A Composite likelihood approach for dynamic structural models

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

October 4, 2016

Abstract

We describe how to use a composite likelihood approach to ameliorate several estimation, computational and inferential problems that emerge when using misspecified dynamic stochastic general equilibrium models. We show how to perform Bayesian inference with the composite likelihood and discuss the differences with finite mixture models. We present examples where the methodology has the potential to resolve well known problems. We show how misspecification can be reduced with the approach and provide an example to illustrate its properties in practice.

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

JEL Classification: C10, E27, E32.

*We thank Mark Watson, Barbara Rossi and Refet Gurkenyak and the participants to the ESEM 2016 invited session on 'New development in the analysis of Macroeconomic data' for comments and suggestions. Canova acknowledges the financial support from the Spanish Ministerio de Economia y Competitividad through the grants ECO2012-33247, ECO2015-68136-P and FEDER, UE. The views presented in this paper do not reflect those of the Federal Reserve Bank of Richmond, or the Federal Reserve system.

1 Introduction

In macroeconomics it is now standard to construct and analyze the properties of dynamic general equilibrium (DSGE) models. Until a decade ago, most of the models used for economic analysis, policy and forecasting exercises had parameters which were formally or informally calibrated. Now, it is more common to conduct inference conditional on the estimates obtained with classical or Bayesian estimation approaches.

It is well known, however, that the estimation of DSGE models it is difficult because of population and sample identification problems, see e.g. Canova and Sala (2009), Komunjer and Ng (2011), Qu and Thachenko (2013); of singularity problems (the number of shock is generally smaller than number of endogenous variables), see e.g. Guerron Quintana (2010), Canova et al (2014), Qu (2015); of informational deficiencies problems (models are constructed to explain only a portion of the data), see Boivin and Giannoni (2006), Canova (2014), or Pagan (2016); of numerical difficulties (if the model is of large scale or if the data is short or of poor quality) and of latent variable problems. All these issues may make inference whimsical.

In addition, estimation is typically performed with full information likelihood-based methods, see Andreasan et al. (2014) for an exception. For standard frequentist asymptotic theory to apply, likelihood-based estimation requires regularity conditions which are often violated in practice. In addition, when T is short or the parameter space is large, asymptotic approximations provide a poor characterization of the actual distribution of estimates. Bayesian methods help if T is short, but it is tricky to specify joint priors when the parameter space is large and, as indicated by Del Negro and Schorfheide (2008), independence is suboptimal.

Perhaps more importantly, while at the specification stage there is a general consensus that all available models are misspecified in, at least, some dimensions, at the estimation stage this fact is neglected. However, likelihood-based inference is conditional on the estimated model being correctly specified.

The recent literature has been trying to deal with some the individual problems, although not much has been done to take into account misspecification, short samples, or large scale problems, but no unified approach to deal with all these issues exists. This paper proposes an approach based on the *composite likelihood* which can potentially address all these problems. The procedure it is easy to implements, works well when the full likelihood may be problematic to construct and use, produces estimators with nice shrinkage properties and, in its Bayesian version, it has an appealing sequential learning interpretation.

The composite likelihood is a limited information object constructed combining marginal or conditional likelihoods. It can be treated as a quasi-likelihood function and employed to estimate the parameters of structural models, both in frequentist and Bayesian setups. Since the score and the information matrix can be easily defined, standard asymptotic theory can be applied as the composite likelihood has a normal shape as either the number of observations or the number of marginal/conditional components grows. Bayesian inference can also be performed as long as the prior makes the compos-

ite posterior a proper density function. The idea that a prior can be used to regularize objects which are not density functions is well established in the econometric literature, see e.g. Kim (2002) or Chernozukov and Hong (2003), and Christiano et al. (2011) have used this idea to provide posterior estimates of structural parameters of a DSGE when the objective function is a weighted average of impulse responses. A composite likelihood approach has been used to solve a variety of complicated problems in fields as diverse as spatial statistics, multivariate extremes, psycometrics, genetics/genomics, see e.g. Varin et al. (2011). Applications to economic problems, however, have been limited: except for Engle et al. (2008) and Qu (2015), the approach has been largely ignored in the literature. The paper describes how to use the composite likelihood approach to either solve or ameliorate the above mentioned problems, shows how the procedure helps to robustify estimates of the structural parameters in a variety of interesting economic problems, highlights how to perform composite posterior inference, and provides an application where its use helps to improve the quality of the inference a researcher is able to draw about one interesting economic parameter.

The rest of the paper is organized as follows. The next section presents the composite likelihood, shows some of its asymptotic properties, provides a Bayesian methodology to construct small sample, exact distributions of the parameters of interest and provides intuition on how the methodology can be applied to the estimation of the parameters of structural models. Section 3 presents a number of examples where the procedure can be used to i) obtain shrinkage estimates of the parameters appearing in multiple (nested and non-nested) misspecified structural models; ii) improve their (sample and population) identification properties, iii) provide a tractable approach to solve computational and singularity problems; iv) exploit information coming either from the cross-section or from different levels of data aggregation; v) produce more stable estimates of parameters present in large scale models. Section 4 discusses an application dealing with the estimation of the slope of Phillips curve. Section 5 concludes.

2 The composite likelihood

Suppose an unknown data generating process (DGP) produces the density $f(y_t, \theta)$ for an $m \times 1$ vector of observable time series y_t, where θ is a $q \times 1$ vector of parameters. The DGP could be unknown because we do not have enough information to construct f or because f is computationally intractable; for example, it is highly nonlinear; it involves integrals on latent (state) variables; or m is too large to be handled with existing computers. Suppose that for some events $A_1, \dots A_K$, we can construct subdensities $f(y_{it} \in A_i, \theta, \eta_i)$. These subdensities could be marginal or conditional depending on the specification of the problem and each event has implications for a subvector of the observables y_{it} of length T_i. The elements of y_{it} do not have to be mutually exclusive across events; for example, the inflation rate could appear in each y_{it}. Each event is associated with an extended vector of parameters $\psi_i = [\theta, \eta_i]'$, where η_i are (nuisance) event specific. We represent the information generated by A_i with the tuple $(y_{it}, \mathrm{T}_i, \psi_i)$.

For a given set of weights ω_i, the composite likelihood is

$$C(\theta, \eta, y) = \Pi_{i=1}^{k} \ f(y_{it} \in A_i, \theta, \eta_i)^{\omega_i} = \Pi_{i=1}^{k} \mathcal{L}(\theta, \eta_i | y_{it} \in A_i)^{\omega_i} \tag{1}$$

$C(\theta, \eta, y)$ is not a likelihood function and thus the properties of θ_{CL},the maximum composite likelihood estimator, are not immediate. If $y_{[1,t]} = (y_1, \ldots, y_t)$ is independent sample from $f(y_t, \theta)$, θ_{CL} is consistent and

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

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

$$\begin{aligned}
G(\theta) &= H(\theta) J(\theta)^{-1} H(\theta); \ H(\theta) \neq J(\theta) \text{ Godambe information} && (3)\\
J(\theta) &= var_\theta u(\theta, \eta_i, y_{[1,t]}) \text{ Variability matrix} && (4)\\
H(\theta) &= E_\theta[- \bigtriangledown_\theta \, u(\theta, \eta_i, y_{[1,t]})] \text{ Sensitivity matrix} && (5)\\
u(\theta, y) &= \sum_i \omega_i \bigtriangledown_\theta \, l_i(\theta, \eta_i, y_{[1,t]}) \text{ Composite scores} && (6)
\end{aligned}$$

and $\bigtriangledown_\theta l_i(\theta, \eta_i, y_{[1,t]}$ denotes the score associated with each log-likelihood. If T is fixed, but $K \to \infty$, we need assumptions on how the K events are constructed. For example, if they are independent, then (2) still holds. This is true even when $\{y_t\}_{t=1}^{T}$ is a single time series and has correlated observations (see Engle et al., 2008). Note that a standard Newey-West correction to $G(\theta)$ can be made if $y_{[1,t]}$ is not an independent sample.

