# Linear Regression with Weak Exogeneity*

Anna Mikusheva†, Mikkel Sølvsten‡

August 2023

**Abstract**

This paper studies linear time series regressions with many regressors. Weak exogeneity is the most used identifying assumption in time series. Weak exogeneity requires the structural error to have zero conditional expectation given the present and past regressor values, allowing errors to correlate with future regressor realizations. We show that weak exogeneity in time series regressions with many controls may produce substantial biases and even render the least squares (OLS) estimator inconsistent. The bias arises in settings with many regressors because the normalized OLS design matrix remains asymptotically random and correlates with the regression error when only weak (but not strict) exogeneity holds. This bias's magnitude increases with the number of regressors and their average autocorrelation. To address this issue, we propose an innovative approach to bias correction that yields a new estimator with improved properties relative to OLS. We establish consistency and conditional asymptotic Gaussianity of this new estimator and provide a method for inference.

Keywords: time series regression, weak exogeneity, many controls, feedback bias, OLS inconsistency, bias correction, valid inference

JEL Codes: C13, C22

*We are grateful to Isaiah Andrews, Morten Ø. Nielsen, Jack Porter, and Jim Stock for useful discussions. Harvey Barnhard, Bas Sanders, and Chris Walker provided excellent research assistance.

†Department of Economics, M.I.T., 50 Memorial Drive, E52-526, Cambridge, MA, 02142, United States. E-mail: amikushe@mit.edu.

‡Department of Economics and Business Economics, Aarhus University, Fuglsangs Allé 4, Building 2621, B 13a, 8210 Aarhus V, Denmark. E-mail: miso@econ.au.dk. This research was supported by grants from the Danish National and Aarhus University Research Foundations (DNRF Chair #DNRF154 and AUFF Grant #AUFF-E-2022-7-3)

# 1 Introduction

Structural estimation in macroeconomics, finance and other economic fields studying dynamic models often employs time series data. The most used identifying assumption for structural estimation in time series settings is weak exogeneity. Weak exogeneity postulates that the structural shock has zero conditional expectation given the present and past regressor values. It is a less restrictive assumption than strict exogeneity, which additionally requires that the shocks have zero conditional expectation given future values of regressors. Strict exogeneity is implausible in most settings due to *feedback*, i.e., the outcome variable in one period affects the values of the regressors in future periods.[1] Specifically, if the lagged outcome variable is among the regressors, strict exogeneity cannot hold.

Another common feature of modern structural regressions is the presence of many regressors, all of which may be autocorrelated. Various motivations for the use of many regressors in time series are that the economic system generating the data is partially observed (Zellner and Palm, 1974; Wallis, 1977), additional controls in local projections may ensure uniformity (Jordà, 2005; Montiel Olea and Plagborg-Møller, 2021), and long memory may be arising from an underlying high-dimensional model (Schennach, 2018; Chevillon et al., 2018). See also Bauwens et al. (2023) for two applications.

We show that these two features – weak exogeneity and many autocorrelated regressors – can produce substantial biases and even lead to inconsistency of the ordinary least squares (OLS) estimator. The large bias in OLS may arise even when all variables are stationary (i.e., no unit roots or strong persistence is needed) and when the feedback effect violating strict exogeneity is limited to just one period. Finite sample unbiasedness of OLS relies heavily on strict exogeneity. It is well understood that OLS is biased in most time series regressions, but there is also a common belief that the biases are relatively small and of second order (see, e.g., Bao and Ullah, 2007). Our results show that such beliefs are unwarranted and that the bias in OLS can be a first-order issue.

This paper contains several results. First, we derive a formula for the asymptotically non-negligible part of the OLS bias, explain which features of the data are responsible for it, and provide a tool to assess the potential for OLS bias in a given time series application. Second,

[1] The formalization of feedback is typically ascribed to Granger (1969), while Engle et al. (1983) provide an early rigorous distinction between weak and strict exogeneity. See also Sims (1972); Chamberlain (1982) for further discussions and an empirical example.

we propose a new estimator which is asymptotically unbiased if the data contains a one-period violation of strict exogeneity. Third, we derive the asymptotic distribution for this new estimator and discuss how to conduct inferences using our new approach. Surprisingly, bias correction does not necessarily trade off with increased variance. Based on our simulations, the standard deviations of OLS and our new estimator are very close to each other with no clear ordering. Finally, we show how our new estimator generalizes to settings where strict exogeneity is violated by multiple periods of feedback effects.

The bias of OLS arises in a setting with many regressors because the properly normalized design matrix, $\frac{1}{T}X^0X$, remains asymptotically random even in large samples. Weak (but not strict) exogeneity allows for correlation between the randomness in the design matrix and the numerator of the OLS estimator, which leads to a bias. Specifying the feedback structure allows us to derive a formula for the leading term of the bias. Specifically, once we assume that strict exogeneity is violated by one-period feedback from the outcome variable to the next period's regressors, we show that the OLS bias is aligned with the feedback direction. The size of the bias increases with the number of regressors and their one-period ahead linear predictability.

We propose a new estimator which eliminates the bias asymptotically and is consistent under the same assumptions that may lead to inconsistency of OLS. Our proposal mimics an instrumental variables (IV) estimator with an intentionally endogenous 'technical' instrument: a linear combination of the regressors and their leads (future values). The main insight is that future values of the regressors in the instrument induce an endogeneity bias along the feedback direction only, the same direction along which the OLS is biased. It is therefore possible to pick the weights in the linear combination to ensure that the bias stemming from the endogenous instrument offsets the bias originating from weak exogeneity.

An important feature of our bias correction is that it is constructed based on knowledge about the regressors only. The correction is the same for any outcome variable and does not require knowledge about or estimation of the direction of the feedback mechanism. We show that the new estimator is consistent and, after proper normalization, asymptotically Gaussian when there is a one-period violation of strict exogeneity. The main results can be generalized to the case where the violation of strict exogeneity is for a finite number of periods, that is when the outcome variable has feedback effects on the regressors for $L$ periods. In such a case, the bias of OLS contains $L$ terms, corresponding to the $L$ directions

of feedback.

We conduct a simulation study aimed at assessing how common and how large the OLS bias is in typical macroeconomic regressions. We take a large collection of US macroeconomic indexes observed at quarterly frequency for 200 periods, extract its business cycle part, and randomly draw a regression from this data set. We show that the time series dependence and magnitude of feedback typical for these macroeconomic data produce empirically important OLS biases. For example, in a typical regression with 25 regressors we find a bias in the feedback direction equal to half of the standard deviation, while in regression with 50 regressors this bias approximately equals one standard deviation. Depending on the number of regressors, approximately 6-21percent of coefficients display a statistically significant difference between the OLS estimator and our proposed estimator. Our proposed IV-type estimator provides full correction of the bias.

Our results are related to three distinct strands of literature. First, there is a classical literature on linear equations in time series which shows the inappropriateness of the Generalized Least Squares (GLS) estimator in linear models with only weak exogeneity. For example, Hansen and West (2002) points out that GLS mixes up the timing of observations and – in the case of weakly exogenous regressors – leads to significant biases. The authors argue against using GLS in macroeconomic data. A concern that weak exogeneity may lead to large biases in OLS was raised by Stambaugh (1999), who considered a regression model with a very persistent (near-unit-root) regressor. Also in the absence of persistence, we show that even larger biases arise when the number of regressors is large. The second literature is that on estimation of dynamic effects in panel data, where the presence of fixed effects (many regressors) produces a sizable bias in the coefficient on the lagged outcome variable (the weakly exogenous regressor) (Nickell, 1981). Unlike the solutions proposed in that literature (e.g., Arellano and Bond, 1991), our solution for the time series context does not rely on the knowledge that only a single known regressor fails to be strictly exogenous nor does it require that the many regressors in the model are fixed effects for mutually exclusive groups. Finally, the underlying algebraic source of the bias issues, as well as some asymptotic statements related to Gaussianity of quadratic forms, are connected to the issues arising in linear models with many instruments and/or many regressors – see Hansen et al. (2008); Chao et al. (2012); Kline et al. (2020).

The rest of the paper is organized as follows. Section 2 derives the formula for the leading

term of the OLS bias, provides intuition for the findings, and discusses features of the data responsible for the bias. Section 3 introduces a new asymptotically unbiased estimator under the assumption of one-period feedback and establishes its consistency. Section 4 establishes asymptotic Gaussianity of the newly proposed estimator and suggests a valid inference procedure. Section 5 extends some results to settings with multi-period feedback. Section 6 contains simulation studies assessing the empirical relevance of the discussed issues in typical macroeconomic data sets. All proofs are in Appendix A.

**Notation:** For any vector $x$, $\|x\| = \sqrt{x'x}$ gives the $L_2$ norm of $x$. For any matrix $A$ (not necessarily square) $\text{rk}(A)$ is the rank of $A$, $\|A\| = \sup_x \frac{\|Ax\|}{\|x\|}$ denotes the operator norm (the positive square root of the largest eigenvalue of $A'A$), and $\|A\|_F = \sqrt{\text{tr}(A'A)}$ denotes the Frobenius norm. Throughout, we let $c$ and $C$ be positive and finite (generic) constants, that may have different values in different equations, but do not depend on the sample size $T$. The $q \quad q$ identity matrix is $I_q$ while $I = I_T$.

# 2 Inconsistency of OLS

## 2.1 Model and assumptions

Consider a linear time series regression

$$y_t = x_t'\beta + \varepsilon_t, \qquad t \in \{1, \ldots, T\},$$

where the regressors $x_t \in \mathbb{R}^K$ are weakly exogenous and the number of regressors $K$ is large. We model this by assuming that $K$ may diverge proportionally to $T$ or slower. All features of the data generating process are implicitly indexed by $T$, but we drop this index for compactness of notation. The object of interest is the linear contrast $\theta = r'\beta$ for a known (non-random) $K$ vector $r$. A leading case is $r'\beta = \beta_1$.

The assumption of weak exogeneity imposes:

$$\mathsf{E}[\varepsilon_t | x_t, x_{t-1}, \ldots] = 0.$$

Weak exogeneity is considerably less restrictive than strict exogeneity, which assumes that

$\mathsf{E}[\varepsilon_t \mid X] = 0$, with $X$ denoting the complete $T \times K$ set of regressors (including past, present, and future). In the context of time series economic data, strict exogeneity is seldom plausible. In contrast, weak exogeneity allows the structural shock to influence future regressor values through feedback. The likelihood of such an effect is substantial, as variables employed in macro estimation often evolve simultaneously within dynamic processes. When outcome variable lags are incorporated as regressors, strict exogeneity is certainly violated. For a textbook discussion regarding the plausibility of weak and strict exogeneity in various applications, see Stock and Watson (2019, ch. 16).

The standard argument for unbiasedness of the OLS estimator $r'\hat{\beta}^{\text{OLS}} = r'(X'X)^{-1}X'y$ heavily relies on strict exogeneity. Although it is well-known that OLS exhibits bias under weak exogeneity (see, e.g., Hamilton, 1994, chapter 8.2), a common belief persists that this bias is small and asymptotically negligible. Here we claim that this widespread belief is both incorrect and misleading: the bias of OLS can be substantial, potentially leading to inconsistency.

Standard statements regarding OLS consistency in a time series context hinge on the assumption that the normalized design matrix concentrates around its expectation: $\left\|\frac{1}{T}X'X - Q\right\| \xrightarrow{p} 0$, where $Q$ is non-singular (see, e.g., Hamilton, 1994, Assumption 8.6). For example, Gupta and Seo (2023) assumes that $K^3/T \to 0$ to invoke a Law of Large Numbers for the normalized design matrix. However, the normalized design matrix remains asymptotically random when the number of regressors (and thus design matrix dimensionality) grows fast enough. Under strict exogeneity the randomness of $\frac{1}{T}X'X$ does not cause a problem as one can condition on $X$ in the analysis, thus treating the design matrix as non-random. This approach is infeasible under weak exogeneity. In such instances, $\frac{1}{T}X'X$ is not just a random matrix, but is also correlated with $X'y$. This correlation produces a bias that may become the leading term in the asymptotic analysis of OLS.

The size and form of the OLS bias depend on the feedback mechanism, or how past errors in the outcome variable affect future regressor values. We start by assuming one-period linear feedback, i.e., the present error term $\varepsilon_t$ only affects the regressors in the subsequent period, namely $x_{t+1}$. Section 5 extends our results to feedback lasting a finite number of periods.

**Assumption 1.** *(i) The observed regressors $x_t$ can be decomposed as $x_t = \tilde{x}_t + \alpha\varepsilon_{t-1}$, where the $T \times K$ matrix $\tilde{X} = [\tilde{x}_1, \ldots, \tilde{x}_T]'$ has full rank.*

*(ii) The errors* $\{\varepsilon_t\}_{t=0}^{T}$ *are i.i.d. conditionally on* $\tilde{X}$ *with* $\mathsf{E}[\varepsilon_t|\tilde{X}] = 0$, $c < \sigma^2 := \mathsf{E}[\varepsilon_t^2|\tilde{X}] < C$, *and* $\mathsf{E}[\varepsilon_t^4|\tilde{X}] < C$ *almost surely.*

*(iii) The size of the non-random vectors* $\alpha, r \in \mathsf{R}^K$ *relative to the strictly exogenous design matrix* $\frac{1}{T}\tilde{X}'\tilde{X}$ *is bounded:* $\alpha'(\frac{1}{T}\tilde{X}'\tilde{X})^{-1}\alpha = O_p(1)$ *and* $r'(\frac{1}{T}\tilde{X}'\tilde{X})^{-1}r = O_p(1)$.

*(iv) The number of regressors* $K$ *may diverge with sample size* $T$ *such that* $\frac{K}{T} < 1 \quad c.$

Part (i) describes a specific violation of strict exogeneity where the present error term only affects the regressors in the subsequent period. If $\alpha = 0$, then all regressors $x_t = \tilde{x}_t$ are strictly exogenous. The amount of assumptions imposed on the strictly exogenous part of the regressors $\{\tilde{x}_t\}_{t=1}^{T}$ is kept to a minimum, the main part of which is a full rank condition. Specifically, we do not assume stationarity or impose any moment conditions on $\tilde{X}$. Such generality is possible mostly because our analysis is done conditionally on $\tilde{X}$. Part (ii) is a standard set of assumptions on the error terms in homoskedastic regression models. Part (iii) places a very loose bound on the magnitudes of $\alpha$ and $r$ relative to the scaled design matrix $\frac{1}{T}\tilde{X}'\tilde{X}$. This condition is, for example, satisfied if $\|\alpha\|$ and $\|r\|$ are bounded and the smallest eigenvalue of $\frac{1}{T}\tilde{X}'\tilde{X}$ is separated from zero. Part (iv) allows a wide range for the number of regressors, including a small fixed number. It requires that the degrees of freedom $T \quad K$ diverges to infinity as the sample size increases.

The violation of strict exogeneity in Assumption 1 seems minimal as it is solely due to feedback from the dependent variable $y_t$ to the one-period-ahead regressors $x_{t+1}$ and the magnitude of the feedback is bounded. However, this violation is enough to produce inconsistency of the OLS estimator.

## 2.2 Parameter estimator

In order to provide some intuition for how and why the OLS bias arises, we consider a special case. Suppose, therefore, that only the first regressor fails to be strictly exogenous. Then, the first regressor experiences a one-period feedback: $x_{1t} = \tilde{x}_{1t} + a\varepsilon_{t\ \ 1}$, while the $T \quad (K \quad 1)$ matrix of all other regressors, $X_{\ \ 1} = \tilde{X}_{\ \ 1}$, is strictly exogenous. Let $X_1$ be the $T \quad 1$ vector with elements $x_{1t}$ and $\tilde{X}_1$ its strictly exogenous part. Furthermore, we take $\tilde{X}$ as fixed and suppose that $\frac{1}{T}\tilde{X}'\tilde{X} = I_K$.

In this special case, we now derive the bias of the OLS estimator for the first coefficient $\beta_1$. According to the Frisch-Waugh-Lovell theorem, we have

$$\hat{\beta}_1^{\text{OLS}} = \frac{X_1^{\prime} M_{\ 1} y}{X_1^{\prime} M_{\ 1} X_1} \quad \text{and} \quad \hat{\beta}_1^{\text{OLS}} \quad \beta_1 = \frac{X_1^{\prime} M_{\ 1} \varepsilon}{X_1^{\prime} M_{\ 1} X_1},$$

where the partialling out operator $M_{\ 1} = I \quad \tilde{X}_{\ 1}(\tilde{X}^{\prime}{}_1 \tilde{X}_{\ 1})^{\ 1} \tilde{X}^{\prime}{}_1$ is the projection on the space orthogonal to $\tilde{X}_{\ 1}$. Notice that $M_{\ 1} = \left(M_{st}\right)$ is fixed (or can be conditioned on) so that all randomness comes from $\varepsilon$. The denominator of $\hat{\beta}_1^{\text{OLS}} \quad \beta_1$ has a standard asymptotic behavior and once normalized by $\frac{1}{T}$ converges to a non-random and non-zero limit.

