Business cycle measurement with some theory *

Fabio Canova
ICREA-UPF, CREI, CREMeD, CEPR; †

Matthias Paustian
Bank of England

First draft: November 2007;
This draft July 21, 2011

Abstract

A method to evaluate cyclical models not requiring knowledge of the DGP and the exact specification of the aggregate decision rules is proposed. We derive robust restrictions in a class of models; use some to identify structural shocks in the data and others to evaluate the class or contrast sub-models. The approach has good properties, even in small samples, and when the class of models is misspecified. The method is used to sort out the relevance of a certain friction (the presence of rule-of-thumb consumers) in a standard class of models.

JEL classification: E32, C32.

Keywords: Sign restrictions, shock identification, model validation, misspecification.

*We thank an anonymous referee, Heather Anderson, Luca Benati, Adrian Pagan, Frank Schorfheide for comments and the participants of seminars at CREI, ECB, EUI, Yale, University of Zurich, Bank of Hungary, Riksbank, Banque of France, Birbeck, the conferences "How much structure in empirical models" , Barcelona, and "Monetary policy in open economies", Sydney, the 6th Dynare conference for suggestions. This paper replaces an earlier manuscript by the first author circulated with the title "Validating Monetary DSGE's with VARs". Canova acknowledges the financial support of the Spanish Ministry of Education grant ECO2009-08556 (Specification and Estimation of Models for Policy Analysis and Forecasting) and of the Barcelona Graduate School of Economics. The views expressed in this paper are not necessarily those of the Bank of England.

†Corresponding author: Department of Economics, UPF, Ramon Trias Fargas 25-27, 08005, Barcelona (Spain), +3493521926, fabio.canova@upf.edu.

1 Introduction

Dynamic stochastic general equilibrium (DSGE) models are nowadays regarded as the benchmark business cycle models for policy analysis and forecasting, both in academic and policy institutions. Their popularity is due to their attractive theoretical aspects and to the good forecasting performance relative to single equation structural models or multiple equations time series specifications.

Existing business cycle models are, however, not problem free. Theoretically, many important features are modelled as black-box mechanisms and questions about their policy invariance have been raised (see e.g. Chari et al., 2009, or Chang et al., 2010); ad-hoc frictions are routinely added to match patterns found in the data, and crucial properties are derived without any reference to parameter or model uncertainty. Empirically, the problems are numerous and varied. Model misspecification is an important concern for classical estimation and generates numerical difficulties for Bayesian estimation. Identification problems make results difficult to interpret (see Canova and Sala, 2009, Iskrev, 2007, and Canova and Gambetti, 2010). The severe mismatch between theoretical and empirical concepts of business cycles (see Canova, 2009), on the other hand, renders structural estimation and policy conclusions generically whimsical. The empirical validation of business cycle models is also difficult: models impose fragile restrictions on the magnitude of interesting statistics and evaluation techniques for misspecified, hard to identify models are underdeveloped. With a few notable exceptions (Del Negro and Schorfheide, 2004, and, 2009), existing work relies on likelihood ratio statistics or marginal likelihood comparisons. Both approaches focus on statistical fit rather than fundamental economic differences, are sensitive to misspecification of aspects of the models not directly tested and computationally intensive.

This paper presents a methodology to validate classes of potentially misspecified business cycle models and to select sub-models in a class. The approach does not rely on statistical measures of fit and thus does not require estimation of often weakly identified structural parameters. Instead, it employs the flexibility of SVAR techniques against model misspecification, the insights of computational experiments (see e.g. Kydland

and Prescott, 1996) and pseudo-Bayesian predictive analysis (see e.g. Canova, 1995) to probabilistically evaluate the class, to discriminate among locally alternative data generating processes (DGP), and to provide information useful to respecify theoretical structures, if needed. Dedola and Neri (2007), Pappa (2009), Peersmann and Straub (2009), Lippi and Nobili (2010) among others, have used this methodology to answer interesting economic questions. What the paper provides is a formal presentation of the methodology, an assessment of its properties in simple experimental designs, and an application studying the role of rule-of-thumb consumers in generating realistic consumption responses to government expenditure shocks.

The analysis starts from a class of models which has an approximate state space representation once (log-)linearized around the steady state. We examine the dynamics of the endogenous variables in response to the disturbances for alternative members of the class using a variety of parameterizations and alternative specifications of non-essential (nuisance) aspects of the class. While magnitude restrictions depend on specification details, the sign of the responses is much more robust to parameter and specification uncertainty. A subset of theoretically robust restrictions is then used to identify structural disturbances in the data and the dynamic responses of unrestricted variables are employed to evaluate the discrepancy between the class and the data or to select a member within the class.

The methodology has a number of advantages. First, it allows for misspecification in the structure to affect the likelihood function as long as it leaves the sign of the responses used for identification and testing unchanged. Thus, it is applicable to a richer class of problems than existing methods. Second, it can be employed to validate classes of models featuring more endogenous variables than shocks or rudimentarily specified dynamics. Third, by focusing shock identification and model testing on robust model-based qualitative restrictions, the approach gives economic content to identification restrictions used in SVARs analyses and de-emphasizes the importance of a good calibration in testing the validity of a theory. Fourth, the procedure does not require optimization routines nor complex integration exercises and allows researchers to make identification and testing stronger or weaker depending on the needs of the analysis.

The approach can recover the sign of the impact response of unrestricted variables to the identified shocks, capture the qualitative features of the conditional dynamics, and exclude with high probability candidate DGPs in relevant designs. This occurs even when sample uncertainty exists, the empirical model is misspecified or the chosen class leaves important aspect of the DGP out. Finally, the approach can distinguish between sub-models in situations where standard approaches fail.

As an illustration, the methodology is used to gauge the frictions consistent with the observed transmission mechanism in the class of models with rule-of-thumb agents, suggested by Gali et al. (2007). The presence of a large number of non-optimizing consumers is insufficient to make consumption responses to government spending shocks positive. However, the robust restrictions the theory imposes can be employed to estimate the sign, the magnitude and the shape of consumption responses in the data. Since the share of non-optimizing agents needed to match the qualitative and quantitative features of conditional consumption dynamics in the data is unrealistically large, the validity of this class of models is called into question.

The rest of the paper is organized as follows. Section 2 illustrates the robust restrictions and the testable implications a class of models delivers. Section 3 describes the testing methodology. Section 4 studies the properties of the procedure. Section 5 evaluates a class of business cycle models. Section 6 concludes.

2 From the theory to the data

To illustrate the fundamental restrictions a theoretical structure imposes on the data and the nature of the testing exercise, the class of New-Keynesian models without capital, employed e.g. by Erceg et. al. (2000), Rabanal and Rubio-Ramirez (2005) among others, is considered.

The equilibrium conditions, with variables in log-deviations from the steady state, are in table 1.a. (T.1) is an Euler equation, (T.2) is a wage Phillips curve, (T.3) is a price Phillips curve, (T.4) is a Taylor rule, (T.5) defines the real wage and equation (T.6) is a production function. The economy is driven by four mutually uncorrelated, zero mean disturbances. The productivity shock e_t^z and the preference shock e_t^b have autocorrelation coefficients ρ_z and ρ_b, respectively. The monetary shock e_t^R and the markup shock e_t^μ are iid. The standard deviations of the innovations are $(\sigma_z, \sigma_b, \sigma_R, \sigma_\mu)$.

The goal is to derive restrictions which are robust to parameter variations, independent of the specification of nuisance features, and common to the sub-models in the class to identify shocks in the data and to test the validity of the class; and restrictions which are robust to parameter variations, independent of the specification of nuisance features but different across sub-models to select members of the class.

The structure represented in (T.1)-(T.6) is labeled M. The sub-models of interest are: a flexible price, sticky wage model ($\zeta_p = 0$) (labelled M1); a sticky price, flexible wage model ($\zeta_w = 0$) (labelled M2); a model with no indexation ($\mu_p = 0, \mu_w = 0$) (labelled M3); a model with infinitely elastic labor supply ($\sigma_l = 0$) (labelled M4). Nuisance features in the class are the specification of habit and of nominal rigidities. In the basic specification, habit is additive and Calvo lotteries are used. As an alternative, multiplicative habit (labelled N1) and quadratic adjustment costs to prices and wages (labelled N2) are considered.

To obtain robust restrictions, a uniform distribution over an interval is specified for each structural parameter, chosen to be large enough to include theoretically reasonable values - see third column of Table 1.b. For example, the interval for the risk aversion coefficient contains the values used in the calibration literature (typically 1 or 2) and the higher values employed in the asset pricing literature (see e.g. Bansal and Yaron, 2004), while the intervals for stickiness and indexation parameters include, roughly, the universe of possible values considered in the literature. While the interval for each parameter is independently and subjectively selected, in line with standard prior predictive analysis (see e.g. Geisser, 1980 or Kadane, 1980), one could make the ranges correlated and data based using the approach of Del Negro and Schorfheide (2008). The former approach is preferable here since it provides information about the range of possible outcomes the model can produce, prior to the use of any data. A large number of parameter vectors is drawn from these intervals, impulse responses are computed for each draw and pointwise 90 percent response intervals are extracted. 90 percent intervals trade-off two opposing forces: the desire to make the analysis as robust as possible (which would suggest choosing large intervals); the awareness that, if the class is misspecified, no restriction will hold with probability one (which would suggest choosing small intervals).

2.1 The restrictions

Figure 1 shows the range of dynamic outcomes for the nominal rate, the real wage, price inflation rate, output, and hours for model M in response to monetary shocks. The magnitude of the responses depends on the parameterization. The sign of several dynamic responses is also fragile: the zero line is often included in the 90 percent interval at medium and long horizons. The sign of impact responses is instead robust: the impact interval for the nominal rate is positive; those for output, inflation and hours are negative.

Are the signs of the impact response intervals independent of the specification of nuisance features? Are they maintained in sub-models of interest? Table 2 reports the signs of the impact intervals in the general model, in the four submodels of interest, and in each of the two alternative specifications of nuisance features; a '+' ('-') indicates robustly positive (negative) responses; a '?' non-robust responses.

