Working Paper Series

A Composite Likelihood Approach for Dynamic Structural Models

WP 18-12

Fabio Canova
BI Norwegian Business School,
CAMP, and CEPR

Christian Matthes
Federal Reserve Bank of Richmond

FEDERAL RESERVE BANK OF RICHMOND®
Richmond ▪ Baltimore ▪ Charlotte

A composite likelihood approach for dynamic structural models

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

Working Paper No. 18-12

Abstract

We describe how to use the composite likelihood to ameliorate estimation, computational, and inferential problems in dynamic stochastic general equilibrium models. We present a number of situations where the methodology has the potential to resolve well-known problems. In each case we consider, we provide an example to illustrate how the approach works and its properties in practice.

Key words: Dynamic structural models, composite likelihood, identification, singularity, large scale models, panel data.

JEL Classification: C10, E27, E32.

*We thank Bruce Hansen, Mark Watson, Ulrich Mueller, Barbara Rossi, Refet Gurkenyak, Ferre de Graeve, Geert Dhaene, Mikkel Plagborg-Moller, and the participants at the 2016 ESEM invited session on 'New developments in the analysis of Macroeconomic data', the 2016 ESOBE conference, the 3rd Henan University -Infer macroeconomic workshop, the SBIES conference, the Barcelona GSE forum and the EABCN conference Time varying models for monetary policy and financial stability, EC^2 conference, Amsterdam, and seminars at Singapore National University, Katholic University Leuven, the Riksbank, the Bank of Finland 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 use them for policy purposes. Until a decade ago, most analyses were performed using parameters formally or informally calibrated. Nowadays, it is more common to conduct inference using parameters estimated with classical or Bayesian full information likelihood methods; see Andreasen et al. (2014) for an exception.

Estimation of DSGE models is difficult, making estimation results whimsical. There are population and sample identification problems, see e.g., Canova and Sala (2009), Komunjer and Ng (2011), Qu and Tkachenko (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); informational deficiencies (models are constructed to explain only a portion of the data), see Boivin and Giannoni (2006), Canova (2014), or Pagan (2016), that restrict the class of models for which the likelihood can be computed. Computational complications, due, for example, to the presence of latent variables that require the computationally challenging integration of the joint likelihood of the endogenous variables of the model, and numerical difficulties are also well-known. Both become particularly acute when the model is of large scale or when the data are short or of poor quality.

Inference in estimated DSGE models is also troublesome. Standard frequentist asymptotic theory needs regularity conditions, which are often violated in practice. 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), assuming prior independence results in an overall prior does not fully reflect priors beliefs held by researchers. Perhaps more importantly, standard likelihood-based inference is conditional on the estimated model being correctly specified.

Policymakers are keenly aware of both estimation and inferential problems and, when choosing policy actions, tend to informally pool results obtained from different models. Furthermore, when there are structural instabilities in the data-generating process (DGP), it may be attractive to use a number of models to robustify counterfactual exercises and improve forecasting performance, see e.g., Aiolfi et al. (2010).

This paper is concerned with the estimation problems that researchers working with DSGE models face. We propose a method that may help to solve some of the above-mentioned difficulties and automatically provide estimates of the parameters that formally combine the information present in different models using shrinkage-type estimators. The approach we suggest is based on the *composite likelihood*, a limited information objective function, well-known in the statistical literature but very sparsely used in economics (Engle et al., 2008; Qu, 2015; Eisenstein et al., 2017).

In the original formulation of Besag (1974) and Lindsay (1980), the composite likelihood is constructed combining marginal or conditional likelihoods of the true DGP and employed because the likelihood of the full model is computationally intractable or features unmanageable integrals. When marginals or conditionals are used, the com-

posite likelihood estimator is consistent and asymptotic normal, as either the number of observations or the number of composite likelihood components grows. 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 structural or statistical models, which are not necessarily marginal or conditional partitions of the DGP. Thus, standard composite likelihood properties do not necessarily apply. Nevertheless, it is still possible to conduct formal inference and produce estimators with desirable properties.

We describe how to construct and use the composite likelihood in a large class of situations relevant to macroeconomists. We briefly discuss asymptotic inference in our non-standard setup and how such an objective function can be treated as a quasi-likelihood to conduct Bayesian inference. Kim (2002), Chernozukov and Hong (2003) and Marin (2011) have used similar ideas in different contexts. However, to the best of our knowledge, no author has constructed composite Bayesian estimators and used the setup to analyze structural macroeconometric models as we do. We provide a sequential, adaptive learning interpretation to our posterior estimators and discuss the differences with standard Bayesian estimators and to other combination devices present in the literature.

We show how the approach (in either its classical or Bayesian version) can be used to potentially address the estimation and inferential problems noted in this introduction. We present examples indicating that the composite likelihood constructed using the information present in distinct models helps 1) to ameliorate population and sample identification problems, 2) to solve singularity problems, 3) to produce more stable estimates of the parameters of large scale structural models, 4) to robustly estimate the parameters appearing in multiple models and 5) to combine information coming from different sources, frequencies, and levels of aggregation.

The rest of the paper is organized as follows. The next section presents the traditional composite likelihood approach and introduces our setup. Section 3 discusses quasi-Bayesian estimation and inference. Section 4 presents a number of examples highlighting how the methodology can address standard estimation problems. Section 5 concludes. The appendices provide details for arguments discussed in the text and the equations of the models used in our examples.

2 The composite likelihood

The original composite likelihood formulation has been suggested to deal with situations where the likelihood of a model is either difficult to construct because of latent variables or hard to manipulate because the covariance matrix of the observables is nearly singular. In some applications, see Engle et al., (2008), the likelihood is conceptually tractable, but the dimensionality of the parameter space makes maximum likelihood computations complex and unappealing.

In all these situations, it might be preferable to use an objective function which has smaller informational content than the likelihood but is easier to work with. One such objective function, originally proposed by Linsday, 1980, is a weighted average of marginal or conditional distributions of submodels ('events' in the terminology used by this literature).

Suppose a known DGP produces a density $F(y_t, \psi)$ for an $m \times 1$ vector of observables y_t, where ψ is a $q \times 1$ vector. Partition $\psi = [\theta, \eta]$ where, by convention, θ is the vector of parameters estimated by composite likelihood methods, and η is a vector of model-specific nuisance parameters. Let $\{A_i, i = 1, ...K\}$ be a set of marginal or conditional events of y_t, and let $f(y_{it} \in A_i, \theta, \eta_i)$ be the subdensities of $F(y_t, \psi)$ corresponding to these events [1]. Each A_i defines a submodel, with implications for a subvector y_{it} of length T_i and is associated with the vector $\psi_i = [\theta, \eta_i]'$, where η_i are (nuisance) event specific parameters. Let $\phi = (\theta, \eta_1, \ldots \eta_K)$. Given a vector of weights ω_i, the composite likelihood is

$$CL(\phi, y_{1t}, \ldots, y_{KT}) = \Pi_{i=1}^{K} \quad f(y_{it} \in A_i, \theta, \eta_i)^{\omega_i}. \tag{1}$$

Clearly, $CL(\phi, y_{1t}, \ldots, y_{KT})$ is not a likelihood function. Nevertheless, if $y_{[1,t]} = (y_1, \ldots, y_t)$ is an independent sample from $F(y_t, \psi)$ and ω_i are fixed quantities, ϕ_{CL}, the maximum composite likelihood estimator of ϕ, satisfies $\phi_{CL} \xrightarrow{P} \phi$ and

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

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

$$\begin{aligned}
G &= HJ^{-1}H; \text{ Godambe information} && (3)\\
J &\equiv var_\phi u(\phi, y_{[1,t]}); \quad \text{Variability matrix} && (4)\\
H &\equiv -E_\phi[\bigtriangledown_\theta u(\theta, \eta_1, \ldots, \eta_K, y_{[1,t]})]; \quad \text{Sensitivity matrix} && (5)\\
u(\phi, y|\omega_1, \ldots, \omega_K) &= \sum_i \omega_i \bigtriangledown_\phi l_i(\phi, y_{[1,t]}); \text{ Composite scores} && (6)
\end{aligned}$$

and $\bigtriangledown_\phi l_i(\phi, y_{[1,t]})$ denotes the score associated with the log of $f(y_{it} \in A_i, \theta, \eta_i)$. Thus, θ_{CL} is constructed using the information present in all submodels, with ω_i determining how important each model is.

Note the composite likelihood ignores the potential dependence across A_i, i.e., submodels may feature common equations, and the fact that y_{it} may not be mutually exclusive across i, i.e., the same variable (say, the inflation rate) may appear in the observables of each submodel [2].

Consistency obtains because each element in (2) is an unbiased estimating function and a weighted avarage of unbiased estimating functions is unbiased. Asymptotic

[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.

[2] If T is fixed but the different A_i are independent, then (2) still holds when $K \to \infty$,and a standard Newey-West correction to $J(\theta)$ can be used if $y_{[1,t]}$ is not an independent sample.

normality holds because the sampling distribution of the maximum likelihood estimator of each submodel can be approximated quadratically around the mode. Note that the asymptotic covariance matrix is $HJ^{-1}H$ and that in general $H \neq J$. Since it differs from the Fisher information matrix, I, θ_{CL} is not efficient.

The choice of weights is typically left to the investigator, and, for example, one may choose ω_i to improve efficiency. Optimal weights can be obtained by 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 (Pauli et al., 2011). Alternatively, one could set $\omega_i = \frac{1}{K}, \forall i$, to minimize the researcher input; or use a data-based approach to their selection. For example, one could set $\omega_i = \frac{\exp(\chi_i)}{1+\sum_{i=1}^{K-1}\exp(\chi_i)}$, where χ_i is a function of some statistics of past data $\chi_i = f_i(Y_{1,[1:\tau]}, ..., Y_{K,[1:\tau]})$. If these statistics are updated over time, ω_i could also be made time varying. There is a large forecasting literature (see e.g. Aiolfi et al., 2010) that 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 computationally demanding. In this case, a two-step estimation approach is possible where η_i is separately estimated for each $\log f(y_{it} \in A_i, \theta, \eta_i)$ 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 for θ_{CL} in this case need to be adjusted to account for the fact that η_i is estimated.

2.1 A composite DSGE setup

Our setup differs from the traditional one in several respects. First, we treat the DGP as unknown. There are many reasons for such a choice. For example, we may not have enough information to construct $F(y_t|\psi)$; we could write a VAR representation for y_t but not the structural model that generated it; or we do not have an analytic expression for $F(y_t|\psi)$, but only the first few terms of its Taylor expansion. Another reason for treating $F(y_t|\psi)$ as unknown is that the dimension of y_t may be large and a researcher may have an idea of how portions of y_t could have been generated but not know yet how to link them in a coherent way.

Second, $f(y_{it} \in A_i, \theta, \eta_i)$ are approximations to the DGP and thus are neither marginal nor conditional representations. Formally, the quality of the approximation of $f(y_{it} \in A_i, \theta, \eta_i)$ is measured by the distance of $G_i(\phi)^{-1}I_i(\phi)$ from the identity matrix - when the approximation is exact, $G_i(\phi)^{-1} = I_i(\phi)$. To be concrete, in one leading example we have in mind, A_i are different structural models, e.g., 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 is the vector of the structural parameters common to all models, e.g. the risk-aversion coefficient, or the Frisch elasticity, while η_i could be other structural parameters of the models, e.g. a LTV ratio, a Calvo parameter, or reduced-form mongrels used to approximate features of the DGP,

e.g., the parameter regulating habit in consumption. In another leading example, we have in mind $F(y_t|\psi)$ 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 represents either the order of the approximation employed or an indicator function describing which parameter is allowed to change.

In all these situations, different models are treated as approximations because they disregard aspects of the DGP; take short cuts to modeling the complexities of the DGP; or condition on features which may be present or absent from the DGP.

A final case of interest is one where $f(y_{it} \in A_i, \theta, \eta_i)$ represents 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-equation 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, $F(y_t, \psi)$ 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 are used; ii) data at a particular aggregation level (e.g. firm, industry, regional, etc.) are employed. Alternatively, $F(y_t, \psi)$ 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-WWII, etc.).

A third important difference from the traditional setup is that models we consider need not be statistically compatible with each other. Compatibility implies that asymptotically, $\theta_{i,ML}$ converges to the same value for each i. This is easy to show when $f(y_{it} \in A_i, \theta, \eta_i)$ are marginals or conditionals. Because of this potential incompatibility, the estimators for θ we construct need not enjoy the standard properties of composite likelihood estimators.

Researchers working with DSGE models are generally free to choose what goes in θ and in η_i. This allows substantial flexibility because even though some parameters might be common to all models, researchers might prefer not to estimate a common value. For example, when using different statistical models, and when A_i represents different levels of data aggregation, one could make the parameter regulating the complementarity of government expenditure and private consumption common, while making the parameters regulating the process for goverment expenditure submodel specific.

