A composite likelihood approach for dynamic structural models

Fabio Canova, BI Norwegian Business School, CAMP, and CEPR
Christian Matthes, Federal Reserve Bank of Richmond *

July 21, 2017

Abstract

We describe how to use the composite likelihood to ameliorate estimation, computational and inferential problems in misspecified dynamic stochastic general equilibrium models. We show how to perform quasi-Bayesian inference, how to construct density forecasts and scenario analyses, and discuss differences with finite mixture models and Bayesian model averaging. We present situations where the methodology has the potential to reduce model misspecification and provide an example to illustrate its properties in practice.

Key words: Dynamic structural models, composite likelihood, misspecification, Bayesian inference.

JEL Classification: C10, E27, E32.

*We thank Bruce Hansen, Mark Watson, Ulrich Mueller, Barbara Rossi, Refet Gurkenyak, Ferre de Grauve, Geert Dhaene and the participants at the 2016 ESEM invited session on 'New development in the analysis of Macroeconomic data', the 2016 ESOBE conference, the 3rd Henan University -Infer macroeconomic workshop, the SBIES conference, the Barcelona GSE Summer Forum and the EABCN conference on 'Time-varying Models for Monetary Policy and Financial Stability' as well as seminar participants at Singapore National University, Catholic University Leuven, and the Riksbank for comments and suggestions. Canova acknowledges the financial support from the Spanish Ministerio de Economia y Competitividad through the grants ECO2012-33247, ECO2015-68136-P and FEDER, UE. The views presented in this paper do not reflect those of the Federal Reserve Bank of Richmond, or the Federal Reserve system.

1 Introduction

In macroeconomics it is standard to construct dynamic stochastic general equilibrium (DSGE) models and analyze their properties. Until a decade ago, most analyses were performed using parameters formally or informally calibrated. Nowadays, it is more common to conduct inference using parameter estimated with classical or Bayesian full information likelihood methods, see Andreasen et al. (2014) for an exception.

Estimation of DSGE models is however difficult. There are population and sample identification problems, see e.g. Canova and Sala (2009), Komunjer and Ng (2011), Qu and Thachenko (2013); singularity problems (the number of shocks is generally smaller than number of endogenous variables), see e.g. Guerron Quintana (2010), Canova et al. (2014), Qu (2015); and informational deficiency problems (models are constructed to explain only a portion of the data), see Boivin and Giannoni (2006), Canova (2014), or Pagan (2016). In addition, there are computation and numerical difficulties, which become particularly acute when the model is of large scale or when the data is short or of poor quality. All these issues may make estimation results whimsical.

Inference in estimated DSGE models is also complicated. For standard frequentist asymptotic theory to apply, one needs regularity conditions, which are often violated in practice. For example, asymptotic approximations poorly characterize the actual distribution of estimates in relevant macroeconomic samples. Bayesian methods may help when the sample size is short, but it is tricky to specify joint priors when the parameter space is large and, as indicated by Del Negro and Schorfheide (2008), independence is suboptimal. Perhaps more importantly, while at the designing stage there is a general consensus that all models are misspecified in, at least, some dimensions, misspecification is forgotten at the estimation stage. However, standard likelihood-based inference is conditional on the estimated model being correctly specified.

Policymakers are keenly aware of this problem and of structural instabilities in the data generating process (DGP). For this reason, they are often interested in pooling results obtained from different models so as to robustify counterfactuals exercises and improve forecasting performance, see e.g. Aiolfi et al. (2010). However, rather than combining models' outputs, one may want to combine models prior to estimation so as to provide robust parameter estimates.

This paper proposes a *composite likelihood* approach to estimation and inference in DGSE models. We contribute to the literature from three different perspective. First, we show how to deal with DSGE model misspecification and provide a simple diagnostic to detect it and to compare misspecification in alternative models. While there are papers which deals with related issues (see e.g. Canova et al., 2016; or Inoue and Rossi, 2015), as far as we know, we are the first to explicitly deal with misspecification in DGSE models. Second, we describe how to construct and use the composite likelihood in the standard macroeconomics situations where all available models are misspecified descriptions of the DGP and how to conduct Bayesian inference in such an object. To the best of our knowledge, applications of the composite likelihood methodology to economic problems has been limited to Engle et al. (2008) and Qu (2015). Third, we

show how the approach can be used to potentially address the estimation and inferential problems noted in this introduction The approach is simple to implement, produces estimators with nice shrinkage properties and, in its Bayesian version, has a number of appealing features.

In the original formulation of Besag (1974) and Lindsay (1980), the composite likelihood is constructed combining marginal or conditional likelihoods of the true DGP. It is typically employed because the likelihood of the full model is computationally intractable or features complicated integrals, due to the presence of latent variables. A composite likelihood approach has been used to solve a number of complicated problems in fields as diverse as spatial statistics, multivariate extremes, psycometrics, genetics/genomics, see e.g. Varin et al. (2011).

In our setup, the composite likelihood combines the likelihood of distinct misspecified structural or underspecified statistical models. Because of this feature, standard results regarding the properties of composite likelihood estimators do not necessarily apply. We describe how to conduct asymptotic inference with such a composite likelihood and show how exact small sample statements about the common parameters of interest can be obtained in a Bayesian setting. We treat the composite likelihood as a quasi-likelihood function and, as in Kim (2002) or Chernozukov and Hong (2003) use priors to make the composite posterior a proper density function.

We show that posterior inference is easy to perform and describe a multiple-block MCMC algorithm that can be used to draw posterior sequences of the parameters. Because in our setup models are misspecified and the same time series can be repeatedly used in constructing the composite likelihood, the asymptotics of Bayesian MCMC estimators are non-standard and we present a simple variation of the MCMC algorithm that achieves the correct asymptotic coverage when all parameters are well identified. We describe how to conduct density forecasts, impulse responses and counterfactual exercises using the composite posterior and compare our approach with the one obtained with other pooling devices, such as mixture models or Bayesian model averaging (BMA).

We present examples indicating that the composite likelihood constructed using the information present in distinct models helps 1) to robustify parameter estimates and to reduce individual model misspecification, 2) to ameliorate population and sample identification problems, 3) to solve singularity problems, 4) to combine information coming from different sources, frequencies and levels of aggregation, and 5) to produce more stable estimates of the parameters of large scale structural models.

We apply the methodology to improve the quality of the inference a researcher is able to draw from the data about the slope of the Phillips curve when five different models are available and the slope of the Phillips curve is a crucial ingredient to determine the transmission of monetary policy disturbances.

The rest of the paper is organized as follows. The next section presents the traditional composite likelihood approach, introduces our setup, and compares our approach to a finite mixture model. Section 3 discusses Bayesian estimation and some non-standard features of our estimation problem. Section 4 deals with the problem

of constructing composite posterior statistics Section 5 has a number of examples highlighting how the methodology can address standard estimation problems. Section 6 shows how model misspecification can be reduced by the composite posterior and presents a misspecification diagnostic, which can be used in practical applications. Section 7 presents the results of estimating the slope of Phillips curve using a variety of structural models. Section 8 concludes. Five appendices provide details for arguments discussed in the text.

2 The composite likelihood

The original composite likelihood formulation has been suggested to deal with situations where the likelihood of the model is either difficult to construct, because of latent variables, or hard to manipulate, because the model is of large scale and the covariance matrix of the observables is nearly singular.

In these situations, it might be preferable to use a weighted average of marginal or conditional distributions of submodels (or 'events' in the nomenclature of the composite likelihood literature), see e.g. Lindsay (1980). This weighted average provides a limited information object that, under regularity conditions, allow consistent estimation of the parameters common to all submodels. In addition, as long as composite likelihood components are marginal or conditionals, the maximum composite likelihood estimator is asymptotically normal (e.g. Varin, et al., 2011). Since the components entering the composite likelihood are generally correlated, it is not necessarily true that adding components will improve the informational content of the composite likelihood and component-selection criteria have been designed to choose the optimal amount of "information" present in the composite likelihood (see e.g. Gao and Song, 2011).

In our setup, submodels are not necessarily marginal or conditionals versions of a known DGP. Still, it is possible to conduct frequentist inference with the composite likelihood; use a Bayesian framework to construct small sample distributions of the object of interests; and more importantly for our purposes, i) address standard DSGE estimation problems and ii) robustify inference.

2.1 The standard setup

Suppose a known DGP produces a density $F(y_t)$ for an $m \times 1$ vector of observables y_t. When y_t is of high dimensions or the DGP contains latent variables, it may be difficult to use $F(y_t, \psi)$, a parametric version of $F(y_t)$, where ψ is a $q \times 1$ vector, for estimation purposes. Let $\psi = [\theta, \eta]$, where by convention θ is the vector of parameters estimated by composite likelihood methods, and η is a vector of other parameters.

Let A(i), i=1, ...K be a set of marginal or conditional events of y(t) and let $f(y_{it} \in A_i, \theta, \eta_i)$ the subdensities of $F(y_t, \psi)$ corresponding to these events [1]. Each A_i

[1]Marginal or conditional integrate out all elements of y_t not in y_{it} or condition on some y_{jt} that are not in y_{it}. For ease of reading, the integrals and conditioning sets are left implicit.

defines a submodel, with implications for a subvector y_{it} of length T_i and is associated with a vector of parameters $\psi_i = [\theta, \eta_i]'$, where η_i are (nuisance) event specific. We represent the information generated by A_i with the tuple (y_{it}, T_i, ψ_i).

Given a vector of weights ω_i, the composite likelihood is

$$CL(\theta, \eta_1, \ldots, \eta_K, y_{1t}, \ldots, y_{KT}) = \Pi_{i=1}^{K} \ f(y_{it} \in A_i, \theta, \eta_i)^{\omega_i} \equiv \Pi_{i=1}^{K} \mathcal{L}(\theta, \eta_i | y_{it} \in A_i)^{\omega_i} \tag{1}$$

$CL(\theta, \eta_1, \ldots, \eta_K, y_{1t}, \ldots, y_{KT})$ is not a likelihood function. Nevertheless, if $y_{[1,t]} = (y_1, \ldots, y_t)$ is an independent sample from $F(y_t)$ and the weights ω_i are fixed, it can be shown that θ_{CL}, the maximum composite likelihood estimator, is consistent and

$$\sqrt{T}(\theta_{CL} - \theta) \xrightarrow{D} N(0, G^{-1}) \tag{2}$$

for T going to infinity and K fixed (see e.g. Varin, et al., 2011) where

$$G = HJ^{-1}H; \text{ Godambe information} \tag{3}$$

$$J \equiv var_\theta u(\theta, \eta_1, ..., \eta_K, y_{[1,t]}) \text{ Variability matrix} \tag{4}$$

$$H \equiv -E_\theta[\bigtriangledown_\theta u(\theta, \eta_1, \ldots, \eta_K, y_{[1,t]})] \text{ Sensitivity matrix} \tag{5}$$

$$u(\theta, \eta_1, \ldots, \eta_k, y | \omega_1, \ldots, \omega_K) = \sum_i \omega_i \bigtriangledown_\theta l_i(\theta, \eta_i, y_{[1,t]}) \text{ Composite scores} \tag{6}$$

$H \neq J$ and $\bigtriangledown_\theta l_i(\theta, \eta_i, y_{[1,t]})$ denotes the score associated with the log of $f(y_{it} \in A_i, \theta, \eta_i)$. If T is fixed, but $K \to \infty$, we need assumptions on how A_i are constructed. For example, if they are independent, then (2) still holds. Note that a standard Newey-West correction to $J(\theta)$ can be used if $y_{[1,t]}$ is not an independent sample.

Consistency obtains because each element in (2) is an unbiased estimating function. Since the ML estimator obtained with each component converges, as T increases, to the true parameter vector, $\theta_{CL} \xrightarrow{P} \theta$. Asymptotic normality holds because the sampling distribution of the maximum likelihood estimator of each submodel can be approximated quadratically around the mode. Note that θ_{CL} is not fully efficient - the Godambe matrix, $G(\theta)$ equals to the Fisher information matrix, $I(\theta)$, only if the composite likelihood is the likelihood of the true model. Carefully choosing ω_i may improve efficiency and optimal weights can be obtained by either minimizing the distance between $G(\theta)$ and $I(\theta)$ or by making sure that the composite likelihood ratio statistics has an asymptotic χ^2 distribution (see Pauli et al., 2011).

If consistency is all that one cares, one could set $\omega_i = \frac{1}{K}, \forall i$, or using a data-based approach, $\omega_i = \frac{\exp(\chi_i)}{1+\sum_{i=1}^{K-1}\exp(\chi_i)}$, where χ_i are a function of some statistics of past data $\gamma_i = f_i(Y_{1,[1:\tau]}, ..., Y_{K,[1:\tau]})$. If these statistics are updated over time, ω_i could also be made time varying. There a large forecasting literature (see e.g. Aiolfi et al., 2010) which can be used select training sample-based estimates of ω_i.

When K or the number of nuisance parameters η_i is large, joint estimation of $(\theta, \eta_1, \ldots, \eta_K)$ may be demanding and a two-step estimation approach is possible. For example, if ω_i is treated as fixed, η_i could be estimated separately from each $\log f(y_{it} \in$

A_i, θ, η_i) and plugged in the composite likelihood, which is then optimized with respect to θ, see e.g. Pakel et al. (2011). Consistency of θ_{CL} is unaffected as long as η_i are consistently estimated, but asymptotic standard errors need to be adjusted.

2.2 Our setup

Our setup differs from the traditional one in two important respects. First, $F(y_t)$ is treated as unknown. When dealing with economic data, there may be good reasons for making such an assumption. For example, we may not have enough information to construct F; we could write the VAR represention for y_t but not the structural model which generated it; or we do not have an analytic expression for $F(y_t)$ but only the first few terms of its Taylor expansion. Another reason for treating F as unknown is that the number of variables may be large and a researcher may have an idea of how portions of y_t could have been generated, but does not know yet how to link them in a coherent way.

Second, $f(y_{it} \in A_i, \theta, \eta_i)$ are neither marginal nor conditional densities. Instead, they represent misspecified approximations of the unknown DGP. To be concrete, in one leading example we have in mind, A_i are different structural models, e.g. a basic RBC model, a RBC model with financial frictions, a New Keynesian model with sticky price, a new Keynesian model with labor market frictions, etc., y_{it} is the data generated by these models, and $f(y_{it} \in A_i, \theta, \eta_i)$ the associated densities. Here, θ are the structural parameters common to all submodels, e.g. the risk aversion coefficient, or the labor share, while η_i could be structural parameters, with a DGP counterpart, e.g. a LTV ratio or a Calvo parameter, or reduced form parameters used to approximate features of the DGP. A variant of this example would be a situation where one considers one specific dynamic equilibrium model, but allows for various models of expectation formation such as rational expectations, least squares learning etc. The common equilibrium conditions can then be used with expectation formation mechanism to form various models In another leading example, $F(y_t)$ is a large scale structural model, for example, a multi-country model of trade interdependencies or a multi-country asset pricing model, and $f(y_{it} \in A_i, \theta, \eta_i)$ are structural models describing bilateral blocks or country specific portfolios. In a third case of interest $f(y_{it} \in A_i, \theta, \eta_i)$ are the densities generated by different approximate (perturbed or projected) solutions of a model or the densities of linear solutions, where only the k-th component of parameter vector is allowed to be time varying. Here A_i represent either the order of the approximation employed, or an indicator function describing which parameter is allowed to change.

In all three cases, different structural models are treated as misspecified subdensities because they disregard aspects of the DGP; take short cuts to modelling the complexities of the DGP; or condition on features which may be present or absent from the DGP. Formally, there is $(\theta, \eta_i) \in \Theta \times E_i$ such that $F(y_t) = f(y_{it}, \theta, \eta_i), \forall i$.

A final case of interest is one where $f(y_{it} \in A_i, \theta, \eta_i)$ represent different *statistical* models. We term models 'statistical' if they are obtained from the same theoretical model but feature different observables. For instance, a standard three-equations New-

Keynesian model could be estimated using inflation, the nominal interest rate, and a measure of output; or inflation, the nominal interest rate, and a measure of consumption - in the model, consumption and output are equal. By extension, $F(y_t)$ could be the density of an aggregate model and $f(y_{it} \in A_i, \theta, \eta_i)$ the densities obtained when i) data from cross sectional unit i is used; ii) data from a particular aggregation level (e.g. firm, industry, regional, etc.) is employed. Alternatively, $F(y_t)$ could be the density obtained using the full sample and $f(y_{it} \in A_i, \theta, \eta_i)$ the densities constructed using different subsamples (say, pre-WWI, interwar, post-WWI, etc.). As shown in section 5, the composite likelihood produces shrinkage estimators for θ in these situations, which combine unit specific and average information contained in the remaining models.

