Linear Regression with Weak Exogeneity*

Anna Mikusheva[†], Mikkel Sølvsten[‡]

February 2025

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 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 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. 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, Ulrich Mueller, Morten Ø. Nielsen, Jack Porter, and Jim Stock for useful discussions. Harvey Barnhard, Baiyun Jing, 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 autocorrelate. 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). Carrasco and Rossi (2016) argue that using many predictors with rolling window (small sample size) guards against instability of forecasting ability in macro forecasting. In the undergraduate textbook by Stock and Watson (2019), the primary example in chapter 16 contains a sub-sample analysis that runs the ordinary least squares (OLS) regression with 32 regressors using 204 observations. See also Bauwens et al. (2023) for two applications.

We show that these two features — weak exogeneity and many autocorrelated regressors — can produce substantial asymptotic biases and even lead to inconsistency of the OLS estimator. A sizable asymptotic 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. The 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 are some statements that the biases are relatively small and of second order (see, e.g., Bao and Ullah, 2007). Our results show that the bias in OLS can be a

[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.

first-order issue when the number of regressors is a non-negligible fraction of the sample size, as is often the case empirically.

This paper contains several results. First, we derive a formula for the asymptotically non-negligible part of the OLS bias, explain which data features cause it, and provide a tool to assess the potential for OLS bias in a given time series application. Second, we propose a new estimator, which is asymptotically unbiased if the data features 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 decreased precision; the simulated standard deviations of OLS and our new estimator are very close, with no explicit ordering. Finally, we show how our new estimator generalizes when multiple periods of feedback effects violate strict exogeneity.

The asymptotic bias of OLS arises in a setting with many regressors because the normalized design matrix, $X'X/T$, remains asymptotically random even in large samples. Weak (but not strict) exogeneity allows randomness in the design matrix to correlate with the OLS estimator's numerator, leading to an asymptotic bias. Specifying the feedback structure lets us derive a formula for the leading term of the bias. Specifically, assuming strict exogeneity is violated by one-period feedback from the outcome variable to the next period's regressors, we show that the OLS asymptotic bias aligns with the feedback direction. The asymptotic bias size increases with the number of regressors and their one-period ahead linear predictability.

We propose a new estimator that 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 central 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 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 essential feature of our bias correction is that it relies solely on knowledge of the regressors. The correction is identical for any outcome variable and requires no knowledge or estimation of the feedback direction. We show that the new estimator is consistent and, after normalization, asymptotically Gaussian under a one-period violation of strict exogeneity.

The main results can be generalized to the case where strict exogeneity is violated for a finite number of periods, i.e., when the outcome variable has feedback effects on the regressors for L periods. In such a case, the asymptotic bias of OLS contains L terms corresponding to the L feedback directions. Our results are also generalizable to a case of infinite, fast-decaying feedback, which covers vector autoregression (VAR) equation-by-equation estimation and regressions with a lagged dependent variable as a regressor. VAR and Augmented Distributed Lag (ADL) regressions tend to include very many regressors for modest sample sizes; thus, the bias issue is potentially very severe there. We should note that our results do not cover local projection regressions, as those have a special structure of the error terms.

We conduct a simulation study to assess how common and large the OLS bias is in typical macroeconomic regressions. We take an extensive collection of US macroeconomic indexes observed quarterly for 200 periods, extract their business cycle part, and randomly draw a regression from this data set. We show that the time series dependence and feedback magnitude typical for these macroeconomic data produce empirically relevant OLS biases. For example, in a typical regression with 25 regressors, we find a bias for the contrast along the feedback direction equal to half of the standard deviation. In comparison, in regression with 50 regressors, this bias approximately equals one standard deviation. Depending on the number of regressors, approximately 6–21 percent of coefficients display a statistically significant difference between OLS and our proposed IV-type estimator. Our estimator fully corrects the bias.

Our results relate to three distinct strands of literature. First, there is a classic literature on linear equations in time series. Sawa (1978) derived the exact bias of OLS in a simple autoregression with Gaussian errors, noting "the least squares estimate is seriously biased for the sample of two-digits sizes." In a setting with few regressors, Kiviet et al. (1999); Bao and Ullah (2007) derive second-order bias formulas (or bias of order $1/T$) where the lagged outcome variable is the only regressor violating strict exogeneity. Stambaugh (1999), who considered a regression model with a very persistent (near-unit-root) regressor, also raised a concern that weak exogeneity may lead to substantial biases in OLS. Hansen and West (2002) shows the inappropriateness of the Generalized Least Squares (GLS) estimator in linear models with weak exogeneity as GLS mixes the timing of observations and leads to substantial biases.

The second literature is that on the 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 asymptotic bias issues, as well as some asymptotic statements related to Gaussianity of quadratic forms, are connected to the problems arising in linear models with many instruments and/or many regressors – see, e.g., 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 the asymptotic Gaussianity of the newly proposed estimator and suggests a valid inference procedure. Section 5 extends some of these 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. The Supplemental Appendix contains additional theoretical results and simulations referenced in the text.

Notation For any vector x, $\|x\|^2 = x'x$. For any matrix A, $\mathrm{rk}(A)$ is the rank of A, $\|A\| = \sup_x \|Ax\|/\|x\|$ is the operator norm, and $\|A\|_F = \sqrt{\mathrm{tr}(A'A)}$ is the Frobenius norm. We let $c < 1$ and C be strictly positive and finite (generic) constants that do not depend on the sample size T and may differ across appearances. The $q \times q$ 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:

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

Weak exogeneity is considerably less restrictive than strict exogeneity, which assumes that $\mathbb{E}[\varepsilon_t \,|\, X] = 0$, with X denoting the $T \times K$ set of all regressors (including past, present, and future). In the context of time series economic data, strict exogeneity is rarely plausible. In contrast, weak exogeneity allows the structural shock to influence future regressor values through feedback. Such feedback is very likely as variables employed in macro estimation often evolve as a joint dynamic process and affect each other either instantaneously (as x_t affects y_t) or with some lag. When the regressors incorporate lagged outcome variables, violations of strict exogeneity are certain. For further discussion regarding the plausibility of weak and strict exogeneity in various applications, see Stock and Watson (2019, ch. 16).

Standard arguments for unbiasedness of the OLS estimator $r'\hat{\beta}^{\mathrm{OLS}} = r'(X'X)^{-1}X'y$ rely heavily on strict exogeneity. Although it is well-known that OLS exhibits bias under weak exogeneity (see, e.g., Hamilton, 1994, ch. 8.2), it is mostly ignored in the literature. Here, we claim that the bias of OLS can be substantial, potentially leading to inconsistency.

Standard statements regarding OLS consistency in time series hinge on the assumption that the normalized design matrix concentrates around its expectation: $\|X'X/T - Q\| \xrightarrow{p} 0$, where Q is non-singular (see, e.g., Hamilton, 1994, Assumption 8.6). For example, Gupta and Seo (2023) assumes $K^3/T \to 0$ 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 the design matrix dimensionality) grows fast enough. Under strict exogeneity, the randomness of $X'X/T$ is not problematic 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, $X'X/T$ is not just a random matrix; it also correlates with $X'y/T$. This correlation produces a bias that may persist in large samples.

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 ε_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 $\mathbb{E}[\varepsilon_t|\tilde{X}] = 0$, $c < \sigma^2 := \mathbb{E}[\varepsilon_t^2|\tilde{X}] < C$, and $\mathbb{E}[\varepsilon_t^4|\tilde{X}] < C$ almost surely.

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

(iv) The number of regressors K may diverge with sample size T such that $K/T < 1 - c$.

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. Section 5 generalizes this assumption to allow for multi-period feedback, and it can be generalized further to infinite-order feedback as long as the feedback magnitude decays sufficiently fast.[2] Such a generalization covers the case with a lagged outcome variable used as a regressor as it leads to an infinite, geometrically decaying feedback. The usual VAR model is a special case of this. Linear feedback is a natural starting point. It arises if, for example, one assumes all regressors and errors are jointly Gaussian (an assumption that originally motivated the OLS estimator). Furthermore, the vast majority of models estimated using macroeconomic data are linear due to the limited length of time series data.

The assumptions on the strictly exogenous part of the regressors $\{\tilde{x}_t\}_{t=1}^T$ are kept to a minimum, the central part of which is full rank. Specifically, we do not assume stationarity or impose moment conditions on $\tilde{X}$. Such generality is possible primarily because our analysis is done conditionally on $\tilde{X}$. One could impose additional assumptions on the evolution of $\tilde{x}_t$ and potentially use them to improve the efficiency of estimators. However, such assumptions increase the potential for misspecification and severely reduce the applicability of the results.

Part (ii) is a standard set of assumptions on the error terms in homoskedastic regression models. Allowing for autocorrelation of error terms is a possible generalization, though we leave it to future research. For such a generalization, one should focus on a distinction between feedback and autocorrelation. Specifically, one could decompose the regression error

[2] The results are available from the authors upon request.

into a sum of two parts: one that produces feedback but no autocorrelation (as in the current paper) and one that autocorrelates but does not produce any feedback. The feedback leads to a bias in OLS, while the autocorrelated part does not cause bias but necessitates different formulae for standard errors of linear estimators.

Part (iii) imposes a loose bound on the magnitudes of α and r relative to the scaled design matrix $\tilde{X}'\tilde{X}/T$. This condition is, for instance, satisfied if $\|\alpha\|$ and $\|r\|$ are bounded and the smallest eigenvalue of $\tilde{X}'\tilde{X}/T$ is separated from zero. In a cross-sectional setting, if all elements of $\tilde{X}$ are i.i.d. with mean zero and finite fourth moment, the smallest eigenvalue of $\tilde{X}'\tilde{X}/T$ converges almost surely to the lower bound of support in Marchenko-Pastur's law and thus is separated from zero. Generalizations to time series appear in Hachem et al. (2016). Part (iv) accommodates a widely ranging amount of regressors, including a small fixed number. It implies degrees of freedom $T-K$ diverges to infinity as 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 in the OLS estimator.

2.2 Parameter estimator

We consider a special case to provide intuition for how and why the OLS bias arises. Suppose, therefore, that only the first regressor fails to be strictly exogenous. That is, the first regressor experiences one-period feedback: $x_{1t} = \tilde{x}_{1t} + a\varepsilon_{t-1}$. All other regressors only contain a strictly exogenous component: $x_{kt} = \tilde{x}_{kt}$ for $2 \leq k \leq K$. In matrix form, this case lets us write $X = (X_1, X_{-1})$ and $\tilde{X} = (\tilde{X}_1, \tilde{X}_{-1})$ with X_1 the $T \times 1$ vector with elements x_{1t} and, crucially, $X_{-1} = \tilde{X}_{-1}$. Furthermore, we take $\tilde{X}$ as fixed and suppose that $\tilde{X}'\tilde{X}/T = I_K$.

In this special case, we now derive the asymptotic bias of the OLS estimator for the first coefficient β_1. According to the Frisch-Waugh-Lovell theorem, we have

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

where the partialling-out operator $M_{-1} = I - \tilde{X}_{-1}(\tilde{X}_{-1}'\tilde{X}_{-1})^{-1}\tilde{X}_{-1}'$ 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 ε. The denominator of $\hat{\beta}_1^{\text{OLS}} - \beta_1$ has a standard asymptotic behavior and, once divided by 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 correlates with ε. This dependence leads to a non-zero expectation of the numerator in $\hat{\beta}_1^{\text{OLS}} - \beta_1$:

$$\mathbb{E}\Big[X_1' M_{-1}\varepsilon \mid \tilde{X}\Big] = \mathbb{E}\Big[\tilde{X}_1' M_{-1}\varepsilon \mid \tilde{X}\Big] + a\mathbb{E}\Big[\sum\nolimits_{s,t} M^*_{st}\varepsilon_{s-1}\varepsilon_t \mid \tilde{X}\Big] = a\sigma^2 \sum\nolimits_t M^*_{tt-1}.$$

While the weakly exogenous regressor x_{1t} does not correlate with the contemporaneous error ε_t, partialling the remaining regressors out mixes the timing of the observations. It makes the partialled-out regressor correlated with the contemporaneous error. This derivation suggests a form for the leading term of the asymptotic bias of $\hat{\beta}_1^{\text{OLS}}$. The asymptotic 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} (namely, $\sum_t M^*_{tt-1}$). 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 $\theta = r'\beta$, the asymptotic bias depends on the weight placed on β_1.

The insight from the preceding special case generalizes. To see this, note that 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 $K \times K$ matrix with full rank that linearly transforms the regressors, then OLS serves as an estimator of $A^{-1}\beta$. To achieve the contrast θ, we should apply $A'r$ as the weighting. Any OLS scenario simplifies to the one discussed earlier when we choose $A = (\tilde{X}'\tilde{X}/T)^{-1/2}$ and select the square root to ensure that $A'\alpha$ is proportional to the first basis vector. These insights and certain technical derivations lead to the following characterization of the asymptotic bias in OLS.

