NBER WORKING PAPER SERIES

ESTIMATING DSGE MODELS:
RECENT ADVANCES AND FUTURE CHALLENGES

Jesús Fernández-Villaverde
Pablo A. Guerrón-Quintana

Working Paper 27715
http://www.nber.org/papers/w27715

NATIONAL BUREAU OF ECONOMIC RESEARCH
1050 Massachusetts Avenue
Cambridge, MA 02138
August 2020

We thank David Childers and Peifan Wu for useful comments and Sara Casella for superb research assistance. The views expressed herein are those of the authors and do not necessarily reflect the views of the National Bureau of Economic Research.

Estimating DSGE Models: Recent Advances and Future Challenges
Jesús Fernández-Villaverde and Pablo A. Guerrón-Quintana
NBER Working Paper No. 27715
August 2020
JEL No. C11,C13,E30

ABSTRACT

We review the current state of the estimation of DSGE models. After introducing a general framework for dealing with DSGE models, the state-space representation, we discuss how to evaluate moments or the likelihood function implied by such a structure. We discuss, in varying degrees of detail, recent advances in the field, such as the tempered particle filter, approximated Bayesian computation, the Hamiltonian Monte Carlo, variational inference, and machine learning, methods that show much promise, but that have not been fully explored yet by the DSGE community. We conclude by outlining three future challenges for this line of research.

Jesús Fernández-Villaverde
Department of Economics
University of Pennsylvania
The Ronald O. Perelman Center
 for Political Science and Economics
133 South 36th Street Suite 150
Philadelphia, PA 19104
and CEPR
and also NBER
jesusfv@econ.upenn.edu

Pablo A. Guerrón-Quintana
Boston College
Department of Economics
140 Commonwealth Avenue
Malloney Hall 325
Chestnut Hill, MA 02467
pguerron@gmail.com

1 Introduction

Dynamic stochastic general equilibrium (DSGE) models have become one of the central tools of macroeconomics. The class of DSGE economies is not defined by a particular set of assumptions, but by an approach to the construction of macroeconomic models. Without being exhaustive, there are DSGE models with fully flexible prices (the subclass of real business cycle models in the tradition of Kydland and Prescott, 1982) or with nominal rigidities (the New Keynesian models à la Woodford, 2003 and Christiano et al., 2005, so prevalent in central banks for the analysis of monetary policy). There are DSGE models with a representative household or with heterogeneous households and firms (Kaplan et al., 2018, and Khan and Thomas, 2007). There are DSGE models with infinitely lived agents and with finitely lived ones (Nishiyama and Smetters, 2014). There are DSGE models with complete financial markets or with financial frictions and incomplete markets (Fernández-Villaverde et al., 2019). There are DSGE models with standard CRRA utility functions and with a wide variety of "exotic" preferences, such as recursive utility functions or ambiguity aversion (van Binsbergen et al., 2012; Ilut and Schneider, 2014). There are DSGE models with rational expectations or with learning (Primiceri, 2006). There are DSGE models with full rationality or with behavioral biases (Gabaix, Forthcoming). The possibilities are endless.

The common thread behind all of these DSGE models is an emphasis on the explicit description of all the details behind the preferences, technology, and information sets of agents, a focus on the dynamic consequences of stochastic shocks, careful attention to general equilibrium interactions, and the first-order importance of a quantitative assessment of the properties of the model and its fit to the data beyond a purely qualitative gauging of its implications.

This last point is particularly salient for this paper: macroeconomists working with DSGE models are keenly concerned with the features of the data that their models can (and, often, cannot) account for. While the research agenda started highlighting calibration as an alternative to formal econometric methods (Hansen and Prescott, 1995), it was soon apparent that we could apply suitably adapted econometric tools to these models. For instance, many of the equilibrium conditions of a DSGE model, such as the Euler equation relating marginal utilities today with marginal utilities tomorrow, can be thought of as a moment condition and the parameters governing them can be estimated by matching moments of the model with analogous moments from the data (Hansen, 1982). Similarly, researchers learned how to build the likelihood function of a DSGE model and either maximize it or use it to derive a posterior for Bayesian analysis (Fernández-Villaverde, 2010). Two vital complementary improvements were the hardware improvements (including the arrival of massive parallelization at low prices) and the development of better computation techniques to solve DSGE models, which allowed economists to work with richer environments.

Formal econometric tools enjoy several decisive advantages. A first advantage is that they provide a general framework for determining the model's parameter values. Calibration, as proposed by Kydland and Prescott (1982), is easy to apply to models with a few well-identified parameters. However, once the models become more complex, researchers face too many degrees of freedom: i) there are plenty of potential moments to match (this is also a problem for methods of moments); ii) micro estimates become harder to import into macro contexts (Browning et al., 1999); and iii) many combinations of parameter values "fit" the data. In comparison, the likelihood function offers a clean solution to all of these concerns, as it embodies all the relevant information existing in the data (Berger and Wolpert, 1988). Even when the DSGE model is not well-identified (a common occurrence; Canova and Sala, 2009, Iskrev, 2010, and Komunjer and Ng, 2011), the likelihood function gives us a range of parameter values and bounds on outcomes of interest implied by them. Far from being a weakness of econometric methods, the rise in the awareness of a lack of (or weak) identification that these methods bring is one of their strengths. There are limits to our knowledge. It is better to tackle those limits than to live under the false impression of certainty that a quick-and-dirty calibration begets. For instance, we can build economic policies that incorporate lack-of-identification results by factoring in robustness considerations in our choices of fiscal stimuli or policy interest rates.

A second advantage is the econometric tools' ability to forecast, assess the model's fit to the data, and compare models rigorously. Calibration, beyond its inability to engage in any forecasting, offers only heuristic approaches to gauge how a model fits the data and decide which of two or more models is superior in its ability to account for the data dynamics. While heuristic assessments are not without value (and, for simple cases, they offer an attractive combination of insight and speed in implementation), they often leave the researcher at a loss when dealing with more complex models.

Similarly, estimation allows us to recover the conditional distribution of latent variables of interest. These variables, such as shocks or unobserved states, are the object of interest in many exercises related to policy assessment, counterfactual analysis, and forecasting.

If one is intrigued by the previous arguments, the natural question is: how do we estimate DSGE models in "real life"? We answer this question by building on Fernández-Villaverde (2010), Herbst and Schorfheide (2015), and Fernández-Villaverde et al. (2016), which reviewed the state-of-the-art in the solution and estimation of DSGE models a few years ago. Many of the ideas in these surveys are still fully applicable, and, here, we will introduce only the bare minimum notation to review them. Here, we will extend those surveys by discussing some of the most recent advances in the field, such as the tempered particle filter, approximated Bayesian computation, the Hamiltonian Monte Carlo, variational inference, and machine learning, methods that show much promise, but that have not yet been fully explored by the DSGE community. Since there is much to cover, let us start without further ado.

2 A general framework for estimating DSGE models

We introduce a high-level notation to explain how to estimate a large class of DSGE models. The key to our approach is to use a state-space representation, a formalism that originated in optimal control theory (Kalman, 1960), but that has gained widespread use across many fields. In any DSGE model, we have a vector of states $S_t \in \mathbb{R}^m$ that describes the economy's situation at period t. A state can be, among others, a scalar (e.g., the aggregate capital in the economy), a population distribution (e.g., the measure of households over their individual states), or a probability distribution (e.g., the beliefs of an agent regarding future events). Dealing with complex states, such as distributions, brings computational challenges, but it does not require much effort in terms of notation.

The states are buffeted by shocks, such as random changes to technology, preferences, fiscal and monetary policy, health conditions, etc. We stack all the shocks in a vector W_t. We do not impose normality or any other constraint on these shocks except that they are i.i.d. We can capture persistence and time-varying moments with additional states in S_t, such as in Fernández-Villaverde et al. (2011).

Finally, we have a vector of parameters $\theta \in \mathbb{R}^d$. These parameters determine the preferences, technology, information sets, and the economy's fiscal and monetary policy rules. For simplicity, we assume that the parameters are fixed over time, but with extra notation, we can allow them to vary over time (Fernández-Villaverde and Rubio-Ramírez, 2008).

Putting all these elements together, we get the first leg of the state-space representation, the transition equation:

$$S_t = f\left(S_{t-1}, W_t; \theta\right). \tag{1}$$

An alternative way to think about equation (1) is as a conditional probability distribution for S_t, $p\left(S_t | S_{t-1}; \theta\right)$, where the conditioning is on the value of the states at $t-1$.

The second leg of the state-space representation is a measurement equation:

$$Y_t = g\left(S_t, V_t; \theta\right), \tag{2}$$

where $Y_t \in \mathbb{R}^n$ are the observables and V_t is a set of shocks. These shocks might be defined within the model (such as a shock to some variable that does not feed back into the states) or outside the model (for example, measurement errors on observables). As before, equation (2) induces a density, $p\left(Y_t | S_t; \theta\right)$, for Y_t conditional on S_{t-1}.

Combining equations (1) and (2), we get $Y_t = g\left(f\left(S_{t-1}, W_t; \theta\right), V_t; \theta\right)$ and, implicitly, the density $p\left(Y_t | S_{t-1}; \theta\right)$ of observables conditional on S_{t-1}. This last density embodies the idea that a DSGE model is nothing more than a restriction on a general stochastic process for Y_t.

The functions $f(\cdot)$ and $g(\cdot)$ are, in general, unknown and cannot be found explicitly. Instead, we need to solve the DSGE model and, with such a solution, build them. Simultaneously, there are occasions when the components of $f(\cdot)$ and $g(\cdot)$ may be trivial. For example, a state may be directly observable, and the corresponding dimension of $g(\cdot)$ would just be the identity function. Fernández-Villaverde et al. (2016) is an updated survey of the main existing solution methods for DSGE models. Fernández-Villaverde and Valencia (2018) outline how to parallelize these methods.[1]

All of these conditional densities can be exploited to take the model to the data. We can use them, for instance, to build moments for Y_t implied by the model and estimate θ by minimizing the distance between these moments and the data analogs. The moments can be direct (means, variances, correlations) or indirect (the model's impulse-response functions, or IRFs). In the latter case, the researcher also needs to estimate the data IRFs, for example, through a structural vector autoregression. Andreasen et al. (2018) show how to build the moments and IRFs of DSGE models using closed-form formulae from a perturbation solution and provide a software toolbox for doing so efficiently.

An alternative to estimation by moments is to use the conditional densities to evaluate the likelihood function of a sequence of observations $y^T = \{y_1, y_2, ..., y_T\}$ at θ, $p\left(y^T; \theta\right)$, as follows. First, given the Markov structure of equations (1) and (2), we write:

$$\begin{aligned} p\left(y^T | \theta\right) &= p\left(y_1 | \theta\right) \prod_{t=2}^{T} p\left(y_t | y^{t-1}; \theta\right) \\ &= \int p\left(y_1 | s_1; \theta\right) dS_1 \prod_{t=2}^{T} \int p\left(y_t | S_t; \theta\right) p\left(S_t | y^{t-1}; \theta\right) dS_t \end{aligned}$$

where lower case letters denote realizations of a random variable. Therefore, if we have access to the sequence $\{p\left(S_t | y^{t-1}; \theta\right)\}_{t=1}^{T}$ and the initial distribution of states $p\left(S_1; \theta\right)$, we can evaluate the likelihood of the model.

Finding the sequence $\{p\left(S_t | y^{t-1}; \theta\right)\}_{t=1}^{T}$ can be recursively accomplished using the Chapman-

[1] While reviewing solution methods is beyond our scope, the speed and accuracy of these methods are crucial. Speed is a vital consideration because we will need to evaluate moments or the likelihood function of the model for many different combinations of parameter values. Accuracy in the solution is required to avoid getting incorrect point estimates.