From (3) one can immediately see that θ_{CL} is not fully efficient. However, careful choice of ω_i may help to improve its efficiency. If one treats ω_i as fixed, one standard a-priori choice is $\omega_i = \frac{1}{K}, \forall i$. Alternatively, one could use a data-based approach to select them, e.g. set $\omega_i = \frac{\exp(\gamma_i)}{1+\sum_{i=1}^{K-1} \exp(\gamma_i)}$, where γ_i are a function of some statistics (MSE, MAD, etc.) of past data $\gamma_i = f_i(Y_{1,[1:\tau]}, ..., Y_{K,[1:\tau]})$. Note that with this latter choice, ω_i could be made time varying. There a large forecasting literature dealing with the choice of weights (see e.g. Aiolfi et al., 2010) which could be used here. As it will be clear below, our approach is to treat ω_i as random variables and jointly estimate the posterior distribution of (ψ_i, ω_i).

When K or the number of nuisance parameters η_i is large, a two step estimation approach is possible. Here, η_i are estimated separately from each $\log f(y_{it} \in A_i, \theta, \eta_i)$ and plugged in the composite likelihood, which is then optimized for θ, see e.g. Pakel et al. (2011). Consistency of θ_{CL} is unaffected as long as η_i are consistently estimated. One can also design information criteria to optimally select the number of events K Thus:

$$\begin{aligned}
AIC_{CL} &= -2C(\theta_{CL}, \eta_{CL}, y) + 2dim(\theta) && (7)\\
BIC_{CL} &= -2C(\theta_{CL}, \eta_{CL}, y) + 2dim(\theta) \log K && (8)
\end{aligned}$$

where $dim(\theta) = tr\{H(\theta) G(\theta)^{-1}\}$ can be optimized in the usual way. These criteria can also be used for model averaging exercises or for selecting tuning parameters in shrinkage methods (see Gao and Song, 2011).

2.1 A Bayesian setup

Because we are interested in small sample exact distributions for the common parameter vector θ, we treat ω as a hyperparameter vector with prior p(ω). We combine the composite likelihood (1) with a prior for (ψ, ω) and construct the posterior for (ψ, ω). Because the posterior distribution is not available in a closed form, we describe a Metropolis-within-Gibbs approach to numerically compute sequences for (ψ, ω) generated from this posterior. For each event A_i, and given (y_{it}, T_i, ψ_i), we assume the likelihood $\mathcal{L}(\psi_i|y_{i,T_i})$ for each event can be constructed and, given $\omega = [\omega_1 \dots \omega_K]'$, where for each i, $0 < \omega_i < 1$, $\sum_i \omega_i = 1$, one can obtain the composite likelihood(1).

We assume that for each A_i, the priors for structural and nuisance parameters are independent and of the form:

$$p(\psi_i) = p(\eta_i)p(\theta)$$

The composite posterior kernel is:

$$\mathcal{L}(\psi_1|Y_{1,T_1})^{\omega_1}p(\psi_1)^{\omega_1}p(\omega_i)\dots\mathcal{L}(\psi_K|Y_{K,T_K})^{\omega_K}p(\psi_K)^{\omega_K}p(\omega_k) = \Pi_i\mathcal{L}(\psi_i|Y_{i,T_i})^{\omega_i}p(\psi_i)^{\omega_i}p(\theta)p(\omega_i) \tag{9}$$

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

For computational and efficiency reasons, it may be preferable to use a $2K+1$ block Metropolis-within-Gibbs algorithm, where we sample the parameters in different blocks separately. Chib and Ramamurthy (2010) and Herbst and Schorfheide (2015) have also suggested drawing DSGE parameters in blocks. However, while they randomly split up the parameter vector in different blocks at each iteration, the blocks here are predetermined by the K events of interest.

2.2 Estimation Algorithm

The algorithm consists of four steps:

1. Start with $\Omega^0 = [\eta_1^0 \dots \eta_K^0, \theta^0, \omega_1^0 \dots \omega_K^0]$.
   For $iter = 1 : draws$ do steps 2-4
2. For $i = 1 : K$ draw η_i^* from a symmetric proposal P^{η_i}. Set $\eta^{iter} = \eta_i^*$ with probability
$$\alpha_i = \min\left(1, \frac{\mathcal{L}(\left[\eta_i^*, \theta^{iter-1}|Y_{i,T_i}\right]^{\omega_i}p(\eta_i^*)^{\omega_i}}{\mathcal{L}(\left[\eta_i^{iter-1}, \theta^{iter-1}|Y_{i,T_i}\right]^{\omega_i}p(\eta_i^{iter-1})^{\omega_i}}\right)$$
   Note: if $\omega_i = 1$, some i, α_i is the standard MH acceptance probability for model i.
3. Draw θ^* from a symmetric proposal P^θ. Set $\theta^{iter} = \theta^*$ with probability
$$\beta = \min\left(1, \frac{\mathcal{L}(\left[\eta_1^{iter}, \theta^*|Y_{1,T_1}\right]^{\omega_1}\dots\mathcal{L}(\left[\eta_K^{iter}\theta^*|Y_{K,T_K}|\right]^{\omega_K}p(\theta^*)}{\mathcal{L}(\left[\eta_1^{iter}, \theta^{iter-1}|Y_{1,T_1}\right]^{\omega_1}\dots\mathcal{L}(\left[\eta_i^{iter}, \theta^{iter-1}|Y_{K,T_K}\right]^{\omega_K}p(\theta^{iter-1})}\right)$$

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

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

When the K events feature no nuisance parameters, step 2. disappears from the algorithm. Similarly, if the ω_i are treated as fixed, step 4 disappears. Notice that when $\omega_i = 0, i \neq k$, $\omega_k = 1$, the algorithm collapses into a standard Metropolis MCMC. A standard random walk proposal for (θ, η_i) seems to work well in practice; a multivariate logistic proposal for ω is advisable.

2.3 A sequential learning interpretation

Suppose, for the sake of illustration, that ω_i are fixed. It is easy to give a sequential, adaptive learning interpretation to the composite posterior kernel (9) and thus to the Bayesian estimators we construct. Suppose K=2. Then

$$\begin{aligned}\breve{g}(\psi_{1....}\psi_K|Y_{1,T_1}, ...Y_{K,T_K}) &= \\ \mathcal{L}(\theta, \eta_1|Y_{1,T_1})^{\omega_1} p(\eta_1)^{\omega_1} p(\eta_2|Y_{2,T_2})^{\omega_2} \{[p(\theta|Y_{2,T_2}, \eta_2) ML(Y_{2,T_2})]^{\omega_2} p(\theta)^{\omega_1}\} &\qquad (10)\end{aligned}$$

where $ML(Y_{2,T_2}) = \int \mathcal{L}(Y_{2,T_2}|\psi_2)^{\omega_2} p(\psi_2)^{\omega_2} d\psi_2$ is the marginal likelihood associated with A_2.