Because all models we consider are approximations to the DGP, likelihood estimators obtained in each of them will be inconsistent and, thus, the composite likelihood estimator will be inconsistent. Following White (1982) and Domowitz and White (1982), one can show that, under regularity conditions, $\phi_{i,ML}$, the likelihood estimator in model A_i, converge, as $T \to \infty$ to the pseudo-parameter vector, ϕ_0, which minimizes the Kullback-Leibler (KL) distance from the true DGP and that $\sqrt{T}(\phi_{i,ML} - \phi_0) \sim N(0, G_i^{-1})$, where G_i is the Godambe information matrix for model

i.

The weighting scheme that the composite likelihood employs defines a density for a different misspecified model (the weighted average of the K submodels). When the weights w_i are constant, ϕ_{CL} approaches asymptotically $\phi_{0,CL}$, the minimizer of the KL distance between the density of the combination of models and the DGP. Note that $\phi_{0,CL}$ is not, in general, a weighted average of $\phi_{0,i}$, because models are not necessarily independent. Mimicking the argument used for each model i, one can show that $\sqrt{T}(\phi_{CL} - \phi_{0,CL}) \sim N(0, G^{-1})$ where G is the Godambe information computed using the composite likelihood (see Canova and Matthes (2017) for details).

3 Quasi-Bayesian estimation

Because we are interested in obtaining a small sample distribution of ϕ, rather than its asymptotic approximation, and in treating ω as a random variable with a prior distribution (to be interpreted as the investigator prior assessment of the likelihood of model i), we estimate $(\theta, \omega_i, \eta_i, i = 1, \dots, K)$ by quasi-Bayesian methods. Note that what we are after is different from what finite mixture models (see e.g., Waggoner and Zha, 2011) or Bayesian model average (BMA) exercises do. In BMA, each model is estimated separately and their predictions combined using posterior weights; in our setup, all models are jointly estimated and the predictions can be combined, if that is of interest. In finite mixture models, $y_{1t}, = \dots = y_{Kt}$ and $T_1 = \dots = T_K$ and the (time-varying) weight determines at each t how important is y_t for the estimations of the parameters of model i. In our setup $y_{1t} \neq \dots \neq y_{Kt}$ and $T_1 \neq \dots \neq T_K$ and, as shown below, parameter information is adaptively and sequentially updated as we add models to the composite pool.

For each i, the prior for the parameters is of the form

$$p(\theta, \eta_i) = p(\theta)p(\eta_i|\theta). \tag{7}$$

In the spirit of Del negro and Schorfheide (2008), We allow the prior for η_i to depend on θ, which is advisable if the composite pools features distinct structural models and, a priori, we want these models to be on equal ground when matching certain statistics of the data. If $p(\omega) \equiv p(\omega_1, \dots \omega_K)$ is the prior for the vector of weights, the composite posterior kernel is:

$$\begin{aligned} \tilde{p}(\theta, \eta_1, ...\eta_K, \omega_1, \dots, \omega_K|Y_{1,t_1}, \dots, Y_{k,T_k}) &= \\ \mathcal{L}(\theta, \eta_1|Y_{1,T_1})^{\omega_1} p(\theta, \eta_1)^{\omega_1} \dots \mathcal{L}(\theta, \eta_K|Y_{K,T_K})^{\omega_K} p(\theta, \eta_K)^{\omega_K} p(\omega) &= \\ \Pi_i \mathcal{L}(\theta, \eta_i|Y_{i,T_i})^{\omega_i} p(\eta_i|\theta)^{\omega_i} p(\theta) p(\omega), & \end{aligned} \tag{8}$$

which can be used to obtain posteriors for (ϕ, ω) , as in Kim (2002) or Chernozukov and Hong (2003). The appendix presents regularity conditions needed for standard MCMC techniques to apply; the algorithm we employ to draw posterior sequences for the parameters; and the adjustments one may want to implement for the posterior

percentiles to take into account the fact y_{it} may not be mutually exclusive across i (along the lines of Mueller (2013), Qu (2015), or Ribatet et al (2012)) or that the models may be more generally misspecified.

3.1 A sequential learning interpretation

It is easy to give a sequential, adaptive learning interpretation to the composite posterior kernel (8) and to the Bayesian estimators for θ one obtains. For the sake of illustration, suppose that ω_i is fixed and K=2. The composite posterior kernel $\check{p}$ is

$$\begin{aligned} \check{p}(\theta, \eta_{1....}\eta_2|Y_{1,T_1}, Y_{2,T_2}) &= \\ \mathcal{L}(Y_{1,T_1}|\theta,\eta_1)^{\omega_1} p(\eta_1|\theta)^{\omega_1} p(\eta_2|Y_{2,T_2},\theta)^{\omega_2} \{[p(\theta|Y_{2,T_2}) ML(Y_{2,T_2})]^{\omega_2} p(\theta)^{\omega_1}\} & \quad (9) \end{aligned}$$

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

As (9) makes clear, the posterior kernel can be obtained in two stages. In the first stage, the prior for ψ_2 and the likelihood for model 2 are used to construct $\mathrm{p}(\theta|Y_{2,T_2})$. This conditional posterior, weighted by the marginal likelihood of the model 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. Then model 2 fits Y_{2,T_2} well. If $\omega_1 = \omega_2$, the prior for model 1 will more heavily reflect $p(\theta|Y_{2,T_2})$ relative to the initial prior $p(\theta)$. On the other hand, if $ML(Y_{2,T_2})$ is low, $p(\theta|Y_{2,T_2})$ has low weight relative to $p(\theta)$ when setting up the prior for model 1. In general, the prior that θ receives in each stage of the learning process depends on the relative weights assigned to the current and to all previous models and on their relative fit for θ. Thus, a composite Bayesian approach to estimation can be interpreted as an adaptive sequential learning process where the information contained in models whose density poorly relates to the observables is appropriately downweighted.

Note that the prior for stage 2 is not the posterior for stage 1 as in a standard Bayesian setup but rather a weighted average of the initial prior and of the posterior obtained at stage 1, where the latter is discounted by the fit at that stage. This is why the approach is adaptive. Also, even though only Y_{2,T_2} contains information for η_2, its posterior may be updated when using Y_{1,T_1} since the posterior for θ sequentially changes. Also, since Y_{2,T_2} does not contain information for η_1, $p(\eta_1|\theta)$ will be unchanged after estimation is performed with model 2.

Finally, note that while with a composite posterior there is an automatic discounting whenever a model does not fit the data well, regardless of whether ω_i is treated as a parameter or a random variable. Del Negro et al. (2016) have shown that a finite mixture have this property only if ω is random.

4 Addressing estimation, computational, and inferential problems

This section shows how the composite likelihood may help to deal with standard problems encountered in the estimation of DSGE models. While the improvements we discuss are specific to the models and the parameterization used, the insights they provide go beyond the model economies we deal with. When possible, we explicitly state the conditions under which the composite likelihood provides a "better" objective function than the likelihood of a single model in the sense that $det(G(\phi)^{-1}I(\phi)) < min_i det(G_i(phi)^{-1}I_i(\phi))$.

The first example discusses how small sample identification problems can be resolved by using the composite likelihood constructed using different structural models. The intuition this example provides applies also to situations when different statistical models are used to compute the composite likelihood or when the same model is used with different samples of data. The second example demonstrates how the approach can ameliorate population identification problems. The third example deals with singularity issues; the fourth with the problem of estimating the parameters of a large-scale structural model. The fifth example demonstrates how to robustly estimate structural parameters appearing in different models. The last example shows how the composite likelihood may be used to partially pool the information contained in panels of data with potentially heterogeneous dynamics.

4.1 Reducing sample identification problems

In macroeconomics it is common to work with relatively small samples of time series. Long data series are generally unavailable and, when they exist, definitional changes or structural breaks make it unwise to use the full sample for estimation purposes. In addition, the phenomena one is interested in characterizing (say, the zero lower bound on interest rates) may be present only in the most recent portion of the sample. In this section, we show how the composite likelihood could help reduce the severity of small sample problems.

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

$$y_{At} = \rho_A y_{At-1} + \sigma_A e_t \tag{10}$$

$$y_{Bt} = \rho_B y_{Bt-1} + \sigma_B u_t \tag{11}$$

where e_t and u_t are both iid (0,I). Suppose that $\rho_B = \delta\rho_A, \sigma_B = \gamma\sigma_A$, y_{At} and y_{Bt} are scalars, that we have $T_A(T_B)$ observations on y_{At} (y_{Bt}) with T_A small, and that we are interested in estimating $\theta = (\rho_A, \sigma_A)$. For the sake of the presentation, let δ, γ be known.

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

$$\log L_A \propto -T_A \log \sigma_A - \frac{1}{2\sigma_A^2} \sum_{t=1}^{T_A} (y_{At} - \rho_A y_{At-1})^2 \tag{12}$$

$$\log L_B \propto -T_B \log(\sigma_A \gamma) - \frac{1}{2\sigma_A^2 \gamma^2} \sum_{t=1}^{T_B} (y_{Bt} - \rho_A \delta y_{Bt-1})^2 \tag{13}$$

which can be easily maximized with respect to ρ_A, σ_A. For $0 < \omega < 1$, the log composite likelihood is

$$\log CL = \omega \log L_A + (1-\omega) \log L_B \tag{14}$$

.

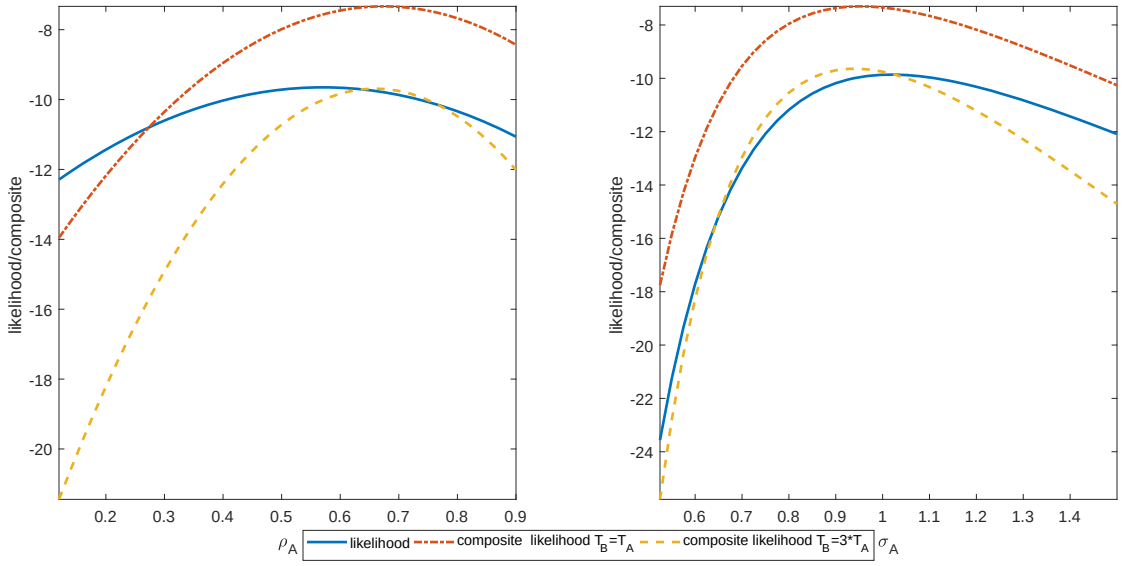

Figure 1: Likelihood and composite likelihood, small T.

We set $\rho_A = 0.7, \sigma_A = 1.0$, $\delta = 1.2$, $\gamma = 0.8$ $T_A = 20, T_B = 20$ (or $T_B = 60$), and plot in Figure 1 the univariate contours in the (ρ_A, σ_A) dimensions, when (12) and (14) are used. In the latter case, we set $\omega = 0.7$. Figure 1 highlights two facts. First, the composite likelihood has more curvature then the likelihood constructed using Y_{tA} only, even when $T_A = T_B$. Second, the mode of the composite likelihood is closer to the true values. Note that, as T_B increases ($T_B = 60$), the composite likelihood becomes more bell-shaped around the true value and almost symmetric in shape.

As we show in section 4.5, differences between the likelihood constructed using y_{At} and the composite likelihood have to do with three quantities $\zeta_1 = \frac{1-\omega}{\omega}\frac{\delta}{\gamma^2}$, $\zeta_2 = \frac{1-\omega}{\omega}\frac{\delta^2}{\gamma^2} = \zeta_1 \delta$, and $\xi = (T_A + T_B \frac{1-\omega}{\omega\gamma^2})^{-1}$. ζ_1 and ζ_2 control the relative shape of the

composite likelihood, while ξ, the effective sample size, controls both the relative height and the relative shape of the composite likelihood. Since all three quantities depend on ω, γ, δ, these parameters regulate the amount of information that y_{Bt} provides for ρ_A, σ_A. For example, if $\omega = 0.5$ and $\gamma = 1.0$, the effective sample size used to construct the composite likelihood is $T_A + T_B$, making this function higher than the likelihood constructed using T_A alone. In addition, the higher is γ, the less informative is y_{Bt} for the estimation of ρ_A, σ_A - model B provides information that twists the composite likelihood away from the true value. Similarly, the lower is δ, the lower will be the informational content of y_{Bt} for the parameters of interest. Thus, the composite likelihood gives importance to y_{Bt} if it is generated by a model with higher persistence and lower standard deviation than the model for Y_{At}. Such a scheme is reasonable since the higher the serial correlation, the more important low frequency information is; and the lower the standard deviation is, the lower the noise in y_{Bt} is.

