Identification-robust inference with simulation-based pseudo-matching *

Bertille Antoine † Lynda Khalaf ‡ Maral Kichian §

Zhenjiang Lin ¶

Abstract

We develop a general simulation-based inference procedure for partially specified models. Our procedure is based on matching auxiliary statistics to simulated counterparts where nuisance parameters are calibrated neither assuming identification of parameters of interest nor a one-to-one binding function. The conditions underlying asymptotic validity of our (pseudo-)simulators in conjunction with appropriate bootstraps are characterized beyond the strict and exact calibration of the parameters of the simulator. Our procedure is illustrated through impulse-response (IR) matching. In addition to usual Wald-type statistics that combine structural or reduced form IRs, we analyze local projections IRs through a factor-analytic measure of distance adapted from Bai (2013) which eschews the need to define a weighting matrix.

*Antoine, Khalaf and Kichian acknowledge funding from SSHRC-Insight grant.

†B. Antoine (Corresponding author): Simon Fraser University, 8888 University Drive, Burnaby, BC V5A 1S6, CANADA. *Email:* Bertille_Antoine@sfu.ca

‡L. Khalaf: Carleton University, Canada. *Email:* Lynda.Khalaf@carleton.ca

§M. Kichian: University of Ottawa, Canada. *Email:* Maral.Kichian@uOttawa.ca

¶Z. Lin: University of Nottingham Ningbo China. *Email:* Zhenjiang.Lin@nottingham.edu.cn

Keywords: Approximate calibration, Weak identification, Bootstrap, Impulse response matching.

JEL Classification: C32, C52, C53, E30, E50.

1 Introduction

Simulation-based matching [SBM] methods such as the simulated method of moments or indirect inference [Gourieroux, Monfort and Renault (1993), Smith (1993), and Gallant and Tauchen (1996)] have become increasingly sought-after, as they reflect the structure of popular economic models. Indeed, despite increasingly complex distributional foundations and resulting intractable likelihoods, many models in economics are still formulated in such a way that generating conformable data - given a particular value of underlying parameters - remains relatively easy.

A recognition of this feature has spurred a large applied and theoretical literature on SBM. See *e.g.* Guay and Scaillet (2003), on threshold models; Genton and Ronchetti (2003) and Dridi, Guay and Renault (2007), on integrating deviations from the assumed model; Czellar and Ronchetti (2010) and Dovonon and Hall (2018), on asymptotic refinements; Calvet and Czellar (2015) and the recent survey by Meenagh, Minford, Wickens and Xu (2019), on applied equilibrium settings; Gourieroux, Phillips and Yu (2010) and Khalaf and Saunders (2019), on dynamic panels; Forneron and Ng (2018) and Kaji, Manresa, and Pouliot (2020), on reconciling machine learning with SBM. In the present paper, we develop a general procedure for inference using SBM that does not assume identification of parameters of interest, and which applies when nuisance parameters cannot be easily partialled out.

Indirect inference relies on a binding function that links a model of interest to an auxiliary one for which a simple estimator is available and that is used for matching

purposes. More broadly, SBM relies on a distance (according to some metric) between a (chosen) vector of auxiliary statistics and its simulated counterpart, drawn from a fully specified structure - the so-called *simulator* - that characterizes the considered model and is indexed by a finite dimensional parameter θ. Despite widespread interest and on-going advances, SBM methods are still generally validated in identified contexts and considered from a minimum-distance perspective, particularly when the simulator may be misspecified.

Yet, the complex models that motivate the use of SBM in the first place are, for the most part, hard to identify. Further, in practice, the *simulator* may or may not hold *exactly*, or alternatively, θ may only be partially recoverable from the available sample. Said differently, $\theta = (\theta_1, \theta_2)$ where θ_1 is the parameter of interest on which the considered data is likely to be informative, while the nuisance parameter θ_2 is hard or impossible to recover from this same dataset. Dridi, Guay and Renault (2007) introduce the concept of partial encompassing to validate the calibration of θ_2. In contrast to its standard minimum distance foundation, we generalize the partial encompassing principle to accommodate possible weak identification of θ_1 and possible *mis-calibration* of θ_2.

To be clear, we propose and validate tests of $\mathcal{H}_0 : \theta_1 = \theta_{1,0}$ based on matching observed to simulated criteria, allowing for mis-calibration of θ_2 in a likelihood-free environment. The objective is inference on θ_1 by inverting the proposed distance-based tests, while allowing for θ_2 to be calibrated exactly (as is standard), but also *approximately*, which admits the possibility that the calibration may not be correct. We will refer to our generic procedure as *pseudo-matching*, in which case we formally define a valid approximate calibration as follows. We introduce *adaptive neighborhoods* around the calibrated parameters that are compatible with our likelihood-free and test inversion approach in the sense that they preserve the level of the proposed

test asymptotically. These neighborhoods are specifically defined via the considered auxiliary statistics, and maintain an equicontinuity condition on the distribution of the criterion function under $\mathcal{H}_0$, thus adapting to the inverted criterion as well as to the tested parameter value $\theta_{1,0}$.

To do this, we generalize available theory on the parametric bootstrap [Dufour (2006) and Bergamelli, Bianchi, Khalaf and Urga (2019)] to test statistics that depend on nuisance parameters (here θ_2) that cannot be partialled out nor evacuated through finite-sample or asymptotic invariance arguments. The procedures we propose involve multilevel (double or triple) bootstraps to reflect conditioning on the hypothesized value of the auxiliary statistics which is derived by simulation. For references on the multilevel parametric bootstrap with invariant measures, where invariance yields exchangeability of observed and bootstrap statistics under $\mathcal{H}_0$, see Khalaf and Saunders (2020), Khalaf and Peraza-Lopez (2020) or Beaulieu, Dufour and Khalaf (2007, 2010, 2014). Since invariance to θ_2 is excluded for the problems we consider here, our results provide generic conditions that extend such procedures beyond invariant simulators.

Taken collectively, the conditions underlying the asymptotic validity of our inference procedure - and associated simulators in conjunction with multilevel bootstraps - extend SBM and the partial indirect inference approach in several important directions: (i) our criterion function taken as a general (univariate) transformation of the vector of auxiliary statistics - such as (but not only) a square norm - may not be one-to-one; (ii) our inference procedure is identification and mis-calibration robust; (iii) the simulator behaves like a pseudo-true model without imposing a unique pseudo-true value.

On the practical side, we consider impulse-response (IR) matching for DSGE models. We first consider Wald-type statistics that combine structural or reduced form IRs as done in Christiano, Eichenbaum and Evans (2005), Hall, Inoue, Nason and Rossi (2012), Inoue and Killian (2013), Guerron-Quintana, Inoue and Killian (2017).

In addition, we analyze IRs obtained by considering multiple horizons autoregressions through a factor-analytic measure of distance adapted from Bai (2013) which has the advantage of circumventing the need for a weighting matrix; see Dufour and Renault (1998) and Dufour, Pelletier and Renault (2006). Such regressions have also been called local projections: see Jordà (2005) and Jordà and Kozicki (2011), as well as the recent work by Plagborg-Møller and Wolf (2020).

In a laboratory environment, we consider the stylized DSGE model used in Fernandez-Villaverde, Rubio-Ramirez and Schorfheide (2016) allowing for exact and approximate calibration; such cases of mis-calibration are particularly relevant for empirical macroeconomic policy analysis. Despite inevitable identification and sensitivity concerns, our results reinforce arguments in the profession favouring calibration. Concretely, our simulations document: (i) potentially severe mis-calibration costs which underscore the usefulness of our proposed multi-scenario solution that allows for several calibrated values to be considered, and (ii) useful testable directions that remain immune to approximate calibration. In addition, our findings further illustrate: the importance of the direction of mis-calibration (and not only its magnitude); the importance of imposing the null hypothesis on weighing matrices of robust statistics; and the informational content of MA dynamics, inherent to local projections.

The rest of the paper is organized as follows. Section 2 introduces our general framework and motivates our inference strategy through the structural (multivariate) linear regression model. In section 3, we present our new simulation-based inference procedure and characterize its asymptotic validity. We highlight the special case of Wald-type simulation-based inference, and formalize a sensitivity analysis when "set calibration" of the nuisance parameters is considered to allow for several candidate calibrated values. In section 4, the relevance of our inference procedure is illustrated in a simulation study of a stylized DSGE model. Section 5 concludes. Graphs and tables

of results as well as the proofs of our theoretical results are gathered in the Appendix.

2 Framework and Motivation

We start by introducing our general framework and motivating our inference strategy through the structural (multivariate) linear regression model.

2.1 General framework

Let θ_1 be a p_1-vector of structural parameters of interest. We propose a test of $\mathcal{H}_0 : \theta_1 = \theta_{1,0}$ (where $\theta_{1,0}$ is a known vector of size p_1) with the objective of inverting this test for inference on θ_1 (or some known function of θ_1). Our test relies on a (chosen) q-vector of auxiliary statistics denoted $\hat{g}(Y_T)$ where $Y_T = (y_1, \cdots, y_T)$ is the sample of observables of size T; and, more specifically, on matching the above vector of auxiliary statistics to its counterpart obtained by simulation. In order to simulate auxiliary statistics, a knowledge of p_2 additional (nuisance) parameters θ_2 is often needed. We have in mind frameworks where θ_2 is calibrated; given that it cannot easily be estimated or partialled out, it is then calibrated. We specifically allow such calibration to be either exact (as done in standard practice) or to be approximate, in the sense that it may not be necessarily correct. Such situations are not uncommon in practice, especially as models' complexity increases: for example, it is the case for DSGE models such as the one considered in our simulation study in section 4.

Auxiliary statistics are often chosen to summarize key features of the data or key model implications. For example, in our simulation study in section 4 we focus on auxiliary statistics that are impulse-responses of the underlying DSGE model that we consider. Knowing θ_1 and θ_2, our *simulator* allows us to simulate the auxiliary statistics: the validity of our matching inference procedure relies on the *compatibility* between the

simulated auxiliary statistics and $\hat{g}(Y_T)$, which is formally characterized in the next section. It is however important to mention that our simulator is not assumed to coincide with the true model: we rather have in mind a parametric model that is not *too far* from our main model. For example, our simulator can be obtained as

- a parametric approximation - such as Quasi-maximum Likelihood - of the semi-parametric model of interest;
- an approximation - such as a linearization - of the main (nonlinear) model;
- a reduced form model of the main structural model of interest.

Hereafter, we denote the full vector of p structural parameters $\theta = (\theta_1', \theta_2')'$ (with $p = p_1 + p_2$).

Definition 1 *(Simulator)*
For a fixed vector $\overline{\theta}$ of size p, $\tilde{Y}_{T,h}(\overline{\theta}) \equiv \tilde{Y}_{T,h}(\epsilon_h, \overline{\theta})$ denotes a sample of size T of data simulated under $\overline{\theta}$ with errors ϵ_h drawn from an assumed distribution F_ϵ, and $\tilde{g}_{T,h}(\overline{\theta}) \equiv \tilde{g}_{T,h}(\tilde{Y}_{T,h}(\epsilon_h, \overline{\theta}))$ is the vector of auxiliary statistics computed using $\tilde{Y}_{T,h}(\overline{\theta})$.

Averaging over H simulated q-vectors of auxiliary statistics denoted $\tilde{g}_{T,h}(\overline{\theta})$ (with $h = 1, \cdots, H$) yields

$$\overline{\tilde{g}}_{T,H}(\overline{\theta}) \;=\; \frac{1}{H}\sum_{h=1}^{H} \tilde{g}_{T,h}(\overline{\theta})\,. \tag{1}$$

Our inference strategy is based on matching the data-based statistic $\hat{g}(Y_T)$ with its simulated counterpart $\overline{\tilde{g}}_{T,H}(\theta)$ through a criterion function $Q(.)$. Our general regularity conditions formalized in the next section allow a unified approach and many possible objective functions $Q(.)$ such as

- Wald- or Score-type inference associated with some standardized square norm of the difference between $\hat{g}(Y_T)$ and $\overline{\tilde{g}}_{T,H}(\theta)$.

- Quasi-Likelihood-Ratio-type inference associated with some ratio of $\hat{g}(Y_T)$ and $\overline{\tilde{g}}_{T,H}(\theta)$.

We could, alternatively, focus on testing directly restrictions on the population analogue of $\hat{g}(Y_T)$. However, this would require the definition of such population analogues, which is not always clear or well-defined in the complex models we have in mind[1]. In addition, it would also yield an unrestricted and often non-structural interpretation of the parameter θ which may not be the most relevant in practice.

2.2 Motivating example

We conclude this section by presenting a simple framework that calls for a likelihood-free approach and where a natural auxiliary statistic is readily available. Our small sample analysis in section 4 goes beyond this simple model and considers instead a stylized DSGE structure. Consider the structural linear (multivariate) regression model:

$$AY_t = BX_t + U_t \qquad \text{with} \qquad t = 1, \cdots, T\,, \tag{2}$$

where the observable dependent variable Y_t is $(n, 1)$, the observable explanatory variable X_t is $(k, 1)$, A and B are matrices of unknown structural parameters of size (n, n) and (n, k), and the unknown error term U_t is $(n, 1)$. In such model, our parameters of interest are B and Σ_U (the covariance matrix of U_t), whereas A is calibrated for economic analysis of interest since it is difficult to estimate directly. By contrast, the

[1] See for example the DSGE model considered in our simulation study in section 4.

natural reduced form of (2) can easily be estimated,

$$Y_t = \Gamma' X_t + V_t \qquad \text{with} \qquad t = 1, \cdots, T\,, \tag{3}$$

where Y_t and X_t are the same observables as above, while Γ is the unknown matrix of size (k, n) and V_t is the unknown error term of size $(n, 1)$.

