Identification and Estimation of Structural Factor Models with External Instruments[*]

Ze-Yu Zhong

Monash University

Department of Econometrics and Business Statistics

Dated: September 2024

Abstract

We develop a new estimator for impulse response functions in structural factor models with the use of external instruments. In contrast to a traditional structural vector autoregression (SVAR), the use of a factor structure naturally deals with the non-fundamentalness and singularity issues that plague SVARs. Additionally, our generalized method of moments approach naturally allows for the joint use of multiple instruments to sharpen inference, an overidentification test for the joint validity of instruments, and an automatic moment selection procedure to select the correct instruments. Simulation results show the improvement in the estimation accuracy of impulse response functions when more than one valid instrument is used, and confirm the size and consistency of theory. We apply the proposed methodology to estimate the effects of a monetary policy shock using a U.S. macroeconomic dataset with the use of popular monetary policy instruments. The results show these monetary policy instruments are all jointly valid, and that their joint use can result in more accurate and reasonable estimates of the impulse response functions.

[*] I acknowledge that this research was supported in part by the Monash eResearch Centre and eSolutions Research Support Services through the use of the MonARCH HPC Cluster. I thank my supervisors Bonsoo Koo and Benjamin Wong for their support, and Xu Han for their comments and code.

1 Introduction

Since the seminal paper of Sims (1980), structural vector autoregressive (SVAR) models have remained a popular and indispensable tool for identifying and estimating the effects of macroeconomic shocks. A wide literature on SVAR modelling now exists (see the survery by Kilian and Lütkepohl, 2017), which has documented three challenging limitations that can preclude practitioners from estimating the true impulse responses to structural shocks. First, SVAR models often suffer from nonfundamentalness due to their inherent inability to include large information sets in a small equation system. Indeed, policymakers have access to information sets which span over literally hundreds of variables, and it is thus extremely difficult for the included variables in an SVAR system to span the space spanned by the true structural shocks, (Sims, 1992; Bernanke et al., 2005). Second, economic theory often implies the SVAR system specified is singular; that is, the number of variables included are driven by a smaller number of structural shocks. This point is often overlooked in the literature, as the number of structural shocks is usually implicitly set to be the same number of variables included as part of the specification. Third, the validity of the identification strategy required for an SVAR may be difficult to formally test for, in addition to each strategy typically requiring the development of a specific estimation—and therefore asymptotic theory—for inferential results. This latter point is particularly important, as this means that strategies such as the inclusion of factors (as in the case of a factor-augmented vector autoregression model) require the careful adaptation of existing theory in order to deal with nuisance parameters.

Ideally, practitioners would employ an estimation framework that addresses all three limitations in a holistic way, as any one of these precludes the ability to recover the true impulse responses. Instead however, these limitations have been typically dealt with in distinct, complex, and oftentimes opposing ways. For example, nonfundamentalness is typically addressed by including more variables that need to be justified to represent theoretical constructs (such as using GDP to represent “economic activity”), or factors estimated from a large dataset in a factor-augmented vector autoregression (FAVAR), (e.g. Bernanke et al., 2005). However, as noted earlier, macroeconomic theory often implies that a small number structural shocks drive a large number of variables, (e.g. Sargent and Sims, 1977; Geweke, 1977) but including more variables therefore increases the chance on the specified system being singular. This problem is not even resolved by the FAVAR model. As we will show, most FAVARs do not effectively distinguish between the so-called dynamic factors and the static factors (i.e. the stacked lags of dynamic factors) whose dimension is typically larger. This directly implies that the static factors, and therefore the FAVAR system, are necessarily singular, which requires special treatment, often in the form of an additional dimensional reduction step, (e.g. Chan et al., 2020). Indeed, the direct augmentation of factors to a typical SVAR system has been noted by Stock and Watson (2005) to be inconsistent with the primitives of the dynamic factor model (DFM) used to compute factors in the first place, and it is therefore unclear whether this could produce a compelling basis for structural identification at all. Furthermore, these issues are compounded with the often wide, yet controversial, choices of identification strategies available to practitioners. Examples include short-run/long-run exclusion restrictions, sign restrictions, and identification via heteroskedasticity, (Stock and Watson, 2016). Among these, however, the method of using external instruments (or proxies) has increasingly gained popularity for their parsimonious set of identifying assumptions and ability to incorporate further external information into models. However, formal extensions of external instruments to a data-rich environment generally remain rare, and still suffer from limitations.

Our contribution to the literature is to provide a framework that addresses these three challenging limitations in a holistic approach. To do so, we propose the use of a structural factor model (SFM) which naturally deals with the problems of nonfundamentalness and covariance singularity, in conjunction with the use of multiple external instruments, which naturally allow for testing and selecting valid identifying restrictions. The SFMs we work with were introduced by Stock and Watson (2005), who were inspired by their success in macroeconomic forecasting. Unlike FAVAR models, SFMs are directly formulated from the factor structure and aim to combine the attractive features of large dimensional factor models and existing identification strategies employed in SVARs. Since their introduction, they have received increasing attention in the literature for their ability to estimate more reasonable and efficient impulse responses, which have been argued to be a direct consequence of their ability to parsimoniously summarise large information sets - something that is generally impossible or difficult to do within a standard SVAR framework. Similar to FAVAR models, the factor structure in an SFM implies the appearance of many extra nuisance parameters, necessitating the development of estimator and relevant asymptotic theory that is often specific to the identification scheme. To this end, we contribute to the literature by developing an asymptotic theory that explicitly acknowledges the random rotation problem included in factor models, similar to Bai and Ng (2006) and Yamamoto and Hara (2022). Furthermore, unlike existing attempts within the SFM literature, we establish the validity of identification and estimation of impulse responses with the use of multiple different instruments via the use of a generalised method-of-moments framework. Altogether, these provide the familiar theoretical basis for us to develop analytical formulas for statistical inference, such as confidence intervals, and overidentification and automatic moment selection procedures for testing/choosing valid instruments. A Monte Carlo experiment shows that the resulting estimators, tests, and selection criteria exhibit good finite sample performance.

Our work is related to the broader SVAR literature that achieves identification with external instruments, and structural factor models that attempt to extend existing identification strategies with factor models. For the former, the literature on identification in SVAR models is extensive, for which a comprehensive summary of mainstream identification approaches is provided by Kilian and Lütkepohl (2017). Since then, within the external instruments approach there have been further developments. Stock and Watson (2018) provide a comparison of local projection instrument variable (LP-IV) and SVAR-IV estimators, and show that SVAR-IV estimators do not require the strict lead-lag exogeneity assumption of LP-IV. Montiel Olea et al. (2021) derive the asymptotic theory for SVAR-IV. Cheng et al. (2021) derive the asymptotic theory for a generalised method-of-moments estimator that is robust to non-stationarity, but do not pursue overidentification or moment selection procedures. Schlaak et al. (2023) combine identification via heteroskedasticity and external instruments to sharpen inference. However, their framework is focused on using heteroskedasticity to achieve exact identification; overidentification using external instruments then proceeds in a proxy-SVAR framework. Importantly, their empirical study still focuses on using one instrument at a time. For the latter, the literature for SFMs is less developed and has typically focused on older identification strategies. Stock and Watson (2005) and Forni and Gambetti (2010a) use an SFM and employ a slow-fast identification, though neither provide formal theoretical justification. Forni et al. (2009) show that the presence of a factor structure in the data typically implies nonfundamentalness in fixed dimensional SVARs. Han (2015) and Han (2018) develops inferential theory for the identification of impulses using a diverging and finite number of zero restrictions respectively. Yamamoto and Hara (2022) develop inferential theory for the identification of impulse responses using heteroskedasticity in a FAVAR model. Forni and Gambetti (2010b) utilise a SFM with sign restrictions, the possibility of which is investigated by Gafarov et al. (2018).

To the best of our knowledge, there exist only a handful of papers that combine a structural factor model with external instrument identification. Stock and Watson (2012) focus on one instrument at a time. Stock and Watson (2016), in their review, propose the use of a normalisation scheme that allows for direct application of an SVAR identification scheme with an SFM. Both of these only provide an estimation algorithm with little formal theoretical treatment of the proposed estimators. Han (2024) proposes a unifying framework for the global identification of structural impulse responses in factor models, but assumes that the number of static factors is equal to the number of primitive shocks, ignoring singularity issues that could occur. Our theory differs from the pre-existing literature in that we provide a formal theoretical treatment of the identification of impulse responses through the use of a factor structure that summarises a data-rich environment, a latent factor process that distinguishes between the static factor and primitive shocks, and a generalised method-of-moments approach which allows for the joint use of multiple instruments and leads to standard overidentification and instrument selection procedures to ensure that the identification conditions are valid. These features of our proposed framework allow us to respectively deal with the problems of nonfundamentalness, covariance singularity, and identification issues that plague typical SVAR models, in a holistic fashion.

In an empirical application on quarterly U.S. macroeconomic data, we apply the proposed method to study the dynamic causal effects of a monetary policy shock, and the validity of many popular monetary policy instruments proposed by the literature. We find evidence that all the monetary policy instrument considered are jointly valid, and that their joint use leads to more efficient and reasonable impulse responses. In particular, we show that using one instrument at a time is more prone to recovering puzzling responses.

The rest of the paper is organised as follows. Section 2 lays out the model setup, identification strategy, and estimation. Section 3 presents the asymptotic theory. Section 4 conducts the Monte Carlo study to confirm finite sample behaviour. Section 5 presents the empirical application. Section 6 concludes. All proofs are relegated to the Appendices. For notation, $P_Z = Z(Z^\top Z)^{-1} Z^\top$ and $M_Z = I - P_Z$ denote the projection and residual maker matrices for any matrix Z, respectively, $\|Z\| = \left[tr(Z^\top Z)\right]^{1/2}$ denotes the Euclidean norm, and $\xrightarrow{p}$ and $\xrightarrow{d}$ denote convergence in probability and distribution, respectively.

2 Identification of Dynamic Responses in Structural Factor Models

2.1 Model Setup

Consider the following structural model for $t = 1, \dots, T$,

$$X_t = \Lambda F_t + e_t, \tag{2.1}$$

$$F_t = \sum_{j=1}^{p} \Phi_j F_{t-j} + G\eta_t, \tag{2.2}$$

$$\eta_t = A\zeta_t, \tag{2.3}$$

where $X_t = [x_{1t}, \dots, x_{Nt}]^\top$ is an N-dimensional vector, F_t is an r-dimensional set of unobserved factors, Λ is the corresponding $N \times r$ factor-loading matrix, and $e_t = [e_{1t}, \dots, e_{Nt}]^\top$ is an N-dimensional idiosyncratic error term. The matrix G is an $r \times q$ matrix of rank q which maps the q-dimensional reduced form shocks η_t to the lags of the factors, ζ_t are the structural shocks subject to the identification condition $E[\zeta_t \zeta_t^\top] = I_q$, and A is a $q \times q$ nonsingular matrix. Unlike many existing studies, we set focus on the case of $q \leq r$ to allow for dynamic factors. This is important, because the case of $q < r$ corresponds to a singular covariance structure in the static factors, rendering the FAVAR, and even many existing SFM approaches untenable. The assumption of stationarity in F_t implies

$$(I_r - \Phi_1 L - \cdots - \Phi_p L^p)^{-1} = \sum_{s=1}^{\infty} \Psi_s L^s, \tag{2.4}$$

where Ψ_s is the coefficient matrix of the vector-moving average representation of Equation (2.2).

Let $\mathcal{F}_t = \left(F_t^\top, \ldots, F_{t-p}^\top\right)^\top$ collect the lags of F_t and $\Phi = [\Phi_1, \ldots, \Phi_p]$ collect the corresponding coefficient matrices. Plugging Equations (2.2) and (2.3) into Equation (2.1), we have

$$\begin{aligned} X_t &= \Pi \mathcal{F}_t + \Theta \eta_t + e_t \qquad (2.5) \\ &= \Pi \mathcal{F}_t + \Gamma \zeta_t + e_t, \qquad (2.6) \end{aligned}$$

where $\Pi = \Lambda\Phi$, $\Theta = \Lambda G$ and $\Gamma = \Theta A$. The matrix representations of Equations (2.5) and (2.6) follow as

$$\begin{aligned} X &= \mathcal{F}\Pi^\top + \eta\Theta^\top + e \\ &= \mathcal{F}\Pi^\top + \zeta\Gamma^\top + e \qquad (2.7) \end{aligned}$$

where $X = [X_{p+1}, \ldots, X_T]^\top$, $\mathcal{F} = [\mathcal{F}_{p+1}, \ldots, \mathcal{F}_T]^\top$, $\eta = [\eta_{p+1}, \ldots, \eta_T]^\top$, $\zeta = [\zeta_{p+1}, \ldots, \zeta_T]^\top$, and $e = [e_{p+1}, \ldots, e_T]^\top$.

The OLS estimators for $\lambda_i^\top$ and Λ are, respectively,

$$\begin{aligned} \widehat{\lambda}_i &= \frac{1}{T} \sum_{t=1}^{T} \widehat{F}_t X_{it}, \\ \widehat{\Lambda} &= \frac{1}{T} \sum_{t=1}^{T} X_t \widehat{F}_t^\top. \qquad (2.8) \end{aligned}$$

To estimate the reduced form shocks, first note that the dynamic factor model in Equation (2.7) implies a factor structure in the reduced form shocks, i.e. $X - \mathcal{F}\Pi^\top = \eta\Theta^\top + e$ itself exhibits a factor structure. Let

$$\widehat{X} = M_{\widehat{\mathcal{F}}} X \qquad (2.9)$$

be a corresponding estimate of $\eta\Theta^\top + e$. It follows that the reduced form shocks can then be estimated via a second-stage principal components estimator. We set the estimated reduced shocks $\widehat{\eta}$ equal to $\sqrt{T-p}$ times the eigenvectors corresponding to the first q eigenvalues of the $(T-p) \times (T-p)$

covariance matrix $\widehat{X}\widehat{X}^\top$. The OLS estimator for G can be computed as

$$\widehat{G} = \widehat{F}^\top \widehat{\eta} \left(\widehat{\eta}^\top \widehat{\eta}\right)^{-1} = \frac{1}{T-p} \sum_{t=p+1}^{T} \widehat{F}_t \widehat{\eta}_t^\top, \tag{2.10}$$

which follows because $\widehat{\mathcal{F}}$ and $\widehat{\eta}$ are orthogonal by design, and $\widehat{\eta}^\top \widehat{\eta}/(T-p) = I_q$ by eigenidentity. The estimator for Θ can be computed as

$$\widehat{\Theta} = \widehat{\Lambda}\widehat{G}, \tag{2.11}$$

where we additionally use $\widehat{\theta}_i$ to denote the transposition of the ith row of $\widehat{\Theta}$. The estimator for Φ via OLS is given by

$$\widehat{\Phi} = \widehat{F}^\top \widehat{\mathcal{F}} \left(\widehat{\mathcal{F}}^\top \widehat{\mathcal{F}}\right)^{-1} = \left(\sum_{t=p+1}^{T} \widehat{F}_t \widehat{\mathcal{F}}_t\right) \left(\widehat{\mathcal{F}}^\top \widehat{\mathcal{F}}\right)^{-1}. \tag{2.12}$$

Given $\widehat{\Phi}$, the estimates for $\widehat{\Psi}_s$, $s = 1, 2, \ldots$ follow by inverting the lag polynomial in Equation (2.4).

Remark. *We focus on the principal components of $\widehat{X}\widehat{X}^\top$ to estimate the reduced form shocks. An alternative way to estimate η_t is to conduct a spectral decomposition of the variance of the residuals $\widehat{\varepsilon}_t = \widehat{F}_t - \widehat{\Phi}\widehat{\mathcal{F}}$ (see, e.g. Forni et al., 2009; Forni and Gambetti, 2010a). Specifically, let $\widehat{\Sigma}_\varepsilon$ be the sample covariance matrix of $\widehat{\varepsilon}$, $\widehat{\mathcal{D}}$ be a diagonal matrix consisting of the first q eigenvalues of $\widehat{\Sigma}_\varepsilon$ in descending order, and $\widehat{\mathcal{S}}$ be the corresponding associated eigenvectors. Then, $\breve{\eta}_t \equiv \widehat{\mathcal{D}}^{-1/2}\widehat{\mathcal{S}}\widehat{\varepsilon}_t$ is an alternative estimator for η_t because $\widehat{\Sigma}_\varepsilon$ is of rank q, asymptotically. Due to the use of a principal components fit, it is likely that the theory developed for this paper can also be adapted for $\breve{\eta}_t$. The simulation results of Han (2018) show that $\widehat{\eta}_t$ tends to produce more accurate estimates for η_t as measured by trace R^2 statistics, and we therefore leave the use of $\breve{\eta}_t$ to future research.*

2.2 Identification and Estimation with External Instruments

Suppose, without loss of generality, that we are interested in the effects of the first structural shock. The preceding model setup dictates that the impulse response function (IRF) to the entire N panel of time series X_t to a one unit increase in the first structural shock is given by

$$\frac{\partial X_t}{\partial \zeta_{1,t-s}} = \Lambda \Psi_s G a_1, \tag{2.13}$$

where a_1 denotes the first column of A, where its columns are partitioned as $A = \begin{bmatrix} a_1 & \dots & a_q \end{bmatrix}$. In the special case of $s = 0$, we set $\Psi_s = I_r$, so the contemporaneous response simplifies to

$$\frac{\partial X_t}{\partial \zeta_{1,t}} = \Lambda G a_1 = \Theta a_1. \tag{2.14}$$

The estimators for Λ, Ψ_s and G are described earlier, and thus it remains to find an appropriate estimator for a_1 to compute the IRF.

It is well known that principal components estimators are only consistent up to a rotation. Specifically, the principal components-based estimator $\widehat{\eta}_t$ is only able to estimate its unobserved counterpart η_t up to a rotation, which we denote as H_η, i.e. $\widehat{\eta}_t$ estimates $H_\eta^\top \eta_t$. The presence of this rotational basis H_η will generally affect the distribution of the individual components, which enter into the expression for the impulse response functions. We show that, however, the resulting estimators of the impulse response functions are not affected by this rotation, and thus the identification of the impulse response functions themselves is not affected. Identification proceeds by requiring $q-1$ restrictions to identify a_1 (assuming that the first element is fixed to unity). Unlike the typical SVAR-IV case, the use of principal components estimator $\widehat{\eta}_t$, which recovers $H_\eta^\top \eta_t$, implies that we are instead identifying $a_1^* = H_\eta^\top a_1$.