If all regressors were strictly exogenous ($X_1 = \tilde{X}_1$), then the OLS estimator would have been properly centered. However, $X_1$ contains randomness that is correlated with $\varepsilon$. This leads to a non-zero expectation of the numerator in $\hat{\beta}_1^{\text{OLS}} \quad \beta_1$:

$$\mathsf{E}[X_1^{\prime} M_{\ 1} \varepsilon \ \mathsf{j} \ \tilde{X}] = \mathsf{E}[\tilde{X}_1^{\prime} M_{\ 1} \varepsilon \ \mathsf{j} \ \tilde{X}] + a\mathsf{E}[\sum_{s,t} M_{st} \varepsilon_s \ {}_1 \varepsilon_t \ \mathsf{j} \ \tilde{X}] = a\sigma^2 \sum_t M_{tt \ 1}.$$

While the weakly exogenous regressor $x_{1t}$ is not correlated with the contemporaneous error $\varepsilon_t$, the process of partialling out the remaining regressors mixes up the timing of the observations and makes the partialled out regressor correlated with the contemporaneous error. This derivation suggest a form for the leading term of the bias of $\hat{\beta}_1^{\text{OLS}}$. We may notice that the bias depends on $a$ which governs the magnitude of feedback in the regressors and on the trace of the lower diagonal of the projection matrix $M_{\ 1}$ ($\sum_t M_{tt \ 1}$). As discussed below, this lower trace may be on the order of the number of regressors and increases with their average autocorrelation. It is also possible to show that the remaining OLS coefficients (e.g., $\hat{\beta}_2^{\text{OLS}}$) have an asymptotically negligible bias (although they have a finite sample bias). Thus for any linear combination $\theta = r^{\prime}\beta$, the bias depends on the weight placed on $\beta_1$.

The insight from the preceding special case can be extended. In fact, the OLS estimator of a linear contrast remains invariant under linear transformations of the regressors. For instance, if we perform a regression of $Y$ on $XA$, where $A$ is a $K \quad K$ matrix with full rank that linearly transforms the regressors, then the OLS serves as an estimator of $A^{\ 1}\beta$. To achieve the contrast $\theta$, we should apply $A^{\prime} r$ as the weighting.

Any OLS scenario simplifies to the one discussed earlier when we choose $A = (\frac{1}{T}\tilde{X}^{\prime}\tilde{X})^{\ 1/2}$, with the square root selected to ensure that $A^{\prime}\alpha$ is proportional to the first basis vector. These

insights, coupled with certain technical derivations, lead to the following characterization of the OLS bias.

**Theorem 1** (Inconsistency of OLS estimator)**.** *Suppose Assumption 1 holds. Then,*

$$r'\hat{\beta}^{\text{OLS}} \quad r'\beta = \sigma^2 r'\bar{S}^{-1}\alpha \sum_{t=2}^{T} \tilde{M}_{tt-1} + o_p(1),$$

*where* $\bar{S} = \tilde{X}'\tilde{X} + \alpha\alpha'\sigma^2(T \quad K)$ *and* $\tilde{M} = I \quad \tilde{X}(\tilde{X}'\tilde{X})^{-1}\tilde{X}'$.

Theorem 1 presents the leading term of the OLS bias under Assumption 1. Note that this formula uses the lower diagonal trace $\sum_t \tilde{M}_{tt-1}$ of the projection matrix orthogonal to $\tilde{X}$ instead of $\sum_t M_{tt-1}$. Theorem 1 establish asymptotic negligibility of this difference.

It is worth discussing the size of the lower trace of the projection matrix as well as the size of the biases we may see in applications. Consider the term $\sum_t \tilde{M}_{tt-1} = \sum_t \tilde{P}_{tt-1}$, where $\tilde{P} = \tilde{X}(\tilde{X}'\tilde{X})^{-1}\tilde{X}'$. Note that this quantity is unchanged by any full rank rotation of the regressors. For simplicity, suppose that the regressors are generated by a stationary and ergodic process with $\frac{1}{T}\tilde{X}'\tilde{X} = I_K$. Then

$$\sum_t \tilde{M}_{tt-1} = \frac{1}{T}\sum_t \tilde{X}_t'\tilde{X}_{t-1} \qquad \mathsf{E}\tilde{X}_t'\tilde{X}_{t-1}.$$

Thus, the lower diagonal trace of $\tilde{M}$ measures a linear connection between $\tilde{X}_t$ and $\tilde{X}_{t-1}$. We can also notice that $\frac{1}{T}\sum_t \tilde{X}_t'\tilde{X}_t = \mathsf{E}\tilde{X}_t'\tilde{X}_t = K$, since we are dealing with $K$-dimensional vectors. In time series settings we tend to work with data where the regressors are autocorrelated. This way we should expect $\sum_t \tilde{M}_{tt-1} \qquad \rho K$, where $\rho$ is the average (over different regressors) scalar measure of regressor predictability.

We can notice that $\bar{S} \qquad \tilde{X}'\tilde{X}$ which leads to Assumption 1(iii) implying $r'\bar{S}^{-1}\alpha = O_p(1/T)$. Let us momentarily assume $Tr'\bar{S}^{-1}\alpha$ converges to a constant $r_\alpha$. When $K$ grows proportionally to the sample size while $\rho$ remains bounded away from zero, the leading bias term becomes approximately $\sigma^2 r_\alpha \frac{\rho K}{T}$. This renders the OLS estimator for $r'\beta$ inconsistent. Notably, the potential bias term is more pronounced for regressors with higher first-order autocorrelation and a larger number of regressors.

Although $\frac{1}{T}\sum_{t=2}^{T} \tilde{M}_{tt-1}$ (the lower diagonal trace of $\tilde{M}$ over the sample size) is typically unavailable in practice as the strictly exogenous part of the regressors is unobservable, the

asymptotically equivalent lower diagonal trace of $M = I \quad X(X^{0}X)^{\ 1}X^{0}$ over the sample size can be calculated from available data. If this quantity is non-negligible in an application, even a small violation of strict exogeneity may result in substantial biases.

Lastly, the extent of bias depends on the alignment between the contrast $r$ and the feedback direction $\alpha$, making the feedback direction the most affected contrast direction. Contrasts in all directions orthogonal to $\alpha$ (using the scalar product weighted by $\bar{S}^{\ 1}$) experience only negligible bias. In our special case where only the first regressor is weakly exogenous, this observation corresponds to the OLS estimator of $\beta_1$ being the sole estimator with significant bias. Thus, for any linear contrast $r^{0}\beta$, the bias depends on the weight placed on $\beta_1$. In real applications, the feedback direction is unknown and challenging to estimate empirically, given that $\alpha$ is a $K \quad 1$ vector.

**Simulations** We demonstrate the potential for OLS bias through a small scale simulation that varies the number of regressors and their short-term dependence. Data is generated following Assumption 1. The outcome vector is generated as $y = X\beta + \varepsilon$ with $\varepsilon \quad N(0, I)$ and $\beta = 0$. The design matrix is generated as $x_{1t} = \tilde{x}_{1t} + a\varepsilon_{t\ 1}$ and $X_{\ 1} = \tilde{X}_{\ 1}$, where $\tilde{X}$ is generated as a rotated VAR(1) process with $\tilde{X}\tilde{X}^{0}/T = I_K$, independent from $\varepsilon$. Specifically, we generate $V_t = \rho V_{t\ 1} + u_t$ with $\{u_t\}_{t=1}^{T}$ *i.i.d.* $N(0, I_K)$ and define $\tilde{X} = V(V^{0}V/T)^{\ 1/2}$, where the square root comes from Cholesky decomposition. Across simulations, we fix the sample size at $T = 200$ and the coefficient on the feedback mechanism at $a = 1.5$. Simulation results are summarized in Figure 1 with the left panel showing results for the number of regressors $K$ between 4 and 100 (fixing $\rho$ at 0.8).[2] The right panel reports the results for the auto-correlation in regressors $\rho$ between 0 and 0.98 (fixing $K$ at 50). We report simulated values of absolute bias and standard deviation for the first coordinate of OLS together with the mean absolute value of the ratio of the lower trace of $M$ to the sample size. Additionally, we report the same summary statistics for the new estimator proposed in Section 3.

The results presented in Figure 1 suggest that the bias of the OLS estimator can easily surpass its standard deviation, leading to highly unreliable statistical inferences. For example, even a regression with just 20 regressors may exhibit an OLS bias comparable in magnitude to the standard deviation in a sample of size 200 — a very common setting in macroeconomic applications. The observable lower trace of $M$ divided by the sample size

[2] All results are presented as sixth order polynomial fits to the actual results across $K$.

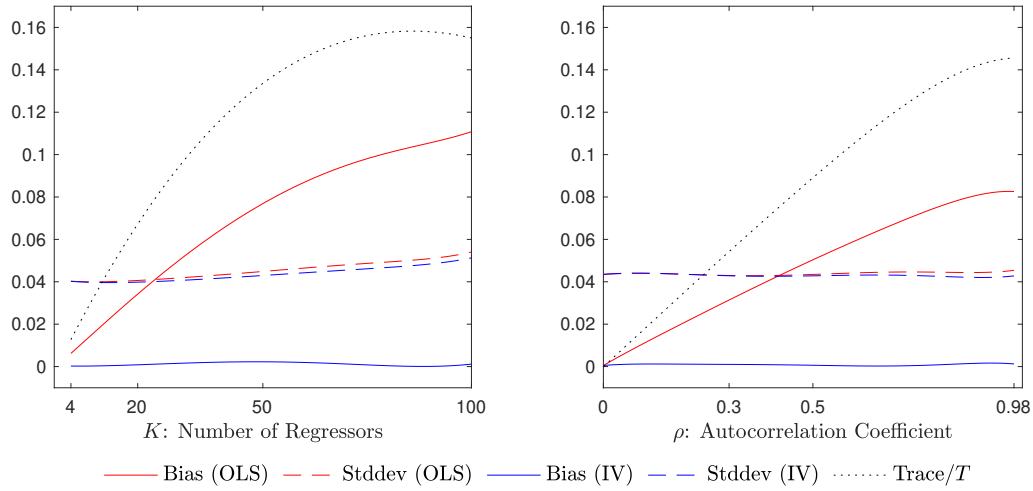


Figure 1: Absolute Bias and Standard Deviation of OLS and IV with $T = 200$

provides a highly predictive measure of the magnitude of the bias in the most affected direction. Notably, both the bias in the most affected direction and the lower diagonal trace tend to increase with the number of regressors and the first auto-correlation of the regressors.[3]

Figure 2 presents results for the same simulation design but with a sample size of $T = 800$, while the number of regressors varies from 16 to 400. Here, we observe that the bias reaches the same level as in Figure 1 when the number of regressors is the same fraction of the sample size, while the standard deviations drops two-fold. This demonstrates the inconsistency of the OLS for the worst direction when the number of regressors $K$ grows proportionally to $T$. In essence, the estimator concentrates around an incorrect value as the sample size increases.

## 2.3 Variance estimator

In order to make reliable statistical inferences in a linear regression we usually need an estimator for the variance of the error term $\sigma^2$. It is well known that in a regression with many regressors and strict exogeneity one has to adjust properly for the degrees of freedom in order to correct for over-fitting. The most commonly applied OLS estimator of the error variance uses this adjustment $\hat{\sigma}^2 = \frac{y'My}{T-K}$. It is not obvious ex ante whether this estimator

[3]One may be worried that the observed biases are due to persistence in the regressors as the coefficient $\rho$ in an AR(1) process measures not only short-term dependence but also the long-run persistence. We re-run simulations generating $V_t = \rho u_{t-1} + u_t$ as an MA(1) process. The results are presented in Appendix B and are essentially identical to those reported in Figure 1.

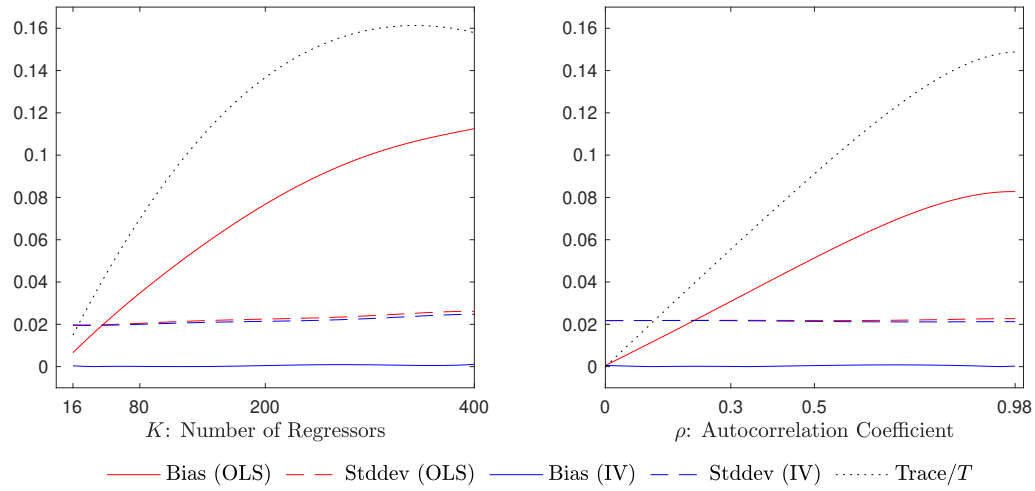


Figure 2: Absolute Bias and Standard Deviation of OLS and IV with $T = 800$

retains consistency with many regressors and only weak exogeneity. On the one hand, a large bias of the OLS estimator for coefficients raises concerns about the consistency of the variance estimator. On the other hand, the bias arises only in the direction of the feedback but not in all directions orthogonal to the feedback, which are numerous. Theorem 2 answers the question of the consistency of the OLS variance estimator.

**Theorem 2** (Inconsistency of OLS variance)**.** *Suppose Assumption 1 holds. Then,*

$$\frac{\hat{\sigma}^2}{\sigma^2} = 1 \quad \frac{\sigma^2 \alpha' \bar{S}^{-1} \alpha}{T \quad K} \left( \sum_{t=2}^{T} \tilde{M}_{tt-1} \right)^2 + o_p(1).$$

Theorem 2 states that the OLS variance of the error estimate tends to be biased downward and reports an overly optimistic measure of fit in a setting with one-period feedback and many regressors. The size of the bias can be judged by the trace of the lower diagonal of the projection matrix $M$. When the number and predictability of the regressors are high enough to imply that the OLS estimator of the linear contrast in the worst direction (the feedback direction) is inconsistent, the OLS estimator of variance is inconsistent as well. Despite the inconsistency result, the biases for the OLS variance we observe in simulations tend to be relatively minor.

# 3 A consistent IV estimator

## 3.1 The idea of the proposed estimator

Let us introduce a $T \quad T$ shift matrix $D$ (or lag operator matrix) that shifts the time series index back by one; its only non-zero elements are $D_{t,t\ 1} = 1$ for all $t$. The transpose $D^{\prime}$ is the lead operator, moving the time-series index forward by one. Notice that the lower trace appearing in the bias of OLS can be written as $\sum_{t=2}^{T} \tilde{M}_{tt\ 1} = \mathrm{tr}(D^{\prime}\tilde{M})$. Consider also a $T \quad T$ matrix $\Gamma$ measurable with respect to $\tilde{X}$ such that $\|\Gamma\| \quad 1 \quad c$.

We propose an IV-inspired estimator that relies on an endogenous instrument $Z = (I \quad \Gamma^{\prime})X$:

$$\hat{\beta}^{\mathrm{IV}}(\Gamma) = (Z^{\prime}X)^{\ 1}Z^{\prime}y = (X^{\prime}(I \quad \Gamma)X)^{\ 1}X^{\prime}(I \quad \Gamma)Y.$$

The key insight behind the estimator is that we use a deliberately invalid instrument created in a way so the 'invalid instrument' bias offsets the bias in OLS. In order to eliminate the asymptotic bias we require that $\Gamma$ solves the following (non-linear) one-dimensional equation:

$$\mathrm{tr}\big[D^{\prime}(I \quad \Gamma)\tilde{M}_{\Gamma}\big] = 0 \tag{1}$$

where $\tilde{M}_{\Gamma} = I \quad \tilde{X}(\tilde{X}^{\prime}(I \quad \Gamma)\tilde{X})^{\ 1}\tilde{X}^{\prime}(I \quad \Gamma) = I \quad \tilde{X}(\tilde{Z}^{\prime}\tilde{X})^{\ 1}\tilde{Z}^{\prime}$ is an oblique projection off $\tilde{X}$ in the direction of $\tilde{Z} = (I \quad \Gamma^{\prime})\tilde{X}$. Further below, we establish that the new estimator $\hat{\beta}^{\mathrm{IV}}(\Gamma)$ is a properly centered $\sqrt{T}$ asymptotically Gaussian (conditionally on $\tilde{X}$) estimator under Assumption 1. The main theoretical result about bias is obtained for a general $\Gamma$ with the leading example of $\Gamma = \gamma D$ for $|\gamma| < 1 \quad c$.

Let us give some intuition of why this approach would work. Consider again the special case when all regressors but the first are strictly exogenous. Namely $x_{1t} = \tilde{x}_{1t} + a\varepsilon_{t\ 1}$, while $X_{\ 1} = \tilde{X}_{\ 1}$ and consider the first coefficient $\beta_1$ only, as it is the most biased direction. Consider also the case of $\Gamma = \gamma D$, thus $z_t = x_t \quad \gamma x_{t+1}$. In this setting all but the first instruments are strictly exogenous, while the first instrument $z_{1t} = \tilde{z}_{1t} + \varepsilon_{t\ 1} \quad \gamma\varepsilon_t$ is endogenous as it correlates with the contemporaneous regression error. The direction of the instrument endogeneity in general coincides with the feedback direction, as transformation $\Gamma$ mixes up timing of the observations but preserves the feedback direction.