Our simulation-based matching inference procedure requires three main ingredients:

(i) the auxiliary statistic used to summarize the information contained in the observables;

(ii) the simulator used to simulate analogues of the auxiliary statistic;

(iii) the criterion function used to compare the auxiliary statistic to its simulated counterpart.

To fix ideas, we end this section by reviewing possible choices for each of these three ingredients.

(i) The auxiliary statistic is often chosen as a convenient way to "summarize" some key features of the information available from the observables, $Z_T = \{(Y_t, X_t), t = 1, \cdots, T\}$: for example, one may focus on the OLS estimator of Γ,

$$\hat{\Gamma}_T = (X'X)^{-1} X'Y \equiv \hat{g}_T(Z_T)$$

with Y the (T, n) matrix with t-th row Y_t' and X the (T, k) matrix with t-th row X_t'. Alternatively, other estimators may be considered, as well as other auxiliary statistics that may focus on different features of the observables such as the variance-covariance matrix of the error term, or some impulse-response functions to only name a few.

(ii) The simulator can be chosen as a parametric model that is used to simulate an analogue of Z_T. For given $\overline{\theta} = \texttt{vec}(\overline{A}, \overline{B})$, we can generate $\tilde{Z}_{T,h}(\overline{\theta}) = \{(\tilde{Y}_{t,h}, X_t), t = 1, \cdots, T\}$ for $h = 1, \cdots, H$,

$$\overline{A}\tilde{Y}_{t,h} = \overline{B}X_t + \overline{U}_t\,, \qquad \overline{U}_t \sim \mathcal{N}(0, \overline{\Sigma}_u) \qquad \text{with } t = 1, \cdots, T\,,$$

and $\overline{\Sigma}_u$ an estimator of the variance of U_t obtained for given $(\overline{A}, \overline{B})$. The auxiliary statistic computed over each simulated path $\tilde{Z}_{T,h}$ is

$$\tilde{\Gamma}_{T,h} = (X'X)^{-1}X'\tilde{Y}_h = \hat{g}_T(\tilde{Z}_{T,h}(\overline{\theta})) \equiv \tilde{g}_{T,h}(\overline{\theta})\,,$$

and its average over the H simulated paths is

$$\overline{\tilde{\Gamma}}_{T,H} = \frac{1}{H}\sum_{t=1}^{T}\tilde{\Gamma}_{T,h} \equiv \overline{\tilde{g}}_{T,H}(\overline{\theta})\,.$$

(iii) The criterion function $Q(.)$ compares the auxiliary statistic computed over the observed data and the simulated one, respectively $\hat{g}_T(Z_T)$ and $\overline{\tilde{g}}_{T,H}(\overline{\theta})$. For example:

- Quasi-Likelihood-Ratio-type inference using the Wilks distance,

$$Q_{LR}(\hat{g}_T(Z_T), \overline{\tilde{g}}_{T,H}(\overline{\theta})) = Q_{LR}(\hat{\Gamma}_T, \overline{\tilde{\Gamma}}_{T,H}) = \frac{\det\left[(Y - X\overline{\tilde{\Gamma}}_{T,H})'(Y - X\overline{\tilde{\Gamma}}_{T,H})\right]}{\det\left[(Y - X\hat{\Gamma}_T)'(Y - X\hat{\Gamma}_T)\right]}$$

- Wald-type inference using

$$Q_W(\hat{g}_T(Z_T), \overline{\tilde{g}}_{T,H}(\overline{\theta})) = T\left[\hat{g}_T(Z_T) - \overline{\tilde{g}}_{T,H}(\overline{\theta})\right]'\hat{\Sigma}_T^{-1}(\overline{\theta})\left[\hat{g}_T(Z_T) - \overline{\tilde{g}}_{T,H}(\overline{\theta})\right]$$

where $\hat{\Sigma}_T(\overline{\theta})$ is an estimator of the (asymptotic) variance covariance matrix

of $[\hat{g}_T(Z_T) - \overline{\tilde{g}}_{T,H}(\overline{\theta})]$ which can be obtained by bootstrap as explained in section 3.1.

– or using the Bai statistic,

$$Q_B(\hat{\Gamma}_T, \overline{\tilde{\Gamma}}_{T,H}) \;=\; \log|\tilde{\Sigma}_{T,H}| + \texttt{trace}(\hat{\Sigma}_T \tilde{\Sigma}_{T,H}^{-1})$$

$$\text{with} \quad \hat{\Sigma}_T = \frac{(Y - X\hat{\Gamma}_{T,H})'(Y - X\hat{\Gamma}_{T,H})}{(T-k)} \quad \text{and} \quad \tilde{\Sigma}_{T,H} = \frac{(Y - X\overline{\tilde{\Gamma}}_{T,H})'(Y - X\overline{\tilde{\Gamma}}_{T,H})}{(T-k)}$$

Many other criterion functions have been proposed in the literature, and we do not provide an exhaustive review of them in this paper.

3 Simulation-based matching inference under $\mathcal{H}_0$

In this section, we establish the asymptotic validity of our simulation-based matching inference procedure. We start by formally detailing the implementation of our simulation-based inference procedure and then we present regularity assumptions that will ensure its asymptotic validity.

3.1 Implementation

Our proposed simulation-based inference procedure relies on inverting a test statistic for the null hypothesis that fixes θ_1 at a known value, $\mathcal{H}_0 : \theta_1 = \theta_{1,0}$, while θ_2 is calibrated at $\overline{\theta}_2$. The following algorithm describes our simulation-based procedure.

Algorithm 1 *(General implementation)*

1. *Using the sample of T observations, compute the vector of auxiliary statistics $\hat{g}_T(Y_T)$.*

2. *For given $\overline{\theta}_0 = (\theta_{1,0}' \; \overline{\theta}_2')'$ with $\theta_{1,0} \in \Theta_1$, use the simulator to generate H independent*

vectors of auxiliary statistics, $\tilde{g}_{T,h}(\overline{\theta}_0)$ *with* $h = 1, \cdots, H$*, and compute*

$$Q_T(\overline{\theta}_0) \equiv Q_T(\hat{g}_T(Y_T), \overline{\tilde{g}}_{T,H}(\overline{\theta}_0)) \quad \textit{with} \quad \overline{\tilde{g}}_{T,H}(\overline{\theta}_0) = \frac{1}{H}\sum_{h=1}^{H} \tilde{g}_{T,h}(\overline{\theta}_0) \tag{4}$$

3. *For* $\overline{\theta}_0 = (\theta'_{1,0} \ \overline{\theta}'_2)'$*:*

   (a) *Use the simulator to generate* B *independent simulated samples* $\tilde{Y}_{T,b}(\overline{\theta}_0)$ *(with* $b = 1, \cdots, B$*) leading to* B *independent realizations of the vector of auxiliary statistics* $\tilde{g}_{T,b}(\overline{\theta}_0)$*, and compute* $Q_{T,b}(\overline{\theta}_0) \equiv Q_T(\tilde{g}_{T,b}(\overline{\theta}_0), \overline{\tilde{g}}_{T,H}(\overline{\theta}_0))$*.*

   (b) *Compute the bootstrap p-value for* Q*,* $\hat{p}_{TB}(Q_T(\overline{\theta}_0)|\theta_{1,0}, \overline{\theta}_2)$*, where*

$$\hat{p}_{TB}(x|\theta) = \frac{1}{B+1}\sum_{b=1}^{B} [1(Q_{T,b}(\theta) \geq x) + 1] \quad \textit{with } \theta = (\theta'_1 \ \theta'_2)'. \tag{5}$$

4. *Decide on the rejection/non rejection of the null hypothesis by comparing the bootstrap p-value to the chosen significance level.*

When Q_T is a quadratic criterion function of the difference between the auxiliary statistic and its simulated counterpart - so-called Wald-type inference - one also needs to estimate the associated weighting matrix which corresponds to the variance-covariance matrix of this difference. The following updates Algorithm 1 to explicitly incorporate the computation of the weighting matrix which is done imposing the null hypothesis at no additional computational cost. Imposing the null for such scaling is akin to Anderson-Rubin type practices rather than standard Wald ones that typically rely on unrestricted estimates; see Guerron-Quintana, Inoue, and Kilian (2017) for the latter.

Algorithm 2 *(Special case: Wald-type implementation)*

1. *Using the sample of* T *observations, compute the vector of auxiliary statistics* $\hat{g}_T(Y_T)$*.*

2. *For a given* $\overline{\theta}_0 = (\theta'_{1,0}\ \overline{\theta}'_2)'$*, use the simulator to generate* H *independent vectors of auxiliary statistics,* $\tilde{g}_{T,h}(\overline{\theta}_0)$ *with* $h = 1, \cdots, H$*, and compute* $W_T(\overline{\theta}_0)$ *as*

$$\begin{aligned} W_T(\overline{\theta}_0) &= \left[\hat{g}_T(Y_T) - \overline{\tilde{g}}_{T,H}(\overline{\theta}_0)\right]' \left[\hat{S}_{T,0}\right]^{-1} \left[\hat{g}_T(Y_T) - \overline{\tilde{g}}_{T,H}(\overline{\theta}_0)\right] \\ \textit{where}\quad \overline{\tilde{g}}_{T,H}(\overline{\theta}_0) &= \frac{1}{H}\sum_{h=1}^{H} \tilde{g}_{T,h}(\overline{\theta}_0) \\ \textit{and}\quad \hat{S}_{T,0} &= \frac{1}{H}\sum_{h=1}^{H} \left[\tilde{g}_{T,h}(\overline{\theta}_0) - \overline{\tilde{g}}_{T,H}(\overline{\theta}_0)\right]' \left[\tilde{g}_{T,h}(\overline{\theta}_0) - \overline{\tilde{g}}_{T,H}(\overline{\theta}_0)\right] \end{aligned}$$

3. *For* $\overline{\theta}_0 = (\theta'_{1,0}\ \overline{\theta}'_2)'$

   (a) *Use the simulator to generate* B *independent samples* $\tilde{Y}_{T,b}(\overline{\theta}_0)$ *(with* $b = 1, \cdots, B$*) leading to* B *independent realizations of the vector of auxiliary statistics* $\tilde{g}_{T,b}(\overline{\theta}_0)$*, and compute*

$$W_{T,b}(\overline{\theta}_0) = \left[\tilde{g}_{T,b}(\overline{\theta}_0) - \overline{\tilde{g}}_{T,H}(\overline{\theta}_0)\right]' \left[\hat{S}_{T,0}\right]^{-1} \left[\tilde{g}_{T,b}(\overline{\theta}_0) - \overline{\tilde{g}}_{T,H}(\overline{\theta}_0)\right]$$

   (b) *Compute the bootstrap p-value for* W*,* $\hat{p}^{(W)}_{TB}(W_T(\overline{\theta}_0)|\theta_{1,0}, \overline{\theta}_2)$*, where*

$$\hat{p}^{(W)}_{TB}(x|\theta_1, \overline{\theta}_2) = \frac{1}{B+1}\sum_{b=1}^{B} \left[1(W_{T,b}(\overline{\theta}_0) \geq x) + 1\right] \tag{6}$$

4. *Decide on the rejection/non rejection of the null hypothesis by comparing the bootstrap p-value to the chosen significance level.*

3.2 Asymptotic validity

In order to establish the asymptotic validity of our simulation-based matching inference, we will use the following high-level regularity assumption.

Assumption 1 *(High-level regularity conditions)*

For given $\theta_{1,0}$*, there exists* $\theta_{2,0} \equiv \theta_{2,0}(\theta_{1,0})$ *such that with* $\theta_0 \equiv (\theta'_{1,0}, \theta'_{2,0})'$

(i) $Q_T(\hat{g}_T(Y_T), \overline{\tilde{g}}_{T,H}(\theta_0)) \xrightarrow{P} Q_0$*;*

(ii) D_0 *is a subset of* $\mathbb{R}$ *such that for some* $I_0 \in \mathbb{N}$*,*

$$P\left[Q_T(\hat{g}_T(Y_T), \overline{\tilde{g}}_{T,H}(\theta_0)) \in D_0 \text{ and } Q_0 \in D_0 \text{ for all } T \geq I_0\right] = 1\,;$$

(iii) $\forall x \in D_0$*,* $\forall \eta > 0$*, and given any sequence* $\theta_{2,T} \xrightarrow{T} \theta_{2,0}$*, there exists an open neighborhood* $B(x,\eta)$ *such that, with* $\theta_T = (\theta'_{1,0}, \theta'_{2,T})'$

$$\limsup_T \left\{ \sup_{y \in B(x,\eta) \cap D_0} |Z_T(y|\theta_T) - Z_T(y|\theta_0)| \right\} \leq \eta\,,$$

where $Z_T(y|\theta) \equiv P(Q_T(\hat{g}_T(Y_T), \overline{\tilde{g}}_{T,H}(\theta)) \leq y)$*.*

The convergence in probability maintained in the first condition is natural in the context of our simulation-based (or indirect) inference where the criterion $Q_T(.)$ is user-chosen[2]. For example, in DSGE models, inference is often obtained through Wald-type impulse response matching with a linear auxiliary model where all statistics converge. The remaining conditions can be understood as a local equicontinuity condition (at $\theta = \theta_0$) on the sequence of distributions functions $Z_T(y|\theta)$. Such local equicontinuity condition holds whenever $Z_T(y|\theta)$ converges to a distribution which is continuous in (y, θ), or whenever $Z_T(y|\theta)$ admits an expansion around a pivotal distribution; see e.g. Dufour (2006) for related discussions.