This discussion highlights an interesting trade-off that the composite likelihood exploits: y_{Bt} may give information for the parameters of interest, but may also twist its shape away from the true values In this example, better local identification could be attained if (y_{At}, y_{Bt}) are jointly used in estimation whenever ω, γ, and T_B are such that the effective sample size $\xi > T_A$ and ζ_1, ζ_2 are different from zero. If γ is small, that is, if y_{tB} is less volatile than y_{tA}, or if ω is not too large, that is, if the degree of trust a researcher has in model B is not negligible, the log composite likelihood (14) will be more peaked around the mode than the likelihood (12).

So far models A and B are different structural models. However, the same argument is applicable when A and B are two statistical models or when they are the same structural model and y_{At} and y_{Bt} represent the same time series in different samples. In the first case, the use of information coming from different time series may make the composite likelihood more peaked around the true value than the likelihood of each model, much in the same spirit as a data-rich approach to estimation may provide better information about structural parameters (see e.g. Boivin and Giannoni, 2006). In the second case, the use of, say, pre-break data may help to sharpen structural inference, even if the pre-break data pulls the composite likelihood away from the current sample likelihood, as long as the weights are appropriately chosen. Baumeister and Hamilton (2015) suggested a procedure to reduce the information contained in earlier subsamples that mimics a composite likelihood estimator in this situation.

We also would like to stress that T_A and T_B may be not only of different lengths but also recorded at different frequencies (e.g., coming from a quarterly and an annual model). The composite likelihood is a flexible tool that exploits the available information to reduce small sample (local) identification problems.

4.2 Ameliorating population identification problems

This subsection presents an example where estimation is difficult because some parameters are underidentified and others weakly identified *in population;* it also shows how the use of a composite likelihood approach can help remedy these problems.

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

$$\begin{aligned} R_{At} &= \tau E_t \pi_{At+1} + e_{1t} && (15)\\ y_{At} &= \delta E_t y_{At+1} - \sigma(R_{At} - E_t \pi_{At+1}) + e_{2t} && (16)\\ \pi_{At} &= \beta E_t \pi_{At+1} + \gamma y_{At} + e_{3t} && (17) \end{aligned}$$

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 determinate solution of (15)-(17) is

$$\begin{bmatrix} R_{At} \\ y_{At} \\ \pi_{At} \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 \\ \sigma & 1 & 0 \\ \sigma\gamma & \sigma & 1 \end{bmatrix} \begin{bmatrix} e_{1t} \\ e_{2t} \\ e_{3t} \end{bmatrix} \equiv A e_t. \quad (18)$$

Clearly, β is underidentified - it disappears from (18) - 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 (18) if σ is sufficiently small, making the likelihood flat in the γ dimension.

Population underidentification of β implies, for example, that when (15)-(17) is the data generating process, applied investigators can not distinguish if the Philips curve is forward looking or not, nor can they measure the degree of forward lookingness, even when $T \to \infty$. Weak population identification of γ implies that it is hard to pin down the effects of the output gap (marginal costs) on inflation, regardless of the magnitude of the 'true' slope of the Phillips curve. Problems of this type are common in DSGE models (see Canova and Sala, 2009) and make estimation results whimsical.

Suppose we have available another model (call it, B) usable for inference. For example, consider a single-equation Phillips curve with exogenous marginal costs:

$$\begin{aligned} \pi_{Bt} &= \beta E_t \pi_{Bt+1} + \gamma y_{Bt} + u_{2t} && (19)\\ y_{Bt} &= \rho y_{Bt-1} + u_{1t} && (20) \end{aligned}$$

where $\rho > 0$ measures the persistence of the output gap (marginal costs). Note that (19) has the same format as (17), so that β and γ have the same economic interpretation but the process generating y_t is different. Suppose that model A is considered more trustworthy and an applied investigator acknowledges this by setting $\omega >> 1 - \omega$. By repeatedly substituting forward and letting ℓ be the lag operator, the solution to (19)-(20) is

$$\begin{bmatrix} (1 - \rho\ell) y_{Bt} \\ (1 - \rho\ell) \pi_{Bt} \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ \frac{\gamma}{1-\beta\rho} & 1 - \rho\ell \end{bmatrix} \begin{bmatrix} u_{1t} \\ u_{2t} \end{bmatrix}. \quad (21)$$

Clearly, unless the process for the output gap is iid ($\rho = 0$), the log-likelihood of model B has information about β. Thus, one would be able to identify (and estimate) β from the composite likelihood but not from the likelihood of model A, avoiding observational equivalence problems. 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 $\rho \neq 0$. Hence,

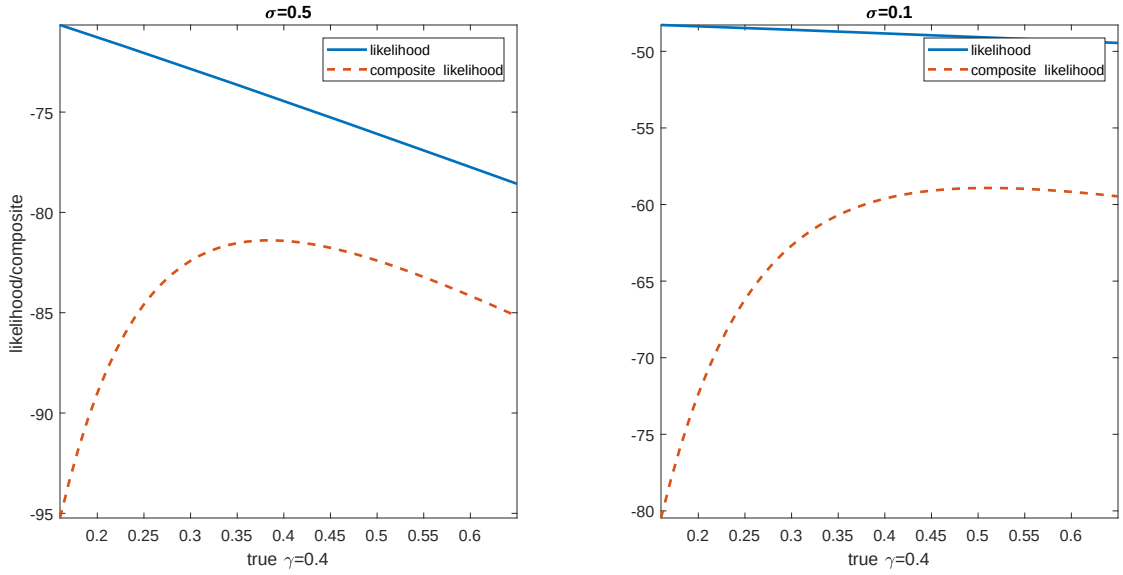

Figure 2: Likelihood and composite likelihood, weak identification.

small variations γ may lead to sufficiently large variations in the decision rule (21) and thus in the composite likelihood, even when $1-\omega$ is small.

We illustrate the argument in Figure 2. We plot the likelihood of model A and the composite likelihood as function of γ when $\sigma = 0.5$ or $\sigma = 0.1$. The DGP has $\gamma = 0.4$, $\beta = 0.99, \rho = 0.8$, and we present the shape of the composite likelihood when $\omega = 0.85$. As discussed, the likelihood of model A is flat around the true value of γ when σ is small, and adding information from the second model helps to improve the identification of γ. The same outcome obtains when $\sigma = 0.5$ as the likelihood constructed from y_{At} is not quadratic in γ.

It should be clear that the argument we make here is independent of the size of the effective sample 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 we have implicitly assumed that the variances of (e_{2t}, e_{3t}) and of (u_{1t}, u_{2t}) are of the same order of magnitude (in Figure 2, they are all equal to 1). When this is not the case, two distinct forces are at play: the relative noise present in the two models is weighted against the relative information present in the decision rules.

It goes without saying that adding models with Philips curves that are non-comparable to those of model A is unlikely to reduce population identification problems. In other words, if model B data have been generated from a mechanism that is different than that of model A or, if the mechanism is the same but the values for β and γ are very different, the biases introduced using model B data may be large relative to the improved curvature. Hence, population identification improvements can be obtained only after carefully examining the shape of the likelihood of the additional model(s)

one may want to consider.

In sum, these two subsection have shown that the composite likelihood may improve parameter identification when the sample is short or when parameters are weakly identified in population. This happens when the additional data used in the composite likelihood adds information to the likelihood of model A for the parameters of interest. This additional information is easily measurable in practice: it will be reflected in the height and the curvature of the composite likelihood, which will be more bell shaped and symmetric than the likelihood of the baseline model. We recommend applied investigators to plot likelihood and composite likelihoods as we have done in Figures 1 and 2 as a routine practice. This will help them to understand whether an additional model should be used in the investigation.

4.3 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 of reduced rank and the likelihood function can not be constructed and optimized. There are many approaches to get around this problem. One could select a subvector of the observables matching the dimension of the shock vector informally (see Guerron Quintana, 2010) or formally (see Canova et al., 2014) and use the log-likelihood of this subvector for estimation. Alternatively, one could add measurement errors to some or all the observables - so as to make the number of shocks (structural and measurement) larger or equal to the number observables (see Ireland, 2004). One could also increase the number of structural shocks, for example, by transforming parameters into disturbances (the discount factor becomes a preference shock, etc.) until shocks and endogenous variables match.

An alternative way to deal with singularity problems is to construct a composite likelihood weighting 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 $e_t \sim iid(0, \sigma^2)$, $\alpha < 1$, and that stock prices are the discounted sum of future dividends. The solution for stock 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) has unitary rank. Thus, one has to decide whether d_t or p_t should be used to construct the likelihood and to estimate the common parameters $\theta = (\alpha, \sigma^2)$.

In this example, adding 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 stock prices becomes non-normal, complicating estimation. When the composite likelihood is employed, the joint information present in (d_t, p_t) can be used to identify and estimate θ (and β, if it is of interest). Optimization makes stock prices and dividends contain different information. Choosing one endogenous variable for estimation, throws away part of the information the model provides. By combining all available information, the composite likelihood may provide sharper estimates of the parameters of interest.

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

$$\log L(\alpha, \sigma^2|\tilde{d}_t) = -0.5T\log(2\pi) - \sum_{t=1}^{T}\log\varsigma_t - 0.5\sum_{t=1}^{T}\frac{\tilde{d}_t^2}{\varsigma_t^2} \tag{22}$$

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

$$\begin{aligned}\tilde{d}_t &= d_t - \alpha\frac{1+\alpha^2+\alpha^4+\ldots+\alpha^{2(t-2)}}{1+\alpha^2+\alpha^4+\ldots+\alpha^{2(t-1)}}\tilde{d}_{t-1} &(23)\\ \varsigma_t^2 &= \sigma^2\frac{1+\alpha^2+\alpha^4+\ldots+\alpha^{2t}}{1+\alpha^2+\alpha^4+\ldots+\alpha^{2(t-1)}} &(24)\end{aligned}$$

and

$$\log L(\beta, \alpha, \sigma^2|\tilde{p}_t) = -0.5T\log(2\pi) - \sum_{t=1}^{T}\log\upsilon_t - 0.5\sum_{t=1}^{T}\frac{\tilde{p}_t^2}{\upsilon_t^2} \tag{25}$$

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

$$\begin{aligned}\tilde{p}_t &= p_t^* - \gamma\frac{1+\gamma^2+\gamma^4+\ldots+\gamma^{2(t-2)}}{1+\gamma^2+\gamma^4+\ldots+\gamma^{2(t-1)}}\tilde{p}_{t-1} &(26)\\ \upsilon_t^2 &= \sigma^2\frac{1+\gamma^2+\gamma^4+\ldots+\gamma^{2t}}{1+\gamma^2+\gamma^4+\ldots+\gamma^{2(t-1)}} &(27)\end{aligned}$$

where $\gamma = \frac{\alpha}{(1-\beta\alpha)}$ and $p_t^* = \frac{p_t}{1-\beta\alpha}$. For illustration, set $\sigma^2 = 1$, $\beta = 0.99$, and focus attention on α. The first-order conditions that a maximum likelihood estimator solves are $\frac{\partial \log L(\tilde{d}_t)}{\partial\alpha} = 0$ and $\frac{\partial \log L(\tilde{p}_t)}{\partial\alpha} = 0$. For a given ω assigned to $\tilde{d}_t$, the composite likelihood is a weighted sum of (22) and (25). While there are no closed expressions for either the maximum likelihood or the maximum composite likelihood estimators of α, we can still infer what (22) and (25) employ to estimate α using simulated data.

