A composite likelihood approach for dynamic structural models

Fabio Canova, BI Norwegian Business School, CAMP, and CEPR
Christian Matthes, Indiana University *

December 20, 2020

Abstract

We explain how to use the composite likelihood function to ameliorate estimation, computational, and inferential problems in dynamic stochastic general equilibrium models. We combine the information present in different models or data sets to estimate the parameters common across models. We provide intuition for why the methodology works and alternative interpretations of the estimators we construct and of the statistics we employ. We present a number of situations where the methodology has the potential to resolve well-known problems and to provide a justification for existing practices that pool different estimates. In each case, 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 the editor (Nezih Guner), the anonymous referees, 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.

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.[1]

Estimation of DSGE models, however, is difficult. 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 (2018); and 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 the presence of latent variables and numerical difficulties are also well-known. Both problems become particularly acute when the model is of large scale or the data is 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 help when the sample size is short, but it is tricky to specify priors for large parameter vectors. As indicated by Del Negro and Schorfheide (2008), assuming prior independence among the components of the parameter vector produces a joint prior that does not reflect researchers' beliefs. Perhaps more importantly, likelihood-based inference is conditional on the estimated model being correctly specified.

Policymakers are keenly aware of these problems and tend to informally pool estimates obtained from different models when choosing policy actions. Furthermore, they find it attractive to use more than one model to robustify counterfactual exercises and to improve medium term forecasting performance, see e.g., Aiolfi et al. (2010).

This paper is concerned with the estimation, computational and inferential problems applied DSGE researchers face. We propose a method that can deal with the challenges mentioned in this introduction. The approach employs the *composite likelihood*, a limited

[1] Andreasen et al. (2018) is an exception, as they focus on Generalized Methods of Moments (GMM) estimation of nonlinear DSGE models.

information objective function, well-known in the statistical literature but sparsely used in macroeconomics, see e.g. Engle et al. (2008), Qu (2018), and Chan et al. (2018).

To give some perspective on what we do, it is useful to state how the existing statistical literature has used composite likelihood methods and how our approach differs. To be specific, we let $L(x, y|\theta, \epsilon)$ be the likelihood of the *known* process generating data (DGP), where the shocks ϵ produce fluctuations in the endogenous variables x, y, given parameter vector θ. In the original work of Besag (1974) and Lindsay (1980), $L(x, y|\theta, \epsilon)$ is analytically intractable, for example, because below a threshold x is not observable, or challenging to compute numerically, because high-dimensional integration is needed. Thus, to estimate θ they suggest to use an alternative objective function that is easier to work with. Their insight is that, in many relevant cases, $L(x|y, \theta, \epsilon)$ and $L(y|x, \theta, \epsilon)$ (the conditional likelihoods) are easy to obtain by, for example, splitting the model in blocks and calculating the likelihood of each block separately. In other situations, $L(x|\theta, \epsilon)$ and $L(y|\theta, \epsilon)$ (the marginal likelihoods) could be calculated. For instance, if $x = [x_1, \ldots, x_t]$ and $y = [x_{t+1}, \ldots, x_T]$, this involves computing the likelihood of each subsample. The objective function they employ for estimation is the composite likelihood, which geometric combines conditional (marginal) likelihoods of submodels (subsamples), where ω_j is an arbitrary weight a researcher assign to the conditional (marginal) likelihood j. Clearly, the composite likelihood is a limited information object, since it neglects, e.g., the fact that the two blocks are part of a unique model, or that x, y may be be correlated, but has nice statistical properties, as it produces consistent and asymptotically normal estimators, and sound inference, see e.g., Varin et al. (2011). For this reason, a composite likelihood approach has been used to solve a number of analytical and computational problems in fields as diverse as spatial statistics, multivariate extremes, psycometrics, and genetics.

This paper demonstrates that a version of the composite likelihood approach is also useful to address many computational, numerical, and inferential problems that plague the empirical DSGE literature. In particular, we show how the framework allows to address population and sample identification and singularity problems. We also highlight how it helps to estimate large scale models, without compromising their general equilibrium nature. Furthermore, we exemplify how the approach helps to robustify estimates and inference. Finally, we illustrate how it rationalizes existing "pooling" practices used in the literature. We provide various intuitive interpretations for the composite esti-

mators and for the composite statistics we construct, and compare our approach with other combination devices present in the policy arena. In general, we are the first to show that composite techniques can be useful in a number of applied macroeconomic applications and provide a toolkit that can be employed in a variety of practical and relevant situations.

The composite setup we employ in this paper differs from the traditional approach in several aspects. First, the DGP and thus the likelihood associated with it are treated as unknown. Second, our composite likelihood combines distinct structural or statistical models, which are not necessarily marginal or conditional partitions of the DGP. Third, models entering the composite likelihood need not be compatible, in the sense that the maximum likelihood estimator of the subset of parameters appearing in all models need not to converge to the same value. Fourth, we employ the methodology to estimate the subset of the structural parameters which are common across models.

To give a specific example, suppose the mechanism generating inflation in the data is unknown and an applied researcher has available, say, two models that explain inflation dynamics, both of which are statistically sufficiently well specified to be potentially usable for estimation. Suppose that a model explains inflation dynamics focusing on macro-financial interactions, while the other exploits the trade-off between firms' market share and pricing frictions. We treat both models as approximation to the true DGP, because each setup generally highlights just one feature that may be present in the data and leaves out other potentially interesting aspects that may account for the data. Because the two models are misspecified in different ways, estimates of the parameters which are common to the two models, say, the elasticity of labor supply, may differ even asymptotically. Thus, policymakers are faced with a difficult situation if, for example, the effectiveness of a fiscal expansion depends on the estimate of elasticity of labor supply. How do we proceed? We compute the likelihood of each model, combine the two likelihoods into a composite likelihood and estimate the parameters common across models jointly using the restrictions the two models provide.

Because of the four differences highlighted above, standard asymptotic results do not apply to our composite estimators. Still, when the weights assigned to the models are fixed, our estimators have desirable properties and may improve over the likelihood estimator of each model as measured by the mean square error (MSE) or by the Kullback-Leibner (KL) divergence. When the weights are random, quasi-Bayesian methods can

be employed to jointly estimate the common parameters and the weights. In this case, the posterior of the weights allows us to rank the quality of the models entering the composite pool, see Canova and Matthes (2018).

We show in a simple example that our methodology exploits the cross equation restrictions appearing in all the models and this helps with identification and singularity, produces robust estimates of structural parameters which are common across models, and provides a way to reduce the dimensionality of a model without drastically affecting estimation. Intuitively, the procedure works because idiosyncratic peculiarities and biases obtained estimating each model separately will wash out as long as the models are sufficiently different in the way they put structure on the data.

It is important to stress that the approach does not force a parameter to have the same value in all models. If models are so different in their structure that a parameter has different interpretations or different features, a researcher may choose not to restrict it to be common across models, even if it has the same name. In that case, composite estimators reduce to likelihood (posterior) estimators, model by model. Perhaps equally important, the approach does not produce mean estimates when parameters have different population values in different models or when different and possibly heterogeneous sources of information are fused in estimation. Instead, it provides estimates for the parameters appearing in a "normalized" unit.

The rest of the paper is organized as follows. The next section introduces the idea, highlights differences with the traditional setting, and gives an interpretation of composite estimators when the weights are fixed quantities. Section 3 presents four examples showing how the methodology can reduce sample and population identification problems, resolve singularity issues, and provide a convenient way to estimate the parameters of a large scale model. Section 4 allows the weights to be random and discusses quasi-Bayesian estimation; it also provides an interpretation of quasi-posterior estimators, and explains how to construct composite posterior statistics. Section 5 has two additional examples. One highlights how to make estimation and inference robust; the other shows how to merge multiple sources of information for the estimation of a crucial structural parameter. Section 6 concludes. Five on-line appendices provide technical details, the equations of the models used in the examples, the results of a small Monte Carlo exercise, and some sensitivity analysis for the first example of section 5.

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 and straightforward to compute, but the dimensionality of the parameter space makes full likelihood computations unappealing. In these situations, it might be preferable to use an objective function with smaller informational content than the full likelihood but that is easier to work with. One such function, originally proposed by Besag (1974) and Linsday (1980), is a weighted average of marginal or conditional distributions of submodels ('events' in the terminology used by this literature). Formally, suppose a known DGP produces a parametric density $F(y_t|\psi)$ for an $m \times 1$ vector of observables y_t, where ψ is a $q \times 1$ vector of parameters. Partition $\psi = [\theta, \eta]$ where, by convention, θ is the vector of parameters estimated with composite methods, and η is a vector of nuisance parameters. Let $\{A_i, i = 1, ...K\}$ be a finite 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 [2]. 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 $\phi = (\theta, \eta_1, \dots \eta_K)$ and $\tilde{y}_t = (y_{1t}, \dots, y_{Kt})$. Given a vector of weights $0 < \omega_i \leq 1$, $\sum_i \omega_i = 1$, the composite likelihood is

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

Clearly, (1) is a misspecified representation of the density of y_t: it ignores the potential dependence across A_i, i.e. that submodels may feature common equations; and the fact that y_{it} may not be mutually exclusive, i.e. the same variable may appear in the vector of observables for different i. Still, the estimator of θ obtained maximizing (1) for a given ω is consistent and asymptotic normal (see e.g. Varin et al., 2011). Intuitively, consistency obtains because each element in (1) is an unbiased limited information object, in the sense that it does not distort the features of the DGP; and geometrically averaging these objective functions will not change this feature. Asymptotic normality holds because the sampling distribution of the maximum likelihood estimator of each submodel

[2] Marginal or conditional distributions 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.

can be approximated quadratically around the same mode. Since each model has the same central tendency but (potentially) different spreads, composite estimators have a quadratic asymptotic distribution around the common mode and spread reflecting the relative precision of the information contained in different submodels.

2.1 A composite DSGE setup

The setup we consider differs from the standard one in several respects. First, we treat the DGP as unknown for many reasons. 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 does not know yet how to link them in a coherent way.

Second, $f(y_{it} \in A_i, \theta, \eta_i)$ are neither marginal nor conditional representations of the DGP. Instead, they are the densities produced by different models a researcher may wish to entertain to study an issue of interest. We assume that all models are relevant, in the sense that they have insights about the phenomena of interest, effectively making K finite and small, but only approximate the DGP, in the sense that for each (θ, η_i) the Kullback-Leibler divergence of $f(y_{it} \in A_i, \theta, \eta_i)$ from $F(y|\psi)$ is strictly positive.