Kolmogorov equation:

$$p\left(S_{t+1}|y^{t};\theta\right)=\int p\left(S_{t+1}|S_{t};\theta\right)p\left(S_{t}|y^{t};\theta\right)dS_{t} \tag{3}$$

and Bayes' theorem:

$$p\left(S_{t}|y^{t};\theta\right)=\frac{p\left(y_{t}|S_{t};\theta\right)p\left(S_{t}|y^{t-1};\theta\right)}{p\left(y_{t}|y^{t-1};\theta\right)} \tag{4}$$

where

$$p\left(y_{t}|y^{t-1};\theta\right)=\int p\left(y_{t}|S_{t};\theta\right)p\left(S_{t}|y^{t-1};\theta\right)dS_{t}.$$

is the conditional likelihood. This recursion is started with $p\left(S_1;\theta\right)$.

The Chapman-Kolmogorov equation forecasts the density of states tomorrow given the observables up to today. This conditional density is the density of states tomorrow conditional on S_t given by the transition equation (1) times the density of S_t given the observables up to today, integrated over all possible states. Bayes' theorem delivers the density of states today given the observables up to today by updating the distribution of states $p\left(S_t|y^{t-1};\theta\right)$ when a new observation arrives given its probability $p\left(y_t|S_t;\theta\right)$.

While equations (3) and (4) are conceptually simple, solving all of the required integrals in them by hand is impossible beyond a few textbook examples. We next outline the main approaches for doing so.

3 Evaluating the likelihood function

3.1 The Kalman filter

Often, equations (3) and (4) are linear or, more likely, we can find a linearized version of them that is close to the original formulation under an appropriate metric (see Fernández-Villaverde et al., 2016). Also, most DSGE models (but not all!) assume that the shocks W_t and V_t follow a normal distribution. Then, we can write

$$\begin{aligned}
s_t &= As_{t-1}+B\varepsilon_t && (5)\\
y_t &= Cs_t+D\varepsilon_t && (6)\\
\varepsilon_t &\sim \mathcal{N}\left(0,I\right)
\end{aligned}$$

where, for notational simplicity, we stack W_t and V_t in the vector ε_t, add the required zero columns in the matrices A, B, C, and D to make the system consistent, and scale B and D to induce the right covariance matrix of the shocks.

Equations (5) and (6) are linear transformations of ε_t conditional on s_{t-1}. Since the linear transformation of a normal random variable is still normally distributed, $p\left(S_t|S_{t-1};\theta\right)$ and $p\left(Y_t|S_t;\theta\right)$ are normal densities. Because the mean and variance are sufficient statistics for a normal distribution, keeping track of the sequence $\left\{p\left(S_t|y^{t-1};\theta\right)\right\}_{t=1}^{T}$ is equivalent to keeping track of the sequence of conditional means and variances of S_t.

Kalman (1960) developed a simple recursive procedure for this tracking, which became known as the Kalman filter (KF), that can be implemented in a few lines of software (see Harvey 1990 for more details).

To explain how the KF works, we define the conditional expectations of the states $s_{t|t-1} = \mathbb{E}\left(s_t|Y_{t-1}\right)$ and $s_{t|t} = \mathbb{E}\left(s_t|Y_t\right)$, where $Y_t = \{y_1, y_2, ..., y_t\}$ and the subindex tracks the conditioning set (i.e., $t|t-1$ means the expectation at t conditional on information until $t-1$). Also, we define the covariance matrices of the state $P_{t-1|t-1} = \mathbb{E}\left(s_{t-1} - s_{t-1|t-1}\right)\left(s_{t-1} - s_{t-1|t-1}\right)'$ and $P_{t|t-1} = \mathbb{E}\left(s_{t-1} - s_{t|t-1}\right)\left(s_{t-1} - s_{t|t-1}\right)'$.

With this notation, we can manipulate equations (5) and (6) and derive the one-step-ahead forecast error, $\eta_t = y_t - Cs_{t|t-1}$, and its variance $V_y = CP_{t|t-1}C' + DD'$. Because of the linearity of equations (5) and (6) and the normality of innovations, η_t is white noise and the loglikelihood of y_t is:

$$\log p\left(y_t|\theta\right) = -\frac{n}{2}\log 2\pi - \frac{1}{2}\log\left|V_y\right| - \frac{1}{2}\eta_t V_y^{-1}\eta_t.$$

The task is, therefore, to recursively compute $s_{t|t-1}$, $s_{t|t}$, $P_{t|t-1}$, and $P_{t|t}$. With these objects, we can compute (3.1) for all $y^T = \{y_1, y_2, ..., y_T\}$ and get the loglikelihood function:

$$\log p\left(y^T|\theta\right) = \sum_{t=1}^{T}\log p\left(y_t|\theta\right)$$

How do we compute $s_{t|t-1}$, $s_{t|t}$, $P_{t|t-1}$, and $P_{t|t}$? Forecasting $s_{t|t-1}$ (i.e., implementing the Chapman-Kolmogorov equation in terms of means) is straightforward: $s_{t|t-1} = As_{t-1|t-1}$. Updating $s_{t|t}$ given a new observation (i.e., implementing Bayes' theorem in terms of means) is also direct:

$$s_{t|t} = s_{t|t-1} + K_t\eta_t = s_{t|t-1} + K_t(y_t - Cs_{t|t-1}), \tag{7}$$

where K_t is the Kalman gain at time t, which minimizes $P_{t|t}$ with the first-order condition:

$$\frac{\partial Tr\left(P_{t|t}\right)}{\partial K_t} = 0$$

and the solution $K_t = \left[P_{t|t-1}C' + BD'\right]\left[V_y + CBD' + DB'C'\right]^{-1}$. We minimize $Tr\left(P_{t|t}\right)$ because we want to update the estimate of the states to have the smallest unbiased covariance matrix. We can obtain the same K_t by alternative routes, such as using the rules of composition of normal distributions.

With some algebra, we can derive the forecast $P_{t|t-1} = AP_{t-1|t-1}A' + BB'$ and the update:

$$P_{t|t} = (I - K_tC)\,P_{t|t-1}\,(I - C'K_t') + K_tDD'K_t' - K_tDB' - BD'K_t' + K_tCBD'K_t' + K_tDB'C'K_t'.$$

The recursion for $s_{t|t-1}$, $s_{t|t}$, $P_{t|t-1}$, and $P_{t|t}$, which takes a fraction of a second in a standard laptop, requires initial values to be started. A conventional choice for DSGE models is to compute, given the values of θ, the mean and variance of the ergodic distribution of S_t, and use those as the starting points of the recursion. Andreasen et al. (2018) provide the formulae of these ergodic moments for a general class of DSGE models.

3.2 Nonlinear filters

The linearity of the state-space representation and normal shocks are restrictive assumptions. For many DSGE models, such as standard real business cycle or New Keynesian models, linearizing the model's equilibrium conditions is not a bad approximation (see, for quantitative evidence, Aruoba et al., 2006). However, in many applications, we cannot rely on linearizing the equilibrium conditions of the model.

A transparent case is when we have time-varying uncertainty, a booming research line during the last 15 years (Fernández-Villaverde and Guerrón-Quintana, 2020). When the variance of the shocks that hit the economy changes over time, we need nonlinear solution methods because the linearized solution is certainty equivalent and, therefore, it misses all the changes in behavior induced by varying second moments. Similarly, DSGE models with "exotic" preferences (van Binsbergen et al., 2012; Ilut and Schneider, 2014) also require nonlinear solution methods. In a linearized world, the decision rules implied by those "exotic" preferences collapse to those coming from a standard CRRA utility function.

When the state-space representation is not linear or when the shocks are not normal, the conditional densities of states do not belong to well-behaved parametric families. Thus, the KF approach of tracking the sufficient statistics of the normal distribution breaks down.

For many years, researchers attempted to extend the KF to nonlinear setups, for example, by keeping track of additional moments or by linearizing the state-space representation around an estimate of the current means and covariances. Those extensions perform poorly when applied to DSGE models.[2] In economics, researchers have found that particle filters (PFs), a

[2] Some more recent extensions, such as the unscented Kalman filter (UKF), have gained popularity in engineering (Wan and Van Der Merwe, 2000). In unpublished work, we applied the UKF to nonlinear DSGE

class of sequential Monte Carlo algorithms (Chopin and Papaspiliopoulos, 2020), deliver good results, although at some computational cost.

3.2.1 The bootstrap particle filter

A PF builds from the idea that, if we can draw from the conditional densities of states $\{p(S_t|y^{t-1};\theta)\}_{t=1}^{T}$, we can approximate the unknown density by an empirical distribution of N draws $\left\{\left\{s_{t|t-1}^{i}\right\}_{i=1}^{N}\right\}_{t=1}^{T}$ from the sequence $\{p(S_t|y^{t-1};\theta)\}_{t=1}^{T}$ generated by simulation. This set of draws is often called a swarm of particles (hence the name of the filter). Under technical conditions, the law of large numbers gives us:

$$p\left(y^T|\theta\right) \simeq \frac{1}{N}\sum_{i=1}^{N} p\left(y_1|s_{0|0}^{i};\theta\right)\prod_{t=2}^{T}\frac{1}{N}\sum_{i=1}^{N} p\left(y_t|s_{t|t-1}^{i};\theta\right) \tag{8}$$

The task is to efficiently draw from $\{p(S_t|y^{t-1};\theta)\}_{t=1}^{T}$ for DSGE models. Fernández-Villaverde and Rubio-Ramírez (2007), in the first application of this technique to the estimation of DSGE models, propose a simple bootstrap particle filter (BPF) structured around sequential sampling (Robert and Casella, 2005):

Proposition 1 *Let $\left\{s_{t|t-1}^{i}\right\}_{i=1}^{N}$ be a draw from $p(S_t|y^{t-1};\theta)$. Let the sequence $\{\widetilde{s}_t^i\}_{i=1}^{N}$ be a draw with replacement from $\left\{s_{t|t-1}^{i}\right\}_{i=1}^{N}$ where the resampling probability is given by the importance weights:*

$$\omega_t^i = \frac{p\left(y_t|s_{t|t-1}^{i};\theta\right)}{\sum_{i=1}^{N} p\left(y_t|s_{t|t-1}^{i};\theta\right)}, \tag{9}$$

Then $\{\widetilde{s}_t^i\}_{i=1}^{N}$ is a draw from $p(S_t|y^t;\theta)$.

Proposition 1 uses a draw $\left\{s_{t|t-1}^{i}\right\}_{i=1}^{N}$ from $p(S_t|y^{t-1};\theta)$ to draw $\left\{s_{t|t}^{i}\right\}_{i=1}^{N}$ from $p(S_t|y^t;\theta)$. By doing so, it updates the conditional distribution of the state by incorporating y_t, as Bayes' theorem requires. Resampling ensures that the sequences of draws do not deviate arbitrarily far away from the true value of the states, losing all information in the simulation.

Once we have $\left\{s_{t|t}^{i}\right\}_{i=1}^{N}$, we draw N vectors of exogenous shocks to the model from their distributions and apply the law of motion for states to generate $\left\{s_{t+1|t}^{i}\right\}_{i=1}^{N}$. This Chapman-Kolmogorov forecast step allows us to go back to Proposition 1, but now having moved from conditioning on $t|t-1$ to conditioning on $t+1|t$.

models with disappointing results. However, other researchers might deliver better performance with further tuning of the algorithm.

The pseudo-code below summarizes the algorithm:

```
Step 0, Initialization: Set $t \rightsquigarrow 1$. Sample $N$ values $\left\{s^i_{0|0}\right\}_{i=1}^N$ from $p\left(S_0;\theta\right)$.
Step 1, Prediction: Sample $N$ values $\left\{s^i_{t|t-1}\right\}_{i=1}^N$ using $\left\{s^i_{t-1|t-1}\right\}_{i=1}^N$, the law of motion for states and the distribution of shocks $\varepsilon_t$.
Step 2, Filtering: Assign to each draw $\left(s^i_{t|t-1}\right)$ the weight $\omega^i_t$ in Proposition 1.
Step 3, Sampling: Sample $N$ times with replacement from $\left\{s^i_{t|t-1}\right\}_{i=1}^N$ using the probabilities $\{\omega^i_t\}_{i=1}^N$. Call each draw $\left(s^i_{t|t}\right)$. If $t < T$ set $t \rightsquigarrow t+1$ and go to step 1. Otherwise stop.
```

Next, we substitute the resulting N draws $\left\{\left\{s^i_{t|t-1}\right\}_{i=1}^N\right\}_{t=1}^T$ into equation (8) and we get an estimate of $p\left(y^T|\theta\right)$. Künsch (2005) shows how laws of large numbers and a central limit theorem apply to this estimate under weak technical conditions.