In all the examples, $f(y_{it} \in A_i, \theta, \eta_i)$ ignores the potential dependence across A_i. In addition, while in the traditional setup, submodels are compatible, in the sense that, asymptotically, $\theta_{i,ML}$ converge to the same true value, this need to be the case here. Finally, the elements of y_{it} need not be mutually exclusive across i; for example, the inflation rate could appear in each y_{it}.

Researchers working with DSGE models are generally free to choose what goes in θ and in η_i and even though some parameters might be common to all submodels, researchers might prefer not to estimate a common value. When using different statistical models, one could estimate an independent set of parameters for each A_i, or impose that some or all them are common. For example, when A_i represents different level of data aggregation, one could make the marginal propensity to consume common, while the parameters regulating the process for income may be submodel specific.

Because of misspecification, maximum likelihood estimators obtained from each $f(y_{it} \in A_i, \theta, \eta_i)$ will be inconsistent and, as a consequence, the composite likelihood estimator will be inconsistent. Following seminal work by White (1982) and Domowitz and White (1982), it is possible to show that, under regularity conditions, $\theta_{i,ML}$, the likelihood estimator of obtained in submodel A_i, converges, as $T \to \infty$ to the pseudo-parameter vector, θ_0, which minimizes the Kullback-Liebler (KL) distance from the true DGP. Moreover, $\sqrt{T}(\theta_{i,ML} - \theta_0) \sim N(0, G_i^{-1})$, where G_i is the Godambe information matrix for that submodel.

The weighting scheme that the composite likelihood employs, defines a density for a different misspecified model (the weighted average of the K submodels). As long as the weights w_i are constant, θ_{CL} approaches asymptotically the pseudo parameter value, say $\theta_{0,CL}$, minimizing the distance between the density of the combination of submodels and the DGP. $\theta_{0,CL}$ is not, in general, a weighted average of $\theta_{0,i}$, because submodels are not necessarily independent.

Mimicking the argument used for each submodel, one can show that $\sqrt{T}(\theta_{CL} - \theta_{0,CL}) \sim N(0, G^{-1})$ where G is the Godambe information computed using the composite likelihood. Thus, when misspecification is present the pivot of the asymptotic distribution is the minimizer of the KL distance, rather than the true parameter vector; the Godambe information matrix is evaluated at the minimizer of the KL distance, rather than the true parameter vector (see Appendix A for details).

2.3 A comparison with finite mixture models

The composite likelihood weights the likelihood of different submodels. An alternative way to pool the information contained in different submodels comes from ideas of Geweke and Amisano (2011), Billio et al. (2013), or Del Negro et al. (2016). The work of Waggoner and Zha (2012) is also relevant, if one thinks of A_i as switching specifications. In this literature, the relevant object is the likelihood of the mixture of the models which, for each t, is $L(\theta, \eta_1....\eta_k|y_{1t}...y_{Kt}) = \Sigma_{i=1}^{K}\omega_i L(\theta, \eta_i|y_{it})$ so that

$$\log L = \sum_{t=1}^{\tau} \log L(\theta, \eta_1....\eta_k|y_{1t}...y_{Kt}) \tag{7}$$

where $\tau = min(T_i)$. Simple manipulations reveal that (7) and the log of (1) differ by a Jensen's inequality term [2].

While a-priori both pooling devices are appealing, for the problems we are interested in, the composite likelihood has superior properties and added flexibility. From a computational point of view, when the decision rules have an autoregressive structure, estimators of θ have a closed form expression in the composite likelihood case, but not in the finite mixture case. In addition, in the finite mixture setup, it must be the case that $y_{it} = y_{jt}$ and $T_i = T_j$, since submodels represent alternatives that could have generated the observable data. These restrictions are unnecessary in the composite likelihood formulation and this gives flexibility to the methodology - see section 5 for examples.

If the weights in the composite likelihood approach are fixed, we can interpret the estimates for common parameters across models as being robust estimates in the sense that the parameters have to be consistent with the data for all models under consideration. To see this, note that in the extreme case where a parameter is inconsistent with the data for one model in the sense that the likelihood for that model is 0, the resulting log composite likelihood is minus infinity . This is not the case for finite mixture models.

3 Bayesian estimation

Because we are interested in the exact small sample distribution of θ, rather than its large sample approximation, we estimate it using Bayesian methods. Also, we treat ω_i

[2] When $K = 2$ and $T_1 = T_2 = 2$ the composite log-likelihood is

$$\log L = \omega(\log L_{11} + \log L_{12}) + (1-\omega)(\log L_{21} + \log L_{22})$$

while the log-likelihood in the mixture model is

$$\log L = \log(\omega L_{11} + (1-\omega)L_{21}) + \log(\omega L_{12} + (1-\omega)L_{22})$$

Suppose $\omega = 1-\omega$. Using $\log \sum_{t=1}^{T} x_t \equiv \log x_1 + \log(1 + \sum_{t=2}^{T} \frac{x_t}{x_1})$, one has $\log \sum_{t=1}^{2} x_t = \log x_1 + \log(1 + \frac{x_2}{x_1})$ and this differs from $\sum_{t=1}^{2} \log x_t = \log x_1 + \log x_2$, since $\log(1 + \frac{x_2}{x_1}) \approx \frac{x_2}{x_1}$ if $\frac{x_2}{x_1}$ is small.

as a random variable with a prior distribution (to be interpreted as the investigator prior assessment of the likelihood of submodel i) and jointly estimate the posterior of the common parameters θ, of the nuisance parameters ψ_i, and of the weights ω_i.

Given (y_{it}, T_i), we assume that $\sup_{\theta,\eta_i} f(y_{it} \in A_i, \theta, \eta_i) < b_i \leq B < \infty$, a condition generally satisfied for DSGE or VAR models, that $\mathcal{L}(\theta, \eta_i | y_{i,T_i})$ can be constructed for each A_i and that the composite likelihood $\mathcal{L}(\theta, \eta_1, ...\eta_K, \omega_1, \ldots, \omega_K | y_{1,T_i}, \ldots, y_{K,T_k})$ can be computed for $0 < \omega_i < 1$, $\sum_i \omega_i = 1$. For each A_i, the priors for the parameters are of the form:

$$p(\theta, \eta_i, \omega_i) = p(\omega_i)p(\theta)p(\eta_i|\theta) \tag{8}$$

We allow for an endogenous prior selection for η_i, in the spirit of Del Negro and Schorfheide (2008), which is advisable if the composite likelihood pools distinct structural models and we want to put these models on the same ground as far as matching certain statistics of the data. The composite posterior kernel is:

$$\begin{aligned} \check{p}(\theta, \eta_1, ...\eta_K, \omega_1, \ldots, \omega_K | Y_{1,t_1}, \ldots, Y_{k,T_k}) &= \\ \mathcal{L}(\theta, \eta_1 | Y_{1,T_1})^{\omega_1} p(\theta, \eta_1)^{\omega_1} p(\omega_i) \ldots \mathcal{L}(\theta, \eta_K | Y_{K,T_K})^{\omega_K} p(\theta, \eta_K)^{\omega_K} p(\omega_k) &= \\ \Pi_i \mathcal{L}(\theta, \eta_i | Y_{i,T_i})^{\omega_i} p(\eta_i)^{\omega_i} p(\theta) p(\omega_i) & \end{aligned} \tag{9}$$

which can be used to estimate the parameters, as described in Kim (2002) or Chernozukov and Hong (2003).

For computational and efficiency reasons, rather than using a brute force Metropolis approach to compute the posterior of θ, it may be advisable to employ a $2K + 1$ block Metropolis-within-Gibbs algorithm. Chib and Ramamurthy (2010) and Herbst and Schorfheide (2015) have also suggested drawing DSGE parameters in blocks. However, while they randomly split up the parameter vector in different blocks at each iteration, the blocks here are predetermined by the K submodels of interest.

In misspecified models estimated via Bayesian methods and when y_{it} are not mutually exclusive, MCMC-based estimates of posterior percentiles do not asymptotically approach frequentist percentiles, see Mueller (2013). After describing the algorithm, we will explain how to correct MCMC percentiles in order to have the required asymptotic coverage.

3.1 Estimation Algorithm

The algorithm consists of four steps:

1. Start with some $[\eta_1^0 \ldots \eta_K^0, \theta^0, \omega_1^0 \ldots \omega_K^0]$.
   For $iter = 1 : draws$ do steps 2-4
2. For $i = 1 : K$, draw η_i^* from a symmetric proposal P^{η_i}. Set $\eta^{iter} = \eta_i^*$ with probability

$$\min\left(1, \frac{\mathcal{L}(\left[\eta_i^*, \theta^{iter-1}\right] | Y_{i,T_i})^{\omega_i^{iter-1}} p(\eta_i^* | \theta^{iter-1})^{\omega_i^{iter-1}}}{\mathcal{L}(\left[\eta_i^{iter-1}, \theta^{iter-1}\right] | Y_{i,T_i})^{\omega_i^{iter-1}} p(\eta_i^{iter-1} | \theta^{iter-1})^{\omega_i^{iter-1}}}\right) \tag{10}$$

3. Draw θ^* from a symmetric proposal P^θ. Set $\theta^{iter} = \theta^*$ with probability

$$\min\left(1, \frac{\mathcal{L}(\left[\eta_1^{iter}, \theta^*\right]|Y_{1,T_1})^{\omega_1^{iter-1}} \dots \mathcal{L}(\left[\eta_K^{iter}\theta^*\right]|Y_{K,T_K})^{\omega_K^{iter-1}} p(\theta^*)}{\mathcal{L}(\left[\eta_1^{iter}, \theta^{iter-1}\right]|Y_{1,T_1})^{\omega_1^{iter-1}} \dots \mathcal{L}(\left[\eta_K^{iter}, \theta^{iter-1}\right]|Y_{K,T_K})^{\omega_K^{iter-1}} p(\theta^{iter-1})}\right) \tag{11}$$

4. For $i = 1 : K$ draw , draw ω_i^* from a symmetric proposal P^ω. Set $\omega^{iter} = \omega^* = (\omega_1^* ... \ \omega_k^*)$ with probability

$$\min\left(1, \frac{\mathcal{L}(\left[\eta_1^{iter}, \theta^{iter}\right]|Y_{1,T_1})^{\omega_1^*} \dots \mathcal{L}(\left[\eta_K^{iter}\theta^{iter}\right]|Y_{K,T_K})^{\omega_K^*} p(\omega^*)}{\mathcal{L}(\left[\eta_1^{iter}, \theta^{iter}\right]Y_{1,T_1})^{\omega_1^{iter-1}} \dots \mathcal{L}(\left[\eta_i^{iter}, \theta^{iter}\right]|Y_{K,T_K})^{\omega_K^{iter-1}} p(\omega^{iter-1})}\right) \tag{12}$$

Note that in (10) only the likelihood of model A_i matter because η_i only appears in that likelihood. A few interesting special cases are nested in the algorithm. For example, when the K submodels feature no nuisance parameters, as in the case when the composite likelihood is constructed using statistical models, steps 2.-3. can be combined in a single step. On the other hand, if ω_i's are treated as fixed, step 4 disappears. Notice also that when $\omega_i = 0, i \neq k$, $\omega_k = 1$, the algorithm collapses into a standard Block Gibbs-Metropolis MCMC. A standard random walk proposal for (θ, η_i) seems to work well in practice; a multivariate logistic proposal or an independent Dirichlet proposal (if only a few submodels are considered) are natural choices for ω_i.

3.2 Adjusting percentiles of the MCMC distribution

The estimation problem is non-standard since submodels are misspecified and y_{it} are not necessarily mutually exclusive across i and estimation may be performed repeatedly using the same time series in the composite likelihood conditioning set. For example, if all K submodels feature a nominal interest rate, the nominal rate series may be used K times. Naive implementation of the MCMC approach produces marginal posterior percentiles for θ which are too concentrated, because the composite likelihood treats y_{it} as if they were independent across i. As shown in appendix B, under regularity conditions and as $T \to \infty$, the posterior distribution will approach a normal distribution, but the asymptotic covariance matrix is the sensitivity matrix H, rather than the Godambe matrix. Hence, MCMC percentiles need to be adjusted to obtain asymptotic coverage which are consistent with the amount of information present in the data. To be clear: the adjustment is not designed to approximate the true (unknown) posterior but to insure that composite posterior inference is asymptotically appropriate.

We follow Ribatet et al. (2012) and modify the MCMC algorithm described in the previous subsection by adding two steps.[3] The first involves computing the "sandwich" matrix, $H(\theta)J(\theta)^{-1}H(\theta)$ where $H(\theta) = -E(\nabla_2 p_c(\theta|Y))$ and $J(\theta) = Var[\nabla p_c(\theta|Y)]$ via maximization of the composite posterior p_c. The second step involves adjusting the accepted draws using

$$\tilde{\theta}^j = \hat{\theta} + V^{-1}(\theta^j - \hat{\theta}) \tag{13}$$

[3]Qu (2015) uses a similar adjustment.

where $\hat{\theta}$ is the posterior mode, $V = C^T HC$ and $C = M^{-1}M_A$ is a semi-definite square matrix; $M_A^T M_A = HJ^{-1}H, M^T M = H$ and M_A and M are obtained via singular value decompositions [4].

The adjustment works well only when θ is well identified from the composite posterior and if the composite posterior has a unique maximum. As Canova and Sala (2009) have shown, such properties may not hold in a number of DSGE models. Because of this problem we recommend users to report both standard and adjusted percentiles.

3.3 A sequential learning interpretation

For the sake of illustration, suppose that ω_i are fixed. It is easy to give a sequential, adaptive learning interpretation to the composite posterior kernel (9) and to the Bayesian estimators we construct. When K=2, the composite posterior kernel $\check{p}$ is

$$\begin{aligned} \check{p}(\theta, \eta_{1....}\eta_2|Y_{1,T_1}, Y_{2,T_2}) \quad = \\ \mathcal{L}(Y_{1,T_1}|\theta,\eta_1)^{\omega_1} p(\eta_1|\theta)^{\omega_1} p(\eta_2|Y_{2,T_2},\theta)^{\omega_2} \{[p(\theta|Y_{2,T_2}) ML(Y_{2,T_2})]^{\omega_2} p(\theta)^{\omega_1}\} \end{aligned} \tag{14}$$

where $ML(Y_{2,T_2}) = \int \mathcal{L}(Y_{2,T_2}|\psi_2)p(\psi_2)$ is the marginal likelihood of model 2.

As (14) makes it clear, the posterior kernel can be obtained in two stages. In the first stage the prior for ψ_2 and the likelihood for submodel 2 are used to construct $\mathrm{p}(\theta|Y_{2,T_2})$. This conditional posterior, weighted by the marginal likelihood of the submodel 2, is geometrically combined with the prior $p(\theta)$ for the next estimation stage of θ. Suppose that $ML(Y_{2,T_2})$ is high, i.e. submodel 2 fits Y_{2,T_2} well. If $\omega_1 = \omega_2$, the prior for 1 will more heavily reflect the posterior of θ obtained from 2 relative to the initial prior $p(\theta)$. If $ML(Y_{2,T_2})$ is low, the posterior for θ obtained from submodel 2 has low weight relative to $p(\theta)$ when setting up the prior for submodel 1. In general, the prior for θ in each stage depends on the relative weights assigned to the current and to the previous submodels and on the fit of all previous submodels. Thus, a composite Bayesian approach to estimation can be interpreted as an adaptive sequential learning process, wherethe information contained in submodels whose density poorly explain observable data is downweighted.

Notice that even though only Y_{2,T_2} contains information for η_2, the posterior for this parameter may be sequentially updated, since the posterior for θ sequentially changes. However, since Y_{2,T_2} does not contain information for η_1, $p(\eta_1)$ will be left unchanged after estimation is performed with submodel 2.

Note that while here there is an automatic discounting whenever a submodel does not fit the data well regardless of whether ω_i is treated as a parameter or a random variable, a finite mixture has this property only if ω is random (see Del Negro et al., 2016).

---

[4] Rather than finding H and J once, prior running the MCMC algorithm, one could think of performing adaptive adjustment where $C(\theta|Y)$ is adjusted as $C(\theta^j|\theta^{j-1}, Y)$ (see Ribatet et al, 2012, p. 826). In this way MCMC draws are adaptively centered, which should make the draws more accurate and convergence faster, but a numerical optimization is needed at each step of the MCMC procedure.

