Inference for Impulse Responses under Model Uncertainty

Maastricht University

Lenard Lieb, Stephan Smeekes

Inference for Impulse Responses under Model Uncertainty

RM/17/022

GSBE

Maastricht University School of Business and Economics
Graduate School of Business and Economics

P.O Box 616
NL- 6200 MD Maastricht
The Netherlands

Inference for Impulse Responses under Model Uncertainty*

Lenard Lieb† Stephan Smeekes‡

Abstract

In many macroeconomic applications, impulse responses and their (bootstrap) confidence intervals are constructed by estimating a VAR model in levels - thus ignoring uncertainty regarding the true (unknown) cointegration rank. While it is well known that using a wrong cointegration rank leads to invalid (bootstrap) inference, we demonstrate that even if the rank is consistently estimated, ignoring uncertainty regarding the true rank can make inference highly unreliable for sample sizes encountered in macroeconomic applications. We investigate the effects of rank uncertainty in a simulation study, comparing several methods designed for handling model uncertainty. We propose a new method - Weighted Inference by Model Plausibility (WIMP) - that takes rank uncertainty into account in a fully data-driven way and outperforms all other methods considered in the simulation study. The WIMP method is shown to deliver intervals that are robust to rank uncertainty, yet allow for meaningful inference, approaching fixed rank intervals when evidence for a particular rank is strong. We study the potential ramifications of rank uncertainty on applied macroeconomic analysis by re-assessing the effects of fiscal policy shocks based on a variety of identification schemes that have been considered in the literature. We demonstrate how sensitive the results are to the treatment of the cointegration rank, and show how formally accounting for rank uncertainty can affect the conclusions.

JEL Classification: C15; C32; C52; E62.

Keywords: Impulse response analysis; cointegration; model uncertainty; bootstrap inference; fiscal policy shocks.

*We thank Marco Avarucci, Nalan Bastürk, Hanno Reuvers and Peter Schotman for their very helpful discussions and suggestions. We also thank conference and seminar participants at the CFE 2015, London, the NESG 2016, Leuven, and the econometrics seminar at the University of Cologne for their constructive comments. The second author thanks the Netherlands Organization for Scientific Research (NWO) for financial support.

†Department of Economics, Maastricht University, P.O. Box 616, 6200 MD Maastricht, The Netherlands. E-mail: L.Lieb@maastrichtuniversity.nl

‡Department of Quantitative Economics, Maastricht University, P.O. Box 616, 6200 MD Maastricht, The Netherlands. E-mail: S.Smeekes@maastrichtuniversity.nl

1 Introduction

Vector autoregressions (VAR) and, more importantly, their implied impulse responses (IR) are essential tools for applied macroeconomists to investigate the dynamic propagation of (structural) shocks. While VARs fitted to macroeconomic data can incorporate information about unit roots and possible cointegration relations, this evidence is regularly ignored in applied work and inference for IR coefficients is usually based on the VAR specification in levels or first-differences. A common argument for the specification in levels is that estimation by ordinary least-squares (OLS) and the associated traditional approach to inference – for example via an asymptotically normal (Lütkepohl, 1990) or a bootstrap (Kilian, 1998b) approximation – 'allows' for the presence of cointegration. Indeed the level specification results in consistent estimates of the VAR parameters regardless of the true underlying cointegration relations, and, for a fixed horizon, such inferential procedures remain valid for inference on IR coefficients. However, albeit asymptotically valid, confidence intervals may have poor coverage in small samples when the data are highly persistent and when considering responses at "longer" horizons (Kilian and Chang, 2000). Phillips (1998) shows theoretically that if one (or more) unit roots are present, confidence bands based on the normal approximation become invalid at "(very) long horizons", while Inoue and Kilian (2002) and Mikusheva (2012) show that the bootstrap also becomes invalid at such increasing horizons.

These seemingly contradicting theoretical results depend on the asymptotic framework considered; or more precisely on the notion of "(very) long horizon". If the considered horizon is kept fixed while the sample size is growing, one arrives at standard asymptotic results. However, if the horizon is modelled as a constant proportion of the sample size, the asymptotic distribution becomes non-standard if (near) unit root(s) are present. Of course, one can view the level specification as a particular form of misspecification in the presence of one or more unit roots; analogously, a wrongly specified vector error correction (VECM) formulation of the VAR suffers from similar shortcomings. Similarly, it is well known in the bootstrap literature that misspecification of the cointegration rank leads to an invalid bootstrap procedure (Choi, 2005; Inoue and Kilian, 2002; Mikusheva, 2012).

Within this growing horizon framework, Pesavento and Rossi (2006) construct confidence intervals for "long-horizon" IRs using local-to-unity asymptotics. The resulting confidence bands differ substantially from those obtained through traditional approaches, and suffer in turn from size distortions in short to medium horizons. Mikusheva (2012) proposes a procedure that works uniformly well over the entire parameter space and the entire trajectory of the IRs, but her approach only allows for the construction of uniformly valid inference if at most one "uncertain" (unit) root is present in the VAR. Similar settings and problems are considered by Gospodinov (2004, 2010), Gospodinov et al. (2011), Pesavento and Rossi (2007) and Wright (2000) among others, but all consider at most one unknown root near unity. This setting does not allow for uncertainty about the number of cointegrating relations (if any),

which we face in practice. Gospodinov et al. (2013) do consider the more general setting in an extensive simulation study and conclude that the applied researcher is best advised to estimate the system in levels and construct inference in a traditional way. Jardet et al. (2013) propose an averaging approach for impulse responses of potentially cointegrated VAR models. While they allow for uncertainty regarding the order of integration, their approach still requires a pre-selection of rank, and does not deal with inference explicitly.

In this paper we re-assess the construction of bootstrap confidence intervals for IRs in persistent, possibly non-stationary VARs. Our main intention is to provide the applied researcher with a more reliable and robust alternative to the traditional "levels" approach, independent of the IR horizon of interest. We approach the issue of choosing the cointegration rank from a model selection perspective, and draw inspiration from (bootstrap) methods initially designed to overcome model selection uncertainty in different contexts. In particular, we adapt the endogenous lag selection procedure of Kilian (1998a), the model averaging estimators of Hjort and Claeskens (2003) and the bagging approach proposed by Efron (2014) to the rank selection problem in VECMs. As elaborated by Leeb and Pötscher (2005), inference after model selection is difficult, and there is no guarantee that the above-mentioned methods can solve the problems in our setting.

Therefore, we draw inspiration from the Post-Selection Inference (PoSI) approach of Berk et al. (2013) proposed explicitly for dealing with inference after model selection to propose a novel way of constructing confidence bands by combining intervals of models for any rank. In our approach, labeled as *Weighted Inference by Model Plausibility* (WIMP), upper and lower bounds of all associated fixed-rank intervals are combined depending on the relative evidence for, or plausibility of, each model. Unlike many approaches considered in the VAR literature, our method does not require any pre-selection of ranks; that is, no pre-testing or selection using economic theory is needed. Instead, the method is fully agnostic about the cointegration rank and is fully data-driven. We provide some simple theoretical results establishing pointwise asymptotic validity of our method under general conditions. Our WIMP intervals tend to deliver coverage probabilities close to or higher than nominal levels across the entire trajectory of the IRs, even for "difficult" situations where cointegrating relations are very weak. Simulation-based evidence also suggests that the WIMP intervals generally outperform all other considered methods, including the traditional "level" approach to inference.

While we focus on frequentist inference in this paper, it is worth mentioning that rank uncertainty could also be tackled in a Bayesian VAR framework. However, in many Bayesian applications, uncertainty regarding the cointegration rank is often not taken into account explicitly. Although conceptually different, the Bayesian approach to cointegration is often similar in nature to the construction of classical (likelihood-based) inference. That is, the posterior distribution of (impulse response) parameters is often derived conditional on a predetermined rank, selected using the marginal likelihood or other model comparison approaches

(see for example Del Negro and Schorfheide, 2011, for a recent survey). However, several approaches incorporating uncertainty about the cointegration rank when analyzing VARs have been suggested in the Bayesian literature. For instance, Villani (2001) or Strachan and van Dijk (2007) propose a Bayesian model averaging scheme, similar in spirit to the approach discussed in Section 3.1.3 below. Alternatively, some authors have suggested various priors on the cointegration relations obtained using economic theory (see e.g. Del Negro et al. 2007 or Giannone et al. 2016 and references therein), which is a different conceptual approach than our fully data-driven, agnostic approach. Moreover, an explicit (theoretical) investigation of the (joint) posterior distribution of impulse responses of VARs under uncertainty on the (co-)integration relations is, however, limited also in the Bayesian literature.

Since uncertainty about the true cointegration rank is mostly ignored in applied macroeconomic research, we investigate to what extend our more robust approach(es) may change the interpretation of results in practice. More specifically, we re-evaluate the effects of fiscal policy based on four influential structural VAR frameworks. Considering Blanchard and Perotti's (2002) recursive identification strategy, Mountford and Uhlig's (2009) sign-restriction approach, Ramey's (2011) narrative VAR framework, and Mertens and Ravn's (2013; 2014) proxy-VAR, we find that neglecting rank uncertainty might lead to misleading results. As a companion to this paper, a ready-to-use MATLAB toolbox for the WIMP approach combined with various SVAR identification schemes is available online.[1]

The remainder of this paper is organized as follows. In Section 2 we discuss standard (bootstrap) approaches to inference in cointegrated VARs and illustrate empirically potential ramifications of rank misspecification. Section 3 first discusses several approaches considered in the literature about model uncertainty and their adaptations to account for rank uncertainty, and next introduces the WIMP method. The performance of the suggested methods is investigated by simulation in Section 4. Fiscal policy under rank uncertainty is analyzed in Section 5. Section 6 concludes.

2 Bootstrap Inference for Impulse Responses

2.1 The Cointegrated VAR Model and Impulse Responses

Consider the k-dimensional structural vector autoregressive (SVAR) time series process $y_t = (y_{1,t}, \ldots, y_{K,t})'$ observed at $t = 1, \ldots, T$:

$$B_0 y_t = \sum_{j=1}^{p} B_j y_{t-j} + \varepsilon_t, \tag{1}$$

[1] http://researchers-sbe.unimaas.nl/stephansmeekes

where ε_t is a K-dimensional vector of contemporaneously and serially uncorrelated, weakly stationary structural shocks[2] and B_0 is the invertible contemporaneous impact matrix. Premultiplying both sides of (1) with B_0^{-1}, we obtain the reduced-form VAR

$$y_t = \sum_{j=1}^{p} A_j y_{t-j} + u_t, \tag{2}$$

where $A_j = B_0^{-1} B_j$ and $u_t = B_0^{-1} \varepsilon_t$.

Define the lag polynomial $A(z)$ as $A(z) = I_k - \sum_{j=1}^{p} A_j z^j$, such that we can write $A(L) y_t = u_t$, where L is the lag operator $L^j y_t = y_{t-j}$. We now formulate assumptions that allow y_t to be (co)integrated with r cointegrating relations, which we label the '$I(1,r)$ conditions' as in Cavaliere et al. (2012).

Assumption 1 ($I(1,r)$ conditions)

(i) $A(z)$ has exactly $K - r$ roots equal to 1 and all other roots are outside the unit circle.

(ii) Defining $\Pi = A(1)$, we have that $\Pi = \alpha\beta'$ for $K \times r$ matrices α and β with full column rank, with the implicit definition that $\alpha\beta' = 0$ when $r = 0$.

If y_t satisfies the $I(1,r)$ conditions, we can write y_t as a VECM

$$\Delta y_t = \Pi y_{t-1} + \sum_{j=1}^{p-1} \Gamma_j \Delta y_{t-j} + u_t, \qquad t = 1, \ldots, T, \tag{3}$$

where $\Gamma_j = -\sum_{i=j+1}^{p} A_j$ for $j = 1, \ldots, p-1$.

We can invert the VAR model (2) to obtain the moving average representation

$$y_t = \sum_{j=0}^{t-1} \Psi_j u_{t-j} = \sum_{j=0}^{t-1} \Psi_j B_0^{-1} \varepsilon_{t-j} \tag{4}$$

where the Ψ_j matrices contain the reduced-form (i.e. forecast error) impulse responses and $\Phi_j = \Psi_j B_0^{-1}$ the structural impulse responses. For ease of notation later on, we directly link the impulse responses to the VECM parameters. Let $\theta = vec(\Pi, \Gamma_1, \ldots, \Gamma_{p-1})$ denote the vector of VECM parameters. Then we can define

$$\Psi_j = f_j(\theta), \qquad j = 0, \ldots, t-1,$$

where the nonlinear functions $f_j(\cdot)$ are defined implicitly through inverting the VAR model.

[2]For simplicity we assume that there is an equal number of structural shocks as variables in the system. Our model can easily be generalized to allow for a smaller number of structural shocks at the expense of complications involving the identification of the shocks. To prevent these from detracting from our main object of study, and given that these generalizations suffer from ignoring rank uncertainty in the same way as our simpler model, we abstract from this generalization in the paper.

In order to obtain structurally interpretable shocks and consequently their impulse responses $\Phi_j = \Psi_j B_0^{-1}$, we transform the estimated reduced-form errors to uncorrelated shocks. However, as B_0 is not identified, we cannot obtain Φ_j in a unique way, and estimating the structural shocks and their impulse responses requires imposing a particular identification scheme. For that purpose, let P be a $K \times K$ matrix such that $PP' = \Sigma_u$, where the specific form of P depends on the identification method. Then define the identified structural impulse responses as $\Phi_j = \Psi_j P$, and similarly

$$\Phi_j = f_j(\theta)P, \qquad j = 0, \ldots, t-1.$$

In Section 5 we discuss several ways to identify the structural shocks.[3]

2.2 Inference Conditional on a Selected Rank

We can estimate the VECM (3) for a given rank r using the Gaussian quasi maximum likelihood estimator of Johansen (1995) to obtain estimates $\hat{\Pi}^{(r)} = \hat{\alpha}^{(r)}\hat{\beta}^{(r)\prime}, \hat{\Gamma}_1^{(r)}, \ldots, \hat{\Gamma}_p^{(r)}$ and $\hat{\Sigma}_u^{(r)}$, where the superscript (r) emphasizes that estimation is conditional on r. To account for deterministic components, we can first regress y_t on a constant and possibly a linear time trend to obtain the detrended series $\tilde{y}_t = y_t - \hat{\mu}_0 - \hat{\mu}_1 t$ for $t = 1, \ldots, T$ and estimate the VECM without deterministic components on $\tilde{y}_t$.[4]

From inverting the VAR representation of the model, we can straightforwardly obtain the estimates of the moving average terms, $\hat{\Psi}_0^{(r)}, \ldots, \hat{\Psi}_h^{(r)}$, where h is the (maximum) horizon we are interested in. Letting

$$\hat{\theta}^{(r)} = vec(\hat{\Pi}^{(r)}, \hat{\Gamma}_1^{(r)}, \ldots, \hat{\Gamma}_p^{(r)}),$$

we can define the estimated impulse responses as

$$\hat{\Psi}_j^{(r)} = f_j(\hat{\theta}^{(r)}), \qquad j = 0, \ldots, h.$$

and

$$\hat{\Phi}_j^{(r)} = f_j(\hat{\theta}^{(r)})\hat{P}^{(r)}, \qquad j = 0, \ldots, h,$$

where $\hat{P}^{(r)}$ is an estimate of P such that $\hat{P}^{(r)}\hat{P}^{(r)\prime} = \hat{\Sigma}_u^{(r)}$

[3] As the impulse responses only depend on the cointegration parameters β through their product with the loadings α, that is through the error correction term $\Pi = \alpha\beta'$, we are not concerned with identification of β, unlike the setting where inference on the long run relations themselves is the objective.

[4] One could also directly incorporate deterministic components in the VECM (cf. Johansen, 1995). However, one then has to decide how the deterministic components affect the long run and short run components separately, resulting in a multitude of different specifications. Our simpler, robust, strategy corresponds to the typical approach taken in most empirical studies.

Now consider a *target impulse response* ζ, which would typically be an element of either Ψ_j or Φ_j for a fixed j; that is, we take

$$\zeta = \psi_{j,a,b} \qquad \text{or} \qquad \zeta = \phi_{j,a,b}, \tag{5}$$

where the subscript 'a, b' indicates the (a, b)-th element of the matrix. It might also be a combination of elements; for example, if one wants to perform simultaneous inference across horizons, using the ideas proposed in Bruder and Wolf (2017) and Lütkepohl et al. (2015, Section 3.6), we could take

$$\zeta = \max_{0\leq j\leq h} \psi_{j,a,b}, \qquad \zeta = \max_{0\leq j\leq h} \phi_{j,a,b}, \tag{6}$$

or its studentized versions. Similarly, one could take the Wald statistics of Inoue and Kilian (2016) as targets. The bootstrap algorithm works the same regardless of the specific target. All targets have in common that they are functions of the VAR model parameters. This way we can write both the true and estimated target impulse response as

$$\zeta = \bar{f}(\theta), \qquad \hat{\zeta}^{(r)} = \bar{f}(\hat{\theta}^{(r)}), \tag{7}$$

where the function $\bar{f}(\cdot)$ depends on the target.

We next describe a bootstrap algorithm that can be used to construct bootstrap confidence intervals for ζ. For the sake of expositional clarity, we restrict ourselves to a fairly simple, straightforward algorithm based on the bootstrap percentile method (Hall, 1992), which has regularly been considered in the literature, see e.g. Benkwitz et al. (2001).

Algorithm 1: Bootstrap Confidence Interval under Rank r

1. Let $\tilde{y}_t = y_t - \hat{\mu}_0 - \hat{\mu}_1 t$ for $t = 1, \ldots, T$ and estimate the VECM under rank r and obtain the residuals

$$\hat{u}_t = \Delta\tilde{y}_t - \hat{\Pi}^{(r)}\tilde{y}_{t-1} - \sum_{j=1}^{p-1} \hat{\Gamma}_j^{(r)} \Delta\tilde{y}_{t-j}, \qquad t = p+2, \ldots, T.$$

2. Use a bootstrap method to obtain bootstrap errors $\{u_t^*\}_{t=p+2}^T$ from the residuals $\{\hat{u}_t\}_{t=p+2}^T$.

3. Build the bootstrap sample $\{y_t^*\}_{t=1}^T$ recursively as

$$y_t^* = y_{t-1}^* + \hat{\Pi}^{(r)} y_{t-1}^* + \sum_{j=1}^{p-1} \hat{\Gamma}_j^{(r)} \Delta y_{t-j}^* + u_t^*, \qquad t = p+2, \ldots, T,$$

   using initial values $y_1^*, \ldots, y_{p+1}^*$.

4. Detrend the bootstrap sample to obtain $\tilde{y}_t^* = y_t^* - \hat{\mu}_0^* - \hat{\mu}_1^* t$ for $t = 1, \ldots, T$. Estimate the VECM under rank r on $\{\tilde{y}_t^*\}_{t=1}^T$ to obtain $\hat{\theta}^{(r)*}$. Obtain the bootstrap target impulse response as $\hat{\zeta}^{(r)*} = \bar{f}(\hat{\theta}^{(r)*})$.

5. Repeat Steps 2 to 4 B times. Let $q^*(\gamma)$ denote the γ-quantile of the B centered bootstrap statistics $\hat{\zeta}^{(r)*} - \hat{\zeta}^{(r)}$. Construct a $(1-\gamma)$-confidence interval for ζ as $\left[L^{(r)}(\gamma), U^{(r)}(\gamma)\right]$, where

$$L^{(r)}(\gamma) = \hat{\zeta}^{(r)} - q^*(1-\gamma/2) \qquad \text{and} \qquad U^{(r)}(\gamma) = \hat{\zeta}^{(r)} - q^*(\gamma/2). \tag{8}$$

Depending on the specific assumptions made on $\{u_t\}$, a variety of different bootstrap methods, such as i.i.d., wild or block bootstrap, can be used in Step 2 of Algorithm 1; we provide further details in Section 3.2.2. Similarly, different initializations in Step 3 can be used. For the simulation study and application in this paper, we use the i.i.d. bootstrap in Step 2 and initialize the bootstrap sample in step 3 by setting $y_t^* = y_t$ for $t = 1, \ldots, p+1$.

Note that $\{\hat{u}_t\}$ in Step 1 and, most importantly, $\{y_t^*\}$ in Step 3 also depend on the chosen rank r. To lighten the notation we choose not to index these formally by r, instead only emphasizing the dependence on the chosen rank r through the estimated bootstrap VAR parameters $\hat{\theta}^{(r)*}$ and target bootstrap impulse response $\hat{\zeta}^{(r)*}$. Although many variations of the bootstrap algorithm exist in the literature, such as the bias correction proposed in Kilian (1998b), all these bootstrap methods have in common that they require fixing the rank r. In particular, in generating the bootstrap sample (our step 3), it seems unavoidable to make a choice to impose a specific rank. This adds a second layer of potential rank misspecification next to the estimators themselves, which turns out to lead to further complications if one wants to account for rank uncertainty, as we discuss in Section 3 below. Before going into methods accounting for rank uncertainty however, we now first illustrate the perils of rank misspecification.

2.3 Effects of Rank Misspecification

Algorithm 1 assumes knowledge of the true cointegrating rank, labeled as r_0; if $r \neq r_0$, inference on ζ will be inappropriate, in particular for longer horizons. If the chosen rank r is smaller than the true rank, the estimated IRs converge to 'pseudo-true' values $\theta_j^{(r)}$ which are different from the true ones. This arises because the VAR parameters converge to their pseudo-true values which satisfy the (incorrect) rank restriction, c.f. Cavaliere et al. (2012). While in this case bootstrap inference remains valid for the pseudo-true parameters, these parameters can be substantially different from the true IRs, making their interpretation and therefore inference somewhat meaningless, in particular as one typically tries to uncover structural effects which requires knowledge of true parameters.

On the other hand, if $r > r_0$, as for instance in the VAR in levels specification, the short

(fixed j) and medium ($j/n \to 0$) horizon IRs are estimated consistently, but long-run ($j \sim n$) IRs are inconsistent and even random (Phillips, 1998). The inconsistency is caused by the domination of the error correction terms for the long-run IRs, and their insufficient estimation accuracy under rank misspecification. The same occurs for bootstrap inference; while valid for short and medium horizon IRs, it becomes invalid in the long-run, as demonstrated in different contexts by Choi (2005), Inoue and Kilian (2002) and Mikusheva (2012).

Figure 1 illustrates potential consequences of rank uncertainty for the construction of inference in practice. Displayed in the left panel are confidence intervals for output responses to a government spending shock identified as in Blanchard and Perotti (2002) for all possible numbers of cointegration relations.[5] Clearly, the assessment of the effectiveness of the spending policy varies drastically with the chosen cointegration rank, indicating that choosing the wrong rank hampers the interpretation of results – for long but equally so for short horizons. One could argue that with proper rank estimation, the most appropriate of these intervals can be selected. However, as demonstrated in the right panel, if evidence for a particular rank is weak, different but equally well established "respectable" rank selection procedures may suggest different models, providing little guidance for the applied researcher.

Finally, note that the unrestricted VAR in levels gives substantially different (and narrower) intervals than the VAR models with reduced rank, even the model with the next highest rank ($r = 9$). Of course, if the true model is indeed a VAR of full rank, all variables are stationary and no (co)integration would be present. However, many macroeconomic series are commonly accepted to have unit roots, which is backed up by ADF tests on our dataset, thus making the level specification unlikely to be the most appropriate. In this case, a reduced-rank VAR model would be more appropriate and constructing inference based on the VAR in levels would be invalid and could, in this example, lead to a misguided interpretation of the IRs.

The strategy to use the VAR in levels based on a robustness argument therefore appears questionable, while rank selection techniques also do not appear to give conclusive answers. It is therefore crucial to take rank uncertainty into account when conducting inference for impulse responses.

3 Inference Accounting for Rank Uncertainty

In this section we discuss several ways of accounting for rank uncertainty, first utilizing existing methods from the model uncertainty literature, before discussing a new principle.

[5]The VAR specification and the data are described in Section 5.

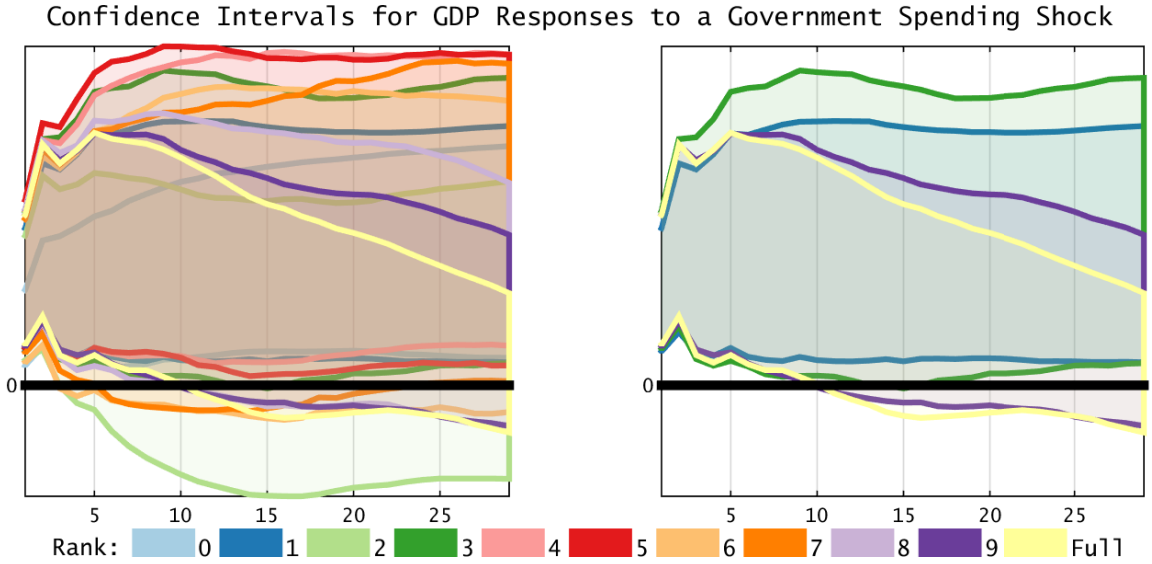

Figure 1: Left panel: Bootstrap 95% confidence intervals of the output response to a government spending shock for every rank specification. Right panel: Bootstrap 95% confidence intervals of the output response to a government spending shock implied by the trace test ($r = 3$), AIC ($r = 9$), BIC ($r = 1$), and the unrestricted VAR

3.1 Adaptations of Existing Model Uncertainty Methods

The perils of ignoring model uncertainty when performing model selection are well known in the statistical literature about model selection. For instance, in a sequence of papers, Leeb and Pötscher (see for example Leeb and Pötscher, 2005) highlight the risk of treating a selected model as a known and correct when performing inference, pointing out that even consistent model selection is no justification for treating the selected model as known. While this post-model selection inference problem is hard to solve, various methods have been proposed to at least mitigate the problem. Here we highlight some of these methods and show how they can be adapted to the problem at hand.

3.1.1 Rank Estimation

The most straightforward way to deal with rank uncertainty is to pre-estimate the rank, and then perform inference for the impulse responses conditional on the estimated rank. While this seems, given the discussion in the previous section, not always an advisable strategy, rank estimation underlies many of the methods considered afterwards. We therefore first discuss how to perform rank estimation and how it can be seen as a model selection problem.

Let the function $M_r(Y_T) : Y_T \mapsto 0, 1 \ldots, K$ be a rank selection procedure that determines the cointegration rank based on the sample $Y_T = (y_1, \ldots, y_T)'$, such that the estimated rank $\hat{r}$ is determined as $\hat{r} = M_r(Y_T)$. The estimated rank can then be imposed in the VECM

estimation to obtain the estimated target impulse responses as $\hat{\zeta}^{(\hat{r})}$.

Several methods can be considered in practice for estimation of the rank. The most common is to perform a sequence of sequential tests in the likelihood framework of Johansen (1995), in particular using the trace or eigenvalue test statistics. Instead of the standard critical values, one can also use one of its many bootstrap extensions (Cavaliere et al., 2010a,b, 2012; Swensen, 2006). Either way, due to the nature of hypothesis testing, this estimation strategy will not lead to consistent estimation of the rank (unless the significance level is chosen to decrease with sample size); the probability of selecting a rank that is too high converges to the chosen significance level instead of to zero.