It is important to emphasize that Assumption 1 is not an identification assumption: e.g. $\theta_{1,0}$ corresponds to some value maintained by the null hypothesis, while $\theta_{2,0}$ will likely depend on $\theta_{1,0}$; however, we do not maintain the uniqueness of such $\theta_{2,0}$. Similarly, the set D_0 is adaptive with respect to the null in the sense that it likely

[2]Regularity conditions that do not maintain such convergence in probability of Q_T can be derived at the price of strengthening *(iii)* into a global equicontinuity condition; see Dufour (2006, section 6).

depends on $\theta_{1,0}$. Assumption 1 rather ensures that, under $\mathcal{H}_0$, the *adequacy* between the data-based statistics and its simulation-based counterpart is not impaired by the chosen (calibrated) value of the nuisance parameters. Such calibration may either be exact when $\theta_2 = \theta_{2,0}$ or approximate when considering sequences $\theta_{2,T} \xrightarrow{T} \theta_{2,0}$. These sequences enable us to define adaptive neighborhoods through the criterion function $Q_T(.)$. Alternatively, we could also consider fixed neighborhoods and replace *(iii)* by

(iii)* $\forall x \in D_0$, $\forall \eta > 0$, $\exists \delta > 0$ *and an open neighborhood* $B(x, \eta)$ *such that, with* $\theta = (\theta'_{1,0}, \theta'_2)'$

$$\|\theta_2 - \theta_{2,0}\| \leq \delta \;\Rightarrow\; \limsup_T \left\{ \sup_{y \in B(x,\eta) \cap D_0} |Z_T(y|\theta) - Z_T(y|\theta_0)| \right\} \leq \eta \,. \qquad (7)$$

Intuitively, the introduction of a ball of size δ around $\theta_{2,0}$ as done in condition (iii^*) assumes that the *direction* of the approximate calibration does not matter, as long as it is of a small enough magnitude δ. This is in contrast with the sequence $\{\theta_{2,T}\}$ considered in condition (iii) since $\{\theta_{2,T}\}$ approaches $\theta_{2,0}$ along a specific direction. The results of our simulation study suggest that, in addition to the magnitude of the departure from $\theta_{2,0}$, the direction of departure plays an important role in the sense that some directions of departure appear more harmful than others: for example, when calibrating the autocorrelation parameter of some macroeconomic shock in the DSGE model considered in section 4, a (slight) under-calibration of such parameter is much more consequential than an over-calibration. As a result, (iii) appears to be more in line with our simulation results and we continue working with Assumption 1 from now on.

The next result formalizes the asymptotic validity of our simulation-based matching inference procedure in a broad and unified framework under Assumption 1.

Theorem 1 *(Asymptotic validity)*

Assume that the regularity conditions stated in Assumption 1 hold.

Then, under $\mathcal{H}_0 : \theta_1 = \theta_{1,0}$*, for* $0 < \alpha < 1$*,*

$$\lim_{T\to\infty} P\left[\hat{p}_{TB}(Q_T(\theta_{0,T})|\theta_{1,0}, \theta_{2T}) \leq \alpha\right] = \alpha$$

where $\hat{p}_{TB}(x|\theta_{1,0}, \theta_{2T})$ *is defined in (5) with* $Q_T(.)$ *as in (4),* $\theta_{0,T} = (\theta'_{1,0}\ \theta'_{2T})'$*,* θ_{2T} *as in Assumption 1(iii) and* B *chosen so that* $\alpha(B+1)$ *is an integer.*

Our main result generalizes the (asymptotic) validity of Monte Carlo tests established by Dufour (2006) to allow statistics of interest to depend on nuisance parameters (here θ_2) that are not estimated, partialled out, or known under $\mathcal{H}_0$. More specifically, we calibrate these nuisance parameters, either exactly at $\theta_{2,0}$ or approximately[3] along sequences $\theta_{2,T}$ converging to $\theta_{2,0}$. Such an assumption generalizes the concept of *partial encompassing* (or exact calibration of nuisance parameters) introduced by Dridi, Guay and Renault (2007). Our result also generalizes - in part - recent results obtained by Bergamelli, Bianchi, Khalaf and Urga (2019) for testing multiple nulls simultaneously since their underlying parameter of interest is estimated.

3.3 Sensitivity analysis of the nuisance parameters

When calibrating (nuisance) parameters that are otherwise difficult to estimate or partial out, applied researchers often have several parameter values in mind that are either obtained from major published studies, or correspond to different economic analyses of interest. And, indeed, researchers often try them all to see whether the associated inference results on the other (structural) parameters of interest change much or not. Our simulation-based matching inference allows us to easily accommodate

[3]The nuisance parameters could also be approximately calibrated by considering fixed balls around $\theta_{2,0}$ as done in condition (iii^*) stated on page 15.

and formalize such sensitivity analysis by considering the supremum of the (bootstrap) p-values computed with each chosen calibrated value. In practice, researchers may not be considering more than a couple of such values, in which case the supremum may not actually be too costly.

We start by presenting an updated version of Algorithm 1 that explicitly accounts for such "set calibration" of the nuisance parameters θ_2, while the null hypothesis fixes θ_1 at a known value, $\mathcal{H}_0 : \theta_1 = \theta_{1,0}$,

Algorithm 3 *(General implementation over a set of calibrated values)*

Let Θ_2 be the set of J candidates $\overline{\theta}_{2,j}$, $j = 1, \cdots, J$ considered for the calibration of parameters θ_2.

1. *For each j, define $\overline{\theta}_{j,0} = (\theta'_{1,0}\ \overline{\theta}'_{2,j})'$ with given $\theta_{1,0} \in \Theta_1$ and follow Algorithm 1 (or 2) to compute the associated bootstrap p-value $\hat{p}_{TB}(Q_T(\overline{\theta}_{j,0})|\theta_1, \overline{\theta}_{2,j})$ and the supremum*

$$\hat{p}^*_{TB}(\theta_1, \Theta_2) = \sup_{\overline{\theta}_{2,j} \in \Theta_2} \left[\hat{p}_{TB}(Q_T(\overline{\theta}_{j,0})|\theta_1, \overline{\theta}_{2,j})\right] \tag{8}$$

2. *Decide on the rejection/non rejection of the null hypothesis by comparing the sup bootstrap p-value $\hat{p}^*_{TB}(.)$ to the chosen significance level.*

The asymptotic validity of our simulation-based matching inference over a set of calibrated values directly follows from our results obtained in the previous section. The following assumption updates Assumption 1 accordingly.

Assumption 2 *(Regularity conditions under set calibration)*

(i) For given $\theta_{1,0} \in \Theta_1$, let Θ_{2T} denote a finite set of J approximate calibrated parameter values denoted $\theta_{2,j,T}$ (with $j = 1, \cdots, J$) such that $\theta_{2,j,T} \xrightarrow{T} \theta_{2,j,0}(\theta_{1,0})$, and define $\theta_{j,0} \equiv (\theta'_{1,0}, \theta_{2,j,0}(\theta_{1,0})')'$.

(ii) $\forall j = 1, \cdots, J$, $Q_T(\hat{g}_T(Y_T), \overline{\tilde{g}}_{T,H}(\theta_{j,0})) \xrightarrow{P} Q_0$*; and,* D_0 *is a subset of* $\mathbb{R}$ *such that for some* $I_0 \in \mathbb{N}$,

$$P\left[Q_T(\hat{g}_T(Y_T), \overline{\tilde{g}}_{T,H}(\theta_{j,0})) \in D_0 \text{ and } Q_0 \in D_0 \text{ for all } T \geq I_0\right] = 1\,.$$

(iii) $\forall x \in D_0$, $\forall \eta > 0$, *there exists an open neighborhood* $B(x, \eta)$ *such that,* $\forall j = 1, \cdots, J$ *and* $\theta_{j,T} = (\theta'_{1,0}, \theta'_{2,j,T})'$

$$\limsup_T \left\{ \sup_{y \in B(x,\eta) \cap D_0} |Z_T(y|\theta_{j,T}) - Z_T(y|\theta_{j,0})| \right\} \leq \eta\,,$$

where $Z_T(y|\theta) \equiv P(Q_T(\hat{g}_T(Y_T), \overline{\tilde{g}}_{T,H}(\theta)) \leq y)$.

Under Assumption 2, each approximate calibrated value $\theta_{2,j,T}$ converges towards a limit point $\theta_{2,j}$ which ensures the adequacy between the data-based statistics and its simulation-based counterpart as originally formalized in Assumption 1. Assumption 2 does not maintain that these limit points are all equal, and emphasizes once again that there could be multiple values of θ_2 that ensures the above-mentioned adequacy. As a result, none of our regularity condition should be interpreted as an identification assumption.

The next result formalizes the asymptotic validity of our simulation-based matching inference procedure over a set of calibrated values under Assumption 2.

Corollary 2 *(Asymptotic validity over a set of calibrated values)*
Assume that the regularity conditions stated in Assumption 2 hold.
Then, under $\mathcal{H}_0 : \theta_1 = \theta_{1,0}$, *for* $0 < \alpha < 1$,

$$\lim_{T \to \infty} P\left[\hat{p}^*_{TB}(\theta_{1,0}, \Theta_{2T}) \leq \alpha\right] = \alpha$$

where $\hat{p}^*_{TB}(\theta_{1,0}, \Theta_{2T})$ is defined in (8), Θ_{2T} is the finite set of approximate calibrated values for θ_2 and B chosen so that $\alpha(B+1)$ is an integer.

4 Simulation study

We illustrate the reliability and the applicability of our inference procedure in a baseline macro model taken from Fernandez-Villaverde, Rubio-Ramirez and Schorfheide (2016, Chapter 8) and adapted from Del Negro and Schorfheide (2008).

4.1 Model and notations

The stylized DSGE (Dynamic Stochastic General Equilibrium) model consists of several sectors including households, intermediate and final goods producers, and a monetary authority. A Calvo assumption is used to introduce nominal rigidity in prices, and firms that cannot re-optimize their prices at a given time adjust these by the steady-state inflation rate.

Denoting the log deviation of a variable w_t from its steady-state by $\hat{w}_t$, the log-linearized equilibrium conditions of the model for output, X_t, labor share, lsh_t, inflation, π_t and interest rate, R_t, are given by:

$$\begin{aligned} \hat{x}_t &= E_t[\hat{x}_{t+1}] - (\hat{R}_t - E_t[\hat{\pi}_{t+1}]) + E_t[z_{t+1}], \\ \widehat{lsh}_t &= \hat{x}_t + \phi_t, \\ \hat{\pi}_t &= \beta E_t[\hat{\pi}_{t+1}] + \frac{(1-\zeta_p\beta)(1-\zeta_p)}{\zeta_p}(\widehat{lsh}_t + \lambda_t), \\ \hat{R}_t &= \frac{1}{\beta}\hat{\pi}_t + \sigma_R \epsilon_{R,t}. \end{aligned} \tag{9}$$

In the above equations, β is the stochastic discount rate and the probability with which a given firm is unable to re-optimize its price is given by ζ_p. In addition, four exogenous

shocks influence the dynamics of the variables. These include a technology shock, z_t, a price markup shock, λ_t, a shock that affects the preference for leisure, ϕ_t, and a monetary policy shock, $\epsilon_{R,t}$. Except for the monetary policy shock, which is assumed to be independently and identically normally distributed with mean zero and variance 1, the remaining shocks are assumed to follow autoregressive processes. Thus, for each shock $i = z, \lambda, \phi$, the autoregression coefficient is ρ_i and the standard deviation is σ_i. The unknown structural parameters of the model are collected in the vector

$$\theta = [\zeta_p, \beta, \gamma, \lambda, \pi^*, \rho_\phi, \rho_\lambda, \rho_z, \sigma_\phi, \sigma_\lambda, \sigma_z, \sigma_R]',$$

where γ is the growth rate of technology, λ is the steady-state markup charged by the intermediate goods producers, and π^* is the steady-state inflation rate. The steady-states for the interest rate and for the labor share can be obtained from the expressions $\bar{R} = \pi^*\gamma/\beta$, and, $\bar{lsh} = 1/(1+\lambda)$, respectively.

This baseline model is designed to a have a state-space representation which is used to obtain our sample of observables Y_T; see Table 4 in Fernandez-Villaverde, Rubio-Ramirez and Schorfheide (2016, Chapter 8). To simplify exposition, we focus on delivering inference on only one parameter of the model, namely the Calvo parameter ζ_p (the probability that firms do not re-optimize the prices they charge), while the remaining structural parameters are calibrated (see Table 1) to values suggested in the literature (e.g. in Fernandez-Villaverde, Rubio-Ramirez and Schorfheide (2016) and Del Negro and Schorfheide (2008)).