4 Composite posterior statistics

It is relatively easy to use our framework to construct predictions of future observations. Let $\tilde{y}_{t+l}$ be future values of the variables appearing in all submodels.[5] Let $f(\tilde{y}_{t+l}|y_{it}, \theta, \eta_i)$ be the prediction of $\tilde{y}_{t+l}$ made by submodel A_i, given θ, η_i and let $f^{cl}(\tilde{y}_{t+l}|y_{1t}, \ldots, y_{kt}, \theta, \eta_1, \ldots, \eta_K, \omega_1, \ldots \omega_K) = \prod_{i=1}^{K} f(\tilde{y}_{t+l}|y_{it}, \theta, \eta_i)^{\omega_i}$. f^{cl} is a geometric pool (with weights ω_i) of K misspecified predictions of $\tilde{y}_{it+l}$. Then

$$\begin{aligned} p(\tilde{y}_{t+l}|y_{1t}, \ldots, y_{Kt}, \omega_1, \ldots \omega_K) \propto \int \ldots \int & f^{cl}(\tilde{y}_{t+l}|y_{1t}, \ldots, y_{Kt}, \theta, \eta_1, \ldots, \eta_K, \omega_1, \ldots, \omega_K) \\ & p(\theta, \eta_1, \ldots, \eta_K|\omega_1, \ldots, \omega_K, y_{1t}, \ldots, y_{Kt}) d\theta d\eta_1 \ldots d\eta_K \\ = & \int \ldots \int \prod_i p(\tilde{y}_{t+l}, \theta, \eta_i|y_{it})^{\omega_i} d\theta d\eta_1 \ldots d\eta_K \quad (15) \end{aligned}$$

is the composite predictive density of $\tilde{y}_{t+l}$, given the data and the weights, where $p(\tilde{y}_{t+l}, \theta, \eta_i|y_{it})^{\omega_i} \equiv (f(\tilde{y}_{t+l}|y_{it}, \theta, \eta_i) p(\theta, \eta_1, \ldots \eta_K|\omega, y_{1t}, \ldots, y_{Kt}))^{\omega_i}$.

As it is clear from (15), the composite predictive density differs from the true predictive density in two respects. First, the prediction function is based on the composite prediction pool density rather than the true prediction density. Second, the composite prediction pool density is integrated with respect to the composite posterior density of the parameters rather than their true posterior density.

To compute (15), one could use the mode or the posterior mean of the marginal of ω_i, depending on the loss function employed. As an alternative, one could integrate (15) with respect to the posterior of ω as well. Given that in many applications it makes sense to condition the prediction on estimates of the weights (which represents the posterior probability associated with each model), we believe (15) is more appealing for standard applications.

Since $f(\tilde{y}_{t+l}|y_{it}, \theta, \eta_i)$ is straightforward to compute (in our case the submodels will be linear Gaussian state space models), (15) can be approximated by first generating draws from the composite posterior, computing the predictive density for each draw in each A_i, multiplying the weighted predictions and, finally, averaging results across draws of $(\theta, \eta_1, \ldots, \eta_K)$ [6].

The problem of combining forecast densities is well studied in the literature (see e.g. Geweke and Amisano, 2011 or Del Negro et al., 2016). Two typical approaches are suggested: linear pooling, which lead to finite mixtures predictive densities such as BMA or static pools, and logarithmic pooling, which is what a composite predictive density approach produces. Logarithmic pooling gives predictive densities which are generally unimodal and less dispersed than linear pooling and have the property of being invariant to the arrival of new information (updating the components of the

[5] The variables of interest could be unobserved state variables, in which case $l = 0$ or even $l < 0$ could be considered.

[6] The prediction problem offers an alternative setup to choose ω_i if one treats them as fixed. In fact, one can choose $\omega_i, i = 1, \ldots, K$ to minimize the Kullback-Leibler (KL) risk measure between the composite and the true predictive density (see e.g. Vidoni, 2012).

composite likelihood commutes with the pooling operator). Relative to naive pooling of submodels' predictive densities, the composite predictive density uses the information in all submodels for estimation as summarized by the composite posterior.

Note that the logarithmic combination formula we present can be obtained as the solution to a well known constrained optimization problem in information theory (see Cover and Thomas, 2006). Appendix C provides the link between the two approaches.

In analogy with the prediction problem, one can compute statistics of interest by geometrically weighting the densities of outcomes obtained with each submodels and the composite posterior for the parameters. Take, for example, the computation of the responses of the subset of variables present in all submodels to a shock also present in all models.

Given $(\theta, \eta_i$, responses to shock j for submodel A_i can be computed setting all other structural shocks to 0 - which reasonable given that most of the model considered are linear. The density of outcome paths, computed randomizing (θ, eta_i) from their posterior, is the impulse response of interest.

The kernel of the composite posterior responses can then be computed analogously to (15), with the density of outcome paths replacing predictive densities. Again, the composite expression defines a logarithmic pool of responses for each value of $(\omega_1, \ldots, \omega_K)$ [7]

Counterfactuals can be similarly computed. Let $\bar{y}_{kt+l}$ be a selected path for the future in the k-th element of $\tilde{y}_{t+l}$. Using $f(\bar{y}_{kt+l}|y_{it}, \epsilon^j_{it+l}, \theta, \eta_i)$ for submodel A_i, one can find the path of ϵ^j_{it+l} consistent with the assumed $\bar{y}_{kt+l}$. With this path one can then compute $f(\bar{y}_{k't+l}|y_{it}, \epsilon^j_{it+l}, \theta, \eta_i)$, for $k' \neq k$. The composite counterfactual can be computed as in (15), geometrically weighting the resulting paths across models.

5 Using the composite likelihood to address estimation problems

This section provides examples showing how a composite likelihood approach can deal with standard problems encountered in the estimation of DSGE models. The first example discusses how to estimate parameters appearing in multiple misspecified models and shows how small sample identification problems can be resolved. The next example demonstrates how the approach can ameliorate population identification problems; the third example deals with singularity issues; the fourth with the problem of estimating

[7] In the case each submodel has a linear Gaussian state space format and fixed weights. the composite posterior of the impulse responses can be computed by first defining a fine grid of values for impulse responses (i.e. the difference in outcome paths). Then, for each horizon l, and each model A_i, one computes the model specific density of the difference in outcome paths, geometrically weight them across models and store this value. By repeating this exercise for a large number of parameter values drawn from the composite posterior (which can be approximated with the Metropolis-within-Gibbs algorithm of section 3) and computing the average value across parameter draws, we obtain a Monte Carlo approximation to the composite posterior density of the impulse responses.

the parameters of a large dimensional model. The last example shows how a composite likelihood approach may help to deal with short samples of data.

5.1 Estimating structural parameters appearing in multiple misspecified models

Suppose we have two structural misspecified models (A, B), with parameters $\psi_A = (\theta, \eta_A), \psi_B = (\theta, \eta_B)$ and implications for (y_{At}, y_{Bt}), two different subvectors of y_t. Assume that $f(y_{At}, \theta, \eta_A)$ and $f(y_{tB}, \theta, \eta_B)$ are produced by the decision rules:

$$y_{At} = \rho_A y_{At-1} + \sigma_A e_t \tag{16}$$

$$y_{Bt} = \rho_B y_{Bt-1} + \sigma_B u_t \tag{17}$$

where e_t and u_t are iid(0,I). For the sake of illustration, suppose that $\rho_B = \delta\rho_A, \sigma_B = \gamma\sigma_A$, that y_{At} and y_{Bt} are scalars, and that we have T_A observations on y_{At} and T_B observations on $y_{Bt}, T_B \leq T_A$. Thus $\theta = (\rho_A, \sigma_A), \eta_B = (\delta, \gamma), \eta_A = \emptyset$.

The (normal) log-likelihood functions of each model are:

$$\log L_A \propto -T_A \log \sigma_A - \frac{1}{2\sigma_A^2} \sum_{t=1}^{T_A} (y_{At} - \rho_A y_{At-1})^2 \tag{18}$$

$$\log L_B \propto -T_B \log \sigma_B - \frac{1}{2\sigma_B^2} \sum_{t=1}^{T_B} (y_{Bt} - \rho_B y_{Bt-1})^2 \tag{19}$$

and for $0 < \omega 1$, the log composite likelihood is

$$\log CL = \omega \log L_A + (1-\omega) \log L_B \tag{20}$$

Maximization of (20) with respect to θ leads to:

$$\rho_A = (\sum_{t=1}^{T_A} y_{At-1}^2 + \zeta_2 \sum_{t=1}^{T_B} y_{Bt-1}^2)^{-1} (\sum_{t=1}^{T_A} y_{At} y_{At-1} + \zeta_1 \sum_{t=1}^{T_B} y_{Bt} y_{Bt-1}) \tag{21}$$

where $\zeta_1 = \frac{1-\omega}{\omega} \frac{\delta}{\gamma^2}$, $\zeta_2 = \frac{1-\omega}{\omega} \frac{\delta^2}{\gamma^2} = \zeta_1 \delta$ and

$$\sigma_A^2 = \frac{1}{\xi} (\sum_{t=1}^{T_A} (y_{At} - \rho_A y_{At-1})^2 + \frac{1-\omega}{\omega\gamma^2} \sum_{t=1}^{T_B} (y_{Bt} - \delta\rho_A y_{Bt-1})^2) \tag{22}$$

where $\xi = (T_A + T_B \frac{1-\omega}{\omega\gamma^2})^{-1}$ is the effective sample size. The estimators of ρ_A and of σ_A^2 obtained using just model A or model B log-likelihoods are

$$\rho_{AA} = (\sum_{t=1}^{T_A} y_{At-1}^2)^{-1} (\sum_{t=1}^{T_A} y_{At} y_{At-1}); \quad \rho_{AB} = \delta^{-1} (\sum_{t=1}^{T_B} y_{Bt-1}^2)^{-1} (\sum_{t=1}^{T_B} y_{Bt} y_{Bt-1}) \tag{23}$$

and

$$\sigma_{AA}^2 = \frac{1}{T_A}\sum_{t=1}^{T_A}(y_{At} - \rho_{AA}y_{At-1})^2; \quad \sigma_{AB}^2 = \frac{1}{T_B}\sum_{t=1}^{T_B}(y_{Bt} - \delta\rho_{AB}y_{Bt-1})^2 \tag{24}$$

As (21)-(22) clearly show the estimator for θ obtained from a composite likelihood approach combine the information coming from y_{At} and y_{Bt}, with model B playing the role of a prior for model A. The formulas in (21) and (22) are also similar to those i) obtained in least square problems with uncertain linear restrictions (Canova, 2007, Ch.10), ii) derived using a prior-likelihood approach, see e.g. Lee and Griffith (1979) or Edwards (1969) and iii) implicitly produced by a DSGE-VAR setup (see Del Negro and Schorfheide, 2004), where T_B are the additional observations added to the original T_A data points. Note that (γ, δ) are estimated using only the information in y_{tB}.

In general, when the decision rules feature an autoregressive structure, the composite likelihood shrinks the information contained in y_{At} and the amount of shrinkage depends, among other things, on the informational content of y_{tB} for θ, as measured by the magnitude of (γ, δ, ω). The higher is ω the less important is the information present in y_{tB}; similarly, the larger is γ, the larger is the variance of the shocks in the decision rules of model B and the lower the information content of y_{Bt}. Conversely, the smaller is δ, the lower will be the shrinkage toward the information contained in model B. Thus, when estimating common parameters, the composite likelihood gives larger importance to data assumed to be generated by a model with higher persistence and lower standard deviation and which is a-priori more likely. The reason is straightforward: higher serial correlation implies important low frequency information; lower standard deviation implies lower noise.

When an array of models are available, estimates of θ will be constrained by the structure present in all models. For example, equation (21) now becomes

$$\rho_A = (\sum_{t=1}^{T_A} y_{At-1}^2 + \sum_{i=1}^{K-1}\zeta_{i2}\sum_{t=1}^{T_i} y_{it-1}^2)^{-1}(\sum_{t=1}^{T_A} y_{At}y_{At-1} + \sum_{i=1}^{K-1}\zeta_{i1}\sum_{t=1}^{T_i} y_{it}y_{it-1}) \tag{25}$$

where $\zeta_{i1} = \frac{\omega_i}{\omega_A}\frac{\delta_i}{\gamma_i^2}$, $\zeta_{i2} = \zeta_{i1}\delta_i$. Thus, the composite likelihood robustifies inference, in the sense that estimates of θ are shrunk to be consistent with the data generated by all available models.

Two further aspects of this example are worth some discussion. First, as mentioned, y_{At} and y_{Bt} may be different series. Thus, the procedure can be used to estimate parameters appearing in models featuring different observables. Alternatively, y_{At} and y_{Bt} may represent the same series with different levels of aggregation (say, aggregate vs. individual consumption). In general, y_{At} and y_{Bt} may have common components and some model specific ones. Second, T_A and T_B may be of different length. Hence, the procedure can be used to combine data of various length or the information in models setup up at different frequencies (e.g., a quarterly and an annual model). T_A and T_B may also represent samples for the same vector of variables coming from different time periods, for example, the real rate before and after a financial crisis. Baumeister and

Hamilton (2015) suggested a procedure to downweigh the information contained in earlier subsamples that mimics a composite likelihood estimator in this situation.

The example also allows us to discuss how the approach may help to reduce small sample identification problems. Suppose the density of y_A is well behaved, but because T_A is short, (18) is flat, making θ may be poorly identified using y_{At}. It is easy to show θ could become better identified if (y_{At}, y_{Bt}) are jointly used in estimation. This is because the curvature of the composite likelihood depends on the effective sample size ξ, which is a function of T_A and $T_B\frac{1-\omega}{\omega\gamma^2}$. Thus, for example, if γ is small, that is y_{tB} is less volatile than y_{tA}, or if ω is large, that is the degree of a-priori trust a researcher has in model B is high enough, $\xi >> T_A$ and the log composite likelihood (20) will be more peaked around the mode than the log-likelihood (18).

5.2 Ameliorating population identification problems

The previous subsection suggested that the composite likelihood may improve parameter identification when a short sample is available. This subsection discusses an example where some parameters are underidentified and others weakly identified *in population* and shows how a composite likelihood approach can remedy these problems.

Consider a canonical three-equations New Keynesian model (call it model A)

$$\begin{aligned} R_{At} &= \tau E_t\pi_{At+1} + e_{1t} && (26)\\ y_{At} &= \delta E_t y_{At+1} - \sigma(R_{At} - E_t\pi_{At+1}) + e_{2t} && (27)\\ \pi_{At} &= \beta E_t\pi_{At+1} + \gamma y_{At} + e_{3t} && (28) \end{aligned}$$

where R_{At} is the nominal rate, y_{At} the output gap and π_{At} the inflation rate;(e_{1t}, e_{2t}, e_{3t}) are mutually uncorrelated structural disturbances, $(\tau, \delta, \sigma, \beta, \gamma)$ are structural parameters, and E_t is the conditional expectations operator. The solution of the model is

$$\begin{bmatrix} R_{At} \\ y_{At} \\ \pi_{At} \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 \\ \sigma & 1 & 0 \\ \sigma\gamma & \sigma & 1 \end{bmatrix} \begin{bmatrix} e_{1t} \\ e_{2t} \\ e_{3t} \end{bmatrix} \equiv Ae_t \qquad (29)$$

Clearly, β is underidentified (it disappears from (29)) and the slope of the Phillips curve γ may not be well identified from the likelihood of $(R_{At}, y_{AT}, \pi_{At})$ if σ is small. In fact, large variations in γ may induce small variations in the decision rules (29) if σ is sufficiently small, making the likelihood flat in the γ dimension.

