DISCUSSION PAPER SERIES

DP17433
(v. 2)

THE LONG-RUN EFFECTS OF GOVERNMENT SPENDING

Juan Antolin-Diaz and Paolo Surico

MACROECONOMICS AND GROWTH AND MONETARY ECONOMICS AND FLUCTUATIONS

CEPR

ISSN 0265-8003

THE LONG-RUN EFFECTS OF GOVERNMENT SPENDING

Juan Antolin-Diaz and Paolo Surico

Discussion Paper DP17433
First Published 03 July 2022
This Revision 25 November 2022

Centre for Economic Policy Research
33 Great Sutton Street, London EC1V 0DX, UK
Tel: +44 (0)20 7183 8801
www.cepr.org

THE LONG-RUN EFFECTS OF GOVERNMENT SPENDING

Abstract

Military spending has sizable effects on long-run growth because it shifts the composition of public spending towards R&D. This boosts innovation and private investment in the medium-term, and increases productivity and output at longer horizons. Public R&D expenditure stimulates long-run growth even when it is not associated with war spending. In contrast, the effects of public investment are shorter-lived and the impact of public consumption is modest at most horizons. We reach these conclusions using Bayesian Vector Auto Regressions (BVAR) with up to sixty lags and 125 years of quarterly data for the United States, including newly reconstructed series of government spending broken down into its main categories since 1890.

JEL Classification: E31, E62, O40

Keywords: Government R&D, Long-run, TFP, Innovation, Output multiplier, inflation

Juan Antolin-Diaz - jantolindiaz@london.edu
London Business School

Paolo Surico - psurico@london.edu
London Business School

Acknowledgements
We are grateful to David Abraham, Michele Andreolli, Tim Besley, Olivier Blanchard, Gabriel Chodorow-Reich, Thomas Drechsel, Luca Gambetti, Ethan Ilzetzki, Diego K¨anzig, Emi Nakamura, Elias Papaioannou, Ivan Petrella, Mikkel Plagborg- Møller, Franck Portier, Valerie Ramey, Giovanni Ricco, Juan Rubio-Ramirez, Andrew Scott, Anna Rogantini Picco, Jon Steinsson, ChristianWolf and seminar participants at the Barcelona Summer Forum 2022, the Monetary Economics workshop of the NBER Summer Institute 2022, Rutgers University and the Sveriges Riksbank for very useful comments and suggestions. Paolo Surico gratefully acknowledges financial support from the European Research Council (Grant No. 771976). Address for correspondence: London Business School, Regent's Park, NW1 4SA London, United Kingdom. Email addresses: jantolindiaz@london.edu and psurico@london.edu. The figures of this paper are best appreciated on screen or when printed in colour.

The Long-Run Effects of Government Spending*

Juan Antolin-Diaz
London Business School

Paolo Surico
London Business School and CEPR

November 2022

Abstract

Military spending has sizable effects on long-run growth because it shifts the composition of public spending towards R&D. This boosts innovation and private investment in the medium-term, and increases productivity and output at longer horizons. Public R&D expenditure stimulates long-run growth even when it is not associated with war spending. In contrast, the effects of public investment are shorter-lived and the impact of public consumption is modest at most horizons. We reach these conclusions using Bayesian Vector Auto Regressions (BVAR) with up to sixty lags and 125 years of quarterly data for the United States, including newly reconstructed series of government spending broken down into its main categories since 1890.

Keywords: government R&D, long-run, TFP, innovation, output multiplier, inflation.

JEL Classification Codes: E32, E62, O40.

*We are grateful to David Abraham, Michele Andreolli, Tim Besley, Olivier Blanchard, Gabriel Chodorow-Reich, Thomas Drechsel, Luca Gambetti, Ethan Ilzetzki, Diego Känzig, Emi Nakamura, Elias Papaioannou, Ivan Petrella, Mikkel Plagborg-Møller, Franck Portier, Valerie Ramey, Giovanni Ricco, Juan Rubio-Ramirez, Andrew Scott, Anna Rogantini Picco, Jon Steinsson, Christian Wolf and seminar participants at the Barcelona Summer Forum 2022, the Monetary Economics workshop of the NBER Summer Institute 2022, Rutgers University and the Sveriges Riksbank for very useful comments and suggestions. Paolo Surico gratefully acknowledges financial support from the European Research Council (Grant No. 771976). Address for correspondence: London Business School, Regent's Park, NW1 4SA London, United Kingdom. Email addresses: jantolindiaz@london.edu and psurico@london.edu. The figures of this paper are best appreciated on screen or when printed in colour.

1 Introduction

Can government spending stimulate long-run growth? Large increases in public expenditure —typically associated with defense buildups around wars— have often been credited with the development of new technologies. For instance, the Manhattan project during WWII led to the development of nuclear energy, the establishment of the Defense Advanced Research Projects Agency (DARPA) in the late 1950s is linked to the creation of the internet, and NASA's moon-landing program of the 1960s spurred several advances in aeronautics and satellite technology, such as GPS. Despite this anecdotal evidence, the macroeconomics literature has not yet established a causal link between large government programs and long-run productivity, innovation and growth at the aggregate level.

Using the series of military spending news constructed by Ramey and Zubairy (2018) (which builds on Ramey and Shapiro, 1998; Ramey, 2011b), we find that the effects of an unanticipated increase in defense spending on the macroeconomy are large, and extend well beyond the frequencies typically studied in business-cycle analyses. The output multiplier (i.e. the dollar increase in GDP that results from a dollar increase in government spending) is around one at business-cycle frequencies, but raises significantly above one in the long-run. Total Factor Productivity (TFP), innovation, private consumption and private investment all fall in the short-run, recover in the medium-term and then increase persistently at longer horizons. Interestingly, prices pick up strongly in the first four years after the military spending shock, but stabilize or decline thereafter.

As for the transmission mechanism, we present evidence that military spending affects long-run growth because it shifts the composition of public spending towards R&D. While in the short-run government consumption, investment in equipment and structures, and R&D all increase following military spending news, R&D is the only category that is significantly higher after ten to fifteen years from the initial shock. To tease out any different effects across components, we use an alternative strategy which identifies the shock that maximizes the variance of each spending category within the first year after the shock. We find that permanent increases in output are associated with shocks that expand the share of government spending going to R&D in the long-run.

Finally, we scrutinize our newly identified 'public R&D shock' and show not only that is weakly correlated with war spending but also that its historical evolution aligns well with narrative evidence on large R&D federal spending programs, including the Manhattan project, DARPA, the Moon-landing program and Reagan's 'Star Wars' initiative. Furthermore, we document that an exogenous increase in public R&D leads to long-run responses in output, TFP, innovation and investment that are even larger and more persistent than the effects of military spending on these variables. Our results highlight a new channel through which fiscal policy can support long-run economic growth in peacetime.

Identifying long-run effects requires long, high-quality historical data and empirical methods designed to capture low-frequency correlations. As for the historical data, we have digitized archival statistics and drawn upon narrative evidence to construct new quarterly series of U.S. government spending since 1890, by main categories: consumption expenditure, equipment & infrastructure investment, and R&D. We have also constructed quarterly series for aggregate hours worked, total factor productivity, private investment and consumption, building on existing and unpublished annual data. This allows us to examine the effects of government spending at high-, medium- and long-term frequencies over a period of 125 years spanning major military conflicts and public spending programs, financial crises and recessions, monetary policy and fiscal policy regimes.

As for the empirical method, we rely on Bayesian Vector Autoregressions (BVAR) with a very long lag structure to compute dynamic causal effects. This approach allows us to capture the gradual patterns of technological diffusion after increases in R&D. It also connects us with the debate in empirical macro about the relative merits between VARs and direct single-equation regressions, known as 'local projections' (Jordà, 2005; Kilian and Lütkepohl, 2017; Nakamura and Steinsson, 2018). Recent work has highlighted the intimate connection between the two approaches, and in particular their coincidence up to the lag-order of the VAR (Plagborg-Møller and Wolf, 2021). Moreover, Li, Plagborg-Møller, and Wolf (2021b) highlight the non-trivial bias-variance trade-off inherent to the choice between methods, and the attractiveness of shrinkage estimators in this context. We set the lag order of the VAR equal to 60 quarters, our maximum horizon of interest in the impulse responses, and employ shrinkage to maximize the marginal likelihood of the model (as in Giannone et al., 2015), balancing these statistical considerations. We show that once we allow for the same lag structure and shrinkage, Bayesian LPs and BVARs produce very similar results.

Related Literature. A voluminous empirical literature has studied the macroeconomic effects of government spending on output over the business-cycle. A key challenge is to isolate movements in public expenditure that are exogenous to economic conditions. Leading approaches have used narrative evidence (Ramey and Shapiro, 1998), timing restrictions (Blanchard and Perotti, 2002), sign restrictions (Mountford and Uhlig, 2009) and geographical variation (Nakamura and Steinsson, 2014; Chodorow-Reich, 2019). In two comprehensive reviews, Ramey (2011a, 2019) summarizes the literature and concludes that the short-run government spending multiplier lies between 0.6 and 1.5, across the reviewed papers. Our focus on the long-run is a distinctive feature relative to earlier studies.

An important strand of research focuses on the impact of public spending on productivity. Moretti et al. (2019) and Deleidi and Mazzucato (2021) find that military expenditure fosters private innovation while Gruber and Johnson (2019), Gross and Sampat (2020), Diebolt and Pellier (2020) and Ilzetzki (2022) document the long-lasting effects of the two World Wars on U.S. patenting and productivity. Kantor and Whalley (2022) show that the Space Race with the Soviet Union of the 1960s had long-run effects on manufacturing growth across U.S. counties. Our historical analysis extends these event studies to a much longer sample and horizon, using a different identification; furthermore, it shows that public R&D can stimulate long-run productivity and output even in peacetime.

Our results also speak to the public infrastructure research surveyed by Ramey (2020). Fernald (1999) and Leff Yaffe (2020) find that the U.S. interstate highway programme boosted industry-level productivity, while Donaldson and Hornbeck (2016) and Hornbeck and Rotemberg (2021) estimate that the U.S. national railroad network improved market access. We complement these studies by showing that public investment in equipment & infrastructure tends to have smaller long-run effects than public R&D.

A growing literature, surveyed by Cerra et al. (2022), studies the long-run effects of demand shocks. Comin and Gertler (2006) and Beaudry et al. (2020) lay out models with strong internal propagation mechanisms in which non-technology shocks have effects beyond the business cycle. Benigno and Fornaro (2017) focus on stagnation traps triggered by weak aggregate demand. Jordà et al. (2020) exploit the international finance trilemma to identify the long-run effects of monetary policy. Akcigit et al. (2022) study the impact of income taxes on innovation and researchers' mobility across U.S. states. Cloyne et al. (2022) estimate the long-run responses of R&D, productivity and GDP to corporate and personal tax changes. Our analysis offers a novel evaluation of the long-run effects of government spending on the aggregate economy.

Structure of the paper. In Section 2, we present the VAR and LP specifications, the historical data and the identification strategy. The main findings on output and the fiscal multiplier are reported in Section 3 while, in Section 4, we investigate the transmission mechanism working through the different categories of private and public spending. In Section 5, we estimate the long-run effects of public R&D and contrast them with those triggered by public consumption and investment. In Section 6, we present an extensive range of sensitivity exercises that assess the role of wars, variable transformation, model specification and shock identification. Conclusions are discussed in Section 7. In the Appendix, we provide details on the estimation and present further analyses.

2 Empirical Framework

In this section, we motivate the empirical model and the estimation strategy that we propose, including prior and lag length selection. We then present the historical data for the United States and review the identification of government spending shocks based on the military spending news constructed by Ramey (2011b) (which in turn builds upon Ramey and Shapiro, 1998) and extended back in time by Ramey and Zubairy (2018). We complement their dataset with extended series for business investment, productivity, innovation, consumption, and government spending broken down into its three main categories.

2.1 Model Specification and Estimation

We use a Vector Autoregressive (VAR) model to conduct inference on the effects of government spending on economic activities. The model can be written as:

$$\mathbf{y'}_t \mathbf{A}_0 = \sum_{\ell=1}^{p} \mathbf{y}'_{t-\ell} \mathbf{A}_\ell + \mathbf{c} + \boldsymbol{\varepsilon}'_t \quad \text{for } 1 \leq t \leq T \tag{1}$$

where $\mathbf{y}_t$ is an $n \times 1$ vector of variables, $\boldsymbol{\varepsilon}_t$ is an $n \times 1$ vector of structural shocks, and $\mathbf{A}_\ell$ is an $n \times n$ matrix of parameters for $0 \leq \ell \leq p$ with $\mathbf{A}_0$ invertible. The vector of parameters $\mathbf{c}$ has dimension $1 \times n$, the letter p refers to the lag length, whereas T denotes the sample size. The vector $\boldsymbol{\varepsilon}_t$, conditional on past information and the initial conditions $\mathbf{y}_0, \ldots, \mathbf{y}_{1-p}$, is Gaussian with zero mean and covariance matrix $\mathbf{I}_n$, the $n \times n$ identity matrix.

Denoting $\mathbf{A}'_+ \equiv \left[\mathbf{A}'_1 \ \cdots \ \mathbf{A}'_p \ \mathbf{c}'\right]$, the reduced-form representation implied by Equation (1) is $\mathbf{y'}_t = \sum_{\ell=1}^{p} \mathbf{y}'_{t-\ell}\mathbf{B}_\ell + \mathbf{d} + \mathbf{u}'_t$ for $1 \leq t \leq T$, or more compactly $\mathbf{y}'_t = \mathbf{x}'_t\mathbf{B} + \mathbf{u}'_t$, where $\mathbf{x}'_t = \left[\mathbf{y}'_{t-1}, \ldots, \mathbf{y}'_{t-p}, 1\right]$, $\mathbf{B} = \mathbf{A}_+\mathbf{A}_0^{-1}$, $\mathbf{d} = \mathbf{c}\mathbf{A}_\mathbf{0}^{\mathbf{-1}}$, $\mathbf{u}'_t = \boldsymbol{\varepsilon}'_t\mathbf{A}_0^{-1}$, and $\mathbb{E}\left[\mathbf{u}_t\mathbf{u}'_t\right] = \boldsymbol{\Sigma} = (\mathbf{A}_0\mathbf{A}'_0)^{-1}$. The matrices $\mathbf{B}$ and $\boldsymbol{\Sigma}$ are the reduced-form parameters, while $\mathbf{A}_0$ and $\mathbf{A}_+$ are the structural parameters. Similarly, $\mathbf{u}'_t$ are the reduced-form innovations, while $\boldsymbol{\varepsilon}'_t$ are the structural shocks. The shocks are orthogonal and have an economic interpretation, while the innovations are typically correlated and have no interpretation.

In the VAR setting, impulse-response functions (IRFs), and related objects of interest such as government spending multipliers, forecast error variance decompositions, etc., are computed by recursively iterating on the VAR coefficients, $\boldsymbol{\Theta} = (\mathbf{A}_0, \mathbf{A}_+)$.[1] However, in recent years it has become increasingly popular to compute IRFs using direct regressions of the variable of interest in period $t+h$ on a measure of an identified shock at time t, as well as on control variables. As shown by Jordà (2005), these "local

[1] For instance, given a value $\boldsymbol{\Theta}$ of the structural parameters, the impulse-response of the i-th variable to the j-th structural shock at horizon k corresponds to the element in row i and column j of the matrix $\mathbf{L}_k(\boldsymbol{\Theta})$, defined recursively by

$$\mathbf{L}_0(\boldsymbol{\Theta}) = \left(\mathbf{A}_0^{-1}\right)', \quad \mathbf{L}_k(\boldsymbol{\Theta}) = \sum_{\ell=1}^{k} \left(\mathbf{A}_\ell\mathbf{A}_0^{-1}\right)' \mathbf{L}_{k-\ell}(\boldsymbol{\Theta}), \text{ for } 1 \leq k \leq p,$$

$$\mathbf{L}_k(\boldsymbol{\Theta}) = \sum_{\ell=1}^{p} \left(\mathbf{A}_\ell\mathbf{A}_0^{-1}\right)' \mathbf{L}_{k-\ell}(\boldsymbol{\Theta}), \text{ for } p < k < \infty.$$

projections" can be written as:

$$y_{i,t+h} = \alpha_h + \beta_h \hat{\varepsilon}_t^1 + \boldsymbol{\psi}_h(L)\mathbf{z}_t' + \nu_{t+h} \quad \text{for } h = 0, 1, \ldots, H \tag{2}$$

where $\hat{\varepsilon}_t^1$ is a proxy for the identified shock. For comparability and without loss of generality, we assume that the shock in the local projection (2) corresponds to the first shock in the VAR (1).

There has been considerable debate in the literature about the relative advantages of VAR versus LP estimates of impulse responses. Plagborg-Møller and Wolf (2021), Montiel Olea and Plagborg-Møller (2021), and Li et al. (2021b) clarify important conceptual and practical aspects and conclude that the two approaches estimate the same impulse responses in population. In particular, their estimands approximately coincide up to horizon p (the maximum lag length of the VAR). Furthermore, standard confidence intervals based on lag-augmented local projections have correct asymptotic coverage, uniformly, over the persistence in the data generating process and over a wide range of horizons. Finally, in small-sample applications, a trade-off emerges between the higher bias of low-order VARs and the higher variance of LPs, such that shrinkage estimators —e.g. Bayesian VARs or penalized LPs (Barnichon and Brownlees, 2019)— become attractive. In our context, which features non-stationary variables and cointegrated relationships, Bayesian VARs are an effective tool to address the finite sample bias that characterize autoregressions containing unit roots via priors elicited on the system as a whole (Doan et al., 1984; Sims et al., 1990; Sims, 1993; Sims and Zha, 1998; Giannone et al., 2015, 2019). This compares favourably with single-equation methods like LPs.

Our focus on long-run dynamics requires a careful consideration of the small sample bias-variance trade-off highlighted by Li et al. (2021a). To balance these two considerations, we set the lag length of our baseline VAR to $p = 60$. This choice fulfills our desire to look at horizons well beyond the eight years traditionally associated with business-cycle frequencies and potentially capture long lags in the diffusion of technological advances in response to increases in R&D spending. In Appendix G, we show however that our results do not depend on a specific number of lags as long as this number is large enough.

As for inference, we take a Bayesian approach and apply priors that shrink coefficients towards zero at a rate that exponentially increases with the more distant lags, in the spirit of the "Minnesota" priors of Doan et al. (1984) and Sims (1993). The generous choice of lag length brings the impulse responses of the VAR close to what would have been obtained with lag-augmented LPs, whereas the use of shrinkage allows us to mitigate the increase in variance stemming from the very large number of parameters involved. Moreover, by placing more shrinkage on more distant lags, the Minnesota prior can be viewed as a conservative approach to draw inference on long-run impulse responses: the data needs to speak strongly about the presence of low-frequency effects to counteract the a-priori view that these are absent. Further details on the specification of the prior are given below.

2.2 Prior Specification and Posterior Sampling

We will use a Normal-Inverse Wishart prior over the reduced form parameters, $(\mathbf{B}, \mathbf{\Sigma})$. This family of distributions is conjugate for this class of models and is the standard choice in empirical work due to its computational tractability (see, for instance Uhlig, 2005; Giannone et al., 2015). Denoting $\mathbf{b} = \mathtt{vec}(\mathbf{B})$, the prior distribution is $NIW(\underline{\nu}, \underline{\mathbf{S}}, \underline{\mathbf{b}}, \underline{\mathbf{V}})$ As discussed above, we employ the 'Minnesota' priors proposed by Doan et al. (1984), which shrink the VAR coefficients towards simple univariate specifications. In particular, the degrees of freedom of the prior covariance matrix are set to $\underline{\nu} = n + 2$, with $\underline{\mathbf{S}}$ a diagonal matrix.[2] As for the autoregressive coefficients, the

[2]As common, we set $\underline{\mathbf{S}}_{i,i}$ to the residual variance of a univariate AR(1) estimated on the full sample.

prior has the following mean and variance:

$$\mathbb{E}\left[(\mathbf{B}_\ell)_{i,j}|\mathbf{\Sigma}\right] \qquad \begin{cases} \delta & \text{if } j = 1 \text{ and } \ell = 1 \\ 0 & \text{otherwise} \end{cases} \tag{3}$$

$$\texttt{cov}\left((\mathbf{B}_\ell)_{i,j},(\mathbf{B}_m)_{r,k}|\mathbf{\Sigma}\right) \quad \begin{cases} \lambda^2 \frac{1}{\ell^2} \frac{\Sigma_{i,h}}{\psi_j/((\underline{\nu}-n-1)} & \text{if } j = k \text{ and } \ell = m \\ 0 & \text{otherwise} \end{cases} \tag{4}$$

The parameter δ, which is the mean of the autoregressive coefficient corresponding to the first lag, is set to 1 for trending variables, to 0.9 for stationary but persistent variables, and to 0 for other variables. In addition, because our dataset will contain a mix of stationary and non-stationary variables, we will combine the Minnesota prior with the 'Single Unit Root' prior proposed by Sims (1993) and Sims and Zha (1998). This prior addresses the problem of the excessive explanatory power of initial conditions and deterministic components, which translates into downward bias in the persistence of autoregressive coefficients (see Sims and Uhlig, 1991; Sims, 2000; Jarocinski and Marcet, 2015; Giannone et al., 2019). These combination of priors is standard and widely used in empirical macroeconomics.

As discussed by Del Negro and Schorfheide (2011), among others, the hyperparameter λ controls the overall tightness of the Minnesota prior, whereas a scalar hyperparameter θ controls the tightness of the Dummy Initial Observation prior. The term $\frac{1}{\ell^2}$ implies that more distant lags are shrunk at an exponentially increasing rate towards zero. Therefore, the Minnesota prior penalizes rich large structures and favors models with shorter lags and "smooth" impulse responses. Because of this, the choice of the tightness of the prior becomes especially important for our results about any possible long-run effect. On the one hand, if λ is large, the prior is too lose and the large number of parameters means that the long-run effects will be estimated imprecisely. On the other hand, as $\lambda \to 0$, the long-run effects are dogmatically shrunk towards zero for stationary variables and the data has no chance to speak about the more distant future. In turn, a tighter θ implies a prior that favors unit roots and cointegration in the system as a whole. Giannone et al. (2015) propose a theoretically-grounded methodology to optimally choose the hyperparameters of the prior, based on maximization of the marginal likelihood. Based on this procedure, we select $\lambda = 0.44$ and $\theta = 0.001$ for our baseline estimates, and we will explore the results of tighter or looser choices in detail in Appendix I. The conjugate nature of the prior allows us to sample from the posterior distribution in a straightforward way, using the standard algorithm described in Appendix B.

2.3 Bayesian Local Projections

We also compare the results of our Bayesian VAR to those based on local projections (LP). As we will see, just like for the case of the VAR, augmenting the LPs with a large amount of lags is important to control for long-range predictability in our application. Therefore, Bayesian shrinkage is needed also in the LP to reduce the variance of the estimates given the large number of parameters. To maximize comparability across the two models, we estimate equation (2) with Bayesian methods, implementing a prior on the coefficients for the lags that has the same mean and variance as in equations (3)-(4). It is important to note that while the two approaches will converge to the same results in large samples, the prior acts in a different way in the VAR and the LP. To see this, recall that one can think of both the recursive VAR identification and the lag augmented LP in terms of a two-stage approach, in which military news is first regressed on p lags of the endogenous variables and itself, and then a second stage in which the impulse responses to the first-step regression residuals are calculated, either by iterating on the VAR coefficients, or by direct projection in the LPs. The prior in the VAR applies shrinkage to the coefficients on the lagged control variables, which are then used to calculate the IRFs, therefore implicitly shrinking the latter. The Bayesian LP approach applies shrinkage to the coefficients on the control variables but does not discipline the shape of the IRFs. Moreover, the single-equation LPs is unable to incorporate the Single Unit Root prior that acts on the system as a whole.[3]

Finally, as discussed by Miranda-Agrippino et al. (2021), the Gaussian likelihood of model (2) is misspecified due to the presence of serial correlation in the residuals at $h > 1$. We follow these authors in interpreting it instead as the likelihood of a misspecified auxiliary model. However, unlike Miranda-Agrippino et al. (2021), we rely on the analysis in Montiel Olea and Plagborg-Møller (2021), who show that lag-augmentation in LPs, as we do here, obviates the need to adjust the covariance matrix for the presence of unmodeled serial correlation. Accordingly, in our baseline estimates, we report standard Bayesian posterior density intervals.[4]

2.4 Data and Identification

Our starting point is the seven variables dataset in Ramey and Zubairy (2018), which spans the sample 1890Q1 to 2015Q4: the present discounted value of military news (Ramey, 2011b), government spending, real GDP, the log GDP deflator, the short-term interest rate, the surplus-to-GDP ratio and the Debt-to-GDP ratio. In drawing inference about the long-run, our baseline approach is to express non-stationary variables in log-levels. Sims et al. (1990) show that, even in the presence of cointegration, this specification leads to consistent estimates. When computing government spending multipliers, however, the log-level specification requires scaling the impulse responses by the steady state value of Y/G. As discussed by Owyang et al. (2013) and Ramey

[3]Moreover, given that the LP represents only one equation of the VAR, the prior we impose is in fact an independent Normal-Inverse Wishart rather than the standard conjugate Normal-Inverse Wishart we use in the VAR. Our approach thus differs from the one proposed by Miranda-Agrippino and Ricco (2021), who center the priors for the LP coefficients around the IRFs produced by a low-order VAR. In our application, there is no particular reason to believe a-priori that a low order VAR is a reasonable approximation of the data, especially in the long-run.