Coding the BPF is straightforward and easy to parallelize. Unfortunately, the resulting approximation of the likelihood contains significant numerical error because of the noise introduced by the simulation and is computationally costly. How can we reduce the noise and improve the accuracy of the evaluation of the likelihood?

3.2.2 The tempered particle filter

A promising alternative to the BPF is the tempered particle filter (TPF) introduced by Herbst and Schorfheide (2019). The TFP starts with oversized shocks in the measurement equation and computes an approximate of the likelihood. Then, it iteratively reduces the covariance matrix of the measurement shocks updating the likelihood at each iteration. After N_ϕ steps, the tempered measurement shocks coincide with the original ones.

We return to our state-space representation, with the transition equation:

$$S_t = f\left(S_{t-1};\theta\right) + W_t, \tag{10}$$

and measurement equation

$$Y_t = g\left(S_t, ;\theta\right) + V_t. \tag{11}$$

where $W_t \sim \mathcal{N}(0,\Sigma_W)$ and $V_t \sim \mathcal{N}(0,\Sigma_V)$. In comparison with equations (1) and (2), we assume that the shocks are normal and enter linearly into (10) and (11). This is just for convenience. We could generalize these assumptions with heavier notation.

Recall that, in the BPF, we built the importance weights according to equation (9). Given the normality of V_t and how they enter linearly in (11), these weights satisfy:

$$\omega_t^i \propto (2\pi)^{d/2}|\Sigma_V|^{-1/2} \exp\left\{-\frac{1}{2}(y_t - g\left(s_{t|t-1}^i;\theta\right)' \Sigma_V^{-1}(y_t - g\left(s_{t|t-1}^i;\theta\right))\right\}. \tag{12}$$

The weights in equation (12) become more equal as the variance of the measurement errors increases. However, blindly raising these errors has the unappealing effect of reducing the explanatory power of the structural shocks of the model. Hence, the TPF proposes to work with an inflated covariance matrix and a sequence of approximated likelihoods:

$$p_n(y_t|s_{t|t-1};\theta) \propto \phi_n^{d/2}|\Sigma_V|^{-1/2} \exp\left\{-\frac{1}{2}(y_t - g(s_{t|t-1}^{n-1};\theta))'\phi_n\Sigma_u^{-1}(y_t - g(s_{t|t-1}^{n-1};\theta))\right\}. \tag{13}$$

The superindex $n = 1, \cdots, N_\phi$ indexes the steps of the tempering process. If $N_\phi = 1$, the TPF collapses to the BPF. Moreover, the tempering sequence satisfies $0 < \phi_1 < \phi_2 < \cdots < \phi_{N_\phi} = 1$. After the last step, we end up with an approximated likelihood featuring less numerical error than its equivalent from the BPF.

The TPF consists of three stages. First, we have the **correction** stage. Suppose that we enter into the tempering stage n with the swarm of particles $\{s_{t|t-1}^{i,n-1}\}$. Define the weights:

$$\widetilde{\omega}_t^{i,n}(\phi_n) = \frac{p_n(y_t|s_t^{i,n-1};\theta)}{p_{n-1}(y_t|s_t^{i,n-1};\theta)},$$

where $p_n(y_t|s_t^{i,n-1};\theta)$ is given by equation (13). Since the proportionality constants cancel out, these weights are easy to compute. Our notation makes explicit the weights' dependence on the tempering sequence.

Next, we define the weights:

$$\omega_t^{i,n}(\phi_n) = \frac{\widetilde{\omega}_t^{i,n}(\phi_n)}{\sum_{i=1}^N \widetilde{\omega}_t^{i,n}(\phi_n)}.$$

and the inefficiency ratio $InEff(\phi_n) = \frac{1}{N}\sum_{i=1}^N \left[\omega_t^{i,n}(\phi_n)\right]^2$.

Herbst and Schorfheide (2019) select the sequence of weights ϕ_n to achieve a targeted inefficiency ratio $\rho^* > 1$. That is, if at stage n, $InEff(1) \leq \rho^*$, we stop and set $\phi_n = 1$, $N_\phi = n$ and $\omega_t^{i,n} = \omega_t^{i,n}(1)$. In contrast, if the threshold is not met, $InEff(1) > \rho^*$, we set the tempering parameter ϕ_n as the solution to the equation $InEff(\phi_n) = \rho^*$, and $\omega_t^{i,n} = \omega_t^{i,n}(\phi_n)$.

The next stage is **resampling**. We resample the swarm of particles $\{s_{t|t-1}^{i,n-1}\}$ with probabilities $\{\omega_t^{i,n}\}$, resulting in a new swarm $\{\hat{s}_{t|t-1}^{i,n}\}$ with weight $\omega_t^{i,n} = 1$. This step guarantees a unique solution of the equation $InEff(\phi_n) = \rho^*$.

The final stage is **mutation**. Here, we use a random walk Metropolis Hastings (to be described in Section 4) with N^{MH} steps to mutate the particles $\{\hat{s}^{i,n}_{t|t-1}\}$ into $\{s^{i,n}_{t|t-1}\}$. This step avoids the algorithm reproducing the BPF as the tempering sequence approaches $\phi_n = 1$. As mentioned above, the final result is a more accurate evaluation of $p\left(y^T|\theta\right)$ than the one obtained from a BPF.

3.3 Approximate Bayesian computation

An alternative to the use of PFs is approximate Bayesian computation (ABC). ABC algorithms deal with the cases where evaluating the likelihood $p\left(y^T|\theta\right)$ is impossible or simply too computationally expensive, even for PFs. ABC proposes a series of likelihood-free methods, such as rejection samplers and perturbation kernels. Useful introductions to the field are Marjoram et al. (2003), Sisson et al. (2007), and Marin et al. (2012). Important asymptotic results are reported by Li and Fearnhead (2018). ABC has not been applied often to the estimation of DSGE models, although Scalone (2018) is an example of its potential usefulness in the field, particularly when dealing with small samples.

4 Markov chain Monte Carlo methods

In the previous section, we saw how we can move from the state-space representation (1) and (2) to a likelihood function $p(\theta|y^T)$. Next, we can either maximize this likelihood function or combine it with a prior using Bayes' theorem:

$$p(\theta|y^T) = \frac{p(y^T|\theta)p(\theta)}{p(y^T)}.$$

With the posterior in hand, we can evaluate expectations of interest, such as the mean, variance, and quantiles of the parameters, build credible intervals, and perform model comparison. Bayesian methods are particularly attractive to DSGE models because i) they can incorporate prior information about the parameter values (for example, from previous studies or the estimates in other countries); ii) the posterior gives us a much wider assessment of the uncertainty existing in the data than a point estimate and its standard deviation; iii) they deal transparently with lack of (or weak) identification; and iv) they scale well to large dimensions, even in those situations where alternative approaches break down.

Unfortunately, finding the posterior $p(\theta|y^T)$ is usually hard. Since the 1990s, the standard approach to tackling this problem has been to use Markov chain Monte Carlo (MCMC) methods. An MCMC builds a Markov chain such that an empirical distribution coming from simulating it converges to the target density of interest (in our case, the posterior) and, thus, has the same moments. MCMC methods can be used to sample from any target distribution of interest, such as those built to minimize moment conditions or other estimating functions in a frequentist set-up (Chernozhukov and Hong, 2003).

The field of MCMC is so vast that it is impossible even to outline most of the ideas in the area. A standard handbook, Brooks et al. (2011), fills 619 dense pages of material and, yet, it misses important recent developments. Instead, we will introduce the basic MCMC method, the Metropolis-Hastings algorithm, to give a flavor of what MCMCs are about. Next, we will describe the Hamiltonian Monte Carlo, a powerful second-generation MCMC that can deal with highly dimensional problems.

4.1 The Metropolis-Hastings algorithm

Given a state of a Markov chain θ_{i-1}, the Metropolis-Hastings algorithm generates draws θ_i^* from a proposal density $q(\theta_{i-1}, \theta_i^*)$ and accepts or rejects them depending on how $p(y^T|\theta_i^*)$ compares with $p(y^T|\theta_{i-1})$. If the θ_i^* moves the chain toward areas of a higher posterior, we accept the draw, and θ_i^* becomes the new state of the chain. If the θ_i^* moves the chain toward areas of a lower posterior, we accept the draw with probability less than 1. Otherwise, the chain stays at θ_{i-1}. The intuition of why we do this is simple: we always want to travel to areas of higher density, but, if the draw proposes exploring areas of lower density, we should travel to them with some probability to avoid getting stuck at local minima.[3]

The pseudo-code for the Metropolis-Hastings algorithm is:

```
Step 0, Initialization: Set $i \rightsquigarrow 0$ and an initial $\theta_i$. Solve the model for $\theta_i$ and build the state-space representation. Evaluate $p(\theta_i)$ and $p(y^T|\theta_i)$. Set $i \rightsquigarrow i+1$.
Step 1, Proposal draw: Get a draw $\theta_i^*$ from a proposal density $q(\theta_{i-1}, \theta_i^*)$.
Step 2, Solving the model: Solve the model for $\theta_i^*$ and build the state-space representation.
Step 3, Evaluating the proposal: Evaluate $p(\theta_i^*)$ and $p(y^T|\theta_i^*)$.
Step 4, Accept/reject: Draw $\chi_i \sim U(0,1)$. If $\chi_i \leq \frac{p(y^T|\theta_i^*)p(\theta_i^*)q(\theta_{i-1},\theta_i^*)}{p(y^T|\theta_{i-1})p(\theta_{i-1})q(\theta_i^*,\theta_{i-1})}$, set $\theta_i = \theta_i^*$, otherwise $\theta_i = \theta_{i-1}$.
Step 5, Iteration: If $i < I$, set $i \rightsquigarrow i+1$ and go to step 1. Otherwise stop.
```

[3] Notice that the Gibbs sampler is a particular case of the Metropolis-Hastings algorithm. Since the Gibbs sampler is less useful for estimating DSGE models because specifying densities conditional on some parameters is difficult, we will skip a further discussion of it. See, for more details, Casella and George (1992).

Step 4, Accept/reject, is the key to the algorithm. We compute the ratio of the posteriors (after cancelling constants) multiplied by the ratio $\frac{q\left(\theta_{i-1},\theta_i^*\right)}{q\left(\theta_i^*,\theta_{i-1}\right)}$. If the numerator is higher than the denominator, as we explained above, we accept the draw and the chain moves to θ_i^*. Otherwise, we accept θ_i^* only with some probability and, with the complementary probability, we keep the chain at its current location by setting $\theta_i = \theta_{i-1}$.

Every step of the algorithm is simple to code except for the need to specify a proposal density $q\left(\cdot,\cdot\right)$ and to set I, the length of the chain. In fact, the code for a Metropolis-Hastings can be recycled from the estimation of one model to the next with next-to-no changes.

Concerning the proposal density, a standard practice is to choose a random walk proposal, $\theta_i^* = \theta_{i-1} + c\kappa_i$, $\kappa_i \sim \mathcal{N}\left(0, \Sigma_\kappa\right)$, where Σ_κ is a covariance matrix (often, an estimate of the posterior covariance matrix obtained in a preliminary run of the Markov chain) and c is a scaling factor picked to get the appropriate acceptance ratio of proposals (i.e., the percentage of times that the chain moves). Roberts et al. (1997) demonstrate that the asymptotically optimal acceptance rate is 0.234 under quite general conditions. One can hit this acceptance rate by playing with c during tuning runs of the Markov chain.