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

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

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

Theorem 1 presents the leading term of the OLS asymptotic bias. We use the term *asymptotic bias* to describe a quantity that, if subtracted from the estimator, results in a consistent (often termed *asymptotically unbiased*) estimator. The asymptotic bias is a random variable since it depends on the exogenous part of the regressor $\tilde{X}$. If the asymptotic bias is non-negligible as the sample size increases, the OLS estimator is inconsistent. If it is asymptotically non-negligible relative to the OLS standard errors, standard t-statistic-based inferences (tests, confidence sets) have asymptotically incorrect size or coverage.

Note that the asymptotic bias formula uses the lower diagonal trace $\sum_t \tilde{M}_{tt-1}$ of the projection matrix orthogonal to $\tilde{X}$ instead of $\sum_t M^*_{tt-1}$ as in the special case. The following Lemma establishes the asymptotic negligibility of this difference.

Lemma 1. *Suppose Assumption 1, part (i), holds, and recall the notation for the three closely related projection matrices* $M = I_T - X(X'X)^{-1}X' = (M_{st})$, $\tilde{M} = I_T - \tilde{X}(\tilde{X}'\tilde{X})^{-1}\tilde{X}' = (\tilde{M}_{st})$, *and* $M_{-1} = I - \tilde{X}_{-1}(\tilde{X}'_{-1}\tilde{X}_{-1})^{-1}\tilde{X}'_{-1} = (M^*_{st})$. *Then,*

$$\left|\sum\nolimits_t \tilde{M}_{tt-1} - \sum\nolimits_t M_{tt-1}\right| \leq 2 \quad \textit{and} \quad \left|\sum\nolimits_t \tilde{M}_{tt-1} - \sum\nolimits_t M^*_{tt-1}\right| \leq 1.$$

It is worth discussing the size of the lower trace of the projection matrix and of the biases we may observe in applications. Consider the term $\sum_t \tilde{M}_{tt-1} = -\sum_t \tilde{P}_{tt-1}$, where $\tilde{P} = \tilde{X}(\tilde{X}'\tilde{X})^{-1}\tilde{X}'$. This quantity is unchanged under any full-rank rotation of the regressors. For simplicity, suppose here that the regressors satisfy $\tilde{X}'\tilde{X}/T = I_K$ (perhaps after suitable rotation). Then $\sum_t \tilde{M}_{tt-1} = -\frac{1}{T}\sum_t \tilde{x}'_t\tilde{x}_{t-1}$. Thus, the lower-diagonal trace of $\tilde{M}$ equals the negative sum of the sample first-order autocorrelation of the regressors and measures a linear connection between consecutive realizations of the K-dimensional regressor $\tilde{x}_t$ and $\tilde{x}_{t-1}$.

In time series settings, we often work with data containing autocorrelated regressors. Consequently, the average of the sample autocorrelations $\hat{\rho} = -\frac{1}{K}\sum_t \tilde{M}_{tt-1}$ can be nontrivial and even large. Deriving statements about the large-sample behavior of $\hat{\rho}$ is impossible without restrictive assumptions on the regressors. A key challenge is that commonly used regressors often exhibit highly heterogeneous persistence. On the technical side, a major difficulty arises because the matrix $\tilde{X}'\tilde{X}/T$ does not concentrate asymptotically. For instance, if all elements of $\tilde{X}$ are i.i.d. mean zero variables with finite fourth moment, and K/T is a constant fraction, the spectral density of $\tilde{X}'\tilde{X}/T$ converges to Marchenko-Pastur's law (Marchenko and Pastur, 1967); generalizations of this result to time series are largely absent.

As a proof of concept that $\hat{\rho}$ can remain asymptotically bounded away from zero, Supplemental Appendix A provides a time series with short dependence where the average population first-order autocorrelation equals 1/2 and $-\frac{1}{K}\sum_t \tilde{M}_{tt-1} \to^p 1/2$. In this example, the asymptotic bias in the worst direction is $\tau K/T$, where $\tau \neq 0$ is a constant determined by the primitives of the data-generating process. This example shows that the OLS estimator of $\alpha'\beta$ becomes inconsistent when K grows proportionally to the sample size, and standard t-statistic-based inferences become asymptotically invalid if K^2/T is separated from zero.

We observe that $\bar{S} \geq \tilde{X}'\tilde{X}$ which implies that Assumption 1(iii) leads to $r'\bar{S}^{-1}\alpha = O_p(1/T)$. Assume momentarily that $Tr'\bar{S}^{-1}\alpha$ converges to a constant r_α. The leading bias term then becomes approximately $-\sigma^2 r_\alpha \cdot \hat{\rho}K/T$. Notably, this potential bias is more pronounced for regressors with higher in-sample first-order autocorrelation and a larger number of regressors. Although analytical results about the size of $-\hat{\rho}K/T = \frac{1}{T}\sum_{t=2}^T \tilde{M}_{tt-1}$ are typically unavailable; the asymptotically equivalent quantity $\frac{1}{T}\sum_{t=2}^T M_{tt-1}$ 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. Simulation results reported below and in Section 6 show that the bias of OLS can be substantial both in simple examples and in data-generating processes calibrated to US macroeconomic data.

Lastly, the extent of asymptotic bias depends on the alignment between the contrast r and the feedback direction α, making the feedback direction the most affected contrast direction, captured by r_α above. Contrasts in all directions orthogonal to α (using the scalar product weighted by $\bar{S}^{-1}$) experience only negligible asymptotic bias. In our special case where only the first regressor is weakly exogenous, this observation corresponds to the OLS estimator of β_1 being the sole estimator with significant asymptotic bias. Thus, for any linear contrast $r'\beta$, the asymptotic bias depends on the weight placed on β_1. In applications, the feedback direction is unknown and challenging to estimate empirically, as α is a $K \times 1$ 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 $\varepsilon \sim N(0, I)$ 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 $\tilde{X}\tilde{X}'/T = I_K$, independent from ε. 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 $\tilde{X} = V(V'V/T)^{-1/2}$, 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 ρ at 0.8).[3] The right panel reports the results for the autocorrelation in regressors ρ 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 bias and standard deviations for the new estimator proposed in Section 3. Specifically, the label IV refers to the estimator defined by equations (1) and (3).

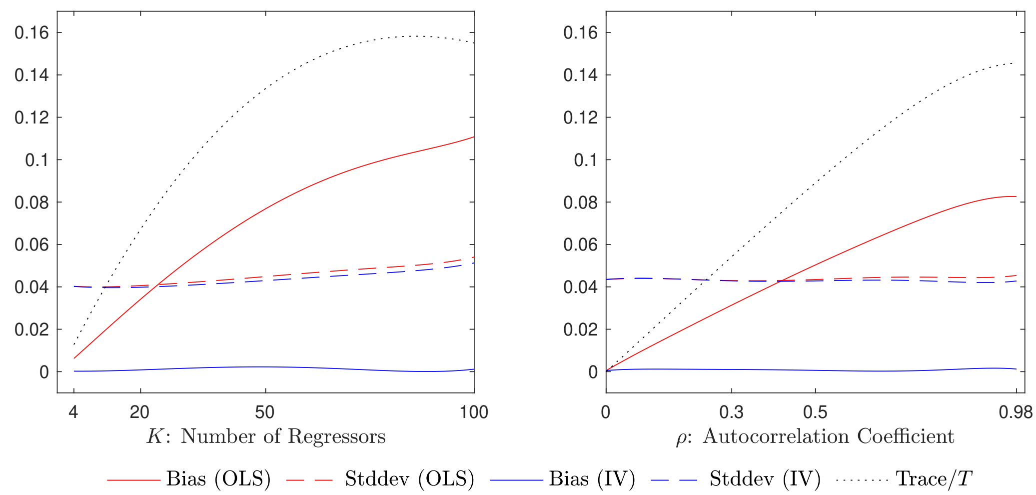

Figure 1: Absolute Bias and Standard Deviation of OLS and IV with $T = 200$

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 prevalent setting in macroeconomic applications. The observable lower trace of M divided by the sample size 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 autocorrelation of the regressors.[4]

[3]The results are presented as sixth-order polynomial fits to the actual results across K.

[4]One may worry that the observed biases are due to persistence in the regressors as the coefficient ρ in an AR(1) process measures both short-term dependence and long-run persistence. We also conduct simulations generating $V_t = \rho u_{t-1} + u_t$ as an MA(1) process. The Supplemental Appendix presents the results, which

2.3 Variance estimator

To make reliable statistical inferences in a linear regression, we usually need an estimator for the variance of the error term σ^2. In regression with many regressors and strict exogeneity, it is well-known that one must adjust appropriately for the degrees of freedom to correct for over-fitting. The most commonly applied OLS estimator of the error variance uses this adjustment $\hat{\sigma}^2 = y'My/(T-K)$. It is not apparent, ex-ante, whether this estimator retains consistency with many regressors and only weak exogeneity. On the one hand, a large asymptotic bias of the OLS estimator for coefficients raises concerns about the consistency of the variance estimator. On the other hand, asymptotic 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,*

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

Theorem 2 states that the OLS estimator of the error variance is asymptotically 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 also inconsistent. Despite this inconsistency result, the biases 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 $T \times T$ 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' is the lead operator, moving the time-series index forward by one. Notice that the lower trace

are essentially identical to those reported in Figure 1. The Supplemental Appendix also contains simulations with a larger sample size $T = 800$, which clearly traces the inconsistency of OLS when K is proportional to T.

appearing in the bias of OLS satisfies the representation $\sum_{t=2}^{T}\tilde{M}_{tt-1} = \mathrm{tr}(D'\tilde{M})$. Consider also a $T \times T$ matrix Γ measurable with respect to $\tilde{X}$ such that $\|\Gamma\| \le 1 - c$.

We propose an IV-inspired estimator relying on an endogenous instrument $Z = (I - \Gamma')X$:

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

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. To eliminate the asymptotic bias, we require that Γ solves the following (non-linear) one-dimensional equation:

$$\mathrm{tr}\big[D'(I-\Gamma)\tilde{M}_{\Gamma}\big] = 0 \tag{2}$$

where $\tilde{M}_{\Gamma} = I - \tilde{X}(\tilde{X}'(I-\Gamma)\tilde{X})^{-1}\tilde{X}'(I-\Gamma) = I - \tilde{X}(\tilde{Z}'\tilde{X})^{-1}\tilde{Z}'$ is an oblique projection off $\tilde{X}$ in the direction of $\tilde{Z} = (I-\Gamma')\tilde{X}$. 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 proven for a general Γ with the leading example of $\Gamma = \gamma D$ for $|\gamma| < 1 - c$.

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 β_1 only, as it is the most biased direction. Consider also the case of $\Gamma = \gamma D$, thus $z_t = x_t - \gamma x_{t+1}$. In this setting, all but the first instruments are strictly exogenous, while the first instrument $z_{1t} = \tilde{z}_{1t} + \varepsilon_{t-1} - \gamma\varepsilon_t$ is endogenous as it correlates with the contemporaneous regression error. The direction of the instrument endogeneity, in general, coincides with the feedback direction, as the transformation Γ mixes the timing of the observations but preserves the feedback direction.

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

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

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:

$$\begin{aligned}\mathbb{E}\Big[Z_1' M_{Z_{-1},X_{-1}} \varepsilon \mid \tilde{X}\Big] &= \mathbb{E}\Big[\tilde{Z}_1' M_{Z_{-1},X_{-1}} \varepsilon \mid \tilde{X}\Big] + a\mathbb{E}\Big[\sum\nolimits_{s,t} M_{st}^* (\varepsilon_{s-1} - \gamma\varepsilon_s)\varepsilon_t \mid \tilde{X}\Big] \\ &= a\sigma^2 \sum\nolimits_t (M_{tt-1}^* - \gamma M_{tt}^*).\end{aligned}$$

Here we used $M_{Z_{-1},X_{-1}} = (M_{st}^*)$ for shortness of notation. We aim to choose γ to render the last sum equal to zero. Finding such a root should generally be feasible, given that the diagonal elements of projection matrices tend to dominate those on the lower diagonal. Lemma 6 shows 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 $a\sigma^2 \operatorname{tr}[D'(I-\Gamma)\tilde{M}_\Gamma]$. By selecting Γ to solve equation (2), we therefore achieve an asymptotically unbiased estimator.