Figure 3 plots the likelihood contour in the α dimension, when (22), (25), or the composite likelihood are used, and the true α is either 0.7 or 0.1. When the true $\alpha = 0.1$ (22) and (25) are similar. Thus, when dividends and stock prices are almost serially uncorrelated, they have the same information and the shape of both likelihood functions primarily reflects the volatility of the generating shock. When $\alpha = 0.7$, the two likelihood functions differ. In particular, the likelihood function of stock prices is bell shaped around the true value, while the likelihood function of dividends is not. Thus, the likelihood of stock prices contains information about the persistence of the generating process that the likelihood of dividends does not generally have.

The composite likelihood, which, in this case, is constructed equally weighting the two likelihoods, captures both the serial correlation and the variability properties of the DGP, it is more bell shaped than each of the likelihoods and is centered around the true value when $\alpha = 0.7$. Because when $\alpha = 0.1$, (22) and (25) have similar information, neither the shape nor the location improves when the composite likelihood is used. Clearly, depending on the value of ω, either the serial correlation or the variance properties of (d_t, p_t) or both will be employed for identification and estimation.

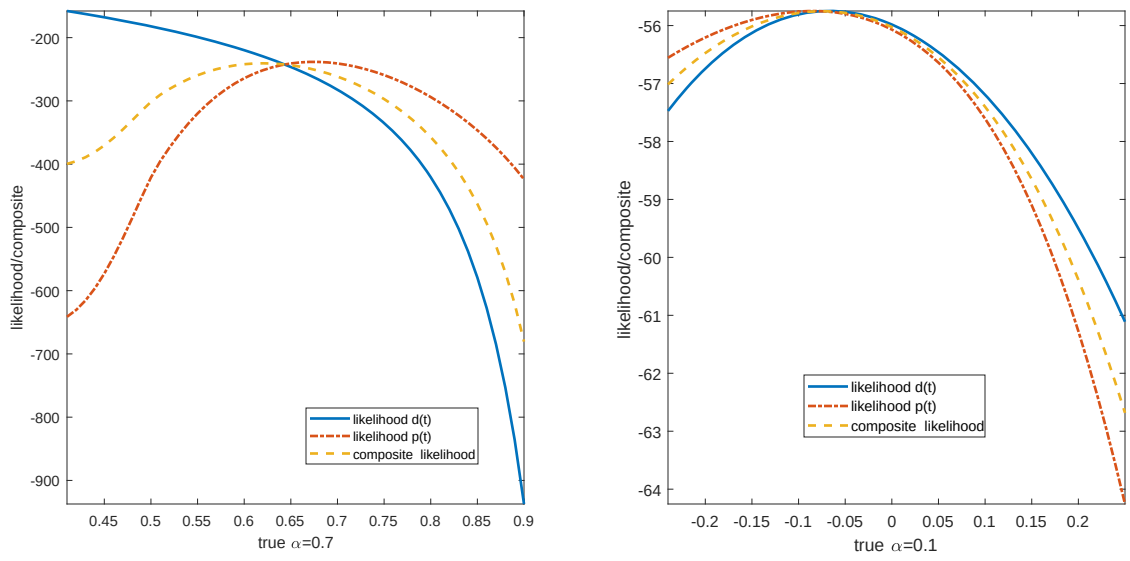

Figure 3: Likelihood and composite likelihood, singularity.

In general, when the equations of a singular model provide different information, it is a-priori difficult to choose which ones to use in estimation. The composite likelihood combines the information contained in different equations.

4.4 Dealing with large scale structural models

While in academics models are kept small for analytical tractability and to enhance intuition, large scale models are common in policy institutions. Such models can be more detailed and realistic, but estimating their parameters is computationally a daunting task and estimates obtained are often unreasonable. We show here how the composite likelihood can be used to make the estimation of the structural parameters of a large scale model more manageable.

Suppose the decision rules of a model are $y_t = A(\theta)y_{t-1}+e_t$, where e_t iid $\mathrm{N}(0,\Sigma(\theta))$, θ is a vector of structural parameters, y_t is of large dimension, and, to keep the presentation simple, we let $\dim(\mathrm{y}_t) = \dim\,(e_t)$.

The likelihood function is

$$L(\theta|y_t) = (2\pi)^{-T/2}|\Sigma(\theta)|^{T/2}\exp\{(y_t - A(\theta)y_{t-1})\Sigma(\theta)^{-1}(y_t - A(\theta)y_{t-1})'\} \qquad (28)$$

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 near collinear or if there are near singularities in the model due, for example, to the presence of an expectational link between long and short term interest rates.

Another case when the computation of (28) is difficult is when there are latent endogenous variables. If $y_t = (y_{1t}, y_{2t})$, where y_{2t} is non-observable,

$$L(\theta|y_{1t}) = \int L(\theta|y_{1t}, y_{2t})dy_2 \tag{29}$$

and, when y_{2t} is of large dimension, (29) may be intractable.

Rather than using (28) or (29) as objective functions or as inputs in Bayesian calculation, we can take a limited information point of view and produce estimates of the parameters using objects that are simpler to construct and use (see earlier work by Pakel et al., 2011).

Suppose we partition $y_t = (y_{1t}, y_{2t}, \dots y_{Kt})$, where y_{it} and y_{jt} are not necessarily independent. Then two such objects are:

$$CL_1(\theta|y_t) = \sum_{i=1}^{K} \omega_i \log L(\theta|y_{it}) \tag{30}$$

$$CL_2(\theta|y_{it}) = \sum_{i=1}^{K} \omega_i \log L(\theta|y_{it}, \bar{y}_{-it}) \tag{31}$$

where y_{-it} indicates any combination of the vector y_t, which excludes the i-th combination, and the bar indicates a given value.

CL_1 is obtained by neglecting the correlation structure among y_{it}. Thus, blocks of the model are treated as if they provide independent information for θ, even though this is not necessarily the case. For example, in a multi-country symmetric model, y_{it} could correspond to the observables of country i; in a closed economy model, it could correspond to different sectors of the economy. CL_2 is obtained by conditionally blocking groups of variables. In the multi-country example, one would construct the likelihood of each country's variables y_{it}, given the vector of the variables of all other countries y_{-it}, and then compute a weighted average. Which composite likelihood one uses depends on the problem and the tractability of conditional vs. marginal likelihoods.

To compare the likelihood of the full model and a particular composite likelihood, we consider a simple consumption-saving problem where there are many countries i, and consumers receive income from different countries but are forced to save domestically. The solution when preferences are quadratic, $\beta(1 + r) = 1$, and the income process in each county is transitory is

$$c_{it} = \frac{r}{r+1} a_{it} + \frac{r}{1-\rho+r} w_{it} \tag{32}$$

$$a_{it+1} = (1 + r)(a_{it} + w_{it} - c_{it}) \tag{33}$$

$$y_{it} = \rho y_{it-1} + \sigma_i e_{it} \tag{34}$$

$$w_{it} = \sum_{j=1}^{K} \zeta_{ij} y_{jt} \tag{35}$$

where $0 < \zeta_{ij} < 1$ and $\sum_i \zeta_{ij} = 1, \sum_j \zeta_{ij} = 1$, y_{it} is domestic income, w_{it} is total income in country i, c_{it}, is consumption and a_{it} is asset holdings of country i, and e_{it} iid $(0,1)$, $i = 1, 2, \ldots, K$.

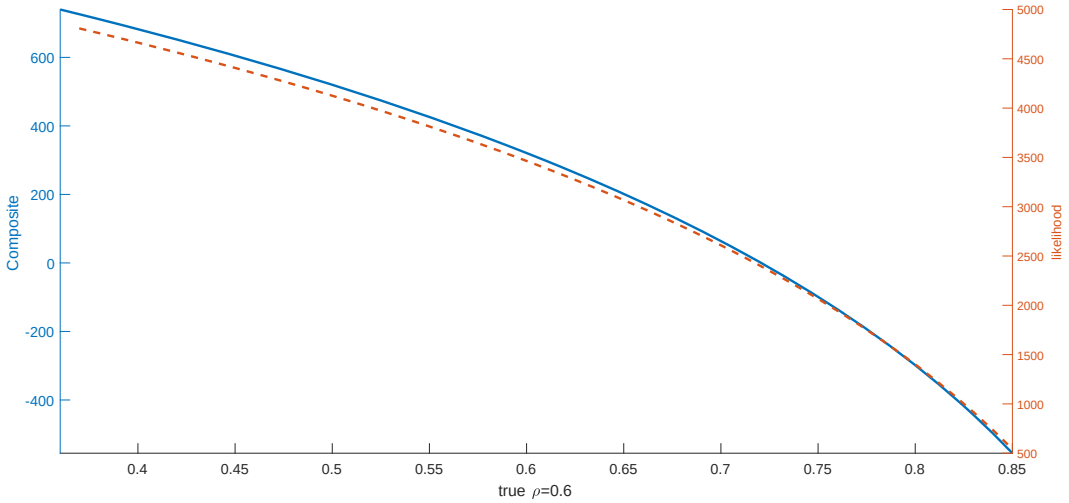

Figure 4: Likelihood and composite likelihood, large scale model.

Suppose that rather than constructing the likelihood using (32)-(35) jointly for the K countries, one constructs the likelihood of the model of each country (i.e. neglecting (35) and using y_{it} in place of w_{it} in the first two equations) and equally weighs the K likelihoods to construct a composite likelihood. Three types of misspecification are present in the composite likelihood: consumption and asset holdings are functions of total income, rather than domestic income; the volatility of domestic income is higher than the volatility of total income; the ω weights should reflect ζ_{ij} rather than being constant. Clearly if $\zeta_{ij} = \zeta_i = 1, \forall j$, and the volatility of the income process across i is the same, and the loss of information in the composite likelihood relative to the full likelihood is minimal.

Figure 4 plots the shape of the likelihood of the full model and the composite likelihood in the ρ dimension when $K = 3, \beta = 0.99, \rho = 0.6, \sigma_i = [0.1, 0.2, 0.3], \omega = 1/K$, $r = 1/\beta - 1, \beta = 0.99$ $\zeta = \begin{bmatrix} 0.5 & 0.25 & 0.25 \\ 0.25 & 0.5 & 0.25 \\ 0.25 & 0.25 & 0.5 \end{bmatrix}$ using consumption data only when T=1000. The likelihood function is not quadratic in ρ, as it is clear from inspection - the marginal propensity to consume out of transitory income increases as ρ moves from -1 to 1 - and the composite likelihood inherits this property. Nevertheless, although the scale is different, the likelihood and the composite likelihood have very similar shapes. Thus, the information loss due to the use of a limited information object like the composite likelihood is small in this case.

4.5 Estimating a parameter appearing in different models

Likelihood-based estimates are seldomly used directly in policy exercises but instead twisted to reflect a-priori information not included in the estimation ("your boss' prior") or informally averaged taking the output of many models into account. Such an approach is consistent with the idea that models are misspecified, that averaging safeguards against structural breaks, time variations, etc., and that "judgement" is important when evaluating the appeal of certain counterfactual exercises.

In practice, two approaches are common in the literature: i) models are separately estimated, counterfactuals are constructed in each model, and then averaged using user-based weights; ii) estimates from different models are informally averaged, and one counterfactual is constructed using the average estimates in the "most-likely" model. This section shows how the composite likelihood can be used to formally construct counterfactuals when a number of structural models are available. The composite likelihood formally averages the inputs of such a process. Canova and Matthes (2017) show that when a Bayesian approach is used, one can pick the model with the highest posterior mode for ω and construct counterfactuals using composite estimates of the parameters and that model. Alternatively, one can use composite estimates in different models and weight counterfactuals from different models with the posterior mode of ω. Thus a composite likelihood approach provides a formal approach that justifies both approaches used in the literature.

Suppose the decision rules that two such models generate are given by (10) and (11). Maximization of (14) with respect to θ leads to:

$$\rho_A = (\sum_{t=1}^{T_A} y_{At-1}^2 + \zeta_2 \sum_{t=1}^{T_B} y_{Bt-1}^2)^{-1} (\sum_{t=1}^{T_A} y_{At} y_{At-1} + \zeta_1 \sum_{t=1}^{T_B} y_{Bt} y_{Bt-1}) \tag{36}$$

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

$$\sigma_A^2 = \frac{1}{\xi} (\sum_{t=1}^{T_A} (y_{At} - \rho_A y_{At-1})^2 + \frac{1-\omega}{\omega\gamma^2} \sum_{t=1}^{T_B} (y_{Bt} - \delta\rho_A y_{Bt-1})^2) \tag{37}$$

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

$$\rho_{AA} = (\sum_{t=1}^{T_A} y_{At-1}^2)^{-1} (\sum_{t=1}^{T_A} y_{At} y_{At-1}); \quad \rho_{AB} = \delta^{-1} (\sum_{t=1}^{T_B} y_{Bt-1}^2)^{-1} (\sum_{t=1}^{T_B} y_{Bt} y_{Bt-1}) \tag{38}$$

and

$$\sigma_{AA}^2 = \frac{1}{T_A} \sum_{t=1}^{T_A} (y_{At} - \rho_{AA} y_{At-1})^2; \quad \sigma_{AB}^2 = \frac{1}{T_B} \sum_{t=1}^{T_B} (y_{Bt} - \delta\rho_{AB} y_{Bt-1})^2 \tag{39}$$

As (36)-(37) clearly show, θ_{CL} combines the information coming from y_{At} and y_{Bt}, with model B playing the role of a prior for model A. The formulas in (36) and (37)

are in fact 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 is the number of additional observations added to the original T_A data points. Note that if (γ, δ) are unknown and estimated jointly with ρ_A, σ_A^2 using the composite likelihood, they will reflect only the information contained in y_{BT}.