With respect to I, we can monitor whether the chain has converged by looking at recursive means of the parameter values and checking that those means have stabilized. Fast convergence is easier to obtain if we start the chain from a "good" θ_0, in the sense of being close to the target density's mean. For DSGE models, a good default choice is the values θ_0 that come out of a standard calibration where θ_0 matches some empirical moments. Also, a percentage of the initial draws is often disregarded as a burn-in.

The performance of the RWMH can be improved by "blocking" the parameters. We can partition the parameter vector into j subsets, $\theta = \{\theta^1, ..., \theta^j\}$. We want partitions where the parameters are strongly correlated within blocks and weakly correlated across blocks. Then, we get a proposal for ${\theta^1}_i$ conditional on $\{\theta_{i-1}^2, ..., \theta_{i-1}^j\}$, accept or reject it, get a proposal for ${\theta^2}_i$ conditional on $\{\theta_i^1, ..., \theta_{i-1}^j\}$, accept or reject it and, so on, cycling over all the blocks in a Gibbs-step manner. Blocking reduces the persistence of the RWMH, a serious concern in high-dimensional parameter spaces. Other possibilities to improve performance include adaptive MCMCs (Andrieu and Thoms, 2008; Strid et al., 2010) and gradient and Hessian-based MCMCs (Herbst, 2011).

4.2 The Hamiltonian Monte Carlo

A drawback of the RWMH algorithm is that it spends much time outside the typical set of the posterior, that is, the part of the parameter space containing the relevant information to compute the expectations we care about in Bayesian analysis. To understand this statement, we need to introduce additional notation.

For $\epsilon > 0$ and any I, we define the typical set A_ϵ^I with respect to the target posterior $p(\theta|y^T)$ as:

$$\left\{(\theta_0, \theta_1, ..., \theta_I) : \left| -\frac{1}{I+1}\sum_{i=0}^{I} \log p(\theta_i|y^T) - h(\theta) \right| \leq \epsilon \right\},$$

where $h(\theta) = -\int p(\theta_i|y^T) \log p(\theta_i|y^T) d\theta$ is the differential entropy of the parameters with respect to their posterior density (i.e., a measure of how concentrated or disperse a density is; Cover and Thomas, 2012, ch. 8).

By a weak law of large numbers, $Pr(A_\epsilon^I) > 1 - \epsilon$ for I sufficiently large. That is, A_ϵ^I is "typical" because it includes "most" sequences of θ_i's that are distributed according to $p(\theta|y^T)$. This result shows why A_ϵ^I is, indeed, the relevant region to compute moments of the posterior. Since "most" sequences belong to the typical set, moments of the posterior will depend on those "most" sequences.

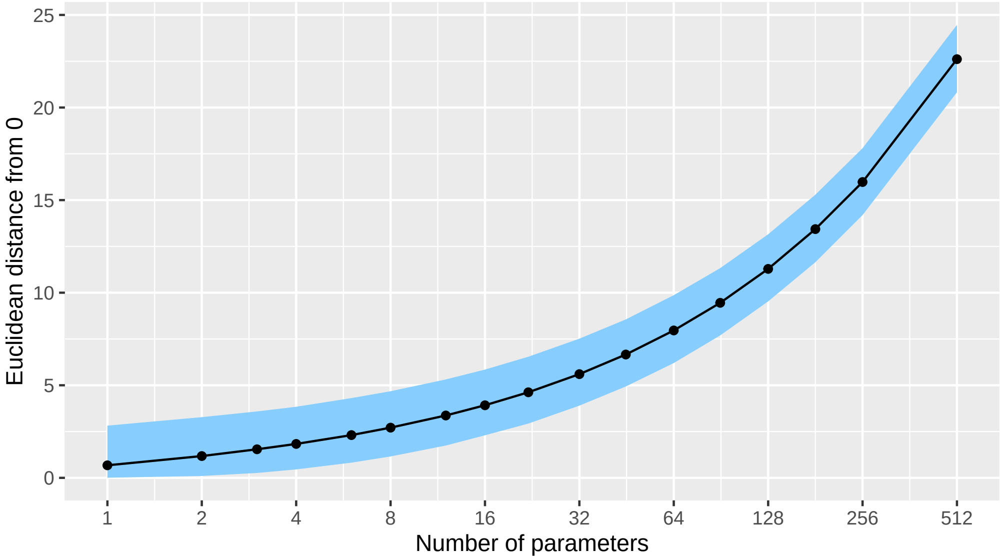

Figure 1: The typical set from a multivariate normal

Two properties of the typical set are surprising but crucial. The first property is that, in general, the typical set is not the region where the posterior density is the highest. To make this point clear, let us assume that the posterior $p(\theta|y^T)$ of our DSGE model is a multivariate normal with zero mean and unit covariance matrix. Let us also vary the dimensionality of θ from 1 to 512 and, for each dimension, draw 100,000 times from $p(\theta|y^T)$. Figure 1 plots the Euclidean distance between a draw and the zero vector, the mode of the posterior for all dimensions.[4] In the circled line, we plot the median distance. The band represents the

[4] Figure 1 is borrowed, with some minor changes, from the excellent explanation of typical sets by Bob Carpenter at `https://mc-stan.org/users/documentation/case-studies/curse-dims.html`.

99% interval of the draws. Clearly, as we increase the dimensionality of the problem, most sequences diverge from the peak of $p(\theta|y^T)$. The second property is that, by concentration of measure, the typical set will be a narrower and narrower band as the number of parameters grows. Again, Figure 1 illustrates this point, as the 99% confidence interval shrinks as dimensions grow.

The RWMH wastes many iterations because the jumps in the proposal density are blind to any information regarding the typical set of the posterior. Thus, most draws of the chain are either away from A_ϵ^I or only induce small movements within A_ϵ^I, $p(\theta|y^T)$ is not explored efficiently, and convergence to the ergodic distribution is slow. This problem is salient when solving the DSGE model is costly and, hence, we cannot run the RWMH for a long time.

A solution to this problem that has gained much popularity during the last decade is the Hamiltonian Monte Carlo (HMC). This method is an MCMC algorithm that improves efficiency by exploiting information from the posterior's gradient and using it to force the Markov chain to spend more time in the typical set (Neal, 2011).

However, we cannot use the gradient of the posterior directly, because it would push the jumps toward the mode of the posterior and stay there, away from the typical set. Instead, the HMC supplements the gradient with an extra force, *momentum*, so that the jumps gravitate around the mode and stay inside the typical set, even if the jump size is large. Thus, we can reduce the correlation between successive values of the parameters in the Markov chain while keeping a high acceptance probability.

It turns out that this idea –a pull toward a center counteracted by a momentum– also appears in many physical contexts and, thus, we can use the framework of Hamiltonian mechanics designed to study these dynamics (Betancourt, 2017).

In particular, we augment the vector of parameters, $\theta \in \mathbb{R}^d$, with an auxiliary momentum variable $\mathbf{p} \in \mathbb{R}^d$ with density $\mathcal{N}(0, M)$. Then, the Hamiltonian associated with the posterior of the DSGE model is:

$$\mathcal{H}(\theta, \mathbf{p}) = -\log p(\theta|y^T) + \frac{1}{2} log((2\pi)^d |M|) + \frac{1}{2}\mathbf{p}' M^{-1}\mathbf{p}.$$

where $-\log p(\theta|y^T)$ is the analog of a potential energy function, $\frac{1}{2}\mathbf{p}'M^{-1}\mathbf{p}$ is a kinetic energy term, and $\frac{1}{2}log((2\pi)^d|M|)$ is a scaled mass matrix. A Markov chain $\{\theta_i, \mathbf{p}_i\}_{i=1}^I$ drawn from this Hamiltonian has a stationary marginal density $p(\theta|y^T)$.

How do we sample from $\mathcal{H}(\theta, \mathbf{p})$? Girolami and Calderhead (2011) propose a Gibbs-sampling scheme. In the first step, we draw from the normal distribution that we have specified for $\mathbf{p}$,

$$\mathbf{p}_{i+1} \sim p(\mathbf{p}_{i+1}|\theta_i) = \mathcal{N}(0, M).$$

This step is straightforward with standard software. In the second step, we draw:

$$\theta_{i+1} \sim p(\theta_{i+1}|\mathbf{p}_{i+1}).$$

To do so, we start from $\mathbf{p}(0) = \mathbf{p}_{i+1}$ and $\theta(0) = \theta_i$ and run the Störmer-Verlet integrator for L iterations:

$$\begin{aligned}\mathbf{p}(\tau + \epsilon/2) &= \mathbf{p}(\tau) + \epsilon\nabla_\theta \log p(\theta(\tau)|y^T) & (14)\\ \theta(\tau + \epsilon) &= \theta(\tau) + \epsilon M^{-1}\mathbf{p}(\tau + \epsilon/2) & (15)\\ \mathbf{p}(\tau + \epsilon) &= \mathbf{p}(\tau + \epsilon/2) + \epsilon\nabla_\theta \log p(\theta(\tau + \epsilon))/2. & (16)\end{aligned}$$

Steps (14) and (15) can be combined to form a discrete preconditioned Langevin diffusion. We call the values $(\theta^*, \mathbf{p}^*)$ at the end of the L-th iteration a proposal. Then, we apply a Metropolis step, where the probability of keeping the proposal $(\theta^*, \mathbf{p}^*)$ is:

$$\min(1, exp(\mathcal{H}(\theta_i, \mathbf{p}_{i+1}) - \mathcal{H}(\theta^*, \mathbf{p}^*))).$$

To implement the HMC, we need values for L, ϵ, and M. L and ϵ can be fine-tuned in each application. If L is too small, the HMC behaves like a RWMH. If L is too large, the algorithm wastes computations. To avoid this fine-tuning, Homan and Gelman (2014) propose the No-U-Turn Sampler (NUTS), a recursive algorithm that builds a set of likely candidate points that spans a wide swath of $p(\theta|y^T)$ and stops when it starts to retrace its steps (hence, its name). With respect to M, Neal (2011) suggests using $M^{-1} = \hat{\Sigma}$, which implies that the momentum variables have covariance $\hat{\Sigma}^{-1}$. One can obtain $\hat{\Sigma}$ by running a RWMH for a few steps and get an approximate shape of the posterior.

But the real bottleneck of the HMC is that, in each iteration of the Störmer-Verlet integrator, we need to evaluate the gradient $\nabla_\theta \log p(\theta|y^T)$ twice. Sometimes this is simple. For instance, if we have a linearized DSGE model, the Kalman filter gives us not only an evaluation of the likelihood but also an easy-to-evaluate gradient (Watson, 1989). However, there are also many situations where evaluating the gradient is too costly or unfeasible. For example, the BPF that we described above gives an evaluation of the likelihood function that is not differentiable with respect to the parameters because, by changing the parameter values by a small ϵ, we would potentially change the resampling of particles.

Strathmann et al. (2015) have proposed the kernel Hamiltonian Monte Carlo to solve this bottleneck. This sampler is an adaptive MCMC algorithm that "learns" the gradient structure by using a surrogate function, f, and the history of the Markov chain, which can be obtained from an initial RWMH run. The surrogate must satisfy $\nabla f \approx \nabla_\theta \log p(\theta|y^T)$.

In practice, suppose we have access to a random sample $\{\theta_i\}_{i=1}^I$ from the Markov chain and let $k(\theta_i, x) = \exp(-\sigma^{-1}||\theta_i - x||_2^2)$ denote the Gaussian kernel. Then, $f(x) = \sum_{i=1}^I \alpha_i k(\theta_i, x)$, where the parameters α_i are obtained by minimizing the empirical score matching objective. Strathmann et al. (2015) show that $\hat{\alpha} = -\frac{\sigma}{2}(C + \lambda \mathbb{I})^{-1}b$, where b and C are parameters that depend on the kernel, $\mathbb{I}$ is the identity matrix, and λ a regularization parameter. The approximated gradient can be easily computed and, with it, we can run the Störmer-Verlet integrator to obtain the proposal $(\theta^*, \mathbf{p}^*)$.

