Federal Reserve Bank of Dallas
Globalization and Monetary Policy Institute
Working Paper No. 189
http://www.dallasfed.org/assets/documents/institute/wpapers/2014/0189.pdf

Assessing Bayesian Model Comparison in Small Samples*

Enrique Martínez-García
Federal Reserve Bank of Dallas

Mark A. Wynne
Federal Reserve Bank of Dallas

August 2014

Abstract

We investigate the Bayesian approach to model comparison within a two-country framework with nominal rigidities using the workhorse New Keynesian open-economy model of Martínez-García and Wynne (2010). We discuss the trade-offs that monetary policy-characterized by a Taylor-type rule faces in an interconnected world, with perfectly flexible exchange rates. We then use posterior model probabilities to evaluate the weight of evidence in support of such a model when estimated against more parsimonious specifications that either abstract from monetary frictions or assume autarky by means of controlled experiments that employ simulated data. We argue that Bayesian model comparison with posterior odds is sensitive to sample size and the choice of observable variables for estimation. We show that posterior model probabilities strongly penalize overfitting which can lead us to favor a less parameterized model against the true data-generating process when the two become arbitrarily close to each other. We also illustrate that the spill-overs from monetary policy across countries have an added confounding effect.

JEL codes: C11, C13, F41

* Enrique Martínez-García, Research Department, Federal Reserve Bank of Dallas, 2200 N. Pearl Street, Dallas, TX 75201. 214-922-5262. enrique.martinez-garcia@dal.frb.org. Mark A. Wynne, Research Department, Federal Reserve Bank of Dallas, 2200 N. Pearl Street, Dallas, TX 75201. 214-922-5159. Mark.A.Wynne@dal.frgb.org. We would like to thank Nathan Balke, María Teresa Martínez-García and Valentín Martínez Mira for helpful suggestions. Diego Vilán was a co-author in a related project and contributed to the early stages of development of this paper. We gratefully acknowledge the outstanding research assistance provided by Valerie Grossman, the help of Kuhu Parasrampuria, and the Federal Reserve Bank of Dallas's support. The views in this paper are those of the authors and do not necessarily reflect the views of the Federal Reserve Bank of Dallas or the Federal Reserve System.

1 Introduction

Bayesian methods have become a standard part of the toolkit in quantitative macroeconomics. They are commonly used to estimate the parameters and assess the fit of a given model, but they are also widely employed for comparison across competing models. We can think of a model as a parameterized probability distribution (based on a given theory of how the economy works) that characterizes the data-generating process (DGP) from which the observables that constitute our data are drawn. Hence, by model comparison we mean the evaluation of $k \geq 2$ competing parameterized probability distributions—the models M_1,..., M_k—representing different theories based on the observed empirical distribution of the data. In other words, model comparison provides guidance on which of the existing theories better accounts for the observed data.

Model selection is a related decision-theory problem that specifies a loss function as a metric to judge the differences across models against the data and pick among competing theories. It is known that under a $0-1$ loss function it is optimal to select the model with the highest posterior probability (see, e.g., Kass and Raftery (1995)). Model averaging is another related notion that incorporates model uncertainty by averaging across all possible k models using the weights to reflect how likely each model is given the observed data (see, e.g., Hoeting et al. (1999)). Selecting the incorrect model or assigning too large a probability, though, can result in misleading inferences and even in the implementation of sub-optimal policies meant to correct for the effect of frictions or economic distortions that may not even be present in the 'true' DGP underlying the data. So, this begs the question, when are Bayesian model comparisons more prone to fail to detect the true DGP (or its closest match among the available models)?

The Bayesian approach to model comparison consists in placing probabilities on a number of competing models and evaluating the posterior probability of each model (see, e.g., Kass and Raftery (1995) and An and Schorfheide (2007)). The significance of posterior model probabilities for making comparison across competing models is largely based on the desirable asymptotic properties of these posterior probabilities derived under fairly general regularity conditions. Fernández-Villaverde and Rubio-Ramírez (2004) show that, as the sample size grows arbitrarily large, the Bayesian parameter point estimates converge to their pseudo-true values. They also show that the best model under the Kullback-Leibler distance criterion—the model closest to the 'true' DGP in the Kullback-Leibler sense—is the one with the highest posterior model probability. Moreover, these asymptotic properties hold even if the models being compared are non-nested, non-linear, and do not even include a model for the 'true' DGP.

In this paper, we illustrate the *less-desirable* small sample properties of Bayesian posterior model probabilities. We work with simulated data in controlled experiments and make our case using a standard log-linearized two-country New Open-Economy Macro (NOEM) model with nominal rigidities as the 'true' DGP. We compare the NOEM model against three alternative (log-linear) specifications that either assume flexible prices (instead of nominal rigidities), posit a closed-economy setting for each country (autarky) or both. All three competing models are nested in the NOEM model and the dimensionality of their parameter space is lower. We consider these three alternative specifications because they evoke important concerns for policy-making—such as the role of globalization (openness to trade) and monetary policy in the presence/absence of nominal rigidities.

We design a number of experiments to illustrate how model comparison depends not only on the length of the time series used for estimation, but also on the selection of the observable macro variables on which the compared models are estimated. We show that in small samples the Bayesian posterior model probabilities

are more likely to favor a more parsimonious specification over the NOEM model (our 'true' DGP) when the simulated data are generated under a parameterization that brings the probability distribution of the DGP *close* to that of some of the alternative model specifications (theories) under consideration.

In our particular illustrations, that means posterior model probabilities can favor a closed-economy model whenever the degree of trade openness is *low enough* or can favor a model that abstracts from nominal rigidities whenever monetary policy is *near-optimal* and the degree of price stickiness is *low*. More generally, our work suggests that model comparison, model selection and model averaging can be distorted in economically relevant ways whenever model comparison strongly penalizes the more richly parameterized models. Furthermore, what the preferred model ends up being is not straightforward as their implied probability distributions tend to be nonlinear in the parameters—and there may be more than one model that appears empirically *close* to the 'true' one.

The remainder of the paper proceeds as follows: Section 2 outlines the workhorse model of Martínez-García and Wynne (2010), and describes its building blocks. Several alternative nested specifications are proposed for model comparison, whereby monetary policy effectiveness changes by removing features such as household's preference for imported varieties or rigidities in firm price-setting behavior. In Section 3 we illustrate our findings showing that in small samples posterior model probabilities may fail to pick the more-heavily parameterized NOEM model against the alternative nested specifications, even though the NOEM model is the 'true' DGP for the data. These confounding results also appear when we try an alternative selection of observables. In Section 4 we discuss our findings, make recommendations for applied work with these techniques, and draw policy implications for the class of open-economy models that we investigate. Section 5 provides a brief summary of the technical insights gained from our exercise and its policy implications, and concludes. We also provide a companion on-line Appendix for the interested reader where further detail on the model and the implementation strategy is given (see Martínez-García and Wynne (2014)).

2 Economic Model

We adopt the model of Martínez-García and Wynne (2010). This is a two-country, symmetric New Open-Economy Macro (NOEM) model with complete asset markets and nominal rigidities in the spirit of Clarida et al. (2002), subject to country-specific productivity and monetary shocks. The stylized model abstracts from capital accumulation, and assumes a cashless economy and perfectly flexible exchange rates. Labor is immobile across countries, but all varieties of goods produced in each country can be traded. The model provides a tractable economic environment that departs from monetary neutrality and allows international spill-overs to be transmitted through trade.

The model features two standard distortions in the goods markets—monopolistic competition in production and constrained price-setting behavior subject to Calvo (1983) contracts and producer currency pricing (as in Clarida et al. (2002)):

∘ The introduction of an optimal labor subsidy for firms funded with lump-sum (non-distortionary) taxes eliminates the mark-up distortion caused by monopolistic competition. It also ensures that the deterministic steady state of the model is the same under either flexible prices or nominal rigidities. Hence, the key assumption on which the non-neutrality of monetary policy hinges is price stickiness modelled à la Calvo

(1983).

∘ The law of one price holds at the variety level because all prices are set in the producer's own currency. Deviations from purchasing-power parity (PPP) arise solely due to differences in preferences that result in the composition of the consumption basket varying across countries, with local households consuming a larger share of the locally-produced varieties than of the imported ones (home bias). The degree of openness to trade that allows for the endogenous propagation of country-specific shocks internationally is also directly tied to the appetite of households for imported goods.[1]

Since the setup of the model we use is otherwise extensively discussed in Martínez-García and Wynne (2010), here we shall put the emphasis instead on the key equations of its log-linearized representation and their economic interpretation. The companion on-line appendix (i.e., Martínez-García and Wynne (2014)) provides further details on the building blocks of the model as well as on our approach to data simulation, Bayesian estimation and Bayesian model comparison.

The workhorse (log-linearized) model.[2] The basic structure of the New Keynesian model is given by a log-linearized system of three-equations—which includes a Phillips curve, an IS curve, and an interest rate-based monetary policy rule—that characterize the dynamics of output, inflation, and the short-term nominal interest rate. Goodfriend and King (1997), Clarida et al. (1999), and Woodford (2003) among others contributed to the derivation of those equations from explicit optimizing behavior on the part of firms (price-setters) and households in the presence of nominal rigidities.

Clarida et al. (2002) extends the three-equation workhorse New Keynesian model to a two-country setting. Building on that contribution, Martínez-García and Wynne (2010) show that the same basic structure of three log-linearized equations can be generalized to describe the dynamics of output, inflation, and the short-term rate when a country is open to trade. The monetary policy rule remains focused on domestic objectives even in the open-economy model in the environment presented by Martínez-García and Wynne (2010)—but both the Phillips curve and the IS curve differ from their closed-economy counterparts due to the interactions across countries that take place through trade and the resulting spillovers into inflation and aggregate demand. The model of Martínez-García and Wynne (2010) showcases for us the interconnectedness that arises through trade in goods, while keeping most of the simplicity and tractability of the workhorse (closed-economy) New Keynesian model.

In the framework of Martínez-García and Wynne (2010), the open-economy Phillips curve can be written for each country as follows,

$$\begin{aligned}
\widehat{\pi}_t &\approx \beta \mathbb{E}_t \left(\widehat{\pi}_{t+1}\right) + ... \\
&\quad \Phi \left[(1-\xi) \left(\varphi + \left(\frac{\sigma - (\sigma-1)(1-2\xi)}{\sigma - (\sigma-1)(1-2\xi)^2} \right) \right) \widehat{x}_t + \xi \left(\varphi + \left(\frac{\sigma + (\sigma-1)(1-2\xi)}{\sigma - (\sigma-1)(1-2\xi)^2} \right) \right) \widehat{x}_t^* \right], \quad (1) \\
\widehat{\pi}_t^* &\approx \beta \mathbb{E}_t \left(\widehat{\pi}_{t+1}^*\right) + ... \\
&\quad \Phi \left[\xi \left(\varphi + \left(\frac{\sigma + (\sigma-1)(1-2\xi)}{\sigma - (\sigma-1)(1-2\xi)^2} \right) \right) \widehat{x}_t + (1-\xi) \left(\varphi + \left(\frac{\sigma - (\sigma-1)(1-2\xi)}{\sigma - (\sigma-1)(1-2\xi)^2} \right) \right) \widehat{x}_t^* \right], \quad (2)
\end{aligned}$$

[1] We distinguish here between the endogenous international propagation that comes from trade and the purely-exogenous international propagation that arises—even in the absence of trade—from the specification of correlated exogenous shock processes in both countries. By endogenous international propagation we refer more precisely to the effect that a shock impacting the foreign country has on the domestic macro aggregates as a result of the domestic economic agents' response to that shock.

[2] All variables are defined in logs as deviations from steady-state.

where $\widehat{\pi}_t$ and $\widehat{\pi}_t^*$ denote Home and Foreign inflation (that is, quarter-over-quarter changes in the consumption price index), and $\widehat{x}_t$ and $\widehat{x}_t^*$ define the Home and Foreign output gaps or slack (that is, the deviations of output from its potential under flexible prices). The composite coefficient $\Phi \equiv \left(\frac{(1-\alpha)(1-\beta\alpha)}{\alpha}\right)$ is the common term on the slope of the open-economy Phillips curve, $0 < \beta < 1$ is the subjective intertemporal discount factor, and $0 < \alpha < 1$ is the Calvo price stickiness parameter. The differences in slope coefficients for domestic and foreign slack that arise in $(1)-(2)$ are related to the inverse of the Frisch elasticity of labor supply $\varphi > 0$, the elasticity of intratemporal substitution between Home and Foreign goods $\sigma > 0$, and the share of imported goods in the consumption basket $0 \leq \xi \leq \frac{1}{2}$.[3]

Price stickiness breaks monetary policy neutrality in the short-run, establishing a Phillips curve relationship between nominal (inflation) and real variables (slack). The assumption that household preferences for consumption goods are defined over imported as well as domestic varieties is what gives rise to the *global slack hypothesis* in this framework—that is, to the idea that in a world open to trade the relevant trade-off for monetary policy captured by the Phillips curve is between a country's inflation and global (rather than local) slack. Not surprisingly, the structural parameters α and ξ feature prominently among the structural parameters that determine the slope of the open-economy Phillips curve in $(1)-(2)$. These parameters characterize respectively the fraction of firms that cannot update their prices in any given period (price stickiness) and the import shares (openness), although the role each plays in the dynamics of the model is different.

∘ The parameter α enters through the common term for the slope Φ. This structural parameter captures the degree of price stickiness, and price stickiness is the key distortion that introduces monetary non-neutrality. Under flexible prices (absent nominal rigidities), monetary policy has no real effects. Therefore, the real effects of monetary policy in the model tend to be negligible as α becomes arbitrarily close to zero—since a larger fraction of firms becomes unconstrained to change prices every period.

