Innovations meet Narratives*

–improving the power-credibility trade-off in macro–

Régis Barnichon[(a)] *and* *Geert Mesters*[(b)]

[(a)] Federal Reserve Bank of San Francisco and CEPR

[(b)] Universitat Pompeu Fabra, Barcelona School of Economics and CREI

January 16, 2025

Abstract

In macro, instances of clear and indisputable exogenous variation are rare and researchers often face a difficult trade-off between credibility and efficiency. In this work, we introduce a new method —innovation-powered IV—, which allows to reduce the confidence intervals of a credible but low power identification scheme (e.g., a narrative instrument) by leveraging the high power of a possibly mis-specified parametric identification assumption (e.g., a short run restriction). The method delivers large reductions in confidence intervals for the causal effects of monetary and fiscal policy, with gains of around 40 percent compared to state of the art narratively-identified estimates.

JEL classification: C14, C32, E32, E52.

Keywords: Narrative identification, time series identification, missing data, monetary and fiscal policy.

*We thank Richard Crump, Macro Del Negro, Kirill Evdokimov, Yuriy Gorodnichenko, Emi Nakamura, Michal Kolesár, Toru Kitagawa, Mikkel Plagborg-Møller, Ulrich Müller, Adam Shapiro, Sanjay Singh for insightful comments and seminar participants at BI Norwegian, Bocconi, Brown, UC Berkeley, Danmarks Nationalbank, EC2 Amsterdam, FRB New York, FRB San Francisco, NBER Summer Institute (Forecasting and Empirical Methods), NBER/NSF Time Series Conference, Norges Bank, SETA2024, Princeton and UCM3. Mesters acknowledges financial support from the Severo Ochoa Programme for Centres of Excellence in R&D (Barcelona School of Economics CEX2019-000915-S), funded by MCIN/AEI/10.13039/50110001103 and the 2021 ERC Starting Grant POLICYMETRICS. The views expressed in this paper are the sole responsibility of the authors and do not necessarily reflect the views of the Federal Reserve Bank of San Francisco or the Federal Reserve System.

1 Introduction

Causal inference is an important task for empirical macro, but it is also a particularly challenging one. In macro time series, instances of clear and indisputable exogenous variation are rare and researchers often face a difficult trade-off between credibility and efficiency. On the one hand, focusing on only the most convincing sources of exogenous variation will yield credible but also very imprecise causal estimates. On the other hand, including a large number of events will yield precise but little credible estimates.

This trade-off is most evident in the two main identification approaches that have dominated macro over the past 40 years: the narrative identification approach pioneered by Friedman and Schwartz (1963) and Romer and Romer (1989), and the parametric identification approach pioneered by Sims (1980).

On one side of the spectrum, the narrative approach consists in exploiting real time narrative records (e.g., newspapers, government reports, or policy meeting transcripts) in order to identify shocks —exogenous movements in a variable of interest— and construct a narrative instrument series—.[1] This approach is arguably the gold standard in terms of credibility and transparency, but it is constrained by an important limitation: instances of uncontroversial exogenous variation are rare.[2] This can lead to an efficiency problem, i.e., large and uninformative confidence bands and/or estimates that are overly sensitive to specification choices.

On the other side of the spectrum, the parametric identification approach estimates causal effects by imposing restrictions on parameters in time series models that are informed by economic theory, institutional characteristics or statistical features of the data.[3] Unlike the narrative approach, the parametric identification approach isolates shocks at each point in time leading to (ceteris paribus) more precise estimates. Unfortunately, this higher efficiency can come at the cost of lower credibility, as parametric identifying restrictions have not gone unchallenged (e.g., Nakamura and Steinsson, 2018).

In this work, we show how the combination of two large literatures —narrative identification and parametric identification— can substantially improve the power-credibility trade-off in macro: approaching the high efficiency of parametric identification while preserving the

[1]Intuitively, the idea is to exploit the rich information contained in the narrative accounts in order to overcome reverse causality and omitted variable biases. Popular narratively identified series include the oil shocks of Hamilton (1983, 1985), the monetary event series of Romer and Romer (1989), the government spending news shocks of Ramey (2011) and Ramey and Zubairy (2018), the tax shocks of Romer and Romer (2010); Cloyne (2013), and the financial shocks of Reinhart and Rogoff (2009), Jalil (2015) and Romer and Romer (2017), among many others.

[2]As we will see, the typical narrative instrument series has 80% of its values equal to zero (see Table 1).

[3]Prominent examples include structural VARs or local projections with short run, long run, heteroskedastic or non-Gaussian restrictions, or functional form restrictions in DSGE models (e.g., Herbst and Schorfheide, 2016; Ramey, 2016; Kilian and Lütkepohl, 2017).

high credibility of narrative identification. More generally, we propose a method that exploits endogenous but widely available IVs in order to increase the power of exogenous but sparse IVs.

Our starting point is to explicitly acknowledge the difficultly of identifying exogenous variation in macro time series. To that effect, we split the data into two (not necessarily consecutive) samples: a *good sample* in which a researcher possesses a valid instrument, for instance a set of dates where the narrative identification is deemed credible, and a *bad sample* where credible identification is missing. The key insight underlying our method is that it is possible to use the bad sample to improve the precision of causal estimates. Specifically, the good sample is used for the identification of the parameter of interest, while the bad sample is used to improve efficiency.

To make things clear, say we are interested in estimating a parameter θ capturing the effect of some policy variable p_t on some outcome variable y_t. A valid instrumental variable z_t is available but only over the good sample, while an *innovation* to p_t, call it v_t, is available over the full sample.[4] Our proposed method —Innovation-Powered IV (IPIV)— can be seen as proceeding in two steps. In a first step, we use the innovation v_t to construct an estimator for θ. This innovation-based, or "reduced-form", estimator will have low variance since it exploits the entire sample, but it will likely be biased since the innovation v_t is not a valid instrument —the parametric identification scheme is not fully credible—. Thus, in a second step we use the valid instrument z_t to estimate (and then correct) the bias of this reduced-form estimator. While this bias correction will have a cost in terms higher variance, we can guarantee asymptotic variance reductions relative to a conventional IV estimator based on z_t alone.

The efficiency gains from IPIV stem from two forces: (a) the use of a larger sample to construct reduced-form estimates, and (b) the use of innovations to best approach the efficiency of a valid instrument available over the *full* sample.

First, like our method, any conventional IV estimator can be seen as a two step process that (i) constructs an initial reduced-form estimate, and (ii) uses the IV to estimate and correct the bias of that reduced-form estimate (e.g. Hausman, 1978). Compared to regular IV however, IPIV uses a larger sample to construct a more precise reduced-form estimate, and this ultimately leads to a more precise θ estimate. The key underlying assumption for this to work is conditional selection ignorability (e.g. Wooldridge, 2007). Intuitively, as long as the credible instrument periods arrive at random conditional on controls, (appropriately re-weighted) covariances between macro variables will be the same on average in the good

[4]A typical example of an innovation to p_t would be the residual of a time series regression of p_t on some contemporaneous and lagged variables, as in Sims (1980)'s recursive ordering for instance, but any other parametric identification scheme can be used.

and in the bad samples, and the bad sample can be used to improve efficiency.[5] Importantly, conditional selection ignorability does allow for a non-random arrival of credible identification periods; for instance it may be easier to identify shocks depending on the state of the economy or on the size of the underlying shocks.

The second factor underlying IPIV's performance is the use of innovations: the closer v_t is to being a valid instrument, i.e., the better the parametric identification scheme, the larger are the efficiency gains relative to a conventional IV estimator based on z_t alone. Intuitively, the smaller the share of residual endogeneity in v_t, the smaller the bias correction term, and thus the larger the relative gains from improving the precision of the reduced-form estimator. In the limit where an innovation has no residual endogeneity left —the parametric identification scheme is valid and the bias is zero—, we obtain the efficiency of a conventional IV estimator when the IV is available over the *full* sample.

Innovation powering is particularly promising for macro applications, precisely because the two forces underlying the gains in efficiency are most at work in macro: (a) the good sample with a credible IV is typically small in macro, averaging only about 20 percent of the full sample size, and (b) a large macro-econometric literature has proposed perhaps not fully credible but at least "close-to-credible" identification schemes, ensuring that the innovations contain little residual endogeneity. Combining (a) and (b), for an innovation approaching a valid IV, IPIV will thus approach a theoretical limit of a 5-fold (1/0.2) reduction in variance compared to a conventional narrative IV.

From an econometrics perspective, we show that innovation powering can be cast in the generalized method of moment (GMM) framework of Hansen (1982), where the moment conditions include the conventional IV moment conditions based on the credible instrument z_t as well as the "innovation powered moment conditions" based on using the time series innovations v_t as instruments (after bias correcting). The model can be completed by including moments that determine the selection probabilities for having a credible instrument at a particular date, allowing the arrival of credible identification to depend on the state of the economy. Further, since narrative instruments are often weak in practice, our preferred inference approach follows the weak instrument literature and we develop an innovation powered version of the Anderson and Rubin (1949) statistic, see also Stock and Wright (2000) and Andrews and Guggenberger (2019).

We illustrate innovation powering by revisiting the seminal narrative works of Romer and Romer (1989, 2023) and Ramey and Zubairy (2018). Using innovations constructed from standard time series models à la Sims (1980), we find that IPIV delivers large reductions in confidence intervals: the weak IV robust innovation-powered Anderson-Rubin bands are

[5] The selection ignorability assumption can be tested, and we provide a specification test, analog to the covariate balance checks used in causal inference, to assess its validity (e.g. Imbens and Rubin, 2015).

40%-50% smaller in length when compared to the conventional AR bands. In fact, armed with the higher efficiency of IPIV we derive new results on the effects of monetary and fiscal policy.

First, we address a critique of the narrative approach going back to Hoover and Perez (1994): in finite sample, some of the narratively identified shocks can co-coincide (by chance) with other structural shocks. When there are relatively few narrative shocks, such finite sample confounding can substantially bias the estimated effect of interest. We show that finite sample confounding does indeed occur for both the Romer and Romer (2023) and the Ramey and Zubairy (2018) shock proxy series, but we also show that IPIV allows to draw informative inference even after removing contentious dates. In particular, we show that Romer and Romer (1989)'s results on the real effects of monetary policy are *robust* to the Hoover and Perez (1994) critique.

Second, we address external validity concerns regarding the use of historical data and major wars to infer the size of the spending multiplier (Gorodnichenko, 2014). The post-1952 sample excludes historical data and major wars and thus corresponds more closely to the multiplier of a contemporaneous peace-time US economy. With IPIV, we find that the Ramey and Zubairy (2018) shock proxy series over 1952-2015 *does contain* enough information to draw informative inference on the size of the government spending multiplier, with a 5-year cumulative multiplier between 0.6 and 1.7 at the 95 percent confidence level.

Our paper relates to two strands of literature. First, earlier work aimed at combining narrative identification with more efficient time series models is presented in Ludvigson, Ma and Ng (2017) and Antolín-Díaz and Rubio-Ramírez (2018) who suggest restricting the shocks identified in a structural VAR to conform with the narrative evidence. A prior robust Bayesian version of this idea, with frequentist guarantees, is developed in Giacomini, Kitagawa and Read (2023), see also Giacomini, Kitagawa and Read (2022) and Plagborg-Møller (2022) for more discussion. Unlike these papers, innovation powering (i) does not require correctly specifying an invertible SVAR model, (ii) allows the external instruments to be contaminated with measurement error and (iii) can account for non-random selection of the good sample, i.e., that the arrival of credible identification periods may depend, for instance, on the size of the shocks.

Second, innovation powering can be viewed as an alternative, and more efficient, way of handling missing instruments. While the existing narrative literature imputes zeros on dates where the narrative is inconclusive, innovation powering instead imputes innovation-based predictions. As such it is related to the missing data literature (e.g. Little and Rubin, 2020), and most closely on the augmented inverse probability weighted estimators developed in Robins, Rotnitzky and Zhao (1994, 1995) which employ a similar bias correction mechanism. More generally, such estimators are widely used for estimating treatment effects in the causal inference literature (e.g. Imbens and Rubin, 2015). We instead work in an aggregate macro time series setting where instruments are often weak, so that we combine ideas from the missing data literature with time series prediction routines and weak identification robust methods (e.g. Andrews, Stock and Sun, 2019). Further, in our setting the instrumental variable is missing and not the explanatory or outcome variable, see also Chaudhuri and Guilkey (2016); Abrevaya and Donald (2017).

The idea of performing a bias correction after imputing missing data (i.e. Robins, Rotnitzky and Zhao, 1994, 1995) was recently exploited by Angelopoulos et al. (2023); Angelopoulos and Zrnic (2024) for settings where imputations are done by modern machine learning methods: they dubbed their approach prediction powered inference. In our time series setting, we instead use innovations from macro time series models for imputing missing IVs and analogously label our method innovation powered inference.

Finally, missing values in time series analysis are often handled using state space methods (e.g. Durbin and Koopman, 2012). This approach is also possible when treating narrative instruments as missing: we can order the narrative instrument first in an SVAR model, cast the SVAR model in state space form and use the Kalman filter to predict the missing values based on past observations. However, a state space approach would require that (i) the model is correctly specified, (ii) a model for the measurement error is specified and (iii) the narrative events arrive at random. Our innovation powering method avoids these modeling choices and remains closer in assumptions and spirit to the conventional IV methodology in the narrative literature (e.g. Ramey and Zubairy, 2018).

The remainder of this paper is organized as follows. The next section presents a simple illustrative example that highlights the main ideas. Section 3 presents the general dynamic framework and Section 4 discusses the empirical results. Section 5 concludes.

2 Motivating example

In this section, we provide a simple static example to convey the main ideas and intuition behind innovation powering. Consider a simple IS/MP macroeconomic model

$$y_t = \theta p_t + \xi_t \tag{IS}$$

$$p_t = \phi y_t + \varepsilon_t \tag{MP}$$

where y_t is the output gap and p_t is the central bank policy rate, so that the first equation can be seen as an (IS) curve (in which case, $\theta < 0$) and the second equation is an (MP) curve —the central bank policy rule—. The error ε_t can then be seen as a monetary shock and ξ_t as a shock to the natural rate of interest. We observe the variables for a sample of periods collected in the set $\mathcal{N}$ with size $n = |\mathcal{N}|$.

We are interested in learning θ, the slope of the (IS) curve, or equivalently the effect of a unit change in the policy rate on the output gap. Because ξ_t and p_t are correlated through the policy rule, conventional least squares cannot be used for estimation. Instead, we will look for exogenous changes in p_t, i.e., monetary shocks ε_t, that can be used as an instrumental variable (IV) for p_t.

The narrative identification approach

A leading (non-parametric) approach for identifying exogenous changes in p_t is the narrative approach.[6] It consists in using narrative accounts (official records, newspaper articles, speeches, etc..) in order to isolate exogenous movements in p_t, i.e. movements in p_t that are independent of ξ_t. Formally, we can view the narrative identification approach as a function $f(\cdot)$ that constructs a shock proxy series z_t defined as

$$z_t = f(\varepsilon_t, \zeta_t) \ , \qquad t \in \mathcal{N} \ , \tag{1}$$

where $f(\varepsilon_t, \zeta_t)$ is a function of the shock of interest ε_t and some measurement error ζ_t, which is independent from the structural shocks.

In its current practice, the narrative approach $f(\cdot)$ constructs z_t as follows: (i) it constructs an estimate for ε_t on dates when the narrative account is sufficiently informative, (ii) it sets z_t to zero on all other dates, i.e., when the narrative account is inconclusive. These two steps ensure that z_t is a valid instrument for θ: $\mathbb{E}(z_t p_t) \neq 0$ and $\mathbb{E}(z_t(y_t - p_t\theta)) = 0$, and we can estimate θ in (IS) by using an IV estimator with instrument z_t.

The narrative approach has been a major step forward for identification in macro, but it is constrained by an important limitation: instances of clear and indisputable ε_t shocks are *rare*. This has a number of implications.

First, the narrative records are uninformative in most cases, and most values of z_t end up being coded as zeroes. In a survey of the recent literature exploiting narrative identification, the share of imputed zeroes is about 80 percent, sometimes even much higher, see Table 1.[7] While this does not invalidate z_t as an instrumental variable, this means that in practice

[6] While we focus on the narrative approach to construct the instrument, our innovation powering approach applies more broadly to other "non-parametric" identification methods; for instance the focus on large shocks, the use of discontinuities, or the use of high-frequency variation, see e.g. Nakamura and Steinsson (2018). The key feature is that these "non-parametric" identification approaches can typically only isolate a limited subset of exogenous events in a time series.

[7] Model (1) can generate the many zeros in different ways. Examples include: $z_t = \zeta_t \varepsilon_t$ where ζ_t is a binary measurement error series; this implies that the credible narrative dates arrive at random. Or, perhaps more realistically, a model of the form $z_t = (\varepsilon_t + \zeta_t)\,\mathbf{1}(|\varepsilon_t| \geq \tau)$ with ζ_t iid measurement error and τ some threshold. In that model, only large shocks get detected by the narrative approach, and the shocks are measured with classical measurement error. We do not take a stand on the specific model for z_t.

researchers exploit only a small share of the variation in p_t in order to infer θ. This can lead to an *efficiency* problem, i.e., large error bands.