[4]For completeness, we have verified that adjusting for residuals serial correlation produces less accurate estimates but does not overturn the significance of our results. More specifically, the estimated short-run and long-run multipliers are still significant and statistically different one from the other at the 90% level.

and Zubairy (2018), multiplier estimates can be quite sensitive to this conversion factor measured from historical averages. Accordingly, we also compute output multipliers from alternative models in which GDP and government spending are scaled either by GDP in the previous quarter (as in Barro and Redlick, 2011), or by the measure of GDP trend proposed by Ramey and Zubairy (2018): a sixth-degree polynomial for log GDP, 1889Q1-2015Q4, excluding 1930Q1–1946Q4. The baseline transformation includes an intercept, and thus implicitly controls for a linear trend; the second transformation is akin to estimating the VAR in differences, hence removing a stochastic trend; the third transformation has the disadvantage of purging low-frequency movements in potential output that may be particularly important to account for the long-run effects of government spending: we include it mostly for comparability with the estimates in Ramey and Zubairy (2018).

We extend the baseline data along several dimensions. First, we construct new quarterly series of private consumption and investment. We obtain unpublished annual estimates of investment since 1901 by the Bureau of Economic Analysis. Before that, we rely on the Macrohistory Database of Jordà et al. (2017), which also offers a measure of annual private consumption since 1890. We interpolate these series to quarterly frequency using the consumption and investment series from NIPA (after 1947), Gordon (2007) (between 1919 and 1940) and real GDP (before 1919 and from 1941 to 1946). Second, we construct quarterly measures of hours worked and Total Factor Productivity (TFP). The annual hours and productivity series comes from Bergeaud et al. (2016). We adjust TFP for capital and labor utilization following Imbs (1999). We interpolate this using the quarterly series of adjusted-TFP in Fernald (2012) (after 1947) and real GDP (before 1947). The data on patents are by IFI CLAIMS Patent Services via Google Patents Public Data.

In addition, we construct new historical series of public consumption and investment, distinguishing between expenditure in Equipment and Infrastructure

(E&I) and in Research and Development (R&D). Official NIPA estimates start in 1929. We reconstruct the series of public investment and its components for the period 1890-1929 by digitizing detailed government outlays data from both the *Historical Statistics of the United States* (Census, 1949) and the annual *Statistical Abstracts* published by the census. We rely on the narrative evidence in Bush (1954) and Dupree (1986) to classify investment into E&I and R&D. Finally, we interpolate the resulting annual series using quarterly government spending, and back out public consumption as residual. Further details are provided in Appendix A.

In all cases, when moving from annual to quarterly frequency, we use the method by Chow and Lin (1971). It is worth emphasizing that the impulse responses at long horizons, which is the primary focus of our analysis, depend mainly on the low-frequency properties of the data, which in turn are pinned down by the properties of the annual series. With the exception of the reconstructed government spending series, these annual series are mostly available from existing sources, which we take at face value. The interpolation from annual to quarterly frequency is likely to affect mostly the high-frequency properties of the data (i.e. within the year) and, as such, the specific method or the series used for interpolation is unlikely to have an effect on the estimated IRFs at longer horizons.

To identify the structural parameters of the VAR, we follow the approach labeled as "internal instruments" by Plagborg-Møller and Wolf (2021), and also used by Ramey (2011b). This approach includes the instrumental variable (in our case the military spending news series) in the VAR and identifies the shock of interest by ordering the instrument first in a Cholesky decomposition. As Plagborg-Møller and Wolf (2021) point out, this approach yields valid impulse response estimates even if the shock of interest is non-invertible or if the instrumental variable is contaminated with measurement error that is unrelated to the shock of interest.[5]

[5]Furthermore, the use of a quarterly VAR(60) featuring both the instrument and the endogenous variables of interest as well as the focus on horizons up to 60 quarters ensure that our set-up meets the conditions for consistency

3 The Effects of Military Spending

In this section, we report our main results, which are based on the quarterly VAR described in the previous section using sixty lags. We begin by analyzing the impulse responses to a military spending shock and then move to the estimates of the (present value) output multipliers across forecast horizons, up to sixty quarters. In the next section, we present the results of an extended VAR, where we add newly constructed time series of investment, productivity, innovation, government R&D, and the three main components of government spending since 1890Q1 to shed light on the transmission mechanism of military shocks.

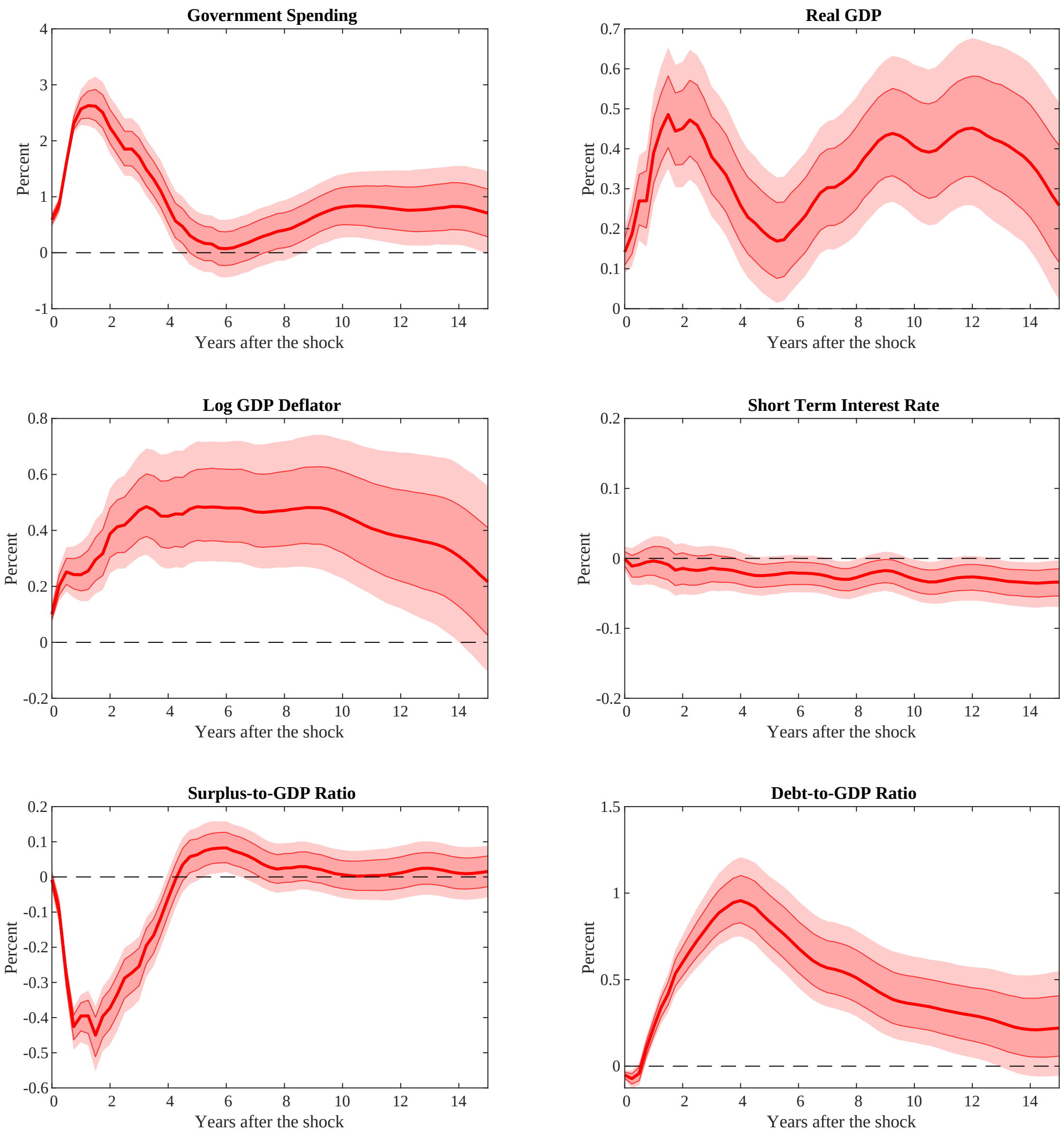
3.1 Impulse Response Analysis

A simple way to summarize the estimates of a VAR is to report impulse responses of the endogenous variables to the identified shock of interest. We select a forecast horizon of 60 quarters to match the number of lags chosen in the estimated VAR(60) and report point-wise 68% and 90% posterior credible sets (as shaded areas). For ease of interpretation, the military spending news shock is normalized so as to increase government spending by 1% of GDP over the first year after the shock. The top row of Figure 1 presents the responses of government spending and real GDP, the middle row refers to the log of the GDP deflator and the short-term nominal interest rate whereas the bottom row focuses on the government balance sheet: fiscal deficit and public debt, both expressed as a share of GDP.

The main findings from our VAR(60) can be summarized as follows. During the first four years after the shock, government spending increases sharply and then reverts to zero, triggering an equally persistent increase in GDP, a notable fiscal deterioration with government debt peaking around 1.5% of GDP, and a significant price spike above

and efficiency of the impulse response estimates provided by Baek and Leeb (2021).

Figure 1: IMPULSE RESPONSES TO MILITARY NEWS SHOCK



Notes. The impulse responses are based on an estimated VAR with sixty lags of military spending news, government spending, real per-capita GDP, GDP deflator, short-term interest rate, fiscal deficit to GDP ratio and government debt to GDP ratio. Government spending, GDP and the GDP deflator enter the VAR in log-levels. Military spending news is ordered first in the Cholesky factorization. The size of the shock is normalized such as to increase government spending by 1 percent of GDP over the first year after the shock, based on a median G/Y ratio of 19%. The darker (lighter) shaded areas represent the central 68% (90%) high posterior density (HPD) intervals. The darker solid lines are the median estimates. Results are based on 5000 posterior draws.

0.8% (or 0.2% inflation per year). At frequencies between 2 and 8 years, government spending falls back to its initial level, causing a short-lived decline in output. This is associated with a switch towards fiscal surplus that contributes to revert the path of the debt-to-GDP ratio.[6]

In the long-run, conventionally defined as frequencies beyond eight years, the response of government spending becomes significant again but its peak is now a fraction of what was at shorter horizon. The fiscal surplus is no longer statistically different from zero and public debt slowly returns to pre-shock levels. In contrast, GDP witnesses a second boom that is not only as large in magnitude as the first peak but appears more persistent in duration. By the end of the forecast horizon, prices and, to a lesser extent, output move back toward their initial levels, while the effects on the short-term nominal interest rate are negligible throughout.[7]

For completeness, we report the Forecast Error Variance Decomposition (FEVD) in Appendix C. This reveals that the military spending news shock explains between 30 and 40% of the variance of government spending at business cycle frequencies, and around 20% of long-run fluctuations. Government spending appears to account for a nontrivial fraction of the variance of real GDP and the price level, at about 15% and 18%, respectively. This is consistent with the evidence in Rossi and Zubairy (2011) of an important role for fiscal policy in explaining U.S. macroeconomic fluctuations over the medium-term.

In summary, we estimate significant long-run effects of government spending on both output and prices. Unlike the short-run dynamics where the movements in

[6]The sequence of fiscal surpluses in Figure 1 associated with the government spending contractions between year 4 and 10 are notably smaller that the fiscal deficits triggered by the initial government spending expansion. This suggests that the (second wave of) GDP response plays a significant role in reducing the debt-to-GDP ratio to pre-shock levels, consistent with the evidence in Hall and Sargent (2011).

[7]Using the yield on 10-year government bonds instead of the short-term rate in the VAR produces very similar findings. This is shown in Appendix D. As noted by Meltzer (2004), until the Treasury-Fed accord of 1951, the Fed pegged interest rates at a low level to facilitate the financing of government debt during wartime. Friedman and Schwartz (1963) argue that the Fed choice of not controlling the growth of the monetary base over this period contributed to fueling inflation. These historical accounts are consistent with the responses of prices and interest rates in Figure 1 and Figure D.1, respectively.

government spending appear larger than the response of GDP, the lower frequency estimates suggest a large long-run multiplier as the effects on output are associated with far smaller changes in government spending at longer horizons. In the next part of this section, we corroborate this conjecture by formally computing the multiplier across forecast horizons.

3.2 The Government Spending Multiplier in the Short- and Long-Run

In the previous section, we have estimated a larger (smaller) output response at longer (shorter) horizons relative to the smaller (larger) lower-frequency (higher-frequency) movements in government spending. In this section, we formally quantify these relative effects by computing the fiscal multiplier of government spending on output across forecast horizons. This is interesting for at least two reasons. First, government spending may have different effects at different horizons and comparing the multipliers at high-, business-cycle and low-frequencies within the same estimated model can help shed light on this issue. Second, as noted by Ramey (2019), different studies often compute the multiplier at different horizons and reporting how the estimates of this statistics vary with the forecast horizon may help reconcile seemingly conflicting findings in the literature.

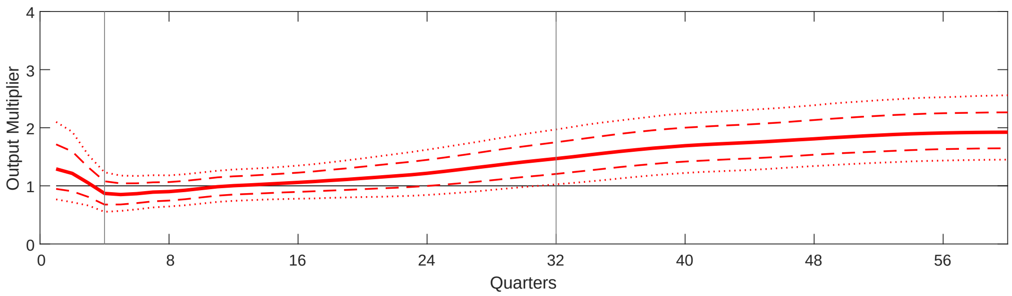
In line with earlier work, we define the output multiplier for each horizon h as the ratio between the cumulative impulse response of real GDP to military spending news up to horizon h and the cumulative impulse response of government spending to the same shock over the same horizon. Following Mountford and Uhlig (2009), we use the sample average nominal interest rate to discount the estimates between one and h quarters ahead. In Figure 2, we display the present value multiplier for each horizon between $h = 0$ (i.e. the impact multiplier) and $h = 60$ (i.e. the long-run multiplier). Panel (a) refers to the specification in log-levels and uses the historical median of $G/Y = 19\%$ to transform the estimated elasticities into multipliers. Panel (b) refers to the specification in which output and government spending are both scaled by Y_{t-1}. Panel (c) is based on a model where both government spending and real GDP are scaled by potential output, as defined by Ramey and Zubairy (2018). The latter two strategies provide direct estimates of the multipliers and do not rely on the government spending-output ratio.

The estimates in Figure 2 reveals that the government spending multiplier on output is, on impact, about 1.35, with most of the distribution mass above one. After the first year, however, the output multiplier decreases to values around 1 according to the model in which variables enter in log-levels, and below 1 in the specification where variables are expressed relative to lagged GDP or potential output, consistent with the evidence in Hall (2009), Barro and Redlick (2011) and Ramey and Zubairy (2018). These estimates are relatively stable over the following three to five years before growing with the forecast horizon. The posterior median of the multiplier takes values above one at frequencies beyond thirty-two quarters and peaks at the significantly larger value of 2 in the forecasts fifteen years ahead. Interestingly, despite very similar median estimates, both the log-level and the previous-quarter-GDP specifications lead to more accurate inference about the long-run multiplier than the one that removes potential output.

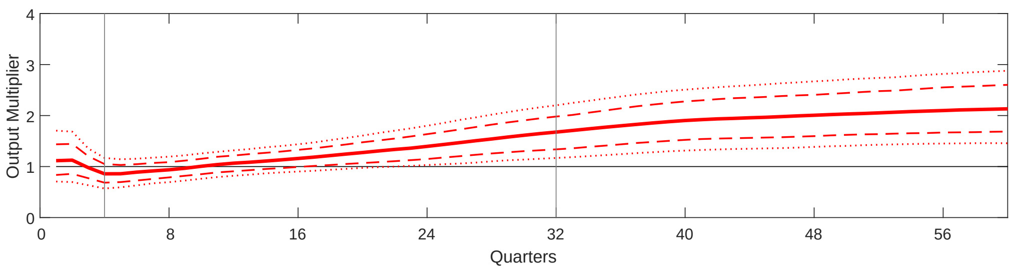
In summary, the findings of this section suggest two main conclusions about the effects of government spending on output. First, on impact and at business-cycle frequencies (i.e. from 6 to 32 quarters) the multipliers span the range of point estimates available in the fiscal policy literature, between 0.6 and 1.5, thereby offering a possible reconciliation of apparently conflicting results in earlier empirical macro studies. Second, while the multipliers at business-cycle frequencies tend to exhibit values below or around one, the multipliers at low-frequencies (i.e. beyond 32 quarters) display much larger values and eventually exceed one significantly in the long-run.

Figure 2: THE GOVERNMENT SPENDING MULTIPLIER ACROSS HORIZONS

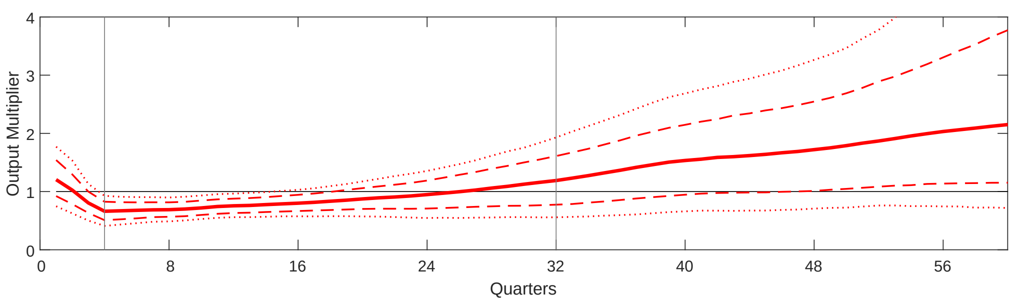
(a) Variables in Log-levels



(b) Variables scaled by previous-quarter GDP



(c) Variables scaled by potential GDP



Notes. The present value multiplier at each horizon h is computed as the ratio of the integral up to horizon h of the output response and the integral up to horizon h of government spending response to a military spending news shock, discounted using the steady-state interest rate. The estimates are based on VARs with sixty lags. In the top panel, government spending and output enter the VAR in log-levels and the multipliers are obtained using the elasticity formula and the historical median G/Y ratio of 19%. In the middle panel, output and government spending are both divided by Y_{t-1}. In the bottom panel, they are expressed in percent of potential output as defined in Ramey and Zubairy (2018). The broken (dotted) lines represent the central 68% (90%) HPD interval. The solid line stands for the median estimate. Results are based on 5000 posterior draws, discarding explosive roots in the stationary specifications.

4 Inspecting the Mechanism

In the previous section, we uncovered a significant long-run output response to a military spending shock. To shed light on the transmission mechanism of this shock, in this section we look at the effects of government spending shocks on private sector outcomes and public spending categories.

To mitigate the curse of dimensionality, in each specification, we augment our baseline VAR(60) with at most two variables at a time, which also enter the model in log-levels. For the private sector models, we consider the following three pairs of additions: (i) labour productivity and hours worked (which substitute for GDP); (ii) total factor productivity adjusted for capital utilization and patents; (iii) private consumption and private investment. For the public sector specifications, we add in turn: (iv) public consumption expenditure, (v) public investment in Equipment & Structure and (vi) public expenditure in R&D.

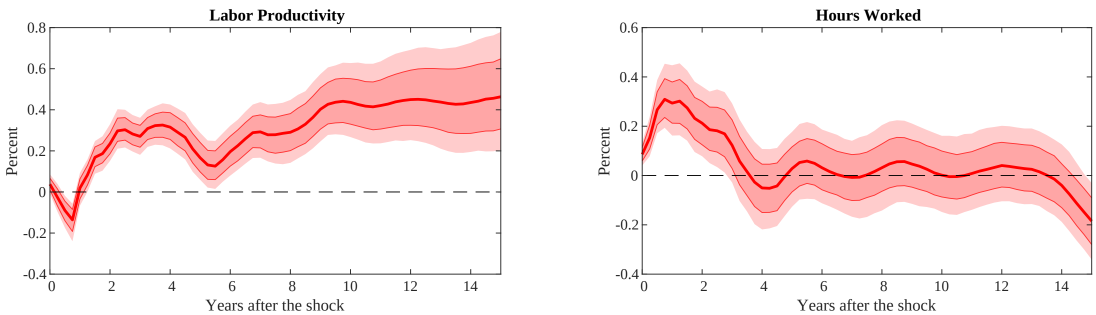
4.1 Private Sector

In Figure 3, we report the posterior credible sets around the responses to a government spending shock, based on the three extended VARs(60) for the private sector described above. The top row focuses on the specification with labour productivity and hours worked, the middle row refers to the model with adjusted TFP and innovation (as measured by patents), and the bottom row reports the estimates for consumption and investment changes.

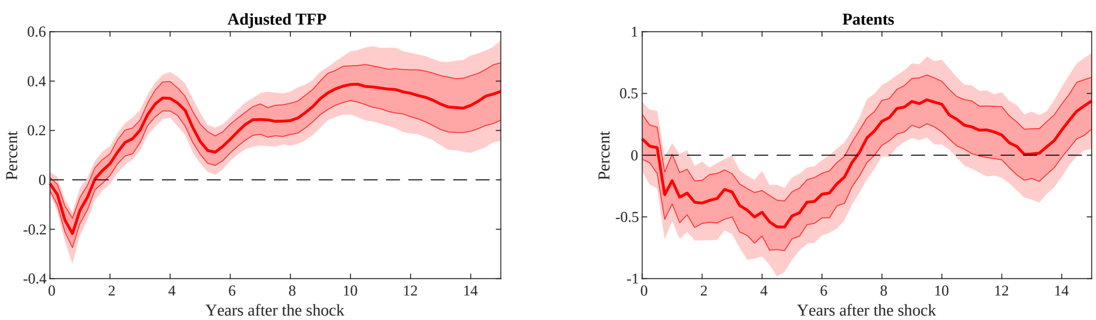
Several results emerge from Figure 3. First, after a short-lived contraction in the top row, labour productivity experiences a sustained increase, which peaks significantly at the end of the forecast horizon. In contrast, hours worked rise on impact, peak in their first year (consistent with the short-term decline in labour productivity) and record small and insignificant changes from three years after the shock onwards.