To be concrete, in one leading example we have in mind, A_i are different structural macroeconomic models, e.g., a RBC model with financial frictions, a New Keynesian model with sticky prices, a New Keynesian model with labor market frictions, etc.; y_{it} is the data generated by these models, and $f(y_{it} \in A_i, \theta, \eta_i)$ are the associated densities. Here θ is a vector of structural parameters common to across models, e.g. the risk-aversion coefficient, or the Frisch elasticity, while η_i are either model specific structural parameters, e.g. a loan-to-value ratio, a Calvo parameter; or reduced-form mongrels used to approximate features of the DGP, e.g., the consumption habit parameter. In another leading example, we have in mind $F(y_t|\psi)$ is the density of a large-scale DGP, 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 or by different

order of (perturbed) approximations; or the densities of linear solutions, where the m-th component of parameter vector is time varying, see e.g. Canova et al. (2020). Here $y_{it} = y_t$ and A_i represents the approximation method or the approximation order employed, or an indicator function describing which parameter is allowed to change.

We treat the K models as approximations of $F(y_t|\psi)$ because they disregard aspects of the DGP; they take short cuts to modeling the complexities of the DGP; or condition on features which may be present or absent from the DGP. For each of these models, we assume a researcher can form the likelihood function, using the optimal decision rules and Kalman, particle, or other standard filters [3]. We geometrically average these likelihoods for estimation and inference, just as the composite likelihood literature has averaged marginal or conditionals likelihoods of a known 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 structural 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 DGP and $f(y_{it} \in A_i, \theta, \eta_i)$ the densities obtained when i) data from cross sectional unit i is used; ii) data at a particular aggregation level (e.g. firm, industry, regional, etc.) is employed. It could also be the density obtained using the full sample of data and $f(y_{it} \in A_i, \theta, \eta_i)$ the densities constructed using different subsamples (say, pre-WWI, interwar, post-WWII, etc.). In all these cases, the researcher has one structural model, but different y_{it} measures are available. As it will be clear later, the K measurements can be treated as different "models" and the likelihood obtained with each y_{it} averaged in a composite objective function. Also in this situation, "models" are treated as approximations since, for example, they may neglect the presence of common shocks across i or omit features across time.

A third important difference with the traditional setup is that the models we consider need not be compatible with each other. Compatibility insures that asymptotically, $\theta_{i,ML}$ converges to the same θ for each i and clearly holds when $f(y_{it} \in A_i; \theta, \eta_i)$ are marginals or conditionals. Because of this potential incompatibility, composite estimators need

[3] Alternatively, one could assume that moment conditions are available and use these to form an approximate likelihood for each i, along the lines of Chernozukov and Hong (2003).

not enjoy the standard properties. Nevertheless, following earlier work by White (1982), one can show that the composite likelihood estimator in our setup is consistent for a pseudo value θ_0, which minimizes the distance between $CL(\tilde{y}_t, \phi)$ and $F(y_t, \psi)$ and is asymptotically normal around it. Details are in appendix A.

Researchers working with DSGE models are generally free to choose what goes in θ and in η_i. In particular, even though some parameter might appear in all models, researchers might prefer not to estimate a common value because, for instance, it may have a different interpretation in different models. For example, consider the persistence of the income process ρ_y. When a partial equilibrium perspective is adopted, this parameter is well defined since income is exogenous. When a general equilibrium perspective is employed the persistence of the income process is endogenous and regulated by the persistence of Total factor productivity (TFP) and the dynamics of capital and labor. If partial and general equilibrium models are jointly used in the composite likelihood, imposing one value for ρ_y may be unappealing and a researcher may decide to make ρ_y model specific [4]. A researcher may also leave a parameter model specific, even if it appears in all i, when models have orthogonal structural features. For example, if models i and j have different product market structures, it could be unwise to force commonality in estimation for, say, a markup parameter.

When $\theta \neq \emptyset$, composite estimates of (θ, η_i) are restricted by the information contained in the K models. To see why this is the case note that, for each i, likelihood estimators of (θ, η_i) solve the score conditions $s_i^1(\theta, \eta_i) = \frac{\partial L(\theta, \eta_i)}{\partial \theta} = 0, s_i^2(\theta, \eta_i) = \frac{\partial L(\theta, \eta_i)}{\partial \eta_i} = 0$, and that, for each i, $dim(s_i(\theta, \eta_i)) = dim(\theta, \eta_i)$, $s_i = [s_i^1, s_i^2]$. The score conditions that the composite likelihood estimator of θ solves are $\sum_i \omega_i \frac{\partial L(\theta, \eta_i)}{\partial \theta} = 0$. Thus, a composite estimator mimics an over-identified GMM estimator where ω_i are the weights the orthogonality conditions of model i receive in the composite objective function. When ω_i is fixed, mean square (MSE) gains can be obtained as long as the models entering the composite likelihood are sufficiently idiosyncratic in the way they approximate the DGP. Thus, nested models or models which are similar in all respects but represent, say, product market frictions differently (e.g. Calvo or Rotemberg pricing) will not provide the variety needed for improvements to materialize. Notice that when composite estimation is used and $(\theta, \eta_1, \ldots, \eta_K)$ are jointly estimated $\frac{\partial CL(\theta, \eta_i)}{\partial \eta_i}$ differs from $\frac{\partial L(\theta, \eta_i)}{\partial \eta_i}$, leading to

[4] Even in this case, one can assume the same prior distribution for ρ_y across models, thus guaranteeing some a-priori model compatibility without imposing that the parameter speaks to similar economic concepts.

different η_i estimators.

When $\theta = \emptyset$, $\phi_i = \eta_i$ and composite estimation replicates likelihood estimates of η_i, separately for each i. Thus, the composite likelihood does not impose restrictions other than those of likelihood of each model, when there is no common parameter.

At times "models" may feature common parameters, but their population values may be different. For example, an elasticity of substitution may appear in a model which is used to explain the dynamics in a set of countries, but the population value it assumes in each country may be different. As discussed in section 5, the procedure we employ is applicable also in this case, as long as one model is used for normalization and the others appropriately reparametrized.

3 Addressing estimation and computational problems

This section shows how the composite likelihood may help 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 apply generally.

The first example discusses how small sample identification problems can be resolved using the composite likelihood constructed using different structural models. The intuition applies also to situations when different statistical models are used or when there is a single model and the composite likelihood is constructed 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 example shows how to estimate the parameters of a large-scale structural model. In all the examples, we treat ω_i as fixed; there is at least one parameter common across models; and theory tells us that it has the same interpretation and the same value across models.

3.1 Reducing sample identification problems

In macroeconomics it is common to work with relatively small samples of time series. Long data is generally unavailable and, when it exists, definitional changes or structural breaks make it unwise to use the full sample for estimation purposes. In addition, the phenomena of interest (say, the effects of the zero lower bound on interest rates) may

be present only in the most recent portion of the sample. We show how the composite likelihood could reduce the severity of small sample problems.

Consider two structural models (say, 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 the observable y_t. Assume that y_{At} and y_{tB} are produced by the decision rules:

$$y_{At} = \rho_A y_{At-1} + \sigma_A e_t \tag{2}$$

$$y_{Bt} = \rho_B y_{Bt-1} + \sigma_B u_t \tag{3}$$

where e_t and u_t are both iid (0,I). While (2)-(3) are chosen for the sake of exposition, it is worth emphasizing that the linear solution of a DSGE model has the same format, where y_t includes the states and the controls and $\psi_i = \psi(\gamma_i)$ are functions of the structural parameters γ_i. Thus, the conclusions we derive are applicable to a large class of models.

Suppose that $\rho_B = \delta\rho_A, \sigma_B = \gamma\sigma_A$, where $(\delta \neq 0, \gamma \neq 0)$; let y_{At} and y_{Bt} be scalars; assume we have $T_A(T_B)$ observations on y_{At} (y_{Bt}) with T_A small, and that we care about $\theta = (\rho_A, \sigma_A)$. The (normal) log-likelihood functions of each model are:

$$\log L_i \propto -T_i \log \sigma_i - \frac{1}{2\sigma_i^2} \sum_{t=1}^{T_i} (y_{it} - \rho_i y_{it-1})^2 \quad i = A, B \tag{4}$$

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

$$\log CL = \omega \log L_A + (1 - \omega) \log L_B \tag{5}$$

We assume the DGP has $\rho_A = 0.7, \sigma_A = 1.0$, $\delta = 1.2$, $\gamma = 0.8$ (so that $\rho_B = 0.84, \sigma_B = 0.8$) and choose $T_A = 20, T_B = 20$. In the composite likelihood we set $\omega = 0.7$. Figure 1 plots the univariate contours of (4) and (5) in the (ρ_A, σ_A) dimensions. Note that the composite likelihood has more curvature then the likelihood constructed using y_{At} only and that its mode is closer to the true vector. Moreover, when T_B increases $(T_B = 60)$, the composite likelihood becomes more bell-shaped around the true value and almost symmetric in shape.

As shown in section 5, differences in the estimates obtained with (4) and (5) have to do with three quantities $\zeta_1 = \frac{1-\omega}{\omega}\frac{\delta}{\gamma^2}$, $\zeta_2 = \frac{1-\omega}{\omega}\frac{\delta^2}{\gamma^2} \equiv \zeta_1\delta$, and $\xi^{-1} = (T_A + T_B\frac{1-\omega}{\omega\gamma^2})$. ζ_1 and ζ_2 control the shape of the composite likelihood, while ξ^{-1}, the effective sample size, controls both the height and the shape of the composite likelihood. In turn, these quantities depend on (ω, γ, δ). Thus, these parameters regulate the amount of information that

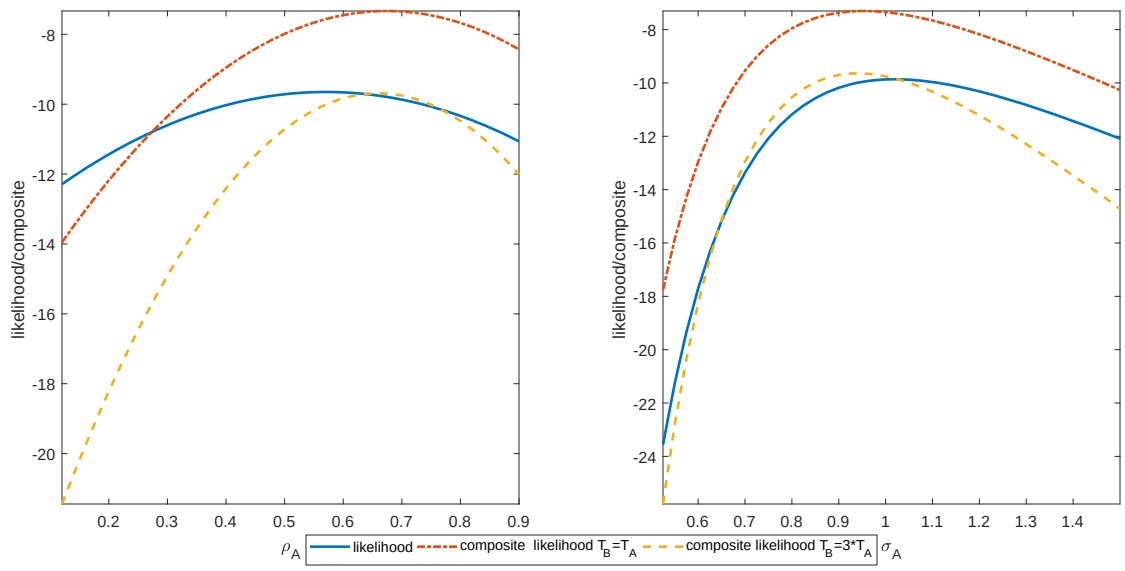

Figure 1: Likelihood and composite likelihood, small T.

y_{Bt} provides for (ρ_A, σ_A). For example, with the parameterization used, the effective sample size is $T_A + 0.67T_B$, making (5) higher than (4). Other things being equal, increasing γ makes y_{Bt} less informative - model B provides noisy information for (ρ_A, σ_A) - and decreasing δ, reduces the informational content of y_{Bt} because it becomes less persistent. Thus, the composite likelihood cares about 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: the higher the serial correlation, the more important low frequency information is; and the lower the standard deviation, 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 the composite shape away from the true values. Thus, identification improvements are not guaranteed. In this example, better local identification could be attained when (y_{At}, y_{Bt}) are jointly used in estimation if ω, γ, and T_B are such that $\xi^{-1} > T_A$ and ζ_1, ζ_2 are different from zero. If γ is small, that is, if y_{Bt} is less volatile than y_{At}, or if ω is not too large, that is, if the degree of trust a researcher has in model B is not negligible, (5) will be more peaked around the mode than (4).