Table 1: Zero Events in Narrative Studies

Paper	# obs	# zeros	% zeros
Bi and Zubairy (2023)	590	464	79%
Carriere-Swallow, David and Leigh (2021)†	28	22	81%
Cloyne (2013)	248	94	38%
Cloyne, Dimsdale and Postel-Vinay (2023)	89	69	77%
Fieldhouse and Mertens (2023)	292	228	78%
Guajardo, Leigh and Pescatori (2014)†	32	22	68%
Gil et al. (2019)	120	87	72%
Hamilton (1985)	140	121	86%
Jalil (2015)	90	83	92%
Ramey and Shapiro (1998)	200	197	98%
Ramey (2011)	280	194	69%
Ramey and Zubairy (2018)	504	396	79%
Romer and Romer (2023)	852	842	99%
Romer and Romer (1989)	852	845	99%
Romer and Romer (2010)	252	207	82%
Romer and Romer (2017)†	91	81	88%
Rojas, Vegh and Vuletin (2022)†	106	86	81%
Average			80%

Notes: Number and percentage of zeros across different narrative series. Here † indicates that the reported values are per unit averages from the panel data setting considered. We omit narrative accounts reported in books, e.g. Reinhart and Rogoff (2009) and Alesina, Favero and Giavazzi (2019), as they include several datasets. Inspection reveals that the fractions of zeros in such texts are not different from the ones reported.

Second, with few non-zero entries, the narrative instrument z_t may suffer from finite sample confounding, i.e., an unlucky finite sample correlation between some correctly identified subset of shocks ε_t and other structural shocks. For instance, Hoover and Perez (1994) argue that some of the monetary shocks of Romer and Romer (1989) coincide (by chance) with oil shocks.

Third, even with extensive documentation, researchers engaged in narrative identification must often use judgment calls, which has the benefit of hindsight. This could lead to subconscious endogeneity biases, see for instance the critique of the Friedman-Schwartz dates in Romer and Romer (1989) and the discussion in Romer and Romer (2023).

Together, low power and endogeneity biases can compound each other: a researcher willing to only entertain uncontroversial dates without possible confounders or judgment calls can be left with very few exogenous variations.

Good and Bad samples

To explicitly take into account the difficult nature of identification in macro time series, we split the data into two (not necessarily consecutive) samples: a *good sample* $\mathcal{G}$ in which a researcher possesses a valid instrument, for instance a set of dates where the narrative identification is deemed credible, and a *bad sample* $\mathcal{B}$ where credible identification is missing.[8]

In this simple illustrative example, we treat the selection of the good sample as random, that is the probability of credibly identifying an exogenous shock —a date in the set $\mathcal{G}$— does not depend on the state of the economy and

$$P(t \in \mathcal{G} | y_t, p_t, z_t) = \pi \tag{2}$$

where $\pi > 0$ is the selection probability of being in the good sample. The assumption of ignorable selection —i.e. a random arrival assumption— is commonly made in the narrative SVAR literature (Antolín-Díaz and Rubio-Ramírez, 2018; Giacomini, Kitagawa and Read, 2023) and is helpful to convey the intuition in this illustrative section.[9] Importantly however, our approach can accommodate more elaborate selection models, allowing π to depend on the state of the economy.[10] This extension is worked out in full detail in the main treatment below.

With a valid instrument z_t only available on $\mathcal{G}$, we can still construct a consistent IV estimator for θ. In fact, with the ignorable selection assumption this is simply the conventional IV estimator based on the periods $\mathcal{G}$ alone, i.e.

$$\hat{\theta}_{\text{iv}} = \frac{\widehat{\text{Cov}}_{\mathcal{G}}(y_t, z_t)}{\widehat{\text{Cov}}_{\mathcal{G}}(p_t, z_t)} , \tag{3}$$

where $\widehat{\text{Cov}}_{\mathcal{G}}(a, b)$ denotes the sample covariance between a and b computed over $\mathcal{G}$.

The low power of the narrative approach can be clearly seen from (3): with only a few credible exogenous dates, the sample $\mathcal{G}$ used to estimate the sample covariances is small, and so the variance of $\hat{\theta}_{\text{iv}}$ is high.

[8] We have that $\mathcal{G} \cup \mathcal{B} = \mathcal{N}$, $\mathcal{G} \cap \mathcal{B} = \emptyset$, $n_{\mathcal{G}} = |\mathcal{G}|$ and $n_{\mathcal{B}} = |\mathcal{B}|$.

[9] The wording "ignorable selection" is adopted from Wooldridge (2010), see the discussion in Section 19.8.

[10] For instance it is plausible that large shocks are easier to detect, or that monetary shocks may be easier to identify when inflation is at a high level (e.g., Romer and Romer, 1989). In each case π depends on p_t, y_t.

Powering IV: from $\mathcal{G}$ to $\mathcal{N}$

In this paper, we propose a method to improve the efficiency of the narrative approach, and more generally of non-parametric identification methods. Our starting point is a simple add-and-subtract operation that lets us rewrite the IV estimator as

$$\hat{\theta}_{\text{iv}} = \frac{\widehat{\text{Cov}}_{\mathcal{G}}(y_t, p_t) + \widehat{\text{Cov}}_{\mathcal{G}}(y_t, z_t - p_t)}{\widehat{\text{Cov}}_{\mathcal{G}}(p_t, z_t)} \,, \tag{4}$$

where we added and subtracted p_t in the numerator. In (4), note how the numerator of the IV estimator is the sum of two covariance terms: (i) the reduced-form covariance $\text{Cov}(y_t, p_t)$, which by itself leads to a biased estimate of θ, and (ii) the term $\text{Cov}(y_t, z_t - p_t)$, which can be seen as a bias correction term.

Expression (4) highlights the first key insight of this paper: under the ignorable selection assumption, we have $\widehat{\text{Cov}}_{\mathcal{N}}(y_t, p_t) - \widehat{\text{Cov}}_{\mathcal{G}}(y_t, p_t) \xrightarrow{p} 0$ as $n \to \infty$. This means that it is possible to use the full sample $\mathcal{N}$ instead of $\mathcal{G}$ to estimate the reduced-form covariance $\widehat{\text{Cov}}(y_t, p_t)$ and thereby reduce the variance of this term.[11] We will refer to this substitution as $\mathcal{N}$-powering and define

$$\hat{\theta}_{\text{iv},\mathcal{N}} = \frac{\widehat{\text{Cov}}_{\mathcal{N}}(y_t, p_t) + \widehat{\text{Cov}}_{\mathcal{G}}(y_t, z_t - p_t)}{\widehat{\text{Cov}}_{\mathcal{G}}(p_t, z_t)} \,. \tag{5}$$

While changing $\widehat{\text{Cov}}_{\mathcal{G}}(y_t, p_t)$ to $\widehat{\text{Cov}}_{\mathcal{N}}(y_t, p_t)$ can seem minor, this simple substitution can lead to large variance reductions when the sample size of $\mathcal{G}$ is small relative to the full sample size (i.e., $n_{\mathcal{G}}/n$ is small), which is precisely the case in macro. Indeed, by switching from $\mathcal{G}$ to $\mathcal{N}$, the reduction in the variance of $\widehat{\text{Cov}}(y_t, p_t)$ will be in the order of $n_{\mathcal{G}}/n$, and with $n_{\mathcal{G}}/n \approx 0.2$ in macro samples (Table 1), this represents an 80 percent (five-fold) reduction in variance.

The assumption of random selection of the $\mathcal{G}$ sample —i.e. ignorable selection (2)— is key for $\mathcal{N}$-powering. Intuitively, under ignorable selection, the sample $\mathcal{G}$ (i.e., the set of dates where the instrument is credible) is randomly drawn from the full sample $\mathcal{N}$, and covariances computed over the good sample $\mathcal{G}$ and over the full sample $\mathcal{N}$ will be asymptotically equal.

[11] Formally, let $s_t = 1$ if $t \in \mathcal{G}$ and $s_t = 0$ else. We have (assuming demeaned variables for simplicity)

$$\widehat{\text{Cov}}_{\mathcal{N}}(y_t, p_t) - \widehat{\text{Cov}}_{\mathcal{G}}(y_t, p_t) = \frac{1}{n}\sum_{t \in \mathcal{N}} y_t p_t - \frac{n}{n_{\mathcal{G}}}\frac{1}{n}\sum_{t \in \mathcal{N}} s_t y_t p_t \xrightarrow{p} \mathbb{E}(y_t p_t) - \mathbb{E}(s_t y_t p_t)/\pi$$

as $n \to \infty$ with $n_{\mathcal{G}}/n \to \pi$. Convergence follows from a standard law of large numbers provided that (y_t, p_t, s_t) satisfy regularity conditions. Subsequently the random arrival assumption implies that $P(s_t = 1|y_t, p_t, z_t) = \mathbb{E}(s_t|y_t, p_t, z_t) = \pi$ and we have $\mathbb{E}(y_t p_t) - \mathbb{E}(s_t y_t p_t)/\pi = \mathbb{E}(y_t p_t) - \mathbb{E}(\mathbb{E}(s_t|y_t, p_t, z_t) y_t p_t)/\pi = \mathbb{E}(y_t p_t) - \mathbb{E}(y_t p_t)\pi/\pi = 0$.

A good example of random shock detection would be a case where the timing of the identified shock is driven by factors independent of the current economic outlook, for instance from international geopolitical considerations (as in the case of war shocks). Ignorable selection will not hold however if large shocks are more likely to be detected. In that case, the covariances between macro variables will be different in the selected sample than in the full sample; simply because the $\mathcal{G}$ sample now features a bigger share of large shocks. In that case, the solution consists in re-weighting the observations to make the $\mathcal{G}$-sample as-good-as-randomly selected conditional on controls: this is the role of inverse probability weighting (e.g., Wooldridge, 2010, Chapter 19), and we will incorporate this generalization in the main treatment.

With all that said, note that $\mathcal{N}$-powering leaves the second term —the bias correction term— unaffected, so that the overall variance reduction in $\hat{\theta}_{\text{iv}}$ will depend on the relative contribution of the bias correction term to the variance of $\hat{\theta}_{\text{iv}}$. By scaling p_t appropriately, we can minimize the bias-correction term and thereby amplify the effect of $\mathcal{N}$-powering. For that reason, our $\mathcal{N}$-powered estimator will be of the form

$$\hat{\theta}_{\text{ip}} = \frac{\widehat{\text{Cov}}_{\mathcal{N}}(y_t, \gamma p_t) + \widehat{\text{Cov}}_{\mathcal{G}}(y_t, z_t - \gamma p_t)}{\widehat{\text{Cov}}_{\mathcal{G}}(p_t, z_t)}, \tag{6}$$

where γ is chosen to minimize the asymptotic variance of $\hat{\theta}_{\text{ip}}$. In fact, with an appropriate rescaling of p_t, we will see that we can guarantee that $\mathcal{N}$-powering delivers asymptotic efficiency gains.

Powering IV with innovations

To maximize the gains from $\mathcal{N}$-powering, there is a better choice than using p_t in the add-and-subtract rewriting of the IV estimator. Instead of rewriting the IV estimator with the variable p_t, one can use a variable v_t that is (i) available every period just like p_t (so as to allow for $\mathcal{N}$-powering), and (ii) has less residual endogeneity than p_t, i.e., has a higher correlation with the shock we want to identify, here the monetary shock. This is the second key insight of this paper: the use of time series innovations to improve the efficiency of valid (but sparse) instrumental variables.

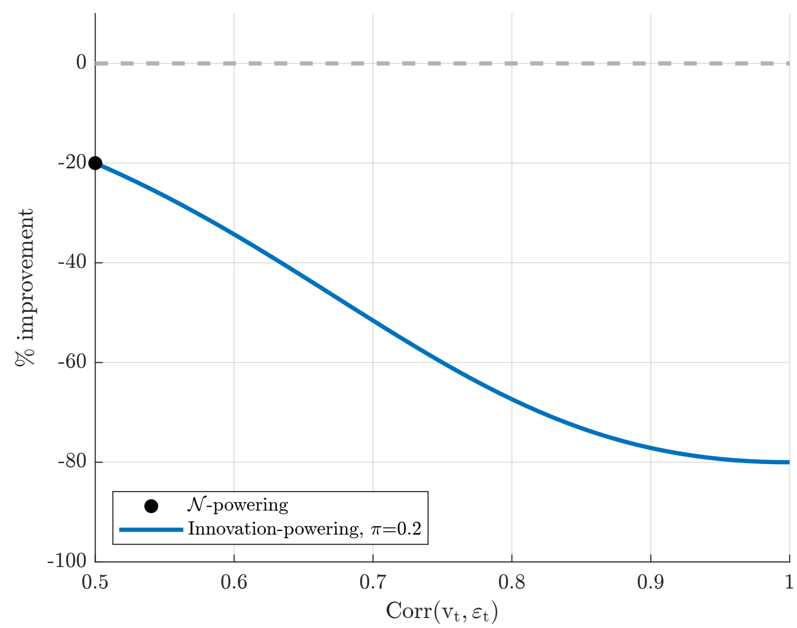
Our general Innovation-Powered IV (IPIV) estimator is then

$$\hat{\theta}_{\text{ip}} = \frac{\widehat{\text{Cov}}_{\mathcal{N}}(y_t, \gamma v_t) + \widehat{\text{Cov}}_{\mathcal{G}}(y_t, z_t - \gamma v_t)}{\widehat{\text{Cov}}_{\mathcal{G}}(p_t, z_t)} \ . \tag{7}$$

While v_t can a priori be any variable available over the whole $\mathcal{N}$ sample, a good candidate for v_t is a variable that has as little residual endogeneity as possible, i.e., that has the highest share of its variance coming from the shock of interest (ε_t). To see that, note that the variance cost of the bias correction step —the variance coming from the second term in (7)— depends on the difference $z_t - \gamma v_t$.[12] The larger the correlation between v_t and ε_t, the larger the correlation between v_t and z_t (since z_t proxies for ε_t), and thus the more effective is γ at lowering the variance of the bias correction term. When this is the case, the relative contribution of $\widehat{\mathrm{Cov}}_{\mathcal{G}}(y_t, z_t - \gamma v_t)$ to the overall variance of $\hat{\theta}_{\text{ip}}$ is smaller, and $\mathcal{N}$-powering —a reduction in the variance of $\widehat{\mathrm{Cov}}(y_t, \gamma v_t)$, the first term in (7)— will yield a larger reduction in the total variance of $\hat{\theta}_{\text{ip}}$.

For these reasons, a prime candidate for v_t is a time series innovation to p_t, i.e., the residual from a time series regression that soaked up most of (but not necessarily all of) the confounding factors, here ξ_t. Prominent candidates for such innovations are Sims (1980) or Romer and Romer (2004) in the context of monetary policy. More generally, there exists a large literature on parametric identification (e.g., Ramey, 2016) that can be used to construct suitable innovations.

Figure 1: Variance reduction from Innovation Powering



Notes: Asymptotic variance reduction of innovation powering relative to baseline IV as a function of $\rho = \mathrm{Corr}(v_t, \varepsilon_t)$ for $\pi = 0.2$. The percentage improvement is $\frac{1-\pi}{1+3(1-\rho)^2/\rho^2}$, when $\sigma_\varepsilon = \sigma_\xi$ and (ξ_t, ε_t) is normally distributed. The filled dot denotes the gains based on using p_t as innovation and taking $\phi = 1$ such that $\mathrm{Corr}(y_t, p_t) = 0.5$.

Figure 1 illustrates the additional variance reduction obtained by using v_t —innovations to p_t— instead of p_t, showing the percentage variance reduction in $\hat{\theta}_{\text{ip}}$ as a function of the

[12] This can be seen from the asymptotic variance of $\hat{\theta}_{\text{ip}}$: $\mathrm{Var}(\hat{\theta}_{\text{ip}}) = \frac{\lim_{n\to\infty} \mathbb{E}(\widehat{\mathrm{Cov}}_{\mathcal{N}}(\xi_t, \gamma p_t) + \widehat{\mathrm{Cov}}_{\mathcal{G}}(\xi_t, z_t - \gamma p_t))^2}{\mathrm{Cov}(p_t, z_t)^2}$.

correlation between the innovation v_t and the monetary shock ε_t. Starting from a baseline improvement of 20 percent when using p_t directly, we can see increasing variance reductions as we use innovations with lower residual endogeneity (i.e., higher correlation with ε_t). When the innovation has no residual endogeneity left —the parametric identification scheme is valid and $\mathrm{Corr}(v_t, \varepsilon_t) = 1$—, we obtain a 5-fold reduction in variance when $\pi = 0.2$. Intuitively, the smaller the amount of residual endogeneity, the closer is the innovation to being a valid IV, and thus the more innovation powering allows to approach the efficiency of an IV available over the full sample.