3.2 Consistency of estimator

Let γ_0 be the solution to (2) among matrices $\Gamma_0 = \gamma_0 D$. As we only observe X, knowing $\tilde{X}$ is equivalent to knowing the feedback direction α. Therefore, solving equation (2) is infeasible in practice. Let instead $\hat{\gamma}$ and $\hat{\Gamma} = \hat{\gamma} D$ be the solution to the empirically feasible equation:

$$\operatorname{tr}\big[D'(I-\hat{\Gamma})M_{\hat{\Gamma}}\big] = 0, \quad \text{where} \quad M_{\hat{\Gamma}} = I - X(X'(I-\hat{\Gamma})X)^{-1}X'(I-\hat{\Gamma}). \tag{3}$$

In practice, in order to find $\hat{\gamma}$, we minimize the function

$$f(\gamma) = \big\{\operatorname{tr}\big[D'(I-\gamma D)(I - X(X'(I-\gamma D)X)^{-1}X'(I-\gamma D))\big]\big\}^2$$

using standard non-linear optimization algorithms. In all our simulations, the convergence is very fast. The following Lemma gives sufficient conditions for the existence and uniqueness of a solution to equation (2).

Lemma 2. *Suppose that $\tilde{X}$ is a $T \times K$ matrix of rank K and $\Gamma = \gamma D$.*

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

(ii) If $|\operatorname{tr}(D'\tilde{M})| \leq \mu^2 K$ and $K < T/[1 + (1+\mu)^2]$ for some $\mu \in [0, 1]$, then there exists a unique γ solving equation (2) such that $|\gamma| < \mu/(1+\mu)$.

One may state an analog of Lemma 2 for equation (3) as well due to their analog structure. According to Lemma 2, equation (2) 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 Γ to future research. The proof of Lemma 2 reveals that though equation (2) is non-linear, its solution can be found relatively fast as a fixed point of a contraction.

One may think $K < T/5$ in Lemma 2(i) is restrictive. It arises because we impose no restrictions on $\tilde{X}$ other than full rank. Thus, we accommodate a wide range of idiosyncrasies in the regressors' distribution, 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 autocorrelation of the regressors, μ, eases this restriction considerably, as shown in Lemma 2(ii). If the original potential for bias is small, the γ that removes the bias is small as well. Specifically, if $\operatorname{tr}[D'\tilde{M}] = 0$, the original OLS has no (asymptotic) bias, and our estimator defaults back to OLS.

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

Theorem 3. *Suppose Assumption 1 holds.*

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

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

where $\bar{S}_\Gamma = \tilde{X}'(I - \Gamma)\tilde{X} + \sigma^2 \alpha\alpha' \operatorname{tr}\big[(I-\Gamma)\tilde{M}_\Gamma\big]$. Specifically, it follows that $r'\hat{\beta}^{\mathrm{IV}}(\Gamma)$ is consistent for $r'\beta$ when Γ solves equation (2).

(ii) If $K < T/5$, then $\hat{\gamma} - \gamma_0 = O_p(1/T)$, and $r'\hat{\beta}^{\mathrm{IV}}(\hat{\Gamma}) - r'\hat{\beta}^{\mathrm{IV}}(\Gamma_0) = o_p(1/\sqrt{T})$.

The result of Theorem 1 is a special case of Theorem 3(i) for $\Gamma = 0$. Using Γ that solves equation (2) produces a consistent estimator for any reasonable contrast, and the proposed solution does not depend on the contrast of interest. The ideal Γ_0 solving equation (2) can be random but is strictly exogenous, as it depends only on the strictly exogenous part of the regressors $\tilde{X}$. The solution to equation (3), $\hat{\Gamma}$, is random and depends on the error terms ε. However, as stated in Theorem 3(ii), the feasible estimator $r'\hat{\beta}^{\mathrm{IV}}(\hat{\Gamma})$ is consistent and has the same asymptotic distribution as the infeasible one using the ideal Γ_0.[5]

An appealing feature of our proposal is that the estimator is linear in the outcome variable. It eliminates bias by working only with the regressors. The same bias correction works for any outcome variable satisfying the weak exogeneity assumption with one-period feedback to regressors. Our solution does not require estimating the feedback direction α.

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.

Alternative approaches The large number of regressors is a primary driver of the size of the OLS bias. There are several proposals in the literature on how to reduce the dimensionality of the regressors, including Principle Components Analysis (PCA) (see, e.g., Carrasco and Rossi, 2016) or LASSO regularization (Medeiros and Mendes, 2016). The validity of these approaches for estimation and inference involving a scalar contrast requires quite strong assumptions like sparsity (for LASSO) or that the influence is fully captured by factors (for PCA). These assumptions should be imposed on one of two relations: the one between outcome and regressors and the relation between the regressor of interest and all other regressors. These are assumptions that may not be easy to justify. Supplemental Appendix C.4 contains simulation results showing the challenges with PCA when these assumptions fail.

Another approach is to recognize that the bias of the OLS arises from the feedback mechanism; controlling for this mechanism eliminates the bias. This insight leads to suggestions such as running an OLS regression that includes either lags of the outcome variable and the regressors or a lag of the residual.[6] Another option is to consider a leave-one-out approach. We explore the performance of these approaches in Supplemental Appendix C.4 and find

[5] The consistency of the new estimator relies heavily on Assumption 1, specifically one-period feedback. Simulations exploring robustness to violations of this assumption are in Supplemental Appendix C.2

[6] We are grateful to the anonymous referee for this suggestion.

that our estimator outperforms both in terms of bias and variance.

3.3 Consistency of variance estimator

For some matrix Γ, introduce $T - K_\Gamma = \mathrm{tr}\big[(I-\Gamma)\tilde{M}_\Gamma\big]$ and define the variance estimator

$$\hat{\sigma}^2(\Gamma) = \frac{y'(I-\Gamma)M_\Gamma y}{T - K_\Gamma}.$$

Using equation (5) in the Appendix, we find that the value of $\hat{\sigma}^2(\Gamma)$ is invariant to the value of β so that $\hat{\sigma}^2(\Gamma) = \varepsilon'(I-\Gamma)M_\Gamma\varepsilon/(T-K_\Gamma)$. Lemma 4(ii) from the Appendix implies that $\hat{\sigma}^2(\Gamma)$ possesses a very desirable property; it is non-negative for any realization of the data. If Γ solves equation (2) then $\hat{\sigma}^2(\Gamma)$ is consistent for σ^2.

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

$$\frac{\hat{\sigma}^2(\Gamma)}{\sigma^2} = 1 - \frac{\sigma^2\alpha'\bar{S}_\Gamma^{-1}\alpha}{T-K_\Gamma}\,\mathrm{tr}\big[D'(I-\Gamma)\tilde{M}_\Gamma\big]\,\mathrm{tr}\big[D(I-\Gamma)\tilde{M}_\Gamma\big] + o_p(1).$$

Specifically, it follows that $\hat{\sigma}^2(\Gamma)$ is consistent for σ^2 when Γ is such that equation (2) holds.

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

4 Inference

This section shows that the proposed IV estimator is asymptotically Gaussian conditionally on $\tilde{X}$ under some additional assumptions. We also suggest standard errors that, when paired with our estimator, provide asymptotically valid inferences, that is, confidence sets and hypothesis tests with asymptotically correct coverage and size, respectively.

There are two theoretical and practical challenges we encounter. The first is to correctly account for the asymptotic importance of a quadratic form. Specifically, the bias of OLS arises due to the presence of a quadratic form in errors, which has a non-trivial mean, as was noticed in Sawa (1978). 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 grows to infinity under a condition of eigenvalue negligibility. We refer the reader to Brown (1971); de Jong (1987); Chao et al. (2012); Anatolyev (2019); 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 the 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. Hansen et al. (2008) shows the importance of adjusting standard errors for the 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 zero on average, making the asymptotic variance depend on skewness and kurtosis of errors. The issue also arises when conducting proper inference for estimators of the limited information maximum likelihood type with many instruments (as in Hansen et al. (2008)). Below, we consider two instances where we can easily handle the first challenge and ignore the second challenge.

4.1 Inference when $K/T \to 0$

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

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

where $\hat{\Sigma}_T = \hat{\sigma}^2(\hat{\Gamma})\|r'(X'(I-\hat{\Gamma})X)^{-1}X'(I-\hat{\Gamma})\|^2$.

The condition $\max_t \|(\tilde{X}'\tilde{X})^{-1/2}\tilde{x}_t\| = o_p(1)$ is a relatively standard negligibility condition often invoked to ensure asymptotic Gaussianity of OLS via the 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. Meanwhile, the average of these diagonal elements equals K/T, so this condition can hold only when the number of regressors is moderately large ($K/T \to 0$). In such cases, both challenges described at the beginning of this section become asymptotically negligible. The standard errors in Theorem 5 resemble usual IV-type standard errors but use the newly proposed variance estimator $\hat{\sigma}^2(\hat{\Gamma})$.

4.2 Inference with Gaussian errors

Theorem 6. *Suppose Assumption 1 holds and ε_1 is Gaussian conditionally on $\tilde{X}$. Assume that Γ solves equation (2) and $\|\Gamma\| < 1 - c$. Then, as $T \to \infty$,*

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

where Σ_T is measurable with respect to $\tilde{X}$. With probability asymptotically approaching one, $\Sigma_T \leq (1+\psi)\hat{\sigma}^2(\Gamma)\|r'(X'(I-\Gamma)X)^{-1}X'(I-\Gamma)\|^2$, where $\psi = \frac{|\operatorname{tr}(B^2)|}{\operatorname{tr}(B'B)}$ and $B = D'(I-\Gamma)\tilde{M}_\Gamma$.

Theorem 6 allows the number of regressors to grow proportionally with the sample size. In this case, the standard errors stated in Theorem 5 do not correctly account for the uncertainty induced by the quadratic form. Similar issues arise in the many instruments literature, which proposes new estimators of asymptotic variances (Hansen et al. (2008)). In the current setup, the missing term in the asymptotic variance formula depends on the importance of the feedback and is hard to estimate. Instead, we provide an upper bound on the asymptotic variance, making the confidence sets asymptotically valid but conservative. In our simulations, we noticed that the correction ψ tends to be tiny and does not change the standard errors much. We advise to calculate ψ as a robustness check.

We use the assumption that errors are Gaussian in two distinct ways. First, it allows us to bypass the Lindeberg-type negligibility condition for the main linear term, as any linear combination of errors with weights depending on $\tilde{X}$ is conditionally Gaussian in finite samples. If errors are not Gaussian, one may instead impose a high-level assumption $\max_t w_t^2 / \sum_t w_t^2 \to 0$ for $w_t = r'(\tilde{Z}'\tilde{X})^{-1}\tilde{Z}_t$.

Second, Gaussianity removes the need to estimate the skewness and kurtosis of the error term. The bias removal procedure ensures the trace of $B = D'(I-\Gamma)\tilde{M}_\Gamma$ equals zero, leading to $\mathbb{E}[\varepsilon' B \varepsilon] = \sum_t B_{tt}\mathbb{E}\varepsilon_t^2 = 0$. That is, the diagonal sum $\sum_t B_{tt}\varepsilon_t^2$ is zero on average, though not for every realization. The variance of this diagonal sum, $Var(\sum_t B_{tt}\varepsilon_t^2) = \sum_t B_{tt}^2 \mathbb{E}(\varepsilon_t^2 - \sigma^2)^2$ is negligible when $K/T \to 0$. When K grows proportionally to T, the variance of the diagonal sum is correctly accounted for by the variance formula provided in the theorem only when $\mathbb{E}(\varepsilon_t^2 - \sigma^2)^2 = 2\sigma^4$, which holds for the Gaussian distribution. If errors have positive excess kurtosis, one must add a positive term to the asymptotic variance. However, the correction term tends to be small relative to the total variance as $\sum_t B_{tt}^2$ tends to be a tiny fraction of $\operatorname{tr}(B'B) = \sum_{ts} B_{ts}^2$. Simulations in Supplemental Appendix C.3 suggest that our proposed inference method maintains good size properties, even when errors are non-Gaussian.

5 Extension to multiple periods

In the previous sections, we assumed that the violation of strict exogeneity happens for one period only. Some results generalize 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 $\tilde{X}$ with $\mathbb{E}[\varepsilon_t|\tilde{X}] = 0$, $c < \sigma^2 := \mathbb{E}[\varepsilon_t^2|\tilde{X}] < C$, and $\mathbb{E}[\varepsilon_t^4|\tilde{X}] < C$ almost surely.

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

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

The vector α_ℓ describes how the shock to the outcome variable affects the regressors ℓ 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 ℓ-th period feedback measured as $\alpha_\ell'(\tilde{X}'\tilde{X}/T)^{-1}\alpha_\ell$ should become negligible for large enough ℓ in typical macroeconomic settings.

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

$$r'\hat{\beta}^{\mathrm{IV}}(\Gamma) - r'\beta = \sigma^2 \sum_{\ell=1}^{L} r'\bar{S}_\Gamma^{-1}\alpha_\ell \operatorname{tr}\left[(D')^\ell (I - \Gamma)\tilde{M}_\Gamma\right] + o_p(1), \tag{4}$$

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. The special case of $\Gamma = 0$ shows that the bias of OLS is a linear combination of L terms involving 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' autocorrelations of order ℓ and are expected to decay for stationary regressors. Combined with the anticipated decrease in the size of α_ℓ, the first few terms in the bias formula should capture most of the bias in stationary applications.