The same argument applies when, rather than structural models, A and B are two statistical models, or when there is a single structural model, but y_{At} and y_{Bt} represent

the same y_t in different samples. When A and B are statistical models, information coming from different time series may make the composite likelihood more peaked around the true value than each likelihood, much in the same spirit as a data-rich approach to estimation may provide more precise information about structural parameters than a standard approach (see e.g. Boivin and Giannoni, 2006). When y_{At}, y_{Bt} are different samples for the same variable say, post-break and pre-break data, the composite likelihood may be more informative about θ than the likelihood of the post-break sample as long as the weights are appropriately chosen. Intuitively, rather than dropped, pre-break information is weighted to trade-off sharpness and distortions in the objective function. Thus, y_{Bt} plays the role of a " training" sample and composite estimates combine likelihood estimates produced by y_{At} with objective prior estimates obtained with Y_{Bt}. Baumeister and Hamilton (2019) suggested a procedure, altering the information contained in earlier subsamples relatively to the current one, that closely mimics a composite likelihood setup when $\omega \neq 1 - \omega$.

Note 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). When two such models are combined, the effective sample size ξ^{-1} will generally increase, making the composite likelihood more peaked and more concentrated that the likelihood of, say, the quarterly model. For this to happen, it is sufficient to have ω, δ, γ such that ζ_1 and ζ_2 are different from zero.

To conclude there are many ways to reduce small sample (local) identification problems with the composite likelihood: one could use different structural models, different data, or the same data in different samples or at different frequencies. In all these situations, if the additional data is informative and the weights appropriately chosen, the composite likelihood has better properties than the likelihood for the θ vector.

3.2 Ameliorating population identification problems

This subsection presents an example where estimation is difficult because some parameters are underidentified and others weakly identified *in population* and shows that a composite approach can remedy these problems.

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

$$R_{At} = \tau E_t \pi_{At+1} + e_{1t} \tag{6}$$

$$y_{At} = \delta E_t y_{At+1} - \sigma(R_{At} - E_t \pi_{At+1}) + e_{2t} \tag{7}$$

$$\pi_{At} = \beta E_t \pi_{At+1} + \gamma y_{At} + e_{3t} \tag{8}$$

where R_{At} is the nominal interest rate, y_{At} the output gap, and π_{At} the inflation rate;(e_{1t}, e_{2t}, e_{3t}) are mutually uncorrelated disturbances, $(\tau, \delta, \sigma, \beta, \gamma)$ are structural parameters, and E_t is the conditional expectations operator. The determinate solution of (6)-(8) is

$$\begin{bmatrix} R_{At} \\ y_{At} \\ \pi_{At} \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 \\ \sigma & 1 & 0 \\ \sigma\gamma & \sigma & 1 \end{bmatrix} \begin{bmatrix} e_{1t} \\ e_{2t} \\ e_{3t} \end{bmatrix} \equiv A e_t. \tag{9}$$

Clearly, β is not identified, as it disappears from (9); and the slope of the Phillips curve γ may be hard to identify from the likelihood of $(R_{At}, y_{At}, \pi_{At})$. In fact, if σ is sufficiently small, large variations in γ may induce small variations in the decision rules (9), making the likelihood flat in the γ dimension.

Population non-identification of β implies, for example, that when (6)-(8) generate the data, 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 output gap changes on inflation, regardless of the magnitude of the 'true' Phillips curve slope. Problems of this type are common in DSGE models (see Canova and Sala, 2009).

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

$$\pi_{Bt} = \beta E_t \pi_{Bt+1} + \gamma y_{Bt} + u_{2t} \tag{10}$$

$$y_{Bt} = \rho y_{Bt-1} + u_{1t} \tag{11}$$

where $\rho > 0$ measures the persistence of the output gap process. (10) has the same format as (8), 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 (10)-(11) is

$$\begin{bmatrix} (1-\rho\ell) y_{Bt} \\ (1-\rho\ell)\pi_{Bt} \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ \frac{\gamma}{1-\beta\rho} & 1-\rho\ell \end{bmatrix} \begin{bmatrix} u_{1t} \\ u_{2t} \end{bmatrix}. \tag{12}$$

Clearly, unless $\rho = 0$, the log-likelihood of model B has information for 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, small variations γ may lead to sufficiently large variations in the decision rule (12) and thus in the composite likelihood. Note that both improvements occur even when $1 - \omega$ is small.

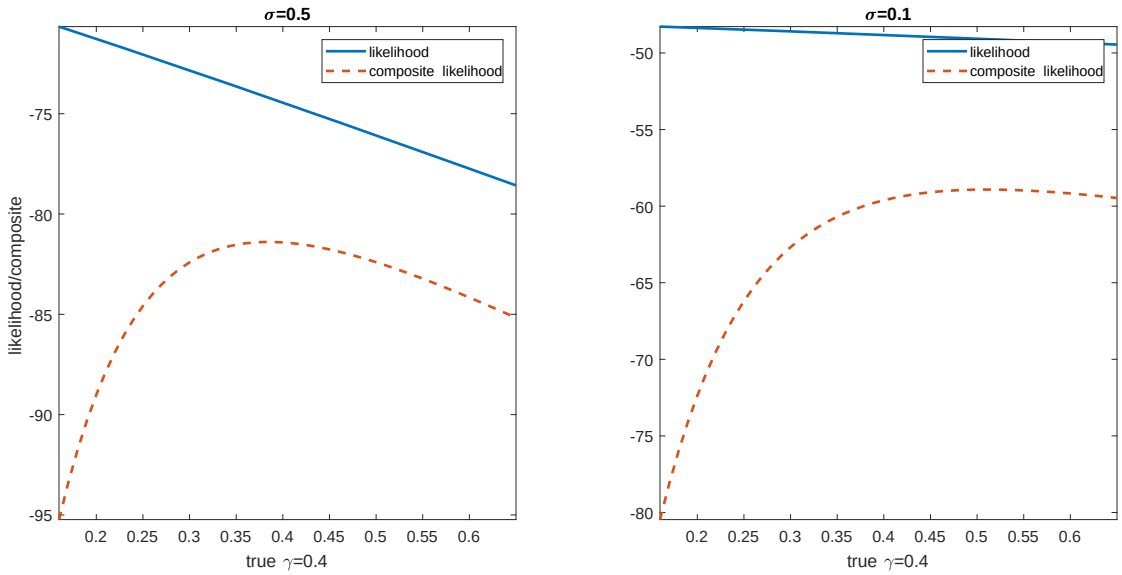

Figure 2: Likelihood and composite likelihood, weak identification.

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, when we use the composite likelihood, we set $\omega = 0.85$. As expected, the likelihood of model A is flat around the true value of γ when σ is small, and adding information from model B helps to improve the identification of γ. A similar situation arises when $\sigma = 0.5$ as the likelihood constructed from y_{At} is not quadratic in γ. Adding model B information makes the composite likelihood close to quadratic.

The point we make here is independent of the effective sample size ξ^{-1}. Since the identification problems we discuss occur in population, having a large or a small ξ^{-1} 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 information present in the decision rules is weighted against the relative noise contained in the two models. When noise in model B is large, the composite curvature approaches the likelihood curvature.

It should be obvious that a random selection of model B is unlikely to provide population identification gains. In particular, adding models with Phillips curves that are non-comparable to those of model A will not work. This would be the case, e.g., if γ is generated by a different mechanism (for example, via the market share model of Gilchrist et al., 2017), effectively making it a different parameter; or if the biases introduced with model B are large relative to the improved curvature. Hence, population identification improvements can be obtained only after carefully examining the structure and the likelihood shape of the additional model(s) one may want to consider.

In sum, a shrewd use of the composite likelihood may improve parameter identification when the sample is short or when parameters are weakly identified in population. For this to happen, the additional models (the additional data) must add 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 that applied investigators plot the likelihoods and composite likelihood, as we have done in Figures 1 and 2, as a routine practice. It will help to understand which models may help to regularize the objective function.

3.3 Solving singularity problems

DSGE models are typically singular. That is, since they generally feature more variables than shocks, the theoretical covariance matrix of the endogenous variables 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 endogenous variables as 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 so as to make the number of shocks (structural and measurement) larger or equal to the number observables (see Ireland, 2004). One could also artificially increase the number of 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 the singularity problem is to construct a composite likelihood weighting non-singular submodels, see also Qu (2018). 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 the stock prices is $p_t = (1 - \beta\alpha)e_t - \alpha e_{t-1}$, where $\beta < 1$ is the discount factor of the investor. Since e_t drives both dividends and stock prices, the covariance matrix of (d_t, p_t) has rank one. Thus, either d_t or p_t must be used to construct the likelihood and to estimate $\theta = (\alpha, \sigma^2)$.

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 contain different information than dividends. Choosing one particular variable for estimation, throws away valuable information. By combining all equations, the composite likelihood provides sharper estimates of θ.