In current work, Childers et al. (2020) show to apply the HMC to DSGE models, in particular when we have a state-space representation that is differentiable.

5 Variational inference

Variational inference (VI) is an approach to statistical computations where we replace the probability distributions implied by a model of interest (a DSGE or any other econometric model) with tractable and easy-to-evaluate functions. This goal is crucial in the Bayesian approach, where the model's posterior density can be unwieldy. While VI was developed in the machine learning community, the main idea has broad applicability in econometrics. Experience has shown that, in many problems, VI can be faster and easier to scale than MCMC, although its theoretical properties are less well understood (Blei et al., 2017).

VI builds on the tradition of variational methods, a class of approximation techniques in applied mathematics that convert a complex model into a simpler one. Suppose, for example, that we want to compute the logarithm of x. This operation is costly, and we might want to avoid it due to computational constraints (for instance, because we do it repeatedly in a loop). A variation approach replaces the original nonlinear problem with a transformed optimization problem that is linear on x: $\ln(x) = \min_\lambda [\lambda x - \ln(\lambda) - 1]$. The first-order condition for this problem is $\lambda = x^{-1}$. Direct substitution shows the equivalence between the two problems. There are situations where solving the minimization problem (or getting a good approximation to it) can be faster than evaluating $\ln(x)$.

To extend this idea to the estimation of DSGE models, denote the joint distribution of the model by $p(y^T, \theta) = p(y^T|\theta) \times p(\theta)$, with a gradient $\nabla_\theta \log p(y^T, \theta)$. Also, consider a situation where evaluating the likelihood of the model or drawing from it is cumbersome.

Instead of dealing with the posterior density $p(\theta|y^T)$, VI works with an approximating density $q(\theta; \phi)$ that is easier to handle. More formally, VI looks for $q(\cdot; \phi)$ by minimizing the Kullback-Leibler (KL) divergence between $q(\cdot; \phi)$ and $p(\theta|y^T)$:

$$KL\left(q(\theta;\phi)||p(\theta|y^T)\right) = \mathbb{E}_{q(\theta)}[\log q(\theta;\phi)] - \mathbb{E}_{q(\theta)}[\log p(\theta|y^T)],$$

with respect to the auxiliary approximation parameters ϕ.

By Bayes' theorem:

$$p(\theta|y^T) = \frac{p(y^T|\theta)p(\theta)}{p(y^T)} = \frac{p(y^T,\theta)}{p(y^T)},$$

we get:

$$KL\left(q(\theta;\phi)||p(\theta|y^T)\right) = \mathbb{E}_{q(\theta)}[\log q(\theta;\phi)] - \mathbb{E}_{q(\theta)}[\log p(y^T,\theta)] + \log p(y^T), \tag{17}$$

since, given some data, $p(y^T)$ is a constant.

Tackling equation (17) is not feasible because it requires the evaluation of $p(y^T)$, the marginal distribution of y^T. Instead, we can switch signs, drop $p(y^T)$, and maximize the proxy:

$$\mathcal{L} = \mathbb{E}_{q(\theta)}[p(y^T,\theta)] - \mathbb{E}_{q(\theta)}[\log q(\theta;\phi)]. \tag{18}$$

with respect to ϕ. Clearly, $\mathcal{L} = -KL\left(q(\theta;\phi)||p(\theta|y^T)\right) + \log p(y^T)$, that is, $\mathcal{L}$ is the negative KL divergence plus $p(y^T)$, a term that is independent of ϕ. $\mathcal{L}$ is called the evidence lower bound, or ELBO for short (also known as the variational lower bound). Its name derives from the fact that $\mathcal{L}$ provides a lower bound for the marginal likelihood of the data. Since, by Gibbs' inequality, $KL(q(\theta;\phi)||p(\theta|y^T) \geq 0$, we must have that $\log p(y^T) \geq \mathcal{L}$.

VI proceeds by maximizing $\mathcal{L}$ subject to $supp(q(\theta;\phi)) \subseteq supp(p(\theta|y^T)$. Once we have $q(\cdot;\phi^*)$, where ϕ^* is the argmax of $\mathcal{L}$, we can employ it like any other posterior. In other words, while MCMC algorithms are built around the idea of sampling the posterior by building a Markov chain with the appropriate ergodic distribution, VI focuses on optimizing an approximation to such a posterior.

Unfortunately, maximizing the ELBO is not straightforward because it demands the evaluation of expectations in equation (18). Also, if we are using a derivative-based optimization algorithm (such as a Quasi-Newton), we need $\nabla_\theta \log p(y^T,\theta)$.[5]

The literature has provided several alternatives for working around the ELBO (Blei et al., 2017). We focus on an option usable in state-space representations (Archer et al., 2015). Let $p(s|y;\theta)$ denote the posterior distribution of states of the DSGE model that we get from filtering its state-space representation and let $q(s|y;\phi)$ be its tractable approximation.

Maximizing (18) with respect to ϕ to find $q(\cdot;\phi)$ will usually require computing:

$$\nabla \mathbb{E}_{q(s|y;\phi)}[\log p(s,y|\theta) - \log q(s|y;\phi)] \tag{19}$$

as an input into a derivative-based optimization algorithm. To do so, Archer et al. (2015)

[5] If the model is not differentiable –or its derivatives are too cumbersome to evaluate–, one can think about non-derivative-based optimization algorithms such as the Nelder-Mead method or genetic approaches.

advocate the use of the "reparameterization" trick: $z = g(s, \epsilon; \phi)$. Here, ϵ is an easy-to-sample random variable with distribution $p(\epsilon)$. For example, one could use:

$$x = \mu(y|\phi) + \Sigma(y|\phi)\epsilon, \tag{20}$$

where ϵ is drawn from a multivariate normal with mean 0 and the identity covariance matrix. We can select flexible functions for the $\mu(\cdot)$ and $\Sigma(\cdot)$ such as Chebyshev polynomials or neural networks to capture intricate behaviors of the posterior densities that we want to approximate. Since neural networks are universal nonlinear approximators (Barron, 1993) and break the curse of dimensionality (Bach, 2017), they are a particularly attractive choice for this step.

With equation (20), we can draw L samples from $p(\epsilon)$ and compute a stochastic approximation to the gradient $\nabla \mathbb{E}_{q(z|x;\phi)}[f_{\{\theta,\phi\}}] \approx \frac{1}{L}\sum_{\ell=1}^{L} \nabla f_{\{\theta,\phi\}}(g(s, \epsilon^{\ell}; \phi))$, where $f_{\{\theta,\phi\}}(s) = \log p(s, y|\theta) - \log q(s|y; \phi)$. With this approximated gradient, we can search for the duple $\{\theta, \phi\}$ that maximizes the ELBO.

6 Our application

To illustrate our arguments, we estimate a canonical DSGE model.[6] We have a representative household that consumes, saves, holds money, and supplies labor. A final good firm produces output using a continuum of intermediate goods. These intermediate goods, in turn, are produced by monopolistic competitors subject to nominal price rigidities à la Calvo. The representative household is the owner of all of these firms. The model is closed with a government that sets up monetary and fiscal policy. Since there are trends in the data, we introduce two unit roots, one in the level of neutral technology and one in the investment-specific technology, that induce stochastic long-run growth.

6.1 The model

The representative household. The representative household has a lifetime utility function on consumption, c_{jt}, real money balances, m_{jt}/p_t (where p_t is the price level), and hours worked, l_{jt}:

$$\mathbb{E}_0 \sum_{t=0}^{\infty} \beta^t d_t \left\{ \log\left(c_t - hc_{t-1}\right) + \upsilon \log\left(\frac{m_t}{p_t}\right) - \varphi_t \psi \frac{l_t^{1+\vartheta}}{1+\vartheta} \right\}. \tag{21}$$

In equation (21), β is the discount factor, h controls habit persistence, ϑ is the inverse of the Frisch labor supply elasticity, d_t is an intertemporal preference shock with law of motion

[6] The reader can find additional details regarding the model and the algebraic derivations at https://www.sas.upenn.edu/~jesusfv/ARE_DSGE.pdf.

$\log d_t = \rho_d \log d_{t-1} + \sigma_d \varepsilon_{d,t}$, where $\varepsilon_{d,t} \sim \mathcal{N}(0,1)$ and φ_t is a labor supply shock with law of motion $\log \varphi_t = \rho_\varphi \log \varphi_{t-1} + \sigma_\varphi \varepsilon_{\varphi,t}$, where $\varepsilon_{\varphi,t} \sim \mathcal{N}(0,1)$. We should understand both shocks as a stand-in for more complex mechanisms, such as financial frictions, demographic shifts, or changes in risk attitudes.

The household's budget constraint is:

$$\begin{aligned} & c_t + x_t + \frac{m_t}{p_t} + \frac{b_{t+1}}{p_t} + \int q_{t+1,t} a_{t+1} d\omega_{t+1,t} \\ = {} & w_t l_t + \left(r_t u_t - \mu_t^{-1} \Phi\left[u_t\right] \right) k_{t-1} + \frac{m_{t-1}}{p_t} + R_{t-1} \frac{b_t}{p_t} + a_t + T_t + \digamma_t . \end{aligned}$$

In terms of uses –the left-hand side of equation (6.1)– and beyond consumption, the household can save in physical capital, by investing x_t in new capital, holding real money balances, purchasing government debt, b_t, and trading in Arrow securities. More concretely, a_{t+1} is the amount of those securities that pays one unit of consumption in event $\omega_{t+1,t}$ purchased at time t at (real) price $q_{t+1,t}$. This price will be such that, in equilibrium, the net supply of the Arrow securities would be zero.

In terms of resources –the right-hand side of equation (6.1)– the household gets income by renting its labor supply at the real wage w_t and its capital at real rental price r_t. The household chooses the utilization rate of capital, $u_t > 0$, given the depreciation cost $\mu_t^{-1}\Phi\left[u_t\right]$, where μ_t is an investment-specific technological shock that we will introduce below and $\Phi = \gamma_1(u-1) + \frac{\gamma_2}{2}(u-1)^2$. We interpret $u_t = 1$ as the "normal" utilization rate. The household also has access to its money balances, the government debt (with a nominal gross interest rate of R_t), the Arrow security that pays in the actually realized event, the lump-sum transfers (of taxes if negative) from the government, T_t, and the profits of the economy's firms, $\digamma_t$.

Capital follows $k_t = \left(1 - \delta \Phi\left[u_t\right]\right) k_{t-1} + \mu_t \left(1 - S\left[\frac{x_t}{x_{t-1}}\right]\right) x_t$, where δ is the depreciation rate when $u_t = 1$ and $S\left[\cdot\right]$ is an adjustment cost function. We assume that $S\left[\Lambda_x\right] = 0$, $S'\left[\Lambda_x\right] = 0$, and $S''\left[\cdot\right] > 0$ where Λ_x is the growth rate of investment along the balanced growth path (BGP) of the economy.

The investment-specific technological shock evolves as $\mu_t = \mu_{t-1} \exp\left(\Lambda_\mu + z_{\mu,t}\right)$ where $z_{\mu,t} = \sigma_\mu \varepsilon_{\mu,t}$ and $\varepsilon_{\mu,t} \sim \mathcal{N}(0,1)$, where μ_t is, in equilibrium, the inverse of the price of new capital in consumption terms.

For future reference, define λ_t as the Lagrangian multiplier associated with the budget constraint, Q_t the Lagrangian multiplier associated with installed capital, and the (marginal) Tobin's Q as $q_t = \frac{Q_t}{\lambda_t}$.