To summarize, the variance reduction from IPIV comes from three key features: (i) $\mathcal{N}$-powering, which leverages a mild (and testable) ignorable selection assumption to exploit a much larger sample size than permitted by credible macro instruments, (ii) a weighting parameter γ, which maximizes the efficiency gains from $\mathcal{N}$-powering, and (iii) the use of time series innovations —plausible (but possibly endogenous) candidates for structural shocks— to further amplify the gains from $\mathcal{N}$-powering. Quantitatively, the efficiency gains are especially promising for two reasons specific to macro: (a) credible identification is hard in the time series so that the good sample $\mathcal{G}$ is typically small relative to the full sample ($n_{\mathcal{G}}/n$ small) and (b) a large macro-econometric literature has devised plausible time series identification schemes to construct good candidate innovations, i.e., innovations with little residual endogeneity.

Last, we note that innovation powering only helps with the numerator of the IV estimator. The denominator in (7) is unaffected by innovation powering, which means that IV strength is not affected by innovation powering. $\widehat{\mathrm{Cov}}_{\mathcal{G}}(p_t, z_t)$ appears in the denominator of (7), which means that identification *strength* of θ is only determined by the narrative instrument — innovation powering reduces variance but does not solve the weak instrument problem—, and the main treatment will rely on weak instrument robust methods for conducting inference.

3 General framework

We now present our innovation powering methodology in the context of a general stationary macro environment where we are interested in estimating a dynamic causal effect using a narrative shock proxy as instrument, or more generally using a credible but sparse instrument. Relative to the illustrative example, we (i) formalize inference for a class of innovation powered IV estimators with general selection processes for the $\mathcal{G}$ sample, (ii) develop weak instrument robust confidence bands, and (iii) provide a specification test for checking the (conditional) ignorable selection assumption underlying $\mathcal{N}$-powering.

3.1 Model and assumptions

The object of interest is the dynamic causal effect $\theta_h \in \Theta \subset \mathbb{R}$ defined in the linear model

$$y_{t+h} = \theta_h p_t + \boldsymbol{\beta}_h' \mathbf{x}_t + \xi_{t+h} \ , \qquad \text{for } t \in \mathcal{N} \ , \tag{8}$$

where y_{t+h} is the h period ahead outcome variable, p_t is the explanatory variable of interest, with θ_h capturing its causal effect, and $\mathbf{x}_t$ denotes a vector of pre-determined control variables whose effect is captured by $\boldsymbol{\beta}_h$. Typically, $\mathbf{x}_t$ includes a constant and some lags of other observable variables.[13]

For exposition purposes we start by projecting out the control variables to obtain

$$y_{t+h}^{\perp} = \theta_h p_t^{\perp} + \xi_{t+h}^{\perp} \ , \qquad \text{for } t \in \mathcal{N} \ , \tag{9}$$

where $l_{t+j}^{\perp} = l_{t+j} - \text{Proj}(l_{t+j}|\mathbf{x}_t)$ for $l = y, p, \xi, z$ and $j = 0, 1, 2, \ldots$, with $\text{Proj}(a|b)$ denoting the linear projection of a on b. Throughout the paper, we will treat the orthogonalized variables as observed, noting that no changes occur when replacing the population projection by its empirical counterpart.

Narrative identification

Our identification approach is based on the existence of a narrative series z_t, which proxies for a structural shock of interest. In our simple example, the shock of interest was a monetary shock, but the shock could be an oil shock, a fiscal shock, a TFP shock, among many others (see e.g., Ramey, 2016).[14] More generally, we think of the narrative identification approach as constructing a proxy for exogenous changes in a variable of interest.

We impose the following exogeneity assumption

Assumption 1 (IV moment condition). $\mathbb{E}(z_t^{\perp}(y_{t+h}^{\perp} - \theta_h p_t^{\perp})) = 0$.

Assumption 1 imposes that $z_t^{\perp}$ —the narrative series after projecting out controls— is uncorrelated with the error term. This is the conventional starting point for existing narrative studies (e.g. Ramey and Zubairy, 2018).

Different from previous work, we will only use the instrument $z_t^{\perp}$ on a selection of "good" dates defined by the set $\mathcal{G} \subset \mathcal{N}$. The disjoint set $\mathcal{B}$ includes the bad dates where the

[13]Equation (8) can be derived from an underlying structural vector moving average model (e.g. Stock and Watson, 2018), but such additional structure is not necessary. In fact, in the supplementary material we show that several interesting underlying structural models satisfy (8) and the other assumptions that we impose in this section.

[14]We adopt Ramey (2016)'s definition of structural shocks: structural disturbances in a simultaneous equation system. Structural shocks are (i) exogenous with respect to the other current and lagged endogenous variables in the model; (ii) uncorrelated with other exogenous shocks; and (iii) represent either unanticipated movements in exogenous variables or news about future movements in exogenous variables.

instrument z_t is unknown or non-credible and thus treated as missing.[15] We denote by s_t an indicator variable capturing dates when the instrument is credible, i.e.,

$$s_t = \begin{cases} 1 & \text{if} \quad t \in \mathcal{G} = \{t \in \mathcal{N} : z_t^{\perp} \text{ credible}\} \\ 0 & \text{if} \quad t \in \mathcal{B} \end{cases} \quad . \tag{10}$$

Unlike in the simple example, we now allow the arrival of credible dates for z_t to depend on the state of the economy.[16] To formalize this let $\mathbf{q}_t$ contain the unique elements of $(y_{t+h}^{\perp}, p_t^{\perp}, \mathbf{v}_t)$, where $\mathbf{v}_t$ are the innovations that we formally introduce below.

Assumption 2 (Conditional ignorable selection)**.** *We have that $\mathbf{q}_t$ is observed on every $\mathcal{N}$ period and*

$$P(s_t = 1 | \mathbf{q}_t, z_t^{\perp}) = P(s_t = 1 | \mathbf{q}_t) \equiv \pi(\mathbf{q}_t) \ , \tag{11}$$

for all $\mathbf{q}_t$ and

$$\pi(\mathbf{q}_t) = \kappa(\mathbf{q}_t; \boldsymbol{\delta}) > 0 \ , \tag{12}$$

where κ is differentiable with respect to parameters $\boldsymbol{\delta} \in \Delta \subset \mathbb{R}^{d_\delta}$.

Assumption 2 is a (conditional) ignorable selection assumption: the probability of credibly identifying a shock —of selecting a date in the $\mathcal{G}$ sample— can depend on the state of the economy, but once we account for the relevant macro variables (i.e., $\mathbf{q}_t$), the instrument $z_t^{\perp}$ has no effect on the selection of $\mathcal{G}$. This assumption is conceptually similar as in other studies that rely on inverse probability weighting (Robins, Rotnitzky and Zhao, 1994; Hirano, Imbens and Ridder, 2003; Wooldridge, 2007).

To illustrate what this assumption rules out we note that a narrative instrument $z_t^{\perp}$ can be decomposed into two parts: (i) a component that depends on the shock of interest and (ii) a component that depends on additional measurement error (see equation (1)). Since $p_t^{\perp}$, which is included in $\mathbf{q}_t$, depends on the shock of interest (by construction), the only part of $z_t^{\perp}$ that is not accounted for by $\mathbf{q}_t$ is the measurement error. As such, by ruling out that the instrument $z_t^{\perp}$ determines the selection of $\mathcal{G}$, we are ruling out that the measurement error in the narrative instrument determines selection. This is a mild assumption, which allows, for example, selection to be determined by the size of the shock.

Assumption 2 will ensure that $\mathcal{N}$-powering is possible. Intuitively, if selection is ignorable (conditional on covariates), then a similar bias correction step as in section 2 is possible as (inverse-probability weighted) covariances computed on the $\mathcal{G}$ sample will be asymptotically

[15] Since our method will only exploit the instrument z_t in the good sample $\mathcal{G}$, whether or not the moment condition (1) holds on the bad sample $\mathcal{B}$ is irrelevant for practical purposes.

[16] We treat s_t as a random variable in the main text. In the appendix, we consider the case where s_t is deterministic. This happens, for instance, when shock proxies are only observed on a sub-period because of data availability, as in the case with high-frequency monetary surprises (Gertler and Karadi, 2015).

equal to covariances computed on the $\mathcal{N}$ sample.

The second part of Assumption 2 posits a parametric model (with parameters $\boldsymbol{\delta}$) for the selection probabilities. A common modeling choice would be to take $\kappa(\mathbf{q}_t;\boldsymbol{\delta})$ as a logit or probit model. In principle, non-parametric methods can also be adopted for estimating $\pi(\mathbf{q}_t)$. However, for most macro studies the sample size is prohibitively small and parametric models are typically adopted (e.g. Hamilton and Jordà, 2002; Angrist, Jordà and Kuersteiner, 2018).

Time series innovations

We now discuss the innovations $\mathbf{v}_t \in \mathbb{R}^{d_v}$ and their construction.

In general, we require that the innovations are available on all time periods and that they are good predictors for the structural shock of interest, but they do not need to be orthogonal to $\xi_{t+h}^{\perp}$. While constructing $\mathbf{v}_t$ is inherently an application specific task, there is a general route that follows naturally from a large macro literature on the parametric identification of structural shocks. While we will not take these identifying schemes as necessarily valid, they will serve as great candidates for innovations.

Indeed, to maximize the gains from $\mathcal{N}$-powering, the innovation $\mathbf{v}_t$ should be as close as possible to the structural shock of interest; the shock that $z_t^{\perp}$ is proxying for. Since structural shocks can be thought of as structural disturbances in a simultaneous equation system, innovations in structural time series model are natural candidates for $\mathbf{v}_t$. The time series models could be fully fledged DSGE models, structural VARs or local projections (among others). In this general treatment, we will not discuss all possible options but consider the general case where the innovations are the residuals from a linear projection, or

$$\mathbf{v}_t = (v_{1t}, \ldots, v_{d_v t})' \,, \qquad \text{with} \quad v_{it} \in \{\mathbf{w}_{t+h}^s - \mathrm{Proj}(\mathbf{w}_{t+h}^s | \mathbf{x}_t^s), h = 0, 1, \ldots, H\} \,, \tag{13}$$

where H is some positive integer and $\mathbf{x}_t^s$ includes a set of control variables. Specification (13) includes recursive identification, identification through controls (e.g., Romer and Romer, 2004), or long-run restrictions, see Plagborg-Møller and Wolf (2021).

In addition, note that (13) allows for the use of future variables to help better predict the shock of interest, as would be the case when shocks are recoverable from past *and* future data (e.g. Chahrour and Jurado, 2022). The collection in (13) comprises a large class of macro models that can be used to obtain innovations, but we stress that other methods for constructing innovations could easily be incorporated.

3.2 Innovation powered inference

We now discuss inference for θ_h using innovation powering and the estimation of the nuisance parameters $\boldsymbol{\delta}$, the parameters of the selection model. In contrast to the simple example we cast this task within the framework of the generalized method of moments (Hansen, 1982), allowing us to exploit general results from this literature for establishing formal asymptotic guarantees for our method. We first introduce the moment conditions that underlie our methodology, and then discuss the construction of IPIV estimates and weak instrument robust confidence bands.

Moment conditions underlying innovation powering

The moment conditions underlying the estimation of θ_h and $\boldsymbol{\delta}$ are summarized in the following lemma.

Lemma 1. *Given Assumptions 1,2 for any $\boldsymbol{\gamma} \in \mathbb{R}^{d_v}$ we have that*

$$\begin{aligned} \mathbb{E}g(\mathbf{d}_t, \theta_h, \boldsymbol{\delta}; \boldsymbol{\gamma}) &= 0 \ , \\ g(\mathbf{d}_t, \theta_h, \boldsymbol{\delta}; \boldsymbol{\gamma}) &= z_t^{\perp}(y_{t+h}^{\perp} - \theta_h p_t^{\perp}) s_t / \kappa(\mathbf{q}_t; \boldsymbol{\delta}) + \boldsymbol{\gamma}' \mathbf{v}_t (y_{t+h}^{\perp} - \theta_h p_t^{\perp})(1 - s_t/\kappa(\mathbf{q}_t; \boldsymbol{\delta})) \end{aligned} \tag{14}$$

$$\begin{aligned} \mathbb{E}m(\mathbf{q}_t, \boldsymbol{\delta}) &= 0 \ , \\ m(\mathbf{q}_t, \boldsymbol{\delta}) &= \boldsymbol{\kappa}^{(1)}(\mathbf{q}_t; \boldsymbol{\delta})(s_t - \kappa(\mathbf{q}_t; \boldsymbol{\delta}))/(\kappa(\mathbf{q}_t; \boldsymbol{\delta})(1 - \kappa(\mathbf{q}_t; \boldsymbol{\delta}))) \ , \end{aligned} \tag{15}$$

where $\boldsymbol{\kappa}^{(1)}(\mathbf{q}_t; \boldsymbol{\delta}) = \partial \kappa(\mathbf{q}_t; \boldsymbol{\delta})/\partial \boldsymbol{\delta}$.

The proof is provided in the appendix.

Lemma 1 contains two moment conditions. The first condition is the Innovation Powering (IP) moment condition (14). Its first term is similar to an IV moment condition based on the valid instrument $z_t^{\perp}$ except that $z_t^{\perp}$ is only credibly observed on periods where $s_t = 1$; hence the scaling by $s_t/\kappa(\mathbf{q}_t; \boldsymbol{\delta})$ which accounts for selection of $\mathcal{G}$. The second term is also similar to an IV moment condition except that the innovation $\mathbf{v}_t$ is used as instrument. Since the innovation is not a valid instrument, a naive IV moment condition based on $\mathbf{v}_t$ would yield a biased estimate. To correct for the bias, we add-and-subtract that same moment condition on the full sample $\mathcal{N}$ and on the good sample $\mathcal{G}$: this is the role of the term $(1 - s_t/\kappa(\mathbf{q}_t; \boldsymbol{\delta}))$. Under the selection assumption 2 these two moments are equal and will cancel out, leaving the IV moment condition for $z_t^{\perp}$ for identification.

The efficiency gains from IPIV can be readily seen from a moment combination perspective. Indeed, (14) can be viewed as a linear combination of two types of moment conditions

$$\mathbb{E}(z_t^{\perp}(y_{t+h}^{\perp} - \theta_h p_t^{\perp}) s_t/\kappa(\mathbf{q}_t; \boldsymbol{\delta})) = 0 \qquad \text{and} \qquad \mathbb{E}(\mathbf{v}_t (y_{t+h}^{\perp} - \theta_h p_t^{\perp})(1 - s_t/\kappa(\mathbf{q}_t; \boldsymbol{\delta}))) = 0 \ , \tag{16}$$

Compared to a conventional IV estimator, which only exploits the first moment condition involving $z_t^\perp$, IPIV also exploits the moment conditions involving the innovations $\mathbf{v}_t$: this is the innovation powering part. If $z_t^\perp$ is correlated with $\mathbf{v}_t$, then these additional moment conditions are non-redundant (in the sense of Breusch et al. (1999)) and incorporating them will reduce the asymptotic variance of the conventional IV estimator (Breusch et al., 1999).[17] The moments weighting vector $\boldsymbol{\gamma}$ allows to place more or less weight on these additional moment conditions in order to maximize the gains from $\mathcal{N}$-powering. We will come back to the optimal $\boldsymbol{\gamma}$ shortly.

The other moment condition (15) in Lemma 1 corresponds to the first order conditions for the sample selection model, a binary response model (e.g. Wooldridge, 2010, Chapter 15). Depending on the choice for κ these can correspond to the first order conditions of a logit model, or other binary response models. As such, estimation and inference for the selection model of Assumption 2 is standard.

Innovation powered IV estimators

For the parameters of the selection model we define the estimate $\widehat{\boldsymbol{\delta}}$ as the solution to the sample moments

$$\frac{1}{n}\sum_{t\in\mathcal{N}} m(\mathbf{d}_t^\pi, \widehat{\boldsymbol{\delta}}) = 0 \ . \tag{17}$$

Using $\widehat{\boldsymbol{\delta}}$, we can estimate θ_h by solving[18]

$$\frac{1}{n}\sum_{t\in\mathcal{N}} g(\mathbf{d}_t, \widehat{\theta}_h(\boldsymbol{\gamma}), \widehat{\boldsymbol{\delta}}; \boldsymbol{\gamma}) = 0 \ ,$$

which has an analytical solution

$$\hat{\theta}_h(\boldsymbol{\gamma}) = \frac{\sum_{t\in\mathcal{N}} z_t^\perp y_{t+h}^\perp s_t/\kappa(\mathbf{d}_t^\pi; \widehat{\boldsymbol{\delta}}) + \boldsymbol{\gamma}'\mathbf{v}_t y_{t+h}^\perp(1 - s_t/\kappa(\mathbf{d}_t^\pi; \widehat{\boldsymbol{\delta}}))}{\sum_{t\in\mathcal{N}} z_t^\perp p_t^\perp s_t/\kappa(\mathbf{d}_t^\pi; \widehat{\boldsymbol{\delta}}) + \boldsymbol{\gamma}'\mathbf{v}_t p_t^\perp(1 - s_t/\kappa(\mathbf{d}_t^\pi; \widehat{\boldsymbol{\delta}}))} \ . \tag{18}$$

The *innovation powered instrumental variable* (IPIV) estimator is defined for any vector $\boldsymbol{\gamma} \in \mathbb{R}^{d_v}$. The intuition is the same as in the simple example. With $\boldsymbol{\gamma} = \mathbf{0}$ we recover the baseline IV estimate, while a nonzero $\boldsymbol{\gamma}$ puts weight on the innovation-based moments in

[17] It is worth noting that we could also treat $\mathbb{E}(z_t^\perp(y_{t+h}^\perp - \theta_h p_t^\perp)s_t/\pi) = 0$ and $\mathbb{E}(\mathbf{v}_t(y_{t+h}^\perp - \theta_h p_t^\perp)(1 - s_t/\pi)) = 0$ as $d_v + 1$ separate moment conditions within a conventional generalized method of moments framework. However, since the gradient of the innovation based moment conditions with respect to θ_h is always zero there is no further advantage of separating the moments over choosing $\boldsymbol{\gamma}$ optimally.