The bias formula (4) is derived for any IV-type estimator for an $\tilde{X}$-measurable $T \times T$ matrix Γ with $\|\Gamma\| < 1 - c$. Theorem 7 suggests that if Γ solves a system of L equations $\operatorname{tr}\left[(D')^\ell (I - \Gamma)\tilde{M}_\Gamma\right] = 0$ for $\ell = 1, ..., L$ then $r'\hat{\beta}^{\mathrm{IV}}(\Gamma)$ is a consistent estimator. A natural suggestion is to search for Γ in the class of matrices $\Gamma = \sum_{\ell=1}^{L} \gamma_\ell D^\ell$ where the L parameters $\{\gamma_\ell\}_\ell$ solve the system of equations. We leave the questions of considering other classes of matrices and establishing guarantees for the existence of a solution to future research. Supplemental Appendix C.2 contains some simulation results in a setting with two-period feedback, when the IV estimator zeros down only the first lower diagonal.

Theorem 7 can be further generalized to the case of infinite feedback (under a summability condition on the feedback magnitudes). This generalization uses an argument that approximates a model with infinite lags by a model with a finite but slowly growing number of lags. Specifically, Theorem 7 generalizes to a model where a lagged dependent outcome serves as a regressor (which yields a model with infinite feedback and geometrically decaying feedback magnitudes).[7] Supplemental Appendix B explains how our setting with infinite geometrically decaying feedback includes typical VAR models. VAR models are often estimated via equation-by-equation OLS, which may result in significant biases when the number of included variables multiplied by the order of the VAR constitutes a substantial fraction of the sample size. Our setting does not cover local projection, as errors in those regressions have a moving average structure, requiring further generalization of our framework.

Formulae for the OLS bias in time series have been derived before in special cases. Specifically, Sawa (1978) and Nankervis and Savin (1988) derived the exact finite-sample bias in an AR(1) model with Gaussian errors and argued that the bias is sizable when the sample size is double-digit. Kiviet et al. (1999) considered a model with Gaussian innovations that has a small number of strictly exogenous regressors and derived a formula for asymptotic $(1/T)$ bias from using the lagged outcome as a regressor. They argued that the bias can be significant in small samples. Stambaugh (1999) showed that the weak exogeneity bias is large when a single weakly exogenous regressor is persistent. All these results are special cases of our formula when applied to infinite geometrically decaying feedback.

[7] These results are available upon request from the authors but are not included here, as they constitute a separate paper.

6 Simulations

This section aims 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 second order: stationary, I(1), or I(2) processes. A prevailing (but not uniformly accepted) practice is to transform all variables to stationarity before running a regression 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 stationarity or make decisions about variable stationarity varies widely across the literature, with many such decisions based on both statistical tests and expert judgments. In order to unify the pre-treatment of variables, given that a large fraction of regressions is aimed at business-cycle parameters, we apply Hamilton (2018) transformation to all variables in the data set. Specifically, we define each variable's cyclical component as a two-year-ahead forecast error to this variable based on a univariate AR(4) regression. According to Hamilton (2018), this filtering transforms many types of stationary and up to second-order integrated variables into stationary ones and extracts their business cycle component.

For each K in $\{5, 15, 25, \dots, 85, 100\}$, we perform 100 experiments where we randomly draw K distinct variables from the transformed data set, denote them X_{r} and an additional variable y_{r}. We calculate, through simulations, the biases and standard deviations of OLS and our proposed estimator (referred to as IV) for the linear contrast in the feedback direction in the regression of y_{r} on X_{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 $\beta = (X_{\mathrm{r}}'X_{\mathrm{r}})^{-1}X_{\mathrm{r}}'y_{\mathrm{r}}$, $\sigma^2 = e'e/(T-K)$ for $e = y_{\mathrm{r}} - X_{\mathrm{r}}\beta$, and $\alpha = X_{\mathrm{r}}'D'e/(e'e)$. We simulate samples as $X = X_{\mathrm{r}} + D'\varepsilon\alpha'$ and $y = X\beta + \varepsilon$ where $\varepsilon \sim N(0, \sigma^2 I)$. We simulate errors here as i.i.d. even though the empirical regression residuals typically exhibit substantial serial correlation. We calculate the bias and standard deviation for the OLS and IV estimators of the linear contrast with $\theta = \alpha'\beta$.

The left panel of Figure 2 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 $M_{\mathrm{r}} = I - X_{\mathrm{r}}(X_{\mathrm{r}}'X_{\mathrm{r}})^{-1}X_{\mathrm{r}}'$ that is $\mathrm{tr}(D'M_{\mathrm{r}})/T$. The right panel contains the results of the experiments that fall into the 90th percentile of the bias.

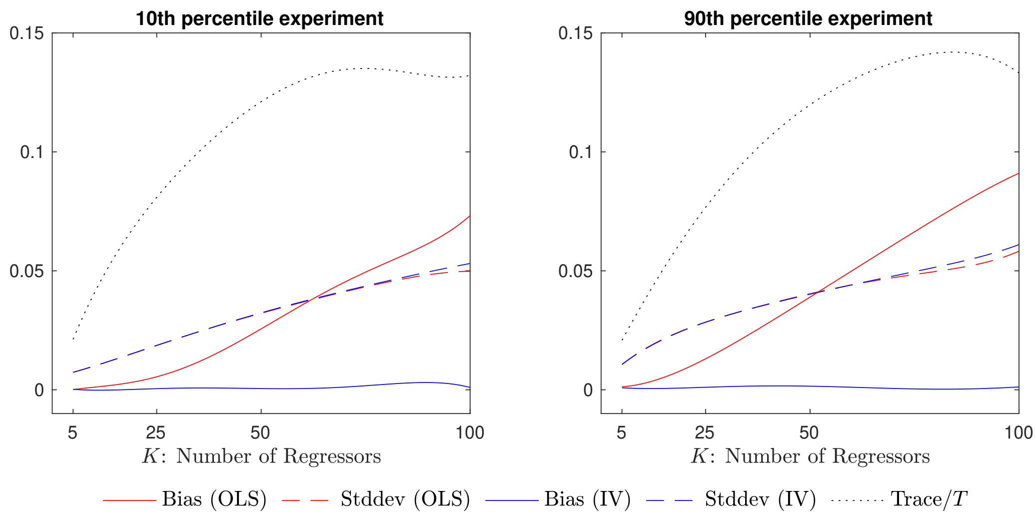

Figure 2: Absolute Bias and Standard Deviation of OLS and IV with $T = 200$

Figure 2 shows that a typical macroeconomic data set demonstrates enough time series dependence or short-term linear predictability that it creates a potential for substantial 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 is about half of the standard deviation and approximately equal to the standard deviation with 50 regressors. Biases of this size lead to invalid statistical inferences when relying on the OLS. Figure 3 reports the size distortions in the experiments described above for the 5% tests about the linear contrast in the direction of the feedback. Our newly proposed estimator completely corrects the bias without any significant change to the standard deviation and restores the correct size for statistical tests.

Another observation from Figure 2 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

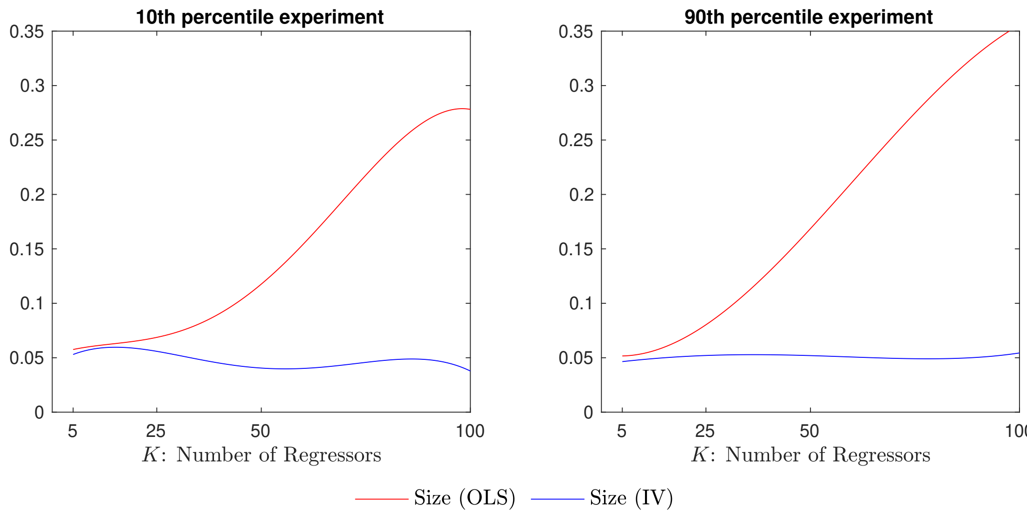

Figure 3: Size of Nominal 5% two-sided tests using OLS and IV with $T = 200$

that in all of our experiments and simulations, we encountered no problems finding the solution to equation (3), which supports our assertion that the sample size requirement imposed in Lemma 2 is sufficient but not necessary for the existence of the solution.

Finally, we note that the results presented in Figure 2 are both qualitatively and quantitatively similar to the left panels of Figure 1. Our theoretical results are relatively agnostic about the time series properties of the regressors. Specifically, we make no assumptions about stationarity or origin of the regressors. Within the framework allowed for in Assumption 1, the data-generating processes underlying Figure 2 mimic the time series behavior and feedback size/direction of a 'typical' macroeconomic application.[8]

It is important to note that the data-generating processes for the experiments reported in Figures 2 and 3 satisfy Assumption 1. We now explore the sensitivity of our results to violations of Assumption 1. We consider an empirically motivated heteroskedasticity by changing only one aspect of the simulations above: the regression errors are simulated as heteroskedastic, $\varepsilon \sim N(0, \text{diag}(\sigma_1^2, \ldots, \sigma_T^2))$, with variances depending on the realization of the regressors in an unspecified way observed in the data. Specifically,

$$\sigma_t^2 = s \begin{cases} e_t^2, & \text{if } e_t^2 \leq 3\sigma^2, \\ 3\sigma^2, & \text{if } e_t^2 > 3\sigma^2, \end{cases}$$

[8] As previously noted the processes do not mimic the serial correlation of the prediction errors.

where s is such that $\sigma^2 = \frac{1}{T}\sum_{t=1}\sigma_t^2$ and e_t are the residuals in the US data. We censor the extreme outliers at $3\sigma^2$ and match the observed unconditional variance.

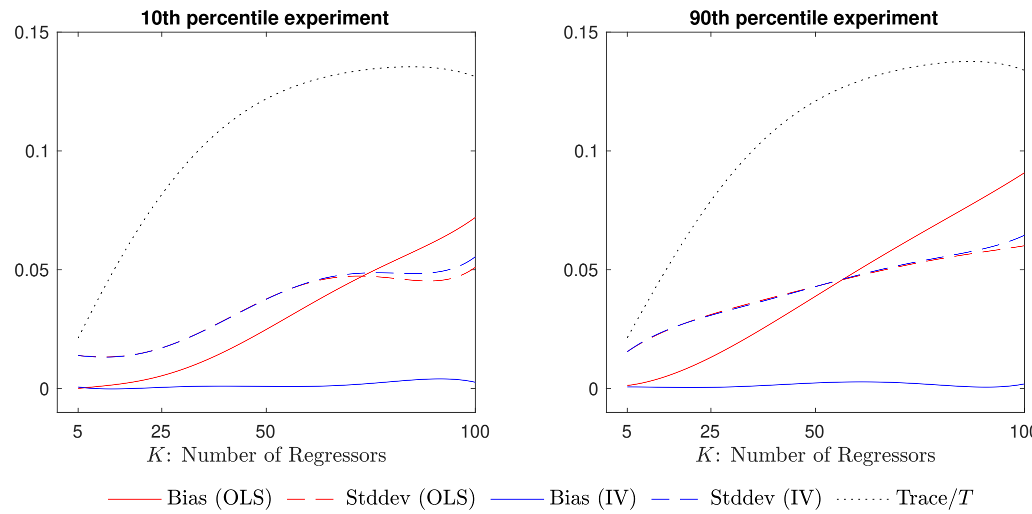

Figure 4: Absolute Bias and Standard Deviation of OLS and IV with heteroskedasticity

Reporting in Figure 4 is analogous to that in Figure 2, and the results look qualitatively very similar. Theoretically, one can adapt our formula for the OLS bias stated in Theorem 1 to conditional heteroskedasticity, but Assumption 1 is essential for the newly proposed IV estimator. However, simulations show that in practice, the proposed estimator corrects bias exceptionally well, even in the presence of empirically relevant heteroskedasticity.