As (10) makes it clear, the posterior kernel can be obtained in two stages. In the first stage the prior for ψ_2 and the likelihood for event A_2 are used to construct the conditional posterior $p(\theta|Y_{2,T_2}, \eta_2)$. This conditional posterior, weighted by the marginal likelihood of A_2, is combined with the prior $p(\theta)$ for the next estimation stage of θ. Suppose that $ML(Y_{2,T_2})$ is high, i.e. the specification for event A_2 fits Y_{2,T_2} well. If $\omega_1 = \omega_2$, the prior for event A_1will more heavily reflect the posterior of A_2 relative to the initial prior $p(\theta)$. Suppose instead that the specification used for A_2 fits Y_{2,T_2} poorly. In this case, the posterior for A_2 will have a low weight relative to $p(\theta)$ when setting up the prior for event A_1. In other words, the approach implicitly discounts information contained in events whose subdensity poorly explain the data observable for those events. In general, the prior for θ in each estimation stage depends on the relative weights assigned to the current and to the previous events, on the fits of the specifications for all previous events, and on the nuisance parameters estimated up to the current stage. Thus, a composite Bayesian approach to estimation can be interpreted as an adaptive sequential learning process. Here, the prior for the current stage is not the posterior at the previous stage as in standard Bayesian setups, but rather a weighted average of the initial prior and of the posterior obtained at the previous stage, where the latter will be discounted by the fit of the specification at that stage.

Notice that since only Y_{2,T_2} contains information for η_2 the posterior for this parameter will not be updated at stage 1. Similarly, since Y_{2,T_2} does not contain information for η_1, $p(\eta_1)$ will be left unchanged after stage 2 estimation.

2.4 A comparison with finite mixture models

The composite likelihood weights the likelihood of different events. An alternative way to pool the information contained in different events comes from ideas of Geweke and Amisano (2011) or Del Negro et al. (2014). The work of Waggoner and Zha (2012) is also relevant in this respect, if one thinks of events as switching specifications. In this literature, the relevant object for the analysis is the likelihood of the mixture of the models, which for each t is $L(\theta, \eta_1....\eta_k|y_{1t}...y_{kt},) = \Sigma_{i=1}^{K}\omega_i L(\theta, \eta_i|y_{it})$ so that

$$\log L = \sum_{t=1}^{\tau} \log L(\theta, \eta_1....\eta_k|y_{1t}...y_{kt}) \tag{11}$$

where $\tau = min(T_i)$.

Simple manipulation of (11) and(1) reveals that the log-likelihood of the two models differ by a Jensen's inequality term: in the composite likelihood approach the objective function is a convex combination of (conditional or marginal) log-likelihoods; in the finite mixture approach, the objective function is the log-likelihood of a mixture of conditionals [1]

While a-priori both specifications have appealing features, for the problems we are interested in, the composite likelihood has superior properties and added flexibility. From a computation point of view, when the decision rules have an autoregressive structure, estimators of θ may have a closed form expression in the composite log-likelihood case, but not in the finite mixture case. Thus, iterative procedures need to be employed. In addition, in the finite mixture setup, it must be the case that $y_{it} = y_{jt}$ and $T_i = T_j$, since events represent alternative models that could have generated the observable data. As we will see below, this is not necessarily the case in the composite likelihood formulation and gives the method much more flexibility. Third, the interpretation of ω differs: in the composite likelihood ω_i represents the proportion of observations one would take from model i out of the effective sample size ξ. In the

[1] The difference between a finite mixture and a composite likelihood formulations can be easily seen when $K = 2$ and $T_A = T_B = 2$. Then, the composite log-likelihood is

$$\log L = \omega(\log L_{A1} + \log L_{A2}) + (1-\omega)(\log L_{B1} + \log L_{B2})$$

while the log-likelihood in the mixture model is

$$\log L = \log(\omega L_{A1} + (1-\omega)L_{B1}) + \log(\omega L_{A2} + (1-\omega)L_{B2})$$

Suppose $\omega = 1-\omega$. Then, (1) and (11) differ by a Jensen's inequality term. Using $\log \sum_{t=1}^{T} x_t \equiv \log x_1 + \log(1+\sum_{t=2}^{T}\frac{x_t}{x_1})$, one has $\log \sum_{t=1}^{2} x_t = \log x_1 + \log(1+\frac{x_2}{x_1})$ and this differ from $\sum_{t=1}^{2}\log x_t = \log x_1 + \log x_2$, since $\log(1+\frac{x_2}{x_1}) \approx \frac{x_2}{x_1}$ if $\frac{x_2}{x_1}$ is small.

mixture model ω_i represents the probability that one observation comes from model i. Thus, for ω to have the same interpretation, we need the effective sample size ξ to be large and ergodicity to apply. Finally, while in the Bayesian composite likelihood approach there is an automatic discounting whenever a specification does not fit well the data of an event, regardless of whether ω_i is treated as a parameter or a random variable, a finite mixture does not necessarily discount the posterior obtained from specification which fits worse in estimation. It is only when ω is treated as a random variable that this becomes true.

2.5 What are $f(y_{it} \in A_i, \theta, \eta_i)$ in a DSGE context?

To understand how to apply composite likelihood ideas to the estimation DSGE models, one needs to understand what the events A_i are and how the resulting $f(y_{it} \in A_i, \theta, \eta_i)$ relate to the DGP.

In a leading example we have in mind, A_i are different economic models and $f(y_{it} \in A_i, \theta, \eta_i)$ are the densities associated with these K different structures, i.e. the densities associated with a basic RBC model, a RBC model with financial frictions, a New Keynesian model with sticky price, a new Keynesian model with sticky wages, etc.. Here θ are the parameters which are common across all models, e.g. the risk aversion coefficient, the Frisch elasticity, or the labor share in production, while η_i are the parameters specific to the model, e.g. a LTV ratio parameter or a Calvo parameter. These models are treated as subdensities because they are obtained either by disregarding aspects of the unknown DGP, or by explicitly conditioning on certain features, e.g. certain markets are treated as competitive or non-competitive. Thus, all these models are mispecified in a sense that will be made clear below. In another leading example we have in mind, $f(y_t, \theta)$ is a large scale DGP (for example, a multi-country model of trade interdependencies or a multi-asset pricing model) and $f(y_{it} \in A_i, \theta, \eta_i)$ are components of the model obtained using sub-elements of y_t, for example, country or asset specific blocks or bilateral blocks. $f(y_{it} \in A_i, \theta, \eta_i)$ could also represent different *statistical models* obtained from the same theoretical DGP but featuring different observables. For instance, a standard three-equations New-Keynesian model could be estimated using inflation, the nominal interest rate, and a measure of output; or inflation, the nominal interest rate, and a measure of consumption. These two set of observables constitute what we call two different statistical models. By extension, $f(y_t, \theta)$ could also be an aggregate model and $f(y_{it} \in A_i, \theta, \eta_i)$ the density obtained when data from cross sectional unit i are used; when data y_{it} is aggregated at different levels (e.g. firm, industry, regional, etc.), or when K different samples (say, pre-WWI, interwar, post-WWI, etc.) are used. As shown later on, in all these situation a composite likelihood approach produces shrinkage estimators for the common parameters θ. Alternatively, $f(y_{it} \in A_i, \theta, \eta_i)$ could be the densities generated by different approximate solutions of a model, where the events A_i represent the order of the perturbation or of the projection employed; or the densities of linear solutions, where only the k-th component of θ is allowed to be time varying. In all the example,

$f(y_{it} \in A_i, \theta, \eta_i)$ ignores the potential dependence of events $A_i, i = 1, ...K$. Thus, the estimators obtained are of limited information type and may lead to inefficient inference from a frequentist point of view. Moreover, since they feature nuisance parameters η_i,inference for θ may be affected.