[18] The $\widehat{\boldsymbol{\delta}}$ parameters can be of independent interest as they allow to asses the plausibility of the random arrival assumption that underlies the usage of narrative evidence in SVAR models (e.g. Antolín-Díaz and Rubio-Ramírez, 2018). Indeed, positive predictive ability for any of the slope parameters of the binary response model implies that the random arrival assumption is violated.

order reduce the variance of $\hat{\theta}_h$. We derive the optimal $\boldsymbol{\gamma}^*$ by minimizing the asymptotic variance of $\hat{\theta}_h(\boldsymbol{\gamma})$ and maximizing the power of the weak identification robust tests defined below. The details for this derivation are reported in the appendix where we define $\widehat{\boldsymbol{\gamma}}(\hat{\theta}_h)$ (see equation (28)) as the estimate for the efficient weights $\boldsymbol{\gamma}^*$.

With the efficient weights in hand we can obtain the two-step efficient IPIV estimate using (18), i.e. $\hat{\theta}_h^{\text{2step}} \equiv \hat{\theta}_h(\widehat{\boldsymbol{\gamma}}(\hat{\theta}_h))$, where $\hat{\theta}_h$ is any initial IPIV estimate. Also, similar as for any GMM estimator we can iterate between updating the estimate for the causal effect, i.e. $\hat{\theta}_h(\widehat{\boldsymbol{\gamma}})$, and the estimate for the weights, i.e. $\widehat{\boldsymbol{\gamma}}(\hat{\theta}_h)$, until convergence leading to an iterated IPIV estimator, i.e. $\hat{\theta}_h^{\text{iter}}$, see Hansen and Lee (2021) for a recent analysis. Alternatively, we can parameterize the weights in terms of θ_h and consider the continuous updating estimator:

$$\frac{1}{n}\sum_{t\in\mathcal{N}} g(\mathbf{d}_t, \hat{\theta}_h^{\text{cue}}, \widehat{\boldsymbol{\delta}}; \widehat{\boldsymbol{\gamma}}(\hat{\theta}_h^{\text{cue}})) = 0 \ , \tag{19}$$

see Hansen, Heaton and Yaron (1996). Note that by parametrizing the weights the sample moment condition becomes nonlinear and numerical methods need to used to find $\hat{\theta}_h^{\text{cue}}$. The two-step, iterated and continuous updating estimators all have the same asymptotic properties, but in finite sample the continuous updating estimator performs most reliably and we recommend using this estimator.

Innovation powered weak identification robust confidence bands

Since narrative shock proxies are often weak instruments, we will construct weak instrument robust confidence bands by inverting tests for

$$H_0 : \theta_h = \bar{\theta}_h \qquad \text{against} \qquad H_1 : \theta_h \neq \bar{\theta}_h \ , \tag{20}$$

for some constant $\bar{\theta}_h$. Our preferred test statistic is an innovation powered version of the Anderson and Rubin (1949) statistic which was further developed for GMM by Stock and Wright (2000), see also Andrews and Guggenberger (2019).

The *Innovation Powered Anderson-Rubin* (IPAR) test statistic is given by

$$\text{AR}_{\text{IP}}(\bar{\theta}_h) = \frac{1}{\sqrt{n}}\sum_{t\in\mathcal{N}} g(\mathbf{d}_t, \bar{\theta}_h; \widehat{\boldsymbol{\gamma}}(\bar{\theta}_h)) / \sqrt{\widehat{\omega}(\bar{\theta}_h, \widehat{\boldsymbol{\gamma}}(\bar{\theta}_h))} \ , \tag{21}$$

where $\widehat{\boldsymbol{\gamma}}(\bar{\theta}_h)$ is the efficient weight evaluated at $\bar{\theta}_h$ and $\widehat{\omega}(\bar{\theta}_h, \widehat{\boldsymbol{\gamma}}(\bar{\theta}_h))$ is the estimate for the long run variance of $\frac{1}{\sqrt{n}}\sum_{t\in\mathcal{N}} g(\mathbf{d}_t, \bar{\theta}_h; \widehat{\boldsymbol{\gamma}}(\bar{\theta}_h))$. The exact expression is provided in equation (29) in the appendix.

The IPAR test statistic (21) can be interpreted as the rescaled innovation powered sample moments — evaluated under the null at $\bar{\theta}_h$ — standardized by the standard deviation estimate. The benefit of this test statistic is that its limiting distribution is the same regardless of the strength of the narrative instruments, while remaining able to distinguish between H_0 and H_1. In other words, it can inform which parameters θ_h are plausible. Formally we have the following result.

Theorem 1. *Given Assumptions 1-2 and Assumption R (stated in the appendix) we have that under H_0*

$$\lim_{n\to\infty} P(|\mathrm{AR}_{\mathrm{IP}}(\bar{\theta}_h)| > z_{\alpha/2}) = \alpha$$

where z_α denotes the 1-α quantile of the standard normal distribution.

The theorem shows that the null rejection probability is equal to the nominal level of the test α when we reject the null using critical values from the standard normal distribution. This result can then be used to define the following confidence set for θ_h.

$$\mathrm{CS}_h^{\mathrm{IP}} = \{\theta_h \in \Theta : \mathrm{AR}_{\mathrm{IP}}(\bar{\theta}_h) \leq z_{\alpha/2}\} \ . \tag{22}$$

In practice, we compute $\mathrm{CS}_h^{\mathrm{IP}}$ by searching over the parameter space Θ and collecting all values for which the IPAR test statistic takes values below $z_{\alpha/2}$.

Checking bias cancellation

A necessary condition for the consistency of IPIV is that the innovation powering part of the moment condition (14) is equal to zero, i.e. that the bias from using innovations as instruments drops out with the add-and-subtract operation. Formally, Assumption 2 imposes that for any $\tilde{\theta}_h \in \Theta$ we must have

$$\mathbb{E}(\mathbf{v}_t(y_{t+h}^{\perp} - \theta_h p_t)(1 - s_t/\kappa(\mathbf{q}_t; \boldsymbol{\delta}))) = 0 \ .$$

A sufficient condition is that the correlations between the innovations and the macro variables y and p are the same whether we use the sample $\mathcal{G}$ or the sample $\mathcal{B}$ after correcting for selection. To evaluate this assumption we consider verifying the moment conditions

$$\mathbb{E}(\mathbf{v}_t k_t(1 - s_t/\kappa(\mathbf{q}_t; \boldsymbol{\delta}))) = 0 \qquad \text{with} \quad k_t \in \{y_{t+h}^{\perp}, p_t^{\perp}\} \ . \tag{23}$$

We can test these moment conditions by adding them to moment condition (15) in Lemma 1 and computing Hansen (1982) J test for over-identification. If we reject the null this implies that there is a structural difference in the innovation-macro correlations between $\mathcal{G}$ and $\mathcal{B}$ after correction for selection and innovation powering most likely does not provide consistent estimates. Similar balance checks are often conducted when using matching for causal inference (e.g. Imbens and Rubin, 2015).

3.3 Summary of simulation study

We summarize the results from a simulation study that we conducted to evaluate the gains in efficiency from innovation powering. A detailed account of the study is given in the supplementary material. We simulated data from the static model of section 2 and various SVAR models calibrated to the empirical studies considered below. For each we considered multiple specifications: varying selection probabilities, instrument strengths and innovation strengths. The following three findings are most important.

1. For the vast majority of designs the length of the innovation powered AR confidence set is smaller when compared to the standard AR confidence set. Pending on the strength of the correlation between the narrative instrument and the innovations the relative improvement in confidence set length can be as large as 80%. Notably, this is observed for realistically calibrated models such as the empirical SVAR designs where prediction errors are used as innovations.

2. The size of the innovation powered AR test remains good as long as there are sufficiently many informative narrative periods. A concrete recommendation is to use innovation powering only when $n_{\mathcal{G}} \geq 50$, as the effective convergence rate of our method remains $\sqrt{n_{\mathcal{G}}}$ which needs to be sufficiently large for the asymptotic approximation to be accurate.

3. The IPIV point estimates are also generally more accurate when compared to their conventional IV counterparts. The differences in mean squared errors are often large mimicking the reductions in the IPAR confidence bands.

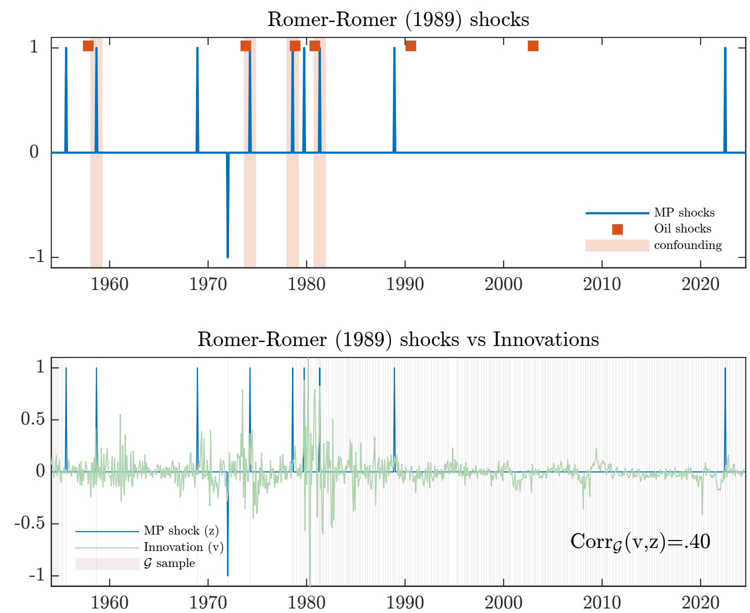
4 Empirical studies

In this section we revisit two landmark narrative studies: Romer and Romer (1989, 2023) on the effect of monetary policy and Ramey and Zubairy (2018) on the size of the government spending multiplier. In doing so, we will highlight the two limitations of the narrative approach —low power and finite sample confounding—, and show that innovation powering can substantially boost power, allowing us to draw robust results on the effects of monetary and fiscal policy.

4.1 The effects of monetary policy

In this first application, we revisit the seminal work of Romer and Romer (1989, 2023), henceforth RR, on the effects of monetary policy. To identify the effect of monetary policy actions Romer and Romer (1989) used the "Record of Policy Actions of the Federal Open Market Committee" (now commonly called "Minutes") to identify dates when the FOMC changed its policy rate for reasons unrelated to the state of the business cycle. Specifically, RR identify dates when the economy was roughly at potential, but the FOMC chose to adjust money growth or interest rates in order to affect the long-run level of inflation. The study was revisited and updated in Romer and Romer (2023) using the more detailed "Transcripts of FOMC meetings", resulting in a series z_t with eleven monetary events —ten ones and one minus one— over 1947-2023. The top panel of Figure 2 plots the narrative monetary dates.

Figure 2: NARRATIVE SHOCKS AND TIME SERIES INNOVATIONS, MONETARY POLICY



Top panel: the blue line depicts the monetary shocks (MP shocks, z_t) identified by Romer and Romer (1989, 2023). The orange squares mark dates with oil shocks (Oil shocks) identified by Hamilton (1983, 2003). The shaded bars ("confounding") depict times when a Romer-Romer monetary shock is within two quarters of an oil shock. *Bottom panel*: the blue line depicts the same MP shocks, the green line depicts the innovation (v_t) to the fed funds rate obtained similarly to Sims (1980), and the grey bars mark the $\mathcal{G}$ sample.

Following Stock and Watson (2018), we will use z_t as an instrumental variable for the fed funds rate. Treating z_t as an IV allows to quantify the effects of monetary policy in units of changes in the policy rate, i.e., to measure the effect of a unit change in the policy rate on economic variables —a key object of interest for policy makers (e.g., Barnichon and Mesters, 2023)—.[19]

[19]Using z_t as an IV also allows for measurement error in z_t, which is attractive when the narrative series is a binary variable or an index —the narrative records being not precise enough to measure exactly the size of the monetary intervention. That said, if the objective is solely to test the classical dichotomy hypothesis —whether money can have real effects—, it is sufficient to use z_t as a right-hand side variable in an OLS

We estimate the local projection (8), that is

$$y_{t+h} = \theta_h p_t + \boldsymbol{\beta}_h' \mathbf{x}_t + \xi_{t+h} \ , \qquad \text{for } t \in \mathcal{N} \ ,$$

with CPI inflation or unemployment as the dependent variable y_{t+h}, p_t is the fed funds rate and the controls $\mathbf{x}_t$ include a constant and twelve lags of p_t and y_t. Our monthly data sample spans 1954M7-2019M12 as we exclude the COVID period. Since the fed funds rate is only available after 1954M7, we do not include Romer and Romer's October 1947 monetary event.[20]

Low power and confounding

Figure 3 plots the estimated effects of a unit increase in the fed funds rate on unemployment and inflation, along with the 95 percent confidence bands, first assuming that the narrative monetary shocks are strong IVs, and then constructing weak-IV robust AR confidence bands following Andrews, Stock and Sun (2019). While the strong-IV bands can give the impression of relatively tight error-bands, the IV is in fact weak and the weak-IV robust bands are much larger,[21] meaning that the effects of monetary policy are not estimated very precisely. Intuitively, this is a result of the low power of the narrative approach: the approach uses here only 9 exogenous events to infer the effects of monetary policy, i.e., a small share of the total variation in the policy rate.

In addition, the low power can be compounded by finite sample confounding. As argued by Hoover and Perez (1994): four RR monetary events coincide (by chance) with oil shocks (Hamilton, 1983, 2003), see Figure 2 (orange squares), making it difficult to establish whether the estimated effect of monetary policy is in fact capturing the effect of oil shocks. Ideally, one would discard these controversial monetary events (highlighted by shaded bands in Figure 2), but this worsens the low power problem, so that the error bands become completely uninformative; with lengths excluding 20 percentage points of inflation or unemployment, see Figure 3 (bottom row). This application illustrates the difficult trade-off between credibility and efficiency in a time series context.

Innovation powering

Innovation powering can improve the power–credibility trade-off, and we will illustrate this point by using time series monetary innovations from Sims (1980)'s identification scheme.

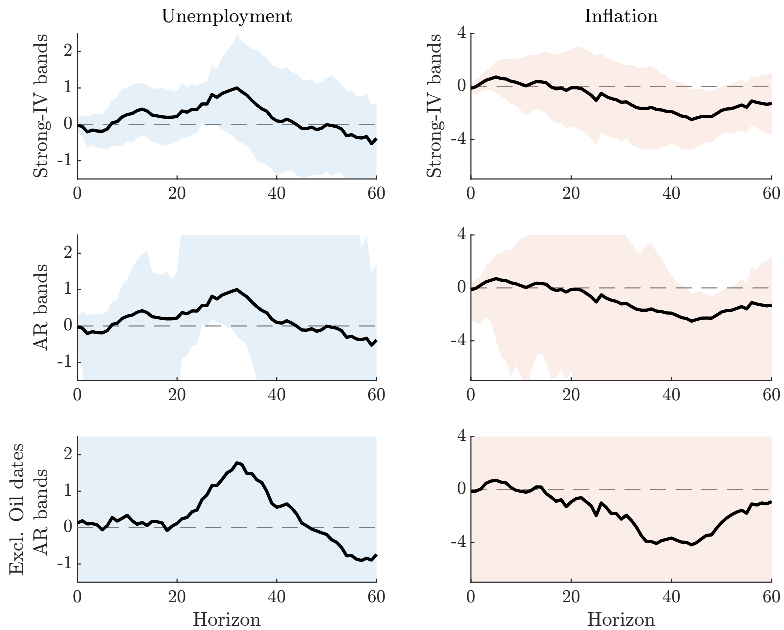
---

regression. This was the route followed by Romer and Romer (1989).

[20]Using instead the 3-months treasury bill with a sample 1945-2019 as in Romer and Romer (2023) gives similar results.

[21]The effective F-statistic of Olea and Pflueger (2013) is only 6.23, a case of weak instruments.

Figure 3: Effects of Monetary Policy, standard IV



Notes: The panels show the regular IV point estimates for unemployment and inflation along with 95% confidence bands computed under: a conventional normal approximation valid under a strong instrument assumption (top panel), the AR confidence bands (middle panel) and excluding oil dates (bottom panel).