In order to deliver inference on ζ_p, our approach allows us to harvest information from the model that comes as (chosen) functions of the model parameters (so-called auxiliary statistics). In our DSGE model, we focus on the information contained in the impulse responses (up to four quarters) of the four endogenous variables of the model following

a 1-unit increase in the monetary policy shock. In the literature, impulse responses have been assessed using notably three different functions, and, accordingly, we consider all three as well[4]: VAR-based reduced-form impulse response functions (RIRF), VAR-based Cholesky-orthogonalized structural impulse response functions (SIRF), and local projection based impulse response function (LPIRF). In addition, to assess the distances between impulse responses computed with observed and simulated data, we consider the following:

- Wald-type inference on SIRF (or RIRF) using 2 possible weighting matrices: either population weights or parametric bootstrap weights;
- Bai-type inference on LPIRF with and without MA correction[5].

The implementation of these inference procedures is detailed in the Supplementary Appendix.

4.2 Simulation results

In our experiments, we consider the data generating process (DGP) described in (9) with different values for the underlying parameter θ. We are interested in testing hypotheses on the parameter θ_1 which here is the Calvo parameter ζ_p, the probability with which a given firm is unable to reoptimize its price, while the remaining parameters θ_2 are calibrated.

- **Experiment #1:** size and power with exact calibration.

[4]It is important to mention that, in our stylized model, the impulse responses are known analytically. However, our inference procedure is implemented without relying on these analytical expressions.

[5]We thank Oscar Jordà for suggesting the MA correction: typically, the moving average errors are corrected by recursively including the residuals; see also Algorithm #3 in the Supplementary Appendix.

We generate a sample of 100 observations according to (9) with ζ_p taking one of nine possible values $\zeta_{p,0}$ from 0.1 to 0.9, while the remaining parameters are set according to Table 1. For each of these 9 DGPs, we test the following null hypotheses on ζ_p, at the 5% significance level, namely:

$$\mathcal{H}_0 : \zeta_p = c\,,$$

where c takes 11 values between 0.01 and 0.99. The remaining parameters θ_2 are correctly calibrated. The associated empirical rejection probabilities (obtained with 1,000 replications) are represented in Figures 1 to 4 where each curve corresponds to a different DGP according to $\zeta_{p,0}$. We consider respectively the following four inference procedures: Bai-LPIRF, Bai-LPIRF with MA correction, Wald-SIRF with parametric bootstrap weights, and Wald-RIRF with parametric bootstrap weights.

Figure 1 focuses on Bai-LPIRF. It is worth noting that size is controlled across all DGPs. This is expected because the experiment is restricted to exact matching, yet the point is emphasized because the statistic as proposed ignores the MA errors that are inherent to local projections. Because we draw from the model with no nuisance parameters (in the context of exact calibration), the Monte-Carlo method will deliver size-correct p-values even when MA effects are ignored. Is there information in the error dynamics of local projections? The power analysis provides useful insights into this question. Indeed, the exact calibration case is an interesting experimental environment to assess whether, in addition to the comparative advantage of each statistic, IRs hold useful information on the Calvo parameter.

When the tested value of ζ_p (that is, c) exceeds the *truth* $[c > \zeta_{p,0}]$, we observe good power for all considered DGPs (that is, for all considered *truths*). In stark contrast, we find no power when the tested value c is lower than the *truth* $[c < \zeta_{p,0}]$: the test

appears in fact to be biased (power less than size) against such alternatives. This contrast holds over all considered DGPs, and overall, the power curves do not differ much across DGPs.

The striking difference in power that we observe may point to reasons related to the structure of the macroeconomic model. The Calvo parameter is known to be quite hard to pin down, and unbounded confidence sets are not to be ruled out; see e.g. Mavroeidis, Plagborg-Møller, and Stock (2014). Yet, our results in Figures 2 to 4 suggest otherwise.

Figure 2 depicts size-power curves for an LR-type statistic - as in Figure 1 - after being corrected for MA errors. We find that bias is completely corrected for $c < \zeta_{p,0}$, with no power loss for the $c > \zeta_{p,0}$ direction. Despite clear improvements on the former direction, we still find a notable asymmetry in power: the power curves for the $c < \zeta_{p,0}$ cases are visibly flatter and peak at a maximum of 40% compared to a fast convergence to one for the $c > \zeta_{p,0}$ case. Nevertheless, the information content of the MA correction represents an interesting result we would like to emphasize.

The MA correction remains costly data-wise since a few lags are needed for its implementation, which may matter for the small sample sizes that we consider in this experiment. In fact, the Wald-type statistics depicted in Figures 3 and 4 confirm that power symmetry is attainable. A qualification is worth raising here, since contrary to conventional practice from a general Wald-type perspective, the weighting matrix we use is obtained *drawing from the null model.* Concretely, since a preliminary simulation is needed to center the impulse responses, there is no added cost if this simulation is also used to scale the deviations.

The main difference between Figures 3 and 4 is that the former relies on structural impulse responses whereas the latter depicts reduced form measures. For comparison purposes with Figure 1, recall that the LPIRFs are (by construction) reduced form

measures. With one exception, we find no crucial difference between structural and reduced form measures. This result is noteworthy in view of current practices in macroeconomics. The exception is again for alternatives of the form $c < \zeta_{p,0}$ with $c \leq 0.5$, as the maximum power we observe for these cases is higher with structural IRs. The power advantage is not substantial, but remains noticeable: one way to interpret this finding is that structural measures add power when it is most needed.

- **Experiment #2:** deviation from the null.

In this experiment, we test the same null hypothesis throughout,

$$\mathcal{H}_0 : \zeta_p = 0.65 \,,$$

with the remaining parameters calibrated according to Table 1. We generate a sample of 100 observations according to (9) where we change one parameter at a time from its value under the null. Specifically, we consider changes in the following seven parameters: the Calvo parameter ζ_p, as well as the autoregression coefficients of the leisure shock, price markup shock, and technology shock, ρ_ϕ, ρ_λ, ρ_z, and their corresponding standard deviation, σ_ϕ, σ_λ, σ_z. We collect the associated empirical rejection probabilities computed for the following 4 inference procedures: Wald-SIRF with parametric bootstrap weights; Wald-RIRF with parametric bootstrap weights; Bai-LPIRF; and Bai-LPIRF with MA correction. In Figures 5 and 6, we display 7 graphs, each corresponds to the component of θ that deviates from its value under $\mathcal{H}_0$; for example, the graph located on the left of the top row collects rejection probabilities when the sample is generated with ζ_p taking values between 0.1 and 0.9 (while the remaining parameters are set according to Table 1); the graph located in the middle of the top row displays empirical rejection probabilities when the sample is generated with $\zeta_p = 0.65$ when ρ_ϕ takes values between 0.1 and 0.9 (while the remaining parameters are set according to

Table 1). Notice that the only difference between Figures 5 and 6 is the null value for the persistence parameter $\rho_{\lambda,0}$, respectively set at 0.88 and 0.33[6]. Our setup allows a direct comparison of the power of the different inference procedures when considering deviations from the null along one specific direction.

Several results emerge from these comparisons. First, LR-type statistics lack power for $0.65 < \zeta_p$ as can be seen in the top left graph. In addition, the MA correction to the Bai-LPIRF does not seem to pay off broadly; in fact, in most cases it seems to cost some power. Also noted is the better performance of the Wald-type statistics where the structural approach dominates, albeit marginally. That being said, Wald statistics neither dominate for all parameters nor for all directions.

Another result which challenges conventional wisdom is that structural impulse responses do not convey much more information than reduced form ones. For inference on σ_λ and σ_z, the structural approach actually costs power drastically to the left of the tested value.

For the persistence parameters, we observe power asymmetry for all statistics when the true parameter is lower than the tested value. This result becomes most evident when we interpret the difference between Figures 5 and 6 on ρ_λ. In the former, the true value of ρ_λ is 0.88, in contrast to 0.33 in the latter. We actually find that none of the statistics have any power to detect departures of ρ_λ in Figure 5, in contrast to good performance to the right of 0.33 in Figure 6.

We also observe asymmetry in the ranking of alternative statistics in both Figures 5 and 6, depending on whether the calibrated value is lower or higher than the true one. When both LR- and Wald-types of statistics have power, the LR-type ones dominate although not uniformly. When LR-type statistics have practically no power, there are

[6]A number of estimates (typically Bayesian) have been obtained for this parameter in the context of different medium-size DSGE models, and these vary substantially. For instance, Del Negro and Shorfheide (2008) find a posterior estimate value of 0.88 for their 'agnostic' scenario, whereas Ferroni, Grassi and Leon-Ledesma (2019) find a posterior estimate value of 0.35 in their preferred scenario.

various cases where their Wald-type counterparts are (quite) informative. However, no distinctive clear pattern emerges that could shed light on such differences.

Power asymmetries and sharp differences in power such as what we document for ρ_λ between Figures 5 and 6 reinforce arguments in the profession favouring calibration, and provide useful guidelines for such practices.

- **Experiment #3:** size and power with approximate calibration.

To illustrate and justify these guidelines, we report in Tables 2 to 4 simulations results with incorrectly calibrated parameters. We specifically investigate the impact of approximately calibrating one parameter, either ρ_ϕ (the autoregression coefficient of the shock that affects the preference for leisure) or ρ_λ (the autoregression coefficient of the price markup shock), on the size and power of our test. Recall from Figure 5 and associated discussions in Experiment #2 that our inference procedures do not have much power on ρ_λ. It would then make sense, from a practical point of view, to calibrate it.

We generate a sample of 100 observations according to (9) with ζ_p taking value $\zeta_{p,0} = 0.65$ while the remaining parameters are set according to Table 1. For such a DGP, we test the following null hypotheses on ζ_p, at the 5% significance level, namely:

$$\mathcal{H}_0 : \zeta_p = c\,,$$

where c takes 11 values between 0.01 and 0.99. The remaining parameters θ_2 are correctly calibrated except for either ρ_ϕ or ρ_λ.

(i) Miscalibration of ρ_ϕ, the autoregression coefficient of the shock that affects the preference for leisure:

ρ_ϕ is calibrated at a value between 0.1 and 0.9 whereas the sample is generated with $\rho_{\phi,0} = 0.3$. The associated rejection probabilities (obtained with 1,000 replications) are represented in Table 2 when inference is conducted using Bai-LPIRF with and without MA correction.

Focusing first on the size of the test which is obtained when testing $\mathcal{H}_0 : \zeta_p = 0.65$, we notice that the size remains well-controlled and somewhat insensitive to ρ_ϕ when over-calibrated, that is calibrated above 0.3 - even up to 0.9. However, when considering under-calibrated ρ_ϕ values, the test becomes over-sized more quickly: for example, with ρ_ϕ calibrated at 0.1 the rejection probability is 0.062 while it is 0.052 when ρ_ϕ is calibrated at 0.9 when considering Bai-LPIRF without MA correction. This is even more pronounced when considering Bai-LPIRF with MA correction: for example, with ρ_ϕ calibrated at 0.1 the rejection probability is 0.176 while it is close to 0 when ρ_ϕ is calibrated at 0.9.

Focusing now on testing $\mathcal{H}_0 : \zeta_p = c \neq 0.65$, the power properties of the test using Bai-LPIRF without MA correction do not seem to be much affected by the incorrect calibration of ρ_ϕ: there is much power to detect $\zeta_p > 0.65$ but little to none to detect $\zeta_p < 0.65$. For example, when testing $\zeta_p = 0.8$, the rejection probabilities are close to 1 for all calibrated values of ρ_ϕ between 0.1 and 0.9; similarly, when testing $\zeta_p = 0.6$ the rejection probabilities are close to 0 for all calibrated values of ρ_ϕ between 0.1 and 0.9. The power properties of the test using Bai-LPIRF with MA correction are much more sensitive to the incorrect calibration: in general, the power decreases when ρ_ϕ is over-calibrated and increases when it is under-calibrated.

(ii) Miscalibration of ρ_λ, the price markup shock autoregression coefficient:

ρ_λ is calibrated at a value between 0.1 and 0.95 whereas the sample is generated with either $\rho_{\lambda,0} = 0.3$ or $\rho_{\lambda,0} = 0.88$. The associated rejection probabilities (obtained with

1,000 replications) are represented in Table 3 when inference is conducted using Bai-LPIRF without MA correction.

Similarly to the results obtained with miscalibration of ρ_ϕ, the size of the test remains well-controlled when ρ_λ is over-calibrated, but it appears severely distorted when ρ_λ is under-calibrated. In addition, the power properties of the test do not seem to be much affected by the incorrect calibration of ρ_λ: there is much power to detect $\zeta_p > 0.65$ but little to none to detect $\zeta_p < 0.65$.

More generally, our results illustrate the importance of the direction of miscalibration - and not only of the magnitude of the miscalibration: e.g. a (slight) under-calibration of the autocorrelation parameter ρ_ϕ or ρ_λ is much more consequential than a (large) over-calibration. This is in-line with our asymptotic framework (see Assumption 1) that maintains equicontinuity along a specific direction of approximate calibration.