Oblique projections underlie the geometry of IV estimation similarly to how orthogonal projections explain the geometry of OLS. An oblique projection is defined as $M_{Z,X} = I \quad X(Z'X)^{-1}Z'$, where $X$ and $Z$ are of the same dimension and $Z'X$ is invertible. Oblique projections satisfy idempotency, $M^2_{Z,X} = M_{Z,X}$, but not symmetry, $M'_{Z,X} \neq M_{Z,X}$. One can easily show that the Frisch-Waugh-Lowell theorem holds for oblique projections as well. Specifically, let $Z = [Z_1; Z_{-1}]$ where the dimensions of the corresponding Z's and X's coincide and all proper matrices are invertible. Then

$$\hat{\beta}_1^{\text{IV}}(\Gamma) = \frac{Z_1' M_{Z_{-1}, X_{-1}} Y}{Z_1' M_{Z_{-1}, X_{-1}} X_1} \quad \text{and} \quad \hat{\beta}_1^{\text{IV}}(\Gamma) \quad \beta_1 = \frac{Z_1' M_{Z_{-1}, X_{-1}} \varepsilon}{Z_1' M_{Z_{-1}, X_{-1}} X_1}.$$

Notice that since $X_{-1}$ is strictly exogenous, $Z_{-1}$ is strictly exogenous as well. If we condition on $\tilde{X}$ we may then treat $M_{Z_{-1}, X_{-1}}$ as fixed. We look at the numerator:

$$\mathsf{E}[Z_1' M_{Z_{-1}, X_{-1}} \varepsilon] = \mathsf{E}[\tilde{Z}_1' M_{Z_{-1}, X_{-1}} \varepsilon] + a\mathsf{E}[\sum_{s,t} M_{st}(\varepsilon_{s-1} \quad \gamma\varepsilon_s)\varepsilon_t] = a\sigma^2 \sum_t (M_{tt-1} \quad \gamma M_{tt}).$$

Here we used $M_{Z_{-1}, X_{-1}} = (M_{st})$ for shortness of notation. Our goal is to choose $\gamma$ in a manner that renders the last sum equal to zero. This should generally be feasible, given that the diagonal elements of projection matrices tend to dominate those on the lower diagonal. In the proof, we show that changing $M_{Z_{-1}, X_{-1}}$ to $\tilde{M}_\Gamma$ in the bias expression introduces only an asymptotically negligible difference. Thus, the expectation of the numerator is asymptotically equivalent to $a\sigma^2 \operatorname{tr}[D'(I \quad \Gamma)\tilde{M}_\Gamma]$. By selecting $\Gamma$ to solve equation (1), we achieve an asymptotically unbiased estimator.

## 3.2 Consistency of estimator

Let $\gamma_0$ denote the solution to (1) among matrices $\Gamma_0 = \gamma_0 D$. Since we only observe regressors $X$, knowing $\tilde{X}$ is equivalent to knowing the feedback direction $\alpha$. This makes solving equation (1) infeasible in practice. Let $\hat{\gamma}$ and $\hat{\Gamma} = \hat{\gamma} D$ be the solution to the empirically feasible equation:

$$\operatorname{tr}\big[D'(I \quad \hat{\Gamma})M_{\hat{\Gamma}}\big] = 0 \quad \text{where} \quad M_{\hat{\Gamma}} = I \quad X(X'(I \quad \hat{\Gamma})X)^{-1}X'(I \quad \hat{\Gamma}). \tag{2}$$

The following Lemma gives sufficient conditions for the existence and uniqueness of a solution to equation (1).

**Lemma 1.** *Suppose that $\tilde{X}$ is a $T$ $K$ matrix of rank $K$ and $\Gamma = \gamma D$.*

*(i) If $K < T/5$, then there exists a unique $\gamma$ 2 $[-1/2, 1/2]$ solving equation (1).*

*(ii) If $|\mathrm{tr}(D^{\prime}\tilde{M})|$ $\mu^2 K$ and $K < T/[1 + (1+\mu)^2]$ for some $\mu$ 2 $[0, 1]$, then there exists a unique $\gamma$ solving equation (1) such that $|\gamma| < \mu/(1+\mu)$.*

One may state an analog of Lemma 1 for equation (2) as well due to their analog structure. According to Lemma 1, equation (1) typically can be solved in a one-dimensional family of transformations $\Gamma = \gamma D$. It is possible to search for a solution in other classes of matrices; we leave the question of finding an optimal class of transformations $\Gamma$ to future research. The proof of Lemma 1 reveals that though equation (1) is non-linear, its solution can be found relatively fast as a fixed point of a contraction.

One may think that $K < T/5$ in Part (i) of Lemma 1 is a restrictive condition. It arises because we do not put any restriction on $\tilde{X}$ other than full rank. This accommodates an extensive range of idiosyncrasies in the generation of regressors, surpassing those encountered in typical macroeconomic time series data sets. Placing some restrictions on regressors may weaken the restriction on the sample size significantly. For example, putting a bound on the average auto-correlation of the regressors, $\mu$, eases this restriction considerably as shown in Part (ii) of Lemma 1. We also notice that if the original potential for bias is small, then the $\gamma$ that removes the bias is small as well. Specifically, if $\mathrm{tr}[D^{\prime}\tilde{M}] = 0$ so the original OLS has no (asymptotic) bias, our estimator defaults back to OLS.

The following theorem establishes the asymptotic bias of IV for a generic choice of $\Gamma$ and shows consistency for the specific choice of $\hat{\Gamma} = \hat{\gamma} D$.

**Theorem 3.** *Suppose Assumption 1 holds.*

*(i) If $\Gamma$ is $\tilde{X}$-measurable and $\|\Gamma\| < 1$ $c$, then*

$$r^{\prime}\hat{\beta}^{\mathrm{IV}}(\Gamma) \quad r^{\prime}\beta = \sigma^2 r^{\prime}\bar{S}_{\Gamma}^{-1}\alpha\,\mathrm{tr}\big[D^{\prime}(I \quad \Gamma)\tilde{M}_{\Gamma}\big] + o_p(1),$$

*where $\bar{S}_{\Gamma} = \tilde{X}^{\prime}(I \quad \Gamma)\tilde{X} + \sigma^2\alpha\alpha^{\prime}\,\mathrm{tr}\big[(I \quad \Gamma)\tilde{M}_{\Gamma}\big]$. Speci cally, it follows that $r^{\prime}\hat{\beta}^{\mathrm{IV}}(\Gamma)$ is consistent for $r^{\prime}\beta$ when $\Gamma$ solves equation (1).*

*(ii) If* $K < T/5$*, then* $\hat{\gamma} \quad \gamma_0 = O_p(\frac{1}{T})$*, and* $r^{\prime}\hat{\beta}^{\mathrm{IV}}(\hat{\Gamma}) \quad r^{\prime}\hat{\beta}^{\mathrm{IV}}(\Gamma_0) = o_p(\frac{1}{\sqrt{T}})$*.*

The result of Theorem 1 is a special case of Part (i) of Theorem 3 for $\Gamma = 0$. Using $\Gamma$ that solves equation (1) produces a consistent estimator for any reasonable contrast, and the proposed solution does not depend on the contrast of interest. The ideal $\Gamma_0$ solving equation (1) can be random but is strictly exogenous, as it depends only on the strictly exogenous part of the regressors $\tilde{X}$. Solution $\hat{\Gamma}$ to equation (2) is random and depends on error term $\varepsilon$. However, as stated in part (ii) of Theorem 3 the feasible estimator $r^{\prime}\hat{\beta}^{\mathrm{IV}}(\hat{\Gamma})$ is consistent and has the same asymptotic distribution as the infeasible one using the ideal $\Gamma_0$.

An appealing feature of our proposed solution is that our estimator is linear in the outcome variable. Our proposal eliminates bias by working only with the regressors. The same bias correction would work for any outcome variables that satisfy the weak exogeneity assumption with one-period feedback to regressors. Our solution does not require one to estimate the direction of feedback $\alpha$.

Figure 1 shows that in the simulations the proposed IV estimator fixes the bias of OLS with essentially no increase in the standard deviation of the estimator.

## 3.3 Consistency of variance estimator

Let us introduce $T \quad K_\Gamma = \mathrm{tr}\left[(I \quad \Gamma)\tilde{M}_\Gamma\right]$ and define an estimator for variance $\hat{\sigma}^2(\Gamma) = \frac{y^{\prime}(I \quad \Gamma)M_\Gamma y}{T \quad K_\Gamma}$ for some matrix $\Gamma$. Using equation (4) in the Appendix, we find that the value of $\hat{\sigma}^2(\Gamma)$ is invariant to the value of $\beta$ and $\hat{\sigma}^2(\Gamma) = \frac{\varepsilon^{\prime}(I \quad \Gamma)M_\Gamma \varepsilon}{T \quad K_\Gamma}$. Part (ii) of Lemma 3 from the Appendix implies that $\hat{\sigma}^2(\Gamma)$ possesses a very desirable property for a variance estimator, namely, it is non-negative for any realization of the data. If $\Gamma$ solves equation (1) then $\hat{\sigma}^2(\Gamma)$ is consistent for $\sigma^2$.

**Theorem 4.** *Suppose Assumption 1 holds,* $\Gamma$ *is* $\tilde{X}$*-measurable, and* $k\Gamma k < 1 \quad c$*. Then*

$$\frac{\hat{\sigma}^2(\Gamma)}{\sigma^2} = 1 \quad \frac{\sigma^2\alpha^{\prime}\bar{S}_\Gamma{}^1\alpha}{T \quad K_\Gamma}\,\mathrm{tr}\left[D^{\prime}(I \quad \Gamma)\tilde{M}_\Gamma\right]\mathrm{tr}\left[D(I \quad \Gamma)\tilde{M}_\Gamma\right] + o_p(1).$$

*Speci cally, it follows that* $\hat{\sigma}^2(\Gamma)$ *is consistent for* $\sigma^2$ *when* $\Gamma$ *is such that equation* (1) *holds.*

As with Theorem 3, the theoretically desirable $\Gamma$ that solves (1) is not known, but one can search for an empirically feasible value $\hat{\Gamma}$ that solves (2). One can easily use Part (ii) of Theorem 3 to extend the argument of Theorem 4 and also show that $\hat{\sigma}^2(\hat{\Gamma})$ is consistent.

# 4 Inference

In this section we show that under some additional assumptions the proposed IV estimator is asymptotically Gaussian conditionally on $\tilde{X}$. We also suggest standard errors that can be paired with our estimator in order to achieve asymptotically valid inferences, that is, to construct confidence sets and/or to produce reliable t-tests.

There are two theoretical and practical challenges we encounter. The first is a need to correctly account for the asymptotic importance of a quadratic form. Specifically, the bias of OLS arises due to presence of a quadratic form in errors, which has a non-trivial mean. We construct our estimator to guarantee a zero mean of its corresponding quadratic form and thus eliminate the bias asymptotically. The quadratic form in error terms with zero mean is asymptotically Gaussian when its rank is growing to infinity under a condition of eigenvalue negligibility. We refer the reader to Anatolyev (2019); Chao et al. (2012); Sølvsten (2020); Kline et al. (2020) for examples of the Central Limit Theorems for quadratic forms. The naive standard errors tend to improperly account for asymptotic uncertainty of a quadratic form. The asymptotic importance of such quadratic forms has appeared previously in the literature on linear models with many instruments and/or many regressors. For example, Hansen et al. (2008) show the importance to adjust standard errors for presence of a quadratic form in many instrument settings. See also Anatolyev (2019) for a comprehensive survey of the issue. The second challenge is that the quadratic form we end up with has non-zero diagonal elements that are on average zero. This makes the asymptotic variance depend on skewness and kurtosis of errors that are hard to estimate. This is similar to the issue of proper inference for LIML-type estimators with many instruments (as in Hansen et al. (2008)). Below we consider two instances when we can handle the first challenge with relative ease and ignore the second challenge.

## 4.1 Inference when $K/T \to 0$

**Theorem 5.** *Suppose Assumption 1 holds,* $\max_t \left\| (\tilde{X}'\tilde{X})^{-1/2}\tilde{x}_t \right\| = o_p(1)$. *Then, as* $T \to \infty$,

$$\frac{r'\hat{\beta}^{\mathrm{IV}}(\hat{\Gamma}) - r'\beta}{\sqrt{\hat{\Sigma}_T}} \to N(0,1)$$

*where* $\hat{\Sigma}_T = \hat{\sigma}^2(\hat{\Gamma})kr^{\prime}(X^{\prime}(I \quad \hat{\Gamma})X)^{-1}X^{\prime}(I \quad \hat{\Gamma})k^2$.

Condition $\max_t \left\|(\tilde{X}^{\prime}\tilde{X})^{-1/2}\tilde{x}_t\right\| = o_p(1)$ is a relatively standard negligibility condition often invoked in order to obtain asymptotic Gaussianity of OLS using Lindeberg CLT (see, e.g., Koenker and Machado, 1999, page 334). It implies among other things that the maximal diagonal element $\tilde{P}_{tt}$ of the projection matrix $\tilde{P}$ is asymptotically negligible. At the same time, the average of these diagonal elements is equal to $K/T$, and thus, this condition can hold only when we have a moderately large number of regressors ($K/T \to 0$). When the number of regressors is moderately large, both challenges described at the beginning of this section are asymptotically negligible. The standard errors suggested by Theorem 5 look like the usual IV-type standard errors with one important change, they use the newly proposed estimator of variance $\hat{\sigma}^2(\hat{\Gamma})$.

## 4.2 Inference with Gaussian errors

**Theorem 6.** *Suppose Assumption 1 holds and* $\varepsilon_1$ *is Gaussian conditionally on* $\tilde{X}$*. Assume that* $\Gamma$ *solves equation* (1) *and* $k\Gamma k < 1 \quad c$*. Then, as* $T \to 1$*,*

$$\frac{r^{\prime}\hat{\beta}^{\mathrm{IV}}(\Gamma) \quad r^{\prime}\beta}{\sqrt{\Sigma_T}} \Rightarrow N(0,1)$$

*where* $\Sigma_T$ *is measurable with respect to* $\tilde{X}$*. With probability asymptotically approaching one,* $\Sigma_T \quad (1+\psi)\hat{\sigma}^2(\Gamma)kr^{\prime}(X^{\prime}(I \quad \Gamma)X)^{-1}X^{\prime}(I \quad \Gamma)k^2$*, where* $\psi = \frac{j\,\mathrm{tr}(B^2)j}{\mathrm{tr}(B^{\prime}B)}$ *and* $B = D^{\prime}(I \quad \Gamma)\tilde{M}_{\Gamma}$.

Theorem 6 allows the number of regressors to grow proportionally with the sample size. The assumption that errors are Gaussian is used in several ways. First, it allows us to avoid imposing the Lindeberg-type negligibility condition, as any linear combination of errors with weights depending on $\tilde{X}$ is conditionally Gaussian in finite-samples. Secondly, it removes the need to estimate skewness and kurtosis.

However, in the case when $K$ grows proportionally to the sample size, the standard errors stated in Theorem 5 do not correctly account for the uncertainty induced by the quadratic form. The missing term in the asymptotic variance formula depends on the importance of the feedback and is hard to estimate. We instead provide an upper bound on the asymptotic variance that makes the confidence sets asymptotically valid but conservative. In our simu-

lation we noticed that the correction $\psi$ tends to be tiny and does not change the standard errors much. We advise the calculation of it as a robustness check.

# 5 Extension to multiple periods

In the previous section we maintained the assumption that the violation of strict exogeneity happens for one period only. Some results can be extended to feedback lasting a fixed finite number of periods.

**Assumption 2.** *(i) The observed regressors $x_t$ can be decomposed as $x_t = \tilde{x}_t + \sum_{\ell=1}^{L} \alpha_\ell \varepsilon_{t-\ell}$ where the $T \times K$ matrix $\tilde{X} = [\tilde{x}_1, \ldots, \tilde{x}_T]'$ has full rank.*

*(ii) The errors $\{\varepsilon_t\}_{t=1-L}^{T}$ are i.i.d. conditional on $\tilde{X}$, $\mathsf{E}[\varepsilon_t | \tilde{X}] = 0$, $c < \sigma^2 := \mathsf{E}[\varepsilon_t^2 | \tilde{X}] < C$, and $\mathsf{E}[\varepsilon_t^4 | \tilde{X}] < C$ almost surely.*

*(iii) The non-random vectors $\alpha_1, \ldots, \alpha_L, r \in \mathsf{R}^K$ satisfy $\alpha_\ell'(\frac{1}{T}\tilde{X}'\tilde{X})^{-1}\alpha_\ell = O_p(1)$ for $\ell = 1, \ldots, L$ and $r'(\frac{1}{T}\tilde{X}'\tilde{X})^{-1}r = O_p(1)$.*

*(iv) $L$ is fixed and $K/T < 1 - c$.*

Vectors $\alpha_\ell$ describe how the shock to the outcome variable affects the regressors $\ell$ periods later. The direction of the feedback may vary freely with the lag as the regressors differ in the speed of reaction/adjustment. However, we probably should expect that the size of the $\ell$-period after feedback measured as $\alpha_\ell'(\frac{1}{T}\tilde{X}'\tilde{X})^{-1}\alpha_\ell$ should become negligible for large enough $\ell$ in typical macroeconomic settings.