Researchers working with DSGE models are generally free to choose what goes in θ and in η_i - even though some parameters might be common to all events, researchers might prefer not to estimate a common value. In the case A_i represents distinct economic models, θ could be the risk aversion parameter or the persistence of technology shocks, while η_i could be the Calvo price (wage) parameter. In the case A_i represents different statistical models, one could estimate one set of parameters for each statistical model, or impose that some or all them are common. When A_i represents different level of data aggregation, one could make, e.g., the marginal propensity to consume common, while the parameters regulating the process for income to be left event specific. Clearly, one can think of cases when the parameters are common to a subset of the K events one wishes to consider.

3 Using the composite likelihood for structural inference

This section provides examples showing the value of a composite likelihood approach when dealing with standard problems encountered in the estimation of DSGE models. The first example discusses the issue of estimating parameters appearing in multiple misspecified models; the next two examples show how the approach can ameliorate sample and population identification problems; the fourth example deals with singularity issues; the fifth the problem of estimating the parameters of a large dimensional model. The last example shows how the composite likelihood approach helps to deal with short samples.

3.1 Estimating structural parameters appearing in multiple misspecified models

Suppose we have two structural models (A, B), which may have implications for different variables (y_{At}, y_{Bt}), and may feature common parameters (such as utility function parameters or policy rule coefficients) and model specific parameters. Both models are treated as misspecified in the sense that $f(y_A, \theta, \eta_A) \neq \int f(y, \theta) dy_B$, and $f(y_B, \theta, \eta_B) \neq \int f(y, \theta) dy_A$, for some subvectors y_{At} and y_{Bt} of y_t.

Assume that the decision rules of the two models are:

$$y_{At} = \rho_A y_{At-1} + \sigma_A e_t \tag{12}$$

$$y_{Bt} = \rho_B y_{Bt-1} + \sigma_B u_t \tag{13}$$

where e_t and u_t are iid(0,I). For the sake of illustration, suppose that $\rho_B = \delta\rho_A, \sigma_B = \gamma\sigma_A$, that y_{At} and y_{Bt} are scalars, and that we have T_A observations on y_{At} and T_B observations on $y_{Bt}, T_B \leq T_A$ and that we are interested in $\theta = (\rho_A, \sigma_A)$.The (normal) log-likelihood functions associated with each model are:

$$\log L_A \propto -T_A \log \sigma_A - \frac{1}{2\sigma_A^2}\sum_{t=1}^{T_A}(y_{At} - \rho_A y_{At-1})^2 \tag{14}$$

$$\log L_B \propto -T_B \log \sigma_B - \frac{1}{2\sigma_B^2}\sum_{t=1}^{T_B}(y_{Bt} - \rho_B y_{Bt-1})^2 \tag{15}$$

and the composite likelihood is

$$\log C = \omega \log L_A + (1-\omega)\log L_B \tag{16}$$

where we interpret ω as the degree of a-priori trust a researcher has in model A.

Maximization of (16) leads to:

$$\rho_A = (\sum_{t=1}^{T_A} y_{At-1}^2 + \phi_2 \sum_{t=1}^{T_B} y_{Bt-1}^2)^{-1}(\sum_{t=1}^{T_A} y_{At}y_{At-1} + \phi_1 \sum_{t=1}^{T_B} y_{Bt}y_{Bt-1}) \tag{17}$$

where $\phi_1 = \frac{1-\omega}{\omega}\frac{\delta}{\gamma^2}$, $\phi_2 = \frac{1-\omega}{\omega}\frac{\delta^2}{\gamma^2} = \phi_1\delta$ and

$$\sigma_A^2 = \frac{1}{\xi}(\sum_{t=1}^{T_A}(y_{At} - \rho_A y_{At-1})^2 + \frac{1-\omega}{\omega\gamma^2}\sum_{t=1}^{T_B}(y_{Bt} - \delta\rho_A y_{Bt-1})^2) \tag{18}$$

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

$$\rho_{AA} = (\sum_{t=1}^{T_A} y_{At-1}^2)^{-1}(\sum_{t=1}^{T_A} y_{At}y_{At-1}); \quad \rho_{AB} = \delta^{-1}(\sum_{t=1}^{T_B} y_{Bt-1}^2)^{-1}(\sum_{t=1}^{T_B} y_{Bt}y_{Bt-1}) \tag{19}$$

and

$$\sigma_{AA}^2 = \frac{1}{T_A}\sum_{t=1}^{T_A}(y_{At} - \rho_{AA}y_{At-1})^2; \quad \sigma_{AB}^2 = \frac{1}{T_B}\sum_{t=1}^{T_B}(y_{Bt} - \delta\rho_{AB}y_{Bt-1})^2) \tag{20}$$

Equations (17) and (18) provide intuition on what a composite likelihood approach does in this setup. In fact, for estimation of the common parameters θ, model B plays the role of a prior for model A. The parameters specific to model B (δ, γ) are instead estimated using only the information present in model B data. The formulas in (17) and (18) 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 are the additional observations added to the T_A available to estimate θ.

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

When an array of models are considered, estimates of the common parameters θ will be constrained by the structure present in all models. For example, equation (17) becomes

$$\rho_A = (\sum_{t=1}^{T_A} y_{At-1}^2 + \sum_{i=1}^{K-1} \phi_{i2} \sum_{t=1}^{T_i} y_{it-1}^2)^{-1} (\sum_{t=1}^{T_A} y_{At} y_{At-1} + \sum_{i=1}^{K-1} \phi_{i1} \sum_{t=1}^{T_i} y_{it} y_{it-1}) \qquad (21)$$

where $\phi_{i1} = \frac{\omega_i}{\omega_A} \frac{\delta_i}{\gamma_i^2}$, $\phi_{i2} = \phi_{i1} \delta_i$. Thus, the composite likelihood robustifies inference, in the sense that estimates of the common parameters are shrunk to be consistent with the data generated by all the models available for inference. Later we present an example where ω, rather than representing the a-priori trust an investigator has on each model, is a vector of unknown parameters. There we show that model misspecification can be reduced by a careful choice of ω.

Two further aspects of this example are worth some discussion. First, y_{At} and y_{Bt} may be different series. Thus, the procedure can be used to estimate common parameters in models featuring different observables. Alternatively, y_{At} and y_{Bt} may represent the same series with different levels of aggregation (say, aggregate vs. individual consumption). This feature makes the procedure stands apart from finite mixture models where, as we have seen, y_{At} and y_{Bt} must contain the same series. In general, y_{At} and y_{Bt} may have common components (say, output and inflation) and some model specific ones (say, the trade balance or asset prices). Second, T_A and T_B may be of different length. Hence, the procedure allows us to combine data of various length or the information present in models setup up at different frequencies (e.g., a quarterly and an annual model). T_A and T_B may also represent samples for the same vector of economic variables coming from different subsamples, for example, consumption and the real rate before and after a financial crisis. The setup we consider is sufficiently general to account for all these possibilities.