The final good producer. A perfectly competitive final good producer combines intermediate goods with the technology $y_t^d = \left(\int_0^1 y_{it}^{\frac{\varepsilon-1}{\varepsilon}} di\right)^{\frac{\varepsilon}{\varepsilon-1}}$, where ε is the elasticity of substitution.

Given the intermediate goods prices p_{ti} and the final good price p_t, the demand function for every intermediate good i is $y_{it} = \left(\frac{p_{it}}{p_t}\right)^{-\varepsilon} y_t^d$, where y_t^d is the aggregate demand in the economy (to be defined below). Integrating over i and using the zero profit condition for the final good producer, we get $p_t = \left(\int_0^1 p_{it}^{1-\varepsilon} di\right)^{\frac{1}{1-\varepsilon}}$.

Intermediate good producers. Each intermediate good producer i has a technology $y_{it} = A_t k_{it-1}^{\alpha} \left(l_{it}^d\right)^{1-\alpha}$, where k_{it-1} is the capital rented by the firm, l_{it}^d is the labor input rented, and A_t is the technology level that evolves as $A_t = A_{t-1} \exp\left(\Lambda_A + z_{A,t}\right)$, where $z_{A,t} = \sigma_A \varepsilon_{A,t}$ and $\varepsilon_{A,t} \sim \mathcal{N}(0,1)$.

Also, define $z_t = A_t^{\frac{1}{1-\alpha}} \mu_t^{\frac{\alpha}{1-\alpha}}$, a variable that encodes the joint effect of both technology shocks along the BGP of the economy. Simple algebra tells us that $z_t = z_{t-1} \exp\left(\Lambda_z + z_{z,t}\right)$ where $z_{z,t} = \frac{z_{A,t} + \alpha z_{\mu,t}}{1-\alpha}$ and $\Lambda_z = \frac{\Lambda_A + \alpha \Lambda_\mu}{1-\alpha}$.

Intermediate good producers solve a two-stage problem. First, taking the input prices w_t and r_t as given, firms rent l_{it}^d and k_{it-1} to minimize real cost. This problem delivers a marginal cost:

$$mc_t = \left(\frac{1}{1-\alpha}\right)^{1-\alpha} \left(\frac{1}{\alpha}\right)^{\alpha} \frac{w_t^{1-\alpha} r_t^{\alpha}}{A_t}$$

The marginal cost is the same for all intermediate good producers, as they all have access to the same technology and take input prices as given.

Second, intermediate good producers choose the price that maximizes discounted real profits subject to Calvo pricing. In each period, a fraction $1 - \theta_p$ of firms can change their prices. All other firms only index their prices by past inflation, $\Pi_{t-1} = \frac{p_{t-1}}{p_{t-2}}$ by a degree $\chi_p \in [0,1]$, where $\chi_p = 0$ is no indexation and $\chi_p = 1$ is total indexation. The marginal value of a dollar of future profits, $\lambda_{t+\tau}/\lambda_t$, comes from the fact that the household owns the firm and we have complete markets. For this problem to be well posed, we assume that $\left(\beta\theta_p\right)^{\tau} \lambda_{t+\tau}$ goes to zero sufficiently fast in relation to inflation.

After some algebra, this pricing problem is characterized by two auxiliary variables:

$$g_t^1 = \lambda_t mc_t y_t^d + \beta\theta_p \mathbb{E}_t \left(\frac{\Pi_t^{\chi}}{\Pi_{t+1}}\right)^{-\varepsilon} g_{t+1}^1,$$

$$g_t^2 = \lambda_t \Pi_t^* y_t^d + \beta\theta_p \mathbb{E}_t \left(\frac{\Pi_t^{\chi}}{\Pi_{t+1}}\right)^{1-\varepsilon} \left(\frac{\Pi_t^*}{\Pi_{t+1}^*}\right) g_{t+1}^2,$$

where $\Pi_t^* = \frac{p_t^*}{p_t}$ is the ratio of the reset price of firms that can change their prices over p_t. Given the Calvo assumption, the price index evolves as $1 = \theta_p \left(\frac{\Pi_{t-1}^{\chi}}{\Pi_t}\right)^{1-\varepsilon} + \left(1-\theta_p\right) \Pi_t^{*1-\varepsilon}$.

The government. The economy has a government that determines monetary and fiscal policy. In terms of monetary policy, the government sets the nominal interest rate following a Taylor rule:

$$\frac{R_t}{R} = \left(\frac{R_{t-1}}{R}\right)^{\gamma_R} \left(\left(\frac{\Pi_t}{\Pi}\right)^{\gamma_\Pi} \left(\frac{\frac{y_t^d}{y_{t-1}^d}}{\Lambda_{y^d}}\right)^{\gamma_y}\right)^{1-\gamma_R} \exp\left(m_t\right)$$

through open market operations. The variable Π is the target level of inflation (equal to inflation in the BGP), R is the BGP nominal gross return of capital (equal to the BGP real gross returns of capital plus Π), and Λ_{y^d} is the BGP growth rate of y_t^d. The term $m_t = \sigma_m \varepsilon_{mt}$ is a random shock to monetary policy, where $\varepsilon_{mt} \sim \mathcal{N}(0,1)$. In terms of fiscal policy, government consumes $g_t = \widetilde{g}_t z_t$, where $\widetilde{g}_t$ follows $\log \widetilde{g}_t = (1-\rho_g)\log g + \rho_g \log \widetilde{g}_{t-1} + \sigma_g \varepsilon_{g,t}$, where $\varepsilon_{g,t} \sim \mathcal{N}(0,1)$.

Aggregation. Aggregate demand is $y_t^d = c_t + g_t + x_t + \mu_t^{-1}\Phi\left[u_t\right]k_{t-1}$. Aggregate supply is $y_t^s = \frac{1}{v_t^p} A_t \left(u_t k_{t-1}\right)^\alpha \left(l_t^d\right)^{1-\alpha}$, where $v_t^p = \int_0^1 \left(\frac{p_{it}}{p_t}\right)^{-\varepsilon} di$ is the loss of output created by the price rigidities and the resulting misallocation of inputs. By the properties of Calvo's pricing, $v_t^p = \theta_p \left(\frac{\Pi_{t-1}^\chi}{\Pi_t}\right)^{-\varepsilon} v_{t-1}^p + \left(1-\theta_p\right)\Pi_t^{*-\varepsilon}$.

Equilibrium and solution. A definition of equilibrium in this economy is standard and we can skip it. Also, since we have two unit roots in the model, we need to make the model stationary before solving it. To do so, we define $\widetilde{c}_t = \frac{c_t}{z_t}$, $\widetilde{\lambda}_t = \lambda_t z_t$, $\widetilde{r}_t = r_t \mu_t$, $\widetilde{q}_t = q_t \mu_t$, $\widetilde{x}_t = \frac{x_t}{z_t}$, $\widetilde{w}_t = \frac{w_t}{z_t}$, $\widetilde{w}_t^* = \frac{w_t^*}{z_t}$, $\widetilde{k}_t = \frac{k_t}{z_t \mu_t}$, and $\widetilde{y}_t^d = \frac{y_t^d}{z_t}$ and re-write all of the relevant equilibrium conditions accordingly. Next, we solve for the steady state of these rescaled equilibrium conditions and we log-linearize them around such a steady state (Fernández-Villaverde et al., 2016).

The endogenous states are $state_t = \left(\widehat{\Pi}_t, \widehat{g}_t^1, \widehat{g}_t^2, \widehat{\widetilde{k}}_t, \widehat{R}_t, \widehat{\widetilde{y}}_t^d, \widehat{\widetilde{c}}_t, \widehat{v}_t^p, \widehat{\widetilde{q}}_t, \widehat{\widetilde{x}}_t, \widehat{\widetilde{\lambda}}_t, \widehat{\widetilde{z}}_t\right)'$, the exogenous states are $exo_t = \left(z_{\mu,t}, \widehat{d}_t, \widehat{\varphi}_t, z_{A,t}, m_t\right)'$, and the shocks are $\varepsilon_t = \left(\varepsilon_{\mu,t}, \varepsilon_{d,t}, \varepsilon_{\varphi,t}, \varepsilon_{A,t}, \varepsilon_{m,t}, \varepsilon_{g,t}\right)'$. Then, after log-linearization, we have:

$$state_t = PP * state_{t-1} + QQ * exo_t \tag{22}$$

and

$$exo_t = NN * exo_{t-1} + \Sigma^{1/2} * \varepsilon_t. \tag{23}$$

where PP, QQ, NN, and Σ are matrices that involve complex nonlinear relations of the parameters of the model and that we obtain from solving the model. Stacking equations (22) and (23) together gives us the a transition equation (5).

6.2 Estimation

Since we have six exogenous shocks (two demand shocks to preferences, two supply shocks to technology, a monetary policy shock, and a fiscal policy shock), we select six time series to match: inflation, the federal funds rate, real wages growth, output per capita growth, per capita hours worked, and the inverse relative price of investment with respect to the price of consumption growth.[7] Then $\mathbb{Y}_t = \left(\log \Pi_t, \log R_t, \triangle \log w_t, \triangle \log y_t, \log l_t, \triangle \log \mu_t^{-1}\right)'$, and the measurement equation is:

$$\mathbb{Y}_t = CCstate_t + DDexo_t. \tag{24}$$

Researchers have some degrees of freedom in determining which series to use for estimation. Since the choice is consequential (Guerrón-Quintana, 2010), we should pick time series that are informative about the parameters of interest. Selecting these time series requires a combination of trial-and-error and experience.

Param	Description	Domain	Density	Mean	SD
Steady-state-related parameters					
$100(1/\beta - 1)$	β is discount factor	$(0,1)$	Gamma	0.25	0.1
g	SS govt expenditure/GDP	$(0,1)$	Beta	0.3	0.1
$100(\Pi^* - 1)$	Target inflation	$\mathbb{R}$	Gamma	0.95	0.1
Endogenous propagation parameters					
h	Habit persistence	$(0,1)$	Beta	0.7	0.1
α	Cobb-Douglas labor	$(0,1)$	Normal	0.3	0.025
κ	Investment adjustment cost	$\mathbb{R}$	Normal	4	1.5
θ_P	Fraction of firms with fixed prices	$(0,1)$	Beta	0.5	0.1
χ_P	Price indexation	$(0,1)$	Beta	0.5	0.15
γ_R	Taylor rule coefficient past rates	$(0,1)$	Beta	0.75	0.1
γ_Π	Taylor rule coefficient inflation	$\mathbb{R}_+$	Normal	1.5	0.25
γ_Y	Taylor rule coefficient demand	$\mathbb{R}_+$	Normal	0.12	0.05
Exogenous shocks parameters					
ρ_D	Persistence demand shock	$(0,1)$	Beta	0.5	0.2
ρ_ϕ	Persistence labor supply shock	$(0,1)$	Beta	0.5	0.2
ρ_G	Persistence govt consumption shock	$(0,1)$	Beta	0.5	0.2
Λ_μ	Long-run growth investment specific	$\mathbb{R}$	Gamma	0.0034	0.001
Λ_A	Long-run growth productivity	$\mathbb{R}$	Gamma	0.00178	0.00075
σ_D	s.d. demand shock innovation	$\mathbb{R}_+$	InvGamma	0.1	1
σ_ϕ	s.d. labor supply shock innovation	$\mathbb{R}_+$	InvGamma	0.1	1
σ_G	s.d. govt consumption shock innovation	$\mathbb{R}_+$	InvGamma	0.1	1
σ_μ	s.d. investment-specific shock	$\mathbb{R}_+$	InvGamma	0.1	1
σ_A	s.d. long-run productivity	$\mathbb{R}_+$	InvGamma	0.1	1
σ_M	s.d. monetary shock	$\mathbb{R}_+$	InvGamma	0.1	1

Table 1: Prior Distribution for Structural Parameters

[7] Data come from the Federal Reserve Bank of St. Louis' FRED database, at a quarterly frequency from 1959:Q1 to 2016:Q1. More than six observables will make the model stochastically singular, i.e., it would give zero probability to nearly all observations. If we wanted to use more observables, we could add other shocks or introduce measurement errors. After all, NIPA is imprecise because of the limitations in the resources of statistical agencies and the conceptual difficulties involved in measuring many goods and services.