Many impact responses have robust signs, both across sub-models and choices of nuisance features. For example, positive markup shocks increase production costs for any of the specifications and parameterizations, making production, the real wage and employment contract and inflation and the nominal rate increase. To test the validity of this class one could use, e.g., the restrictions that markup shocks produce on nominal rate, inflation, output and real wages to identify these disturbances in the data and then examine whether the hours impact response interval is negative, as theory predicts. How many restrictions are used to identify and how many to test is question dependent. More identification restrictions avoid shocks confusion (for example, if only restrictions on output and inflation are used, markup and technology shocks are indistinguishable). More restrictions at the testing stage make the validation exercise sharper.

The impact response of the real wage to monetary disturbances is of interest since the sign of the interval differs for sub-models in the class featuring alternative nominal frictions. In sub-model M1 (flexible prices and sticky wages), workers are off their labor supply schedule and from the firm's labor demand schedule, $w_t = -\frac{\alpha}{1-\alpha} y_t$, making real wages positively comove contemporaneously with monetary shocks. In sub-model M2 (sticky prices, flexible wages), workers are on their labor supply schedule and, on impact, $w_t = \left(\frac{\sigma_c}{1-h} + \frac{\sigma_l}{1-\alpha}\right) y_t$, so that real wages are instantaneously negatively related to monetary shocks. Thus, to contrast sticky wages vs. sticky prices in the data, one could identify monetary shocks using the robust restrictions that the theory imposes on all variables but real wages and then examine whether real wages instantaneously fall or increase. Clearly, for testing to be meaningful, real wages need to be correctly measured, but such a problem is not specific to the approach proposed here.

Distinguishing between sticky price and sticky wage models is difficult using unconditional measures of wage cyclicality because there are shocks which can instantaneously drive real wages up and down in each sub-model. Formal likelihood comparison may not be helpful either, because price and wage stickiness parameters may be only weakly identified (see Del Negro and Schorfheide, 2008 or Canova and Sala, 2009). The fundamental differences in the propagation mechanism emphasized here may help to resolve the issue.

The methodology can also be employed to select classes of models featuring alternative transmission properties. In this case, one would derive robust restrictions for each class; estimate partially identified VARs using common restrictions; and select a candidate using restrictions differing in the two classes.

3 The mechanics of the evaluation approach

The approach presumes that current business cycle models are still too stylized and feature too many black-box frictions to be taken seriously, even as an approximation to part of the DGP of the actual data (a point also made by Chari et al., 2009). This misspecification need not vanish by adding measurement errors or tagging artificial dynamics to the model, making standard measures of fit inadequate. By focusing on fundamental features of the propagation of shocks and using robust implications to distinguish alternatives, the methodology sidesteps potential misspecification problems. To formally describe the approach, let

$$F(w_t^s(\alpha_0(\theta), \alpha_1(\theta))|\epsilon_t, g, \mathcal{M}) \equiv F^s(\theta) \tag{1}$$

be a set of continuous model-based functions, computable conditional on the structural disturbances ϵ_t, using models in the class $\mathcal{M}$, featuring the nuisance aspects g. $F^s(\theta)$ could include impulse responses, conditional cross correlations, distributions of conditional turning points, etc., and depends on the model-produced series w_t^s via the coefficients of VAR representation of the decision rules, where $\alpha_0(\theta)$ is the matrix of contemporaneous coefficients, $\alpha_1(\theta)$ the matrix of lagged coefficients and θ the structural parameters. Let

$$F(w_t(\alpha_0, \alpha_1)|u_t) \equiv F(\alpha_0, \alpha_1) \tag{2}$$

be the corresponding set of data-based functions, conditional on the reduced form shocks u_t, where α_0, α_1 are the contemporaneous and lagged parameters of the VAR representation of the data. Both θ and α_0, α_1 are treated as random variables. As it will be clear, identification and sampling variability make α_0, α_1 random. The class $\mathcal{M}$ is assumed to be broad enough to include sub-models with interesting economic features. The nuisance features g are not of direct interest but may affect the time series properties of w_t^s. The class $\mathcal{M}$ is misspecified in the sense that even if there exists a θ_0 such that $\alpha_0 = \alpha_0(\theta_0)$ or $\alpha_1 = \alpha_1(\theta_0)$, $w_t^s(\theta_0) \neq w_t$. Thus, important aspects of the data (such as shocks, frictions or variables) may be left out of the class.

Among all possible $F^s(\theta)$ functions, attention is restricted to the subset $\tilde{F}^s(\theta)$ which are robust to parameter variations and to the specification of nuisance features: the $J_1 \times 1$ vector $\tilde{F}_1^s(\theta) \subset \tilde{F}^s(\theta)$ is used for shock identification and the $J_2 \times 1$ vector $\tilde{F}_2^s(\theta) \subset \tilde{F}^s(\theta)$ for evaluation purposes, $\tilde{F}_1^s(\theta) \neq \tilde{F}_2^s(\theta)$. $\tilde{F}^s(\theta)$ is termed robust if $sgn(F^s(\theta_1)) = sgn(F^s(\theta_2))$, $\forall\ \theta_1, \theta_2 \in [\theta_l, \theta_u]$, where sgn is the sign of F^s; θ_l, θ_u are the upper and lower range of economically reasonable parameter values and the above holds for all interesting specifications of g. $\tilde{F}_1^s(\theta)$ must hold for all $\mathcal{M}_j \in \mathcal{M}$, while depending on what it is tested, $\tilde{F}_2^s(\theta)$ may contain functions whose sign does not depend on the sub-model (if generic fit is evaluated) or depends on $\mathcal{M}_j$ (if sub-models are compared). The economic question dictates what $\tilde{F}_1^s(\theta)$ and $\tilde{F}_2^s(\theta)$ will be.

To compute $\tilde{F}^s(\theta)$, one can follow Canova (1995), draw θ from some prior distribution, solve the model, and store $F^s(\theta)$ at every draw. With the ordered output, one can then extract a credible interval and check if it is entirely on one side of zero or compute the probability that $\tilde{F}^s(\theta)$ is on one side of the zero line. To make sure that $\tilde{F}_1^s(\theta)$ holds in the data, the covariance matrix of the reduced form shocks Σ_u is rotated until $sgnF(w_{1t}^s(\alpha_0(\theta), \alpha_1(\theta))|\epsilon_t, g, \mathcal{M}) = sgnF(w_{1t}(\alpha_0, \alpha_1)|u_t)$ where $A_0 A_0' = \Sigma_u$, $\alpha_0 = A_0 H$, $HH' = I$, $\alpha_1 = \alpha_0^{-1} A_1$, where both A_1 and Σ_u are drawn their empirical based distribution, and w_{1t} is the subset of w_t over which restrictions are imposed. An algorithm to efficiently generate H is provided by Rubio et al. (2010). There maybe many, one or no H with the required characteristics. If no H exists, one can impose the restrictions on another subset of w_{1t}, if available, or use another set of $\tilde{F}_1^s(\theta)$. If all interesting options are exhausted and still no H is found, one can stop the evaluation process - the robust restrictions that the class of models impose have no counterpart in the data. When $k = 1, 2, \ldots, K$ H matrices are found, all the generated (α_0, α_1) are stored.

Model evaluation then consists in probabilistic statements concerning the features of $\tilde{F}_2(w_{2t}(\alpha_0, \alpha_1)|u_t)$. For example, one can compute the probability that $sgn\tilde{F}_2(w_{2t}(\alpha_0, \alpha_1)|u_t) - sgn\tilde{F}_2^s(w_{2t}^s(\alpha_0(\theta), \alpha_1(\theta))|\epsilon_t, g, \mathcal{M}) = 0$ and $w_{2t} \neq w_{1t}$ is a subset of w_t. Alternatively, one could compute the degree of overlap between the distribution of $\tilde{F}_2^s(\theta)$ and of $\tilde{F}_2(\alpha_0, \alpha_1)$, where the distributions are obtained using the random draws of θ and of (α_0, α_1) obtained in the previous steps. If only one H is available, A_1 and Σ_u are fixed at their sample point estimate, one useful summary statistics is the probability that $\tilde{F}_2^s(\theta) \leq \tilde{F}_2(\alpha_0, \alpha_1)$ where θ are drawn from $[\theta_l, \theta_u]$. Simple graphical devices, such as plots of the 90% bands in theory and in the data, could also give a good idea of the likelihood of the restrictions.

To select among candidates the probability that $sgn\tilde{F}_2(w_{2t}(\alpha_0, \alpha_1)|u_t) - sgn\tilde{F}_2^s(w_{2t}^s(\alpha_0(\theta), \alpha_1(\theta))|\epsilon_t, g, \mathcal{M}_j) = 0$ for each $\mathcal{M}_j$ could be constructed and the sub- model with the highest probability chosen. Alternatively, one could plot credible intervals for the sub-models of interest and take the one where the overlap with the theory is largest.

3.1 Discussion

The sign of the responses is used to derive robust constraints for two reasons: theory does not impose robust magnitude restrictions; and even if it did, magnitude restrictions need not hold in the data if the class of models is misspecified. Typically, impact restrictions are of interest, since as shown in section 2, the sign of the responses at longer horizons is generally not robust. When informational delays are present in theory, restrictions at longer horizons could be considered. Conditional functions, such as impulse responses, are preferred since they are more informative than unconditional moments about the features of $\mathcal{M}$.

The methodology is flexible and can be adapted to the need of the analysis. In fact, the identification process may involve more or less restrictions and one or more disturbances can be considered. Since standard rank and order conditions are not applicable to our case, how minimal this set of restrictions must be is generally unknown. Some indications on to proceed in practice are provided in the next section. Contrary to traditional practices, the identification restrictions are explicitly derived from a class of models and only robust constraints are considered. Thus, the procedure relies only on generic conditional dynamics and refrains from conditioning on a member of the class or on its parameterization.