Alternatively, one can use an information criterion as proposed by Phillips (1996), Chao and Phillips (1999), Cheng and Phillips (2009) and Cheng and Phillips (2012). This has two advantages compared to the sequential testing approach. First, rank selection and lag length selection can be done in a single step. Second, depending on the penalty function chosen in the information criterion, it is possible to estimate the rank consistently. A recent alternative is provided by Liao and Phillips (2015) who propose to select the rank and lag length simultaneously by penalized reduced rank regression. An advantage of this approach is that model selection and estimation are performed simultaneously, thus needing only a single step for the full estimation from start to end.

Irrespective of the chosen selection method, standard inference is based on the selected rank, treating it as known. This is often justified by the consistency of the rank selection method, but even in those cases where it is indeed consistent, ignoring the selection step leads to invalid inference as referred to earlier (Leeb and Pötscher, 2005). In particular if the data do not provide clear and strong evidence for one particular cointegrating rank, this approach will fail to deliver reliable confidence intervals. We therefore next consider methods that explicitly take rank uncertainty into account in the inference procedure.

3.1.2 Endogenous Rank Selection

Kilian (1998a) proposes the *endogenous lag selection* bootstrap method for autoregressive models where the autoregressive lag length is re-estimated within the bootstrap to account for the model selection uncertainty. We adapt his approach to rank selection, labeling this approach *Bootstrap Endogenous Rank Selection (BERS)*. Specifically, we consider the following modification to our bootstrap algorithm.

Algorithm 2: Bootstrap Endogenous Rank Selection (BERS)
Choose a rank selection method $M_r(\cdot)$, and let $\hat{r} = M_r(Y_T)$. Perform Steps 1-3 of Algorithm 1 with $r = \hat{r}$ or $r = K$. Next, replace Step 4 by

4. Let $\hat{r}^* = M_r(Y_T^*)$, where $Y_T^* = (y_1^*, \ldots, y_T^*)'$. Estimate the VECM with rank $\hat{r}^*$ on the bootstrap sample $(y_t^*)_{t=1}^T$ (after detrending) to obtain $\hat{\theta}^{(\hat{r}^*)*}$. Obtain the bootstrap target impulse responses as $\hat{\zeta}^{(\hat{r}^*)*} = \bar{f}(\hat{\theta}_j^{(\hat{r}^*)*})$.

Perform Step 5 as in Algorithm 1.

We can choose to generate the bootstrap sample Y_T^* with the "neutral" maximum rank K or the estimated rank $\hat{r}$. While Kilian (1998a) reports that this choice has little consequence for lag selection, this is very different for rank selection. After all, if the rank used to generate Y_T^* is not correct, we still face all the problems with the bootstrap as we described before. Hence, while some rank uncertainty is taken into account, the validity of this approach still hinges on the correct rank being used for the generation of the bootstrap data, which as we argued before, is impossible to guarantee.

3.1.3 Model Averaging

One of the most popular approaches to account for model uncertainty is to use model averaging (Hjort and Claeskens, 2003). By combining estimators from different models (and potentially weighting by evidence for these models), model uncertainty is taken into account. Given that the decision of which model to use is discrete, and therefore the selected model may change abruptly for a slight variation in the sample, the resulting estimators after model selection may be quite unstable and exhibit a large variability. By constructing weighted averages of the estimators arising from the individual models, one smoothes out the changes in the estimator, resulting in more stable estimators that typically display lower variability.

For this purpose we define the *Model Averaging (MA)* impulse response estimator

$$\hat{\zeta}^{MA} = \sum_{r=0}^{K} W_K(r)\hat{\zeta}^{(r)}, \quad W_K(r) = \frac{W(Y_T, r)}{\sum_{s=0}^{K} W(Y_T, s)}, \tag{9}$$

where $W(Y_T, r) : Y_T \times \{0, 1, \ldots, K\} \mapsto [0, 1]$ is a function that determines a weight for rank r based on the sample Y_T.

Unlike the typical application of model averaging, which often focuses on improving accuracy of point estimators in a mean squared error sense, we are not interested in the averaged point estimators. Instead, we only take the MA estimator as an input into our bootstrap scheme in order to construct confidence intervals: By using the more stable MA estimator, we may hope that the confidence intervals are more robust to rank misspecification. The bootstrap scheme can straightforwardly be adapted to incorporate this estimator in Step 4 of either Algorithm 1 or 2, depending on whether one wants endogenously determined weights in the bootstrap or not.

Typical weights in the model averaging literature are exponential weights based on information criteria such as BIC. However, in our simulations we find that such standard weighting schemes give weights that are too close to each other and do not differ much from simple unweighted averages. Given the widely varying behavior of impulse responses under different ranks, such weights are therefore not the most useful ones in our setting. Instead, we ad-

vocate using weights that are derived directly from cointegration tests, following the spirit of Sobreira and Nunes (2012), but rather than their KPSS type weights, we opt for weights based on the trace test statistic proposed by Johansen (1995). Details about the weights and their properties can be found in Lemma 1 in Section 3.2.2.

In a similar framework, Jardet et al. (2013) propose an averaging approach for impulse responses of potentially cointegrated VAR models based on a very specific set of weights. While they allow for uncertainty regarding the order of integration, their approach only averages two estimators: the one obtained from the VAR in levels, and one obtained from a cointegrated VAR where the number of cointegrating relations is pre-determined by pre-testing or economic theory. It can therefore not account for the general case where we are agnostic about the number of cointegration relations.

While such model averaging explicitly takes model uncertainty into account, it still relies on an explicit choice of the cointegration rank in the bootstrap algorithm to do inference. Hence, even while the weight construction can be endogenized in the bootstrap in the same way as for rank selection, the bootstrap DGP relies on the choice of a single cointegration rank. As such it still does not fully account for rank uncertainty in our context.

3.1.4 Bagging

We now take a first step in endogenizing the rank uncertainty in the bootstrap DGP itself, by bootstrapping a bagging estimator. The bagging estimator is constructed by averaging the bootstrap estimates over an initial bootstrap procedure in which the cointegration rank is re-estimated for every bootstrap sample. Bagging was originally proposed by Breiman (1996) to improve estimation accuracy of unstable estimators. Bühlmann and Yu (2002) analyzed bagging formally and found that it can lead to a variance reduction of estimation after hard decisions, such as an initial model selection. As the model averaging described above, bagging smoothes those hard decisions yielding more accurate estimators. Efron (2014) considers bagging in the context of post-selection inference, rather than point estimation, and we build on his approach here.

As bagging is essentially the simulation equivalent of model averaging, with the weights implicitly determined by how often each rank is selected within the bootstrap, it is subject to the same critique. However, one can modify the bagging algorithm to endogenize rank uncertainty in the bootstrap DGP by performing a second-level bootstrap in which we draw new bootstrap samples from the first-level bootstrap samples. By determining the rank of the second-level bootstrap DGPs from the first-level bootstrap samples, the ranks are randomized according to their evidence in the (simulated) sample. This allows to take the uncertainty into account when constructing the bootstrap confidence intervals based on the second-level bootstrap samples. While this does not fully solve the bootstrap invalidity problem (bootstrap samples are still generated under incorrect ranks, especially in the first step), the method has

the potential to alleviate the problem.

There is a computational problem with this method though, as one has B_1 iterations in the first bootstrap and B_2 in each second-level bootstrap, such that a full double bootstrap requires $B_1(1 + B_2)$ iterations which quickly becomes computationally infeasible. For this purpose we implement the Fast Double Bootstrap (FDB) developed by Davidson and MacKinnon (2002), which requires drawing only a single second-level bootstrap sample for every first-level bootstrap sample, which means the computation cost of the FDB is only double ($2B_1$) that of a regular bootstrap. The algorithm below describes the method, labeled as *FDB bagging (FDBb)*, in detail.

Algorithm 3: FDB bagging (FDBb)
Choose a rank selection method $M_r(\cdot)$, and perform steps 1-4 of Algorithm 2. Next:

5. Perform a second bootstrap procedure on the bootstrap sample $\{y_t^*\}_{t=1}^T$ to obtain double-bootstrap impulse responses. For every bootstrap sample $\{y_t^*\}_{t=1}^T$, only *one* second-level bootstrap sample has to be drawn. Specifically, take the following steps:

   (i) Estimate the VECM with rank $\hat{r}^* = M_r(Y_T^*)$ and obtain the residuals

   $$\hat{u}_t^* = \Delta\tilde{y}_t^* - \hat{\Pi}^{(\hat{r}^*)*}\tilde{y}_{t-1}^* + \sum_{j=1}^{p} \hat{\Gamma}_j^{(\hat{r}^*)*}\Delta\tilde{y}_{t-j}^*, \qquad t = p+2, \ldots, T. \tag{10}$$

   (ii) Construct the second-level bootstrap errors $\{u_t^{**}\}_{t=p+2}^T$ from $\{\hat{u}_t^*\}_{t=p+2}^T$ using the same bootstrap method as for the first level, and build the second-level bootstrap sample $\{y_t^{**}\}_{t=1}^T$ recursively as

   $$y_t^{**} = y_{t-1}^{**} + \hat{\Pi}^{(\hat{r}^*)*}y_{t-1}^{**} + \sum_{j=1}^{p} \hat{\Gamma}_j^{(\hat{r}^*)*}\Delta y_{t-j}^{**} + u_t^{**}, \qquad t = p+2, \ldots, T, \tag{11}$$

   with initial values $y_1^{**}, \ldots, y_{p+1}^{**}$.

   (iii) Estimate the cointegration rank $\hat{r}^{**} = M_r(Y_T^*)$ and use it to obtain $\hat{\zeta}^{(\hat{r}^{**})**}$.

6. Repeat Steps 1 to 5 B times. Let $\hat{\zeta}_1^{(\hat{r}^*)*}, \ldots, \hat{\zeta}_B^{(\hat{r}^*)*}$ denote the ordered sequence of the first-level bootstrap estimates obtained over the B bootstrap replications. The *bagging* estimator of the impulse response is then defined as

   $$\hat{\zeta}^{\text{bag}} = B^{-1}\sum_{b=1}^{B} \hat{\zeta}_b^{(\hat{r}^*)*}. \tag{12}$$

   Let $q^{**}(\gamma)$ denote the γ-quantile of the B centered second-level bootstrap statistics

$\hat{\zeta}^{(\hat{r}^{**})**} - \hat{\zeta}^{(\hat{r}^{*})*}$. Construct a $(1-\gamma)$-confidence interval for ζ as

$$\left[\hat{\zeta}^{(\hat{r})} - q^{**}(1-\gamma/2), \hat{\zeta}^{(\hat{r})} - q^{**}(\gamma/2)\right].$$

3.2 Weighted Inference by Model Plausibility

None of the methods described above fully address the post-model selection inference problem. To work towards a more satisfactory solution, we now combine the ideas discussed above with new concepts arising from the recent statistical literature that directly addresses the post-model selection inference problem.

We would like to build on the idea of averaging or weighting models to account for rank uncertainty. However, as elaborated on in the previous section, such weighting is typically designed for point estimation and translating it to confidence intervals, as needed here, is not so straightforward. In order to make the transition, we take inspiration from the perspective taken by Berk et al. (2013), who view the issue of constructing valid post-model selection inference as a simultaneous inference problem: by controlling for performing inference in all models simultaneously, the specific model selected by a model selection procedure is covered by construction. In our notation, and following their approach, we could construct intervals $[\hat{\zeta}^{\hat{r}} - q^{\text{PoSI}}(1-\gamma/2), \hat{\zeta}^{\hat{r}} - q^{\text{PoSI}}(\gamma/2)]$ such that

$$\mathbb{P}\left(q^{\text{PoSI}}(\gamma/2) \leq \hat{\zeta}^{(r)} - \zeta^{(r)} \leq q^{\text{PoSI}}(1-\gamma/2), \quad \forall r \in \{0, K\}\right) \to 1-\gamma$$

as $T \to \infty$. It is important to note that here $\zeta^{(r)} = \bar{f}(\theta^{(r)})$ is a *pseudo-true* parameter defined in terms of $\theta^{(r)}$, the pseudo-true parameters of the model (2) under the restriction that rank r is imposed – see Lemma 1 and its proof in Cavaliere et al. (2012) for a formal definition. These parameters represent the probability limits of the estimators of (2) under the restriction of imposing rank r, and can informally be seen as those parameters which minimize a distance to the true parameters under the restriction that the cointegration rank is r. If $r < r_0$, the true parameter cannot be recovered, and therefore the pseudo-true parameter will be different.

For our purposes, there is a fundamental problem with the *sub-model* view of Berk et al. (2013) where the pseudo-true parameters are the objects of interests. In the context of structural impulse responses, the sub-model view has little relevance, as it cannot uncover any structural effects. We therefore need the *full model* view, in which it is assumed that one of the models is the true (structural) one. Denoting this extension of the PoSI approach as PoSI_0, we seek to control

$$\mathbb{P}\left(q^{\text{PoSI}_0}(\gamma/2) \leq \hat{\zeta}^{(r)} - \zeta \leq q^{\text{PoSI}_0}(1-\gamma/2), \quad \forall r \in \{0, K\}\right) \to 1-\gamma$$

as $T \to \infty$. This implies that we require that the distance between every fixed-rank estimate $\hat{\zeta}^{(r)}$ and the true impulse response ζ is taken into account in constructing the confidence

intervals, rather than the much shorter distance between $\hat{\zeta}^{(r)}$ and its probability limit or pseudo-true impulse response $\zeta^{(r)}$, resulting in rather wide intervals. The seemingly only way to control this quantity is to construct confidence intervals for every rank separately, and then take the union of these, which will typically result in very wide intervals that are useless in practice.

We have not yet considered any evidence on the plausibility of each rank, that can be extracted from the data. If this information can incorporated into our inferential procedure, we may be able to achieve intervals that are still useful in applications, as the impact of ranks that the data deem very implausible can be eliminated, or at least reduced. We therefore augment the PoSI view of simultaneous inference by a weighting scheme akin to model averaging, except that we apply the weighting not to the estimators but directly to the bounds of the intervals. The direct weighting of the inference output, in this case the interval bounds, by evidence of the plausibility of each model, leads us to label our approach as *Weighted Inference by Model Plausibility (WIMP)*.

3.2.1 The WIMP Principle

Define the most plausible model - according to a certain plausibility measure based on the data - as the *reference model*, and denote the corresponding confidence interval arising from this model (ignoring model uncertainty) as the *reference interval*. As input to the WIMP procedure we consider all *model intervals*, which are defined as the confidence intervals obtained by assuming any particular model as the true one. In our case these would be the intervals obtained by imposing all the $K+1$ different cointegrating ranks. Before going into the details of our application, we now describe the general conditions that any prudent WIMP scheme must adhere to:

WIMP Prudence Conditions

1. The WIMP confidence interval must always cover at least the reference interval. That is, any non-reference model can only lead to widening the WIMP interval compared to the reference interval.

2. If two models are equally plausible, the model interval bounds which are furthest away from the reference model must contribute the most to widening the WIMP interval.

3. If the bounds of two model intervals are equally far away from the reference interval, the most plausible model must contribute the most to widening the WIMP interval for a given distance of the bounds from the reference interval.

4. The WIMP confidence interval may not be wider than the interval obtained by joining all individual model intervals.

The first condition is required to avoid invalid intervals, in whatever way validity is measured. If it is possible to obtain a confidence interval which is more narrow than the "standard" interval assuming no model uncertainty, the WIMP interval can never be guaranteed to contain an adequate coverage probability. The second condition ensures that the locations of intervals in relation to the reference interval are properly taken into account for equally plausible models. Compare two equally plausible models with almost identical intervals, to two equally plausible models with very different intervals. Any prudent method of accounting for model uncertainty must result in wider intervals for the second case than for the first case. The third condition implies that one has to take more plausible models more strongly into account than implausible models. In particular, this condition allows to reduce the impact of implausible models that may have very different intervals than the reference model but are so implausible, that there is little to no uncertainty about them. Finally, the fourth condition ensures that the WIMP intervals do not become too conservative. While the first and fourth condition impose hard (but sensible) restrictions on the WIMP intervals, the second and third conditions allow for quite some variation in the procedure. Finding a right balance between conservatism and interval length is therefore of great practical importance, and varies per setting.

We now turn to our specific implementation of the WIMP Prudence Conditions. Let $W_K(r)$ be model plausibility weights assigned to all ranks $r = 0, \ldots, K$ and define $X(r,s) = \frac{W_K(r)}{W_K(s)}$ as the relative plausibility of rank r compared to rank s. Letting $R = \arg\max_{0 \le r \le K} W_K(r)$ be the most plausible or reference rank, we define the WIMP interval as $\left[L^{\text{WIMP}}(\gamma), U^{\text{WIMP}}(\gamma)\right]$ with

$$\begin{aligned} L^{\text{WIMP}}(\gamma) &= \min_{r=0,\ldots,K} \left\{ L^{(R)}(\gamma) - X(r,R) \left[L^{(r)}(\gamma) - L^{(R)}(\gamma) \right]^- \right\}, \\ U^{\text{WIMP}}(\gamma) &= \max_{r=0,\ldots,K} \left\{ U^{(R)}(\gamma) + X(r,R) \left[U^{(r)}(\gamma) - U^{(R)}(\gamma) \right]^+ \right\}, \end{aligned} \tag{13}$$

where $x^+ = \max(x, 0)$, $x^- = -\min(x, 0)$ and $L^{(r)}(\gamma)$ and $U^{(r)}(\gamma)$ are the lower and upper bounds respectively of the confidence intervals with fixed rank r as defined in (8).

The term $\left[L^{(r)}(\gamma) - L^{(R)}(\gamma)\right]^-$ (respectively $\left[U^{(r)}(\gamma) - U^{(R)}(\gamma)\right]^+$) ensures that only lower bounds smaller (upper bounds larger) than those of the reference interval are taken into account; for lower bounds larger (upper bounds smaller) than those of the reference interval, this term is simply zero. Together with $X(r,s) \ge 0$, this implies that the WIMP interval always contains the reference interval, hence Condition 1 is satisfied. Condition 2 is also trivially satisfied as this term increases when the lower (upper) bound of the rank r interval is further away from the reference interval.

The shape of $X(r,s)$ determines how strongly less plausible models are taken into account and can be different from the linear function of $W_K(r)$ imposed above. As long as $X(r,s)$ is an increasing function of $W_K(r)$, more plausible ranks are given more importance and Condition 3 is satisfied; varying $X(r,s)$ and $W_K(r)$ allows one to change the balance between

conservatism and interval length. Finally, with respect to Condition 4, note that as long as $X(r,s) \leq 1$, the WIMP interval can never be wider than the interval obtained by combining the smallest lower bound with the largest upper bound.[6]

Two final remarks about the WIMP principle are in order. First, although we focus here exclusively on the case of rank uncertainty, other types as uncertainty, such as about the lag order or the deterministic components can be incorporated into the WIMP procedure as well. For instance, if one want to allow for P different lag orders in addition to the $K+1$ ranks, one needs weights that measure the plausibility of each of the $(K+1)P$ different models resulting from combining the different ranks and lag orders. In this paper we focus on rank uncertainty only as it has a far bigger and more fundamental impact than (slight) lag misspecification. Moreover, successful methods exist for accounting for lag uncertainty, such as Kilian's (1998a) endogenous lag selection bootstrap. One may therefore also opt for accounting for lag order uncertainty through the fixed rank intervals that form the input to the WIMP.

Second, note that the WIMP intervals are not built directly around a single point estimator for ζ. While all $K+1$ fixed-rank estimators are incorporated through their respective confidence intervals, we do not directly obtain a corresponding point estimate for ζ. Of course, if there is a desire to pair the confidence interval with a point estimator, one can do so, in which case the model averaging estimator with the same weights $W_K(\cdot)$ as used for the WIMP intervals is the most natural candidate.[7]

3.2.2 Asymptotic Properties

In this section we derive asymptotic properties of the WIMP intervals. We mainly do so under general high-level assumptions on the tests and bootstrap method available, but we will also provide some details about how these assumptions can be verified in our application. We first turn to the pointwise asymptotic validity of our method.

Theorem 1. *Let Y_T be generated according to (2), and let $\Theta^{(r)}$ denote the parameter space of θ such that the $I(1,r)$ conditions are satisfied. Then assume that*

(i) As $T \to \infty$, $W_K(r) \xrightarrow{p} \mathbb{1}(r = r_0)$, where $\mathbb{1}(\mathcal{A})$ is equal to 1 if $\mathcal{A}$ is true, and 0 otherwise;

(ii) For $r = r_0$ the bootstrap confidence interval has correct coverage; that is, as $T \to \infty$, we

[6] If some of the individual model intervals are disjoint, the "maximal" WIMP interval as constructed in (13) is larger than the union of these intervals, apparently violating Condition 4. However, this is an intentional violation. Such a disjointed confidence *set* is not a confidence *interval* any more, and therefore is rather awkward to interpret. The natural modification of this set, that yields an interval again, would be to "fill the gaps" and extend it from the lowest lower bound to the highest upper bound, which is exactly what the WIMP construction does automatically.

[7] As expected from the model averaging literature, unreported simulations in the same setup as considered in Section 4 show that this estimator performs very well in terms of mean squared error when compared to fixed-rank estimators. Of course, its performance purely as a point estimator is different from its performance as basis for inference, as we shall see in Section 4.

have that

$$\mathbb{P}\left(L^{(r_0)}(\gamma) \leq \zeta \leq U^{(r_0)}(\gamma)\right) \to 1-\gamma \qquad \forall \theta \in \Theta^{(r_0)} \quad \forall r_0 \in \{0, K\}.$$

Then

$$\mathbb{P}\left(L^{\text{WIMP}}(\gamma) \leq \zeta \leq U^{\text{WIMP}}(\gamma)\right) \to 1-\gamma \qquad \forall \theta \in \Theta^{(r_0)}, \quad \forall r_0 \in \{0, K\}.$$

as $T \to \infty$.

Proof. By Assumption (i), we have that $\mathbb{P}(R = r_0) \to 1$ and consequently that $X(r, R) \xrightarrow{p} \mathbb{1}(r = r_0)$. It therefore follows directly that $L^{\text{WIMP}}(\gamma) = L^{(R)} \xrightarrow{p} L^{(r_0)}(\gamma)$ and $U^{\text{WIMP}}(\gamma) \xrightarrow{p} U^{(r_0)}(\gamma)$. The result then follows from assumption (ii). □

Assumption (ii), which implies asymptotic validity of the intervals under a known rank, has been verified for many bootstrap methods under different assumptions on $\{u_t\}$ (or equivalently $\{\varepsilon_t\}$). For instance, if we assume that $\{u_t\}$ is i.i.d. with sufficiently many moments existing, one can show that the i.i.d. bootstrap version of Algorithm 1 satisfies assumption (ii), c.f. Kilian (1998b) and Cavaliere et al. (2012). Inoue and Kilian (2016) also formulate general assumptions to assure bootstrap validity, while alternative methods that allow for heteroskedasticity are considered by Brüggemann et al. (2016). The WIMP principle can be applied to any of these methods and targets; in fact, it does not even require bootstrap confidence intervals, but can equally well be applied to any asymptotically valid inference method.

We now propose a concrete weighting scheme that satisfies Assumption (i) in Theorem 1. Following the spirit of Sobreira and Nunes (2012), we base our weights on cointegration tests. Rather than their KPSS type weights, we opt for weights based on the trace test statistic proposed by Johansen (1995), which, as a "standard" cointegration test, has intuitive appeal and is available in all standard econometric and statistical software.[8]

Lemma 1. *Let* $J_T(r) = -T\sum_{i=r+1}^{K} \ln(1-\hat{\lambda}_i)$ *denote the trace test of Johansen (1995) for testing* $H_0 : r_0 \leq r$. *For constants* $c_1 > 0$ *and* $0 < c_2 < 1$, *define*

$$\begin{aligned} W(Y_T, r) &= e^{-c_1 T^{-c_2} J_T(r)} && \text{for } r = 0. \\ W(Y_T, r) &= e^{-c_1 T^{-c_2} J_T(r)} - e^{-c_1 T^{-c_2} J_T(r-1)} && \text{for } r = 1, \ldots, K-1, \\ W(Y_T, r) &= 1 - e^{-c_1 T^{-c_2} J_T(r-1)} && \text{for } r = K, \end{aligned} \tag{14}$$

and $W_K(r) = W(Y_T, r) / \sum_{r=0}^{K} W(Y_T, r)$. *Then* $W_K(r) \xrightarrow{p} \mathbb{1}(r = r_0)$ *as* $T \to \infty$.

[8] We also explored Johansen's (1995) maximum eigenvalue test statistic, which similarly satisfies assumption (i) in Theorem 1. Numerical experiments showed virtually no difference with the trace test.

Proof. It follows from Johansen (1995) and Bernstein and Nielsen (2014) that for all $r \geq r_0$, $J_T(r) = O_p(1)$, such that $T^{-c_2} J_T(r) \xrightarrow{p} 0$, while for $r < r_0$, we have that $J_T(r)/T$ is tight, such that $T^{-c_2} J_i(r) = T^{1-c_2} J_T(r)/T \xrightarrow{p} \infty$. Therefore we have that

$$e^{-c_1 T^{-c_2} J_T(r)} \xrightarrow{p} \mathbb{1}(r \geq r_0) \quad \Rightarrow \quad W_T(r) \xrightarrow{p} \mathbb{1}(r = r_0). \qquad \square$$

While the results above establish the pointwise asymptotic validity of our proposed scheme, it should be noted that this does not imply uniform validity, that is, the property

$$\liminf_{T\to\infty} \inf_{\theta\in\Theta} \mathbb{P}\left(L^{\text{WIMP}}(\gamma) \leq \zeta \leq U^{\text{WIMP}}(\gamma)\right) \geq 1-\gamma, \quad \text{where } \Theta = \bigcup_{r=0}^{K} \Theta^{(r)}.$$

Uniform validity is a more informative property about finite sample behavior of the intervals, as it does not rely on the *oracle property* that the true rank is always selected asymptotically, as is assumed in Assumption (i) in Theorem 1. In particular for small deviations from a certain rank, the weights are unlikely to pick this up, so the oracle property in Assumption (i) is a poor approximation to finite sample performance and can be very misleading. In fact, the same pointwise reasoning underlies the use of consistency of information criteria like BIC as a valid approach to model uncertainty, and should therefore be treated with caution, see e.g. Leeb and Pötscher (2005).

However, while clearly of great interest, uniform validity is very hard to establish for the cointegrated VAR based on bootstrap inference, as it requires the consideration of sequences of local deviations from certain ranks, under which the bootstrap is known to have problems. So far uniform results have only been established in the presence of a single local-to-unit root (cf. Mikusheva, 2007, 2012), while more general results are needed for our setting, and are to the best of our knowledge unavailable. Establishing a full uniform asymptotic theory is therefore outside the scope of this paper and left for future research. Here we focus on evaluating the small sample properties of the WIMP method for situations where rank uncertainty is present. Note that even though the asymptotic validity of our WIMP implementation is based on the same oracle properties used to validate consistent rank selection, unlike these methods our WIMP intervals do explicitly take rank uncertainty into account, and are always wider in finite samples than the fixed-rank intervals. We therefore expect that the WIMP intervals will be much more reliable in small samples when even minor rank uncertainty is present.

4 Monte Carlo Simulations

In this section we investigate the performance of the various methods discussed above by simulation. We assess coverage probabilities (CP) of confidence bands for *forecast error impulse responses*, and hence evaluate intervals for the moving average parameters. As such

we base our analysis fully on the reduced-form VAR, and do not consider structural VARs. We intentionally abstract from the identification problem in structural VARs, since the structural moving average parameters are linear combinations of their reduced-form counterparts, and one can expect that the performance of one inferential procedure for reduced-form parameters is inherited by the structural parameters.[9]

The data generating process (DGP) for the Monte Carlo experiment is a three-dimensional VAR of order one inspired by Phillips (1998), given by $y_t = (I_3 + \Pi)y_{t-1} + \epsilon_t$, with $\epsilon_t \sim$ *i.i.d.* $\mathcal{N}(0, I_3)$ for all t. The cointegration matrix is specified as $\Pi = d_1\alpha_1\beta_1' + d_2\alpha_2\beta_2'$, where $\alpha_1 = (0, 1, 0)'$, $\alpha_2 = (0, 0, 1)'$, $\beta_1 = (2, -1, 0)'$, and $\beta_2 = (1, -1, -1)'$. We consider two versions of the above process when simulating data.