Figure 3: Effects of Military Shocks on Private Sector Outcomes

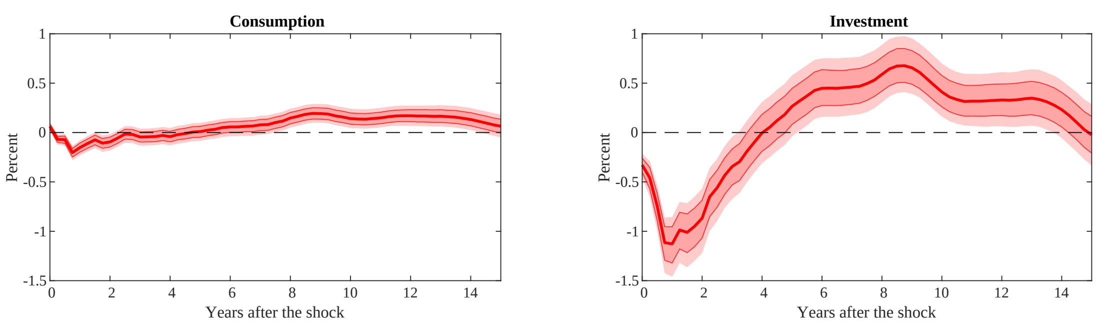
(a) Labor Productivity and Hours



(b) Total Factor Productivity (TFP) and Patents



(c) Private Consumption and Private Investment



Notes. The impulse responses are based on an estimated VAR with sixty lags of military spending news, government spending, real GDP, GDP deflator, short-term interest rate, fiscal deficit to GDP ratio, government debt to GDP ratio. Government spending, GDP and the GDP deflator enter the VAR in log-levels. Military spending news is ordered first in the Cholesky factorization. The size of the shock is normalized such as to increase total government spending by 1 percent of GDP over the first year after the shock. The darker (lighter) shaded area represents the central 68% (90%) HPD band. The darker solid line stands for the median estimates. Results are based on 5000 posterior draws.

The second row reveals that the effects of government spending shocks on labour productivity are mirrored by the dynamics of adjusted TFP. The latter displays a very significant and persistent response, which is only a touch smaller in the long-run than its labour-only counterpart. In contrast, the right panel makes clear that patents are crowded out in the first few years after the shock; their response, however, turns positive and significant in the medium- to long-term, consistent with the evidence by Diebolt and Pellier (2020).

Finally, the bottom row shows that an increase in public spending crowds out private consumption and private investment in the short-run, as reported also by Ramey (2011b). Four years after the shock, however, both types of expenditure increase significantly, peaking after about 9 years and returning to their pre-shock trends by the end of the forecast horizon. The magnitude of the investment response is about five times the size of the consumption effect, possibly reflecting its more volatile nature and smaller GDP share.

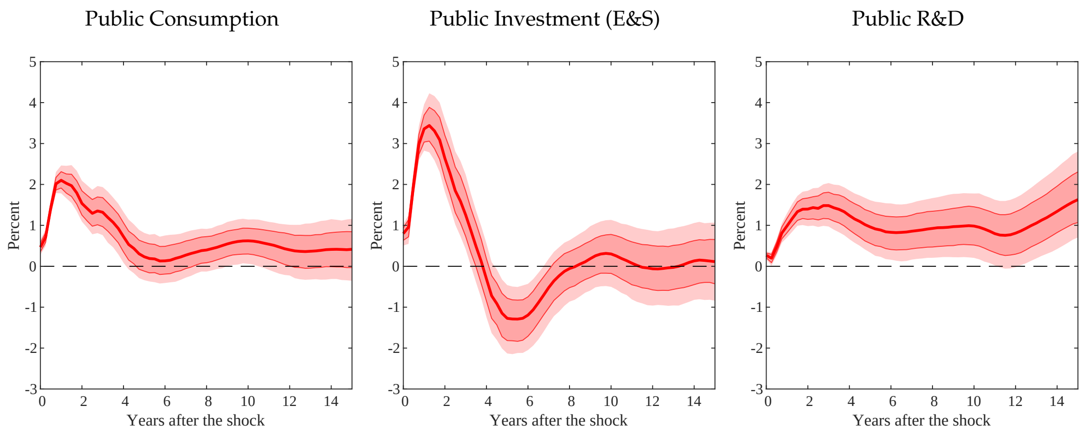
In summary, government spending causes a short-lived rise in hours worked and a temporary crowding out of innovation, private investment and consumption. In the medium- to long-run, however, they all experience significant and sustained increases, which feed into large and very persistent effects on labour and total factor productivity.

4.2 Public Sector

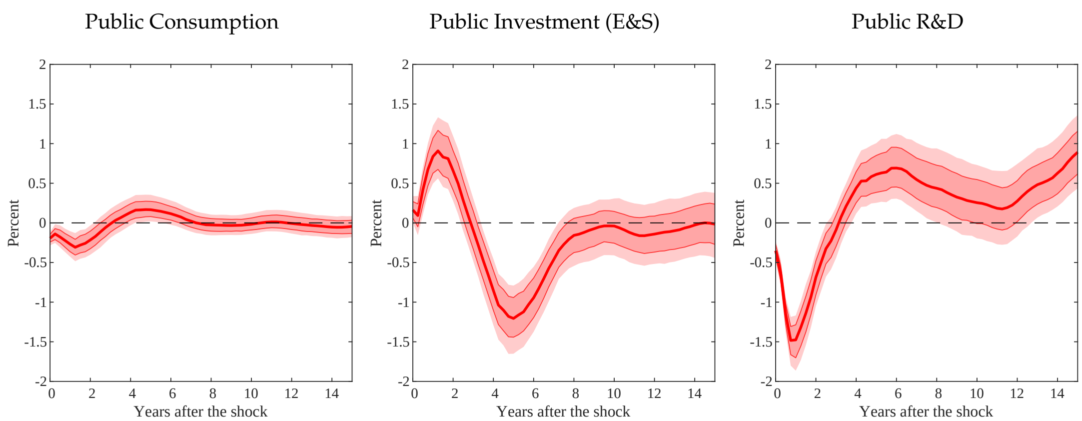
The findings in the previous section are consistent with an important role played by private investment, innovation and productivity in shaping the long-run response of output to a government spending shock. In this section, we ask whether the particular composition of public spending triggered by a defense budget increase may also play a role. To this end, we run three separate specifications in which we augment the baseline VAR(60) of Section 3 with our newly constructed historical series of public consumption, government investment and public R&D, respectively.

Figure 4: Effects of Military Shocks on Public Spending Components

(a) Responses of Public Spending Components



(b) Responses as a Share of Total Government Spending



Notes. The impulse responses are based on an estimated VAR with sixty lags of military spending news, real government spending per capita, real GDP per capita, GDP deflator, short-term interest rate, fiscal deficit to GDP ratio, and government debt to GDP ratio. In each of the columns, real public consumption per capita, real public non-R&D investment per capita and real public R&D expenditure per capita, respectively, are added in turn to the VAR. Each government spending category, total government spending, output and the GDP deflator enter the VAR in log-levels. Military spending news is ordered first in the Cholesky factorization. The size of the shock is normalized such as to increase total government spending by 1 percent of GDP over the first year after the shock. The top (bottom) panel refers to the response of each public spending category in log-level (as share of total government spending). The darker (lighter) shaded area represents the central 68% (90%) HPD band. The darker solid line stands for the median estimates. Results are based on 5000 posterior draws.

The results of these public-sector-augmented VARs(60) are reported in Figure 4. The first row depicts the response of the *log-level* of each category to the military spending news while the second row reports the response of each category as a *share* of total government spending. The top panels reveal three main findings. First, the responses of these three public spending categories are highly correlated: military spending triggers a joint increase in public consumption, investment and R&D.[8] Second, public investment is the category that responds most in the short-run. Third, government R&D expenditure is the only component that displays a large and significant long-run response.

To appreciate the relative contribution of each category, in the second row of Figure 4, we look at the responses of public consumption, public investment and public R&D as *shares* of total government spending. Given the data are in logarithms, these are computed as the difference between the impulse response of each spending category and the impulse response of total government spending at each horizon.[9] Three results stand out from this exercise. First, following a military spending shock, there are little movements in the consumption share, except for two small drops at the beginning and at the end of the forecast horizon. Second, in the short-run, the composition of public spending shifts significantly towards investment and away from R&D. Third, in sharp contrast, the medium- and long-run witness a significant increase in the share of public R&D, which is offset by a decline in the public investment share and, to a lesser extent, in the consumption share.

One interpretation of the responses of the different public spending categories is that public investment in E&S plays a far more important role in explaining the

[8]We interpret this finding as a cautionary note against counterfactual exercises that try to isolate the effects of a specific public spending category by setting to zero at all times the responses of all other components of the government budget. In the context of military spending (and possibly also of other large public programs), this 'counterfactual' mix is actually well outside the distribution of historical combinations of public spending components.

[9]Over our long sample, consumption, investment in E&S and R&D expenditure account, on average, for about 77%, 20% and 3% of total government spending. During the post-WWII period, the average share of public R&D has increased to around 5%, offset almost entirely by a decline in the share of public investment.

short-run effects of government spending whereas public R&D expenditure plays a far more important role in accounting for the *long-run* effects of government spending. In the next section, we will corroborate this interpretation by identifying the dynamic effects of exogenous movements in each public spending category.

5 What Drives the Long-Run Effects of Government Spending?

In the previous section, we have shown that military spending has very persistent effects: (i) on public R&D (but not on public consumption or investment), and (ii) on productivity and output. In this section, we ask whether changes in R&D expenditure may be driving the long-run effects of government spending. The ideal experiment would consist in 'shocking' public R&D while keeping fixed both public consumption and public investment. But this has virtually never happened in our long historical sample, as government spending typically involves a simultaneous expansion in all three categories.[10] In so far as the correlation is not perfect, however, we can use a statistical approach to tease out the effects of each public spending category.

5.1 Identifying a Public R&D Shock

Our starting point is the observation that, historically, the major shifts in public R&D spending have been unrelated to business-cycle conditions. In Appendix E, we discuss the narrative evidence around large public R&D programs and argue that, over our long sample, these have been, in fact, motivated by military rivalries (with Germany until WWII and the Soviet Union afterwards), scientific progress and the ideological priorities of the different administrations, rather than by the endogenous policy response to the state of the U.S. economy.

[10] Interestingly, the evidence in Figure 4 reveals that military spending comes close to an ideal (long-run) experiment, as it is associated with a significant long-run response of public R&D but very small and insignificant long-run responses of public consumption and investment. However the short-run dynamics are very different.

In addition, the timing and implementation lag associated with large public R&D programs extend well beyond the business-cycle frequencies or the terms of office of the different administrations. These considerations suggest that, after controlling for the lags of other macro variables, innovations to public R&D expenditure may be regarded as exogenous to current or prospective economic conditions, in the spirit of the short-run restrictions on government spending proposed by Blanchard and Perotti (2002) or the narrative identification for income tax changes pioneered by Romer and Romer (2010).

In practice, we drop military spending news from the model, and add public R&D, patents and TFP. We then identify exogenous changes in public R&D by searching for the shock that explains the maximum share of the public R&D innovation variance over the first year, following Uhlig (2003).[11] We focus on the first year, rather than the first quarter, because much of our historical data have been interpolated from annual series and the interpolation method might spuriously affect some of the high-frequencies correlations. As discussed below, we obtain very similar results if we focus instead on lower frequencies.

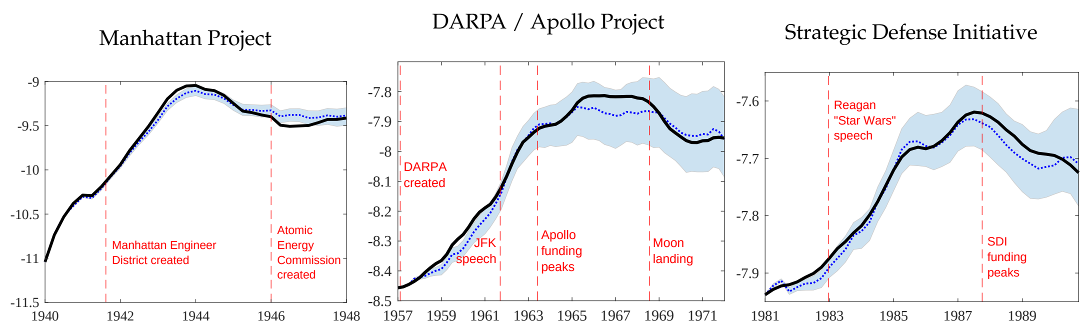
Before presenting the impulse response analysis, we find it useful to verify whether our newly identified shock can match the historical evolution of large federal R&D programs, as discussed in Appendix E. In Panel (a) of Figure 5, we present the historical decomposition of public R&D around three key historical events: (i) the Manhattan project, from its establishment in 1941 to its dissolution with the foundation of the Atomic Energy Commission in 1946, (ii) the creation of DARPA in 1958 and the Moon-landing project from 1961 to 1969, and (iii) Reagan's Strategic Defense Initiative from 1983 to 1987. In each sub-panel, the solid black line represents the historical increase in public R&D while the dotted blue line, and associated 68% posterior bands, refers to the

[11] The 'max-share' method generalizes to a desired frequency the well-known Cholesky decomposition. The latter imposes the far more restrictive restriction that the identified shock explains the entirety of the variance of the variable of interest on impact. The 'max-share' method has been shown to be more robust than the Cholesky factorization in a variety of empirical settings (see, e.g. Kurmann and Otrok, 2013; Francis et al., 2014).

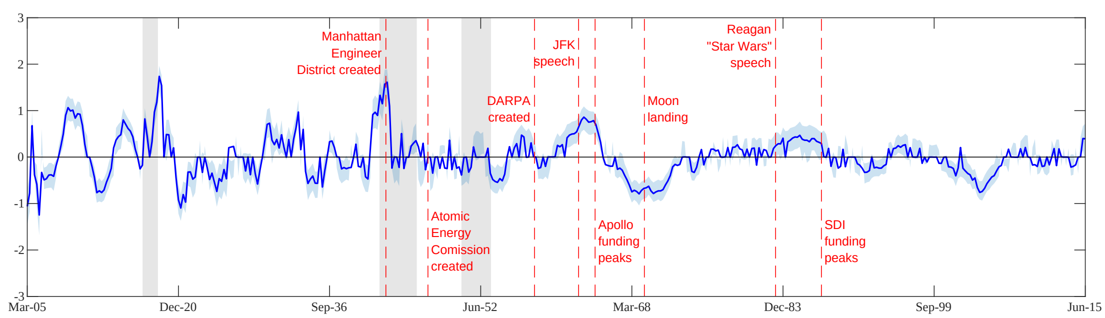
part explained by our public R&D shock. In all cases, the movements in government R&D attributed to the shock aligns very closely with the actual increases around the three events. We interpret this as suggestive evidence that our shock captures the exogenous nature of military or ideologically driven surges in public R&D.[12]

Figure 5: Historical Analysis of Public R&D and Public R&D Shocks

(a) Historical Decomposition of Public R&D Expenditure Around Key Events



(b) Time Series of Public R&D Shocks (eight quarter moving-average)



Notes. Panel (a) plots the historical decomposition of the public R&D series around three key historical events: (i) the Manhattan project, (ii) DARPA and the Apollo program, (iii) the Strategic Defense Initiative. In each of the sub-panels, the solid black line is the historical increase in real per capita R&D spending by the government. The dotted blue line, and associated 68% posterior bands, show the part of the increase in R&D that can be explained by the effects of the exogenous public R&D shock identified using the max-share method at the one-year forecast horizon. Panel (b) plots the history of identified public R&D together with 68% posterior bands. To facilitate visualization, the shock is plotted as an 8-quarter moving average. Shaded areas represent major wars.

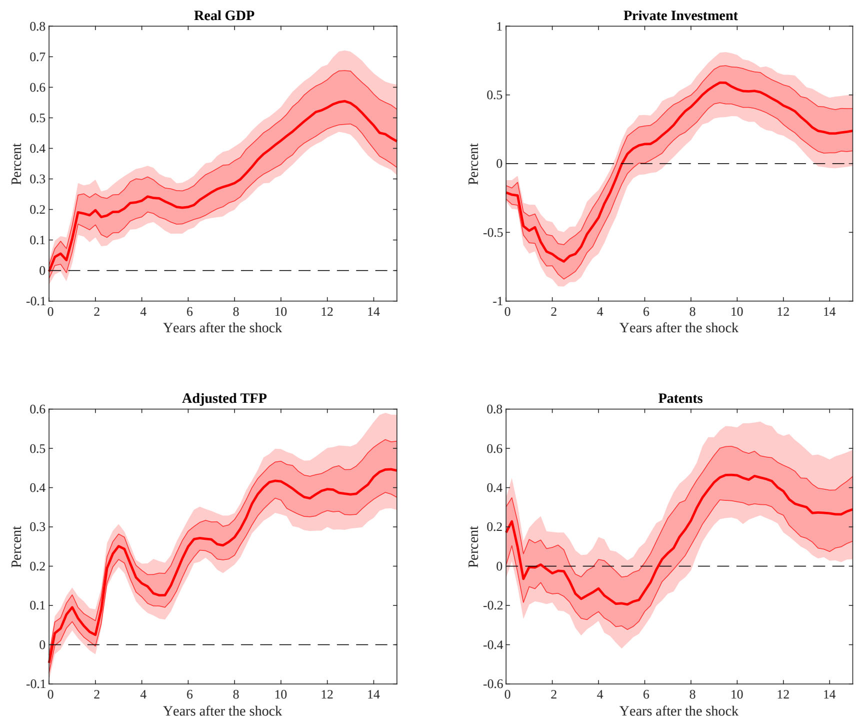
[12] In Appendix F, we show that —in sharp contrast to the results in this section— the military spending shocks cannot explain the lion share of movements in public R&D expenditure around these three key historical events.

In Panel (b) of Figure 5, we report the time series of the identified public R&D shock, together with 68% posterior bands. The shock is plotted as an eight quarter moving-average. Two findings are worth noting. First, there are clusters of positive shocks around the three major public R&D programs, denoted by the vertical dashed lines. Similarly, a cluster of negative shocks is visible around the wind down of the Apollo project. Second, the timing of these programs does not always coincide with major wars, marked as shaded areas in Figure 5. For instance, while WWI and WWII led to large increases in defense-related R&D, the Korean war did not. In other words, the R&D shock seems distinct from the military news shock. Indeed, their correlation is only 0.17.

In Figure 6, we report the impulse responses to the public R&D shock. In keeping with previous charts, the shock is scaled so as to increase total government spending by 1% of GDP over the first year. At shorter horizons, the increase in output is much more muted than for the military spending shock and does not display any hump shape. At longer horizons, however, the size and persistence of the effects on output become much larger, with a peak after about 12 years. The responses of private investment, TFP and patents display dynamics that are qualitatively similar to the ones produced by the military spending shock. For all variables, however, the public R&D shock causes a smaller short-run crowding out effect, which is even no longer statistically significant for patents.

In summary, our identified public R&D shock aligns very well with the narrative account around large public R&D programs in the economic history of the United States. Furthermore, we estimate that the long-run effects on output, productivity, private investment and innovation generated by an exogenous increase in public R&D are very similar to, if not stronger than, those triggered by the military spending shock, despite a modest correlation between the two of only 0.17.

Figure 6: IMPULSE RESPONSES TO PUBLIC R&D SHOCK



Notes. The impulse responses are based on an estimated VAR with sixty lags of real public R&D per capita, real total government spending per capita, real GDP per capita, GDP deflator, short-term interest rate, fiscal deficit to GDP ratio, government debt to GDP ratio, real private investment per capita, total factor productivity and patents. Public R&D, total government spending, GDP and the GDP deflator enter the VAR in log-levels. The public R&D shock is identified using the max-share method at the one-year forecast horizon. The size of the shock is normalized such as to increase government spending by 1 percent of GDP over the first year after the shock, based on a median G/Y ratio of 19%. The darker (lighter) shaded areas represent the central 68% (90%) high posterior density (HPD) intervals. The darker solid lines are the median estimates. Results are based on 5000 posterior draws.

5.2 The Role of Public Consumption and Public Investment

In the previous section, we have identified the effects of public R&D on the economy by searching for the shock that explains most of the public R&D variance during the first year after the shock. For sake of comparability, in this section we adopt an identical strategy for the other two components of government spending, and isolate the innovations to public consumption and public investment, respectively, that maximize the share of the forecast error variance of each spending component at the one year horizon.

It is worth noting, however, that both shocks are in fact associated with significant contemporaneous movements in public R&D, which makes it hard to interpret them as 'pure' innovations (i.e. everything else equal) to public consumption and investment. On the other hand, each innovation brings about movements in public R&D of different sizes, and therefore we can exploit this variation to assess whether the strength of the output responses is correlated with the relative strength or "intensity" of the changes in public R&D.

In Panel (a) of Figure 7, we report the output responses to shocks that maximize the one-year ahead error variance of public consumption, public investment, and public R&D, respectively.[13] Across all specifications, the shocks are normalized such that total government spending moves by 1% of GDP over the first year; hence, the three columns can also be thought of as varying the intensity of each spending category. The main finding is that the 'consumption-intensive' shock leads to a smaller output response than the 'investment-intensive' shock, which in turn triggers a smaller response than the 'R&D-intensive' shock.

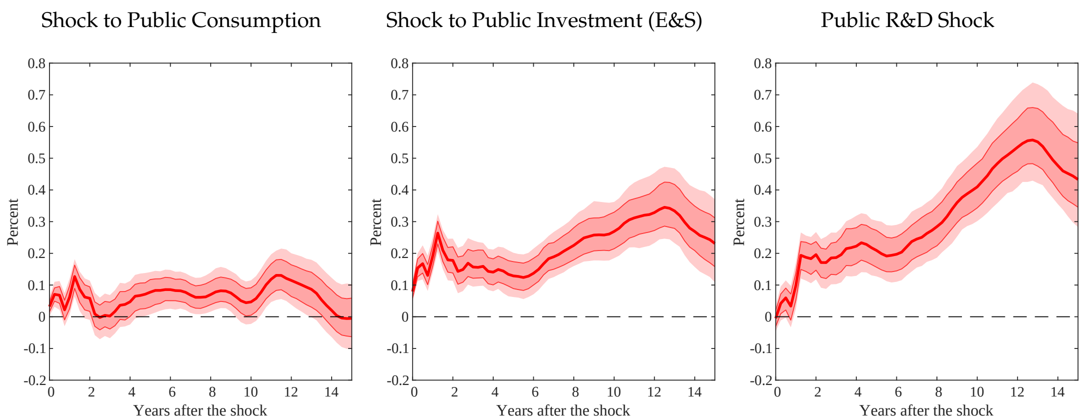
To explore these results further, in Panel (b), we look at the effects of each shock on public R&D as a share of total government spending. The shock to public consumption leads to a drop in the R&D share in the short-run and a muted response thereafter. This is associated with modest long-run effects on output in Panel (a). The shock to public investment in the middle column also leads to a short-run decline in the R&D share; this is however quickly reversed and then replaced by a persistent increase, which is mirrored by the output response in the top row.[14] On the other hand, the R&D shock is characterized by both the strongest R&D share response *and* the strongest long-run output response.

[13] The chart in the third column of Figure 7 is therefore a repetition of the top-left panel in Figure 6.

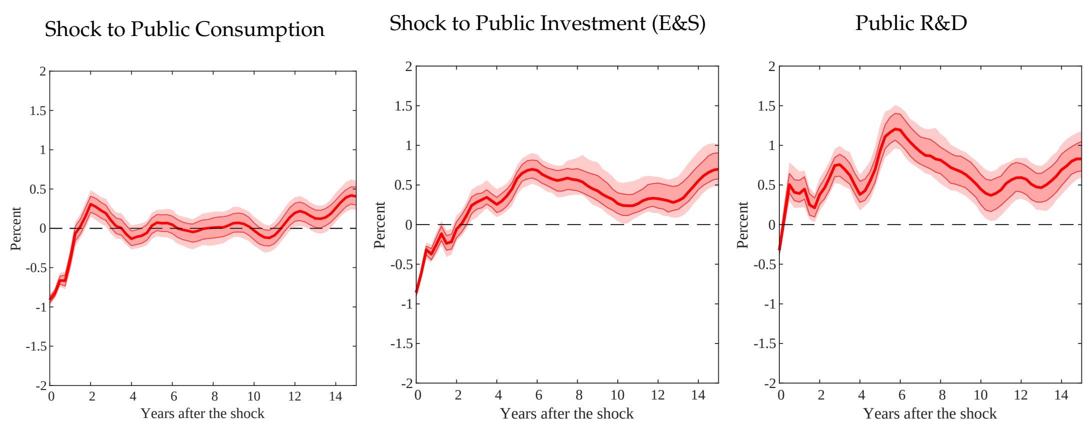
[14] This is likely to reflect the patterns of military spending ramp which, as discussed around Figure 4, lead to large short-run responses of investment and a longer-run increase in the share of R&D. Unsurprisingly, the output response to the public investment shock is more similar to the output response to the military spending shock.