(iii) Robustness to the miscalibration of ρ_λ, the price markup shock autoregression coefficient:

So far, our results under miscalibration highlight the very serious consequences of miscalibrating - and, more specifically, of under-calibrating ρ_ϕ or ρ_λ. They also point to the sensible approach often adopted by empirical macroeconomists to *"go for more persistence when in doubt"*. The severe consequences of miscalibration reinforce the usefulness of our multi-scenario sup-p-value approach introduced in section 3.3 that consists in computing the supremum of the p-values obtained for each candidate calibrated value.

We implement the robust approach with respect to the calibration of ρ_λ when testing ζ_p. The true values of ζ_p and ρ_λ are set at $\zeta_{p,0} = 0.65$ and $\rho_{\lambda,0} = 0.88$. To compute the empirical rejection probabilities reported in Table 4, we consider two candidates for the calibration of ρ_λ: in panel A, 0.8 and 0.9; and in panel B, 0.7 and 0.95. The

size of the test remains well-controlled. In addition, the power properties of the test do not seem to be much affected by our robust procedure: there is still much power to detect $\zeta_p > 0.65$, but none to detect $\zeta_p < 0.65$.
Overall, we find that the power cost is minimal considering the severe over-rejections due to ignoring calibration effects in finite samples.

The above simulation studies demonstrate that Bai's statistic with Jorda's projection method generally performs well when compared to the other considered options. The method is sensitive to several miscalibrated parameters of the model, although this sensitivity is asymmetric and considerably more present when calibrating at lower values than the truth: e.g. calibrating a persistence parameter below its actual (unknown) value. Since a considerable amount of dynamics is injected into DSGE models via its 'external' sources, namely the model's shock processes, our simulation studies indicate the importance of properly pinning down the parameter values of these processes for testing θ (which, in our simplified case, consists of the Calvo parameter ζ_p). Indeed, if the latter are to be calibrated, they should be fixed at the highest economically-acceptable values so that the test remains valid and powerful. In addition, our multi-scenario sup-p-value approach appears promising. It can be interpreted as a way to operationalize - and formalize - the sensitivity analysis routinely done in empirical macro when calibrating the parameters of the shocks that are closing the model.

5 Conclusion

Our paper proposes a general procedure for inference with auxiliary statistics that does not require identification, and that is moreover compatible with - possibly multiple - calibration. Conditions underlying asymptotic validity are characterized beyond the

strict and exact calibration of intervening parameters. On the practical side, we focus on IR matching for DSGE models, including a local projection perspective. A laboratory analysis with a prototypical model illustrates the usefulness of calibration despite serious potential cost that can provably be addressed using our proposed methodology. Results further document useful testable directions that suggest practical calibration practices, shed light on the importance of MA dynamics in local projections and importantly, on the information content of structural IRs.

Acknowledgements

We thank conference participants (at EC2 2019 in Oxford, CIREQ 2019 in Montreal, SVEC 2019 in Vancouver, NBER 2019 Workshop on Methods and Applications for Dynamic Stochastic Equilibrium Models, 2020 World Congress of the Econometric Society), and seminar participants (at Warwick U., U. of Surrey, U. of Bergamo, Durham U.) for helpful comments and discussions.

References

[1] J. Bai, *Fixed-effects dynamic panel models, a factor analytical method*, Econometrica **81** (2013), 285–314.

[2] M.-C. Beaulieu, J.-M. Dufour., and L. Khalaf, *Multivariate tests of mean-variance efficiency with possibly non-Gaussian errors: an exact simulation-based approach*, Journal of Business and Economic Statistics **25** (2007), 398–441.

[3] ———, *Identification-robust estimation and testing of the zero-beta CAPM*, Review of Economic Studies **80** (2013), 892–924.

[4] ———, *Exact confidence set estimation and goodness-of-fit test methods for asymmetric heavy tailed stable distributions*, Journal of Econometrics **181** (2014), 3–14.

[5] M. Bergamelli, A. Bianchi, L. Khalaf, and G. Urga, *Combining p-values to test for multiple structural breaks in cointegrated regressions*, Journal of Econometrics **211** (2019), 461–482.

[6] L.E. Calvet and V. Czellar, *Through the looking glass: Indirect inference via simple equilibria*, Journal of Econometrics **185** (2015), no. 2, 343 – 358.

[7] L.J. Christiano, M. Eichenbaum, and C.L. Evans, *Nominal rigidities and the dynamic effects of a shock to monetary policy*, Journal of Political Economy **113** (2005), 1–45.

[8] V. Czellar and E. Ronchetti, *Accurate and Robust Tests for Indirect Inference*, Biometrika **97** (2010), 621–630.

[9] P. Dovonon and A. R. Hall, *The asymptotic properties of GMM and Indirect Inference under second-order identification*, Journal of Econometrics **205** (2018), 76–111.

[10] R. Dridi, A. Guay, and E. Renault, *Indirect inference and calibration of dynamic stochastic general equilibrium models*, Journal of Econometrics **136** (2007), 397–430.

[11] J.-M. Dufour, *Monte Carlo tests with nuisance parameters: A general approach to finite-sample inference and nonstandard asymptotics*, Journal of Econometrics **133** (2006), 443–477.

[12] J.-M. Dufour, D. Pelletier, and E. Renault, *Short-Run and Long-Run Causality in Time Series: Inference*, Journal of Econometrics **132** (2006), 337–362.

[13] J.-M. Dufour and E. Renault, *Short-Run and Long-Run Causality in Time Series: Theory*, Econometrica **66** (1998), 1099–1125.

[14] J. Fernandez-Villaverde, J.F. Rubio-Ramirez, and F. Schorfheide, *Solution and estimation methods for DSGE models*, Handbook of Macroeconomics **2** (2016).

[15] F. Ferroni, S. Grassi, and M. Leon-Ledesma, *Selecting Structural Innovations in DSGE models*, Journal of Applied Econometrics **34** (2019), 205–220.

[16] J.J. Forneron and S. Ng, *The ABC of simulation estimation with auxiliary statistics*, Journal of Econometrics **205** (2018), 112–139.

[17] R. Gallant and G. Tauchen, *Which moments to match?*, Econometric Theory **12** (1996), 657–681.

[18] M. G. Genton and E. Ronchetti, *Robust Indirect Inference*, Journal of the American Statistical Association **98** (2003), 67–76.

[19] C. Gouriéroux, A. Monfort, and E. Renault, *Indirect inference*, Journal of Applied Econometrics **85** (1993), 85–117.

[20] G. Gouriéroux, P. Phillips, and J. Yu, *Indirect inference of dynamic panel models*, Journal of Econometrics **157** (2010), 68–77.

[21] A. Guay and O. Scaillet, *Indirect inference, nuisance parameter, and threshold moving average models*, Journal of Business and Economic Statistics **21** (2003), no. 1, 122–32.

[22] P. Guerron-Quintana, A. Inoue, and L. Kilian, *Impulse response matching estimators for DSGE models*, Journal of Econometrics **196** (2017), 144–155.

[23] A. Hall, A. Inoue, J. Nason, and B. Rossi, *Information criteria for impulse response function matching estimation of DSGE models*, Journal of Econometrics **170** (2012), 499–518.

[24] A. Inoue and L. Kilian, *Inference on impulse response functions in structural VAR models*, Journal of Econometrics **177** (2013), 1–13.

[25] O. Jordà, *Estimation and inference of impulse responses by local projections*, American Economic Review **95** (2005), 161–182.

[26] O. Jordà and S. Kozicki, *Estimation and inference by the method of projection minimum distance: An application to the new Keynesian hybrid Phillips curve*, International Economic Review **52** (2011), no. 2, 461–487.

[27] T. Kaji, E. Manresa, and G. Pouliot, *An adversarial approach to structural estimation*, Working paper (2020).

[28] L. Khalaf, Z. Lin, and A. Reza, *Beyond the Linearized Straight Jacket: Finite Sample Inference for - Possibly Singular - DSGE Models*, Working paper (2019).

[29] L. Khalaf and B. Peraza-Lopez, *Simultaneous Indirect Inference, Impulse Responses and ARMA models*, Econometrics **forthcoming** (2012).

[30] L. Khalaf and C. Saunders, *Monte Carlo Two-Stage Indirect Inference (2SIF) for Autoregressive Panels*, Journal of Econometrics **218** (2020), 419–434.

[31] S. Mavroeidis, M. Plagborg-Møller, and J.H. Stock, *Empirical Evidence on Inflation Expectations in the New Keynesian Phillips Curve*, Journal of Economic Literature **52** (2014), 124–188.

[32] D. Meenagh, P. Minford, M. Wickens, and Y. Xu, *Testing DSGE models by indirect inference: a survey of recent findings*, Open Economies Review **30** (2019), 593–620.

[33] M. Del Negro and F. Schorfheide, *Forming priors for DSGE models (and how it affects the assessment of nominal rigidities)*, Journal of Monetary Economics **55** (2008), no. 7, 1191–1208.

[34] M. Plagborg-Møller and C.K. Wolf, *Local Projections and VARs Estimate the Same Impulse Responses*, Econometrica **Forthcoming** (2020).

[35] A. Smith, *Estimating nonlinear time series models using simulated vector autoregressions*, Journal of Applied Econometrics **8** (1993), S63–S84.

Appendix

A Proof of the theoretical results

A.1 Proof of Theorem 1

We start by showing the convergence of the bootstrap p-values, by showing:

$$|P(\hat{p}_{TB}(Q_T(\theta_{0,T})|\theta_{0,T}) \leq \alpha) - P(\hat{p}_{TB}(Q_T(\theta_{0,T})|\theta_0) \leq \alpha)| \underset{T\to\infty}{\overset{P}{\to}} 0$$

Our proof borrows and extends parts of the proof of Theorem 1 in Bergamelli, Bianchi, Khalaf and Urga (2019). From Assumption 1(i), we have:

$$\begin{aligned}
& Q_T \equiv Q_T(\hat{g}_T(Y_T), \overline{\tilde{g}}_{T,H}(\theta_T)) \underset{T\to\infty}{\overset{P}{\to}} Q_0 \\
\Rightarrow\ & (Q_T, \theta_T) \underset{T\to\infty}{\overset{P}{\to}} (Q_0, \theta_0) \quad \text{from Assumption 1(iii) and } \theta_T \text{ deterministic} \\
\Rightarrow\ & \text{There exists a subsequence } (Q_{T_k}, \theta_{T_k})_k \text{ with } T \geq I_0 \text{ and } k = 1, 2, \cdots \text{ s.t.} \\
& (Q_{T_k}, \theta_{T_k}) \underset{k\to\infty}{\overset{P}{\to}} (Q_0, \theta_0) \qquad (10)
\end{aligned}$$

Consider now the following two events:

$$\begin{aligned}
A_0 &= \{\omega \in \Omega \ s.t.\ Q_T(\omega) \in D_0 \text{ and } Q_0(\omega) \in D_0 \text{ and } T \geq I_0\} \\
C_0 &= \{\omega \in \Omega \ s.t.\ \lim_{k\to\infty} Q_{T_k}(\omega) = Q_0(\omega) \text{ and } Q_0(\omega) \in D_0 \text{ and } \lim_{k\to\infty} \theta_{T_k} = \theta_0\}
\end{aligned}$$

From Assumption 1(ii), $P(A_0) = 1$. Together with (10), this implies that $P(C_0) = 1$. Now let $\eta > 0$.