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

$$\log L(\alpha, \sigma^2|\tilde{d}_t) = -0.5T\log(2\pi) - \sum_{t=1}^{T}\log\varsigma_t - 0.5\sum_{t=1}^{T}\frac{\tilde{d}_t^2}{\varsigma_t^2} \tag{13}$$

where $\tilde{d}_t = d_t - \alpha\frac{1+\alpha^2+\alpha^4+...+\alpha^{2(t-2)}}{1+\alpha^2+\alpha^4+...+\alpha^{2(t-1)}}\tilde{d}_{t-1}$, $\varsigma_t^2 = \sigma^2\frac{1+\alpha^2+\alpha^4+...+\alpha^{2t}}{1+\alpha^2+\alpha^4+...+\alpha^{2(t-1)}}$ and

$$\log L(\beta, \alpha, \sigma^2|\tilde{p}_t) = -0.5T\log(2\pi) - \sum_{t=1}^{T}\log\upsilon_t - 0.5\sum_{t=1}^{T}\frac{\tilde{p}_t^2}{\upsilon_t^2} \tag{14}$$

where $\tilde{p}_t = p_t^* - \gamma\frac{1+\gamma^2+\gamma^4+...+\gamma^{2(t-2)}}{1+\gamma^2+\gamma^4+...+\gamma^{2(t-1)}}\tilde{p}_{t-1}$, $\upsilon_t^2 = \sigma^2\frac{1+\gamma^2+\gamma^4+...+\gamma^{2t}}{1+\gamma^2+\gamma^4+...+\gamma^{2(t-1)}}$ and $\gamma = \frac{\alpha}{(1-\beta\alpha)}$ and $p_t^* = \frac{p_t}{1-\beta\alpha}$. For illustration, we set $\sigma^2 = 1$, $\beta = 0.99$, and focus attention on the maximum and composite likelihood estimation of α. While there are no closed form expressions for maximum likelihood or maximum composite likelihood estimators, we can still infer what (13) and (14) employ to estimate α using artificial data.

Figure 3 plots the likelihood contour in the α dimension, when (13), (14), or the composite likelihood are used, and the true α is either 0.7 or 0.1. When $\alpha = 0.1$ (13) and (14) are similar. Thus, when dividends and stock prices are almost serially

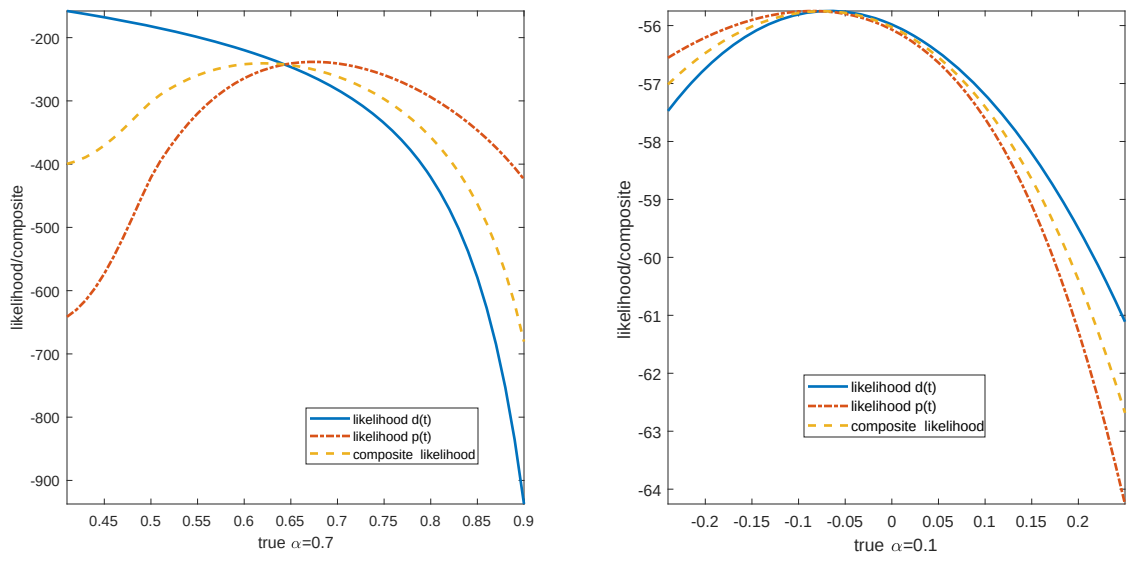

Figure 3: Likelihood and composite likelihood, singularity.

uncorrelated, they have similar information and the shape of the likelihood functions primarily reflects the volatility of e_t. When $\alpha = 0.7$, 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 which is absent from the likelihood of dividends.

The composite likelihood, constructed using $\omega = 0.5$, captures both the serial correlation and the variability properties of the DGP, it is more bell shaped than each likelihood, and centered around the true value when $\alpha = 0.7$. In fact, when T=500, $\alpha_{ML} = 0.4$ (standard deviation 0.1) and $\alpha_{CL} = 0.65$ (standard deviation 0.06). Clearly, the value of ω regulates whether the serial correlation, the variance properties of (d_t, p_t), or both matter for estimation. When $\alpha = 0.1$, (13) and (14) have similar information. Thus, neither the shape nor the location improves when the composite likelihood is used.

In general, when the equations of a singular model provide separate information about the parameters of interest, it is a-priori difficult to choose which ones to use in estimation. The composite likelihood eliminates the dilemma, weighting the information contained in all equations in a meaningful way.

3.4 Estimating large scale structural models

While in academics models are kept small to enhance intuition, large scale models are common in policy institutions. Large models are more detailed and realistic, but estimating their parameters is computationally a daunting task and estimates obtained are often unreasonable. Thus, the models used in the policy process are often informally calibrated, apart for the parameters of the shock processes, which are estimated by likelihood methods. We show how the composite likelihood can make the estimation of the structural parameters of a large scale model possible and 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 (y_t) = dim (e_t). The likelihood function is

$$L(\theta|y_t) = (2\pi)^{-T/2}|\Sigma(\theta)|^{T/2}\exp\{(y_t - A(\theta)y_{t-1})\Sigma(\theta)^{-1}(y_t - A(\theta)y_{t-1})'\} \tag{15}$$

If $dim(y_t)$ is large, computation of $\Sigma(\theta)^{-1}$ may be demanding. Furthermore, if some of the elements of y_t are nearly collinear or if there are near singularities due, for example, to the presence of an expectational link between long and short term interest rates, numerical difficulties may emerge. (15) is also difficult to compute when there are latent endogenous variables. If $y_t = (y_{1t}, y_{2t})$, and y_{2t} is non-observable, the likelihood of y_{1t} is

$$L(\theta|y_{1t}) = \int L(\theta|y_{1t}, y_{2t})dy_2 \tag{16}$$

which may be intractable when y_{2t} is of large dimension.

Rather than using (15) or (16) for estimation, one can take a limited information point of view and estimate the parameters using objects that are simpler to construct (see also 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:

$$CL_1(\theta|y_{it}) = \sum_{i=1}^{K}\omega_i \log L(\theta|y_{it}) \tag{17}$$

$$CL_2(\theta|y_{it}) = \sum_{i=1}^{K}\omega_i \log L(\theta|y_{it}, \bar{y}_{-it}) \tag{18}$$

where y_{-it} indicates any combination of the vector y_t, excluding the i-th combination, and bars indicate a given value. The first expression averages marginal likelihoods (integrating out all variables but y_{it}), whereas the second averages conditional likelihoods.

CL_1 is obtained by neglecting the correlation structure among y_{it}. Thus, blocks of equations are treated as if they provide independent information for θ, even though this is generally false. For example, in a multi-country 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 weight them for estimation purposes. Which composite likelihood one uses depends on the problem and the tractability of conditional vs. marginal likelihoods.

To compare how the composite likelihood relates to the full likelihood of a particular model, we consider a simple consumption-saving problem where there are many countries, indexed by i, 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 processes are transitory is

$$\begin{aligned} c_{it} &= \frac{r}{r+1}a_{it} + \frac{r}{1-\rho+r}Y_{it} && (19)\\ a_{it+1} &= (1+r)(a_{it}+Y_{it}-c_{it}) && (20)\\ y_{it} &= \rho y_{it-1} + \sigma_i e_{it} && (21)\\ Y_{it} &= \sum_{j=1}^{K}\zeta_{ij}y_{jt} && (22) \end{aligned}$$

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

Suppose that rather than constructing the likelihood using (19)-(22) jointly for the K countries, one constructs the likelihood of the model of each country (i.e. neglecting (22) and using y_{it} in place of Y_{it} in the first two equations) and sets $\omega_i = 1/K$ to construct a composite likelihood. Three types of distortions are present: consumption and asset holdings are functions of domestic income, rather than total income; the volatility of domestic income is higher than the volatility of total income; ω should be a function of ζ_{ij} rather than constant. Clearly if $\zeta_{ij}=\zeta_i=1, \forall j$, and the volatility of y_{it} is the same in all i, the information loss is minimal.

Figure 4 plots the shape of the full likelihood and the composite likelihood in the ρ dimension using consumption data only when T=1000, $K=3, \beta=0.99, \rho=0.6, \sigma_i=$

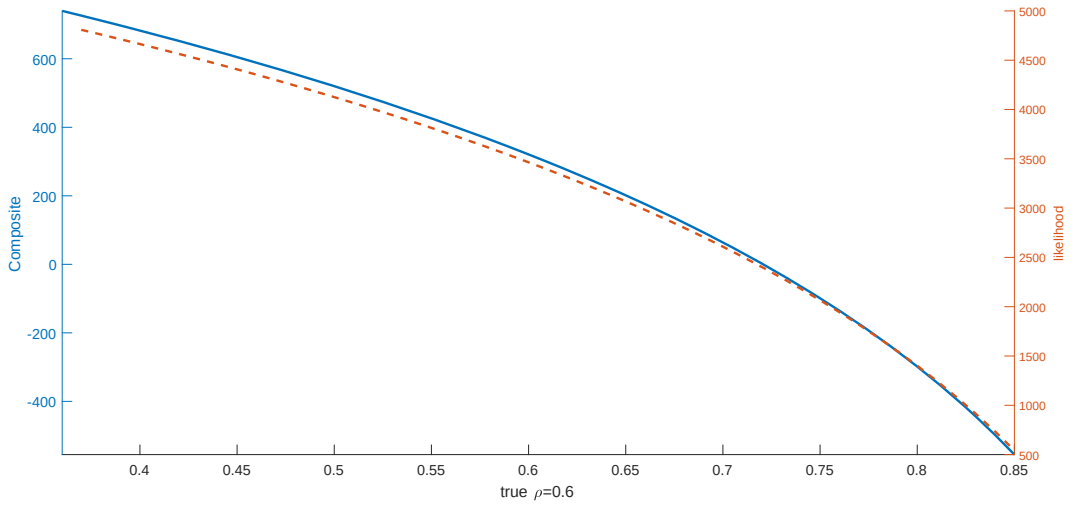

Figure 4: Likelihood and composite likelihood, large scale model.

$[0.1, 0.2, 0.3]$, $r = 1/\beta - 1, \beta = 0.99$, $\zeta = \begin{bmatrix} 0.5 & 0.25 & 0.25 \\ 0.25 & 0.5 & 0.25 \\ 0.25 & 0.25 & 0.5 \end{bmatrix}$. The likelihood function is not quadratic in ρ since the marginal propensity to consume out of transitory income increases as ρ moves from -1 to 1, and the composite likelihood inherits this property. However, although the scale is different, the two functions have similar shapes making the information loss from using the composite likelihood small. In fact, in the setup we use $\rho_{ML} = 0.71$ and $\rho_{CL} = 0.75$, both with a standard deviation equal to 0.02.