We performed additional sensitivity checks and placed the results in the Supplemental Appendix. Specifically, we explored a violation of Assumption 1 by simulating the data with two-period feedback. In such a case, the OLS produces significant biases in two directions corresponding to two feedback directions. The IV estimator successfully eliminates the bias in the direction of the first feedback and reduces the bias in another direction. We also explored a sensitivity to the Gaussianity assumption stated in Theorem 6, and found that the proposed inference work quite well when the errors are not Gaussian.

Figure 5 answers the question of how different the results of the OLS and our newly proposed estimators are in a randomly selected regression based on typical macroeconomic data. As in the experiments described above, we select $(y_{\mathrm{r}}, X_{\mathrm{r}})$ at random. 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

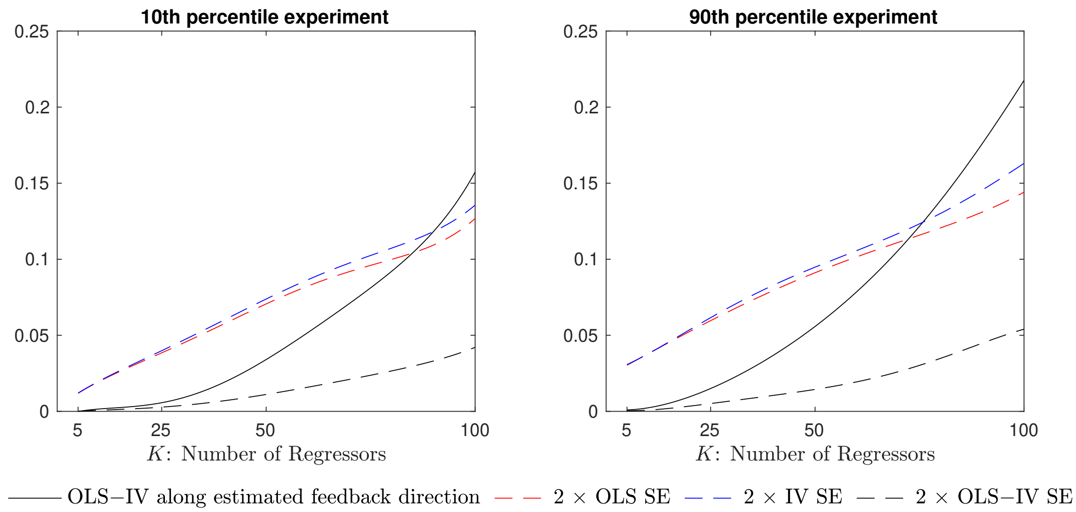

Figure 5: Difference between IV and OLS along estimated feedback direction, $T = 200$

experiment corresponding to the 10th percentile of the absolute difference on the left panel and the 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 the OLS and IV standard errors, and finally, double the standard deviations of the difference.[9] We report doubled standard errors to relate the results directly to the corresponding t statistics.

As shown in Figure 5, the difference between the OLS and IV estimators is statistically significant in all experiments reported. Since the validity of OLS with many regressors is only known for the strictly exogenous case (a special case of Assumption 1 with $\alpha = 0$), the observed difference between estimators are largely due to bias and cannot be attributed to randomness in realizations. An alternative explanation is that the data reject a hypothesis of no feedback in almost all regressions we considered, indicating that feedback mechanisms are prevalent in macroeconomic applications. Comparing the two estimators' differences to doubled standard errors reveals that when the number of regressors is above 70, the IV estimate falls outside of the OLS confidence set, and the OLS estimate falls outside the IV confidence set. This discrepancy underscores the potential for substantial disagreement between the two methods. Finally, the standard deviation of the OLS-IV difference is much smaller than the standard error of either estimator — more than five times smaller in some

[9]One can derive the distribution of the difference under the assumption of strict exogeneity using ideas that are analogous to those stated in Section 4.

cases. This stark contrast shows that stochastic deviations in the two estimators are tightly aligned, and most of their difference arises from the bias, not variance.

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

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

NOTE: $t_\Delta = |\hat{\beta}^{\text{OLS}} - \hat{\beta}^{\text{IV}}|/se(\hat{\beta}^{\text{OLS}} - \hat{\beta}^{\text{IV}})$. The average is 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 $\text{tr}(D'M_\text{r})/T$ 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 with more regressors, 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 estimates. The average is over the K coefficients in each regression and over the 100 random regressions with K regressors drawn from the macroeconomic database described above. While the fraction of coefficients with a statistically significant difference between OLS and IV is declining with K, the absolute number of such coefficients increases. We conclude that while most directions/coefficients are immune to the biases, a non-trivial fraction of coefficients is 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.

Brown, B. M. (1971). Martingale central limit theorems. *The Annals of Mathematical Statistics*, 59–66.

Carrasco, M. and B. Rossi (2016). In-sample inference and forecasting in misspecified factor models. *Journal of Business & Economic Statistics 34*(3), 313–338.

Chamberlain, G. (1982). The general equivalence of Granger and Sims causality. *Econometrica 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.

de Jong, P. (1987). A central limit theorem for generalized quadratic forms. *Probability Theory and Related Fields 75*(2), 261–277.

Engle, R. F., D. F. Hendry, and J.-F. Richard (1983). Exogeneity. *Econometrica 51*(2), 277–304.

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

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

Hachem, W., A. Hardy, and J. Najim (2016). Large complex correlated wishart matrices: Fluctuations and asymptotic independence at the edges. *Annals of Probability 44*(3), 2264–2348.

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

Kiviet, J. F., G. D. Phillips, and B. Schipp (1999). Alternative bias approximations in first-order dynamic reduced form models. *Journal of Economic Dynamics and Control 23*(7), 909–928.

Kline, P., R. Saggio, and M. Sølvsten (2020). Leave-out estimation of variance components. *Econometrica 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.

Marchenko, V. A. and L. A. Pastur (1967). Distribution of eigenvalues for some sets of random matrices. *Matematicheskii Sbornik 114*(4), 507–536.

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).

Medeiros, M. C. and E. F. Mendes (2016). l1-regularization of high-dimensional time-series models with non-gaussian and heteroskedastic errors. *Journal of Econometrics 191*(1), 255–271.

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

Nankervis, J. C. and N. E. Savin (1988). The exact moments of the least-squares estimator for the autoregressive model corrections and extensions. *Journal of Econometrics 37*(3), 381–388.

Nickell, S. (1981). Biases in dynamic models with fixed effects. *Econometrica 49*(6), 1417–1426.

Sawa, T. (1978). The exact moments of the least squares estimator for the autoregressive model. *Journal of Econometrics 8*(2), 159–172.

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

A.1 Auxilliary lemmas

The results of the following lemma are well-known but included here for ease of reference.

Lemma 3. *(i) For a symmetric matrix A and positive semi-definite (psd) matrix B, we have the bounds $\lambda_{\min}(A)\operatorname{tr}(B) \leq \operatorname{tr}(AB) \leq \lambda_{\max}(A)\operatorname{tr}(B)$.*

(ii) For any square matrix A, let λ be the smallest eigenvalue of $\frac{A+A'}{2}$. If $\lambda > 0$, then $\|A^{-1}\| \leq 1/\lambda$.

(iii) For any compatible matrices A and B, we have $|\operatorname{tr}(AB)| \leq \|A\|_F\|B\|_F$.

Proof of Lemma 3. (i) As A is symmetric, there exists U with $U'U = I$ such that $D = UAU'$ is a diagonal matrix with eigenvalues of A along its diagonal. Let $F = UBU'$. Note that $F_{ii} \geq 0$ since B is psd and $\operatorname{tr}(B) = \operatorname{tr}(F) = \sum_i F_{ii}$. Now,

$$\operatorname{tr}(AB) = \operatorname{tr}(UAU'UBU') = \sum_i D_{ii}F_{ii} \leq \max_i D_{ii} \sum_i F_{ii} = \lambda_{\max}(A)\operatorname{tr}(B),$$

and $\sum_i D_{ii}F_{ii} \geq \min_i D_{ii} \sum_i F_{ii} = \lambda_{\min}(A)\,\mathrm{tr}(B)$. (ii) Now, $\lambda\|x\|^2 \leq x'(A+A')x/2 = x'Ax \leq \|x\|\|Ax\|$. As $\lambda > 0$, it follows that $Ax \neq 0$ when $x \neq 0$, so A is invertible. For $x \neq 0$, we then have $\lambda \leq \|Ax\|/\|x\| = \|y\|/\|A^{-1}y\|$ where $y = Ax$. Thus, $\|A^{-1}\| = \sup_{y\neq 0}\|A^{-1}y\|/\|y\| \leq 1/\lambda$. (iii) $|\mathrm{tr}(AB)| = \left|\sum_{ij} A_{ij}B_{ij}\right| \leq \sqrt{\sum_{ij} A_{ij}^2}\sqrt{\sum_{ij} B_{ij}^2} = \|A\|_F\|B\|_F$. □

Lemma 4. *Suppose $X \in \mathbb{R}^{T\times K}$ has full rank and $\Gamma \in \mathbb{R}^{T\times T}$ has $\|\Gamma\| < 1 - c$. Then (i) the matrices $I - \Gamma$, $I - P\Gamma$, $X'(I - \Gamma)X$ and $I + A_\Gamma M$ are invertible where $A_\Gamma = (I - \Gamma)^{-1}\Gamma$; (ii) the matrix $(I - \Gamma)M_\Gamma + M_\Gamma'(I - \Gamma')$ is positive semi-definite and $c < \frac{T-K_\Gamma}{T-K} < C$ where $T - K_\Gamma = \mathrm{tr}[(I - \Gamma)M_\Gamma]$; (iii) the following identities hold:*

$$(I - \Gamma)M_\Gamma = M(I + A_\Gamma M)^{-1}, \tag{5}$$

$$P_\Gamma = (I - P\Gamma)^{-1}P(I - \Gamma), \tag{6}$$

$$M_\Gamma = (I - P\Gamma)^{-1}M, \tag{7}$$

$$\hat\beta^{\mathrm{IV}}(\Gamma) = (X'X)^{-1}X'(I + A_\Gamma M)^{-1}y. \tag{8}$$

Proof of Lemma 4. (i) Lemma 3(ii), the triangle inequality, and $\|\Gamma\| < 1$ yields invertibility of $I - \Gamma$, $I - P\Gamma$, $X'(I - \Gamma)X$, and $I - \Gamma P$. $I + A_\Gamma M$ is invertible since

$$I + A_\Gamma M = (I - \Gamma)^{-1}(I - \Gamma + \Gamma M) = (I - \Gamma)^{-1}(I - \Gamma P).$$

(iii) Next, we note that

$$\begin{aligned}(X'X)^{-1}X' &= (X'X)^{-1}\left[X'(I - \Gamma)X\left(X'(I - \Gamma)X\right)^{-1}\right]X' \\ &= (X'X)^{-1}X'(I - \Gamma)P_\Gamma(I - \Gamma)^{-1}.\end{aligned} \tag{9}$$

Pre-multiplying (9) by X and using $P_\Gamma = PP_\Gamma$ gives us $P = (I - P\Gamma)P_\Gamma(I - \Gamma)^{-1}$ and hence (6). Therefore, we also have (7):

$$M_\Gamma = I - P_\Gamma = (I - P\Gamma)^{-1}(I - P\Gamma - P(I - \Gamma)) = (I - P\Gamma)^{-1}M.$$

Using the "push-through" identity $(I - P\Gamma)^{-1}P = P(I - \Gamma P)^{-1}$ on (6) similarly yields (5):

$$M_\Gamma = I - P(I + A_\Gamma M)^{-1} = (I + A_\Gamma)M(I + A_\Gamma M)^{-1} = (I - \Gamma)^{-1}M(I + A_\Gamma M)^{-1}.$$

Reusing that $P_\Gamma = P(I + A_\Gamma M)^{-1}$ we get (8) from $\hat{\beta}^{\mathrm{IV}}(\Gamma) = \left(X'X\right)^{-1} X' P_\Gamma y$ and $XP = X$.
(ii) Positive semi-definiteness comes from (5):

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

and psd of $(I-\Gamma')^{-1}+(I-\Gamma)^{-1}$. The rate condition follows from $\mathrm{tr}(M) = T-K$, Lemma 3(i),

$$T - K_\Gamma = \mathrm{tr}[M(I + A_\Gamma M)^{-1}] = \mathrm{tr}\left[\tfrac{(I+A_\Gamma M)^{-1}+(I+MA_\Gamma')^{-1}}{2} M\right],$$

and that the eigenvalues of $\frac{(I+A_\Gamma M)^{-1}+(I+MA_\Gamma')^{-1}}{2}$ are in $\frac{1-\|\Gamma\|}{1+\|\Gamma\|}$ to $\frac{1+\|\Gamma\|}{1-\|\Gamma\|}$ with $\|\Gamma\| < 1-c$. $\square$

Lemma 5. *Suppose $y_t = x_t'\beta + \varepsilon_t$, $x_t = \tilde{x}_t + \sum_{\ell=1}^{L} \alpha_\ell \varepsilon_{t-\ell}$ where the $T \times K$ matrix $\tilde{X} = [\tilde{x}_1, \ldots, \tilde{x}_T]'$ has full rank, and $\alpha_1, \ldots, \alpha_L, r$ are $K\times 1$ vectors. Then, there exists an invertible $K \times K$ matrix Θ mapping $\{\tilde{X}, r, \beta, \{\alpha_l\}_{l=1}^{L}\}$ to $\{\tilde{X}\Theta, \Theta' r, \Theta^{-1}\beta, \{\Theta'\alpha_l\}_{l=1}^{L}\}$, that satisfies:*

(i) $\Theta' r$ and $\{\Theta'\alpha_\ell\}_{\ell=1}^{L}$ are spanned by the first $L+1$ basis vectors so that $r'\Theta = (r_*', \mathbf{0}_{K-L-1}')$ and $\alpha_\ell'\Theta = (\alpha_{*,\ell}', \mathbf{0}_{K-L-1}')$ with $r_*, \alpha_{*,1}, \ldots, \alpha_{*,L} \in \mathbb{R}^{L+1}$;

(ii) $\Theta'\tilde{X}_1'\tilde{X}_1\Theta/T = I_{L+1}$ and $\Theta'\tilde{Z}_2'\tilde{X}_1\Theta = 0$ for $\tilde{X} = [\tilde{X}_1, \tilde{X}_2]$, and $\tilde{Z} = [\tilde{Z}_1, \tilde{Z}_2] = (I - \Gamma')\tilde{X}\Theta$, where $\tilde{X}_1$ and $\tilde{Z}_1$ each have $L+1$ columns.