DGP1: Setting $d_1 = 0.05$ and $d_2 = 0.02$ implies that the model has one root at unity and two roots close to one at 0.98 and 0.95. Thus, we have two "*weak*" cointegration relations.

DGP2: Setting $d_1 = d_2 = 1$ implies a VAR with one unit root and two roots at zero, thus two "*strong*" cointegration relations. This is the original setting considered by Phillips (1998).

We evaluate CPs of 95% confidence intervals for each response and horizon ($h = 1, 2, ..., 60$) for $T = 100, 200$. The results are based on 1000 MC simulations and 399 bootstrap replications. To compute the WIMP intervals we set $c_1 = 1$ and $c_2 = 0.5$ for the weights in (14).[10] As mentioned above we do not consider identification of structural IRs. We also abstract from lag length selection (we fix $p = 1$),[11] deterministic components, and small sample bias correction (Kilian, 1998). All simulations were done in MATLAB.

Figure 2 and 3 display CPs of the various inferential procedures discussed above for DGP1 for $T = 100$ and $T = 200$. Based on the two model selection criteria employed, we can partly confirm the findings of Gospodinov et al. (2013). That is, if evidence for a particular rank is weak, pre-testing seems not to deliver more accurate inference than (bootstrap) CIs based on unrestricted OLS. This holds for both sample sizes considered. However, these two frequently used approaches can both not be considered as reliable strategies for the construction of inference – minimum CPs are well below 60%. Surprisingly, even when the true model specification is imposed (which could be considered to be the *oracle* method), CPs are generally not closer to the nominal level either; both in short and long horizon.

[9] Except for SVARs identified through long-run restrictions, the exact persistence properties of the underlying reduced-form process are of no direct relevance for identification.

[10] This choice of parameters seems to be the most natural for the weights in (14). We did not experiment with changing these values, as the performance in the simulations was already quite satisfactory. It is likely that by careful tuning these parameters, even better performance can be obtained. However, the optimal choice will typically be highly case-dependent, and optimal values should therefore be treated with caution. Instead we prefer to report results for a natural albeit naive choice of parameters without claiming any optimality.

[11] Unreported results with $p = 3$ show the same patterns as $p = 1$.

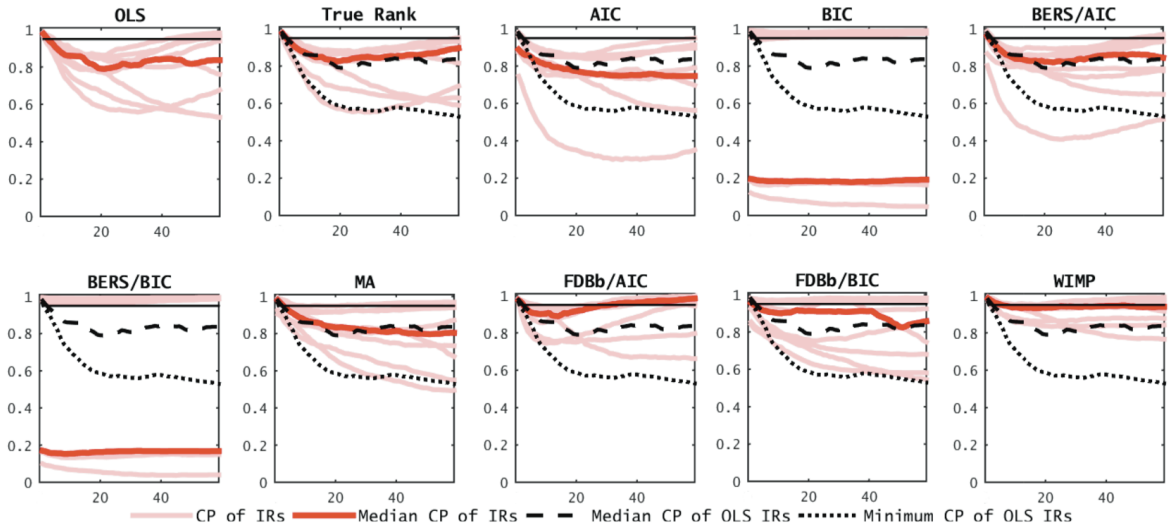

Figure 2: DGP1: Empirical coverage rates for ten inference methods for $T = 100$.
'**OLS**' refers to the (unrestricted) VAR in levels estimated by OLS; '**True Rank**' refers to the VECM estimated with knowledge of the true rank '**AIC**' and '**BIC**' refer to the rank estimation of Section 3.1.1 using AIC and BIC, respectively; '**BERS/AIC**' and '**BERS/BIC**' refer to the Bootstrap Endogenous Rank Selection of Section 3.1.2 with respectively AIC and BIC used for rank selection; '**MA**' refers to the model averaging method of Section 3.1.3 with weights as in (14); '**FDBb/AIC**' and '**FDBb/BIC**' refer to the FDB bagging method of Section 3.1.4 with respectively AIC and BIC used for rank selection; '**WIMP**' refers to the WIMP method of Section 3.2 with weights as in (14).
The pink lines show CPs for all nine impulse responses; the **red** line is the median of these per horizon. For ease of comparison, the median and minimum coverage of the OLS intervals is always reported in **black**.

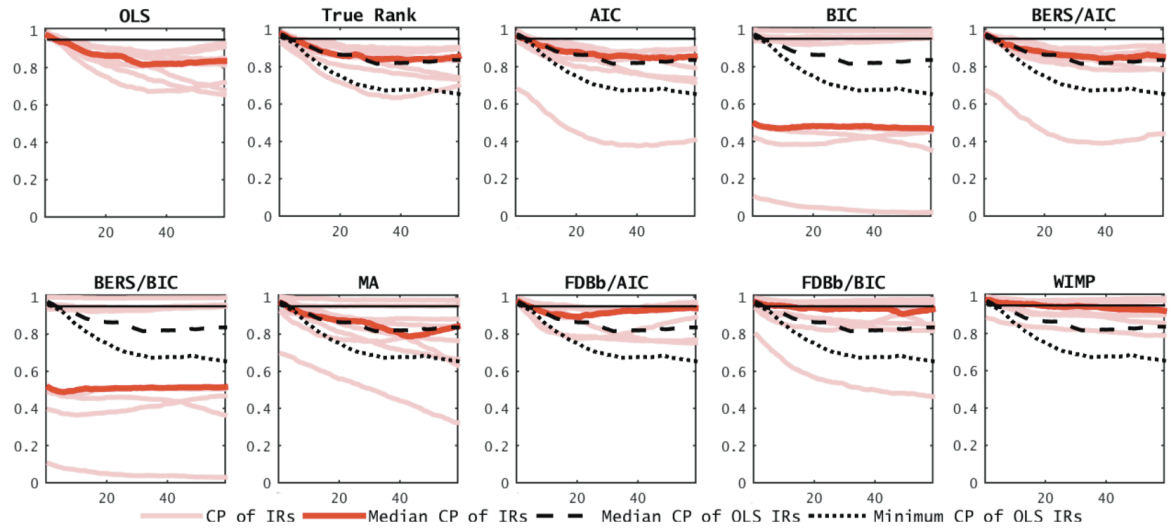

Figure 3: DGP1: Empirical coverage rates for the various inference methods for $T = 200$. See Figure 2 for details.

Endogenous rank selection does not seem to improve the performance compared to the pre-testing procedure. FDB bagging does give CPs closer to nominal level, in particular when based on AIC. However, the WIMP intervals outperform all other methods, and deliver CPs that are on average quite close to the 95% nominal level.

Figure 4 presents the corresponding average width of the bootstrap intervals over all horizons for the five most relevant methods. There are several interesting observations to make from this figure. First, note that even though FDB bagging and WIMP produce much more accurate intervals than OLS or imposing the true rank, they actually do not produce intervals that are much wider. It of course makes perfect sense that they deliver wider intervals, as the intervals of the other methods are too narrow, but the limited extent to which they are wider indicates the methods are not overly conservative. Second, even though the WIMP method produces more accurate intervals than FDB bagging, intervals are not wider. This shows that the mechanism imposed in the WIMP to reduce the impact of implausible models works well in practice.

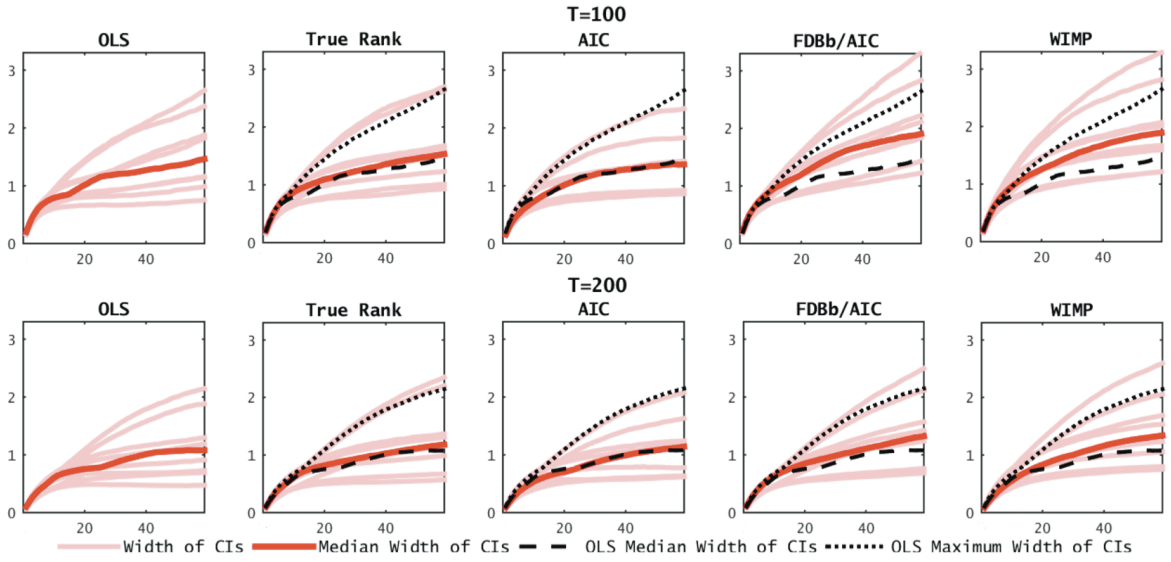

Figure 4: DGP1: Average width of 95% bootstrap CIs for various inference methods for $T = 100$ and $T = 200$. For details see Figure 2.

It stands to reason that if evidence for a specific cointegration relation is strong, rank pre-estimation could result in more reliable inference than unrestricted OLS and may outperform the WIMP intervals which – despite weighting down implausible ranks – are inherently more conservative. We investigate this further by turning to DGP2. Figure 5 displays CPs for the case of strong cointegration relations. Indeed, CPs implied by model selection based on AIC and BIC are much closer to the nominal level than those entailed by OLS. Bootstrap intervals based on unrestricted estimation can again not be considered as reliable, with minimum CPs around 60% for both sample sizes. Imposing the true rank delivers CPs close to but still

below the nominal level. As in the weak cointegration setting, the WIMP intervals again outperform all other approaches and even deliver CPs closer to nominal level than those implied by the correct rank specification. It is noticeable that the WIMP intervals do not produce overly conservative inference when evidence for a particular rank is strong, but result in CPs very close to the 95% level. This is also reflected in the average width (over 1000 MC simulations) of the CIs displayed in Figure (6). WIMP intervals are (if at all) only marginally wider than those implied by the correct rank specification, and are even much narrower than some of the intervals based on the unrestricted model. Finally, note that the WIMP intervals are now also much narrower than some of the FDB bagging intervals while having superior coverage. Concluding, the WIMP intervals allow for meaningful inference in practical sample sizes irrespective of the degree of rank uncertainty.

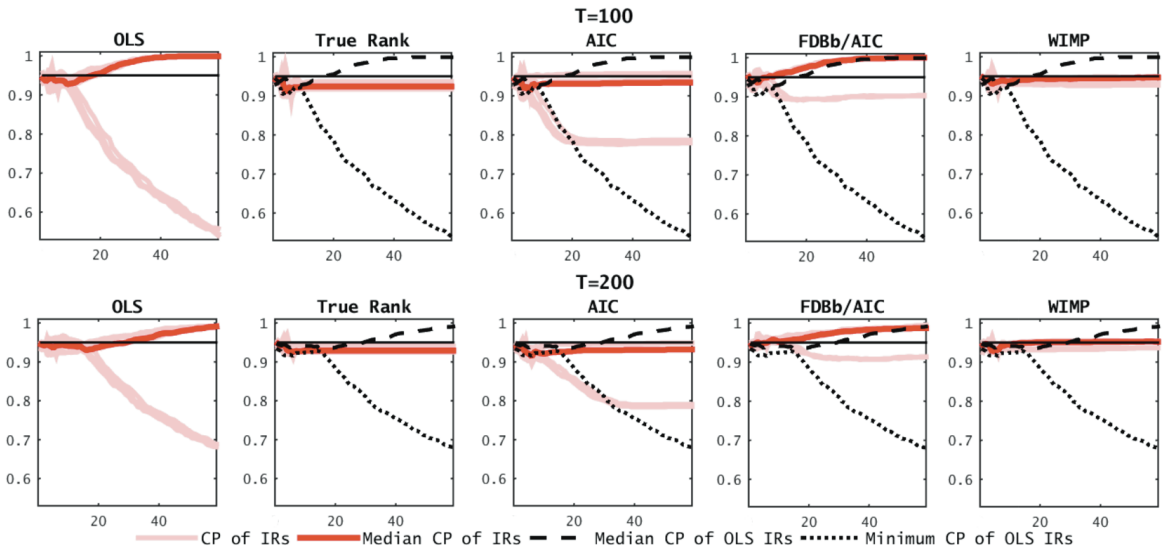

Figure 5: DGP2: Empirical coverage rates for various inference methods for $T = 100$ and $T = 200$. For details see Figure 2.

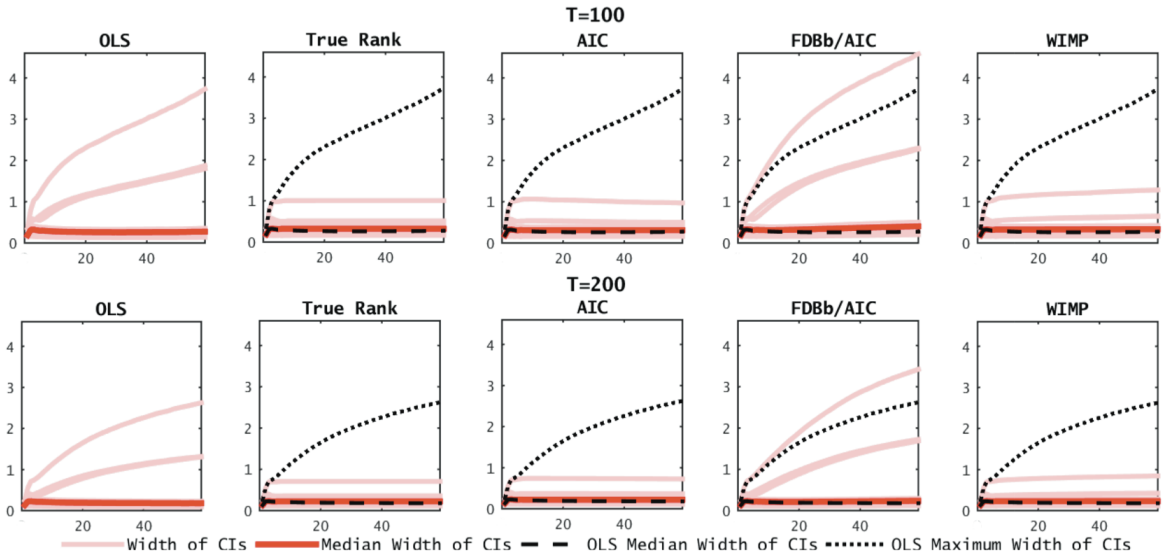

Figure 6: DGP2: Average width of 95% bootstrap CIs for various inference methods for $T = 100$ and $T = 200$. For details see Figure 2.

5 Fiscal Policy Shocks and Rank Uncertainty

In this section we study the potential ramifications of rank uncertainty on applied macroeconomic analysis. By using our proposed approaches to construct inference accounting for rank uncertainty, we aim at assessing the robustness of results usually obtained from unrestricted VARs. While there are countless VAR-based studies that use impulse response analysis to investigate the propagation of structural economic shocks, we focus in the following on fiscal policy shocks.

The novelty of this paper is methodological and we do not complement the literature on identification of structural VARs. This is why we dispense with a detailed literature review summarizing important contributions on VAR-based policy analysis and only focus on evaluating seminal papers, reflecting various ways of identification. We also skip a detailed discussion of different identification approaches and their respective merits.[12] Instead, our goal is to demonstrate that the problem – and our solution – is present regardless of the identification scheme. For this purpose it suffices for us to focus on several seminal papers that consider different identification schemes.

In this paper we do not want to engage in a discussion about the exact size of the fiscal multiplier. We rather want to emphasize the amplified uncertainty associated with its estimation under unknown cointegration relationships. For that reason we omit any discussion on point estimates and focus solely on inference, and highlight the role of (ignoring) rank uncertainty.

Our aim is also not to challenge (widely accepted) empirical findings on the effects of economic policies, but to provide the applied researcher with tools that might help to construct more reliable inference. For that reason, we refrain from a simple replication exercise comparing different inferential approaches, and we want to stress that our goal is certainly not to contrast our findings to the original papers. Instead, we use the same reduced-form VAR and (subsets of) the same dataset across all applications, in order to move away from the original papers and only contrast results based on different identification procedures. By "homogenizing" the underlying models and data used, we construct a coherent structure in which the effects of rank uncertainty can properly be investigated, and which is of interest in itself.

Fiscal policy can relate to both the expenditure and revenue side of the government's budget. Measuring the effect of active spending policies as well as the consequences of tax changes has been an active field of economic research since decades. One of the first influential contributions using VAR-based impulse responses to assess the effect of government purchases is Blanchard and Perotti (2002). The authors identify spending shocks by a recursive identification scheme. With government spending ordered first, this translates into the

[12] For a detailed exposition we refer to Ramey (2016) for a recent survey on various identification approaches and results in the literature.

assumption that government purchases are predetermined within the quarter.

Due to their assumed independence from general macroeconomic conditions, Ramey and Shapiro (1998) construct narrative records based on military buildups to identify truly exogenous spending changes. Those narrative time series have been embedded in several VAR studies and used to identify spending shocks by ordering this series first in a Cholesky-identified VAR. Among the most prominent studies following this approach is Ramey (2011). In her paper she revisits the construction of the government spending news variable, filtering out possible distortions due to anticipation effects.

Narrative series have also been used to identify tax changes. In a series of papers Mertens and Ravn (2011, 2012, 2013, 2014) construct various "dis-aggregates" of the Romer and Romer (2009) measures of legislated changes in federal tax liabilities. More specifically, Mertens and Ravn distinguish between announced and unannounced tax changes, or between personal and corporate taxes. Moreover, the authors do not view those narrative series as a direct measure of "tax-shocks" but rather as an external *proxy* which is correlated with the unknown structural shocks.[13] Thus, instead of including the narrative variable in the VAR, one can obtain the structural shock of interest by regressing the narrative *proxy* on the reduced-form residuals.

Yet another structural VAR identification approach imposes signs on the impulse responses to a particular shock for a certain horizon. Mountford and Uhlig (2009) identify a contractionary tax-shock as a shock, which leads to non-negative responses in government revenue during the first year after impact. Additionally, this tax-shock is identified by requiring it to be orthogonal to a business cycle shock and a monetary policy shock – both identified through signs.[14] In particular, the orthogonality to business cycle fluctuations aims at controlling for movements in the government's budget caused by automatic stabilizers.

We compare uncertainty associated with the estimated impulse responses resulting from the above mentioned four identification approaches using the same data, and the same specification (as far as possible) of the underlying (reduced-form) VAR. That is, we use Blanchard and Perotti's (2002) structural VAR approach as well as Ramey's (2011) strategy to incorporate her narrative series in a VAR to identify the effect of government spending. Further, we use Mountford and Uhlig's (2009) sign-restriction scheme and Mertens and Ravn's (2014) proxy-VAR to assess the effect of tax-shocks.

The choice of variables and the sample period is largely determined by the "highest minimal requirement" across the above identification approaches. The benchmark VAR is estimated in logs of GDP, logs of private consumption, logs of non-residential investment, logs of

[13] See also Stock and Watson (2012) and Montiel-Olea et al. (2016).

[14] All three shocks are identified sequentially by maximizing a penalty function which rewards responses in the desired direction and penalizes the others. Business cycle shocks are identified by assuming that they lead to co-movements in the same direction of output, consumption, investment, and government revenue. A contractionary monetary policy shocks affect responses in reserves and prices negatively and the interest rate positively.

total government spending, logs of (federal) tax receipts, logs of total non-borrowed reserves, real wages, a price index and the GDP deflator.[15] We use Ramey's (2011) news variable and Mertens and Ravn's (2011; 2012; 2014) unanticipated tax-change proxy. The data is quarterly, sampling from 1950/Q1-2006/Q4. The VAR representation in levels includes an intercept and a deterministic linear time trend. Four lags are included.

We construct inference using the residual-based bootstrap algorithm presented in Algorithm 1, incorporated in the methods discussed in Section 3, detrending on both an intercept and linear trend.[16] While Ramey's (2011) news series is included in the VAR, and thus, bootstrapped "endogenously", we jointly draw (with replacement) from the reduced-form residuals and Mertens and Ravn's (2012; 2014) external variable to account for uncertainty in estimating the effects of tax-shocks using this proxy.

In order to make results somewhat comparable, impulse responses are normalized such that the point estimate of the response of the policy instruments has a peak at unity across different identification approaches (see for example Ramey, 2011). As a measure of uncertainty we plot 68% confidence intervals, which is standard in the fiscal policy literature.[17]

Figure 7 and Figure 8 display unrestricted VAR in levels (estimated by OLS), FDB bagging (with AIC selection), and WIMP confidence bands (using the same specifications as in Section 4) of impulse responses due to a government spending shock. For the recursive VAR as in Blanchard and Perotti (2002), all three measures of uncertainty suggest that government spending shocks generate an initial boost in GDP. While the FDBb intervals indicate a rather moderate increase relative to the OLS intervals, the WIMP intervals imply maximum multiplier effects greater in range (roughly between 0.7 and 1.5). Considering impulse responses following Ramey's news shocks, it seems to be less clear whether government spending stimulates output or not. While the OLS confidence bands (and to a lesser extend the FDBb bands) support findings in the literature suggesting a short-lived boost in GDP, the WIMP intervals indicate greater uncertainty associated with the output response. Indeed, "robust" spending peak multipliers range between 0 and 3.3, such that a reliable conclusion on the effectiveness of spending policies cannot be made in this case.

Confidence intervals of impulse responses following a contractionary tax-shock are displayed in Figures 9 and 10. Qualitatively, responses of GDP and its main aggregates are rather similar across both identification approaches and across all three inferential procedures: Output, consumption, and investment decrease significantly. The long-lived contraction in economic activity is, however, accompanied by an equally lengthy decline in government spending, which hinders the interpretation of the identified shocks as "pure" tax-shocks.

[15] A detailed description of the data is given in the appendix.

[16] We did not find strong evidence of heteroskedasticity in the reduced-form residuals and refrain from using a robust bootstrap procedure such as the moving block bootstrap (Brüggemann et al., 2016). All approaches outlined in this paper could be easily extended in this way.

[17] The data set as well as a MATLAB toolbox for the WIMP method with the identification schemes used used in this section are available at http://researchers-sbe.unimaas.nl/stephansmeekes.

Quantitatively however, the implied response of output is much greater in the proxy VAR framework compared to the SVAR one. Intervals for peak multipliers include -6 for the former, and -3 for the latter.

Similar to the responses due to a government spending shock, the FDBb intervals are not necessarily wider than the OLS intervals. However, when considering the impact on output, and in contrast to scenario investigated above, the two intervals do not intersect at times and the FDBb intervals imply a significantly smaller impact on economic activity. This holds for both the shocks of Mountford and Uhlig (2009) and Mertens and Ravn (2012, 2014). Reflecting potentially more conservative inference, the WIMP intervals are wider, often encompassing the OLS intervals. Yet the WIMP intervals indicate that OLS-based inference rather underrates the effect of the identified tax-shocks on almost all variables. Generally, tax-shocks estimated by the proxy VAR imply greater effects on economic activity than those identified through sign-restrictions. Moreover, the comparison with the spending shocks, supports some results in the literature suggesting that tax-cuts may be more effective in stimulating the economy. Indeed, comparing peak multipliers displayed in Figure 11 reveals that evidence suggesting that multipliers exceed unity is much stronger for tax-cut policies than for spending policies. Based on the results for Ramey's news shock, multipliers due to expansionary spending policies might even not be significant at all.

In general, the above results illustrate that ignoring uncertainty about the co-integration relations in the data, may lead to ambiguous interpretation of statistical significance. Incorporating this uncertainty via our proposed WIMP approach allows for a more confident interpretation of the results.

6 Discussion

In this paper we have shown empirically and through a simulation study that ignoring uncertainty about cointegration relations may lead to unreliable inference for (structural) impulse responses. Since the commonly used specification of the VAR in levels ignores any evidence for cointegration in the data, associated inference captures uncertainty only poorly. Also, model selection techniques, such as rank pre-estimation by sequential testing or information criteria, seem to deliver reliable inference only if evidence for the true cointegration rank is strong. In this paper we propose a novel data-driven approach to robust inference for impulse responses in the presence of uncertainty regarding the cointegration rank. Our WIMP approach is shown both by simulation and empirically to still be able to deliver meaningful (i.e. not too wide) confidence intervals while being robust to rank uncertainty. As such it provides a reliable and simple alternative to the unreliable standard approaches.

Practical implementation of the WIMP approach only requires fixed-rank (bootstrap) intervals plus the sequence of trace tests for all rank tests, which are both readily available

in any standard statistical software. While a toolbox for the WIMP methods used in our application is directly available, our approach can also easily be implemented for any desired SVAR analysis, as the fixed-rank intervals used as input for the WIMP can be based on any appropriate method, both in terms of inference method such as the bootstrap and identification scheme. Finally, the computational cost of the method is fairly low; on any modern computer bootstrap intervals for a fixed rank are fast to compute, and given that in this kind of VAR model the number of variables (and hence the number of ranks) has to be relatively low to avoid the curse of dimensionality, doing so for all ranks should pose no problem.

While the prudent construction of inference is particularly important for impulse responses, our proposed WIMP procedure may equally well be beneficial when used in a different VAR context, such as forecasting. While forecast combinations across different models are well accepted as point forecasts, our WIMP method allows to construct corresponding interval forecasts that account for model uncertainty. More generally, the approach can be adapted to a variety of model selection problems, as long as we can assess the relative evidence for a particular model against a modest number of alternatives. While in theory it can be applied to high-dimensional problems as well, computationally the method is particularly suited for low-dimensional problems where the number of models is relatively small. While this is a limitation of the method, it is inherent to the simultaneous inference philosophy behind, which also holds for the PoSI method of Berk et al. (2013). Exploring the usefulness and limitations of the WIMP in more general settings is therefore an interesting avenue for future research.

A Appendix: Data

All data is quarterly, sampling from 1950/Q1-2006/Q4. We composed the data from three sources: The Bureau of Economic Analysis' *U.S. National Income and Product Accounts* (NIPA) (bea.gov/national), The Bureau of Labor Statistics (BLS) (bls.gov), and *FRED Economic Database* hosted by the Federal Reserve Bank of St. Louis (fred.stlouisfed.org).

GDP is taken from NIPA table 1.1.5.

Investment is *gross private non-residential investment*, NIPA table 1.1.5.

Government spending is *government expenditure and gross investment*, NIPA table 3.9.5.

Government revenue is *Federal government current tax receipts* plus *contributions for social insurance* minus *income taxes from federal reserve banks*, all in NIPA table 3.2.

Real wages are *nonfarm business sector: real compensation per hour*, from the BLS.

GDP deflator is taken from NIPA table 1.1.9

Federal funds rate is taken from FRED, series code: (*fedfunds*)

Adjusted reserves is taken from FRED, series code: (*ADJRESSL*)

GDP and its components, government revenue, and adjusted reserves are transformed into real per capita values using the GDP deflator and a population measure (NIPA table 7.1).

References

Benkwitz, A., H. Lütkepohl, and J. Wolters (2001). Comparison of bootstrap confidence intervals for impulse responses of German monetary systems. *Macroeconomic Dynamics 5*, 81–100.