∘ The parameter ξ appears in the composite terms that differentiate the slope for domestic and foreign slack. This structural parameter determines the import share (the extent of trade openness), and explains deviations from PPP in the model. In a closed-economy setting or under autarky, there is no endogenous mechanism for the international transmission of shocks. Even if trade were permitted in this model, an analogous situation would arise with no endogenous international propagation of shocks if all imports were excluded from the consumption basket—that is, when $\xi = 0$.[4] Therefore, international propagation tends to be attenuated as the import share ξ becomes arbitrarily close to zero.

The open-economy IS equations in $(3)-(4)$ illustrate how the output gaps, $\widehat{x}_t$ and $\widehat{x}_t^*$, are tied to shifts in consumption demand over time and across countries,

$$\begin{aligned}(1-2\xi)\,\mathbb{E}_t\left[\widehat{x}_{t+1}-\widehat{x}_t\right] \;\approx\; & (1-\xi)\left(\sigma-(\sigma-1)(1-2\xi)\right)\left[\widehat{r}_t-\widehat{\overline{r}}_t\right]-... \\ & \xi\left(\sigma+(\sigma-1)(1-2\xi)\right)\left[\widehat{r}_t^*-\widehat{\overline{r}}_t^*\right], \qquad (3)\\ (1-2\xi)\,\mathbb{E}_t\left[\widehat{x}_{t+1}^*-\widehat{x}_t^*\right] \;\approx\; & -\xi\left(\sigma+(\sigma-1)(1-2\xi)\right)\left[\widehat{r}_t-\widehat{\overline{r}}_t\right]+... \\ & (1-\xi)\left(\sigma-(\sigma-1)(1-2\xi)\right)\left[\widehat{r}_t^*-\widehat{\overline{r}}_t^*\right], \qquad (4)\end{aligned}$$

[3] The inverse of the intertemporal elasticity of substitution is equal to 1 under the assumption of log-utility on consumption.

[4] In that case, there would be no reason for these countries to trade with each other and in equilibrium there would be no exchange of goods anyway because the households of one country would not demand imports from the other country.

where the real interest rates in the Home and Foreign country are defined by the Fisher equation as $\widehat{r}_t \equiv \widehat{i}_t - \mathbb{E}_t\left[\widehat{\pi}_{t+1}\right]$ and $\widehat{r}_t^* \equiv \widehat{i}_t^* - \mathbb{E}_t\left[\widehat{\pi}_{t+1}^*\right]$ respectively, and $\widehat{i}_t$ and $\widehat{i}_t^*$ are the Home and Foreign short-term nominal interest rates. The natural real rates that would prevail under flexible prices are denoted $\widehat{\overline{r}}_t$ for the Home country and $\widehat{\overline{r}}_t^*$ for the Foreign country. Price stickiness introduces in the IS equations a wedge between the real interest rate (the actual opportunity cost of consumption today versus consumption tomorrow) and the natural real rate of interest that captures its distortionary effects on aggregate demand as shown in $(3)-(4)$. However, the Calvo parameter α, which determines the degree of nominal rigidities present, does not appear explicitly in the equations. In turn, the appetite for imported goods ξ plays a prominent role in the open-economy IS equations as it affects the contributions of the demand distortions arising in the local and export markets to the output gap of each given country.

The Home and Foreign Taylor (1993)-type monetary policy rules complete the specification of the NOEM model. Monetary policy pursues the goal of domestic stabilization (even in a fully integrated world) and, hence, solely responds to changes in the local economic conditions as determined by each country's inflation and output gap. As is commonly done in the literature, we assume *intrinsic* or endogenous inertia in the policy rules described in $(5)-(6)$ resulting from policy-makers intentionally smoothing out their policy response to changing economic conditions,

$$\widehat{i}_t \approx \rho_i \widehat{i}_{t-1} + (1-\rho_i)\left[(1+\psi_\pi)\,\widehat{\pi}_t + \psi_x \widehat{x}_t\right] + \widehat{\varepsilon}_t^m, \tag{5}$$

$$\widehat{i}_t^* \approx \rho_i \widehat{i}_{t-1}^* + (1-\rho_i)\left[(1+\psi_\pi)\,\widehat{\pi}_t^* + \psi_x \widehat{x}_t^*\right] + \widehat{\varepsilon}_t^{m*}, \tag{6}$$

where $\widehat{\varepsilon}_t^m$ and $\widehat{\varepsilon}_t^{m*}$ are the Home and Foreign monetary policy shocks modelled with a bivariate normal distribution with zero mean and positive covariance across countries. The policy parameters $\psi_\pi > 0$ and $\psi_x > 0$ represent the sensitivity of the monetary policy rule to movements in inflation and the output gap respectively, while $0 \leq \rho_i < 1$ represents the policy smoothing parameter.

The natural rates $\widehat{\overline{r}}_t$ and $\widehat{\overline{r}}_t^*$ can be expressed as functions of expected changes in Home and Foreign potential output, i.e.,

$$\begin{aligned}\widehat{\overline{r}}_t \approx\ & (1-\xi)\left(\frac{\sigma-(\sigma-1)(1-2\xi)}{\sigma-(\sigma-1)(1-2\xi)^2}\right)\left(\mathbb{E}_t\left[\widehat{\overline{y}}_{t+1}\right]-\widehat{\overline{y}}_t\right)+\ldots \\ & \xi\left(\frac{\sigma+(\sigma-1)(1-2\xi)}{\sigma-(\sigma-1)(1-2\xi)^2}\right)\left(\mathbb{E}_t\left[\widehat{\overline{y}}_{t+1}^*\right]-\widehat{\overline{y}}_t^*\right),\end{aligned} \tag{7}$$

$$\begin{aligned}\widehat{\overline{r}}_t^* \approx\ & \xi\left(\frac{\sigma+(\sigma-1)(1-2\xi)}{\sigma-(\sigma-1)(1-2\xi)^2}\right)\left(\mathbb{E}_t\left[\widehat{\overline{y}}_{t+1}\right]-\widehat{\overline{y}}_t\right)+\ldots \\ & (1-\xi)\left(\frac{\sigma-(\sigma-1)(1-2\xi)}{\sigma-(\sigma-1)(1-2\xi)^2}\right)\left(\mathbb{E}_t\left[\widehat{\overline{y}}_{t+1}^*\right]-\widehat{\overline{y}}_t^*\right),\end{aligned} \tag{8}$$

reflecting the fact that real rates respond to expected changes in—rather than the level of—real economic activity as measured by potential output. Potential output refers to the output that would have been produced under flexible prices, and accordingly $\widehat{\overline{y}}_t$ and $\widehat{\overline{y}}_t^*$ denote the corresponding Home and Foreign potential output in the model. Home and Foreign potential output can be expressed solely in terms of real

shocks since monetary shocks have no real effects absent nominal rigidities, i.e.,

$$\begin{aligned}\widehat{\overline{y}}_t &\approx \left(1+(\sigma-1)\left(\frac{2\xi\,(1-\xi)}{\varphi\left(\sigma-(\sigma-1)\,(1-2\xi)^2\right)+1}\right)\right)\widehat{a}_t-...\\ &\quad (\sigma-1)\left(\frac{2\xi\,(1-\xi)}{\varphi\left(\sigma-(\sigma-1)\,(1-2\xi)^2\right)+1}\right)\widehat{a}^*_t, \qquad (9)\\ \widehat{\overline{y}}^*_t &\approx -(\sigma-1)\left(\frac{2\xi\,(1-\xi)}{\varphi\left(\sigma-(\sigma-1)\,(1-2\xi)^2\right)+1}\right)\widehat{a}_t+...\\ &\quad \left(1+(\sigma-1)\left(\frac{2\xi\,(1-\xi)}{\varphi\left(\sigma-(\sigma-1)\,(1-2\xi)^2\right)+1}\right)\right)\widehat{a}^*_t, \qquad (10)\end{aligned}$$

where $\widehat{a}_t$ and $\widehat{a}^*_t$ denote the corresponding Home and Foreign productivity shocks in the model.

The natural rates of interest and potential output are invariant to monetary policy or the monetary policy shocks. In turn, the natural rates only depend on productivity shocks that are modelled as a VAR(1) without spill-overs but with positive covariance across countries of their innovations. Natural rates and potential output summarize the dynamics of a competing, nested model that abstracts from nominal rigidities—in effect, a stylized International Real Business Cycle (IRBC) model without capital accumulation. The model presented here also nests naturally another competing class of models which assume a closed economy whenever we set the import share ξ equal to zero. We include all those nested variants in our Bayesian model comparison exercise.

Moreover, $\widehat{y}_t = \widehat{\overline{y}}_t + \widehat{x}_t$ and $\widehat{y}^*_t = \widehat{\overline{y}}^*_t + \widehat{x}^*_t$ are respectively the actual Home and Foreign output variables. Domestic terms of trade (defined as the price of imports relative to the price of exports) is proportional to the output differential across countries, $\widehat{tot}_t \approx \left(\frac{1}{\sigma-(\sigma-1)(1-2\xi)^2}\right)(\widehat{y}_t-\widehat{y}^*_t)$, capturing the relative scarcity of Home- versus Foreign-produced goods. The domestic trade balance $\widehat{tb}_t \equiv \widehat{y}_t - \widehat{c}_t \approx \xi\left(\frac{\sigma+(\sigma-1)(1-2\xi)}{\sigma-(\sigma-1)(1-2\xi)^2}\right)(\widehat{y}_t-\widehat{y}^*_t)$ is proportional to the output differential across countries illustrating the net movement in goods that takes place across borders whenever relative scarcity of Home- versus Foreign-produced goods arises in order to intratemporally smooth consumption.[5] For further details on the trade features of this class of open-economy New Keynesian models, the interested reader is referred to Martínez-García and Søndergaard (2009).

Model Solution. We can replace (7) − (10) into (1) − (6) to express the system of equations that characterizes the model as follows,

$$M\widehat{Z}_t = N\mathbb{E}_t\left[\widehat{Z}_{t+1}\right]+Q\widehat{\varepsilon}_t, \qquad (11)$$

[5] The terms of trade and the trade balance are related within the model as follows,

$$\widehat{tb}_t \approx (\sigma+(\sigma-1)\,(1-2\xi))\,\xi\widehat{tot}_t.$$

Hence, the so-called Harberger–Laursen–Metzler (HLM) effect arises naturally within this model: an improvement in a country's terms of trade raises current income, but current consumption increases less than current income causing private savings to increase and improving the trade balance (given a marginal propensity to consume less than unity).

where

$$\begin{aligned}\widehat{Z}_t &= \left(\widehat{\pi}_t, \widehat{\pi}^*_t, \widehat{y}_t, \widehat{y}^*_t, \widehat{i}_{t-1}, \widehat{i}^*_{t-1}, \widehat{a}_{t-1}, \widehat{a}^*_{t-1}\right)', \\ \widehat{\varepsilon}_t &= \left(\widehat{\varepsilon}^a_t, \widehat{\varepsilon}^{a*}_t, \widehat{\varepsilon}^m_t, \widehat{\varepsilon}^{m*}_t\right)',\end{aligned}$$

and M, N and Q are conforming matrices. For reasonable parameter values, the matrix M is invertible and (11) can be re-written as,

$$\widehat{Z}_t = \Gamma \mathbb{E}_t \left[\widehat{Z}_{t+1}\right] + \Psi \widehat{\varepsilon}_t, \tag{12}$$

where $\Gamma = M^{-1}N$ and $\Psi = M^{-1}Q$. Blanchard and Kahn (1980) provide conditions under which a unique stable solution exists for (12). Although it is not easy to derive analytically the parameter restrictions that guarantee existence and uniqueness, numerical experiments show that the policy parameter ψ_π is key and also that the lower bound on ψ_π above which the model attains determinacy depends on the policy parameter ψ_x. In an open-economy model with interest rate smoothing in the monetary policy rule, the Taylor principle (i.e., $\psi_\pi > 1$) remains broadly consistent with satisfying the Blanchard-Kahn condition for determinacy for a wide range of plausible values of the structural parameters of the model. We parameterize the model for simulation to ensure existence and uniqueness of the solution, and we accordingly set the range of priors for estimation to avoid as much as possible the regions of the parameter space that result in indeterminacy or no-solution.

We partition $\widehat{Z}_t$ into two blocks with $\widehat{Z}_{1t} = (\widehat{\pi}_t, \widehat{\pi}^*_t, \widehat{y}_t, \widehat{y}^*_t)'$ and $\widehat{Z}_{2t} = \left(\widehat{i}_{t-1}, \widehat{i}^*_{t-1}, \widehat{a}_{t-1}, \widehat{a}^*_{t-1}\right)'$. Assuming the Blanchard-Kahn condition is indeed satisfied and imposing $\lim_{J \to +\infty} \Gamma^J \mathbb{E}_t \left[\widehat{Z}_{1t+J}\right] = 0$, we solve (12) to characterize the solution of the NOEM in state space form as follows,

$$\widehat{Z}_{2t} = A_1(\theta)\, \widehat{Z}_{2t-1} + B_1(\theta)\, \widehat{\varepsilon}_t, \tag{13}$$

$$\widehat{Z}_{1t} = C_1(\theta)\, \widehat{Z}_{2t} + D_1(\theta)\, \widehat{\varepsilon}_t, \tag{14}$$

where $A_1(\theta)$, $B_1(\theta)$, $C_1(\theta)$ and $D_1(\theta)$ are conforming matrices, and θ is the vector of structural parameters of the model that enter those matrices. Fernández-Villaverde et al. (2007) explore the link between Dynamic Stochastic General Equilibrium (DSGE) models and state space representations like this one. The solution in $(13) - (14)$ shows that inflation and output in both countries, $\widehat{Z}_{1t}$, can be characterized as linear functions of a vector of state variables, $\widehat{Z}_{2t}$, and structural shock innovations, $\widehat{\varepsilon}_t$. Since the vector of structural shock innovations, $\widehat{\varepsilon}_s$, is normally distributed, then the Gaussian state-space representation of the solution in $(13) - (14)$ implies that inflation and output are also normally-distributed processes (see Hamilton (1994) for further discussion on the Gaussian state-space model).