- By Assumption 1(iii), for any $x \in D_0$, given $\theta_T \to \theta_0$, there exists $B(x, \eta)$ and $T(x, \eta)$

s.t. for any $T > T(x, \eta)$,

$$|Z_T(y|\theta_T) - Z_T(y|\theta_0)| \leq \eta \quad \forall y \in B(x, \eta) \cap D_0$$

- In addition, $\forall \omega \in C_0$, there exists k_0 s.t. $\forall k \geq k_0$

$$Q_{T_k}(\omega) \in B(Q_0(\omega), \eta) \cap D_0$$

- Overall, $\forall \omega \in C_0$, $\forall T_k > \max(T_{k_0}, T(Q_0(\omega), \eta))$

$$\begin{aligned}
& \left|Z_{T_k}(Q_{T_k}(\omega)|\theta_{0,T_k}) - Z_{T_k}(Q_{T_k(\omega)}|\theta_0)\right| \leq \eta \\
\Rightarrow \quad & \lim_{k\to\infty} |Z_{T_k}(Q_{T_k}(\omega)|\theta_{0,T_k}) - Z_{T_k}(Q_{T_k}(\omega)|\theta_0)| = 0 \quad \text{for } \omega \in C_0 \\
\Rightarrow \quad & \lim_{k\to\infty} |P(\hat{p}_{T_k B}(Q_{T_k}(\omega)|\theta_{0,T_k}) \leq \alpha) - P(\hat{p}_{T_k B}(Q_{T_k}(\omega)|\theta_0) \leq \alpha)| = 0 \quad \text{for } \omega \in C_0 \qquad (11)
\end{aligned}$$

which follows from:

$$\begin{aligned}
P(\hat{p}_{TB}(Q_T(\omega)|\theta_{0,T}) \leq \alpha) &= P(\sum_b 1(Q_{T,b} \geq Q_T(\omega)) \leq \alpha(B+1) - B) \\
&= \sum_{k=0}^{\lfloor \alpha(B+1)-B \rfloor} \left(\begin{array}{c} B \\ k \end{array} \right) P^k(Q_{T,b} \geq Q_T(\omega)) \left[1 - P(Q_{Tb} \geq Q_T(\omega))\right]^{B-k} \\
&= \sum_{k=0}^{\lfloor \alpha(B+1)-B \rfloor} \left(\begin{array}{c} B \\ k \end{array} \right) \left[1 - Z_T(Q_T(\omega)|\theta_{0,T})\right]^k \left[Z_T(Q_T(\omega)|\theta_{0,T})\right]^{B-k}
\end{aligned}$$

$$(11) \quad \Rightarrow \quad \lim_{k\to\infty} |P(\hat{p}_{T_k B}(Q_{T_k}(\omega)|\theta_{0,T_k}) \leq \alpha) - P(\hat{p}_{T_k B}(Q_{T_k}(\omega)|\theta_0) \leq \alpha)| = 0 \quad a.s.$$

This shows that any subsequence of

$$|P(\hat{p}_{TB}(Q_T(\omega)|\theta_{0,T}) \leq \alpha) - P(\hat{p}_{TB}(Q_T(\omega)|\theta_0) \leq \alpha)| \;, \quad T \geq I_0$$

contains a further subsequence which converges a.s. to zero. This implies the convergence of bootstrap p-values,

$$|P(\hat{p}_{TB}(Q_T(\omega)|\theta_{0,T}) \leq \alpha) - P(\hat{p}_{TB}(Q_T(\omega)|\theta_0) \leq \alpha)| \underset{T\to\infty}{\overset{P}{\to}} 0\,. \tag{12}$$

To conclude, we apply Proposition 2.4 in Dufour (2004) which states that

$$P(\hat{p}_{TB}(Q_T(\omega)|\theta_0) \leq \alpha) = \alpha\,,$$

and follows from the exchangeability of the observed and simulated statistics and the fact that $\alpha(B+1)$ is an integer. Substituting back into (12) proves the expected result,

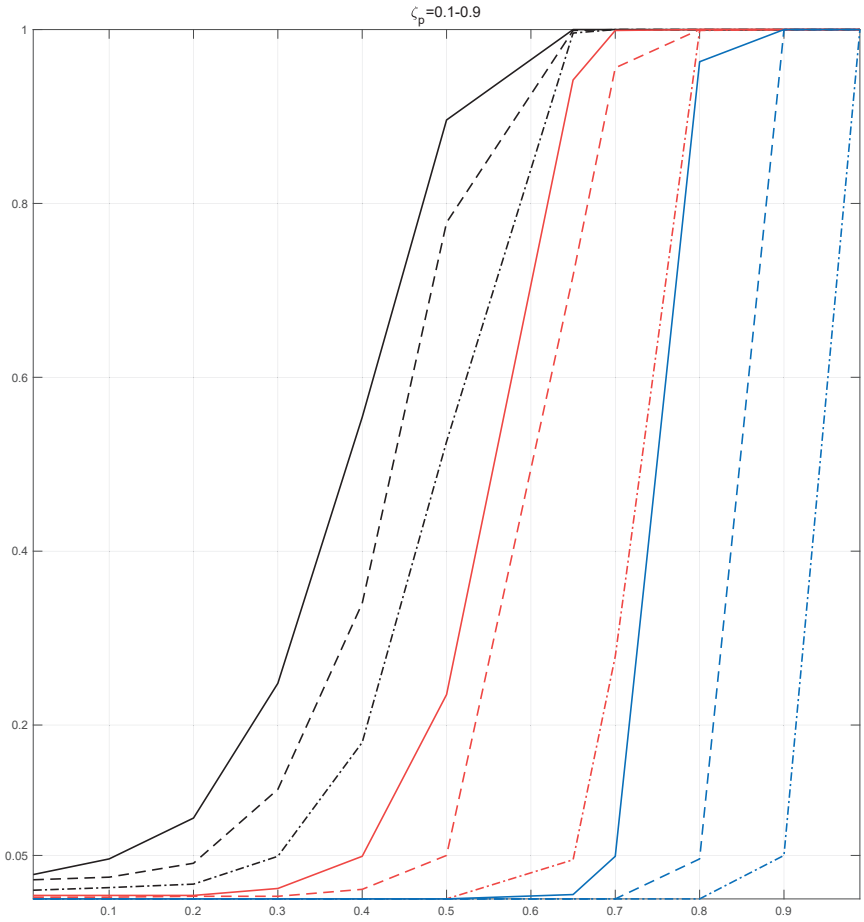
$$\lim_{T\to\infty} P(\hat{p}_{TB}(Q_T(\omega)|\theta_{0,T}) \leq \alpha) = \alpha\,.$$

B Tables and Figures of the simulation study

Table 1: Calibrated Parameter Values

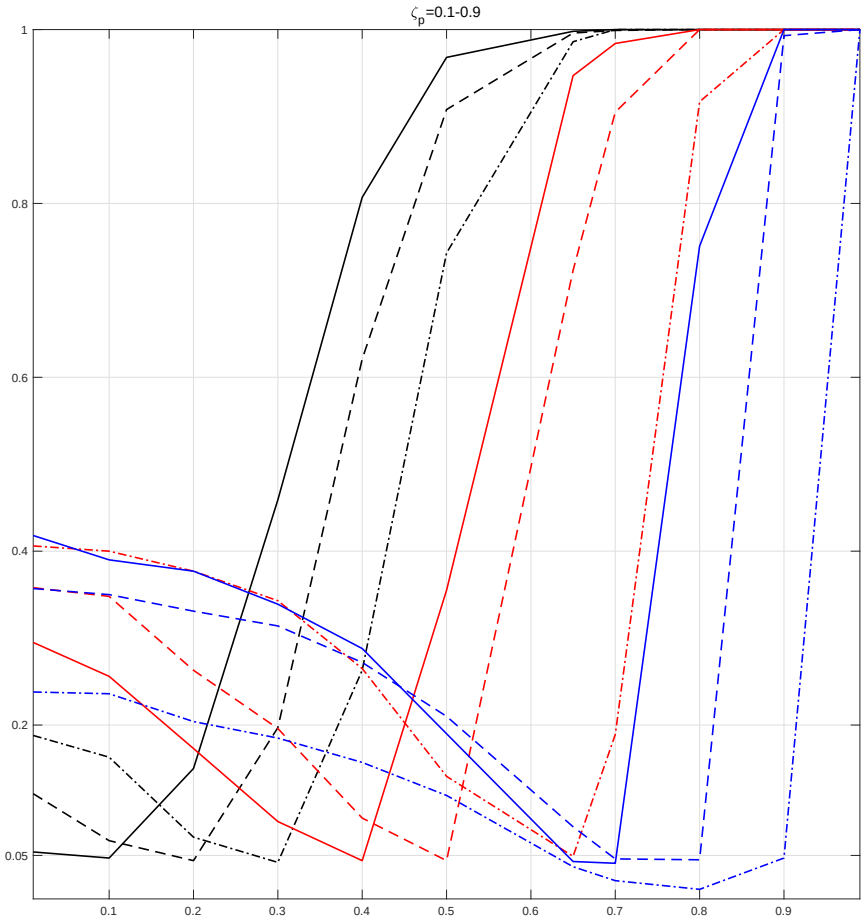
	Parameter	**Value**
β	stochastic discount rate	0.98
γ	growth rate of technology	1.005
λ	steady-state intermediate goods markup	0.15
π^*	steady-state inflation rate	1.005
ρ_z	autoregression parameter of the technology shock	0.13
ρ_λ	autoregression parameter of the price markup shock	0.88
ρ_ϕ	autoregression parameter of the shock that affects the preference for leisure	0.30
σ_z	standard deviation of the technology shock	1.50
σ_λ	standard deviation of the price markup shock	0.50
σ_ϕ	standard deviation of the shock that affects the preference for leisure	3.00
σ_R	standard deviation of the monetary policy shock	1.00

Figure 1: Experiment #1, Empirical rejection frequencies for testing $\mathcal{H}_0 : \zeta_p = c$ at the 5% significance level with the Bai-LPIRF inference procedure as a function of c. The remaining parameters are correctly calibrated; we match IRs at horizons 1 to 4, the sample size is 100 and there are 1,000 replications.



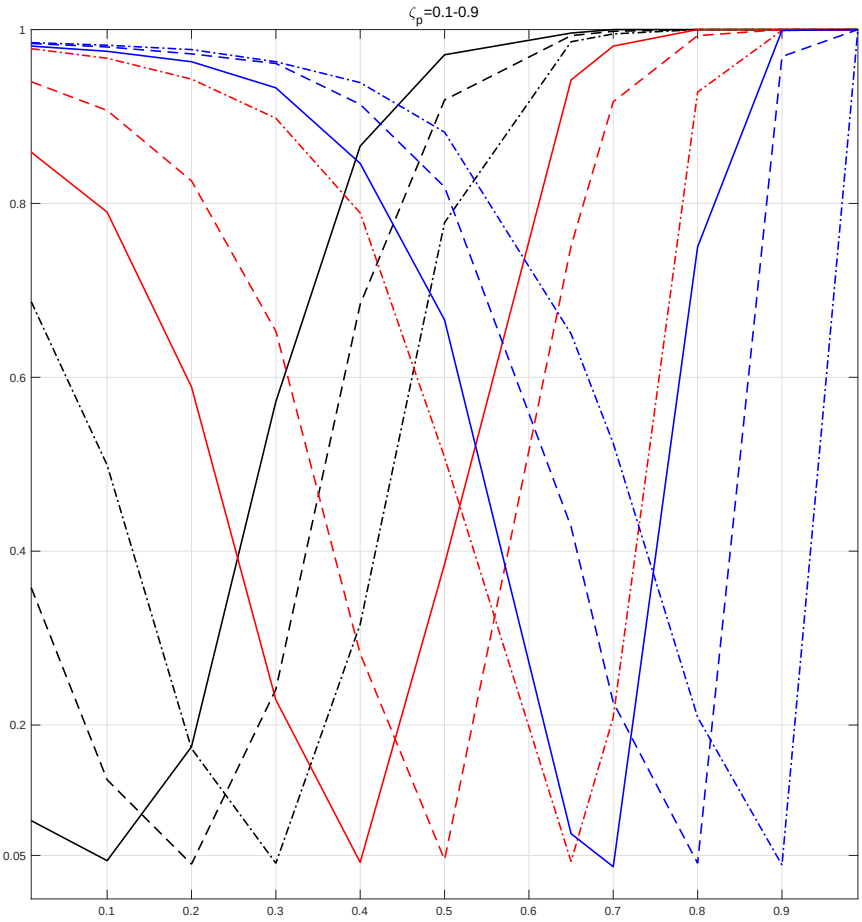
Note: Each curve corresponds to a different DGP obtained with the following values of $\zeta_{p,0}$: Solid Black, $\zeta_{p,0}$=0.1; Dashed Black, $\zeta_{p,0}$=0.2; Dash-dot Black, $\zeta_{p,0}$=0.3; Solid Red, $\zeta_{p,0}$=0.4; Dashed Red, $\zeta_{p,0}$=0.5; Dash-dot Red, $\zeta_{p,0}$=0.65; Solid Blue, $\zeta_{p,0}$=0.7; Dashed Blue, $\zeta_{p,0}$=0.8; Dash-dot Blue: $\zeta_{p,0}$=0.9.

Figure 2: Experiment #1, Empirical rejection frequencies for testing $\mathcal{H}_0 : \zeta_p = c$ at the 5% significance level with the Bai-LPIRF with MA correction inference procedure as a function of c. The remaining parameters are correctly calibrated; we match IRs at horizons 1 to 4, the sample size is 100 and there are 1,000 replications.