Berk, R., L. Brown, A. Buja, K. Zhang, and L. Zhao (2013). Valid post-selection inference. *Annals of Statistics 41*, 802–837.

Bernstein, D. and B. Nielsen (2014). Asymptotic theory for cointegration analysis when the cointegration rank is deficient. Economic Working Papers 2014-W06, Nuffield College, University of Oxford.

Blanchard, O. and R. Perotti (2002). An empirical characterization of the dynamic effects of changes in government spending and taxes on output. *The Quarterly Journal of Economics 117*(4), 1329–1368.

Breiman, L. (1996). Bagging predictors. *Machine Learning 24*, 123–140.

Bruder, S. and M. Wolf (2017). Balanced bootstrap joint confidence bands for structural impulse response functions. Technical Report No. 246, University of Zurich.

Brüggemann, R., C. Jentsch, and C. Trenkler (2016). Inference in VARs with conditional heteroskedasticity of unknown form. *Journal of Econometrics 191*, 69–85.

Bühlmann, P. and B. Yu (2002). Analyzing bagging. *Annals of Statistics 30*, 927–961.

Cavaliere, G., A. Rahbek, and A. M. R. Taylor (2010a). Cointegration rank testing under conditional heteroskedasticity. *Econometric Theory 26*, 1719–1760.

Cavaliere, G., A. Rahbek, and A. M. R. Taylor (2010b). Testing for co-integration in vector autoregressions with non-stationary volatility. *Journal of Econometrics 158*, 7–24.

Cavaliere, G., A. Rahbek, and A. M. R. Taylor (2012). Bootstrap determination of the co-integration rank in vector autoregressive models. *Econometrica 80*, 1721–1740.

Chao, J. C. and P. C. B. Phillips (1999). Model selection in partially nonstationary vector autoregressive processes with reduced rank structure. *Journal of Econometrics 91*, 227—271.

Cheng, X. and P. C. B. Phillips (2009). Semiparametric cointegrating rank selection. *Econometrics Journal 12*, S83–S104.

Cheng, X. and P. C. B. Phillips (2012). Cointegrating rank selection in models with time-varying variance. *Journal of Econometrics 142*, 201–211.

Choi, I. (2005). Inconsistency of bootstrap for nonstationary, vector autoregressive processes. *Statistics & Probability Letters 75*, 39–48.

Davidson, R. and J. G. MacKinnon (2002). Fast double bootstrap tests of nonnested linear regression models. *Econometric Reviews 21*, 419–429.

Del Negro, M. and F. Schorfheide (2011). Bayesian macroeconometrics. In J. Geweke, G. Koop, and H. van Dijk (Eds.), *The Oxford Handbook of Bayesian Econometrics*, pp. 293–389. Oxford University Press.

Del Negro, M., F. Schorfheide, F. Smets, and R. Wouters (2007). On the fit of New Keynesian models. *Journal of Business & Economic Statistics 25*, 123–143.

Efron, B. (2014). Estimation and accuracy after model selection. *Journal of the American Statistical Association 109*, 991–1007.

Giannone, D., M. Lenza, and G. E. Primiceri (2016). Priors for the long run. CEPR Discussion Paper 11261, Centre for Economic Policy Research.

Gospodinov, N. (2004). Asymptotic confidence intervals for impulse responses of near-integrated processes. *Econometrics Journal 7*, 505–527.

Gospodinov, N. (2010). Inference in nearly nonstationary SVAR models with long-run identifying restrictions. *Journal of Business & Economic Statistics 28*, 1–12.

Gospodinov, N., A. M. Herrera, and E. Pesavento (2013). Unit roots, cointegration, and pretesting in VAR models. In T. B. Fomby, L. Kilian, and A. Murphy (Eds.), *VAR Models in Macroeconomics - New Developments and Applications: Essays in Honor of Christopher A. Sims*, Volume 32 of *Advances in Econometrics*, pp. 81–115. Emerald Group Publishing Limited.

Gospodinov, N., A. Maynard, and E. Pesavento (2011). Sensitivity of impulse responses to small low-frequency comovements: reconciling the evidence on the effects of technology shocks. *Journal of Business & Economic Statistics 29*, 455–467.

Hall, P. (1992). *The bootstrap and Edgeworth expansions*. New York: Springer-Verlag.

Hjort, N. L. and G. Claeskens (2003). Frequentist model average estimators. *Journal of the American Statistical Association 98*, 879–899.

Inoue, A. and L. Kilian (2002). Bootstrapping autoregressive processes with possible unit roots. *Econometrica 70*, 377–391.

Inoue, A. and L. Kilian (2016). Joint confidence sets for structural impulse responses. *Journal of Econometrics 192*, 421–432.

Jardet, C., A. Monfort, and F. Pegoraro (2013). No-arbitrage near-cointegrated VAR(p) term structure models, term premia and {GDP} growth. *Journal of Banking and Finance 37*, 389–402.

Johansen, S. (1995). *Likelihood-Based Inference in Cointegrated Vector Autoregressive Models*. Oxford University Press.

Kilian, L. (1998a). Accounting for lag order uncertainty in autoregressions: the endogenous lag order bootstrap algorithm. *Journal of Time Series Analysis 19*, 531–548.

Kilian, L. (1998b). Small-sample confidence intervals for impulse response functions. *Review of Economics and Statistics 80*, 218–230.

Kilian, L. and P.-L. Chang (2000). How accurate are confidence intervals for impulse responses in large VAR models? *Economics Letters 69*, 299–307.

Leeb, H. and B. M. Pötscher (2005). Model selection and inference: Facts and fiction. *Econometric Theory 21*, 21–59.

Liao, Z. and P. C. B. Phillips (2015). Automated estimation of vector error correction models. *Econometric Theory 31*, 581–646.

Lütkepohl, H. (1990). Asymptotic distributions of impulse response functions and forecast error variance decompositions of vector autoregressive models. *Review of Economics and Statistics 72*, 116–125.

Lütkepohl, H., A. Staszewska-Bystrova, and P. Winker (2015). Comparison of methods for constructing joint confidence bands for impulse response functions. *International Journal of Forecasting 31*, 782–798.

Mertens, K. and M. O. Ravn (2011). Understanding the aggregate effects of anticipated and unanticipated tax policy shocks. *Review of Economic Dynamics 14*, 27–54.

Mertens, K. and M. O. Ravn (2012). Empirical evidence on the aggregate effects of anticipated and unanticipated US tax policy shocks. *American Economic Journal: Economic Policy 4*, 145–181.

Mertens, K. and M. O. Ravn (2013). The dynamic effects of personal and corporate income tax changes in the United States. *American Economic Review 103*, 1212–1247.

Mertens, K. and M. O. Ravn (2014). A reconciliation of SVAR and narrative estimates of tax multipliers. *Journal of Monetary Economics 68*, 1–19.

Mikusheva, A. (2007). Uniform inference in autoregressive models. *Econometrica 75*, 1411–1452.

Mikusheva, A. (2012). One-dimensional inference in autoregressive models with the potential presence of a unit root. *Econometrica 80*, 173–212.

Montiel-Olea, J. L., J. H. Stock, and M. W. Watson (2016). Uniform inference in SVARs identified with external instruments. Mimeo.

Mountford, A. and H. Uhlig (2009). What are the effects of fiscal policy shocks? *Journal of Applied Econometrics 24*, 960–992.

Pesavento, E. and B. Rossi (2006). Small-sample confidence intervals for multivariate impulse response functions at long horizons. *Journal of Applied Econometrics 21*, 1135–1155.

Pesavento, E. and B. Rossi (2007). Impulse response confidence intervals for persistent data: What have we learned? *Journal of Economic Dynamics & Control 31*, 2398–2412.

Phillips, P. C. B. (1996). Econometric model determination. *Econometrica 64*, 763–812.

Phillips, P. C. B. (1998). Impulse response and forecast error variance asymptotics in non-stationary VARs. *Journal of Econometrics 83*, 21–56.

Ramey, V. A. (2011). Identifying government spending shocks: It's all in the timing. *Quarterly Journal of Economics 126*(1), 1–50.

Ramey, V. A. (2016). Macroeconomic shocks and their propagation. NBER Working Papers 21978, National Bureau of Economic Research.

Ramey, V. A. and M. D. Shapiro (1998). Costly capital reallocation and the effects of government spending. *Carnegie-Rochester Conference Series on Public Policy 48*(1), 145–194.

Romer, C. D. and D. H. Romer (2009). A narrative analysis of postwar tax changes. Mimeo, University of California, Berkeley.

Sobreira, N. and L. C. Nunes (2012). Testing for broken trends in multivariate time series. Mimeo, Nova School of Business and Economics.

Stock, J. H. and M. W. Watson (2012). Disentangling the channels of the 2007-2009 recession. NBER Working Papers 18094, National Bureau of Economic Research.

Strachan, R. W. and H. K. van Dijk (2007). Bayesian model averaging in vector autoregressive processes with an investigation of stability of the US great ratios and risk of a liquidity trap in the USA, UK and Japan. Econometric Institute Research Papers EI 2007-11, Erasmus University Rotterdam.

Swensen, A. R. (2006). Bootstrap algorithms for testing and determining the cointegration rank in VAR models. *Econometrica 74*, 1699–1714.

Villani, M. (2001). Bayesian prediction with cointegrated vector autoregressions,. *International Journal of Forecasting 17*, 585–605.

Wright, J. H. (2000). Confidence intervals for univariate impulse responses with a near unit root. *Journal of Business & Economic Statistics 18*, 368–373.

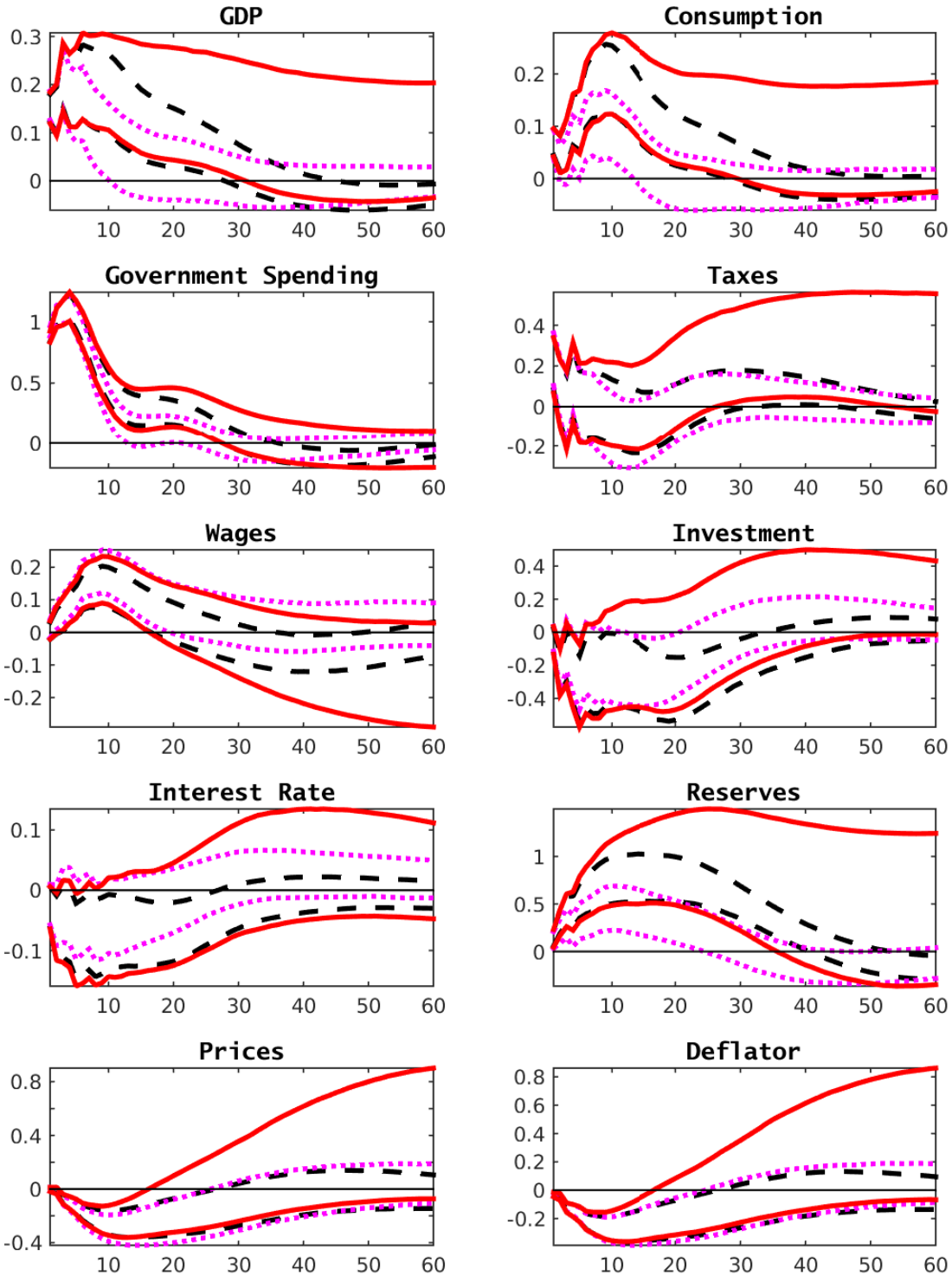

Figure 7: 68% confidence intervals of impulse responses to a government spending shock identified as in Blanchard and Perotti (2002). **Black** dashed lines are OLS intervals, pink dotted lines are FDBb/AIC intervals, **red** solid lines are WIMP intervals.

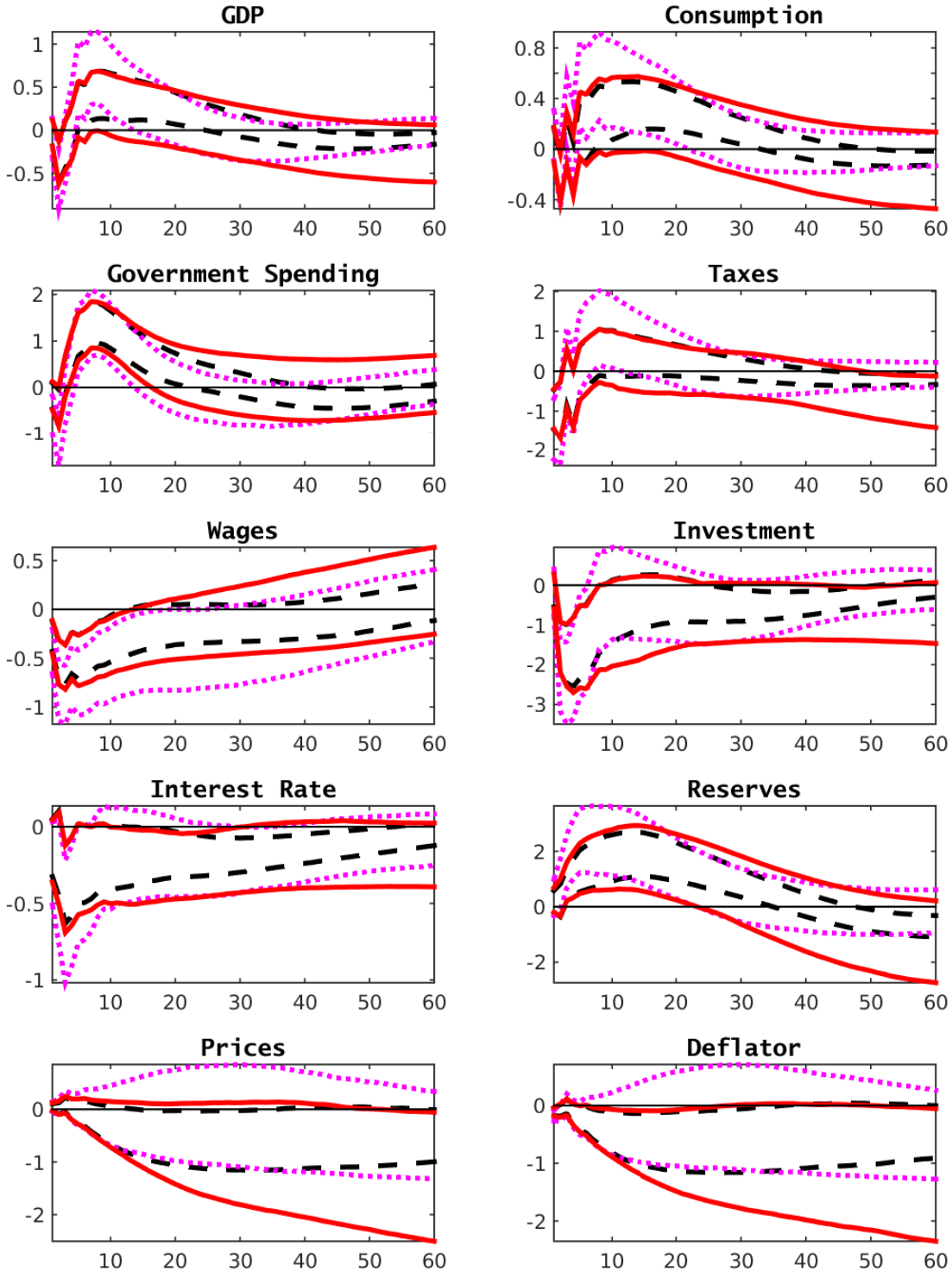

Figure 8: 68% confidence intervals of impulse responses to a government spending shock identified as in Ramey (2011). For details see Figure 7.

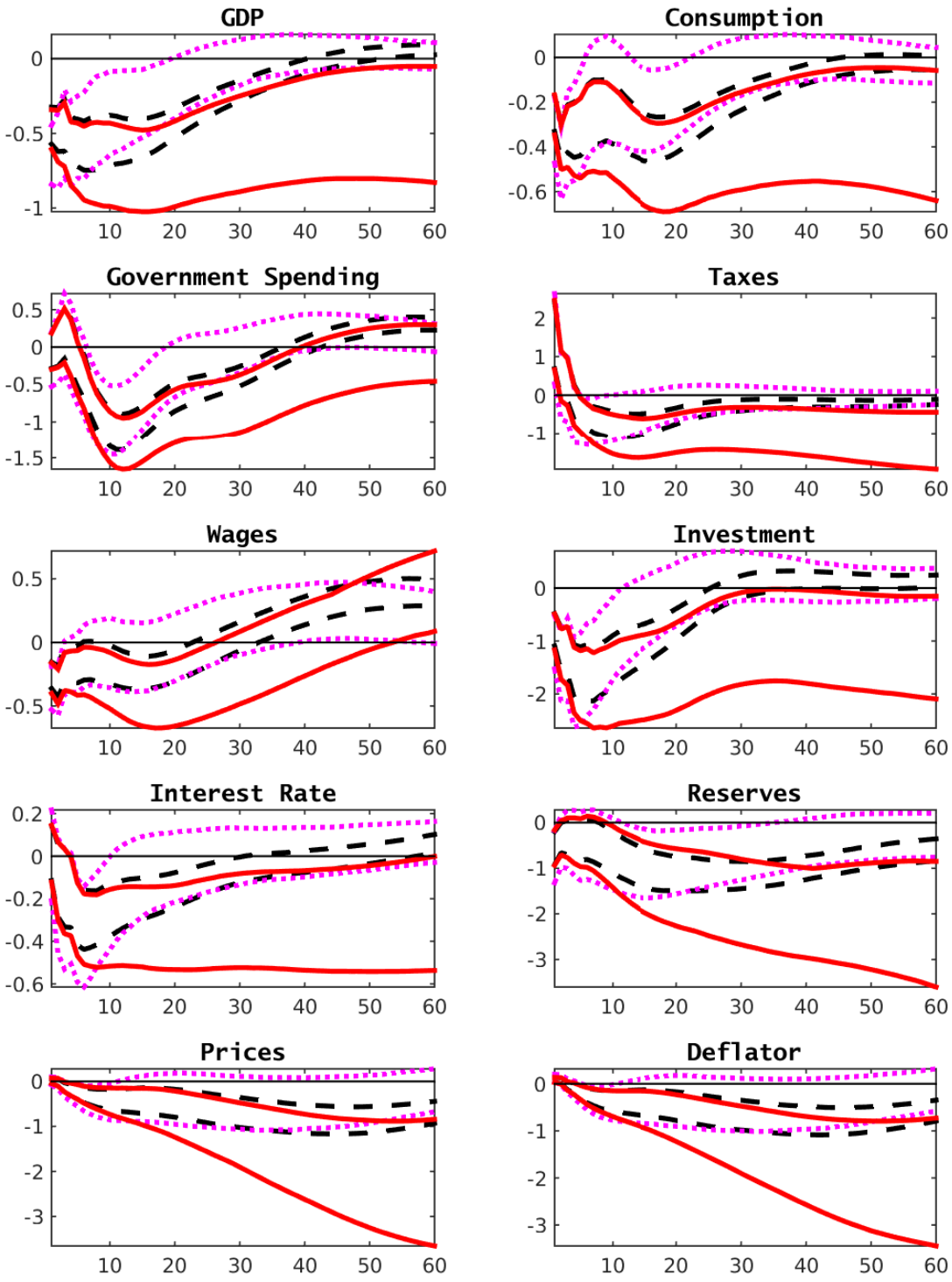

Figure 9: 68% confidence intervals of impulse responses to a tax-shock identified as in Mountford and Uhlig (2009). For details see Figure 7.

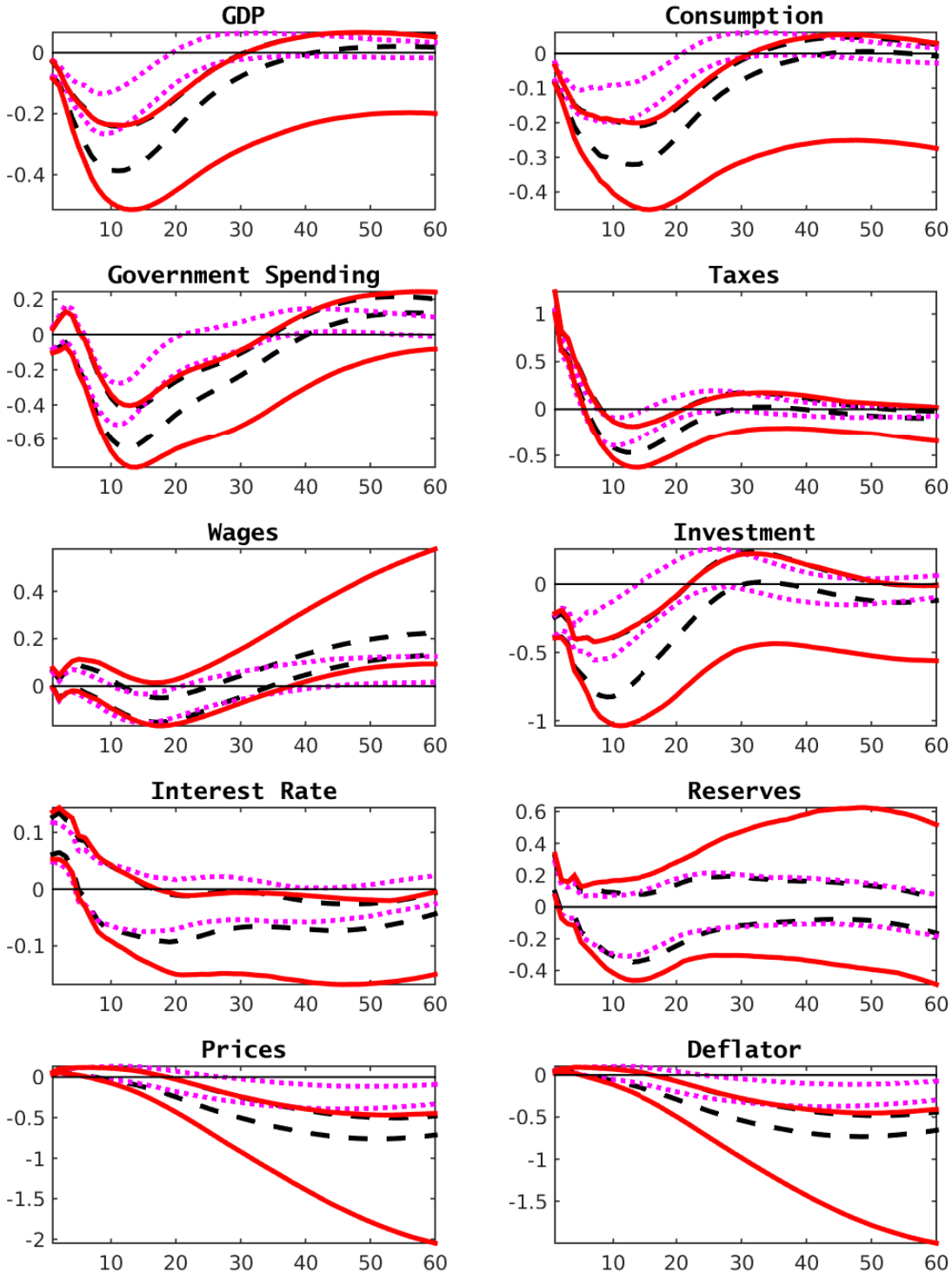

Figure 10: 68% confidence intervals of impulse responses to a tax-shock identified as in Mertens and Ravn (2012, 2014). For details see Figure 7.

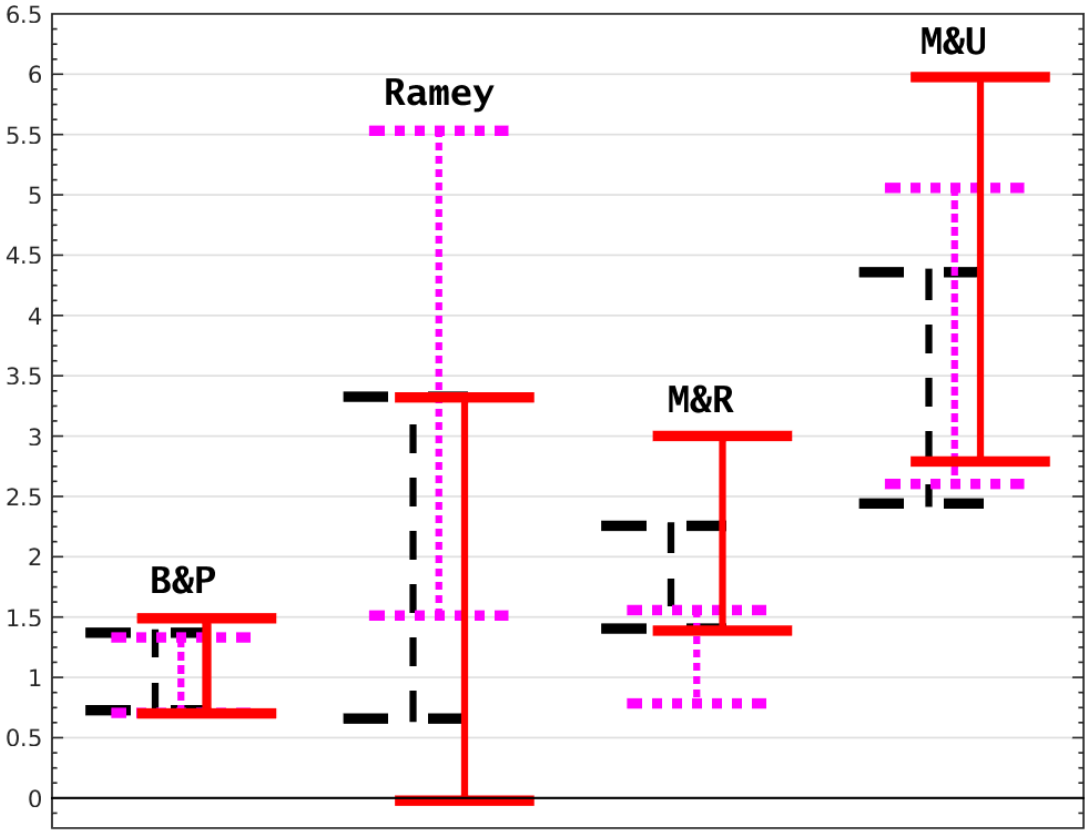

Figure 11: 68% confidence intervals of peak multipliers implied by government spending and tax-cut shocks in the analyses based on Blanchard and Perotti (2002) [**B&P**], Ramey (2011) [**Ramey**], Mountford and Uhlig (2009) [**M&U**] and Mertens and Ravn (2012, 2014) [**M&R**]. **Black** dashed lines are OLS intervals, **pink** dotted lines are FDBb/AIC intervals, **red** solid lines are WIMP intervals.