Figure 7: IMPULSE RESPONSE TO PUBLIC SPENDING CATEGORY SHOCKS

(a) Responses of GDP to Public Spending Category Shocks



(b) Responses of Research and Development, as Share of Total Government Spending



Notes. The impulse responses are based on an estimated VAR with sixty lags of real government spending per capita, real GDP per capita, GDP deflator, short-term interest rate, fiscal deficit to GDP ratio, and government debt to GDP ratio. In each of the columns real public consumption per capita, real public non-R&D investment per capita and public R&D expenditure per capita, respectively, are added in turn to the VAR. Each public spending category, total government spending, GDP and the GDP deflator enter the VAR in log-levels. The shock to each public spending category is identified using the max-share method at the one-year forecast horizon for that category. The size of the shock is normalized such as to increase total government spending by 1 percent of GDP over the first year after the shock. The top (bottom) panel refers to the response of GDP (public R&D as share of total government spending) to shocks to each public spending category. The darker (lighter) shaded area represents the central 68% (90%) HPD band. The darker solid line stands for the median estimates. Results are based on 5000 posterior draws.

In summary, military spending shifts the composition of public spending towards R&D. A shock that rises the relative intensity of public R&D leads to strong responses of investment, productivity, innovation and output. The latter is larger than the output responses to either more 'public consumption-intensive' or more 'public investment-intensive' shocks. We interpret this as suggestive evidence that public R&D is a key driver of the long-run effects of government spending documented in this paper.

6 Further Results

In this section, we show that our estimated long-run effects are not driven by any specific war episode, and that we obtain similar results when we focus on samples/public spending categories that are not dominated by military conflicts. Furthermore, we assess the role of lag length selection and variable transformation in shaping our long-run results. Finally, we show that none of our conclusions is overturned when we use Bayesian Local Projections, vary the tightness of the priors or adopt a different identification of the public R&D shock, in the spirit of Blanchard and Quah (1989). In contrast, we find that a monetary policy shock —isolated using a standard recursive identification— has no long-run effects on output or productivity.

6.1 Not Only Wars

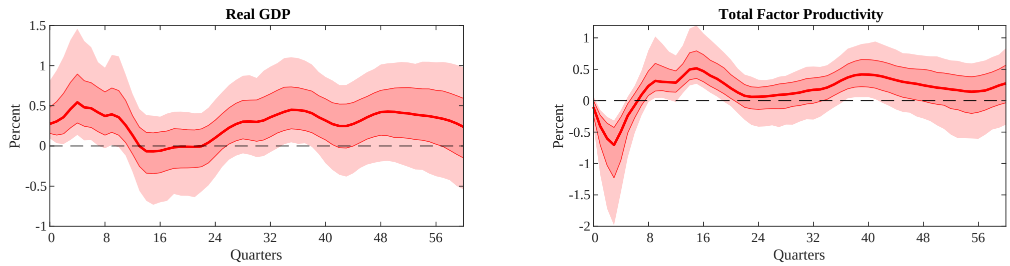
As argued by Friedman (1952), exploiting wars and military spending for identifying the effects of government expenditure is attractive for at least two reasons. First, the variation in military spending associated with wars (abroad) is typically independent from the state of the (domestic) business-cycle and thus should prevent reverse causality feedbacks running from GDP to government spending. Second, these public spending swings tend to be large in historical perspective, thereby offering sufficient variation in the leading variable. On the other hand, using wars as source of exogenous variation poses some external validity challenges on whether a specific episode or a specific public spending category drives the empirical findings and on whether these generalize to a peacetime period.

In the top three rows of Figure 8, we present results from three separate VARs(60) in which we have censored to zero the observations in the military spending news series associated with the following three pairs of military conflicts: (i) WWI and WWII, (ii) WWI and the Korean war, (iii) WWII and the Korean war. The first column refers to the GDP response while the second column is about the impact on TFP. In none of these exercises, the exclusion of any pair of these war-induced military spending overturns our main conclusions: the long-run effects on GDP and productivity are still large and significant. On the other hand, excluding all three war episodes at once (not reported) produces small and insignificant output responses at longer horizons. In other words, each and every one of these war-induced, large government spending increases seems sufficient to elicit significant long-run effects, though none of them is actually necessary.

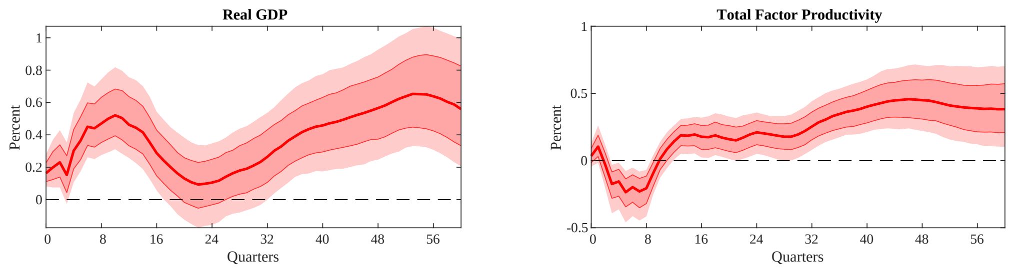
A related question is whether government spending can stimulate long-run growth also in peacetime. In the previous section, we have already shown that a surge in public R&D has long-lasting effects on innovation, productivity and output, but that our newly identified 'public R&D shock' has a correlation of only 0.17 with the military spending shock of Ramey and Zubairy (2018). A possible concern, however, is that a sizable share of variation in public R&D may come from the 'Manhattan project', which in turn was spurred by WWII. To investigate this, in the fourth row of Figure 8, we restrict our sample to the post-WWII period and focus on the effects of a public R&D shock identified as in the previous section: the long-run responses of output and productivity are large and significant also in the post-WWII sample. We interpret these findings as suggestive evidence that government spending (in R&D) can stimulate long-run growth even when it is not associated with a military conflict.

Figure 8: SENSITIVITY TO WARS

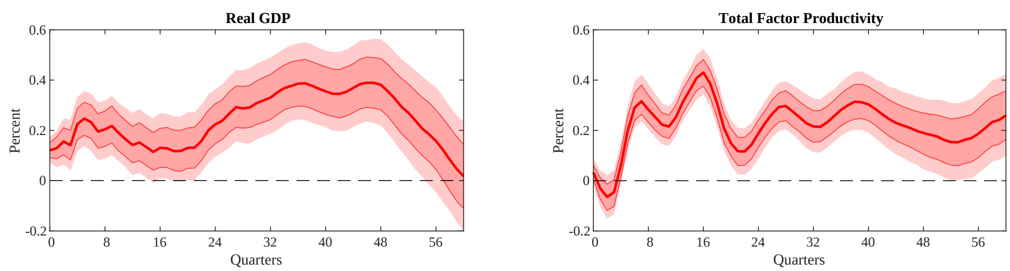
(a) Military Spending Shock Excluding Both World Wars



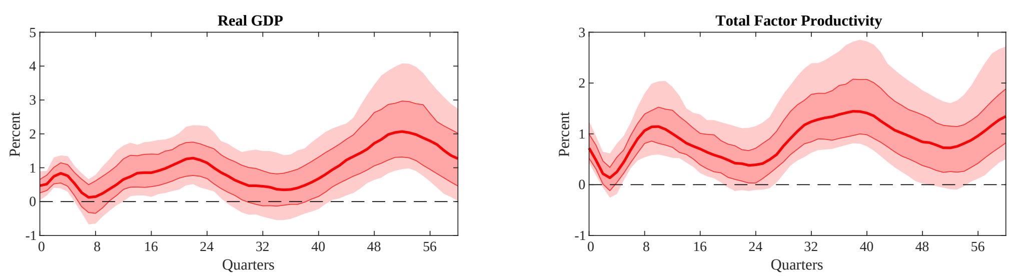
(b) Military Spending Shock Excluding World War I and Korea



(c) Military Spending Shock Excluding World War II and Korea



(d) Public R&D Shock using only Post-WWII Data



Note: The solid lines represent the median posterior response. The darker shadow area represents the 68^{th} posterior credible intervals, while the lighter shadow are represents the 95^{th} posterior credible intervals. All specifications use sixty lags of military spending news, government spending, GDP, GDP deflator, short-term rate, deficit to GDP ratio and debt to GDP ratio, with the exception of the last row which excludes military spending news.

6.2 Treatment of Trends and Lag Length Selection

All the results so far have been based on a VAR with the main variables in log-levels and using sixty lags. The empirical macro literature, however, has often used a much shorter lag selections, with four lags being the most popular choice in analyses with quarterly data. Moreover, detrending output and government spending by a measure of potential output, as in Ramey and Zubairy (2018), is also a popular strategy among fiscal policy studies. Accordingly, in Figure 9, we look at the consequences and interactions of these choices.

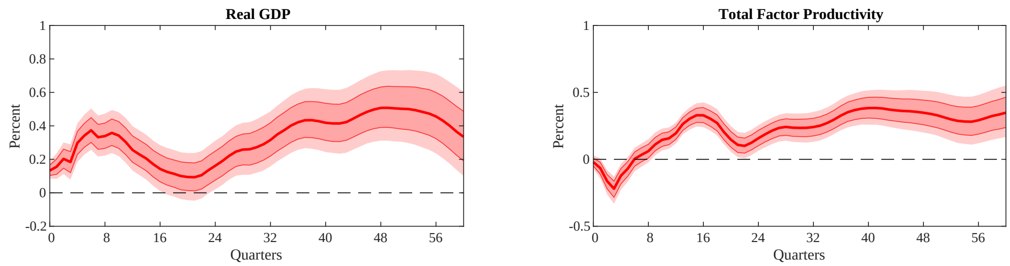
In Panel (a), we record estimates from the baseline VAR specification in levels, using 60 lags. In Panel (b), we retain the variables in levels, but employs only $p = 4$ lags. While the impulse responses are much smoother and the second hump in GDP is understated in the second row, the broad qualitative patterns, including a persistent response of GDP and TFP, are still present. In Panel (c), we look instead at the IRFs after removing potential from output, using 60 lags. The two-humped GDP response is visible, although by the end of the horizon the effects on GDP move towards zero, i.e. this transformation is filtering out any effect of the shock on the level of potential output.[15] In Panel (d), we consider the de-trended specification using only four lags. This specification completely misses the long-run effects of government spending on output and productivity.

The results in this section echo Ramey (2016) that specifications in levels are probably a safer option, particularly when the analysis focuses on the long-run: the low frequency components of the data can be broadly recovered even by using only a small number of lags. However, the lag-truncation bias can be very significant and severely distort the inference one can draw on medium-term dynamics, in a way that can be particularly severe when using detrended data.

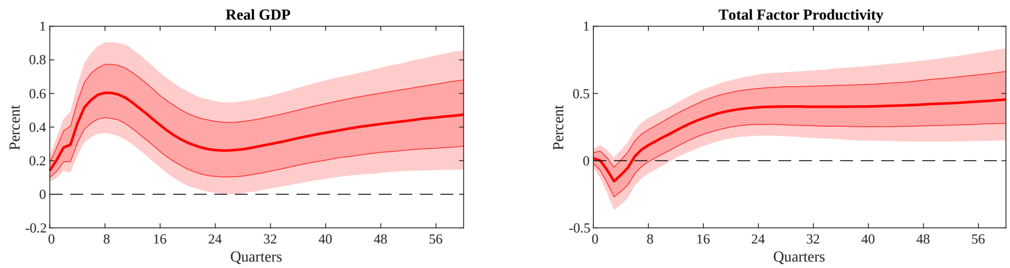
[15]Notice that this transformation is akin to taking a two-sided filter of the data, so implicitly it is using information from the future and this may introduce additional distortions to the time series properties of the data.

Figure 9: IMPULSE RESPONSE FUNCTIONS UNDER ALTERNATIVE SPECIFICATIONS

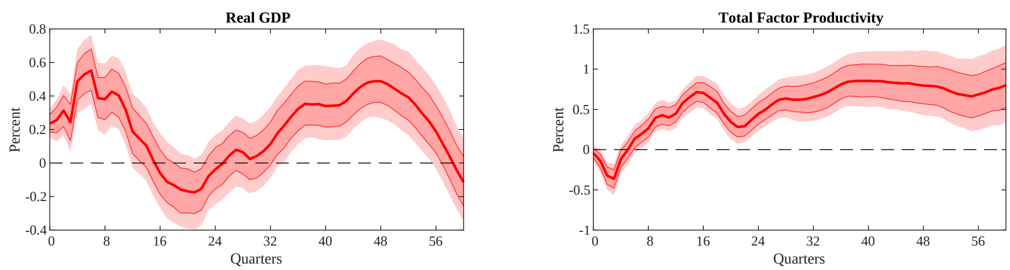
(a) Variables in levels ($p = 60$)



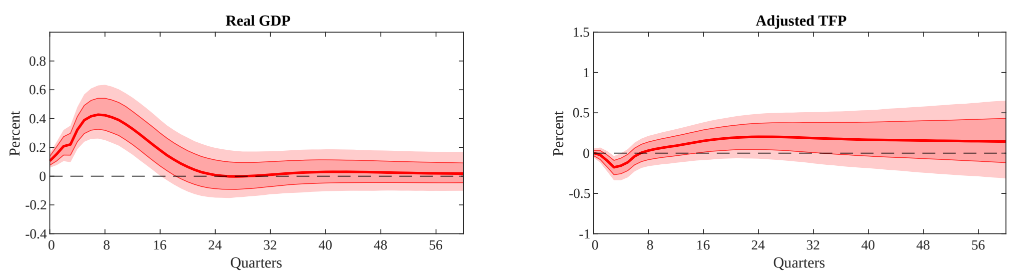
(b) Variables in Levels ($p = 4$)



(c) Detrending by Potential Output ($p = 60$)



(d) Detrending by Potential Output ($p = 4$)



Note: The solid lines represent the median posterior response. The darker shadow area represents the 68^{th} posterior credible intervals, while the lighter shadow are represents the 95^{th} posterior credible intervals. Results are based on 5000 posterior draws.

6.3 Sensitivity Analysis

In this section, we assess the robustness of our main findings to an alternative estimation method, identification strategy of public R&D shocks, prior tightness and to considering another popular macroeconomic shock.

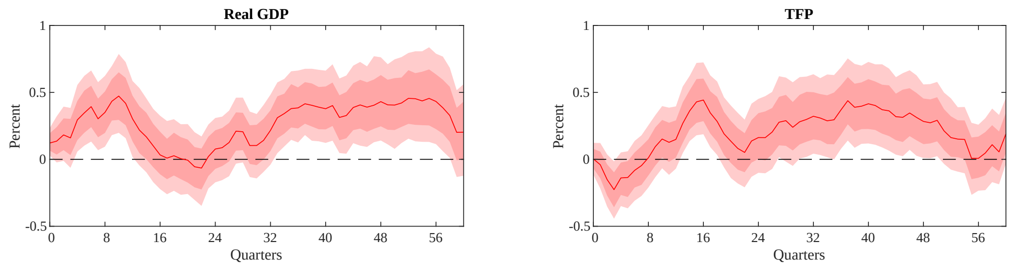
Local Projections. In a recent paper, Plagborg-Møller and Wolf (2021) demonstrate that, in large samples, VAR and Local Projections (LP) estimate the same impulse responses whenever the number of lags in the VAR is as large as the forecast horizon. Furthermore, Montiel Olea and Plagborg-Møller (2021) show that to obtain robust inference about the long-run, local projections should be augmented with a sufficient number of lags of all relevant control variables. The impulse responses from a LP specification with sixty lags of the controls are reported in the top row of Figure 10 and confirms, by and large, the estimates of the VAR(60) in Figure 1: the long-run impact of government spending on GDP and TFP is large and significant also using local projections. The full set of impulse responses are reported in Appendix H.

An Alternative Identification of Public R&D shocks. In Section 5, we have isolated exogenous movements in 'public R&D' by extracting the shock that explains most of the public R&D variance during the first year after the shock. Given our focus on long-run dynamics, however, it may be argued that a plausible alternative could be to identify the shock that explains most of the long-run fluctuations in public R&D, as measured for instance by frequencies beyond 32 quarters. This identification is very much related to the long-run restrictions popularized by Blanchard and Quah (1989), and therefore the resulting series could be interpreted as a sort of 'publicly-funded technology shock'.[16]

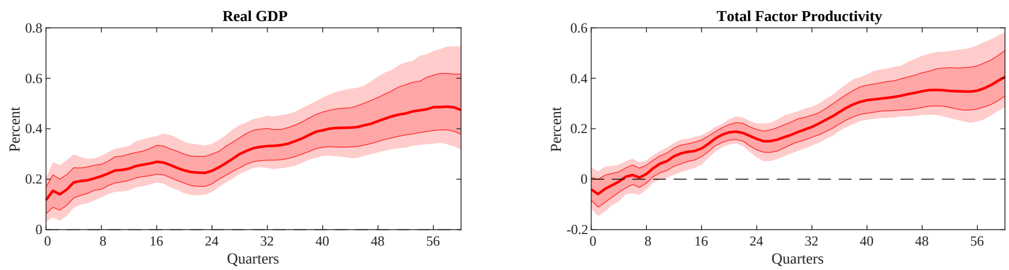
[16] As all trending variables enter the model in log-level, one may be concerned about reverse causality: the long-run movements in public R&D may simply reflect the trend in output. To ameliorate this concern, we estimate VARs(60) where public R&D enter as share of either total government spending or GDP (rather than in log-level). The public R&D shock is then identified as the shock that explains most of the long-run fluctuations in the public R&D share. The

Figure 10: IMPULSE RESPONSE FUNCTIONS UNDER ALTERNATIVE SPECIFICATIONS

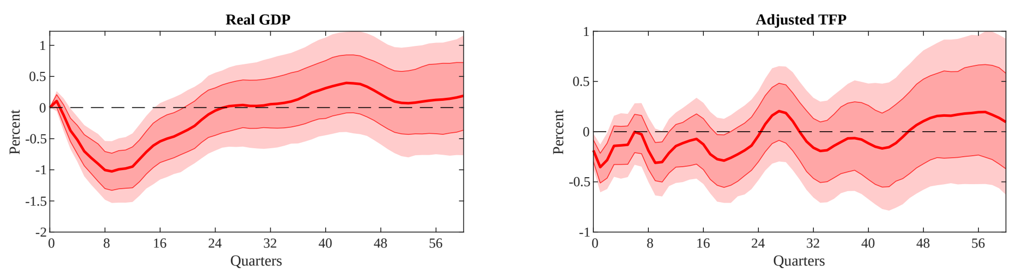
(a) Bayesian Local Projections



(b) Alternative Identification of Public R&D Shock



(c) Placebo Identification: Monetary Policy Shock



Note: The solid lines represent the median posterior response. The darker shadow area represents the 68^{th} posterior credible intervals, while the lighter shadow are represents the 95^{th} posterior credible intervals. All specifications use sixty lags of military spending news, real government spending per capita, real GDP per-capita, GDP deflator, short-term rate, deficit to GDP ratio and debt to GDP ratio. Government spending, GDP and the GDP deflator enter the VAR in log-levels. Results are based on 5000 posterior draws.

In the second row of Figure 10, we show that this alternative, long-run identification produces very similar results relative to the responses estimated in Figure 6 using short-run restrictions. More specifically, an exogenous increase in public R&D triggers very persistent effects on output and TFP, with sizable peaks towards the end of the forecast

impulses responses of output and productivity are very similar to those reported in the second row of Figure 10.

horizon. We conclude that our estimates of the long-run effects of public R&D are insensitive to the specific frequencies upon which these shocks are isolated, possibly reflecting the relatively a-cyclical nature of public R&D expenditure.

The Effects of Monetary Policy Shocks. Another possible concern is that the rich parameterization may have introduced some spurious cycles in the VAR(60) reduced-form estimates. Alternatively, a propagation mechanism a la Comin and Gertler (2006) or Beaudry et al. (2020) may drive such a large share of low-frequency variation in the data that any shock would produce highly persistent dynamics. In either case, it would be misleading to infer that government spending is responsible for the estimated long-run effects on output.

To evaluate this hypothesis, we use our baseline VAR(60) to identify a series of monetary policy shocks employing a Cholesky factorization where real GDP and the GDP deflator are ordered before the short-term interest rate. The idea behind this identification, which has a long tradition in empirical macro (Christiano et al., 2005), is that while monetary policy responds to contemporaneous developments in output and prices, it takes at least a quarter for the effects of central bank interventions to transmit to the macroeconomy.[17]

The estimated impulse responses to a monetary policy shock are presented in the third row of Figure 10 and they closely resemble those typically found in the empirical monetary literature (Christiano et al., 2005). The estimates of this structural VAR(60) point to significant short-term contractions in output and productivity but exhibit no second wave of effects at longer horizons.[18] We conclude that the long-run effects

[17]Relative to equally popular approaches such as those based on narrative evidence and the Greenbook forecasts (Romer and Romer, 2004) or on high frequency movements of interest rate futures around policy announcements (Gürkaynak et al., 2005), the recursive identification has the notable advantage of being readily implementable in our long sample, over which neither the Fed internal forecasts nor the interest rate futures are available.

[18]The results in this section are not necessarily inconsistent with those in Jordà et al. (2020). First, these authors look at an international panel of 17 advanced economies whereas we focus on the U.S. only. Second, and most importantly, Jordà et al. (2020) isolate the exogenous component of monetary policy via the trilemma in international finance while we use a more conventional Cholesky identification, whose only purpose is to show one example in which the type of contemporaneous zero restrictions used in the main analysis can produce small and insignificant long-run effects.

that we have documented in this paper are likely to reflect a genuine (low-frequency) feature of the U.S. government spending data rather than an artifact of our richly parameterized model, or a systematic response of output to any type of shocks.

6.4 The Role of the Priors

The results in this and previous sections have used the prior tightness selection in Giannone et al. (2015), who propose to treat λ and θ in equation (4) as a hyperparameters to be estimated hierarchically. The authors recommend setting these to the value that maximizes the marginal likelihood of the model. Since the latter is closely related to the one-step ahead out-of-sample forecast error, this selection strategy is attractive because it targets values of λ and θ that is optimal at a horizon (i.e. one quarter ahead) which is not the focus of our analysis (i.e. the long-run). But how much does the prior selection influence the posterior distributions? This question can be answered by varying the tightness of the priors.

The findings of this exercise are recorded in Appendix I. We report the impact of tightening the prior further on both output and productivity. Unsurprisingly, as the hyperparameter λ of the Minnesota prior is tightened, the impulse responses become progressively smoother and the long-run effects less significant. A the same time, the confidence bands narrow.[19] On the other hand, it is remarkable to note that despite the heavy tightness imposed by the lower λs, the posterior estimates in the bottom rows still point to some non-negligible and significant effects of government spending on output and productivity beyond business-cycle frequencies. On the other hand, relaxing the hyperparameter θ from its baseline value of 0.001 leads to smaller long-run effects of output and productivity; but even for the relatively uninformative value of $\theta = 1$, the second hump of GDP and productivity is present and strongly significant.

[19] The exponential discounting of the lag structure embedded in the square of the parameter ℓ in equation (4) implies that the prior variance for these coefficients on lag $\ell = 60$ of all variables is scaled down by a factor of $3600 = (\ell^2)/1$ in the case of the uninformative prior variance $\lambda = 1$, by $22500 = (\ell^2)/0.4^2$ for the informative scenario of $\lambda = 0.4$, by $90000 = (\ell^2)/0.2^2$ for the 'conservative' case of $\lambda = 0.2$ and by $360000 = (\ell^2)/0.1^2$ for the 'dogmatic' case of $\lambda = 0.1$.