Suppose we have available another model (call it model B), known to be more misspecified relative to the New Keynesian model - we acknowledge this by selecting $\omega >> 1-\omega$. For example, consider a single equation Phillips curve with exogenous marginal costs:

$$\begin{aligned} \pi_{Bt} &= \beta E_t\pi_{Bt+1} + \gamma y_{Bt} + u_{2t} && (30)\\ y_{Bt} &= \rho y_{Bt-1} + u_{1t} && (31) \end{aligned}$$

where $\beta < 1$ and $\rho > 0$ measures the persistence of marginal costs. By repeatedly substituting forward and letting $x_t \equiv(1-\rho\ell)y_{Bt}$, $w_t \equiv(1-\rho\ell)\pi_{Bt}$, where ℓ is the lag operation, the solution is

$$\begin{bmatrix} x_t \\ w_t \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ \frac{\gamma}{1-\beta\rho} & 1-\rho\ell \end{bmatrix} \begin{bmatrix} u_{1t} \\ u_{2t} \end{bmatrix} \tag{32}$$

Because the log-likelihood of model B has information about β, one would be able to identify (and estimate) it from the composite likelihood. In addition, in model B the curvature of the likelihood in the γ dimension depends on $\frac{1}{1-\beta\rho}$ which, in general, is greater than one for $\beta \neq 0$. Hence, small variations γ may lead to sufficiently large variations in the decision rule (32) and thus in the composite likelihood, even when $1-\omega$ is small. Thus, shrinking model A toward model B which is more misspecified but has sharper information about the parameters of interest may help in terms of identification and estimation. Note that the argument is independent of the size of effective sample size ξ: since the identification problems we discuss occur in population, having a large or a small ξ is irrelevant. It should also be emphasized that the above argument implicitly assumes that the variances of $(\mathrm{e}_{2t}, e_{3t})$ and of (u_{1t}, u_{2t}) are of the same order of magnitude.

5.3 Solving singularity problems

DSGE models are typically singular. That is, since they generally feature more endogenous variables than shocks, the theoretical covariance matrix of the observables is singular and the likelihood function can not be constructed and optimized. To solve this problem the literature has suggested to i) select a subvector of the observables matching the dimension of the shock vector either informally (see Guerron Quintana, 2010) or formally (see Canova et al., 2014) and use the log-likelihood of this subvector for estimation; ii) add measurement errors to some or all the observables - so as to make the number of shocks (structural and measurement) equal or larger to the number observables; iii) increase the number of structural shocks, for example, by transforming parameters into disturbances (the discount factor becomes a preference shock, etc.). An alternative is to construct a composite likelihood using non-singular submodels, see also Qu (2015). To illustrate the approach, we use a stylized asset pricing example.

Suppose that the dividend process is $d_t = e_t - \alpha e_{t-1}$, where $e_t \sim iid(0, \sigma^2)$, and $\alpha < 1$, and that stock prices are the discounted infinite sum of future dividends. The solution for stock prices is $p_t = (1-\beta\alpha)e_t - \alpha e_{t-1}$, where $\beta < 1$ is the discount factor. Since e_t drives both dividends and stock prices, the covariance matrix of (d_t, p_t) is singular. Thus, one has to decide whether d_t or p_t should ve used to construct the likelihood and to estimate the common parameters (α, σ^2). Here adding a measurement error is difficult to justify, since neither dividends nor stock prices are subject to revisions, and making β a random variable is unappealing because the density of p_t becomes non-normal. When the composite likelihood is employed, the joint information in (d_t, p_t) is used to identify and estimate (α, σ^2) and β, if it is of interest. Optimization makes

stock prices contain different information relative to dividends; the composite likelihood provides a flexible tool to combine all available information to estimate parameters.

Following Hamilton (1994, p. 129), the likelihood functions of d_t and p_t are

$$\log L(\alpha, \sigma^2|\tilde{d}_t) = -0.5T\log(2\pi) - 0.5\sum_{t=1}^{T}\log\varsigma_t - 0.5\sum_{t=1}^{T}\frac{\tilde{d}_t^2}{\varsigma_t} \tag{33}$$

where $\tilde{d}_t$ and ς_t can be recursively computed as:

$$\begin{aligned}\tilde{d}_t &= d_t - \alpha\frac{1+\alpha^2+\alpha^4+\ldots+\alpha^{2(t-2)}}{1+\alpha^2+\alpha^4+\ldots+\alpha^{2(t-1)}}\tilde{d}_{t-1} &(34)\\ \varsigma_t &= \sigma^2\frac{1+\alpha^2+\alpha^4+\ldots+\alpha^{2t}}{1+\alpha^2+\alpha^4+\ldots+\alpha^{2(t-1)}} &(35)\end{aligned}$$

and

$$\log L(\beta, \alpha, \sigma^2|\tilde{p}_t) = -0.5T\log(2\pi) - 0.5\sum_{t=1}^{T}\log\upsilon_t - 0.5\sum_{t=1}^{T}\frac{\tilde{p}_t^2}{\upsilon_t} \tag{36}$$

where $\tilde{p}_t$ and υ_t can be recursively computed as:

$$\begin{aligned}\tilde{p}_t &= p_t - \frac{\gamma^2}{\zeta}\frac{1+\gamma^2+\gamma^4+\ldots+\gamma^{2(t-2)}}{1+\gamma^2+\gamma^4+\ldots+\gamma^{2(t-1)}}\tilde{p}_{t-1} &(37)\\ \upsilon_t &= \sigma^2(1-\beta\zeta)^2\frac{1+\gamma^2+\gamma^4+\ldots+\gamma^{2t}}{1+\gamma^2+\gamma^4+\ldots+\gamma^{2(t-1)}} &(38)\end{aligned}$$

and $\gamma^2 = \frac{\alpha^2}{(1-\beta\alpha)^2}$. For illustration, set $\sigma^2 = 1$ and focus attention on α. The first order conditions that a maximum likelihood estimator solves are

$$\begin{aligned}\frac{\partial\log L(\tilde{d}_t)}{\partial\alpha} &= -0.5\sum_t\frac{\partial\log\varsigma_t}{\partial\alpha} - 0.5\sum_t\frac{\partial\log\tilde{d}_t}{\partial\alpha}\frac{1}{\varsigma_t} + 0.5\sum_t\frac{\partial\log\varsigma_t}{\partial\alpha}\frac{\tilde{d}_t}{\varsigma_t^2} &(39)\\ \frac{\partial\log L(\tilde{p}_t)}{\partial\gamma} &= -0.5\sum_t\frac{\partial\log\upsilon_t}{\partial\alpha} - 0.5\sum_t\frac{\partial\log\tilde{p}_t}{\partial\alpha}\frac{1}{\upsilon_t} + 0.5\sum_t\frac{\partial\log\upsilon_t}{\partial\alpha}\frac{\tilde{p}_t}{\upsilon_t^2} &(40)\end{aligned}$$

For a given ω assigned to $\tilde{d}_t$, the composite likelihood is a weighted sum of (39)) and (40). While there are no closed expressions for either the maximum likelihood or the maximum composite likelihood estimators of α, which would allow a direct comparison of their properties, we can still infer what ((39) and (40) employ to estimate α and what a composite likelihood does using simple algebra. Appendix D shows that α will be identified and estimated more from the serial correlation properties of the data if $\tilde{p}_t$ is used to construct the likelihood function and more from the variance properties of the data if $\tilde{d}_t$ is used to construct the likelihood function. Hence, estimates obtained from (39) and (40) will generally be different.

The composite likelihood provides a compromise between these two types of information. Depending on the value of ω, either the serial correlation or the variance

properties of (d_t, p_t) or both will be employed. Clearly, if the low frequency components of $\tilde{p}_t$ are poorly characterized because, for example, the sample is short or because α is close to zero, the composite likelihood provides a better objective function to identify and estimate the parameters than each of the individual likelihood functions. In addition, if d_t is smooth and p_t is highly volatile, the composite likelihood will give a more stable estimate of α than the individual likelihood functions.

5.4 Dealing with large scale structural models

This section analyzes the situation when one needs to estimate the parameters of a large scale model but computation or numerically problems makes the task difficult. Suppose the decision rules of a model are $y_t = A(\theta)y_{t-1} + e_t$, where e_t iid N(0,$\Sigma(\theta)$), θ is a vector of structural parameters, y_t is of large dimension and dim(y_t) = dim (e_t).

The likelihood function is

$$L(\theta|y) = (2\pi)^{-T/2}|\Sigma|^{T/2}\exp\{(y_t - A(\theta)y_{t-1})\Sigma(\theta)^{-1}(y_t - A(\theta)y_{t-1})'\} \tag{41}$$

If $dim(y_t)$ is large, computation of $\Sigma(\theta)^{-1}$ may be demanding. Furthermore, numerical difficulties may emerge if some of the variables in y_t are collinear or if there are near singularities in the model (for example, if we have a long term and a short term interest rate). Furthermore, if $y_t = (y_{1t}, y_{2t})$, where y_{2t} are non-observables,

$$L(\theta|y_1) = \int L(y_1|y_2, A(\theta), \Sigma(\theta))g(\theta|y_2)dy_2 \tag{42}$$

where $g(\theta|y_2)$ is the posterior computed using y_2, may be intractable.

Rather than computing (41) or (42), we can take a limited information point of view and produce estimates the parameters using objects which are simpler to construct. If we partition $y_t = (y_{1t}, y_{2t}, \ldots y_{Kt})$, two such objects are:

$$CL_1(\theta|y_t) = \sum_{i=1}^{K} \omega_i \log L(\theta|y_{it}) \tag{43}$$

$$CL_2(\theta|y_{it}) = \sum_{i=1}^{K} \omega_i \log L(\theta|y_{it}, y_{-it}) \tag{44}$$

where y_{-it} indicates any combination of the vector y_t, which excludes the i-th combination.

CL_1 is obtained neglecting the correlation structure between y_{it}. Thus, the blocks are treated as providing independent information, even though this is not necessarily true. For example, in a multi-country symmetric model, y_{it} could correspond to the observables of country i; in a closed economy model, they could correspond to different sectors of the economy. CL_2 is obtained by conditionally blocking groups of variables. In the multi-country example, we can construct the likelihood of each country variables y_{it}, given the vector of the variables of all other countries y_{-it}. Which composite likelihood one would use depends on the problem and the tractability of conditional vs. marginal likelihoods.

5.5 Dealing with short T when a panel is available.

The setup we consider can easily account for the situation where there is a single structural model, for example, an asset pricing equation, or a consumption function equation and the observable data comes from different units (consumers, portfolios, countries), as discussed by Pakel et al. (2011) or from different level of aggregation (firm, industry, sector). Here $\hat{y}_{1t}, \hat{y}_{2t}, ...\hat{y}_{Kt}$ represent the same observables obtained from unit (level of aggregation) i=1,2...K and the composite log-likelihood is

$$CL(\theta|\hat{y}_{1t}, \hat{y}_{2t}, ...\hat{y}_{Kt}) = \sum_{i=1}^{K} \omega_i \log L(\theta|\hat{y}_{it}) \tag{45}$$

(45) neglects the correlation structure across units, but pools information about the common parameters from the available cross section. Given a linear autoregressive structure for the decision rules for each i, the pooling procedure produces estimators for θ which are similar to those derived by Zellner and Hong (1989): they combine individual and a (weighted) average of the information present in the cross section of data. This is clear when looking at (25), once it is realized that $\sum_{i=2}^{K} \zeta_{i2} \sum_{t=1}^{T_i} y_{it-1}^2$ represents a weighted average of the information present in the data of the units other than the first one. Such a pooling approach is likely to be preferable when each $\hat{y}_{it}$ is short. Note that when the composite likelihood is used the cross sectional information is not exactly pooled: the degree of shrinkage depends on the precision of various sources of information. Thus, the composite likelihood uses at least as much information as the individual likelihoods; stochastically exploits commonalities in cross section if they exist; and may lead to improved estimates of the vector of common parameters θ.

6 Reducing misspecification

The shrinkage estimators that the composite likelihood (posterior) generates may help to reduce biases in likelihood (posterior) estimates obtained with misspecified models. The reason is simple: when a model is misspecified, information contained in additional (misspecified) models restrict the range of values that common parameters can take. Thus, the quality of the estimates may improve, both in terms of location and dispersion. To make this intuition operative, we need a diagnostic that compares the misspecification present in the composite likelihood and in each submodel.

As argued in White (1982), White and Domowitz (1982), if the model is correctly specified, the information matrix can be expressed either via the Hessian $H(\theta)$ or via the outer product matrix $J(\theta)$. Thus, to examine if a model is misspecified, one can test $D(\theta_{ML}) = H(\theta_{ML}) - J(\theta_{ML}) = 0$ which, under regularity conditions and appropriately scaled, has a $\chi^2(q)$ distribution, where $q \leq 0.5p(p+1)$ and $p = dim(\theta, \eta)$. Since each $f(y_{it} \in A_i, \theta, \eta_i)$ is assumed to be misspecified, one can first use the above procedure to rank submodels using the value of the χ^2 statistic. The composite likelihood then reduces the misspecification present in individual models if $D(\theta_{CL}) < \min_{A_i} D(\theta_{ML,i})$.

To check for this possibility, one could compare $\text{tr}D(\theta_{ML,i})$ to $\text{tr}D(\theta_{CL})$ (or $|D(\theta_{ML,i})|$ to $|D(\theta_{CL})|$), using

$$DD = 100 * (tr[D(\theta_{CL})]/tr[D(\theta_{ML,i})] - 1.0) \tag{46}$$

(46) gives, in percentage terms, the reduction in the distance to the true model achieved by using the composite likelihood

To show in which practical situations gains more likely to occur, we run a simulation exercise. The DGP is a univariate AR(2) $y_t = \rho_1 y_{t-1} + \rho_2 y_{t-2} + e_t,\ e_t \sim (0, \sigma^2)$ and the models used for estimation are an AR(1): $y_t = \rho_1 y_{t-1} + u_t$ and an MA(1): $y_t = \epsilon_t + \alpha_1 \epsilon_{t-1}$. We focus attention on the relationship between true and estimated σ^2, the variance of the estimated error term, which is common across models. Because of misspecification, $\sigma^2_{u,ML}, \sigma^2_{\epsilon,ML}$ are likely to be biased. The question is whether the composite posterior, which considers the information in the two models, gives better estimates of σ^2 than the AR(1) or the MA(1), by how much, and in what conditions.

We first treat (ω_1, ω_2) as fixed and let ω be the weight for the AR(1) model. We construct composite posterior estimates: i) equally weighting the two models, ii) using weights based on the MSE or the Marginal likelihood obtained for each model in a training sample; iii) using weights that optimize the composite marginal likelihood in a training sample. The training sample has 100 observations and the estimation sample 50 observations. Since there are only two parameters in the AR(1) and MA(1) models, and three in the composite likelihood, this is actually a medium sized sample.

Table 1: Estimates of σ^2

$y_t = \rho_1 y_{t-1} + \rho_2 y_{t-2} + e_t,\ e_t \sim N(0, \sigma^2)$, T=50						
DGP	AR(1)	MA(1)	CL, equal weights	CL, ML weigths	CL, MSE weights	CL, optimal weights
$\sigma^2 = 0.5, \rho_1 = 0.7, \rho_2 = -0.1$	0.36(0.03)	0.36 (0.03)	0.38 (0.03)	0.37 (0.03)	0.36 (0.03)	0.48 (0.04)
$\sigma^2 = 0.5, \rho_1 = 0.5, \rho_2 = 0.2$	0.35 (0.03)	0.36 (0.03)	0.37 (0.03)	0.36 (0.03)	0.35 (0.03)	0.47 (0.04)
$\sigma^2 = 0.5, \rho_1 = 0.6, \rho_2 = 0.35$	0.36 (0.03)	0.40 (0.03)	0.40 (0.03)	0.41 (0.03)	0.37 (0.03)	0.49 (0.04)
$\sigma^2 = 1.0, \rho_1 = 0.7, \rho_2 = -0.1$	0.61 (0.04)	0.35 (0.05)	0.62 (0.04)	0.62 (0.04)	0.60 (0.04)	0.78 (0.05)
$\sigma^2 = 1.0, \rho_1 = 0.5, \rho_2 = 0.2$	0.60 (0.04)	0.61 (0.04)	0.61 (0.04)	0.62 (0.04)	0.60 (0.04)	0.78 (0.05)
$\sigma^2 = 1.0, \rho_1 = 0.6, \rho_2 = 0.35$	0.62(0.04)	0.38 (0.05)	0.67 (0.04)	0.67 (0.04)	0.61 (0.04)	0.76 (0.05)
$\sigma^2 = 2.0, \rho_1 = 0.7, \rho_2 = -0.1$	0.95 (0.04)	0.45 (0.04)	0.96 (0.06)	0.96 (0.04)	0.93 (0.04)	1.14 (0.05)
$\sigma^2 = 2.0, \rho_1 = 0.5, \rho_2 = 0.2$	0.93 (0.04)	0.43 (0.04)	0.95 (0.04)	0.95 (0.04)	0.94 (0.04)	1.14 (0.05)
$\sigma^2 = 2.0, \rho_1 = 0.6, \rho_2 = 0.35$	0.01(0.001)	0.01 (0.001)	1.02 (0.008)	1.02 (0.008)	0.99 (0.008)	1.15 (0.05)

ML is the marginal likelihood. The MSE and the ML for the AR(1) and the MA(1) are computed in a sample of 100 observations prior to the T=50 data points used to construct the composite likelihood (CL). The last column is obtained choosing weights to maximize the marginal composite likelihood over the initial 100 points. In parentheses are standard errors of the estimates

We consider a number of configurations for $(\rho_1, \rho_2, \sigma^2)$. Table 1 reports a subset

of the results we obtain: for each parameter vector, we report the posterior mean and the posterior standard deviation of σ^2 in the AR(1) and MA(1) models and in the four composite posterior pools. In all cases, the prior for the AR (MA) parameters is truncated normal but loose (mean zero and variance one) and the prior for σ is relatively flat in the positive orthant.