Inference for Impulse Responses under Model Uncertainty*

Lenard Lieb† Stephan Smeekes‡

Abstract

In many macroeconomic applications, impulse responses and their (bootstrap) confidence intervals are constructed by estimating a VAR model in levels - thus ignoring uncertainty regarding the true (unknown) cointegration rank. While it is well known that using a wrong cointegration rank leads to invalid (bootstrap) inference, we demonstrate that even if the rank is consistently estimated, ignoring uncertainty regarding the true rank can make inference highly unreliable for sample sizes encountered in macroeconomic applications. We investigate the effects of rank uncertainty in a simulation study, comparing several methods designed for handling model uncertainty. We propose a new method - Weighted Inference by Model Plausibility (WIMP) - that takes rank uncertainty into account in a fully data-driven way and outperforms all other methods considered in the simulation study. The WIMP method is shown to deliver intervals that are robust to rank uncertainty, yet allow for meaningful inference, approaching fixed rank intervals when evidence for a particular rank is strong. We study the potential ramifications of rank uncertainty on applied macroeconomic analysis by re-assessing the effects of fiscal policy shocks based on a variety of identification schemes that have been considered in the literature. We demonstrate how sensitive the results are to the treatment of the cointegration rank, and show how formally accounting for rank uncertainty can affect the conclusions.

JEL Classification: C15; C32; C52; E62.

Keywords: Impulse response analysis; cointegration; model uncertainty; bootstrap inference; fiscal policy shocks.

*We thank Marco Avarucci, Nalan Bastürk, Hanno Reuvers and Peter Schotman for their very helpful discussions and suggestions. We also thank conference and seminar participants at the CFE 2015, London, the NESG 2016, Leuven, and the econometrics seminar at the University of Cologne for their constructive comments. The second author thanks the Netherlands Organization for Scientific Research (NWO) for financial support.

†Department of Economics, Maastricht University, P.O. Box 616, 6200 MD Maastricht, The Netherlands. E-mail: L.Lieb@maastrichtuniversity.nl

‡Department of Quantitative Economics, Maastricht University, P.O. Box 616, 6200 MD Maastricht, The Netherlands. E-mail: S.Smeekes@maastrichtuniversity.nl

1 Introduction

Vector autoregressions (VAR) and, more importantly, their implied impulse responses (IR) are essential tools for applied macroeconomists to investigate the dynamic propagation of (structural) shocks. While VARs fitted to macroeconomic data can incorporate information about unit roots and possible cointegration relations, this evidence is regularly ignored in applied work and inference for IR coefficients is usually based on the VAR specification in levels or first-differences. A common argument for the specification in levels is that estimation by ordinary least-squares (OLS) and the associated traditional approach to inference – for example via an asymptotically normal (Lütkepohl, 1990) or a bootstrap (Kilian, 1998b) approximation – 'allows' for the presence of cointegration. Indeed the level specification results in consistent estimates of the VAR parameters regardless of the true underlying cointegration relations, and, for a fixed horizon, such inferential procedures remain valid for inference on IR coefficients. However, albeit asymptotically valid, confidence intervals may have poor coverage in small samples when the data are highly persistent and when considering responses at "longer" horizons (Kilian and Chang, 2000). Phillips (1998) shows theoretically that if one (or more) unit roots are present, confidence bands based on the normal approximation become invalid at "(very) long horizons", while Inoue and Kilian (2002) and Mikusheva (2012) show that the bootstrap also becomes invalid at such increasing horizons.

These seemingly contradicting theoretical results depend on the asymptotic framework considered; or more precisely on the notion of "(very) long horizon". If the considered horizon is kept fixed while the sample size is growing, one arrives at standard asymptotic results. However, if the horizon is modelled as a constant proportion of the sample size, the asymptotic distribution becomes non-standard if (near) unit root(s) are present. Of course, one can view the level specification as a particular form of misspecification in the presence of one or more unit roots; analogously, a wrongly specified vector error correction (VECM) formulation of the VAR suffers from similar shortcomings. Similarly, it is well known in the bootstrap literature that misspecification of the cointegration rank leads to an invalid bootstrap procedure (Choi, 2005; Inoue and Kilian, 2002; Mikusheva, 2012).

Within this growing horizon framework, Pesavento and Rossi (2006) construct confidence intervals for "long-horizon" IRs using local-to-unity asymptotics. The resulting confidence bands differ substantially from those obtained through traditional approaches, and suffer in turn from size distortions in short to medium horizons. Mikusheva (2012) proposes a procedure that works uniformly well over the entire parameter space and the entire trajectory of the IRs, but her approach only allows for the construction of uniformly valid inference if at most one "uncertain" (unit) root is present in the VAR. Similar settings and problems are considered by Gospodinov (2004, 2010), Gospodinov et al. (2011), Pesavento and Rossi (2007) and Wright (2000) among others, but all consider at most one unknown root near unity. This setting does not allow for uncertainty about the number of cointegrating relations (if any),

which we face in practice. Gospodinov et al. (2013) do consider the more general setting in an extensive simulation study and conclude that the applied researcher is best advised to estimate the system in levels and construct inference in a traditional way. Jardet et al. (2013) propose an averaging approach for impulse responses of potentially cointegrated VAR models. While they allow for uncertainty regarding the order of integration, their approach still requires a pre-selection of rank, and does not deal with inference explicitly.

In this paper we re-assess the construction of bootstrap confidence intervals for IRs in persistent, possibly non-stationary VARs. Our main intention is to provide the applied researcher with a more reliable and robust alternative to the traditional "levels" approach, independent of the IR horizon of interest. We approach the issue of choosing the cointegration rank from a model selection perspective, and draw inspiration from (bootstrap) methods initially designed to overcome model selection uncertainty in different contexts. In particular, we adapt the endogenous lag selection procedure of Kilian (1998a), the model averaging estimators of Hjort and Claeskens (2003) and the bagging approach proposed by Efron (2014) to the rank selection problem in VECMs. As elaborated by Leeb and Pötscher (2005), inference after model selection is difficult, and there is no guarantee that the above-mentioned methods can solve the problems in our setting.

Therefore, we draw inspiration from the Post-Selection Inference (PoSI) approach of Berk et al. (2013) proposed explicitly for dealing with inference after model selection to propose a novel way of constructing confidence bands by combining intervals of models for any rank. In our approach, labeled as *Weighted Inference by Model Plausibility* (WIMP), upper and lower bounds of all associated fixed-rank intervals are combined depending on the relative evidence for, or plausibility of, each model. Unlike many approaches considered in the VAR literature, our method does not require any pre-selection of ranks; that is, no pre-testing or selection using economic theory is needed. Instead, the method is fully agnostic about the cointegration rank and is fully data-driven. We provide some simple theoretical results establishing pointwise asymptotic validity of our method under general conditions. Our WIMP intervals tend to deliver coverage probabilities close to or higher than nominal levels across the entire trajectory of the IRs, even for "difficult" situations where cointegrating relations are very weak. Simulation-based evidence also suggests that the WIMP intervals generally outperform all other considered methods, including the traditional "level" approach to inference.

While we focus on frequentist inference in this paper, it is worth mentioning that rank uncertainty could also be tackled in a Bayesian VAR framework. However, in many Bayesian applications, uncertainty regarding the cointegration rank is often not taken into account explicitly. Although conceptually different, the Bayesian approach to cointegration is often similar in nature to the construction of classical (likelihood-based) inference. That is, the posterior distribution of (impulse response) parameters is often derived conditional on a predetermined rank, selected using the marginal likelihood or other model comparison approaches

(see for example Del Negro and Schorfheide, 2011, for a recent survey). However, several approaches incorporating uncertainty about the cointegration rank when analyzing VARs have been suggested in the Bayesian literature. For instance, Villani (2001) or Strachan and van Dijk (2007) propose a Bayesian model averaging scheme, similar in spirit to the approach discussed in Section 3.1.3 below. Alternatively, some authors have suggested various priors on the cointegration relations obtained using economic theory (see e.g. Del Negro et al. 2007 or Giannone et al. 2016 and references therein), which is a different conceptual approach than our fully data-driven, agnostic approach. Moreover, an explicit (theoretical) investigation of the (joint) posterior distribution of impulse responses of VARs under uncertainty on the (co-)integration relations is, however, limited also in the Bayesian literature.

Since uncertainty about the true cointegration rank is mostly ignored in applied macroeconomic research, we investigate to what extend our more robust approach(es) may change the interpretation of results in practice. More specifically, we re-evaluate the effects of fiscal policy based on four influential structural VAR frameworks. Considering Blanchard and Perotti's (2002) recursive identification strategy, Mountford and Uhlig's (2009) sign-restriction approach, Ramey's (2011) narrative VAR framework, and Mertens and Ravn's (2013; 2014) proxy-VAR, we find that neglecting rank uncertainty might lead to misleading results. As a companion to this paper, a ready-to-use MATLAB toolbox for the WIMP approach combined with various SVAR identification schemes is available online.[1]

The remainder of this paper is organized as follows. In Section 2 we discuss standard (bootstrap) approaches to inference in cointegrated VARs and illustrate empirically potential ramifications of rank misspecification. Section 3 first discusses several approaches considered in the literature about model uncertainty and their adaptations to account for rank uncertainty, and next introduces the WIMP method. The performance of the suggested methods is investigated by simulation in Section 4. Fiscal policy under rank uncertainty is analyzed in Section 5. Section 6 concludes.

2 Bootstrap Inference for Impulse Responses

2.1 The Cointegrated VAR Model and Impulse Responses

Consider the k-dimensional structural vector autoregressive (SVAR) time series process $y_t = (y_{1,t}, \ldots, y_{K,t})'$ observed at $t = 1, \ldots, T$:

$$B_0 y_t = \sum_{j=1}^{p} B_j y_{t-j} + \varepsilon_t, \tag{1}$$

[1] http://researchers-sbe.unimaas.nl/stephansmeekes

where ε_t is a K-dimensional vector of contemporaneously and serially uncorrelated, weakly stationary structural shocks[2] and B_0 is the invertible contemporaneous impact matrix. Premultiplying both sides of (1) with B_0^{-1}, we obtain the reduced-form VAR

$$y_t = \sum_{j=1}^{p} A_j y_{t-j} + u_t, \tag{2}$$

where $A_j = B_0^{-1} B_j$ and $u_t = B_0^{-1} \varepsilon_t$.

Define the lag polynomial $A(z)$ as $A(z) = I_k - \sum_{j=1}^{p} A_j z^j$, such that we can write $A(L) y_t = u_t$, where L is the lag operator $L^j y_t = y_{t-j}$. We now formulate assumptions that allow y_t to be (co)integrated with r cointegrating relations, which we label the '$I(1,r)$ conditions' as in Cavaliere et al. (2012).

Assumption 1 ($I(1,r)$ conditions)

(i) $A(z)$ has exactly $K - r$ roots equal to 1 and all other roots are outside the unit circle.

(ii) Defining $\Pi = A(1)$, we have that $\Pi = \alpha\beta'$ for $K \times r$ matrices α and β with full column rank, with the implicit definition that $\alpha\beta' = 0$ when $r = 0$.

If y_t satisfies the $I(1,r)$ conditions, we can write y_t as a VECM

$$\Delta y_t = \Pi y_{t-1} + \sum_{j=1}^{p-1} \Gamma_j \Delta y_{t-j} + u_t, \qquad t = 1, \ldots, T, \tag{3}$$

where $\Gamma_j = -\sum_{i=j+1}^{p} A_j$ for $j = 1, \ldots, p-1$.

We can invert the VAR model (2) to obtain the moving average representation

$$y_t = \sum_{j=0}^{t-1} \Psi_j u_{t-j} = \sum_{j=0}^{t-1} \Psi_j B_0^{-1} \varepsilon_{t-j} \tag{4}$$

where the Ψ_j matrices contain the reduced-form (i.e. forecast error) impulse responses and $\Phi_j = \Psi_j B_0^{-1}$ the structural impulse responses. For ease of notation later on, we directly link the impulse responses to the VECM parameters. Let $\theta = vec(\Pi, \Gamma_1, \ldots, \Gamma_{p-1})$ denote the vector of VECM parameters. Then we can define

$$\Psi_j = f_j(\theta), \qquad j = 0, \ldots, t-1,$$

where the nonlinear functions $f_j(\cdot)$ are defined implicitly through inverting the VAR model.

[2]For simplicity we assume that there is an equal number of structural shocks as variables in the system. Our model can easily be generalized to allow for a smaller number of structural shocks at the expense of complications involving the identification of the shocks. To prevent these from detracting from our main object of study, and given that these generalizations suffer from ignoring rank uncertainty in the same way as our simpler model, we abstract from this generalization in the paper.

In order to obtain structurally interpretable shocks and consequently their impulse responses $\Phi_j = \Psi_j B_0^{-1}$, we transform the estimated reduced-form errors to uncorrelated shocks. However, as B_0 is not identified, we cannot obtain Φ_j in a unique way, and estimating the structural shocks and their impulse responses requires imposing a particular identification scheme. For that purpose, let P be a $K \times K$ matrix such that $PP' = \Sigma_u$, where the specific form of P depends on the identification method. Then define the identified structural impulse responses as $\Phi_j = \Psi_j P$, and similarly

$$\Phi_j = f_j(\theta)P, \qquad j = 0, \ldots, t-1.$$

In Section 5 we discuss several ways to identify the structural shocks.[3]

2.2 Inference Conditional on a Selected Rank

We can estimate the VECM (3) for a given rank r using the Gaussian quasi maximum likelihood estimator of Johansen (1995) to obtain estimates $\hat{\Pi}^{(r)} = \hat{\alpha}^{(r)}\hat{\beta}^{(r)\prime}, \hat{\Gamma}_1^{(r)}, \ldots, \hat{\Gamma}_p^{(r)}$ and $\hat{\Sigma}_u^{(r)}$, where the superscript (r) emphasizes that estimation is conditional on r. To account for deterministic components, we can first regress y_t on a constant and possibly a linear time trend to obtain the detrended series $\tilde{y}_t = y_t - \hat{\mu}_0 - \hat{\mu}_1 t$ for $t = 1, \ldots, T$ and estimate the VECM without deterministic components on $\tilde{y}_t$.[4]

From inverting the VAR representation of the model, we can straightforwardly obtain the estimates of the moving average terms, $\hat{\Psi}_0^{(r)}, \ldots, \hat{\Psi}_h^{(r)}$, where h is the (maximum) horizon we are interested in. Letting

$$\hat{\theta}^{(r)} = vec(\hat{\Pi}^{(r)}, \hat{\Gamma}_1^{(r)}, \ldots, \hat{\Gamma}_p^{(r)}),$$

we can define the estimated impulse responses as

$$\hat{\Psi}_j^{(r)} = f_j(\hat{\theta}^{(r)}), \qquad j = 0, \ldots, h.$$

and

$$\hat{\Phi}_j^{(r)} = f_j(\hat{\theta}^{(r)})\hat{P}^{(r)}, \qquad j = 0, \ldots, h,$$

where $\hat{P}^{(r)}$ is an estimate of P such that $\hat{P}^{(r)}\hat{P}^{(r)\prime} = \hat{\Sigma}_u^{(r)}$

[3] As the impulse responses only depend on the cointegration parameters β through their product with the loadings α, that is through the error correction term $\Pi = \alpha\beta'$, we are not concerned with identification of β, unlike the setting where inference on the long run relations themselves is the objective.

[4] One could also directly incorporate deterministic components in the VECM (cf. Johansen, 1995). However, one then has to decide how the deterministic components affect the long run and short run components separately, resulting in a multitude of different specifications. Our simpler, robust, strategy corresponds to the typical approach taken in most empirical studies.

Now consider a *target impulse response* ζ, which would typically be an element of either Ψ_j or Φ_j for a fixed j; that is, we take

$$\zeta = \psi_{j,a,b} \qquad \text{or} \qquad \zeta = \phi_{j,a,b}, \tag{5}$$

where the subscript 'a, b' indicates the (a, b)-th element of the matrix. It might also be a combination of elements; for example, if one wants to perform simultaneous inference across horizons, using the ideas proposed in Bruder and Wolf (2017) and Lütkepohl et al. (2015, Section 3.6), we could take

$$\zeta = \max_{0\leq j\leq h} \psi_{j,a,b}, \qquad \zeta = \max_{0\leq j\leq h} \phi_{j,a,b}, \tag{6}$$

or its studentized versions. Similarly, one could take the Wald statistics of Inoue and Kilian (2016) as targets. The bootstrap algorithm works the same regardless of the specific target. All targets have in common that they are functions of the VAR model parameters. This way we can write both the true and estimated target impulse response as

$$\zeta = \bar{f}(\theta), \qquad \hat{\zeta}^{(r)} = \bar{f}(\hat{\theta}^{(r)}), \tag{7}$$

where the function $\bar{f}(\cdot)$ depends on the target.

We next describe a bootstrap algorithm that can be used to construct bootstrap confidence intervals for ζ. For the sake of expositional clarity, we restrict ourselves to a fairly simple, straightforward algorithm based on the bootstrap percentile method (Hall, 1992), which has regularly been considered in the literature, see e.g. Benkwitz et al. (2001).

Algorithm 1: Bootstrap Confidence Interval under Rank r

1. Let $\tilde{y}_t = y_t - \hat{\mu}_0 - \hat{\mu}_1 t$ for $t = 1, \ldots, T$ and estimate the VECM under rank r and obtain the residuals

$$\hat{u}_t = \Delta\tilde{y}_t - \hat{\Pi}^{(r)}\tilde{y}_{t-1} - \sum_{j=1}^{p-1} \hat{\Gamma}_j^{(r)} \Delta\tilde{y}_{t-j}, \qquad t = p+2, \ldots, T.$$

2. Use a bootstrap method to obtain bootstrap errors $\{u_t^*\}_{t=p+2}^T$ from the residuals $\{\hat{u}_t\}_{t=p+2}^T$.

3. Build the bootstrap sample $\{y_t^*\}_{t=1}^T$ recursively as

$$y_t^* = y_{t-1}^* + \hat{\Pi}^{(r)} y_{t-1}^* + \sum_{j=1}^{p-1} \hat{\Gamma}_j^{(r)} \Delta y_{t-j}^* + u_t^*, \qquad t = p+2, \ldots, T,$$

   using initial values $y_1^*, \ldots, y_{p+1}^*$.

4. Detrend the bootstrap sample to obtain $\tilde{y}_t^* = y_t^* - \hat{\mu}_0^* - \hat{\mu}_1^* t$ for $t = 1, \ldots, T$. Estimate the VECM under rank r on $\{\tilde{y}_t^*\}_{t=1}^T$ to obtain $\hat{\theta}^{(r)*}$. Obtain the bootstrap target impulse response as $\hat{\zeta}^{(r)*} = \bar{f}(\hat{\theta}^{(r)*})$.

5. Repeat Steps 2 to 4 B times. Let $q^*(\gamma)$ denote the γ-quantile of the B centered bootstrap statistics $\hat{\zeta}^{(r)*} - \hat{\zeta}^{(r)}$. Construct a $(1-\gamma)$-confidence interval for ζ as $\left[L^{(r)}(\gamma), U^{(r)}(\gamma)\right]$, where

$$L^{(r)}(\gamma) = \hat{\zeta}^{(r)} - q^*(1-\gamma/2) \qquad \text{and} \qquad U^{(r)}(\gamma) = \hat{\zeta}^{(r)} - q^*(\gamma/2). \tag{8}$$

Depending on the specific assumptions made on $\{u_t\}$, a variety of different bootstrap methods, such as i.i.d., wild or block bootstrap, can be used in Step 2 of Algorithm 1; we provide further details in Section 3.2.2. Similarly, different initializations in Step 3 can be used. For the simulation study and application in this paper, we use the i.i.d. bootstrap in Step 2 and initialize the bootstrap sample in step 3 by setting $y_t^* = y_t$ for $t = 1, \ldots, p+1$.

Note that $\{\hat{u}_t\}$ in Step 1 and, most importantly, $\{y_t^*\}$ in Step 3 also depend on the chosen rank r. To lighten the notation we choose not to index these formally by r, instead only emphasizing the dependence on the chosen rank r through the estimated bootstrap VAR parameters $\hat{\theta}^{(r)*}$ and target bootstrap impulse response $\hat{\zeta}^{(r)*}$. Although many variations of the bootstrap algorithm exist in the literature, such as the bias correction proposed in Kilian (1998b), all these bootstrap methods have in common that they require fixing the rank r. In particular, in generating the bootstrap sample (our step 3), it seems unavoidable to make a choice to impose a specific rank. This adds a second layer of potential rank misspecification next to the estimators themselves, which turns out to lead to further complications if one wants to account for rank uncertainty, as we discuss in Section 3 below. Before going into methods accounting for rank uncertainty however, we now first illustrate the perils of rank misspecification.

2.3 Effects of Rank Misspecification

Algorithm 1 assumes knowledge of the true cointegrating rank, labeled as r_0; if $r \neq r_0$, inference on ζ will be inappropriate, in particular for longer horizons. If the chosen rank r is smaller than the true rank, the estimated IRs converge to 'pseudo-true' values $\theta_j^{(r)}$ which are different from the true ones. This arises because the VAR parameters converge to their pseudo-true values which satisfy the (incorrect) rank restriction, c.f. Cavaliere et al. (2012). While in this case bootstrap inference remains valid for the pseudo-true parameters, these parameters can be substantially different from the true IRs, making their interpretation and therefore inference somewhat meaningless, in particular as one typically tries to uncover structural effects which requires knowledge of true parameters.

On the other hand, if $r > r_0$, as for instance in the VAR in levels specification, the short

(fixed j) and medium ($j/n \to 0$) horizon IRs are estimated consistently, but long-run ($j \sim n$) IRs are inconsistent and even random (Phillips, 1998). The inconsistency is caused by the domination of the error correction terms for the long-run IRs, and their insufficient estimation accuracy under rank misspecification. The same occurs for bootstrap inference; while valid for short and medium horizon IRs, it becomes invalid in the long-run, as demonstrated in different contexts by Choi (2005), Inoue and Kilian (2002) and Mikusheva (2012).

Figure 1 illustrates potential consequences of rank uncertainty for the construction of inference in practice. Displayed in the left panel are confidence intervals for output responses to a government spending shock identified as in Blanchard and Perotti (2002) for all possible numbers of cointegration relations.[5] Clearly, the assessment of the effectiveness of the spending policy varies drastically with the chosen cointegration rank, indicating that choosing the wrong rank hampers the interpretation of results – for long but equally so for short horizons. One could argue that with proper rank estimation, the most appropriate of these intervals can be selected. However, as demonstrated in the right panel, if evidence for a particular rank is weak, different but equally well established "respectable" rank selection procedures may suggest different models, providing little guidance for the applied researcher.

Finally, note that the unrestricted VAR in levels gives substantially different (and narrower) intervals than the VAR models with reduced rank, even the model with the next highest rank ($r = 9$). Of course, if the true model is indeed a VAR of full rank, all variables are stationary and no (co)integration would be present. However, many macroeconomic series are commonly accepted to have unit roots, which is backed up by ADF tests on our dataset, thus making the level specification unlikely to be the most appropriate. In this case, a reduced-rank VAR model would be more appropriate and constructing inference based on the VAR in levels would be invalid and could, in this example, lead to a misguided interpretation of the IRs.

The strategy to use the VAR in levels based on a robustness argument therefore appears questionable, while rank selection techniques also do not appear to give conclusive answers. It is therefore crucial to take rank uncertainty into account when conducting inference for impulse responses.

3 Inference Accounting for Rank Uncertainty

In this section we discuss several ways of accounting for rank uncertainty, first utilizing existing methods from the model uncertainty literature, before discussing a new principle.

[5]The VAR specification and the data are described in Section 5.

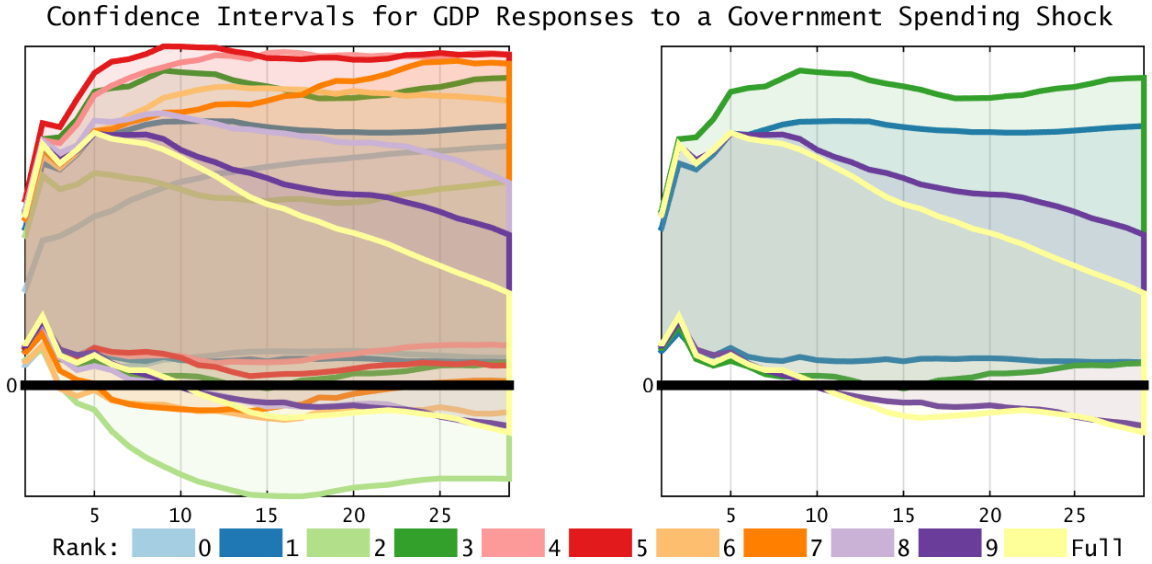

Figure 1: Left panel: Bootstrap 95% confidence intervals of the output response to a government spending shock for every rank specification. Right panel: Bootstrap 95% confidence intervals of the output response to a government spending shock implied by the trace test ($r = 3$), AIC ($r = 9$), BIC ($r = 1$), and the unrestricted VAR

3.1 Adaptations of Existing Model Uncertainty Methods

The perils of ignoring model uncertainty when performing model selection are well known in the statistical literature about model selection. For instance, in a sequence of papers, Leeb and Pötscher (see for example Leeb and Pötscher, 2005) highlight the risk of treating a selected model as a known and correct when performing inference, pointing out that even consistent model selection is no justification for treating the selected model as known. While this post-model selection inference problem is hard to solve, various methods have been proposed to at least mitigate the problem. Here we highlight some of these methods and show how they can be adapted to the problem at hand.

3.1.1 Rank Estimation

The most straightforward way to deal with rank uncertainty is to pre-estimate the rank, and then perform inference for the impulse responses conditional on the estimated rank. While this seems, given the discussion in the previous section, not always an advisable strategy, rank estimation underlies many of the methods considered afterwards. We therefore first discuss how to perform rank estimation and how it can be seen as a model selection problem.

Let the function $M_r(Y_T) : Y_T \mapsto 0, 1 \ldots, K$ be a rank selection procedure that determines the cointegration rank based on the sample $Y_T = (y_1, \ldots, y_T)'$, such that the estimated rank $\hat{r}$ is determined as $\hat{r} = M_r(Y_T)$. The estimated rank can then be imposed in the VECM

estimation to obtain the estimated target impulse responses as $\hat{\zeta}^{(\hat{r})}$.

Several methods can be considered in practice for estimation of the rank. The most common is to perform a sequence of sequential tests in the likelihood framework of Johansen (1995), in particular using the trace or eigenvalue test statistics. Instead of the standard critical values, one can also use one of its many bootstrap extensions (Cavaliere et al., 2010a,b, 2012; Swensen, 2006). Either way, due to the nature of hypothesis testing, this estimation strategy will not lead to consistent estimation of the rank (unless the significance level is chosen to decrease with sample size); the probability of selecting a rank that is too high converges to the chosen significance level instead of to zero.