7 Conclusions

What are the long-run effects of government spending? Despite the resurgence in fiscal research spurred by the financial crisis of 2007-09 and the policy debate triggered by the global pandemic of 2020-22, this question seems to have so far eluded empirical research. In this paper, we use 125 years of U.S. quarterly data —including newly constructed series of public spending by main categories— and time series models with up to sixty lags to shed light on this issue. We argue that the combination of historical data, a generous lag length selection and Bayesian shrinkage makes our framework well-suited to draw inference about long-run dynamics, while retaining the ability to look also at the short-run.

We uncover four main regularities. First, fiscal policy can stimulate long-run growth when it tilts the share of public spending towards R&D, as it does for instance during military conflicts. However, we also find that an exogenous increase in public R&D expenditure can have long-run effects on output even when it is not systematically associated with war spending. Second, in contrast, government investment has shorter-lived effects whereas the impact of public consumption on economic growth is modest at most horizons. Third, while government spending crowds out innovation, private investment and private consumption in the short-run, it crowds them in over the medium-term, feeding into a sustained increase in total factor productivity at longer horizons. As a result, the government spending multiplier on output is around one at business-cycle frequencies but raises above one in the long-run. Finally, an increase in government spending of 1% of GDP triggers a sustained inflation spell of around 0.2% per year for about four years, bringing the price index back to the pre-shock trend after more than ten years since the shock. Our analysis seems to uncover a novel mechanism through which fiscal policy can stimulate long-run economic growth.

References

AKCIGIT, U., J. R. GRIGSBY, T. NICHOLAS, AND S. STANTCHEVA (2022): "Taxation and Innovation in the 20th Century," *Quarterly Journal of Economics*, forthcoming.

BAEK, C. AND B. LEEB (2021): "A Guide to Autoregressive Distributed Lag Models for Impulse Response Estimations," mimeographed, University of California at Berkeley and Hong Kong University.

BARNICHON, R. AND C. BROWNLEES (2019): "Impulse response estimation by smooth local projections," *Review of Economics and Statistics*, 101, 522–530.

BARRO, R. J. AND C. J. REDLICK (2011): " Macroeconomic Effects From Government Purchases and Taxes *," *The Quarterly Journal of Economics*, 126, 51–102.

BEAUDRY, P., D. GALIZIA, AND F. PORTIER (2020): "Putting the Cycle Back into Business Cycle Analysis," *American Economic Review*, 110, 1–47.

BENIGNO, G. AND L. FORNARO (2017): "Stagnation Traps," *The Review of Economic Studies*, 85, 1425–1470.

BERGEAUD, A., G. CETTE, AND R. LECAT (2016): "Productivity Trends in Advanced Countries between 1890 and 2012," *Review of Income and Wealth*, 62, 420–444.

BLANCHARD, O. AND R. PEROTTI (2002): "An Empirical Characterization of the Dynamic Effects of Changes in Government Spending and Taxes on Output," *Quarterly Journal of Economics*, 117, 1329–68.

BLANCHARD, O. J. AND D. QUAH (1989): "The Dynamic Effects of Aggregate Demand and Supply Disturbances," *American Economic Review*, 79, 655–673.

BUSH, V. (1954): "Science, the endless frontier," in *Science, the Endless Frontier*, Princeton University Press.

CENSUS (1949): *Historical statistics of the United States, 1789 - 1957*, US Department of Commerce, Bureau of the Census.

CERRA, V., A. FATÁS, AND S. SAXENA (2022): "Hysteresis and Business Cycles," *Journal of Economic Literature*, forthcoming.

CHODOROW-REICH, G. (2019): "Geographic Cross-Sectional Fiscal Spending Multipliers: What Have We Learned?" *American Economic Journal: Economic Policy*, 11, 1–34.

CHONG, Y. Y. AND D. F. HENDRY (1986): "Econometric Evaluation of Linear Macro-Economic Models," *The Review of Economic Studies*, 53, 671–690.

CHOW, G. C. AND A.-L. LIN (1971): "Best linear unbiased interpolation, distribution, and extrapolation of time series by related series," *The review of Economics and Statistics*, 372–375.

CHRISTIANO, L., M. EICHENBAUM, AND C. EVANS (2005): "Nominal Rigidities and the Dynamic Effects of a Shock to Monetary Policy," *Journal of Political Economy*, 113, 1–45.

CLOYNE, J., J. MARTINEZ, H. MUMTAZ, AND P. SURICO (2022): "Short-Term Tax Cuts, Long-Term Stimulus," mimeographed, University of California at Davis, London Business School and Queen Mary University of London.

COMIN, D. AND M. GERTLER (2006): "Medium-Term Business Cycles," *American Economic Review*, 96, 523–551.

DEL NEGRO, M. AND F. SCHORFHEIDE (2011): "Bayesian macroeconometrics," *The Oxford handbook of Bayesian econometrics*, 293, 389.

DELEIDI, M. AND M. MAZZUCATO (2021): "Directed innovation policies and the supermultiplier: An empirical assessment of mission-oriented policies in the US economy," *Research Policy*, 50, 104151.

DIEBOLT, C. AND K. PELLIER (2020): "Patents in the Long Run : Theory, History and Statistics," *in: History Mathematics*, 8, 80–119.

DOAN, T., R. LITTERMAN, AND C. SIMS (1984): "Forecasting and conditional projection using realistic prior distributions," *Econometric Reviews*, 3, 1–100.

DONALDSON, D. AND R. HORNBECK (2016): "Railroads and American Economic Growth: A "Market Access" Approach," *The Quarterly Journal of Economics*, 131, 799–858.

DUPREE, A. H. (1986): *Science in the federal government*, Johns Hopkins University Press.

FERNALD, J. G. (1999): "Roads to Prosperity? Assessing the Link between Public Capital and Productivity," *American Economic Review*, 89, 619–638.

——— (2012): "A Quarterly, Utilization-Adjusted Series on Total Factor Productivity," Federal Reserve Bank of San Francisco Working Paper Series 2012-19.

FRANCIS, N., M. T. OWYANG, J. E. ROUSH, AND R. DICECIO (2014): "A flexible finite-horizon alternative to long-run restrictions with an application to technology shocks," *Review of Economics and Statistics*, 96, 638–647.

FRIEDMAN, M. (1952): "Price, Income, and Monetary Changes in Three Wartime Periods," *The American Economic Review*, 42, 612–625.

FRIEDMAN, M. AND A. J. SCHWARTZ (1963): *A Monetary History of the United States, 1867-1960*, Princeton University Press.

Giannone, D., M. Lenza, and G. E. Primiceri (2015): "Prior selection for vector autoregressions," *Review of Economics and Statistics*, 97, 436–451.

——— (2019): "Priors for the long run," *Journal of the American Statistical Association*, 114, 565–580.

Gordon, R. J. (2007): *The American business cycle: Continuity and change*, vol. 25, University of Chicago Press.

Gross, D. P. and B. N. Sampat (2020): "Organizing Crisis Innovation: Lessons from World War II," Working Paper 27909, National Bureau of Economic Research.

Gruber, J. and S. Johnson (2019): *Jump-starting America: how breakthrough science can revive economic growth and the American dream*, New York, NY : PublicAffairs.

Gürkaynak, R. S., B. Sack, and E. Swanson (2005): "Do Actions Speak Louder Than Words? The Response of Asset Prices to Monetary Policy Actions and Statements," *International Journal of Central Banking*, 1.

Hall, G. J. and T. J. Sargent (2011): "Interest Rate Risk and Other Determinants of Post-WWII US Government Debt/GDP Dynamics," *American Economic Journal: Macroeconomics*, 3, 192–214.

Hall, R. E. (2009): "By How Much Does GDP Rise If the Government Buys More Output?" *Brookings Papers on Economic Activity*, 40, 183–249.

Hornbeck, R. and M. Rotemberg (2021): "Railroads, Market Access, and Aggregate Productivity Growth," mimeographed, University of Chicago and New York University.

Ilzetzki, E. (2022): "Learning by Necessity: Government Demand, Capacity Constraints, and Productivity Growth," mimeographed, London School of Economics and Political Science.

Imbs, J. M. (1999): "Technology, growth and the business cycle," *Journal of Monetary Economics*, 44, 65–80.

Jarocinski, M. and A. Marcet (2015): "Contrasting Bayesian and Frequentist Approaches to Autoregressions: The Role of the Initial Condition," Tech. rep., Working Papers 776, Barcelona Graduate School of Economics.

Jordà, Ò. (2005): "Estimation and Inference of Impulse Responses by Local Projections," *American Economic Review*, 95, 161–182.

Jordà, Ò., M. Schularick, and A. M. Taylor (2017): "Macrofinancial history and the new business cycle facts," *NBER macroeconomics annual*, 31, 213–263.

Jordà, Ò., S. R. Singh, and A. M. Taylor (2020): "The Long-Run Effects of Monetary Policy," Working Paper 26666, National Bureau of Economic Research.

KANTOR, S. AND A. WHALLEY (2022): "Moonshot: Public RD and Growth," mimeographed, Florida State University and University of Calgary.

KILIAN, L. AND H. LÜTKEPOHL (2017): *Structural vector autoregressive analysis*, Cambridge University Press.

KURMANN, A. AND C. OTROK (2013): "News shocks and the slope of the term structure of interest rates," *American Economic Review*, 103, 2612–32.

LEFF YAFFE, D. (2020): "The Interstate Multiplier," mimeographed, University of California at San Diego.

LI, D., M. PLAGBORG-MØLLER, AND C. K. WOLF (2021a): "Local Projections vs. VARs: Lessons From Thousands of DGPs," *arXiv preprint arXiv:2104.00655*.

LI, D., M. PLAGBORG-MØLLER, AND C. K. WOLF (2021b): "Local Projections vs. VARs: Lessons From Thousands of DGPs," mimeographed, Princeton University and University of Chicago.

MELTZER, A. H. (2004): *A History of the Federal Reserve, Volume 1: 1913-1951*, no. 9780226520001 in University of Chicago Press Economics Books, University of Chicago Press.

MIRANDA-AGRIPPINO, S. AND G. RICCO (2021): "The transmission of monetary policy shocks," *American Economic Journal: Macroeconomics*, 13, 74–107.

MIRANDA-AGRIPPINO, S., G. RICCO, ET AL. (2021): *Bayesian local projections*, University of Warwick, Department of Economics.

MONTIEL OLEA, J. L. AND M. PLAGBORG-MØLLER (2021): "Local Projection Inference Is Simpler and More Robust Than You Think," *Econometrica*, 89, 1789–1823.

MORETTI, E., C. STEINWENDER, AND J. VAN REENEN (2019): "The Intellectual Spoils of War? Defense RD, Productivity and International Spillovers," Working Paper 26483, National Bureau of Economic Research.

MOUNTFORD, A. AND H. UHLIG (2009): "What are the effects of fiscal policy shocks?" *Journal of Applied Econometrics*, 24, 960–992.

MÜLLER, U. K. AND M. W. WATSON (2018): "Long-Run Covariability," *Econometrica*, 86, 775–804.

NAKAMURA, E. AND J. STEINSSON (2014): "Fiscal Stimulus in a Monetary Union: Evidence from US Regions," *American Economic Review*, 104, 753–92.

——— (2018): "Identification in macroeconomics," *Journal of Economic Perspectives*, 32, 59–86.

NATIONAL RESEARCH COUNCIL (1995): *Allocating federal funds for science and technology*, National Academies Press.

OWYANG, M. T., V. A. RAMEY, AND S. ZUBAIRY (2013): "Are government spending multipliers greater during periods of slack? Evidence from twentieth-century historical data," *AEA Papers and Proceedings*, 103, 129–34.

PLAGBORG-MØLLER, M. AND C. K. WOLF (2021): "Local Projections and VARs Estimate the Same Impulse Responses," *Econometrica*, 89, 955–980.

RAMEY, V. A. (2011a): "Can Government Purchases Stimulate the Economy?" *Journal of Economic Literature*, 49, 673–85.

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

——— (2016): "Macroeconomic shocks and their propagation," *Handbook of macroeconomics*, 2, 71–162.

——— (2019): "Ten Years after the Financial Crisis: What Have We Learned from the Renaissance in Fiscal Research?" *Journal of Economic Perspectives*, 33, 89–114.

——— (2020): "The Macroeconomic Consequences of Infrastructure Investment," in *Economic Analysis and Infrastructure Investment*, National Bureau of Economic Research, Inc, NBER Chapters.

RAMEY, V. A. AND M. SHAPIRO (1998): "Costly Capital Reallocation and the Effects of Government Spending," *Carnegie Rochester Conference on Public Policy*, 48, 145–194.

RAMEY, V. A. AND S. ZUBAIRY (2018): "Government Spending Multipliers in Good Times and in Bad: Evidence from US Historical Data," *Journal of Political Economy*, 126, 850–901.

ROMER, C. D. AND D. H. ROMER (2004): "A New Measure of Monetary Shocks: Derivation and Implications," *American Economic Review*, 94, 1055–1084.

——— (2010): "The macroeconomic effects of tax changes: estimates based on a new measure of fiscal shocks," *American Economic Review*, 100, 763–801.

ROSSI, B. AND S. ZUBAIRY (2011): "What Is the Importance of Monetary and Fiscal Shocks in Explaining U.S. Macroeconomic Fluctuations?" *Journal of Money, Credit and Banking*, 43, 1247–1270.

SIMS, C. AND H. UHLIG (1991): "Understanding Unit Rooters: A Helicopter Tour," *Econometrica*, 59, 1591–99.

SIMS, C. A. (1993): "A nine-variable probabilistic macroeconomic forecasting model," in *Business cycles, indicators, and forecasting*, University of Chicago press, 179–212.

——— (2000): "Using a Likelihood Perspective to Sharpen Econometric Discourse: Three Examples," *Journal of Econometrics*, 95, 443–462.

SIMS, C. A., J. H. STOCK, AND M. W. WATSON (1990): "Inference in linear time series models with some unit roots," *Econometrica: Journal of the Econometric Society*, 113–144.

SIMS, C. A. AND T. ZHA (1998): "Bayesian methods for dynamic multivariate models," *International Economic Review*, 949–968.

UHLIG, H. (2003): "What moves real GNP?" Tech. rep.

——— (2005): "What are the effects of monetary policy on output? Results from an agnostic identification procedure," *Journal of Monetary Economics*, 52, 381–419.

WELCH, I. AND A. GOYAL (2008): "A comprehensive look at the empirical performance of equity premium prediction," *The Review of Financial Studies*, 21, 1455–1508.

Online Appendix to
"The Long-Run Effects of Government Spending"

by Juan Antolin-Diaz (LBS) and Paolo Surico (LBS and CEPR)

Contents

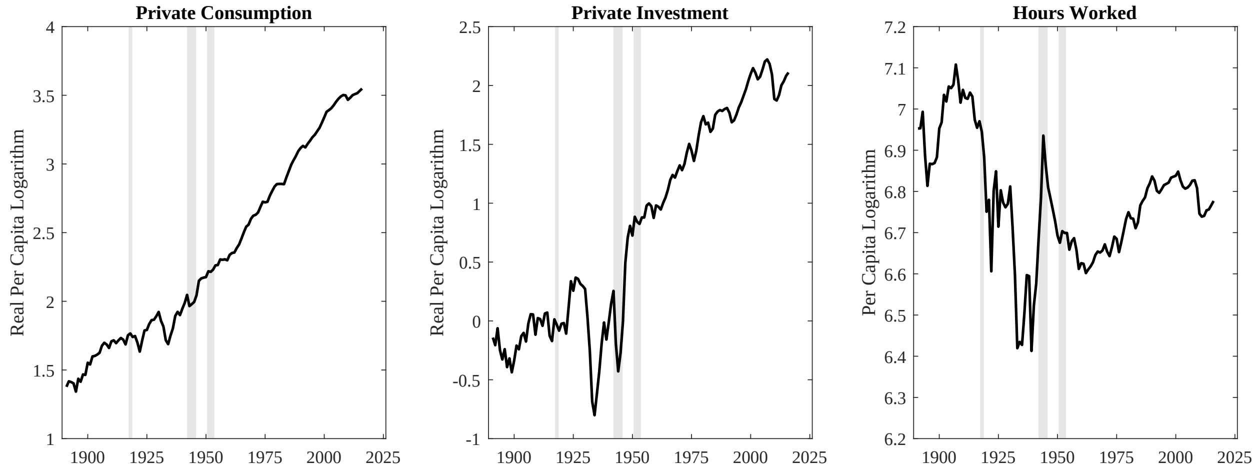
A Data construction

The data set for our baseline VAR and LP specifications comes from Ramey and Zubairy (2018) and contains seven variables from 1890Q1 to 2015Q5: the present discounted value of military news (Ramey, 2011b), government spending, real GDP, the log GDP deflator, the short-term interest rate, the surplus-to-GDP ratio and the Debt-to-GDP ratio. We use two main transformations of the data. Either we express real GDP per-capita and real government spending per-capita in logarithm or, following Ramey and Zubairy (2018), we scale them by a measure of GDP trend, estimated as a sixth-degree polynomial for the log of GDP, from 1889q1 through 2015q4, excluding 1930Q1–1946Q4.

We extend the dataset in Ramey and Zubairy (2018) in a number of dimensions that we describe in turn. We first extend backwards the time series for the short-term nominal interest rate, using data from Welch and Goyal (2008) for the New York Fed commercial paper rate. We obtain the long-term (10-year) interest rate from the same source. Private consumption and investment are based on unpublished annual estimates of investment by the Bureau of Economic Analysis, available since 1901. Before that, we rely on the Macrohistory Database of Jordà et al. (2017). These authors provide series for real GDP, real consumption of goods (including durables), and the investment-to-output ratio, from which levels of investment can be calculated. We then interpolate the annual series to quarterly frequency in the following way: from 1919-1940, we exploit quarterly series on consumption and investment from Gordon (2007) to interpolate the annual series using the method in Chow and Lin (1971). For the period when these are not available (1889-1918 and 1941-1946), we use quarterly real GDP from Ramey and Zubairy (2018) to perform the interpolation. After 1947, we employ the official NIPA estimates for quarterly consumption and investment. The results are displayed in Figure A.1.

We also construct new time series that break down government spending into its consumption and investment components. Annual series of government investment are available from the BEA since 1914, but we found that, because of rounding, they are inaccurate until the official NIPA estimates start in 1929. Therefore, we reconstruct the series of public investment and its components for the period 1890-1929 by manually transcribing detailed government outlays data from both the *Historical Statistics of the United States* (Census, 1949) and the annual *Statistical Abstracts* published

Figure A.1: U.S. CONSUMPTION, INVESTMENT, AND HOURS: 1890-2015

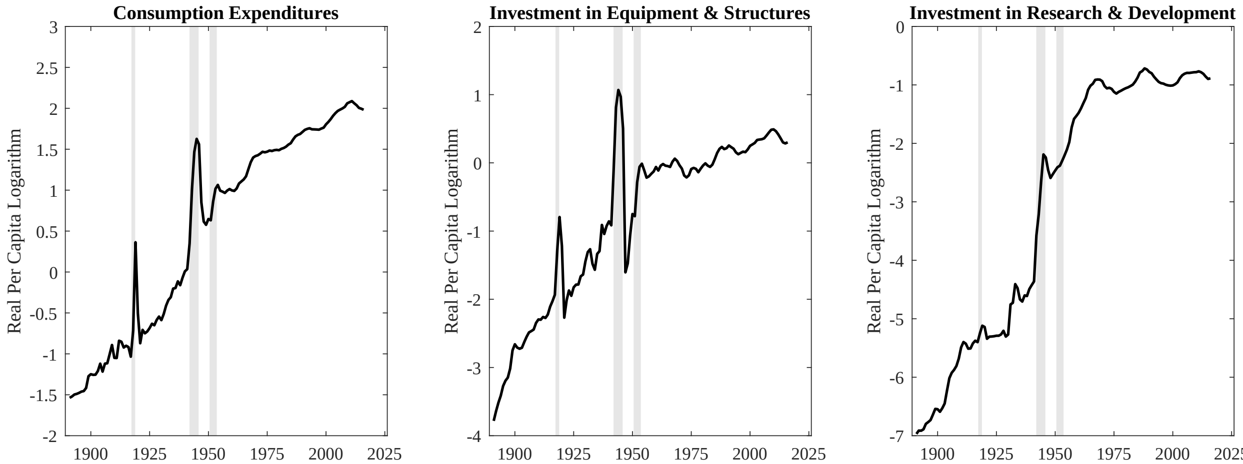


Notes. Private consumption and investment are deflated using the GDP deflator. All variables are scaled by the civilian population. Shaded areas represent the three major wars in the sample: World War I, World War II, and the Korean War.

by the census. We transcribe separately data for Federal and State and Local investment. First, the *Historical Statistics*, Chapter P, p.314, provides data points for State and Local "capital outlays" for the years 1890, 1902, 1913. We linearly interpolate observations between these years. For Federal investments in each year between 1890 and 1929, the *Statistical Abstracts* provides detailed annual breakdowns of federal government expenditures by use over the prior ten years. We transcribe this breakdown and sum up all categories of each department that appear to refer to investment, either in Equipment & Structures or in Research & Development.

To classify R&D investments, we rely on the narrative evidence in Bush (1954) and Dupree (1986) to allocate amounts across government departments. In particular, we cross check that total R&D spending matches the estimates reported by (Dupree, 1986, pp. 331-333). We also cross check that the sum of these categories is a good match to the official total amount for the years when they overlap. These estimates refer to the year ending on June 30, and thus we average with the next year to obtain an approximation of spending on the calendar year ending in December. After adding the Federal total to the State and Local investment constructed above, we obtain an annual investment series for the total government sector for 1890-1930, which we splice with the official BEA estimate starting in 1914. We then interpolate to quarterly frequency using the series of total government spending, and finally back out government consumption as a residual. Figure A.2 displays the three resulting series for Government spending components, in real, per capita terms.

Figure A.2: MAIN CATEGORIES OF U.S. GOVERNMENT SPENDING: 1890-2015



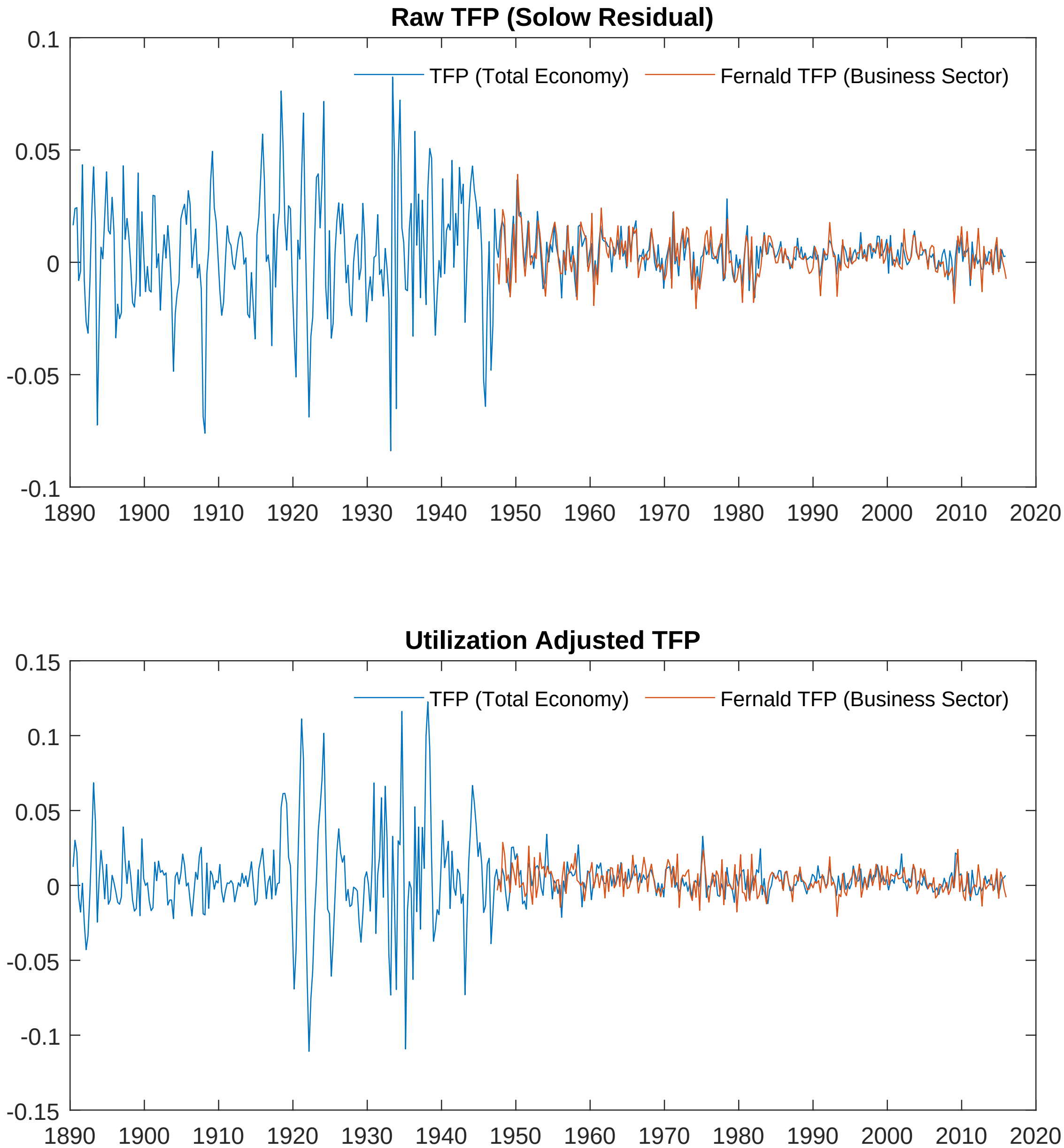
Notes. All components of government spending are deflated using the GDP deflator and scaled by the civilian population. Shaded areas represent the three major wars in the sample: World War I, World War II, and the Korean War.