Proof of Lemma 5. Let $\Theta = \Theta_0^{-1}\Theta_1$, where $K \times K$ matrix Θ_0 is the symmetric square root of $[\tilde{X}_1, \tilde{Z}_2]'[\tilde{X}_1, \tilde{Z}_2]/T$ and $\Theta_1 R_1$ is the QR decomposition of $\Theta_0^{-1}[r, \alpha_1, \ldots, \alpha_L]$. Specifically, $\Theta_1' = \Theta_1^{-1}$ is a $K \times K$ matrix and R_1 is spanned by the first $L+1$ basis vectors. We then have that $\Theta'[r, \alpha_1, \ldots, \alpha_L] = \Theta_1'\Theta_0^{-1}[r, \alpha_1, \ldots, \alpha_L] = \Theta_1'\Theta_1 R_1 = R_1$ while we also have $\Theta'[\tilde{X}_1, \tilde{Z}_2]'[\tilde{X}_1, \tilde{Z}_2]\Theta/T = \Theta_1'\Theta_0^{-1}\Theta_0^2\Theta_0^{-1}\Theta_1 = I_K$. $\square$

Lemma 6. *Suppose X and Z are $T \times K$ matrices with $Z'X$ invertible. Let $X = [X_1, X_2]$ and $Z = [Z_1, Z_2]$, where X_ℓ and Z_ℓ are $T \times K_\ell$ with $K_1 + K_2 = K$. Define the oblique projections $P_Z = X(Z'X)^{-1}Z' = I - M_Z$ and $P_2 = X_2(Z_2'X_2)^{-1}Z_2' = I - M_2$.*

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

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

(ii) $M_Z = M_2 - M_2X_1(Z_1'M_2X_1)^{-1}Z_1'M_2$.

(iii) If $Z = (I - \Gamma')X$ for $T \times T$ matrix Γ with $\|\Gamma\| < 1 - c$ and A is a $T \times T$ matrix, then

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

If $\|\Gamma\| = 0$, then the statement above holds with $C = 1$.

Proof of Lemma 6. Letting $\Delta = (Z_1' M_2 X_1)^{-1}$, the matrix block inversion formula gives

$$(Z'X)^{-1} = \begin{bmatrix} \Delta & -\Delta Z_1' X_2 (Z_2' X_2)^{-1} \\ -(Z_2' X_2)^{-1} Z_2' X_1 \Delta & (Z_2' X_2)^{-1} \{I + Z_2' X_1 \Delta Z_1' X_2\} \end{bmatrix} \tag{10}$$

(i) From (10) we have

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

(ii) Denote $\delta = X_1 \Delta Z_1'$. Using (10) above:

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

(iii) We impose without loss of generality that $X_2' X_1 = 0$ and $X_1' X_1 = I_{K_1}$. This entails no loss since (5) and (7) yields $Z_2' M_2 = 0$ and $M_2 X_2 = 0$, which in turn implies that (M_Z, M_2) is invariant under the transformation $[X_1, X_2] \mapsto [M^* X_1 (X_1' M^* X_1)^{-1/2}, X_2]$ where $M^* = I - X_2 (X_2' X_2)^{-1} X_2' = I - P^*$. From (ii), Lemma 3(iii), and $\|\Psi\|_F \leq \|\Psi\| \sqrt{\mathrm{rk}(\Psi)}$:

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

Equation (7) gives $M_2 = (I - P^* \Gamma)^{-1} M^*$ and therefore $\|M_2\| \leq \frac{1}{1 - \|\Gamma\|} < C$. Together with $X_2' X_1 = 0$, (7) also yields

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

and therefore

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

As $X_1'X_1 = I_{K_2}$, the eigenvalues of the last matrix are larger than $\frac{1-\|\Gamma\|}{1+\|\Gamma\|}$. Therefore, Lemma 3(ii) and $\|\Gamma\| < 1 - c$ implies that $\|(Z_1'M_2X_1)^{-1}\| < C$. $\square$

Proof of Lemma 1. Let us apply the linear transformation Θ_0 described in Lemma 5 to the regressors. Notice that projections $\tilde{M}$ and M are invariant towards this transformation. For the transformed data projection off all but the first regressor M_{-1} differ from both $\tilde{M}$ and M by not projecting one regressor, since $X_{-1} = \tilde{X}_{-1}$, and M_{-1} coincides with what in Lemma 6 labeled as M_2. Applying part (iii) of Lemma 6 with $\Gamma = 0$, we find:

$$|\mathrm{tr}(D(M - M_{-1}))| \leq 1, \text{ and } \left|\mathrm{tr}(D(\tilde{M} - M_{-1}))\right| \leq 1.$$

This implies the statement of Lemma 1. $\square$

Lemma 7. *Suppose $\tilde{X}$ is a $T \times k$ matrix with $\tilde{X}'\tilde{X}/T = I_k$, 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 ℓ, and $\{\varepsilon_t\}_{t=1-k}^T$ are i.i.d. conditionally on $\tilde{X}$ with $\mathbb{E}[\varepsilon_t \,|\, \tilde{X}] = 0$, $0 < \sigma^2 = \mathbb{E}[\varepsilon_t^2 \,|\, \tilde{X}] < C$ and $\mathbb{E}[\varepsilon_t^4 \,|\, \tilde{X}] < C$. 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) $\varepsilon'A\varepsilon - \mathbb{E}[\varepsilon'A\varepsilon \,|\, \tilde{X}] = O_p(\|A\|_F)$.*

Proof of Lemma 7. (i) Note that:

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

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

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

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

$$\begin{aligned}
\mathbb{E}\Big[\Big(\varepsilon' A\varepsilon - \tilde{\mathbb{E}}[\varepsilon' A\varepsilon]\Big)^2\Big] &= \sigma^4 \sum_t \sum_{s\neq t} (A_{ts}A_{st} + A_{ts}^2) + \sum_t A_{tt}^2 \tilde{\mathbb{E}}\big[(\varepsilon_t^2 - \sigma^2)^2\big] \\
&\leq C\left(\sum_t \sum_{s\neq t} (A_{ts}A_{st} + A_{ts}^2) + \sum_t A_{tt}^2\right) \leq 2C \sum_{t,s} A_{ts}A_{st} = 2C\|A\|_F^2,
\end{aligned}$$

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

Lemma 8. *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 8. Due to Lemma 3(i) and equation (6) we have, for any matrix A,

$$|\operatorname{tr}(AP)| = \left|\operatorname{tr}\left(\frac{A+A'}{2}P\right)\right| \leq \left\|\frac{A+A'}{2}\right\| \operatorname{tr}(P) \leq \|A\|K; \tag{11}$$

$$|\operatorname{tr}(AP_\Gamma)| = |\operatorname{tr}((I-\Gamma)A(I-P\Gamma)^{-1}P)| \leq \frac{1+\|\Gamma\|}{1-\|\Gamma\|}\|A\|K. \tag{12}$$

(i) $B = D'(I-\Gamma) - D'(I-\Gamma)P_\Gamma$, thus $B_{tt} = -\gamma - F_{tt}$, where $F = D'(I-\Gamma)P_\Gamma$. The condition $\operatorname{tr}(B) = O(K)$ implies $-\gamma T = \sum_t F_{tt} + O(K)$ and $\sum_t B_{tt}^2 = \sum_t F_{tt}^2 - T\gamma^2 + O(K) \leq \sum_t F_{tt}^2 + O(K)$. Consider diagonal elements of the matrix $F = AP_\Gamma$ with $A = D'(I-\Gamma)$ and $\|A\| \leq 1 + |\gamma|$:

$$\begin{aligned}
\sum_t F_{tt}^2 = \sum_t (AP_\Gamma)_{tt}^2 = \sum_t \left(\sum_s A_{ts}P_{\Gamma,st}\right)^2 &\leq \sum_t \left\{\sum_s A_{ts}^2 \sum_s P_{\Gamma,st}^2\right\} \\
= \sum_t (AA')_{tt}(P_\Gamma P_\Gamma')_{tt} &\leq \|A\|^2 \operatorname{tr}(P_\Gamma P_\Gamma') \leq CK.
\end{aligned}$$

(ii) Next we have

$$\begin{aligned}
\operatorname{tr}(B^2) &= \operatorname{tr}(D'(I-\Gamma)(I-P_\Gamma)D'(I-\Gamma)(I-P_\Gamma)) \\
&= \operatorname{tr}[D'(I-\Gamma)D'(I-\Gamma)] + \operatorname{tr}\big[D'(I-\Gamma)\left\{-2D'(I-\Gamma) + P_\Gamma D'(I-\Gamma)\right\}P_\Gamma\big].
\end{aligned} \tag{13}$$

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

$$\operatorname{tr}(D'(I-\Gamma)D'(I-\Gamma)) = T\gamma^2 \leq \sum_t F_{tt}^2 = O(K).$$

The second term in (13) is $O(K)$ due to (12). (iii) We use that for any matrix A, $\operatorname{tr}(A'A) = \sum_{t,s} A_{ts}^2 \geq |\sum_{t,s} A_{ts}A_{st}| = |\operatorname{tr}(A^2)|$:

$$\begin{aligned}\operatorname{tr}(B'B) &= \operatorname{tr}(M_\Gamma'(I-\Gamma')(I-\Gamma)M_\Gamma) \geq (1-|\gamma|)^2 \operatorname{tr}(M_\Gamma' M_\Gamma) \\ &\geq (1-|\gamma|)^2 \operatorname{tr}(M_\Gamma^2) = (1-|\gamma|)^2 \operatorname{tr}(M_\Gamma) = (1-|\gamma|)^2(T-K). \qquad \square\end{aligned}$$

Lemma 9. *Suppose that $\tilde{X}$ is a $T \times K$ matrix of rank K and $\Gamma = \gamma D$.*

(i) Equation (2) holds if and only if γ is a fixed point of the transformation f given by

$$f(\gamma) = \operatorname{tr}(D'\tilde{M}_\Gamma)/(T-K). \tag{14}$$

(ii) If $|\operatorname{tr}(D'\tilde{M})| \leq \mu^2 K$ 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 μ.