This simple example also allows us to discuss how a composite likelihood approach may help to reduce small sample identification problems. This is a situation where

the likelihood of model A is well behaved,but because T_A is short, it may be flat and part of the domain and $\theta = (\rho_A, \sigma_A)$ may be poorly identified using y_{At}. It is easy to show θ could become better identified if (y_{At},y_{Bt}) are jointly used in estimation. This is because the curvature of the composite likelihood depends on the effective sample size is ξ which, in turn, is a function of T_A and $T_B \frac{1-\omega}{\omega\gamma^2}$. Thus, for example, if γ (or ω) is small, that is the data generated by model B to be less volatile than the data generated by model A or the degree of a-priori trust a researcher has in model B is high enough, $\xi >> T_A$ and the composite likelihood will be more peaked than the likelihood constructed with y_{At} only around the mode.

3.2 Ameliorating population identification problems

The previous subsection discussed how a composite likelihood approach may help to improve parameter identification when the sample size associated with the baseline model makes the likelihood in the dimensions of interest flat. This subsection presents an example where some parameters are underidentified and others only weakly identified *in population* in a baseline model and shows how a composite likelihood approach can remedy these problems.

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

$$R_{At} = \tau E_t \pi_{At+1} + e_{1t} \tag{22}$$
$$y_{At} = \delta E_t y_{At+1} - \sigma(R_{At} - E_t \pi_{At+1}) + e_{2t} \tag{23}$$
$$\pi_{At} = \beta E_t \pi_{At+1} + \gamma y_{At} + e_{3t} \tag{24}$$

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

$$\begin{bmatrix} R_{At} \\ y_{At} \\ \pi_{At} \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 \\ \sigma & 1 & 0 \\ \sigma\gamma & \sigma & 1 \end{bmatrix} \begin{bmatrix} e_{1t} \\ e_{2t} \\ e_{3t} \end{bmatrix} \equiv A e_t \tag{25}$$

Clearly, β is underidentified (it disappears from the solution) and that the slope of the Phillips curve γ may not be well identified from the likelihood of the model if σ is small, regardless of the size of T_A. In fact, large variations in γ may induce small variations in the decision rules (25) if σ is sufficiently small, making the likelihood flat in the γ dimension.

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

$$\pi_{Bt} = \beta E_t \pi_{Bt+1} + \gamma y_{Bt} + u_{2t} \tag{26}$$
$$y_{Bt} = \rho y_{Bt-1} + u_{1t} \tag{27}$$

where $\beta < 1$ and $\rho \neq 0$, measures the persistence of marginal costs. By repeatedly substituting forward we have

$$\pi_{Bt} = \frac{\gamma}{1-\beta\rho} y_{Bt} + u_{2t} \tag{28}$$

$$y_{Bt} = \rho y_{Bt-1} + u_{1t} \tag{29}$$

We can rewrite (28) and (29) in terms of $x_t \equiv (1-\rho\ell) y_{Bt}$, $w_t \equiv (1-\rho\ell)\pi_{Bt}$ as

$$\begin{bmatrix} x_t \\ w_t \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ \frac{\gamma}{1-\beta\rho} & 1-\rho\ell \end{bmatrix} \begin{bmatrix} u_{1t} \\ u_{2t} \end{bmatrix} \tag{30}$$

where ℓ is the lag operator.

Because the log-likelihood of model B has information about β, one would be able to identify (and estimate) β from the composite likelihood. In addition, since in model B the curvature of the likelihood in the γ dimension depends on $\frac{1}{1-\beta\rho}$ which, in general, is greater than one for $\beta \neq 0$. Hence, small variations γ may lead to sufficiently large variations in (30) and thus in the composite likelihood, even when $1-\omega$ is small. In this particular example, shrinking model A toward a model which is more misspecified but has sharper information about the parameters of interest may be beneficial in terms of identification and estimation. It should be emphasized that all the arguments are independent of the size of effective sample size ξ: since the identification problems we discuss occur in population, having a large or a small ξ is irrelevant. It should also be emphasized that the above argument implicitly assumes that the variances of $(\mathrm{e}_{2t}, e_{3t})$ are of the same order of magnitude as the variances of (u_{1t}, u_{2t}).

3.3 Solving singularity problems

DSGE models are typically singular. That is, since they typically features more endogenous variables than shocks, the theoretical covariance matrix of the observables is singular and the likelihood function of the model can not be constructed and optimized. There are two types of responses to this problem in the literature : one is to select a subvector of the observables that matches the dimension of the shock vector either informally (see Guerron Quintana, 2010) or formally (see Canova et al. , 2014) and then use the model to construct the log-likelihood for this subvector. The second is to add measurement errors to some or all the observables so as to make the number of shocks (structural and measurement) equal to the number observables or to increase the number of structural shocks in the model, for example by transforming parameters into disturbances (the discount factor becomes a preference shock, the elasticity of substitution between goods in an aggregator becomes a markup shock, etc.). A more appealing alternative is to construct the composite likelihood using non-singular submodels, see also Qu (2015). To illustrate the approach and the consequences of using a composite likelihood, we use a stylized asset pricing example.

Suppose that the dividend process is given by $d_t = e_t - \zeta e_{t-1}$, where $e_t \sim iid(0, \sigma^2)$, and $\zeta < 1$, and that stock prices are the discounted infinite sum of future dividends.

The solution for stock prices in terms of the dividend innovation is $p_t = (1 - \beta\zeta)e_t - \zeta e_{t-1}$,where $\beta < 1$ is the discount factor. Since the same shock e_t drives both dividends and stock prices, the covariance matrix of (d_t, p_t) is singular. Thus, one has to decide either whether the information in d_t or in p_t is used to construct the likelihood and to estimate the common parameters (ζ, σ^2) - clearly, if the dividend process is used, β is underidentified - unless we add a measurement error (which is difficult to justify since neither dividends nor stock prices are subject to revisions) or make β a random variable. When the composite likelihood is employed, information present in both series is used to identify and estimate (ζ, σ^2) and β, if it is of interest. The optimization process implies that dividends and stock prices series contain different types of information; the composite likelihood combines these types of information and thus provides a more flexible way to use all available information to estimate parameters.

Following Hamilton (1994, p. 129), the exact likelihood functions of the two observables are

$$\log L(\tilde{d}_t|\zeta, \sigma^2) = -0.5T\log(2\pi) - 0.5\sum_{t=1}^{T}\log\varsigma_t - 0.5\sum_{t=1}^{T}\frac{\tilde{d}_t^2}{\varsigma_t} \tag{31}$$

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

$$\begin{aligned}\tilde{d}_t &= d_t - \zeta\frac{1+\zeta^2+\zeta^4+\ldots+\zeta^{2(t-2)}}{1+\zeta^2+\zeta^4+\ldots+\zeta^{2(t-1)}}\tilde{d}_{t-1} &(32)\\ \varsigma_t &= \sigma^2\frac{1+\zeta^2+\zeta^4+\ldots+\zeta^{2t}}{1+\zeta^2+\zeta^4+\ldots+\zeta^{2(t-1)}} &(33)\end{aligned}$$

and

$$\log L(\tilde{p}_t|\beta, \zeta, \sigma^2) = -0.5T\log(2\pi) - 0.5\sum_{t=1}^{T}\log\lambda_t - 0.5\sum_{t=1}^{T}\frac{\tilde{p}_t^2}{\lambda_t} \tag{34}$$