Model Simulation. We use the same benchmark parameterization of the model described in Martínez-García and Wynne (2010) with only a small modification: in our exercise, we assume log-utility on consumption and accordingly set the elasticity of intertemporal substitution to one. We explore the sensitivity of standard Bayesian model comparison with respect to the value of the parameters ξ and ψ_π replacing in each case the parameterization used in Martínez-García and Wynne (2010) with points along an interval that spans for each the region of interest of the parameter space. We provide further details on the choice of parameter values and intervals in the companion on-line appendix (i.e., Martínez-García and Wynne (2014)).

These two parameters—one structural ξ, one policy ψ_π—are crucial in the model for different reasons.

The structural parameter ξ defines how *close* the countries are to autarky and, as we have indicated before, it plays a significant role in the specification of the open-economy Phillips curves and IS equations. The structural parameter α indicates the degree of nominal rigidity and this friction is the reason why monetary policy has real effects in the model. The parameter α directly affects the overall slope of the open-economy Phillips curve, in a similar way as in the closed-economy case. However, we recognize that the distortion that arises is conditional not only on the structure of the economy (for instance, the degree of integration through trade ξ or the size of the nominal rigidity α) but—most importantly—on monetary policy. Since the policy parameter ψ_π determines the tolerance for inflation of the domestic policy-makers in this environment, it has a direct influence on how much slack accumulates—measured by the output gap. It also affects how *close* monetary policy is to attain the optimal allocation under flexible prices. We choose to focus on this policy parameter here rather than directly on α.

We use the log-linear approximation of the workhorse model of Martínez-García and Wynne (2010)—henceforth, the NOEM model—as our DGP and simulate data at each point of the relevant interval of the parameter space for each of the two parameters under consideration. We keep in all cases the realization of the shocks invariant and all other structural parameters unchanged at their benchmark values. We simulate the full model over $11,000$ periods, and drop the first $1,000$ observations of each series to exclude any effect of the initial conditions on the simulation. We also select three sub-samples of 160 observations each, which correspond to 40 years of quarterly observations—a plausible upper bound length for many time series of international macro data which often can be much shorter than that. The simulation is implemented with code written for Dynare (see, e.g., Adjemian et al. (2011)). Working with simulated rather than actual data allows us a more precise assessment of the Bayesian posterior mode probabilities and their sensitivity to implementation, as we always know the true DGP.

Model Estimation. The $10,000$-period long simulated sample allows us to illustrate the asymptotic behavior of the posterior model probability, while the simulated sub-samples of 160 observations illustrate the small sample inference problems that could arise in the data. Bayesian estimation and model comparison is implemented with the Dynare software too. We assume a uniform prior over all competing models: the NOEM model (the true DGP, M_1), a variant with flexible prices and openness to trade (M_2), a variant with nominal rigidities and autarky derived under the assumption $\xi = 0$ (M_3), and a variant with flexible prices and $\xi = 0$ (M_4).

The system of equations that characterizes M_1 and the variants M_2, M_3, and M_4 as special cases of the specification for the NOEM model (M_1) can be found in the companion on-line appendix (i.e., Martínez-García and Wynne (2014)). The solution for each model variant $k = 1, 2, 3, 4$ fits into the Gaussian state-space representation form given in $(13) - (14)$ where $A_k(\theta)$, $B_k(\theta)$, $C_k(\theta)$ and $D_k(\theta)$ are the corresponding conforming matrices for each. The set of structural parameters θ is common to all models, but not all of the parameters affect the dynamics in each of the k specifications and this is reflected in the matrices $A_k(\theta)$, $B_k(\theta)$, $C_k(\theta)$ and $D_k(\theta)$ accordingly. We compute the marginal density of each model with a Laplace approximation after estimating these four nested variants of the model including the true DGP (M_1). The Laplace approximation works rather well in practice, in particular for highly-peaked, unimodal posterior densities.

As is conventionally done in the Bayesian literature, rather than imposing 'non-informative' priors on the structural parameters we choose fairly 'informative' priors to incorporate other sources of information and to reflect current views on the structural parameters themselves. The prior mean is set to match the

true parameter value of the DGP used to simulate the data (which corresponds with the parameterization indicated above). For the parameters of interest ξ and ψ_π (and also for α), the mean of the prior is set to vary along the interval that we evaluate. The shape and the dispersion of the prior distributions are fixed in all our experiments.[6] The same priors for the parameters are used in the estimation of the four competing models that we compare. All our choices on the prior distributions are summarized in Table 1. Further discussion on the rationale behind the selection of prior distributions can be found in the companion on-line appendix (i.e., Martínez-García and Wynne (2014)).

[Insert Table 1 about here.]

3 Findings

Sample Size of Observables for Estimation. The key features of the NOEM model, the DGP for the simulated data (model M_1), that distinguish it from the competing models M_2, M_3, and M_4 are openness to trade and monetary non-neutrality due to the presence of nominal rigidities. Conventional practice would be to include a selection of nominal and real variables for both the Home and Foreign country in the estimation in order to facilitate the empirical assessment of these four models. We also require that the observables be measured variables in all competing models. In order to avoid stochastic singularity in Bayesian estimation, we must have the same number of observable variables as structural shocks. Since we have monetary and productivity shocks that are country-specific in all models considered, we choose to estimate all competing theories with four observable variables: Home and Foreign output as well as Home and Foreign inflation. However, we argue that a standard choice of variables such as the one postulated here—while reasonable *ex ante*—has implications for Bayesian model comparison that are worth considering further.

Monetary policy under flexible prices and a zero import share $\xi = 0$ (model M_4) or with flexible prices and open to trade (model M_2) has no real effects, therefore influencing nominal variables only. Home and foreign inflation offers insights only on the differences in the implementation of monetary policy across countries. In other words, nominal variables do not help us distinguish between autarky (M_4) and openness to trade (M_2) under flexible prices. The economies represented by models M_2 and M_4 are already at their respective potential and the output gap is naturally zero—but they still differ in the allocations they attain. As can be inferred from equations $(5) - (6)$, productivity shocks from both countries endogenously influence the potential attained by each in M_2 but not in M_4 where only local productivity shocks matter. Still, potential output comoves across countries in model M_4 if solely because of the exogenous covariance of the productivity innovations—so comovement by itself does not rule out an autarky solution. Naturally, a lower import share ξ in model M_2 tends to result in output allocations that are increasingly *more similar* between models M_2 and M_4, making it harder to tell them apart based on the selected macro observables.

The optimal monetary policy for the workhorse closed-economy New Keynesian model with nominal rigidities is to set inflation at zero.[7] This policy prescription seemingly carries over to the open-economy model under autarky. The optimal monetary policy for the case with nominal rigidities and a zero import

[6] In keeping the priors invariant, we tie our hands to facilitate model comparison and preclude the priors themselves from becoming a source of additional degrees of freedom to *fine-tune* the estimation and the computation of posterior model probabilities.

[7] Assuming, as Martínez-García and Wynne (2010) do, that an optimal labor subsidy for firms funded with lump-sum (non-distortionary) taxes is set to eliminate the mark-up distortion.

share $\xi = 0$ (model M_3) is to set inflation to zero in both countries ensuring that the economy attains the same allocation as under flexible prices and autarky (model M_4). Therefore, a more aggressive policy response to inflation—a higher value of ψ_π to keep inflation at bay—should result in allocations that are increasingly *more similar* between models M_3 and M_4, making it harder to tell them apart based on the observed nominal and real macro data.

Setting the inflation rate on domestic consumption to zero in both countries for the NOEM model with trade (model M_1) is not necessarily going to attain the same allocation as under flexible prices and trade (model M_2). One way to illustrate this point is through closer inspection of the inflation rate. We can express the Home and Foreign inflation rate of consumption goods $\widehat{\pi}_t \approx \widehat{\pi}_t^H + \xi \Delta \widehat{tot}_t$ and $\widehat{\pi}_t^* \approx \widehat{\pi}_t^{F*} - \xi \Delta \widehat{tot}_t$ respectively, where $\widehat{\pi}_t^H$ and $\widehat{\pi}_t^{F*}$ denote Home and Foreign inflation of the locally-produced goods (that is, quarter-over-quarter changes in the output price index) and $\Delta \widehat{tot}_t$ represents changes in the terms of trade. Setting monetary policy to bring down inflation to zero in both countries (i.e., $\widehat{\pi}_t \approx \widehat{\pi}_t^* \approx 0$) does not ensure that the rate of inflation on the locally-produced goods $\widehat{\pi}_t^H$ and $\widehat{\pi}_t^{F*}$ becomes zero as well, except in the case where $\xi = 0$ (model M_3 with nominal rigidities or model M_4 under flexible prices).

As noted before, the terms of trade capture the relative scarcity of Home- versus Foreign-produced goods, so the inflation rate on consumption goods can be further re-written as follows,

$$\widehat{\pi}_t \approx \widehat{\pi}_t^H + \xi \left(\frac{1}{\sigma - (\sigma - 1)(1 - 2\xi)^2} \right) (\Delta \widehat{y}_t - \Delta \widehat{y}_t^*), \tag{15}$$

$$\widehat{\pi}_t^* \approx \widehat{\pi}_t^{F*} - \xi \left(\frac{1}{\sigma - (\sigma - 1)(1 - 2\xi)^2} \right) (\Delta \widehat{y}_t - \Delta \widehat{y}_t^*), \tag{16}$$

which implies that inflation and the growth differential across countries must be related in equilibrium.[8] We cannot assume that optimal monetary policy implies that $(\Delta \widehat{y}_t - \Delta \widehat{y}_t^*) \approx 0$ because output is driven by country-specific shocks which generally do not produce identical growth rates for both countries in every period. As a result, in the NOEM model with trade (model M_1), a monetary policy set to bring down consumption inflation towards zero would generally not attain a zero inflation rate on the locally-produced goods $\widehat{\pi}_t^H$ and $\widehat{\pi}_t^{F*}$—so long as local firms are subject to price stickiness and cannot all adjust their prices, there will be some loss relative to the flexible price case. However, the distortion that remains from implementing such a monetary policy tends to diminish the smaller the value of the import shares ξ is.

Based on that, setting inflation at zero in both countries under the NOEM model specification (model M_1) should result in an allocation that is increasingly *close* to the allocation attained under flexible prices and trade (model M_2) as the import share ξ becomes arbitrarily close to zero (as we assume for models M_3 under price stickiness and M_4 under flexible prices). Therefore, a more aggressive policy response to inflation—a higher value of ψ_π to keep inflation at bay—should result in allocations that are increasingly *more similar* between all models—M_1, M_2, M_3, and M_4—as the import share ξ becomes arbitrarily close to zero, making it harder to tell them apart based on any of the observed macro data that we have.[9]

[8] In the model, the inflation differential is the same whether measured in terms of consumption or the output of the locally-produced goods—i.e. $\widehat{\pi}_t - \widehat{\pi}_t^* \approx \widehat{\pi}_t^H - \widehat{\pi}_t^{F*}$. It is worth noting as well that the model implies growth differentials across countries should be reflected in the differentials between the inflation rates calculated on consumption and output (akin to the CPI and the GDP deflator respectively), i.e. $\widehat{\pi}_t - \widehat{\pi}_t^H \approx \xi \left(\frac{1}{\sigma - (\sigma-1)(1-2\xi)^2} \right) (\Delta \widehat{y}_t - \Delta \widehat{y}_t^*)$ and $\widehat{\pi}_t^* - \widehat{\pi}_t^{F*} \approx -\xi \left(\frac{1}{\sigma - (\sigma-1)(1-2\xi)^2} \right) (\Delta \widehat{y}_t - \Delta \widehat{y}_t^*)$.

[9] The fact that we include inflation among our observables, however, may help distinguish the models with price stickiness

Experiment with the Policy Parameter ψ_π. To investigate Bayesian model comparison, we evaluate the implications of increasing the similarity between all four competing models with the selection of observables indicated before. We simulate data and compare all four models on an interval that spans ψ_π between 0 and 6 under the benchmark parameterization that sets the import share in the consumption basket ξ at a low value of 0.06 in order to increase the similarity between all four competing models as tolerance for inflation declines—that is, as ψ_π increases while keeping the Calvo-parameter α unchanged at 0.75. All the posterior model probabilities from this experiment are summarized in Figure 1.

[Insert Figure 1 about here.]

As expected, posterior model probabilities favor the true DGP (the NOEM model, M_1) as the sample size gets asymptotically large—which is what we find with a long sample of $10,000$ simulated observations. Interestingly, the international transmission mechanism is weak enough that under reasonable parameterizations of the monetary policy rule the closed-economy model M_3 can still appear as the preferred one. Moreover, we show that the more parsimonious model M_3 may get the upper hand in samples of 160 simulated observations based on the computed posterior model probabilities—see sub-sample 3 in Figure 1—whenever monetary policy is more aggressive.

The crucial difference between these two models is that M_1 (the true model) features nominal rigidities that result in monetary policy non-neutrality and is open to trade, while monetary policy remains neutral in M_3 but there is no endogenous transmission of shocks across countries as households' in each do not demand imported varieties of goods from each other. Therefore, if we were to wrongly conclude on the basis of the evidence available that M_3 is preferred by the data, we may also wrongly conclude that a loosening of monetary policy has no real effects and spillovers on the economic activity of the other country (when it actually does!).

The implications of that policy *mistake*, of course, would only become obvious if the policy change were to be implemented and take effect. This may result in an incorrect identification of the source of business cycle fluctuations as endogenous international spillovers would be attributed in the M_3 model to the country-specific shocks and, in particular, to the exogenous covariance of the innovations. It would be too late to find out *ex post* that the expected dynamic implications of this policy shock were predicated on a misspecified model that did not take into account households' true preferences for imported varieties from other countries and the potential impact that intratemporal smoothing consumption through trade can have.