The evaluation process is similar to the one employed in computational experiments where some moments are used to calibrate the structural parameters and others to check the goodness of the theory. Here a subset of the robust sign restrictions are employed to identify structural disturbances; the signs (and the shapes) of the dynamic responses of unrestricted variables are used to check the quality of the model's approximation to the data or to select a sub-model in the class. Two aspects are different: qualitative rather than quantitative restrictions are employed here at both stages; the evaluation process is probabilistic and takes into account both identification and sampling uncertainty.

Researchers are often concerned with the relative likelihood of sub-models in a class differing in terms of microfoundations, frictions, or functional forms. While the likelihood function need not be informative about these differences, our approach can, whenever sub-models differ in the sign (or the shape) of certain responses. For example, it is well known that sticky and flexible price versions of the same class of model produce different signs for the instantaneous response of hours to technology shocks. Once restrictions which are common to the two sub-models are used to identify technological disturbances, the response of hours can be used to discriminate the two theories. If sub-models differ in a number of implications, a weighted average of the relevant probabilities can be used to select the sub-model with the smaller discrepancy with the data. Candidate sub-models could be nested and or non-nested: the method works in both setups.

The approach compares favorably to existing methods for at least three reasons. First, the use of robust identification and testing restrictions shields researchers from model and parameter misspecification. Clearly, one cannot rule out the possibility that some type of misspecification changes the sign of key impulse responses; but qualitative restrictions on the sign of conditional moments tend to hold across many forms of misspecification. Second, the computational burden is smaller than the one involved in classical or Bayesian Likelihood-based evaluation techniques. Distributions of outcomes in theory are obtained when robust restrictions are sought; distributions of data outputs are obtained during the identification process and both require simple Monte Carlo exercises. Finally, the statistics one constructs can help to respecify the class, if the match with the data is unsatisfactory. For example, shape differences may suggest what type of amplification mechanism may be missing and sign differences the frictions that need to be introduced.

3.2 The relationship with the literature

The methodology is related to early work by Canova, Finn and Pagan, (1994) and Canova (1995), and to the recent strand of literature identifying VAR disturbances using sign restrictions (see Canova and De Nicolo', 2002, or Uhlig, 2005). It is also related to Del Negro and Schorfheide (2004) and (2009), who use the data generated by a cyclical model as a prior for reduced form VARs. Two differences set the approaches apart: the analysis here is conditional on a general class, rather than on a single model; qualitative rather than quantitative restrictions are used. This focus allows generic forms of model misspecification to be present and vastly extends the range of structures for which model evaluation becomes possible.

Corradi and Swanson (2007) developed a procedure to test misspecified models. Their approach is considerably more complicated, requires knowledge of the DGP and is not necessarily informative about the economic reasons for the discrepancy between the model and the data. Fukac and Pagan (2010) suggest to evaluate business cycle models using limited information methods but consider quantitative restrictions on single equations of the model while the focus here is on qualitative implications induced by certain disturbances. Finally, Chari, et. al. (2007) evaluate business cycle models using reduced form "wedges". Relative to their work, a structural conditional approach and probabilistic measures of fit for model comparison exercises are employed. The emphasis on model evaluation techniques which do not employ statistical measures of fit is also present in Kocherlakota (2007), who shows that when the available candidates are all misspecified the best fitting model is not necessarily the more accurate for policy and inferential exercises.

4 The evaluation procedure in controlled experiments

To examine the properties of the procedure in realistic settings, either the small scale class of models described in section 2, or the larger scale version employed by Smets and Wouters (2003) are used as experimental DGPs. The analysis proceeds in two steps: in the first the properties of the procedure are investigated in population; in the second sampling and specification uncertainty are added to the setup.

4.1 Population analysis

Starting with the class of section 2, the flexible price, sticky wage sub-model M1 is selected as the DGP. The parameters used in simulating "pseudo-actual" data are the fourth column of table 1.b and similar to the estimates of Rabanal and Rubio-Ramirez (2005). The researcher knows (T.1)-(T.6) and its solution, meaning that both the model dynamics A_1 and the covariance matrix of the reduced form errors Σ_u are known. We ask whether the responses of the real wage can be recovered with high probability employing different subsets of the robust restrictions, in alternative VAR systems, and identifying shocks either jointly or separately. The matrix of impact coefficients is obtained as follows: i) a large number of normal matrices with zero mean, unitary variance is drawn; ii) the QR decomposition is used to construct impact responses as $\alpha_0 = S * Q$, where $SS' = \Sigma$; iii) the responses satisfying the required restrictions are kept. To make results stable, draws are made until 10000 candidates satisfying the restrictions are found. Thus here, $\tilde{F}(\alpha_0, \alpha_1)$ reflects only identification but not sampling uncertainty.

4.1.1 Can we recover the true model?

In the baseline case, the empirical model includes 5 variables: the nominal rate, output, inflation, hours and the real wage. Since the economy features 4 structural shocks, a measurement error is attached to the law of motion of the real wage when simulating data. Disturbances are identified (a) jointly, using robust impact restrictions on all variables but the real wage; (b) jointly, using robust impact restrictions on all variables but hours and the real wage; (c) individually, the markup shock; (d) individually, the monetary shock. In (c) and (d), robust impact restrictions on all variables but the real wage are used. In addition to the basic DGP, setups where either the standard deviation of monetary shocks or the standard deviation of the markup shocks is 10 times larger are examined, and for each configuration, the four experiments are repeated. Table 3 reports the percentage of correctly signed impact real wage responses.

The procedure recognizes the qualitative features of the DGP with high probability, in the ideal conditions considered here. Two features of table 3 deserve attention. First, the number of shocks identified seems to matter in some cases. For instance, in a 5 variable VAR and when a large standard deviation for markup shocks is assumed, moving from identification scheme (d), which imposes restrictions only on responses to monetary shocks, to identification scheme (a), which restricts responses to four structural shocks, raises the fraction of correctly signed responses to monetary shocks by 3 percentage points. In general, the benefit from identifying additional shocks when the economic interest is only in one particular structural shock depends on the DGP and seems to be larger when the variability of the shocks is more heterogeneous.

Second, as in Paustian (2007), the relative strength of the shock signal matters. For instance, when the standard deviation of the monetary shock increases tenfold, the fraction of correctly identified real wage responses to monetary shocks rises from about 72% to about 90% under identification scheme (d). Conversely, if the relative strength of the monetary shock signal is reduced, by increasing the standard deviation of the markup shock tenfold, the fraction of correctly signed responses to monetary shocks falls from roughly 74% to roughly 52%, again under identification scheme (d). On the other hand, the real wage effects of markup and taste shocks are easy to measure because their signal is relatively strong, making conclusions largely independent of the number of restrictions used and the number of shocks identified.

Studies of the transmission of monetary shocks are abundant in the last 15 years and several researchers have used sign restrictions to identify these shocks in the data. Since such disturbances are likely to have relatively small variability, their transmission properties could be mismeasured, unless a sufficiently large number of restrictions is employed. In general, since the relative volatility of many structural shocks is unknown, being too agnostic in the identification process may have important costs for inference.

The same conclusions hold when hours is dropped from the VAR. A 4 variable VAR is fundamentally different from a 5 variable VAR since, in the latter, a state variable is missing - the observed real wage is a contaminated signal of the true one. Ravenna (2007) and Chari et. al. (2008) indicated that such an omission may be dangerous for inference if standard structural VARs are estimated. When robust sign restrictions on the impact response are used for identification, such an omission is less crucial.

4.1.2 Can we exclude alternative models?

As Table 2 shows, a sticky price, flexible wage sub-model (M2) and a flexible price, sticky wage sub-model (M1) are local to each other as far as the sign of impact responses is concerned. The procedure can recover the sign of the real wage response to monetary shocks well when M1 is the DGP. Would the answer change if M2 and the parameterization listed in the last column of table 1 characterizes the DGP? Can the sign of the impact responses of the real wage to monetary shocks uncover the correct DGP with high probability?

The answer is positive. In the three experiments considered (identifying all shocks using the impact restrictions on output, inflation, hours and the nominal rate; identifying monetary, taste and supply shocks using impact restrictions on output, inflation and the nominal rate; and identifying only monetary shocks) the percentage of incorrectly recognized cases ranges between 0.4 and 1.3 percent. Could this conclusion be due to the selection of the parameters of the DGP? To examine this possibility, two other experiments are considered. First, the standard deviation of either the monetary or the markup shock is increased by a factor of ten. The conclusions are broadly unchanged: the fraction of impact real wage responses to monetary shocks that is incorrectly signed never exceeds 8 percent. Second, the parameters are randomly and uniformly drawn from the intervals shown in table 1.b. - in this case, 200 parameter vectors are drawn, setting $\theta_w = 0$ for every draw, and for each vector, 10000 identification matrices are considered. When only monetary shocks are identified, the sign of the impact real wage response is incorrectly identified, on average, 3.21 percent of the times. Thus, the exact parameterization has little influence on the results.

Why is the procedure successful in both capturing the DGP and in excluding local sub-models as potential data generators? While the range of impact real wage responses to monetary shocks obtained randomizing the parameters of the DGP in M1 and M2 is relatively large, the degree of overlap of the distribution of responses is minimal. Thus, one can tell apart the two sub-models with high probability because theory has sharp and alternative implications for the real wage responses to monetary shocks. The answer would be different if the implications of different sub-models were more mudded. For example, the response of the real wage to technology shocks in M2 is not robust and the percentage of incorrect cases exceeds 25 percent under some identification configurations. Hence, only robust restrictions should be used for testing purposes.

These results are interesting also from a different perspective. Canova and Sala (2009) and Iskrev (2007) showed that classical econometric approaches have difficulties in separating sticky price and sticky wage models, because the distance function constructed using dynamic responses or the likelihood function are flat in the parameters controlling price and wage stickiness. Del Negro and Schorfheide (2008) report similar difficulties when Bayesian methods are used. The semi-parametric approach described here, which does not require structural parameter estimation, can potentially resolve the issue.