There are location gains when using the composite posterior; they are larger when the DGP is persistent or has a large volatility, and seem insensitive to the choice of weights. As often documented in the forecasting combination literature (see Aiolfi et al, 2010), choosing equal weights is as good as choosing them either based on MSE or the marginal likelihood in a training sample (compare columns 4, 5 and 6). However, choosing the weights to optimize the composite likelihood in the training sample, seems to give an important hedge to the approach: location gains are large and they increase, the smaller is the volatility of the DGP. It is important to stress that the approach employed in column 7 is feasible even when the models feature different observables, while this is not the case for the results produced in, e.g., column 5.

Relative to the AR(1) or MA(1) models, reduction gains in the spread of the posterior are small. Two reasons account for this outcome. First, we only consider two models; in analogy with a portfolio decision problem, reduction in the dispersion are more likely to occur the larger the number of models is. Second, mean estimates of σ^2 obtained with the AR(1) and the MA(1) models do not differ much for several parameter configuration. Thus, dispersion gains are relatively small.

Table 2: Determinant misspecification diagnostic

DGP	MA(1)	CL, equal weights	CL, ML weigths	CL, MSE weights	CL, optimal weights
$\sigma^2 = 0.5, \rho_1 = 0.7, \rho_2 = -0.1$	-0.0151	-0.0958	-0.0958	-0.0927	-0.0921
$\sigma^2 = 0.5, \rho_1 = 0.5, \rho_2 = 0.2$	0.0016	-0.0951	-0.0951	-0.0916	-0.0924
$\sigma^2 = 0.5, \rho_1 = 0.6, \rho_2 = 0.35$	0.0761	-0.0959	-0.0953	-0.0956	-0.0954
$\sigma^2 = 1.0, \rho_1 = 0.7, \rho_2 = -0.1$	0.0262	-0.0963	-0.0963	-0.0941	-0.0937
$\sigma^2 = 1.0, \rho_1 = 0.5, \rho_2 = 0.2$	0.0931	-0.0950	-0.0950	-0.0921	-0.0917
$\sigma^2 = 1.0, \rho_1 = 0.6, \rho_2 = 0.35$	0.0367	-0.0952	-0.0952	-0.0915	-0.0923
$\sigma^2 = 2.0, \rho_1 = 0.7, \rho_2 = -0.1$	-0.0889	-0.0991	-0.0992	-0.0991	-0.0991
$\sigma^2 = 2.0, \rho_1 = 0.5, \rho_2 = 0.2$	2.0275	-0.0893	-0.0899	-0.0863	-0.0856
$\sigma^2 = 2.0, \rho_1 = 0.6, \rho_2 = 0.35$	-0.0606	-0.0875	-0.0876	-0.0768	-0.0792

ML is the marginal likelihood. The MSE and the ML for the AR(1) and the MA(1) are computed in a sample of 100 observations prior to the T=50 data points used to construct the composite likelihood (CL). The last column is obtained choosing weights to maximize the marginal composite likelihood over the initial 100 points. Numbers are multiplied by 10000.

Table 2 reports the determinant diagnostic, when the AR(1) model is used to benchmark the misspecification. Clearly, for many configurations the MA(1) model is more misspecified than the AR(1). The combination pools always reduce the misspecification

present in the AR(1) - all entries in columns 2-6 are negative - but which combination pools one uses seems less important as magnitude differences are small. Overall, the simple equal weighting scheme does well, regardless of the parameters of the DGP.

Figure 1 plots the marginal composite posterior of σ^2 obtained when the data has been generated by $y_t = 0.6y_{t-1}+0.35y_{t-2},\ e_t, \sim N(0, 0.5)$ in three cases: equally weighting the AR(1) and the MA(1) models; optimally selecting ω to maximize the composite marginal likelihood in the training sample; and letting ω be a random variable with a normal prior distribution centered at 0.5 and standard deviation equals to 0.1, When equally weighting the two models or optimally selecting ω and the mode is somewhat lower than the true value. However, the posterior of σ^2 obtained with optimal weights is more skewed and displays a longer right tail. The posterior of σ^2 obtained with random weights, on the other hand, is centered at the true value (mode=0.502) but has larger dispersion relative to the other two posterior distributions- the presence of an additional parameter to be estimated explains why.

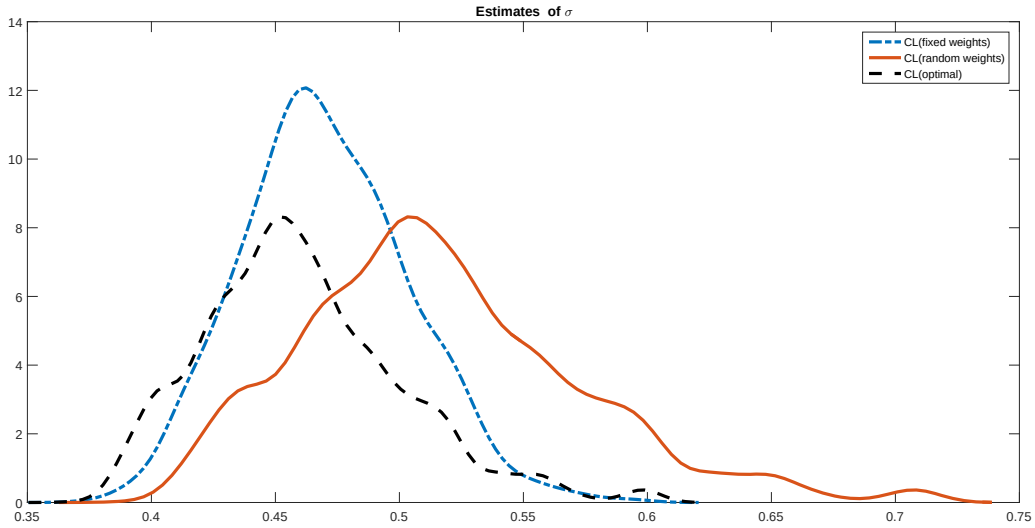


Figure 1: Composite posterior density of σ^2 and two posterior mixtures

Figure 2 plots the posterior of ω: it is heavily skewed and with a long left tail. In accordance with previous results, the AR(1) model receives more weight relative to the MA(1), since the DGP has stronger persistence. Thus, while there are location gains from having a random ω, at least in this example, taking ω random could increase the dispersion of the posterior of the common parameter. Overall, having a random ω is probably more appealing when composite posterior gains obtained using fixed weights are parametrization dependent.

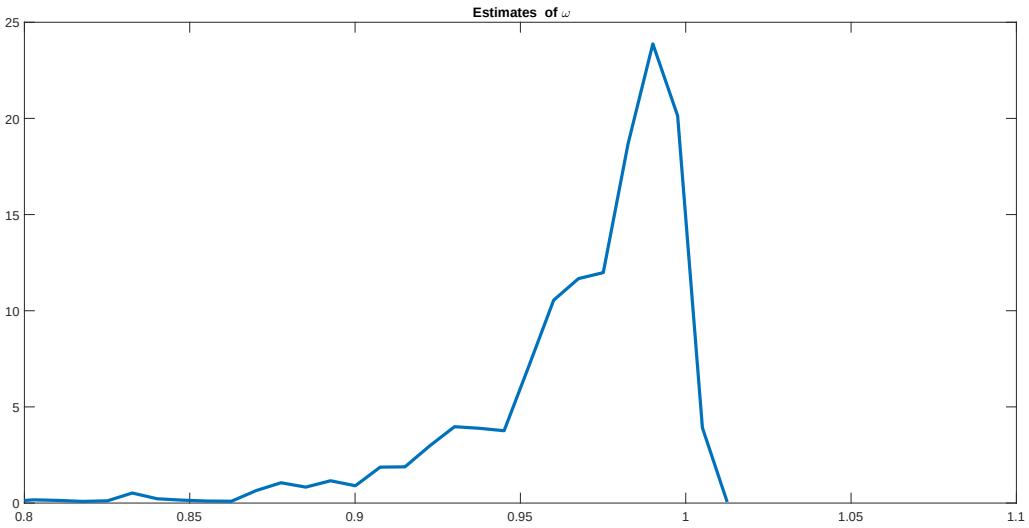


Figure 2: Posterior density of ω

7 Estimating the slope of the Phillips curve

The conventional wisdom in monetary economics is that the slope of the New Keynesian Phillips curve is historically small (see Smets and Wouters, 2007, or Altig et al., 2011). Thus, large changes in firms' marginal costs imply small changes in the aggregate inflation rate. In addition, there is evidence that the slope has further decreased since 2009 (see e.g. Coibon and Gorodnichenko, 2013), perhaps because financial constraints imply a trade-off between pricing decisions and firms' market share (see e.g. Gilchrist et al., 2015). However, Schorfheide (2008), surveying estimates of the slope of the Phillips curve obtained in DSGE models, documents large cross study variations and associates differences to i) the specification of the model, ii) the observability of marginal costs, iii) the number and the type of variables used in estimation.

Because the distribution of this parameter is crucial for forecasting and counterfactual exercises, this section examines how the composite posterior distribution of the slope of the Phillips curve, obtained employing a variety of models, looks like relative to the posterior distribution obtained with a i) single model and ii) ex-post averaging the posteriors of different models. Moreover, it measures the consequences of these differences for economic inference.

We construct the composite posterior using five different DSGE models: a small scale New Keynesian model with sticky prices but non-observable marginal costs, where the variables used in estimation are detrended output Y, demeaned inflation π, and demeaned nominal rate R, as in Rubio and Rabanal (2005); a small scale New Keynesian model with sticky prices and sticky wages, and observable marginal costs, where the variables used in estimation are detrended Y, demeaned π, demeaned R and de-

Table 2:Percentiles of the posterior of the slope of the Philips curve

	5%	50%	95%
Prior	0.01	0.80	1.40
Basic NK	0.06	0.18	0.49
Basic NK with nominal wages	0.05	0.06	0.07
SW with capital and adj.costs	0.04	0.05	0.07
Search	0.44	0.62	0.86
BGG	0.13	0.21	0.35
CL	0.18	0.26	0.40
CL (corrected)	0.18	0.28	0.44

The table reports posterior percentiles of the slope of the Phillips curve for the prior, for a three variable New Keynesian model (Basic NK); for a four variable New Keynesian model (Basic NK with nominal wage); for a medium scale New Keynesian model with seven observables (SW with capital and adj. costs), for the four variable search and matching model (Search) and the three variable financial friction model (BGG). The rows with CL report composite posterior percentiles obtained with MCMC draws and adjusted for repeated use of the same time series. The estimation sample is 1960:1-2005:4.

trended nominal wage W, again as in Rubio and Rabanal (2005); a medium scale New Keynesian model with sticky prices, sticky wages, habit in consumption and investment adjustment costs, where the variables used in estimation are detrended Y, detrended consumption, detrendend investment, demeaned π, demeaned R, detrended hours, and detrended nominal wage W, as in Justiniano et al. (2010); a New Keynesian model with search and matching labour market frictions, where the variables used in estimation are detrended Y, demeaned π, demeaned R and detrended real wage w, as in Christoffel and Kuester (2008); and a version of the Bernanke, Gertler, and Gilchrist (1999) model, estimated with detrended output Y, demeaned inflation π, and demeaned nominal rate R. In this last model, part of the parameters governing the financial frictions are calibrated, as in Cogley et al (2011), to sidestep the issue of what data series should be used to match the model-implied spread. In all cases, the estimation sample is 1960:1-2005:4 and a quadratic trend is used to detrend the data. The series used are from the Smets and Wouters (2007) database and the equations of each model are reported in appendix F.

We report results assuming random weights, where the prior is Dirichlet with parameters 250*[1/4;1/3;1/7;1/4;1/3]. Results obtained using a looser prior (Dirichlet with parameters 40*[1/4;1/3;1/7;1/4;1/3]), and fixed equal or unequal weights are similar and available on request [8]

[8]While it is not the case in our specific example, it may be that in some applications the posterior weight for some model goes to zero implying that the parameters of that model become under-identified when the composite likelihood is used in estimation. In this case, one can use a two-step approach, where the prior for the nuisance parameters is made data-based using the posterior for each model estimated on a

Table 2 displays posterior percentiles for the slope of the Phillips curve we obtained using the likelihood of each model separately and the composite likelihood. For the first three models the median is low. For the remaining two models, the posterior median is higher but, for the model with search and matching friction, the posterior has larger dispersion. Nevertheless, the posterior one obtains hardly overlaps with those of the first three models. Thus, in agreement with Schorfheide, estimation results depend on the model employed, the nuisance features it includes, the observability of marginal costs, and the variables used in the estimation.

The composite posterior has median value of 0.26, and a credible 90 percentile ranging from 0.18 to 0.40, which is smaller than the range obtained with a number of individual models. When we correct the percentiles for the fact that each model is misspecified and the same time series are repeatedly used in the estimation process, the location and the spread of the posterior distribution hardly change, indicating that, at least for this particular application, the non-standard features of our estimation procedure are not crucial for the results.

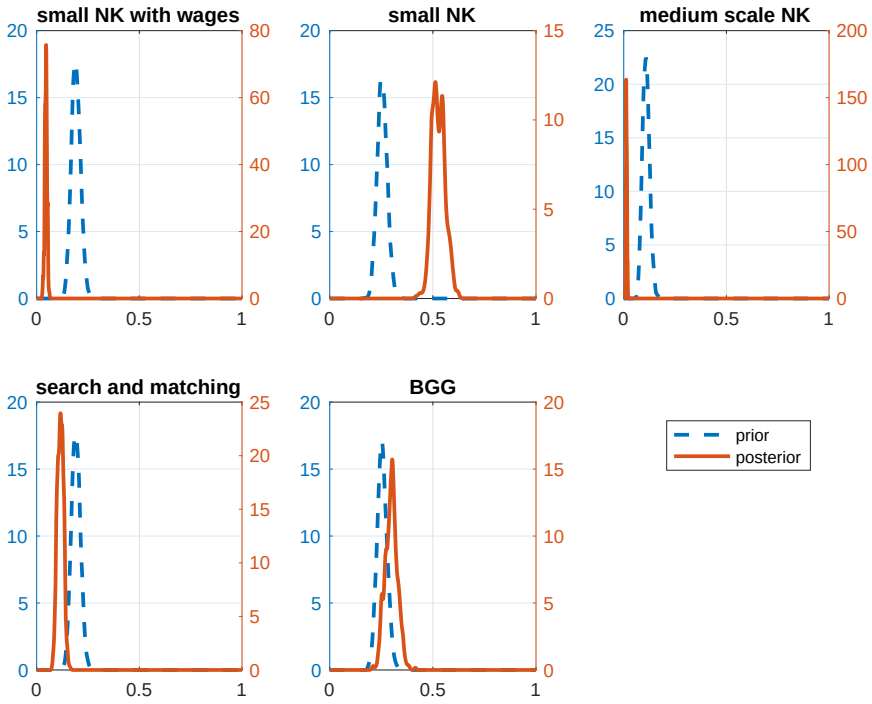


Figure 3: Prior and posterior densities of ω

Figure 3 reports the prior and the posterior weights on each of the models. The data is informative about the quality of each model and both the location and the spread of the prior distribution are updated in the estimation process. Interestingly,

training sample. This effectively avoids under-identification and makes the priors for the nuisance parameters endogenous.

while the location of posterior for the models with financial frictions and with labour market friction are the least affected by the estimation process, the location of the small NK model with observable marginal costs and of the medium scale NK model all decrease relative to the prior and the one of the basic NK model increases. In general, posterior spreads are tighter than prior spread, indicating that the data is informative about the weights (see Mueller, 2012). Overall, the composite posterior estimates of the Phillips curve reflect, to a large extent, the information present in the small scale New Keynesian model and, to a less extent, in the BGG and the search and matching model.