Experiment with the Structural Parameter ξ. A smaller share of imported goods in the consumption basket under the NOEM model (model M_1) shuts down the key channel for the endogenous international transmission of shocks, resulting in an allocation closer to autarky (as in model M_3). Therefore, a lower value of ξ should result in allocations that make it increasingly more difficult to distinguish between models M_1 and M_3. Similar to what we did for the policy parameter, we simulate data and compare all four alternative models on an interval that spans ξ between 0 and $\frac{1}{2}$. All the posterior model probabilities from this experiment are summarized in Figure 2.

[Insert Figure 2 about here.]

from those under flexible prices if inflation is set to zero under flexible prices. However, more generally, flexible prices only imply that the output gap ought to be zero but it does not constraint inflation. If we adopt the same monetary policy specification as for the NOEM model, this pins down a non-zero inflation rate that becomes increasingly less informative about the presence of nominal rigidities in the model as the allocation of all four models becomes more similar.

While the asymptotic results validate the true DGP (the NOEM model, M_1) when we look at a long sample of $10,000$ observations, we see that again it is possible to argue in favor of the more parsimonious closed-economy model M_3 in 160-observations samples based on the computed posterior model probabilities—see Figure 2—whenever the import shares are small enough. As before, we would be missing out the endogenous transmission mechanism that comes from trade by selecting model M_3. The crucial difference from the policy-makers point of view between these two models is that M_1 (the true model) defines the relevant trade-off for monetary policy to be between domestic inflation and global slack (the *global slack hypothesis*) while model M_3 represents the standard closed-economy view which postulates that the monetary policy trade-off that arises from nominal rigidities is between domestic inflation and domestic output. Selecting the wrong model in this case would result in an incorrect identification of the sources of business cycle fluctuations and how they are transmitted across countries, as we argued before. However, it can also lead policy-makers to ignore the role and consequences of foreign factors in the dynamics of inflation when setting monetary policy or in evaluating a policy change.

One of the major concerns for us would be that model comparison in small samples may contribute to such policy *mistakes*. However, we also recognize that this is a selection error that could have easily been avoided just by looking at trade itself. Since model M_1 implies non-zero imports while model M_3 imposes zero imports, both predictions are incompatible and so one of the two models can be easily refuted in the data. Therefore, more generally we expect that the selection of observables for estimation can presumably help us avoid some of these *mistakes*.

Selection of Observables for Estimation. In the benchmark implementation described so far, we make model comparisons based on Home and Foreign output, as well as Home and Foreign inflation. Now we experiment with the selection of an alternative set of four observables replacing Foreign output with the terms of trade for Bayesian estimation. Guerron-Quintana (2010) shows that Bayesian estimation and structural identification can be sensitive to the selection of observables, and not too surprisingly we find that posterior model probabilities are also sensitive to our selection of observables in small samples.

Using terms of trade data as an observable is meant to reveal further information about the trade channel for the international transmission of shocks. Figures 3 and 4 replicate the experiments behind Figures 1 and 2 respectively—everything remains the same in our implementation and estimation, except for the fact that we are using now a different set of observables to estimate each model. The evidence confirms that the posterior model probabilities are unperturbed by the alternative combinations of observables used for estimation when arbitrarily large samples are available. However, we see that the information content of the terms of trade can work to either revert or worsen the erroneous preference documented earlier toward the more parsimonious, closed-economy models that may arise in small samples.

[Insert Figures 3 and 4 about here.]

Using terms of trade data as an observable, we also investigate in Figure 5 a range of values for the parameter α that determines the degree of price stickiness present in the economy. In this case, posterior model probabilities favor the true DGP (the NOEM model, M_1) as the sample size gets arbitrarily large—but, interestingly, we find that $10,000$ quarterly observations ($2,500$ years of data!) may not be *large enough* to pick the true DGP if α is *too low*. We also show that the international transmission mechanism is weak enough that under reasonable parameterizations of the stickiness parameter α, the closed-economy model

M_3 can become the preferred one in small samples (see sub-sample 2 in Figure 5). Moreover, we also show that the flexible price specifications M_2 and M_4 may get the upper hand as well in sub-samples of 160 observations when α is low—see sub-samples 1 and 3 in Figure 5.

The crucial difference between these two alternative models (M_2 and M_4) and model M_1 (the true model) is that the true DGP features nominal rigidities that result in monetary policy non-neutrality, while monetary policy is neutral whenever prices are flexible. Therefore, if we were to wrongly conclude on the basis of the evidence available that either M_2 or M_4 are preferred by the data over M_1, we may also wrongly conclude that a loosening of monetary policy has no real effects on economic activity (when it actually does!). The implications of that policy *mistake*, of course, would only become obvious if a policy change were to be implemented and take effect by which time it would already be too late to find out that this policy choice was predicated on a misspecified model.

[Insert Figures 5 about here.]

Common sense suggests that one may want to experiment with a number of possible combinations of observable variables as a robustness check. In practice, though, the selection may be already significantly limited due to data problems (quality) and due to availability limitations. However, exploring alternative sets of observables whenever that is feasible is only a practical recommendation that can help us determine how robust the support for a particular model is—it does not say anything about the deeper question of how we should choose the model favored by the data whenever alternative combinations of observables produce contradictory evidence (that is, when they produce significantly different posterior model probabilities). It does not offer us further guidance on how to select the *appropriate* set of observables for a given model either.

In our exercise, however, we have the advantage—uncommon in applied macroeconometrics work—to know the true DGP underlying the data and so we can dig a little deeper into these results based on simulated data. The macro observables that are common to all four models are all ultimately related to two core variables per country that characterize the dynamics of the NOEM model—inflation and the output gap (which given a specification for potential output can be related to the observable measure of output), whose dynamic path is characterized by a solution of the form presented in $(13)-(14)$. Not surprisingly, as different models become more similar in the path they imply for output and inflation, they also appear closer when we use an alternative set of observables that are in effect linear combinations of output and inflation themselves together with the structural shock innovations. Naturally, if Bayesian model comparison methods fail to select the correct specification in small samples with the standard selection of observables that includes the core variables, they may tend to produce false signals in small samples with other alternative selection of observable variables—as we have seen here.

Variable selection, in this case, could contribute to attenuate the problem or simply help us detect whether a selection problem exists (when it gives different predictions with alternative observables), but it cannot in general avoid the problem entirely for us. In other words, when model specifications become arbitrarily *close*, the selection of observables for estimation cannot help us consistently avoid the preference toward more parsimonious specifications that we have found in our small sample experiments.

4 Discussion

We have a collection of $k \geq 2$ models each of which is fully-described with a parameterized joint probability density over the vector of observable (endogenous) variables Z, i.e.

$$M_i = \{f_i(z \mid \theta_i) : \theta_i \in \Theta_i\}, \ \forall i = 1, ..., k, \tag{17}$$

where θ_i is the vector of unknown parameters of model M_i, Θ_i is the parameter space and $d_i = \dim(\Theta_i)$ its dimension, $f_i(z \mid \theta_i)$ is the parameterized probability density, and z is a given realization of the vector of observable variables Z.

The likelihood function for model M_i, given n observations of the observable variables $z^n = (z_1, ..., z_n)$, is the probability of z^n occurring under the probability density that describes model M_i given the vector of parameters θ_i, i.e. $L_i(\theta_i) \equiv f_i(z^n \mid \theta_i)$. We refer to the log-likelihood function for model M_i as $l_i(\theta_i)$ and represent it as follows,

$$l_i(\theta_i) \equiv \ln f_i(z^n \mid \theta_i) = \sum\nolimits_{j=1}^{n} \ln f_i(z_j \mid \theta_i), \ \forall i = 1, ..., k. \tag{18}$$

We assign prior probabilities, $\Pr(M_i)$, to all model specifications $i = 1, ..., k$, and also prior probabilities to the parameters θ_i that characterize each model, $f_i(\theta_i)$.

The marginal likelihood $m_i \equiv f_i(z^n \mid M_i)$ of any model M_i, $i = 1, ..., k$, is referred to as the model evidence, and it is defined by the expectation taken over the likelihood function $L_i(\theta_i) \equiv f_i(z^n \mid \theta_i)$ with respect to the prior distribution of the parameters $f_i(\theta_i)$, i.e.

$$m_i \equiv f_i(z^n \mid M_i) = \int_{\theta_i} f_i(z^n \mid \theta_i) f_i(\theta_i) \, d\theta_i, \ \forall i = 1, ..., k. \tag{19}$$

The posterior probability for model M_i can be calculated using Bayes' Theorem as,

$$\Pr(M_i \mid Z^n = z^n) = \frac{f_i(z^n \mid M_i) \Pr(M_i)}{\sum_{p=1}^{k} f_p(z^n \mid M_p) \Pr(M_p)} = \frac{m_i \Pr(M_i)}{\sum_{p=1}^{k} m_p \Pr(M_p)}, \ \forall i = 1, ..., k, \tag{20}$$

where the marginal likelihood m_i providing evidence for model M_i times the prior assigned to that particular model is normalized with respect to the model evidence times the model prior of all $k (\geq 2)$ models under consideration.

The Bayesian posterior odds for model M_1 versus the alternative model M_i, $i = 2, ..., k$, summarize the relative support that the data provides for one specification over the other with the ratio of their posterior probabilities, i.e.,

$$\frac{\Pr(M_1 \mid Z^n = z^n)}{\Pr(M_i \mid Z^n = z^n)} = \frac{m_1}{m_i} \frac{\Pr(M_1)}{\Pr(M_i)}, \ \forall i = 2, ..., k. \tag{21}$$

The posterior odds in favor of model M_1 against an alternative specification M_i, $i = 2, ..., k$, can be expressed as the product of the prior odds $\frac{\Pr(M_1)}{\Pr(M_i)}$ in favor of M_1 times the corresponding Bayes Factor defined by the ratio $B_{1i} = \frac{m_1}{m_i}$, as can be seen in (21). The marginal likelihood is key to calculate the Bayes Factor—which is the quotient of the marginal likelihoods of the two alternative models. Therefore, marginal likelihoods are also crucial to derive the Bayesian posterior odds in (21) conventionally used for Bayesian model selection.

Though there are alternative approaches to compute the Bayes Factor and the Bayesian posterior odds—such as the (generalized) Savage-Dickey density ratio discussed by Verdinelli and Wasserman (1995)—the method based on the marginal likelihood remains the most common in applied macro-econometrics work with Bayesian techniques. We rely on the computation of the marginal likelihood in all our experiments as well. Here we review some aspects of the estimation and the computation of the marginal likelihoods that can directly affect the assessment of competing models based on Bayesian posterior odds, with special emphasis on understanding what contributes to explain the *false signals* in model selection that we have encountered in our experiments with small samples.

Interpreting Our Findings: The Role of Sample Size. After specifying the priors over the models and over the model parameters, the practical difficulty in calculating posterior probabilities or posterior odds is computing the marginal likelihood defined in (19). Only in very special cases we can calculate the marginal likelihood analytically—most notably for the exponential likelihood family with conjugate priors, as in the case of Gaussian linear models (see, e.g., Zellner (1971)). In practice, analytical solutions are often intractable and computational methods are needed.

Among the different methods available to approximate the marginal likelihood, we can list: (a) asymptotic approximations (Laplace's method, Schwarz Criterion, BIC); (b) numerical integration (e.g., Gaussian quadrature), importance sampling and annealed importance sampling (see, e.g., Geweke (1989), Neal (2001)); (c) posterior distribution simulations (e.g., Markov Chain Monte Carlo (MCMC) methods like the Metropolis-Hastings algorithm and the Gibbs sampler); and (d) variational inference (see, e.g., Corduneanu and Bishop (2001)), expectation propagation (see, e.g., Minka (2001)).

We use the Laplace approximation to compute the marginal likelihood of a given specification and derive the Bayesian posterior odds for model comparison. Asymptotic approximation methods such as Laplace's method rely on normal asymptotic approximations of the marginal likelihood. These methods work well in most familiar problems, are accurate, easy to compute and fast. They provide adequate approximations especially for well-behaved posterior densities that are highly-peaked and unimodal, since asymptotic approximations rely on a normal density to approximate the posterior density.

The Gaussian state space representation in $(13)-(14)$ implies that the likelihood of the model, $L_i(\theta_i) = f_i(z^n \mid \theta_i)$, is characterized by a normal distribution under the DGP as well as under any of the alternative specifications we propose for model comparison. Hence, in our case the likelihood of the models we investigate is known to be Gaussian. The specification of the prior distributions for the model parameters then plays an important role to retain the highly-peaked and unimodal shape on the posterior density and, therefore, to ensure that the Laplace approximation is reasonably accurate.

In our illustrations of Bayesian model comparison, the choice of the Laplace approximation method appears reasonable on grounds of computational accuracy. For other models, however, approximation methods may not attain accurate estimates of the marginal likelihood. In that case, alternative ways to compute the marginal likelihood should be pursued in order to avoid model selection errors due to inaccurate estimates of the marginal likelihood. An evaluation of the advantages and disadvantages of alternative methods to compute the marginal likelihood—especially when models are less well-behaved than the ones considered here—goes beyond the scope of this paper. We leave it for future research.

Apart from the reasonable accuracy attained in our exercise, we also discuss this asymptotic approximation method in greater detail here to gain further insight on the role of sample size and the penalization of

overfitting that is inherent in Bayesian posterior odds calculations.