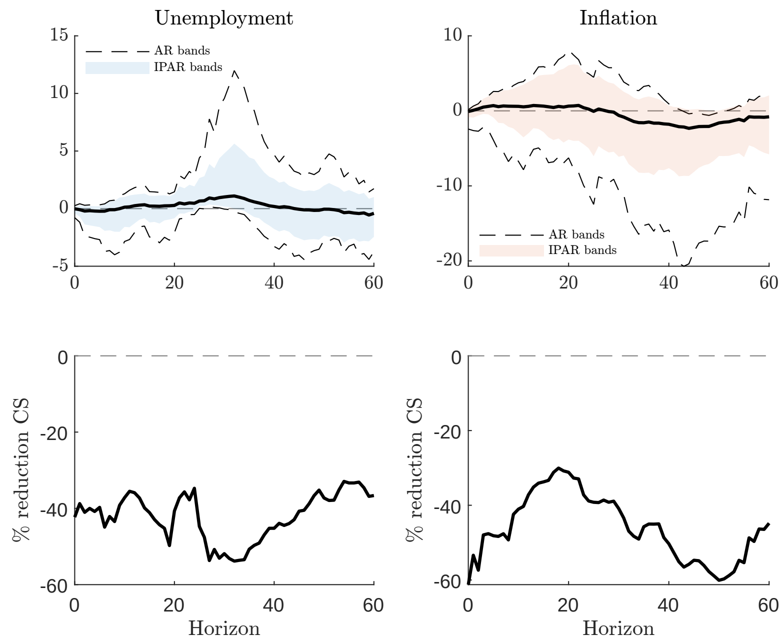
Specifically, implementing the IPIV method requires two inputs: (i) deciding which observations in the narrative series z_t are useful for identification, i.e., deciding on the samples $\mathcal{G}$ and $\mathcal{B}$ and on the accompanying selection model, and (ii) finding a good candidate for the innovations v_t.

First, we set the "good" sample $\mathcal{G}$ to include all RR monetary events ($+1$ or -1) as well as months when there was no FOMC meeting (scheduled or unscheduled) to set z_t to 0. All other observations are placed in $\mathcal{B}$. This gives us $n_{\mathcal{G}} = 156$ good periods and $n_{\mathcal{B}} = 593$, which implies that 79% of values of z_t are effectively treated as missing. In the baseline specification, we treat the entries of $\mathcal{G}$ as arriving as random, i.e., $\pi(\mathbf{q}_t) = \pi$. Robustness checks to alternative selection models are presented in the appendix.

Second, as innovations we use the standard "recursive identification scheme" of Sims (1980), i.e., we use the residual of a regression of the policy rate p_t on contemporaneous and lags of inflation and unemployment, as well as lags of the policy rate. Importantly, recall that our approach does not require this "recursive identification scheme" to be correct; it only needs to correlate with monetary shocks. The correlation between innovations and RR shocks on the good sample $\mathcal{G}$ is 0.40. Finally, we do not reject the null of bias cancellation between the monetary innovation (v_t) and the fed funds rate p_t, unemployment or inflation

(y_{t+h}) with p-values above 0.3 for the specification test based on (23).[22]

Figure 4: Effects of Monetary Policy, Innovation-Powered IV



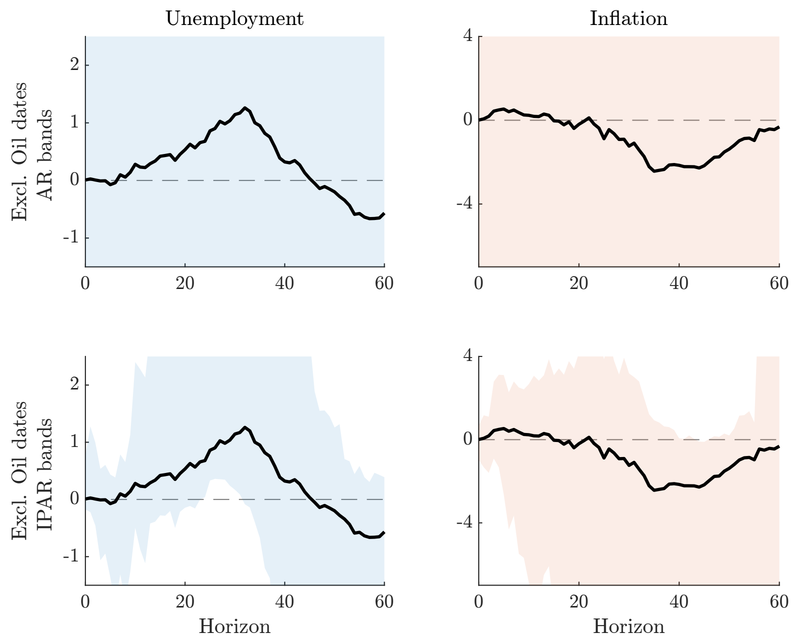
Notes: The top row shows the IPIV point estimate ($\hat{\theta}_h^{\text{cue}}$, thick black line) for unemployment and inflation, together with the innovation-powered AR bands (IPAR, CS_h^{IP}) at 95% confidence levels. The bottom row reports the difference in the lengths of the 95% bands for the regular IV estimator ("AR bands") and the Innovation-Powered IV estimator ("IPAR bands" CS_h^{IP} based on the $\text{AR}_{\text{IP}}(\bar{\theta}_h)$ statistic)

Figure 4 shows that IPIV delivers large reductions in confidence bands. The top row shows our innovation-powered IV estimates with 95 percent Innovation-Powered AR (IPAR) confidence bands along with the AR bands for conventional IV (dashed lines); the left panels shows the results for inflation and the right panels shows the results for unemployment. The bottom row, plots the percentage differences in the length of the confidence sets. IPIV reduces the length of the confidence sets by about 40-50 percent.

Armed with these large power gains, we can reconsider the Hoover-Perez critique and estimate the effects of monetary policy *without* the controversial oil/monetary dates, i.e., treating the four contentious dates as missing. Figure 5 plots the results. Innovation-

[22] In the appendix, we consider an alternative candidate for the innovations v_t: the Romer and Romer (2004) monetary shocks, henceforth RR04. RR04 estimate a time series model of the fed funds rate (a policy rule), where they use internal forecast from the Fed Board staff to control for the Fed's information set, a more credible identification strategy than Sims (1980)'s original approach. As a side comment, we note that the comparison between RR89 and RR04 is a great illustration of the power-credibility trade-off in time series macro. On the one hand, RR89 is a very credible approach to isolate exogenous changes in the policy rate, but the method can only isolate a few episodes: credibility is very high but power is lower. On the other hand, RR04 identify monetary innovations at every single dates but some of these innovations may not be entirely exogenous: power is high but credibility is lower. See the Appendix for more discussion.

Figure 5: Effects of Monetary Policy, Innovation-powered IV excl. Oil dates



Notes: Impulse responses of unemployment and inflation to a unit increase in the fed funds rate. The shaded area denote the 95 percent confidence sets without innovation power (AR bands, top panel) and with innovation powering (IPAR bands, bottom panel).

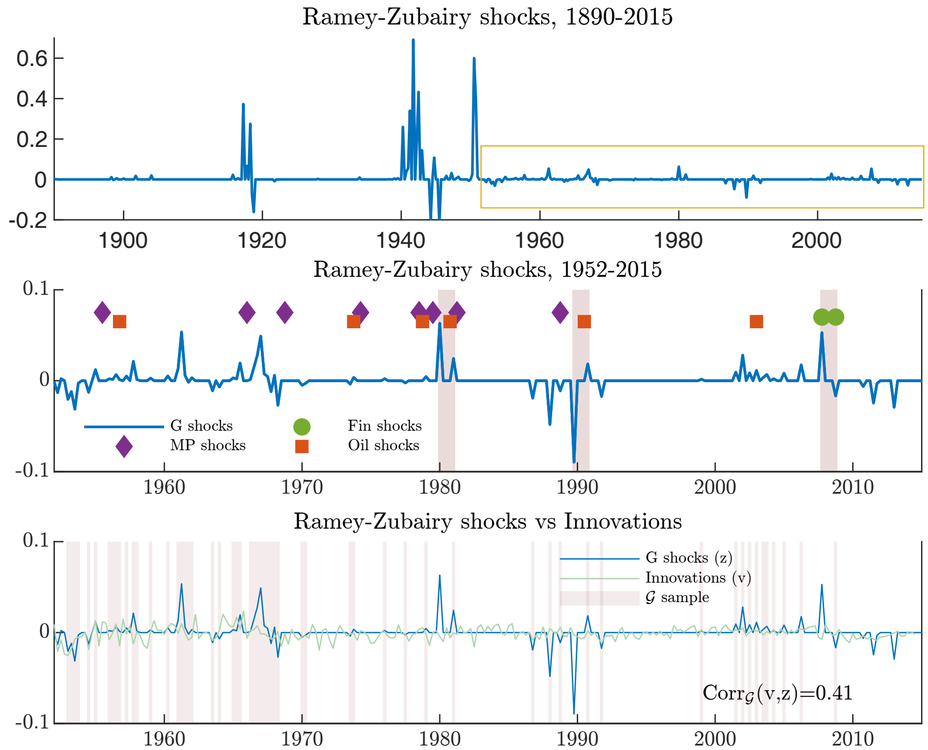
powering leads to large reductions in the length of the confidence sets, and we are able to obtain significant point estimates for the effects of monetary policy after 20 to 30 quarters, making the Romer and Romer (2023) evidence *robust* to the Hoover-Perez critique.

4.2 The government spending multiplier

In the second application, we revisit the landmark work of Ramey and Zubairy (2018), who use information from periodicals to narratively identify news shocks to defense expenditures over 1890-2015.

This study is a major step forward in the measurement of the government spending multiplier, but one recurrent question has been that of external validity; to what extent can we consider that the multiplier estimates are representative of the multiplier of a contemporaneous peace-time US economy? First, the use of long historical sample can mix different policy regimes, economies with very different underlying structure (e.g., from an industry-dominated economy to service-dominated one), as well as secular changes in the size and composition of government spending, see the discussion of Gorodnichenko (2014). Second, in the specific case of the Ramey-Zubairy (RZ) shocks, a few large war events —World War I, World War II, Korean war— are dominating the sample (see Figure 6), and these shocks could tilt estimates towards the multiplier of a war-time economy.

Figure 6: NARRATIVE SHOCKS AND TIME SERIES INNOVATIONS, FISCAL POLICY



Notes: Top panel: Ramey-Zubairy shocks over 1890-2015, with the orange rectangle depicting the 1952-2015 sample. Middle panel: Ramey-Zubairy ("G") shocks over 1952-2015. The markers (resp. diamond, circle, square) depicts other major shocks: monetary ("MP"), financial ("Fin") and oil ("Oil"). The red bars depict the large Ramey-Zubairy shocks that are within two quarters of another shock. Bottom panel: the blue line depicts the same G shocks, the green line depicts the innovation (v_t) to spending, and the grey bars mark the $\mathcal{G}$ sample.

Lower power and confounding

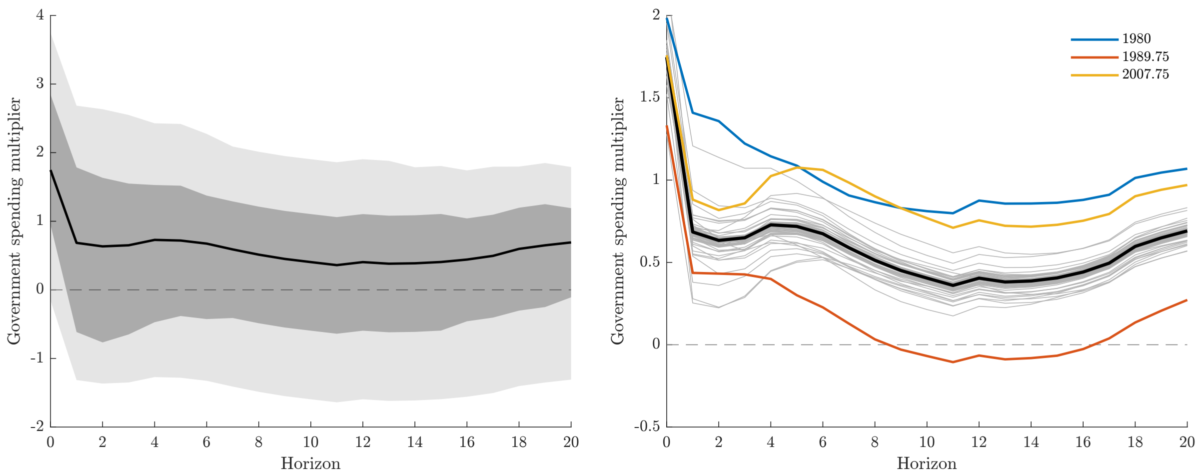
To address these issues of external validity, one would ideally estimate the multiplier from post-1952 data alone, as the post-1952 sample excludes the major war episodes and as such may be more representative of a peace-time modern economy. Figure 7a presents the result from such an exercise, running the same regression as Ramey and Zubairy (2018), i.e., local projections (8) of cumulative spending on cumulative gdp growth:

$$\underbrace{\sum_{j=0}^{h} \nu_{t+j}}_{=y_{t+h}} = \theta_h \underbrace{\sum_{j=0}^{h} g_{t+j}}_{p_t} + \boldsymbol{\beta}_h' \mathbf{x}_t + \xi_{t+h}$$

using the RZ shock proxy as instrument for p_t.

Unfortunately, the shorter 1952-2015 sample leaves the multiplier estimates very imprecise with the 95 percent confidence bands ranging from -1.5 to 2 at a 5 year horizon (Figure 7a).

Figure 7: Government spending multiplier: IV estimates



(a) Multiplier uncertainty (b) Multiplier sensitivity

Notes: Government spending multiplier IV estimates based on RZ shock proxy over 1952-2015. *Left panel:* point estimate (thick black line) with the 67 and 95 percent AR confidence sets (shaded bands). *Right panel:* Each grey line reports the estimated multiplier after setting one (and only one) of the non-zero RZ shocks to zero, and the colored lines depict three of these estimates with the corresponding date of the zeroed shock.

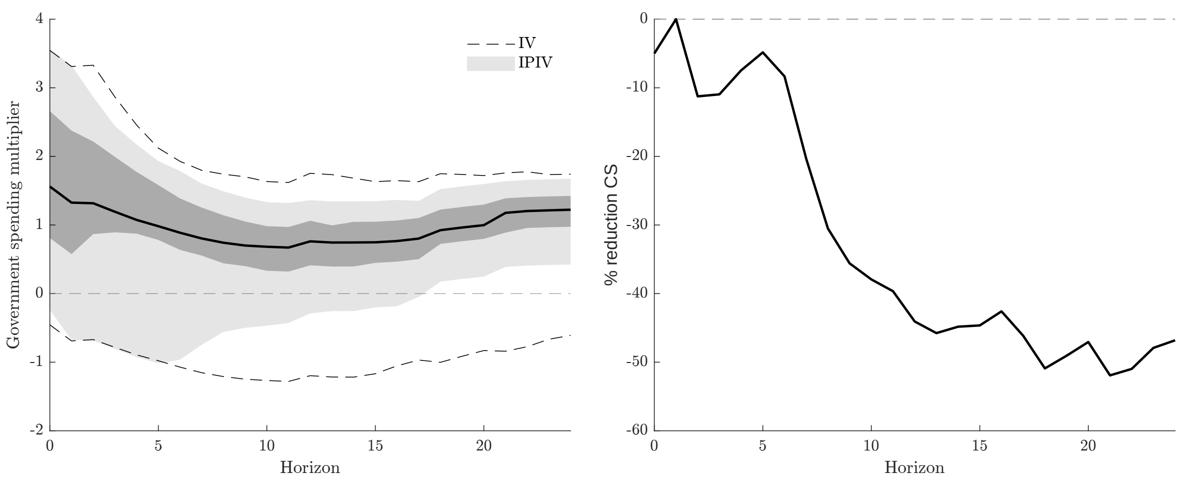
In addition, and compounding this low power problem, finite sample confounding may also be an issue, the Hoover and Perez (1994) critique again: note how three relatively large defense spending shocks overlap with other structural shocks (Figure 6, middle panel): the RZ 1980q1 shock coincides with both the Iran/Iraq war oil shock as well as the appointment of Volcker as Fed chairman, the 1990q4 RZ shock coincides with the oil shock from the Kuwait invasion, and the 2007q4 RZ shock coincides with the financial crisis.[23] This confounding may distort the multiplier estimates.

To assess this possibility, Figure 7b reports multiplier estimates when we leave out one (and only one) of the non-zero RZ shocks at a time (there is a total of 68 non-zero entries). Confirming our suspicion, the three major confounded dates —1980q1, 1990q4 and 2007q4— are most influential for the multiplier estimate. For instance, discarding the 2007q4 positive defense spending shock immediately raises the multiplier point estimate by a *full* unit, from

[23]In figure 6, the oil shocks are the 1956 Suez crisis, the 1973 Arab-Israel war, the 1978 Iranian revolution, the 1980 Iran-Iraq war, the 1990 Persian Gulf war and the 2003 second Persian Gulf war (Hamilton, 1983, 2003). The monetary shocks are from Romer and Romer (1989), and the financial shocks are the BNP Paribas shock in 2007Q3 and Lehman Brothers shock in 2008Q4.