We are interested in identifying a_1^* with external instruments $Z_t \in R^k$, which satisfy i) $E(Z_t \zeta_{1t}) = \alpha \neq 0_k$, and ii) $E(Z_t \zeta_{jt}) = 0_k$ for $j \neq 1$, which are the instrument relevance and exogeneity conditions. We emphasise that these conditions are with respect to only the *contemporaneous* shocks - these SVAR-IV conditions permit the instruments to be correlated with lagged values of the non-target shocks, and is thus far less restrictive compared to the Local Projection (LP-IV) approach as noted by Stock and Watson (2018). Under these conditions, the instruments satisfy

$$\begin{aligned} E(H_\eta^\top \eta_t Z_t^\top) &= E(H_\eta^\top A \zeta_t Z_t^\top) \\ &= H_\eta^\top a_1 \alpha^\top \\ &= a_1^* \alpha^\top \in R^{q \times k}, \end{aligned} \tag{2.15}$$

and are thus able the identify a_1^* up to a scale. The case of $k = 1$ instrument corresponds to $q - 1$ restrictions and suffices to just identify a_1^*; the system is overidentified if $k > 1$. In the traditional SVAR setting, the reduced form shocks are estimated without the effect of H_η and thus the moment conditions are $E(\eta_t Z_t^\top) = a_1 \alpha^\top$; estimation then proceeds by regressing each reduced form shock on the first using the instrument(s) Z_t via two stage least squares (2SLS) as in Ramey (2016), or a generalised method-of-moments approach as in Cheng et al. (2021).

Without loss of generality, we normalise the first element of a_1^* to be 1.[1] This allows us to remove the constant 1 and define the parameter

$$\delta = \left[a_{12}^*, \ldots, a_{1q}^*\right]^\top \in R^{q-1}. \tag{2.16}$$

With $a_1^* = 1$, Equation (2.15) is therefore equivalent to the moment conditions

$$\begin{aligned} &E\left[\left((H_\eta^\top \eta)_{-1t} - \delta \eta_{1t}^*\right) \otimes Z_t\right] \\ =&E\left[\left(\eta_{-1t}^* - \delta \eta_{1t}^*\right) \otimes Z_t\right] = \mathbf{0} \in R^{k(q-1)}, \end{aligned} \tag{2.17}$$

where η_{1t}^* is the first element of $H_\eta^\top \eta_t$ and η_{-1t}^* is the rest of $H_\eta^\top \eta_t$ with η_{1t}^* removed. Let $\widehat{\eta}_{1t}$ and $\widehat{\eta}_{-1t}$ denote the principal components-based estimated counterparts of η_{1t}^* and η_{-1t}^*, respectively. We estimate δ by minimising the generalised method-of-moments (GMM) criterion

$$\mathcal{Q}_T(\delta) = \bar{g}_T(\delta)^\top W_T \bar{g}_T(\delta) \tag{2.18}$$

[1] This corresponds to an additional scale assumption that $H_{\eta,1}^\top \alpha_1 = 1$, and is analogous to the innocuous identification condition of setting the first element of a_1 to one in the case of η_t being observed or estimated without the effects of H_η, as is the case in a traditional SVAR setting. In practice, any normalisation can be used afterwards, such as the unit-effect normalisation.

using the empirical moments

$$\bar{g}_T(\delta) = \frac{1}{T-p}\sum_{t=p+1}^{T}[(\tilde{\eta}_{-1,t} - \delta\tilde{\eta}_{1,t}) \otimes Z_t] \tag{2.19}$$

and a weighting matrix W_T. The first-order condition yields the GMM estimator

$$\widehat{\delta} = (\mathcal{A}_T W_T \mathcal{A}_T^\top)^{-1}\mathcal{A}_T W_T \mathcal{G}_T \tag{2.20}$$

where

$$\mathcal{A}_T = I_{q-1} \otimes \left(\frac{1}{T-p}\sum_{t=p+1}^{T}\tilde{\eta}_{1,t}Z_t^\top\right), \quad \text{and} \quad \mathcal{G}_T = \frac{1}{T-p}\sum_{t=p+1}^{T}\left(\tilde{\eta}_{-1,t} \otimes Z_t\right). \tag{2.21}$$

If $W_T = I_{q-1} \otimes \left(\frac{1}{T-p}\sum_{t=p+1}^{T} Z_t Z_t^\top\right)^{-1}$ then $\widehat{\delta}$ corresponds to the equation by equation 2SLS estimator that is typically considered by the literature. By defining V_δ as the variance covariance matrix of $\mathcal{G}_T$, an optimal two-step GMM estimator $\widehat{\delta}^o$ can be estimated as follows. In the first step, we use either $I_{(q-1)k}$ or $I_{q-1} \otimes \left(T^{-1}\sum_{t=1}^{T} Z_t Z_t^\top\right)^{-1}$ as the weighting matrix and compute the GMM estimator $\tilde{\delta}$. In the second step, we compute the feasible weight estimate as

$$\widehat{V}_\delta = [\mathbb{S}_{\widehat{\delta}} \otimes I_k]\widehat{\Sigma_i^{(1)}}[\mathbb{S}_{\widehat{\delta}} \otimes I_k]^\top, \tag{2.22}$$

where $\mathbb{S}_{\widehat{\delta}}$ is a $(q-1) \times q$ matrix such that $\mathbb{S}_{\widehat{\delta}}\widehat{\eta}_t = \widehat{\eta}_{-1t} - \widehat{\delta}\widehat{\eta}_{1t}$, which by definition is equal to $[\widehat{\delta} \vdots I_{q-1}(1 : q-1)]$ where $I_{q-1}(1 : q-1)$ collects the $q-1$ matrix of I_{q-1}, and $\widehat{\Sigma_i^{(1)}}$ is a feasible estimate of the variance of the instruments, detailed in Section 3.4. Note that, due to the effects of the generated regressor, this is different to the implicit 2SLS weighting matrix $I_{q-1} \otimes \left(\frac{1}{T}\sum_{t=1}^{T} Z_t Z_t^\top\right)^{-1}$, even in the absence of conditional heteroskedasticity.

2.3 Comparison With Existing Approaches in Factor Models

The identification scheme in this paper uses the information from an external instrument of the structural shock, which is widely considered to be parsimonious in terms of identifying assumptions. The literature on combining identification with external instruments with a factor structure has seen increasing, though still limited, attention.

This approach was initially proposed by Stock and Watson (2012). However, their methodology in implementing the identification condition differs somewhat - instead of regressing the reduced form shocks on each other in a typical IV regression as we have, they opt to regress the instrument on all remaining reduced form shocks. Although both approaches are consistent at recovering the same structural shock (see Montiel Olea et al., 2021, for a proof in an SVAR-IV context), our adoption of a generalised method-of-moments framework allows researchers to easily adapt the wide array of tools within that literature. This can be seen in how Stock and Watson (2012) only report the estimated structural shock as estimated by each instrument one at a time, and investigate joint validity of instruments by reporting their correlations, an approach that precludes the ability to formally *test* joint validity. The theoretical validity of this approach in the context of factor models is also unclear; Stock and Watson (2016) provide only an unjustified bootstrap algorithm to calculate inferential quantities such as confidence intervals.

Stock and Watson (2016) justify this by proposing a “named factor” normalisation which allows for the direct implementation of existing SVAR identification methods. However, this requires that the space of the innovations to the first r common components span the space of the innovations of the remaining variables. Although the theoretical assumptions this additionally imposes are mild, in practice this is sensitive to the choice of named factor variables; one needs to ensure that the set of named variables 1) be sufficiently heterogeneous, 2) are sufficiently representative of the remaining groups of variables and 3) have innovations to their common components that sufficiently span the space of the innovations to the factors. As explained by Han (2024), this normalisation is mostly applicable in cases when the large dataset employed does not adequately capture the true structural shock, such as in the case of oil shocks.

3 Asymptotic Theory

3.1 Assumptions

To analyse the properties of the proposed estimators, we make the following assumptions.

Assumption 1. *There exists a positive constant $M < \infty$ such that:*

a) $E\|f_t\|^4 < M$, $\frac{1}{T}\sum_{t=1}^{T} f_t f_t^\top \xrightarrow{p} \Sigma_F$, *and* $\frac{1}{T}\sum_{t=p+1}^{T} \mathcal{F}_t\mathcal{F}_t^\top \xrightarrow{p} \Sigma_{\mathcal{F}}$ *for some positive definite matrices* Σ_F *and* $\Sigma_{\mathcal{F}}$.

b) $E(\zeta_t\zeta_t^\top) = I_q$, $E\|\zeta_t\|^4 < M$, $E(\zeta_s\zeta_t^\top) = 0$ *for any* $s \neq t$, *and* $\frac{1}{T-p}\sum_{t=p+1}^{T} \zeta_t\zeta_t^\top \xrightarrow{p} I_q$.

c) $E\left\|\frac{1}{\sqrt{T}}\sum_{t=p+1}^{T} \zeta_t\mathcal{F}_t^\top\right\|^2 < M$.

Assumption 2. *There exists a positive constant M such that:*

a) $E\|\lambda_i\|^4 \le M$, $\left\|\Lambda^\top\Lambda/N\right\| - \Sigma_\Lambda \xrightarrow{p} 0$ *for some* $\Sigma_\Lambda > 0$.

b) $\mathrm{rank}(G) = q$, $\|G\| \le M$, *and* $\|\Phi\| \le M$.

c) All of the roots of $|I_q - \Phi_1 L - \cdots - \Phi_p L^p| = 0$ *are outside the unit circle.*

d) The matrices $\Sigma_F\Sigma_\Lambda$ *and* $G^\top\Sigma_\Lambda G$ *have distinct eigenvalues.*

Assumption 3. *There exists some positive constant $M < \infty$ such that for all N and T:*

a) $E(e_{it}) = 0, E|e_{it}|^8 \le M$.

b) $E(e_s^\top e_t/N) = E(N^{-1}\sum_{i=1}^{N} e_{is}e_{it}) = \gamma_N(s,t)$, $|\gamma_N(s,s)| \le M$ *for all* s, *and*
$T^{-1}\sum_{t=1}^{T}\sum_{s=1}^{T} |\gamma_N(s,t)| \le M$.

c) $E(e_{it}e_{jt}) = \tau_{ij,t}$, *with* $|\tau_{ij,t}| < \tau_{ij}$ *for some* τ_{ij} *and for all* t. *In addition,*
$N^{-1}\sum_{i=1}^{N}\sum_{j=1}^{N} |\tau_{ij}| \le M$.

d) $E(e_{it}e_{js}) = \tau_{ij,ts}$, *and* $(NT)^{-1}\sum_{i=1}^{N}\sum_{j=1}^{N}\sum_{t=1}^{T}\sum_{s=1}^{T} |\tau_{ij,ts}| \le M$.

e) For every (t,s), $E\left|N^{-1/2}\sum_{i=1}^{N}[e_{is}e_{it} - E(e_{is}e_{it})]\right|^4 \le M$.

Assumption 4. *The variables* $\{\lambda_i\}$, $\{\zeta_t\}$, *and* $\{e_{it}\}$ *are mutually independent groups.*

Assumption 5. *There exists an $M < \infty$ such that for all T and N, and for every $t \le T$ and $i \le N$ such that:*

a) $\sum_{s=1}^{T} |\gamma_N(s,t)| \le M$.

b) $\sum_{k=1}^{N} |\tau_{ki}| \leq M$.

Assumption 6. *There exists an $M < \infty$ such that for all N and T:*

a) *For each t,* $E\left\| \frac{1}{NT} \sum_{s=1}^{T} \sum_{k=1}^{N} F_s[e_{ks}e_{kt} - E(e_{ks}e_{kt})] \right\|^2 \leq M$, *and* $E\left\| \frac{1}{NT} \sum_{s=p+1}^{T} \sum_{k=1}^{N} \zeta_s[e_{ks}e_{kt} - E(e_{ks}e_{kt})] \right\|^2 \leq M$.

b) $E\left\| \frac{1}{\sqrt{NT}} \sum_{t=1}^{T} \sum_{k=1}^{N} F_t \lambda_k^\top e_{kt} \right\|^2 \leq M$.

c) $E\left\| \frac{1}{\sqrt{TN}} \sum_{t=p+1}^{T} \sum_{i=1}^{N} \lambda_i e_{i,t-j} \mathcal{F}_t^\top \right\|^2 \leq M$ *and* $E\left\| \frac{1}{\sqrt{TN}} \sum_{t=p+1}^{T} \sum_{i=1}^{N} \lambda_i e_{i,t-j} \zeta_t^\top \right\|^2 \leq M$ *for* $j = 0, 1, \ldots, p$.

d) *For each t,* $E\left\| \frac{1}{\sqrt{N}} \sum_{i=1}^{N} \lambda_i e_{it} \right\|^2 \leq M$.

e) *For each i,* $E\left\| \frac{1}{\sqrt{T}} \sum_{t=1}^{T} F_t e_{it} \right\|^4 \leq M$ *and* $E\left\| \frac{1}{\sqrt{T}} \sum_{t=p+1}^{T} \zeta_t e_{it} \right\|^4 \leq M$. *For each i and for* $j = 1, \ldots, p$, $E\left\| \frac{1}{\sqrt{T}} \sum_{t=p+1}^{T} F_{t-j}^\top e_{it} \right\|^4 \leq M$.

Assumption 7. *For $i = 1, \ldots, N$,*

$$\frac{1}{\sqrt{T}} \begin{bmatrix} \operatorname{vec}\left(Z^\top \eta - E(Z^\top \eta)\right) \\ F^\top e_i \\ \operatorname{vec}\left(\mathcal{F}^\top \eta\right) \end{bmatrix} \xrightarrow{d} N(\mathbf{0}_{(qk+r+rpq)\times 1}, \Sigma_i).$$

Assumption 8. *The structural shock ζ_t is linked to the reduced form error by the linear transformation $\eta_t = A\zeta_t$, for some nonsingular matrix A, and $E(Z_t \zeta_t^\top) = [\alpha, 0_{k\times(q-1)}]$, where $\alpha \neq 0_k$.*

Assumptions 1 to 6 are either straight from, or slight modifications of, Assumptions A-G of Bai (2003) and Assumptions 1-6 of Han (2018). Assumption 1 (a) regulates the moments of the static factors. The positive definiteness of $\Sigma_{\mathcal{F}}$ is the same as Assumption A10 of Amengual and Watson (2007). Assumption 1 (b) restricts the structural shocks to be serially uncorrelated and have an identity covariance matrix. Assumption 1 (c) is not restrictive because the structural shocks ζ_t and lags of F_t are commonly assumed to be uncorrelated in the VAR literature. Assumption 2 (a) follows from Assumption B of Bai (2003). Assumption 2 (d) is similar to Assumption G of Bai (2003), and ensures the existence of the probability limits of the rotation matrices H_F and H_η. Assumptions 3 and 5 allows for weak serial and cross-sectional correlation in the errors, corresponding to Assumptions C and E of Bai (2003). Assumption 4 is similar to Assumption D of Bai and Ng (2004). Assumption 6 is not stringent because all of the sums involve zero mean random variables. It is close to Assumption F of Bai (2003) and Assumption 6 of Han (2015). Assumption 7 are central limit theorems which can be obtained under primitive assumptions, (e.g. Theorem 5.15 or Theorem 7.1.2 of White, 1984; Brockwell and Davis, 1991, respectively). Assumption 8 formalises the instrument relevance and exogeneity conditions with respect to contemporaneous shocks. Note that $\zeta_t = A^{-1}\eta_t$ implies $E(\zeta_t|Z_{t-1}, Z_{t-2}, \dots) = 0$, i.e. that the structural shock is uncorrelated with lags of the instruments. This is consistent with the structural VAR literature, where the structural shocks are interpreted as unanticipated, and therefore, unpredictable conditional on historical information. Similar to Stock and Watson (2018), we allow Z_t to be correlated with lags of ζ_t, which is a much looser condition than is required for a local projection (LP-IV) approach.

It is well known that the principal components estimator is only able to estimate the true factors up to a rotational basis. Let $X^0 \equiv [X_1, \dots, X_T]^\top$ be the full data matrix. We define the following normalisation bases for F_t and η_t

$$H_F = \left(\frac{\Lambda^\top \Lambda}{N}\right)\left(\frac{F^\top \widehat{F}}{T}\right)\widehat{V}_F^{-1} \quad \text{and} \tag{3.1}$$

$$H_\eta = \left(\frac{\Theta^\top \Theta}{N}\right)\left(\frac{\eta^\top \widehat{\eta}}{T-p}\right)\widehat{V}_\eta^{-1} \tag{3.2}$$

where $\widehat{V}_F$ is an $r \times r$ diagonal matrix consisting of the first r largest eigenvalues of $X^0 X^{0\top}/(NT)$ in descending order, and $\widehat{V}_\eta$ is a $q \times q$ diagonal matrix consisting of the first q eigenvalues of $\widehat{X}\widehat{X}^\top/(N(T-p))$ in descending order. Their corresponding probability limits are

$$\bar{H}_F = \operatorname{plim} H_F \quad \text{and} \tag{3.3}$$

$$\bar{H}_\eta = \operatorname{plim} H_\eta, \tag{3.4}$$

which can be shown by Lemma A3 and Proposition 1 of Bai (2003).

