5	No	Yes	Yes
Table-22	8	Yes	No	No
Table-23	8	Yes	No	Yes
Table-24	8	Yes	Yes	No
Table-25	8	Yes	Yes	Yes
Table-26	8	No	No	No
Table-27	8	No	No	Yes
Table-28	8	No	Yes	No
Table-29	8	No	Yes	Yes

				$E(\widehat{\beta}_1)$				RMSE$(\widehat{\beta}_1)$			
ρ	r	T	N	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 18: Simulations in Non-Linear Design & Non-AR Errors ($\boldsymbol{\gamma} = 0.5$, 500 Reps)
[Notes: True value of β_1 is 2. We cannot estimate 2SLS when $N > T$.]

				$E(\widehat{\beta}_1)$				RMSE($\widehat{\beta}_1$)			
ρ	r	T	N	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 19: Simulations in Non-Linear Design & AR Errors ($\boldsymbol{\gamma} = 0.5$, 500 Reps)
[Notes: True value of β_1 is 2. We cannot estimate 2SLS when $N > T$.]

				$E(\widehat{\beta}_1)$				RMSE$(\widehat{\beta}_1)$			
ρ	r	T	N	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 20: Simulations in Non-Linear Design & Non-AR Errors ($\boldsymbol{\gamma} = 0$, 500 Reps)
[Notes: True value of β_1 is 2. We cannot estimate 2SLS when $N > T$.]

				$E(\widehat{\beta}_1)$				RMSE$(\widehat{\beta}_1)$			
ρ	r	T	N	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 21: Simulations in Non-Linear Design & AR Errors ($\boldsymbol{\gamma} = 0$, 500 Reps)
[Notes: True value of β_1 is 2. We cannot estimate 2SLS when $N > T$.]

				$E(\widehat{\beta}_1)$				RMSE$(\widehat{\beta}_1)$			
ρ	r	T	N	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 22: Simulations in Linear Design & Non-AR Errors ($\boldsymbol{\gamma} = 0.5$, 500 Reps)
[Notes: True value of β_1 is 2. We cannot estimate 2SLS when $N > T$.]

				$E(\widehat{\beta}_1)$				RMSE$(\widehat{\beta}_1)$			
ρ	r	T	N	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 23: Simulations in Non-Linear Design & Non-AR Errors ($\boldsymbol{\gamma} = 0.5$, 500 Reps)
[Notes: True value of β_1 is 2. We cannot estimate 2SLS when $N > T$.]

				$E(\widehat{\beta}_1)$				RMSE($\widehat{\beta}_1$)			
ρ	r	T	N	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 24: Simulations in Linear Design & AR Errors ($\boldsymbol{\gamma} = 0.5$, 500 Reps)
[Notes: True value of β_1 is 2. We cannot estimate 2SLS when $N > T$.]

				$E(\widehat{\beta}_1)$				RMSE$(\widehat{\beta}_1)$			
ρ	r	T	N	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 25: Simulations in Non-Linear Design & AR Errors ($\boldsymbol{\gamma} = 0.5$, 500 Reps)
[Notes: True value of β_1 is 2. We cannot estimate 2SLS when $N > T$.]

				$E(\widehat{\beta}_1)$				RMSE$(\widehat{\beta}_1)$			
ρ	r	T	N	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 26: Simulations in Linear Design & Non-AR Errors ($\boldsymbol{\gamma} = 0$, 500 Reps)
[Notes: True value of β_1 is 2. We cannot estimate 2SLS when $N > T$.]

				$E(\widehat{\beta}_1)$				RMSE$(\widehat{\beta}_1)$			
ρ	r	T	N	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 27: Simulations in Non-Linear Design & Non-AR Errors ($\boldsymbol{\gamma} = 0$, 500 Reps)
[Notes: True value of β_1 is 2. We cannot estimate 2SLS when $N > T$.]

				$E(\widehat{\beta}_1)$				RMSE($\widehat{\beta}_1$)			
ρ	r	T	N	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 28: Simulations in Linear Design & AR Errors ($\boldsymbol{\gamma} = 0$, 500 Reps)
[Notes: True value of β_1 is 2. We cannot estimate 2SLS when $N > T$.]

				$E(\widehat{\beta}_1)$				RMSE($\widehat{\beta}_1$)			
ρ	r	T	N	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 29: Simulations in Non-Linear Design & AR Errors ($\boldsymbol{\gamma} = 0$, 500 Reps)
[Notes: True value of β_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.

				$E(\widehat{\beta}_1)$				RMSE$(\widehat{\beta}_1)$			
ρ	r	T	N	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 30: PIR Simulations in Design-I (4.1) & Non-AR Errors ($\boldsymbol{\gamma} = 0$, 500 Reps)
[Notes: True value of β_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.

				$E(\widehat{\beta}_1)$				RMSE$(\widehat{\beta}_1)$			
ρ	r	T	N	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 31: PIR Simulations in Design-II (4.2) with Non-AR Errors ($\boldsymbol{\gamma} = 0.5$, 500 Reps)
[Notes: True value of β_1 is 2. We cannot estimate 2SLS when $N > T$.]

				$E(\widehat{\beta}_1)$				RMSE$(\widehat{\beta}_1)$			
ρ	r	T	N	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 32: PIR Simulations in Design-III (4.3) & Non-AR Errors ($\boldsymbol{\gamma} = 0.5$, 500 Reps)
[Notes: True value of β_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)'s Post Lasso IV (PLIV) 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-33 shows the performance of the OLS method and PLIV method of Belloni *et al.* (2012). The true value of the parameter of interest β_1 is 2. In the table-33, 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	ρ	T	N	$E(\widehat{\beta}_1)$ OLS	$E(\widehat{\beta}_1)$ PLIV	RMSE($\widehat{\beta}_1$) OLS	RMSE($\widehat{\beta}_1$) PLIV
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 33: Expected value and RMSE of estimates of OLS and PLIV method of Belloni *et al.* (2012)
[Notes: True value of β_1 is 2. ρ represents extent of endogeneity, $\rho = 0$ means no-endogenity. Errors and factors are AR(1) with AR(1) coefficient being 0.5. Other Parameters: $r = 3$, number of replication =500]