When an array of models are available, θ_{CL} will be constrained by the structure present in all models. For example, equation (36) becomes

$$\rho_A = (\sum_{t=1}^{T_A} y_{At-1}^2 + \sum_{i=1}^{K-1} \zeta_{i2} \sum_{t=1}^{T_i} y_{it-1}^2)^{-1} (\sum_{t=1}^{T_A} y_{At} y_{At-1} + \sum_{i=1}^{K-1} \zeta_{i1} \sum_{t=1}^{T_i} y_{it} y_{it-1}) \tag{40}$$

where $\zeta_{i1} = \frac{\omega_i}{\omega_A} \frac{\delta_i}{\gamma_i^2}$, $\zeta_{i2} = \zeta_{i1} \delta_i$. (40) has the same format as the estimator suggested by Zellner and Hong (1989), and combines unit specific and average information contained in the cross section of models. Thus, the composite likelihood makes inference more robust, in the sense that estimates of θ are shrunk to be consistent with the data generated by all available models.

Note that y_{At} and y_{Bt} may be different series. Thus, the procedure can be used to estimate parameters appearing in models featuring different observables or 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. The approach works in all these situations.

We illustrate the idea when a researcher is interested in estimating the slope of Phillips curve. The conventional wisdom 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., 2016). However, Schorfheide (2008), surveying estimates of the slope of the Phillips curve obtained in DSGE models, documents large cross-study variations and associates the differences to i) the specification of the model, ii) the observability of marginal costs, and iii) the number and the type of variables used in estimation.

Because of its importance for forecasting and counterfactual exercises, we examine how the composite posterior distribution of the Phillips curve looks relative to the posterior distribution obtained with i) single models and ii) ex-post averaging the posteriors of different models. We then construct the responses of the ex-ante real rate to monetary shocks in a variety of situations.

We consider five different 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 detrended 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 W, as in Justiniano et al. (2010); a New Keynesian model with search and matching labor 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 Y, demeaned π, and demeaned 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 which 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 B. Note that the models do not use the same observables, so standard Bayesian model averaging is not possible in our case.

We assume that the prior for ω is Dirichlet with parameters 250*[1/4;1/3;1/7;1/4;1/3].[3] [4]

Table 2 displays some percentiles of the posterior of the slope of the Phillips curve obtained either with the likelihood of each model separately or the composite likelihood. For the first three models the median value is low and having non-observable marginal costs increases the location of the posterior distribution. For the other two models, the posterior median is higher and, for the model with search and matching friction, the posterior also has larger spread. Note that the posteriors of these two models hardly overlap 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 a 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. Correcting the posterior percentiles (as suggested by Mueller, 2013) leaves the location and the spread of the posterior distribution unchanged.

Figure 5 plots the prior and the posterior ω for each model. Interestingly, the location of posterior for the models with financial and labor market friction is the least affected by the estimation process. On the other hand, for the small NK model

[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 (set to the mean of the prior density we use as a benchmark) are similar and available on request.

[4]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. When this happens, a two-step approach can be used, where the prior for the nuisance parameters is made data-based using the posterior for each model estimated on a training sample. This effectively avoids under-identification and makes the priors for the nuisance parameters endogenous.

Table 1: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 misspecification (such as repeated use of the same time series across models). The estimation sample is 1960:1-2005:4.

with observable marginal costs and the medium scale NK model the posterior median decreases relative to the prior median, and the opposite is true for the basic NK model. Also, posterior spreads are tighter than the prior spread, indicating that the data are 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 readers may be surprised about the fact that the standard medium scale New Keynesian model, which is the workhorse used in many policy institutions, has the lowest posterior probability among our five models. Recall that the posterior for ω reflects the information of each model for the slope of the Phillips curve. Thus, figure 5 indicates that the medium scale NK model does not provide independent information relative to the pool of the other models for this parameter.

Figure 6 presents the composite posterior distribution for the slope of the Phillips curve we obtain together with two alternative naive posterior combinations: one that equally weights the posteriors obtained with the five models separately; and one which weights the posteriors obtained with the five models by the mode of ω for that model. Clearly, combining ex-post estimates generate distributions whose locations are different and generally lower. In addition, ex-post combinations produce multimodal posteriors: there is a sharp peak at 0.05, and a secondary, more round, one at 0.15.

Figure 7 reports the responses of the ex-ante real rate to a 25 annualized basis points monetary policy shock in four situations: using the estimates obtained in the model with the largest modal value of ω (the small NK model); using the two ex-post

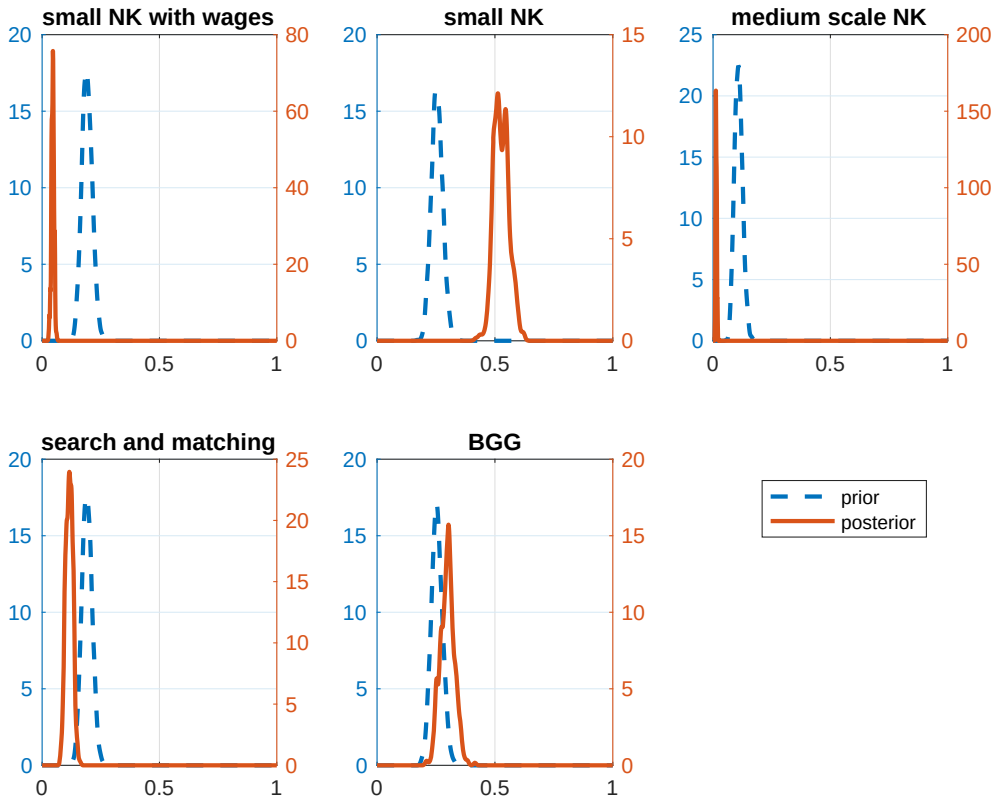

Figure 5: Prior and posterior densities of ω

combinations previously discussed, and using composite posterior estimates in each model and then weighting the resulting impulse responses using the posterior mean of ω for each model.

The mean impact is estimated to be 45-50 basis points, and the composite impact response is intermediate among the values we present. Uncertainty is substantial, and while the composite responses are a-posteriori different from zero, the 68% credible set includes the point estimates of all models. At horizons larger than one, the composite posterior real rate band becomes tighter and the responses obtained with the naive equal weighting scheme fall outside the credible composite posterior interval. Note also that the composite posterior real rate responses are much less persistent relative to any other alternative and after four quarters the real rate responses are negligible.

4.6 Exploiting panel information in estimation.

A composite likelihood setup can also easily deal with the situation where there is a single structural model, for example, an asset pricing model, but the observable data

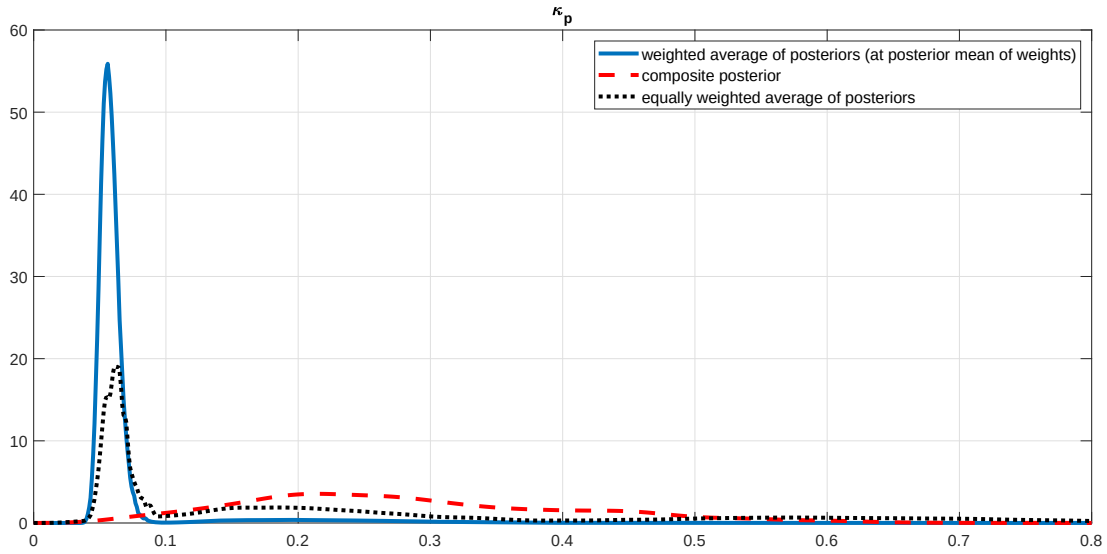

Figure 6: Composite posterior and two naive posterior mixtures

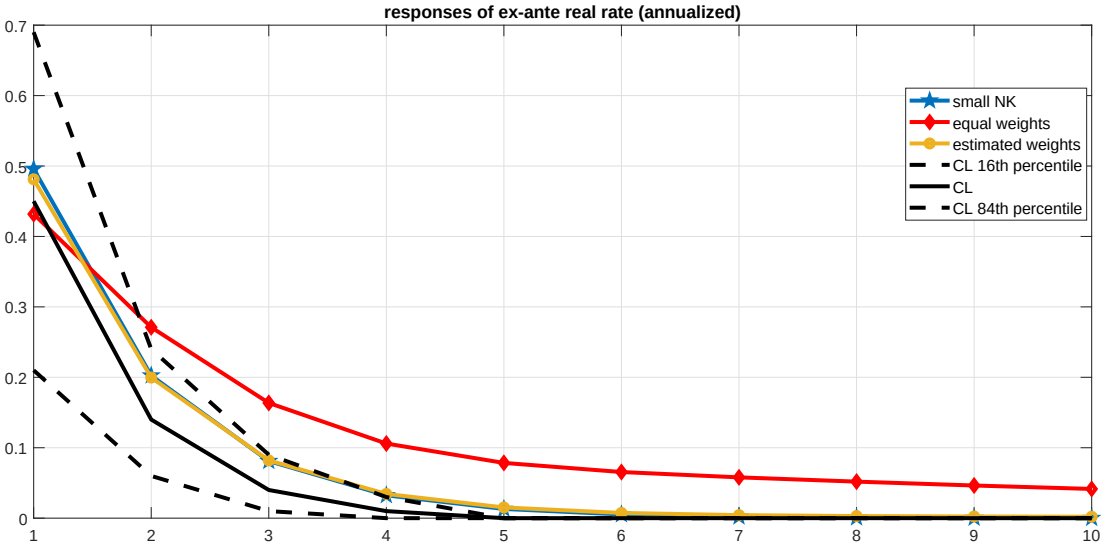

Figure 7: Real rate responses to a monetary shock

come from either different units (which could be, for example, many consumers or many countries); or from different levels of data aggregation (firm, industry, sector, region).

Earlier work by Chamberlain (1984, p.1272) has used similar ideas to estimate the

parameters of a reduced form model when a panel is available but the cross-sectional data are not necessarily homogenous. In the composite likelihood setup, we treat time series for different cross-sectional units as different "models" and combine their information to estimate common structural parameters.