Analogously, we can define $H_{\mathcal{F}}$ such that $\widehat{\mathcal{F}}$ is a consistent estimator for $H_{\mathcal{F}}^\top \mathcal{F}_t$. Recall that $\widehat{\mathcal{F}}_t = \left[\widehat{F}_{t-1}^\top, \dots, \widehat{F}_{t-p}^\top\right]^\top$ and $\mathcal{F} = \left[F_{t-1}^\top, \dots, F_{t-p}^\top\right]^\top$. It follows that the rotational basis $H_{\mathcal{F}}$ can be defined as

$$H_{\mathcal{F}} \equiv I_p \otimes H_F \tag{3.5}$$

so that $\widehat{\mathcal{F}}$ is a consistent estimator for $H_{\mathcal{F}}^{\top}\mathcal{F}_t$. The probability limit of $H_{\mathcal{F}}$ is

$$\bar{H}_{\mathcal{F}} = I_p \otimes \bar{H}_F. \tag{3.6}$$

Remark. *We focus on the setup where the static factors F_t are unobserved. If some of the factors are treated as observed, then the model becomes a factor-augmented VAR (FAVAR) model, (e.g. Bernanke et al., 2005; Bai et al., 2016). Specifically, in the FAVAR setup, $\widehat{F}_t$ can be constructed by stacking the observed factors (regressors) and the estimated factors. Generally, the introduction of observed factors results in $\frac{1}{T}\sum_{t=1}^{T}\widehat{F}_t\widehat{F}_t^{\top}$ no longer being an identity matrix; the corresponding loading matrix should then be estimated by least squares as $\widehat{\Lambda} = X^{\top}\widehat{F}\left(\widehat{F}^{\top}\widehat{F}\right)^{-1}$. The rotational basis H_F then needs to be redefined as $\begin{bmatrix} I & 0 \\ 0 & H_F^u \end{bmatrix}$ where the identity matrix is the same dimension as the observed factors, and H_F^u is the normalisation basis for the unobserved factors defined in a similar manner to Equation* (3.1)*, i.e. H_F is defined in a suitable way that keeps the observed factors unchanged, but rotates the columns of the unobserved factors. Therefore, the theory developed in this paper can also be applied to FAVAR models with some minor adjustments.*

Remark. *In addition, the X_t series that are used for factor estimation need not be identical to the series whose impulse responses we are interested in. This can occur, for example, if a subset of X_t corresponding to non-aggregate series is used to estimate the factors as is commonly done (e.g. Stock and Watson, 2002, 2012, 2016). The corresponding loading matrix is still estimated by least squares as $\widehat{\Lambda}$, and the theory developed in this paper remains applicable.*

3.2 Asymptotic Distribution of Structural Parameters

We begin by deriving the asymptotic distribution of $\widehat{\delta}$, which is necessary to analyse $\widehat{a}$ and, therefore, the IRF. Let $\mathbb{S}_\delta$ be the infeasible counterpart of $\mathbb{S}_{\widehat{\delta}}$, i.e. a $(q-1) \times q$ matrix such that

$$\mathbb{S}_\delta \eta_t^* = \eta_{-1t}^* - \delta \eta_{1t}^*, \tag{3.7}$$

which by definition is equal to

$$\mathbb{S}_\delta = \left[\delta \ \vdots \ I_{q-1}(1:q-1)\right], \tag{3.8}$$

where $I_{q-1}(1:q-1)$ collects the last $q-1$ matrix of I_{q-1}.

Theorem 1. *Under Assumptions 1 to 8, and the conditions that $W_T \xrightarrow{p} W$, and $\sqrt{T}/N \to 0$ as $N, T \to \infty$,*

a) $\widehat{\delta}$ is a consistent estimator of δ, and

$$\sqrt{T}\left(\widehat{\delta} - \delta\right) \xrightarrow{d} \left(\mathcal{A} W \mathcal{A}^\top\right)^{-1} \mathcal{A} W N\left(0_{kq\times 1}, \left(\mathbb{S}_\delta \bar{H}_\eta^\top \otimes I_k\right) \Sigma_i^{(1)} \left(\mathbb{S}_\delta \bar{H}_\eta^\top \otimes I_k\right)^\top\right),$$

where $\mathcal{A} = I_{q-1} \otimes \mathbb{S}_1 \bar{H}_\eta^\top E(\eta_{1t} Z_t^\top)$, $\mathbb{S}_1 = [1, 0_{1\times(q-1)}]$, and $\Sigma_i^{(1)}$ is the upper left block of Σ_i.

b) The optimal choice of the weighting matrix is V_δ^{-1}, where $V_\delta = \mathcal{C}\Sigma_i^{(1)}\mathcal{C}^\top$ and $\mathcal{C} = \left[\mathbb{S}_\delta \bar{H}_\eta^\top \otimes I_k\right]$.

c) $\widehat{V}_\delta \xrightarrow{p} V_\delta$.

d) $\sqrt{T}\left(\widehat{\delta}^o - \delta\right) \xrightarrow{d} N\left(0, [\mathcal{A} V_\delta^{-1} \mathcal{A}^\top]^{-1}\right)$.

Theorem 1 shows that $\widehat{\delta}$ is consistent for δ and has a standard asymptotic normal distribution. Theorems 1 (b) and 1 (c) show the form of the infeasible weight matrix and the consistency of the feasible weight matrix, respectively. The use of the optimal weight matrix results in the optimal two-step GMM estimator $\widehat{\delta}^o$ in Theorem 1 (d), which follows from typical GMM arguments.

3.3 Asymptotic Distributions of Impulse Response Functions

In this subsection, we present the asymptotic distributions of the estimators of the IRFs. The IRFs are a function of $\widehat{a}_1, \widehat{\Lambda}, \widehat{\Psi}_s, \widehat{G}$ and $\widehat{\Theta}$. Because $\widehat{a}_1 = \left(1, \widehat{\delta}^\top\right)^\top$, the asymptotic properties of $\widehat{\delta}$ in Section 3.2 can be used by defining $\bar{\mathbb{S}}_1$ as the last $q-1$ columns of I_q, so that $\widehat{a} - a^* = \bar{\mathbb{S}}_1 \begin{bmatrix} 0 \\ \widehat{\delta} - \delta \end{bmatrix}$. Thus, we derive the asymptotic representations of the remaining terms and then combine these results to obtain the asymptotic distributions of the IRFs.

Proposition 1. *Under Assumptions 1 to 6,* $\widehat{G} - H_F^\top G \Sigma_\eta H_\eta = O_p\left(\frac{1}{\delta_{NT}^2}\right)$.

Proposition 2. *Under Assumptions 1 to 6, if* $\sqrt{T}/N \to 0$ *as* $N, T \to \infty$*,*

$$\sqrt{T}\left(\widehat{\theta}_i - H_\eta^{-1}\theta_i\right) = \left(H_\eta^{-1} G^\top H_F\right) \sqrt{T}\left(\widehat{\lambda}_i - H_F^{-1}\lambda_i\right) + o_p(1). \tag{3.9}$$

Proposition 3. *Under Assumptions 1 to 8, if* $\sqrt{T}/N \to 0$ *as* $N, T \to \infty$*,*

a)

$$\begin{aligned} \sqrt{T}\begin{bmatrix} \widehat{a}_1 - a_1^* \\ \widehat{\lambda}_i - H_F^{-1}\lambda_i \\ \operatorname{vec}(\widehat{\Psi}_s^\top - H_F^{-1}\Psi_s^\top H_F) \end{bmatrix} &= B_s \frac{1}{\sqrt{T}} \begin{bmatrix} \operatorname{vec}\left(Z^\top \eta - E(Z^\top \eta)\right) \\ F^\top e_i \\ \operatorname{vec}\left(\mathcal{F}^\top \eta\right) \end{bmatrix} + o_p(1) \\ &\xrightarrow{d} N(0_{(q+r+r^2)\times 1}, B_s \Sigma_i B_s^\top), \end{aligned} \tag{3.10}$$

where

$$B_s = \begin{bmatrix} \bar{\mathbb{S}}_1 \left(\mathcal{A} W \mathcal{A}^\top\right)^{-1} \mathcal{A} W (\mathbb{S}_\delta \bar{H}_\eta^\top \otimes I_k) & 0_{q\times r} & 0_{k\times rpq} \\ 0_{r\times qk} & \bar{H}_F^\top & 0_{r\times rpq} \\ 0_{r^2\times qk} & 0_{r^2\times r} & \bar{R}_s\left[\bar{H}_F G \otimes \left(\Sigma_{\mathcal{F}} \bar{H}_{\mathcal{F}}\right)^{-1}\right] \end{bmatrix},$$

$$\bar{R}_s = \sum_{j=1}^{s} \left(\bar{H}_F^\top \Psi_{j-1} \bar{H}_F^{-\top} \otimes \left[\bar{H}_F^{-\top} \Psi_{s-j}^\top \bar{H}_F, \bar{H}_F^{-\top} \Psi_{s-j-1}^\top \bar{H}_F, \ldots, \bar{H}_F^{-\top} \Psi_{s-j-p+1}^\top \bar{H}_F\right]\right),$$

$$\Psi_0 = I_r,$$

$$\Psi_s = 0_{r\times r} \quad \textit{for} \quad s < 0,$$

b)

$$\sqrt{T}\begin{bmatrix}\widehat{a}_1 - a_1^* \\ \widehat{\lambda}_i - H_F^{-1}\lambda_i\end{bmatrix} = B_0\frac{1}{\sqrt{T}}\begin{bmatrix}\mathrm{vec}\left(Z^\top\eta - E(Z^\top\eta)\right) \\ F^\top e_i\end{bmatrix} + o_p(1)$$
$$\xrightarrow{d} N\left(0_{(q+r)\times 1}, B_0\Sigma_i^{(1)}B_0^\top\right),$$

where $B_0 = \begin{bmatrix}\mathbb{S}_1\left(\mathcal{A}W\mathcal{A}^\top\right)^{-1}\mathcal{A}W\left(\mathbb{S}_\delta\bar{H}_\eta^\top\otimes I_k\right) & 0_{q\times r} \\ 0_{r\times qk} & \bar{H}_F^\top\end{bmatrix}.$

Proposition 1 shows that $\sqrt{T}\left(\widehat{G} - H_F^\top G\Sigma_\eta H_\eta\right)$ has a degenerate limiting distribution if $\sqrt{T}/N \to 0$, therefore $\widehat{G}$ can be directly replaced by $H_F^\top G\Sigma_\eta H_\eta$ as if $\widehat{G}$ is observed when N is large relative to T. Proposition 2 is used for obtaining the asymptotic representations of the contemporaneous impulse responses. Proposition 3 (a) implies the asymptotic distribution of the IRFs. Proposition 3 (b) is simply Proposition 3 (a) but without the effects of $\widehat{\Psi}_s$, and is used for simplifying the results for the contemporaneous IRFs. Theorem 1 and Propositions 1 to 3 together are applied to obtain the asymptotic distributions of the dynamic IRFs over time, as summarised in the following theorem.

Theorem 2. *Under Assumptions 1 to 8, and the conditions that* $W_T \xrightarrow{p} W$, *and* $\sqrt{T}/N \to 0$ *as* $N, T \to \infty$,

a) For the contemporaneous IRFs of X_{it} *to* ζ_{1t},

$$\sqrt{T}\left(\widehat{\theta}_i^\top\widehat{a}_1 - \theta_i^\top a_1\right) = \sqrt{T}\bar{Q}_{1,i}\begin{bmatrix}\widehat{a}_1 - a_1^* \\ \widehat{\lambda}_i - H_F^{-1}\lambda_i\end{bmatrix} + o_p(1)$$
$$\xrightarrow{d} N\left(0, \bar{Q}_{1,i}B_0\Sigma_i B_0^\top\bar{Q}_{1,i}^\top\right),$$

where $\bar{Q}_{1,i} = \left(\theta_i^\top\bar{H}_\eta^{-\top}C_1 + a_1 G^\top\bar{H}_F C_2\right)$, $C_1 = [I_q \vdots 0_{q\times r}]$, *and* $C_2 = [0_{r\times q} \vdots I_r]$.

b) For the IRFs of X_{it} to $\zeta_{1,t-s}, (s \geq 1)$,

$$\sqrt{T}\left(\widehat{\lambda}_i^\top \widehat{\Psi}_s \widehat{G} \widehat{a}_1 - \lambda_i^\top \Psi_s G a_1\right) = \sqrt{T} \bar{Q}_{2,i} \begin{bmatrix} \widehat{a}_1 - a_1^* \\ \widehat{\lambda}_i - H_F^{-1}\lambda_i \\ \mathrm{vec}\left(\widehat{\Psi}_s^\top - H_F^{-1}\Psi_s H_F\right) \end{bmatrix} + o_p(1)$$

$$\xrightarrow{d} N(0, \bar{Q}_{2,i} B_s \Sigma_i B_s^\top \bar{Q}_{2,i}^\top),$$

where

$$\bar{Q}_{2,i} = \lambda_i^\top \Psi_s G \Sigma_\eta \bar{H}_\eta C_3 + a_1^\top G^\top \Psi_s^\top \bar{H}_F C_4 + \left(\lambda_i^\top \bar{H}_F^{-\top} \otimes a_1^\top G^\top \bar{H}_F\right) C_5,$$

and $C_3 = [I_q \vdots 0_{q\times r} \vdots 0_{q\times r^2}]$, $C_4 = [0_{r\times q} \vdots I_r \vdots 0_{r\times r^2}]$ and $C_5 = [0_{r^2\times q} \vdots 0_{r^2\times r} \vdots I_{r^2}]$.

Theorem 2 establishes the consistency and asymptotic normality of the dynamic IRFs. Note that despite our estimator $\widehat{a}_1$ recovering $a_1^* = H_\eta^\top a_1$ and therefore being subject to the effect of the principal components rotation, our resulting estimators for the IRFs can consistently estimate the true impulse responses without the effect of any rotations. Additionally, we do not need H_η or H_F to estimate their asymptotic variances. Hence, Theorem 2 is the main result necessary for frequentist inference and construction of valid confidence intervals in empirical analysis. We discuss practical implementation of the covariance matrices in the next subsection.

Remark. *The structural factor model can also offer a way to test conventional identifying restrictions employed in SVAR models, such as short-run and long-run exclusion restrictions on the impulse responses. It is straightforward to implement any tests for zero impulse responses using Theorem 2. However, note that this approach does not establish a consistent estimator for A per se; and hence simply gives some theoretical justification to the often ad-hoc practice of checking for "reasonable" impulse responses often employed by the literature, (e.g. Bernanke et al., 2005).*

3.4 Covariance Matrix Estimation

We next detail feasible estimation of the covariance matrices for the IRFs in Theorem 2. First note that the idiosyncratic errors can be consistently estimated as

$$\widehat{e}_t = X_t - \widehat{\Lambda}\widehat{F}_t, \tag{3.11}$$

and let $\widehat{e}_{it}$ denote the ith element of $\widehat{e}_t$. We define an estimator for Σ_i as

$$\widehat{\Sigma}_i = \frac{1}{T-p}\sum_{t=p+1}^{T} \xi_{it}\xi_{it}^\top, \tag{3.12}$$

where

$$\xi_{it} = \begin{bmatrix} \mathrm{vec}\left(Z_t^\top \widehat{\eta}_t - \frac{1}{T-p}(Z^\top \widehat{\eta})\right) \\ \widehat{F}_t^\top \widehat{e}_{it} \\ \mathrm{vec}\left(\widehat{\mathcal{F}}_t^\top \widehat{\eta}_t\right) \end{bmatrix},$$

and

$$\widehat{R}_s = \sum_{j=1}^{s}\left(\widehat{\Psi}_{j-1} \otimes \left[\widehat{\Psi}_{s-j}^\top, \widehat{\Psi}_{s-j-1}^\top, \ldots, \widehat{\Psi}_{s-j-p+1}^\top\right]\right)$$

with $\widehat{\Phi}_0 = I_r$ and $\widehat{\Psi}_s = 0_{r\times r}$ for $s < 0$. Similarly, an estimator for $\Sigma_i^{(1)}$ can be defined as

$$\widehat{\Sigma}_i^{(1)} = \frac{1}{T-p}\sum_{t=p+1}^{T} C_6\xi_{it}\xi_{it}^\top C_6^\top, \tag{3.13}$$

where $C_6 = \begin{bmatrix} I_q & 0_{q\times r} & 0_{q\times rpq} \\ 0_{r\times qk} & I_r & 0_{r\times rpq} \end{bmatrix}$. When e_t is serially correlated, the HAC estimators for the asymptotic variances can be readily constructed following the arguments of Bai (2003) and Han and Inoue (2015); cross-sectional correlation in e_t can be additionally accommodated via a CS-HAC estimator following Bai and Ng (2006) and Gonçalves and Perron (2020). Estimators for B_0 and B_s for $s = 1, \ldots, h$ follow by appropriate replacement of their unknown quantities with their feasible counterparts

$$\widehat{B}_0 = \begin{bmatrix} \bar{\mathbb{S}}_1 \left(\mathcal{A}_T W_T \mathcal{A}_T\right)^{-1} \mathcal{A}_T W_T \left(\mathbb{S}_{\widehat{\delta}} \otimes I_k\right) & 0_{q\times r} \\ 0_{r\times qk} & I_r \end{bmatrix} \tag{3.14}$$

$$\widehat{B}_s = \begin{bmatrix} \bar{\mathbb{S}}_1 \left(\mathcal{A}_T W_T \mathcal{A}_T\right)^{-1} \mathcal{A}_T W_T \left(\mathbb{S}_{\widehat{\delta}} \otimes I_k\right) & 0_{q\times r} & 0_{q\times rpq} \\ 0_{q\times qk} & I_r & 0_{r\times rpq} \\ 0_{r^2\times qk} & 0_{r^2\times r} & \widehat{R}_s \left(\widehat{G} \otimes \left(\frac{\widehat{\mathcal{F}}^\top \widehat{\mathcal{F}}}{T-p}\right)^{-1}\right) \end{bmatrix}. \tag{3.15}$$

The constant matrices $\bar{Q}_{j,i}$ for $j = 1,2$ can be estimated by replacing the unknown parameters with their consistent counterparts. Based on Theorems 1 and 2 and Propositions 1 to 3, we know that $\widehat{\theta}_i, \widehat{\lambda}_i, \widehat{\Psi}_s, \widehat{G}$, and $\widehat{a}_1$ consistently estimate $\bar{H}_\eta^{-\top}\theta_i, \bar{H}_F^{-1}\lambda_i, \bar{H}_F^\top \Psi_s \bar{H}_F^{-\top}, \bar{H}_F^\top G \Sigma_\eta \bar{H}_\eta$, and $\bar{H}_\eta^\top a_1$, respectively. Hence, we propose the following estimators for the constant matrices

$$\widehat{Q}_{1,i} = \left[\widehat{\theta}_i^\top C_1 + \widehat{a}_1 \widehat{G}^\top C_2\right],$$
$$\widehat{Q}_{2,i} = \left[\widehat{\lambda}_i^\top \widehat{\Psi}_s \widehat{G} C_3 + \widehat{a}_1^\top \widehat{G}^\top \widehat{\Psi}_s^\top C_4 + \left(\widehat{\lambda}_i^\top \otimes \widehat{a}_1^\top \widehat{G}^\top\right) C_5\right].$$

3.5 Overidentification and Automatic Selection of External Instruments

A major advantage of the GMM-based framework that we adopt is the possibility of 1) testing the joint validity of external instruments and 2) automatic selection of valid external instruments.

We first present the J-test for the joint validity of external instruments as

$$J_T \equiv T\mathcal{Q}_T(\widehat{\delta}), \tag{3.16}$$

where $\mathcal{Q}_T(\widehat{\delta})$ is the GMM-criterion function $\mathcal{Q}_T$ evaluated at $\widehat{\delta}$ where the weight W_T is chosen optimally. We show that J_T has a standard asymptotic $\chi^2_{(k-1)(q-1)}$ distribution.