0.3 to 1.3.[24] The other two dates have similarly large effects. As in the monetary case, an ideal solution would thus consist in treating these confounded dates as missing, but this will worsen the low power problem. We thus turn to innovation powering.

Figure 8: Government spending multiplier: Innovation-Powered IV



(a) IPIV estimates (b) Variance reduction

Left panel: Government spending multiplier IPIV estimates based on RZ shock proxy over 1952-2015 with 67 and 95 percent IPAR confidence sets ("IPIV", shaded bands) and 95 percent AR confidence sets for the regular IV estimator ("IV", dashed lines). *Right panel:* difference in the lengths of the 95% bands for the regular IV estimator ("AR bands") and the Innovation-Powered IV estimator ("IPAR bands" CS_h^{IP} based on the $AR_{IP}(\bar{\theta}_h)$ statistic).

Innovation powering

Since the RZ series proxies for news shocks to future defense spending, a good candidate for innovation powering is a series that correlates with news shocks to defense spending. We thus start from the one-year ahead realized change in government spending, and we project out lags of spending and tax revenues to minimize residual endogeneity. This constitutes our innovation v_t.[25] Figure 6 (bottom row) depicts the innovation series along with RZ the shocks.

[24] Intuitively, the 2007q4 *positive* defense spending shocks coincide with the (much larger) *negative* financial shock, a shock that lowered GDP substantially. Because of this confounding, the effect of government spending on output is biased downward, leading to a large increase in the multiplier point estimate once we treat the 2007q4 RZ shock as missing. For the same reason, the 1980q1 positive government spending shock is confounded with the negative oil shock and the negative monetary shock, both of which bias the multiplier estimate downward. Conversely, the negative 1990q4 spending shock is confounded with a negative oil shock, which combined give a large fall in output, and this biases the multiplier estimate upwards.

[25] Alternatively, we explored using Blanchard and Perotti (2002) (BP)'s innovations lead by four quarters. Power improvements were similar.

We set the "good" sample $\mathcal{G}$ to include all non-zero RZ dates, except for the three dates with possible confounding that we set to missing. All other observations are set to missing. This gives us $n_{\mathcal{G}} = 65$ good periods and $n_{\mathcal{B}} = 191$, which implies that 75% of values are treated as missing. As selection model for $\pi(\mathbf{q}_t)$, we use a linear logit model in $y^{\perp}_{t+h}, p^{\perp}_t, v_t$, and we cannot reject the null of bias cancellation between spending and output (p_t, y_{t+h}) or innovations (v_t) with p-values above 0.8 for the bias cancellation tests.

Figure 8 plots the multiplier estimates and confidence sets from innovation powering. Again, IPIV delivers large reduction in confidence bands, in the order of 40-50 percent after 12 quarters. Instead of the uninformative bands (between -0.5 and 2) in Figure 7, the multiplier at the 5 year horizon is now between 0.6 and 1.7 at 95 percent confidence, and 1 and 1.5 at 68 percent. In other words, the post-1952 Ramey-Zubairy news shocks series *does contain* enough information to draw informative inference on the size of the government spending multiplier, allowing to address both externally validity questions as well as Hoover-Perez type critiques.

5 Conclusion

In this work, we introduced a new method —innovation-powered IV—, which allows to reduce the confidence intervals of a credible but low power identification scheme (e.g., a narrative instrument) by leveraging the high power of a possibly mis-specified parametric identification assumption (e.g., a short run restriction). The method delivers large reductions in confidence intervals for the causal effects of monetary and fiscal policy, with gains of around 40 percent compared to state of the art narrative estimates.

Going forward, innovation powering could be used in other macro settings where narrative instruments are used for identification, for instance for the estimation of structural macro equation like the Phillips curve or the Euler equation (Barnichon and Mesters, 2020; Lewis and Mertens, 2023). Further, while linearity assumptions are common when estimating dynamic causal effects in macro, linearity is not a requirement for innovation powering and the method can be easily modified for estimating non-linear effects, or more generally to exploit other nonlinear moment equations (e.g. Hansen and Singleton, 1982; Gonçalves et al., 2024).

Finally, while we mainly focused on the case where z_t is taken from a narrative study, our methodology applies equally for any instrument series that is credible but sparse, and our approach offers a way to exploits endogenous IVs that are traditionally ignored because invalid. Provided that these endogenous IVs are widely available, they can be used to increase the power of exogenous but sparse IVs.

References

Abrevaya, Jason, and Stephen G. Donald. 2017. "A GMM Approach for Dealing with Missing Data on Regressors." *The Review of Economics and Statistics*, 99(4): 657–662.

Alesina, Alberto, Carlo Favero, and Francesco Giavazzi. 2019. *Austerity: When It Works and When It Doesn't.* Princeton University Press.

Anderson, Theodore W., and Herman Rubin. 1949. "Estimation of the Parameters of a Single Equation in a Complete System of Stochastic Equations." *Ann. Math. Statist.*, 20(1): 46–63.

Andrews, Donald W. K. 1991. "Heteroskedasticity and Autocorrelation Consistent Covariance Matrix Estimation." *Econometrica*, 59(3): 817–858.

Andrews, Donald W. K., and Patrik Guggenberger. 2019. "Identification- and singularity-robust inference for moment condition models." *Quantitative Economics*, 10(4): 1703–1746.

Andrews, Isaiah, James Stock, and Liyang Sun. 2019. "Weak Instruments in IV Regression: Theory and Practice." *Annual Review of Economics*, 11: 727–753.

Angelopoulos, Anastasios N., and Tijana Zrnic. 2024. "PPI++: Efficient Prediction-Powered Inference." Working paper.

Angelopoulos, Anastasios N., Stephen Bates, Clara Fannjiang, Michael I. Jordan, and Tijana Zrnic. 2023. "Prediction-powered inference." *Science*, 382(6671): 669–674.

Angrist, Joshua D., Òscar Jordà, and Guido M. Kuersteiner. 2018. "Semiparametric Estimates of Monetary Policy Effects: String Theory Revisited." *Journal of Business & Economic Statistics*, 36(3): 371–387.

Antolín-Díaz, Juan, and Juan F. Rubio-Ramírez. 2018. "Narrative Sign Restrictions for SVARs." *American Economic Review*, 108(10): 2802–29.

Barnichon, Regis, and Geert Mesters. 2020. "Identifying Modern Macro Equations using Old Shocks." *Quarterly Journal of Economics*, 135: 2255–2298.

Barnichon, Regis, and Geert Mesters. 2023. "A Sufficient Statistics Approach for Macroeconomic Policy." *American Economic Review*, 113(11): 2809–45.

Bi, Huixin, and Sarah Zubairy. 2023. "Public Pension Reforms and Retirement Decisions: Narrative Evidence and Aggregate Implications." *American Economic Journal: Economic Policy*, 15(4): 142–82.

Blanchard, Olivier, and Roberto Perotti. 2002. "An Empirical Characterization of the Dynamic Effects of Changes in Government Spending and Taxes on Output." *The Quarterly Journal of Economics*, 117: 1329–1368.

Breusch, Trevor, Hailong Qian, Peter Schmidt, and Donald Wyhowski. 1999. "Redundancy of moment conditions." *Journal of Econometrics*, 91(1): 89–111.

Carriere-Swallow, Yan, Antonio C. David, and Daniel Leigh. 2021. "Macroeconomic Effects of Fiscal Consolidation in Emerging Economies: New Narrative Evidence from Latin America and the Caribbean." *Journal of Money, Credit and Banking*, 53(6): 1313–1335.

Chahrour, Ryan, and Kyle Jurado. 2022. "Recoverability and expectations-driven fluctuations." *The Review of economic studies*, 89(1): 214–239.

Chaudhuri, Saraswata, and David K. Guilkey. 2016. "GMM with Multiple Missing Variables." *Journal of Applied Econometrics*, 31(4): 678–706.

Cloyne, James. 2013. "Discretionary Tax Changes and the Macroeconomy: New Narrative Evidence from the United Kingdom." *American Economic Review*, 103(4): 1507–28.

Cloyne, James, Nicholas Dimsdale, and Natacha Postel-Vinay. 2023. "Taxes and Growth: New Narrative Evidence from Interwar Britain." *The Review of Economic Studies*, rdad081.

de Jong, Robert M., and James Davidson. 2000. "Consistency of Kernel Estimators of Heteroscedastic and Autocorrelated Covariance Matrices." *Econometrica*, 68(2): 407–423.

Durbin, J., and S. J. Koopman. 2012. *Time Series Analysis by State Space Methods; 2nd Edition.* Oxford:Oxford University Press.

Fieldhouse, Andrew J., and Karel Mertens. 2023. "The Returns to Government R&D: Evidence from U.S. Appropriations Shocks." working paper.

Friedman, Milton, and Anna Jacobson Schwartz. 1963. *A monetary history of the United States, 1867-1960.* Princeton University Press.

Gertler, Mark, and Peter Karadi. 2015. "Monetary Policy Surprises, Credit Costs, and Economic Activity." *AEJ: Macroeconomics*, 7: 44–76.

Giacomini, Raffaella, Toru Kitagawa, and Matthew Read. 2022. "Narrative Restrictions and Proxies." *Journal of Business & Economic Statistics*, 40(4): 1415–1425.

Giacomini, Raffaella, Toru Kitagawa, and Matthew Read. 2023. "Identification and Inference Under Narrative Restrictions." working paper.

Gil, Paula, Francisco Martí, Richard Morris, Javier J. Pérez, and Roberto Ramos. 2019. "The output effects of tax changes: narrative evidence from Spain." *SERIEs*, 10(1): 1869–4195.

Gonçalves, Sílvia, Ana María Herrera, Lutz Kilian, and Elena Pesavento. 2024. "State-dependent local projections." *Journal of Econometrics*, 244(2): 105702.

Gorodnichenko, Yuriy. 2014. "Comments on "Government Spending Multipliers in Good Times and in Bad: Evidence from U.S. Historical Data" by Ramey and Zubairy." NBER EFG Meeting.

Guajardo, Jaime, Daniel Leigh, and Andrea Pescatori. 2014. "Expansionary austerity? International evidence." *Journal of the European Economic Association*, 12(4): 949–968.

Hamilton, James D. 1983. "Oil and the Macroeconomy since World War II." *Journal of Political Economy*, 91(2): 228–248.

Hamilton, James D. 1985. "Historical Causes of Postwar Oil Shocks and Recessions." *The Energy Journal*, 6(1): 97–116.

Hamilton, James D. 2003. "What is an oil shock?" *Journal of Econometrics*, 113(2): 363–398.

Hamilton, James D., and Òscar Jordà. 2002. "A Model of the Federal Funds Rate Target." *Journal of Political Economy*, 110(5): 1135–1167.

Hansen, Bruce E., and Seojeong Lee. 2021. "Inference for Iterated GMM Under Misspecification." *Econometrica*, 89(3): 1419–1447.

Hansen, Lars Peter. 1982. "Large sample properties of generalized method of moments estimators." *Econometrica*, 50: 1029–1054.

Hansen, Lars Peter, and Kenneth J Singleton. 1982. "Generalized instrumental variables estimation of nonlinear rational expectations models." *Econometrica*, 1269–1286.

Hansen, Lars Peter, John Heaton, and Amir Yaron. 1996. "Finite-Sample Properties of Some Alternative GMM Estimators." *Journal of Business & Economic Statistics*, 14(3): 262–280.

Hausman, J. A. 1978. "Specification Tests in Econometrics." *Econometrica*, 46(6): 1251–1271.

Herbst, Edward, and Frank Schorfheide. 2016. *Bayesian Estimation of DSGE Models.* . 1 ed., Princeton University Press.

Hirano, Keisuke, Guido W. Imbens, and Geert Ridder. 2003. "Efficient Estimation of Average Treatment Effects Using the Estimated Propensity Score." *Econometrica*, 71(4): 1161–1189.

Hoover, Kevin D, and Stephen J Perez. 1994. "Post hoc ergo propter once more an evaluation of 'does monetary policy matter?'in the spirit of James Tobin." *Journal of monetary Economics*, 34(1): 47–74.

Imbens, Guido W., and Donald B. Rubin. 2015. *Causal Inference for Statistics, Social, and Biomedical Sciences.* Cambridge University Press.

Jalil, Andrew J. 2015. "A New History of Banking Panics in the United States, 1825-1929: Construction and Implications." *American Economic Journal: Macroeconomics*, 7(3): 295–330.

Kilian, Lutz, and Helmut Lütkepohl. 2017. *Structural Vector Autoregressive Analysis.* Cambridge University Press.

Lewis, Daniel J., and Karel Mertens. 2023. "Dynamic Identification Using System Projections and Instrumental Variables." working paper.

Little, Roderick J. A., and Donald B. Rubin. 2020. *Statistical Analysis with Missing Data.* John Wiley & Sons, Inc.

Ludvigson, Sydney C, Sai Ma, and Serena Ng. 2017. "Shock Restricted Structural Vector-Autoregressions." National Bureau of Economic Research Working Paper 23225.

Nakamura, Emi, and Jón Steinsson. 2018. "Identification in Macroeconomics." *Journal of Economic Perspectives*, 32(3): 59–86.

Newey, Whitney K., and Daniel McFadden. 1994. "Chapter 36 Large sample estimation and hypothesis testing." In . Vol. 4 of *Handbook of Econometrics*, 2111 – 2245. Elsevier.

Newey, Whitney K., and Kenneth D. West. 1987. "A Simple, Positive Semi-Definite, Heteroskedasticity and Autocorrelation Consistent Covariance Matrix." *Econometrica*, 55(3): 703–708.

Olea, José Luis Montiel, and Carolin Pflueger. 2013. "A robust test for weak instruments." *Journal of Business & Economic Statistics*, 31(3): 358–369.

Plagborg-Møller, Mikkel. 2022. "Discussion of "Narrative Restrictions and Proxies" by Raffaella Giacomini, Toru Kitagawa, and Matthew Read." *Journal of Business & Economic Statistics*, 40(4): 1434–1437.

Plagborg-Møller, Mikkel, and Christian K. Wolf. 2021. "Local Projections and VARs Estimate the Same Impulse Responses." *Econometrica*, 89(2): 955–980.

Ramey, Valerie. 2016. "Macroeconomic Shocks and Their Propagation." In *Handbook of Macroeconomics.* , ed. J. B. Taylor and H. Uhlig. Amsterdam, North Holland:Elsevier.

Ramey, Valerie A. 2011. "Identifying Government Spending Shocks: It's all in the Timing." *The Quarterly Journal of Economics*, 126(1): 1–50.

Ramey, Valerie A., and Matthew D. Shapiro. 1998. "Costly capital reallocation and the effects of government spending." *Carnegie-Rochester Conference Series on Public Policy*, 48: 145–194.

Ramey, Valerie A., and Sarah Zubairy. 2018. "Government Spending Multipliers in Good Times and in Bad: Evidence from U.S. Historical Data." *Journal of Political Economy*, 126.

Reinhart, Carmen M, and Kenneth S Rogoff. 2009. “This time is different.” In *This Time Is Different.* princeton university press.

Robins, James M., Andrea Rotnitzky, and Lue Ping Zhao. 1994. “Estimation of Regression Coefficients When Some Regressors Are Not Always Observed.” *Journal of the American Statistical Association*, 89(427): 846–866.

Robins, James M., Andrea Rotnitzky, and Lue Ping Zhao. 1995. “Analysis of Semiparametric Regression Models for Repeated Outcomes in the Presence of Missing Data.” *Journal of the American Statistical Association*, 90(429): 106–121.

Rojas, Diego, Carlos Vegh, and Guillermo Vuletin. 2022. “The macroeconomic effects of macroprudential policy: Evidence from a narrative approach.” *Journal of International Economics*, 139: 103644.

Romer, Christina D, and David H Romer. 1989. “Does monetary policy matter? A new test in the spirit of Friedman and Schwartz.” *NBER macroeconomics annual*, 4: 121–170.

Romer, Christina D., and David H. Romer. 2004. “A New Measure of Monetary Shocks: Derivation and Implications.” *American Economic Review*, 94: 1055–1084.

Romer, Christina D., and David H. Romer. 2010. “The Macroeconomic Effects of Tax Changes: Estimates Based on a New Measure of Fiscal Shocks.” *American Economic Review*, 100(3): 763–801.

Romer, Christina D., and David H. Romer. 2017. “New Evidence on the Aftermath of Financial Crises in Advanced Countries.” *American Economic Review*, 107(10): 3072–3118.

Romer, Christina D., and David H. Romer. 2023. “Presidential Address: Does Monetary Policy Matter? The Narrative Approach after 35 Years.” *American Economic Review*, 113(6): 1395–1423.

Sims, Christopher A. 1980. “Macroeconomics and reality.” *Econometrica*, 1–48.

Stock, James H., and Jonathan H. Wright. 2000. “GMM with Weak Identification.” *Econometrica*, 68(5): 1055–1096.