4.1.3 Summarizing the shape of the dynamic responses

So far the sign of the impact response of a variable left unrestricted in the identification process is used to test the propagation mechanism of a sub-model. For many purposes this restricted focus is sufficient: business cycle theories do not typically have robust implications for the magnitude or the persistence of the responses to shocks. At times, however, the shape of the dynamic responses may be of interest. Alternatively, one may want to extend the testing to multiple horizons (if robust restrictions exist) and ask, for example, whether there exists a location measure that reasonably approximates, say, certain conditional multipliers.

Figure 2 plots the median of the set of identified real wage responses to shocks, horizon by horizon, and the true real wage responses in the basic setup, case (a) of table 3. The median is a good measure of the impact response of real wages to all shocks, both in a qualitative and in a quantitative sense. It also captures the sign of the dynamics well, but it is an imperfect estimator of the magnitude of the conditional real wage dynamics, at least as far as the responses to monetary shocks are concerned. Relative to other location measures, it is slightly better than the average response and very similar to the trimmed mean (computed dropping the top and the bottom 25 percent of the responses).

Fry and Pagan (2007) criticized the practice of using the median of the distribution as a location measure when structural disturbances are identified with sign restrictions. Since the median at each horizon may be obtained from different candidate draws, identified shocks may be correlated. As an alternative, they suggest to use the single identification matrix that comes closest to producing the median impulse response for all variables. The correlation among identified shocks, computed using the median, ranges from 0.59 to 0.89 in absolute value depending on the experimental design. Therefore, Fry and Pagan's concern seems legitimate. However, as figure 2 shows, the alternative median is not a uniformly superior summary measure and its correlation with the true disturbances is generally low.

Several exercises were conducted to check the performance of the median in other experimental designs. The results suggest that (i) identifying more shocks or increasing the strength of the variance signal improves its performance; (ii) the dimensionality of the VAR is irrelevant for the dynamic properties of the median; and (iii) using model M1 or M2 as the DGP leaves the conclusions unchanged.

4.2 Does sampling uncertainty matter?

The ideal conditions considered so far are useful to understand the properties of the procedure but unlikely to hold in practice. What happens if the autoregressive parameters A_1 and the covariance matrix of the shocks Σ_u are estimated prior to the identification exercise?

To capture estimation uncertainty, 200 replications of each experiment previously run are considered. In each replication, data is simulated, keeping the parameters fixed, and drawing shocks (and measurement error) from iid normal distributions with zero mean and standard deviations, as reported in table 1.b. Samples with 80, 160 and 500 points are considered. For each replication, a BVAR is estimated with a close to non-informative conjugate Normal-Wishart prior . An arbitrary fixed lag length is chosen because it is typical to do so in practice even though it adds misspecification - the decision rules imply that a VAR(∞) should be used. What happens if the lag length is optimally selected with BIC is also considered. The joint posterior of the dynamic parameters A_1, the covariance matrix Σ_u, and the identification matrices H

is sampled until 2000 draws satisfying the restrictions are found for each replication. Table 4 reports the median value across replications of the probability that the impact response of the real wage to monetary shocks has the correct sign. Here the DGP is a sticky wage, flexible price model with one measurement error; a BVAR with the nominal rate, output, inflation, hours, and the real wage is estimated and shocks are identified imposing sign restrictions on the impact responses of the nominal rate, output, inflation and hours. Additional statistics for this experiment are in the accompanying materials (Appendix A) [1].

Three features of table 4 stand out. First, sample uncertainty is small relative to identification uncertainty (see Kilian and Murphy, 2009, for related evidence) and the recognition probabilities do not clearly increase with the sample size, for each lag length. Second, changing the lag length of the VAR has little consequences on the outcomes. Since the same patterns are present when the lag length of the VAR is selected with BIC, none of the problems highlighted by Chari, et al. (2008) appear to be present here. Third, the number of shocks which are identified has minor consequences on the quality of the outcomes.

All other conclusions obtained in population hold also here. For example, the number of variables included in the VAR has little effect on the conclusions, and changing the variability of shocks produces the same results found in population. The DGP can be recognized and local sub-models can be excluded with high probability by looking at the impact response of the real wage to monetary shocks. Finally, the performance of the median, as a summary measure for the true responses, is broadly unaffected.

4.3 Using the wrong model for inference

We have argued that misspecification is generically less of a problem for the approach. To show that this is indeed the case, the procedure is next applied to a class of models which leaves out important aspects of the true DGP. For that purpose, data is generated from a version of the Smets and Wouters (SW) (2003) class of models and used to test the validity of the restrictions imposed by the class of models of section 2. The smaller class has less shocks (investment specific, labor supply and government expenditure shocks

[1] Supplementary materials are available at JME in Science Direct.

are missing) than the SW class and the costs of adjusting investment and production frictions (fixed costs and variable capacity utilization) are disregarded. Since these differences are problematic for likelihood based methods, it is interesting to examine how large are the distortions that the approach would produce. The log-linearized optimality conditions, the parameter intervals used to derive robust restrictions and the parameters of the DGP are in the accompanying materials (Appendix B).

To begin with, it is useful check what robust restrictions the SW class imposes on output, inflation, the nominal rate, real wages and hours for each of the seven disturbances of the class. Table 5 reports the signs of the 90 percent impact response intervals. Interestingly, the sign of the intervals in responses to markup, monetary, taste and TFP disturbances are the same as in the basic model (compare with table 2) and robust across interesting sub-models. Thus, inference would not be necessarily distorted if a class of models which leaves out shocks and frictions present in the DGP is used to derive robust restrictions.

Table 5 also indicates that these restrictions alone may not be sufficient to uniquely obtain these four disturbances. In fact, in a five variable VAR, identified shocks may capture, in principle, any of the seven true structural shocks. For example, taste shocks could be confused with government expenditure shocks (four of the five signs are identical and for the fifth some confusion is possible), while markup and technology shocks may reflect investment specific shocks. To check the extent of the problem, the proportion of correctly signed real wage responses to shocks in population is computed. Some contamination is present, but it is remarkably small. For example, when markup, monetary, taste and technology shocks are identified using 16 impact restrictions, the probabilities of correctly signing the impact real wage response are 98.1, 98.7, 90.7 and 98.8, respectively. When only three shocks are identified using 12 impact restrictions, the probabilities are 98.6 for supply shocks, 99.5 for monetary shocks and 91.0 for taste shocks.

Since theory offers no guideline on the number of shocks to be included in a class of models, how can one limit shock confusion? Shrewdly choosing the variables of the VAR helps. As the last row of table 5 shows, if the labor productivity-real wage gap is added and the nominal rate is dropped from the list of observables, the seven shocks produce mutually exclusive patterns of signs on the contemporaneous responses of the variables of interest. Thus, shock confusion is unlikely even if the smaller class is used for inference.

4.4 Testing multiple restrictions

With the SW DGP one can also illustrate how the use of multiple restrictions - some of which may not be directly of interest - can strengthen testing in relevant practical situations. For the class considered, the instantaneous response of hours is robustly negative to TFP shocks if some price rigidities are present and robustly positive to labor supply, investment and markup shocks, regardless of the extent of price rigidities. The first implication is typically evaluated in the empirical literature, but hardly anyone seems to care about the other implications of the theory. However, jointly imposing the four restrictions may give sharper answers when price rigidities are weak, even if the latter restrictions are not of interest. To show this, data is simulated from the SW class using the same parameters as before except that $\zeta_p = 0.3$ and $\mu_p = 0$. The probabilities that the impact response of hours is negative in response to TFP shocks and that the impact response of hours is negative in response to TFP shocks and positive in response to investment, labor supply and markup shocks are then computed.

The former probability is 61 percent indicating that, when price stickiness is low, it is difficult to distinguish presence or absence of price rigidities. This probability increases to 83 percent when the four restrictions are jointly imposed - the difference is due to rotations matrices that imply negative hours responses to TFP shocks but also negative hours responses to any of the other shocks. Thus, when the data does not speak loud about the question of interest, imposing a larger set of restrictions can sharpen inference.

4.5 Advice to the users

The procedure has good properties in all the experiments. However, three ingredients are needed to give the methodology its best chance of success. First, it is important not to be too agnostic in the identification process. Sign restrictions are weak and this makes identification uncertainty important (see Manski and Nagy, 1998 for a similar result in micro settings). Thus, it is generally easier to recognize the DGP when more variables are restricted, for a given number of identified shocks, or more shocks are identified. Since theoretical sign restrictions at horizons larger than the impact one are often whimsical, constraints on the dynamic responses should be avoided at the identification stage. Similarly, sharper answers can be obtained if a number of robust restrictions, some which are of interest, some which are not, are jointly tested.

The experiments also showed that credible intervals tend to be large - this is expected given that the methodology delivers partially identified empirical models (see Moon and Schorfheide, 2009). Nevertheless, the probabilistic summary statistics employed are informative about the features of the DGP, even when asymptotically-based standard normal tests are not. If one insists on using the latter, a sufficient number of restrictions and smaller confidence intervals should be employed at the inferential stage.

Second, estimation biases should be, when possible, reduced since they may compound with identification uncertainty. In the experiments, estimation biases were small, even in small samples, but this need not to be the case for every possible design. A loose but informative prior was sufficient to reduce them. Other approaches, such as Kilian (1999), may work as well.

Third, inference is very reliable when the analysis focuses on the dynamics induced by shocks with stronger relative variance signal. However, even when the shock signal is weak, systematic mistakes are absent. While pathological examples can always be constructed (see Paustian, 2007, or Fry and Pagan, 2007), relative variance differences become a serious problem only in extreme circumstances. When interesting shocks are suspected to generate a weak relative signal, it is recommended to employ plenty of identification restrictions and to consider a class of models with a sufficiently rich shock structure. These two conditions were sufficient to ensure a good performance in all experiments we ran.

If a small scale class of models is used in the analysis, the choice of variables to be included in the VAR should be guided not only by economic but also by identification considerations. If the shocks produce mutually exclusive patterns of robust signs for the impulse responses of the selected variables in theory, it is unlikely that the identified shocks mix true shocks of different type, making aggregation issues (see e.g. Faust and Leeper, 1997) less important.