Alternatively, one can use an information criterion as proposed by Phillips (1996), Chao and Phillips (1999), Cheng and Phillips (2009) and Cheng and Phillips (2012). This has two advantages compared to the sequential testing approach. First, rank selection and lag length selection can be done in a single step. Second, depending on the penalty function chosen in the information criterion, it is possible to estimate the rank consistently. A recent alternative is provided by Liao and Phillips (2015) who propose to select the rank and lag length simultaneously by penalized reduced rank regression. An advantage of this approach is that model selection and estimation are performed simultaneously, thus needing only a single step for the full estimation from start to end.

Irrespective of the chosen selection method, standard inference is based on the selected rank, treating it as known. This is often justified by the consistency of the rank selection method, but even in those cases where it is indeed consistent, ignoring the selection step leads to invalid inference as referred to earlier (Leeb and Pötscher, 2005). In particular if the data do not provide clear and strong evidence for one particular cointegrating rank, this approach will fail to deliver reliable confidence intervals. We therefore next consider methods that explicitly take rank uncertainty into account in the inference procedure.

3.1.2 Endogenous Rank Selection

Kilian (1998a) proposes the *endogenous lag selection* bootstrap method for autoregressive models where the autoregressive lag length is re-estimated within the bootstrap to account for the model selection uncertainty. We adapt his approach to rank selection, labeling this approach *Bootstrap Endogenous Rank Selection (BERS)*. Specifically, we consider the following modification to our bootstrap algorithm.

Algorithm 2: Bootstrap Endogenous Rank Selection (BERS)
Choose a rank selection method $M_r(\cdot)$, and let $\hat{r} = M_r(Y_T)$. Perform Steps 1-3 of Algorithm 1 with $r = \hat{r}$ or $r = K$. Next, replace Step 4 by

4. Let $\hat{r}^* = M_r(Y_T^*)$, where $Y_T^* = (y_1^*, \ldots, y_T^*)'$. Estimate the VECM with rank $\hat{r}^*$ on the bootstrap sample $(y_t^*)_{t=1}^T$ (after detrending) to obtain $\hat{\theta}^{(\hat{r}^*)*}$. Obtain the bootstrap target impulse responses as $\hat{\zeta}^{(\hat{r}^*)*} = \bar{f}(\hat{\theta}_j^{(\hat{r}^*)*})$.

Perform Step 5 as in Algorithm 1.

We can choose to generate the bootstrap sample Y_T^* with the "neutral" maximum rank K or the estimated rank $\hat{r}$. While Kilian (1998a) reports that this choice has little consequence for lag selection, this is very different for rank selection. After all, if the rank used to generate Y_T^* is not correct, we still face all the problems with the bootstrap as we described before. Hence, while some rank uncertainty is taken into account, the validity of this approach still hinges on the correct rank being used for the generation of the bootstrap data, which as we argued before, is impossible to guarantee.

3.1.3 Model Averaging

One of the most popular approaches to account for model uncertainty is to use model averaging (Hjort and Claeskens, 2003). By combining estimators from different models (and potentially weighting by evidence for these models), model uncertainty is taken into account. Given that the decision of which model to use is discrete, and therefore the selected model may change abruptly for a slight variation in the sample, the resulting estimators after model selection may be quite unstable and exhibit a large variability. By constructing weighted averages of the estimators arising from the individual models, one smoothes out the changes in the estimator, resulting in more stable estimators that typically display lower variability.

For this purpose we define the *Model Averaging (MA)* impulse response estimator

$$\hat{\zeta}^{MA} = \sum_{r=0}^{K} W_K(r)\hat{\zeta}^{(r)}, \quad W_K(r) = \frac{W(Y_T, r)}{\sum_{s=0}^{K} W(Y_T, s)}, \tag{9}$$

where $W(Y_T, r) : Y_T \times \{0, 1, \ldots, K\} \mapsto [0, 1]$ is a function that determines a weight for rank r based on the sample Y_T.

Unlike the typical application of model averaging, which often focuses on improving accuracy of point estimators in a mean squared error sense, we are not interested in the averaged point estimators. Instead, we only take the MA estimator as an input into our bootstrap scheme in order to construct confidence intervals: By using the more stable MA estimator, we may hope that the confidence intervals are more robust to rank misspecification. The bootstrap scheme can straightforwardly be adapted to incorporate this estimator in Step 4 of either Algorithm 1 or 2, depending on whether one wants endogenously determined weights in the bootstrap or not.

Typical weights in the model averaging literature are exponential weights based on information criteria such as BIC. However, in our simulations we find that such standard weighting schemes give weights that are too close to each other and do not differ much from simple unweighted averages. Given the widely varying behavior of impulse responses under different ranks, such weights are therefore not the most useful ones in our setting. Instead, we ad-

vocate using weights that are derived directly from cointegration tests, following the spirit of Sobreira and Nunes (2012), but rather than their KPSS type weights, we opt for weights based on the trace test statistic proposed by Johansen (1995). Details about the weights and their properties can be found in Lemma 1 in Section 3.2.2.

In a similar framework, Jardet et al. (2013) propose an averaging approach for impulse responses of potentially cointegrated VAR models based on a very specific set of weights. While they allow for uncertainty regarding the order of integration, their approach only averages two estimators: the one obtained from the VAR in levels, and one obtained from a cointegrated VAR where the number of cointegrating relations is pre-determined by pre-testing or economic theory. It can therefore not account for the general case where we are agnostic about the number of cointegration relations.

While such model averaging explicitly takes model uncertainty into account, it still relies on an explicit choice of the cointegration rank in the bootstrap algorithm to do inference. Hence, even while the weight construction can be endogenized in the bootstrap in the same way as for rank selection, the bootstrap DGP relies on the choice of a single cointegration rank. As such it still does not fully account for rank uncertainty in our context.

3.1.4 Bagging

We now take a first step in endogenizing the rank uncertainty in the bootstrap DGP itself, by bootstrapping a bagging estimator. The bagging estimator is constructed by averaging the bootstrap estimates over an initial bootstrap procedure in which the cointegration rank is re-estimated for every bootstrap sample. Bagging was originally proposed by Breiman (1996) to improve estimation accuracy of unstable estimators. Bühlmann and Yu (2002) analyzed bagging formally and found that it can lead to a variance reduction of estimation after hard decisions, such as an initial model selection. As the model averaging described above, bagging smoothes those hard decisions yielding more accurate estimators. Efron (2014) considers bagging in the context of post-selection inference, rather than point estimation, and we build on his approach here.

As bagging is essentially the simulation equivalent of model averaging, with the weights implicitly determined by how often each rank is selected within the bootstrap, it is subject to the same critique. However, one can modify the bagging algorithm to endogenize rank uncertainty in the bootstrap DGP by performing a second-level bootstrap in which we draw new bootstrap samples from the first-level bootstrap samples. By determining the rank of the second-level bootstrap DGPs from the first-level bootstrap samples, the ranks are randomized according to their evidence in the (simulated) sample. This allows to take the uncertainty into account when constructing the bootstrap confidence intervals based on the second-level bootstrap samples. While this does not fully solve the bootstrap invalidity problem (bootstrap samples are still generated under incorrect ranks, especially in the first step), the method has

the potential to alleviate the problem.

There is a computational problem with this method though, as one has B_1 iterations in the first bootstrap and B_2 in each second-level bootstrap, such that a full double bootstrap requires $B_1(1 + B_2)$ iterations which quickly becomes computationally infeasible. For this purpose we implement the Fast Double Bootstrap (FDB) developed by Davidson and MacKinnon (2002), which requires drawing only a single second-level bootstrap sample for every first-level bootstrap sample, which means the computation cost of the FDB is only double ($2B_1$) that of a regular bootstrap. The algorithm below describes the method, labeled as *FDB bagging (FDBb)*, in detail.

Algorithm 3: FDB bagging (FDBb)
Choose a rank selection method $M_r(\cdot)$, and perform steps 1-4 of Algorithm 2. Next:

5. Perform a second bootstrap procedure on the bootstrap sample $\{y_t^*\}_{t=1}^T$ to obtain double-bootstrap impulse responses. For every bootstrap sample $\{y_t^*\}_{t=1}^T$, only *one* second-level bootstrap sample has to be drawn. Specifically, take the following steps:

   (i) Estimate the VECM with rank $\hat{r}^* = M_r(Y_T^*)$ and obtain the residuals

   $$\hat{u}_t^* = \Delta\tilde{y}_t^* - \hat{\Pi}^{(\hat{r}^*)*}\tilde{y}_{t-1}^* + \sum_{j=1}^{p} \hat{\Gamma}_j^{(\hat{r}^*)*}\Delta\tilde{y}_{t-j}^*, \qquad t = p+2, \ldots, T. \tag{10}$$

   (ii) Construct the second-level bootstrap errors $\{u_t^{**}\}_{t=p+2}^T$ from $\{\hat{u}_t^*\}_{t=p+2}^T$ using the same bootstrap method as for the first level, and build the second-level bootstrap sample $\{y_t^{**}\}_{t=1}^T$ recursively as

   $$y_t^{**} = y_{t-1}^{**} + \hat{\Pi}^{(\hat{r}^*)*}y_{t-1}^{**} + \sum_{j=1}^{p} \hat{\Gamma}_j^{(\hat{r}^*)*}\Delta y_{t-j}^{**} + u_t^{**}, \qquad t = p+2, \ldots, T, \tag{11}$$

   with initial values $y_1^{**}, \ldots, y_{p+1}^{**}$.

   (iii) Estimate the cointegration rank $\hat{r}^{**} = M_r(Y_T^*)$ and use it to obtain $\hat{\zeta}^{(\hat{r}^{**})**}$.

6. Repeat Steps 1 to 5 B times. Let $\hat{\zeta}_1^{(\hat{r}^*)*}, \ldots, \hat{\zeta}_B^{(\hat{r}^*)*}$ denote the ordered sequence of the first-level bootstrap estimates obtained over the B bootstrap replications. The *bagging* estimator of the impulse response is then defined as

   $$\hat{\zeta}^{\text{bag}} = B^{-1}\sum_{b=1}^{B} \hat{\zeta}_b^{(\hat{r}^*)*}. \tag{12}$$

   Let $q^{**}(\gamma)$ denote the γ-quantile of the B centered second-level bootstrap statistics

$\hat{\zeta}^{(\hat{r}^{**})**} - \hat{\zeta}^{(\hat{r}^{*})*}$. Construct a $(1-\gamma)$-confidence interval for ζ as

$$\left[\hat{\zeta}^{(\hat{r})} - q^{**}(1-\gamma/2), \hat{\zeta}^{(\hat{r})} - q^{**}(\gamma/2)\right].$$

3.2 Weighted Inference by Model Plausibility

None of the methods described above fully address the post-model selection inference problem. To work towards a more satisfactory solution, we now combine the ideas discussed above with new concepts arising from the recent statistical literature that directly addresses the post-model selection inference problem.

We would like to build on the idea of averaging or weighting models to account for rank uncertainty. However, as elaborated on in the previous section, such weighting is typically designed for point estimation and translating it to confidence intervals, as needed here, is not so straightforward. In order to make the transition, we take inspiration from the perspective taken by Berk et al. (2013), who view the issue of constructing valid post-model selection inference as a simultaneous inference problem: by controlling for performing inference in all models simultaneously, the specific model selected by a model selection procedure is covered by construction. In our notation, and following their approach, we could construct intervals $[\hat{\zeta}^{\hat{r}} - q^{\text{PoSI}}(1-\gamma/2), \hat{\zeta}^{\hat{r}} - q^{\text{PoSI}}(\gamma/2)]$ such that

$$\mathbb{P}\left(q^{\text{PoSI}}(\gamma/2) \leq \hat{\zeta}^{(r)} - \zeta^{(r)} \leq q^{\text{PoSI}}(1-\gamma/2), \quad \forall r \in \{0, K\}\right) \to 1-\gamma$$

as $T \to \infty$. It is important to note that here $\zeta^{(r)} = \bar{f}(\theta^{(r)})$ is a *pseudo-true* parameter defined in terms of $\theta^{(r)}$, the pseudo-true parameters of the model (2) under the restriction that rank r is imposed – see Lemma 1 and its proof in Cavaliere et al. (2012) for a formal definition. These parameters represent the probability limits of the estimators of (2) under the restriction of imposing rank r, and can informally be seen as those parameters which minimize a distance to the true parameters under the restriction that the cointegration rank is r. If $r < r_0$, the true parameter cannot be recovered, and therefore the pseudo-true parameter will be different.

For our purposes, there is a fundamental problem with the *sub-model* view of Berk et al. (2013) where the pseudo-true parameters are the objects of interests. In the context of structural impulse responses, the sub-model view has little relevance, as it cannot uncover any structural effects. We therefore need the *full model* view, in which it is assumed that one of the models is the true (structural) one. Denoting this extension of the PoSI approach as PoSI_0, we seek to control

$$\mathbb{P}\left(q^{\text{PoSI}_0}(\gamma/2) \leq \hat{\zeta}^{(r)} - \zeta \leq q^{\text{PoSI}_0}(1-\gamma/2), \quad \forall r \in \{0, K\}\right) \to 1-\gamma$$

as $T \to \infty$. This implies that we require that the distance between every fixed-rank estimate $\hat{\zeta}^{(r)}$ and the true impulse response ζ is taken into account in constructing the confidence

intervals, rather than the much shorter distance between $\hat{\zeta}^{(r)}$ and its probability limit or pseudo-true impulse response $\zeta^{(r)}$, resulting in rather wide intervals. The seemingly only way to control this quantity is to construct confidence intervals for every rank separately, and then take the union of these, which will typically result in very wide intervals that are useless in practice.

We have not yet considered any evidence on the plausibility of each rank, that can be extracted from the data. If this information can incorporated into our inferential procedure, we may be able to achieve intervals that are still useful in applications, as the impact of ranks that the data deem very implausible can be eliminated, or at least reduced. We therefore augment the PoSI view of simultaneous inference by a weighting scheme akin to model averaging, except that we apply the weighting not to the estimators but directly to the bounds of the intervals. The direct weighting of the inference output, in this case the interval bounds, by evidence of the plausibility of each model, leads us to label our approach as *Weighted Inference by Model Plausibility (WIMP)*.

3.2.1 The WIMP Principle

Define the most plausible model - according to a certain plausibility measure based on the data - as the *reference model*, and denote the corresponding confidence interval arising from this model (ignoring model uncertainty) as the *reference interval*. As input to the WIMP procedure we consider all *model intervals*, which are defined as the confidence intervals obtained by assuming any particular model as the true one. In our case these would be the intervals obtained by imposing all the $K+1$ different cointegrating ranks. Before going into the details of our application, we now describe the general conditions that any prudent WIMP scheme must adhere to:

WIMP Prudence Conditions

1. The WIMP confidence interval must always cover at least the reference interval. That is, any non-reference model can only lead to widening the WIMP interval compared to the reference interval.

2. If two models are equally plausible, the model interval bounds which are furthest away from the reference model must contribute the most to widening the WIMP interval.

3. If the bounds of two model intervals are equally far away from the reference interval, the most plausible model must contribute the most to widening the WIMP interval for a given distance of the bounds from the reference interval.

4. The WIMP confidence interval may not be wider than the interval obtained by joining all individual model intervals.

The first condition is required to avoid invalid intervals, in whatever way validity is measured. If it is possible to obtain a confidence interval which is more narrow than the "standard" interval assuming no model uncertainty, the WIMP interval can never be guaranteed to contain an adequate coverage probability. The second condition ensures that the locations of intervals in relation to the reference interval are properly taken into account for equally plausible models. Compare two equally plausible models with almost identical intervals, to two equally plausible models with very different intervals. Any prudent method of accounting for model uncertainty must result in wider intervals for the second case than for the first case. The third condition implies that one has to take more plausible models more strongly into account than implausible models. In particular, this condition allows to reduce the impact of implausible models that may have very different intervals than the reference model but are so implausible, that there is little to no uncertainty about them. Finally, the fourth condition ensures that the WIMP intervals do not become too conservative. While the first and fourth condition impose hard (but sensible) restrictions on the WIMP intervals, the second and third conditions allow for quite some variation in the procedure. Finding a right balance between conservatism and interval length is therefore of great practical importance, and varies per setting.

We now turn to our specific implementation of the WIMP Prudence Conditions. Let $W_K(r)$ be model plausibility weights assigned to all ranks $r = 0, \ldots, K$ and define $X(r,s) = \frac{W_K(r)}{W_K(s)}$ as the relative plausibility of rank r compared to rank s. Letting $R = \arg\max_{0 \le r \le K} W_K(r)$ be the most plausible or reference rank, we define the WIMP interval as $\left[L^{\text{WIMP}}(\gamma), U^{\text{WIMP}}(\gamma)\right]$ with

$$\begin{aligned} L^{\text{WIMP}}(\gamma) &= \min_{r=0,\ldots,K} \left\{ L^{(R)}(\gamma) - X(r,R) \left[L^{(r)}(\gamma) - L^{(R)}(\gamma) \right]^- \right\}, \\ U^{\text{WIMP}}(\gamma) &= \max_{r=0,\ldots,K} \left\{ U^{(R)}(\gamma) + X(r,R) \left[U^{(r)}(\gamma) - U^{(R)}(\gamma) \right]^+ \right\}, \end{aligned} \tag{13}$$

where $x^+ = \max(x, 0)$, $x^- = -\min(x, 0)$ and $L^{(r)}(\gamma)$ and $U^{(r)}(\gamma)$ are the lower and upper bounds respectively of the confidence intervals with fixed rank r as defined in (8).

The term $\left[L^{(r)}(\gamma) - L^{(R)}(\gamma)\right]^-$ (respectively $\left[U^{(r)}(\gamma) - U^{(R)}(\gamma)\right]^+$) ensures that only lower bounds smaller (upper bounds larger) than those of the reference interval are taken into account; for lower bounds larger (upper bounds smaller) than those of the reference interval, this term is simply zero. Together with $X(r,s) \ge 0$, this implies that the WIMP interval always contains the reference interval, hence Condition 1 is satisfied. Condition 2 is also trivially satisfied as this term increases when the lower (upper) bound of the rank r interval is further away from the reference interval.

The shape of $X(r,s)$ determines how strongly less plausible models are taken into account and can be different from the linear function of $W_K(r)$ imposed above. As long as $X(r,s)$ is an increasing function of $W_K(r)$, more plausible ranks are given more importance and Condition 3 is satisfied; varying $X(r,s)$ and $W_K(r)$ allows one to change the balance between

conservatism and interval length. Finally, with respect to Condition 4, note that as long as $X(r,s) \leq 1$, the WIMP interval can never be wider than the interval obtained by combining the smallest lower bound with the largest upper bound.[6]

Two final remarks about the WIMP principle are in order. First, although we focus here exclusively on the case of rank uncertainty, other types as uncertainty, such as about the lag order or the deterministic components can be incorporated into the WIMP procedure as well. For instance, if one want to allow for P different lag orders in addition to the $K+1$ ranks, one needs weights that measure the plausibility of each of the $(K+1)P$ different models resulting from combining the different ranks and lag orders. In this paper we focus on rank uncertainty only as it has a far bigger and more fundamental impact than (slight) lag misspecification. Moreover, successful methods exist for accounting for lag uncertainty, such as Kilian's (1998a) endogenous lag selection bootstrap. One may therefore also opt for accounting for lag order uncertainty through the fixed rank intervals that form the input to the WIMP.

Second, note that the WIMP intervals are not built directly around a single point estimator for ζ. While all $K+1$ fixed-rank estimators are incorporated through their respective confidence intervals, we do not directly obtain a corresponding point estimate for ζ. Of course, if there is a desire to pair the confidence interval with a point estimator, one can do so, in which case the model averaging estimator with the same weights $W_K(\cdot)$ as used for the WIMP intervals is the most natural candidate.[7]

3.2.2 Asymptotic Properties

In this section we derive asymptotic properties of the WIMP intervals. We mainly do so under general high-level assumptions on the tests and bootstrap method available, but we will also provide some details about how these assumptions can be verified in our application. We first turn to the pointwise asymptotic validity of our method.

Theorem 1. *Let Y_T be generated according to (2), and let $\Theta^{(r)}$ denote the parameter space of θ such that the $I(1,r)$ conditions are satisfied. Then assume that*

(i) As $T \to \infty$, $W_K(r) \xrightarrow{p} \mathbb{1}(r = r_0)$, where $\mathbb{1}(\mathcal{A})$ is equal to 1 if $\mathcal{A}$ is true, and 0 otherwise;

(ii) For $r = r_0$ the bootstrap confidence interval has correct coverage; that is, as $T \to \infty$, we

[6] If some of the individual model intervals are disjoint, the "maximal" WIMP interval as constructed in (13) is larger than the union of these intervals, apparently violating Condition 4. However, this is an intentional violation. Such a disjointed confidence *set* is not a confidence *interval* any more, and therefore is rather awkward to interpret. The natural modification of this set, that yields an interval again, would be to "fill the gaps" and extend it from the lowest lower bound to the highest upper bound, which is exactly what the WIMP construction does automatically.

[7] As expected from the model averaging literature, unreported simulations in the same setup as considered in Section 4 show that this estimator performs very well in terms of mean squared error when compared to fixed-rank estimators. Of course, its performance purely as a point estimator is different from its performance as basis for inference, as we shall see in Section 4.

have that

$$\mathbb{P}\left(L^{(r_0)}(\gamma) \leq \zeta \leq U^{(r_0)}(\gamma)\right) \to 1-\gamma \qquad \forall \theta \in \Theta^{(r_0)} \quad \forall r_0 \in \{0, K\}.$$

Then

$$\mathbb{P}\left(L^{\text{WIMP}}(\gamma) \leq \zeta \leq U^{\text{WIMP}}(\gamma)\right) \to 1-\gamma \qquad \forall \theta \in \Theta^{(r_0)}, \quad \forall r_0 \in \{0, K\}.$$

as $T \to \infty$.

Proof. By Assumption (i), we have that $\mathbb{P}(R = r_0) \to 1$ and consequently that $X(r, R) \xrightarrow{p} \mathbb{1}(r = r_0)$. It therefore follows directly that $L^{\text{WIMP}}(\gamma) = L^{(R)} \xrightarrow{p} L^{(r_0)}(\gamma)$ and $U^{\text{WIMP}}(\gamma) \xrightarrow{p} U^{(r_0)}(\gamma)$. The result then follows from assumption (ii). □

Assumption (ii), which implies asymptotic validity of the intervals under a known rank, has been verified for many bootstrap methods under different assumptions on $\{u_t\}$ (or equivalently $\{\varepsilon_t\}$). For instance, if we assume that $\{u_t\}$ is i.i.d. with sufficiently many moments existing, one can show that the i.i.d. bootstrap version of Algorithm 1 satisfies assumption (ii), c.f. Kilian (1998b) and Cavaliere et al. (2012). Inoue and Kilian (2016) also formulate general assumptions to assure bootstrap validity, while alternative methods that allow for heteroskedasticity are considered by Brüggemann et al. (2016). The WIMP principle can be applied to any of these methods and targets; in fact, it does not even require bootstrap confidence intervals, but can equally well be applied to any asymptotically valid inference method.

We now propose a concrete weighting scheme that satisfies Assumption (i) in Theorem 1. Following the spirit of Sobreira and Nunes (2012), we base our weights on cointegration tests. Rather than their KPSS type weights, we opt for weights based on the trace test statistic proposed by Johansen (1995), which, as a "standard" cointegration test, has intuitive appeal and is available in all standard econometric and statistical software.[8]

Lemma 1. *Let* $J_T(r) = -T\sum_{i=r+1}^{K} \ln(1-\hat{\lambda}_i)$ *denote the trace test of Johansen (1995) for testing* $H_0 : r_0 \leq r$. *For constants* $c_1 > 0$ *and* $0 < c_2 < 1$, *define*

$$\begin{aligned} W(Y_T, r) &= e^{-c_1 T^{-c_2} J_T(r)} && \text{for } r = 0. \\ W(Y_T, r) &= e^{-c_1 T^{-c_2} J_T(r)} - e^{-c_1 T^{-c_2} J_T(r-1)} && \text{for } r = 1, \ldots, K-1, \\ W(Y_T, r) &= 1 - e^{-c_1 T^{-c_2} J_T(r-1)} && \text{for } r = K, \end{aligned} \tag{14}$$

and $W_K(r) = W(Y_T, r) / \sum_{r=0}^{K} W(Y_T, r)$. *Then* $W_K(r) \xrightarrow{p} \mathbb{1}(r = r_0)$ *as* $T \to \infty$.

[8] We also explored Johansen's (1995) maximum eigenvalue test statistic, which similarly satisfies assumption (i) in Theorem 1. Numerical experiments showed virtually no difference with the trace test.

Proof. It follows from Johansen (1995) and Bernstein and Nielsen (2014) that for all $r \geq r_0$, $J_T(r) = O_p(1)$, such that $T^{-c_2} J_T(r) \xrightarrow{p} 0$, while for $r < r_0$, we have that $J_T(r)/T$ is tight, such that $T^{-c_2} J_i(r) = T^{1-c_2} J_T(r)/T \xrightarrow{p} \infty$. Therefore we have that

$$e^{-c_1 T^{-c_2} J_T(r)} \xrightarrow{p} \mathbb{1}(r \geq r_0) \quad \Rightarrow \quad W_T(r) \xrightarrow{p} \mathbb{1}(r = r_0). \qquad \square$$

While the results above establish the pointwise asymptotic validity of our proposed scheme, it should be noted that this does not imply uniform validity, that is, the property

$$\liminf_{T\to\infty} \inf_{\theta\in\Theta} \mathbb{P}\left(L^{\text{WIMP}}(\gamma) \leq \zeta \leq U^{\text{WIMP}}(\gamma)\right) \geq 1-\gamma, \quad \text{where } \Theta = \bigcup_{r=0}^{K} \Theta^{(r)}.$$

Uniform validity is a more informative property about finite sample behavior of the intervals, as it does not rely on the *oracle property* that the true rank is always selected asymptotically, as is assumed in Assumption (i) in Theorem 1. In particular for small deviations from a certain rank, the weights are unlikely to pick this up, so the oracle property in Assumption (i) is a poor approximation to finite sample performance and can be very misleading. In fact, the same pointwise reasoning underlies the use of consistency of information criteria like BIC as a valid approach to model uncertainty, and should therefore be treated with caution, see e.g. Leeb and Pötscher (2005).

However, while clearly of great interest, uniform validity is very hard to establish for the cointegrated VAR based on bootstrap inference, as it requires the consideration of sequences of local deviations from certain ranks, under which the bootstrap is known to have problems. So far uniform results have only been established in the presence of a single local-to-unit root (cf. Mikusheva, 2007, 2012), while more general results are needed for our setting, and are to the best of our knowledge unavailable. Establishing a full uniform asymptotic theory is therefore outside the scope of this paper and left for future research. Here we focus on evaluating the small sample properties of the WIMP method for situations where rank uncertainty is present. Note that even though the asymptotic validity of our WIMP implementation is based on the same oracle properties used to validate consistent rank selection, unlike these methods our WIMP intervals do explicitly take rank uncertainty into account, and are always wider in finite samples than the fixed-rank intervals. We therefore expect that the WIMP intervals will be much more reliable in small samples when even minor rank uncertainty is present.

4 Monte Carlo Simulations

In this section we investigate the performance of the various methods discussed above by simulation. We assess coverage probabilities (CP) of confidence bands for *forecast error impulse responses*, and hence evaluate intervals for the moving average parameters. As such

we base our analysis fully on the reduced-form VAR, and do not consider structural VARs. We intentionally abstract from the identification problem in structural VARs, since the structural moving average parameters are linear combinations of their reduced-form counterparts, and one can expect that the performance of one inferential procedure for reduced-form parameters is inherited by the structural parameters.[9]

The data generating process (DGP) for the Monte Carlo experiment is a three-dimensional VAR of order one inspired by Phillips (1998), given by $y_t = (I_3 + \Pi)y_{t-1} + \epsilon_t$, with $\epsilon_t \sim$ *i.i.d.* $\mathcal{N}(0, I_3)$ for all t. The cointegration matrix is specified as $\Pi = d_1\alpha_1\beta_1' + d_2\alpha_2\beta_2'$, where $\alpha_1 = (0, 1, 0)'$, $\alpha_2 = (0, 0, 1)'$, $\beta_1 = (2, -1, 0)'$, and $\beta_2 = (1, -1, -1)'$. We consider two versions of the above process when simulating data.

DGP1: Setting $d_1 = 0.05$ and $d_2 = 0.02$ implies that the model has one root at unity and two roots close to one at 0.98 and 0.95. Thus, we have two "*weak*" cointegration relations.