Let $\hat{y}_{1t}, \hat{y}_{2t}, ...\hat{y}_{Kt}$ represent the subset of the vector of observables of unit (level of aggregation) i=1,2...K that is common across units. The composite log-likelihood is

$$CL(\theta|\hat{y}_{1t}...\hat{y}_{Kt}, \eta_1, \dots \eta_k) = \sum_{i=1}^{K} \omega_i \log L(\theta|\hat{y}_{it}, \eta_i) \tag{41}$$

As in section 4.4, (41) neglects the correlation structure across units and, in particular, the presence of common shocks, but partially pools information about common parameters from the available cross section. Thus, the composite likelihood (41) represents an intermediate case between complete pooling of cross unit information $CL(\theta, \eta|\hat{y}_{1t}...\hat{y}_{Kt}) = \sum_{i=1}^{K} \omega_i \log L(\theta, \eta|\hat{y}_{it})$ and complete heterogeneity $CL(\theta_1, \dots \theta_k, \eta_1, \dots \eta_k|\hat{y}_{1t}...\hat{y}_{Kt}) = \sum_{i=1}^{K} \omega_i \log L(\theta_i, \eta_i|\hat{y}_{it})$. It is similar in spirit to the objective function employed in partial pooling Bayesian literature (e.g., Zellner and Hong, 1989). The main difference with that literature is that in partial pooling exercises all cross-sectional parameters are restricted; here only the subvector θ is restricted across units.

Suppose we have available decision rules like (11) for unit i where now δ_i, γ_i are unit specific, $\delta_1 = \gamma_1 = 1$, while ρ_A, σ_A are common. As seen, for fixed ω, the composite likelihood estimator for ρ_A is

$$\rho_A = (\sum_{t=1}^{T_1} y_{1t-1}^2 + \sum_{i=2}^{K} \zeta_{i2} \sum_{t=1}^{T_i} y_{it-1}^2)^{-1} (\sum_{t=1}^{T_1} y_{1t} y_{1t-1} + \sum_{i=2}^{K} \zeta_{i1} \sum_{t=1}^{T_i} y_{it} y_{it-1}) \tag{42}$$

where $\zeta_{i1} = \frac{\omega_i}{\omega_1} \frac{\delta_i}{\gamma_i^2}$, $\zeta_{i2} = \zeta_{i1}\delta_i$. Clearly, the CL estimator for ρ_A pools cross-sectional information if $\zeta_{ij} = 1, \forall i, j$, and corresponds to the ML estimator obtained with unit 1 data if $\zeta_{ij} = 0, \forall i, j,$. When $\omega_i = 1/K$, ζ_{ij} captures the degree of heterogeneity in the cross section. In general, cross sectional information is not exactly pooled, as for example, in standard panel estimators and the degree of cross-sectional shrinkage depends on the precision of various sources of information. Thus, when dealing with panels of time series, 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 common parameters when the cross sectional data display similarities. The partial pooling approach that the composite likelihood delivers is likely to be preferable when each y_{it} is short, when the heterogeneties in the DGP for θ are unsystematic (if they are systematic, the partial pooling device could be applied to units whose variations are unsystematic), and when the volatility of the endogenous variables is of similar order of magnitude.

To illustrate the use of the composite likelihood in this particular setup, we build on the exercise of Karabarbounis and Neiman (2014). They notice that the labor

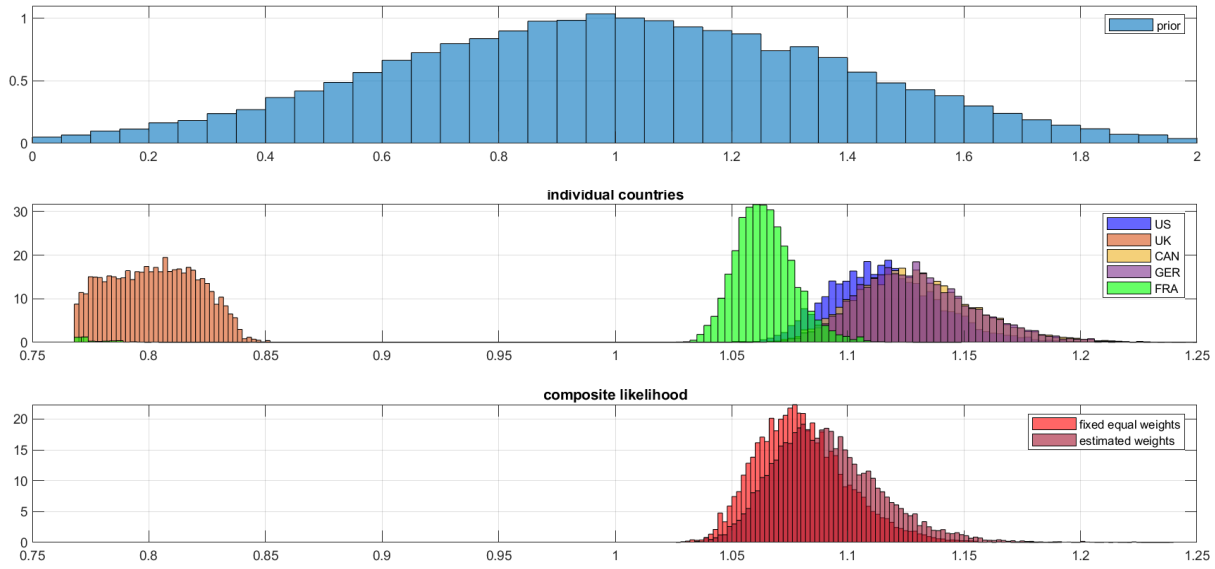

Figure 8: Prior and Posterior distributions for σ

share has dramatically fallen in many countries over the last twenty years and argue that shocks to the relative price of investment, which also decline over time, may be responsible for this decline. Their argument hinges on having the elasticity of substitution between labor and capital in production, σ, to be greater than one. Using their model specification (the log-linearized optimality conditions are in appendix C) and their dataset, we first estimate this parameter separately using data from the US, UK, Canada, Germany and France. We then use the composite likelihood to estimate σ jointly using data from all five countries [5]. In this latter case, all other parameters of the model are allowed to be country specific.

Figure 8 presents the prior for σ (first row), the posterior obtained with the data of the individual countries (second row) and the composite posterior when we use fixed equal weights or random weights. The data is informative about σ for all countries and, except for the UK, the posterior distribution is entirely above one. The two composite posterior distributions are also all above one and tight, despite the fact that UK data receive non-negligible weight in both composite estimation exercise (modal value of the posterior of ω for the Uk is 0.07). US data appear to be most informative and the posterior of ω for the US has mode equal to 0.45.

Figure 9 shows the responses of the labor share, in log deviation from the steady state to a positive shock to the relative price of investment (with mean equal to half of

[5] Although we present results when shocks to the price of investment are stationary, we also perform estimation assuming non-stationary shocks. None of the conclusions we reach depend on this assumption.

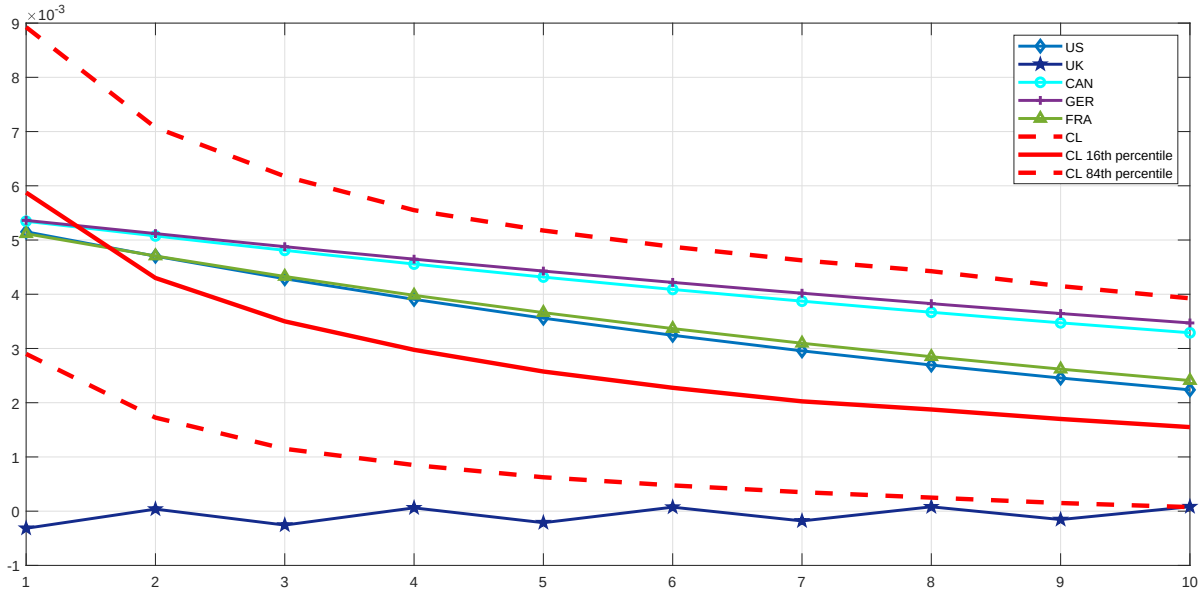

Figure 9: Labor share responses to shocks to the relative price of investment

the estimated US standard deviation) in each of the five countries and with the panel when random weights are used. Indeed, we do find a positive dynamic conditional correlation between shocks to the relative price of investment and the labor share whenever the posterior of σ is above one. For the UK instead, shocks to the relative price of investment have negligible dynamic effects on the labor share.

Thus, our analysis confirms by and large Karabarbounis and Neiman's two main conclusions, i.e., i) the elasticity of substitution between capital and labor is greater than one, and ii) shocks to the relative price of investment are potentially able to explain the fall in the labor share observed in many countries. Nevertheless, we would like to stress that our results are more general because we allow for stochastic heterogeneity across countries, and we use likelihood-based estimators that exploit all the information present in the optimality conditions the theory provides.

5 Conclusions

This paper describes a procedure that has the potential to ameliorate identification and estimation problems in DSGE models. The method may help to solve a number of difficulties researchers typically face and automatically provides estimates of the parameters that formally combine the information present in different models/ different data using shrinkage-like estimators. The procedure helps to robustify estimates of the structural parameters in a variety of interesting economic problems and it is applicable

to many empirical situations of interest.

The approach we suggest is based on the *composite likelihood*, a limited-information objective function, well known in the statistical literature but very sparsely used in economics. In the original formulation, the composite likelihood is constructed combining marginal or conditional likelihoods of the true DGP when the likelihood of the full model is computationally intractable or features unmanageable integrals due to the presence of latent variables. When marginals or conditionals are used, the composite likelihood estimator is consistent and asymptotic normal, as either the number of observations or the number of composite likelihood components grows, but it is not fully efficient.

In our setup, the composite likelihood combines the likelihood of distinct structural or statistical models, none of which are necessarily marginal or conditional partitions of the DGP. Thus, standard composite likelihood properties do not necessarily apply. Still, the approach we propose has desirable statistical properties, is easy to use; and, in its Bayesian version, has an appealing sequential learning interpretation.

We present examples indicating that the composite likelihood constructed using the information present in distinct models helps 1) to ameliorate population and sample identification problems, 2) to solve singularity problems, 3) to produce more stable estimates of the parameters of large-scale structural models, 4) to robustly estimate the parameters appearing in multiple models and select models with different numbers of observables, 5) to combine information coming from different sources and levels of aggregation. In Canova and Matthes (2017), we have shown that a composite likelihood approach can also be used to deal with model misspecification and has a built-in feature that allows researchers to i) examine whether the composite likelihood produces better estimates than the likelihood of a single model, and ii) assess a-posteriori which model is closer to the unknown DGP.

We believe the methodology has potential, and the examples we describe in the text highlight ways in which the flexibility of the approach can be exploited for useful economic applications.

6 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,1, 1341-1393.

Besag, J. (1974): Â"Spatial Interaction and the Statistical Analysis of Lattice Systems,Â' Journal of the Royal Statistical Society (Series B), 36, 192-236.

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.

Canova, F. and C. Matthes (2017) An alternative approach to deal with model misspecification, manuscript.

Chamberlain, G. (1984). Panel Data. In Z. Griliches and M. D. Intriligator (eds.). Handbook of Econometrics, Volume 2 chapter 22, pp. 1247â€"1318. North-Holland, Amsterdam.

Chan, J., Eisentain, E., Hu, C. and G. Koop (2017) Composite likelihood methods for large BVARs with stochastic volatility, manuscript.

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

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.

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

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

Hamilton, J. (1994). Time series analysis. Princeton University Press, Princeton, NJ.

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

Ireland, P. (2004). A method for taking models to the data, Journal of Economic Dynamics and Control, 28, 1205-1226. Justianiano, A. Primiceri, G. and A. Tambalotti (2010). Investment shocks and the business cycle. Journal of Monetary Economics, 57, 132-145.