Laplace's Approximation Method: Accuracy and Sample Size. The Laplace's (or Gaussian) method which we apply in our experiments with Bayesian model comparison is based on the idea that asymptotically the posterior distribution can be approximated with a multivariate Gaussian distribution (see, e.g., Kass et al. (1988)). Let $\widehat{\theta}_i$ be the posterior mode which is defined as the vector of parameters θ_i that maximizes the posterior probability $f_i(\theta_i \mid z^n)$ that characterizes model M_i. The posterior probability is proportional to the likelihood function times the model parameters' priors, i.e. $f_i(\theta_i \mid z^n) \propto f_i(z^n \mid \theta_i) f_i(\theta_i))$, so the optimization required to derive the posterior mode can be defined as,

$$\widehat{\theta}_i \;=\; \arg\max_{\theta_i} \; \{\ln(f_i(z^n \mid \theta_i) f_i(\theta_i))\}, \tag{22}$$

where $h_i(\theta_i) \equiv \ln(f_i(z^n \mid \theta_i) f_i(\theta_i))$ is a log-transformation that also maximizes the posterior probability. The first-order conditions of the maximization problem in (22) imply that $\bigtriangledown h_i\left(\widehat{\theta}_i\right) = 0$.

Expanding $h_i(\theta_i)$ as a quadratic function around $\widehat{\theta}_i$ we obtain that,

$$h_i(\theta_i) \approx h_i\left(\widehat{\theta}_i\right) + \bigtriangledown h_i\left(\widehat{\theta}_i\right)\left(\theta_i - \widehat{\theta}_i\right) - \frac{1}{2}\left(\theta_i - \widehat{\theta}_i\right)' H\left(\widehat{\theta}_i\right)\left(\theta_i - \widehat{\theta}_i\right), \tag{23}$$

where $H\left(\widehat{\theta}_i\right) = -D^2 h_i(\theta_i)$ is the negative Hessian of second derivatives of $h_i(\theta_i)$ evaluated at $\widehat{\theta}_i$. Replacing the first-order conditions from (22) and exponentiating (23) yields an approximation of $f_i(z^n \mid \theta_i) f_i(\theta_i)$ that has the form of a normal density with mean $\widehat{\theta}_i$ and covariance matrix $H\left(\widehat{\theta}_i\right)^{-1}$. Integrating that expression we obtain the Laplace approximation of the marginal likelihood, i.e.,

$$\ln m_i \;\approx\; \ln\left(f_i\left(z^n \mid \widehat{\theta}_i\right)\right) + \ln\left(f_i\left(\widehat{\theta}_i\right)\right) + \frac{d_i}{2}\ln(2\pi) - \frac{1}{2}\ln\left|H\left(\widehat{\theta}_i\right)\right| \equiv \ln m_i|_{Laplace}, \tag{24}$$

where d_i is the dimension of the parameter space Θ_i of model M_i for any $i = 1, ..., k$.

Kass et al. (1988) and Kass et al. (1990) show that, under certain regularity conditions, errors in this approximation are bounded by $O_P\left(n^{-1}\right)$ where n is the number of observations used in the estimation. We can also obtain an $O_P\left(n^{-1}\right)$ approximation of the marginal likelihood with $\widehat{\theta}_i^{MLE}$ being the maximum likelihood estimator (MLE) and $H\left(\widehat{\theta}_i^{MLE}\right)$ being the observed information matrix (that is, the negative of the Hessian matrix evaluated at the MLE estimator, $\widehat{\theta}_i^{MLE}$) in (24). The inverse of the Fisher information matrix (i.e. the inverse of the expected information matrix which converges as n grows to the inverse of the asymptotic covariance matrix) can also be used in (24), but at the expense of incurring a greater approximation error in the computation of the marginal likelihood of order $O_P\left(n^{-\frac{1}{2}}\right)$.

Thus, when Laplace's method is applied to both the numerator and denominator of the Bayes Factors $B_{1i} = \frac{m_1}{m_i}$ in (21) to compare M_1 against any other alternative specification M_i, $i = 2, ..., k$, the resulting approximation of the Bayes Factors retains an approximation error of order $O_P\left(n^{-1}\right)$ (or of order $O_P\left(n^{-\frac{1}{2}}\right)$ if the Fisher information matrix is used).[10] For many problems for which the sample size n is moderate and the likelihood is reasonably approximated by that of a normal distribution, the Laplace method produces accurate and easy to compute approximations of the marginal likelihood and the Bayes Factors.[11]

[10] See, e.g., the discussion on page 778 of Kass and Raftery (1995) of the approximation error of the Bayes Factors of nested models under the Laplace method.

[11] The Gaussian state-space representation of the solution implies the normality of the likelihood for the models investigated in

Hence, the Laplace approximation to computing marginal likelihoods seems reasonable in our illustrations in part because the Gaussian state-space representation of the solution ensures the normality of the likelihood and the posterior densities are expected to be well-behaved and single-peaked. It is reasonable also because the sample sizes more relevant to us are sufficiently large so that the approximation error is negligible and the computed Bayes Factors adequately accurate. In providing guidance on the sample size required to attain an adequate approximation with the Laplace method, we follow the recommendations of Kass and Raftery (1995).

Kass and Raftery (1995) warn us that sample sizes of less than $5d_i$ observations may be insufficient to attain an accurate approximation of the marginal likelihood with the Laplace method, where d_i is the dimension of the parameter space of model M_i. In turn, sample sizes greater than $20d_i$ should be *large enough* to ensure the method works well in most cases in which the likelihood function itself is not too different from that of a normal distribution. However, we must recognize that a sample size of $20d_i$ observations appears increasingly out of reach in practice for most heavily parameterized medium- and large-scale DSGE models.

In the experiments reported in this paper, the most parameterized specification is the DGP (model M_1) which includes 12 parameters (not counting the calibrated intertemporal discount factor, β). All other specifications have fewer than 12 parameters. We set the small sample size in our experiments to $n = 160$ quarterly observations (40 years of quarterly data). This implies that all models under consideration are above the threshold of $5d_i$ observations suggested by Kass and Raftery (1995) and, in fact, come close to the $20d_i$ threshold in our case.

We are neither interested in very long sample sizes that should lead to the correct outcome in model selection but are generally not available for applied work, nor in the very short samples where the posterior densities are still largely dominated by the priors we place on the model parameters. In turn, we examine in our experiments a sample range in between those which is more realistic for applied work (given the length of data that is generally available) and relevant. Our notion of a small sample size in practice satisfies the following broad criteria:

(a) The sample size is *large enough* so that the Laplace approximation works well given an expected approximation error of order $O_P\left(n^{-1}\right)$ (or of order $O_P\left(n^{-\frac{1}{2}}\right)$) and surpasses the lower threshold recommended by Kass and Raftery (1995).

(b) The sample size is *large enough* so that there is enough data to overwhelm the priors.

(c) The sample size is *not too large* so that the penalization for overfitting that we highlight in this paper still has bite to tilt the posterior odds in favor of the most parsimonious specification (and at the expense of selecting the wrong model).

Under this notion of a small sample, the Laplace method suffices for our purpose of providing an accurate assessment of the problem of *false signals* in Bayesian model selection—a problem that arises, as can be seen in our illustrations, whenever very large sample sizes of observations are not available for the estimation and a problem that otherwise would be masked by the priors for very short sample sizes.

Laplace's Approximation Method: Overfitting Penalization and Sample Size. Apart from the appropriateness of the Laplace method given the notion of a small sample that we investigate here, this asymptotic approximation also helps us shed some light on the role that sample size n and the dimensionality of the

this paper. Sample size is a determinant factor on the appropriateness of the normal approximation for the posterior distribution. Slate (1994) provides guidance on the sample size requirements needed to obtain posterior normality and guarantee the accuracy of the Laplace's method for the exponential distribution family. The normal, gamma, and beta among other well-known distributions belong to the exponential family.

parameter space d_i of a given model M_i, $i = 1, ..., k$, play on the calculations of the marginal likelihood, the Bayes Factors and the Bayesian posterior odds for model comparison.

As the sample size n grows, the different terms of the Laplace approximation to the marginal likelihood grow at different rates. The log-likelihood function should grow proportionally to n, the size of the penalization for overfitting that arises from the Hessian term $\ln\left|H\left(\widetilde{\theta}_i\right)\right|$ increases at the rate of $d_i \ln(n)$ which also depends on the dimensionality of the parameter space, while the remaining approximation terms are invariant with sample size but depend on the choice of priors and the dimensionality d_i.

More generally, the different terms of the Laplace approximation of the marginal likelihood in (24) grow with sample size n as indicated here,

$$\ln m_i|_{Laplace} \approx \underbrace{\ln\left(f_i\left(z^n \mid \widetilde{\theta}_i\right)\right)}_{O(n)} + \underbrace{\ln\left(f_i\left(\widetilde{\theta}_i\right)\right)}_{O(1)} + \underbrace{\frac{d_i}{2}\ln(2\pi)}_{O(1)} - \frac{1}{2}\underbrace{\ln\left|H\left(\widetilde{\theta}_i\right)\right|}_{O(d_i \ln n)}. \tag{25}$$

For any given sample size for which this approximation holds, there is a penalty for the dimensionality of the model d_i that comes from the last two terms in the right-hand side of (25) and varies with n. When Laplace's method is applied to both the numerator and denominator of the Bayes Factors $B_{1i} = \frac{m_1}{m_i}$, $i = 2, ..., k$ using (25), the resulting approximation to compare the model evidence of M_1 against that of any other alternative specification M_i, $i = 2, ..., k$ can be expressed as follows,

$$\begin{aligned}\ln B_{1i}|_{Laplace} \approx & \underbrace{\left(l_1\left(\widetilde{\theta}_1\right) - l_i\left(\widetilde{\theta}_i\right)\right)}_{O(n)} + \underbrace{\left(\ln\left(f_1\left(\widetilde{\theta}_1\right)\right) - \ln\left(f_i\left(\widetilde{\theta}_i\right)\right)\right)}_{O(1)} + ... \\ & \underbrace{\left(\frac{(d_1 - d_i)}{2}\ln(2\pi)\right)}_{O(1)} - \frac{1}{2}\underbrace{\left(\ln\left|H\left(\widetilde{\theta}_1\right)\right| - \ln\left|H\left(\widetilde{\theta}_i\right)\right|\right)}_{O((d_1 - d_i)\ln n)}.\end{aligned} \tag{26}$$

Similarly, this can be extended to approximate the Bayesian posterior odds defined in (21).

At moderate sample sizes for which the Laplace approximation seems appropriate, the penalty for overfitting can become the deciding factor to understand why Bayesian model comparison may favor parsimony even at the expense of selecting the wrong model. As the sample size n keeps growing, the differences in the log-likelihood function $\left(l_1\left(\widetilde{\theta}_1\right) - l_i\left(\widetilde{\theta}_i\right)\right)$ should grow proportionally to n while the size of the penalty increases at the rate of $(d_1 - d_i)\ln n$. Hence, the overfitting penalty embedded here is a relatively harder threshold to meet in samples of moderate length such as the ones we explore in all our illustrations whenever the probability densities that characterize each competing model are arbitrarily *close* to each other.

In other words, for moderate sample sizes it might occur that the Bayesian posterior odds favors the less parameterized model if the log-likelihood differences between the models under comparison are too small to outweigh the overfitting penalty found in (25). Otherwise, researchers would require unrealistically large sample sizes to be able to consistently identify the correct model when the correct specification is more heavily parameterized than the alternative. That explains mechanically why in our experiments we validate the asymptotics in Fernández-Villaverde and Rubio-Ramírez (2004) but still find that the more parsimonious model could be the one picked up against the more complex true specification (even when using 40 years of quarterly data for that!).

BIC's Approximation Method: An Alternative Trade-off Between Accuracy at a Given Sample Size and the

Role of Priors for Model Selection. A more efficient, but (in general) less accurate asymptotic approximation is obtained by: (a) using a consistent, likelihood-based estimator $\widetilde{\theta}_i$ to evaluate the approximation (naturally, the MLE estimator, $\widetilde{\theta}_i = \widehat{\theta}_i^{MLE}$ can used for this); (b) retaining only those terms in equation (24) that increase with the sample size n, i.e. dropping $\ln\left(f_i\left(\widetilde{\theta}_i\right)\right) + \frac{d_i}{2}\ln(2\pi)$ which do not increase with n; and (c) using the fact that for large n, the determinant $\left|H\left(\widetilde{\theta}_i\right)\right|$ is proportional to n^{d_i}. This approximation is called the Schwarz criterion and takes the form,

$$\widehat{m}_i \approx l_i\left(\widehat{\theta}_i^{MLE}\right) - \left(\frac{d_i}{2}\right)\ln(n), \tag{27}$$

where $l_i\left(\widetilde{\theta}_i\right) = \ln\left(f_i\left(z^n \mid \widetilde{\theta}_i\right)\right)$ is the log-likelihood function evaluated at the value of the estimator $\widetilde{\theta}_i$. The right-hand side in (27) is equal to the Schwarz criterion for model selection where d_i is the dimension of the parameter space Θ_i of model M_i for any $i = 1, ..., k$. This approximation was first derived by Schwarz (1978) (see also Akaike (1978)).

Kass and Wasserman (1995) show that under regularity conditions similar to those for the Laplace approximation, the Schwarz criterion satisfies that,

$$m_i = \widehat{m}_i + O_P(1), \tag{28}$$

where $\widetilde{\theta}_i$ is a consistent, likelihood-based estimator (or simply the MLE estimator $\widehat{\theta}_i^{MLE}$ as indicated before). Moreover, the relative error of the approximation tends to zero in probability, i.e. $\frac{|\widehat{m}_i - m_i|}{|m_i|} \underset{P}{\to} 0$. Notice that minus twice the Schwarz criterion is the Bayesian Information Criterion (or BIC). Hence, the BIC provides an $O_P(1)$ approximation for the marginal likelihood as well. The Schwarz criterion and by extension the BIC are in effect $O_P(1)$ approximations to the marginal likelihood.