Finally, the quarterly time series for Total Factor Productivity (TFP) has been constructed in two steps. First, we obtain annual measures of hours worked and the capital stock from Bergeaud et al. (2016).[1] These annual time series are interpolated to quarterly frequency. In the case of investment, we interpolate the annual measure of capital stock using the quarterly series of investment constructed above, cumulated using the perpetual inventory method.[2] For hours, we interpolate the annual measure using the unemployment rate series in Ramey and Zubairy (2018). The raw TFP series is then calculated as the Solow residual using quarterly real GDP, hours worked and the capital stock, assuming a Cobb-Douglass production function with constant returns to scale and a capital share of $\alpha = 0.28$. Second, to derive a measure of TFP adjusted for both capital and labour utilization, we use the method described by Imbs (1999) (and also employed by Jordà et al., 2020). This involves calculating steady-state measures of the capital-labor ratio, the consumption-output ratio and hours. We do so by applying the Hodrick-Prescott filter with a smoothing parameter of $\lambda = 1600$.

As shown in Figure A.3, which displays growth rates, and Figure A.4, which depicts log-levels, our historical quarterly time series of adjusted TFP, which refers to the whole economy, moves very closely to the more sophisticated and more data intensive measure proposed by Fernald (2012), which covers the business sector only, over the sample in which the two series overlap. Finally,

[1]We are thankful to Antonin Bergeaud for sharing this data with us.
[2]We assume a depreciation rate of $\delta = 0.1$ per annum.

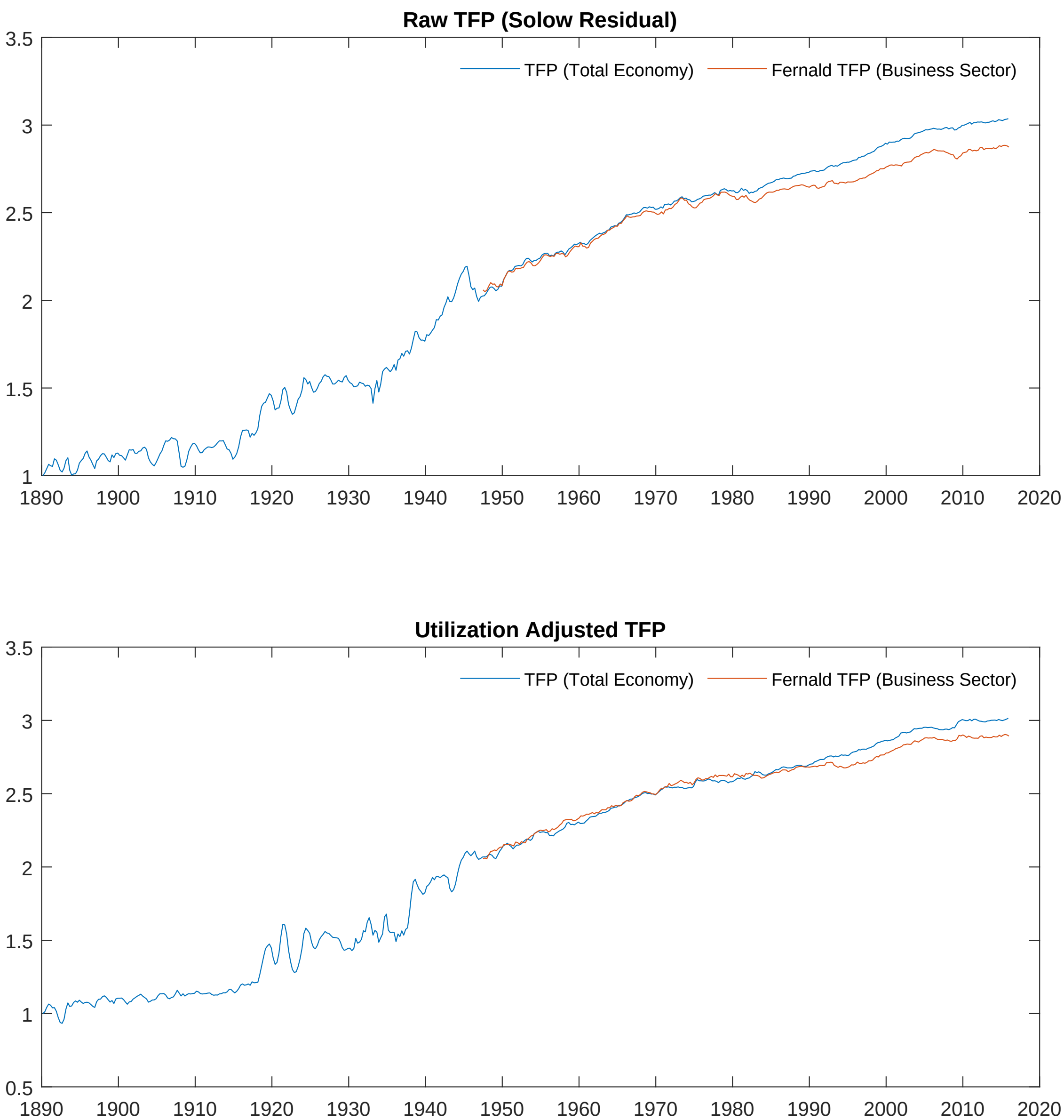
Figure A.3: RAW AND UTILIZATION ADJUSTED TFP GROWTH RATES



Notes. TFP Growth Rates as described in the Text. Top (bottom) row refers to the raw (utilization adjusted) TFP series.

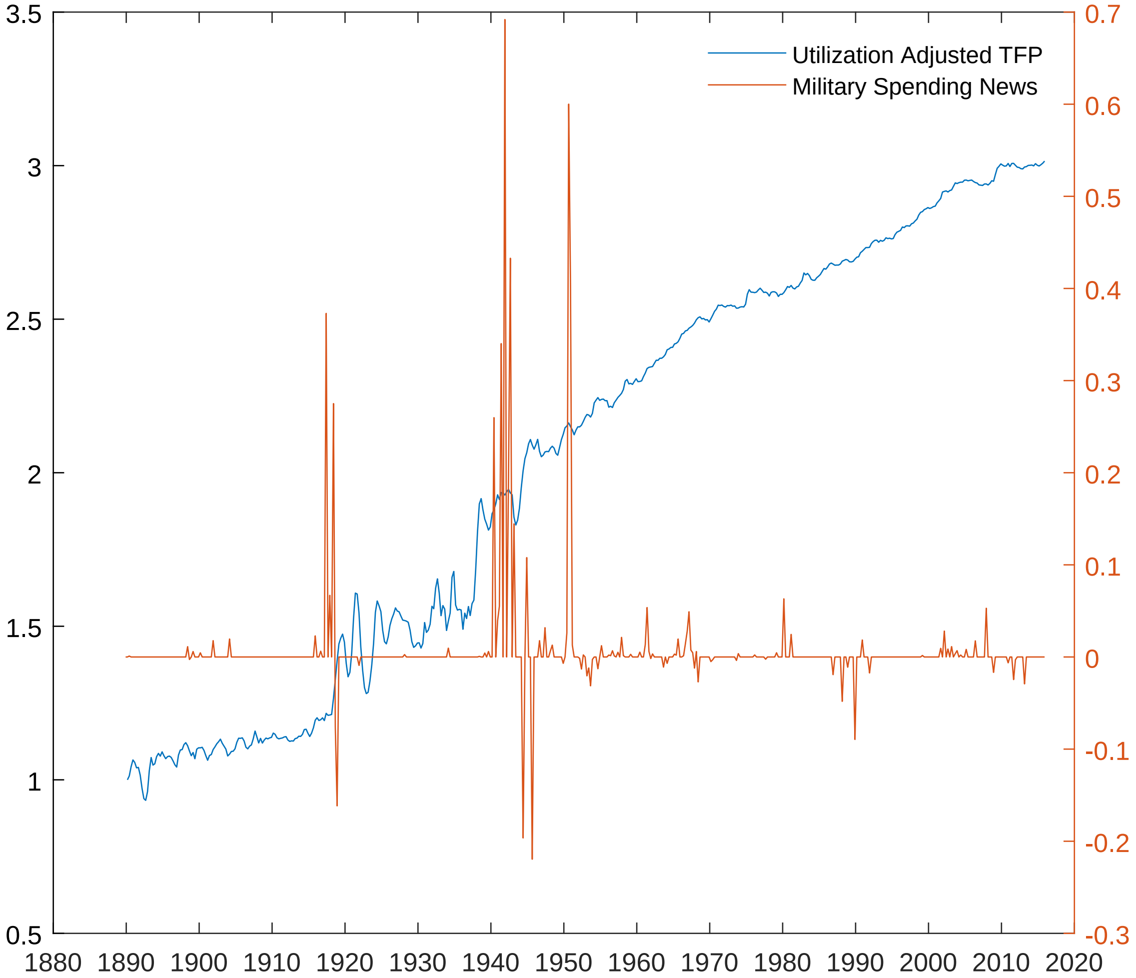
and mostly for completeness, in A.5, we report the quarterly measure of utilization adjusted TFP together with the quarterly time series of military spending news from Ramey and Zubairy (2018). It is interesting to note that our measure of total factor productivity tends to increase persistently after major episodes of military spending buildup, such as the two World Wars and –to a lesser extent– the Korean war, in a way that is visually apparent already at the naked eye. The estimates of our VAR(60) in the main text confirms formally this leading correlation.

Figure A.4: Raw and Utilization Adjusted TFP Levels



Notes. TFP levels as described in the Text. Top (bottom) row refers to the raw (utilization adjusted) TFP series.

Figure A.5: Utilization Adjusted TFP Levels and Military Spending News



Notes. Utilization-adjusted TFP levels as described in the Text. The military spending news as a percentage of GDP (right axis) is from Ramey and Zubairy (2018).

B Estimation algorithm

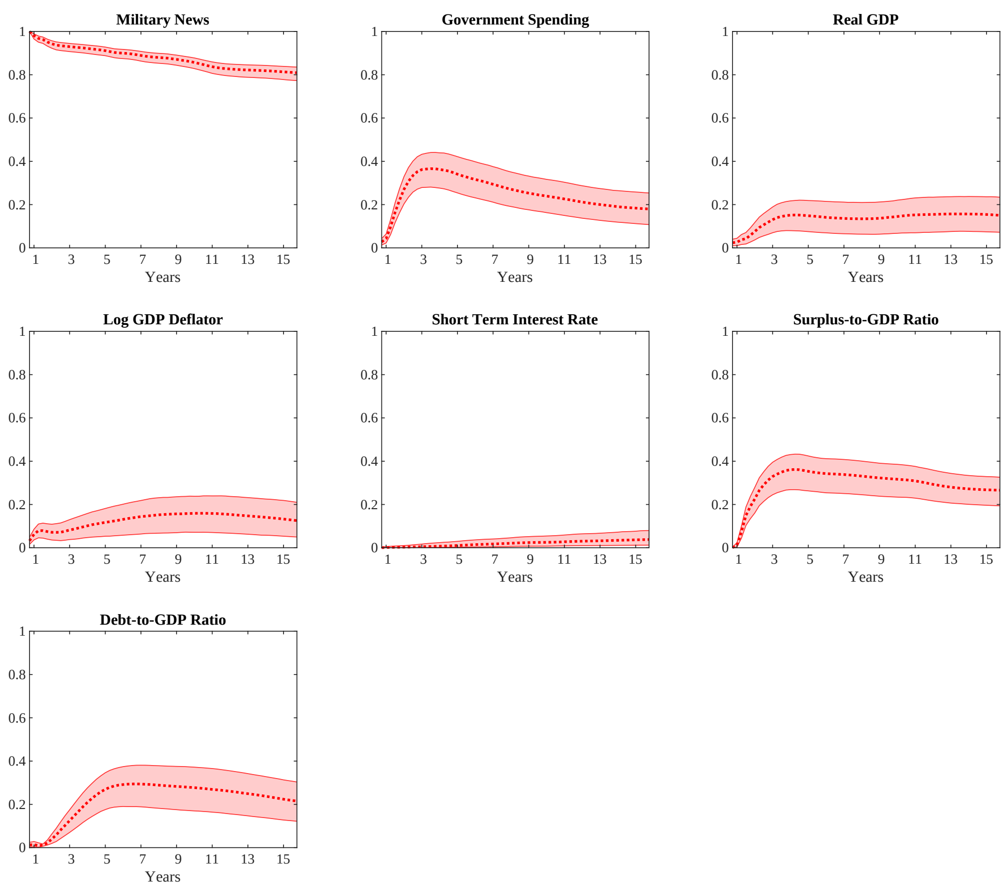
To estimate the VAR model, we can write it in matrix form as $\mathbf{Y} = \mathbf{X}\mathbf{B}' + \mathbf{U}$. Denoting T the length of the sample, n the number of variables, and p the number of lags in the VAR, $\mathbf{Y} = (\mathbf{y}_1', \ldots, \mathbf{y}_T')'$ is a $T \times n$ matrix, $\mathbf{X} = (\mathbf{x}_1', \ldots, \mathbf{x}_T')'$ is a $T \times K$ matrix, where $K = np + 1$, and $\mathbf{U} = (\mathbf{u}_1', \ldots, \mathbf{u}_T')'$ is a $T \times n$ matrix. The vector of innovations $\mathbf{u}_t$ is assumed to be independently and identically distributed $\mathcal{N}(0, \mathbf{\Sigma})$.

The NIW family of distributions is conjugate for this class of models. If the prior distribution over the parameters is $NIW(\underline{\nu}, \underline{\mathbf{S}}, \underline{\boldsymbol{b}}, \underline{\mathbf{V}})$, then the posterior distribution over the parameters is $NIW(\overline{\nu}, \overline{\mathbf{S}}, \overline{\boldsymbol{b}}, \overline{\mathbf{V}})$, where $\overline{\boldsymbol{b}} = \texttt{vec}\left(\overline{\mathbf{B}}\right)$, $\overline{\mathbf{V}} = \left(\underline{\mathbf{V}}^{-1} + \mathbf{X}'\mathbf{X}\right)^{-1}$, $\overline{\mathbf{B}} = \overline{\mathbf{V}}\left(\underline{\mathbf{V}}^{-1}\underline{\mathbf{B}} + \mathbf{X}'\mathbf{X}\hat{\mathbf{B}}\right)^{-1}$, $\hat{\mathbf{B}} = \left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'\mathbf{Y}$, and $\overline{\mathbf{S}} = \hat{\mathbf{S}} + \underline{\mathbf{S}} + \hat{\mathbf{B}}'\mathbf{X}'\mathbf{X}\hat{\mathbf{B}} + \underline{\mathbf{B}}'\underline{\mathbf{V}}^{-1}\underline{\mathbf{B}} - \overline{\mathbf{A}}'\overline{\mathbf{V}}^{-1}\overline{\mathbf{A}}$, $\hat{\mathbf{S}} = \left(\mathbf{Y} - \mathbf{X}\hat{\mathbf{B}}\right)'\left(\mathbf{Y} - \mathbf{X}\hat{\mathbf{B}}\right)$, and $\overline{\nu} = T + \underline{\nu}$. The NIW posterior distributions defined above can be factored into the following conditional and marginal posterior distributions: $\mathcal{N}\left(\overline{\boldsymbol{b}}, \mathbf{\Sigma} \otimes \overline{\mathbf{V}}\right)$ and $p(\mathbf{\Sigma}|\mathbf{y}) \sim \mathcal{IW}\left(\overline{\mathbf{S}}, \overline{\nu}\right)$. This structure allows to independently draw from the posterior.

C Forecast Error Variance Decomposition

In Figure C.1, we report the Forecast Error Variance Decomposition (FEVD) for the baseline results of Figure 1. The darker (lighter) shaded area represents the central 90% posterior band. The darker solid line stands for the median estimates. As can be seen from the figure, at business-cycle frequencies, the military spending news shock explains about 30%-40% of the variance of the unexpected movements in government spending, whereas it explains about 10% of the variance of real GDP and between 10% and 20% of the variance of the price level.

Figure C.1: Forecast Error Variance Decomposition for Military News Shock

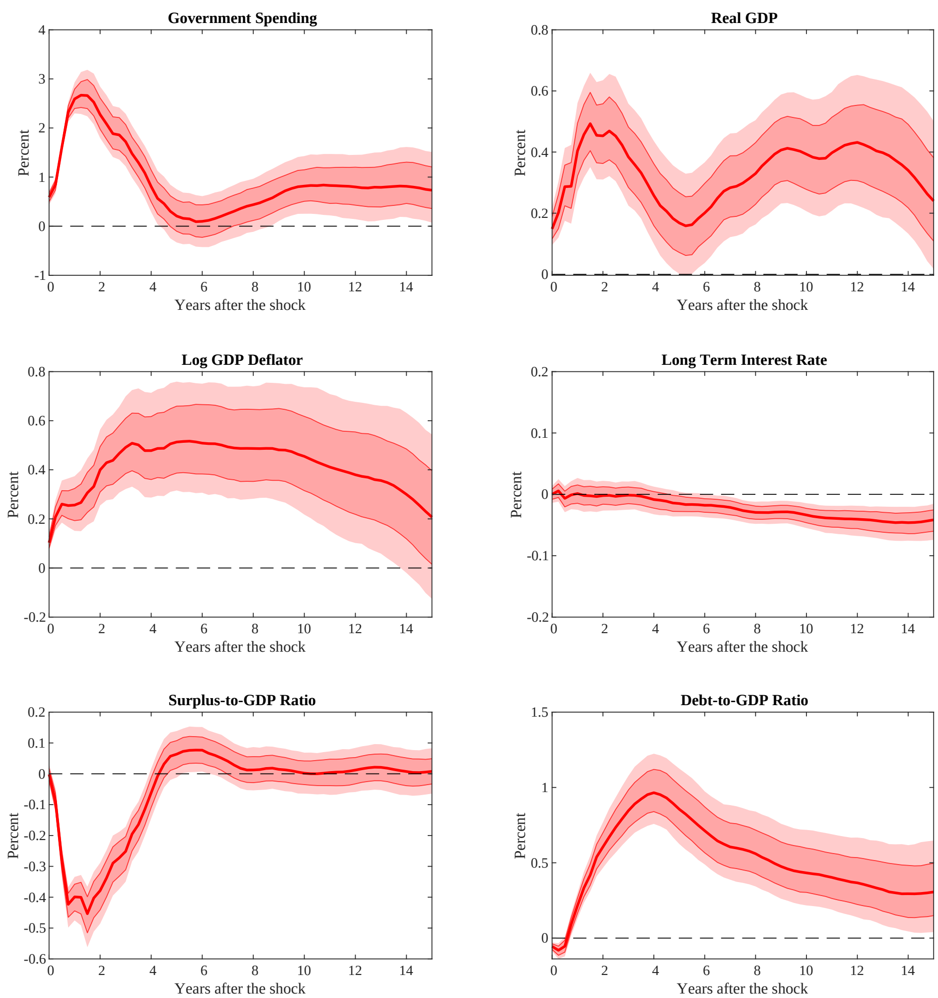


Notes. The FEVD is based on an estimated VAR with sixty lags of military spending news, government spending, real per-capita GDP, GDP deflator, short-term interest rate, fiscal deficit to GDP ratio and government debt to GDP ratio. Military spending news is ordered first in the Cholesky factorization. Output, government spending, and the GDP deflator enter the VAR in log-levels. The size of the shock is normalized such as to increase government spending by 1 percent of GDP over the first year after the shock, based on a median G/Y ratio of 19%. The darker (lighter) shaded area represents the 90% HPD interval. The dotted line stands for the median estimates. Results are based on 5000 posterior draws.

D Adding the long-term interest rate

In this section, we replace the short-term nominal interest rate with the yields on the ten-year government bond in the baseline seven variable VAR(60) of Section 3.1. The results are reported in Figure D.1. The response of the long-term nominal interest rate to the military spending news in Figure D.1 is qualitatively very similar to the effects on the short-term nominal interest rate in Figure 1, with median values that appear more accurately estimated and somewhat larger at longer horizons (around -5 basis points) than their short-term counterparts. The estimated dynamic effects of government spending on the remaining variables of the VAR (including prices) are indistinguishable from those reported in Figure 1 of the main text.

Figure D.1: ADDING THE LONG TERM INTEREST RATE



Notes. See notes to Figure 1.

E A Brief Narrative Account of Major Federal R&D Programs

In this Appendix, we provide a brief narrative account of the major public R&D programs funded in the United States over our long historical sample. Although the data includes spending at both the federal and the state and local levels, the discussion below focuses on federal funding towards R&D because it represents about 90% of the total public expenditure on R&D and it underwent major shifts during the XXth century. In contrast, state and local R&D public funds have grown steadily over time and have not experienced abrupt variations.

From the end of the XIX^{th} century to World War I. Dupree (1986) surveys the history of federal investment in Research & Development, from the creation of the United States until the outbreak of World War II. From 1890 to 1940, R&D expenditure represented 1% or less of the total federal budget. Agricultural and natural-resource oriented research, such as the Geological survey and the weather bureau, were far more dominant targets of public spending at the beginning of the XX^{th} century. Indeed, our reconstructed estimates indicate that in 1900, the Department of Agriculture was responsible for 70% of all federal R&D outlays. Its activities included the establishment of weather stations and laboratories, with the objective of preventing disease and improving farm productivity.

The beginning of the XX^{th} century saw the creation of various federal agencies, whose objective was to provide support to business activities and to address national objectives. Examples include the Public Health Service and, within it, the Hygienic Laboratory, predecessor of the National Institutes of Health, established in 1901. In the same year, the National Bureau of Standards (predecessor of the National Institute of Standards and Technology) was established to maintain standards of weights and measures in the face of rapid technological expansion.

World War I and the interwar period. World War I spurred new research efforts, and for the first time defense and national security started rivaling agricultural research. This includes the creation of the National Advisory Committee for Aeronautics, the predecessor of NASA, formed in 1915. There was not, however, a governmental agency for federal R&D with an organization structure similar to the department of Agriculture, with much of the research done in support of the war efforts being coordinated by the National Research Council, an advisory arm of the National Academy of Sciences. In the meantime, social sciences became more prominent, with the Bureau of the Census and the Bureau of Labor Statistics playing an important role within the departments of commerce and labor. During the New Deal era, federal research in health expanded and federal funding to the Public Health Service increased as part of the Social Security Act. A major achievement was the growth of the National Institutes of Health (NIH), established in 1931, and expanded in 1937 with the creation of the National Cancer Institute.