**Theorem 7.** *Suppose Assumption 2 holds, $\Gamma$ is $\tilde{X}$-measurable, and $\|\Gamma\| < 1 - c$. Then,*

$$r'\hat{\beta}^{\mathrm{IV}}(\Gamma) - r'\beta = \sigma^2 \sum_{\ell=1}^{L} r'\bar{S}_\Gamma^{-1}\alpha_\ell \operatorname{tr}\left[(D')^\ell (I - \Gamma)\tilde{M}_\Gamma\right] + o_p(1), \tag{3}$$

*where $\bar{S}_\Gamma = \tilde{X}'(I - \Gamma)\tilde{X} + \sigma^2 \sum_{j,\ell=1}^{L} \alpha_j \alpha_\ell' \operatorname{tr}\left[(D')^j (I - \Gamma)\tilde{M}_\Gamma D^\ell\right]$.*

Theorem 7 is a direct generalization of Part (i) of Theorem 3. A special case of $\Gamma = 0$ shows that the bias of the OLS is a linear combination of $L$ terms connected to the lower diagonal traces of the projection matrix $\tilde{M}$. Lower diagonal traces, $\operatorname{tr}[(D')^\ell \tilde{M}]$, correspond

to average measures of the regressors' auto-correlations of order $\ell$, and are expected to decay for stationary regressors. Combined with the expected decrease in the size of $\alpha_\ell$, the first few terms in the bias formula should capture most of the bias in stationary applications.

The bias formula (3) is derived for any IV-type estimator for an $\tilde{X}$-measurable $T \quad T$ matrix $\Gamma$ with $k\Gamma k < 1 \quad c$. Theorem 7 suggests that if $\Gamma$ solves a system of $L$ equations $\operatorname{tr}\left((D^{\emptyset})^{\ell}(I \quad \Gamma)\tilde{M}_{\Gamma}\right) = 0$ for $\ell = 1, ..., L$ then $r^{\emptyset}\hat{\beta}^{\mathrm{IV}}(\Gamma)$ is a consistent estimator. A natural suggestion is to search for $\Gamma$ in the class of matrices $\Gamma = \sum_{\ell=1}^{L} \gamma_\ell D^\ell$ choosing $L$ parameters $\gamma_\ell$ in such a way as to solve the system of equations. We leave the question of considering other classes of matrices as well as getting some guarantees for finding a solution to future research.

# 6 Simulations

The goal of this section is to assess the size of the OLS bias in a 'typical' regression using US macroeconomic data. We use the data set from Stock and Watson (2016) containing quarterly observations from 1964 to 2013 ($T = 200$) on 108 US macro indicators. This data set is largely similar to the McCracken and Ng (2020) FRED-QD data set. The data set includes a broad class of variables with diverse time series properties.

Many macro and financial indicators tend to be very persistent and may be integrated up to the second order (have one or two unit roots). A prevailing (but not uniformly accepted) practice is to transform all variables to stationary before running a regression in order to avoid issues of co-integration and near co-integration or biases related to persistent regressors (see Stambaugh (1999) on such biases). The way applied researchers transform variables to stationary or make decisions about variable stationarity varies widely across the literature with many such decisions based both on statistical tests and expert judgements. Given that a large fraction of regressions is aimed at business-cycle parameters and in order to not take a stand and to unify the pre-treatment of variables we apply Hamilton (2018) transformation to all variables in the data set. Specifically, for each variable we define its cyclical component to be a two-year-ahead forecast error to this variable based on a univariate AR(4) regression. According to Hamilton (2018), this filtering transforms all types of stationary and up to second order integrated variables into stationary ones and extracts their business cycle component.

For each $K$ in the set $\{5, 15, 25, \ldots, 85, 100\}$ we perform 100 experiments where we randomly draw $K$ distinct variables from the transformed data set, denote them $X_{\mathrm{r}}$ and an additional variable $y_{\mathrm{r}}$. We calculate through simulations, the biases and standard deviations of the OLS and of our proposed estimator (referred to as IV) for the linear contrast in the feedback direction in the regression of $y_{\mathrm{r}}$ on $X_{\mathrm{r}}$ under the assumption of one-period violation of strict exogeneity. For this, we simulate $N = 1000$ samples from a data generating process satisfying Assumption 1 that preserves the time series behavior of regressors and the feedback size/direction of the observed $(y_{\mathrm{r}}, X_{\mathrm{r}})$. Specifically, we use as true parameters the empirical OLS values $\beta = (X_{\mathrm{r}}^{0}X_{\mathrm{r}})^{\ 1}X_{\mathrm{r}}^{0}y_{\mathrm{r}}$, $\sigma^2 = e^{0}e/(T \quad K)$ for $e = y_{\mathrm{r}} \quad X_{\mathrm{r}}\beta$, and $\alpha = X_{\mathrm{r}}^{0}D^{0}e/(e^{0}e)$. We simulate samples as $X = X_{\mathrm{r}} + D^{0}\varepsilon\alpha^{0}$ and $y = X\beta + \varepsilon$ where $\varepsilon \quad N(0, \sigma^2 I)$. We calculate the bias and standard deviation for the OLS and IV estimators of the linear contrast with $r = \alpha$.

The left panel of Figure 3 depicts the results of the experiments (for different $K$) that fall into the 10th percentile of the OLS bias. For those experiments, we report the OLS bias and standard deviation and the IV bias and standard deviation in the feedback direction alone along with the normalized lower trace of $M_{\mathrm{r}} = I \quad X_{\mathrm{r}}(X_{\mathrm{r}}^{0}X_{\mathrm{r}})^{\ 1}X_{\mathrm{r}}^{0}$, that is $\mathrm{tr}(D^{0}M_{\mathrm{r}})/T$. The right panel contains the results of the experiments that fall into the 90th percentile of the bias.

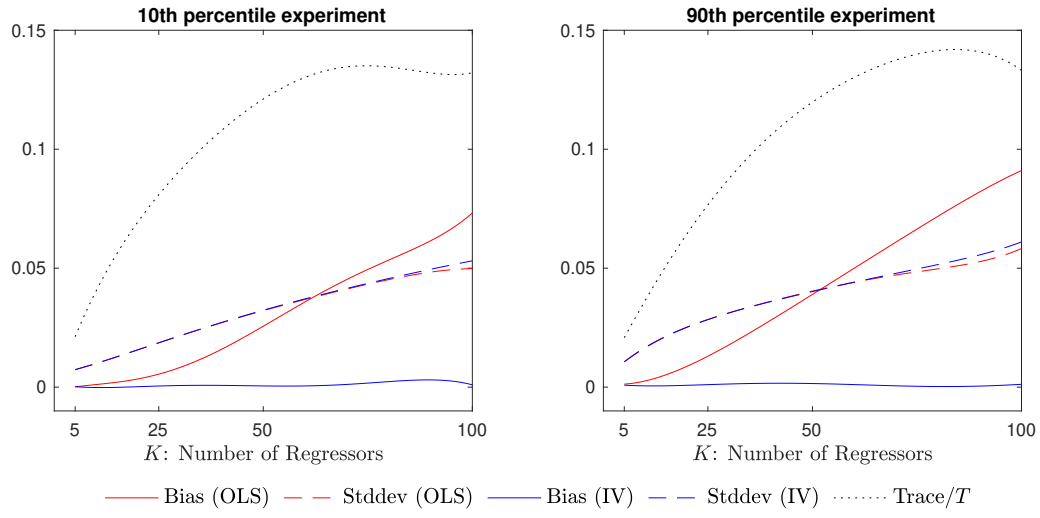


Figure 3: Absolute Bias and Standard Deviation of OLS and IV with $T = 200$

Figure 3 shows that a typical macroeconomic data set demonstrates enough time series

dependence or short-term linear predictability that it creates a potential for significant OLS biases. A typical size/direction of one-period feedback for a randomly picked-up regression using macro indicators is such that for a sample with 200 time periods in a regression with 25 regressors the OLS bias in the feedback direction equals about half of the standard deviation and is approximately equal to the standard deviation when the number of regressors is 50. This leads to invalid statistical inferences when relying on the OLS. Figure 4 reports the size distortions in the experiments described above for the 5% tests about the linear contrast in the direction of the feedback. Our new proposed estimator completely corrects the bias without any significant change to the standard deviation and restores the correct size for statistical tests.

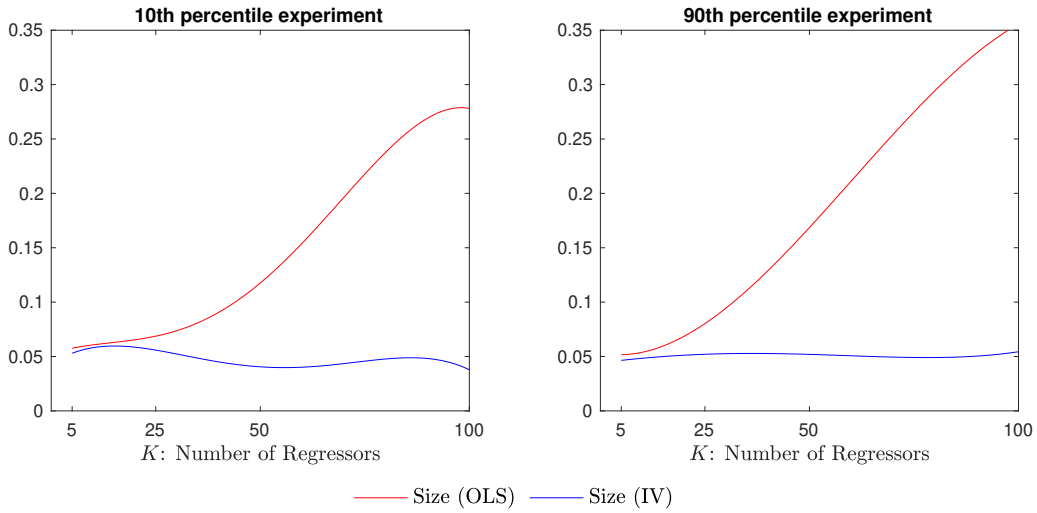


Figure 4: Size of Nominal 5% two-sided tests using OLS and IV with $T = 200$

Another observation from Figure 3 is that the ratio of the lower trace of the regressor projection matrix to the sample size is highly indicative of the size of the worst bias. Applied researchers should be worried when this indicator exceeds 5–10%. It is worth pointing out that in all of our experiments and simulations we encountered no problems of finding the solution to equation (2), which supports our assertion that the sample size requirement imposed in Lemma 1 is sufficient but not necessary for the existence of the solution.

Finally, we note that the results presented in Figure 3 are both qualitatively and quantitatively similar to the left panels of Figures 1 and 6. The difference between figures is that in Figures 1 and 6 we simulated artificial regressors following a simple AR(1) or MA(1) process.

Our theoretical results are quite agnostic about the time series properties of the regressors, specifically, we make no assumptions about stationarity or origin of the regressors. The data generating processes underlying Figure 3 mimics the time series behavior and feedback size/direction of a 'typical' macroeconomic application to the best of our ability.

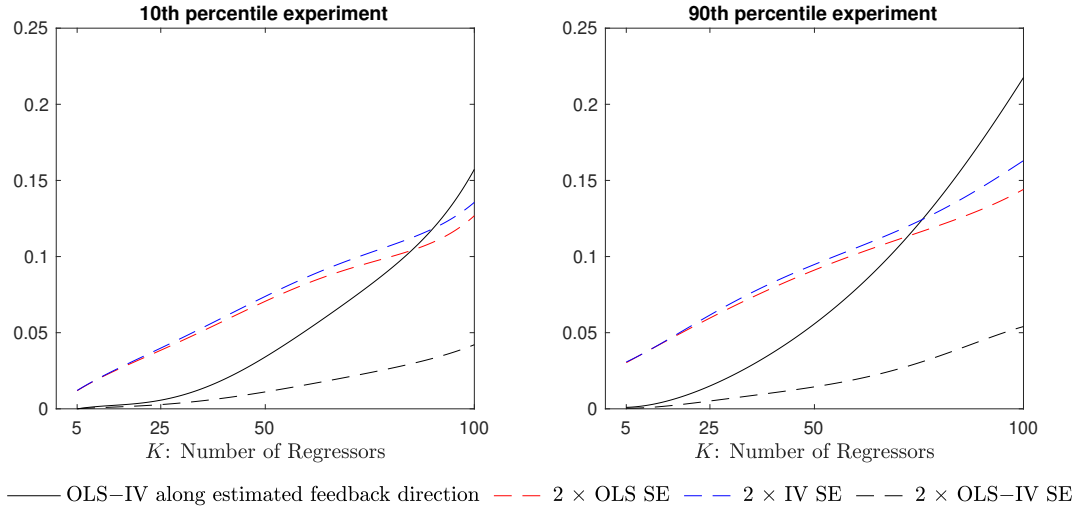


Figure 5: Difference between IV and OLS along estimated feedback direction, $T = 200$

Figure 5 answers the question of how different the results of the OLS and our new proposed estimators are in a randomly selected regression based on typical macroeconomic data. As in the experiments described above we randomly select $(y_{\mathrm{r}}, X_{\mathrm{r}})$ but rather than evaluate the theoretical bias in simulations we calculate the realization of the difference between two estimators for the contrast in the estimated feedback direction. We report the experiment corresponding to 10th percentile of the absolute difference on the left panel and 90th percentile on the right. For the described experiments we report the absolute value of the difference between the OLS and the IV estimators, double of the OLS and IV standard errors as well as the double standard deviations of the difference. We report doubled standard errors to directly relate the results to the corresponding $t$-statistics.

As we see from Figure 5, the difference between the OLS and IV estimators is statistically significant in all experiments we report. This indicates that the observed difference between estimators are mostly due to the biases and cannot be explained by randomness in realizations. An alternative explanation is that a hypothesis of no feedback is rejected in almost all regressions we considered. This is a sign of prevalence of feedback mechanisms in

most macroeconomic applications. Comparing the difference between the two estimators to the doubled standard errors shows that when the number of regressors is above 70 the IV estimate falls outside of the standard OLS confidence set and the OLS estimate falls outside the IV confidence set. This demonstrates the potential for drastic disagreement between the two estimation methods. Finally, we notice that the standard deviation of the OLS versus IV difference is much smaller than the standard error of either the OLS or the IV estimators, in some cases more than five times smaller. This shows that stochastic deviations in two estimators are greatly aligned, and that most of the difference between the two estimators comes from the bias, not the variance.

Table 1: Statistical significance of the differences in OLS and IV coefficients

| K | | 5 | 15 | 25 | 35 | 45 | 55 | 65 | 75 | 85 | 100 |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 100 | ave($t_\Delta > 1.96$) | 21.20 | 13.93 | 11.72 | 11.03 | 10.47 | 8.93 | 8.12 | 7.64 | 6.91 | 6.41 |

NOTE: $t_\Delta = j\hat{\beta}^{\mathrm{LS}} \quad \hat{\beta}^{\mathrm{IV}} j / se(\hat{\beta}^{\mathrm{LS}} \quad \hat{\beta}^{\mathrm{IV}})$. The average is taken over all coefficients in 100 randomly chosen models for each value of $K$.

While Figure 5 reports the difference between the two estimators in the most affected direction, one may also ask how different coefficients on individual regressors are. The average bias of individual coefficients is a counter-play of two forces. On one side, a larger number of regressors leads to a larger lower diagonal trace $\mathrm{tr}(D^0 M_{\mathrm{r}})/T$ and thus to a larger bias in the most affected direction. At the same time, when the dimensionality of regressors is large, (a randomly selected) feedback direction is on average less aligned with any coordinate direction. Thus, the same size of the worst bias results in a smaller average individual coefficient bias when the number of regressors increases, as it spreads out among many individual coefficients.

In Table 1, we report the average fraction of coefficients that display a statistically significant difference between the OLS and the IV estimators. The average is taken over the $K$ coefficients in the regression and over the 100 random regressions with $K$ regressors that we draw from the macroeconomic data base described above. While the fraction of regressors with a statistically significant difference between the OLS and the IV estimators is declining with larger $K$, the absolute number of such regressors increases. We conclude that while most directions/coefficients are immune to the biases, a non-trivial fraction of coefficients are significantly affected.

# References


Anatolyev, S. (2019). Many instruments and/or regressors: A friendly guide. *Journal of Economic Surveys 33* (2), 689–726.

Arellano, M. and S. Bond (1991). Some tests of specification for panel data: Monte Carlo evidence and an application to employment equations. *The Review of Economic Studies 58* (2), 277–297.

Bao, Y. and A. Ullah (2007). The second-order bias and mean squared error of estimators in time-series models. *Journal of Econometrics 140* (2), 650–669.

Bauwens, L., G. Chevillon, and S. Laurent (2023). We modeled long memory with just one lag! *Journal of Econometrics 236* (1), 105467.

Chamberlain, G. (1982). The general equivalence of Granger and Sims causality. *Econometrica: Journal of the Econometric Society 50* (3), 569–581.

Chao, J. C., N. R. Swanson, J. A. Hausman, W. K. Newey, and T. Woutersen (2012). Asymptotic distribution of JIVE in a heteroskedastic IV regression with many instruments. *Econometric Theory 28* (1), 42–86.