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

$$\begin{aligned}\tilde{p}_t &= p_t - \frac{\gamma^2}{\zeta}\frac{1+\gamma^2+\gamma^4+\ldots+\gamma^{2(t-2)}}{1+\gamma^2+\gamma^4+\ldots+\gamma^{2(t-1)}}\tilde{p}_{t-1} &(35)\\ \lambda_t &= \sigma^2(1-\beta\zeta)^2\frac{1+\gamma^2+\gamma^4+\ldots+\gamma^{2t}}{1+\gamma^2+\gamma^4+\ldots+\gamma^{2(t-1)}} &(36)\end{aligned}$$

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

$$\begin{aligned}\frac{\partial\log L(\tilde{d}_t)}{\partial\zeta} &= -0.5\sum_t\frac{\partial\log\varsigma_t}{\partial\zeta} - 0.5\sum_t\frac{\partial\log\tilde{d}_t}{\partial\zeta}\frac{1}{\varsigma_t} + 0.5\sum_t\frac{\partial\log\varsigma_t}{\partial\zeta}\frac{\tilde{d}_t}{\varsigma_t^2} &(37)\\ \frac{\partial\log L(\tilde{p}_t)}{\partial\gamma} &= -0.5\sum_t\frac{\partial\log\lambda_t}{\partial\zeta} - 0.5\sum_t\frac{\partial\log\tilde{p}_t}{\partial\zeta}\frac{1}{\lambda_t} + 0.5\sum_t\frac{\partial\log\lambda_t}{\partial\zeta}\frac{\tilde{p}_t}{\lambda_t^2} &(38)\end{aligned}$$

For a given weight ω assigned to $\tilde{d}_t$, the composite likelihood is a weighted sum of (37) and (38). While there are no closed expressions for either the maximum likelihood or the maximum composite likelihood estimators of ζ, which would allow a direct comparison of the properties of the two estimators, we can still infer what (37) and (38) employ to estimate ζ and what a composite likelihood does using simple algebra. In appendix A we show that ζ will be identified and estimated more from the serial correlation properties of the data if $\tilde{p}_t$ is used to construct the likelihood function and more from the variance properties of the data if $\tilde{d}_t$ is used to construct the likelihood function. Hence, estimates obtained from (38) generally differ from those obtained with (37), because the former weighs more the low frequency components of data.

The composite likelihood provides a compromise between these two types of information. Depending on the value of ω, either the serial correlation properties or the variance properties of (d_t, p_t) or both will be employed. Clearly, if the low frequency components of $\tilde{p}_t$ are poorly characterized because, for example, the sample is short or because ζ is close to zero, the composite likelihood provides a better objective function to identify and estimate ζ than each of the individual likelihood functions. In addition, if d_t is smooth and p_t is highly volatile, the composite likelihood may provide a more stable estimate of ζ than standard individual likelihood functions.

3.4 Dealing with large scale structural models

Combining the examples we have considered so far, we can analyze the situation when one needs to estimate the parameters of a large scale model when the number observable variables potentially exceeds the number of shocks. Suppose the decision rules of the model can be written as $\mathrm{y}_t = A(\theta)y_{t-1} + e_t$, where e_t iid $\mathrm{N}(0,\Sigma(\theta))$, θ is a vector of structural parameters, y_t is of large dimension and, generally, $\dim(\mathrm{y}_t) \geq \dim(e_t)$.

Let $\tilde{y}_t \subset y_t$ be a subset of the variables such that $\dim(\tilde{y}_t) = \dim(e_t)$ and let $\widetilde{A(\theta)}$ be the square version of $\mathrm{A}(\theta)$ corresponding to $\tilde{y}_t$. The likelihood function is

$$L(\tilde{y}|A(\theta),\Sigma(\theta)) = (2\pi)^{-T/2}|\Sigma|^{T/2}\exp\{(\tilde{y}_t - \widetilde{A(\theta)}\tilde{y}_{t-1})\Sigma(\theta)^{-1}(\tilde{y}_t - \widetilde{A(\theta)}\tilde{y}_{t-1})'\} \quad (39)$$

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

$$L(\tilde{y}_1|A,\Sigma) = \int L(\tilde{y}_1|\tilde{y}_2, \widetilde{A(\theta)}, \Sigma(\theta))g(\tilde{y}_2)d\tilde{y}_2 \quad (40)$$

which may be intractable.

Rather than trying to compute the likelihood for $\tilde{y}_{1t}$, we can take a limited information approach and produce estimates the parameters using objects which are simpler to construct. Let $\hat{y}_t$ be the set of observable variables. If we partition $\hat{y}_t = (\hat{y}_{At}, \hat{y}_{Bt}, \ldots \hat{y}_{Kt})$, where $\dim(\hat{y}_{At}) = \dim(\hat{y}_{Bt}) = \ldots = \dim(e_t)$, two such objects one

can consider are:

$$CL_1(\hat{y}_t|A_i(\theta),\Sigma(\theta)) = \sum_{i=1}^{K}\omega_i \log L(\hat{y}_{it}|A_i(\theta),\Sigma(\theta)) \quad (41)$$

$$CL_2(\hat{y}_{it},A_i(\theta),\Sigma(\theta)) = \sum_{i=1}^{K}\omega_i \log L(\hat{y}_{it}|\hat{y}_{-it},A_i(\theta),\Sigma(\theta)) \quad (42)$$

where $A_i(\theta),\Sigma(\theta)$ are the autoregressive and the variance parameters corresponding to $\hat{y}_{it}$ and $\hat{y}_{-it}$ indicates the combinations of the vector $\hat{y}_t$, which exclude the i-th combination.

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

3.5 Dealing with short T when a panel is available.

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

$$CL(\hat{y}_{1t},\hat{y}_{2t},...\hat{y}_{Kt}|A(\theta),\Sigma(\theta)) = \sum_{i=1}^{K}\omega_i \log L(\hat{y}_{it}|A(\theta),\Sigma(\theta)) \quad (43)$$

(17) in this case neglects the correlation structure across units, but pools information about the common parameters from the available cross section. Given a linear autoregressive structure for the decision rules, the pooling procedure produces estimators for θ which are similar to those derived by Zellner and Hong (1989): they weight individual and a (weighted) average of the information present in the cross section of data. This is clear when looking at (21), once it is realized that terms such as $\sum_{i=2}^{K}\phi_{i2}\sum_{t=1}^{T_i}y_{it-1}^2$ represent a weighted average of the information present in the data of the units other than the first one. Such a pooling approach is likely to be superior when each $\hat{y}_{it}$ is short, since the composite likelihood uses the information present in the panel (rather than in single individual time series). However, the cross sectional information is not exactly pooled: the degree of shrinkage depends on the the precision of various sources

of information. Thus, the composite likelihood uses at least as much information as the likelihood of individual units, exploits commonalities in cross section, if they exists and may lead to improved estimates of the vector of common parameters θ.

4 Reducing misspecification

As we have mentioned the shrinkage estimators that the composite likelihood generates can help to reduce biases in likelihood (posterior) estimates obtained with misspecified models. The logic is relatively simple: when the baseline model is misspecified, information contained in additional (misspecificied) models restrict the range of values that the common parameters can take and thus the quality of the estimates may improve both in terms of location and dispersion. This is similar to having N imperfect instruments in IV estimation: estimation with one instrument is likely to be less successful than with N instruments.