World War II and the Manhattan project. The war constituted a revolution in both the scale and the scope of federal R&D. Just before the United States entered the war, President Roosevelt set up the Office of Scientific Research and Development (OSRD), which was responsible for coordinating R&D efforts in support of the war. Large numbers of academic researchers were mobilized to work in their own institutions' laboratories on wartime R&D projects. This was a key difference with World War I, when scientists working on military projects had been recruited by military agencies. Another important innovation was the establishment of R&D contracts as a mechanism to pay for private performance of work whose approach and outcome could not be specified precisely in advance. Importantly, the federal government agreed to compensate university and industry performers for the indirect or over-head costs of R&D undertaken as part of grants and contracts, in addition to paying for direct expenses. Moreover, to carry out the vastly increased scale of R&D during World War II, major investments were made in government research laboratories (National Research Council, 1995). The largest and most notable of all projects was the Manhattan Engineering District, which was responsible for the development of the atomic bomb. At its peak in 1944, the Manhattan project accounted for nearly one-tenth of all public and private R&D performed in the United States.

In the same year, President Roosevelt asked Vannevar Bush, then director of OSRD, to 'export' the wartime R&D experience to a peacetime institution. The celebrated Bush (1954) report was delivered to President Truman in July 1945. It argued that knowledge and scientific research was an essential ingredient for improving the nation well being, health, economic growth, and national security. Moreover, the report stated that the the federal government had an important responsibility to support both scientific research and the training of new scientific talents. The key recommendation of the report was the establishment of a central research funding agency, initially called the National Research Foundation, to implement those responsibilities.

The Post-WWII Scientific Establishment. After the war, and heavily influenced by the vision laid out by the Bush report, the wartime scientific efforts were consolidated through the creation, after much congressional debate, of the National Science Foundation in 1950. Major increases in R&D efforts, including the creation of DARPA and NASA, followed the Soviet launch of the Sputnik satellite in 1958. This event revealed that the United States had fallen behind the Soviets in space technology. In 1961, president Kennedy kick-started the Apollo program by which NASA landed on the moon in 1969. The conclusion of the Apollo program led to a decline in federal R&D spending, which did not reach its 1960s peak in real per capita terms until the 1980s.

Reagan and the Strategic Defense Initiative. The 1980s witnessed large increases in defense R&D by the Reagan administration, including the Strategic Defense Initiative (SDI, popularly known as 'Star Wars'). This also was motivated by concerns about the Soviet Union and a desire to achieve technological superiority. Defense R&D spending peaked again in 1987, having doubled since the beginning of the 1980s, and generally declined through the 1990s after the fall of the USSR.

Health and Defense R&D at the end of the XX^{th} century. At the end of the 1990s, a major shift occurred with the doubling of the budget for medical research at the National Institutes for Health from 1998 to 2003. A third major boom in defense R&D was triggered by 9/11 in 2001 and lasted until the beginning of the Obama administration in 2008.

In summary, the narrative evidence discussed in this Appendix highlights that, especially compared to other types of government spending, public R&D was mostly driven by scientific, military and ideological goals, rather than by the endogenous policy response to the state of the U.S. economy. Accordingly, we propose to identify exogenous movements in public R&D using the short-run restrictions that while no macroeconomic variable can explain a large share of public R&D variation in the short-run (within the first year after the shock), public R&D is allowed (but not required) to have a significant impact on the economy in the short-run.

F Historical Decomposition of Public R&D based on the Military Spending Shocks and Time Series of the Military Spending Shocks

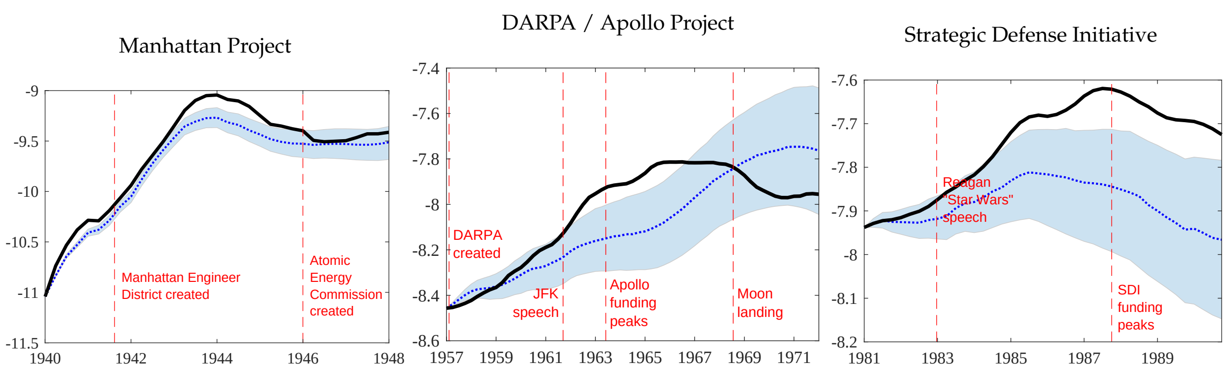
In this appendix, we perform a historical decomposition of the time series public R&D (solid black lines) around three historical major events: (a) the Manhattan Project, (b) the DARPA / Apollo project, (c) the Strategic Defense Initiative. The blue lines and associated 68% central posterior bands in Figure F.1 represent the component of public R&D that can be explained by the military spending shock from a VAR(60) using military spending news, real public R&D per-capita, real GDP per-capita, real government spending per-capita, the GDP deflator, the short-term nominal interest rate, government deficit to GDP ratio and public debt to GDP ratio. The eight quarter moving-average of the time series of the military spending shocks (and associated 68% credible set) is plotted in the second panel.

The top panel of Figure F.1 makes clear that the military spending shock can explain a significant share of the public R&D increase around the Manhattan project, which occurred during World War II and was part of military R&D spending, but can account only for a limited extent of the public R&D increases in the other historical events, which occurred in peacetime. The historical period associated with the DARPA/Apollo episode is a mixture of military and non-military spending, while the Reagan SDI increases are mostly defense spending that occurred, however, entirely during a peacetime period.

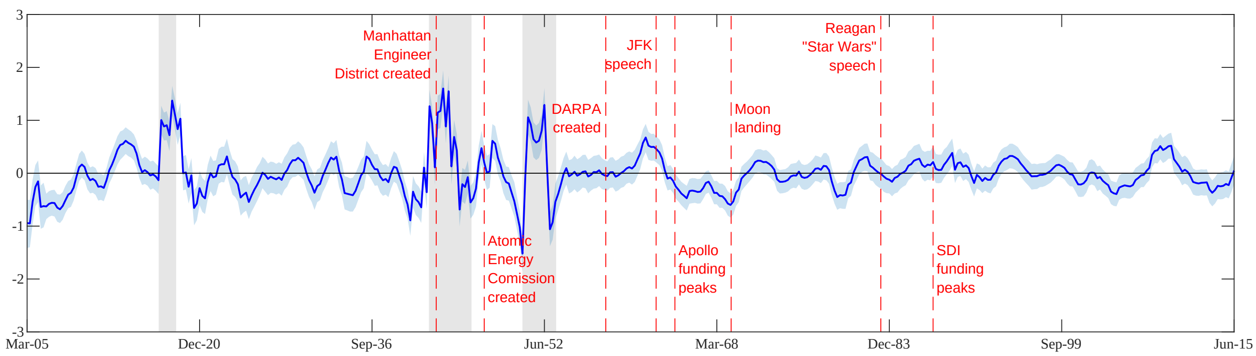
Finally, the time series of the military spending shock behind Figure F.1 and the time series of the public R&D shock behind Figure 5 have a correlation of 0.17.

Figure F.1: Historical Analysis of Public R&D and Military Spending Shocks

(a) Historical Decomposition of Public R&D Expenditure Around Key Events



(b) Time Series of Public R&D Shocks (eight quarter moving-average)



Note: Panel (a) plots the historical decomposition of public R&D around three historical events: (i) the Manhattan project, (ii) DARPA and the Apollo program, (iii) the Strategic Defense Initiative (SDI). In each sub-panel, the solid black line is the historical surge in real per capita R&D spending by the government. The dotted blue line, and associated 68% posterior bands, show the part of the increase in R&D that can be explained by the effects of the *military spending shock*. Panel (b) plots a eight quarter moving-average of the military spending shock together with 68% posterior bands. Shaded areas represent major wars.

G Lag Length Selection

In Section 6.2, we have compared the VAR(4) in log-levels with the VAR(4) in deviations from potential and found that the latter has a much harder time to identify long-run effects. Given the popularity of detrended specifications in empirical macro, in this Appendix, we ask whether a richer lag length may ameliorate the problems of the VAR(4) that expresses output and government expenditure in deviations from potential output. We show that it does, using the three exercises below on: forecast encompassing, multiplier analysis at different horizons using different time-series models and lag length, and low-frequency predictability.

G.1 Forecast encompassing

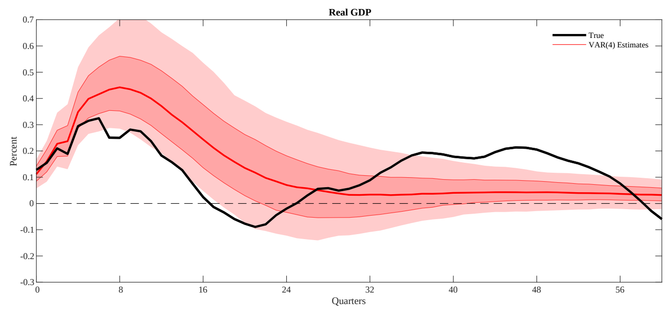
In this section, we look at one possible metrics along which to compare the predictive accuracy of our VAR(60) versus the popular VAR(4) used in the empirical macro literature. This is the encompass strategy in Chong and Hendry (1986), who recommend to estimate two competing models on artificial data generated using the estimates of the other specification to evaluate which model produces forecasts that encompass the forecasts of the other model.

We proceed in two symmetric steps. First, we generate one set of artificial data using the point estimates of the VAR (60) on actual data and then fit, on these artificial data, a VAR(4) specification. Second, we do the reverse and generate another set of artificial data, using this time the point estimates of the VAR(4) on actual data, and then fit on these artificial data a VAR(60). If the impulse responses of the VAR(60) are within the credible sets of estimates for the impulse responses of the VAR(4) when the data generating process is the VAR(4) *but* the impulse responses of the VAR(4) are outside the credible sets of estimates for the impulse responses of the VAR(60) when the data generating process is the VAR(60), then we conclude that (the forecasts of) the VAR(60) encompass (the forecasts of) the VAR(4) but the VAR(4) does not encompass the VAR(60).

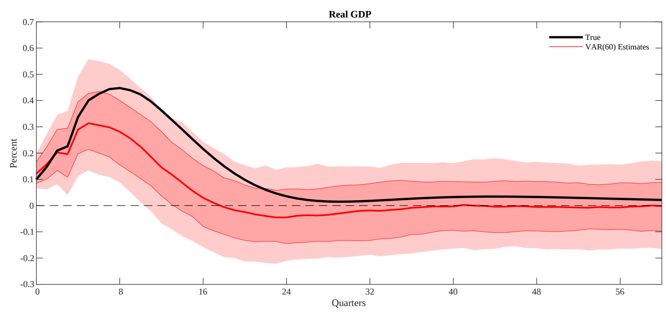
The findings from this exercise are in Figure G.1. The top panel presents the VAR(4) estimates of the output response when the data generating process is the VAR(60) whereas the bottom panel displays the VAR(60) estimates when the data generating process is the VAR(4). The top panel reveals that the VAR(4) estimates (in red) have hard time to match the true impulse response of output (in black) when the data generating process is the VAR(60). This is true not only at most

Figure G.1: ESTIMATES FROM ALTERNATIVE DGPS AND VAR SPECIFICATIONS

(a) VAR(4) estimates when DGP is VAR(60)



(b) VAR(60) estimates when DGP is VAR(4)



Note: The red solid lines represent the median posterior responses. The darker (lighter) shadow area represents the 68^{th} (90^{th}) posterior credible intervals. The black solid line refers to the true impulse response in the data generating process, which is a VAR(60) in the top row and a VAR(4) in the bottom row estimated on actual data on military spending news, government spending, real GDP per-capita, GDP deflator, short-term interest rate, fiscal deficit to GDP ratio and government debt to GDP ratio.

frequencies beyond 32 quarters (i.e. in the long-run) when the true long-run effects on output are well above the small and insignificant effects estimated by the VAR(4), but also, though at a lesser extent, in years 2 and 3 when the VAR(4) estimates are now significantly larger than in the data generating process. In sharp contrast, the bottom panel shows that, when the data generating process is the VAR(4), the true impulse responses are always inside the credible sets of estimates of the VAR(60), especially at lower frequencies.

We conclude that the VAR(60) is able to detect small and insignificant long-run effects on output when, indeed, there are none in the data generating process whereas the VAR(4) is unable to pick up any long-run effect on output when, in fact, these are large and significant in the data generating process. In the language of Chong and Hendry (1986), the forecasts of the VAR(60) encompass those of the VAR(4) but the converse is not true. We interpret these findings as further evidence that a high number of lags is desirable to draw inference on medium-term and long-run dynamics.

G.2 The Multipliers across Forecast Horizons

The empirical literature on the dynamic effects of government spending has presented a plethora of output multipliers that have been typically estimated at different short-run and business cycle horizons across papers. While the main focus of our analysis is to provide novel evidence on the impulse response of output, in Table G.1 of this Appendix we complement the analysis in the main text by computing the output multiplier of government spending at various horizons, including one quarter (Panel A) and four years (Panel B) used in earlier studies as well as fifteen years (Panel C). A main take away from this exercise is that, by providing a systematic analysis of the output multiplier across forecast horizons (within each and every specification), we are able to span the whole range of estimates available in the empirical literature. This suggests that the seemingly conflicting results in earlier work may simply reflect the fact that different studies focus on different forecast horizons.

The left portion of each panel of Table G.1 refers to VAR specifications whereas the sections on the right correspond to LP models. Each row represents a different number of lags for the relevant model in that row, ranking from a minimum of 4 lags to a maximum of 60 lags. The columns (from left to right) stand for the 5^{th}, 50^{th} and 95^{th} of the posterior distribution of the multiplier of interest for the specification in each row whereas the last column in each section, headed with $\mathcal{M} > 1$, records the share of draws for which the multiplier at the horizon in that panel and for that specification is above one.

Two main results emerge from Table G.1. First, independently of whether we use VARs or LPs and independently of the lag length selection, the entries in Panels A and B (i.e. at shorter forecast horizons) are fairly similar to each other. This suggests that omitting higher lags in either VAR or LP specifications is inconsequential for the estimates of the short-run multipliers, despite the strong evidence in Panel C of Table G.1 about the sizable bias that omitting those lags produce when estimating the long-run multiplier. Second, and again very robustly across models and specifications, the impact multiplier one quarter head (Panel A) is about twice as large as the output multipliers at two (not reported) and four years horizons (Panel B), such that the share of posterior draws for which the multiplier is above one range from 64% to 88% after one quarter but is no larger than 8% at the four year horizon. In contrast, using specifications with at least forty lags in

Table G.1: THE PRESENT VALUE MULTIPLIER, $\mathcal{M}$, ACROSS FORECAST HORIZONS

Panel A. Multiplier at 1-quarter horizon

	Vector AutoRegressions (VAR)				Bayesian Local Projections (LP)			
No. lags (p)	$5^{th}pct$	$50^{th}pct$	$95^{th}pct$	$\mathcal{M} > 1$	$5^{th}pct$	$50^{th}pct$	$95^{th}pct$	$\mathcal{M} > 1$
4	0.71	1.34	2.43	80%	0.61	1.26	2.32	73%
10	0.78	1.39	2.37	84%	0.68	1.39	2.68	80%
20	0.80	1.35	2.15	85%	0.64	1.29	2.35	75%
30	0.70	1.21	1.92	74%	0.49	1.17	2.27	64%
40	0.79	1.30	2.05	82%	0.54	1.30	2.77	72%
50	0.85	1.31	2.00	85%	0.60	1.35	3.03	76%
60	0.89	1.35	2.06	88%	0.58	1.40	3.35	76%

Panel B. Multiplier at 4-year horizon

	Vector AutoRegressions (VAR)				Local Projections (LP)			
No. lags (p)	$5^{th}pct$	$50^{th}pct$	$95^{th}pct$	$\mathcal{M} > 1$	$5^{th}pct$	$50^{th}pct$	$95^{th}pct$	$\mathcal{M} > 1$
4	0.46	0.71	0.98	4%	0.63	0.72	0.83	0%
10	0.57	0.78	1.00	5%	0.63	0.73	0.83	0%
20	0.56	0.76	0.95	2%	0.60	0.70	0.80	0%
30	0.49	0.67	0.86	0%	0.59	0.71	0.84	0%
40	0.51	0.76	1.03	8%	0.68	0.81	0.96	2%
50	0.56	0.78	1.01	6%	0.62	0.75	0.90	0%
60	0.50	0.70	0.90	1%	0.56	0.66	0.77	0%

Panel C. Multiplier at 15-year horizon

	Vector AutoRegressions (VAR)				Local Projections (LP)			
No. lags (p)	$5^{th}pct$	$50^{th}pct$	$95^{th}pct$	$\mathcal{M} > 1$	$5^{th}pct$	$50^{th}pct$	$95^{th}pct$	$\mathcal{M} > 1$
4	0.28	0.67	1.19	12%	0.72	0.95	1.21	37%
10	0.52	1.00	1.74	49%	0.85	1.13	1.48	77%
20	0.45	1.04	1.90	54%	0.68	0.93	1.23	34%
30	0.35	0.95	1.73	44%	1.13	1.58	2.29	99%
40	0.53	1.77	4.21	83%	2.73	3.74	5.65	100%
50	0.94	2.30	5.14	94%	2.11	2.96	4.46	100%
60	1.03	2.08	3.90	96%	1.73	2.26	3.01	100%

Notes: The VAR and LP specifications use seven variables: military spending news, government spending, real per-capita GDP, GDP deflator, short-term interest rate, fiscal deficit to GDP ratio and government debt to GDP ratio. Military spending news is ordered first in the Cholesky factorization of the VAR. The definition of variables and the definition of the present value multiplier over 60 quarters horizon, $\mathcal{M}$, follow Ramey and Zubairy (2018) and Mountford and Uhlig (2009). Results are based on 5000 posterior draws, discarding explosive roots. Each row refers to the estimates of a different specification of either the VAR(p) or the LP using p lags of all variables. The number of lags, p, selected in each specification is reported in the first column. The columns $5^{th}pct$, $50^{th}pct$ and $95^{th}pct$ present the 5^{th}, 50^{th} and 95^{th} percentiles of the posterior distribution of $\mathcal{M}$. The columns $\mathcal{M} > 1$ report the share of posterior draws for which the cumulated response of GDP over the cumulated response of government spending is larger than one.

Panel C, we find that the likelihood that the long-run multiplier is above one exceeds 80%.

G.3 Low-frequency predictability

One way to understand the discrepancy between low- and high-order VARs/LPs is to realize that the measurement of the shock that is projected onto the endogenous variables in both methods differs depending on the number of lags used as controls. Both in the VAR, where the military news series is ordered first in a Cholesky factorization, and in the lag-augmented LP, the identified shock corresponds to the residual of a regression of military spending news on p lags of itself and all other variables. If lagged endogenous variables help predict the military spending news, then controlling for those lags will affect the results in a classical omitted variable problem.

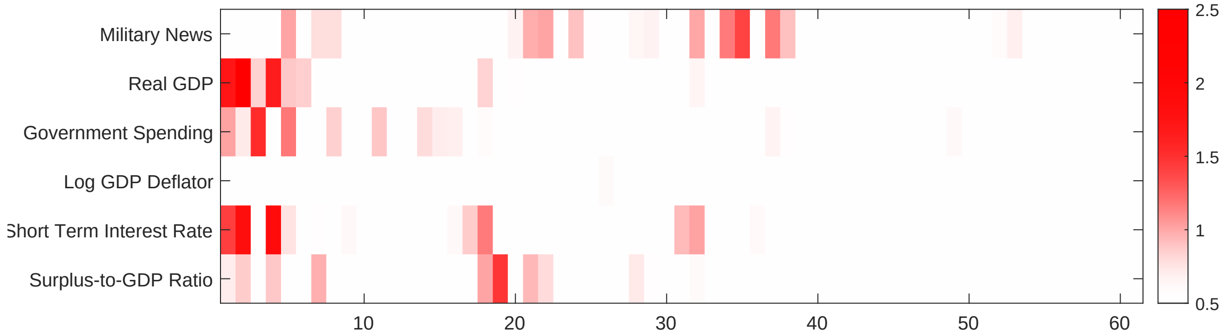
In Figure G.2, we report the outcome of regressing military spending news on 60 lags of itself and all other variables, using the same prior shrinkage described in Section 2.2. The residual of this regression corresponds to the identified shock used by both the VAR(60) and the LP(60). In the figure, column i and row j jointly identify the i-th lag of the j-th variable, and the shade of the color grows with the ratio between the absolute value of the associated regression coefficient posterior mean and its posterior standard deviation. A darker shade indicates greater significance in predicting future military spending news. Two main results can be taken away from Figure G.2. First, the military spending news are highly predictable, especially using lags of the military spending news itself, and every other variable except perhaps the GDP deflator. Second, the most systematic pattern is associated with the high significance of the estimated coefficients on lags up to about 40 quarters. Beyond that, all lagged variables appear to lose their ability to predict future military spending news.

Two comments are worth noting. First, the findings in Figure G.2 suggest that a generous lag length selection is not only desirable (as shown in the main text and this Appendix) but it may be, in fact, also necessary to isolate exogenous movements in military spending news, and therefore in government spending. Second, the omission of longer lags may also be responsible (over and above any possible truncation bias) for the inability of conventional specifications such as the VAR(4) and LP(4) to capture medium-term and long-run dynamics, especially when using detrended data. It should be noted, however, that the omitted variable bias hinted by Figure G.2 seems to have less of an impact at shorter horizons and when the data enter the model in log-levels.[3]

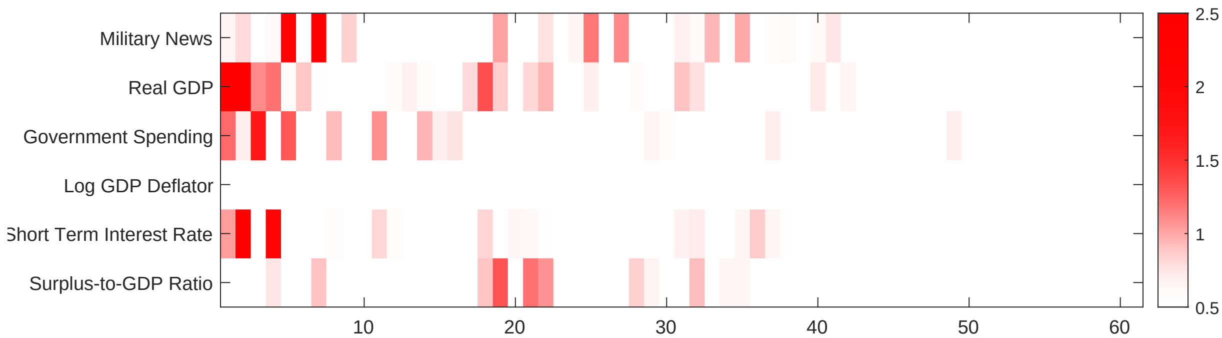
[3]Low-frequency predictability is related to but distinct from the low-frequency covariability in Müller and Watson (2018). In that paper, the authors seek to draw inference on the *contemporaneous* relationship between the low-frequency components of two time-series,

Figure G.2: SIGNIFICANCE OF LAGS

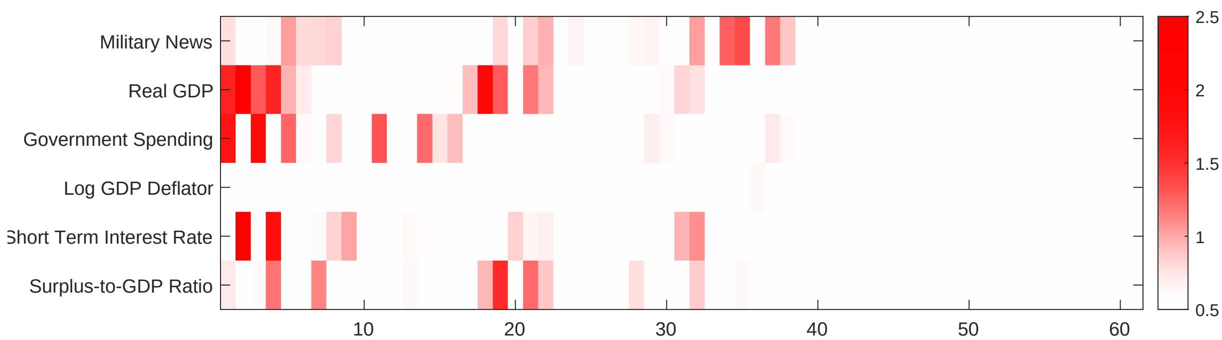
(a) Variables in Log-levels



(b) Variables scaled by previous-quarter GDP



(c) Variables scaled by potential GDP



Notes. Military spending news is projected on sixty lags of itself, government spending, GDP, GDP deflator, treasury bills, deficit to GDP ratio and debt to GDP ratio. Columns refer to the regressors and rows refers to their lags. Darker shades of red indicate higher predicting power as measured by a higher value of the ratio between (the absolute value of) the posterior mean of the estimated coefficient and its posterior standard deviation.