The BIC approximation is appealing for model comparison in a number of respects that we highlight here:

First, it does not depend on the prior assigned to the vector of parameters. So this procedure can be applied to compute the marginal likelihood even when the priors $f_i(\theta_i)$ are difficult to set precisely or are debated in the literature. This is an important consideration in applied work where we often don't have strong reasons to favor one particular prior distribution over others.

Second, the BIC is related to the Minimum Description Length (MDL) stochastic complexity measure proposed by Rissanen (1987). In recent years, MDL has received much attention in the literature on statistical model selection as it allows for a unified treatment of model selection and statistical inference. The MDL measure provides a quantification of the goodness of fit that can be attained with a given probability distribution to account for the statistical regularities observed in the data. From the work of Rissanen (1996) and Qian and Künsch (1998) it follows that the MDL-proposed measure of stochastic complexity of the observed data relative to a given parameterized model can be expressed as minus the maximum log-likelihood plus a model complexity term that is determined by the Fisher information matrix and the MLE estimator of the model parameters. In this sense, the BIC approximation we consider here is minus the MDL measure of stochastic complexity. Hence, our findings using the BIC can be interpreted in light of what the MDL principle stands for as well.

Third, the Laplace and the BIC approximations should be asymptotically equivalent for large sample

sizes, i.e.

$$\ln m_i|_{Laplace} \overset{n\to\infty}{\longrightarrow} \underbrace{l_i\left(\widetilde{\theta}_i\right) - \left(\frac{d_i}{2}\right)\ln(n)}_{=Schwarz\ criterion=-\frac{1}{2}BIC}, \tag{29}$$

under some conditions. The BIC approximation may be viewed as a rough approximation to the log of the marginal likelihood. We say that BIC and the Laplace method are asymptotically correct, though, because they both select a model whose posterior probability is a maximum whenever n becomes sufficiently large. Moreover, as indicated by (29), the BIC and Laplace methods must agree on the selected model as sample size n becomes arbitrarily large.

Fourth, the BIC approximation to the Bayes Factor B_{1i} that compares model M_1 against the alternative model M_i for any $i = 2, ..., k$, i.e.,

$$BIC_{1i} = -2\left[l_1\left(\widetilde{\theta}_1\right) - l_i\left(\widetilde{\theta}_i\right) - \left(\frac{d_1 - d_i}{2}\right)\ln(n)\right] \tag{30}$$

satisfies, as shown by Kass and Raftery (1995), that as $n \to \infty$,

$$\frac{-\frac{1}{2}BIC_{1i} - \ln B_{1i}}{B_{1i}} \underset{P}{\to} 0,\ \forall i = 2, ..., k. \tag{31}$$

In contrast to the Laplace approximation, the relative error of $\exp\left(-\frac{1}{2}BIC_{1i}\right)$ in approximating the Bayes Factor B_{1i} is generally of order $O_P(1)$.[12] For the moderate and large sample sizes n for which this result holds, the error bounds of the approximation would not increase with the sample size itself. This is a rough approximation, but one that should give us a reasonable indication of the evidence for the sample sizes that we use in our illustrations of Bayesian model comparison in this paper.

Under some conditions applying to nested models such as the ones considered in our work, the BIC approximation under unit information priors is accurate to order $O_P\left(n^{-\frac{1}{2}}\right)$ (see Kass and Wasserman (1995) and Kass and Raftery (1995)).[13] Thus, if one is willing to consider these priors as suitable, then the BIC (and the Schwarz criterion) can be thought as providing a reasonably good approximation to the log of the Bayes Factors that is comparable in terms of the accuracy attained for moderate and large sample sizes to that of the Laplace method using the Fisher information matrix.

Fifth, the BIC approximation is quite intuitive and easy to interpret retaining the penalization for

[12] We can re-write the posterior model probability in (20) corresponding to model M_i for any $i = 2, ..., k$ in terms of Bayes Factor with respect to model M_1, B_{i1}, as follows,

$$\Pr(M_i \mid Z^n = z^n) = \frac{B_{i1}m_1 \Pr(M_i)}{\sum_{p=1}^{k} B_{p1}m_1 \Pr(M_p)} = \frac{e^{-\ln B_{1i}} \Pr(M_i)}{\sum_{p=1}^{k} e^{-\ln B_{1p}} \Pr(M_p)},\ \forall i = 2, ..., k,$$

where in the second equality we use the fact that $B_{i1} = \frac{1}{B_{1i}}$ for all $i = 2, ..., k$. Then, it is possible to use the approximation result in (31) to express the posterior model probability in terms of the BIC as defined in (30), i.e.

$$\Pr(M_i \mid Z^n = z^n) \approx \frac{e^{-\frac{1}{2}BIC_{1i}} \Pr(M_i)}{\sum_{p=1}^{k} e^{-\frac{1}{2}BIC_{1p}} \Pr(M_p)} \propto e^{-\frac{1}{2}BIC_{1i}} = e^{\frac{1}{2}BIC_{i1}}.$$

Posterior model probabilities and the BIC are related up to an approximation of order $O_P(1)$ as well, and should be asymptotically equivalent (under weak conditions).

[13] The unit information prior is a data-dependent prior, (typically multivariate Normal) with mean at the MLE estimator, and precision equal to the information provided by one observation.

overfitting indicated before with the Laplace approximation in (25). The BIC approximation contains a term evaluating how much better (or worse) one model with parameters set to their consistent, likelihood-based estimates fits the data relative to an alternative model also evaluated with parameters at their consistent, likelihood-based estimates (i.e. $l_1\left(\widetilde{\theta}_1\right) - l_i\left(\widetilde{\theta}_i\right)$) and another term that punishes the added complexity of one model over the other (i.e. $\left(\frac{d_1-d_i}{2}\right)\ln(n)$).[14]

This confirms the simple interpretation given before of one of the plausible explanations of the *false signals* problem that we have illustrated in the experiments described in the previous section. It suggests posterior model probabilities can favor the wrong model specification in part because of the penalization of complexity that comes with it, as can be inferred from (30).

Other Considerations in Evaluating Bayesian Posterior Odds for Model Comparison. Our discussion thus far provides a qualitative interpretation of the reported findings, but one that is ultimately dependent on the accuracy of the approximation of the marginal likelihood used. We have argued that using the Laplace approximation is reasonable given the characteristics of the solution to the models we are investigating and the fact that we explore moderate and large sample sizes for which the approximation should hold. We conclude that unless we have an arbitrarily large sample size, standard Bayesian posterior odds may still favor parsimony even when the true model specification is more complex. This is well-understood on the basis of the Laplace approximation. What we do not have from this is a quantitative rule to determine the sample size that would be needed to accurately and consistently select the true model overcoming the penalization for overfitting. We are unable to be much more specific than this since assessing the sample sizes required to avoid the problem of *false signals* is likely to be model-dependent, and to vary for different families of probability distributions.

Finally, we discuss a number of related points regarding the implementation of the Bayesian estimation of a model (such as parameter identification, the choice of priors, the selection of observables, etc.) that can affect the fit of the competing specifications under comparison and consequently also lead to erroneous model selection for small sample sizes.

Parameter Identification. Identification can refer to the mapping from the deep parameters of the model to the reduced-form parameters that characterize a unique solution as in (13) − (14). As indicated before, Blanchard and Kahn (1980) provides conditions under which such a unique stable solution exists. In this regard, the conventional practice is to set the range of the prior distributions to avoid or minimize the draws of that come from regions of the parameter space for which no solution exists or where indeterminacy arises. Although the unique solution is linear, the reduced-form parameters are generally non-linear functions of the deep parameters—reflecting the cross-equation restrictions implied by the model.

Identification also refers in our context to the mapping from the solution to the observable data, and the conditions under which a unique likelihood function $L_i(\theta_i) = f_i(z^n \mid \theta_i)$ exists. Identification problems in this latter sense arise if distinct parameter values do not result in different probability distributions of the

[14] The BIC is part of a family of competing penalized likelihood functions that also includes the Akaike Information Criterion, the Deviance Information Criterion (DIC) or the Takeuchi Information Criterion (TIC). These functions differ mostly on the penalty they impose for overfitting. The AIC has a fixed penalty that does not grow with $\ln(n)$, i.e. $l_i\left(\widetilde{\theta}_i\right) - d_i$ where d_i is the dimensionality of the parameter space. Although it can be shown that AIC is optimal in the sense of minimizing the Kullback-Leibler (KL) divergence, when it comes to model selection it is not consistent asymptotically unlike the BIC. For sample sizes of 8 or more observations, BIC has a higher penalty for overfitting than the Akaike Information Criterion (AIC). Hence, since a sample size of less than 8 observations is unrealistic, we can say nonetheless that BIC penalizes complex models more than other well-known model selection criteria such as AIC.

data, i.e. θ_i is identified if $f_i\left(z^n \mid \theta_i^{(1)}\right) = f_i\left(z^n \mid \theta_i^{(2)}\right)$ implies that $\theta_i^{(1)} = \theta_i^{(2)}$ for all z^n (see, e.g., Hsiao (1983), pp. 226-227). If identification fails to hold, no estimation procedure can pin down uniquely the vector of parameters θ_i irrespective of the sample size. Bayesian estimation only circumvents the problem by using priors and, as Canova and Sala (2009) point out, may end up concealing the problems of identification that way.

It is recognized that lack of identification leads to wrong inferences and can significantly affect our estimates of a model (see, e.g., Ríos-Rull et al. (2012), and Martínez-García et al. (2012)). The lack of identification can also be a problem for Bayesian model selection, as we need to compute the marginal likelihood of models to derive their posterior odds with a badly-shaped likelihood function due to lack of identification. The issue is rarely addressed in applied work where identification is not usually explicitly verified before estimation. We argue that checking identification of the model should be standard practice given the potential problems derived from lack of identification.

Several methods already exist to check identification in linearized models using: (i) the autocovariogram (Iskrev (2010), Andrle (2010));[15] (ii) the spectral density (Komunjer and Ng (2011) and Qu and Tkachenko (2012)); and (iii) Bayesian indicators (Koop et al. (2013)). For a review and methodological comparison of these techniques, the interested reader is referred to Mutschler (2014).

Variable Selection. Guerron-Quintana (2010) illustrates how the set of observables chosen for estimation affects the way in which the structural parameters enter into the log-likelihood function and, therefore, conditions the model estimation via likelihood-based methods. Our experiments show that the dangers that Guerron-Quintana (2010) warned us about in regards to estimation also play a role in Bayesian model comparison as differences in the set of observables can affect the differences in the log-likelihood functions across models that we can tease out from the data. Our simulations indicate that the selection of observables might help with model comparison in small samples, but it does not necessarily resolve the problem that arises when more parsimonious specifications are preferred over the more heavily parameterized ones that characterize the true DGP of the observed data.

All our experiments were conducted after estimating the competing model specifications on the same set of observables to maintain comparability. We suggest, however, that data not included in the set of observables can be used for cross-validation. For instance, we use output and inflation to estimate the four models under consideration in the experiments plotted in Figures 1 and 2. Trade data—while not directly used in the estimation—can serve for cross-validation of the model selection implied by Bayesian posterior odds given that preference for a closed-economy specification would be inconsistent with a non-zero trade series.

One could argue that model selection could be refined in the same way—e.g., Bayesian estimation and model comparison is not warranted with closed-economy models when there are open-economy alternatives if the data suggests non-zero trade (irrespective of whether we actually end up using the trade data for the estimation or not). In practice when none of the models available describes the exact DGP underlying the observed data unlike in our experiments. We, nonetheless, suggest that even in those circumstances performing Bayesian model comparison with different sets of observable variables can offer additional insights about the robustness of the evidence in favor of a given model against the alternatives.

[15] For the implementation of the local identification procedure of Iskrev (2010) adopted by the software package Dynare and their implementation of an optional Monte Carlo exploration of the state space of model parameters, see Ratto and Iskrev (2011).

Prior Selection. In this paper we assume 'non-informative' prior probabilities on the models, i.e. $\Pr(M_i) = \frac{1}{k}$ for all $i = 1, ..., k$, and we keep invariant the distribution of priors on parameters $f_1(\theta_i)$ across all model specifications. The set of structural parameters that characterizes each competing model is a subset of the set of parameters for the true model (the DGP), i.e. $\Theta_i \subseteq \Theta_1$ for all $i = \{2, 3, 4\}$. In fact, the set of parameters for each competing model M_i can be described simply as $\Theta_i = \left\{\theta_1 \in \Theta_1 : \theta_{1l} = 0 \text{ for some } l = 1, ..., \bar{l} \text{ where } 1 \leq \bar{l} < d_1\right\}$. Our experiments make the illustration simpler because the distributions of all competing models $f_i(z \mid \theta_i)$ for all $i = \{2, 3, 4\}$ are in effect limiting cases of the distribution of the true model $f_1(z \mid \theta_i)$. We then merely choose points on an interval that parameterize the true DGP (model M_1) *closer and closer* to the probability distribution of at least one of the alternative (more parsimonious) models to highlight the importance of the penalization for overfitting that arises even with moderate sample sizes.[16]

We ignore very short samples where the posterior distribution may still be dominated by the priors, and base our investigation on moderate sample sizes for which the Laplace approximation works reasonably well. We view this as most relevant for applied work, and do not explore the role of priors (and prior selection) further in our current analysis. We leave that for future research.

Nested versus Non-nested Models in Bayesian Model Comparison. When the distributions of the true model and a competing one become arbitrarily close to each other, for a given sample size the differences in the log-likelihood function ought to be smaller between the two models. Then, the penalty for overfitting ends up dominating our results and favoring the more parsimonious one over the more heavily-parameterized one (even if that is the true model). Bayesian model comparison through posterior model probabilities embodies a strong preference toward the lowest dimensional model (Occam's razor) and our experiments show that as a consequence we may fail to find support for the true (more complex) model in small samples in spite of the good asymptotic properties demonstrated in Fernández-Villaverde and Rubio-Ramírez (2004).