In theory disturbances often generate a unique pattern of impact responses for the endogenous variables. In practice responses are not restricted to satisfy this uniqueness condition. Thus, when a subset of the shocks is identified, it is possible that shocks disregarded in the analysis generate similar pattern of responses. This multiplicity has no reason to exist and may make inference weaker than it should. As shown in the accompanying materials (Appendix C), failure to impose the uniqueness condition in identification may lead researchers astray. Thus, unless all shocks are identified, the condition should always be imposed.

Finally, as section 4.3 has shown, misspecification of the class of models does not necessarily imply wrong inference. In addition, the class of models used to derive the restrictions need not have the same number of shocks as the empirical VAR. All that is required is that any shock omitted from the structural model, but potentially present in the data, is not isomorphic to the shocks of interest in terms of signs of impulse responses. Thus, there is no need to arbitrarily add ad-hoc shocks to the structural model to conduct inference and starting from a good fitting class is not a precondition for the methodology to be applied.

5 An example

Standard business cycle models find it difficult to reproduce the private consumption dynamics in response to government expenditure shocks generated by structural VARs (see e.g. Perotti, 2007). However, one should also be aware that the restrictions used in this literature are not explicitly derived from any theoretical specification that is used to interpret the results. Gali et al. (2007) have taken a standard New Keynesian model and argued that adding one particular friction (a portion of non-Ricardian consumers) can make the theory consistent with the VAR evidence. This section investigates three questions. First, does the Gali et al. class of models produce positive consumption responses to spending shocks with high probability?

Second, what do consumption responses in the data look like if robust theoretical sign restrictions are used to identify government spending shocks? Third, what is the likelihood that this class has generated the data?

5.1 The class of models

The log-linearized optimality conditions are in Table 6.a. Equations (T.7)-(T.8) describe the dynamics of Tobin's q, its relationship with investments i_t. The law of motion of capital is in equation (T.9). Equation (T.10) is the Euler equation of optimizing agents. Consumption of the non-Ricardian agents, c_t^r, depends on their labor income obtained from supplying n_t^r hours at wage w_t, net of paying taxes t_t^r, where α is the share of labor in production, as in equation (T.11). The labor supply schedule for each group is in equation (T.12). Cost minimization implies (T.13) and (T.14), where mc_t is real marginal cost, e_t^z a total factor productivity shock and r_t the rental rate of capital. Output is produced as in (T.15). (T.16) indicates that output is absorbed by aggregate consumption c_t, investment i_t and government spending e_t^g, which is random. The new Keynesian Phillips curve is in equation (T.17), where e_t^u is an iid markup shock, μ_p parameterizes the degree of indexation, $\kappa_p = \frac{(1-\beta\zeta_p)(1-\zeta_p)}{\zeta_p}$, and ζ_p is the Calvo probability of non-changing prices. The monetary policy rule is in equation (T.18) and e_t^R a monetary policy shock. The government budget constraint and the fiscal rule give equation (T.19), where b_t are real bonds. The fiscal rule is in (T.20). In the aggregate, $c_t = \lambda c_t^r + (1-\lambda)c_t^o$, $n_t = \lambda n_t^r + (1-\lambda)n_t^o$, $t_t = \lambda t_t^r + (1-\lambda)t_t^o$, λ is the share of non-Ricardian agents (ROTC), and $t_t^j = \frac{T_t^j - T^j}{Y}$, $j = o, r$.

5.2 Evaluating the friction in theory

The literature often presumes that this class of models produces instantaneously positive consumption responses to government spending shocks when the share of ROTCs is sufficiently large. Is this a robust implication of the theory? To check this, parameters values are drawn uniformly from the intervals in the third column of Table 6.b, except for λ which is fixed at different values on a grid. The first panel of Figure 3, which reports the percentage of draws in which instantaneous consumption responses to government spending shocks are negative for different λ, shows that the percentage increases with the share of ROTC but a large λ is insufficient to robustly produce the desired result. In fact, even when the majority of the consumers are not optimizers there is a non-negligible probability that reasonable parameters configurations induce instantaneous negative consumption responses. The first panel of figure 3 also shows that if a large share of ROTC is combined with large price stickiness, the required result obtains. Thus, while a large value of λ is necessary, it is by no means sufficient. It is only when both λ and ζ_p exceed 0.8 that one can confidently conclude (say, with at least 90 percent probability) that this class has the required feature.

5.3 Deriving robust identification restrictions

Structural parameters are drawn from the intervals presented in the third column of Table 6.b, setting $\beta = 0.99$, endogenously calculating c_y, i_y using steady state conditions, and keeping only those draws producing a determinate rational expectations equilibrium - indeterminacy may occur for certain combinations of λ and ζ_p. The range for most of the parameters is the same as in the experiments of section 4. For the fiscal parameters, large intervals centered around the values used in the literature are selected.

Table 7 presents the sign of the 90 percent impact response intervals of output growth, inflation, hours growth, investment growth to the four shocks. The combination of signs these intervals display is sufficient to mutually distinguish all disturbances. This would not be the case, for example, if the nominal interest rate is used in place of inflation (markup and monetary policy shocks will have similar sign implications). Interestingly, 15 of the 16 sign restrictions displayed in the table remain if a positive correlation in the intervals for (γ_π, γ_y), for (μ_p, ζ_p) and for (ϕ_b, ϕ_g) is allowed. Only the response of inflation to expenditure shocks is signed with less precision (around 65 percent) when γ_π and γ_y are sufficiently positively correlated. Thus, having uncorrelated or correlated intervals makes little difference for the restrictions one derives.

Prior to the testing exercise, it is useful to check in a controlled experimental design whether the approach can distinguish situations with and without non-Ricardian consumers using the restrictions of Table 7. The simulation uses the parameter values presented in the last column of Table 6.b (which are the same as in Gali et al., 2007). It is assumed that the researcher observes data on output growth, inflation, hours growth, investment growth and consumption growth and that the population VAR representation of these variables is known. For illustration, two polar cases are considered: no ROTC, $\lambda = 0$; a large portion of ROTC $\lambda = 0.8$. In both cases, ζ_p is set to 0.75 to make the practical distinction between the two setups empirically relevant. Do the restrictions present in Table 7 allow us to sign the impact consumption growth response to government spending shocks with high probability? Do the dynamic responses of consumption growth in the VAR and in theory look similar? It turns out that in 99.6 percent of the accepted draws consumption falls on impact when $\lambda = 0$ and in 78.2 percent of the accepted draws consumption increase on impact when $\lambda = 0.8$. Furthermore, the median response path of consumption growth tracks the true response almost perfectly in both cases (see second panel of figure 3). Hence, the method can detect both the sign of the impact consumption responses and the shape of its dynamic responses to spending shocks, if the class of models has generated the data and if model-based restrictions are employed to identify spending shocks.

5.4 Is the friction relevant?

A BVAR with a loose Normal Inverted-Wishart prior is estimated using quarterly U.S. data from 1954:1 to 2007:2 obtained from the FRED database. The lag length of the VAR is two as selected by BIC. The BVAR includes, together with government consumption expenditure, output growth, GDP inflation, the growth rate of hours worked in the nonfarm business sector, and the growth rates of private investment and of private consumption. Four shocks are identified, imposing the 16 impact restrictions appearing in Table 7. The joint posterior of the BVAR parameters and orthonormal matrices is sampled until 1000 draws satisfying the restrictions are found. Data based error bands thus reflect sampling and identification uncertainty

The third panel of Figure 3 presents the responses of consumption growth to government spending shocks in the data. When model based robust restrictions are imposed, consumption growth instantaneously increases. The point estimate is 0.25 and it is statistically significant but there is considerable uncertainty concerning the magnitude of the instantaneous consumption multiplier to spending shock (it could be anywhere between 0.06 and 0.45). Thus, the instantaneous consumption responses to spending shocks are comparable to those found in the micro literature for tax shocks (see e.g. Broda and Parker, 2008) Moreover, the increase is very short lived and after one quarter the 68 percent band includes zero.

Is the class of models a good candidate to explain the consumption responses observed in the data? To answer this question, the third panel of Figure 3 superimposes the theoretical consumption responses obtained when $\lambda = 0.8$ and $\zeta_p = 0.75$ while allowing all other parameters to be random. Clearly, the profile of the distribution of the responses in theory and in the data is similar. Instantaneously, the median responses are very close. At short horizons the median of the two distributions have similar size and shape and the probability that the sign of the responses in theory and in the data is the same is 83 percent on impact and 72 percent over 2 horizons. Thus, to match the sign and the shape of the consumption responses observed in the data, considerable price stickiness and an unrealistically large share of ROTC are needed. Since micro evidence suggests moderate price stickiness, these results call into serious question the use of this class for inference and policy analyses [2].

6 Summary and conclusions

A new methodology to examine the validity of business cycle models and to discriminate sub-models is presented. The approach employs the flexibility of SVAR techniques against model misspecification, the insights of computational experiments, and pseudo-Bayesian predictive analysis to link models to the data. Probabilistic measures of fit, which are robust to misspecification of the class and effective in providing information useful to respecify the class, are used to evaluate the discrepancy of the theory.

[2] As noted by Gali et. al., a model with imperfectly competitive labor markets may help to lower the share of rule of thumb consumers required to generate a rise in consumption to spending shocks.

The starting point of the analysis is a class of models which has an approximate state space representation once (log-)linearized around their steady states. The dynamics in response to shocks for alternative members of the class are examined using a variety of parameterizations and for different specifications of nuisance features. A subset of the robust restrictions is used to identify structural disturbances; another subset is used to measure the discrepancy between the class and the data or to discriminate members of the class. In controlled experiments, the approach can recognize the qualitative features of DGP with high probability and can tell apart local sub-models. It also provides a good handle of the quantitative features of the DGP if identification restrictions are abundant and if the relative variance signal of the shock(s) one wishes to identify is sufficiently strong. The methodology is successful even when the VAR is misspecified relative to the aggregate decision rules and when sampling uncertainty is present.