To show in which practical situations this is more likely to occur, we run a simulation exercise. We assume that the DGP is a univariate AR(2) $y_t = \rho_1 y_{t-1} + \rho_2 y_{t-2} + e_t,\ e_t \sim (0, \sigma^2)$. The models we consider for estimation are an AR(1): $y_t = \rho_1 y_{t-1} + u_t$ and an MA(1): $y_t = \epsilon_t + \theta_1 \epsilon_{t-1}$. We focus attention on estimates of σ^2, the variance of the estimated error term, which is common across models. We are interested in examining how posterior estimates relate to the true σ^2. Given that both models are misspecified relative to the DGP, $\sigma^2_{u,ML}, \sigma^2_{\epsilon,ML}$ are likely to display biases. The question of interest is whether the composite posterior, which jointly considers the two models, gives us better estimates of σ^2 than those obtained with the AR(1) or the MA(1) and in what conditions.

We first consider fixed weights and let ω be the weight for the AR(1) model. We present composite posterior estimates obtained in a number of interesting cases: i) equally weighting the two models, ii) using weights based on the MSE or the Marginal likelihood for the two models in a training sample; iii) using the weights that optimize the composite marginal likelihood in a training sample. The training sample consists of 100 observations and the estimation sample of 50 observations; since there are only two parameters to estimate in the AR(1) and MA(1) models, and three when the composite likelihood is used, this is actually a medium sized sample.

We consider a number of configurations for ρ_1, ρ_2, σ^2 in order to gain insights about the cases where a composite likelihood approach helps most. Table 1 reports a subset of the results we obtain: for each DGP configuration, we report the posterior mean and the posterior standard deviation of σ^2 in the AR(1) and MA(1) models and in various composite posterior setups we consider. In all cases, the prior for the AR (MA) parameters is loose (mean equal to zero and variance equal to 1) and the prior for sigma is relatively flat in the positive orthant.

There seems to be location gains when using the composite posterior. The gains are larger whenever the DGP is persistent or has a large volatility and the results seem insensitive to the choice of weights. As often documented in the forecasting combination literature (see Aiolfi et al, 2010), choosing equal weights is as good as

Table 1: Estimates of σ^2

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

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

choosing the weights either based on MSE or the marginal likelihood of the AR(1) and MA(1) models in the training sample (compare columns 4, 5 and 6). However, choosing the weights to optimize the performance of the composite likelihood in the training sample, seems to give an important hedge to the approach: location gains are large and they increase, the smaller is the volatility of the DGP. It is important to stress that the approach employed in column 7 is feasible even when the models feature different observables, while this is not the case for the results produced in column 5. When models feature a common subset of observables, an alternative approach, for example, based on the average log-scores (see Geweke and Amisano 2011) constructed using variables common to all models could be used. The table does not show much gains relative to the AR(1) or MA(1) models as far as the spread of the posterior is concerned. Two reasons account for this outcome. First, we only consider two models; dispersion gains are more likely to occur when the number of models one consider is larger. Second, mean estimates of σ^2 obtained with the AR(1) and the MA(1) models do not differ much for many parameter configuration. Thus, dispersion gains are relatively small.

The first panel of Figure 1 presents the composite posteriors of σ obtained when the data has been generated by $y_t = 0.6y_{t-1} + 0.35y_{t-2},\ e_t, \sim N(0, 0.5)$ in three cases: equally weighting the AR(1) and the MA(1) models; optimally selecting ω to maximize the composite marginal likelihood in the training sample; and letting ω be a random variable with a normal prior distribution centered at 0.5 and standard deviation equal to 0.1,

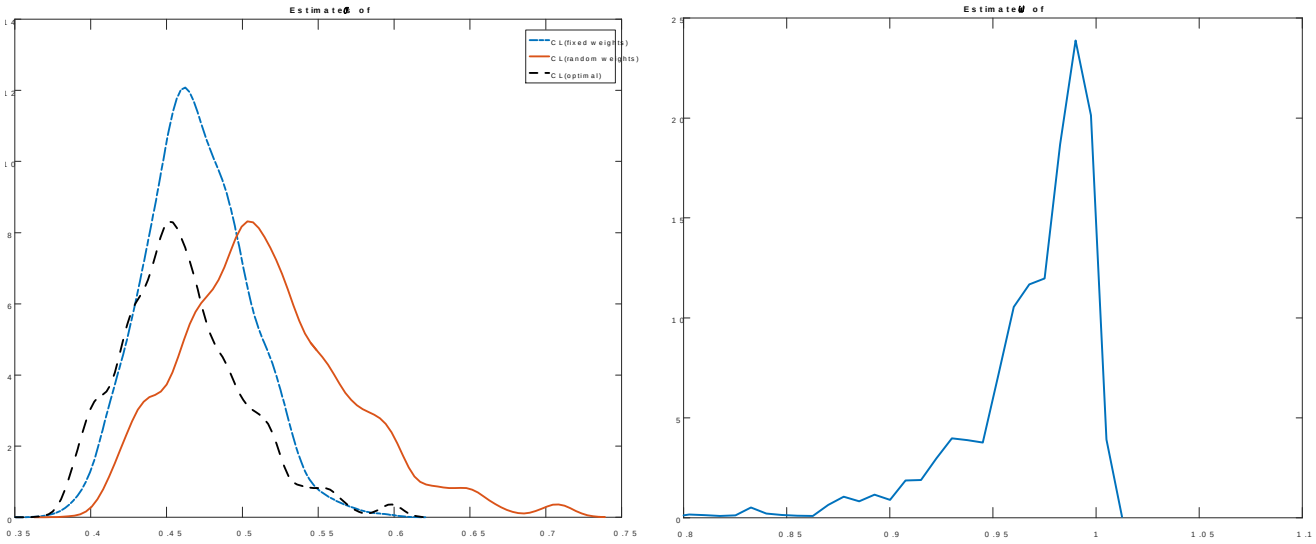

Figure 1: Composite posteriors and posteriors for ω.

The shape of posterior is similar equally weighting the two models or selecting ω optimally. However, the posterior of σ obtained with optimal weights is more leptokurtic and displays much longer tails. The posterior of σ with random weights is centered at the true value (mode=0.502) but has a larger dispersion relative to the other two posterior distributions, due to the fact that there is additional uncertainty in the model (there is one extra random variable) and, as shown in the second panel of Figure 1, the posterior of ω is heavily skewed and has a very long left tail. Thus while there are location gains from having a random ω in this particular case, taking ω random could increase the dispersion of the posterior of the common parameters. Notice that having a random ω is probably more appealing from a theoretical point of view when composite posterior gains obtained using fixed weights are parameterization dependent.

5 An example

Schorfheide (2008) surveyed estimates of the slope of the Phillips curve and tried to explain for the reported differences. He finds that variations in the existing estimates are large and that they depend on i) the exact model specification, iii) the observability of marginal costs, iii) the number and the type of variables used in estimation. Given this background, we are interested in examining how the posterior distribution of the slope of the Phillips curve would look like when composite likelihood approach is employed. We consider four models: a small scale New Keynesian model with sticky

prices where marginal costs are non-observable as in Rubio and Rabanal (2005) where the variables used in estimation are detrended output Y, demeaned inflation π, and demeaned nominal rate R;a small scale New Keynesian model with sticky prices and sticky wages, where marginal costs are observables, again as in Rubio and Rabanal (2005), where the variables used in estimation are detrended Y, demeaned π, demeaned R and detrended nominal wage W; a medium scale New Keynesian model with sticky prices, sticky wages, habit in consumption and investment adjustment costs as in Justiniano et al. (2010), where the variables used in estimation are detrended Y, demeaned π, demeaned R , detrended nominal W, detrended consumption C, detrended investment I, detrended hours worked N; and a monetary search and matching model, as in Christoffel and Kuester (2008) where the variables used in estimation are detrended Y, demeaned π, demeaned R and detrended real wage w. Detrending in all case is done with a quadratic trend and, for comparability, the estimation sample is 1960:1-2005:4 for all four models. The priors for the structural parameters of various models are standard. Because the models feature different numbers of observables, we can not use random weights (since the posterior weight on the larger model will approach 1, simply because the likelihood of a larger model is higher than the likelihood of a smaller model.). In the exercise we report, we fix choose $\omega = (0.25, 0.25, 0.25, 0.25)$; choosing a higher weight on the larger model ($\omega_3 = 0.4, \omega_i = 0.2$, i =1,2,4) produces a composite posterior with similar features.