Our illustrations, however, are largely based on comparisons between nested model specifications. When competing models are non-nested and can be represented by probability distributions that do not overlap, then the posterior probability of the true model converges more quickly to the true one. This fact follows from standard asymptotic theory, as noted in Kass and Wasserman (1995). In those instances, we expect the severity of the *false signals* problem highlighted in this paper to be lessened. The simple logic behind this is that the more dimensions along which two models differ, the easier it becomes to find a way to tell them apart.

5 Concluding Remarks

In this paper we compare models with Bayesian posterior model probabilities working with a stylized specification of an open economy model that generates a short-run relationship between global slack and domestic inflation—the open-economy New Keynesian model of Martínez-García and Wynne (2010). Using a standard parameterization of the model, we generate artificial data which we then use to estimate four competing models (including the true model from which the data is simulated and three nested, simpler variants) with standard Bayesian techniques. We find that Bayesian model comparison based on posterior model prob-

[16] We use two different ways of accomplishing this because when we compare closed-economy versus open-economy models we do so by bringing the import share closer and closer to zero. In the case where we are comparing the NOEM model against the International Real Business Cycle model we do not alter the degree of price stickiness but in turn bring the two distributions closer together by choosing to implement an increasingly more aggressive monetary policy that is closer to the optimal policy.

abilities is sensitive to the choice of observables and to sample size. While asymptotically the posterior probability of the true model converges to one, we show that in small samples (of moderate length) the posterior model probabilities penalization for overfitting may lead us to favor a more parsimonious model instead.

It has been argued in the literature that when the evidence favors the more parsimonious model, the costs in terms of fit cannot be too large as the probability distribution of the preferred model and the true model must be close. We believe, though, that this has consequences that go beyond our ability to fit the data. Selecting the wrong model (model selection) or accounting for model uncertainty (through model averaging) on the basis of posterior model probabilities that seemingly support the wrong specification affects our ability to use these models for things that we care about such as policy analysis or forecasting.

That is particularly important, for example, when we think that Bayesian model comparison may have trouble to find support in the data for frictions in the goods market—nomimal rigidities—if monetary policy is near optimal, even when those frictions are a feature of the economy. This can affect how we evaluate the costs of alternative monetary policies or how we forecast the future path of standard aggregate macro variables as the trade-offs that policy-makers would face hinge on whether those frictions are present or not.

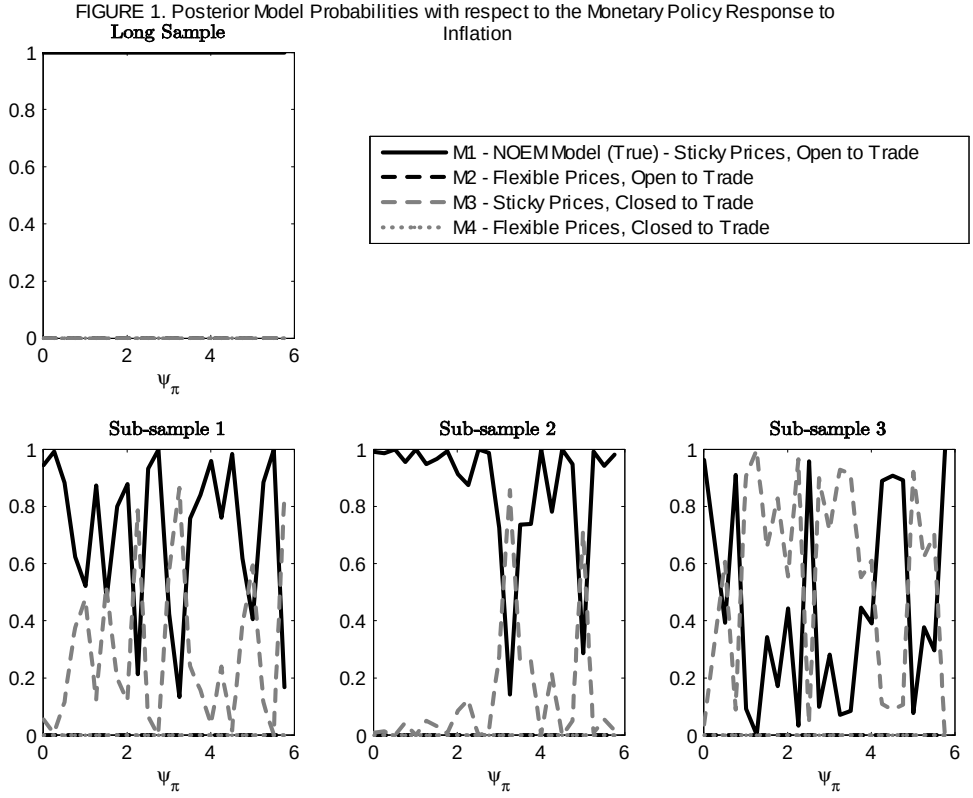
In our view, a strong preference for parsimonious models is not always and everywhere a desirable feature—even if they fit the data well. We see the primary contribution of our paper as illustrating how these 'wrong choices' can occur in small samples and why it matters. We caution that variable selection may not help us eliminate the problem of *false signals* in model selection with small samples. We leave it for future research to investigate the small sample properties of other criteria for model comparison.

Appendix of Tables and Figures

Table 1 - Prior Distributions				
Structural parameters	**Prior Density**	**Domain**	**Prior Mean**	**Prior Std. Dev.**
Non-policy parameters				
β	Fixed	–	0.99	–
φ	Gamma	$\mathbb{R}^+$	2	2
σ	Gamma	$\mathbb{R}^+$	1.5	1
ξ	Beta	$(0, 0.5)$	0.06, range: $(0, 0.5)$	0.01
α	Beta	$(0, 1)$	0.75, range: $(0, 1)$	0.07
Policy parameters				
ρ_i	Beta	$(0, 1)$	0.78	0.1
ψ_π	InvGamma	$\mathbb{R}^+$	0.33, range: $(0, 6)$	2
ψ_x	InvGamma	$\mathbb{R}^+$	1.29	2
Shock parameters				
δ_a	Beta	$(0, 1)$	0.95	0.05
σ_a	InvGamma	$\mathbb{R}^+$	0.7	2
ρ_{a,a^*}	Beta	$(0, 1)$	0.25	0.18
σ_m	InvGamma	$\mathbb{R}^+$	0.38	2
ρ_{m,m^*}	Beta	$(0, 1)$	0.5	0.22

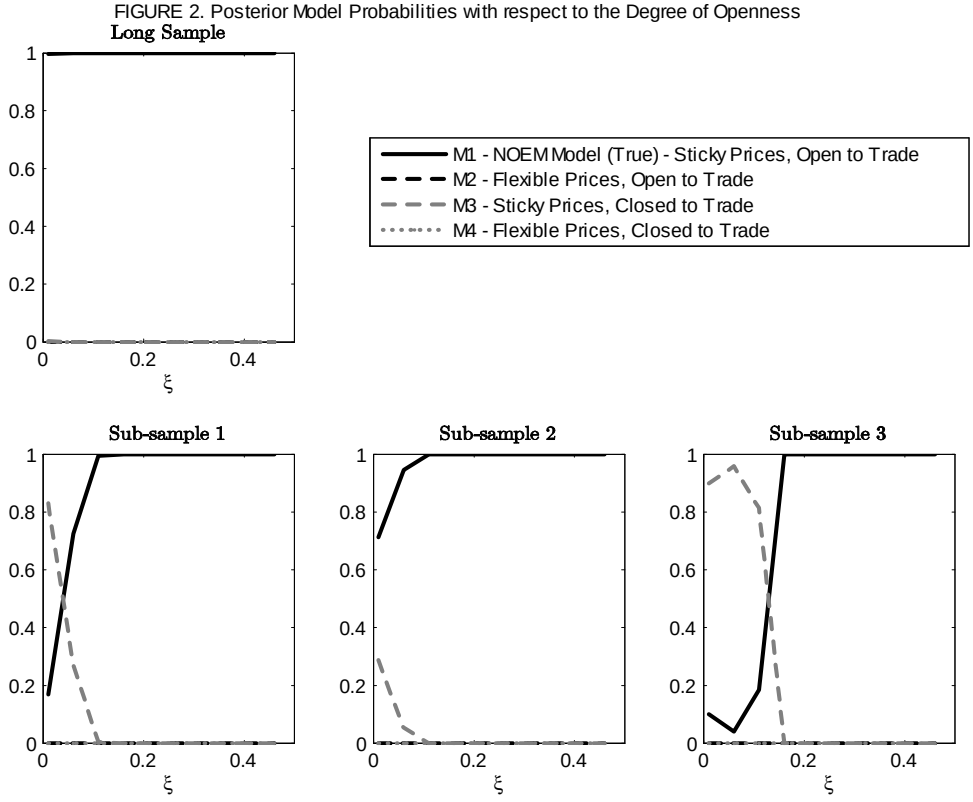
Note: This table reports only the prior mean and prior standard deviation for each model parameter. For any plausible choice of these two moments of the prior there is a mapping onto the prior distribution parameters v and s that matches both of them and fully characterizes the prior distribution itself. For the Normal distribution, the mean is μ=v and the variance is σ^2=s^2. For the Beta distribution, the mean is μ=$v/(v+s)$ and the variance is σ^2=$vs/((v+s)^2(v+s+1))$. For the Gamma distribution, the mean is μ=vs and the variance is σ^2=vs^2. For the Uniform distribution, the upper and lower bound of the support are v and s respectively, while the mean is μ=$(v+s)/2$ and the variance is σ^2=$(v-s)^2/12$. For the Inverse Gamma distribution, the mean is μ=s/(v-1) and the variance is σ^2=$s^2/((v-1)^2(v-2))$.

FIGURE 1. Posterior Model Probabilities with respect to the Monetary Policy Response to Inflation



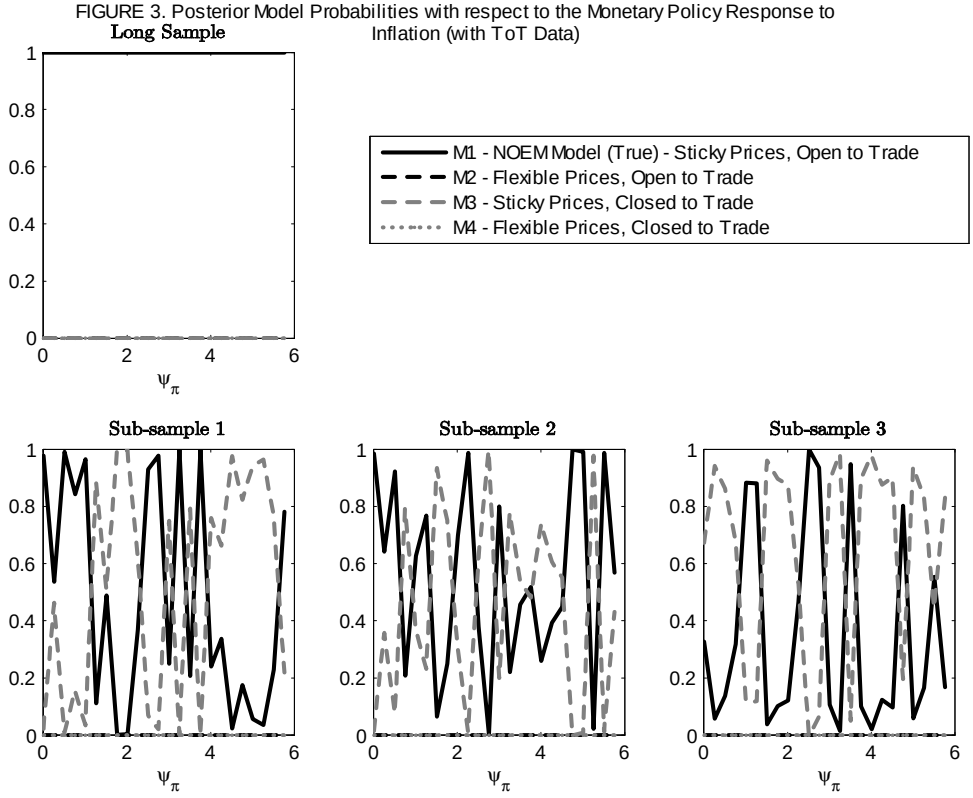
Note: The model is simulated over 10000 periods with code written for Dynare version 4.2.4 and Matlab version 7.13.0.564. The long sample refers to the 10000 observations while the three sub-samples are selected to cover the same three sub-periods including 160 observations each. The set of observables include Home and Foreign inflation, Home and Foreign Output. This figure plots the computed Bayesian posterior model probabilities for an interval over the parameter ψ_π. The code for the simulation is available upon request from the authors.