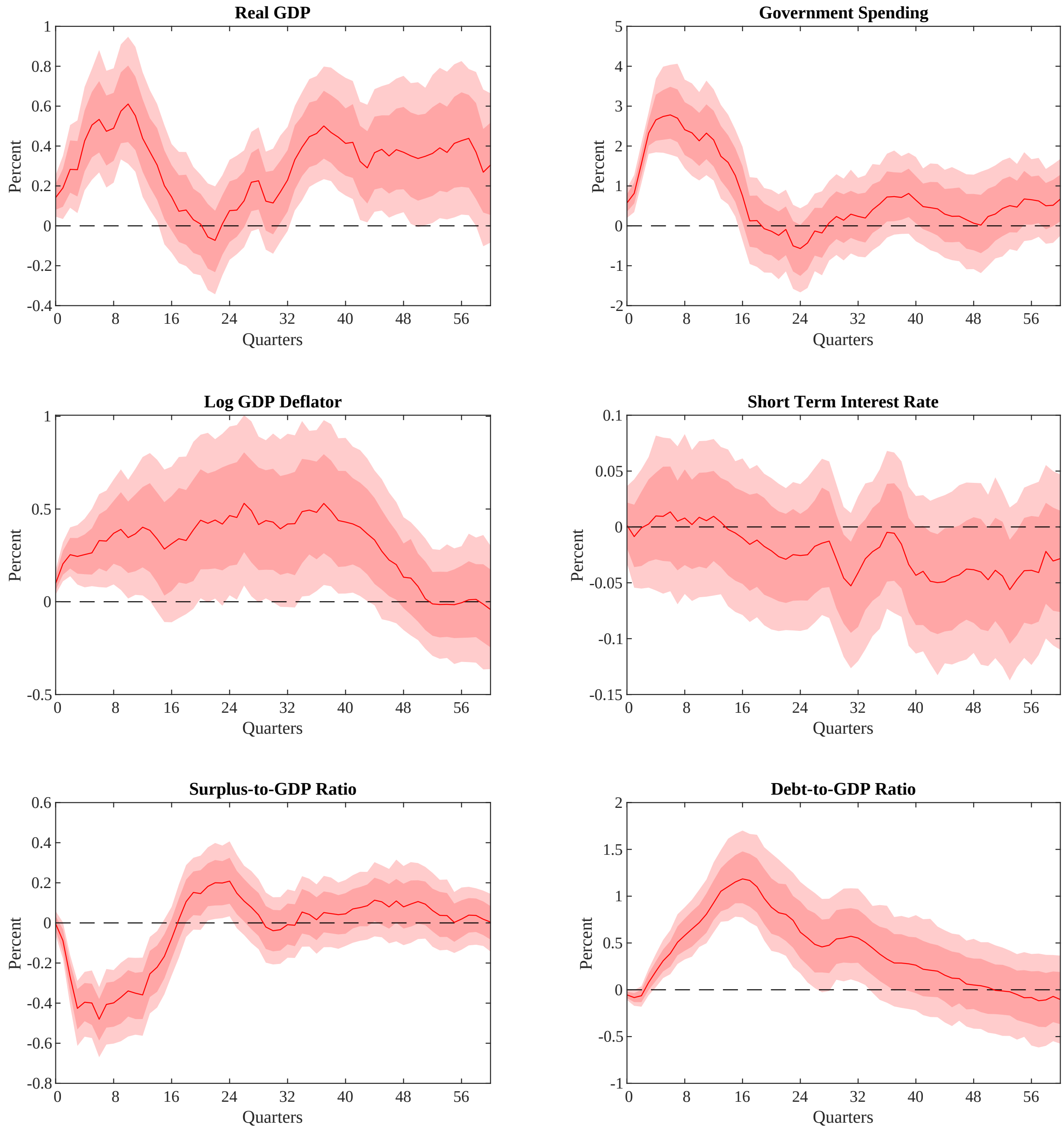
whereas here we are interested in the cross-frequency correlations among a set of variables, at potentially very long leads and lags.

H Results based on Bayesian Local Projections

In this section, we report the full set of results associated with the estimates of the LP(60), whose impulse responses for real GDP and total factor productivity have been reported among the sensitivity checks of Figure 10 (top row). In particular, in Figure H.1 we present also the impulse response for the price level, government spending, the short-term nominal interest rate, the fiscal surplus to GDP ratio and the public debt-to-GDP ratio. All estimates are fairly similar to those based on the VAR(60), consistent with the finding in Plagborg-Møller and Wolf (2021) that VARs and LPs estimate the same impulse responses (in large samples) whenever the span of the forecast horizon is as large as the number of lags used to estimate each model. We observe that the Local Projection specification leads to less precise estimates of the impulse response functions, for reasons described in the main text.

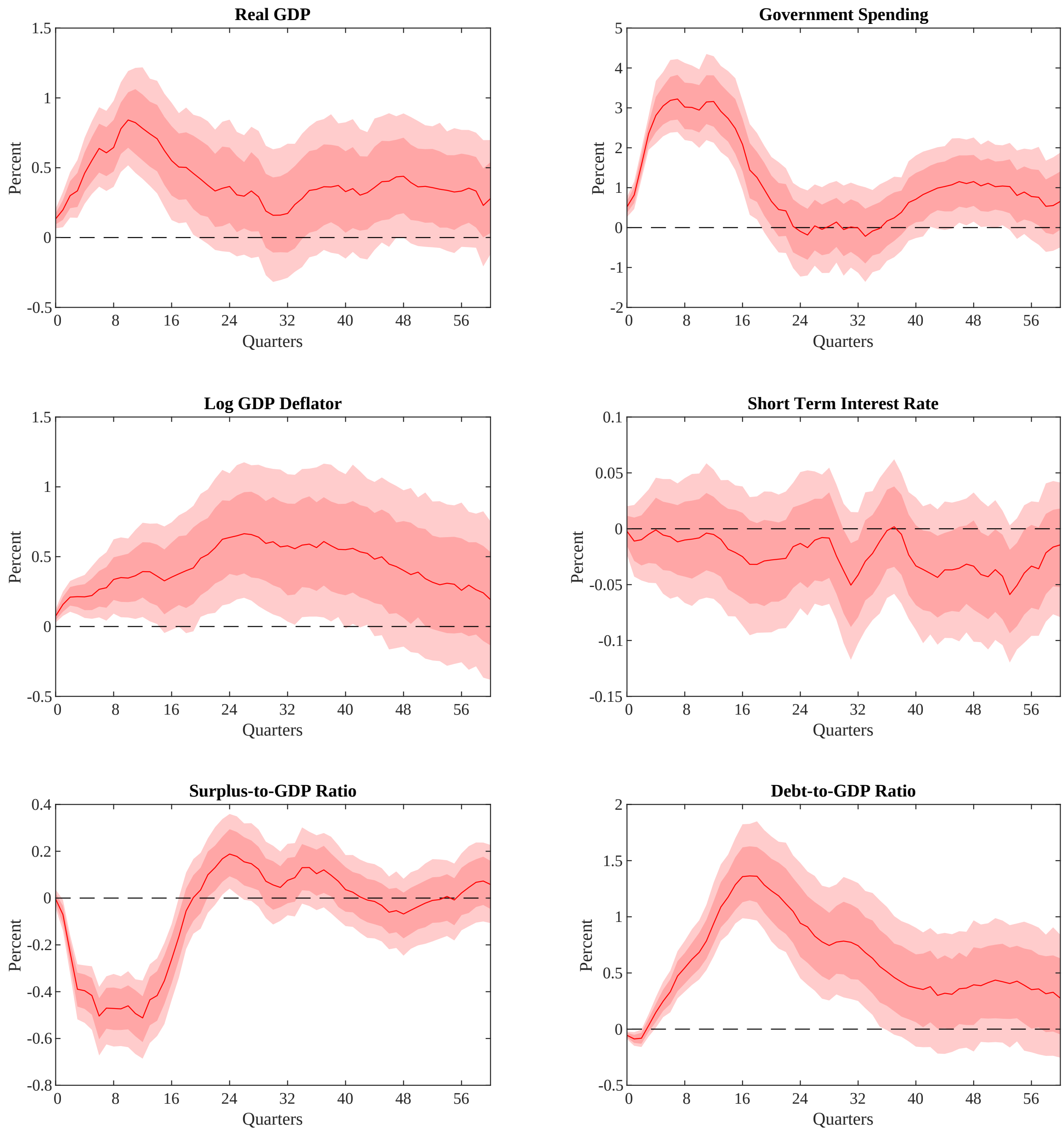
For completeness, we also report the results of a Bayesian LP that uses only four lags in the specification in levels. With a lower number of lags, much of the long run effects on output are smaller ans less significant. Furthermore, the dynamics of government spending and output are different form those obtained with a richer lag structure in Figure H.1. As we have seen in Appendix G.2, these differences lead to long-run multiplier estimates that are below one whenever an insufficient number of lags is used in specifications in which output and government spending are detrended using potential output.

Figure H.1: IMPULSE RESPONSES TO MILITARY NEWS SHOCK USING LP(60)



Notes. The impulse responses are based on an estimated Bayesian LP with sixty lags of military spending news, government spending, real per-capita GDP, GDP deflator, short-term interest rate, fiscal deficit to GDP ratio and government debt to GDP ratio. Contemporaneous military news is taken to be the shock of interest. Output and government spending are expressed in log levels. The size of the shock is normalized such as to increase government spending by 1 percent of GDP over the first year after the shock. The darker (lighter) shaded area represents the 68% (95%) HPD interval. The darker solid line stands for the median estimates. Results are based on 5000 posterior draws.

Figure H.2: IMPULSE RESPONSES TO MILITARY NEWS SHOCK USING LP(4)



Notes. The impulse responses are based on an estimated Bayesian LP with four lags of military spending news, government spending, real per-capita GDP, GDP deflator, short-term interest rate, fiscal deficit to GDP ratio and government debt to GDP ratio. Contemporaneous military news is taken to be the shock of interest. Output and government spending are expressed in log levels. The size of the shock is normalized such as to increase government spending by 1 percent of GDP over the first year after the shock. The darker (lighter) shaded area represents the 68% (95%) HPD interval. The darker solid line stands for the median estimates. Results are based on 5000 posterior draws.

I Sensitivity to the Prior Tightness

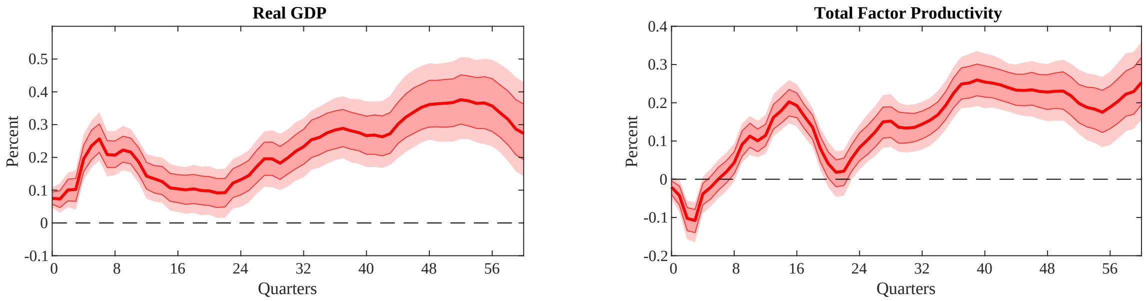
In this section, we present the impulse responses of output and productivity based on variants of the VAR(60) estimated in the main text. The four versions differ in the tightness of the prior hyper-parameters λ (which controls the tightness of the "Minnesota" prior) and θ (which controls the tightness of the "sum of coefficients" prior). The results are reported in Figures I.1 and Figure I.2 below.

In Figure I.1, we perform the analysis that varies the hyper-parameter λ while keeping fixed θ at the baseline value, whereas, in Figure I.2, we conduct the opposite exercise: we vary the hyper-parameter θ while keeping λ fixed at the value estimated using the method in Giannone et al. (2015). Each of the rows starts with relatively uninformative priors, which become progressively tighter going down the figure.

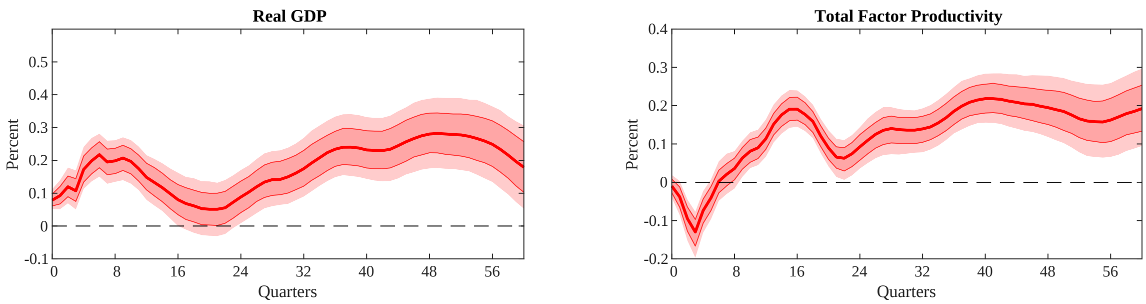
Three main results emerge from this sensitivity analysis. First, progressively tighter priors on the λ hyperparameter of the VAR(60), which are visible going down the rows of Figure I.1, are associated both with smoother shapes of the impulse responses and also with progressively smaller long-run effects. Second, despite this progressively increase in tightness, it is still the case that government spending has non-negligible and significant long-run effects on output and productivity, even in the most conservative specification of $\lambda = 0.1$ in the fourth row. Third, the tightness of the prior hyperparameter θ has some effect on the magnitude of the output and TFP responses at the 15 year horizon in Figure I.2, with stronger effects associated with tighter priors. This is the case, for instance, for the baseline $\theta = 0.001$, which is estimated by maximizing the marginal likelihood as in Giannone et al. (2015). The overall shape and significance of the responses, however, are unchanged even with the relatively loose prior of $\lambda = 1$.

Figure I.1: IMPULSE RESPONSE FUNCTIONS UNDER ALTERNATIVE TIGHTNESS OF PRIOR

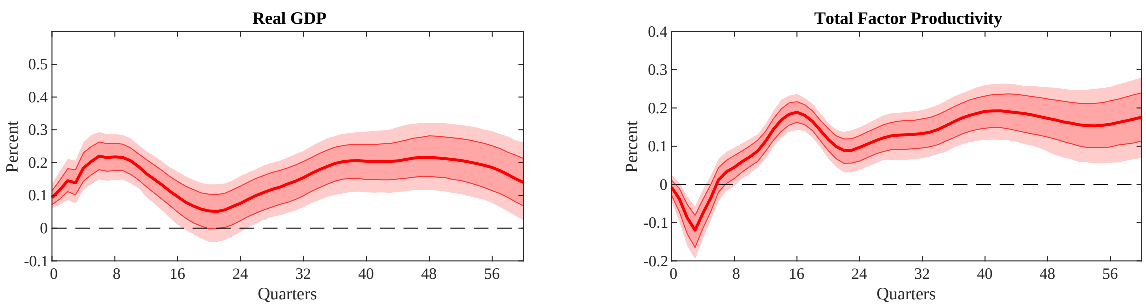
(a) $\lambda = 1$



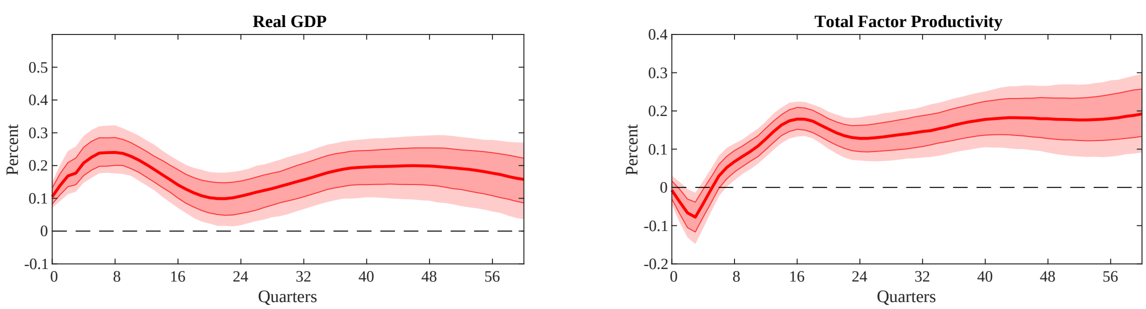
(b) $\lambda = 0.4$



(c) $\lambda = 0.2$



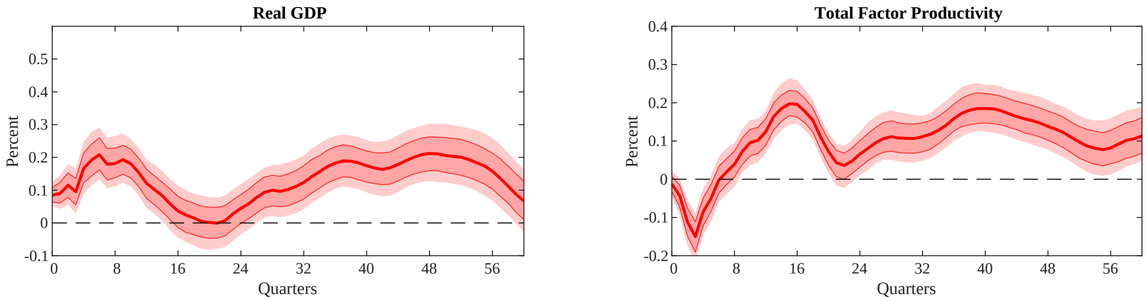
(d) $\lambda = 0.1$



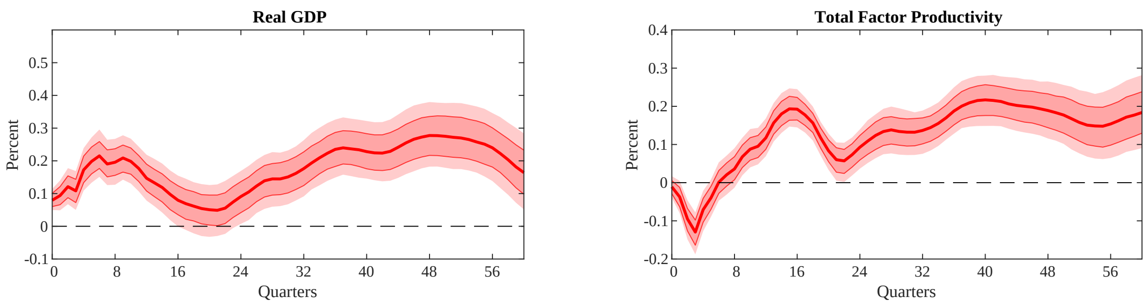
Note: The solid lines represent the median posterior response. The darker (lighter) shadow area represents the 68^{th} (95^{th}) HPD interval. All specifications use sixty lags of military spending news, government spending, GDP, GDP deflator, treasury bills, deficit to GDP ratio and debt to GDP ratio as in the baseline model. In each row, the parameter λ that governs the tightness of the Minnesota prior in equation (4) takes a different value, ranging from 1 in the top row, to 0.4 and 0.2 in the middle rows, and finally 0.1 in the bottom row. In all cases, the prior hyperparameter θ for the "single unit root" dummy is set at the baseline value of $\theta = 0.001$ that we use as baseline specification in the main text.

Figure I.2: IMPULSE RESPONSE FUNCTIONS UNDER ALTERNATIVE TIGHTNESS OF PRIOR

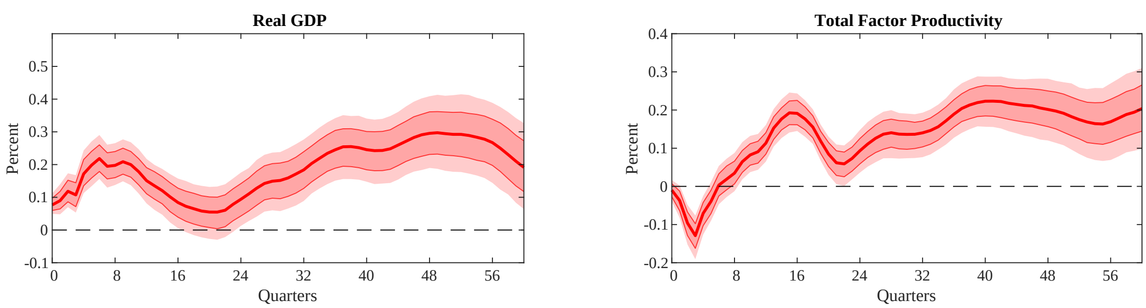
(a) $\theta = 1$



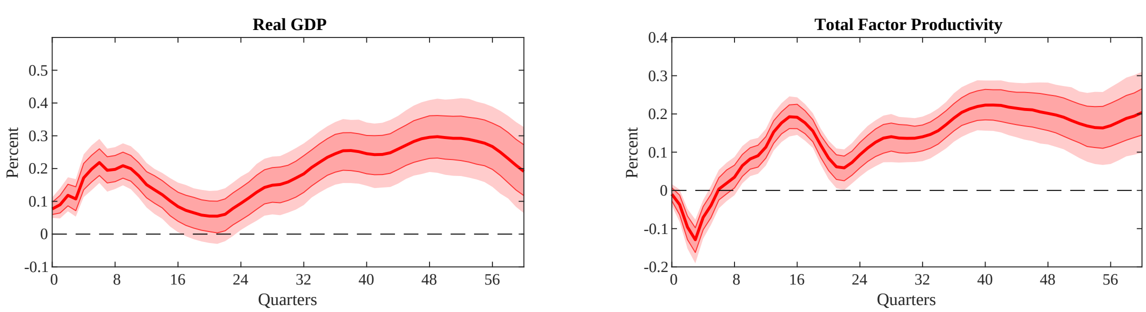
(b) $\theta = 0.1$



(c) $\theta = 0.01$



(d) $\theta = 0.001$



Note: The solid lines represent the median posterior response. The darker (lighter) shadow area represents the 68^{th} (95^{th}) HPD interval. All specifications use sixty lags of military spending news, government spending, GDP, GDP deflator, treasury bills, deficit to GDP ratio and debt to GDP ratio as in the baseline model. In each row, the parameter θ that governs the tightness of the "dummy initial observation" prior in equation (4) takes a different value, ranging from 1 in the top row, to 0.1 and 0.01 in the middle rows, and finally 0.001 in the bottom row. In all cases, the prior hyperparameter λ for the Minnesota prior is set at the baseline value of $\lambda = 0.44$ that we use as baseline specification in the main text.