Some reader may be surprised about the fact that the standard medium scale New Keynesian model, which is the workhorse used in many policy institution, has the lowest posterior probability among our five models. To provide further evidence on its performance, it is useful to compute the misspecification diagnostics presented in section 6. Because, there are lots of parameters and numerical considerations are important, we focus attention on the slope parameter only, keeping all others fixed at the estimated mode. Indeed, we find that relative to the small NK model (which we take as our benchmark), the small scale model with sticky prices and sticky wages and the medium scale New Keynesian model are the most misspecified (the DD statistics are 1117 and 824). In addition, we find that regardless of the prior specification, our composite model is less misspecified relative than the benchmark (the DD statistic is -70) and that the pooling of models provide robustness but not the best possible outcome in this sample (the DD for the model with search and matching friction is -91).

Figure 4 presents the composite posterior distribution for the slope of the Phillips curve we obtain together with two alternative naive posterior combination estimators: one that equally weights the posteriors obtained with the five models; and one which weights the posteriors by the in-sample MSE they produce for inflation and the nominal rate. Two features of the figure deserve attention. First, combining ex-post estimates generate distributions which are different from those obtained with a composite posterior approach; the location is lower and the spread much larger when naive posterior combinations are computed. Second, both naive combinations have multimodal posteriors: a sharper one at 0.05, a secondary one at 0.15, and a small one at 0.6, reflecting the variety of estimates one obtain with the five models under consideration. Thus, naive posterior short cuts may lead to wrong conclusions regarding the pass-through of marginal costs to inflation.

Nowadays, there is a lot of interest in policy institutions about the ability of monetary policy to influence real rates in the long run, via forward guidance and other unconventional tools. The other side of this coin is that monetary policy should have little effect on the real rate in the short run. Figure 5 reports the responses of the ex-ante real rate to a monetary policy shock with a mean of 25 basis points and a standard deviation of half the mean, using parameters estimates obtained with each model and the composite responses constructed as in section 5.[9]

[9]To obtain non-degenerate distributions over which we can form Monte Carlo averages, we can not fix

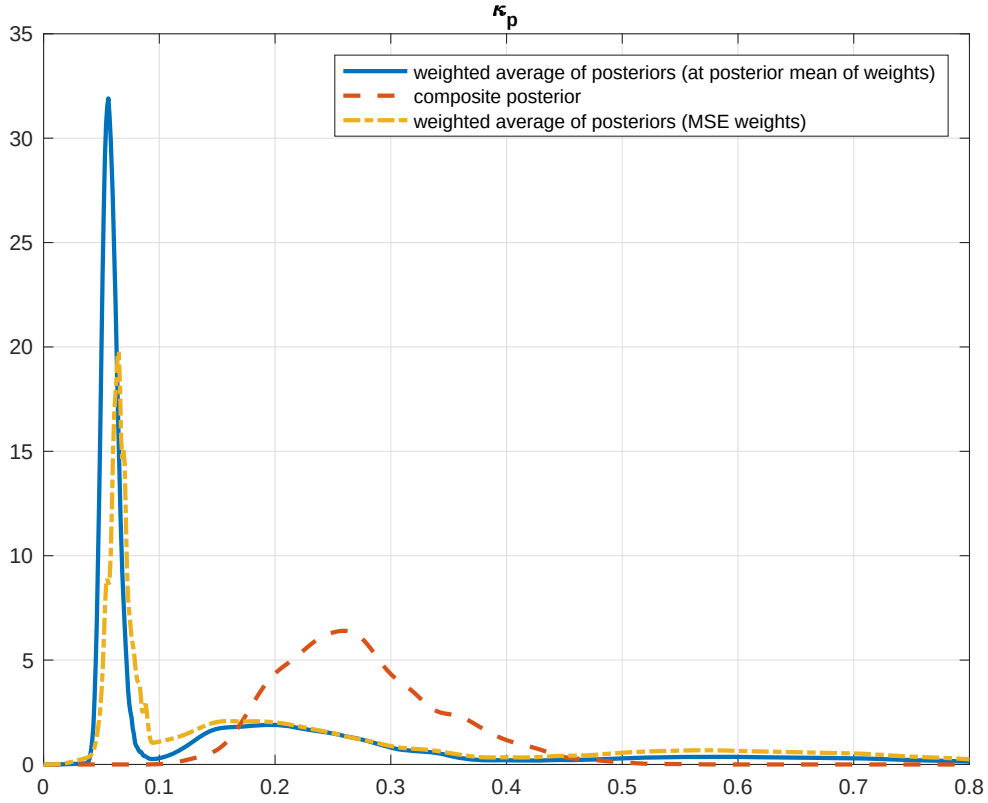


Figure 4: Composite posterior and two naive posterior mixtures

The mean impact is quite uncertain and depending on the model it may range from 20 to 60 basis points, with the models featuring labor and financial frictions on the lower side. The composite impact response is intermediate among individual values (about 45 basis points) and not very precisely estimated - it is significantly different from zero but the 68% credible set includes the point estimate of all models. At horizons larger than the first one, the composite posterior for the real rate response becomes tighter and the responses obtained with the search and matching friction model and the medium scale NK model now fall outside the credible composite posterior interval. Finally, note that, consistent with figure 3, the location of the composite posterior distribution of the real rate responses at all horizons heavily reflects the real rate responses obtained with the small scale NK model without observable marginal costs.

Once we analyze the response of the ex-ante real rate to a monetary policy shock, it is natural to ask how different the estimates from the different models for this real rate are. Why is this important? The ex-ante real rate gives us a measure of the stance of monetary policy: values larger than the steady state (or 'natural rate') imply monetary policy that is tighter than in steady state and vice versa. Figure 6 depicts

---

the size of the shock.

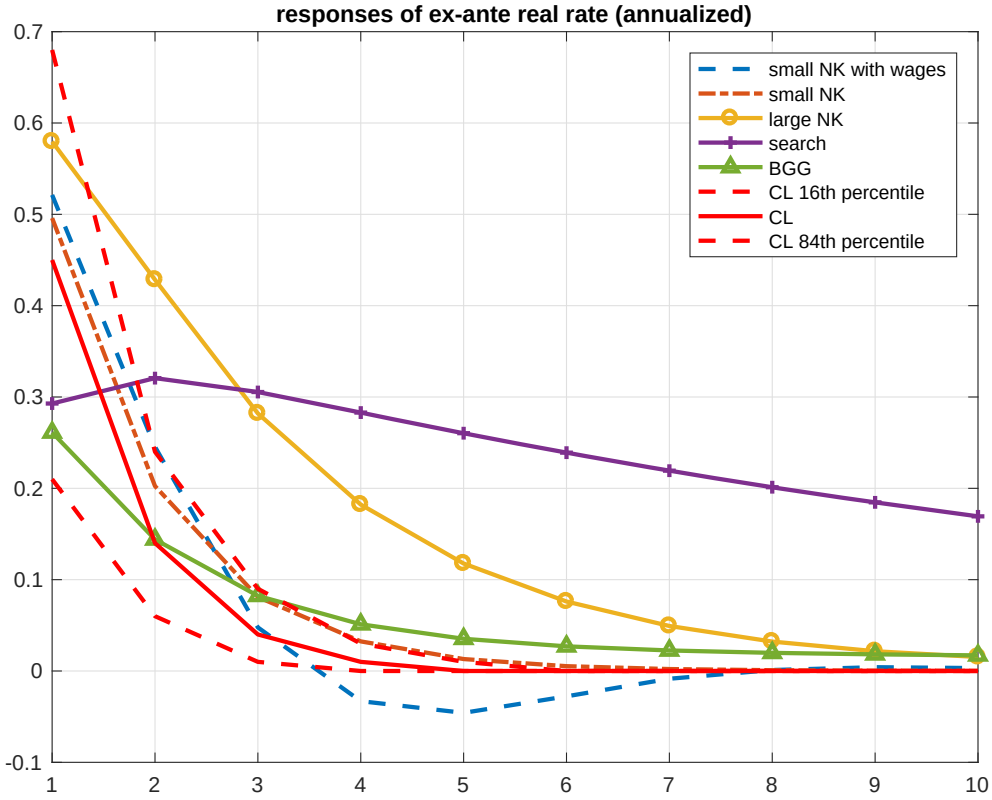


Figure 5: Real rate responses to a monetary shock

the estimated ex-ante real rate in deviations from steady state. We can see that there is substantial variation across models (even though the real rates across models are highly correlated, especially at lower frequencies) - the gray lines depict the medians for each model.[10]

8 Conclusions

This paper describes an estimation procedure which has the potential to ameliorate identification and estimation problems in DSGE models. We also demonstrate that the procedure helps to robustify estimates of the structural parameters in a variety of interesting economic problems and reduce misspecification.

The approach we propose is easy to use, works well in a number of relevant situations, produces estimators with nice small sample properties and, in its Bayesian version we pursue throughout the paper, has an appealing sequential learning interpretation and allows to make posterior statements about models featuring different

[10] As a check, we also computed the ex-post real rate, which is observable with our data set. Our estimate of the ex-ante real rate is highly correlated with this ex-post measure

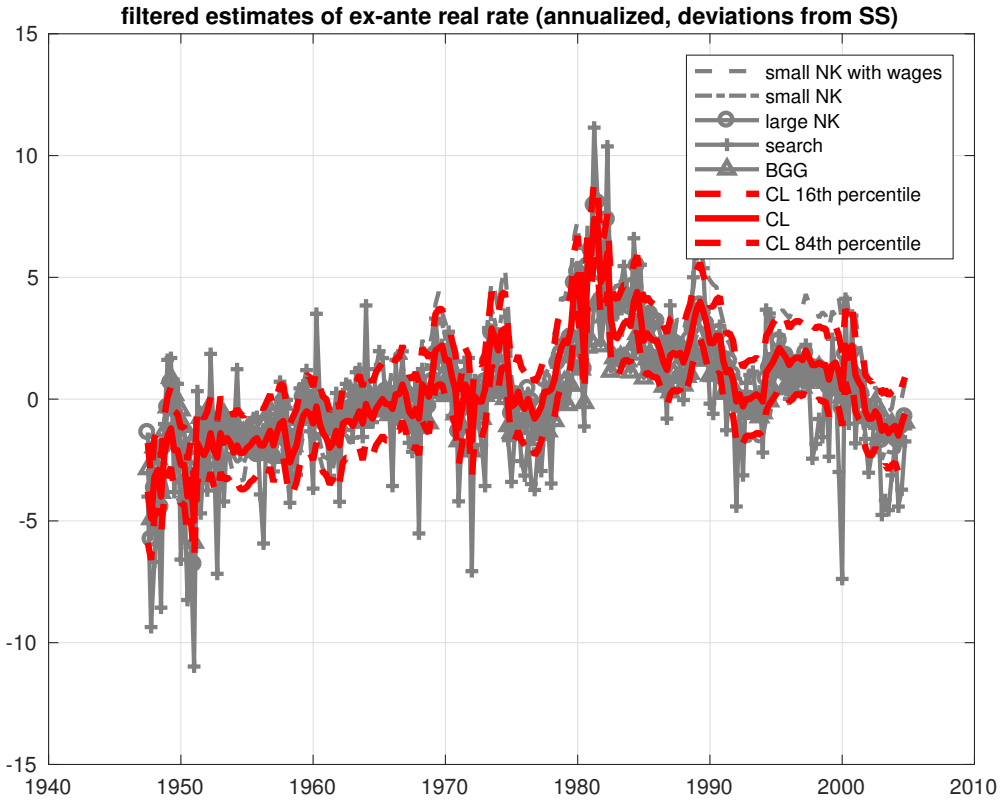


Figure 6: Estimated ex-ante real rates

observables, if that is of interest.

By way of examples we highlight the properties of procedure. In particular, that it can be used to i) obtain shrinkage estimates of the parameters appearing in multiple (nested and non-nested) misspecified structural models; ii) improve the (sample and population) identification properties of troublesome parameters, iii) provide a tractable approach to address computational and singularity problems; iv) exploit information coming from cross-section data or data obtained at different levels of aggregation; v) produce stable estimates of parameters present in large scale models.

.

9 References

Aiolfi, M., Capistran, C., and A. Timmerman (2010). Forecast combinations in Clements, M. and D. Hendry (eds.) Forecast Handbook, Oxford University Press, Oxford.

Altig, D. Christiano, L. Eichembaum, M. and J. Linde (2011) Firm-specific capital, nominal rigidities and the business cycle. Review of Economic Dynamics, 14, 225-247.

Andreasen, M., Fernandez Villaverde, J., and J. Rubio Ramirez (2014). The pruned state space system for Non-Linear DSGE Models: Theory and Empirical Applications, NBER working paper 18983.

Baumeister, C. and J. D. Hamilton (2015). Structural Interpretation of Vector Autoregressions with Incomplete Identification: Revisiting the Role of Oil Supply and Demand Shocks, manuscript.

Bernanke, B., Gertler, M., and S. Gilchrist (1999). The financial accelerator in a quantitative business cycle framework. Handbook of Macroeconomics.

Billio, M., Casarin, R., Ravazzolo, F., and van Dijk, H. K. (2013). BTime-varying combinations of predictive densities using nonlinear filtering. Journal of Econometrics, 177 (2), 213-232.

Boivin, J. and M. Giannoni (2006). Data-rich DSGE models, manuscript.

Canova, F. (2014). Bridging DSGE models and the raw data. Journal of Monetary Economics, 67, 1-15.

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

Canova, F., Ferroni, F., and C. Matthes (2014). Choosing the variables to estimate DSGE models. Journal of Applied Econometrics, 29, 1009-1117.

Chernozhukov, V. and A. Hong (2003). An MCMC approach to classical inference, Journal of Econometrics, 115, 293-346.

Chib, S. and S. Ramamurthy (2010). Tailored Randomized-block MCMC Methods with applications to DSGE models, Journal of Econometrics, 155, 19-38.

Christoffel, K. and K. Kuester (2008). Resuscitating the wage channel in models with unemployment fluctuations. Journal of Monetary Economics, 55, 865-887.

Cogley, T., de Paoli, B., Matthes, C., Nikolov, K., and T. Yates (2011). A Bayesian Approach to Optimal Monetary Policy with Parameter and Model Uncertainty. Journal of Economic Dynamics and Control, 35, 2186-2212.

Coibon, O. and Y., Gorodnichenko (2013). Is the Phillips curve alive and well after all? Inflation expectations and the missing deflation, University of Berkeley, manuscript.

Del Negro, M. and F. Schorfheide (2004). Prior for General equilibrium models for VARs. International Economic Review, 45, 643-573.

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

Del Negro, M., Hasegawa, R., and F. Schorfheide (2016). Dynamic Prediction Pools: An Investigation of Financial Frictions and Forecasting Performance. Journal

of Econometrics, 192, 391-405.

Domowitz, I and H. White (1982). Misspecified models with dependent observations. Journal of Applied Econometrics, 20,35-58

Engle, R. F., Shephard, N. and K. Sheppard, (2008). Fitting vast dimensional time-varying covariance models., Oxford University, manuscript.

Edwards, A.W. F. (1969). Statistical methods in scientific inference, Nature, Land 22, 1233-1237.

Gao, X. and P. Song (2010). Composite Likelihood Bayesian information criteria for model selection in high dimensional data, Journal of the American Statistical Association, 105, 1531-1540.

Gilchrist, S., Sim, J., Schoenle, R., and E. Zackrajsek (2016).Inflation dynamcis during the financial crisis, forthcoming, American Economic Review.

Geweke, J. and G. Amisano (2011). Optimal Prediction Pools, Journal of Econometrics, 164, 130-141.

Guerron Quintana, P. (2010). What do you match does matter: the effect of data on DSGE estimation. Journal of Applied Econometrics, 25, 774-804.

Herbst, E. and F. Schorfheide (2015) Bayesian Estimation of DSGE models, Princeton University Press, Princeton, NJ.

Justianiano, A. Primiceri, G. and A. Tambalotti (2010). Investment shocks and the business cycle. Journal of Monetary Economics, 57, 132-145.

Komunjer, I and S. Ng (2011) Dynamic identification of DSGE models. Econometrica, 79, 1995-2032.

Kim, J.Y. (2002). Limited information likelihood and Bayesian methods. Journal of Econometrics, 108, 175-193.

Lee, L. F. and W. Griffith (1979). The prior likelihood and the best linear unbiased prediction in stochastic coefficients linear models, http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.518.5107&rep=rep1&type=pdf.