DGP2: Setting $d_1 = d_2 = 1$ implies a VAR with one unit root and two roots at zero, thus two "*strong*" cointegration relations. This is the original setting considered by Phillips (1998).

We evaluate CPs of 95% confidence intervals for each response and horizon ($h = 1, 2, ..., 60$) for $T = 100, 200$. The results are based on 1000 MC simulations and 399 bootstrap replications. To compute the WIMP intervals we set $c_1 = 1$ and $c_2 = 0.5$ for the weights in (14).[10] As mentioned above we do not consider identification of structural IRs. We also abstract from lag length selection (we fix $p = 1$),[11] deterministic components, and small sample bias correction (Kilian, 1998). All simulations were done in MATLAB.

Figure 2 and 3 display CPs of the various inferential procedures discussed above for DGP1 for $T = 100$ and $T = 200$. Based on the two model selection criteria employed, we can partly confirm the findings of Gospodinov et al. (2013). That is, if evidence for a particular rank is weak, pre-testing seems not to deliver more accurate inference than (bootstrap) CIs based on unrestricted OLS. This holds for both sample sizes considered. However, these two frequently used approaches can both not be considered as reliable strategies for the construction of inference – minimum CPs are well below 60%. Surprisingly, even when the true model specification is imposed (which could be considered to be the *oracle* method), CPs are generally not closer to the nominal level either; both in short and long horizon.

[9] Except for SVARs identified through long-run restrictions, the exact persistence properties of the underlying reduced-form process are of no direct relevance for identification.

[10] This choice of parameters seems to be the most natural for the weights in (14). We did not experiment with changing these values, as the performance in the simulations was already quite satisfactory. It is likely that by careful tuning these parameters, even better performance can be obtained. However, the optimal choice will typically be highly case-dependent, and optimal values should therefore be treated with caution. Instead we prefer to report results for a natural albeit naive choice of parameters without claiming any optimality.

[11] Unreported results with $p = 3$ show the same patterns as $p = 1$.

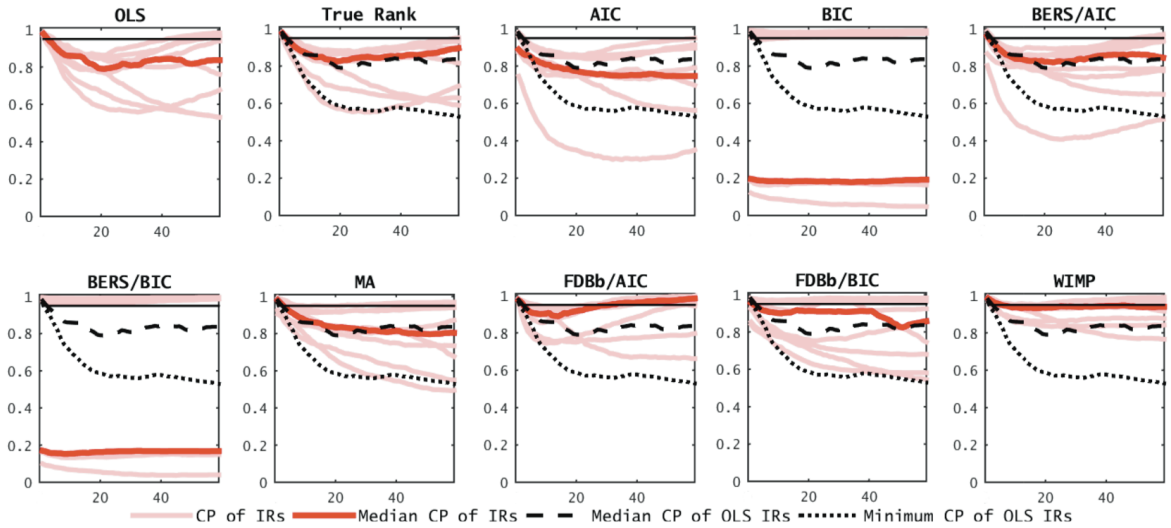

Figure 2: DGP1: Empirical coverage rates for ten inference methods for $T = 100$.
'**OLS**' refers to the (unrestricted) VAR in levels estimated by OLS; '**True Rank**' refers to the VECM estimated with knowledge of the true rank '**AIC**' and '**BIC**' refer to the rank estimation of Section 3.1.1 using AIC and BIC, respectively; '**BERS/AIC**' and '**BERS/BIC**' refer to the Bootstrap Endogenous Rank Selection of Section 3.1.2 with respectively AIC and BIC used for rank selection; '**MA**' refers to the model averaging method of Section 3.1.3 with weights as in (14); '**FDBb/AIC**' and '**FDBb/BIC**' refer to the FDB bagging method of Section 3.1.4 with respectively AIC and BIC used for rank selection; '**WIMP**' refers to the WIMP method of Section 3.2 with weights as in (14).
The pink lines show CPs for all nine impulse responses; the **red** line is the median of these per horizon. For ease of comparison, the median and minimum coverage of the OLS intervals is always reported in **black**.

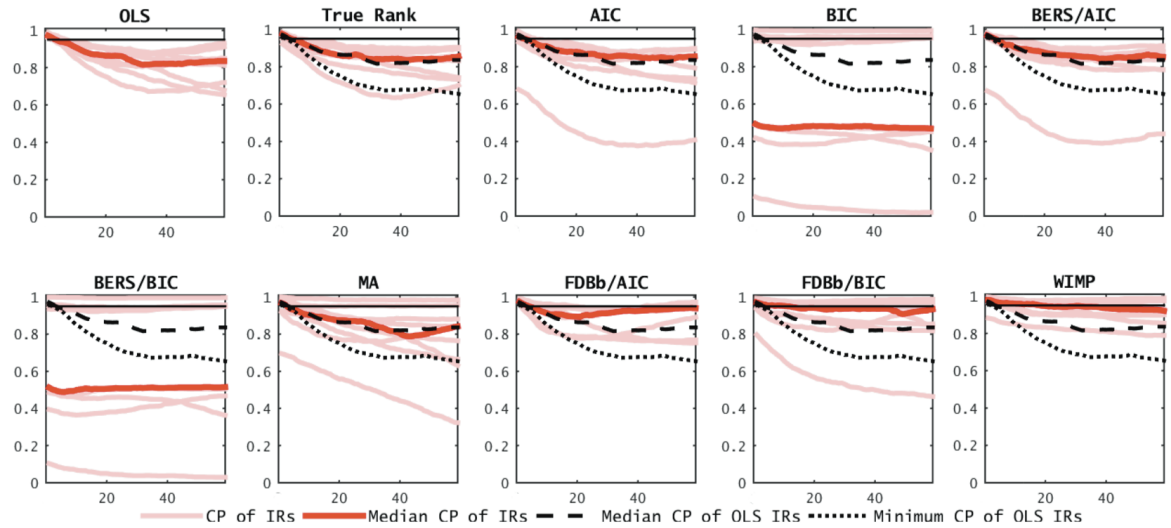

Figure 3: DGP1: Empirical coverage rates for the various inference methods for $T = 200$. See Figure 2 for details.

Endogenous rank selection does not seem to improve the performance compared to the pre-testing procedure. FDB bagging does give CPs closer to nominal level, in particular when based on AIC. However, the WIMP intervals outperform all other methods, and deliver CPs that are on average quite close to the 95% nominal level.

Figure 4 presents the corresponding average width of the bootstrap intervals over all horizons for the five most relevant methods. There are several interesting observations to make from this figure. First, note that even though FDB bagging and WIMP produce much more accurate intervals than OLS or imposing the true rank, they actually do not produce intervals that are much wider. It of course makes perfect sense that they deliver wider intervals, as the intervals of the other methods are too narrow, but the limited extent to which they are wider indicates the methods are not overly conservative. Second, even though the WIMP method produces more accurate intervals than FDB bagging, intervals are not wider. This shows that the mechanism imposed in the WIMP to reduce the impact of implausible models works well in practice.

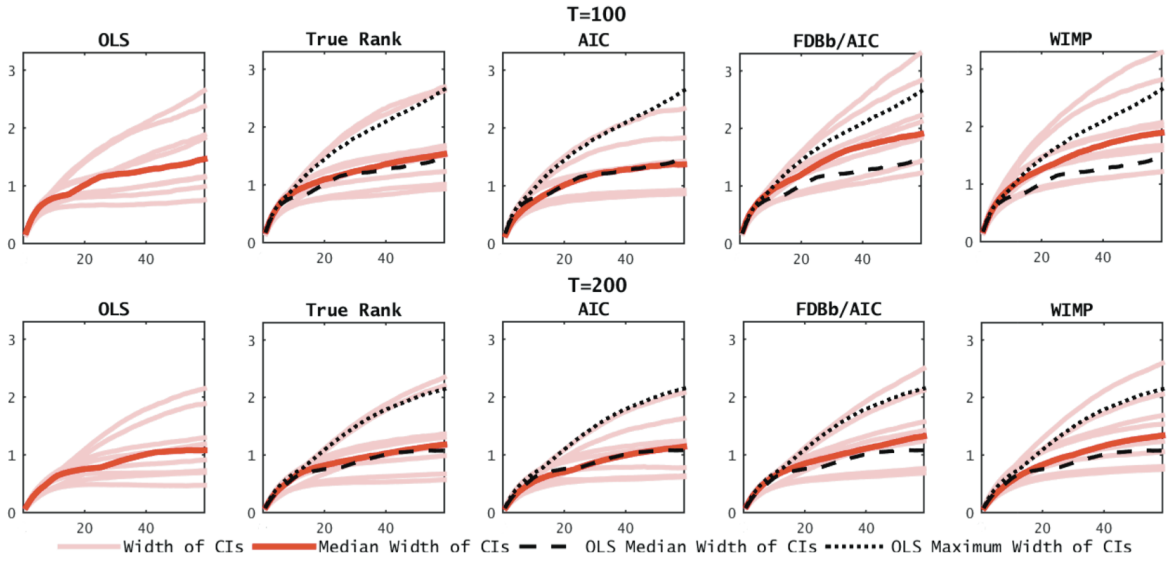

Figure 4: DGP1: Average width of 95% bootstrap CIs for various inference methods for $T = 100$ and $T = 200$. For details see Figure 2.

It stands to reason that if evidence for a specific cointegration relation is strong, rank pre-estimation could result in more reliable inference than unrestricted OLS and may outperform the WIMP intervals which – despite weighting down implausible ranks – are inherently more conservative. We investigate this further by turning to DGP2. Figure 5 displays CPs for the case of strong cointegration relations. Indeed, CPs implied by model selection based on AIC and BIC are much closer to the nominal level than those entailed by OLS. Bootstrap intervals based on unrestricted estimation can again not be considered as reliable, with minimum CPs around 60% for both sample sizes. Imposing the true rank delivers CPs close to but still

below the nominal level. As in the weak cointegration setting, the WIMP intervals again outperform all other approaches and even deliver CPs closer to nominal level than those implied by the correct rank specification. It is noticeable that the WIMP intervals do not produce overly conservative inference when evidence for a particular rank is strong, but result in CPs very close to the 95% level. This is also reflected in the average width (over 1000 MC simulations) of the CIs displayed in Figure (6). WIMP intervals are (if at all) only marginally wider than those implied by the correct rank specification, and are even much narrower than some of the intervals based on the unrestricted model. Finally, note that the WIMP intervals are now also much narrower than some of the FDB bagging intervals while having superior coverage. Concluding, the WIMP intervals allow for meaningful inference in practical sample sizes irrespective of the degree of rank uncertainty.

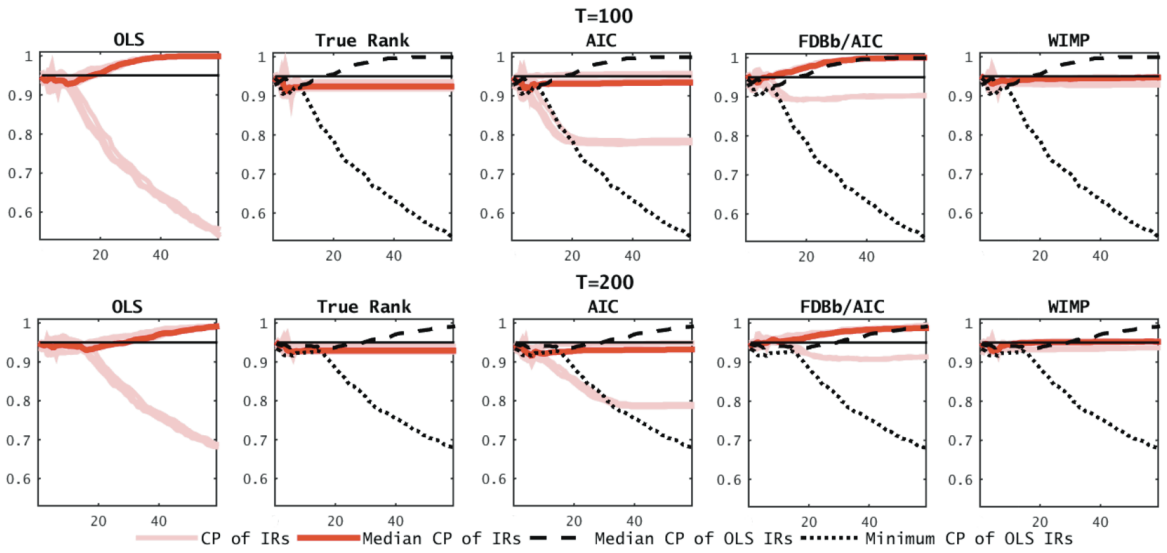

Figure 5: DGP2: Empirical coverage rates for various inference methods for $T = 100$ and $T = 200$. For details see Figure 2.

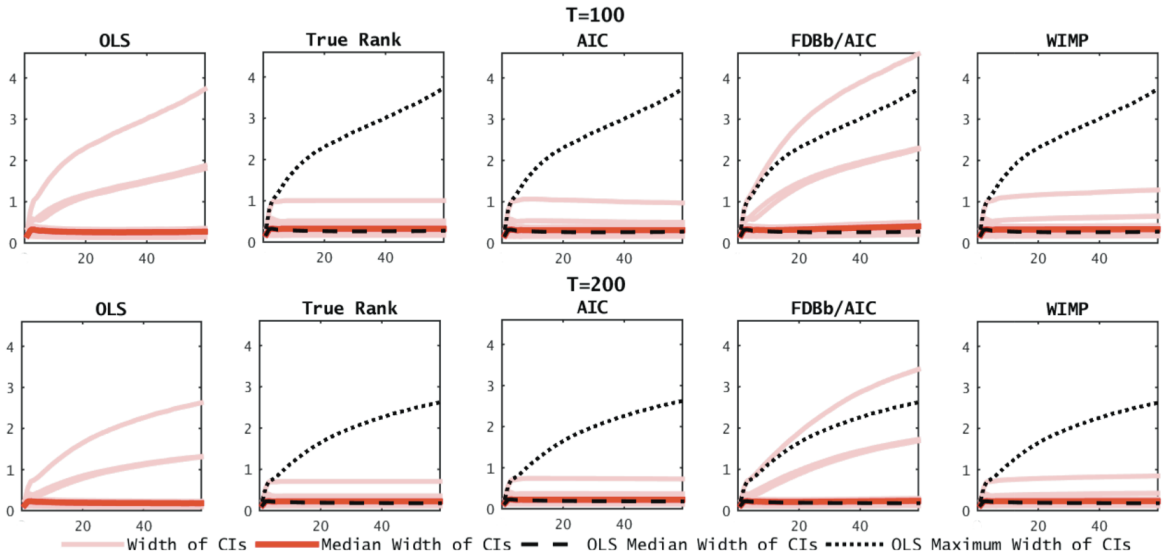

Figure 6: DGP2: Average width of 95% bootstrap CIs for various inference methods for $T = 100$ and $T = 200$. For details see Figure 2.

5 Fiscal Policy Shocks and Rank Uncertainty

In this section we study the potential ramifications of rank uncertainty on applied macroeconomic analysis. By using our proposed approaches to construct inference accounting for rank uncertainty, we aim at assessing the robustness of results usually obtained from unrestricted VARs. While there are countless VAR-based studies that use impulse response analysis to investigate the propagation of structural economic shocks, we focus in the following on fiscal policy shocks.

The novelty of this paper is methodological and we do not complement the literature on identification of structural VARs. This is why we dispense with a detailed literature review summarizing important contributions on VAR-based policy analysis and only focus on evaluating seminal papers, reflecting various ways of identification. We also skip a detailed discussion of different identification approaches and their respective merits.[12] Instead, our goal is to demonstrate that the problem – and our solution – is present regardless of the identification scheme. For this purpose it suffices for us to focus on several seminal papers that consider different identification schemes.

In this paper we do not want to engage in a discussion about the exact size of the fiscal multiplier. We rather want to emphasize the amplified uncertainty associated with its estimation under unknown cointegration relationships. For that reason we omit any discussion on point estimates and focus solely on inference, and highlight the role of (ignoring) rank uncertainty.

Our aim is also not to challenge (widely accepted) empirical findings on the effects of economic policies, but to provide the applied researcher with tools that might help to construct more reliable inference. For that reason, we refrain from a simple replication exercise comparing different inferential approaches, and we want to stress that our goal is certainly not to contrast our findings to the original papers. Instead, we use the same reduced-form VAR and (subsets of) the same dataset across all applications, in order to move away from the original papers and only contrast results based on different identification procedures. By "homogenizing" the underlying models and data used, we construct a coherent structure in which the effects of rank uncertainty can properly be investigated, and which is of interest in itself.

Fiscal policy can relate to both the expenditure and revenue side of the government's budget. Measuring the effect of active spending policies as well as the consequences of tax changes has been an active field of economic research since decades. One of the first influential contributions using VAR-based impulse responses to assess the effect of government purchases is Blanchard and Perotti (2002). The authors identify spending shocks by a recursive identification scheme. With government spending ordered first, this translates into the

[12] For a detailed exposition we refer to Ramey (2016) for a recent survey on various identification approaches and results in the literature.

assumption that government purchases are predetermined within the quarter.

Due to their assumed independence from general macroeconomic conditions, Ramey and Shapiro (1998) construct narrative records based on military buildups to identify truly exogenous spending changes. Those narrative time series have been embedded in several VAR studies and used to identify spending shocks by ordering this series first in a Cholesky-identified VAR. Among the most prominent studies following this approach is Ramey (2011). In her paper she revisits the construction of the government spending news variable, filtering out possible distortions due to anticipation effects.

Narrative series have also been used to identify tax changes. In a series of papers Mertens and Ravn (2011, 2012, 2013, 2014) construct various "dis-aggregates" of the Romer and Romer (2009) measures of legislated changes in federal tax liabilities. More specifically, Mertens and Ravn distinguish between announced and unannounced tax changes, or between personal and corporate taxes. Moreover, the authors do not view those narrative series as a direct measure of "tax-shocks" but rather as an external *proxy* which is correlated with the unknown structural shocks.[13] Thus, instead of including the narrative variable in the VAR, one can obtain the structural shock of interest by regressing the narrative *proxy* on the reduced-form residuals.

Yet another structural VAR identification approach imposes signs on the impulse responses to a particular shock for a certain horizon. Mountford and Uhlig (2009) identify a contractionary tax-shock as a shock, which leads to non-negative responses in government revenue during the first year after impact. Additionally, this tax-shock is identified by requiring it to be orthogonal to a business cycle shock and a monetary policy shock – both identified through signs.[14] In particular, the orthogonality to business cycle fluctuations aims at controlling for movements in the government's budget caused by automatic stabilizers.

We compare uncertainty associated with the estimated impulse responses resulting from the above mentioned four identification approaches using the same data, and the same specification (as far as possible) of the underlying (reduced-form) VAR. That is, we use Blanchard and Perotti's (2002) structural VAR approach as well as Ramey's (2011) strategy to incorporate her narrative series in a VAR to identify the effect of government spending. Further, we use Mountford and Uhlig's (2009) sign-restriction scheme and Mertens and Ravn's (2014) proxy-VAR to assess the effect of tax-shocks.

The choice of variables and the sample period is largely determined by the "highest minimal requirement" across the above identification approaches. The benchmark VAR is estimated in logs of GDP, logs of private consumption, logs of non-residential investment, logs of

[13] See also Stock and Watson (2012) and Montiel-Olea et al. (2016).

[14] All three shocks are identified sequentially by maximizing a penalty function which rewards responses in the desired direction and penalizes the others. Business cycle shocks are identified by assuming that they lead to co-movements in the same direction of output, consumption, investment, and government revenue. A contractionary monetary policy shocks affect responses in reserves and prices negatively and the interest rate positively.

total government spending, logs of (federal) tax receipts, logs of total non-borrowed reserves, real wages, a price index and the GDP deflator.[15] We use Ramey's (2011) news variable and Mertens and Ravn's (2011; 2012; 2014) unanticipated tax-change proxy. The data is quarterly, sampling from 1950/Q1-2006/Q4. The VAR representation in levels includes an intercept and a deterministic linear time trend. Four lags are included.

We construct inference using the residual-based bootstrap algorithm presented in Algorithm 1, incorporated in the methods discussed in Section 3, detrending on both an intercept and linear trend.[16] While Ramey's (2011) news series is included in the VAR, and thus, bootstrapped "endogenously", we jointly draw (with replacement) from the reduced-form residuals and Mertens and Ravn's (2012; 2014) external variable to account for uncertainty in estimating the effects of tax-shocks using this proxy.

In order to make results somewhat comparable, impulse responses are normalized such that the point estimate of the response of the policy instruments has a peak at unity across different identification approaches (see for example Ramey, 2011). As a measure of uncertainty we plot 68% confidence intervals, which is standard in the fiscal policy literature.[17]

Figure 7 and Figure 8 display unrestricted VAR in levels (estimated by OLS), FDB bagging (with AIC selection), and WIMP confidence bands (using the same specifications as in Section 4) of impulse responses due to a government spending shock. For the recursive VAR as in Blanchard and Perotti (2002), all three measures of uncertainty suggest that government spending shocks generate an initial boost in GDP. While the FDBb intervals indicate a rather moderate increase relative to the OLS intervals, the WIMP intervals imply maximum multiplier effects greater in range (roughly between 0.7 and 1.5). Considering impulse responses following Ramey's news shocks, it seems to be less clear whether government spending stimulates output or not. While the OLS confidence bands (and to a lesser extend the FDBb bands) support findings in the literature suggesting a short-lived boost in GDP, the WIMP intervals indicate greater uncertainty associated with the output response. Indeed, "robust" spending peak multipliers range between 0 and 3.3, such that a reliable conclusion on the effectiveness of spending policies cannot be made in this case.

Confidence intervals of impulse responses following a contractionary tax-shock are displayed in Figures 9 and 10. Qualitatively, responses of GDP and its main aggregates are rather similar across both identification approaches and across all three inferential procedures: Output, consumption, and investment decrease significantly. The long-lived contraction in economic activity is, however, accompanied by an equally lengthy decline in government spending, which hinders the interpretation of the identified shocks as "pure" tax-shocks.

[15] A detailed description of the data is given in the appendix.

[16] We did not find strong evidence of heteroskedasticity in the reduced-form residuals and refrain from using a robust bootstrap procedure such as the moving block bootstrap (Brüggemann et al., 2016). All approaches outlined in this paper could be easily extended in this way.

[17] The data set as well as a MATLAB toolbox for the WIMP method with the identification schemes used used in this section are available at http://researchers-sbe.unimaas.nl/stephansmeekes.

Quantitatively however, the implied response of output is much greater in the proxy VAR framework compared to the SVAR one. Intervals for peak multipliers include -6 for the former, and -3 for the latter.

Similar to the responses due to a government spending shock, the FDBb intervals are not necessarily wider than the OLS intervals. However, when considering the impact on output, and in contrast to scenario investigated above, the two intervals do not intersect at times and the FDBb intervals imply a significantly smaller impact on economic activity. This holds for both the shocks of Mountford and Uhlig (2009) and Mertens and Ravn (2012, 2014). Reflecting potentially more conservative inference, the WIMP intervals are wider, often encompassing the OLS intervals. Yet the WIMP intervals indicate that OLS-based inference rather underrates the effect of the identified tax-shocks on almost all variables. Generally, tax-shocks estimated by the proxy VAR imply greater effects on economic activity than those identified through sign-restrictions. Moreover, the comparison with the spending shocks, supports some results in the literature suggesting that tax-cuts may be more effective in stimulating the economy. Indeed, comparing peak multipliers displayed in Figure 11 reveals that evidence suggesting that multipliers exceed unity is much stronger for tax-cut policies than for spending policies. Based on the results for Ramey's news shock, multipliers due to expansionary spending policies might even not be significant at all.

In general, the above results illustrate that ignoring uncertainty about the co-integration relations in the data, may lead to ambiguous interpretation of statistical significance. Incorporating this uncertainty via our proposed WIMP approach allows for a more confident interpretation of the results.

6 Discussion

In this paper we have shown empirically and through a simulation study that ignoring uncertainty about cointegration relations may lead to unreliable inference for (structural) impulse responses. Since the commonly used specification of the VAR in levels ignores any evidence for cointegration in the data, associated inference captures uncertainty only poorly. Also, model selection techniques, such as rank pre-estimation by sequential testing or information criteria, seem to deliver reliable inference only if evidence for the true cointegration rank is strong. In this paper we propose a novel data-driven approach to robust inference for impulse responses in the presence of uncertainty regarding the cointegration rank. Our WIMP approach is shown both by simulation and empirically to still be able to deliver meaningful (i.e. not too wide) confidence intervals while being robust to rank uncertainty. As such it provides a reliable and simple alternative to the unreliable standard approaches.

Practical implementation of the WIMP approach only requires fixed-rank (bootstrap) intervals plus the sequence of trace tests for all rank tests, which are both readily available

in any standard statistical software. While a toolbox for the WIMP methods used in our application is directly available, our approach can also easily be implemented for any desired SVAR analysis, as the fixed-rank intervals used as input for the WIMP can be based on any appropriate method, both in terms of inference method such as the bootstrap and identification scheme. Finally, the computational cost of the method is fairly low; on any modern computer bootstrap intervals for a fixed rank are fast to compute, and given that in this kind of VAR model the number of variables (and hence the number of ranks) has to be relatively low to avoid the curse of dimensionality, doing so for all ranks should pose no problem.

While the prudent construction of inference is particularly important for impulse responses, our proposed WIMP procedure may equally well be beneficial when used in a different VAR context, such as forecasting. While forecast combinations across different models are well accepted as point forecasts, our WIMP method allows to construct corresponding interval forecasts that account for model uncertainty. More generally, the approach can be adapted to a variety of model selection problems, as long as we can assess the relative evidence for a particular model against a modest number of alternatives. While in theory it can be applied to high-dimensional problems as well, computationally the method is particularly suited for low-dimensional problems where the number of models is relatively small. While this is a limitation of the method, it is inherent to the simultaneous inference philosophy behind, which also holds for the PoSI method of Berk et al. (2013). Exploring the usefulness and limitations of the WIMP in more general settings is therefore an interesting avenue for future research.

A Appendix: Data

All data is quarterly, sampling from 1950/Q1-2006/Q4. We composed the data from three sources: The Bureau of Economic Analysis' *U.S. National Income and Product Accounts* (NIPA) (bea.gov/national), The Bureau of Labor Statistics (BLS) (bls.gov), and *FRED Economic Database* hosted by the Federal Reserve Bank of St. Louis (fred.stlouisfed.org).

GDP is taken from NIPA table 1.1.5.

Investment is *gross private non-residential investment*, NIPA table 1.1.5.

Government spending is *government expenditure and gross investment*, NIPA table 3.9.5.

Government revenue is *Federal government current tax receipts* plus *contributions for social insurance* minus *income taxes from federal reserve banks*, all in NIPA table 3.2.

Real wages are *nonfarm business sector: real compensation per hour*, from the BLS.

GDP deflator is taken from NIPA table 1.1.9

Federal funds rate is taken from FRED, series code: (*fedfunds*)

Adjusted reserves is taken from FRED, series code: (*ADJRESSL*)

GDP and its components, government revenue, and adjusted reserves are transformed into real per capita values using the GDP deflator and a population measure (NIPA table 7.1).

References

Benkwitz, A., H. Lütkepohl, and J. Wolters (2001). Comparison of bootstrap confidence intervals for impulse responses of German monetary systems. *Macroeconomic Dynamics 5*, 81–100.

Berk, R., L. Brown, A. Buja, K. Zhang, and L. Zhao (2013). Valid post-selection inference. *Annals of Statistics 41*, 802–837.