4 Quasi-Bayesian estimation

The improvements in the shape of the objective function and computation gains described in section 3 can be obtained even when ω is fixed and the prior for the parameters play no role. However, it is unusual in the applied literature to estimate structural DSGE parameters with frequentist methods because the likelihood is poorly behaved (multiple modes and sharp cliffs are common) and samples are sufficiently short to make asymptotic approximations not very credible. Furthermore, some justification for the selected ω vector is needed and, at the minimum, one needs to show how the conclusions are affected by the choice of ω.

Rather than using a frequentist approach, conditional on a given ω, we use a quasi-

Bayesian approach to construct the joint posterior distributions of $(\theta, \eta_i, \omega_i, i = 1, \ldots, K)$. Such an approach may be preferable, even if the priors for the extended parameter vector are loose, for a number of reasons. As discussed in section 4.1, quasi-Bayesian estimators have a simple and intuitive interpretation. Furthermore, as explained in section 4.3, the composite statistics we construct are pooling devices with desirable statistical features. Perhaps more importantly, when ω_i is random, the quasi-posterior mode of ω_i can be used to rank the quality of the models entering (1), much in the same spirit as a Bayesian model averaging (BMA) weight. As highlighted in section 4.2, the quasi-posterior mode of ω can be employed in situations when BMA can not be computed and a measure of uncertainty can attached to the rankings.

Intuitively, the posterior mode of ω_i can serve as a ranking device because the sample information relevant for calculating the mode reflects the relative one-step mean square forecast error of different models. Thus, when $y_{it} = y_t$ for all i, the posterior of ω assigns higher weights to models closer to the DGP in a Kullback-Leibner sense, see Canova and Matthes (2018) for the details. When $y_{it} \neq y_{jt}, i \neq j$, the weights and the parameters will be jointly selected so that the composite model approximate as best as possible the DGP, again in a Kullback-Leibner sense, see section 4.3.

We specify the prior for the parameters of model i as:

$$p(\theta, \eta_i) = p(\theta)p(\eta_i|\theta, y_{i0}). \tag{23}$$

where, in the spirit of Del Negro and Schorfheide (2008), we allow the prior for η_i to depend on θ and some training sample data y_{i0}. If $p(\omega) \equiv p(\omega_1, \ldots \omega_K)$ is the prior for ω, the composite posterior kernel is:

$$\begin{aligned} \check{p}(\theta, \eta_1, ...\eta_K, \omega_1, \ldots, \omega_K|Y_{1,t_1}, \ldots, Y_{k,T_k}) \quad &= \\ p(\theta)p(\omega)\Pi_i \mathcal{L}(\theta, \eta_i|Y_{i,T_i})^{\omega_i} p(\eta_i|\theta)^{\omega_i}, \end{aligned} \tag{24}$$

which can be used to obtain quasi-posteriors for (ϕ, ω), as in Kim (2002) or Chernozukov and Hong (2003). Thus, even though the composite likelihood is not a likelihood function, (24) permits credible prior updating. Bissiri et al. (2016) demonstrated that a valid update of prior beliefs to a posterior can be made for parameters which are connected to observations through a general loss function. Our composite likelihood is precisely one of these loss functions.Seen through this lens, our Bayesian composite setup is also similar to the framework used in the Bayesian misspecification literature, see e.g.

Walker (2012), and standard results regarding the asymptotic negligibility of the prior, the consistency of the posterior mode for the KL minimizer, and asymptotic normality of quasi-posteriors apply under regularity conditions, see e.g. Fernandez Villaverde and Rubio Ramirez (2004); Klein and Van der Vaart (2012); Clydec and Iversen (2013).

Since there is no closed form solution for the posteriors of the parameters, we use a multiple-block Metropolis-within Gibbs algorithm to compute posterior sequences. The algorithm is summarized in appendix B, together with the adjustments one may want to implement to make sure that MCMC and frequentist percentiles asymptotically match (see also Ribatet et al. 2012, Mueller, 2013, Qu 2018).

4.1 Interpreting quasi-posterior estimates

While it is standard to combine models for forecasting, it is less common to use them jointly for estimation and inference. Thus, it is worth mentioning that the composite posterior kernel (24) and the quasi-Bayesian estimators for θ one obtains have a simple sequential, adaptive learning interpretation which should make them attractive to applied macroeconomists.

For the sake of illustration, let ω_i be fixed, and K=2. Then

$$\check{p}(\theta, \eta_2, \eta_1 | Y_{1,T_2}, Y_{2,T_1}) = \mathcal{L}(Y_{2,T_2}|\theta, \eta_2)^{\omega_2} p(\eta_2|\theta)^{\omega_2} \{[p(\theta|Y_{1,T_1}) ML(Y_{1,T_1})]^{\omega_1} p(\theta)^{\omega_2}\} p(\eta_1|Y_{1,T_1}\theta)^{\omega_1} \quad (25)$$

where $ML(Y_{1,T_1}) = \int \mathcal{L}(Y_{1,T_1}|\psi_1)p(\psi_1)d\psi_1$ is the marginal likelihood of model 1 and $\psi_1 = (\theta, \eta_i)$.

As (25) makes clear, the posterior kernel can be obtained in two stages. In the first stage, the prior for ψ_1 and the likelihood of model 1 are used to construct $\mathrm{p}(\theta|Y_{1,T_1})$ and $p(\eta_1|Y_{1,T_1})$. The conditional posterior of θ, weighted by the marginal likelihood of the model 1, is geometrically combined with $p(\theta)$ to form a new prior for θ for the next estimation stage. Suppose that $ML(Y_{1,T_1})$ is high. Then model 1 fits Y_{1,T_1} well. If $\omega_1 = \omega_2$, the prior for θ in stage 2 will reflect more heavily $p(\theta|Y_{1,T_1})$ than $p(\theta)$. On the other hand, if $ML(Y_{1,T_1})$ is low, $p(\theta|Y_{1,T_1})$ has low weight relative to $p(\theta)$ when setting up the prior for the second stage. In general, at each stage of the learning process, the prior for θ depends on the relative weights assigned to the current and to all previous models and on their relative fit for θ. Thus, a composite quasi-posterior approach 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.

Contrary to standard learning algorithms, the prior for stage 2 is not the posterior for stage 1 but rather a weighted average of the initial prior and of the posterior of stage 1, where the latter is discounted by its fit. This is why the approach is adaptive. Note that with a composite setup a model is automatically discounted if does not fit the data well, even when ω is fixed. Del Negro et al. (2016) have shown that finite mixtures have this property only if ω is random. Finally, even though only Y_{1,T_1} contains information for η_1, its posterior may be updated when using Y_{2,T_2} since the posterior of θ sequentially changes. Since Y_{1,T_1} does not contain information for η_2, $p(\eta_2|\theta)$ will be unchanged after estimation at stage 1.

When ω is not treated as fixed, the expressions in (25) are cumbersome to derive, since one needs to take the prior for ω into account, but the same intuition applies.

4.2 Discussion

Our quasi-Bayesian composite estimates differ from Bayesian model average (BMA) estimates in several respects. In BMA, each model is estimated separately and estimates of the parameters combined using posterior weights; in our setup, parameters which are common and have the same meaning in all models are estimated with the information provided by all models, while idiosyncratic parameters are estimated with the information contained in the model where they appear. Furthermore, to apply BMA techniques we need to assume $y_{1t} = \ldots = y_{Kt}$, and that the frequency of the data is the same, while this is not required in our setup. Finally, one can not give BMA estimates a sequential learning interpretation. When $T_i = T$ and $y_{it} = y_t$ for all t, BMA and ω weights have a similar meaning and reflect the relative information each model has for y_t. In this case, differences in the statistics produced by the two pooling devices are due to the different information sets used to estimate the parameters. Our quasi-Bayesian composite estimates also differ from naive ex-post combinations which assign equal weights to the estimates of all models, or other ex-post combination devices. Section 5.1 provides some evidence on the properties of different pooling devices.

What happens to quasi-posterior estimators when an 'irrelevant' or a 'nested' model enters the composite likelihood? As we have seen irrelevant models are downweighted since they poorly fit y_{it} and the same will be true of models which are nested in another

and poorly fit the data. A composite estimator tries to identify regions of the parameters space that are consistent with the data and *all* available models and, as we explain in the next subsection, trades off various models' information to achieve the best possible fit. If a model with poor or no information for θ enters the composite likelihood, it will not contribute to the estimation of θ. Thus, while our approach is motivated by the fact that researchers often have a number of theoretical relevant models usable to estimate the parameters of interest, all of which have roughly comparable fits, it is robust to the inclusion of irrelevant models.

While it does not happen in any of the examples we study, it may be the case that in some applications the posterior weight for some model goes to zero, implying that the nuisance parameters of that model become under-identified when the composite likelihood is used for estimation. When this occurs, 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 trick effectively avoids under-identification and makes the priors for the nuisance parameters endogenous.

4.3 Inference

Once composite estimates of (θ, η, ω) are available, one can pick the model with the highest ω mode for inference or produce statistics which combine the the information contained in various models.

Let $m_{it} = h(y_{it}, \theta_{CL}, \eta_{i,CL})$ be a statistic computable from model i, where h is a continuous differentiable function, and $\theta_{CL}, \eta_{i,CL}$ composite estimates of the parameters. m_{it} could be a future value of y_{it}, a counterfactual path of y_{it}, or the response of y_{it} to a structural shock. Rather than choosing the m_{it} produced by the "best" model, one could robustify inference by computing $\tilde{M}_t = \prod_{i=1}^{K} \tilde{m}_{it}^{\bar{\omega}_i}$, where $\bar{\omega}_i$ is, for example, the posterior mode of ω_i. $\tilde{M}_t$ weights the outcomes of each model by their relative posterior probabilities and may be superior to using one $\tilde{m}_{it}$ in situations of instabilities or structural breaks. [5]

[5] Alternatively, h could be the model specific density of the object of interest (such as an impulse response) and one could integrate the geometric average of this density with respect to the composite posterior (see Canova and Matthes, 2018). We use this latter approach in section 5.1, as it naturally delivers an estimate of the entire distribution of the object of interest. The general approach we describe, on the other hand, is particularly useful in situations where the model implied density is costly to compute (as, e.g., in non-linear models).