Mueller, U. K. (2013). Risk of Bayesian Inference in Misspecified Models, and the Sandwich Covariance Matrix. Econometrica, 81, 1805 1849.

Pagan, A. (2016). An unintended consequence of using errors-in-variables shocks in DSGE models?, manuscript.

Pakel, C., Shephard N. and K. Sheppard (2011). Nuisance parameters, composite likelihoods and a panel of GARCH models. Statistica Sinica, 21, 307-329.

Pauli, F., Racugno, W., and L. Ventura (2011). Bayesian composite marginal likelihoods. Statistica Sinica, 21, 149-164.

Qu, Z. and D. Thackenko (2012). Identification and frequency domain QML estimation of linearized DSGE models. Quantitative Economics, 3, 95-132.

Qu, Z. (2015). A Composite likelihood approach to analyze singular DSGE models, Boston University manuscript.

Ribatet, M., Cooley, D. and A. Davison (2012). Bayesian inference from composite likelihoods, with an application to spatial extremes. Statistica Sinica, 22, 813-845.

Rubio Ramirez, J. and P. Rabanal (2005). Comparing New Keynesian models of the business cycle. Journal of Monetary Economics, 52, 1151-1166.

Schorfheide, F. (2008). DSGE model-based estimation of the New Keynesian Phillips curve. Federal Reserve of Richmond, Economic Quarterly, 94(4), 397-433.

Smets, F. and R. Wouters (2007). Shocks and Frictions in US Business Cycles: A Bayesian DSGE Approach. American Economic Review, 97, 586-606.

Varin, C., Read, N. and D. Firth (2011). An overview of Composite likelihood methods. Statistica Sinica, 21, 5-42.

Vidoni, P. (2013) A Note on predictive densities based on composite likelihoods, manuscript.

Waggoner, D. and T. Zha (2012). Confronting model misspecification in macroeconomics. Journal of Econometrics, 146, 329-341.

White, H. (1982). Maximum likelihood estimation of misspecified models. Econometrica, 50, 1-25.

Appendix A

Let y_t be a sample from the density $f(y_t)$ with respect to some σ-measure μ. Suppose a model with the density $g(y_t, \psi)$, where $\psi \in \Psi \subset R^m$ is a vector of parameters, is used and the log-likelihood is $L_g(\psi) = \sum_t \log g(y_t, \psi)$. The model is misspecified because $f(y_t) \neq g(y_t, \psi)$, $\forall \psi$. Let ψ_{ML} be the maximum likelihood estimator, i.e. $\psi_{ML} = sup_\psi L_g(\psi)$. Since $T^{-1}L_g(\psi) \to E(\log g(y_t, \psi))$ by a uniform law of large numbers, ψ_{ML} will be consistent for $\psi_0 = arg\max_\psi E \log g(y_t, \psi)$, where the expectations are taken with respect to the density f. If f is absolutely continuous with respect to g

$$E \log g(y_t, \psi) - E \log f(y_t) = -\int f(y_t) \log \frac{f(y_t)}{g(y_t, \psi)} d\mu(y) = -KL(\psi) \tag{47}$$

where $KL(\psi)$ is the Kullback-Liebler divergence between f and g. Hence ψ_0 is also the minimizer of $KL(\psi)$.

Let $s_t(\psi) = \frac{\partial \ln g(y_t, \psi)}{\partial \psi}$ be the score of observation t and let $h_t(\psi) = \frac{\partial s_t(\psi)}{\partial \psi}$. If the maximum is in the interior $\sum_t s_t(\psi) = 0$, and taking a first order expansion we have

$$0 \approx T^{-0.5} \sum_t s_t(\psi_0) + T^{0.5} \Sigma_1^{-1} (\psi_{ML} - \psi_0) \tag{48}$$

where $\Sigma_1 = -E(h_t(\psi_0)) = \frac{\partial^2 KL(\psi)}{\partial \psi \partial \psi^T}|_{\psi=\psi_0}$. Then, using a central limit theorem for correlated observations we have that $T^{-0.5}(\psi_{ML} - \psi_0) \sim N(0, V)$ where $V = \Sigma_1 \Sigma_2 \Sigma_1$ and $\Sigma_2 = E(s_t(\psi) s_t(\psi)')$.

In standard DSGE applications $s_t(\psi)$ are computed with the Kalman filter and are function of martingale difference processes(the shocks of the model). Thus, the condition $\sum_t s_t(\psi) = 0$ is likely to hold. Further regularity condition (see, e.g. Mueller(2013)) need to be imposed for the argument to go through.

The composite likelihood is weighted average of different models $g(y_t, \psi_i)$ each of which is misspecified. Thus the resulting composite model is in general misspecified with density $\tilde{g}(y_t, \psi_1, \dots \psi_K) = \tilde{g}(y_t, \theta, \eta_1, \dots \eta_K)$. Repeating the argument of the previous paragraph, the composite likelihood estimator θ_{CL} minimizes the $KL(\theta)$ divergence between the $\tilde{g}$ and f. Under regularity conditions θ_{CL} convergences to $\theta_{0,CL}$ and its distribution is normal with zero mean and covariance maatrix $V_{CL} = \Sigma_{1,CL} \Sigma_{2,CL} \Sigma_{1,CL}$ where $\Sigma_{2,CL} = E(s_{t,CL}(\theta, \eta_1, \dots, \eta_K) s_{t,CL}(y_t, \theta, \eta_1, \dots \eta_K)')$, $\Sigma_{1,CL} = \frac{\partial s_{t,CL}(y_t, \theta, \eta_1, \dots \eta_K)}{\partial \theta}$ and $s_{t,CL} = \frac{\partial \tilde{g}(y_t, \theta, \eta_1, \dots, \eta_K)}{\partial \theta}$.

Appendix B

Let θ_{CL} be the maximum composite likelihood estimator of θ and let θ_p be the mode of the prior $p(\theta)$.Let $h(\theta_{CL}) = -\nabla^2_\theta CL(\theta_{CL}|y)$ and $h(\theta_p) = -\nabla^2_\theta \log p(\theta_p)$. Taking a second order expansion of $p_{CL}(\theta|Y)$ we have

$$\begin{aligned} p_{CL}(\theta|Y) &\propto \{CL(\theta_{CL}|y) - 0.5(\theta - \theta_{CL})^T (h_(\theta_{CL})(\theta - \theta_{CL}) + logp(\theta_p) - 0.5(\theta - \theta_p)^T (h_(\theta_p)(\theta - \theta_p)\} \\ &\approx N(\hat{\theta}, h(\theta_{CL}, \theta_p)^{-1}) \end{aligned} \tag{49}$$

where $\hat{\theta} = h(\theta_{CL}, \theta_p)^{-1}(h(\theta_{CL})\theta_{CL} + h(\theta_p)\theta_p)$ and $h(\theta_{CL}, \theta_p) = h(\theta_{CL}) + h(\theta_p)$.

Under regularity conditions, $p(\theta)$ will vanish as $T \to \infty$. Then, almost surely, the strong law of large number implies that

$$T^{-1}h(\theta_{CL}, \theta_p) \rightarrow -E(\triangledown^2 CL(\theta_0|Y)) \equiv H(\theta_0) \quad (50)$$

$$\hat{\theta} = (T^{-1}h(\theta_{CL}, \theta_p))^{-1}(T^{-1}h(\theta_{CL})\theta_{CL} + T^{-1}h(\theta_p)\theta_p) \rightarrow \theta_0 \quad (51)$$

and thus $p_{CL}(\theta|Y) \approx N(\theta_0, T^{-1}H(\theta_0)^{-1})$.

Appendix C

We look for the predictive density solving the following problem

$$\hat{p} = arg\min_p KL(p(z|y), f(z)) \quad (52)$$

where z is any future sequence of y subject to

$$E_p\{\log\frac{f(z|y_i)}{f(z)}\} = E_{Z|Y=y}\{\log\frac{f(z|y_i)}{f(z)}\} \quad i = 1, \ldots, K \quad (53)$$

and the normalization condition $E_p(1) = 1$, where E_p is the expectation with respect to the density $p(z|y)$ and $f(z)$ is any preliminary density of z, for example, its marginal. In words, we seek for the predictive density which is closest in the KL sense to any preliminary marginal density $f(z)$ and reproduces the same conditional expectation as the true density $f(z|y)$ on $log\frac{f(z|y_i)}{f(z)}$ [11]. The solution is $\hat{p}(z|y) = f(z)\exp\{\sum_i \xi_i \log\frac{f(z|y_i)}{f(z)}\} - Z(y_i, \xi)\}$ where $Z(y_i, \xi)$ is a normalizing constant, ξ_i are the Lagrange multipliers on the constraints (53). $\hat{p}(z|y)$ has an exponential tilting format: we tilt $f(z)$ in the directions spanned by $\log\frac{f(z|y_i)}{f(z)}$. If $\xi_i \geq 0, \sum_i \xi_i \leq 1$, then $\hat{p}(z|y)$ is the composite predictive density previously derived with $\omega_i = \xi_i,\ i = 1, \ldots, K$ and $\omega_0 = 1 - \sum_i \xi_i$ where ω_0 is the weight on $f(z)$. Note that in this setup, the weights ω_i satisfy the following (score) equations:

$$\frac{\partial E_{z|Y=y} \log f_p(Z|y, \omega)}{\partial \omega_i} = 0, \quad i = 1, \ldots K \quad (54)$$

Thus, ω_i can be chosen to maximize the conditional expected logarithmic score.

[11] When $f(z)$ is disregarded, the problem becomes one of maximizing the entropy -$E_p[\log p(z|y)]$, subject to the constraints.

Appendix D

Consider observations t=1 and t=2. From (34) and (37) we have

$$\tilde{d}_1 = d_1 \tag{55}$$

$$\tilde{d}_2 = d_2 - \frac{\zeta}{1+\zeta^2} d_1 \tag{56}$$

$$\tilde{p}_1 = p_1 \tag{57}$$

$$\tilde{p}_2 = p_2 - \frac{\zeta}{(1-\beta\zeta)^2+\zeta^2} p_1 \tag{58}$$

Since $\frac{\zeta}{1+\zeta^2} < \frac{\zeta}{(1-\beta\zeta)^2+\zeta^2}$ $\tilde{p}_2$ puts more weights on p_1 relative to p_2 than $\tilde{d}_2$ does on d_1 relative to d_2. By induction, $\tilde{p}_t$ puts more weights on $p_{t-j},\ j > 0$ relative to p_t than does $\tilde{d}_t$ on d_{t-j} relative to d_t. Thus, $\tilde{p}_t$ has a stronger memory than $\tilde{d}_t$.

Similarly, using (35) and (38), for t=1 and t=2 we have

$$\varsigma_1 = \sigma^2(1+\zeta^2) \tag{59}$$

$$\varsigma_2 = \sigma^2 \frac{1+\zeta^2+\zeta^4}{1+\zeta^2} \tag{60}$$

$$\lambda_1 = \sigma^2((1-\beta\zeta)^2+\zeta^2) \tag{61}$$

$$\lambda_2 = \sigma^2(1-\beta\zeta)^2 \frac{1+\frac{\zeta^2}{(1-\beta\zeta)^2}+\frac{\zeta^4}{(1-\beta\zeta)^4}}{1+\frac{\zeta^2}{(1-\beta\zeta)^2}} \tag{62}$$

Clearly $\lambda_1 < \varsigma_1$ and $\lambda_2 < \varsigma_2$. Proceeding by induction, we have that $\lambda_t < \varsigma_t$. Thus, the model for $\tilde{p}_t$ implies larger weighs on $\tilde{p}_t^2$ relative to $\log \lambda_t$ while the model for $\tilde{d}_t$ implies smaller weight on $\tilde{d}_t^2$ relative to $\log \varsigma_t$ at each t. Combining these two results, we have that ζ will be identified and estimated more from the serial correlation properties of the data if $\tilde{p}_t$ is used to construct the likelihood function and more from the variance properties of the data if $\tilde{d}_t$ is used to construct the likelihood function.

Appendix E

There are various alternatives to draw candidate weights $\omega_i,\ i = 1, ..., K$. If only few submodels are considered (say two), then using an independent Dirichlet proposal works well. When the number of submodels is large, one could first employ a logistic transformation for the weights and then use a random walk proposal on the transformed weights. This approach has the disadvantage that the proposal for the weights is no longer a multivariate random walk (in particular, it is no longer symmetric) and this needs to be taken into account when computing the acceptance probability in Metropolis-within-Gibbs algorithm. Furthermore, one needs to compute the Jacobian of the mapping induced by the logistic distribution, which may be tedious to code and may lead to numerical issues because of the non-linear transformation of parameters.

Our preferred approach is to use a proposal density which directly operates on the weights. We call this proposal a 'random-walk Dirichlet' since the expected value of

the proposal is the last accepted draw. Denote by $\omega^{accepted}$ the last accepted vector of weights, by $\omega^{proposal}$ a proposal draw, and by $scaling > 0$ a (scalar) scaling variable, which determines the variance of the proposal. The proposal density is Dirichlet, denoted by $p_D(\omega^{proposal}|\omega^{accepted}, scaling)$ with parameter $scaling*\omega^{accepted}$. The mean of this proposal is independent of $scaling$ and equal to $\omega^{accepted}$. The variance of any element of $\omega^{proposal}$ is a decreasing function of $scaling$. In an initial adaptive phase where draws are discarded before computing posterior quantities of interest, we adjust $scaling$ so as to achieve a reasonable acceptance probability. This proposal density is not symmetric, but no Jacobian needs to be computed.

Appendix F

The optimality conditions used in each of the five models we consider in are as follows.

a) Small scale New Keynesian models

$$
\begin{aligned}
y_t &= E_t y_{t+1} - \sigma\left(r_t - E_t\Delta p_{t+1} + E_t g_{t+1} - g_t\right) && (63)\\
y_t &= a_t + (1-\delta)\, n_t && (64)\\
mc_t &= w_t - p_t + n_t - y_t && (65)\\
mrs_t &= \frac{1}{\sigma} y_t + \gamma n_t - g_t && (66)\\
r_t &= \rho_r r_{t-1} + (1-\rho_t)\left[\gamma_\pi \Delta p_t + \gamma_y y_t\right] + z_t && (67)\\
w_t - p_t &= w_{t-1} - p_{t-1} + \delta w_t - \delta p_t && (68)\\
a_t &= \rho_a a_{t-1} + \epsilon_t^a && (69)\\
g_t &= \rho_g g_{t-1} + \epsilon_t^g && (70)\\
z_t &= \epsilon_t^z && (71)\\
\lambda_t &= \epsilon_t^\lambda && (72)\\
\Delta p_t &= \beta E_t \Delta p_{t+1} + \kappa_p\left(mc_t + \lambda_t\right) && (73)\\
w_t - p_t &= mrs_t && (74)\\
\Delta p_t &= \gamma_b \Delta p_{t-1} + \gamma_f E_t \Delta p_{t+1} + \kappa_p'\left(mc_t + \lambda_t\right) && (75)
\end{aligned}
$$

In the sticky wage model, the wage equation (74) is replaced by:

$$\Delta w_t = \beta E_t \Delta w_{t+1} + \kappa_w\left[mrs_t - (w_t - p_t)\right] \quad (76)$$