FIGURE 2. Posterior Model Probabilities with respect to the Degree of Openness



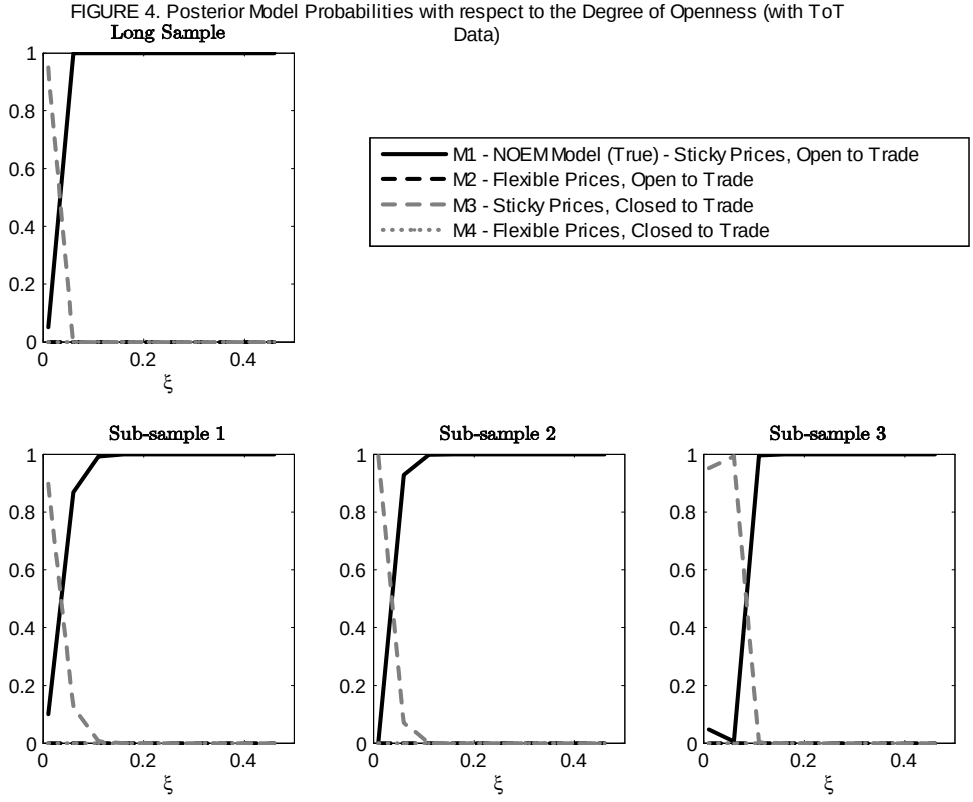
Note: The model is simulated over 10000 periods with code written for Dynare version 4.2.4 and Matlab version 7.13.0.564. The long sample refers to the 10000 observations while the three sub-samples are selected to cover the same three sub-periods including 160 observations each. The set of observables include Home and Foreign inflation, Home and Foreign Output. This figure plots the computed Bayesian posterior model probabilities for an interval over the parameter ξ. The code for the simulation is available upon request from the authors.

FIGURE 3. Posterior Model Probabilities with respect to the Monetary Policy Response to Inflation (with ToT Data)



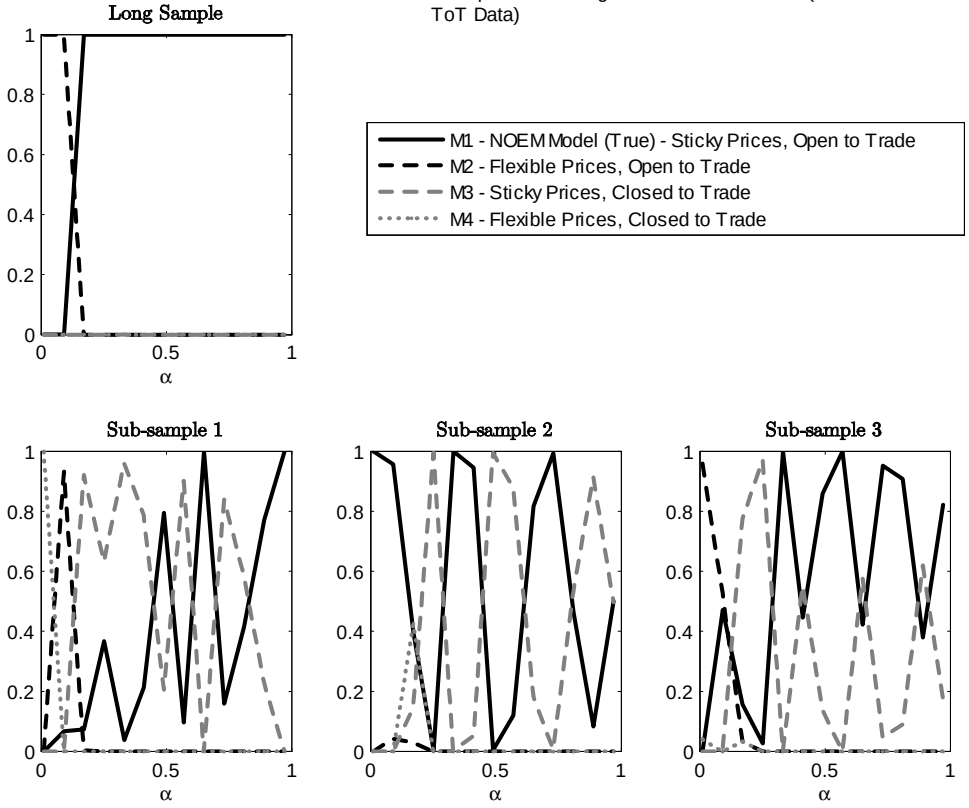
Note: The model is simulated over 10000 periods with code written for Dynare version 4.2.4 and Matlab version 7.13.0.564. The long sample refers to the 10000 observations while the three sub-samples are selected to cover the same three sub-periods including 160 observations each. The set of observables include Home and Foreign inflation, Home Output and terms of trade. This figure plots the computed Bayesian posterior model probabilities for an interval over the parameter ψ_π. The code for the simulation is available upon request from the authors.

FIGURE 4. Posterior Model Probabilities with respect to the Degree of Openness (with ToT Data)



Note: The model is simulated over 10000 periods with code written for Dynare version 4.2.4 and Matlab version 7.13.0.564. The long sample refers to the 10000 observations while the three sub-samples are selected to cover the same three sub-periods including 160 observations each. The set of observables include Home and Foreign inflation, Home Output and terms of trade. This figure plots the computed Bayesian posterior model probabilities for an interval over the parameter ξ. The code for the simulation is available upon request from the authors.

FIGURE 5. Posterior Model Probabilities with respect to the Degree of Price Stickiness (with ToT Data)



Note: The model is simulated over 10000 periods with code written for Dynare version 4.2.4 and Matlab version 7.13.0.564. The long sample refers to the 10000 observations while the three sub-samples are selected to cover the same three sub-periods including 160 observations each. The set of observables include Home and Foreign inflation, Home Output and terms of trade. This figure plots the computed Bayesian posterior model probabilities for an interval over the parameter α. The code for the simulation is available upon request from the authors.

References

Adjemian, S., H. Bastani, M. Juillard, F. Mihoubi, G. Perendia, M. Ratto, and S. Villemot (2011). *Dynare: Reference Manual, Version 4*.

Akaike, H. (1978). A New Look at the Bayes Procedure. *Biometrika 65*, 53–59.

An, S. and F. Schorfheide (2007). Bayesian Analysis of DSGE Models. *Econometric Reviews 26*(2-4), 113–172.

Andrle, M. (2010). A Note on Identification Patterns in DSGE Models. *ECB Working Paper Series no. 1235*.

Blanchard, O. J. and C. M. Kahn (1980). The Solution of Linear Difference Models Under Rational Expectations. *Econometrica 48*(5), 1305–13011.

Calvo, G. A. (1983). Staggered Prices in a Utility-Maximizing Framework. *Journal of Monetary Economics 12*(3), 383–398.

Canova, F. and L. Sala (2009). Back to Square One: Identification Issues in DSGE Models. *Journal of Monetary Economics 56*(4), 431–449.

Clarida, R., J. Galí, and M. Gertler (1999). The Science of Monetary Policy: A New Keynesian Perspective. *Journal of Economic Literature 37*(4), 1661–1707.

Clarida, R., J. Galí, and M. Gertler (2002). A Simple Framework for International Monetary Policy Analysis. *Journal of Monetary Economics 49*(5), 879–904.

Corduneanu, A. and C. Bishop (2001). Variational Bayesian Model Selection for Mixture Distributions. In T. Richardson and T. Jaakkola (Eds.), *Proceedings Eighth International Conference on Artificial Intelligence and Statistics*, pp. 27'34. Morgan Kaufmann Publishers Inc.

Fernández-Villaverde, J. and J. F. Rubio-Ramírez (2004). Comparing Dynamic Equilibrium Models to Data: A Bayesian Approach. *Journal of Econometrics 123*(1), 153–187.

Fernández-Villaverde, J., J. F. Rubio-Ramírez, T. Sargent, and M. Watson (2007). A, B, C, (and D)'s for Understanding VARs. *American Economic Review 97*, 1021–1026.

Geweke, J. (1989). Bayesian Inference in Econometric Models Using Monte Carlo Integration. *Econometrica 57*, 1317–1340.

Goodfriend, M. and R. G. King (1997). The New Neoclassical Synthesis and the Role of Monetary Policy. In *NBER Macroeconomics Annual*, pp. 231–283. NBER.

Guerron-Quintana, P. A. (2010). What You Match Does Matter: The Effects of Data on DSGE Estimation. *Journal of Applied Econometrics 25*(5), 774–804.

Hamilton, J. D. (1994). State-Space Models. In R. Engle and D. McFadden (Eds.), *Handbook of Econometrics*, Volume IV, Chapter 50, pp. 3039–3080. Elsevier Science B.V.

Hoeting, J. A., D. Madigan, A. E. Raftery, and C. T. Volinsky (1999). Bayesian Model Averaging: A Tutorial. *Statistical Science 14*(4), 382–417.

Hsiao, C. (1983). Identification. In Z. Griliches and M. D. Intriligator (Eds.), *Handbook of Econometrics*, Volume 1, Chapter 4. Amsterdam: North'Holland.

Iskrev, N. I. (2010). Local Identification in DSGE Models. *Journal of Monetary Economics 57*(2), 189–202.

Kass, R. E. and A. E. Raftery (1995). Bayes Factors. *Journal of the American Statistical Association 90*(430), 773–795.

Kass, R. E., L. Tierney, and J. B. Kadane (1988). Asymptotics in Bayesian Computation. In J. Bernardo, M. DeGroot, D. Lindley, and A. Smith (Eds.), *Bayesian Statistics 3*. Oxford University Press.

Kass, R. E., L. Tierney, and J. B. Kadane (1990). The Validity of Posterior Asymptotic Expansions Based on Laplace's Method. In S. Geisser, J. S. Hodges, S. J. Press, and A. Zellner (Eds.), *Bayesian and Likelihood Methods in Statistics and Econometrics*. New York: North-Holland.

Kass, R. E. and L. Wasserman (1995). A Reference Bayesian Test for Nested Hypotheses and its Relationship to the Schwarz Criterion. *Journal of the American Statistical Association 90*(431), 928–934.

Komunjer, I. and S. Ng (2011). Dynamic Identification of Dynamic Stochastic General Equilibrium Models. *Econometrica 79*(6), 1995–2032.

Koop, G., M. H. Pesaran, and R. P. Smith (2013). On Identification of Bayesian DSGE Models. *Journal of Business and Economic Statistics 31*(3), 300–314.

Martínez-García, E. and J. Søndergaard (2009). Investment and Trade Patterns in a Sticky-Price, Open-Economy Model. In G. Calcagnini and E. Saltari (Eds.), *The Economics of Imperfect Markets. The Effect of Market Imperfections on Economic Decision-Making*, Series: Contributions to Economics. Heidelberg: Springer (Physica-Verlag). December.

Martínez-García, E., D. Vilán, and M. A. Wynne (2012). Bayesian Estimation of NOEM Models: Identification and Inference in Small Samples. *Advances in Econometrics 28*, 137–199.

Martínez-García, E. and M. A. Wynne (2010). The Global Slack Hypothesis. Federal Reserve Bank of Dallas *Staff Papers, 10*. September.

Martínez-García, E. and M. A. Wynne (2014). Technical Note on 'Assessing Bayesian Model Comparison in Small Samples'. *Globalization and Monetary Policy Institute Working Paper no. 190*. August.

Minka, T. P. (2001). Expectation Propagation for Approximate Bayesian Inference. In *Proceedings of the Seventeenth Conference on Uncertainty in Artificial Intelligence*, pp. 362–369. Morgan Kaufmann Publishers Inc.

Mutschler, W. (2014). Identification of DSGE Models - A Comparison of Methods and the Effect of Second Order Approximation. *Mimeo, University of Münster*.

Neal, R. M. (2001). Annealed Importance Sampling. *Statistics and Computing 11*(2), 125–139.

Qian, G. and H. Künsch (1998). Some Notes on Rissanen's Stochastic Complexity. *IEEE Transactions on Information Theory 42*(2), 782–786.

Qu, Z. and D. Tkachenko (2012). Identification and Frequency Domain Quasi-Maximum Likelihood Estimation of Linearized Dynamic Stochastic General Equilibrium Models. *Quantitative Economics 3*(1), 95–132.

Ratto, M. and N. I. Iskrev (2011). Identification Analysis of DSGE Models with Dynare. *European Commission and Banco de Portugal*. https://www.ifk-cfs.de/fileadmin/downloads/events/conferences/monfispol2011/RATTO_IdentifFinal.pdf.

Rissanen, J. (1987). Stochastic Complexity (with Discussion). *Journal of the Royal Statistical Society, Series B 49*(3), 223–265.

Rissanen, J. (1996). Fisher Information and Stochastic Complexity. *IEEE Transactions on Information Theory 42*, 40–47.

Ríos-Rull, J.-V., F. Schorfheide, C. Fuentes-Albero, M. Kryshko, and R. Santaeulàlia-Llopis (2012). Methods versus Substance: Measuring the Effects of Technology Shocks. *Journal of Monetary Economics 59*(8), 826–846.

Schwarz, G. (1978). Estimating the Dimension of a Model. *The Annals of Statistics 6*(2), 461–464.

Slate, E. (1994). Parameterizations for Natural Exponential Families with Quadratic Variance Functions. *Journal of the American Statistical Association 89*, 1471–1482.

Taylor, J. B. (1993). Discretion versus Policy Rules in Practice. *Carnegie-Rochester Conference Series on Public Policy 39*, 195–214.

Verdinelli, I. and L. Wasserman (1995). Computing Bayes Factors Using a Generalization of the Savage-Dickey Density Ratio. *Journal of the American Statistical Association 90*(430), 614–618.

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

Zellner, A. (1971). *An Introduction to Bayesian Inference in Econometrics*. John Wiley & Sons.