One may wonder what is the meaning of $\tilde{M}_t$ and how it relates to the true statistic m_t. $\tilde{M}_t$ has two simple and natural interpretations. The first comes from noticing that since $\tilde{m}_{it}$ is the Bayesian prediction made by model (agent) i, given the information in y_{it} and the composite estimates, $\tilde{M}_t$ is an "opinion" pool where the predictions made by each model (agent) are combined by a Bayesian planner assigning to each prediction weights that reflect the relative ability of each model to explain y_{it}. The problem of combining disparate pieces of probabilistic information is well studied in the literature and, as shown by Genest and Zidek (1986), the logarithmic pooling we employ preserves external Bayesianity (the property that updating and pooling are interchangeable) without imposing independence among agents' opinions.

$\tilde{M}_t$ can also be thought as an approximation to the true m_t constructed using the "messages" that each model sends about the latent variable m_t. Different models send $\tilde{m}_{it}$ messages using potentially different data. These messages are aggregated assuming that the statistical dependences between them is unknown. As shown by Roche (2019), the aggregation scheme we employ happens to be optimal in an information-theoretical sense, i.e. the log-linear pool minimizes the average Kullback-Leibler divergence to the probabilistic opinions. Thus, the composite statistics we construct are the best consensus mechanism according to a natural criterion among differing agents.

Note that what we are proposing differs from ex-post averaging models' predictions computed with likelihood (posterior) estimates of the parameters $M_t^\dagger = \prod_{i=1}^K m_{it}^{\dagger\,\omega_i} = \prod_{i=1}^K h(y_{it}, \theta_{ML}, \eta_{i,ML})^{\omega_i}$, where ω_i could be arbitrary, forecast, or posterior-based weights. $\tilde{M}_t$ and $M_t^\dagger$ differ because $\psi_{ML} \neq \psi_{CL}$ and the weights will differ.

5 Addressing inferential problems

This section demonstrates how the quasi-Bayesian methodology developed in section 4 helps in inferential exercises. In the first subsection we show how to robustify the estimation of structural parameters and how to produce statistics which formally combine the information contained in models with potentially different observables. The second subsection shows how the methodology may be used to partially pool the information contained in panels of data with potentially heterogeneous dynamic features. The final subsection discusses additional applications where the methodology could help researchers to resolve well known conundrums.

5.1 Robustifying parameter estimates and efficiently combining predictions

Likelihood-based estimates of structural parameters are rarely 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 a number of models into account, before historical decompositions or counterfactual exercises are computed. Such a practice is consistent with the idea that available models are approximations, that information not used in estimation ("judgment") could be important when evaluating the appeal of certain policy choices, and that averaging may reduce misspecification biases.

In practice, when a number of models are available to answer a specific question, two approaches are used in the literature. One either estimates the models separately, conducts policy experiments in each model, and then averages the outcomes using user-selected weights; or constructs one counterfactual exercise using informally averaged estimates of the structural parameters in the "most-likely"model, for example, the one with the largest marginal likelihood. Clearly, the two approaches need not produce the same answer. Furthermore, in neither case, the information present in different models is used to estimate common structural parameters.

This section shows that a composite posterior approach robustifies estimates of the common parameters, by formally combining the information about them present in all models. These estimates have superior statistical properties if the models are carefully chosen and can be used in each model for decompositions or counterfactuals. As mentioned in the section 4, a researcher can then decide to either pick the outcomes obtained with the most likely model or geometrically average the outcomes to robustify inference.

To see why composite estimators of the common parameters robustify inference, suppose K=2, and assume that the decision rules they generate are given by (2) and (3). Maximization of (5) with respect to θ leads to:

$$\rho_A = (\sum_{t=1}^{T_A} y_{At-1}^2 + \zeta_2 \sum_{t=1}^{T_B} y_{Bt-1}^2)^{-1} (\sum_{t=1}^{T_A} y_{At} y_{At-1} + \zeta_1 \sum_{t=1}^{T_B} y_{Bt} y_{Bt-1}) \tag{26}$$

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

$$\sigma_A^2 = \frac{1}{\xi} (\sum_{t=1}^{T_A} (y_{At} - \rho_A y_{At-1})^2 + \frac{1-\omega}{\omega \gamma^2} \sum_{t=1}^{T_B} (y_{Bt} - \delta \rho_A y_{Bt-1})^2) \tag{27}$$

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

$$\rho_{AA} = (\sum_{t=1}^{T_A} y_{At-1}^2)^{-1}(\sum_{t=1}^{T_A} y_{At}y_{At-1}); \quad \rho_{AB} = \delta^{-1}(\sum_{t=1}^{T_B} y_{Bt-1}^2)^{-1}(\sum_{t=1}^{T_B} y_{Bt}y_{Bt-1}) \tag{28}$$

and

$$\sigma_{AA}^2 = \frac{1}{T_A}\sum_{t=1}^{T_A}(y_{At} - \rho_{AA}y_{At-1})^2; \quad \sigma_{AB}^2 = \frac{1}{T_B}\sum_{t=1}^{T_B}(y_{Bt} - \delta\rho_{AB}y_{Bt-1})^2 \tag{29}$$

As (26)-(27) clearly show, θ_{CL} combines the information present in y_{At} and y_{Bt}, with model B playing the role of a prior for model A. The formulas in (26) and (27) are 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.

If model B is irrelevant ($\delta = 0$), y_{Bt} will not be used in the estimation of ρ_A; and it will affect the estimate of σ_A only through the effective sample size ξ^{-1}. Thus, as discussed in section 4.1, the approach automatically discounts models providing poor information in the dimensions assumed to be common. Moreover, when (γ, δ) are unknown and jointly estimated with ρ_A, σ_A^2 using composite methods they reflect only the information in y_{Bt}.

Asymptotically, the composite estimate of ρ is a linear combination of the true values of ρ_A and ρ_B. However, under in certain situations, the combination weight on ρ_A may be greater than one (and thus the weight on ρ_B less than zero). While a negative weight on ρ_B may be optimal if, for example, y_{At}, y_{Bt} are positively correlated and $\sigma_B^2 >> \sigma_A^2$, it should be made clear that one *should not* employ composite methods to estimate an average of ρ_A and ρ_B. The method is clearly applicable when $\rho_A = \rho_B = \rho$. It can also be applied when $\rho_A \neq \rho_B$ if one reparametrizes the models, setting $\rho_A = \rho$ and $\rho_B = \rho\delta$ and estimate ρ using the information contained in y_{At} and y_{Bt}, precisely as we have done in the above example. When the weights are fixed and chosen to be a function of δ, the composite estimator of ρ will be consistent for ρ_A. With an arbitrary ω, there is no guarantee that the approach will deliver consistent estimates. Still, asymptotic MSE gains may materialize because of the shrinkage property of composite estimators.

A potential concern is that our framework leads to nonsensical results when $\rho_A \neq \rho_B$

and one computes a common ρ value using composite methods [6]. Thus, care is needed and ρ_B must be appropriately reparametrized to make estimation estimation meaningful.

When K models are available, θ_{CL} will be constrained by the structure of all models. For example, equation (26) becomes

$$\rho_A = (\sum_{t=1}^{T_A} y_{At-1}^2 + \sum_{i=1}^{K-1} \zeta_{i2} \sum_{t=1}^{T_i} y_{it-1}^2)^{-1} (\sum_{t=1}^{T_A} y_{At} y_{At-1} + \sum_{i=1}^{K-1} \zeta_{i1} \sum_{t=1}^{T_i} y_{it} y_{it-1}) \tag{30}$$

where $\zeta_{i1} = \frac{\omega_i}{\omega_A} \frac{\delta_i}{\gamma_i^2}$, $\zeta_{i2} = \zeta_{i1}\delta_i$. (30) has the format of a shrinkage estimator: it combines model specific and average information contained in the remaining cross section of models. Hence, a composite approach robustifies inference, by requiring estimates of θ to be consistent with the data generated by all available models. (30) works because model specific biases and noise will be averaged out.

Suppose we are interested in the responses of $x_{t+h}, h = 1, 2, \ldots$ to a shock in e_t, where x_{t+h} is a subvector of y_t common to the two models. Clearly, such a response is $x_{t+h}^A = \sum_{j=0}^{h} (\rho_A)^j (\sigma_A \bar{e})$ in model A and $x_{t+h}^B = \sum_{j=0}^{h} (\rho_A \delta)^j (\sigma_A \gamma \bar{e})$ in model B. If one uses composite estimates of $\rho_A, \sigma_A, \gamma, \delta$, one can choose the predictions of model A or model B depending on their fit. Alternatively, one can combine them and use $x_{t+h} = (x_{t+h}^A)^{\omega} (x_{t+h}^B)^{1-\omega}$.

To illustrate the ideas we discussed, consider the problem of estimating the slope of the 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 have small pass-though to the aggregate inflation rate. However, Schorfheide (2008), surveying estimates obtained in DSGE models, documents large cross-study variations and associates the differences with i) the model specification, ii) the observability of marginal costs, and iii) the number and type of variables used in estimation. Here we examine how the composite posterior distribution of the slope of the Phillips curve looks relative to the posterior distribution obtained with i) single models and ii) ex-post averaging of the posteriors of different models. We then compare the responses of the ex-ante real rate to monetary shocks obtained with the various approaches.

We consider five models: a small scale New Keynesian model with sticky prices but non-observable marginal costs, where the variables used in estimation are detrended out-

[6] Thanks to an anonymous referee for pointing this out

put 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 the same as in Smets & Wouters (2007); 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 1947:1-2004:4 and a quadratic trend is used to detrend the data. Using alternative detrending does not change the conclusions we reach, see appendix D. The series used are from the Smets and Wouters (2007) database; the equations of each model and the specifications for the priors are in appendix D. Because the observables are different in different models, Bayesian model averaging of the estimates is not possible.

Table 1 displays some posterior percentiles of the slope of the Phillips curve obtained using each model likelihood 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. The posteriors of these latter two models hardly overlap with those of the second and third 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 obtained with random weights has a median value of 0.22 and a credible 90 percentile ranging from 0.15 to 0.35, which is smaller than the range obtained with a number of individual models.

Figure 5 plots the prior and the posterior ω for each model. The posterior location of ω 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 with observable marginal costs and the medium scale NK model the posterior median is lower than the

Table 1: Percentiles of the posterior distribution of the slope of the Phillips curve

	5%	50%	95%
Prior	0.05	0.42	1.13
Basic NK	0.04	0.18	0.62
Basic NK with nominal wages	0.05	0.06	0.07
SW with capital and adj.costs	0.07	0.09	0.11
Search	0.16	0.28	0.38
BGG	0.12	0.21	0.36
CL	0.15	0.22	0.35
CL (corrected)	0.07	0.23	0.39

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 unadjusted or adjusted. The estimation sample is 1947:1-2004:4.