The methodology is appealing in several respects. First, it can be used even when the true DGP is not a member of the class of models one considers as long as the restrictions employed for identification and testing are not affected by the misspecification. Second, it does not require the probabilistic structure to be fully specified to be operative. Third, it shields researchers against omitted variable biases and representation problems. Fourth, it can be adapted to the needs of the user and requires limited computer time.

Apart from the illustrative example of section 5, recent work by Dedola and Neri (2007), Pappa (2009) Peersmann and Straub (2009) Lippi and Nobili (2010) among others, indicate the potentials that the methodology possesses, the type of information it provides, and the interaction between theory and empirical work it produces. One interesting extension worth pursuing is transforming the evaluation approach into an estimation procedure, where the initial ranges for the parameters are updated using information similar to the one presented in Section 5. This approach, which provides an indirect way for obtaining interval estimates of the parameters, could become a useful alternative to likelihood based estimation approaches when the objective function is flat in the parameters of interest.

References

Bansal, R., Yaron, A., 2004. Risks for the Long Run: A Potential Resolution of Asset Pricing Puzzles. Journal of Finance LIX, 1481-1509.

Broda, C., Parker, J., 2008. The impact of the 2008 Tax Rebates on consumer spending: preliminary evidence, manuscript.

Canova, F., 1995. Sensitivity Analysis and Model Evaluation in Simulated Dynamic General Equilibrium Economies. International Economic Review 36, 477-501.

Canova, F., 2009. Bridging cyclical DSGE models and the data, manuscript.

Canova, F., De Nicoló, G., 2002. Money Matters for Business Cycle Fluctuations in the G7. Journal of Monetary Economics 49, 1131-1159.

Canova, F., Gambetti, L., 2010. Do expectations matter? The Great Moderation revisited. American Economic Journals: Macroeconomics 2, 185-205.

Canova, F., Sala, L., 2009. Back to square one: identification issues in DSGE models. Journal of Monetary Economics 56, 431-449.

Canova, F., Finn, M. and Pagan, A., 1994. Evaluating a RBC model. In Hargraves, C., (Ed.) Non Stationary Time Series Analysis and Cointegration. Oxford University Press, Oxford.

Chang,Y., Kim, S., Schorfheide, F., 2010. Financial Frictions, Aggregation and Lucas critique, NBER working paper 16401.

Chari, V.V., Kehoe, P., McGrattan, E., 2007. Business cycle accounting. Econometrica 75, 781-836.

Chari, V.V., Kehoe, P., McGrattan, E., 2008. Are structural VAR with long run restrictions useful for developing business cycle theory?. Journal of Monetary Economics 55, 1337-1352.

Chari, V.V., Kehoe, P., McGrattan, E., 2009. New Keynesian Models: Not yet useful for policy analyses. American Economic Journals: Macroeconomics 1, 781-836.

Corradi, V., Swanson, N., 2007. Evaluation of Dynamic Stochastic General Equilibrium Models Based on Distributional Comparison of Simulated and Historical Data. Journal of Econometrics 136, 699-723.

Dedola, L., Neri, S., 2007. What does a technology shock do? A VAR analysis with model-based sign restrictions. Journal of Monetary Economics 54, 512-549.

Del Negro, M., Schorfheide, F., 2004. Priors from General Equilibrium Models for VARs. International Economic Review 95, 643-673.

Del Negro, M., Schorfheide, F., 2008. Forming Priors for DSGE models (and how it affects the assessment of nominal rigidities . Journal of Monetary Economics 55, 1191-1208.

Del Negro, M., Schorfheide, F., 2009. Monetary Policy Analysis with Potentially Misspecified models. American Economic Review 99, 1415-1450.

Erceg, C., Henderson, D., Levin, A., 2000. Optimal Monetary Policy with Staggered Wage and Price Contracts. Journal of Monetary Economics 46, 281–313.

Faust, J., Leeper, E., 1997. Do Long Run Restrictions Really Identify Anything?. Journal of Business and Economic Statistics 15, 345-353.

Fry, R., Pagan, A., 2007. Some Issues in Using Sign Restrictions for Identifying Structural VARs. NCER Working Paper No 14.

Fukac, M., Pagan, A., 2010. Limited Information Estimation and Evaluation of DSGE models. Journal of Applied Econometrics 25, 55-70.

Gali, J., López-Salido, J. D., Vallés, J., 2007. Understanding the Effects of Government Spending on Consumption. Journal of the European Economic Association 5, 227-270.

Geisser, S., 1980. A Predictivist Primer. In Zellner,A. (Ed.), Bayesian Analysis in Econometrics and Statistics. North-Holland, Amsterdam, 363-381.

Kydland, F., Prescott, E., 1996. The Computational Experiment. Journal of Economic Perspectives 10, 69-85.

Kadane, J. B., 1980. Predictive and Structural Methods for Eliciting Prior Distributions. In Zellner, A.

(Ed.) Bayesian Analysis in Econometrics and Statistics. North Holland, Amsterdam, 89-94.

Kilian, L., Murphy, D., 2009. Why agnostic sign restrictions are not enough: understanding the dynamics of oil markets VAR models. Forthcoming, Journal of the European Economic Association.

Kocherlakota, N., 2007. Model fit and model selection. Federal Reserve Bank of St. Louis Review 89, 439-450.

Iskrev, N., 2007. How much do we learn from the estimation of DSGE models? A case study of Identification Issues in DSGE models, manuscript.

Lippi, F., Nobili, A., 2010. Oil and the macroeconomy: A structural VAR analysis with sign restrictions. Forthcoming, Journal of the European Economic Association.

Manski, C., Nagy, D., 1998. Bounding Disagreement about Treatment Effects: A case study of Sentencing and Recidivism. Sociological Methodology 28, 99-137.

Moon, K., Schorfheide, F., 2009. A Bayesian Look at Partially Identified Models, NBER working paper 14882.

Pappa, E., 2009. The Effects of Fiscal Shocks on Employment and Real Wages. International Economic Review 50, 217-244.

Paustian, M., 2007. Assessing Sign Restrictions. B.E. Journals of Macroeconomics, Topics, 7(1), article 23.

Perotti, R., 2007. In Search of a Transmission Mechanism of Fiscal Policy. NBER Macroeconomic Annual 2007, 169-226.

Peersmann, G., R. Straub, 2009. Technology shocks and robust sign restrictions in a Euro area SVAR. International Economic Review 50, 727-750.

Rabanal, P., Rubio-Ramirez, J., 2005. Comparing New Keynesian models of the business cycle: A Bayesian approach. Journal of Monetary Economics 52, 1151-1166.

Ravenna, F., 2007. Vector Autoregressions and Reduced Form Representations of Dynamic Stochastic

General Equilibrium Models. Journal of Monetary Economics 54, 2048-2064.

Rubio-Ramirez, J., Waggoner, D., Zha, T. , 2010. Structural Vector Autoregressions: Theory of Identification and Algorithms for Inference. Review of Economic Studies 77, 665-696.

Smets, F., Wouters, R., 2003. An estimated Dynamic Stochastic General Equilibrium models of the Euro area. Journal of the European Economic Association I, 1123-1175.

Uhlig, H., 2005. What are the Effects of Monetary Policy? Results from an agnostic Identification procedure. Journal of Monetary Economics 52, 381-419.

Table 1.a: The equations of the model

$e_t^b - \frac{\sigma_c}{1-h}(y_t - hy_{t-1}) =$	$E_t[e_{t+1}^b - \frac{\sigma_c}{1-h}(y_{t+1} - hy_t)] + (R_t - E_t\pi_{t+1})$	(T.1)
$\pi_t^w - \mu_w\pi_{t-1} =$	$\kappa_w \left[-(e_t^b - \frac{\sigma_c}{1-h}(y_t - hy_{t-1})) + \sigma_l N_t - w_t\right] + \beta(E_t\pi_{t+1}^w - \mu_w\pi_t)$	(T.2)
$\pi_t - \mu_p\pi_{t-1} =$	$\kappa_p [w_t + n_t - y_t + e_t^\mu] + \beta(E_t\pi_{t+1} - \mu_p\pi_t)$	(T.3)
$R_t =$	$\rho_R R_{t-1} + (1-\rho_R)\left[\gamma_\pi\pi_t + \gamma_y y_t\right] + e_t^R$	(T.4)
$w_t =$	$w_{t-1} + \pi_t^w - \pi_t$	(T.5)
$y_t =$	$e_t^z + (1-\alpha)N_t$	(T.6)

The endogenous variables are y_t: output; N_t: hours worked; R_t: nominal rate; w_t: real wage rate; π_t: price inflation rate; π_t^w: wage inflation rate. The disturbances are: technology shock ($e_t^z = \rho_z e_{t-1}^z + u_t, u_t \sim N(0, \sigma_z^2)$); preference shock ($e_t^b = \rho_b e_{t-1}^b + v_t, v_t \sim N(0, \sigma_b^2)$); monetary policy shock ($e_t^R \sim N(0, \sigma_r^2)$); and price markup shock ($e_t^\mu \sim N(0, \sigma_u^2)$). In equation (T.3) $\kappa_p \equiv \frac{(1-\zeta_p)(1-\beta\zeta_p)}{\zeta_p}\frac{1-\alpha}{(1-\alpha+\alpha\epsilon)}$ and in equation (T.2) $\kappa_w \equiv \frac{(1-\zeta_w)(1-\beta\zeta_w)}{\zeta_w(1+\varphi\sigma_l)}$.

Table 1.b: Supports for the parameters and DGPs used in the experiments.