We estimate the model with Bayesian methods, using the Kalman filter to evaluate the likelihood function and a single-block RWMH algorithm to generate draws from the posterior distribution. The priors we use are summarized in Table 1. Because of space constraints, we do not have time to discuss them in detail beyond pointing out that they are conventional. We calibrate $\vartheta = 1$, $\gamma_2 = 0.01$, $\delta = 0.025$, and $\epsilon = 10$ because these parameters are poorly identified without microdata (γ_1 disappears when we log-linearize the model). Also, notice that only one parameter κ of the investment-adjustment function $S[\cdot]$ appears in the log-linearized solution

We start our chain from the prior mean and variance, adjust tuning constant to get an acceptance rate of around 30%. We run the chain 3,000,000 times and discard the first 10% of draws as a burn-in. We then calculate the posterior mean and in-sample variance-covariance matrix and rerun the RWMH using the new mode and variance for 2,000,000 times (again, with acceptance rate around 30% and discard the first 10% of draws). Finally, we the fix structural parameters at their posterior mean, and run RWMH only on exogenous process parameters for 2,000,000 times (again, with acceptance rate around 30% and discard the first 10% of draws). One can think about this last run as a Gibbs step to improve the convergence of these parameters.

Figure 2 plots the marginal posteriors for the structural parameters not related to the exogenous processes (with vertical lines to denote the mean). As usual in similar exercises, we estimate a high discount factor, large adjustment costs, and high price indexation. In terms of the parameters in the Taylor rule, monetary policy targets an average inflation of a bit above 2.3% per year, and it satisfies Taylor's principle (i.e., $\gamma_\Pi > 1$).

Figure 3 plots the marginal posteriors for the parameters in the exogenous processes. Among the most interesting results, we find volatile preference shocks and a long-lived government expenditure shock. These two findings hint at the importance of demand considerations in our estimated DSGE model.

If we had more space, we could analyze these results in detail, including an assessment of the robustness of the results with respect to our priors, the implications for variance decompositions, the study of moments and IRFs, forecasting, policy counterfactuals, welfare analysis, and model comparison. We hope that, at least, the reader will have a feeling of the wide variety of outputs one can obtain from this class of estimation exercises.

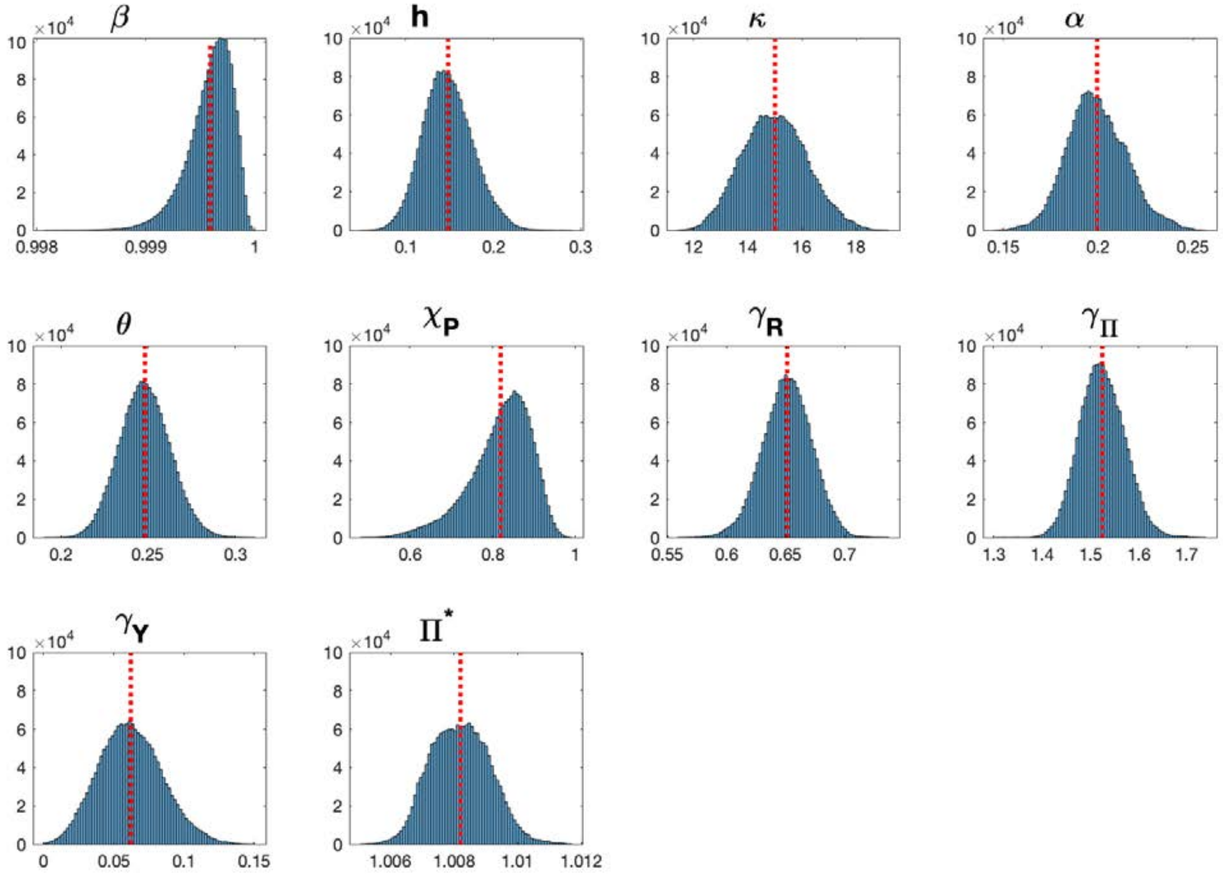

Figure 2: Posterior distribution for structural parameters

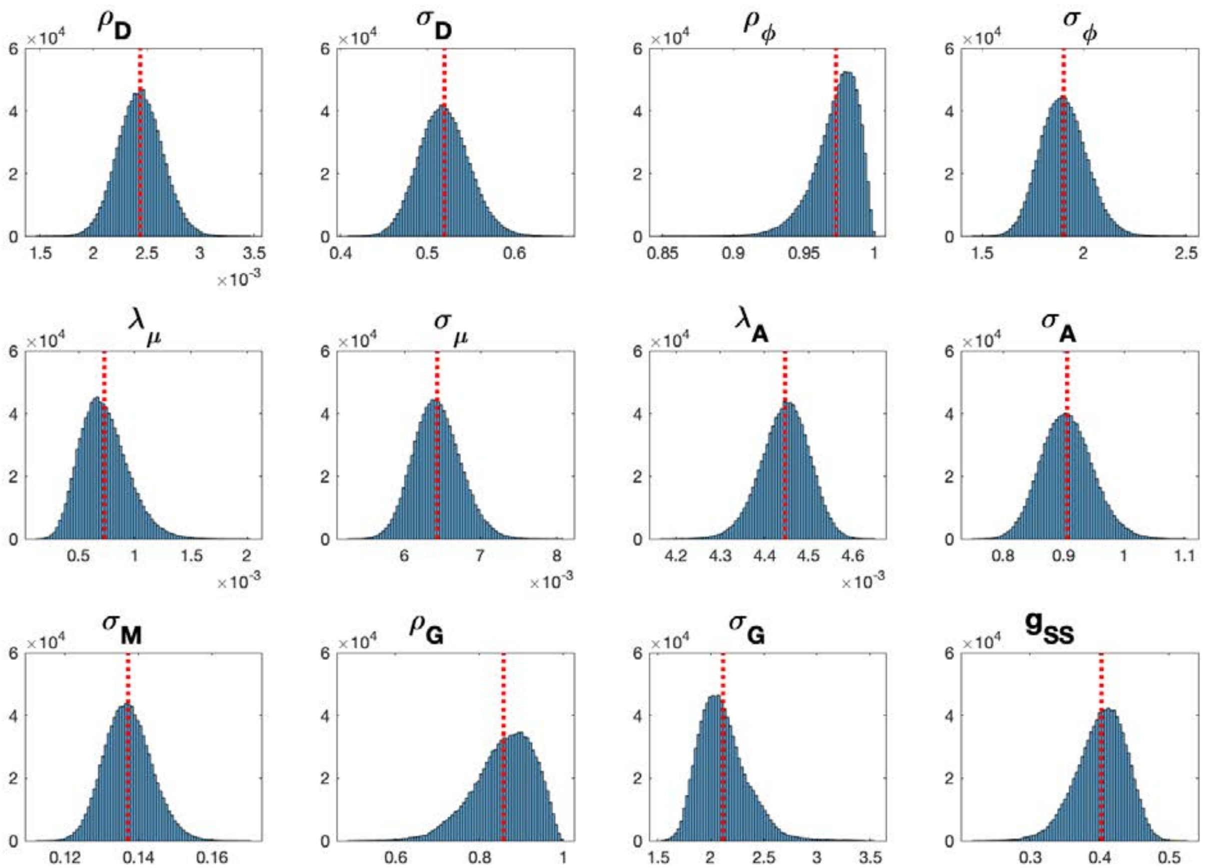

Figure 3: Posterior distribution for exogenous processes only, keeping other parameters fixed

7 Future challenges

The estimation of DSGE models is a vibrant area of research that faces many open challenges. We will conclude by outlining three of them.

First, during the last few years, many macroeconomists have moved from DSGE models in discrete time to models in continuous time (Brunnermeier and Sannikov, 2014, and Achdou et al., 2017). Continuous-time models are often easier to solve and there are powerful mathematical results that apply to them, such as the theory of stochastic differential equations. At the same time, there has been little work done in the structural estimation of DSGE models in continuous time, perhaps because researchers have to deal with time aggregation issues (data come in discrete time units, such as the GDP for 2020:Q2). Fernández-Villaverde et al. (2019) show how to take advantage of the mathematical structure of a continuous-time model to build its associated likelihood with next-to-no computational effort.

Second, DSGE models with heterogeneous agents, such as the HANK class pioneered by Gornemann et al. (2012) and Kaplan et al. (2018), are undergoing a boom. The estimation of these models is difficult because solving them is computationally hard. This is an area, therefore, where some of the ideas introduced in the survey, such as ABC or variational inference, can be a promising area of exploration.

Related to this point is the third challenge for DSGE models: the incorporation of machine learning methods (Goodfellow et al., 2016). Those algorithms can be used in at least two ways. The first way is as a solution method. Machine learning is particularly suitable for approximating high-dimensional functions (such as the state-space representation (1) and (2)) efficiently. For example, deep neural networks break the curse of dimensionality (Bach, 2017) and, therefore, can tackle large DSGE economies. Fernández-Villaverde et al. (2019), Maliar et al. (2019), and Azinović et al. (2020) are recent examples of how to apply these ideas to solve DSGE models with heterogeneous agents.

The second way is to "process" unstructured data (such as text, satellite images, social media activity) and use them as additional observables in DSGE models. For instance, statements about monetary policy from central banks can carry information about the expectations of agents in the economy that are difficult to tease out of NIPA data or even from other rich data sets as in Boivin and Giannoni (2006). The work by Casella et al. (2020) is an illustration of how the estimation of structural DSGE models with unstructured data can be accomplished by merging techniques described in this survey and a Latent Dirichlet allocation for text data in an augmented state-space representation. The posterior distribution of parameters from the resulting representation can be sampled with an MCMC algorithm, and it is readily amenable to massive parallelization.

We hope to see these challenges tackled during the next few years. The combination of better computers, better methods, and better data makes us optimistic about the crop of papers that we will see during the 2020s.