Following this definition of the J-test statistic, we then propose a series of instrument selection criteria for automatic selection of valid instruments. Let c denote the instrument selection vector, which takes values 0 or 1. The number of overidentifying restrictions is therefore $(|c|-1)(q-1)$.

Define the GMM-estimator using the instruments selected by c as

$$\widehat{\delta}(c) = \underset{\delta}{\operatorname{argmin}} \mathcal{Q}_T\left(\delta(c)\right) = \underset{\delta}{\operatorname{argmin}} \bar{g}_{T,c}(\delta)^\top W_T(c)\bar{g}_{T,c}(\delta), \tag{3.17}$$

where $\bar{g}_{T,c}(\delta)$ and $W_T(c)$ are the empirical moments and their weight matrix, defined using only the instruments selected by c. Thus, the corresponding $J_T(c)$ test can also be written as

$$J_T(c) = T\bar{g}_{T,c}\left(\widehat{\delta}(c)\right)^\top W_T(c)\bar{g}_{T,c}\left(\widehat{\delta}(c)\right). \tag{3.18}$$

We consider estimation of c^0 the "correct" selection vector, using an estimator $\widehat{c}$, which has parameter space $\mathscr{C} \in \mathcal{C}$. The space $\mathscr{C}$ contains $c = \mathbf{0}$, and is defined in terms of the selection of the *instruments* in order to exploit the block structure implied by the moment conditions; that is, if the first instrument is invalid, then this implies that all of the first $q-1$ moment conditions are also invalid.

Assumption 9. *Define $\mathscr{Z} = \{c \in \mathcal{C} : c = c^0(\delta) \text{ for some } \delta\}$, the set of selection vectors in $\mathcal{C}$ which select only moment conditions that are zero asymptotically, and $\mathscr{M}\mathscr{Z} = \{c \in \mathscr{Z} : |c| \geq |c^*| \forall c^* \in \mathscr{Z}\}$, the set of selection vectors in $\mathscr{Z}$ that maximise the selected moments out of selection vectors in $\mathscr{Z}$. We require the following conditions:*

a) $\mathscr{M}\mathscr{Z}$ contains a single element c^0.

b) $\bar{g}_{T,c}(\delta)$ has a unique solution δ.

Assumptions 9 (a) and 9 (b) correspond to Assumptions ICc^0 and ICθ^0 of Andrews (1999). Assumption 9 (a) requires that the correct selection vector uniquely selects the maximal number of moment conditions that equal zero asymptotically. Assumption 9 (b) specifies that δ is the "true" value of δ. As we are in a standard GMM context, $\mathscr{M}\mathscr{Z} = \{1_{q-1}\}$ and thus both assumptions hold.

Similarly, we follow Andrews (1999) and propose a set of information criteria which can be used to select for the correct set of moments (instruments). The GMM moment selection criterion chooses the vector in $\widehat{c}_{GMM_{BIC}}, \widehat{c}_{GMM_{AIC}}, \widehat{c}_{GMM_{HQIC}}$ in $\mathcal{C}$ which, respectively, minimise:

$$\begin{aligned} GMM_{BIC} &= J(c) - (|c|-1)(q-1)logT; \\ GMM_{AIC} &= J(c) - 2(|c|-1)(q-1); \\ GMM_{HQIC} &= J(c) - Q(|c|-1)(q-1)loglogT, \end{aligned}$$

for some $Q > 2$ (which we set to 2.01), and where $(|c|-1)(q-1)$ is the number of identifying restrictions.[2] These criteria are the counterparts of the Bayesian Information Criteria, Akaike Information Criteria, and Hannan-Quinn Information Criteria.

Downwards Testing

As an alternative, we next propose a downwards testing procedure that sequentially tests all combinations of the external instruments and asymptotically yields the correct selection of instruments.[3] As stated by Andrews (1999), this downwards testing procedure formalises the ad-hoc approaches used by empirical researchers.

We describe the downwards testing procedure, which is based on the test statistic $J_T(c)$. Starting with vectors $c \in \mathscr{C}$ for which $|c|$ is the largest, we carry out tests with progressively smaller $|c|$ until we find a test that does not reject the null hypothesis; let $\widehat{k}_{DT}$ denote the value of $|c|$ for the first such test that does not reject. Given $\widehat{k}_{DT}$, the downwards testing estimator of $\widehat{c}_{DT}$ is defined to be the vector that minimises $J_T(c)$ over $c \in \mathscr{C}$ with $|c| = \widehat{k}_{DT}$. The downwards testing moment selection procedure thus progresses from the most to least restrictive model.

The consistency of the J-test, moment selection criteria, and downwards testing procedure are summarised in the following theorem.

Theorem 3. *Under Assumptions 1 to 8 and the condition that* $\frac{\sqrt{T}}{N} \to 0$ *as* $N, T \to \infty$*,*

a) $J_T \equiv T\mathcal{Q}_T(\widehat{\delta}) \xrightarrow{d} \chi^2_{(k-1)(q-1)}$*,*

b) Additionally under Assumption 9, for $MSC \in \{GMM_{BIC}, GMM_{AIC}, GMM_{HQIC}\}$*,* $\widehat{c}_{MSC} = c^0$ *w.p.a.* 1*,*

[2] Note that $GMM - AIC$ is inconsistent.

[3] Note that the upwards testing procedure requires some extra assumptions, so we omit it for brevity.

c) Additionally under Assumption 9, $\widehat{c}_{DT} = c^0$ *w.p.a.* 1.

Theorem 3 (a) follows from a standard application of Hansen's J-test. The degrees of freedom corresponds to the fact that each of the k instruments corresponds to $(q-1)$ moments, and only $k = 1$ instrument is required to just identify a_1. Theorem 3 (b) establishes that the model selection criteria $\widehat{c}_{GMM_{BIC}}, \widehat{c}_{GMM_{AIC}}$, and $\widehat{c}_{GMM_{HQIC}}$ are consistent at determining when there are no over-identifying restrictions. Theorem 3 (c) corresponds to Theorem 2 of Andrews (1999), and establishes that the downwards testing estimator $\widehat{c}_{DT}$ is able to determine when there are no over-identifying restrictions, similar to $\widehat{c}_{MSC}$. In practice, over-rejection of the J test in finite samples tends to lead to a higher probability of using only correct moments, but not necessarily all valid moments.

Remark. *An upwards testing procedure can also be considered following Andrews (1999). However, this requires an additional assumption on the parameter space* $\mathscr{C}$ *to ensure that it does not stop at too small a value of* $|c|$. *In addition, although both the upwards and downwards testing procedure are consistent, in finite sample the upwards testing procedure will always select fewer moments than the downwards testing procedure if they do not agree. Thus, we focus on the model selection criteria and downwards testing approaches.*

4 Monte Carlo

4.1 Data Specification

The factor loadings λ_i are drawn from a multivariate normal distribution with mean $\mathbf{0}_r$ and covariance matrix $\Sigma_\Lambda = I_r$.[4] The structural shocks ζ_t are drawn from $N(\mathbf{0}_q, I_q)$.

Similar to Bai and Wang (2015), we specify a VAR process for the dynamic factors as

$$f_t = \phi f_{t-1} + A\zeta_t, \tag{4.1}$$

[4] The set of loadings λ_i is set to be a vector of ones, in order to ensure that the first impulse response function is not too small, which can cause some numerical issues when implementing the unit effect normalisation.

where $\phi = 0.7$ and $A = \begin{bmatrix} 1 & 0 & 0 \\ 1 & 1 & 0 \\ 1 & 0 & 1 \end{bmatrix}$. The static factors are stacked as $F_t = \left[f_t^\top, f_{1,t-1}, \dots, f_{r-q,t-1}^\top\right]^\top$ where $r \leq 2q$ in order to include some lags of f_t as static factors. Equation (4.1) implies

$$F_t = \Phi_1 F_{t-1} + GA\zeta_t, \tag{4.2}$$

where $\Phi = \begin{bmatrix} \phi I_r & 0_{q\times(r-q)} \\ I_{(r-q)\times q} & 0_{(r-q)\times(r-q)} \end{bmatrix}$ where $I_{(r-q)\times q}$ denotes the first $(r-q)$ rows of I_q, and $GA = \begin{bmatrix} A \\ 0_{(r-q)\times q} \end{bmatrix}$. We set $r = 5$ and $q = 3$. The observable series are then generated as

$$X_t = \Lambda F_t + e_t. \tag{4.3}$$

To investigate the efficiency gains from overidentification and the size of the proposed J-test, we generate the instrument Z_{jt} that is correlated with the first structural shock at time t by

$$\textbf{DGP 1}: Z_{jt} = \sqrt{1-a^2}w_{jt} + a\zeta_{1t} + \zeta_{q,t-1}, \quad \text{for} \quad j = 1, \dots, k = 4, \tag{4.4}$$

where w_{jt} are i.i.d. standard normal random variables, and a is set equal to $\sqrt{1/2}/2$ so that the correlation between Z_{jt} and ζ_{1t} is equal to 0.25. We set the number of instruments as $k = 1, \dots, 4$ to investigate the benefits of overidentification.

To investigate the power of the overidentification test and consistency of the moment selection procedures, we generate instruments as

$$\begin{aligned} \textbf{DGP 2}: \quad & Z_{jt} = \sqrt{1-a^2}w_{jt} + a\zeta_{1t} + \zeta_{q,t-1}, \quad \text{for} \quad j = 1, 2, \\ & Z_{3t} = \sqrt{1-a^2}w_{3t} + a\zeta_{2t} + \zeta_{q,t-1}, \\ & Z_{4t} = \sqrt{1-a^2}w_{4t} + a\zeta_{3t} + \zeta_{q,t-1}, \end{aligned} \tag{4.5}$$

such that the first two instruments are only correlated with the structural shock of interest and hence valid, but the last two instruments are contaminated with the effects of other structural shocks, and hence invalid. In either specification, the instruments are also correlated with the lags of the qth structural shock. The number of replications is $1,000$.

With the observed data X_t, we estimate the static factors $\widehat{F}$ and loadings $\widehat{\Lambda}$. The MA coefficients $\widehat{\Psi}_s$ are computed using the OLS estimate of $\widehat{\Phi}$. The principal components-based estimates of the reduced form shocks $\widehat{\eta}_t$ and the instruments are then used for the estimation of δ. We implement two-step GMM estimation using the weighting matrix $I_q \otimes \left(\frac{1}{T}\sum_{t=1}^{T} Z_t Z_t^\top\right)^{-1}$ in the first step, then re-estimate δ using the optimal weighting matrix $\widehat{V}^{-1}$ in the second step. The confidence intervals for the structural IRFs are computed based on the asymptotic normal distribution in Theorem 2 and the proposed consistent estimators of the covariance matrices.

Additionally, for a point of comparison, we also identify and estimate the impulse response using an SVAR-IV estimator using only the first four variables in X_t. The implementation of this follows Cheng et al. (2021), the difference from the SFM approach being that the SVAR is estimated in X_{jt} for $j = 1, \ldots, 4$ directly, and the reduced form shocks estimated as the residuals of this system. Note that because $q = 3$, this results in a singular VAR system.

4.2 Results

4.2.1 Efficiency Gains

Table 1 reports the finite-sample RMSEs of the estimated IRFs as measured by a ratio of each overidentified scheme compared to the just-identified scheme using only one instrument. The RMSE ratios are typically decreasing in k for the IRFs, with the effect being more pronounced at the contemporaneous horizon and when $N = T$. This confirms that a more efficient estimate of a_1 via the use of more than one instrument can lead to efficiency gains for the IRFs at both zero and non-zero horizons, and that our asymptotic theory is correct.

Table 2 reports the finite-sample coverage rates of the confidence intervals. In general, these coverage rates are acceptable and close to the nominal level of 95%, particularly at the $h = 3$ horizon. There is some evidence that the coverage ratios are slightly under-estimated for contemporaneous horizons. This is a commonly encountered problem in the factor modelling literature (e.g. Yamamoto

		$h=0$				$h=3$			
T	N	$k=2$	$k=3$	$k=4$	SVAR-IV ($k=4$)	$k=2$	$k=3$	$k=4$	SVAR-IV ($k=4$)
250	125	0.945	0.926	0.919	1.962	1.001	1.007	1.008	1.625
	250	0.949	0.928	0.911	2.006	0.998	0.986	0.978	1.520
500	125	0.961	0.938	0.925	2.020	1.000	1.000	0.996	1.927
	250	0.961	0.948	0.936	2.063	1.001	1.004	1.003	1.823

Note:
Entries report the RMSE of the estimated IRFs of the overidentified system, compared to the RMSE of the IRFs of the just-identified system.

Table 1: RMSE ratios

		$h=0$				$h=3$			
T	N	$k=1$	$k=2$	$k=3$	$k=4$	$k=1$	$k=2$	$k=3$	$k=4$
250	125	0.919	0.897	0.889	0.881	0.954	0.949	0.947	0.944
	250	0.920	0.904	0.895	0.888	0.954	0.949	0.945	0.944
500	125	0.903	0.890	0.883	0.880	0.946	0.943	0.941	0.940
	250	0.909	0.898	0.892	0.889	0.948	0.945	0.943	0.942

Note:
Entries report the coverage probabilities of the IRFs using the proposed asymptotic distributions (nominal 95%).

Table 2: Coverage Probabilities

and Hara, 2022), and is a finite-sample aberration that can be readily addressed by employing a bootstrap procedure similar to that of Yamamoto (2019). We leave this issue for future research. The effective coverage rates all improve at the sample size increases and particularly as N increases, confirming our asymptotic theory.

Table 3 reports the finite-sample performance of the proposed J-test to test the null hypothesis of joint instrument exogeneity. In general, the effect size of the proposed tests is acceptable for the sample sizes considered in simulations.

4.2.2 Overidentification Test and Moment Selection

We investigate the results of the proposed overidentification and moment selection procedures, which correspond to DGP 2. Tables 4 and 5 present the finite sample performance of the J-test for joint exogeneity of the instruments and accuracy of the moment selection procedures, respectively, under DGP 2. It can be seen that, across all specifications, the J-test has high power, and the

T	N	$k=2$	$k=3$	$k=4$
250	125	0.046	0.034	0.035
	250	0.058	0.040	0.021
500	125	0.049	0.047	0.043
	250	0.039	0.035	0.039

Note:
Entries report the rejection frequencies of the J-test (nominal size of 5%).

Table 3: Size of J-test

T	N	Rejection Frequency
250	125	1.000
	250	1.000
500	125	1.000
	250	1.000

Note:
Entries report the rejection frequency of J-test for overidentification, with $k = 4$ instruments.

Table 4: Power of J-test

		Information Criteria			Testing	
T	N	GMM_{BIC}	GMM_{AIC}	GMM_{HQIC}	DT	UT
250	125	1.000	1.000	1.000	1.000	1.000
	250	1.000	1.000	1.000	1.000	1.000
500	125	1.000	1.000	1.000	1.000	1.000
	250	1.000	1.000	1.000	1.000	1.000

Note:
Entries report the frequencies of correct instrument selection. DT and UT denote Downwards and Upwards Testing respectively. Correct instruments are Z_1 and Z_2.

Table 5: Accuracy of Moment Selection Procedures

moment selection procedures are able to accurately select the correct instruments.

5 Empirical Application

5.1 Data and Instruments

We consider the dataset used by Stock and Watson (2012). This dataset consists of quarterly observations from 1959Q1 - 2011Q2 on 200 U.S. macroeconomic time series, grouped into 13 categories and suitably transformed to induce stationarity. Of the 200 series available, we only use the 132 disaggregated series to estimate the factors in order to avoid double counting high level aggregates.[5] We restrict our sample to 1980Q1 - 2007Q2; the start is dictated by the data availability of the instruments, while the end is chosen to avoid the onset of the Global Financial Crisis. We focus on identifying and estimating the dynamic causal effect of a monetary policy shock with the use of various monetary policy instruments. These include the narrative-based instrument of Romer and Romer (2004) computed as the residual of a Fed monetary intentions measure on internal Fed forecasts,[6] a model based instrument in the form of the monetary shocks identified from the SVAR of Bernanke and Mihov (1998), and a collection of monetary surprises identified using

[5] See Stock and Watson (2012) for more details on data cleaning.

[6] We use an updated and extended version of the Romer and Romer (2004) shocks, as constructed by Wieland and Yang (2020).

high(er) frequency data: the changes in federal funds futures around policy announcements using a daily window (Barakchian and Crowe, 2013), a 30-minute window (Gertler and Karadi, 2015), and a 30-minute window with further cleaning of the surprises via a regression on more control variables, (Miranda-Agrippino and Ricco, 2021). This selection of five instruments corresponds to the instrument set used by Schlaak et al. (2023).[7]

5.2 Model Specification

We first estimate the number of factors in the dataset. The $IC_p(2)$ criterion of Bai and Ng (2002) suggests $r = 5$ static factors, though the criteria are quite flat for $4 - 12$ factors. As suggested by Stock and Watson (2016), the sixth to twelfth factors can often help in explaining the majority of variation in many important variables such as labour productivity, hourly compensation, the term spread, and exchange rates. The first nine factors explain about 52.1% of the total variance in the dataset; whereas the contributions of the 10th-12th factors only provide marginal gains totalling 6.79% additional explanatory power. As stated by Han (2015, 2018), it is important to set r high enough such that the space spanned by the static factors can be fully recovered; at the same time, setting r too high (usually more than nine factors) tends to introduce too much extra noise.[8] We therefore proceed with the choice of $r = 9$ static factors for our benchmark model. We fit a VAR(2) process to model the dynamics of F_t, corresponding to $p = 2$ lags in our benchmark analysis, as supported by the BIC. The criterion of Bai and Ng (2007) tends to detect three dynamic factors; we thus set $q = 3$ factors in our benchmark model.

We proceed with the identification of the monetary policy shock by implementing the proposed estimators with all five available instruments.

[7] We do not consider combining the instruments by taking their first eigenvector, as our generalised method-of-moments framework already achieves this in a data-driven manner.

[8] The procedures of Onatski (2010) and Ahn and Horenstein (2013) tend to estimate $r = 1$ factors, which as noted Forni and Gambetti (2014) is at odds with most macroeconomic theory and the theoretical premise of the SVAR literature.