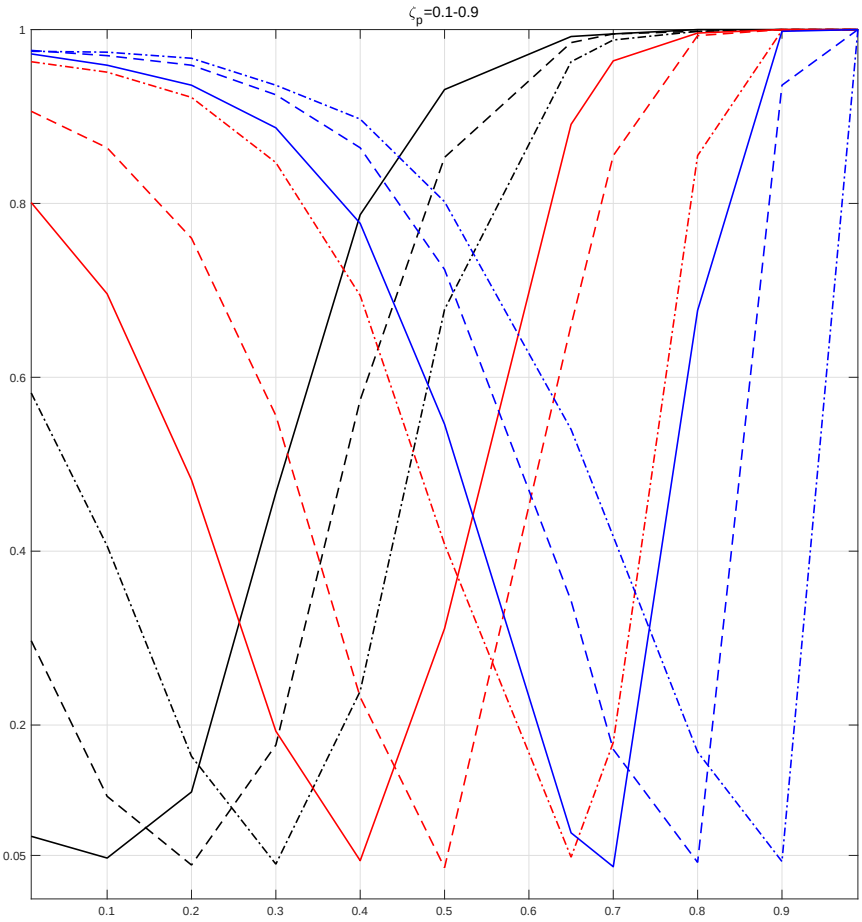
Note: Each curve corresponds to a different DGP obtained with the following values of $\zeta_{p,0}$: Solid Black, $\zeta_{p,0}$=0.1; Dashed Black, $\zeta_{p,0}$=0.2; Dash-dot Black, $\zeta_{p,0}$=0.3; Solid Red, $\zeta_{p,0}$=0.4; Dashed Red, $\zeta_{p,0}$=0.5; Dash-dot Red, $\zeta_{p,0}$=0.65; Solid Blue, $\zeta_{p,0}$=0.7; Dashed Blue, $\zeta_{p,0}$=0.8; Dash-dot Blue: $\zeta_{p,0}$=0.9.

Figure 3: Experiment #1, Empirical rejection frequencies for testing $\mathcal{H}_0 : \zeta_p = c$ at the 5% significance level with the Wald-SIRF inference procedure with parametric bootstrap weights as a function of c. The remaining parameters are correctly calibrated; we match IRs at horizons 1 to 4, the sample size is 100 and there are 1,000 replications.



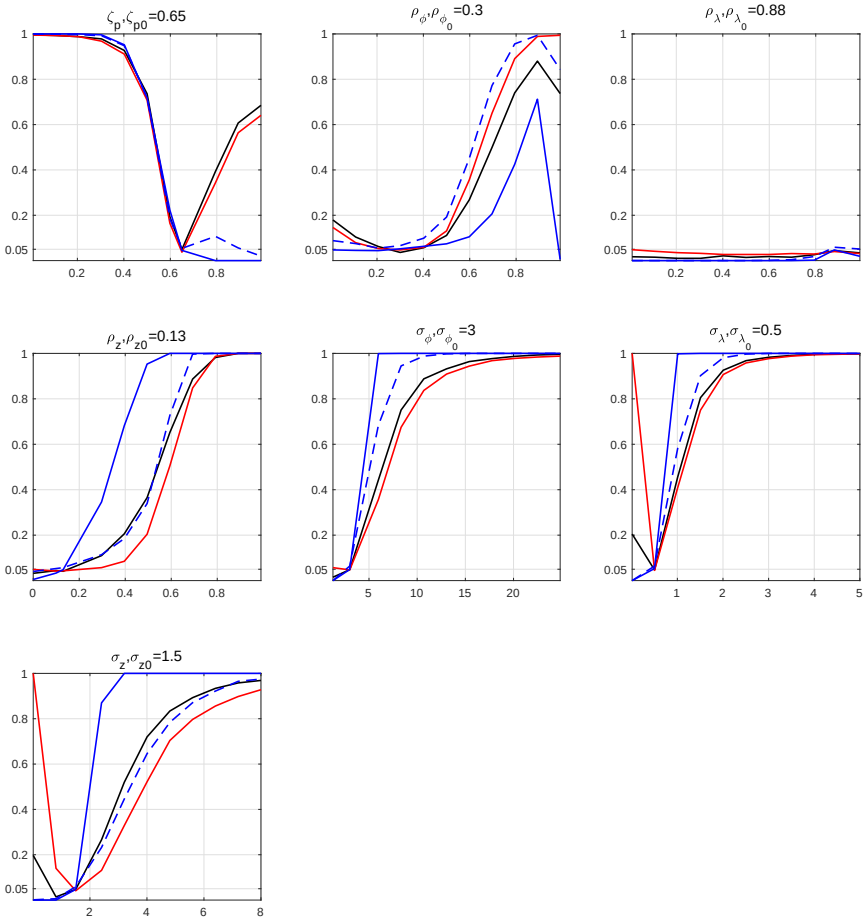
Note: Each curve corresponds to a different DGP obtained with the following values of $\zeta_{p,0}$: Solid Black, $\zeta_{p,0}$=0.1; Dashed Black, $\zeta_{p,0}$=0.2; Dash-dot Black, $\zeta_{p,0}$=0.3; Solid Red, $\zeta_{p,0}$=0.4; Dashed Red, $\zeta_{p,0}$=0.5; Dash-dot Red, $\zeta_{p,0}$=0.65; Solid Blue, $\zeta_{p,0}$=0.7; Dashed Blue, $\zeta_{p,0}$=0.8; Dash-dot Blue: $\zeta_{p,0}$=0.9.

Figure 4: Experiment #1, Empirical rejection frequencies for testing $\mathcal{H}_0 : \zeta_p = c$ at the 5% significance level with the Wald-RIRF inference procedure with parametric bootstrap weights as a function of c. The remaining parameters are correctly calibrated; we match IRs at horizons 1 to 4, the sample size is 100 and there are 1,000 replications.



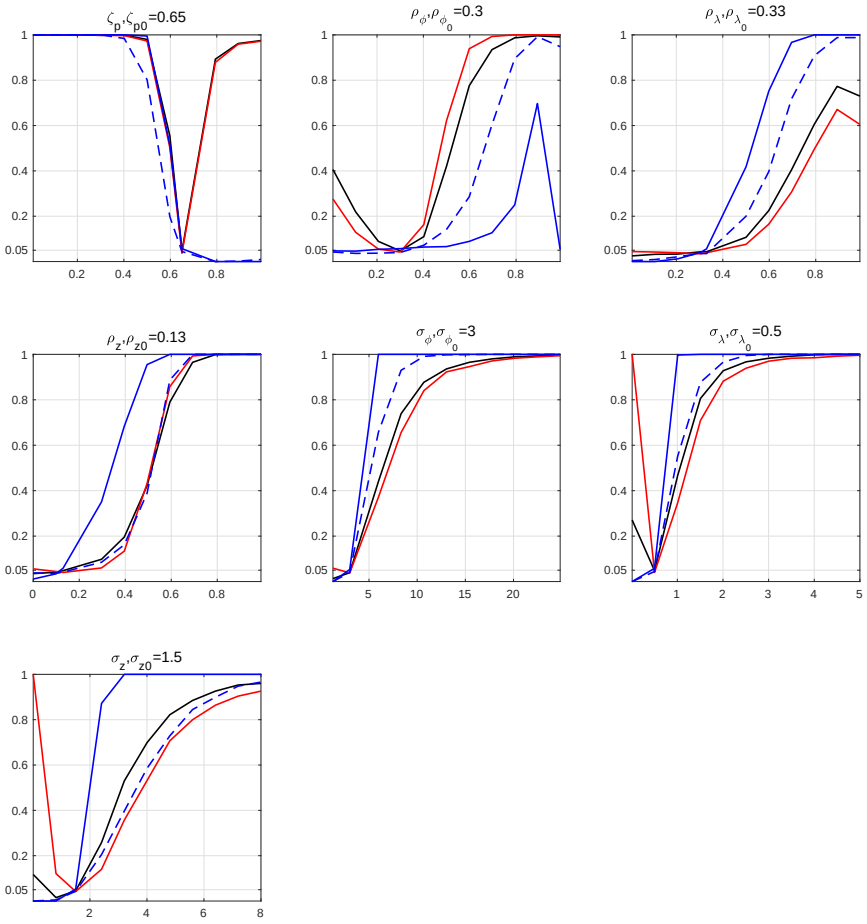
Note: Each curve corresponds to a different DGP obtained with the following values of $\zeta_{p,0}$: Solid Black, $\zeta_{p,0}$=0.1; Dashed Black, $\zeta_{p,0}$=0.2; Dash-dot Black, $\zeta_{p,0}$=0.3; Solid Red, $\zeta_{p,0}$=0.4; Dashed Red, $\zeta_{p,0}$=0.5; Dash-dot Red, $\zeta_{p,0}$=0.65; Solid Blue, $\zeta_{p,0}$=0.7; Dashed Blue, $\zeta_{p,0}$=0.8; Dash-dot Blue: $\zeta_{p,0}$=0.9.

Figure 5: Experiment #2, Empirical rejection frequencies for testing $\mathcal{H}_0 : \zeta_p = 0.65$ at the 5% significance level when the remaining parameters are calibrated according to Table 1. Each graph corresponds to one parameter deviating from its value set under the null: top left ζ_p; top middle ρ_ϕ; top right ρ_λ; middle left ρ_z; middle middle σ_ϕ; middle right σ_λ; bottom left σ_z. (we match IRs at horizons 1 to 4, the sample size is 100 and there are 1,000 replications).



Note: Each curve corresponds to a different inference procedure: Solid Black, Wald-SIRF with parametric bootstrap weights; Solid Red, Wald-RIRF parametric bootstrap weights; Solid Blue, Bai-LPIRF; Dashed Blue, Bai-LPIRF with MA correction.

Figure 6: Experiment #2, Rejection frequencies for testing $\mathcal{H}_0 : \zeta_p = 0.65$ at the 5% significance level when the remaining parameters are calibrated according to Table 1 except ρ_λ which is set at 0.33. Each graph corresponds to one parameter deviating from its value set under the null: top left ζ_p; top middle ρ_ϕ; top right ρ_λ; middle left ρ_z; middle middle σ_ϕ; middle right σ_λ; bottom left σ_z. (we match IRs at horizons 1 to 4, the sample size is 100 and there are 1,000 replications).



Note: Each curve corresponds to a different inference procedure: Solid Black, Wald-SIRF with parametric bootstrap weights; Solid Red, Wald-RIRF parametric bootstrap weights; Solid Blue, Bai-LPIRF; Dashed Blue, Bai-LPIRF with MA correction.

Table 2: Experiment #3, Rejection probabilities of testing $\mathcal{H}_0 : \zeta_p = c$ with approximate calibration of ρ_ϕ when the DGP is generated with $\zeta_p = 0.65$ and $\rho_\phi = 0.3$. The inference procedure is Bai-LPIRF and the remaining parameters are correctly calibrated.

Panel A: inference with Bai-LPIRF without MA correction									
Calibrated ρ_ϕ	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
$H_0 : \zeta_p = 0.40$	0	0	0	0	0	0	0	0	0
$H_0 : \zeta_p = 0.45$	0	0	0	0	0	0	0	0	0
$H_0 : \zeta_p = 0.50$	0	0	0	0	0	0	0	0	0
$H_0 : \zeta_p = 0.55$	0.001	0.001	0.001	0	0	0	0	0	0.001
$H_0 : \zeta_p = 0.60$	0.009	0.008	0.006	0.005	0.004	0.003	0.003	0.003	0.009
$H_0 : \zeta_p = 0.65$	0.062	0.058	0.053	0.0460	0.034	0.030	0.026	0.027	0.052
$H_0 : \zeta_p = 0.70$	0.380	0.358	0.331	0.294	0.274	0.250	0.232	0.237	0.311
$H_0 : \zeta_p = 0.75$	0.868	0.862	0.847	0.825	0.789	0.770	0.75	0.75	0.801
$H_0 : \zeta_p = 0.80$	0.999	0.999	0.999	0.999	0.999	0.998	0.996	0.997	0.998
$H_0 : \zeta_p = 0.85$	1	1	1	1	1	1	1	1	1
$H_0 : \zeta_p = 0.90$	1	1	1	1	1	1	1	1	1
Panel B: inference with Bai-LPIRF with MA correction									
Calibrated ρ_ϕ	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
$H_0 : \zeta_p = 0.40$	0.390	0.345	0.292	0.159	0.063	0.027	0.010	0	0
$H_0 : \zeta_p = 0.45$	0.367	0.319	0.256	0.121	0.082	0.021	0.005	0	0
$H_0 : \zeta_p = 0.50$	0.199	0.201	0.178	0.095	0.053	0.023	0.010	0	0
$H_0 : \zeta_p = 0.55$	0.098	0.095	0.099	0.065	0.064	0.019	0.008	0	0
$H_0 : \zeta_p = 0.60$	0.096	0.034	0.044	0.036	0.037	0.016	0.009	0	0
$H_0 : \zeta_p = 0.65$	0.176	0.094	0.043	0.023	0.032	0.043	0.011	0.003	0
$H_0 : \zeta_p = 0.70$	0.554	0.433	0.250	0.138	0.100	0.093	0.065	0.015	0
$H_0 : \zeta_p = 0.75$	0.847	0.742	0.594	0.489	0.348	0.342	0.258	0.159	0.052
$H_0 : \zeta_p = 0.80$	0.974	0.906	0.921	0.894	0.777	0.674	0.599	0.610	0.237
$H_0 : \zeta_p = 0.85$	0.998	0.998	0.998	0.999	0.997	0.980	0.966	0.960	0.923
$H_0 : \zeta_p = 0.90$	1	1	1	1	1	1	1	1	1

Table 3: Experiment #3, Rejection probabilities of testing $\mathcal{H}_0 : \zeta_p = c$ with approximate calibration of ρ_λ when the DGP is generated with $\zeta_p = 0.65$ and $\rho_\lambda = 0.3$ or 0.88. The inference procedure is Bai-LPIRF and the remaining parameters are correctly calibrated.