References

Achdou, Y., J. Han, J.-M. Lasry, P.-L. Lions, and B. Moll (2017): "Income and Wealth Distribution in Macroeconomics: A Continuous-Time Approach," Working Paper 23732, National Bureau of Economic Research.

Andreasen, M., J. Fernández-Villaverde, and J. Rubio-Ramírez (2018): "The Pruned State-Space System for Non-Linear DSGE Models: Theory and Empirical Applications," *Review of Economic Studies*, 85, 1–49.

Andrieu, C. and J. Thoms (2008): "A Tutorial on Adaptive MCMC," *Statistical Computing*, 18, 343–373.

Archer, E., I. M. Park, L. Buesing, J. Cunningham, and L. Paninski (2015): "Black Box Variational Inference for State Space Models," Tech. rep., Columbia University.

Aruoba, S. B., J. Fernández-Villaverde, and J. F. Rubio-Ramírez (2006): "Comparing Solution Methods for Dynamic Equilibrium Economies," *Journal of Economic Dynamics and Control*, 30, 2477–2508.

Azinović, M., L. Gaegauf, and S. Scheidegger (2020): "Deep Equilibrium Nets," Mimeo, University of Zurich, https://ssrn.com/abstract=3393482.

Bach, F. (2017): "Breaking the Curse of Dimensionality with Convex Neural Networks," *Journal of Machine Learning Research*, 18, 629–681.

Barron, A. R. (1993): "Universal Approximation Bounds for Superpositions of a Sigmoidal Function," *IEEE Transactions on Information Theory*, 39, 930–945.

Berger, J. and R. Wolpert (1988): *The Likelihood Principle*, Institute of Mathematical Statistics, 2nd ed.

Betancourt, M. (2017): "A Conceptual Introduction to Hamiltonian Monte Carlo," Tech. rep., Columbia University.

Blei, D. M., A. Kucukelbir, and J. D. McAuliffe (2017): "Variational Inference: A Review for Statisticians," *Journal of the American Statistical Association*, 112, 859–877.

Boivin, J. and M. Giannoni (2006): "DSGE Models in a Data-Rich Environment," Working Paper 12772, National Bureau of Economic Research.

Brooks, S., A. Gelman, G. Jones, and X. Meng (2011): *Handbook of Markov Chain Monte Carlo*, Chapman & Hall/CRC Handbooks of Modern Statistical Methods, CRC Press.

Browning, M., L. Hansen, and J. Heckman (1999): "Micro Data and General Equilibrium Models," in *Handbook of Macroeconomics*, ed. by J. B. Taylor and M. Woodford, Elsevier, vol. 1, Part A, chap. 08, 543–633, 1 ed.

Brunnermeier, M. K. and Y. Sannikov (2014): "A Macroeconomic Model with a Financial Sector," *American Economic Review*, 104, 379–421.

Canova, F. and L. Sala (2009): "Back to Square One: Identification Issues in DSGE Models," *Journal of Monetary Economics*, 56, 431–449.

Casella, G. and E. I. George (1992): "Explaining the Gibbs Sampler," *The American Statistician*, 46, 167–174.

Casella, S., J. Fernández-Villaverde, and S. Hansen (2020): "Structural Estimation of Dynamic Equilibrium Models with Unstructured Data," Mimeo, University of Pennsylvania.

Chernozhukov, V. and H. Hong (2003): "An MCMC Approach to Classical Estimation," *Journal of Econometrics*, 115, 293 – 346.

Childers, D., J. Fernádez-Villaverde, J. Perla, C. Rackauckas, and P. Wu (2020): "Differentiable State-Space Models with Applications to Estimation using Hamiltonian Monte Carlo," Mimeo, University of Pennsylvania.

Chopin, N. and O. Papaspiliopoulos (2020): *An Introduction to Sequential Monte Carlo*, Springer Verlag.

Christiano, L. J., M. Eichenbaum, and C. L. Evans (2005): "Nominal Rigidities and the Dynamic Effects of a Shock to Monetary Policy," *Journal of Political Economy*, 113, 1–45.

Cover, T. and J. Thomas (2012): *Elements of Information Theory, 2nd edition*, Wiley.

Fernández-Villaverde, J. (2010): "The Econometrics of DSGE Models," *SERIEs: Journal of the Spanish Economic Association*, 1, 3–49.

Fernández-Villaverde, J. and P. A. Guerrón-Quintana (2020): "Uncertainty Shocks and Business Cycle Research," *Review of Economic Dynamics.*

Fernández-Villaverde, J., P. A. Guerrón-Quintana, J. F. Rubio-Ramírez, and M. Uribe (2011): "Risk Matters: The Real Effects of Volatility Shocks," *American Economic Review*, 101, 2530–2561.

Fernández-Villaverde, J., S. Hurtado, and G. Nuño (2019): "Financial Frictions and the Wealth Distribution," Working Paper 26302, National Bureau of Economic Research.

Fernández-Villaverde, J., J. Rubio-Ramírez, and F. Schorfheide (2016): "Solution and Estimation Methods for DSGE Models," in *Handbook of Macroeconomics*, ed. by J. B. Taylor and H. Uhlig, Elsevier, vol. 2, 527–724.

Fernández-Villaverde, J. and J. F. Rubio-Ramírez (2007): "Estimating Macroeconomic Models: A Likelihood Approach," *Review of Economic Studies*, 74, 1059–1087.

——— (2008): "How Structural are Structural Parameters?" in *NBER Macroeconomics Annual 2007*, ed. by D. Acemoglu, K. Rogoff, and M. Woodford, University of Chicago Press, Chicago, vol. 22.

Fernández-Villaverde, J. and D. Z. Valencia (2018): "A Practical Guide to Parallelization in Economics," Working Paper 24561, National Bureau of Economic Research.

Gabaix, X. (Forthcoming): "A Behavioral New Keynesian Model," *American Economic Review.*

Girolami, M. and B. Calderhead (2011): "Riemann Manifold Langevin and Hamiltonian Monte Carlo Methods," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 73, 123–214.

Goodfellow, I., Y. Bengio, and A. Courville (2016): *Deep Learning*, MIT Press.

Gornemann, N., K. Kuester, and M. Nakajima (2012): "Monetary Policy with Heterogeneous Agents," Working Paper 12-21, Federal Reserve Bank of Philadelphia.

Guerrón-Quintana, P. A. (2010): "What You Match Does Matter: The Effects of Observable Variables on DSGE Estimation," *Journal of Applied Econometrics*, 25, 774–804.

Hansen, G. D. and E. C. Prescott (1995): "Recursive Methods for Computing Equilibria of Business Cycle Models," in *Frontiers of Business Cycle Research*, ed. by T. F. Cooley, Princeton University Press, 39–64.

Hansen, L. P. (1982): "Large Sample Properties of Generalized Method of Moments Estimators," *Econometrica*, 50, 1029–54.

Harvey, A. C. (1990): *Forecasting, Structural Time Series Models and the Kalman Filter*, Cambridge University Press.

Herbst, E. (2011): "Gradient and Hessian-based MCMC for DSGE Models," *Unpublished Manuscript, University of Pennsylvania.*

Herbst, E. and F. Schorfheide (2015): *Bayesian Estimation of DSGE Models*, Princeton University Press.

——— (2019): "Tempered Particle Filtering," *Journal of Econometrics*, 210, 26–44.

Homan, M. D. and A. Gelman (2014): "The No-U-Turn Sampler: Adaptively Setting Path Lengths in Hamiltonian Monte Carlo," *Journal of Machine Learning Research*, 15, 1593–1623.

Ilut, C. L. and M. Schneider (2014): "Ambiguous Business Cycles," *American Economic Review*, 104, 2368–99.

Iskrev, N. (2010): "Local identification in DSGE models," *Journal of Monetary Economics*, 57, 189–202.

Kalman, R. (1960): "A New Approach to Linear Filtering and Prediction Problems," *Transactions of the ASME–Journal of Basic Engineering,*, 82, 35–45.

Kaplan, G., B. Moll, and G. L. Violante (2018): "Monetary Policy According to HANK," *American Economic Review*, 108, 697–743.

Khan, A. and J. K. Thomas (2007): "Inventories and the Business Cycle: An Equilibrium Analysis of (S, s) Policies," *American Economic Review*, 97, 1165–1188.

Komunjer, I. and S. Ng (2011): "Dynamic Identification of Dynamic Stochastic General Equilibrium Models," *Econometrica*, 79, 1995–2032.

Künsch, H. R. (2005): "Recursive Monte Carlo Filters: Algorithms and Theoretical Analysis," *Annals of Statistics*, 33, 1983–2021.

Kydland, F. E. and E. C. Prescott (1982): "Time to Build and Aggregate Fluctuations," *Econometrica*, 50, 1345–1370.

Li, W. and P. Fearnhead (2018): "On the asymptotic efficiency of approximate Bayesian computation estimators," *Biometrika*, 105, 285–299.

Maliar, L., S. Maliar, and P. Winant (2019): "Will Artificial Intelligence Replace Computational Economists Any Time Soon?" CEPR Discussion Papers 14024, C.E.P.R. Discussion Papers.

Marin, J.-M., P. Pudlo, C. P. Robert, and R. J. Ryder (2012): "Approximate Bayesian Computational Methods," *Statistics and Computing*, 22, 1167–1180.

Marjoram, P., J. Molitor, V. Plagnol, and S. Tavaré (2003): "Markov Chain Monte Carlo without Likelihoods," *Proceedings of the National Academy of Sciences*, 100, 15324–15328.

Neal, R. M. (2011): *MCMC Using Hamiltonian Dynamics*, CRC Press, chap. chapter 5.

Nishiyama, S. and K. Smetters (2014): "Analyzing Fiscal Policies in a Heterogeneous-Agent Overlapping-Generations Economy," in *Handbook of Computational Economics*, ed. by K. Schmedders and K. L. Judd, Elsevier, vol. 3, 117–160.

Primiceri, G. (2006): "Why Inflation Rose and Fell: Policymakers' Beliefs and US Postwar Stabilization Policy," *Quarterly Journal of Economics*, 121, 867–901.

Robert, C. and G. Casella (2005): *Monte Carlo Statistical Methods*, Springer, 2nd ed.

Roberts, G. O., A. Gelman, and W. R. Gilks (1997): "Weak Convergence and Optimal Scaling of Random Walk Metropolis Algorithms," *Annals of Applied Probability*, 7, 110–120.

Scalone, V. (2018): "Estimating Non-Linear DSGEs with the Approximate Bayesian Computation: An Application to the Zero Lower Bound," Working papers 688, Banque de France.

Sisson, S. A., Y. Fan, and M. M. Tanaka (2007): "Sequential Monte Carlo without Likelihoods," *Proceedings of the National Academy of Sciences*, 104, 1760–1765.

Strathmann, H., D. Sejdinovic, S. Livingstone, Z. Szabo, and A. Gretton (2015): "Gradient-free Hamiltonian Monte Carlo with Efficient Kernel Exponential Families," Tech. rep., University College Londn.

Strid, I., P. Giordani, and R. Kohn (2010): "Adaptive hybrid Metropolis-Hastings samplers for DSGE models," SSE/EFI Working Paper Series in Economics and Finance 724, Stockholm School of Economics.

van Binsbergen, J. H., J. Fernández-Villaverde, R. S. Koijen, and J. F. Rubio-Ramírez (2012): "The Term Structure of Interest Rates in a DSGE Model with Recursive Preferences," *Journal of Monetary Economics*, 59, 634–648.

Wan, E. and R. Van Der Merwe (2000): "The Unscented Kalman Filter for Nonlinear Estimation," *Adaptive Systems for Signal Processing, Communications, and Control Symposium 2000. AS-SPCC. The IEEE 2000*, 153–158.

Watson, M. W. (1989): "Recursive Solution Methods for Dynamic Linear Rational Expectations Models," *Journal of Econometrics*, 41, 65 – 89.

Woodford, M. (2003): *Interest and Prices: Foundations of a Theory of Monetary Policy*, Princeton University Press.