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

Wooldridge, Jeffrey M. 2007. “Inverse probability weighted estimation for general missing data problems.” *Journal of Econometrics*, 141(2): 1281–1301.

Wooldridge, Jeffrey M. 2010. *Econometric Analysis of Cross Section and Panel Data, Second Edition.* MIT Press.

Appendix: econometric details

In this appendix we discuss the econometric details for IPIV inference and we provide the proofs for the results stated in the main text. Some lemmas are deferred to the supplementary material; they are indicated by (S).

A1: Deriving the efficient weights

To derive the formal results we require a set of regularity conditions. To present these in a compact form we define $\boldsymbol{\psi} = (\theta_h, \boldsymbol{\delta}')'$ with parameter space $\Psi = \Theta \times \Delta$, where $\Theta \subset \mathbb{R}$ and $\Delta \subset \mathbb{R}^{d_\delta}$. The average decomposed sample moments are summarized in

$$h_n(\boldsymbol{\psi}) = \frac{1}{n}\sum_{t=1}^{n} h(\mathbf{d}_t, \boldsymbol{\psi}) \text{ , with } h(\mathbf{d}_t, \boldsymbol{\psi}) = \begin{pmatrix} z_t^{\perp}(y_{t+h}^{\perp} - \theta_h p_t^{\perp}) s_t/\kappa(\mathbf{q}_t; \boldsymbol{\delta}) \\ \mathbf{v}_t(y_{t+h}^{\perp} - \theta_h p_t^{\perp})(1 - s_t/\kappa(\mathbf{q}_t; \boldsymbol{\delta})) \\ m(\mathbf{q}_t, \boldsymbol{\delta}) \end{pmatrix} \text{ ,} \tag{24}$$

where $\mathbf{d}_t = (\mathbf{q}_t, z_t^{\perp}, s_t)$. The relevant gradients are defined as

$$\begin{aligned} \mathbf{G}_\theta &= \mathbb{E}(\nabla_{\theta_h} g(\mathbf{d}_t, \theta_h, \boldsymbol{\delta}; \boldsymbol{\gamma})) &&= -\mathbb{E}(z_t^{\perp} p_t^{\perp} s_t/\kappa(\mathbf{q}_t; \boldsymbol{\delta})) \text{ ,} \\ \mathbf{G}_\delta(\boldsymbol{\gamma}) &= \mathbb{E}(\nabla_{\boldsymbol{\delta}} g(\mathbf{d}_t, \theta_h, \boldsymbol{\delta}; \boldsymbol{\gamma})) &&= -\mathbb{E}\left(\frac{(z_t^{\perp} - \boldsymbol{\gamma}'\mathbf{v}_t)(y_{t+h}^{\perp} - \theta_h p_t^{\perp})\kappa^{(1)}(\mathbf{d}_t^{\pi}; \boldsymbol{\delta})}{\kappa(\mathbf{d}_t^{\pi}, \boldsymbol{\delta})}\right) \text{ ,} \\ \mathbf{M} &= \mathbb{E}(\nabla_{\boldsymbol{\delta}'} m(\mathbf{d}_t^{\pi}, \boldsymbol{\delta})) &&= -\mathbb{E}\left(\frac{\boldsymbol{\kappa}^{(1)}(\mathbf{d}_t^{\pi}; \boldsymbol{\delta})\boldsymbol{\kappa}^{(1)}(\mathbf{d}_t^{\pi}; \boldsymbol{\delta})'}{(\kappa(\mathbf{d}_t^{\pi}; \boldsymbol{\delta})(1 - \kappa(\mathbf{d}_t^{\pi}; \boldsymbol{\delta})))}\right) \text{ ,} \end{aligned} \tag{25}$$

where g and m are defined in Lemma 1. Further, the IPIV estimate and selection model estimates are given by $\widehat{\boldsymbol{\psi}}(\boldsymbol{\gamma}) = (\hat{\theta}_h(\boldsymbol{\gamma}), \widehat{\boldsymbol{\delta}}')'$. Finally, throughout we assume that all random variables are defined on some underlying probability space and all considered functions of the random variables are measurable. The required regularity conditions are as follows.

Assumption R.

1. *For some $r > 2$ we have that $\{\mathbf{d}_t\}$ is a strictly stationary process that is α-mixing of size $-r/(r-2)$.*

2. *Ψ is compact and $\boldsymbol{\psi} \in \text{int}(\Psi)$.*

3. *We have $\mathbb{E}(|z_t^{\perp}\xi_{t+h}^{\perp}|^r) < \infty$, $\mathbb{E}(|z_t^{\perp}p_t^{\perp}|^r) < \infty$, $\mathbb{E}(\|\mathbf{v}_t\xi_{t+h}^{\perp}\|^r) < \infty$ and $\mathbb{E}(\|\mathbf{v}_t p_t^{\perp}\|^2) < \infty$*

4. *(i) $\mathbb{E}m(\mathbf{q}_t, \tilde{\boldsymbol{\delta}}) \neq 0$ for all $\tilde{\boldsymbol{\delta}} \neq \boldsymbol{\delta}$, (ii) $\kappa(\mathbf{q}_t; \tilde{\boldsymbol{\delta}})$ is twice continuously differentiable at each $\tilde{\boldsymbol{\delta}} \in \Delta$ with probability 1, (iii) there exist constants c_1, c_2 such that $0 < c_1 \leq \kappa(\mathbf{q}_t; \tilde{\boldsymbol{\delta}}) \leq c_2 < 1$ at each $\tilde{\boldsymbol{\delta}} \in \Delta$ with probability 1, (iv) $\mathbb{E}[\sup_{\tilde{\boldsymbol{\delta}} \in \Delta} \|\kappa^{(1)}(\mathbf{q}_t; \tilde{\boldsymbol{\delta}})\|] < \infty$, (v) let N be a neighborhood of $\boldsymbol{\psi}$, we have $\mathbb{E}[\sup_{\tilde{\boldsymbol{\delta}} \in N} \|\kappa^{(1)}(\mathbf{q}_t; \tilde{\boldsymbol{\delta}})\|^2] < \infty$ and $\mathbb{E}(\sup_{\tilde{\boldsymbol{\delta}} \in N} \|\kappa^{(2)}(\mathbf{q}_t; \tilde{\boldsymbol{\delta}})\|) < \infty$, (vi) $\mathbb{E}(\|\boldsymbol{\kappa}^{(1)}(\mathbf{q}_t; \boldsymbol{\delta})\|^r) < \infty$ and (vii) $\mathbf{M}$ is non-singular.*

5. *$\mathbf{S} = \lim_{n\to\infty} \text{var}(n^{1/2}h_n(\boldsymbol{\psi}))$ is positive definite with partitioning conforming $h(\mathbf{d}_t; \boldsymbol{\psi})$:*

$$\mathbf{S} = \begin{pmatrix} \mathbf{S}_{zz} & \mathbf{S}_{zv} & \mathbf{S}_{zm} \\ \mathbf{S}_{vz} & \mathbf{S}_{vv} & \mathbf{S}_{vm} \\ \mathbf{S}_{mz} & \mathbf{S}_{mv} & \mathbf{S}_{mm} \end{pmatrix} \text{ .}$$

Conditions 1 and 3 restrict the dependence and the moments of the process $\{\mathbf{d}_t\}$ in such a way that a central limit theorem applies to suitable rescaled averages and functions of the process. Condition 2 imposes that the parameter space is compact which can be relaxed using more refined arguments. Condition 4 imposes a set of regularity conditions on the selection model that can be verified for any specific model. Condition 5 ensures that the long run variance of the sample moments is well behaved.

In addition, the following assumption is used to derive a standard asymptotic normal approximation for the distribution of the IPIV estimator.

Assumption I. $\mathbb{E}(z_t^{\perp} p_t^{\perp}) \neq 0$.

This imposes that the narrative instruments are strong: the correlation between the instrument and the variable of interest is bounded away from zero in the limit. The evidence for whether this assumption holds for narrative instrument series is mixed, while some studies like Ramey and Zubairy (2018); Fieldhouse and Mertens (2023) present compelling evidence that suggests that their narrative instruments are strong, others do not present such evidence and our own analysis for monetary policy based on the Romer and Romer (1989, 2023) instrument series shows that here the assumption is too strong. In our empirical work we rely on the IPAR statistic for constructing confidence bands, which implies that we do not require Assumption I as is reflected in the statement of Theorem 1.

Theorem 2. *Given Assumptions 1-2 and Assumptions R and I we have for any* $\boldsymbol{\gamma}_n \xrightarrow{p} \boldsymbol{\gamma}$ *that* $\hat{\theta}_h(\boldsymbol{\gamma}_n) \xrightarrow{p} \theta_h$ *and*

$$\sqrt{n}(\hat{\theta}_h(\boldsymbol{\gamma}_n) - \theta_h) \xrightarrow{d} N(0, V(\boldsymbol{\gamma})) \ ,$$

where $V(\boldsymbol{\gamma}) = \omega(\boldsymbol{\gamma})/\mathbf{G}_\theta^2$ *with*

$$\omega(\boldsymbol{\gamma}) = \boldsymbol{\Omega}_{gg}(\boldsymbol{\gamma}) + \mathbf{G}_\delta(\boldsymbol{\gamma})'\mathbf{M}^{-1}\boldsymbol{\Omega}_{mm}\mathbf{M}^{-1}\mathbf{G}_\delta(\boldsymbol{\gamma}) - 2\mathbf{G}_\delta(\boldsymbol{\gamma})'\mathbf{M}^{-1}\boldsymbol{\Omega}_{mg}(\boldsymbol{\gamma}) \ . \tag{26}$$

with $\boldsymbol{\Omega}(\boldsymbol{\gamma}) = \mathbf{U}(\boldsymbol{\gamma})'\mathbf{S}\mathbf{U}(\boldsymbol{\gamma})$.

The following corollary determines the $\boldsymbol{\gamma}$ weights that minimize $\omega(\boldsymbol{\gamma})$ and the asymptotic variance of the IPIV estimator $\hat{\theta}_h(\boldsymbol{\gamma}_n)$.

Corollary 1. *Given Assumptions 1-2 and Assumption R the optimal weights that minimize* $\omega(\boldsymbol{\gamma})$ *and (under additionally assumption I) the asymptotic variance* $V(\boldsymbol{\gamma})$ *are given by*

$$\begin{aligned}\boldsymbol{\gamma}^* = &-\left(\mathbf{S}_{vv} + \mathbf{G}_{\delta v}'\mathbf{M}^{-1}\boldsymbol{\Omega}_{mm}\mathbf{M}^{-1}\mathbf{G}_{\delta v} - 2\mathbf{G}_{\delta v}'\mathbf{M}^{-1}\mathbf{S}_{mv}\right)^{-1} \\ &\times \left(\mathbf{S}_{vz} + \mathbf{G}_{\delta v}'\mathbf{M}^{-1}\boldsymbol{\Omega}_{mm}\mathbf{M}^{-1}\mathbf{G}_{\delta z} - 2\mathbf{G}_{\delta v}'\mathbf{M}^{-1}\mathbf{S}_{mz}\right) \ ,\end{aligned} \tag{27}$$

where

$$\begin{aligned}\mathbf{G}_{\delta z} &= -\mathbb{E}(z_t^{\perp}(y_{t+h}^{\perp} - \theta_h p_t^{\perp})\kappa^{(1)}(\mathbf{d}_t^{\pi};\boldsymbol{\delta})/\kappa(\mathbf{d}_t^{\pi},\boldsymbol{\delta})) \ , \\ \mathbf{G}_{\delta v}' &= \quad \mathbb{E}(\mathbf{v}_t(y_{t+h}^{\perp} - \theta_h p_t^{\perp})\kappa^{(1)}(\mathbf{d}_t^{\pi};\boldsymbol{\delta})'/\kappa(\mathbf{d}_t^{\pi},\boldsymbol{\delta})) \ .\end{aligned}$$

Importantly, if the strong instruments assumption does not hold $\boldsymbol{\gamma}^*$ continues to minimize $\omega(\boldsymbol{\gamma})$ which is the variance of the sample moment that is used to build the IPAR statistic, see equation (21).

Assumption R.

6. There exists an estimator $\widehat{\mathbf{S}}(\widehat{\boldsymbol{\psi}})$ such that $\widehat{\mathbf{S}}(\widehat{\boldsymbol{\psi}}) \xrightarrow{p} \mathbf{S}$ for any $\widehat{\boldsymbol{\psi}} \xrightarrow{p} \boldsymbol{\psi}$.

This condition imposes that a consistent estimate for the asymptotic variance exists. Primitive assumptions for specific estimators can be found in for example Newey and West (1987), Andrews (1991) and de Jong and Davidson (2000). With this the efficient weights can be consistently estimated by

$$\begin{aligned}\widehat{\boldsymbol{\gamma}}(\hat{\theta}_h) = &- \left(\widehat{\mathbf{S}}_{vv}(\widehat{\boldsymbol{\psi}}) + \widehat{\mathbf{G}}_{\delta v}(\hat{\theta}_h)'\widehat{\mathbf{M}}^{-1}\widehat{\mathbf{S}}_{mm}(\widehat{\boldsymbol{\psi}})\widehat{\mathbf{M}}^{-1}\widehat{\mathbf{G}}_{\delta v}(\hat{\theta}_h) - 2\widehat{\mathbf{G}}_{\delta v}(\hat{\theta}_h)'\widehat{\mathbf{M}}^{-1}\widehat{\mathbf{S}}_{mv}(\widehat{\boldsymbol{\psi}})\right)^{-1} \\ &\times \left(\widehat{\mathbf{S}}_{vz}(\widehat{\boldsymbol{\psi}}) + \widehat{\mathbf{G}}_{\delta v}(\hat{\theta}_h)'\widehat{\mathbf{M}}^{-1}\widehat{\mathbf{S}}_{mm}(\widehat{\boldsymbol{\psi}})\widehat{\mathbf{M}}^{-1}\widehat{\mathbf{G}}_{\delta z}(\hat{\theta}_h) - 2\widehat{\mathbf{G}}_{\delta v}(\hat{\theta}_h)'\widehat{\mathbf{M}}^{-1}\widehat{\mathbf{S}}_{mz}(\widehat{\boldsymbol{\psi}})\right) ,\end{aligned} \tag{28}$$

where $\widehat{\boldsymbol{\psi}} = (\hat{\theta}_h, \widehat{\boldsymbol{\delta}}')'$ is an arbitrary consistent estimator,

$$\begin{aligned}\widehat{\mathbf{G}}_{\delta z}(\hat{\theta}_h) &= -\frac{1}{n}\sum_{t\in\mathcal{N}} s_t z_t^{\perp}(y_{t+h}^{\perp} - \hat{\theta}_h p_t^{\perp})\kappa^{(1)}(\mathbf{d}_t^{\pi};\widehat{\boldsymbol{\delta}})/\kappa^2(\mathbf{d}_t^{\pi},\widehat{\boldsymbol{\delta}}) ,\\ \widehat{\mathbf{G}}_{\delta v}(\hat{\theta}_h)' &= \frac{1}{n}\sum_{t\in\mathcal{N}} s_t \mathbf{v}_t(y_{t+h}^{\perp} - \hat{\theta}_h p_t^{\perp})\kappa^{(1)}(\mathbf{d}_t^{\pi};\widehat{\boldsymbol{\delta}})'/\kappa^2(\mathbf{d}_t^{\pi},\widehat{\boldsymbol{\delta}}) ,\\ \widehat{\mathbf{M}} &= -\frac{1}{n}\sum_{t\in\mathcal{N}} \boldsymbol{\kappa}^{(1)}(\mathbf{d}_t^{\pi};\widehat{\boldsymbol{\delta}})\boldsymbol{\kappa}^{(1)}(\mathbf{d}_t^{\pi};\widehat{\boldsymbol{\delta}})'/(\kappa(\mathbf{d}_t^{\pi};\widehat{\boldsymbol{\delta}})(1-\kappa(\mathbf{d}_t^{\pi};\widehat{\boldsymbol{\delta}})))\end{aligned}$$

and note that $\widehat{\mathbf{S}}(\widehat{\boldsymbol{\psi}})$ partitions conforming $\mathbf{S}$.

Corollary 2. *Given Assumptions 1-2 and Assumption R, we have for any $\widehat{\boldsymbol{\psi}} \xrightarrow{p} \boldsymbol{\psi}$ that*

$$\widehat{\boldsymbol{\gamma}}(\hat{\theta}_h) \xrightarrow{p} \boldsymbol{\gamma}^* ,$$

where $\widehat{\boldsymbol{\gamma}}$ is defined in (28) and $\boldsymbol{\gamma}^*$ in Corollary 1.

The estimated weights are used to compute the IPAR statistic in equation (21). We note that the consistent estimation of these weights does not require the strong instruments assumption I.