Chevillon, G., A. Hecq, and S. Laurent (2018). Generating univariate fractional integration within a large VAR(1). *Journal of Econometrics 204* (1), 54–65.

Engle, R. F., D. F. Hendry, and J.-F. Richard (1983). Exogeneity. *Econometrica: Journal of the Econometric Society 51* (2), 277–304.

Granger, C. W. (1969). Investigating causal relations by econometric models and cross-spectral methods. *Econometrica: Journal of the Econometric Society 37* (3), 424–438.

Gupta, A. and M. H. Seo (2023). Robust inference on infinite and growing dimensional time series regression. *Econometrica: Journal of the Econometric Society*, 1333–1361.

Hamilton, J. D. (1994). *Time series analysis*. Princeton University Press.

Hamilton, J. D. (2018). Why you should never use the Hodrick-Prescott filter. *Review of Economics and Statistics 100* (5), 831–843.

Hansen, B. E. and K. D. West (2002). Generalized method of moments and macroeconomics. *Journal of Business & Economic Statistics 20* (4), 460–469.

Hansen, C., J. Hausman, and W. Newey (2008). Estimation with many instrumental variables. *Journal of Business & Economic Statistics 26* (4), 398–422.

Jordà, Ò. (2005). Estimation and inference of impulse responses by local projections. *Amer-*

*ican Economic Review 95*(1), 161–182.

Kline, P., R. Saggio, and M. Sølvsten (2020). Leave-out estimation of variance components. *Econometrica: Journal of the Econometric Society 88*(5), 1859–1898.

Koenker, R. and J. A. Machado (1999). GMM inference when the number of moment conditions is large. *Journal of Econometrics 93*(2), 327–344.

McCracken, M. W. and S. Ng (2020). FRED-QD: A quarterly database for macroeconomic research. *Federal Reserve Bank of St. Louis* (Working Paper 2020-005).

Montiel Olea, J. L. and M. Plagborg-Møller (2021). Local projection inference is simpler and more robust than you think. *Econometrica: Journal of the Econometric Society 89*(4), 1789–1823.

Nickell, S. (1981). Biases in dynamic models with fixed effects. *Econometrica: Journal of the Econometric Society 49*(6), 1417–1426.

Schennach, S. M. (2018). Long memory via networking. *Econometrica 86*(6), 2221–2248.

Sims, C. A. (1972). Money, income, and causality. *The American Economic Review 62*(4), 540–552.

Sølvsten, M. (2020). Robust estimation with many instruments. *Journal of Econometrics 214*(2), 495–512.

Stambaugh, R. F. (1999). Predictive regressions. *Journal of Financial Economics 54*(3), 375–421.

Stock, J. H. and M. W. Watson (2016). Dynamic factor models, factor-augmented vector autoregressions, and structural vector autoregressions in macroeconomics. In *Handbook of Macroeconomics*, Volume 2, pp. 415–525. Elsevier.

Stock, J. H. and M. W. Watson (2019). *Introduction to econometrics, 4th edition*. Pearson.

Wallis, K. F. (1977). Multiple time series analysis and the final form of econometric models. *Econometrica: Journal of the Econometric Society 45*(6), 1481–1497.

Zellner, A. and F. Palm (1974). Time series analysis and simultaneous equation econometric models. *Journal of Econometrics 2*(1), 17–54.

# Appendix A Proofs

**Notation** For brevity, the expectation $\mathsf{E}[\,]$ is used to denote the conditional expectation $\mathsf{E}[\,j\tilde{X}]$. The proofs use some well known identities involving matrix traces. For matrices of compatible dimensions, $\mathrm{tr}(ABD) = \mathrm{tr}(BDA)$, $\mathrm{tr}(A) = \mathrm{tr}(A^{\prime}) = \mathrm{tr}(A + A^{\prime})/2$. As in the main paper, let $P = I \quad M = X(X^{\prime}X) \ {}^{1}X^{\prime}$ and $P_{\Gamma} = I \quad M_{\Gamma} = X(X^{\prime}(I \quad \Gamma)X) \ {}^{1}X^{\prime}(I \quad \Gamma)$.

## A.1 Auxilliary lemmas

**Lemma 2.** *(i) For a symmetric matrix $A$ and positive semi-de nite (psd) matrix $B$, we have the bounds $\lambda_{\min}(A)\,\mathrm{tr}(B) \quad \mathrm{tr}(AB) \quad \lambda_{\max}(A)\,\mathrm{tr}(B)$.*

*(ii) For any square matrix $A$, let $\lambda$ be the smallest eigenvalue of $\frac{A+A^{\prime}}{2}$. If $\lambda > 0$, then $kA \ {}^{1}k \quad 1/\lambda$.*

*(iii) For any compatible matrices $A$ and $B$, we have $j\mathrm{tr}(AB)j \quad kAk_{F}kBk_{F}$.*

*Proof of Lemma 2.* (i) As $A$ is symmetric, there exists $U$ with $U^{\prime}U = I$ such that $D = UAU^{\prime}$ is a diagonal matrix with eigenvalues of $A$ along its diagonal. Let $F = UBU^{\prime}$. Note that $F_{ii} \quad 0$ since $B$ is psd and $\mathrm{tr}(B) = \mathrm{tr}(F) = \sum_i F_{ii}$. Now,

$$\mathrm{tr}(AB) = \mathrm{tr}(UAU^{\prime}UBU^{\prime}) = \sum\nolimits_i D_{ii}F_{ii} \quad \max_i D_{ii} \sum\nolimits_i F_{ii} = \lambda_{\max}(A)\,\mathrm{tr}(B),$$

and $\sum_i D_{ii}F_{ii} \quad \min_i D_{ii} \sum_i F_{ii} = \lambda_{\min}(A)\,\mathrm{tr}(B)$. (ii) Now, $\lambda kxk^2 \quad x^{\prime}(A+A^{\prime})x/2 = x^{\prime}Ax \quad kxkkAxk$. As $\lambda > 0$, it follows that $Ax \neq 0$ when $x \neq 0$, so $A$ is invertible. For $x \neq 0$, we then have $\lambda \quad kAxk/kxk = kyk/kA \ {}^{1}yk$ where $y = Ax$. Thus, $kA \ {}^{1}k = \sup_{y \neq 0} kA \ {}^{1}yk/kyk$ $1/\lambda$. (iii) $j\mathrm{tr}(AB)j = \left|\sum_{ij} A_{ij}B_{ij}\right| \quad \sqrt{\sum_{ij} A_{ij}^2}\sqrt{\sum_{ij} B_{ij}^2} = kAk_{F}kBk_{F}$. □

**Lemma 3.** *Suppose $X \ 2 \ \mathsf{R}^{T \ K}$ has full rank and $\Gamma \ 2 \ \mathsf{R}^{T \ T}$ has $k\Gamma k < 1 \quad c$. Then (i) the matrices $I \quad \Gamma$, $I \quad P\Gamma$, $X^{\prime}(I \quad \Gamma)X$ and $I + A_{\Gamma}M$ are invertible where $A_{\Gamma} = (I \quad \Gamma) \ {}^{1}\Gamma$;*

*(ii) the following identities hold:*

$$(I \quad \Gamma)M_\Gamma = M(I + A_\Gamma M) \quad ^1, \tag{4}$$

$$P_\Gamma = (I \quad P\Gamma) \quad ^1P(I \quad \Gamma), \tag{5}$$

$$M_\Gamma = (I \quad P\Gamma) \quad ^1M, \tag{6}$$

$$\hat{\beta}^{\text{IV}}(\Gamma) = (X^0X) \quad ^1X^0(I + A_\Gamma M) \quad ^1y. \tag{7}$$

*(iii) the matrix* $(I \quad \Gamma)M_\Gamma + M_\Gamma^0(I \quad \Gamma^0)$ *is positive semi-de nite and* $c < \frac{T \quad K_\Gamma}{T \quad K} < C$ *where* $T \quad K_\Gamma = \text{tr}[(I \quad \Gamma)M_\Gamma]$*;*

*Proof of Lemma 3.* (i) Lemma 2(ii), the triangle inequality, and $k\Gamma k < 1$ yields invertibility of $I \quad \Gamma$, $I \quad P\Gamma$, $X^0(I \quad \Gamma)X$, and $I \quad \Gamma P$. $I + A_\Gamma M$ is invertible since

$$I + A_\Gamma M = (I \quad \Gamma) \quad ^1(I \quad \Gamma + \Gamma M) = (I \quad \Gamma) \quad ^1(I \quad \Gamma P).$$

(ii) Next, we note that

$$\begin{aligned}\left(X^0X\right) \quad ^1X^0 &= \left(X^0X\right) \quad ^1\left[X^0(I \quad \Gamma)X\left(X^0(I \quad \Gamma)X\right) \quad ^1\right]X^0 \\ &= \left(X^0X\right) \quad ^1X^0(I \quad \Gamma)P_\Gamma(I \quad \Gamma) \quad ^1.\end{aligned} \tag{8}$$

Pre-multiplying (8) by $X$ and using $P_\Gamma = PP_\Gamma$ gives us $P = (I \quad P\Gamma)P_\Gamma(I \quad \Gamma) \quad ^1$ and hence (5). Therefore, we also have (6):

$$M_\Gamma = I \quad P_\Gamma = (I \quad P\Gamma) \quad ^1(I \quad P\Gamma \quad P(I \quad \Gamma)) = (I \quad P\Gamma) \quad ^1M.$$

Using the "push-through" identity $(I \quad P\Gamma) \quad ^1P = P(I \quad \Gamma P) \quad ^1$ on (5) similarly yields (4):

$$M_\Gamma = I \quad P(I + A_\Gamma M) \quad ^1 = (I + A_\Gamma)M(I + A_\Gamma M) \quad ^1 = (I \quad \Gamma) \quad ^1M(I + A_\Gamma M) \quad ^1.$$

Reusing that $P_\Gamma = P(I + A_\Gamma M) \quad ^1$ we get (7) from $\hat{\beta}^{\text{IV}}(\Gamma) = \left(X^0X\right) \quad ^1X^0P_\Gamma y$ and $XP = X$.
(iii) Positive semi-definiteness comes from (4):

$$(I \quad \Gamma)M_\Gamma + M_\Gamma^0(I \quad \Gamma^0) = (I + MA_\Gamma^0) \quad ^1M\left\{(I \quad \Gamma^0) \quad ^1 + (I \quad \Gamma) \quad ^1\right\}M(I + A_\Gamma M) \quad ^1$$

and observing that $(I \quad \Gamma^0)^{\ 1} + (I \quad \Gamma)^{\ 1}$ is psd. The rate condition follows from $\mathrm{tr}(M) = T \quad K$, Lemma 2(i),

$$T \quad K_\Gamma = \mathrm{tr}[M(I + A_\Gamma M)^{\ 1}] = \mathrm{tr}\left[\frac{(I+A_\Gamma M)^{\ 1}+(I+MA_\Gamma^0)^{\ 1}}{2} M\right],$$

and that the eigenvalues of $\frac{(I+A_\Gamma M)^{\ 1}+(I+MA_\Gamma^0)^{\ 1}}{2}$ are in $\frac{1 \quad k\Gamma k}{1+k\Gamma k}$ to $\frac{1+k\Gamma k}{1 \quad k\Gamma k}$ with $k\Gamma k < 1 \quad c$. □

**Lemma 4.** *Suppose* $y_t = x_t^0\beta + \varepsilon_t$, $x_t = \tilde{x}_t + \sum_{\ell=1}^L \alpha_\ell \varepsilon_{t\ \ell}$ *where the* $T \quad K$ *matrix* $\tilde{X} = [\tilde{x}_1, \ldots, \tilde{x}_T]^0$ *has full rank, and* $\alpha_1, \ldots, \alpha_L, r$ *are* $K \quad 1$ *vectors. Then, there exists an invertible* $K \quad K$ *matrix* $\Theta$ *mapping* $f\tilde{X}, r, \beta, f\alpha_l g_{l=1}^L g$ *to* $f\tilde{X}\Theta, \Theta^0 r, \Theta^{\ 1}\beta, f\Theta^0\alpha_l g_{l=1}^L g$, *that satis es:*

*(i)* $\Theta^0 r$ *and* $f\Theta^0\alpha_\ell g_{\ell=1}^L$ *are spanned by the rst* $L+1$ *basis vectors so that* $r^0\Theta = (r^0, \mathbf{0}_{K \quad L \quad 1}^0)$ *and* $\alpha_\ell^0\Theta = (\alpha^0{}_{,\ell}, \mathbf{0}_{K \quad L \quad 1}^0)$ *with* $r \ , \alpha \ _{,1}, \ldots, \alpha \ _{,L} \in \mathsf{R}^{L+1}$;

*(ii)* $\Theta^0\tilde{X}_1^0\tilde{X}_1\Theta/T = I_{L+1}$ *and* $\Theta^0\tilde{Z}_2^0\tilde{X}_1\Theta = 0$ *for* $\tilde{X} = [\tilde{X}_1, \tilde{X}_2]$, *and* $\tilde{Z} = [\tilde{Z}_1, \tilde{Z}_2] = (I \quad \Gamma^0)\tilde{X}\Theta$, *where* $\tilde{X}_1$ *and* $\tilde{Z}_1$ *each have* $L+1$ *columns.*

*Proof of Lemma 4.* Let $\Theta = \Theta_0^{\ 1}\Theta_1$, where $K \quad K$ matrix $\Theta_0$ is the symmetric square root of $[\tilde{X}_1, \tilde{Z}_2]^0[\tilde{X}_1, \tilde{Z}_2]/T$ and $\Theta_1 R_1$ is the QR decomposition of $\Theta_0^{\ 1}[r, \alpha_1, \ldots, \alpha_L]$. Specifically, $\Theta_1^0 = \Theta_1^{\ 1}$ is a $K \quad K$ matrix and $R_1$ is spanned by the first $L+1$ basis vectors. We then have that $\Theta^0[r, \alpha_1, \ldots, \alpha_L] = \Theta_1^0\Theta_0^{\ 1}[r, \alpha_1, \ldots, \alpha_L] = \Theta_1^0\Theta_1 R_1 = R_1$ while we also have $\Theta^0[\tilde{X}_1, \tilde{Z}_2]^0[\tilde{X}_1, \tilde{Z}_2]\Theta/T = \Theta_1^0\Theta_0^{\ 1}\Theta_0^2\Theta_0^{\ 1}\Theta_1 = I_K$. □

**Lemma 5.** *Suppose* $X$ *and* $Z$ *are* $T \quad K$ *matrices with* $Z^0X$ *invertible. Let* $X = [X_1, X_2]$ *and* $Z = [Z_1, Z_2]$, *where* $X_\ell$ *and* $Z_\ell$ *are* $T \quad K_\ell$ *with* $K_1 + K_2 = K$. *De ne the oblique projections* $P_Z = X(Z^0X)^{\ 1}Z^0 = I \quad M_Z$ *and* $P_2 = X_2(Z_2^0X_2)^{\ 1}Z_2^0 = I \quad M_2$.

*(i) (Generalized Frisch-Waugh-Lowell) If* $r = [r_1^0; \mathbf{0}_{K_2}]^0$ *with* $r_1 \in \mathsf{R}^{K_1}$, *then*

$$r^0(Z^0X)^{\ 1}Z^0 = r_1^0(Z_1^0M_2X_1)^{\ 1}Z_1^0M_2.$$

*(ii)* $M_Z = M_2 \quad M_2X_1(Z_1^0M_2X_1)^{\ 1}Z_1^0M_2$.