Parameter	Description	Support	DGP1	DGP2
β	Discount factor	0.99	0.99	0.99
ϵ	Elasticity in goods bundler	[5.00, 7.00]	6	6
φ	Elasticity in labor bundler	[5.00, 7.00]	6	6
σ_c	Risk aversion coefficient	[1.00, 5.00]	2.00	2.00
σ_l	Inverse Frish elasticity of labor supply	[0.00, 5.00]	1.74	1.74
h	Habit parameter	[0.00, 0.95]	0	0
ζ_p	Probability of keeping prices fixed	[0.00, 0.90]	0	0.75
ζ_w	Probability of keeping wages fixed	[0.00, 0.90]	0.62	0
μ_p	Indexation in price setting	[0.00, 0.80]	0	0
μ_w	Indexation in wage setting	[0.00, 0.80]	0	0
α	1 - labor share in production function	[0.30, 0.40]	0.36	0.36
ρ_r	Inertia in Taylor rule	[0.25, 0.95]	0.74	0.74
γ_y	Response to output in Taylor rule	[0.00, 0.50]	0.26	0.26
γ_π	Response to inflation in Taylor rule	[1.05, 2.50]	1.08	1.08
ρ_z	Persistence of productivity	[0.50, 0.99]	0.74	0.74
ρ_b	Persistence in taste process	[0.00, 0.99]	0.82	0.82
σ_z	Standard deviation of productivity		0.0388	0.0388
σ_μ	Standard deviation of markup		0.0316	0.0316
σ_b	Standard deviation of preferences		0.1188	0.1188
σ_r	Standard deviation of monetary		0.0033	0.0033
σ_m	Standard deviation of measurement error		0.0010	0.0010

Table 2: Signs of the impact response intervals to shocks.

	Markup shocks							Monetary shocks						
	M	M1	M2	M3	M4	N1	N2	M	**M1**	**M2**	M3	M4	N1	N2
R_t	+	+	+	+	+	+	+	+	+	+	+	+	+	+
w_t	-	-	-	-	-	-	-	?	+	-	?	?	?	?
π_t	+	+	+	+	+	+	+	-	-	-	-	-	-	-
y_t	-	-	-	-	-	-	-	-	-	-	-	-	-	-
n_t	-	-	-	-	-	-	-	-	-	-	-	-	-	-
	Taste shocks							Technology shocks						
	M	M1	M2	M3	M4	N1	N2	M	M1	M2	M3	M4	N1	N2
R_t	+	+	?	+	?	+	+	-	-	-	-	-	-	-
w_t	?	-	?	?	-	?	?	?	+	?	?	+	?	?
π_t	+	+	?	+	?	+	+	-	-	-	-	-	-	-
y_t	+	+	+	+	+	+	+	+	+	+	+	+	+	+
n_t	+	+	+	+	+	+	+	-	-	-	-	-	-	-

A '+' indicates that at least 90 percent of the impact response interval is positive; a '-' that at least 90 percent of the impact response interval is negative; a '?' a response interval which lies on both sides of the zero line. M is the general model; in M1 $\zeta_p = 0$; in M2 $\zeta_w = 0$; in M3 $\mu_p = 0$ and $\mu_w = 0$; in M4 $\sigma_l = 0$. In N1 habit is of multiplicative form and in N2 nominal rigidities are modelled with quadratic adjustment costs.

Table 3: Percentage of cases where the impact real wage response is correctly signed.

	5 variable VAR											
	Basic				Larger monetary shocks				Larger markup shocks			
Identified shocks	(a)	(b)	(c)	(d)	(a)	(b)	(c)	(d)	(a)	(b)	(c)	(d)
Markup	99.9		99.8		99.9		99.9		100		100	
Monetary	73.1	78.6		72.6	93.1	90.1		90.2	55.3	65.2		52.2
Taste	98.3	97.9			99.1	99.3			96.3	94.9		
Technology	99.5				99.6				97			
Supply		99.8				99.9				99.9		
	4 variable VAR											
	Basic				Larger monetary shocks				Larger markup shocks			
Identified shocks	(a)	(b)	(c)	(d)	(a)	(b)	(c)	(d)	(a)	(b)	(c)	(d)
Monetary		78.9		78.1		94.4		90.4		66.2		64.3
Taste		98.7				99.5				94.2		
Supply		99.8	99.6			99.8	99.8			99.9	99.8	

The VAR includes output, real wages, hours, inflation and the nominal rate in the first panel and output, real wages, inflation and the nominal rate in the second panel. In case (a) output, inflation, nominal rate and hours are restricted and shocks are jointly identified; in case (b) output, nominal rate and inflation are restricted and a supply shock, a monetary and a markup shock are identified; in cases (c) and (d) output, inflation, nominal rate and hours are restricted and a markup (supply) or a monetary shock are separately identified. In the second panel the standard deviation of either the monetary shocks is set 10 times larger. In the third panel the standard deviation of either the markup shocks is set 10 times larger.

Table 4: Percentage of correctly signed real wage impact response to monetary shocks.

	All identified			Monetary shocks identified		
	T=80	T=160	T=500	T=80	T=160	T=500
VAR(2)	72	73	73	72	71	71
VAR(4)	73	72	73	72	71	72
VAR(10)	72	74	74	72	71	72
BIC	72	73	72	70	71	73

Median value across 200 Monte Carlo replications. The DGP is a flexible price, sticky wage model and the VAR includes output, real wages, hours, inflation and the nominal rate. $p = 2, 4, 10$ is to the lag length of the VAR. The row labelled "BIC" reports probabilities computed when the lag length of the VAR is selected with BIC.

Table 5: Signs of the impact response intervals to shocks, Smets and Wouter class.

	Markup	Monetary	Taste	Technology	Investment	Labor supply	Government
y_t	+	+	+	+	?	+	+
π_t	-	+	+	-	-	-	?
R_t	-	-	+	-	?	-	+
w_t	+	?	?	?	?	-	?
n_t	+	+	+	-	?	+	+
LP-W gap$_t$	-	?	-	+	+	-	-

A '+' indicates that at least 90 percent of the impact response interval is positive; a '-' that at least 90 percent of the impact response interval is negative; a '?' a response interval which lies on both sides of the zero line.

Table 6.a: The equations of the model

$q_t = \beta E_t q_{t+1} + [1-\beta(1-\delta)]\, E_t r^k_{t+1} - (R_t - E_t \pi_{t+1})$	(T.7)
$i_t - k_{t-1} = \eta q_t$	(T.8)
$k_t = (1-\delta)k_{t-1} + \delta i_t$	(T.9)
$c^o_t = c^o_{t+1} - (R_t - E_t \pi_{t+1})$	(T.10)
$c^r_t = \frac{1-\alpha}{\mu c_y}(w_t + n^r_t) - \frac{1}{c_y} t^r_t$	(T.11)
$w_t = c^j_t + \sigma_l n^j_t \quad j = o, r$	(T.12)
$r_t = mc_t + e^z_t + (1-\alpha)(n_t - k_{t-1})$	(T.13)
$w_t = mc_t + e^z_t - \alpha(n_t - k_{t-1})$	(T.14)
$y_t = e^z_t + (1-\alpha)n_t + \alpha k_{t-1}$	(T.15)
$y_t = c_y c_t + i_y i_t + g_y e^g_t$	(T.16)
$\pi_t - \mu_p \pi_{t-1} = \kappa_p(mc_t + e^u_t) + \beta(E_t \pi_{t+1} - \mu_p \pi_t)$	(T.17)
$R_t = \rho_R R_{t-1} + (1-\rho_R)(\gamma_\pi \pi_t + \gamma_y y_t) + e^R_t$	(T.18)
$b_t = \frac{1}{\beta}[(1-\phi_b)b_{t-1} + (1-\phi_g)e^g_t]$	(T.19)
$t_t = \phi_b b_{t-1} + \phi_g e^g_t$	(T.20)

The disturbances are: technology shock ($e^z_t = \rho_z e^z_{t-1} + u_t, u_t \sim N(0, \sigma^2_z)$); government spending shock ($e^g_t = \rho_g e^g_{t-1} + v_t, v_t \sim N(0, \sigma^2_g)$); monetary policy shock ($e^R_t \sim N(0, \sigma^2_r)$); and price markup shock ($e^\mu_t \sim N(0, \sigma^2_u)$). The compound parameters in equation (T.17) is defined as: $\kappa_p \equiv \frac{(1-\zeta_p)(1-\beta\zeta_p)}{\zeta_p}$.

Table 6.b: Supports for the structural parameters and DGP used in the experiments.

Parameter	Description	Support	DGP
λ	Share of ROTC	[0.00,0.90]	0, 0.80
σ_l	Wage elasticity to hours	[0.00,1.00]	0.2
δ	Depreciation of capital	[0.00,0.05]	0.025
α	Capital share	[0.30,0.40]	0.33
η	Elasticity of i/K to q	[0.50,2.00]	1.0
ζ_p	Price stickiness	[0.00,0.90]	0.75
μ	Gross monopolistic markup	[1.10,1.30]	1.2
ρ_r	Inertia in monetary policy	[0.00,0.90]	0.0
γ_π	policy response to inflation	[1.05,2.50]	1.5
γ_y	Policy response to output	[0.00,0.10]	0.0
μ_p	Indexation in price setting	[0.00,0.80]	0.0
ϕ_b	Fiscal rule response to bonds	[0.25,0.40]	0.33
ϕ_g	Fiscal rule response to expenditure	[0.05,0.15]	0.1
ρ_g	AR(1) parameter government spending	[0.50,0.95]	0.9
ρ_t	AR(1) parameter productivity	[0.50,0.95]	0.9
g_y	Steady state spending share in output	[0.15,0.20]	0.2
σ_u	Standard deviation of markup shocks		0.30
σ_R	Standard deviation of monetary shocks		0.025
σ_z	Standard deviation of TPF shocks		0.07
σ_g	Standard deviation of government shocks		0.10

Table 7: Signs of the impact response intervals to shocks.