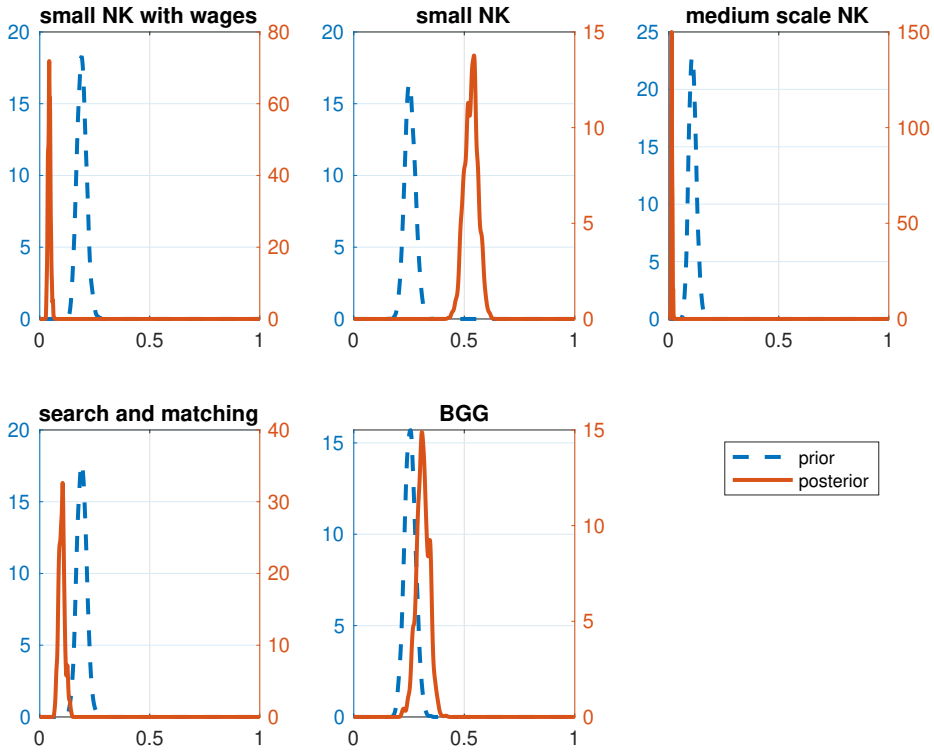

Figure 5: Prior and posterior densities of ω

prior median, and the opposite is true for the basic NK model. Because posterior spreads are tighter than the prior spread, the data is informative about the weights (see Mueller, 2012). Overall, 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 lesser extent, in the BGG and the search and matching model.

Some readers may be surprised that the medium scale New Keynesian model, which is the workhorse used in many policy institutions, has the lowest posterior probability. 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 have independent information for this parameter relative to the other models.

Figure 6 presents the composite posterior distribution for the slope of the Phillips curve together with two naive posterior combinations: one that equally weights the posteriors in the five models; and one which weights the posteriors in the five models by the mode of ω. Combining ex-post posterior estimates generate distributions with lower location. In addition, ex-post combinations produce multimodal posteriors.

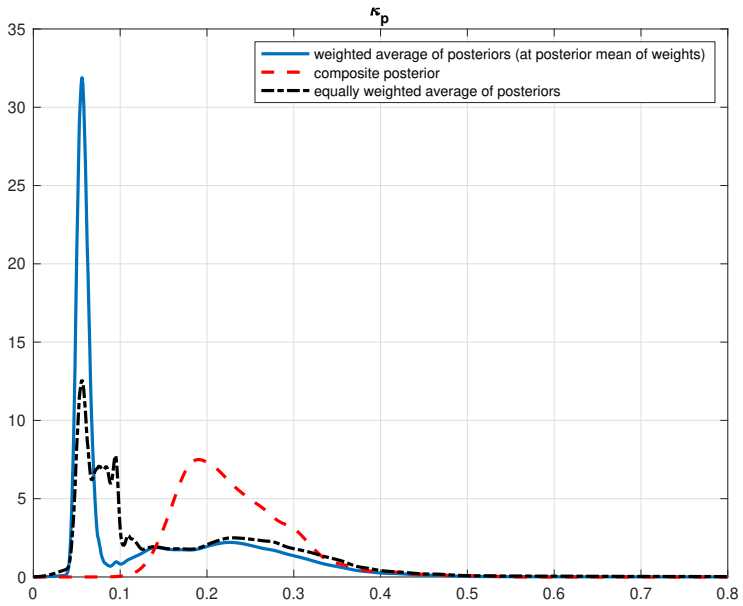

Figure 6: Composite posterior and two naive posterior mixtures

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

combinations previously discussed; and using our composite approach.

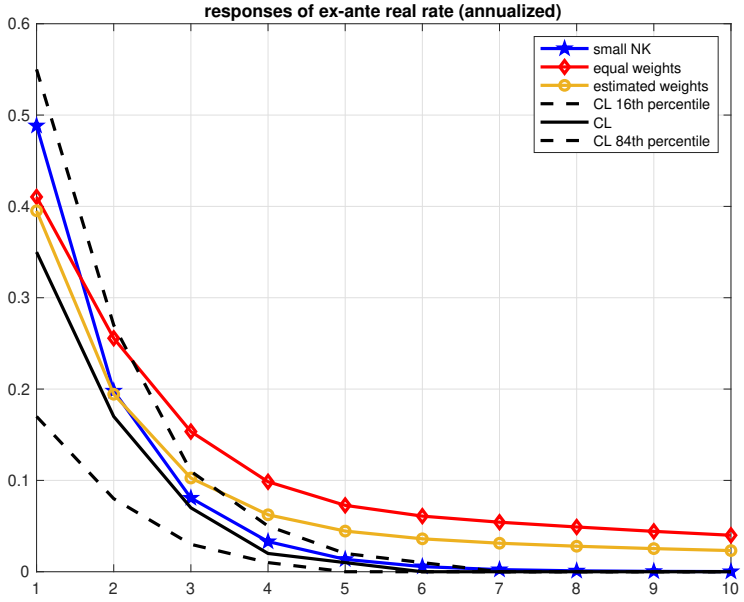

Figure 7: Real rate responses to a monetary shock

The median impact effect is estimated to be 35 basis points. Uncertainty is substantial: while the composite responses are a-posteriori different from zero, the 68% credible set includes the point estimates of all models on impact. At larger horizons, the composite posterior for the real rate responses becomes tighter and the naive equally weighted responses fall outside credible composite posterior intervals. Note also that composite real rate responses are less persistent than other alternatives and, after four quarters, they are basically zero.

5.2 Exploiting cross sectional information

Researchers often have hard time drawing conclusions about an economic phenomena because the results obtained, say, across countries, contradict each other, or because single unit data is not very informative about the issue of interest.

A composite approach is well suited to deal with the situation where there is a single structural model, but data may come from different units (for example, consumers or countries); different levels of aggregation (firms, industries, sectors, or regions); or it is collected at different frequencies (say, weekly or monthly). In this case, we treat time

series for different units (levels of aggregation, frequencies) as different "models"and combine their information to estimate common structural parameters. Chamberlain (1984, p.1272) proposed an estimator for the parameters of a reduced form model, when a panel is available but the cross-section is not necessarily homogeneous, that has the same features as our composite estimator.

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

$$CL(\theta|\hat{y}_{1t}...\hat{y}_{Kt}, \eta_1, \dots \eta_k) = \sum_{i=1}^{K} \omega_i \log L(\theta|\hat{y}_{it}, \eta_i) \tag{31}$$

(31) neglects the correlation structure $\hat{y}_{it}$, in particular, the presence of common shocks that may simultaneously affect all K units, but uses the information present in all units to estimate common parameters. Thus, it defines an objective function that is intermediate between the one obtained with complete pooling of cross model information $CL(\theta, \eta|\hat{y}_{1t}...\hat{y}_{Kt}) = \sum_{i=1}^{K} \omega_i \log L(\theta, \eta|\hat{y}_{it})$ and with complete heterogeneity $CL(\theta_1, \dots \theta_k, \eta_|\hat{y}_{1t}...\hat{y}_{Kt}) = \sum_{i=1}^{K} \omega_i \log L(\theta_i, \eta_i|\hat{y}_{it})$. (31) is similar, in spirit, to the objective function employed in the cross sectional Bayesian literature (e.g., Zellner and Hong, 1989). In this literature, all parameters are restricted; here only θ is restricted across i.

Suppose we have available decision rules like (3) 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 we have seen, for fixed ω, the composite likelihood estimator for ρ_A is

$$\rho_A = (\sum_{t=1}^{T_1} y_{1t-1}^2 + \sum_{i=2}^{K} \zeta_{i2} \sum_{t=1}^{T_i} y_{it-1}^2)^{-1} (\sum_{t=1}^{T_1} y_{1t} y_{1t-1} + \sum_{i=2}^{K} \zeta_{i1} \sum_{t=1}^{T_i} y_{it} y_{it-1}) \tag{32}$$

where $\zeta_{i1} = \frac{\omega_i}{\omega_1} \frac{\delta_i}{\gamma_i^2}$, $\zeta_{i2} = \zeta_{i1}\delta_i$. Clearly, ρ_A pools information from different sources if $\zeta_{ij} = 1, \forall i, j$, and mimics the ML estimator obtained with $\hat{y}_{1t}$ if $\zeta_{ij} = 0, (\delta_i = 0) \forall i, j$. Note that cross model information is not exactly pooled and the degree of cross-model shrinkage of unit specific information depends on the precision of various sources of information. Thus, when dealing with panels of data, the composite approach uses at least as much information as the likelihood of a model; stochastically exploits cross-sectional information; and may lead to improved estimates of the common parameters when different units feature dynamic similarities. The partial pooling approach the composite likelihood delivers is advantageous when $\hat{y}_{1t}$ is short; when the heterogeneities

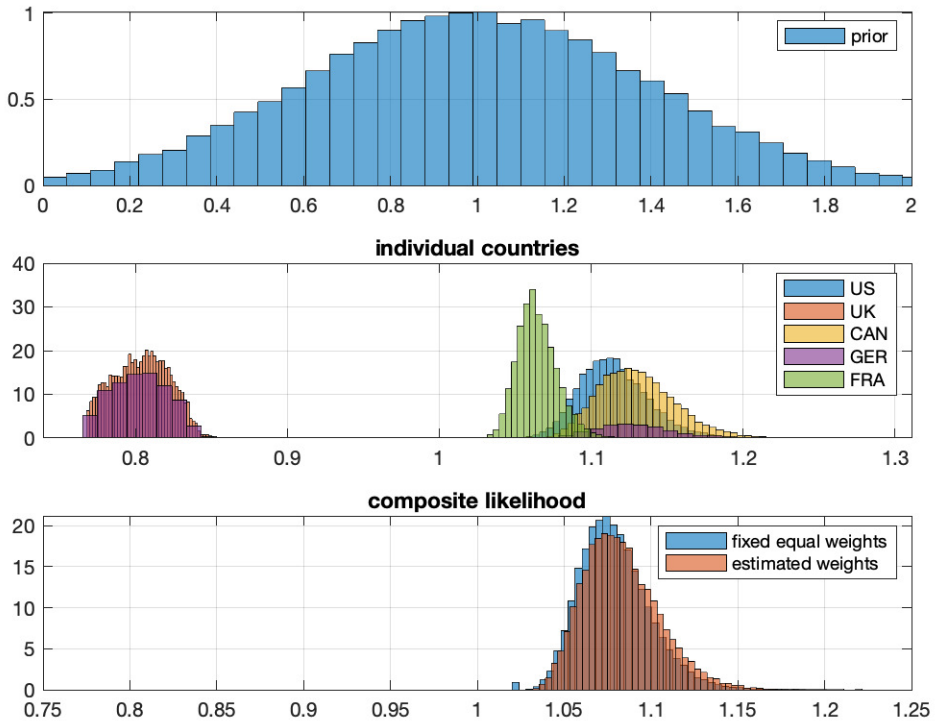

Figure 8: Prior and Posterior distributions for σ

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 across i has similar magnitude.