A2: Variance IPAR statistic

For the IPAR statistic (21) the variance estimate is given by

$$\widehat{\omega}(\theta_h,\boldsymbol{\gamma}) = \widehat{\boldsymbol{\Omega}}_{gg}(\theta_h,\boldsymbol{\gamma}) + \widehat{\mathbf{G}}_{\delta}(\theta_h,\boldsymbol{\gamma})'\widehat{\mathbf{M}}^{-1}\widehat{\boldsymbol{\Omega}}_{mm}\widehat{\mathbf{M}}^{-1}\widehat{\mathbf{G}}_{\delta}(\theta_h,\boldsymbol{\gamma}) - 2\widehat{\mathbf{G}}_{\delta}(\theta_h,\boldsymbol{\gamma})'\widehat{\mathbf{M}}^{-1}\widehat{\boldsymbol{\Omega}}_{mg}(\theta_h,\boldsymbol{\gamma}) , \tag{29}$$

with $\widehat{\mathbf{G}}_{\delta}(\theta_h,\boldsymbol{\gamma}) = \widehat{\mathbf{G}}_{\delta z}(\theta_h) + \widehat{\mathbf{G}}_{\delta v}(\theta_h)\boldsymbol{\gamma}$ and $\widehat{\boldsymbol{\Omega}}(\theta_h,\boldsymbol{\gamma}) = \mathbf{U}(\boldsymbol{\gamma})'\widehat{\mathbf{S}}(\theta_h,\boldsymbol{\delta})\mathbf{U}(\boldsymbol{\gamma})$ with

$$\mathbf{U}(\boldsymbol{\gamma})' = \begin{pmatrix} 1 & \boldsymbol{\gamma}' & 0 \\ 0 & 0 & \mathbf{I}_{d_\delta}\end{pmatrix} \quad \text{and partitioning} \quad \widehat{\boldsymbol{\Omega}}(\boldsymbol{\gamma}) = \begin{pmatrix} \widehat{\boldsymbol{\Omega}}_{gg}(\boldsymbol{\gamma}) & \widehat{\boldsymbol{\Omega}}_{gm}(\boldsymbol{\gamma}) \\ \widehat{\boldsymbol{\Omega}}_{mg}(\boldsymbol{\gamma}) & \widehat{\boldsymbol{\Omega}}_{mm}\end{pmatrix} .$$

To compute the IPAR statistic we evaluate $\widehat{\omega}(\theta_h,\boldsymbol{\gamma})$ at $\bar{\theta}_h$, $\widehat{\boldsymbol{\gamma}}(\bar{\theta}_h)$ and $\widehat{\boldsymbol{\delta}}$.

A3: Proofs

Proof of Lemma 1. Assumption 1 states $\mathbb{E}(z_t^{\perp}(y_{t+h}^{\perp} - \theta_h p_t^{\perp})) = 0$. The law of iterative expectations together with Assumption 2 implies

$$\begin{aligned}\mathbb{E}(z_t^{\perp}(y_{t+h}^{\perp} - \theta_h p_t^{\perp})s_t/\kappa(\mathbf{q}_t, \boldsymbol{\delta})) &= \mathbb{E}(z_t^{\perp}(y_{t+h}^{\perp} - \theta_h p_t^{\perp})\mathbb{E}(s_t|\mathbf{q}_t, z_t^{\perp})/\kappa(\mathbf{q}_t, \boldsymbol{\delta}))\\ &= \mathbb{E}(z_t^{\perp}(y_{t+h}^{\perp} - \theta_h p_t^{\perp})\mathbb{E}(s_t|\mathbf{q}_t)/\kappa(\mathbf{q}_t, \boldsymbol{\delta}))\\ &= \mathbb{E}(z_t^{\perp}(y_{t+h}^{\perp} - \theta_h p_t^{\perp})\kappa(\mathbf{q}_t, \boldsymbol{\delta})/\kappa(\mathbf{q}_t, \boldsymbol{\delta}))\\ &= \mathbb{E}(z_t^{\perp}(y_{t+h}^{\perp} - \theta_h p_t^{\perp})) = 0\ .\end{aligned}$$

To obtain the expression in the Lemma we add and subtract $\boldsymbol{\gamma}'\mathbb{E}(\mathbf{v}_t(y_{t+h}^{\perp} - \theta_h p_t^{\perp}))$ while nothing that under Assumption 2 we have $\boldsymbol{\gamma}'\mathbb{E}(\mathbf{v}_t(y_{t+h}^{\perp} - \theta_h p_t^{\perp})) = \boldsymbol{\gamma}'\mathbb{E}(\mathbf{v}_t(y_{t+h}^{\perp} - \theta_h p_t^{\perp})s_t/\kappa(\mathbf{q}_t, \boldsymbol{\delta}))$, which again follows from the law of iterative expectations. We get

$$\mathbb{E}(z_t^{\perp}(y_{t+h}^{\perp} - \theta_h p_t^{\perp})s_t/\kappa(\mathbf{q}_t, \boldsymbol{\delta})) + \boldsymbol{\gamma}'\mathbb{E}(\mathbf{v}_t(y_{t+h}^{\perp} - \theta_h p_t^{\perp})(1 - s_t/\kappa(\mathbf{q}_t, \boldsymbol{\delta})) = 0\ ,$$

which is equal to $\mathbb{E}g(\mathbf{d}_t, \theta_h, \boldsymbol{\delta}; \boldsymbol{\gamma}) = 0$. Next, the expression for $m(\mathbf{q}_t; \boldsymbol{\delta})$ in (15) follows from Wooldridge (2010, equation (15.18)) and by Assumption 2 this corresponds to the first order conditions of a correctly specified model such that $\mathbb{E}m(\mathbf{q}_t; \boldsymbol{\delta}) = 0$. □

Proof of Theorem 2. Note that

$$\widehat{\boldsymbol{\psi}}(\boldsymbol{\gamma}_n) = \underset{\tilde{\boldsymbol{\psi}} \in \Psi}{\operatorname{argmin}}\, h_n(\tilde{\boldsymbol{\psi}})'\mathbf{W}_n h_n(\tilde{\boldsymbol{\psi}}) \quad \text{where} \quad \mathbf{W}_n = \begin{pmatrix} 1 & \boldsymbol{\gamma}_n' & 0 \\ \boldsymbol{\gamma}_n & \boldsymbol{\gamma}_n\boldsymbol{\gamma}_n' & 0 \\ 0 & 0 & \mathbf{I}_{d_v} \end{pmatrix}\ , \tag{30}$$

and $\mathbf{W}_n$ is positive semi-definite. This clarifies that the IPIV estimator can be written as a GMM estimator. To show consistency Lemma S1 uses the stated assumptions to verify the conditions of Theorem 2.1 in Newey and McFadden (1994), which imply $\widehat{\boldsymbol{\psi}}(\boldsymbol{\gamma}_n) \xrightarrow{p} \boldsymbol{\psi}$ and thus $\hat{\theta}_h(\boldsymbol{\gamma}_n) \xrightarrow{p} \theta_h$. Next, to derive the limiting distribution we compute the first order condition of (30) and rearrange to obtain

$$\sqrt{n}(\hat{\boldsymbol{\psi}}(\boldsymbol{\gamma}_n) - \boldsymbol{\psi}) = -(\mathbf{H}_n(\hat{\boldsymbol{\psi}}(\boldsymbol{\gamma}_n))'\mathbf{W}_n\mathbf{H}_n(\bar{\boldsymbol{\psi}}))^{-1}\mathbf{H}_n(\hat{\boldsymbol{\psi}}(\boldsymbol{\gamma}_n))'\mathbf{W}_n\sqrt{n}h_n(\boldsymbol{\psi})\ ,$$

where $\bar{\boldsymbol{\psi}}$ lies on the line segment between $\hat{\boldsymbol{\psi}}(\boldsymbol{\gamma}_n)$ and $\boldsymbol{\psi}$,

$$\mathbf{H}_n(\boldsymbol{\psi}) = \frac{1}{n}\sum_{t=1}^{n}\mathbf{H}(\mathbf{d}_t, \boldsymbol{\psi}) \qquad \text{and} \qquad \mathbf{H}(\mathbf{d}_t, \boldsymbol{\psi}) = \frac{\partial h(\mathbf{d}_t, \boldsymbol{\psi})}{\partial \boldsymbol{\psi}'}\ ,$$

with exact expressions given in Lemma S2-(iv) which shows that $\mathbb{E}\mathbf{H}(\mathbf{d}_t, \tilde{\boldsymbol{\psi}})$ is continuous in $\tilde{\boldsymbol{\psi}}$ and $\sup_{\tilde{\boldsymbol{\psi}} \in N}\|\mathbf{H}_n(\tilde{\boldsymbol{\psi}}) - \mathbb{E}\mathbf{H}(\mathbf{d}_t; \tilde{\boldsymbol{\psi}})\| \xrightarrow{p} 0$ for some neighborhood N of $\boldsymbol{\psi}$. This, together with the consistency of $\hat{\boldsymbol{\psi}}(\boldsymbol{\gamma}_n)$, is sufficient to ensure that $\mathbf{H}_n(\hat{\boldsymbol{\psi}}(\boldsymbol{\gamma}_n)) \xrightarrow{p} \mathbf{H}$ and $\mathbf{H}_n(\bar{\boldsymbol{\psi}}) \xrightarrow{p} \mathbf{H}$. Further, as $\mathbf{W}_n \xrightarrow{p} \mathbf{W}$ we have $-(\mathbf{H}_n(\hat{\boldsymbol{\psi}}(\boldsymbol{\gamma}_n))'\mathbf{W}_n\mathbf{H}_n(\bar{\boldsymbol{\psi}}))^{-1}\mathbf{H}_n(\hat{\boldsymbol{\psi}}(\boldsymbol{\gamma}_n))'\mathbf{W}_n \xrightarrow{p} -(\mathbf{H}'\mathbf{W}\mathbf{H})^{-1}\mathbf{H}'\mathbf{W}$. Lemma S3 shows that $\sqrt{n}h_n(\boldsymbol{\psi}) \xrightarrow{d} N(0, \mathbf{S})$, such that Slutsky's lemma allows to conclude

$$\sqrt{n}(\hat{\boldsymbol{\psi}}(\tilde{\boldsymbol{\gamma}}_n) - \boldsymbol{\psi}) \xrightarrow{d} N(0, (\mathbf{H}'\mathbf{W}\mathbf{H})^{-1}\mathbf{H}'\mathbf{W}\mathbf{S}\mathbf{W}\mathbf{H}(\mathbf{H}'\mathbf{W}\mathbf{H})^{-1})\ .$$

Finally, we show that the top left entry of the asymptotic variance corresponds to the expression in the theorem. We have that

$$(\mathbf{H}'\mathbf{W}\mathbf{H})^{-1} = \begin{pmatrix} \mathbf{G}_\theta^{-2} + \mathbf{G}_\theta^{-1}\mathbf{G}_\delta'\mathbf{M}^{-2}\mathbf{G}_\delta\mathbf{G}_\theta^{-1} & -\mathbf{G}_\theta^{-1}\mathbf{G}_\delta'\mathbf{M}^{-2} \\ -\mathbf{M}^{-2}\mathbf{G}_\delta\mathbf{G}_\theta^{-1} & \mathbf{M}^{-2} \end{pmatrix}$$

Also, using $\mathbf{W} = \mathbf{U}(\boldsymbol{\gamma})\mathbf{U}(\boldsymbol{\gamma})'$ we have

$$\mathbf{U}(\boldsymbol{\gamma})'\mathbf{S}\mathbf{U}(\boldsymbol{\gamma}) = \begin{pmatrix} \mathbf{S}_{zz} + \tilde{\boldsymbol{\gamma}}'\mathbf{S}_{vz} + \mathbf{S}_{zv}\tilde{\boldsymbol{\gamma}} + \tilde{\boldsymbol{\gamma}}'\mathbf{S}_{vv}\tilde{\boldsymbol{\gamma}} & \mathbf{S}_{zm} + \tilde{\boldsymbol{\gamma}}'\mathbf{S}_{vm} \\ \mathbf{S}_{mz} + \mathbf{S}_{mv}\tilde{\boldsymbol{\gamma}} & \mathbf{S}_{mm} \end{pmatrix} \equiv \begin{pmatrix} \boldsymbol{\Omega}_{gg}(\tilde{\boldsymbol{\gamma}}) & \boldsymbol{\Omega}_{gm}(\tilde{\boldsymbol{\gamma}}) \\ \boldsymbol{\Omega}_{mg}(\tilde{\boldsymbol{\gamma}}) & \boldsymbol{\Omega}_{mm} \end{pmatrix},$$

Putting everything together shows that the top left element is equal to

$$\frac{\boldsymbol{\Omega}_{gg}(\tilde{\boldsymbol{\gamma}}) + \mathbf{G}_\delta'\mathbf{M}^{-1}\boldsymbol{\Omega}_{mm}\mathbf{M}^{-1}\mathbf{G}_\delta - 2\mathbf{G}_\delta'\mathbf{M}^{-1}\boldsymbol{\Omega}_{mg}}{\mathbf{G}_\theta^2}.$$

□

Proof of Corollary 1. In terms of the partitioning of $\mathbf{S}$ we can write $\boldsymbol{\Omega}_{gg} = \mathbf{S}_{zz} + 2\mathbf{S}_{zv}\boldsymbol{\gamma} + \boldsymbol{\gamma}'\mathbf{S}_{vv}\boldsymbol{\gamma}$, $\boldsymbol{\Omega}_{mg} = \mathbf{S}_{mz} + \mathbf{S}_{mv}\boldsymbol{\gamma}$ and $\mathbf{G}_\delta(\boldsymbol{\gamma}) = \mathbf{G}_{\delta z} + \mathbf{G}_{\delta v}\boldsymbol{\gamma}$, where $\mathbf{G}_{\delta z}$ and $\mathbf{G}_{\delta v}$ are defined in the corollary. The variance matrix $V(\boldsymbol{\gamma}) = \omega(\boldsymbol{\gamma})/\mathbf{G}_\theta^2$ only depends on $\boldsymbol{\gamma}$ via ω. We solve $\min_{\boldsymbol{\gamma}} \omega(\boldsymbol{\gamma})$ to find the minimum. Using the notation above we have

$$\begin{aligned}\omega(\boldsymbol{\gamma}) =& \mathbf{S}_{zz} + 2\mathbf{S}_{zv}\boldsymbol{\gamma} + \boldsymbol{\gamma}'\mathbf{S}_{vv}\boldsymbol{\gamma} + (\mathbf{G}_{\delta z} + \mathbf{G}_{\delta v}\boldsymbol{\gamma})'\mathbf{M}^{-1}\boldsymbol{\Omega}_{mm}\mathbf{M}^{-1}(\mathbf{G}_{\delta z} + \mathbf{G}_{\delta v}\boldsymbol{\gamma}) \\ &- 2(\mathbf{G}_{\delta z} + \mathbf{G}_{\delta v}\boldsymbol{\gamma})'\mathbf{M}^{-1}(\mathbf{S}_{mz} + \mathbf{S}_{mv}\boldsymbol{\gamma}),\end{aligned}$$

which shows that $\omega(\boldsymbol{\gamma})$ is quadratic in $\boldsymbol{\gamma}$. Taking the derivative and setting it to zero gives $\boldsymbol{\gamma}^*$ as stated in the Corollary. □

Proof of Corollary 2. Given $\widehat{\boldsymbol{\psi}} \xrightarrow{p} \boldsymbol{\psi}$ the convergence follows directly from the continuous mapping theorem using Lemma S2-(i)-(iii) and Assumption R-6. □

Proof of Theorem 1. Under $H_0 : \theta_h = \bar{\theta}_h$, the numerator of the IPAR statistic satisfies

$$\begin{aligned}\frac{1}{\sqrt{n}}\sum_{t\in\mathcal{N}} g(\mathbf{d}_t, \theta_h; \widehat{\boldsymbol{\gamma}}(\theta_h)) &= (1, \widehat{\boldsymbol{\gamma}}(\theta_h)')\sqrt{n}h_n(\boldsymbol{\psi}) \\ &\xrightarrow{d} (1, \boldsymbol{\gamma}') \times N(0, \mathbf{S}) = N(0, \omega(\boldsymbol{\gamma})),\end{aligned}$$

which follows from Slutsky's lemma using Lemma S3 and Corollary 2. Further, under H_0, the denominator satisfies

$$\widehat{\omega}(\theta_h, \widehat{\boldsymbol{\gamma}}) = (1, \widehat{\boldsymbol{\gamma}}')\widehat{\mathbf{S}}(\theta_h, \widehat{\boldsymbol{\delta}})(1, \widehat{\boldsymbol{\gamma}}(\theta_h)')' \xrightarrow{p} (1, \boldsymbol{\gamma}')\mathbf{S}(1, \boldsymbol{\gamma}')' = \omega(\boldsymbol{\gamma}),$$

which follows from Assumption R-6 and Corollary 2 applied with $\hat{\boldsymbol{\psi}} = (\bar{\theta}_h, \widehat{\boldsymbol{\delta}}')'$ which is consistent under H_0 and Lemma S1. Combining the results using Slutky's lemma implies that under H_0 we have $\mathrm{AR}_{\mathrm{IP}}(\bar{\theta}_h) = \mathrm{AR}_{\mathrm{IP}}(\theta_h) \xrightarrow{d} N(0, 1)$. □