*(iii) If* $Z = (I \quad \Gamma^0)X$ *for* $T \quad T$ *matrix* $\Gamma$ *with* $k\Gamma k < 1 \quad c$ *and* $A$ *is a* $T \quad T$ *matrix, then*

$$|\mathrm{tr}(A(M_Z \quad M_2))| \quad CK_1\|A\|.$$

*Proof of Lemma 5.* Letting $\Delta = (Z_1^0 M_2 X_1)^{-1}$, the matrix block inversion formula gives

$$(Z^0 X)^{-1} = \begin{bmatrix} \Delta & \Delta Z_1^0 X_2 (Z_2^0 X_2)^{-1} \\ (Z_2^0 X_2)^{-1} Z_2^0 X_1 \Delta & (Z_2^0 X_2)^{-1} \{ I + Z_2^0 X_1 \Delta Z_1^0 X_2 \} \end{bmatrix} \tag{9}$$

(i) From (9) we have

$$r^0 (Z^0 X)^{-1} Z^0 = r_1^0 \Delta Z_1^0 \quad r_1^0 \Delta Z_1^0 X_2 (Z_2^0 X_2)^{-1} Z_2^0 = r_1^0 \Delta Z_1^0 M_2.$$

(ii) Denote $\delta = X_1 \Delta Z_1^0$. Using (9) above:

$$P_Z = X(Z^0 X)^{-1} Z^0 = \delta \quad \delta P_2 \quad P_2 \delta + P_2 + P_2 \delta P_2 = I \quad M_2 + M_2 \delta M_2.$$

(iii) We impose without loss of generality that $X_2^0 X_1 = 0$ and $X_1^0 X_1 = I_{K_1}$. This entails no loss since (4) and (6) yields $Z_2^0 M_2 = 0$ and $M_2 X_2 = 0$, which in turn implies that $(M_Z, M_2)$ is invariant under the transformation $[X_1, X_2] \mapsto [M X_1 (X_1^0 M X_1)^{-1/2}, X_2]$ where $M = I - X_2 (X_2^0 X_2)^{-1} X_2^0 = I - P$. From (ii), Lemma 2(iii), and $\|\Psi\|_F \leq \|\Psi\| \sqrt{\mathrm{rk}(\Psi)}$:

$$\begin{aligned} |\mathrm{tr}(A(M_\Gamma - M_2))| &= \left| \mathrm{tr} \left( Z_1^0 M_2 A M_2 X_1 (Z_1^0 M_2 X_1)^{-1} \right) \right| \\ &\leq K_1 \| Z_1^0 M_2 A M_2 X_1 \| \, \| (Z_1^0 M_2 X_1)^{-1} \| \\ &\leq K_1 \| I - \Gamma \| \, \| M_2 \|^2 \, \| A \| \, \| (Z_1^0 M_2 X_1)^{-1} \|. \end{aligned}$$

Equation (6) gives $M_2 = (I - P\Gamma)^{-1} M$ and therefore $\|M_2\| \leq \frac{1}{1 - \|\Gamma\|} < C$. Together with $X_2^0 X_1 = 0$, (6) also yields

$$Z_1^0 M_2 X_1 = X_1^0 (I - \Gamma)(I - P\Gamma)^{-1} M X_1 = X_1^0 (I - \Gamma)(I - P\Gamma)^{-1} X_1,$$

and therefore

$$\tfrac{1}{2} (Z_1^0 M_2 X_1 + X_1^0 M_2^0 Z_1) = \tfrac{1}{2} X_1^0 \left\{ (I - \Gamma)(I - P\Gamma)^{-1} + (I - \Gamma^0 P)^{-1} (I - \Gamma^0) \right\} X_1.$$

As $X_1^0 X_1 = I_{K_2}$, the eigenvalues of the last matrix are larger than $\frac{1 - \|\Gamma\|}{1 + \|\Gamma\|}$. Therefore, Lemma 2(ii) and $\|\Gamma\| < 1 - c$ implies that $\|(Z_1^0 M_2 X_1)^{-1}\| < C$. □

**Lemma 6.** *Suppose $\tilde{X}$ is a $T \times k$ matrix with $\tilde{X}^0 \tilde{X} / T = I_k$, $A$ is a $T \times T$ matrix that is $\tilde{X}$-measurable, $\alpha_0, \ldots, \alpha_k$ are non-random $k \times 1$ vectors with $\|\alpha_\ell\| < C$ for all $\ell$, and $\{\varepsilon_t\}_{t=1}^T \sim_{iid} (0, \sigma^2)$ independent of $\{\tilde{X}_t\}_{t=1}^T$. Then*

*are i.i.d. conditionally on* $\tilde{X}$ *with* $\mathsf{E}[\varepsilon_t \mid \tilde{X}] = 0$, $0 < \sigma^2 = \mathsf{E}[\varepsilon_t^2 \mid \tilde{X}] < C$ *and* $\mathsf{E}[\varepsilon_t^4 \mid \tilde{X}] < C$. *Let* $u_\ell = (\varepsilon_{1\ \ell}, \ldots, \varepsilon_{T\ \ell})'$. *Then, as* $T \to \infty$ *while* $k$ *is fixed: (i)* $\|T^{-1/2}\tilde{X}'A\sum_{\ell=0}^{k} u_\ell \alpha_\ell'\|_F = O_p(\|A\|)$; *(ii)* $\varepsilon'A\varepsilon - \mathsf{E}[\varepsilon'A\varepsilon \mid \tilde{X}] = O_p(\|A\|_F)$.

*Proof of Lemma 6.* (i) Note that:

$$\mathsf{E}\left\|\frac{\tilde{X}'A}{T}\sum_{\ell=0}^{k} u_\ell\alpha_\ell'\right\|_F^2 = \operatorname{tr}\left[\frac{\tilde{X}'\tilde{X}}{T}A\sum_{\ell,j=0}^{k}\mathsf{E}[u_\ell u_j']\alpha_\ell'\alpha_j A'\right] \le \operatorname{tr}\left[\frac{\tilde{X}'\tilde{X}}{T}\right]\left\|A\sum_{\ell,j=0}^{k}\mathsf{E}[u_\ell u_j']\alpha_\ell'\alpha_j A'\right\|$$
$$\le \|A\|^2\left(\sum_{\ell,j=0}^{k}|\alpha_\ell'\alpha_j|\ \left\|\mathsf{E}[u_\ell u_j']\right\|\right) \le Ck^3\|A\|^2.$$

The last inequality uses that $u_\ell$ and $u_j$ has $\mathsf{E}[u_\ell u_j'] = \sigma^2 D^{\ell-j}$ when $\ell > j$, $\mathsf{E}[u_\ell u_j'] = \sigma^2 (D')^{j-\ell}$ when $\ell < j$, and $\mathsf{E}[u_\ell u_\ell'] = \sigma^2 I$. In all cases, $\|\mathsf{E}[u_\ell u_j']\| \le \sigma^2$. (ii) We have

$$\varepsilon'A\varepsilon - \mathsf{E}[\varepsilon'A\varepsilon] = \sum_t\sum_{s\neq t} A_{ts}\varepsilon_t\varepsilon_s + \sum_t A_{tt}(\varepsilon_t^2 - \sigma^2).$$

In the summations above summands are correlated only when $\{t,s\} = \{t',s'\}$. Therefore

$$\mathsf{E}\left[\left(\varepsilon'A\varepsilon - \tilde{\mathsf{E}}[\varepsilon'A\varepsilon]\right)^2\right] = \sigma^4\sum_t\sum_{s\neq t}(A_{ts}A_{st} + A_{ts}^2) + \sum_t A_{tt}^2\tilde{\mathsf{E}}\left[(\varepsilon_t^2 - \sigma^2)^2\right]$$
$$C\left(\sum_t\sum_{s\neq t}(A_{ts}A_{st} + A_{ts}^2) + \sum_t A_{tt}^2\right) \le 2C\sum_{t,s}A_{ts}A_{st} = 2C\|A\|_F^2,$$

for $C = \max\left\{\sigma^4,\ \mathsf{E}[(\varepsilon_t^2 - \sigma^2)^2]\right\}$. We used that $|\sum_t\sum_{s\neq t}A_{ts}A_{st}| \le \sum_t\sum_{s\neq t}A_{ts}^2$. □

**Lemma 7.** *Suppose* $B = D'(I - \Gamma)M_\Gamma$ *where* $\Gamma = \gamma D$, $|\gamma| < 1$ - *c and* $\operatorname{tr}(B) = O(K)$. *Then, (i)* $\sum_t B_{tt}^2 = O(K)$; *(ii)* $\operatorname{tr}(B^2) = \sum_{t,s}B_{ts}B_{st} = O(K)$; *(iii)* $\frac{T-K}{\operatorname{tr}(B'B)} = \frac{T-K}{\sum_{s,t}B_{s,t}^2} = O(1)$.

*Proof of Lemma 7.* Due to Lemma 2(i) and equation (5) we have, for any matrix $A$,

$$|\operatorname{tr}(AP)| = \left|\operatorname{tr}\left(\frac{A+A'}{2}P\right)\right| \le \left\|\frac{A+A'}{2}\right\|\operatorname{tr}(P) \le \|A\|K; \tag{10}$$
$$|\operatorname{tr}(AP_\Gamma)| = |\operatorname{tr}((I-\Gamma)A(I-P\Gamma)^{-1}P)| \le \frac{1+\|\Gamma\|}{1-\|\Gamma\|}\|A\|K. \tag{11}$$

(i) $B = D'(I-\Gamma) - D'(I-\Gamma)P_\Gamma$, thus $B_{tt} = 0 - \gamma F_{tt}$, where $F = D'(I-\Gamma)P_\Gamma$. The

condition $\mathrm{tr}(B) = O(K)$ implies $\gamma T = \sum_t F_{tt} + O(K)$ and $\sum_t B_{tt}^2 = \sum_t F_{tt}^2$ $T\gamma^2 + O(K)$ $\sum_t F_{tt}^2 + O(K)$. Consider diagonal elements of the matrix $F = AP_\Gamma$ with $A = D^0(I$ $\Gamma)$ and $\|A\|$ $1 + |\gamma|$:

$$\sum_t F_{tt}^2 = \sum_t (AP_\Gamma)_{tt}^2 = \sum_t \left( \sum_s A_{ts} P_{\Gamma,st} \right)^2 \quad \sum_t \left\{ \sum_s A_{ts}^2 \sum_s P_{\Gamma,st}^2 \right\}$$
$$= \sum_t (AA^0)_{tt} (P_\Gamma P_\Gamma^0)_{tt} \quad \|A\|^2 \,\mathrm{tr}(P_\Gamma P_\Gamma^0) \quad CK.$$

(ii) Next we have

$$\mathrm{tr}(B^2) = \mathrm{tr}(D^0(I \quad \Gamma)(I \quad P_\Gamma) D^0(I \quad \Gamma)(I \quad P_\Gamma))$$
$$= \mathrm{tr}[D^0(I \quad \Gamma) D^0(I \quad \Gamma)] + \mathrm{tr}\big[D^0(I \quad \Gamma) \left\{ \quad 2D^0(I \quad \Gamma) + P_\Gamma D^0(I \quad \Gamma) \right\} P_\Gamma\big]. \quad (12)$$

Since $\sum_t B_{tt}^2$ $0$ reasoning above gives us that

$$\mathrm{tr}(D^0(I \quad \Gamma) D^0(I \quad \Gamma)) = T\gamma^2 \quad \sum_t F_{tt}^2 = O(K).$$

The second term in (12) is $O(K)$ due to (11). (iii) We use that for any matrix $A$, $\mathrm{tr}(A^0 A) = \sum_{t,s} A_{ts}^2$ $|j \sum_{t,s} A_{ts} A_{st}| = |j\,\mathrm{tr}(A^2)|j|$:

$$\mathrm{tr}(B^0 B) = \mathrm{tr}(M_\Gamma^0 (I \quad \Gamma^0)(I \quad \Gamma) M_\Gamma) \quad (1 \quad |j\gamma|j)^2 \,\mathrm{tr}(M_\Gamma^0 M_\Gamma)$$
$$(1 \quad |j\gamma|j)^2 \,\mathrm{tr}(M_\Gamma^2) = (1 \quad |j\gamma|j)^2 \,\mathrm{tr}(M_\Gamma) = (1 \quad |j\gamma|j)^2 (T \quad K). \qquad \square$$

**Lemma 8.** *Suppose that $\tilde{X}$ is a $T$ $K$ matrix of rank $K$ and $\Gamma = \gamma D$.*

*(i) Equation (1) holds if and only if $\gamma$ is a fixed point of the transformation $f$ given by*

$$f(\gamma) = \frac{\mathrm{tr}(D^0 \tilde{M}_\Gamma)}{T \quad K}. \quad (13)$$

*(ii) If $|\mathrm{tr}(D^0 \tilde{M})| \quad \mu^2 K$ and $K < T/(1 + (1+\mu)^2)$ for some $\mu \in [0,1]$, then $f$ is a contraction on $[\ -\mu, \mu]/(1+\mu)$ with Lipshitz constant strictly less than $\mu$.*

*Proof.* (i) Since $\Gamma = \gamma D$ we have $D^0(I \quad \gamma D)\tilde{M}_\Gamma = D^0 \tilde{M}_\Gamma \quad D^0 D\gamma \tilde{M}_\Gamma$. We can therefore

re-write (1) as: $\mathrm{tr}(D^{\prime}\tilde{M}_{\Gamma}) \quad \gamma(T \quad K) = 0$. This equation is solved if (and only if) $\gamma = f(\gamma)$. (ii) Equation (6) yields $\tilde{M}_{\Gamma} = \tilde{M} + \gamma\tilde{P}D\tilde{M}_{\Gamma}$ and $\boldsymbol{k}\tilde{M}_{\Gamma}\boldsymbol{k} \quad \frac{1}{1 \quad j\gamma j}$. Equation (10) gives $\boldsymbol{j}\,\mathrm{tr}(D^{\prime}\tilde{P}D\tilde{M}_{\Gamma})\boldsymbol{j} \quad K\boldsymbol{k}D\tilde{M}_{\Gamma}D^{\prime}\boldsymbol{k}$. Therefore,

$$j f(\gamma) j \quad \frac{j\,\mathrm{tr}(D^{\prime}\tilde{M})j}{T \quad K} + j\gamma j\frac{j\,\mathrm{tr}(D^{\prime}\tilde{P}D\tilde{M}_{\Gamma})j}{T \quad K} \quad \frac{\mu^2 K}{T \quad K} + j\gamma j\frac{K}{(T \quad K)(1 \quad j\gamma j)} < \frac{\mu}{1+\mu}$$

and using equation (7)

$$\frac{j f(\gamma_1) \quad f(\gamma_2) j}{j\gamma_1 \quad \gamma_2 j} = \frac{j\mathrm{tr}[D^{\prime}(\tilde{M}_{\Gamma_1} \quad \tilde{M}_{\Gamma_2})]j}{j\gamma_1 \quad \gamma_2 j(T \quad K)} = \frac{j\mathrm{tr}[D^{\prime}\tilde{P}D(I \quad \tilde{P}\Gamma_2) \quad {}^{1}\tilde{M}_{\Gamma_1}]j}{T \quad K}$$
$$\frac{K}{T \quad K}\frac{1}{1 \quad j\gamma_1 j}\frac{1}{1 \quad j\gamma_2 j} < \mu$$

where the strict inequalities use $K < T/(1+(1+\mu)^2)$ and $j\gamma j, j\gamma_1 j, j\gamma_2 j \quad \frac{\mu}{1+\mu}$. □

## A.2 Proofs for results stated in the main text

*Proof of Theorem 1.* Special case of Theorem 3(i) with $\Gamma = 0$. □

*Proof of Theorem 2.* Special case of Theorem 4 with $\Gamma = 0$. □

*Proof of Lemma 1.* (i) Special case of (ii) with $\mu = 1$ since $j\mathrm{tr}(D^{\prime}\tilde{M})j = j\mathrm{tr}(D^{\prime}\tilde{P})j \quad K$ follows from (10). (ii) Follows from Lemma 8 and the Banach fixed point theorem. □

*Proof of Theorem 3.* (i) Special case of Theorem 7 with $L = 1$. (ii) Let $f(\gamma)$ be as in equation (13) define its empirical analog $\hat{f}(\gamma) = \frac{tr(D^{\prime}M)}{T \quad K} + \gamma\frac{tr(D^{\prime}PDM_{\Gamma})}{T \quad K}$ where $\Gamma = \gamma D$. Since $K < T/5$, Lemma 8(ii) yields that both $f$ and $\hat{f}$ are contractions on $[\quad 1/2, 1/2]$ with contraction speed bounded by $\frac{1}{2}$ and therefore has unique fixed points $\gamma_0$ and $\hat{\gamma}$ by the Banach fixed point theorem. Furthermore, $j\hat{f}(\hat{\gamma}) \quad \hat{f}(\gamma_0)j \quad \frac{1}{2}j\hat{\gamma} \quad \gamma_0 j$. For any $\gamma$ we have:

$$\left|\hat{f}(\gamma) \quad f(\gamma)\right| \quad \frac{1}{T \quad K}\left(\left|\mathrm{tr}[D^{\prime}(P \quad \tilde{P})] + \gamma\,\mathrm{tr}\left[D^{\prime}PDM_{\Gamma} \quad D^{\prime}\tilde{P}D\tilde{M}_{\Gamma}\right]\right|\right).$$

Consider the transformation $\Theta$ of Lemma 4. Since the projections $P, \tilde{P}, M_{\Gamma}, \tilde{M}_{\Gamma}$ are invariant to linear transformations we may assume that Lemma 4((i) and (ii)) hold with $L = 1$ and $\Theta = I_K$. This implies that $M_2 = \tilde{M}_2$, where $M_2$ is defined as in Lemma 5 using $X$, $(I \quad \Gamma^{\prime})X$, and $K_1 = 2$, while $\tilde{M}_2$ is an analogously defined starting from $\tilde{X}$, $(I \quad \Gamma^{\prime})\tilde{X}$, and $K_1 = 2$.