Note that when $\omega_i = 1/K$, ζ_{ij} reflects the degree of heterogeneity in the panel and has important information for the user. On the other hand, when ω is random, one can use its posterior to evaluate the unit, the level of aggregation or the frequency with most information about the parameters of interest.

To illustrate the use of a composite approach in this situation, we build on the exercise of Karabarbounis and Neiman (2014). They notice that the labor share has dramatically fallen in many countries over the last twenty years and argue that shocks to the relative price of investment, which also declines over time, may be responsible for this fall. Their argument hinges on the elasticity of substitution between labor and capital in production, σ, being greater than one. Using their model specification (the optimality conditions

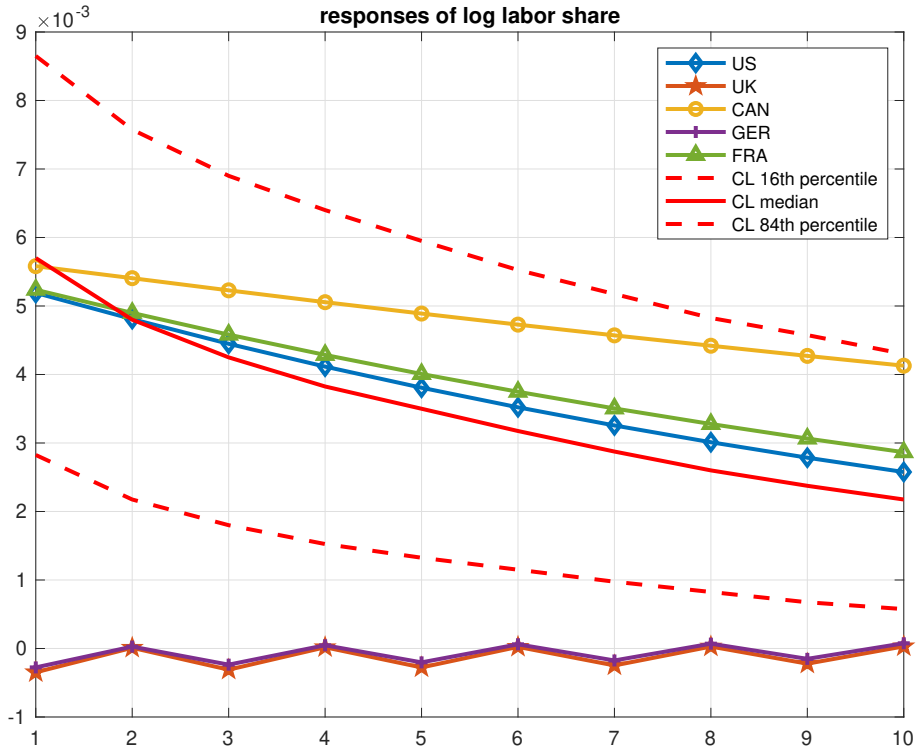

Figure 9: Labor share responses to shocks to the relative price of investment

and the priors are in appendix E) and their dataset, we first estimate σ using US, UK, Canada, Germany, and France data separately. We then use composite methods and data for the five countries to estimate σ, allowing all other structural parameters to be country specific[7].

Figure 8 has the common prior for σ (first row), the single country posteriors (second row) and the composite posterior obtained with cross sectional data when fixed equal weights or random weights are used (third row). The data is informative and, except for the UK and Germany, the posterior distribution about σ is entirely above one. The two composite posterior distributions are also all above one and tight, despite the fact that UK data receives a non-negligible weight (modal posterior value of ω is 0.05). US data appears to be most informative and the posterior of ω 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 in the relative price of investment (with mean equal to half of the

[7]Although we assume that shocks to the price of investment are stationary, none of the results we present are overturned when the price of investment is allowed to have a unit root.

estimated US standard deviation) in each of the five countries and with the panel when random weights are used. Indeed, we 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 and Germany, shocks to the relative price of investment have instead negligible dynamic effects on the labor share.

In sum, our analysis confirms Karabarbounis and Neiman's two main conclusions: i) the elasticity of substitution between capital and labor is greater than one, ii) shocks to the relative price of investment can potentially explain the fall in the labor share observed in many countries. The conclusions we obtain are more general because we allow for stochastic heterogeneity across countries, and use composite estimators that exploit the information present in the optimality conditions the theory provides.

5.3 Other applications

There are at least three other situations where the methodology may help researchers to improve inference. One is when different methods are used to solve a model. For example, it is standard to use the first order perturbed solution of a DSGE model to construct the likelihood, but additional information for the parameters of interest is also present in, say, the second order perturbed solution. Rather than choosing an order of approximation and thus a likelihood for estimation, one could combine the likelihoods obtained with different approximation orders into a composite likelihood and robustify estimation and inference. Alternatively, one could use the posterior of ω to decide which order of approximation provides the most relevant information for parameter estimation.

The second application is to use the approach to combine different approximations of the likelihood function for a given solution method [8]. For example, when second order perturbation solutions are computed, the likelihood approximated with a particle filter or other fully non-linear devices is sensitive to the tuning parameters and the details of the filtering exercise. If different likelihood approximations (say, computed with variants of the Kalman filter) are available, one could combine them and use the composite likelihood to estimate structural parameters. One should expect improved estimates if different approximations err in different directions, since the averaging of information that the composite likelihood produces de-emphasizes idiosyncratic errors.

[8] We thank an anonymous referee for suggesting this possibility

In addition, poorly approximated likelihoods will receive low posterior weights if they provide inferior information for the parameters treated as common.

The approach can also be used to estimate models subject to certain constraints. For example, Acharya et al. (2020) have estimated the parameters of a Heterogeneous Agent New Keynesian (HANK) model, subject to the constraint that the aggregate outcomes the model generates must be consistent with some features of the income distribution. Here the composite likelihood has two pieces: the likelihood of the HANK model; and quasi-likelihood obtained from the income distribution constraints. The weight then represents the relative importance the two types of information have for parameter estimation.

6 Conclusions

This paper describes a procedure to ameliorate identification, estimation and inferential problems in DSGE models. The method helps in a number of situations and automatically provides estimates of the parameters that formally combine the information present in different models or data sets using a shrinkage-like approach. 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 is based on the *composite likelihood*, a limited-information objective function, well known in the statistical literature but very sparsely used in economics. In our setup, the composite likelihood combines the likelihoods of distinct structural or statistical models, none of which is necessarily a marginal or conditional partition of the DGP. Thus, standard composite likelihood properties do not necessarily apply. Still, the approach we propose has desirable statistical properties, it is easy to use, in its quasi-Bayesian version it has an appealing sequential learning interpretation, and provides a way to rank the quality of the available models.

We present examples indicating that using the information present in distinct models helps 1) to ameliorate population and sample identification problems, 2) to solve singularity problems, 3) to produce stable estimates of the parameters of large-scale structural models. It also helps 4) to robustly estimate the parameters appearing in multiple models and to rank models with different observables, 5) to combine information coming from different sources and levels of aggregation for structural estimation.

We believe the methodology has potential in DSGE settings, and the toolkit we provide makes the used of composite methods easy. Furthermore, the examples we have provided highlight ways in which the flexibility of the approach can be exploited in a number of applications.

7 References

Acharya, S., Cai, M., Del Negro, M., Dogra, K., Matlin, E. and R. Scafati (2020). Estimating HANK: Macro Time Series and Micro Moments, Federal Reserve Bank of New York, manuscript.

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 (2018). The pruned state space system for Non-Linear DSGE Models: Theory and Empirical Applications. Review of Economic Studies, 85, 1-49.

Baumeister, C. and J. D. Hamilton (2019). Structural Interpretation of Vector Autoregressions with Incomplete Identification: Revisiting the Roleof Oil Supply and Demand Shocks. American Economic Review, 109(5), 1873-1910.

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). Spatial Interaction and the Statistical Analysis of Lattice Systems, Journal of the Royal Statistical Society (Series B), 36, 192-236.

Bissiri, P. G., C. C. Holmes, and S. G. Walker (2016). A general framework for updating belief distributions, Journal of the Royal Statistical Society: Series B (Statistical Methodology), 78(5), 1103-1130.

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., Ferroni, F., and C. Matthes (2020). Detecting and Analyzing the Effects of time vayring coefficients in DSGE models. International Economic Review, 61, 105-125.

Canova, F. and C. Matthes (2018). An alternative approach to deal with model misspecification, forthcoming. Quantitative Economics.

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

Chan, J., Eisenstat, E., Hu, C. and G. Koop (2018). 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.

Cleydec, M. and E. Iversen (2013). Bayesian model averaging in M-open framework, in P. Damien, P. Dellaportas, N. Polson, and D. Stephens (eds.) Bayesian theory and applications. Oxford Scholarship online.

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.

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.

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.

Fernandez Villaverde, J. and J. Rubio Ramirez (2004). Comparing dynamic equilibrium models to data: a Bayesian approach. Journal of Econometrics, 123: 153-187.

Gilchrist, S., Sim, J., Schoenle, R., and E. Zackrajsek (2017).Inflation dynamics during the financial crisis. American Economic Review, 107(3), 785-823.

Genest, G and J. V. Zidek (1986) Combining probability distributions: A critique and an annotated bibliography. Statistical Science, 1:114-135

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.

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

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

Kleijn, B. and A. Van der Vaart (2012). The Bernstein-Von-Mises theorem under misspecification. Electronic Journal of Statistics, 6: 354–381.

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

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. (1980). Composite Likelihood Methods. Contemporary Mathematics, 80, 221-239.

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

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

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

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

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

Qu, Z. (2018). A Composite likelihood approach to analyze singular DSGE models. Review of Economics and Statistics, 100(5), 916-932.

Ribatet, M., Cooley, D. and A. Davison (2012). Bayesian inference from composite

likelihoods, with an application to spatial extremes. Statistica Sinica, 22, 813-845.

Roche, A. (2019). Composite Bayesian inference. CHUV, Siemens Healthcare, EPFL manuscript.

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.

Walker, S. (2012). Bayesian inference in misspecified models. Journal of statistical planning and inference, 143: 1621-1633.

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.