5.3 Results

Dynamic Causal Effects of Monetary Policy Shocks

We present the results of our benchmark model. Figure 1 shows the cumulative impulse responses of various macroeconomic variables to a standard deviation monetary contraction in the Federal Funds rate as identified, using the benchmark model with all instruments. Although most impulse responses are not statistically significant from zero, it is remarkable that most of the point estimates are consistent with economic theory and the consensus as documented by Christiano et al. (1998). Economic activity as measured by industrial production declines immediately, with the response bottoming after one year. Falls are also evident in earnings, employment, and money variables. The unemployment rate is estimated to increase after a contractionary monetary policy shock, with a persistent effect.

Note that there is a persistent, though generally statistically insignificant puzzle evident. In contrast, Figure 2 presents the cumulative IRFs after a contractionary monetary policy as (just) identified by using each instrument one at a time, in comparison with the benchmark overidentified model. Note that the impulse responses as identified by the Bernanke and Mihov (1998) model-based instrument are omitted due to scaling issues. It is remarkable that although almost all impulse responses from these just identified schemes lie within the 95% confidence interval of the overidentified case, each of them produces responses that are less efficient and/or more puzzling.

For example, the narrative-based instrument of Romer and Romer (2004) produces responses with the correct signs for series such as earnings, housing starts, industrial production, and employment. However, it is also prone to producing greatly puzzling behaviour in prices, exchange rates, and the S&P500. Next, the model-based instrument of Bernanke and Mihov (1998) identifies a contemporaneous response of the Federal Funds Rate to a monetary policy shock to be near zero. Such a result poses significant numerical problems when imposing the unit-effect normalisation and causes significant scaling issues; this result is additionally highly incompatible with macroeconomic theory. We therefore believe that the impulse responses as identified by this model-based measure to be untenable in any practical sense. Finally, the high frequency-based instruments of Barakchian and Crowe (2013); Gertler and Karadi (2015); Miranda-Agrippino and Ricco (2021) are much more

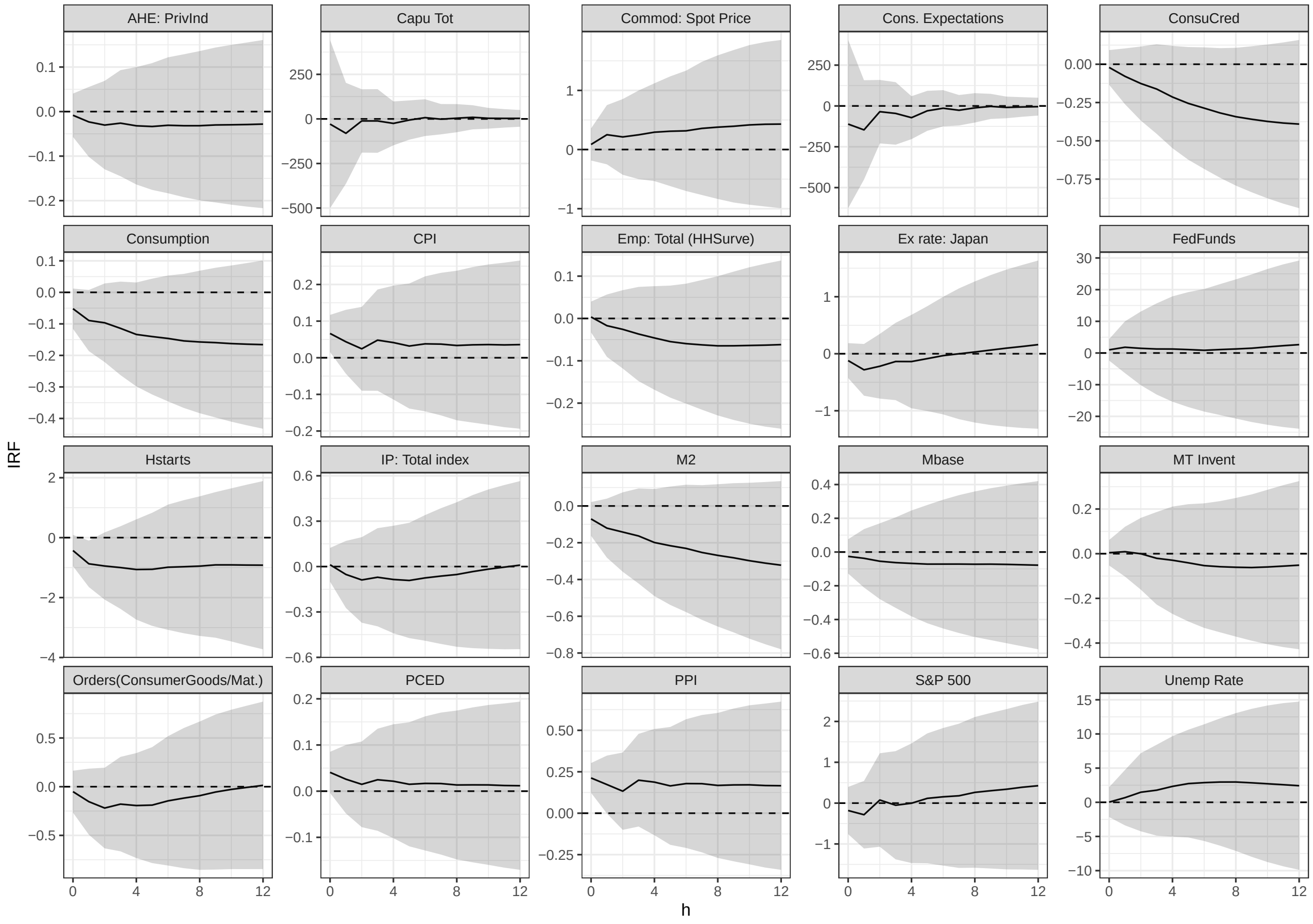

Figure 1: Cumulative IRFs after a contractionary monetary policy shock for the over-identified benchmark, normalised to a 100 basis point movement in the Federal Funds rate.

prone to producing puzzling responses. In particular, all high frequency instruments produce puzzling responses in the key variables of earnings, orders, employment, inventories, consumer credit, and crucially industrial production. Although these responses are not statistically significantly different from zero, the point estimates are still broadly incompatible with the macroeconomic consensus as summarised by Christiano et al. (1998). On the other hand, we find that high frequency instruments tend to alleviate, and, in some cases, eliminate price puzzles altogether. Of these, it is noteworthy that the Gertler and Karadi (2015) instrument tends to produce the largest puzzles in real activity and prices, a phenomenon that Miranda-Agrippino and Ricco (2021) attribute to omitted information effects and alleviate via their instrument. These results are not entirely surprising; as noted by Ramey (2016), these popular instrument variables are unstable and can produce puzzling responses to prices and real variables.

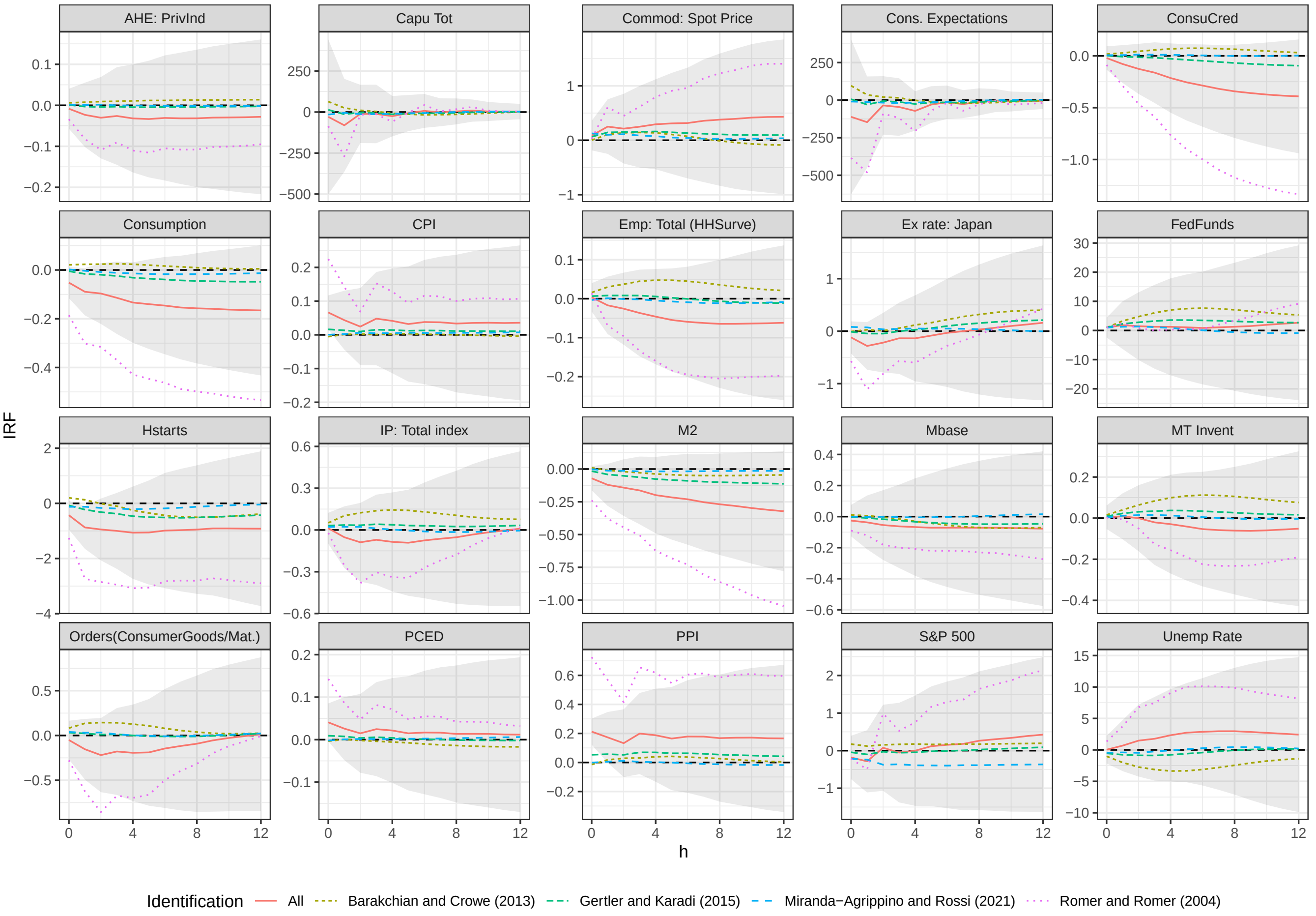

Figure 2: Cumulative IRFs after a contractionary monetary policy shock for the over-identified benchmark (solid curve), and just identified setups each using one instrument at a time, normalised to a 100 basis point movement in the Federal Funds rate.

Evidently, the responses identified by each instrument have their distinct advantages and disadvantages, which, in practice, make economic reconciliation difficult. Thus, the comparison in Figure 2 shows that the proposed overidentification scheme is able to automatically leverage the respective advantages of each instrument, and produce, overall, more reasonable responses.

Which Monetary Policy Instruments are Valid?

Testing the exogeneity and therefore validity of the instruments has so far been largely unresolved in the literature, but can be addressed with the J-test and automatic moment selection criteria we propose. Table 6 presents the results of J-test for joint exogeneity of instruments and automatic moment selection criteria. Across all overidentified specifications, we fail to reject the J-test. Correspondingly, all three model selection criteria are minimised when all five instruments are used;

the downwards testing estimator consequently selects all five instruments.

GMM_{BIC}	GMM_{HQIC}	J_T	J_{crit}	Romer and Romer (2004)	Gertler and Karadi (2015)	Miranda-Agrippino and Rossi (2021)	Bernanke and Mihov (1998)	Barakchian and Crowe (2013)
-8.573	-5.502	0.638	5.991	TRUE	TRUE	FALSE	FALSE	FALSE
-8.824	-5.753	0.386	5.991	TRUE	FALSE	TRUE	FALSE	FALSE
-8.836	-5.765	0.375	5.991	FALSE	TRUE	TRUE	FALSE	FALSE
-15.285	-9.143	3.135	9.488	TRUE	TRUE	TRUE	FALSE	FALSE
-9.200	-6.129	0.011	5.991	TRUE	FALSE	FALSE	TRUE	FALSE
-8.760	-5.689	0.451	5.991	FALSE	TRUE	FALSE	TRUE	FALSE
-17.766	-11.623	0.655	9.488	TRUE	TRUE	FALSE	TRUE	FALSE
-8.878	-5.807	0.333	5.991	FALSE	FALSE	TRUE	TRUE	FALSE
-18.035	-11.893	0.386	9.488	TRUE	FALSE	TRUE	TRUE	FALSE
-15.953	-9.810	2.468	9.488	FALSE	TRUE	TRUE	TRUE	FALSE
-23.846	-14.633	3.785	12.592	TRUE	TRUE	TRUE	TRUE	FALSE
-4.998	-1.927	4.212	5.991	TRUE	FALSE	FALSE	FALSE	TRUE
-5.783	-2.712	3.427	5.991	FALSE	TRUE	FALSE	FALSE	TRUE
-14.073	-7.930	4.348	9.488	TRUE	TRUE	FALSE	FALSE	TRUE
-8.450	-5.379	0.761	5.991	FALSE	FALSE	TRUE	FALSE	TRUE
-13.764	-7.622	4.656	9.488	TRUE	FALSE	TRUE	FALSE	TRUE
-14.377	-8.235	4.043	9.488	FALSE	TRUE	TRUE	FALSE	TRUE
-20.511	-11.298	7.120	12.592	TRUE	TRUE	TRUE	FALSE	TRUE
-5.826	-2.754	3.385	5.991	FALSE	FALSE	FALSE	TRUE	TRUE
-14.060	-7.918	4.360	9.488	TRUE	FALSE	FALSE	TRUE	TRUE
-14.155	-8.012	4.266	9.488	FALSE	TRUE	FALSE	TRUE	TRUE
-23.035	-13.822	4.596	12.592	TRUE	TRUE	FALSE	TRUE	TRUE
-14.780	-8.637	3.641	9.488	FALSE	FALSE	TRUE	TRUE	TRUE
-23.044	-13.830	4.587	12.592	TRUE	FALSE	TRUE	TRUE	TRUE
-21.320	-12.107	6.311	12.592	FALSE	TRUE	TRUE	TRUE	TRUE
-28.955	-16.670	7.887	15.507	TRUE	TRUE	TRUE	TRUE	TRUE

Table 6: Results of J-test for Overidentification and GMM_{MSC} Criteria.

Is Using More Instruments Better?

Given the evidence that all monetary policy instruments are jointly valid, we next investigate the efficiency gains from using more than one instrument. We do this by estimating and comparing the asymptotic variances of the IRFs. Table 7 reports the ratios of the asymptotic standard deviations of the estimated IRFs under the just-identified IRFs obtained by using one instrument compared to the benchmark model overidentified using all instruments; a ratio greater than one means that the overidentified model provides a more efficient estimate. Not all ratios are greater than one, so we additionally compute the means of the ratios to see if overidentification can lead to efficiency gains on average. On average, the benchmark model that uses all instruments tends to produce more efficient estimated for both zero and nonzero horizons compared to all instruments, with the notable exception of Gertler and Karadi (2015). However, although the relative efficiency of the responses of this instrument are on average lower, its behaviour can be quite erratic across different horizons even for the same variable. Therefore, we conclude that the overidentified scheme provides an ideal trade-off between producing the most reasonable impulse responses, and efficiency.

		h				
Instrument	Series	0	1	2	3	4
Barakchian and Crowe (2013)	IP: Total index	13.469	15.811	12.019	10.952	8.082
	Commod: Spot Price	2.192	12.585	7.305	2.023	3.937
	CPI	6.216	6.830	2.954	3.994	3.918
	FedFunds	7.750	14.087	13.601	14.745	14.167
	S&P 500	11.052	5.487	2.460	3.784	2.429
	Ex rate: Japan	1.301	9.709	5.268	11.549	13.178
	Consumption	11.975	5.316	2.344	4.717	1.762
	Mean	7.553	8.720	7.585	7.201	8.007
Bernanke and Mihov (1998)	IP: Total index	27.201	26.339	21.741	25.927	21.605
	Commod: Spot Price	32.133	22.162	18.692	21.629	19.736
	CPI	20.598	18.303	21.087	17.794	14.360
	FedFunds	28.214	23.022	22.871	22.970	22.277
	S&P 500	26.318	14.397	38.714	15.720	22.038
	Ex rate: Japan	22.969	18.982	31.212	21.465	22.962
	Consumption	21.169	15.815	15.703	14.187	12.970
	Mean	22.756	20.545	21.151	19.861	19.043
Gertler and Karadi (2015)	IP: Total index	1.048	1.012	0.815	1.355	0.990
	Commod: Spot Price	1.495	1.112	0.708	0.940	0.895
	CPI	0.529	0.653	0.826	0.620	0.530
	FedFunds	0.937	0.982	0.952	1.011	1.141
	S&P 500	1.473	0.606	1.780	0.645	0.844
	Ex rate: Japan	1.027	0.693	1.312	0.759	0.840
	Consumption	0.901	0.578	0.703	0.574	0.555
	Mean	0.925	0.835	0.837	0.827	0.811
Miranda-Agrippino and Rossi (2021)	IP: Total index	5.555	2.558	2.057	12.454	8.651
	Commod: Spot Price	15.735	8.226	3.382	9.605	6.929
	CPI	3.589	3.379	7.183	5.169	2.358
	FedFunds	8.710	2.295	2.349	4.850	6.428
	S&P 500	13.762	1.413	19.213	4.447	8.660
	Ex rate: Japan	10.731	1.790	12.743	2.236	2.181
	Consumption	3.248	2.834	5.699	2.977	1.641
	Mean	6.720	4.188	5.039	5.385	4.288
Romer and Romer (2004)	IP: Total index	3.501	3.458	2.974	3.350	2.889
	Commod: Spot Price	3.748	2.708	2.813	2.733	2.617
	CPI	2.374	2.507	2.605	2.542	2.227
	FedFunds	3.143	3.111	3.161	3.080	3.143
	S&P 500	3.325	2.234	4.387	2.499	2.535
	Ex rate: Japan	3.117	2.443	3.512	2.621	3.075
	Consumption	3.105	2.249	2.553	2.277	2.063
	Mean	2.956	2.765	2.821	2.733	2.638

Table 7: Ratios of Asymptotic Standard Deviations of Estimated Impulse Response Functions: Just Identified / Over-identified.