Therefore, Lemma 5(iii) implies for any compatible matrix $A$ that

$$\left|\mathrm{tr}\left[A(M_\Gamma \quad \tilde{M}_\Gamma)\right]\right| \quad \left|\mathrm{tr}\left[A(M_\Gamma \quad M_2)\right]\right| + \left|\mathrm{tr}\left[A(\tilde{M}_\Gamma \quad M_2)\right]\right| \quad C\|A\|$$

A similar statement holds with $(P, \tilde{P})$ replacing $(M_\Gamma, \tilde{M}_\Gamma)$. Thus $\|\hat{f}(\gamma) \quad f(\gamma)\| \quad \frac{C}{T \quad K}$. As a result:

$$\|\hat{\gamma} \quad \gamma_0\| = \left|\hat{f}(\hat{\gamma}) \quad f(\gamma_0)\right| \quad \left|\hat{f}(\hat{\gamma}) \quad \hat{f}(\gamma_0)\right| + \left|\hat{f}(\gamma) \quad f(\gamma_0)\right| \quad \frac{1}{2}\|\hat{\gamma} \quad \gamma_0\| + \frac{C}{T \quad K}.$$

This implies $\|\hat{\gamma} \quad \gamma_0\| \quad \frac{1}{2}\frac{C}{T \quad K} = O_p(1/T)$. Equation (7) yields

$$\begin{aligned}\left|r^0(\hat{\beta}^{\mathrm{IV}}(\hat{\Gamma}) \quad \hat{\beta}^{\mathrm{IV}}(\Gamma_0))\right| \quad & \left\|(X^0X) \quad {}^{1/2}r\right\| \left\|(I + A_{\Gamma_0}M) \quad {}^1(A_{\hat{\Gamma}} \quad A_{\Gamma_0})M(I + A_{\hat{\Gamma}}M) \quad {}^1\varepsilon\right\| \\ & \left\|(X^0X) \quad {}^{1/2}r\right\| \left\|(I + A_{\Gamma_0}M) \quad {}^1\right\| \left\|A_{\hat{\Gamma}} \quad A_{\Gamma_0}\right\| \left\|(I + A_{\hat{\Gamma}}M) \quad {}^1\right\| \|\varepsilon\|,\end{aligned}$$

where $A_{\Gamma_0} = (I \quad \Gamma_0) \quad {}^1\Gamma_0$ and $A_{\hat{\Gamma}} = (I \quad \hat{\Gamma}) \quad {}^1\hat{\Gamma}$. We have $\|\varepsilon\| = O_p(\sqrt{p}/T)$ and $\|A_{\hat{\Gamma}} \quad A_{\Gamma_0}\| = O_p(\hat{\gamma} \quad \gamma_0) = O_p(1/T)$. By Assumption 1(iii) $r^0(X^0X) \quad {}^1r = O(1/T)$, while $\left\|(I + A_{\Gamma_0}M) \quad {}^1\right\|$ is uniformly bounded. Thus, $r^0\hat{\beta}^{\mathrm{IV}}(\hat{\Gamma}) \quad r^0\hat{\beta}^{\mathrm{IV}}(\Gamma_0) = O_p(T \quad {}^1) = o_p(1/\sqrt{p}/T)$. □

The proofs of Theorems 4-6 follow after the proof of Theorem 7 as they rely on results established therein.

*Proof of Theorem 7.* Define $R_{\Gamma,\ell} = \mathrm{tr}[(D^0)^\ell(I \quad \Gamma)\tilde{M}_\Gamma]$. Note that $\|r^0\bar{S} \quad {}^1\alpha_\ell, R_{\Gamma,\ell}\|_{\ell=1}^L$ and $r^0\hat{\beta}^{\mathrm{IV}}(\Gamma) \quad r^0\beta$ are invariant under the transformation $\Theta$ of Lemma 4. Thus, we may assume without loss of generality that Lemma 4((i) and (ii)) hold with $\Theta = I_K$ and $Z = (I \quad \Gamma^0)X$. Since $(\alpha_1, \ldots, \alpha_L)$ is spanned by the first $L+1$ basis vectors, we have $M_2 = \tilde{M}_2$, $X_2 = \tilde{X}_2$, and $Z_2 = \tilde{Z}_2$, where $M_2$, $X_2$, and $Z_2$ are defined as in Lemma 5 using $X$, $Z = (I \quad \Gamma^0)X$, and $K_1 = L+1$, while $\tilde{M}_2$, $\tilde{X}_2$, $\tilde{Z}_2$ are analogously defined starting from $\tilde{X}$, $\tilde{Z} = (I \quad \Gamma^0)\tilde{X}$, and $K_1 = L+1$. This also implies that $X_2$, $Z_2$, and $M_2$ are non-random conditionally on $\tilde{X}$. Lemma 5(i) now yields:

$$r^0\hat{\beta}^{\mathrm{IV}}(\Gamma) \quad r^0\beta = r^0(Z_1^0M_2X_1) \quad {}^1Z_1^0M_2\varepsilon. \tag{14}$$

Defining $\bar{S}_2 = \mathsf{E}\left[Z_1^0M_2X_1\|\tilde{X}\right] = \tilde{Z}_1^0\tilde{X}_1 + \sigma^2\sum_{j,\ell=1}^L \alpha \quad {}_j\alpha^0 \quad {}_\ell \mathrm{tr}[(D^0)^j(I \quad \Gamma)M_2D^\ell]$ and $R_{2,\ell} =$

$\mathrm{tr}[(D^{0})^{\ell}(I \quad \Gamma)M_2]$, we show as a **first** step that

$$r^{0}\hat{\beta}^{\mathrm{IV}}(\Gamma) \quad r^{0}\beta = \sigma^2 \sum_{\ell=1}^{L} r^{0}\,\bar{S}_2^{\ 1}\alpha_{\ ,\ell}R_{2,\ell} + o_p(1). \tag{15}$$

Equation (15) follows from equation (14) and the following statements proven below:

$$\left\|(Z_1^{0}M_2X_1 \quad \bar{S}_2)/T\right\|_F = O_p(\sqrt{1/T}), \tag{16}$$

$$\left\|(\bar{S}_2/T)^{\ 1}\right\| \quad \frac{1}{(1 \quad \|\Gamma\|)}, \tag{17}$$

$$\left\|(Z_1^{0}M_2\varepsilon \quad \sigma^2\sum\nolimits_{\ell=1}^{L}\alpha_{\ ,\ell}R_{2,\ell})/T\right\| = O_p(\sqrt{1/T}), \tag{18}$$

and $\|R_{2,\ell}\| \quad T$ and thus $\|\sum_{\ell=1}^{L}\alpha_{\ ,\ell}R_{2,\ell}\| = O(T)$. Let $u_\ell = (\varepsilon_{1\ \ell},\ldots,\varepsilon_{T\ \ell})^{0}$ as in Lemma 6. Equation (16) considers a $(L+1) \quad (L+1)$ matrix with mean of zero:

$$\begin{aligned} Z_1^{0}M_2X_1 \quad \bar{S}_2 &= \left(\tilde{X}_1 + \sum_{j=1}^{L} u_j\alpha^{0}_{\ ,j}\right)^{0}(I \quad \Gamma)M_2\left(\tilde{X}_1 + \sum_{\ell=1}^{L} u_\ell\alpha^{0}_{\ ,\ell}\right) \quad \bar{S}_2 \\ &= \sum_{\ell=1}^{L}\alpha_{\ ,\ell}u_\ell^{0}(I \quad \Gamma)M_2\tilde{X}_1 + \tilde{X}_1^{0}(I \quad \Gamma)M_2\sum_{\ell=1}^{L}u_\ell\alpha^{0}_{\ ,\ell} \\ &+ \sum_{j,\ell=1}^{L}\alpha_{\ ,\ell}\alpha^{0}_{\ ,j}\left[u_\ell^{0}(I \quad \Gamma)M_2u_j \quad \mathsf{E}[u_\ell^{0}(I \quad \Gamma)M_2u_j]\right]. \end{aligned}$$

Notice that for $A = (I \quad \Gamma)M_2$, we have $\|A\| \quad \frac{1+\|\Gamma\|}{1\ \|\Gamma\|}$ and $\|A\|_F = O(^{p}\overline{T})$. Thus, applying Lemma 6((i) and (ii)) we obtain (16). As Lemma 3(iii) yields that $B = (I \quad \Gamma)M_2 + M_2^{0}(I \quad \Gamma^{0})$ is a non-negative definite matrix, we have

$$\begin{aligned} x^{0}(\bar{S}_2 + \bar{S}_2^{0})x = x^{0}\left[\tilde{Z}_1^{0}\tilde{X}_1 + \tilde{X}_1^{0}\tilde{Z}_1\right]x + \sum_{j,\ell=1}^{L} x^{0}\alpha_{\ ,j}\mathsf{E}\left[u_j^{0}Bu_\ell\|\tilde{X}\right]\alpha^{0}_{\ ,\ell}x \\ x^{0}\tilde{X}_1^{0}\left[I \quad (\Gamma+\Gamma^{0})/2\right]\tilde{X}_1x \quad (1 \quad \|\Gamma\|)\|\tilde{X}_1x\|^2 = (1 \quad \|\Gamma\|)T\|x\|^2 \end{aligned}$$

Applying Lemma 2(ii) now implies (17). To prove (18), we note that

$$Z_1^{\prime} M_2 \varepsilon = \tilde{X}_1^{\prime}(I \quad \Gamma) M_2 \varepsilon + \sum_{\ell=1}^{L} \alpha_{\cdot,\ell} u_{\ell}^{\prime}(I \quad \Gamma) M_2 \varepsilon.$$

The first term is $O_p(\sqrt{p}\overline{T})$ due to Lemma 6(i) applied with $A = (I \quad \Gamma) M_2$ and $\alpha_\ell = 0$ for $\ell > 0$. As $\sigma^2 R_{2,\ell} = \mathsf{E}[u_\ell^{\prime}(I \quad \Gamma) M_2 \varepsilon]$, Lemma 6(ii) yields $\sum_{\ell=1}^{L} \alpha_{\cdot,\ell} \left[ u_\ell^{\prime}(I \quad \Gamma) M_2 \varepsilon \quad R_{2,\ell} \right] = O_p(\sqrt{p}\overline{T})$. This leads to (18).

As the **second** and final step of the proof, we show that (15) implies (3). The first difference is that (15) depends on $r^{\prime}$, $\alpha_{\cdot,\ell}$, and the $(L+1) \quad (L+1)$ matrix $\bar{S}_2$, while (3) is described using $r^{\prime} = (r^{\prime}, \mathbf{0}_{K \quad L \quad 1}^{\prime})$, $\alpha_\ell^{\prime} = (\alpha^{\prime}_{\cdot,\ell}, \mathbf{0}_{K \quad L \quad 1}^{\prime})$, and the $K \quad K$ matrix $\bar{S}_\Gamma$. The second difference between these two statements is that (15) employs the oblique projection $\tilde{M}_2$, which projects off the $T \quad (K \quad L \quad 1)$ matrix $\tilde{X}_2$, while (3) employs $\tilde{M}_\Gamma$, which projects off the full $T \quad K$ matrix of regressors $\tilde{X} = [\tilde{X}_1, \tilde{X}_2]$.

For the first of these discrepancies, from Lemma 4((i) and (ii)) we have that the lower left $(K \quad L \quad 1) \quad (L+1)$ blocks of $\alpha_\ell \alpha_j^{\prime}$ and $\tilde{Z}^{\prime}\tilde{X}$ are zero. Thus the upper left $(L+1) \quad (L+1)$ block of $\bar{S}_\Gamma{}^{1}$ is equal to the inverse of the upper left $(L+1) \quad (L+1)$ block of $\bar{S}_\Gamma$. Thus

$$r^{\prime} \bar{S}_\Gamma{}^{1} \alpha_\ell = r^{\prime} \left( \bar{S}_2 + \sigma^2 \sum_{j,\ell=1}^{L} \alpha_{\cdot,j} \alpha^{\prime}_{\cdot,\ell} \Delta_{j\ell} \right)^{1} \alpha_{\cdot,\ell},$$

where $\Delta_{j\ell} = \operatorname{tr}\left[ (D^{\prime})^j (I \quad \Gamma)(\tilde{M}_\Gamma \quad \tilde{M}_2) D^\ell \right]$. Now, Lemma 5(iii) gives $\sqrt{k}\Delta_{j\ell}\sqrt{k} \quad Ck D^j (D^{\prime})^i (I \quad \Gamma)k = O(1)$ so that $\sum_{\ell=1}^{L} (r^{\prime} \bar{S}_\Gamma{}^{1} \alpha_\ell \quad r^{\prime} \bar{S}_2{}^{1} \alpha_{\cdot,\ell}) R_{\Gamma,\ell} = o_p(1)$. For the second discrepancy, we have $R_{\Gamma,\ell} \quad R_{2,\ell} = \operatorname{tr}\left[ (D^{\prime})^\ell (I \quad \Gamma)(\tilde{M}_\Gamma \quad \tilde{M}_2) \right]$ and Lemma 5(iii) similarly yielding $\sqrt{k} R_{\Gamma,\ell} \quad R_{2,\ell} \sqrt{k} = O(1)$. Thus $\sum_{\ell=1}^{L} (R_{\Gamma,\ell} \quad R_{2,\ell}) r^{\prime} \bar{S}_2{}^{1} \alpha_{\cdot,\ell} = o_p(1)$. In conclusion, we have $\sum_{\ell=1}^{L} r^{\prime} \bar{S}_\Gamma{}^{1} \alpha_\ell R_{\Gamma,\ell} \quad r^{\prime} \bar{S}_2{}^{1} \alpha_{\cdot,\ell} R_{2,\ell} = o_p(1)$, so (15) implies (3) and Theorem 7 follows. □

*Proof of Theorem 4.* From equation (4) we have $\hat{\sigma}^2(\Gamma) = \frac{\varepsilon^{\prime}(I \quad \Gamma) M_\Gamma \varepsilon}{T \quad K_\Gamma}$. Note that $\sigma^2$ and $\hat{\sigma}^2(\Gamma)$ are invariant under the transformation $\Theta$ of Lemma 4. Thus, we may assume without loss of generality that Lemma 4((i) and (ii)) hold with $\Theta = I_K$. As in the proof of Theorem 7, we therefore have $M_2 = \tilde{M}_2$. Applying Lemma 5(ii) yields:

$$\hat{\sigma}^2(\Gamma) = \frac{\varepsilon^{\prime}(I \quad \Gamma)\tilde{M}_2 \varepsilon}{T \quad K_\Gamma} \quad \frac{1}{T \quad K_\Gamma} \varepsilon^{\prime}(I \quad \Gamma)\tilde{M}_2 X_1 (Z_1^{\prime}\tilde{M}_2 X_1)^{1} Z_1^{\prime}\tilde{M}_2 \varepsilon. \tag{19}$$

Lemma 6(ii) applied with $A = \frac{(I \quad \Gamma)M_2}{T \quad K_\Gamma}$ implies that $\frac{\varepsilon^{\prime}(I \quad \Gamma)M_2\varepsilon}{\sigma^2(T \quad K_\Gamma)} = 1 + o_p(1)$. For the second term in (19), we use statements (16)–(18) and the second part of the proof of Theorem 7 to obtain the conclusion of Theorem 4. □

*Proof of Theorem 5.* Let $\gamma_0$ be as in Theorem 3. Note that $r^{\prime}\hat{\beta}^{\mathrm{IV}}(\Gamma_0), r^{\prime}\beta$ and $\hat{\Sigma}_T(\Gamma_0) = \hat{\sigma}^2(\Gamma_0)\|r^{\prime}(X^{\prime}(I \quad \Gamma_0)X)^{-1}X^{\prime}(I \quad \Gamma_0)\|^2$ are invariant under the transformation $\Theta$ of Lemma 4. Thus, we may assume without loss of generality that Lemma 4((i) and (ii)) hold with $\Theta = I_K$ and $Z = (I \quad \Gamma_0^{\prime})X$. Following the proof of Theorem 7, we have formula (14), where the denominator satisfies (16) and (17). This implies that

$$r^{\prime}\hat{\beta}^{\mathrm{IV}}(\Gamma_0) \quad r^{\prime}\beta = \big(1 + o_p(1)\big)\left(\sum_t w_t\varepsilon_t + r^{\prime}\bar{S}_2^{-1}\alpha\ \varepsilon^{\prime}B\varepsilon\right), \tag{20}$$

where $w^{\prime} = (w_1, \ldots, w_T) = r^{\prime}\bar{S}_2^{-1}\tilde{Z}_1^{\prime}M_2$, $B = D^{\prime}(I \quad \Gamma_0)M_2$, and $(\bar{S}_2, M_2)$ is defined as in the proof of Theorem 7 with $\Gamma = \Gamma_0$. The weights $\{w_t\}$ and the matrix $B$ are measurable with respect to $\tilde{X}$. Below, we show that

$$\frac{r^{\prime}\hat{\beta}^{\mathrm{IV}}(\Gamma_0) \quad r^{\prime}\beta}{\sqrt{\Sigma_T}} = \frac{\sum_t w_t\varepsilon_t + r^{\prime}\bar{S}_2^{-1}\alpha\ \varepsilon^{\prime}B\varepsilon}{\sqrt{\Sigma_T}} \ \to\ N(0,1), \tag{21}$$

where $\Sigma_T = \sigma^2\sum_t w_t^2 + \sigma^4(r^{\prime}\bar{S}_2^{-1}\alpha\ )^2\,\mathrm{tr}(B^2 + B^{\prime}B)$. We have

$$\varepsilon^{\prime}B\varepsilon = \sum_t B_{tt}\varepsilon_t^2 + \sum_t\sum_{s\neq t}\frac{B_{st} + B_{ts}}{2}\varepsilon_t\varepsilon_s.$$

Lemma 7((i) and (iii)) together with $K/T \ \to\ 0$ imply that $(r^{\prime}\bar{S}_2^{-1}\alpha\ )^2\left(\sum_t B_{tt}\varepsilon_t^2\right)/\sqrt{\Sigma_T} \xrightarrow{p} 0$ and $\frac{\mathrm{tr}(B^2)}{\mathrm{tr}(B^{\prime}B)} \ \to\ 0$.