Bernstein, D. and B. Nielsen (2014). Asymptotic theory for cointegration analysis when the cointegration rank is deficient. Economic Working Papers 2014-W06, Nuffield College, University of Oxford.

Blanchard, O. and R. Perotti (2002). An empirical characterization of the dynamic effects of changes in government spending and taxes on output. *The Quarterly Journal of Economics 117*(4), 1329–1368.

Breiman, L. (1996). Bagging predictors. *Machine Learning 24*, 123–140.

Bruder, S. and M. Wolf (2017). Balanced bootstrap joint confidence bands for structural impulse response functions. Technical Report No. 246, University of Zurich.

Brüggemann, R., C. Jentsch, and C. Trenkler (2016). Inference in VARs with conditional heteroskedasticity of unknown form. *Journal of Econometrics 191*, 69–85.

Bühlmann, P. and B. Yu (2002). Analyzing bagging. *Annals of Statistics 30*, 927–961.

Cavaliere, G., A. Rahbek, and A. M. R. Taylor (2010a). Cointegration rank testing under conditional heteroskedasticity. *Econometric Theory 26*, 1719–1760.

Cavaliere, G., A. Rahbek, and A. M. R. Taylor (2010b). Testing for co-integration in vector autoregressions with non-stationary volatility. *Journal of Econometrics 158*, 7–24.

Cavaliere, G., A. Rahbek, and A. M. R. Taylor (2012). Bootstrap determination of the co-integration rank in vector autoregressive models. *Econometrica 80*, 1721–1740.

Chao, J. C. and P. C. B. Phillips (1999). Model selection in partially nonstationary vector autoregressive processes with reduced rank structure. *Journal of Econometrics 91*, 227—271.

Cheng, X. and P. C. B. Phillips (2009). Semiparametric cointegrating rank selection. *Econometrics Journal 12*, S83–S104.

Cheng, X. and P. C. B. Phillips (2012). Cointegrating rank selection in models with time-varying variance. *Journal of Econometrics 142*, 201–211.

Choi, I. (2005). Inconsistency of bootstrap for nonstationary, vector autoregressive processes. *Statistics & Probability Letters 75*, 39–48.

Davidson, R. and J. G. MacKinnon (2002). Fast double bootstrap tests of nonnested linear regression models. *Econometric Reviews 21*, 419–429.

Del Negro, M. and F. Schorfheide (2011). Bayesian macroeconometrics. In J. Geweke, G. Koop, and H. van Dijk (Eds.), *The Oxford Handbook of Bayesian Econometrics*, pp. 293–389. Oxford University Press.

Del Negro, M., F. Schorfheide, F. Smets, and R. Wouters (2007). On the fit of New Keynesian models. *Journal of Business & Economic Statistics 25*, 123–143.

Efron, B. (2014). Estimation and accuracy after model selection. *Journal of the American Statistical Association 109*, 991–1007.

Giannone, D., M. Lenza, and G. E. Primiceri (2016). Priors for the long run. CEPR Discussion Paper 11261, Centre for Economic Policy Research.

Gospodinov, N. (2004). Asymptotic confidence intervals for impulse responses of near-integrated processes. *Econometrics Journal 7*, 505–527.

Gospodinov, N. (2010). Inference in nearly nonstationary SVAR models with long-run identifying restrictions. *Journal of Business & Economic Statistics 28*, 1–12.

Gospodinov, N., A. M. Herrera, and E. Pesavento (2013). Unit roots, cointegration, and pretesting in VAR models. In T. B. Fomby, L. Kilian, and A. Murphy (Eds.), *VAR Models in Macroeconomics - New Developments and Applications: Essays in Honor of Christopher A. Sims*, Volume 32 of *Advances in Econometrics*, pp. 81–115. Emerald Group Publishing Limited.

Gospodinov, N., A. Maynard, and E. Pesavento (2011). Sensitivity of impulse responses to small low-frequency comovements: reconciling the evidence on the effects of technology shocks. *Journal of Business & Economic Statistics 29*, 455–467.

Hall, P. (1992). *The bootstrap and Edgeworth expansions*. New York: Springer-Verlag.

Hjort, N. L. and G. Claeskens (2003). Frequentist model average estimators. *Journal of the American Statistical Association 98*, 879–899.

Inoue, A. and L. Kilian (2002). Bootstrapping autoregressive processes with possible unit roots. *Econometrica 70*, 377–391.

Inoue, A. and L. Kilian (2016). Joint confidence sets for structural impulse responses. *Journal of Econometrics 192*, 421–432.

Jardet, C., A. Monfort, and F. Pegoraro (2013). No-arbitrage near-cointegrated VAR(p) term structure models, term premia and {GDP} growth. *Journal of Banking and Finance 37*, 389–402.

Johansen, S. (1995). *Likelihood-Based Inference in Cointegrated Vector Autoregressive Models*. Oxford University Press.

Kilian, L. (1998a). Accounting for lag order uncertainty in autoregressions: the endogenous lag order bootstrap algorithm. *Journal of Time Series Analysis 19*, 531–548.

Kilian, L. (1998b). Small-sample confidence intervals for impulse response functions. *Review of Economics and Statistics 80*, 218–230.

Kilian, L. and P.-L. Chang (2000). How accurate are confidence intervals for impulse responses in large VAR models? *Economics Letters 69*, 299–307.

Leeb, H. and B. M. Pötscher (2005). Model selection and inference: Facts and fiction. *Econometric Theory 21*, 21–59.

Liao, Z. and P. C. B. Phillips (2015). Automated estimation of vector error correction models. *Econometric Theory 31*, 581–646.

Lütkepohl, H. (1990). Asymptotic distributions of impulse response functions and forecast error variance decompositions of vector autoregressive models. *Review of Economics and Statistics 72*, 116–125.

Lütkepohl, H., A. Staszewska-Bystrova, and P. Winker (2015). Comparison of methods for constructing joint confidence bands for impulse response functions. *International Journal of Forecasting 31*, 782–798.

Mertens, K. and M. O. Ravn (2011). Understanding the aggregate effects of anticipated and unanticipated tax policy shocks. *Review of Economic Dynamics 14*, 27–54.

Mertens, K. and M. O. Ravn (2012). Empirical evidence on the aggregate effects of anticipated and unanticipated US tax policy shocks. *American Economic Journal: Economic Policy 4*, 145–181.

Mertens, K. and M. O. Ravn (2013). The dynamic effects of personal and corporate income tax changes in the United States. *American Economic Review 103*, 1212–1247.

Mertens, K. and M. O. Ravn (2014). A reconciliation of SVAR and narrative estimates of tax multipliers. *Journal of Monetary Economics 68*, 1–19.

Mikusheva, A. (2007). Uniform inference in autoregressive models. *Econometrica 75*, 1411–1452.

Mikusheva, A. (2012). One-dimensional inference in autoregressive models with the potential presence of a unit root. *Econometrica 80*, 173–212.

Montiel-Olea, J. L., J. H. Stock, and M. W. Watson (2016). Uniform inference in SVARs identified with external instruments. Mimeo.

Mountford, A. and H. Uhlig (2009). What are the effects of fiscal policy shocks? *Journal of Applied Econometrics 24*, 960–992.

Pesavento, E. and B. Rossi (2006). Small-sample confidence intervals for multivariate impulse response functions at long horizons. *Journal of Applied Econometrics 21*, 1135–1155.

Pesavento, E. and B. Rossi (2007). Impulse response confidence intervals for persistent data: What have we learned? *Journal of Economic Dynamics & Control 31*, 2398–2412.

Phillips, P. C. B. (1996). Econometric model determination. *Econometrica 64*, 763–812.

Phillips, P. C. B. (1998). Impulse response and forecast error variance asymptotics in non-stationary VARs. *Journal of Econometrics 83*, 21–56.

Ramey, V. A. (2011). Identifying government spending shocks: It's all in the timing. *Quarterly Journal of Economics 126*(1), 1–50.

Ramey, V. A. (2016). Macroeconomic shocks and their propagation. NBER Working Papers 21978, National Bureau of Economic Research.

Ramey, V. A. and M. D. Shapiro (1998). Costly capital reallocation and the effects of government spending. *Carnegie-Rochester Conference Series on Public Policy 48*(1), 145–194.

Romer, C. D. and D. H. Romer (2009). A narrative analysis of postwar tax changes. Mimeo, University of California, Berkeley.

Sobreira, N. and L. C. Nunes (2012). Testing for broken trends in multivariate time series. Mimeo, Nova School of Business and Economics.

Stock, J. H. and M. W. Watson (2012). Disentangling the channels of the 2007-2009 recession. NBER Working Papers 18094, National Bureau of Economic Research.

Strachan, R. W. and H. K. van Dijk (2007). Bayesian model averaging in vector autoregressive processes with an investigation of stability of the US great ratios and risk of a liquidity trap in the USA, UK and Japan. Econometric Institute Research Papers EI 2007-11, Erasmus University Rotterdam.

Swensen, A. R. (2006). Bootstrap algorithms for testing and determining the cointegration rank in VAR models. *Econometrica 74*, 1699–1714.

Villani, M. (2001). Bayesian prediction with cointegrated vector autoregressions,. *International Journal of Forecasting 17*, 585–605.

Wright, J. H. (2000). Confidence intervals for univariate impulse responses with a near unit root. *Journal of Business & Economic Statistics 18*, 368–373.

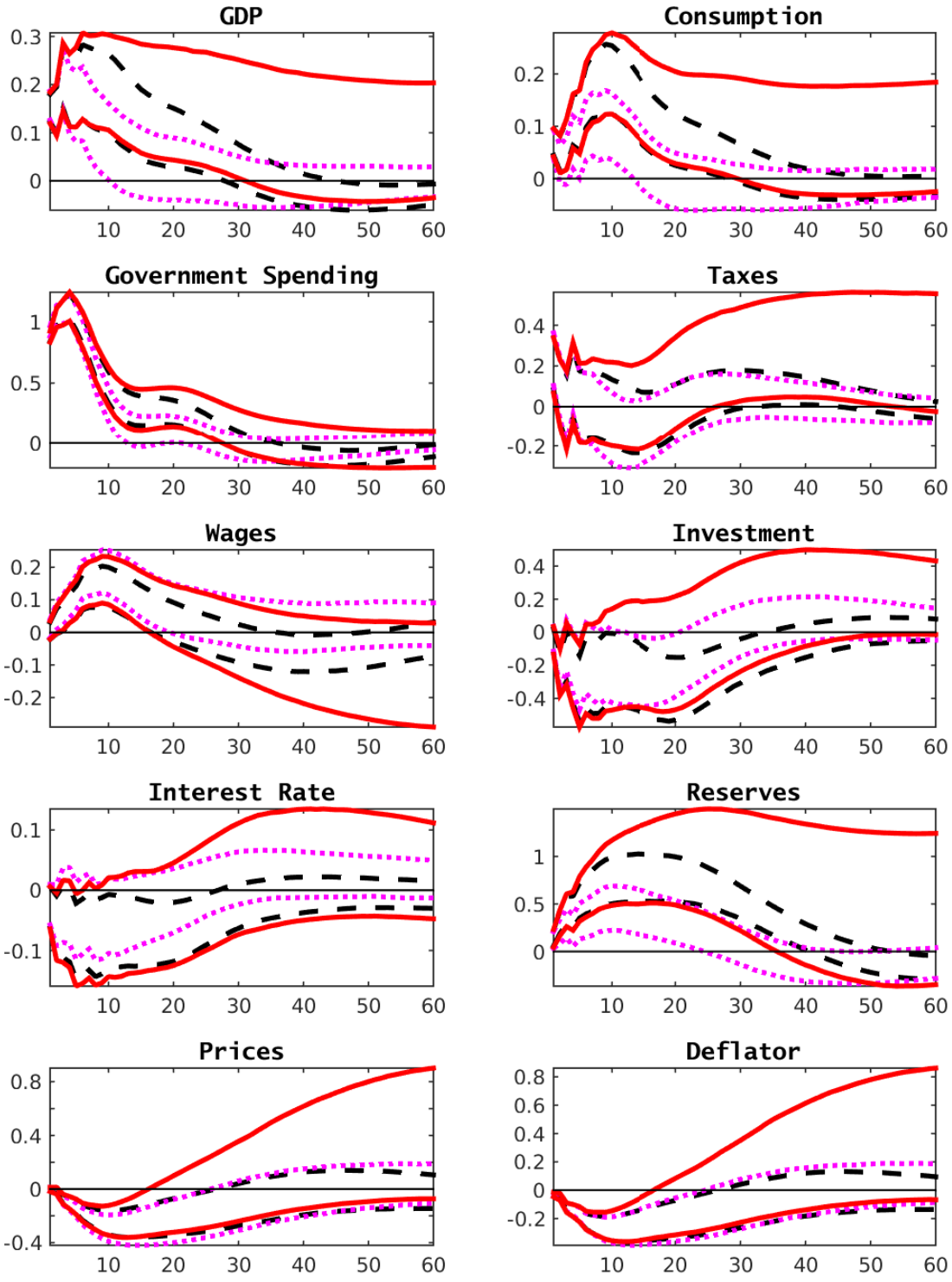

Figure 7: 68% confidence intervals of impulse responses to a government spending shock identified as in Blanchard and Perotti (2002). **Black** dashed lines are OLS intervals, pink dotted lines are FDBb/AIC intervals, **red** solid lines are WIMP intervals.

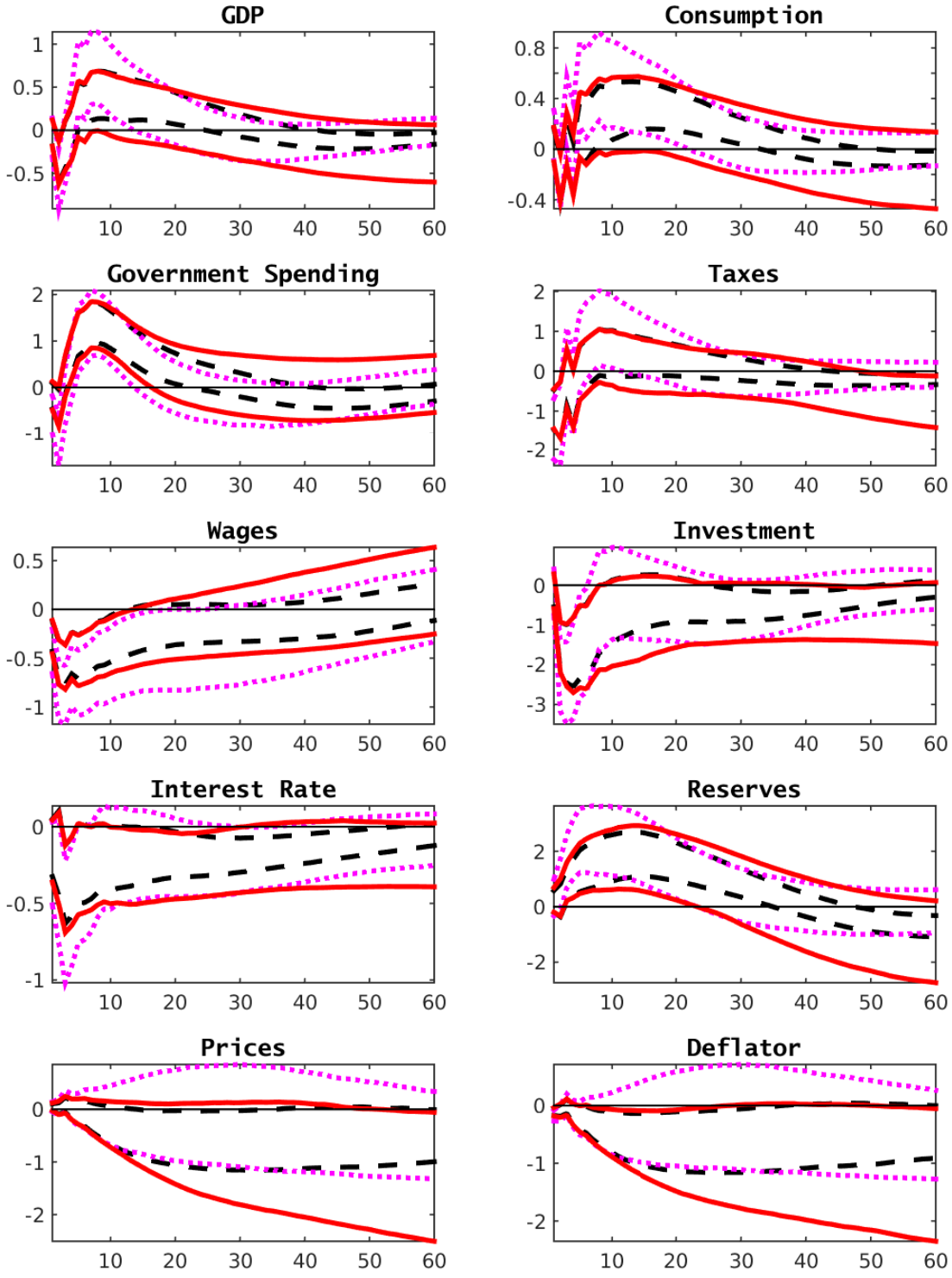

Figure 8: 68% confidence intervals of impulse responses to a government spending shock identified as in Ramey (2011). For details see Figure 7.

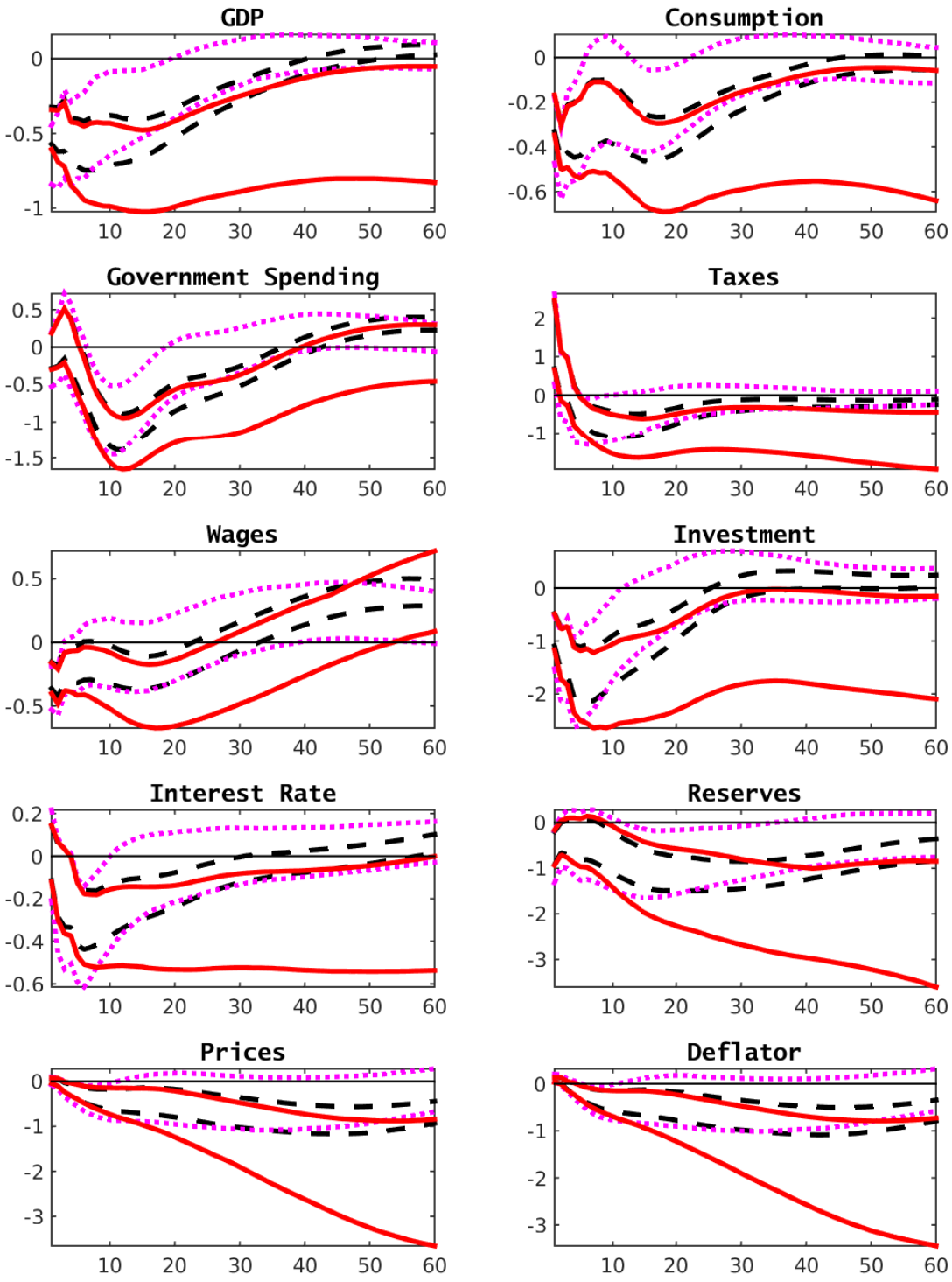

Figure 9: 68% confidence intervals of impulse responses to a tax-shock identified as in Mountford and Uhlig (2009). For details see Figure 7.

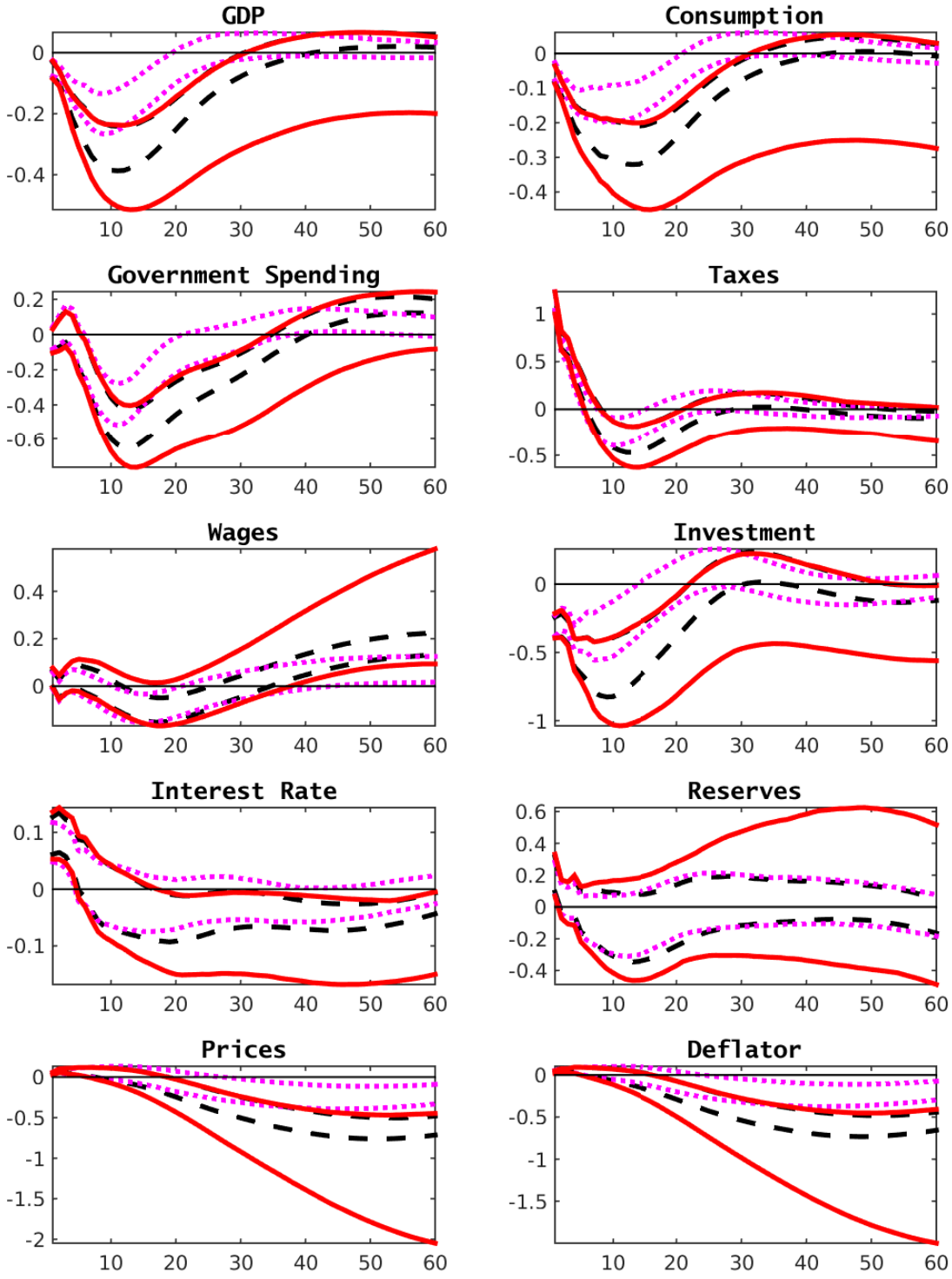

Figure 10: 68% confidence intervals of impulse responses to a tax-shock identified as in Mertens and Ravn (2012, 2014). For details see Figure 7.

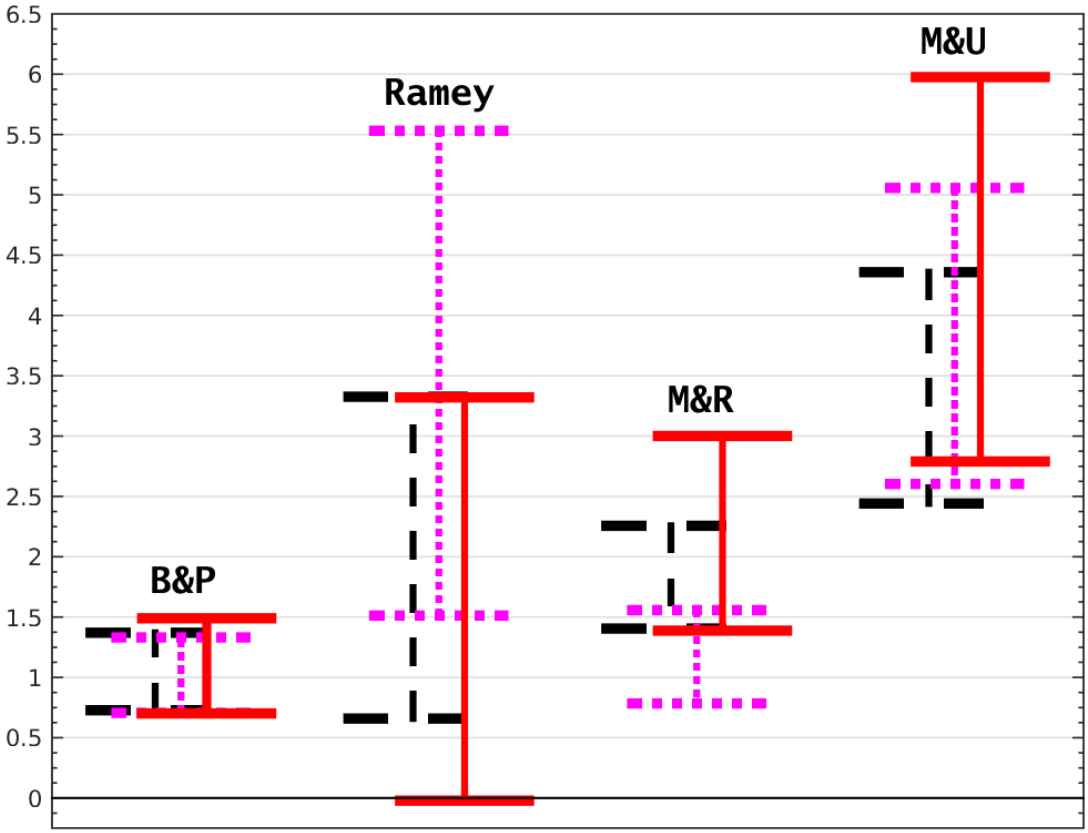

Figure 11: 68% confidence intervals of peak multipliers implied by government spending and tax-cut shocks in the analyses based on Blanchard and Perotti (2002) [**B&P**], Ramey (2011) [**Ramey**], Mountford and Uhlig (2009) [**M&U**] and Mertens and Ravn (2012, 2014) [**M&R**]. **Black** dashed lines are OLS intervals, **pink** dotted lines are FDBb/AIC intervals, **red** solid lines are WIMP intervals.