B) Medium scale New Keynesian model

$$\hat{y}_t = \frac{y+F}{y}\left[\alpha\hat{k}_t + (1-\alpha)\,\hat{L}_t\right] \tag{77}$$

$$\hat{\rho}_t = \hat{w}_t + \hat{L}_t - \hat{k}_t \tag{78}$$

$$\hat{s}_t = \alpha\hat{\rho}_t + (1-\alpha)\,\hat{w}_t \tag{79}$$

$$\hat{\pi}_t = \gamma_f E_t\hat{\pi}_{t+1} + \gamma_b\hat{\pi}_{t-1} + \kappa\hat{s}_t + \kappa\hat{\lambda}_{p,t} \tag{80}$$

$$\hat{\lambda}_t = \frac{h\beta e^{\gamma}}{(e^{\gamma}-h\beta)(e^{\gamma}-h)}E_t\hat{c}_{t+1} - \frac{e^{2\gamma}+h^2\beta}{(e^{\gamma}-h\beta)(e^{\gamma}-h)}\hat{c}_t + \frac{he^{\gamma}}{(e^{\gamma}-h\beta)(e^{\gamma}-h)}\hat{c}_{t-1} \tag{81}$$

$$+ \frac{h\beta e^{\gamma}\rho_z - he^{\gamma}}{(e^{\gamma}-h\beta)(e^{\gamma}-h)}\hat{z}_t + \frac{e^{\gamma}-h\beta\rho_b}{e^{\gamma}-h\beta}\hat{b}_t \tag{82}$$

$$\hat{\lambda}_t = \hat{R}_t + E_t\left(\hat{\lambda}_{t+1} - \hat{z}_{t+1} - \hat{\pi}_{t+1}\right) \tag{83}$$

$$\hat{\rho}_t = \chi\hat{u}_t \tag{84}$$

$$\hat{\phi}_t = (1-\delta)\,\beta e^{-\gamma}E_t\left(\hat{\phi}_{t+1} - \hat{z}_{t+1}\right) + \left(1-(1-\delta)\,\beta e^{-\gamma}\right)E_t\left[\hat{\lambda}_{t+1} - \hat{z}_{t+1} + \hat{\rho}_{t+1}\right] \tag{85}$$

$$\hat{\lambda}_t = \hat{\phi}_t + \hat{u}_t - e^{2\gamma}S''\left(\hat{\imath}_t - \hat{\imath}_{t-1} + \hat{z}_t\right) + \beta e^{2\gamma}S''E_t\left[\hat{\imath}_{t+1} - \hat{\imath}_t + \hat{z}_{t+1}\right] \tag{86}$$

$$\hat{k}_t = \hat{u}_t + \hat{\bar{k}}_{t-1} - \hat{z}_t \tag{87}$$

$$\hat{\bar{k}}_t = (1-\delta)\,e^{-\gamma}\left(\hat{\bar{k}}_{t-1} - \hat{z}_t\right) + \left(1-(1-\delta)\,e^{-\gamma}\right)(\hat{u}_t + \hat{\imath}_t) \tag{88}$$

$$\hat{w}_t = \frac{1}{1+\beta}\hat{w}_{t-1} + \frac{\beta}{1+\beta}E_t\hat{w}_{t+1} - \kappa_w\hat{g}_{w,t} + \tag{89}$$

$$+ \frac{\iota_w}{1+\beta}\hat{\pi}_{t-1} + \frac{1+\beta\iota_w}{1+\beta}\pi_t + \frac{\beta}{1+\beta}E_t\hat{\pi}_{t+1} + \tag{90}$$

$$+ \frac{\iota_w}{1+\beta}z_{t-1} - \frac{1+\beta\iota_w-\rho_z\beta}{1+\beta}z_t + \kappa_w\hat{\lambda}_{w,t} \tag{91}$$

$$\hat{g}_{w,t} = \hat{w}_t - \left(\nu\hat{L}_t + \hat{b}_t - \hat{\lambda}_t\right) \tag{92}$$

$$\hat{R}_t = \rho_R\hat{R}_{t-1} + (1-\rho_R)\left[\phi_\pi\hat{\pi}_t + \phi_X\left(\hat{x}_t - \hat{x}_t^*\right)\right] + \phi_{dX}\left[\left(\hat{x}_t - \hat{x}_{t-1}\right) - \left(\hat{x}_t^* - \hat{x}_{t-1}^*\right)\right] + \hat{\eta}_{mp,t} \tag{93}$$

$$\hat{x}_t = \hat{y}_t - \frac{\rho k}{y}\hat{u}_t \tag{94}$$

$$\frac{1}{g}\hat{y}_t = \frac{1}{g}\hat{g}_t + \frac{c}{y}\hat{c}_t + \frac{i}{y}\hat{\imath}_t + \frac{\rho k}{y}\hat{u}_t \tag{95}$$

c) model with search and matching frictions

$$\widehat{\lambda}_t = E_t\left\{\widehat{\lambda}_{t+1} + \widehat{R_t} + \widehat{\epsilon}_t^b - \widehat{\Pi}_{t+1}\right\} \tag{96}$$

$$\widehat{\lambda}_t = -\frac{\sigma}{1-\varrho}\left(\widehat{c}_t \varrho \widehat{c}_{t-1}\right) \tag{97}$$

$$\widehat{\Pi}_t = \gamma_f E_t\left\{\widehat{\Pi}_{t+1}\right\} + \gamma_b \pi_{t-1} + \kappa_p \widehat{mc}_t \tag{98}$$

$$\widehat{mc}_t = \widehat{x}_t^L \tag{99}$$

$$\widehat{m}_t = \xi \widehat{u}_t + (1-\xi)\,\widehat{v}_t \tag{100}$$

$$\widehat{n}_t = (1-\vartheta)\,\widehat{n}_{t-1} + \frac{m}{n}\widehat{m}_{t-1} \tag{101}$$

$$\widehat{n}_t = \frac{u}{1-u}\widehat{u}_t \tag{102}$$

$$\widehat{q}_t = \widehat{m}_t - \widehat{v}_t \tag{103}$$

$$\widehat{s}_t = \widehat{m}_t - \widehat{u}_t \tag{104}$$

$$\widehat{J^\star}_t + \widehat{\delta}_t^W = \widehat{\Delta}_t^\star + \widehat{\delta}_t^F - \frac{1}{1-\eta}\widehat{\eta}_t \tag{105}$$

$$\widehat{x}_t^L + \widehat{z}_t = (\alpha - 1)\,\widehat{h}_t = \widehat{w}_t \tag{106}$$

$$\widehat{w}_t = \gamma\left[\widehat{w}_{t-1} - \widehat{\Pi}_t\right] + (1-\gamma)\,\widehat{w}_t^\star \tag{107}$$

$$\begin{aligned}\widehat{\delta}_t^F &= [1-\beta(1-\vartheta)\gamma]\left[\frac{-\alpha}{1-\alpha}\widehat{w}_t^\star + \frac{1}{1-\alpha}\left(\widehat{x}_t^L + \widehat{z}_t\right)\right] \\ &+ \beta(1-\vartheta)\gamma E_t\left\{\frac{-\alpha}{1-\alpha}\left[\widehat{w}_t^\star - \widehat{w}_{t+1}^\star - \widehat{\Pi}_{t+1}\right] + \widehat{\delta}_{t+1}^F + \widehat{\lambda}_{t+1} - \widehat{\lambda}_t\right\}\end{aligned} \tag{108}$$

$$\begin{aligned}\delta^W \widehat{\delta}_t^W &= \frac{-\alpha}{1-\alpha} wh\left[\frac{-\alpha}{1-\alpha}\widehat{w}_t^\star + \frac{1}{1-\alpha}\left(\widehat{x}_t^L + \widehat{z}_t\right)\right] \\ &- \frac{-1}{1-\alpha} mrsh\left[\frac{(-1)(1+\varphi)}{1-\alpha}\widehat{w}_t^\star - \widehat{\lambda}_t + \frac{1+\varphi}{1-\alpha}\left(\widehat{x}_t^L + \widehat{z}_t\right)\right] \\ &+ \frac{\beta(1-\vartheta)\gamma}{1-\beta(1-\vartheta)\gamma}\left[\left(\frac{\alpha}{1-\alpha}\right)^2 wh - \frac{(1+\vartheta)}{(1-\alpha)^2} mrsh\right] E_t\left\{\widehat{w}_t^\star - \widehat{w}_{t+1}^\star - \widehat{\Pi}_{t+1}\right\} \\ &+ \beta(1-\vartheta)\gamma\delta^W E_t\left\{\widehat{\lambda}_{t+1} - \widehat{\lambda}_t + \widehat{\delta}_{t+1}^W\right\}\end{aligned} \tag{109}$$

$$\begin{aligned}J\widehat{J}_t^\star &= \frac{wh}{\alpha}\left[-\alpha\widehat{w}_t^\star + \widehat{x}_t^L + \widehat{z}_t\right] \\ &+ \frac{\beta(1-\vartheta)\gamma}{1-\beta(1-\vartheta)\gamma} wh E_t\left\{\widehat{w}_{t+1}^\star + \widehat{\Pi}_{t+1} - \widehat{w}_t^\star\right\} \\ &+ \beta(1-\vartheta) J E_t\left\{\widehat{\lambda}_{t+1} - \widehat{\lambda}_t + \widehat{J}_{t+1}^\star\right\}\end{aligned} \tag{110}$$

$$\begin{aligned}
\Delta\widehat{\Delta}^{\star}t &= wh\frac{1}{1-\alpha}\left[-\alpha\widehat{w}_t^{\star}+\widehat{x}_t^L+\widehat{z}_t\right] \\
&- \frac{1}{1+\varphi}mrsh\left[\frac{1+\varphi}{1-\alpha}\left(-\widehat{w}_t^{\star}+\widehat{x}_t^L+\widehat{z}_t\right)-\widehat{\lambda}_t\right] \\
&+ \frac{\beta\left(1-\vartheta\right)\gamma}{1-\beta\left(1-\vartheta\right)\gamma}\left[\frac{\alpha}{1-\alpha}wh-\frac{1}{1-\alpha}mrsh\right]E_t\left\{\widehat{w}_{t+1}^{\star}+\widehat{\Pi}_{t+1}-\widehat{w}_t^{\star}\right\} \\
&+ \frac{\beta\gamma s}{1-\beta\left(1-\vartheta\right)\gamma}\left[\frac{\alpha}{1-\alpha}wh-\frac{1}{1-\alpha}mrsh\right]E_t\left\{\widehat{w}_{t+1}^{\star}+\widehat{\Pi}_{t+1}-\widehat{w}_t^{\star}\right\} \\
&+ \left(1-\vartheta-s\right)\beta\Delta E_t\left\{\widehat{\lambda}_{t+1}-\widehat{\lambda}_t+\widehat{\Delta}_{t+1}^{\star}\right\} \\
&- \beta\Delta s\widehat{s}_t \qquad (111) \\
-\frac{\kappa}{q}\widehat{q}_t &= \frac{\beta\gamma}{1-\beta\left(1-\vartheta\right)\gamma}whE_t\left\{\widehat{w}_{t+1}^{\star}+\widehat{\Pi}_{t+1}-\widehat{w}_t^{\star}\right\} \\
&+ \beta JE_t\left\{\widehat{\lambda}_{t+1}-\widehat{\lambda}_t+\widehat{J}_{t+1}^{\star}\right\} \qquad (112) \\
y\widehat{y}_t &= c\widehat{c}_t+g\widehat{g}_t+\kappa v\widehat{v}_t+\Phi n\widehat{n}_t \qquad (113) \\
\widehat{y}_t &= \widehat{z}_t+\alpha\widehat{h}_t+\widehat{n}_t \qquad (114) \\
\widehat{\Psi}_t^L &= \frac{\frac{1-\alpha}{\alpha}wh}{\frac{1-\alpha}{\alpha}wh-\Phi}\left[\widehat{w}_t+\widehat{h}_t\right] \qquad (115) \\
\widehat{R}_t &= \gamma_R\widehat{R}_{t-1}+\left(1-\gamma_R\right)\left[\frac{\gamma_\pi}{12}\widehat{\Pi}_{t-1}^a+\frac{\gamma_y}{12}\widehat{y}_t\right]+\widehat{\epsilon}_t^{money} \qquad (116) \\
\widehat{\epsilon}_t^b &= \rho_b\widehat{\epsilon}_{t-1}^b+\xi_t^b, \quad \xi_t^b \overset{iid}{\sim} N\left(0,\sigma_b^2\right) \qquad (117) \\
\widehat{z}_t^b &= \rho_b\widehat{z}_{t-1}^b+\xi_t^z, \quad \xi_t^z \overset{iid}{\sim} N\left(0,\sigma_z^2\right) \qquad (118) \\
\widehat{g}_t^b &= \rho_b\widehat{g}_{t-1}^b+\xi_t^g, \quad \xi_t^g \overset{iid}{\sim} N\left(0,\sigma_g^2\right) \qquad (119) \\
\widehat{\epsilon}_t^{money} &= \xi_t^{money}, \quad \xi_t^{money} \overset{iid}{\sim} N\left(0,\sigma_{money}^2\right) \qquad (120)
\end{aligned}$$

$$\widehat{J}_t^{e,n}+\widehat{\delta}_t^{W,e,n}=\widehat{\Delta}_t^{e,n}+\widehat{\delta}_t^{F,e,n}-\frac{1}{1-\eta}\widehat{\eta}_t^{e,n} \qquad (121)$$

$$\widehat{x}_t^L+\widehat{z}_t+\left(\alpha-1\right)\widehat{h}_t^{e,n}=\widehat{w}_t^{e,n} \qquad (122)$$

$$\begin{aligned}
\delta^W\widehat{\delta}_t^{W,e,n} =& \frac{-\alpha}{1-\alpha}wh\left[\frac{-\alpha}{1-\alpha}\widehat{w}_t^{e,n}+\frac{1}{1-\alpha}\left(\widehat{x}_t^L+\widehat{z}_t\right)\right] \\
&- \frac{-1}{1-\alpha}mrsh\left[\frac{\left(-1\right)\left(1+\varphi\right)}{1-\alpha}\widehat{w}_t^{e,n}-\widehat{\lambda}_t+\frac{1+\varphi}{1-\alpha}\left(\widehat{x}_t^L+\widehat{z}_t\right)\right]
\end{aligned} \qquad (123)$$

$$\begin{aligned}
J\widehat{J}_t^{e,n} =& \frac{wh}{\alpha}\left[-\alpha\widehat{w}_t^{e,n}+\widehat{x}_t^L+\widehat{z}_t\right] \\
&+ \beta\left(1-\vartheta\right)JE_t\left\{\widehat{\lambda}_{t+1}-\widehat{\lambda}_t+\widehat{J}_{t+1}^e\right\}
\end{aligned} \qquad (124)$$

$$\begin{aligned}\Delta\widehat{\Delta}_t^{e,n} =& wh\frac{1}{1-\alpha}\left[-\alpha\widehat{w}_t^{e,n}+\widehat{x}_t^L+\widehat{z}_t\right]\\ &-\frac{1}{1+\varphi}mrsh\left[\frac{1+\varphi}{1-\alpha}\left(-\widehat{w}_t^{e,n}+\widehat{x}_t^L+\widehat{z}_t\right)-\widehat{\lambda}_t\right]\\ &+(1-\vartheta-s)\,\beta\Delta E_t\left\{\widehat{\lambda}_{t+1}-\widehat{\lambda}_t+\widehat{\Delta}_{t+1}^e\right\}\\ &-\beta\Delta s\widehat{s}_t\end{aligned} \tag{125}$$

$$-\frac{\kappa}{q}\widehat{q}_t=\beta JE_t\left\{\widehat{\lambda}_{t+1}-\widehat{\lambda}_t+\widehat{J}_{t+1}^n\right\} \tag{126}$$

$$\widehat{w}_t=(1-\vartheta)\,\widehat{w}_t^e+\vartheta\widehat{w}_t^n \tag{127}$$

d) Model with financial frictions

$$\begin{aligned}
y_t &= \frac{C}{Y}c_t+\frac{I}{Y}i_t+\frac{G}{Y}g_t+\frac{C^e}{Y}c_t^e+\cdots+\phi_t^y && (128)\\
c_t &= -r_{t+1}+E_t\left\{c_{t+1}\right\} && (129)\\
c_t^e &= n_{t+1}+\cdots+\phi_t^{c^e} && (130)\\
E_t\left\{r_{t+1}^k\right\} &= r_{t+1}-\nu\left[n_{t+1}-(q_t+k_{t+1})\right] && (131)\\
r_{t+1}^k &= (1-\epsilon)\left(y_{t+1}-k_{t+1}-x_{t-1}\right)+\epsilon q_{t+1}-q_t && (132)\\
q_t &= \varphi\left(i_t-k_t\right) && (133)\\
y_t &= a_t+\alpha k_t+(1-\alpha)\,\Omega h_t && (134)\\
y_t &= h_t+x_t+c_t+\eta^{-1}h_t && (135)\\
\pi_t &= E_t\left\{\kappa_p\left(-x_t\right)+\gamma_f\pi_{t+1}+\gamma_b\pi_{t-1}\right\} && (136)\\
k_{t+1} &= \delta i_t+(1-\delta)\,k_t && (137)\\
n_{t+1} &= \frac{\gamma RK}{N}\left(r_t^k-r_t\right)+r_t+n_t+\cdots+\phi_t^n && (138)\\
r_t^n &= \rho r_{t-1}^n+(1-\rho)\varsigma\pi_{t-1}+\epsilon_t^{tn} && (139)\\
g_t &= \rho_g g_{t-1}+\epsilon_t^g && (140)\\
a_t &= \rho_a a_{t-1}+\epsilon_t^a && (141)
\end{aligned}$$