6 Conclusion

This paper develops new estimators for the impulse response functions in structural factor models under overidentifying restrictions by using multiple external instruments. Compared with a typical SVAR-IV approach, our framework is able to simultaneously address the challenging issues of nonfundamentalness, covariance singularity, and validity testing of identification restrictions, any of which can prevent the practitioner from recovering the true impulse responses. We establish the asymptotic distributions of the new estimators, and develop test statistics for the joint validity of instruments, and a downwards testing procedure which automatically selects the correct instruments. Our simulation study confirms that the estimated impulse response functions are more accurate than structural factor models that only use one instrument at a time, and the pre-existing SVAR-IV approach, as well as the finite sample properties of the proposed validity tests and moment selection criteria. We apply the framework to identify and estimate the impacts of a contractionary monetary policy shock on a large quarterly U.S. macroeconomic dataset using five commonly used instruments, including narrative-based measures, model-based measures, and monetary surprises identified with high(er) frequency data. We find that, although all of these instruments are jointly valid, using these instruments one by one as is commonly done in the literature can nevertheless produce puzzling and highly inaccurate responses. Instead, our proposed framework that jointly uses all instruments is able to produce responses which are overall more reasonable, and more accurate.

References

Ahn, S. C. and Horenstein, A. R. (2013). Eigenvalue Ratio Test for the Number of Factors. *Econometrica*, 81(3):1203–1227.

Amengual, D. and Watson, M. W. (2007). Consistent Estimation of the Number of Dynamic Factors in a Large N and T Panel. *Journal of Business & Economic Statistics*, 25(1):91–96.

Andrews, D. W. K. (1999). Consistent Moment Selection Procedures for Generalized Method of Moments Estimation. *Econometrica*, 67(3):543–563.

Bai, J. (2003). Inferential Theory for Factor Models of Large Dimensions. *Econometrica*, 71(1):135–171.

Bai, J., Li, K., and Lu, L. (2016). Estimation and Inference of FAVAR Models. *Journal of Business & Economic Statistics*, 34(4):620–641.

Bai, J. and Ng, S. (2002). Determining the Number of Factors in Approximate Factor Models. *Econometrica*, 70(1):191–221.

Bai, J. and Ng, S. (2004). A PANIC Attack on Unit Roots and Cointegration. *Econometrica*, 72(4):1127–1177.

Bai, J. and Ng, S. (2006). Confidence Intervals for Diffusion Index Forecasts and Inference for Factor-Augmented Regressions. *Econometrica*, 74(4):1133–1150.

Bai, J. and Ng, S. (2007). Determining the Number of Primitive Shocks in Factor Models. *Journal of Business & Economic Statistics*, 25(1):52–60.

Bai, J. and Wang, P. (2015). Identification and Bayesian Estimation of Dynamic Factor Models. *Journal of Business & Economic Statistics*, 33(2):221–240.

Barakchian, S. M. and Crowe, C. (2013). Monetary policy matters: Evidence from new shocks data. *Journal of Monetary Economics*, 60(8):950–966.

Bernanke, B. S., Boivin, J., and Eliasz, P. (2005). Measuring the Effects of Monetary Policy: A Factor-Augmented Vector Autoregressive (FAVAR) Approach*. *The Quarterly Journal of Economics*, 120(1):387–422.

Bernanke, B. S. and Mihov, I. (1998). Measuring Monetary Policy*. *The Quarterly Journal of Economics*, 113(3):869–902.

Brockwell, P. J. and Davis, R. A. (1991). *Time Series: Theory and Methods*. Springer Science & Business Media.

Chan, J. C. C., Eisenstat, E., and Strachan, R. W. (2020). Reducing the state space dimension in a large TVP-VAR. *Journal of Econometrics*, 218(1):105–118.

Cheng, X., Han, X., and Inoue, A. (2021). Instrumental Variable Estimation of Structural Var Models Robust to Possible Nonstationarity. *Econometric Theory*, pages 1–30.

Christiano, L., Eichenbaum, M., and Evans, C. (1998). Monetary Policy Shocks: What Have We Learned and to What End? Technical Report w6400, National Bureau of Economic Research, Cambridge, MA.

Forni, M. and Gambetti, L. (2010a). The dynamic effects of monetary policy: A structural factor model approach. *Journal of Monetary Economics*, 57(2):203–216.

Forni, M. and Gambetti, L. (2010b). Macroeconomic Shocks and the Business Cycle: Evidence from a Structural Factor Model.

Forni, M. and Gambetti, L. (2014). Sufficient information in structural VARs. *Journal of Monetary Economics*, 66:124–136.

Forni, M., Giannone, D., Lippi, M., and Reichlin, L. (2009). Opening the Black Box: Structural Factor Models with Large Cross Sections. *Econometric Theory*, 25(5):1319–1347.

Gafarov, B., Meier, M., and Montiel Olea, J. L. (2018). Delta-method inference for a class of set-identified SVARs. *Journal of Econometrics*, 203(2):316–327.

Gertler, M. and Karadi, P. (2015). Monetary Policy Surprises, Credit Costs, and Economic Activity. *American Economic Journal: Macroeconomics*, 7(1):44–76.

Geweke, J. (1977). The dynamic factor analysis of economic time series. *Latent variables in socio-economic models.*

Gonçalves, S. and Perron, B. (2020). Bootstrapping factor models with cross sectional dependence. *Journal of Econometrics*, 218(2):476–495.

Hamilton, J. D. (2020). *Time Series Analysis.* Princeton university press.

Han, X. (2015). Tests for overidentifying restrictions in Factor-Augmented VAR models. *Journal of Econometrics*, 184(2):394–419.

Han, X. (2018). Estimation and inference of dynamic structural factor models with over-identifying restrictions. *Journal of Econometrics*, 202(2):125–147.

Han, X. (2024). Global Identification, Estimation and Inference of Structural Impulse Response Functions in Factor Models: A Unified Framework.

Han, X. and Inoue, A. (2015). Tests for Parameter Instability in Dynamic Factor Models. *Econometric Theory*, 31(5):1117–1152.

Kilian, L. and Lütkepohl, H. (2017). *Structural Vector Autoregressive Analysis*. Cambridge University Press.

Miranda-Agrippino, S. and Ricco, G. (2021). The Transmission of Monetary Policy Shocks. *American Economic Journal: Macroeconomics*, 13(3):74–107.

Montiel Olea, J. L., Stock, J. H., and Watson, M. W. (2021). Inference in Structural Vector Autoregressions identified with an external instrument. *Journal of Econometrics*, 225(1):74–87.

Onatski, A. (2010). Determining the Number of Factors from Empirical Distribution of Eigenvalues. *The Review of Economics and Statistics*, 92(4):1004–1016.

Ramey, V. A. (2016). Macroeconomic Shocks and Their Propagation.

Romer, C. D. and Romer, D. H. (2004). A New Measure of Monetary Shocks: Derivation and Implications. *THE AMERICAN ECONOMIC REVIEW*, 94(4).

Sargent, T. and Sims, C. (1977). Business cycle modeling without pretending to have too much a priori economic theory. Working Paper 55, Federal Reserve Bank of Minneapolis.

Schlaak, T., Rieth, M., and Podstawski, M. (2023). Monetary policy, external instruments, and heteroskedasticity. *Quantitative Economics*, 14(1):161–200.

Sims, C. A. (1980). Macroeconomics and Reality. *Econometrica*, 48(1):1–48.

Sims, C. A. (1992). Interpreting the macroeconomic time series facts: The effects of monetary policy. *European Economic Review*, 36(5):975–1000.

Stock, J. H. and Watson, M. W. (2002). Forecasting Using Principal Components From a Large Number of Predictors. *Journal of the American Statistical Association*, 97(460):1167–1179.

Stock, J. H. and Watson, M. W. (2005). Implications of Dynamic Factor Models for VAR Analysis. Working Paper 11467, National Bureau of Economic Research.

Stock, J. H. and Watson, M. W. (2012). Disentangling the channels of the 2007-09 recession. *Brookings Papers on Economic Activity*, (1):81–135.

Stock, J. H. and Watson, M. W. (2016). Chapter 8 - Dynamic Factor Models, Factor-Augmented Vector Autoregressions, and Structural Vector Autoregressions in Macroeconomics. In Taylor, J. B. and Uhlig, H., editors, *Handbook of Macroeconomics*, volume 2, pages 415–525. Elsevier.

Stock, J. H. and Watson, M. W. (2018). Identification and Estimation of Dynamic Causal Effects in Macroeconomics Using External Instruments. *The Economic Journal*, 128(610):917–948.

White, H. (1984). *Asymptotic Theory for Econometricians.* Academic press.

Wieland, J. F. and Yang, M.-J. (2020). Financial Dampening. *Journal of Money, Credit and Banking*, 52(1):79–113.

Yamamoto, Y. (2019). Bootstrap inference for impulse response functions in factor-augmented vector autoregressions. *Journal of Applied Econometrics*, 34(2):247–267.

Yamamoto, Y. and Hara, N. (2022). Identifying factor-augmented vector autoregression models via changes in shock variances. *Journal of Applied Econometrics*, 37(4):722–745.

A Proofs

A.1 Preliminary

Lemma 1. *Under Assumptions 1 to 4,*

a) $\frac{1}{T}\left\|\widehat{F}_t - H_F^\top F_t\right\|^2 = O_p\left(\frac{1}{\delta_{NT}^2}\right)$ *and* $\frac{1}{T}\left\|\widehat{\mathcal{F}}_f - H_{\mathcal{F}}^\top \mathcal{F}_t\right\|^2 = O_p\left(\frac{1}{\delta_{NT}^2}\right)$.

b) $\widehat{V}_F \xrightarrow{p} V_F$*, where* V_F *is the diagonal matrix consisting of the eigenvalues of* $\Sigma_F \Sigma_\Lambda$ *in descending order.*

c) H_F *and* $H_{\mathcal{F}}$ *are* $O_p(1)$ *and nonsingular.*

Proof of Lemma 1.

a) The first equation is the same as Lemma A1 of Bai (2003). The second equation holds by definition of $H_{\mathcal{F}}$ and the first equation.

b) This is Lemma A3 of Bai (2003).

c) $\|H_F\| \leq \left\|\frac{\widehat{F}^\top \widehat{F}}{T}\right\|^{\frac{1}{2}} \left\|\frac{F^\top F}{T}\right\|^{\frac{1}{2}} \left\|\frac{\Lambda^\top \Lambda}{N}\right\| \left\|V_F^{-1}\right\| = O_p(1)$ by Assumptions 1 and 2 and Lemma 1. The matrices H_F and $H_{\mathcal{F}}$ are nonsingular by Lemma A2 of Han and Inoue (2015).

■

Lemma 2. *Under Assumptions 1 to 6,*

a) $\frac{1}{T}\sum_{t=1}^{T}(\widehat{F}_t - H_F^\top F_t)[F_t^\top, e_{it}, \eta_t^\top] = O_p\left(\frac{1}{\delta_{NT}^2}\right)$ *for* $i = 1, \ldots, N$*,*

b) $\frac{1}{T}(\widehat{\mathcal{F}} - \mathcal{F}H_{\mathcal{F}})^\top \left[\mathcal{F} \vdots \eta\right] = O_p\left(\frac{1}{\delta_{NT}^2}\right)$*.*

Proof of Lemma 2.

a) Lemmas B.1 and B.2 of Bai (2003) imply $\frac{1}{T}\sum_{t=1}^{T}(\widehat{F}_t - H_F^\top F_t)(F_t^\top, e_{it}) = O_p\left(\frac{1}{\delta_{NT}^2}\right)$ for $i = 1, \ldots, N$. Lemma 2 a) of Han (2018) shows that $\frac{1}{T}\sum_{t=1}^{T}(\widehat{F}_t - H_F^\top F_t)\eta_t^\top = O_p\left(\frac{1}{\delta_{NT}^2}\right)$.

b) This is the same as Lemma 2 b) of Han (2018).

■

Lemma 3. *Under Assumptions 1 to 6,* $\frac{1}{T}\sum_{t=p+1}^{T}\left\|\widehat{\eta}_t - H_\eta^\top \eta_t\right\|^2 = O_p\left(\frac{1}{\delta_{NT}^2}\right)$*.*

Proof of Lemma 3. Let $S_1 = \eta\Theta^\top + e$, $S_2 = -P_{\widehat{\mathcal{F}}}(\eta\Theta^\top + e) + M_{\widehat{\mathcal{F}}}\mathcal{F}\Pi^\top$, S_{1t} be the transpose of the tth row of $S1$ and S_{2t} be the transpose of the tth row of S_2. By eigen-identity, we have

$$\widehat{\eta} - \eta H_\eta = \frac{1}{TN}\left(\eta\Theta^\top e^\top + e\Theta\eta^\top + ee^\top + S_1 S_2^\top + S_2 S_1^\top + S_2 S_2\right)\widehat{\eta}\widehat{V}_\eta^{-1} \tag{A.1}$$

where $\widehat{V}_\eta$ denotes the diagonal matrix consisting of the first q eigenvalues of $\widehat{X}\widehat{X}^\top/NT$ in descending order. Hence, we obtain

$$\widehat{\eta}_t - H_\eta^\top \eta_t = \frac{1}{TN}\widehat{V}_\eta^{-1}\widehat{\eta}^\top\left(ee_t + \eta\Theta^\top e_t + e\Theta\eta_t + S_2S_{1t} + S_1S_{2t} + S_2S_{2t}\right). \tag{A.2}$$

It is sufficient to show that

$$\begin{aligned}
\frac{1}{T}\|c_{lt}\|^2 &= O_p\left(\frac{1}{\delta_{NT}^2}\right) \quad \text{for} \quad l = 1, \dots, 4,\\
\frac{1}{T}\sum_{t=1}^T\left\|\frac{1}{TN}\widehat{\eta}^\top S_2S_{1t}\right\|^2 &= O_p\left(\frac{1}{\delta_{NT}^2}\right),\\
\frac{1}{T}\sum_{t=1}^T\left\|\frac{1}{TN}\widehat{\eta}^\top S_1S_{2t}\right\|^2 &= O_p\left(\frac{1}{\delta_{NT}^2}\right), \quad \text{and}\\
\frac{1}{T}\sum_{t=1}^T\left\|\frac{1}{TN}\widehat{\eta}^\top S_2S_{2t}\right\|^2 &= O_p\left(\frac{1}{\delta_{NT}^2}\right).
\end{aligned}$$

The proof of $\frac{1}{T}\sum_{t=1}^T\|c_{lt}\|^2 = O_p\left(\frac{1}{\delta_{NT}^2}\right)$ is the same as the proof of Theorem 1 in Bai and Ng (2002).

Note that, since $\widehat{\mathcal{F}}^\top\widehat{X} = 0$, $\widehat{\mathcal{F}}^\top\widehat{\eta} = \frac{1}{TN}\widehat{\mathcal{F}}^\top\widehat{X}\widehat{X}^\top\widehat{\eta}\widehat{V}_\eta^{-1} = 0$, implying that $\widehat{\eta}^\top P_{\widehat{\mathcal{F}}} = 0$ and $\widehat{\eta}^\top M_{\widehat{\mathcal{F}}} = \widehat{\eta}^\top$.

We have the following identities for the terms $\frac{1}{TN}\widehat{\eta}^\top S_2S_{1t}$, $\frac{1}{TN}\widehat{\eta}^\top S_1S_{2t}$, and $\frac{1}{TN}\widehat{\eta}^\top S_2S_{2t}$:

$$\begin{aligned}
\frac{1}{TN}\widehat{\eta}^\top S_2S_{1t} &= \frac{1}{TN}\widehat{\eta}^\top\left[-P_{\widehat{\mathcal{F}}}\left(\eta\Theta^\top + e\right) + M_{\widehat{\mathcal{F}}}\mathcal{F}\Pi^\top\right](\Theta\eta_t + e_t)\\
&= \frac{1}{TN}\widehat{\eta}^\top\mathcal{F}\Pi^\top\Theta\eta_t + \frac{1}{TN}\widehat{\eta}^\top\mathcal{F}\Pi^\top e_t\\
&= c_{5t} + c_{6t},\\
\frac{1}{TN}\widehat{\eta}^\top S_1S_{2t} &= \frac{1}{TN}\widehat{\eta}^\top\left(\eta\Theta^\top + e\right)\left[-\left(\Theta\eta^\top + e^\top\right)\widehat{\mathcal{F}}\left(\widehat{\mathcal{F}}^\top\widehat{\mathcal{F}}\right)^{-1}\widehat{\mathcal{F}}_t + \Pi\left(\mathcal{F}_t - \mathcal{F}^\top\widehat{\mathcal{F}}\left(\widehat{\mathcal{F}}^\top\widehat{\mathcal{F}}\right)^{-1}\widehat{\mathcal{F}}_t\right)\right]\\
&= c_{7t} + c_{8t} + c_{9t} + c_{10t} + c_{11t} + c_{12t},
\end{aligned}$$

where

$$
\begin{aligned}
c_{7t} &= -\frac{1}{TN}\widehat{\eta}^\top \eta\Theta^\top\Theta\eta^\top\widehat{\mathcal{F}}\left(\widehat{\mathcal{F}}^\top\widehat{\mathcal{F}}\right)^{-1}\widehat{\mathcal{F}}_t,\\
c_{8t} &= -\frac{1}{TN}\widehat{\eta}^\top \eta\Theta^\top e^\top\widehat{\mathcal{F}}\left(\widehat{\mathcal{F}}^\top\widehat{\mathcal{F}}\right)^{-1}\widehat{\mathcal{F}}_t,\\
c_{9t} &= \frac{1}{TN}\widehat{\eta}^\top \eta\Theta^\top\Pi\left[\mathcal{F}_t-\widehat{\mathcal{F}}\left(\widehat{\mathcal{F}}^\top\widehat{\mathcal{F}}\right)^{-1}\widehat{\mathcal{F}}_t\right],\\
c_{10t} &= -\frac{1}{TN}\widehat{\eta}^\top e\Theta\eta^\top\widehat{\mathcal{F}}\left(\widehat{\mathcal{F}}^\top\widehat{\mathcal{F}}\right)^{-1}\widehat{\mathcal{F}}_t,\\
c_{11t} &= -\frac{1}{TN}\widehat{\eta}^\top ee^\top\widehat{\mathcal{F}}\left(\widehat{\mathcal{F}}^\top\widehat{\mathcal{F}}\right)^{-1}\widehat{\mathcal{F}}_t,\\
c_{12t} &= \frac{1}{TN}\widehat{\eta}^\top e\Pi\left[\mathcal{F}_t-\widehat{\mathcal{F}}\left(\widehat{\mathcal{F}}^\top\widehat{\mathcal{F}}\right)^{-1}\widehat{\mathcal{F}}_t\right],\\
\frac{1}{TN}\widehat{\eta}^\top S_2 S_{2t} &= \frac{1}{TN}\widehat{\eta}^\top \mathcal{F}\Pi^\top\left[-\left(\Theta\eta^\top+e^\top\right)\widehat{\mathcal{F}}\left(\widehat{\mathcal{F}}^\top\widehat{\mathcal{F}}\right)^{-1}\widehat{\mathcal{F}}_t+\Pi\left(\mathcal{F}_t-\widehat{\mathcal{F}}\left(\widehat{\mathcal{F}}^\top\widehat{\mathcal{F}}\right)^{-1}\widehat{\mathcal{F}}_t\right)\right]\\
&= c_{13t}+c_{14t}+c_{15t}\\
c_{13t} &= -\frac{1}{TN}\widehat{\eta}^\top \mathcal{F}\Pi^\top\Theta\eta^\top\widehat{\mathcal{F}}\left(\widehat{\mathcal{F}}^\top\widehat{\mathcal{F}}\right)^{-1}\widehat{\mathcal{F}}_t,\\
c_{14t} &= -\frac{1}{TN}\widehat{\eta}^\top \mathcal{F}\Pi^\top e^\top\widehat{\mathcal{F}}\left(\widehat{\mathcal{F}}^\top\widehat{\mathcal{F}}\right)^{-1}\widehat{\mathcal{F}}_t,\\
c_{15t} &= \frac{1}{TN}\widehat{\eta}^\top \mathcal{F}\Pi^\top\Pi\left[\mathcal{F}_t-\widehat{\mathcal{F}}\left(\widehat{\mathcal{F}}^\top\widehat{\mathcal{F}}\right)^{-1}\widehat{\mathcal{F}}_t\right].
\end{aligned}
$$

Han (2018) proves that $\frac{1}{T}\sum_{t=1}^T \|c_{lt}\|^2 = O_p\left(\frac{1}{\delta_{NT}^2}\right)$ for $l = 5, \ldots, 15$. These quantities are used in subsequent proofs, so are provided here for convenience. ■

Lemma 4. *Under Assumptions 1 to 6,*

a) $\frac{1}{T}(\widehat{\eta}-\eta H_\eta)^\top \eta = O_p\left(\frac{1}{\delta_{NT}^2}\right)$,

b) $\frac{1}{T}(\widehat{\eta}-\eta H_\eta)^\top Z = O_p\left(\frac{1}{\delta_{NT}^2}\right)$, *and*

c) $H_\eta^\top = H_\eta^{-1}\Sigma_\eta^{-1} + O_p\left(\frac{1}{\delta_{NT}^2}\right)$.