Proof. (i) Since $\Gamma = \gamma D$ we have $D'(I-\gamma D)\tilde{M}_\Gamma = D'\tilde{M}_\Gamma - D'D\gamma\tilde{M}_\Gamma$. We can therefore re-write (2) as: $\operatorname{tr}(D'\tilde{M}_\Gamma) - \gamma(T-K) = 0$. This equation is solved if (and only if) $\gamma = f(\gamma)$. (ii) Equation (7) yields $\tilde{M}_\Gamma = \tilde{M} + \gamma\tilde{P}D\tilde{M}_\Gamma$ and $\|\tilde{M}_\Gamma\| \leq \frac{1}{1-|\gamma|}$. Equation (11) gives $|\operatorname{tr}(D'\tilde{P}D\tilde{M}_\Gamma)| \leq K\|D\tilde{M}_\Gamma D'\|$. Therefore,

$$|f(\gamma)| \leq \frac{|\operatorname{tr}(D'\tilde{M})|}{T-K} + |\gamma|\frac{|\operatorname{tr}(D'\tilde{P}D\tilde{M}_\Gamma)|}{T-K} \leq \frac{\mu^2 K}{T-K} + |\gamma|\frac{K}{(T-K)(1-|\gamma|)} < \frac{\mu}{1+\mu}$$

and using equation (8)

$$\begin{aligned}&\frac{|f(\gamma_1)-f(\gamma_2)|}{|\gamma_1-\gamma_2|} = \frac{|\operatorname{tr}[D'(\tilde{M}_{\Gamma_1}-\tilde{M}_{\Gamma_2})]|}{|\gamma_1-\gamma_2|(T-K)} = \frac{|\operatorname{tr}[D'\tilde{P}D(I-\tilde{P}\Gamma_2)^{-1}\tilde{M}_{\Gamma_1}]|}{T-K} \\ &\leq \frac{K}{T-K}\frac{1}{1-|\gamma_1|}\frac{1}{1-|\gamma_2|} < \mu\end{aligned}$$

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

A.2 Proofs for results stated in the main text

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

Proof of Theorem 2. A special case of Theorem 4 with $\Gamma = 0$. $\square$

Proof of Lemma 2. (i) Special case of (ii) with $\mu = 1$ since $|\mathrm{tr}(D'\tilde{M})| = |\mathrm{tr}(D'\tilde{P})| \leq K$ follows from (11). (ii) Follows from Lemma 9 and the Banach fixed point theorem. $\square$

Proof of Theorem 3. (i) Special case of Theorem 7 with $L = 1$. (ii) Let $f(\gamma)$ be as in equation (14) define its empirical analog $\hat{f}(\gamma) = \frac{tr(D'M)}{T-K} + \gamma \frac{tr(D'PDM_\Gamma)}{T-K}$ where $\Gamma = \gamma D$. Since $K < T/5$, Lemma 9(ii) yields that both f and $\hat{f}$ are contractions on $[-1/2, 1/2]$ with contraction speed bounded by $\frac{1}{2}$ and therefore has unique fixed points γ_0 and $\hat{\gamma}$ by the Banach fixed point theorem. Furthermore, $|\hat{f}(\hat{\gamma}) - \hat{f}(\gamma_0)| \leq \frac{1}{2}|\hat{\gamma} - \gamma_0|$. For any γ we have:

$$\left|\hat{f}(\gamma) - f(\gamma)\right| \leq \frac{1}{T-K}\left(\left|\mathrm{tr}[D'(P-\tilde{P})] + \gamma\,\mathrm{tr}\left[D'PDM_\Gamma - D'\tilde{P}D\tilde{M}_\Gamma\right]\right|\right).$$

Consider the transformation Θ of Lemma 5. Since the projections $P, \tilde{P}, M_\Gamma, \tilde{M}_\Gamma$ are invariant to linear transformations, we may assume that Lemma 5((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 6 using X, $(I - \Gamma')X$, and $K_1 = 2$, while $\tilde{M}_2$ is an analogously defined starting from $\tilde{X}$, $(I - \Gamma')\tilde{X}$, and $K_1 = 2$. Therefore, Lemma 6(iii) implies for any compatible matrix A that

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

The same statement holds with $(P, \tilde{P})$ replacing $(M_\Gamma, \tilde{M}_\Gamma)$. Thus $|\hat{f}(\gamma) - f(\gamma)| \leq \frac{C}{T-K}$, so

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

This implies $|\hat{\gamma} - \gamma_0| \leq \frac{1}{2}\frac{C}{T-K} = O_p(1/T)$. Equation (8) yields

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

where $A_{\Gamma_0} = (I - \Gamma_0)^{-1}\Gamma_0$ and $A_{\hat{\Gamma}} = (I - \hat{\Gamma})^{-1}\hat{\Gamma}$. We have $\|\varepsilon\| = O_p(\sqrt{T})$ and $\|A_{\hat{\Gamma}} - A_{\Gamma_0}\| = O_p(\hat{\gamma} - \gamma_0) = O_p(1/T)$. By Assumption 1(iii) $r'(X'X)^{-1}r = O(1/T)$, while $\left\|(I + A_{\Gamma_0}M)^{-1}\right\|$ is uniformly bounded. Thus, $r'\hat{\beta}^{\text{IV}}(\hat{\Gamma}) - r'\hat{\beta}^{\text{IV}}(\Gamma_0) = O_p(T^{-1}) = o_p(1/\sqrt{T})$. □

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 $R_{\Gamma,\ell} = \operatorname{tr}[(D')^\ell(I - \Gamma)\tilde{M}_\Gamma]$. Note that $\{r'\bar{S}^{-1}\alpha_\ell, R_{\Gamma,\ell}\}_{\ell=1}^L$ and $r'\hat{\beta}^{\text{IV}}(\Gamma) - r'\beta$ are invariant under the transformation Θ of Lemma 5. Thus, we may assume without loss of generality that Lemma 5((i) and (ii)) hold with $\Theta = I_K$ and $Z = (I - \Gamma')X$. 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 6 using X, $Z = (I - \Gamma')X$, and $K_1 = L + 1$, while $\tilde{M}_2$, $\tilde{X}_2$, $\tilde{Z}_2$ are analogously defined starting from $\tilde{X}$, $\tilde{Z} = (I - \Gamma')\tilde{X}$, and $K_1 = L + 1$. This also implies that X_2, Z_2, and M_2 are non-random conditionally on $\tilde{X}$. Lemma 6(i) now yields:

$$r'\hat{\beta}^{\text{IV}}(\Gamma) - r'\beta = r'(Z_1'M_2X_1)^{-1}Z_1'M_2\varepsilon. \tag{15}$$

Defining $\bar{S}_2 = \mathbb{E}\left[Z_1'M_2X_1 \,|\, \tilde{X}\right] = \tilde{Z}_1'\tilde{X}_1 + \sigma^2 \sum_{j,\ell=1}^L \alpha_{*,j}\alpha_{*,\ell}' \operatorname{tr}[(D')^j(I - \Gamma)M_2D^\ell]$ and $R_{2,\ell} = \operatorname{tr}[(D')^\ell(I - \Gamma)M_2]$, we show as a **first** step that

$$r'\hat{\beta}^{\text{IV}}(\Gamma) - r'\beta = \sigma^2 \sum_{\ell=1}^L r_*'\bar{S}_2^{-1}\alpha_{*,\ell}R_{2,\ell} + o_p(1). \tag{16}$$

Equation (16) follows from equation (15) and the following statements proven below:

$$\left\|(Z_1'M_2X_1 - \bar{S}_2)/T\right\|_F = O_p(\sqrt{1/T}), \tag{17}$$

$$\left\|(\bar{S}_2/T)^{-1}\right\| \leq \frac{1}{(1 - \|\Gamma\|)}, \tag{18}$$

$$\left\|\left(Z_1'M_2\varepsilon - \sigma^2 \textstyle\sum_{\ell=1}^L \alpha_{*,\ell}R_{2,\ell}\right)/T\right\| = O_p(\sqrt{1/T}), \tag{19}$$

and $|R_{2,\ell}| \leq T$ and thus $\|\sum_{\ell=1}^L \alpha_{*,\ell}R_{2,\ell}\| = O(T)$. Let $u_\ell = (\varepsilon_{1-\ell}, \ldots, \varepsilon_{T-\ell})'$ as in Lemma 7.

Equation (17) considers a $(L+1)\times(L+1)$ matrix with mean of zero:

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

Notice that for $A=(I-\Gamma)M_2$, we have $\|A\|\le\frac{1+\|\Gamma\|}{1-\|\Gamma\|}$ and $\|A\|_F = O(\sqrt{T})$. Thus, applying Lemma 7((i) and (ii)) we obtain (17). As Lemma 4(ii) yields that $B=(I-\Gamma)M_2+M_2'(I-\Gamma')$ is a non-negative definite matrix, we have

$$
\begin{aligned}
x'(\bar{S}_2+\bar{S}_2')x &= x'\left[\tilde{Z}_1'\tilde{X}_1+\tilde{X}_1'\tilde{Z}_1\right]x + \sum_{j,\ell=1}^{L}x'\alpha_{*,j}\mathbb{E}\left[u_j'Bu_\ell|\tilde{X}\right]\alpha_{*,\ell}'x \\
&\ge x'\tilde{X}_1'\left[I-(\Gamma+\Gamma')/2\right]\tilde{X}_1x \ge (1-\|\Gamma\|)\|\tilde{X}_1x\|^2 = (1-\|\Gamma\|)T\|x\|^2
\end{aligned}
$$

Applying Lemma 3(ii) now implies (18). To prove (19), we note that

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

The first term is $O_p(\sqrt{T})$ due to Lemma 7(i) applied with $A=(I-\Gamma)M_2$ and $\alpha_\ell=0$ for $\ell>0$. As $\sigma^2R_{2,\ell}=\mathbb{E}[u_\ell'(I-\Gamma)M_2\varepsilon]$, Lemma 7(ii) yields $\sum_{\ell=1}^{L}\alpha_{*,\ell}\left[u_\ell'(I-\Gamma)M_2\varepsilon - R_{2,\ell}\right]=O_p(\sqrt{T})$. This leads to (19).

As the **second** and final step of the proof, we show that (16) implies (4). The first difference is that (16) depends on r_*', $\alpha_{*,\ell}$, and the $(L+1)\times(L+1)$ matrix $\bar{S}_2$, while (4) is described using $r'=(r_*',\mathbf{0}_{K-L-1}')$, $\alpha_\ell'=(\alpha_{*,\ell}',\mathbf{0}_{K-L-1}')$, and the $K\times K$ matrix $\bar{S}_\Gamma$. The second difference between these two statements is that (16) employs the oblique projection $\tilde{M}_2$, which projects off the $T\times(K-L-1)$ matrix $\tilde{X}_2$, while (4) employs $\tilde{M}_\Gamma$, which projects off the full $T\times K$ matrix of regressors $\tilde{X}=[\tilde{X}_1,\tilde{X}_2]$.

For the first of these discrepancies, from Lemma 5((i) and (ii)) we have that the lower left $(K-L-1)\times(L+1)$ blocks of $\alpha_\ell\alpha_j'$ and $\tilde{Z}'\tilde{X}$ are zero. Thus the upper left $(L+1)\times(L+1)$

block of $\bar{S}_\Gamma^{-1}$ is equal to the inverse of the upper left $(L+1)\times(L+1)$ block of $\bar{S}_\Gamma$. Thus

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

where $\Delta_{j\ell} = \mathrm{tr}\left[(D')^j(I-\Gamma)(\tilde{M}_\Gamma - \tilde{M}_2)D^\ell\right]$. Now, Lemma 6(iii) gives $|\Delta_{j\ell}| \le C\|D^j(D')^i(I-\Gamma)\| = O(1)$ so that $\sum_{\ell=1}^{L}(r'\bar{S}_\Gamma^{-1}\alpha_\ell - r'_*\bar{S}_2^{-1}\alpha_{*,\ell})R_{\Gamma,\ell} = o_p(1)$. For the second discrepancy, we have $R_{\Gamma,\ell} - R_{2,\ell} = \mathrm{tr}\left[(D')^\ell(I-\Gamma)(\tilde{M}_\Gamma - \tilde{M}_2)\right]$ and Lemma 6(iii) similarly yielding $|R_{\Gamma,\ell} - R_{2,\ell}| = O(1)$. Thus $\sum_{\ell=1}^{L}(R_{\Gamma,\ell} - R_{2,\ell})r'_*\bar{S}_2^{-1}\alpha_{*,\ell} = o_p(1)$. In conclusion, we have $\sum_{\ell=1}^{L} r'\bar{S}_\Gamma^{-1}\alpha_\ell R_{\Gamma,\ell} - r'_*\bar{S}_2^{-1}\alpha_{*,\ell}R_{2,\ell} = o_p(1)$, so (16) implies (4) and Theorem 7 follows. □

Proof of Theorem 4. From equation (5) we have $\hat{\sigma}^2(\Gamma) = \frac{\varepsilon'(I-\Gamma)M_\Gamma\varepsilon}{T-K_\Gamma}$. Note that σ^2 and $\hat{\sigma}^2(\Gamma)$ are invariant under the transformation Θ of Lemma 5. Thus, we may assume without loss of generality that Lemma 5((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 6(ii) yields:

$$\hat{\sigma}^2(\Gamma) = \frac{\varepsilon'(I-\Gamma)\tilde{M}_2\varepsilon}{T-K_\Gamma} - \frac{1}{T-K_\Gamma}\varepsilon'(I-\Gamma)\tilde{M}_2X_1(Z_1'\tilde{M}_2X_1)^{-1}Z_1'\tilde{M}_2\varepsilon. \tag{20}$$

Lemma 7(ii) applied with $A = \frac{(I-\Gamma)M_2}{T-K_\Gamma}$ implies that $\frac{\varepsilon'(I-\Gamma)M_2\varepsilon}{\sigma^2(T-K_\Gamma)} = 1 + o_p(1)$. For the second term in (20), we use statements (17)–(19) and the second part of the proof of Theorem 7 to obtain the conclusion of Theorem 4. □

Proof of Theorem 5. Let γ_0 be as in Theorem 3. Note that $r'\hat{\beta}^{\mathrm{IV}}(\Gamma_0), r'\beta$ and $\hat{\Sigma}_T(\Gamma_0) = \hat{\sigma}^2(\Gamma_0)\|r'(X'(I-\Gamma_0)X)^{-1}X'(I-\Gamma_0)\|^2$ are invariant under the transformation Θ of Lemma 5. Thus, we may assume without loss of generality that Lemma 5((i) and (ii)) hold with $\Theta = I_K$ and $Z = (I-\Gamma_0')X$. Following the proof of Theorem 7, we have formula (15), where the denominator satisfies (17) and (18). This implies that

$$r'\hat{\beta}^{\mathrm{IV}}(\Gamma_0) - r'\beta = \left(1+o_p(1)\right)\left(\sum\nolimits_t w_t\varepsilon_t + r'_*\bar{S}_2^{-1}\alpha_*\varepsilon'B\varepsilon\right), \tag{21}$$

where $w' = (w_1,\ldots,w_T) = r'_*\bar{S}_2^{-1}\tilde{Z}_1'M_2$, $B = D'(I-\Gamma_0)M_2$, 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

$$\frac{r'\hat{\beta}^{\mathrm{IV}}(\Gamma_0) - r'\beta}{\sqrt{\Sigma_T}} = \frac{\sum_t w_t\varepsilon_t + r_*'\bar{S}_2^{-1}\alpha_*\varepsilon' B\varepsilon}{\sqrt{\Sigma_T}} \Rightarrow N(0,1), \tag{22}$$

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

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

Lemma 8((i) and (iii)) together with $K/T \to 0$ imply that $(r_*'\bar{S}_2^{-1}\alpha_*)^2\left(\sum_t B_{tt}\varepsilon_t^2\right)/\sqrt{\Sigma_T} \xrightarrow{p} 0$ and $\frac{\mathrm{tr}(B^2)}{\mathrm{tr}(B'B)} \to 0$.