Karabarbounis, L. and B. Neiman (2014). The global decline of the labor share. Quarterly Journal of Economics, 129, 61-103

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.

Lindsay, B.G. (1988): Â“Composite Likelihood Methods,Â”Contemporary Mathematics, 80, 221-239.

Marin, J.M., Pudlo, P., Robert, C. and R. Ryder (2012) Approximate Bayesian computational models. Statistics and Computing, 22, 1167-1180.

Mueller, U.K. (2012) Measuring Prior Sensitivity and Prior Informativeness in Large Bayesian Models, Journal of Monetary Economics 59, 581 â€“ 597.

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

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.

Zellner, A. and C. Hong (1989) Forecasting International growth rates using Bayesian shrinkage and other procedures. Journal of Econometrics, 40, 183-202.

Appendix A

The MCMC algorithm Given (y_{it}, T_i), suppose 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 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 computational and efficiency reasons, we employ a $2K + 1$ block Metropolis-within-Gibbs algorithm to derive sequences for the parameters. 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.

The algorithm we use has 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
$$\min\left(1, \frac{\mathcal{L}(\left[\eta_i^*, \theta^{iter-1}\right] | Y_{i,T_i})^{\omega_i^{iter-1}} p(\eta_i^* | \theta^{iter-1})^{\omega_i^{iter-1}}}{\mathcal{L}(\left[\eta_i^{iter-1}, \theta^{iter-1}\right] | Y_{i,T_i})^{\omega_i^{iter-1}} p(\eta_i^{iter-1} | \theta^{iter-1})^{\omega_i^{iter-1}}}\right) \tag{43}$$
3. Draw θ^* from a symmetric proposal P^θ. Set $\theta^{iter} = \theta^*$ with probability
$$\min\left(1, \frac{\mathcal{L}(\left[\eta_1^{iter}, \theta^*\right] | Y_{1,T_1})^{\omega_1^{iter-1}} \ldots \mathcal{L}(\left[\eta_K^{iter} \theta^*\right] | Y_{K,T_K})^{\omega_K^{iter-1}} p(\theta^*)}{\mathcal{L}(\left[\eta_1^{iter}, \theta^{iter-1}\right] | Y_{1,T_1})^{\omega_1^{iter-1}} \ldots \mathcal{L}(\left[\eta_K^{iter}, \theta^{iter-1}\right] | Y_{K,T_K})^{\omega_K^{iter-1}} p(\theta^{iter-1})}\right) \tag{44}$$
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
$$\min\left(1, \frac{\mathcal{L}(\left[\eta_1^{iter}, \theta^{iter}\right] | Y_{1,T_1})^{\omega_1^*} \ldots \mathcal{L}(\left[\eta_K^{iter} \theta^{iter}\right] | Y_{K,T_K})^{\omega_K^*} p(\omega^*)}{\mathcal{L}(\left[\eta_1^{iter}, \theta^{iter}\right] Y_{1,T_1})^{\omega_1^{iter-1}} \ldots \mathcal{L}(\left[\eta_i^{iter}, \theta^{iter}\right] | Y_{K,T_K})^{\omega_K^{iter-1}} p(\omega^{iter-1})}\right) \tag{45}$$

Note that in (43) only the likelihood of model i matters 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, when 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 models are considered) are natural choices for ω_i.

The estimation problem is non-standard since 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. 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. In addition, as we show next 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. For all these reasons, one may want to adjust the percentiles of the posterior to reflect these facts.

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

$$\begin{aligned} p_{CL}(\theta|Y) &\propto \{CL(\theta_{CL}|y) - 0.5(\theta-\theta_{CL})^T(h_(\theta_{CL})(\theta-\theta_{CL}) + logp(\theta_p) - 0.5(\theta-\theta_p)^T(h_(\theta_p)(\theta-\theta_p)\} \\ &\approx N(\hat{\theta}, h(\theta_{CL},\theta_p)^{-1}) \end{aligned} \tag{46}$$

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

$$T^{-1}h(\theta_{CL},\theta_p) \to -E(\nabla^2 CL(\theta_0|Y)) \equiv H(\theta_0) \tag{47}$$

$$\hat{\theta} = (T^{-1}h(\theta_{CL},\theta_p))^{-1}(T^{-1}h(\theta_{CL})\theta_{CL} + T^{-1}h(\theta_p)\theta_p) \to \theta_0 \tag{48}$$

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

Mueller (2103) has argued that in situations like ours, MCMC percentiles should be adjusted to obtain asymptotic coverage which is consistent with the amount of information present in the data. To do so, we follow Ribatet et al. (2012) and Qu (2015) and modify the MCMC algorithm adding two steps. 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

$$\tilde{\theta}^j = \hat{\theta} + V^{-1}(\theta^j - \hat{\theta}) \tag{49}$$

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.

Note that 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. Thus, it may be advisable to report both standard and adjusted percentiles.

Asymptotic properties of estimators of misspecified models 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

$$E \log g(y_t, \psi) - E \log f(y_t) = -\int f(y_t) \log \frac{f(y_t)}{g(y_t, \psi)} d\mu(y) = -KL(\psi) \tag{50}$$

where $KL(\psi)$ is the Kullback-Leibler 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

$$0 \approx T^{-0.5} \sum_t s_t(\psi_0) + T^{0.5} \Sigma_1^{-1} (\psi_{ML} - \psi_0) \tag{51}$$

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 functions of martingale difference processes (the shocks of the model). Thus, the condition $\sum_t s_t(\psi) = 0$ is likely to hold. Further regularity conditions (see, e.g. Mueller, 2013) need to be imposed for the argument to go through.

The composite likelihood is the 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

We present the optimality conditions for each of the five models we consider in section 4.5. In estimation, the priors for the parameters are generally Gaussian and centered at the values used (or estimated) in the original papers, with a standard deviation of at least 25 percent of the mean value. For those parameters that are naturally restricted to be positive or between 0 and 1, we truncate the Gaussian priors, in which case the standard deviation refers to the value before truncation. The only parameter we treat as common across models is the slope of the Phillips curve, for which we assume a prior mean of 0.2 and a prior standard deviation of 0.5 (thus a very loose prior) and truncate the support to be positive. Posterior moments are computed using 50000 draws, which

are generated after a burn-in phase of 10000 draws.

a) Small scale New Keynesian models

$$
\begin{aligned}
y_t &= E_t y_{t+1} - \sigma\left(r_t - E_t \Delta p_{t+1} + E_t g_{t+1} - g_t\right) && (52)\\
y_t &= a_t + (1-\delta)\, n_t && (53)\\
mc_t &= w_t - p_t + n_t - y_t && (54)\\
mrs_t &= \frac{1}{\sigma} y_t + \gamma n_t - g_t && (55)\\
r_t &= \rho_r r_{t-1} + (1-\rho_t)\left[\gamma_\pi \Delta p_t + \gamma_y y_t\right] + z_t && (56)\\
w_t - p_t &= w_{t-1} - p_{t-1} + \delta w_t - \delta p_t && (57)\\
a_t &= \rho_a a_{t-1} + \epsilon_t^a && (58)\\
g_t &= \rho_g g_{t-1} + \epsilon_t^g && (59)\\
z_t &= \epsilon_t^z && (60)\\
\lambda_t &= \epsilon_t^\lambda && (61)\\
\Delta p_t &= \beta E_t \Delta p_{t+1} + \kappa_p\left(mc_t + \lambda_t\right) && (62)\\
w_t - p_t &= mrs_t && (63)\\
\Delta p_t &= \gamma_b \Delta p_{t-1} + \gamma_f E_t \Delta p_{t+1} + \kappa_p'\left(mc_t + \lambda_t\right) && (64)
\end{aligned}
$$

In the sticky wage model, the wage equation (63) is replaced by:

$$\Delta w_t = \beta E_t \Delta w_{t+1} + \kappa_w\left[mrs_t - (w_t - p_t)\right] \qquad (65)$$

b) Medium scale New Keynesian model

$$\hat{y}_t = \frac{y+F}{y}\left[\alpha\hat{k}_t + (1-\alpha)\,\hat{L}_t\right] \tag{66}$$

$$\hat{\rho}_t = \hat{w}_t + \hat{L}_t - \hat{k}_t \tag{67}$$

$$\hat{s}_t = \alpha\hat{\rho}_t + (1-\alpha)\,\hat{w}_t \tag{68}$$

$$\hat{\pi}_t = \gamma_f E_t\hat{\pi}_{t+1} + \gamma_b\hat{\pi}_{t-1} + \kappa\hat{s}_t + \kappa\hat{\lambda}_{p,t} \tag{69}$$

$$\hat{\lambda}_t = \frac{h\beta e^{\gamma}}{(e^{\gamma}-h\beta)(e^{\gamma}-h)}E_t\hat{c}_{t+1} - \frac{e^{2\gamma}+h^2\beta}{(e^{\gamma}-h\beta)(e^{\gamma}-h)}\hat{c}_t + \frac{he^{\gamma}}{(e^{\gamma}-h\beta)(e^{\gamma}-h)}\hat{c}_{t-1} \tag{70}$$

$$+ \frac{h\beta e^{\gamma}\rho_z - he^{\gamma}}{(e^{\gamma}-h\beta)(e^{\gamma}-h)}\hat{z}_t + \frac{e^{\gamma}-h\beta\rho_b}{e^{\gamma}-h\beta}\hat{b}_t \tag{71}$$

$$\hat{\lambda}_t = \hat{R}_t + E_t\left(\hat{\lambda}_{t+1} - \hat{z}_{t+1} - \hat{\pi}_{t+1}\right) \tag{72}$$

$$\hat{\rho}_t = \chi\hat{u}_t \tag{73}$$

$$\hat{\phi}_t = (1-\delta)\,\beta e^{-\gamma}E_t\left(\hat{\phi}_{t+1} - \hat{z}_{t+1}\right) + \left(1-(1-\delta)\,\beta e^{-\gamma}\right)E_t\left[\hat{\lambda}_{t+1} - \hat{z}_{t+1} + \hat{\rho}_{t+1}\right] \tag{74}$$

$$\hat{\lambda}_t = \hat{\phi}_t + \hat{u}_t - e^{2\gamma}S''\left(\hat{\imath}_t - \hat{\imath}_{t-1} + \hat{z}_t\right) + \beta e^{2\gamma}S''E_t\left[\hat{\imath}_{t+1} - \hat{\imath}_t + \hat{z}_{t+1}\right] \tag{75}$$

$$\hat{k}_t = \hat{u}_t + \hat{\bar{k}}_{t-1} - \hat{z}_t \tag{76}$$

$$\hat{\bar{k}}_t = (1-\delta)\,e^{-\gamma}\left(\hat{\bar{k}}_{t-1} - \hat{z}_t\right) + \left(1-(1-\delta)\,e^{-\gamma}\right)\left(\hat{u}_t + \hat{\imath}_t\right) \tag{77}$$

$$\hat{w}_t = \frac{1}{1+\beta}\hat{w}_{t-1} + \frac{\beta}{1+\beta}E_t\hat{w}_{t+1} - \kappa_w\hat{g}_{w,t} + \tag{78}$$

$$+ \frac{\iota_w}{1+\beta}\hat{\pi}_{t-1} + \frac{1+\beta\iota_w}{1+\beta}\pi_t + \frac{\beta}{1+\beta}E_t\hat{\pi}_{t+1} + \tag{79}$$

$$+ \frac{\iota_w}{1+\beta}z_{t-1} - \frac{1+\beta\iota_w-\rho_z\beta}{1+\beta}z_t + \kappa_w\hat{\lambda}_{w,t} \tag{80}$$

$$\hat{g}_{w,t} = \hat{w}_t - \left(\nu\hat{L}_t + \hat{b}_t - \hat{\lambda}_t\right) \tag{81}$$

$$\hat{R}_t = \rho_R\hat{R}_{t-1} + (1-\rho_R)\left[\phi_\pi\hat{\pi}_t + \phi_X\left(\hat{x}_t - \hat{x}_t^*\right)\right] + \phi_{dX}\left[\left(\hat{x}_t - \hat{x}_{t-1}\right) - \left(\hat{x}_t^* - \hat{x}_{t-1}^*\right)\right] + \hat{\eta}_{mp,t} \tag{82}$$

$$\hat{x}_t = \hat{y}_t - \frac{\rho k}{y}\hat{u}_t \tag{83}$$

$$\frac{1}{g}\hat{y}_t = \frac{1}{g}\hat{g}_t + \frac{c}{y}\hat{c}_t + \frac{i}{y}\hat{\imath}_t + \frac{\rho k}{y}\hat{u}_t \tag{84}$$

c) Model with search and matching frictions

$$\widehat{\lambda}_t = E_t\left\{\widehat{\lambda}_{t+1} + \widehat{R_t} + \widehat{\epsilon}_t^b - \widehat{\Pi}_{t+1}\right\} \tag{85}$$