Lemma 5. *Under Assumptions 1 to 6,* $\frac{1}{T}\sum_{t=1}^T \|c_{lt}\|^2 = O_p\left(\frac{1}{\delta_{NT}^2}\right)$ *for* $l = 1, 2, 3, 4, 5, 6, 8, 10, 11, 12, 13, 14$ *and* 15 *where each* c_{lt} *are defined in the proof of Lemma 3.*

Proof of Lemma 4. Recall that $\widehat{\eta}_t - H_\eta^\top \eta_t = \widehat{V}_\eta^{-1}\sum_{l=1}^{15} c_{lt}$, so it suffices to prove that $\frac{1}{T}\sum_{t=1}^T c_{lt}\eta_t = O_p\left(\frac{1}{\delta_{NT}^2}\right)$ for Lemma 4 (a) and $\frac{1}{T}\sum_{t=1}^T c_{lt}Z_t = O_p\left(\frac{1}{\delta_{NT}^2}\right)$ for Lemma 4 (b), where $l = 1, \ldots, 15$.

Lemma 4 (a) has been proven by Han (2018), so we focus on Lemma 4 (b), which can be proven similarly. By the CS-inequality, for $l = 1, 2, 3, 4, 5, 8, 10 - 15$, we have

$$\left\| \frac{1}{T} \sum_{t=1}^{T} c_{lt} Z_t^\top \right\| \leq \left(\sum_{t=1}^{T} \|c_{lt}\|^2 \frac{1}{T} \sum_{t=1}^{T} \|Z_t\|^2 \right)^{1/2} = O_p\left(\frac{1}{\delta_{NT}^2} \right), \tag{A.3}$$

where $\frac{1}{T} \sum_{t=1}^{T} \|c_{lt}\|^2 = O_p\left(\frac{1}{\delta_{NT}^2} \right)$ for $l = 1, 2, 3, 4, 5, 6, 8, 10 - 15$ by Lemma 5. Thus, it remains to prove that $\frac{1}{T} \sum_{t=1}^{T} c_{lt} Z_t^\top = O_p\left(\frac{1}{\delta_{NT}^2} \right)$ for $l = 3, 7, 9$.

The term $\frac{1}{T} \sum_{t=1}^{T} c_{3t} Z_t^\top$ can be bounded by $\frac{1}{TN} \widehat{\eta}^\top \eta \Theta^\top e_t$

$$\begin{aligned} \left\| \frac{1}{T} \sum_{t=1}^{T} c_{lt} Z_t^\top \right\| &\leq \left\| \frac{1}{T} \widehat{\eta}^\top \eta \right\| \left\| \frac{1}{TN} \sum_{t=1}^{T} G^\top \Lambda e_t Z_t^\top \right\| \\ &\leq \left\| \frac{1}{T} \widehat{\eta}^\top \eta \right\| \|G\| \left\| \frac{1}{TN} \sum_{t=1}^{T} \sum_{k=1}^{N} \lambda_k e_{kt} Z_t^\top \right\| = O_p\left(\frac{1}{TN} \right). \end{aligned}$$

Next, the term $\frac{1}{T} \sum_{t=1}^{T} c_{7t} Z_t^\top$ can be rewritten as

$$\begin{aligned} \frac{1}{T} \sum_{t=1}^{T} c_{7t} Z_t^\top &= -\frac{\widehat{\eta}^\top \eta}{TN} \frac{\Theta^\top \Theta}{N} \frac{\eta^\top \widehat{\mathcal{F}}}{T} \left(\frac{\widehat{\mathcal{F}}^\top \widehat{\mathcal{F}}}{T} \right)^{-1} \frac{1}{T} \sum_{t=1}^{T} \widehat{\mathcal{F}}_t Z_t^\top \\ &= O_p\left(\frac{1}{\delta_{NT}} \right) \frac{1}{T} \sum_{t=1}^{T} \widehat{\mathcal{F}}_t Z_t^\top \\ &= O_p\left(\frac{1}{\delta_{NT}^2} \right). \end{aligned}$$

Finally, the term $\frac{1}{T} \sum_{t=1}^{T} c_{9t} Z_t^\top$ can be expressed as

$$\begin{aligned} &\frac{\widehat{\eta}^\top \eta}{T} \frac{\Theta^\top \Pi}{N} \frac{1}{T} \sum_{t=1}^{T} \left[\mathcal{F}_t - \mathcal{F}^\top \widehat{\mathcal{F}} \left(\widehat{\mathcal{F}}^\top \widehat{\mathcal{F}} \right)^{-1} \widehat{\mathcal{F}}_t \right] Z_t^\top \\ =& O_p(1) \frac{1}{T} \sum_{t=1}^{T} H_{\mathcal{F}}^{-\top} \left(H_{\mathcal{F}}^\top \mathcal{F}_t - \widehat{\mathcal{F}}_t \right) Z_t^\top + O_p(1) H_{\mathcal{F}}^{-\top} \frac{\left(\widehat{\mathcal{F}}^\top - H_{\mathcal{F}}^\top \mathcal{F}^\top \right) \widehat{\mathcal{F}}}{T} \left(\frac{\widehat{\mathcal{F}}^\top \widehat{\mathcal{F}}}{T} \right)^{-1} \frac{1}{T} \sum_{t=1}^{T} \widehat{\mathcal{F}}_t Z_t^\top \\ =& O_p\left(\frac{1}{\delta_{NT}^2} \right). \end{aligned}$$

■

Proof of Lemma 4 (c). Lemmas 4 (a) and 3 imply that

$$\begin{aligned}\frac{\widehat{\eta}^\top \widehat{\eta} - H_\eta^\top \eta^\top \eta H_\eta}{T-p} &= \frac{\left(\widehat{\eta} - \eta H_\eta\right)^\top \left(\widehat{\eta} - \eta H_\eta\right) + \left(\widehat{\eta} - \eta H_\eta\right)^\top \eta H_\eta + H_\eta^\top \eta^\top \left(\widehat{\eta} - \eta H_\eta\right)}{T-p} \\ &= O_p\left(\frac{1}{\delta_{NT}^2}\right).\end{aligned}$$

Thus,

$$\begin{aligned}\frac{\left(\widehat{\eta}^\top \widehat{\eta} - H_\eta^\top \eta^\top \eta H_\eta\right)}{T-p} &= I_q - \frac{H_\eta^\top \eta^\top \eta H_\eta}{T-p} \\ &= O_p\left(\frac{1}{\delta_{NT}^2}\right).\end{aligned}$$

Next, $\eta^\top \eta/(T-p) = \Sigma_\eta$ by Assumption 8, which means

$$H_\eta^\top = H_\eta^{-1} \Sigma_\eta^{-1} + O_p\left(\frac{1}{\delta_{NT}^2}\right).$$

■

A.2 Main Proofs

Proof of Theorem 1. Applying the formula for $\widehat{\delta}$, we have

$$\begin{aligned}\sqrt{T}\left(\widehat{\delta} - \delta\right) &= \left(\mathcal{A}_T W_T \mathcal{A}_T^\top\right)^{-1} \mathcal{A}_T W_T \mathcal{G}_T - \sqrt{T}\delta \\ &= \left(\mathcal{A}_T W_T \mathcal{A}_T^\top\right)^{-1} \mathcal{A}_T W_T \mathcal{B}\end{aligned}$$

where

$$
\begin{aligned}
\mathcal{A}_T &= I_{q-1} \otimes \left(\frac{1}{T-p} \sum_{t=p+1}^{T} \widehat{\eta}_{1t} Z_t^\top \right), \\
\mathcal{B} &= \frac{1}{\sqrt{T}} \left[\sum_{t=p+1}^{T} \widehat{\eta}_{-1t} \otimes Z_t - \left(I_{q-1} \otimes \sum_{t=1}^{T} Z_t \widehat{\eta}_{1t} \delta \right) \right] \\
&= \frac{1}{\sqrt{T}} \sum_{t=p+1}^{T} \left(\widehat{\eta}_{-1t} - \delta \widehat{\eta}_{1t} \right) \otimes Z_t \\
&= \frac{1}{\sqrt{T}} \sum_{t=p+1}^{T} \left(\mathbb{S}_\delta \widehat{\eta} \right) \otimes \left(Z_t - \mu_Z \right).
\end{aligned}
$$

We need to study the asymptotic distributions of $\mathcal{A}_T$ and $\mathcal{B}$. For $\mathcal{B}$, decompose it into $\mathcal{B} = \mathcal{B}_1 + \mathcal{B}_2$, where

$$
\begin{aligned}
\mathcal{B}_1 &= \frac{1}{\sqrt{T}} \sum_{t=1}^{T} \left(\mathbb{S}_\delta H_\eta^\top \eta_t \right) \otimes \left(Z_t - \mu_Z \right), \\
\mathcal{B}_2 &= \frac{1}{\sqrt{T}} \sum_{t=1}^{T} \left[\mathbb{S}_\delta \left(\widehat{\eta}_t - H_\eta^\top \eta_t \right) \right] \otimes \left(Z_t - \mu_Z \right).
\end{aligned}
$$

For $\mathcal{B}_1$, note that $E\left[\mathbb{S}_\delta H_\eta^\top \eta_t \otimes \left(Z_t - \mu_Z \right) \right] = 0$ following the moment condition for the estimation of δ and $E(H_\eta^\top \eta_t) = H_\eta^\top E(\eta_t) = 0$. It follows that

$$
\begin{aligned}
\mathcal{B}_1 &= \frac{1}{\sqrt{T}} \sum_{t=p+1}^{T} \left\{ \left(\mathbb{S}_\delta H_\eta^\top \eta_t \right) \otimes \left(Z_t - \mu_Z \right) - E\left[\left(\mathbb{S}_\delta H_\eta^\top \eta_t \right) \otimes \left(Z_t - \mu_Z \right) \right] \right\} \\
&= \left(\mathbb{S}_\delta H_\eta^\top \otimes I_k \right) \left\{ \frac{1}{\sqrt{T}} \sum_{t=p+1}^{T} \eta_t \otimes \left(Z_t - \mu_Z \right) - E\left[\eta_t \otimes \left(Z_t - \mu_Z \right) \right] \right\} \\
&= \left(\mathbb{S}_\delta H_\eta^\top \otimes I_k \right) \frac{1}{\sqrt{T}} \operatorname{vec}\left(Z^\top \eta - E(Z^\top \eta) \right) \\
&\xrightarrow{d} N\left(0_{kq \times 1}, \left(\mathbb{S}_\delta \bar{H}_\eta^\top \otimes I_k \right) \Sigma_i^{(1)} \left(\mathbb{S}_\delta \bar{H}_\eta^\top \otimes I_k \right)^\top \right).
\end{aligned}
$$

To study $\mathcal{B}_2$, rewrite it as

$$\begin{aligned}
\mathcal{B}_2 &= \frac{1}{\sqrt{T}} \sum_{t=p+1}^{T} \left[\mathbb{S}_\delta \left(\widehat{\eta}_t - H_\eta^\top \eta_t \right) \right] \otimes (Z_t - \mu_Z) \operatorname{vec}(1) \\
&= \frac{1}{\sqrt{T}} \sum_{t=p+1}^{T} \operatorname{vec} \left((Z_t - \mu_Z) \left(\widehat{\eta}_t - H_\eta^\top \eta_t \right)^\top \mathbb{S}_\delta \right) \\
&= \operatorname{vec} \left(\frac{1}{\sqrt{T}} (Z - \mu_Z)^\top (\widehat{\eta} - \eta H_\eta) \mathbb{S}_\delta \right) \\
&= O_p \left(\frac{\sqrt{T}}{\delta_{NT}^2} \right), \qquad \text{(A.4)}
\end{aligned}$$

where the last equality follows because $\frac{1}{T} Z^\top (\widehat{\eta} - \eta H_\eta) = O_p \left(\frac{1}{\delta_{NT}^2} \right)$. Therefore, $\mathcal{B}_2$ is asymptotically negligible.

To study the limit of $\mathcal{A}_T$, we have

$$\begin{aligned}
\frac{1}{T-p} \sum_{t=p+1}^{T} \widehat{\eta}_{1t} Z_t^\top &= \frac{1}{T-p} \sum_{t=p+1}^{T} (H_\eta^\top \eta)_{1t} Z_t^\top + \frac{1}{T-p} \sum_{t=p+1}^{T} \left(\widehat{\eta}_{1t} - (H_\eta^\top \eta)_{1t} \right) Z_t^\top \\
&\xrightarrow{p} E(\mathbb{S}_1 H_\eta^\top \eta_t Z_t^\top),
\end{aligned}$$

where $\mathbb{S}_1 = [1, 0_{1\times(q-1)}]$, and $\frac{1}{T-p} \sum_{t=p+1}^{T} \left(\widehat{\eta}_{1t} - (H_\eta^\top \eta)_{1t} \right) Z_t^\top \xrightarrow{p} 0$ using similar arguments used in the proof of $\mathcal{B}_2$. Therefore,

$$\mathcal{A}_T \xrightarrow{p} \mathcal{A} = I_{q-1} \otimes \mathbb{S}_1 \bar{H}_\eta^\top E(\eta_t Z_t^\top). \qquad \text{(A.5)}$$

The weighting matrix W is a full rank matrix and thus invertible. The optimal choice of the weighting matrix follows from standard arguments for GMM estimators.

Collecting the results yields the following distribution as required:

$$\begin{aligned}
\sqrt{T} \left(\widehat{\delta} - \delta \right) &\xrightarrow{p} \left(\mathcal{AWA}^\top \right)^{-1} \mathcal{AWB}_1 \\
&\xrightarrow{d} \left(\mathcal{AWA}^\top \right)^{-1} \mathcal{AW} N \left(0_{kq\times 1}, \left(\mathbb{S}_\delta \bar{H}_\eta^\top \otimes I_k \right) \Sigma_i^{(1)} \left(\mathbb{S}_\delta \bar{H}_\eta^\top \otimes I_k \right)^\top \right),
\end{aligned}$$

where $\mathcal{A} = I_{q-1} \otimes \mathbb{S}_1 H_\eta^\top E(\eta_t Z_t^\top)$, and $\Sigma_i^{(1)}$ is the upper left block of Σ_i. ■

Proof of Proposition 1. By definition,

$$\begin{aligned}
\widehat{G} &= \frac{1}{T-p}\sum_{t=p+1}^{T}\widehat{F}_t\widehat{\eta}_t \\
&= \frac{1}{T-p}\sum_{t=p+1}^{T}H_F^\top F_t\widehat{\eta}_t + \frac{1}{T-p}\sum_{t=p+1}^{T}\left(\widehat{F}_t - H_F^\top F_t\right)\left(\widehat{\eta}_t - H_\eta^\top \eta_t\right)^\top \\
&\quad + \frac{1}{T-p}\sum_{t=p+1}^{T}(\widehat{F}_t - H_F^\top F_t)\eta_t^\top H_\eta \\
&= \frac{1}{T-p}\sum_{t=p+1}^{T}\left(H_F^\top \Phi H_{\mathcal{F}}^{-\top} H_{\mathcal{F}}\mathcal{F}_{t-1}\widehat{\eta}_t + H_F^\top G\eta_t\widehat{\eta}_t\right) + O_p\left(\frac{1}{\delta_{NT}^2}\right) \\
&= \frac{1}{T-p}\sum_{t=p+1}^{T}\left(H_F^\top \Phi H_{\mathcal{F}}^{-\top} H_{\mathcal{F}}\mathcal{F}_{t-1}\widehat{\eta}_t + H_F^\top G\eta_t\eta_t^\top H_\eta\right) + O_p\left(\frac{1}{\delta_{NT}^2}\right). \qquad \text{(A.6)}
\end{aligned}$$

Recall that $\eta^\top\eta/(T-p) \xrightarrow{p} \Sigma_\eta$. By the fact that $\widehat{\mathcal{F}}^\top\widehat{\eta} = 0$, we have

$$\begin{aligned}
\widehat{G} - H_F^\top G\Sigma_\eta H_\eta &= \frac{1}{T-p}H_F^\top\Phi H_{\mathcal{F}}^{-\top}\left(\mathcal{F}H_{\mathcal{F}} - \widehat{\mathcal{F}}\right)^\top + O_p\left(\frac{1}{\delta_{NT}^2}\right) \\
&= \frac{H_F^\top\Phi H_{\mathcal{F}}^{-\top}\left[(\mathcal{F}H_{\mathcal{F}} - \widehat{\mathcal{F}})^\top(\widehat{\eta} - \eta H_\eta) + (\mathcal{F}H_{\mathcal{F}} - \widehat{\mathcal{F}})^\top\eta H_\eta\right]}{T-p} \\
&= O_p\left(\frac{1}{\delta_{NT}^2}\right).
\end{aligned}$$

■

Proof of Proposition 2. Recall that $\widehat{\theta}_i = \widehat{G}^\top\widehat{\lambda}_i$. Therefore,

$$\begin{aligned}
\sqrt{T}\widehat{\theta}_i &= \sqrt{T}\left(H_\eta\Sigma_\eta G^\top H_F^\top\widehat{\lambda}_i\right) + O_p\left(\frac{\sqrt{T}}{\delta_{NT}^2}\right) \\
&= \left(H_\eta\Sigma_\eta G^\top H_F H_F^{-1}\lambda_i\right) + \left(H_\eta\Sigma_\eta G^\top H_F\right)\sqrt{T}\left(\widehat{\lambda}_i - H_F^{-1}\lambda_i\right) + O_p\left(\frac{\sqrt{T}}{\delta_{NT}^2}\right) \\
\sqrt{T}\left(\widehat{\theta}_i - H_\eta^{-1}\theta_i\right) &= \left(H_\eta^{-1}G^\top H_F\right)\sqrt{T}\left(\widehat{\lambda}_i - H_F^{-1}\lambda_i\right) + O_p\left(\frac{\sqrt{T}}{\delta_{NT}^2}\right).
\end{aligned}$$

■

Proof of Proposition 3 (a).