	Markup	Monetary Policy	Technology	Spending
Δy	-	-	+	+
π	+	-	-	+
Δn	-	-	-	+
Δi	-	-	+	-
R	+	+	-	+

A '+' indicates that at least 90 percent of the impact response interval is positive; a '-' that at least 90 percent of the impact response interval is negative; a '?' a response interval which lies on both sides of the zero line. 10000 parameter vectors are drawn from the intervals in table 6

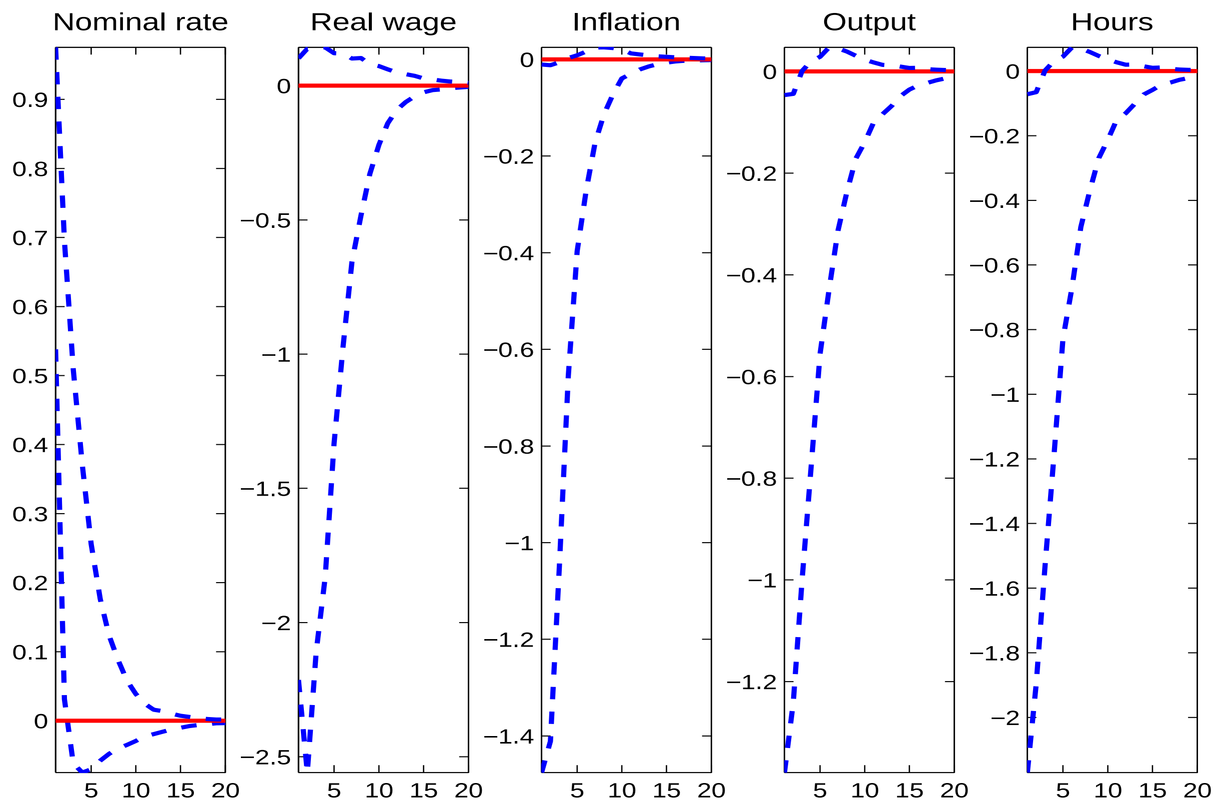
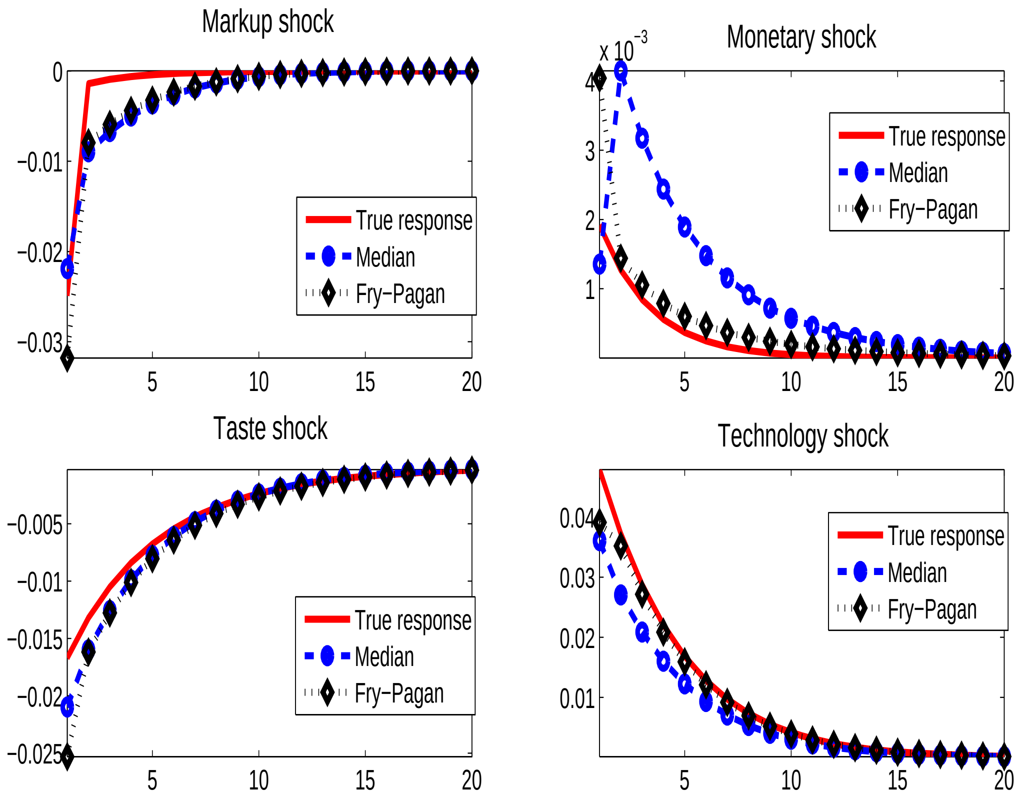


Figure 1:Pointwise 90 percent response intervals to monetary shocks. Model M.


Markup shock
True response
Median
Fry−Pagan
0
−0.01
−0.02
−0.03
5
10
15
20
Monetary shock
x 10^-3
4
3
2
1
True response
Median
Fry−Pagan
5
10
15
20
Taste shock
−0.005
−0.01
−0.015
−0.02
−0.025
True response
Median
Fry−Pagan
5
10
15
20
Technology shock
0.04
0.03
0.02
0.01
True response
Median
Fry−Pagan
5
10
15
20

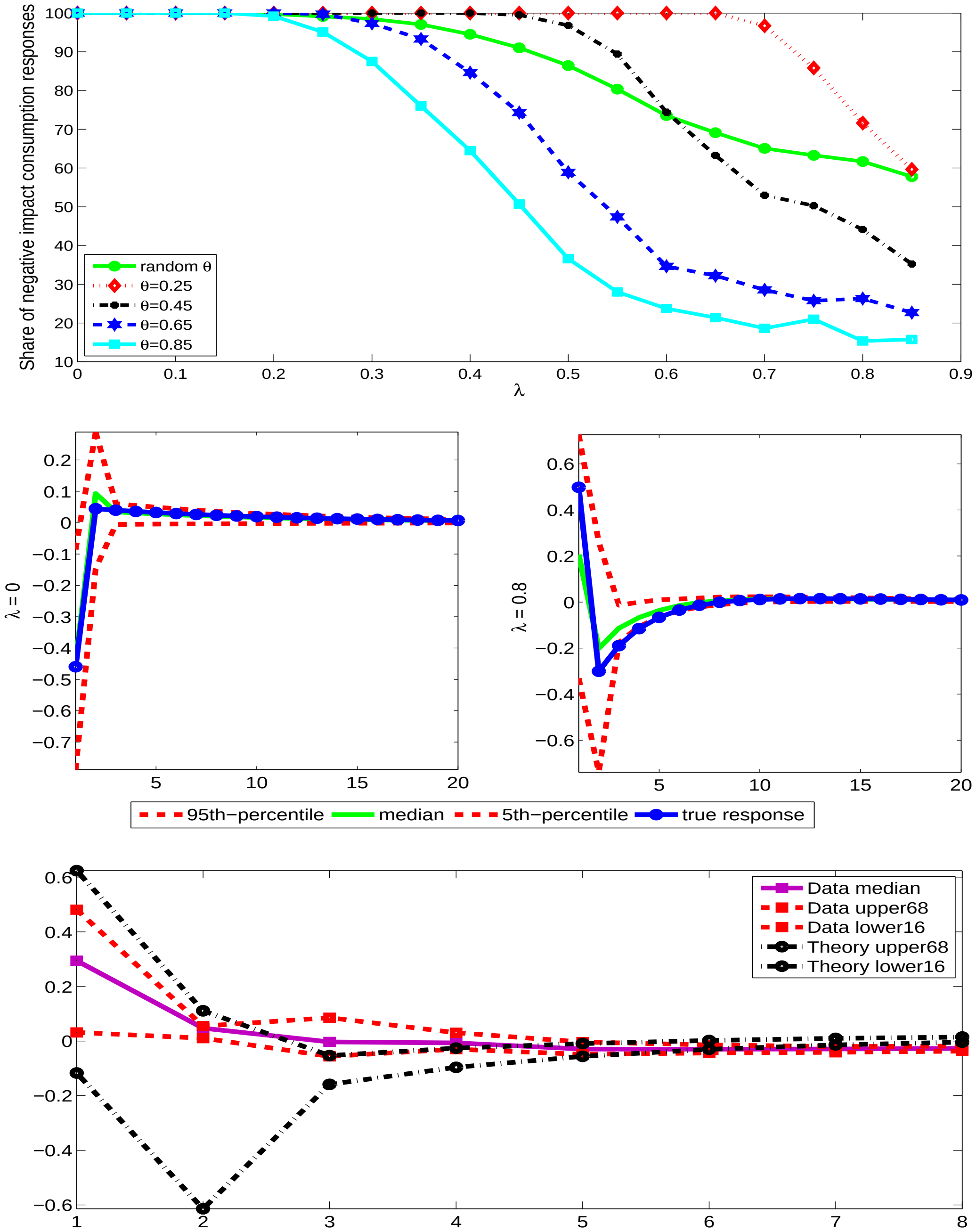


Figure 3: Consumption responses to government spending shocks. First panel theory; second panel simulated data; third panel actual data.