We obtain statement (22) by establishing the four conditions, (i)–(iv), of Sølvsten (2020), Corollary A2.8, located in the Supplemental Appendix of that article. Condition (i) is automatically satisfied if we define $w_{t,T} = \frac{w_t}{\sqrt{\Sigma_T}}$, $M_{st} = \frac{r_*'\bar{S}_2^{-1}\alpha_*}{2\sqrt{\Sigma_T}}(B_{st}+B_{ts})$ 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 6(i), we have for any $\tilde{r}$ of the form $\tilde{r}' = (\tilde{r}_*', \mathbf{0}_{K-L-1}')$ that

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

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 $\tilde{r}_*' = r_*'\bar{S}_2^{-1}(\tilde{Z}_1'M_2\tilde{X}_1)$ we have

$$w' = r_*'\bar{S}_2^{-1}\tilde{Z}_1'M_2 = \tilde{r}_*'(\tilde{Z}_1'M_2\tilde{X}_1)^{-1}\tilde{Z}_1'M_2 = \tilde{r}'(\tilde{Z}'\tilde{X})^{-1}\tilde{Z}'$$

so that $w_t = \tilde{r}'(\tilde{Z}'\tilde{X})^{-1}(\tilde{X}_t - \gamma_0\tilde{X}_{t+1})$. Note, that

$$\begin{aligned}
\max_t |w_t| &\leq \left\|(\tilde{X}'\tilde{X})^{1/2}(\tilde{X}'\tilde{Z})^{-1}\tilde{r}\right\|(1+|\gamma_0|)\max_t\left\|(\tilde{X}'\tilde{X})^{-1/2}\tilde{X}_t\right\|,\\
\sum_t w_t^2 &= \tilde{r}'(\tilde{Z}'\tilde{X})^{-1}\tilde{X}'(I-\Gamma_0)(I-\Gamma_0')\tilde{X}(\tilde{X}'\tilde{Z})^{-1}\tilde{r} \geq (1-|\gamma_0|)^2\left\|(\tilde{X}'\tilde{X})^{1/2}(\tilde{X}'\tilde{Z})^{-1}\tilde{r}\right\|^2.
\end{aligned}$$

$$\max_t|w_{t,T}| \leq \frac{\max_t |w_t|}{\sqrt{\sum_t w_t^2}} \leq \frac{1+|\gamma_0|}{1-|\gamma_0|}\max_t\|(\tilde{X}'\tilde{X})^{-1/2}\tilde{X}_t\| \to 0.$$

For condition (iii), we note that Lemma 8 and $K/T \to 0$ yields

$$\sum_s \sum_{t \neq s} \left[\frac{B_{st} + B_{ts}}{2}\right]^2 = \frac{1}{2}\operatorname{tr}(B^2 + B'B)(1 + o(1)).$$

This yields $\sum_s \sum_{t\neq s} M_{st}^2 \leq 1$. Also, $\left\|(B+B')/2 - diag(B)\right\| \leq \|B\| + \max_t |B_{tt}| = O(1)$. Therefore we have $\|(M_{st})_{s,t}\| \to 0$ and have therefore established (22).

Finally, we prove that $\frac{\hat{\varSigma}_T}{\varSigma_T} \xrightarrow{p} 1$. Reusing the argument in the proof of Theorem 3, we first have that $(\hat{\varSigma}_T - \hat{\varSigma}_T(\Gamma_0))/\varSigma_T = o(1)$. By Lemma 6(ii), we have for $u = (\varepsilon_0, \dots, \varepsilon_{T-1})'$ that

$$\begin{aligned}
\hat{\varSigma}_T(\Gamma_0) &= \hat{\sigma}^2(\Gamma_0)\|r_*'(Z_1'M_2X_1)^{-1}Z_1'M_2\|^2 = [1 + o_p(1)]\sigma^2 r_*'\bar{S}_2^{-1}Z_1'M_2M_2'Z_1(\bar{S}_2')^{-1}r_* \\
&= [1 + o_p(1)]\sigma^2 r_*'\bar{S}_2^{-1}(\tilde{Z}_1' + \alpha_* u(I - \Gamma_0'))M_2M_2'(\tilde{Z}_1 + (I - \Gamma_0')u\alpha_*')(\bar{S}_2')^{-1}r_*,
\end{aligned}$$

where we also used Theorem 4 and (17). From Lemma 8((i) and (iii)), we have

$$\mathbb{E}[u'(I - \Gamma_0')M_2M_2'(I - \Gamma_0)u] = [1 + o(1)]\mathbb{E}[\varepsilon'BB'\varepsilon] = [1 + o(1)]\sigma^2 \operatorname{tr}[B'B].$$

From $K/T \to 0$ we have $\frac{\operatorname{tr}(B^2+B'B)}{\operatorname{tr}(B'B)} \to 1$ and therefore

$$\frac{\hat{\varSigma}_T - \varSigma_T}{\varSigma_T} = \frac{2\sigma^2 r_*'\bar{S}_2^{-1}\alpha_* u(I-\Gamma_0)M_2M_2'\tilde{Z}_1 + \sigma^2(r_*'\bar{S}_2^{-1}\alpha_*)^2(\varepsilon'BB'\varepsilon - \mathbb{E}\varepsilon'BB'\varepsilon)}{\varSigma_T} + o_p(1). \quad (23)$$

Define $R = M_2'\tilde{Z}_1$, $\xi_1 = r_*'\bar{S}_2^{-1}\alpha_* u(I - \Gamma_0)M_2M_2'\tilde{Z}_1$, $\xi_2 = (r_*'\bar{S}_2^{-1}\alpha_*)^2[\varepsilon'BB'\varepsilon - \mathbb{E}\varepsilon'BB'\varepsilon]$.

$$\begin{aligned}
\mathbb{E}[\xi_1^2] &= C(r_*'\bar{S}_2^{-1}\alpha_*)^2\operatorname{tr}(R'B'BR) \leq C(r_*'\bar{S}_2^{-1}\alpha_*)^2\operatorname{tr}(R'R)\|B'B\|; \\
\frac{\xi_1}{\varSigma_T} &= O_p\left(\frac{r_*'\bar{S}_2^{-1}\alpha_*\sqrt{\operatorname{tr}(R'R)\|B'B\|}}{\varSigma_T}\right) \\
&= O_p\left(\sqrt{\frac{\|B'B\|}{\operatorname{tr}(B'B)}}\frac{(r_*'\bar{S}_2^{-1}\alpha_*)^2\operatorname{tr}(B'B) + \operatorname{tr}(R'R)}{\varSigma_T}\right) = O_p\left(\sqrt{\frac{\|B'B\|}{\operatorname{tr}(B'B)}}\right).
\end{aligned}$$

Lemma 7(ii) yields $\varepsilon'BB'\varepsilon - \mathbb{E}\varepsilon'BB'\varepsilon = O_P(\|B'B\|_F)$. Thus

$$\frac{\xi_2}{\varSigma_T} = O_p\left(\frac{\|B'B\|_F}{\operatorname{tr}(B'B)}\right) = O_p\left(\frac{\sqrt{\operatorname{tr}(B'B)\|B'B\|}}{\operatorname{tr}(B'B)}\right) = O_p\left(\frac{\|B\|}{\sqrt{\operatorname{tr}(B'B)}}\right).$$

Thus, by Lemma 8(iii), both terms in (23) are $O_p(1/\sqrt{T})$. $\square$

Proof of Theorem 6. Note that $r'\hat{\beta}^{\text{IV}}(\Gamma), r'\beta$ and $\hat{\Sigma}_T(\Gamma) = \hat{\sigma}^2(\Gamma)\|r'(X'(I-\Gamma)X)^{-1}X'(I-\Gamma)\|^2$ are invariant under the transformation Θ of Lemma 5. Thus, we may assume without loss of generality that Lemma 5((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 (21) with $B = D'(I-\Gamma)M_2$. Due to Gaussianity of the errors, $r_*'\bar{S}_2^{-1}\tilde{Z}_1'M_2\varepsilon = w'\varepsilon$ has a Gaussian distribution conditionally on $\tilde{X}$ with conditional variance $\sigma^2\|w\|^2 = \sigma^2\|r_*'\bar{S}_2^{-1}\tilde{Z}_1'M_2\|^2$. Below, we show that

$$\frac{\varepsilon'B\varepsilon}{\sigma^2\sqrt{\text{tr}(B^2)+\text{tr}(B'B)}} \Rightarrow N(0,1) \tag{24}$$

and that this term is asymptotically independent of the conditionally Gaussian term $w'\varepsilon$.

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

$$\begin{aligned}
\left|\varepsilon'P_wBP_w\varepsilon\right| &= \left(\frac{w'\varepsilon}{\|w\|}\right)^2\left|\frac{w'Bw}{w'w}\right| \le \|B\|\cdot\chi_1^2 = O_p(1); \\
\varepsilon'BP_w\varepsilon &= \frac{w'\varepsilon}{\|w\|}\frac{w'B\varepsilon}{\|w\|} = \frac{w'\varepsilon}{\|w\|}N(0,\frac{w'BB'w}{w'w}) = O_p(1).
\end{aligned}$$

Due to Lemma 8 we have $\frac{T}{\text{tr}(B^2)+\text{tr}(B'B)} = O(1)$. Thus

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

But $\frac{\varepsilon'M_wBM_w\varepsilon}{\sqrt{\text{tr}(B^2)+\text{tr}(B'B)}}$ is independent from $\frac{w'\varepsilon}{\|w\|} \sim N(0,\sigma^2)$. This implies that $\frac{\varepsilon'B\varepsilon}{\sqrt{\text{tr}(B^2)+\text{tr}(B'B)}}$ is asymptotically independent from the first term. Since $\text{tr}(B) = \sum_t B_{tt} = O_p(1)$:

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

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

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

and due to Lemma 8 the right-hand-side grows no slower than of order T. Since $\max_t |B_{t,t}| \leq \|B\| = O(1)$ and the operator norm of $B + B'$ is bounded, conditions (i)–(iv) of Corollary A2.8 in Sølvsten (2020) hold. Therefore (24) holds and we have asymptotic Gaussianity.

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

$$\frac{\hat{\sigma}^2(\Gamma)\|r'(X'(I-\Gamma)X)^{-1}X'(I-\Gamma)\|^2}{\sigma^2\|r_*'\bar{S}_2^{-1}\tilde{Z}_1'M_2\|^2 + \sigma^4(r_*'\bar{S}_2^{-1}\alpha_*)^2\operatorname{tr}(B'B)} \to^p 1.$$

Pre-multiplying this estimator by $1 + \psi = \frac{|\operatorname{tr}(B^2)| + \operatorname{tr}(B'B)}{\operatorname{tr}(B'B)}$ guarantees that the resulting quantity asymptotically weakly exceeds $\varSigma_T = \sigma^2 \sum_t w_t^2 + \sigma^4(r_*'\bar{S}_2^{-1}\alpha_*)^2 \operatorname{tr}(B^2 + B'B)$. $\square$