$$\widehat{\lambda}_t = -\frac{\sigma}{1-\varrho}\left(\widehat{c}_t - \varrho\widehat{c}_{t-1}\right) \tag{86}$$

$$\widehat{\Pi}_t = \gamma_f E_t\left\{\widehat{\Pi}_{t+1}\right\} + \gamma_b \pi_{t-1} + \kappa_p \widehat{mc}_t \tag{87}$$

$$\widehat{mc}_t = \widehat{x}_t^L \tag{88}$$

$$\widehat{m}_t = \xi\widehat{u}_t + (1-\xi)\,\widehat{v}_t \tag{89}$$

$$\widehat{n}_t = (1-\vartheta)\,\widehat{n}_{t-1} + \frac{m}{n}\widehat{m}_{t-1} \tag{90}$$

$$\widehat{n}_t = \frac{u}{1-u}\widehat{u}_t \tag{91}$$

$$\widehat{q}_t = \widehat{m}_t - \widehat{v}_t \tag{92}$$

$$\widehat{s}_t = \widehat{m}_t - \widehat{u}_t \tag{93}$$

$$\widehat{J^\star}_t + \widehat{\delta}_t^W = \widehat{\Delta}_t^\star + \widehat{\delta}_t^F - \frac{1}{1-\eta}\widehat{\eta}_t \tag{94}$$

$$\widehat{x}_t^L + \widehat{z}_t = (\alpha-1)\,\widehat{h}_t = \widehat{w}_t \tag{95}$$

$$\widehat{w}_t = \gamma\left[\widehat{w}_{t-1} - \widehat{\Pi}_t\right] + (1-\gamma)\,\widehat{w}_t^\star \tag{96}$$

$$\begin{aligned}\widehat{\delta}_t^F &= \left[1-\beta(1-\vartheta)\gamma\right]\left[\frac{-\alpha}{1-\alpha}\widehat{w}_t^\star + \frac{1}{1-\alpha}\left(\widehat{x}_t^L + \widehat{z}_t\right)\right] \\ &+ \beta(1-\vartheta)\gamma E_t\left\{\frac{-\alpha}{1-\alpha}\left[\widehat{w}_t^\star - \widehat{w}_{t+1}^\star - \widehat{\Pi}_{t+1}\right] + \widehat{\delta}_{t+1}^F + \widehat{\lambda}_{t+1} - \widehat{\lambda}_t\right\}\end{aligned} \tag{97}$$

$$\begin{aligned}\delta^W\widehat{\delta}_t^W &= \frac{-\alpha}{1-\alpha}wh\left[\frac{-\alpha}{1-\alpha}\widehat{w}_t^\star + \frac{1}{1-\alpha}\left(\widehat{x}_t^L + \widehat{z}_t\right)\right] \\ &- \frac{-1}{1-\alpha}mrsh\left[\frac{(-1)(1+\varphi)}{1-\alpha}\widehat{w}_t^\star - \widehat{\lambda}_t + \frac{1+\varphi}{1-\alpha}\left(\widehat{x}_t^L + \widehat{z}_t\right)\right] \\ &+ \frac{\beta(1-\vartheta)\gamma}{1-\beta(1-\vartheta)\gamma}\left[\left(\frac{\alpha}{1-\alpha}\right)^2 wh - \frac{(1+\vartheta)}{(1-\alpha)^2}mrsh\right]E_t\left\{\widehat{w}_t^\star - \widehat{w}_{t+1}^\star - \widehat{\Pi}_{t+1}\right\} \\ &+ \beta(1-\vartheta)\gamma\delta^W E_t\left\{\widehat{\lambda}_{t+1} - \widehat{\lambda}_t + \widehat{\delta}_{t+1}^W\right\}\end{aligned} \tag{98}$$

$$\begin{aligned}J\widehat{J}_t^\star &= \frac{wh}{\alpha}\left[-\alpha\widehat{w}_t^\star + \widehat{x}_t^L + \widehat{z}_t\right] \\ &+ \frac{\beta(1-\vartheta)\gamma}{1-\beta(1-\vartheta)\gamma}whE_t\left\{\widehat{w}_{t+1}^\star + \widehat{\Pi}_{t+1} - \widehat{w}_t^\star\right\} \\ &+ \beta(1-\vartheta)JE_t\left\{\widehat{\lambda}_{t+1} - \widehat{\lambda}_t + \widehat{J}_{t+1}^\star\right\}\end{aligned} \tag{99}$$

$$\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 (100)\\
-\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 (101)\\
y\widehat{y}_t &= c\widehat{c}_t+g\widehat{g}_t+\kappa v\widehat{v}_t+\Phi n\widehat{n}_t \qquad (102)\\
\widehat{y}_t &= \widehat{z}_t+\alpha\widehat{h}_t+\widehat{n}_t \qquad (103)\\
\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 (104)\\
\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 (105)\\
\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 (106)\\
\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 (107)\\
\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 (108)\\
\widehat{\epsilon}_t^{money} &= \xi_t^{money}, \quad \xi_t^{money}\overset{iid}{\sim}N\left(0,\sigma_{money}^2\right) \qquad (109)
\end{aligned}$$

$$\widehat{J}_t^{e,n}+\widehat{\delta}_t^{W,e,n}=\widehat{\Delta}_t^{e,n}+\widehat{\delta}_t^{F,e,n}-\frac{1}{1-\eta}\widehat{\eta}_t^{e,n} \qquad (110)$$

$$\widehat{x}_t^L+\widehat{z}_t+\left(\alpha-1\right)\widehat{h}_t^{e,n}=\widehat{w}_t^{e,n} \qquad (111)$$

$$\begin{aligned}
\delta^W\widehat{\delta}_t^{W,e,n} =& \frac{-\alpha}{1-\alpha}wh\left[\frac{-\alpha}{1-\alpha}\widehat{w}_t^{e,n}+\frac{1}{1-\alpha}\left(\widehat{x}_t^L+\widehat{z}_t\right)\right] \\
&- \frac{-1}{1-\alpha}mrsh\left[\frac{\left(-1\right)\left(1+\varphi\right)}{1-\alpha}\widehat{w}_t^{e,n}-\widehat{\lambda}_t+\frac{1+\varphi}{1-\alpha}\left(\widehat{x}_t^L+\widehat{z}_t\right)\right]
\end{aligned} \qquad (112)$$

$$\begin{aligned}
J\widehat{J}_t^{e,n} =& \frac{wh}{\alpha}\left[-\alpha\widehat{w}_t^{e,n}+\widehat{x}_t^L+\widehat{z}_t\right] \\
&+ \beta\left(1-\vartheta\right)JE_t\left\{\widehat{\lambda}_{t+1}-\widehat{\lambda}_t+\widehat{J}_{t+1}^e\right\}
\end{aligned} \qquad (113)$$

$$\begin{aligned}\Delta\widehat{\Delta}_t^{e,n} =& wh\frac{1}{1-\alpha}\left[-\alpha\widehat{w}_t^{e,n}+\widehat{x}_t^L+\widehat{z}_t\right]\\ &-\frac{1}{1+\varphi}mrsh\left[\frac{1+\varphi}{1-\alpha}\left(-\widehat{w}_t^{e,n}+\widehat{x}_t^L+\widehat{z}_t\right)-\widehat{\lambda}_t\right]\\ &+(1-\vartheta-s)\,\beta\Delta E_t\left\{\widehat{\lambda}_{t+1}-\widehat{\lambda}_t+\widehat{\Delta}_{t+1}^e\right\}\\ &-\beta\Delta s\widehat{s}_t\end{aligned} \tag{114}$$

$$-\frac{\kappa}{q}\widehat{q}_t=\beta JE_t\left\{\widehat{\lambda}_{t+1}-\widehat{\lambda}_t+\widehat{J}_{t+1}^n\right\} \tag{115}$$

$$\widehat{w}_t=(1-\vartheta)\,\widehat{w}_t^e+\vartheta\widehat{w}_t^n \tag{116}$$

d) Model with financial frictions

$$\begin{aligned}
y_t &= \frac{C}{Y}c_t+\frac{I}{Y}i_t+\frac{G}{Y}g_t+\frac{C^e}{Y}c_t^e+\cdots+\phi_t^y && (117)\\
c_t &= -r_{t+1}+E_t\{c_{t+1}\} && (118)\\
c_t^e &= n_{t+1}+\cdots+\phi_t^{c^e} && (119)\\
E_t\left\{r_{t+1}^k\right\} &= r_{t+1}-\nu\left[n_{t+1}-(q_t+k_{t+1})\right] && (120)\\
r_{t+1}^k &= (1-\epsilon)\left(y_{t+1}-k_{t+1}-x_{t-1}\right)+\epsilon q_{t+1}-q_t && (121)\\
q_t &= \varphi\left(i_t-k_t\right) && (122)\\
y_t &= a_t+\alpha k_t+(1-\alpha)\,\Omega h_t && (123)\\
y_t &= h_t+x_t+c_t+\eta^{-1}h_t && (124)\\
\pi_t &= E_t\left\{\kappa_p\left(-x_t\right)+\gamma_f\pi_{t+1}+\gamma_b\pi_{t-1}\right\} && (125)\\
k_{t+1} &= \delta i_t+(1-\delta)\,k_t && (126)\\
n_{t+1} &= \frac{\gamma RK}{N}\left(r_t^k-r_t\right)+r_t+n_t+\cdots+\phi_t^n && (127)\\
r_t^n &= \rho r_{t-1}^n+(1-\rho)\varsigma\pi_{t-1}+\epsilon_t^{tn} && (128)\\
g_t &= \rho_g g_{t-1}+\epsilon_t^g && (129)\\
a_t &= \rho_a a_{t-1}+\epsilon_t^a && (130)
\end{aligned}$$

Appendix C

The log-linearized optimality conditions that Karabarbounis and Neiman's (2014) model delivers are:
1) production function

$$\hat{Y}_t=Y^{\sigma/(\sigma-1)}[\alpha k(\hat{A}_{kt}+\hat{k}_t)+((1-\alpha)n\hat{A}_{nt})] \tag{131}$$

2) Labor share

$$\frac{\mu s_L}{1-\mu s_L}\hat{s}_{Lt} + \frac{1}{1-\mu s_L}\hat{\mu}_t = (\sigma-1)(\hat{A}_{kt} - \hat{\mu}_t - \hat{R}_t) \tag{132}$$

3) Definition of return to capital

$$\hat{R}_{t+1} = \frac{1}{R}[(1+r)(\hat{Z}_t + \hat{r}_{t+1}) - (1-\delta)\hat{Z}_{t+1}] \tag{133}$$

4) Definition of the real rate

$$\frac{r}{1+r}\hat{r}_{t+1} = -\gamma(\hat{c}_t - \hat{c}_{t+1}) \tag{134}$$

5) Markup

$$\hat{\mu}_t + \frac{s_L}{s_L+s_k}\hat{s}_{Lt} + (1 - \frac{s_L}{s_L+s_k})\hat{s}_{kt} = 0 \tag{135}$$

6) Capital share

$$\hat{s}_{kt} = \hat{R}_t + \hat{K}_t - \hat{Y}_t \tag{136}$$

7) National identity

$$\hat{Y}_t = \frac{c}{y}\hat{C}_t + \frac{k}{y}(\delta\hat{Z}_{it} + \hat{k}_t - (1-\delta)\hat{k}_{t-1}) \tag{137}$$

8) MPK=real wage

$$\frac{\sigma-1}{\sigma}\hat{A}_{kt} + \hat{y}_t - \hat{k}_t = \hat{\mu}_t + \hat{R}_t \tag{138}$$

9) Labor supply

$$\hat{n}_t = 0 \tag{139}$$

The process for the three exogenous variables are:

$$\log Z_t = \rho_1 \log Z_{t-1} + u_{1t} \quad u_{1t} \sim (0, \omega_1) \tag{140}$$

$$\log A_{nt} = \rho_2 \log A_{nt-1} + u_{2t} \quad u_{2t} \sim (0, \omega_2) \tag{141}$$

$$\log A_{kt} = \rho_3 \log A_{kt-1} + u_{3t} \quad u_{3t} \sim (0, \omega_3) \tag{142}$$

We set $\delta = 0.10$ and $\beta = 0.96$. We estimate $\rho_j, \omega_j, j = 1, 2, 3, \gamma, \sigma$. The prior for σ is truncated normal with mean 1 and standard deviation 0.4; the prior for γ is truncated normal with mean 1 and variance 1; the priors for ρ_j are truncated normal with mean 0.9 and variance 0.4; the prior for ω_j are truncated normal with mean 1 and variance 1. The only common parameter we assume across countries is σ. We have estimated the model also under the assumption that γ is also common without appreciable changes in the posteior of σ. To construct the composite likelihood, data for the five countries receives either equal weight (ω=0.20) or the prior for ω is Dirichlet with mean 0.20. We use 50000 draws after an initial burn-in phase of 10000 draws.