Panel A: the DGP is generated with $\rho_\lambda = 0.3$									
Calibrated ρ_λ	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
$H_0 : \zeta_p = 0.01$	0	0	0	0	0	0	0	0	0
$H_0 : \zeta_p = 0.1$	0	0	0	0	0	0	0	0	0
$H_0 : \zeta_p = 0.2$	0	0	0	0	0	0	0	0	0
$H_0 : \zeta_p = 0.3$	0	0	0	0	0	0	0	0	0
$H_0 : \zeta_p = 0.4$	0	0	0	0	0	0	0	0	0
$H_0 : \zeta_p = 0.5$	0	0	0	0	0	0	0	0	0
$H_0 : \zeta_p = 0.65$	0.403	0.18	0.042	0.008	0.001	0	0	0	0
$H_0 : \zeta_p = 0.7$	0.96	0.852	0.637	0.321	0.062	0.006	0	0	0
$H_0 : \zeta_p = 0.8$	1	1	1	1	1	0.995	0.936	0.527	0.016
$H_0 : \zeta_p = 0.9$	1	1	1	1	1	1	1	1	1
$H_0 : \zeta_p = 0.99$	1	1	1	1	1	1	1	1	1
Panel B: the DGP is generated with $\rho_\lambda = 0.88$									
Calibrated ρ_λ	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.88	0.95
$H_0 : \zeta_p = 0.01$	0	0	0	0	0	0	0	0	0
$H_0 : \zeta_p = 0.1$	0	0	0	0	0	0	0	0.001	0.002
$H_0 : \zeta_p = 0.2$	0.002	0.001	0.001	0.001	0	0	0	0	0
$H_0 : \zeta_p = 0.3$	0.024	0.01	0.005	0.003	0.003	0.001	0.001.	0.001	0.001
$H_0 : \zeta_p = 0.4$	0.352	0.228	0.114	0.049	0.016	0.008	0.004	0.001	0.001
$H_0 : \zeta_p = 0.5$	0.943	0.862	0.731	0.531	0.289	0.12	0.021	0.003	0.001
$H_0 : \zeta_p = 0.65$	1	1	1	1	0.999	0.974	0.802	0.05	0.009
$H_0 : \zeta_p = 0.7$	1	1	1	1	1	1	0.991	0.289	0.045
$H_0 : \zeta_p = 0.8$	1	1	1	1	1	1	1	0.998	0.811
$H_0 : \zeta_p = 0.9$	1	1	1	1	1	1	1	1	1
$H_0 : \zeta_p = 0.99$	1	1	1	1	1	1	1	1	1

Table 4: Experiment #3, Rejection probabilities of testing $\mathcal{H}_0 : \zeta_p = c$ with 2 candidates for the approximate calibration of ρ_λ (sup-p-value) when the DGP is generated with $\zeta_p = 0.65$ and $\rho_\lambda = 0.88$. The inference procedure is Bai-LPIRF and the remaining parameters are correctly calibrated.

Panel A: the sup P-value relies on $\rho_\lambda = 0.8$ **and** 0.9	
$H_0 : \zeta_p = 0.4$	0
$H_0 : \zeta_p = 0.5$	0
$H_0 : \zeta_p = 0.65$	0.01
$H_0 : \zeta_p = 0.7$	0.124
$H_0 : \zeta_p = 0.8$	0.991
$H_0 : \zeta_p = 0.9$	1
$H_0 : \zeta_p = 0.99$	1
Panel B: the sup P-value relies on $\rho_\lambda = 0.7$ **and** 0.95	
$H_0 : \zeta_p = 0.4$	0
$H_0 : \zeta_p = 0.5$	0
$H_0 : \zeta_p = 0.65$	0.008
$H_0 : \zeta_p = 0.7$	0.058
$H_0 : \zeta_p = 0.8$	0.82
$H_0 : \zeta_p = 0.9$	1
$H_0 : \zeta_p = 0.99$	1

Supplementary Appendix for: Identification-robust inference with simulation-based pseudo-matching

Bertille Antoine [*] Lynda Khalaf [†] Maral Kichian [‡]

Zhenjiang Lin [§]

October 22, 2020

In the supplementary appendix, we provide details on the implementation of the inference procedures used in the simulation study.

1 Wald-type inference on SIRF (or RIRF)

We start by postulating a reduced-form VAR model of order p to represent the dynamics of the vector of observables in the vector y_t. Here, y_t is ordered as follows: output, labor share, inflation, and interest rate; the lag order of the VAR model is 4. The model is given by:

$$y_t = \Phi_1 y_{t-1} + \Phi_2 y_{t-2} + \ldots + \Phi_p y_{t-p} + \Phi_0 + u_t, \qquad u_t \sim i.i.d.(0, \Sigma) \tag{1}$$

[*]B. Antoine (Corresponding author): Simon Fraser University, 8888 University Drive, Burnaby, BC V5A 1S6, CANADA. *Email:* Bertille_Antoine@sfu.ca

[†]L. Khalaf: Carleton University, *Email:* Lynda.Khalaf@carleton.ca

[‡]M. Kichian: University of Ottawa, *Email:* Maral.Kichian@uOttawa.ca

[§]Z. Lin: University of Nottingham Ningbo China *Email:* Zhenjiang.Lin@nottingham.edu.cn

Reduced-form impulse responses are then obtained from the infinite moving average representation of the model, assuming invertibility. The impulse response of the reduced-form shock $u_{j,t}$ on the variable $y_{i,t}$ at horizon h is defined as

$$RIRF(i,j,h) = \partial y_{i,t+h}/\partial u_{j,t} \tag{2}$$

and given by the appropriate coefficient of the MA representation. It is estimated using the residuals $\hat{u}_{j,t}$ obtained after estimating the MA model,

$$y_t = \Phi^{-1}(L)u_t \equiv \Theta(L)u_t, \qquad u_t \sim i.i.d.(0,\Sigma) \tag{3}$$

Structural impulse responses are obtained as follows. Assuming that the reduced-form errors u_t are linked to the structural model innovations ϵ_t via the equation $Pu_t = \epsilon_t$ with $P\Sigma P' = I$, a Choleski decomposition can be applied to the variance-covariance matrix Σ. The impulse response of the structural shock $\epsilon_{j,t}$ on the variable $y_{i,t}$ at horizon h is defined as

$$SIRF(i,j,h) = \partial y_{i,t+h}/\partial \epsilon_{j,t} \tag{4}$$

and given by the appropriate coefficient in the following model,

$$y_t = \Theta(L)P^{-1}Pu_t \equiv \Psi(L)\epsilon_t, \qquad \epsilon_t \sim i.i.d.(0,I) \tag{5}$$

Wald-type inference on SIRF (or RIRF) is then obtained by implementing the algorithm 2 described in section 3.1.

2 Bai-type inference on LPIRF

Local projection impulse responses are obtained from the following SURE form,

$$Y_h = Y_p B + U, \tag{6}$$

with the $(T, n^*(h+1))$-matrix $Y_h = [Y'_{h,1}, Y'_{h,2}, ..., Y'_{h,T}]'$ (where $Y_{h,t} = [Y'_t, Y'_{t+1}, ..., Y'_{t+h}]$), the $(T, n^*(p+1))$-matrix $Y_p = [Y'_{p,1}, Y'_{p,2}, ..., Y'_{p,T}]'$ (where $Y_{p,t} = [1, Y'_{t-1}, Y'_{t-2}, ..., Y'_{t-p}],$), the $(n^*(p+1), n^*(h+1))$-matrix B of coefficients, and U the $(T, n^*(h+1))$ error term which is obtained from stacking the following h regressions,

$$Y_{t+s} = c^s + Y_{t-1}B_1^{s+1} + Y_{t-2}B_2^{s+1} + ... + Y_{t-p}B_p^{s+1} + u_{t+s}^s, \ s = 0, 1, 2, ..., h \tag{7}$$

where c^s is an n^*-vector of constants, B_i^{s+1} are matrices of coefficients for each lag i and horizon $s+1$. And let $\hat{B}$ denote the least squares estimator of B.
The following algorithm describes how to obtain a "population counterpart" to $\hat{B}$; this is done for given $\theta = \theta_0$ (either from parameters fixed under $\mathcal{H}_0$ or calibrated).

- **Algorithm #1:** "population" counterpart of the auxiliary statistic.

1. For a given θ_0, generate $m = 1, ..., M$ simulated series $\tilde{Y}_{h,m}(\theta_0)$ and $\tilde{Y}_{p,m}(\theta_0)$ and random draws $\{\tilde{\varepsilon}_{t,m}\}_{t=1}^T$;

2. Calculate the equation-by-equation OLS estimates using each simulated dataset and collect them in $\tilde{B}_m(\theta_0)$. Averaging over the M paths yields

$$\bar{B}(\theta_0) \ = \ \frac{1}{M}\sum_{m=1}^{M} \tilde{B}_m(\theta_0).$$

The null distribution of the Bai statistic denoted $\Lambda(\theta_0)$, as well as the statistics intro-

duced above, can then be easily simulated, which justifies the application of Monte-Carlo test (MCT). When applied to the above $\Lambda(\theta_0)$ statistic, for example, the MCT technique can be summarized as follows.

- **Algorithm #2:** Monte-Carlo Test for Bai-LPIRF (without MA correction)

1. Implement the procedure described in Algorithm #1 to obtain $\bar{B}(\theta_0)$ given θ_0.

2. Compute the observed value of the Bai statistic,

$$\Lambda(\theta_0) = \log|\hat{\Sigma}_W^0| + \texttt{trace}(\hat{\Sigma}_W(\hat{\Sigma}_W^0)^{-1}) \tag{8}$$

where $\quad \hat{\Sigma}_W = \dfrac{1}{T-p-h}\displaystyle\sum_{t=p+1}^{T-h} W_t(Y)W_t'(Y) \quad \text{with} \quad W_t(Y) = Y_h - Y_p\hat{B}(\theta_0)$

$$\hat{\Sigma}_W^0 = \frac{1}{T-p-h}\sum_{t=p+1}^{T-h} W_t(Y,\theta_0)W_t'(Y,\theta_0) \quad \text{with} \quad W_t(Y,\theta_0) = Y_h - Y_p\bar{B}(\theta_0)$$

3. Draw N *i.i.d.* samples of size T from the model under $\theta = \theta_0$; these draws should be independent from those underlying $\bar{B}(.)$.

4. Using the same population measure $\bar{B}(\theta_0)$ and the above N simulated paths, obtain N simulated values of the Bai statistic, denoted $\Lambda_1(\theta_0)$, ... , $\Lambda_N(\theta_0)$.

5. Compute a simulated p-value for the test statistic, using the rank of the observed statistic, relative to its simulated counterpart:

$$p_N(\Lambda_0(\theta_0)) \;=\; \frac{NG_N(\Lambda_0(\theta_0)\,)+1}{N+1} \quad \text{with} \quad G_N(\Lambda_0(\theta_0)) = \frac{1}{N}\sum_{i=1}^{N}\mathbf{1}\left(\Lambda_i(\theta_0) \geq \Lambda_0(\theta_0)\right) \tag{9}$$

where $\mathbf{1}(C)$ is the indicator function associated with event C.

The null hypothesis is rejected at level $\alpha^\star$ by the above test if the MCT p-value so

obtained is less than or equal to $\alpha^{\star}$. The MCT critical region is:

$$p_N(\Lambda_0(\theta_0)) \leq \alpha^{\star}, 0 < \alpha^{\star} < 1,$$

which is exact for finite T and N, in the following sense.

The following algorithm describes how to obtain the Jorda's local projections based on (7) with an MA correction.

- **Algorithm #3:** Monte-Carlo Test for Bai-LPIRF with MA correction.

1. Obtain the least squares residuals $\hat{u}_{t+s}^{s}$ from (7) for $s = 0, 1, 2, ..., h-1$.

2. Add the $(s-1)$-th residuals as an independent variable in the s-th regression (7),

$$Y_{t+s} = c^s + Y_{t-1}B_1^{s+1} + Y_{t-2}B_2^{s+1} + ... + Y_{t-p}B_p^{s+1} + \hat{u}_{t+s-1}^{s-1}B_u^{s+1} + v_{t+s}^{s}$$

   and compute the associated least squares estimates $\hat{B}_u^{s+1}$ to obtain

$$Y_{t+s}^{\star} = Y_{t+s} - \hat{u}_{t+s-1}^{s-1}\hat{B}_u^{s+1} \quad \text{for } s = 1, 2, ..., h$$

3. Regress $Y_{t+s}^{\star}$ on $Y_{t-1}, Y_{t-2}, ..., Y_{t-p}$ for $s = 1, 2, ..., h$. Then stack these $s = 1, 2, ..., h$ regressions as in (6) and proceed according to Algorithm #2.