For the estimator $\widehat{a}_1$, its distribution is based on the distribution of $\widehat{\delta}$. By definition, we have

$$\begin{aligned}
\sqrt{T}(\widehat{a} - a_1^*) &= \sqrt{T}\mathbb{S}_1 \begin{bmatrix} 0 \\ \widehat{\delta} - \delta \end{bmatrix} \\
&\xrightarrow{p} \bar{\mathbb{S}}_1 \left(\mathcal{A}\mathcal{W}\mathcal{A}^\top\right)^{-1} \mathcal{A}\mathcal{W}(\mathbb{S}_\delta \bar{H}_\eta^\top \otimes I_k) \frac{1}{\sqrt{T}} \operatorname{vec}\left(Z^\top \eta - E(Z^\top \eta)\right),
\end{aligned}$$

which follows from Theorem 1 (a).

For the OLS estimator of the factor loadings $\widehat{\lambda}_i$, Bai (2003) shows that

$$\widehat{\lambda}_i - H_F^{-1}\lambda_i = \frac{1}{T}\bar{H}_F^\top F^\top e_i + O_p\left(\frac{1}{\delta_{NT}^2}\right). \tag{A.7}$$

For the estimators $\widehat{\Phi}$ and $\widehat{\Psi}$, Han (2018) shows that

$$\begin{aligned}
\sqrt{T}\operatorname{vec}\left(\widehat{\Phi}^\top - H_{\mathcal{F}}^{-1}\Phi^\top H_F\right) &= \left[H_F^\top G \otimes \left(\frac{H_{\mathcal{F}}^\top \mathcal{F}^\top \mathcal{F} H_{\mathcal{F}}}{T-p}\right)^{-1} H_{\mathcal{F}}^\top\right] \times \frac{\sqrt{T}\sum_{t=p+1}^T vec(\mathcal{F}_t^\top \eta_t)}{T-p} \\
&\quad + O_p\left(\frac{\sqrt{T}}{\delta_{NT}^2}\right), \\
\sqrt{T}\operatorname{vec}\left(\widehat{\Psi}_s^\top - H_F^{-1}\Psi_s^\top H_F\right) &= R_s\sqrt{T}\operatorname{vec}\left(\widehat{\Phi}^\top - H_{\mathcal{F}}^{-1}\Phi^\top H_F\right) + o_p(1),
\end{aligned} \tag{A.8}$$

where $\bar{R}_s = \sum_{j=1}^s \left(\bar{H}_F^\top \Psi_{j-1}\bar{H}_F^{-\top} \otimes \left[\bar{H}_F^{-\top}\Psi_{s-j}^\top \bar{H}_F, \bar{H}_F^{-\top}\Psi_{s-j-1}^\top \bar{H}_F, \ldots, \bar{H}_F^{-\top}\Psi_{s-j-p+1}^\top \bar{H}_F\right]\right)$ with $\Psi_0 = I_r$ and $\Psi_s = 0_{r\times r}$ for $s < 0$, which follows by (11.7.1) to (11.7.5) of Hamilton (2020).

Combining the above gives

$$\sqrt{T}\begin{bmatrix} \widehat{a}_1 - a_1^* \\ \widehat{\lambda}_i - H_F^{-1}\lambda_i \\ \operatorname{vec}\left(\widehat{\Psi}^\top - H_F^{-1}\Psi^\top H_F\right) \end{bmatrix} = B_s \frac{1}{\sqrt{T}} \begin{bmatrix} \operatorname{vec}\left(Z^\top \eta - E(Z^\top \eta)\right) \\ F^\top e_i \\ vec(\mathcal{F}^\top \eta) \end{bmatrix} + o_p(1).$$

■

Proof of Proposition 3 (b). This is directly implied by Proposition 3 (a). ■

Proof of Theorem 2 (a). By adding and subtracting terms, we have

$$
\begin{aligned}
&\sqrt{T}\left(\widehat{\theta}_i^\top \widehat{a}_1 - \theta_i a_1\right) \\
&= \sqrt{T}\widehat{\theta}_i^\top \left(\widehat{a} - H_\eta^\top a_1\right) + \sqrt{T}\left(\widehat{\theta}_i^\top - \theta_i^\top H_\eta^{-1}\right) H_\eta^\top a_1 \\
&= \sqrt{T}\widehat{\theta}_i^\top \left(\widehat{a}_1 - a_1^*\right) + \sqrt{T} a_a^\top H_\eta \left(\widehat{\theta}_i - H_\eta^{-1}\theta_i\right) + o_p(1) \\
&= \left[\widehat{\theta}_i^\top [I_q \vdots 0_{q\times r}] + a_1^\top H_\eta H_\eta^{-1} G^\top H_F [0_{r\times qk} \vdots I_r]\right] \sqrt{T} \begin{bmatrix} \widehat{a}_1 - a_1^* \\ \widehat{\lambda}_i - H_F^{-1}\lambda_i \end{bmatrix} + o_p(1).
\end{aligned}
$$

Because $\widehat{\theta}_i^\top$ is an estimate of $\theta_i^\top H_\eta^{-\top}$, taking the probability limit and applying Proposition 3 (b) yields the result. ■

Proof of Theorem 2 (b). By adding and subtracting terms, we have the following asymptotic expansion

$$
\begin{aligned}
&\sqrt{T}\left(\widehat{\lambda}_i^\top \widehat{\Psi}_s \widehat{G} \widehat{a}_1 - \lambda_i^\top H_F^{-\top} H_F^\top \Psi_s H_F^{-\top} H_F^\top G \Sigma_\eta H_\eta H_\eta^{-1} \Sigma_\eta^{-1} a_1\right) \\
&= \sqrt{T}\widehat{\lambda}_i^\top \widehat{\Psi}_s \widehat{G}\left(\widehat{a}_1 - a_1^*\right) + \sqrt{T} a_1^\top G^\top H_F \widehat{\Psi}_s^\top \left(\widehat{\lambda}_i - H_F^{-1}\lambda_i\right) \\
&\quad + \sqrt{T} H_F^{-1}\lambda_i^\top \left(\widehat{\Psi}_s - H_F^\top \Psi_s H_F^{-\top}\right) H_F^\top G a_1^* + o_p(1) \\
&= \sqrt{T}\widehat{\lambda}_i \widehat{\Psi}\widehat{G}\left(\widehat{a}_1 - a_1^*\right) + \sqrt{T} a_1^{*\top} G^\top H_F \widehat{\Psi}_s^\top \left(\widehat{\lambda}_i - H_F^{-1}\lambda_i\right) \\
&\quad + \left(\lambda_i^\top H_F^{-\top} \otimes a_1^{*\top} G^\top H_F\right) \sqrt{T}\,\mathrm{vec}\left(\widehat{\Psi}_s - H_F^\top \Psi_s H_F^{-\top}\right) + o_p(1) \\
&= \Big[\widehat{\lambda}_i^\top \widehat{\Psi}_s \widehat{G}[I_q \vdots 0_{q\times r} \vdots 0_{q\times r^2}] + a_1^{*\top} G^\top H_F \widehat{\Psi}_s^\top [0_{r\times q} \vdots I_r \vdots 0_{r\times r^2}] \qquad \text{(A.9)} \\
&\quad + \left(\lambda_i^\top H_F^{-\top}\right) \otimes \left(a_1^{*\top} G^\top H_F\right) [0_{r^2\times q} \vdots 0_{r^2\times r} \vdots I_{r^2}]\Big] \sqrt{T} \begin{bmatrix} \widehat{a}_1 - a_1^* \\ \widehat{\lambda}_i - H_F^{-1}\lambda_i \\ \mathrm{vec}\left(\widehat{\Psi}_s^\top - H_F^{-1}\Psi_s^\top H_F\right) \end{bmatrix} + o_p(1).
\end{aligned}
$$

Hence, we have

$$
\sqrt{T}\left(\widehat{\lambda}_i^\top \widehat{\Psi}_s \widehat{G} \widehat{a}_1 - \lambda_i^\top \Psi_s G a_1\right) \xrightarrow{d} N\left(0, \bar{Q}_{2,i} B_s \Sigma_i B_s^\top \bar{Q}_{2,i}^\top\right),
$$

where

$$\bar{Q}_{2,i} = \lambda_i^\top \Psi_s G \Sigma_\eta \bar{H}_\eta C_3 + a_1^\top G^\top \Psi_s^\top \bar{H}_F C_4 + \left(\lambda_i^\top \bar{H}_F^{-\top} \otimes a_1^\top G^\top \bar{H}_F\right) C_5,$$

and $C_3 = [I_q \vdots 0_{q\times r} \vdots 0_{q\times r^2}]$, $C_4 = [0_{r\times q} \vdots I_r \vdots 0_{r\times r^2}]$ and $C_5 = [0_{r^2\times q} \vdots 0_{r^2\times r} \vdots I_{r^2}]$. Applying the distribution in Proposition 3 (a) yields the result. ■

Proof of Theorem 3 (a). The proof follows standard methods for proving overidentification tests. Note that, the sample moment condition, derivative, and first order condition are respectively

$$\mathcal{G}_T = \frac{1}{T-p} \sum_{t=p+1}^{T} \left(\widehat{\eta}_{-1t} \otimes Z_t\right), \tag{A.10}$$

$$\mathcal{A}_T = I_{q-1} \otimes \left(\frac{1}{T-p} \sum_{t=p+1}^{T} \widehat{\eta}_{1t} Z_t^\top\right), \tag{A.11}$$

$$\mathcal{A}_T W_T \mathcal{G}_T = 0. \tag{A.12}$$

Let $\widehat{\delta}$ denote a GMM estimator obtained with an optimal weight matrix W_T, i.e. $W_T \xrightarrow{p} W$, which is $\mathcal{B}\Sigma_i^{(1)}\mathcal{B}^\top$. Expand the moment condition about $\mathcal{G}_T(\delta)$ to obtain

$$\mathcal{G}_T(\widehat{\delta}) = \mathcal{G}_T(\delta) + \mathcal{A}_T(\delta^*)\left(\widehat{\delta} - \delta\right) + o\left\|\widehat{\delta} - \delta\right\| \tag{A.13}$$

where $\|\delta^* - \delta\| \leq \left\|\widehat{\delta} - \delta\right\|$. Substituting this back into the first order condition yields

$$\begin{aligned}
\mathcal{A}_T(\widehat{\delta})^\top W_T^{-1} \delta^* \left(\widehat{\delta} - \delta\right) &= -\mathcal{A}_T(\widehat{\delta})^\top W_T^{-1} \mathcal{G}_T(\delta) \\
\left(\widehat{\delta} - \delta\right) &= -\left(\mathcal{A}_T\left(\widehat{\delta}\right)^\top W_T^{-1} \mathcal{A}_T(\delta^*)\right)^{-1} \mathcal{A}_T(\widehat{\delta})^\top W_T^{-1} \mathcal{G}_T(\delta).
\end{aligned}$$

Substituting this back into the Taylor expansion gives

$$\begin{aligned}
\mathcal{G}_T\left(\widehat{\delta}\right) &= \mathcal{G}_T(\delta) + \mathcal{A}_T(\delta^*)\left(\widehat{\delta} - \delta\right) + o_p(1) \\
&= \mathcal{G}_T(\delta) - \mathcal{A}_T(\delta^*)\left(\mathcal{A}_T\left(\widehat{\delta}\right)^\top W_T^{-1} \mathcal{A}_T(\delta^*)\right)^{-1} \mathcal{A}_T\left(\widehat{\delta}\right)^\top W_T^{-1} \mathcal{G}_T(\delta) \\
&= \left(I - \mathcal{A}_T(\delta^*)\left(\mathcal{A}_T\left(\widehat{\delta}\right)^\top W_T^{-1} \mathcal{A}_T(\delta^*)\right)^{-1} \mathcal{A}_T\left(\widehat{\delta}\right)^\top W_T^{-1}\right) \mathcal{G}_T(\delta).
\end{aligned}$$

Because $\mathcal{A}_T\left(\widehat{\delta}\right) \xrightarrow{p} \mathcal{A}_T(\delta)$, $\mathcal{A}_T\left(\delta^*\right) \xrightarrow{p} \mathcal{A}_T(\delta)$ by the proof of Theorem 1 (a), and $W_T \xrightarrow{p} W$, the Cramer-Wold device and Slutsky's Theorem yield

$$\sqrt{T} W_T^{1/2} \mathcal{G}_T\left(\widehat{\delta}\right) \xrightarrow{p} \left(I - W^{-1/2}\mathcal{A}\left(\mathcal{A}^\top W^{-1}\mathcal{A}\right)^{-1}\mathcal{A}^\top W^{-1/2}\right)\mathbf{Z}_T$$

where $\mathbf{Z}_T = \sqrt{T}W^{-1/2}\mathcal{G}\left(\delta\right) \xrightarrow{d} \mathbf{Z} \sim N(0, I)$. It follows that by recognising that plim $\left|T\mathcal{Q}_T(\delta) - \mathbf{Z}_T^\top \mathbf{Z}_T\right| = 0$,

$$T\mathcal{Q}_T\left(\widehat{\delta}\right) \xrightarrow{d} \mathbf{Z}^\top\left(I - W^{-1/2}\mathcal{A}\left(\mathcal{A}^\top W^{-1}\mathcal{A}\right)^{-1}\mathcal{A}^\top W^{-1/2}\right)\mathbf{Z} = \chi^2_{(k-1)(q-1)}.$$

■

Proof of Theorem 3 (b) and Theorem 3 (c).

These follow the Proofs of Theorem 1 and 2 of Andrews (1999), respectively. ■

B Additional Simulation Results

T	N	h	$k=1$	$k=2$	$k=3$	$k=4$
250	125	0	0.919	0.897	0.889	0.881
		1	0.931	0.917	0.909	0.903
		2	0.950	0.942	0.939	0.935
		3	0.954	0.949	0.947	0.944
		6	0.941	0.938	0.938	0.935
		9	0.936	0.931	0.929	0.926
		12	0.929	0.925	0.924	0.922
	250	0	0.920	0.904	0.895	0.888
		1	0.937	0.924	0.916	0.911
		2	0.953	0.946	0.941	0.938
		3	0.954	0.949	0.945	0.944
		6	0.939	0.935	0.932	0.931
		9	0.934	0.927	0.925	0.923
		12	0.931	0.925	0.921	0.919
500	125	0	0.903	0.890	0.883	0.880
		1	0.912	0.901	0.892	0.887
		2	0.943	0.938	0.934	0.932
		3	0.946	0.943	0.941	0.940
		6	0.940	0.937	0.934	0.934
		9	0.931	0.926	0.924	0.923
		12	0.918	0.915	0.913	0.913
	250	0	0.909	0.898	0.892	0.889
		1	0.928	0.922	0.917	0.915
		2	0.947	0.943	0.941	0.940
		3	0.948	0.945	0.943	0.942
		6	0.939	0.935	0.933	0.932
		9	0.926	0.923	0.919	0.918
		12	0.916	0.913	0.911	0.910

Note:
Entries report the coverage rate of the IRFs using the proposed asymptotic distributions (nominal 95%).

Table 8: Coverage Probabilities

T	N	h	$k=2$	$k=3$	$k=4$	SVAR-IV ($k=4$)
250	125	0	0.945	0.926	0.919	1.962
		1	0.980	0.968	0.962	4.299
		2	0.994	0.987	0.982	2.476
		3	1.001	1.007	1.008	1.625
		6	1.010	1.021	1.021	1.157
		9	1.014	1.028	1.026	1.627
		12	1.025	1.038	1.039	2.116
	250	0	0.949	0.928	0.911	2.006
		1	0.975	0.959	0.943	4.236
		2	1.000	0.990	0.981	2.439
		3	0.998	0.986	0.978	1.520
		6	1.024	1.010	1.010	1.126
		9	1.026	1.021	1.017	1.570
		12	1.031	1.026	1.029	2.024
500	125	0	0.961	0.938	0.925	2.020
		1	0.974	0.963	0.953	5.802
		2	0.989	0.987	0.984	3.285
		3	1.000	1.000	0.996	1.927
		6	1.020	1.027	1.017	1.357
		9	1.019	1.021	1.013	2.110
		12	1.022	1.024	1.019	2.878
	250	0	0.961	0.948	0.936	2.063
		1	0.984	0.975	0.969	6.064
		2	0.989	0.988	0.984	3.195
		3	1.001	1.004	1.003	1.823
		6	1.020	1.022	1.020	1.379
		9	1.017	1.022	1.018	2.171
		12	1.016	1.021	1.020	2.955

Note:
Entries report the RMSE ratios of the estimated IRFs of the overidentified system to the just-identified system.

Table 9: RMSE ratios