We obtain statement (21) by establishing the four conditions, (i)–(iv), of Sølvsten (2020), Corollary A2.8, located in the Supplementary Appendix of that article. Condition (i) is automatically satisfied if we define $w_{t,T} = \frac{w_t}{\sqrt{\Sigma_T}}$, $M_{st} = \frac{r^{\prime}\bar{S}_2^{-1}\alpha}{2\sqrt{\Sigma_T}}(B_{st} + B_{ts})$ for $s \neq t$, and $M_{tt} = 0$. Condition (iv) is implied by Assumption 1(ii). To establish condition (ii) we note that by Lemma 5(i), we have for any $\tilde{r}$ of the form $\tilde{r}^{\prime} = (\tilde{r}^{\prime}, \mathbf{0}_{K \quad L}^{\prime}\ _1)$ that

$$\tilde{r}^{\prime}(\tilde{Z}^{\prime}\tilde{X})^{-1}\tilde{Z}^{\prime} = \tilde{r}^{\prime}(\tilde{Z}_1^{\prime}M_2\tilde{X}_1)^{-1}\tilde{Z}_1^{\prime}M_2$$

where we will use that $M_2 = \tilde{M}_2$ since $X_2 = \tilde{X}_2$ is strictly exogenous. Thus for the specific choice of $\tilde{r}$ where $\tilde{r}^0 = r^0 \bar{S}_2{}^1(\tilde{Z}_1^0 M_2 \tilde{X}_1)$ we have

$$w^0 = r^0 \bar{S}_2{}^1 \tilde{Z}_1^0 M_2 = \tilde{r}^0 (\tilde{Z}_1^0 M_2 \tilde{X}_1)\ {}^1 \tilde{Z}_1^0 M_2 = \tilde{r}^0 (\tilde{Z}^0 \tilde{X})\ {}^1 \tilde{Z}^0$$

so that $w_t = \tilde{r}^0 (\tilde{Z}^0 \tilde{X})\ {}^1 (\tilde{X}_t \quad \gamma_0 \tilde{X}_{t+1})$. Note, that

$$\max_t |w_t| \quad \left\| (\tilde{X}^0 \tilde{X})^{1/2} (\tilde{X}^0 \tilde{Z})\ {}^1 \tilde{r} \right\| (1 + |\gamma_0|) \max_t \left\| (\tilde{X}^0 \tilde{X})\ {}^{1/2} \tilde{X}_t \right\|,$$

$$\sum_t w_t^2 = \tilde{r}^0 (\tilde{Z}^0 \tilde{X})\ {}^1 \tilde{X}^0 (I \quad \Gamma_0)(I \quad \Gamma_0^0) \tilde{X} (\tilde{X}^0 \tilde{Z})\ {}^1 \tilde{r} \quad (1 \quad |\gamma_0|)^2 \left\| (\tilde{X}^0 \tilde{X})^{1/2} (\tilde{X}^0 \tilde{Z})\ {}^1 \tilde{r} \right\|^2.$$

Thus,

$$\max_t |w_{t,T}| \quad \frac{\max_t |w_t|}{\sqrt{\sum_t w_t^2}} \quad \frac{1 + |\gamma_0|}{1 \quad |\gamma_0|} \max_t \| (\tilde{X}^0 \tilde{X})\ {}^{1/2} \tilde{X}_t \| \ \to\ 0.$$

For condition (iii), we note that Lemma 7 and $K/T \to 0$ yields

$$\sum_s \sum_{t \neq s} \left[ \frac{B_{st} + B_{ts}}{2} \right]^2 = \frac{1}{2} \operatorname{tr}(B^2 + B^0 B)(1 + o(1)).$$

This yields $\sum_s \sum_{t \neq s} M_{st}^2 \quad 1$. Furthermore,

$$\left\| \frac{B + B^0}{2} \quad diag(B) \right\| \quad \|B\| + \max_t |B_{tt}| = O(1).$$

Therefore we have $\|(M_{st})_{s,t}\| \to 0$ and have therefore established (21).

Finally, we prove that $\frac{\hat{\Sigma}_T}{\Sigma_T} \xrightarrow{p} 1$. Reusing the argument in the proof of Theorem 3, we first have that $(\hat{\Sigma}_T \quad \hat{\Sigma}_T(\Gamma_0))/\Sigma_T = o(1)$. By Lemma 5(ii), we have for $u = (\varepsilon_0, \ldots, \varepsilon_{T\ 1})^0$ that

$$\begin{aligned} \hat{\Sigma}_T(\Gamma_0) &= \hat{\sigma}^2(\Gamma_0) \|r^0 (Z_1^0 M_2 X_1)\ {}^1 Z_1^0 M_2\|^2 = [1 + o_p(1)] \sigma^2 r^0 \bar{S}_2\ {}^1 Z_1^0 M_2 M_2^0 Z_1 (\bar{S}_2^0)\ {}^1 r \\ &= [1 + o_p(1)] \sigma^2 r^0 \bar{S}_2\ {}^1 (\tilde{Z}_1^0 + \alpha\ u(I \quad \Gamma_0^0)) M_2 M_2^0 (\tilde{Z}_1 + (I \quad \Gamma_0^0) u \alpha^0)(\bar{S}_2^0)\ {}^1 r\ , \end{aligned}$$

where we also used Theorem 4 and (16). From Lemma 7((i) and (iii)), we have

$$\mathsf{E}[u^0 (I \quad \Gamma_0^0) M_2 M_2^0 (I \quad \Gamma_0) u] = [1 + o(1)] \mathsf{E}[\varepsilon^0 B B^0 \varepsilon] = [1 + o(1)] \sigma^2 \operatorname{tr}[B^0 B].$$

From $K/T \to 0$ we have $\frac{\mathrm{tr}(B^2+B^{0\prime}B)}{\mathrm{tr}(B^{0\prime}B)} \to 1$ and therefore

$$\frac{\hat{\Sigma}_T - \Sigma_T}{\Sigma_T} = \frac{2\sigma^2 r^{0\prime}\bar{S}_2^{-1}\alpha\; u(I - \Gamma_0)M_2M_2^0\tilde{Z}_1 + \sigma^2(r^{0\prime}\bar{S}_2^{-1}\alpha)^2(\varepsilon^{0\prime}BB^0\varepsilon - \mathsf{E}\varepsilon^{0\prime}BB^0\varepsilon)}{\Sigma_T} + o_p(1). \quad (22)$$

Define $R = M_2^0\tilde{Z}_1$, $\xi_1 = r^{0\prime}\bar{S}_2^{-1}\alpha\; u(I - \Gamma_0)M_2M_2^0\tilde{Z}_1$, $\xi_2 = (r^{0\prime}\bar{S}_2^{-1}\alpha)^2[\varepsilon^{0\prime}BB^0\varepsilon - \mathsf{E}\varepsilon^{0\prime}BB^0\varepsilon]$. Now,

$$\begin{aligned}
\mathsf{E}[\xi_1^2] &= C(r^{0\prime}\bar{S}_2^{-1}\alpha)^2\,\mathrm{tr}(R^{0\prime}B^{0\prime}BR) \le C(r^{0\prime}\bar{S}_2^{-1}\alpha)^2\,\mathrm{tr}(R^{0\prime}R)\|B^{0\prime}B\|; \\
\frac{\xi_1}{\Sigma_T} &= O_p\left(\frac{r^{0\prime}\bar{S}_2^{-1}\alpha\;\sqrt{\mathrm{tr}(R^{0\prime}R)\|B^{0\prime}B\|}}{\Sigma_T}\right) \\
&= O_p\left(\sqrt{\frac{\|B^{0\prime}B\|}{\mathrm{tr}(B^{0\prime}B)}}\frac{(r^{0\prime}\bar{S}_2^{-1}\alpha)^2\,\mathrm{tr}(B^{0\prime}B) + \mathrm{tr}(R^{0\prime}R)}{\Sigma_T}\right) = O_p\left(\sqrt{\frac{\|B^{0\prime}B\|}{\mathrm{tr}(B^{0\prime}B)}}\right).
\end{aligned}$$

Lemma 6(ii) yields $\varepsilon^{0\prime}BB^0\varepsilon - \mathsf{E}\varepsilon^{0\prime}BB^0\varepsilon = O_P(\|B^{0\prime}B\|_F)$. Thus

$$\frac{\xi_2}{\Sigma_T} = O_p\left(\frac{\|B^{0\prime}B\|_F}{\mathrm{tr}(B^{0\prime}B)}\right) = O_p\left(\frac{\sqrt{\mathrm{tr}(B^{0\prime}B)\|B^{0\prime}B\|}}{\mathrm{tr}(B^{0\prime}B)}\right) = O_p\left(\frac{\|B\|}{\sqrt{\mathrm{tr}(B^{0\prime}B)}}\right).$$

Thus, by Lemma 7(iii), both terms in (22) are $O_p(1/\sqrt{T})$. □

*Proof of Theorem 6.* Note that $r^{0\prime}\hat{\beta}^{\mathrm{IV}}(\Gamma), r^{0\prime}\beta$ and $\hat{\Sigma}_T(\Gamma) = \hat{\sigma}^2(\Gamma)\|r^{0\prime}(X^{0\prime}(I - \Gamma)X)^{-1}X^{0\prime}(I - \Gamma)\|^2$ are invariant under the transformation $\Theta$ of Lemma 4. Thus, we may assume without loss of generality that Lemma 4((i) and (ii)) hold with $\Theta = I_K$ and $Z = (I - \Gamma)X$. Proceeding as in the proof of Theorem 5 we arrive at equation (20) with $B = D^{0\prime}(I - \Gamma)M_2$. Due to Gaussianity of the errors, $r^{0\prime}\bar{S}_2^{-1}\tilde{Z}_1^{0\prime}M_2\varepsilon = w^{0\prime}\varepsilon$ has a Gaussian distribution conditionally on $\tilde{X}$ with conditional variance $\sigma^2\|w\|^2 = \sigma^2\|r^{0\prime}\bar{S}_2^{-1}\tilde{Z}_1^{0\prime}M_2\|^2$. Below we show that

$$\frac{\varepsilon^{0\prime}B\varepsilon}{\sigma^2\sqrt{\mathrm{tr}(B^2) + \mathrm{tr}(B^{0\prime}B)}} \Rightarrow N(0,1) \quad (23)$$

and that this term is asymptotically independent from the conditionally Gaussian term $w^{0\prime}\varepsilon$.

Define $P_w = \frac{ww^{0\prime}}{w^{0\prime}w}$ and $M_w = I - P_w$ then $\varepsilon^{0\prime}B\varepsilon = 2\varepsilon^{0\prime}BP_w\varepsilon - \varepsilon^{0\prime}P_wBP_w\varepsilon + \varepsilon^{0\prime}M_wBM_w\varepsilon$. We

notice that

$$\left|\varepsilon' P_w B P_w \varepsilon\right| = \left(\frac{w'\varepsilon}{\|w\|}\right)^2 \left|\frac{w' B w}{w' w}\right| \quad \|B\| \quad \chi_1^2 = O_p(1);$$
$$\varepsilon' B P_w \varepsilon = \frac{w'\varepsilon}{\|w\|}\frac{w' B \varepsilon}{\|w\|} = \frac{w'\varepsilon}{\|w\|} N(0, \frac{w' B B' w}{w' w}) = O_p(1).$$

Due to Lemma 7 we have $\frac{T}{\mathrm{tr}(B^2)+\mathrm{tr}(B'B)} = O(1)$. Thus

$$\frac{\varepsilon' B \varepsilon}{\sqrt{\mathrm{tr}(B^2)+\mathrm{tr}(B'B)}} = \frac{\varepsilon' M_w B M_w \varepsilon}{\sqrt{\mathrm{tr}(B^2)+\mathrm{tr}(B'B)}} + o_p(1)$$

But $\not{p}\frac{\varepsilon' M_w B M_w \varepsilon}{\mathrm{tr}(B^2)+\mathrm{tr}(B'B)}$ is independent from $\frac{w'\varepsilon}{\|w\|} \quad N(0,\sigma^2)$. This implies that $\not{p}\frac{\varepsilon' B \varepsilon}{\mathrm{tr}(B^2)+\mathrm{tr}(B'B)}$ is asymptotically independent from the first term. Since $\mathrm{tr}(B) = \sum_t B_{tt} = O_p(1)$:

$$\varepsilon' B \varepsilon = \sum_t \sum_{s \neq t} \frac{B_{ts}+B_{st}}{2}\varepsilon_t \varepsilon_s + \sum_t B_{tt}(\varepsilon_t^2 \quad \sigma^2) + O_p(1).$$

Using that $\mathsf{E}\varepsilon_t^4 = 3\sigma^4$, one can show that

$$Var(\varepsilon' B \varepsilon) = 2\sigma^4 \sum_t \sum_{s \neq t} \left(\frac{B_{ts}+B_{st}}{2}\right)^2 + 2\sigma^4 \sum_t B_{tt}^2 = \mathrm{tr}(B^2) + \mathrm{tr}(B'B),$$

and due to Lemma 7 the right-hand-side grows no slower than of order $T$. Since $\max_t |B_{t,t}| \|B\| = O(1)$ and the operator norm of $B + B'$ is bounded, conditions (i)–(iv) of Corollary A2.8 in Sølvsten (2020) hold. Therefore (23) holds and we have asymptotic Gaussianity.

By the same argument as in the proof of Theorem 5, we can show that

$$\frac{\hat{\sigma}^2(\Gamma)\|r'(X'(I \quad \Gamma)X)^{-1}X'(I \quad \Gamma)\|^2}{\sigma^2\|r'\bar{S}_2^{-1}\tilde{Z}_1' M_2\|^2 + \sigma^4(r'\bar{S}_2^{-1}\alpha)^2\,\mathrm{tr}(B'B)} \xrightarrow{p} 1.$$

Pre-multiplying this estimator by $1+\psi = \frac{|\mathrm{tr}(B^2)|+\mathrm{tr}(B'B)}{\mathrm{tr}(B'B)}$ guarantees that the resulting quantity asymptotically weakly exceeds $\Sigma_T = \sigma^2\sum_t w_t^2 + \sigma^4(r'\bar{S}_2^{-1}\alpha)^2\,\mathrm{tr}(B^2+B'B)$. □

# Appendix B Additional Simulations

The outcome vector is generated as $y = X\beta + \varepsilon$ with $\varepsilon \quad N(0, I)$ and $\beta = 0$. The design matrix is generated as $x_{1,t} = \tilde{x}_{1,t} + a\varepsilon_{t\ 1}$ and $X_{\ 1} = \tilde{X}_{\ 1}$, where $\tilde{X}$ is generated as a rotated MA(1) process with $\tilde{X}\tilde{X}^{0}/T = I_K$, independent from $\varepsilon$. Specifically, we generate $v_t = \rho u_{t\ 1} + u_t$ with $\{u_t\}_{t=1}^{T}$ *i.i.d.* $N(0, I_K)$ and define $\tilde{X} = V(V^{0}V/T)^{\ 1/2}$, where the square root comes from Cholesky decomposition. Across simulations, we fix the sample size at $T = 200$ and the coefficient on the feedback mechanism at $a = 1.5$. Simulation results are summarized in Figure 6 with the left panel showing results for number of regressors $K$ between 4 and 150 (fixing $\rho$ at 0.8). The right panel reports the results for the autocorrelation in regressors $\rho$ between 0 and 0.98 (fixing $K$ at 50). We report simulated values of absolute bias and standard deviation for the first coordinate of OLS and IV together with the mean absolute value of the ratio of the lower trace of $M$ to the sample size. The results are extremely similar to the results reported in Section 2 both in term of the size of the bias/standard deviations as well as dependence on the number of regressors and their one-period predictability.

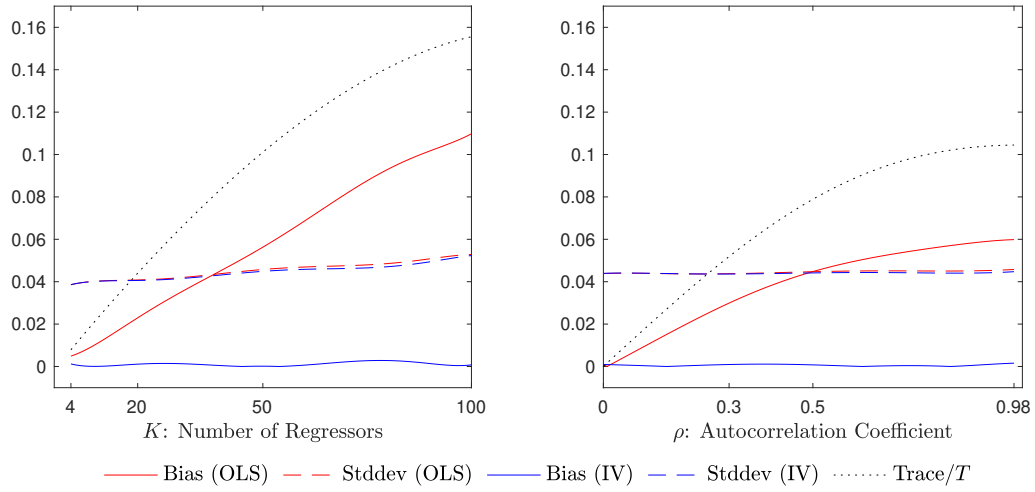


Figure 6: Absolute Bias and Standard Deviation of OLS and IV with T=200