Table 2:Estimates of the slope of the Philips curve

	5%	50%	95%
Prior	0.01	0.80	1.40
Basic NK	0.06	0.18	0.49
Basic NK with nominal wages	0.05	0.06	0.07
SW with capital and adj.costs	0.04	0.05	0.07
Search model	0.44	0.62	0.86
CL	0.13	0.16	0.21

Reported are the posterior estimates of the slope of the Phillips curve for a three vairable 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 7 observables (SW with capital and adj. costs) and the four variable search and matching model. The row with CL reports estimates obtained with the composite likelihood approach.

Table 2 displays some percentiles of the posterior distribution for the slope of the Phillips curve (κ_p) in each of the four models and when the composite likelihood is used to aggregate their information. The posteriors for the first three models have modes close to zero and, as Schorfheide suggested, having non-observables marginal costs tend to increase the location of the mode. The search an matching model instead has a much higher mode and the posterior does not overlap with the posterior obtained with the large scale model or the small scale sticky-price sticky-wage model.

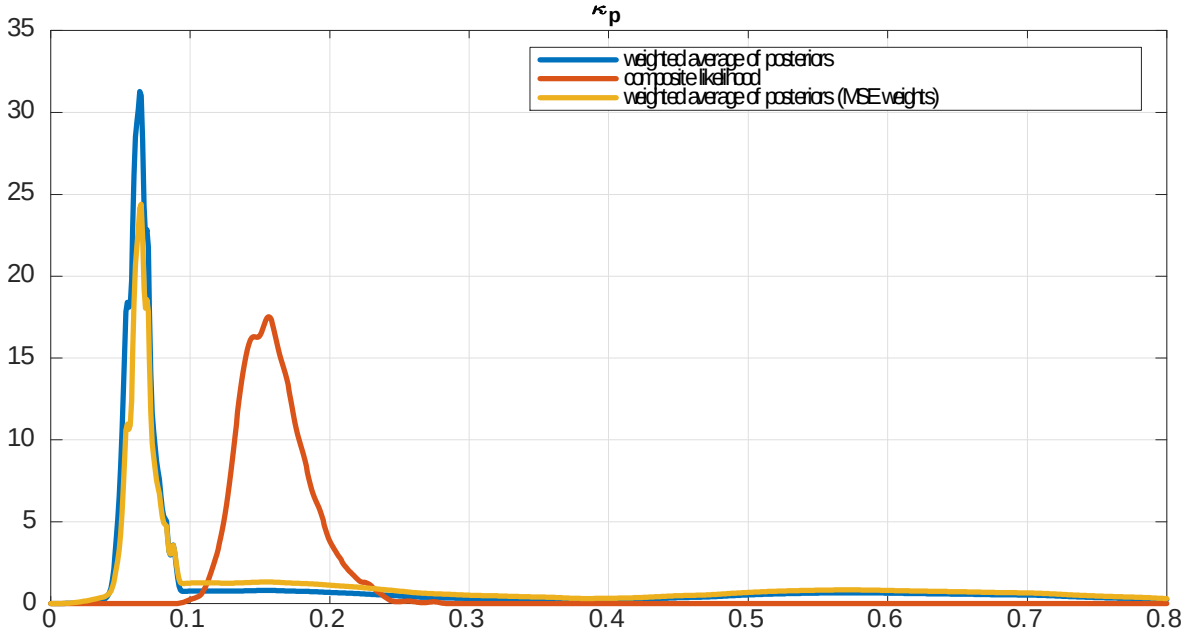

Figure 2: Composite posterior density and naive combination densities.

Thus, again in agreement with Schorfheide the estimation results seem to depend on which observable variable is used in estimation. In many cases, the spread of the posterior is relatively large and in a few cases the amount of overlap is large. The mode of the posterior obtained with the composite likelihood is centered around 0.2, and the posterior has a smaller spread than the one of individual models. Thus, much sharper inference about the effect of, say, increasing marginal costs can be made in any of these models with the composite posterior estimates.

Figure 2 present the composite posterior for the slope of the Phillips curve obtained together with two alternative naive combination posterior estimators: one that equally weights the posteriors of κ_p obtained with the four models; and one which weights the posteriors of κ_p by the in-sample MSE they produce for inflation and the nominal rate. Two features of the figures deserve discussion. First, ex-post combining estimates of κ_p obtained with the four models produce results which are different from those generated by a composite likelihood approach; both location and spread differences are important. Second, the way naive combination estimators are constructed is irrelevant: the combination posterior produced with equal weights or MSE weights are indistinguishable.

6 Conclusions

This paper describes how to use the composite likelihood approach to solve or ameliorate estimation problems in DSGE models, shows how the procedure helps to robustify estimates of the structural parameters in a variety of interesting economic problems, highlights how to perform composite posterior inference, and provides intuition on how the methodology can be applied to the estimation of the parameters of structural models.

We show that the approach it is easy to implements, works well when the full likelihood may be problematic to construct and use, produces estimators with nice shrinkage properties and, in its Bayesian version, it has an appealing sequential learning interpretation.

We presented a number of examples where the procedure can be used to i) obtain shrinkage estimates of the parameters appearing in multiple (nested and non-nested) misspecified structural models; ii) improve their (sample and population) identification properties, iii) provide a tractable approach to solve computational and singularity problems; iv) exploit information coming either from the cross-section or from different levels of data aggregation; v) produce more stable estimates of parameters present in large scale models.

Finally, we shows how inference in misspecified models can be improved and how estimates of the slope of Phillips curve can be robustified using the composite likelihood constructed using multiple nested and non-nested models .

7 References

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

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

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

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

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

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

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

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

Christiano, L., Trabant, M. and K. Walentin (2011). DSGE models for policy analysis in B. Friedman and M. Woodford (eds.) Handbook of Monetary Economics, 3A, Elsevier, North Holland, The Netherland, 285-368.

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

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

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

Del Negro, M., Hasegawa, R., and F. Schorfheide (2016). Dynamic Prediction Pools: An Investigation of Financial Frictions and Forecasting Performance. Journal of Econometrics, 192, 391-405.

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

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

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

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

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

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

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

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

Iskrev, N. (2010). Local identification in DSGE models. Journal of Monetary Economics, 57, 189-202.

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

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.

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

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

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

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

8 Appendix A

Consider observations t=1 and t=2. From (32) and (35) we have

$$\tilde{d}_1 = d_1 \tag{44}$$

$$\tilde{d}_2 = d_2 - \frac{\zeta}{1+\zeta^2} d_1 \tag{45}$$

$$\tilde{p}_1 = p_1 \tag{46}$$

$$\tilde{p}_2 = p_2 - \frac{\zeta}{(1-\beta\zeta)^2+\zeta^2} p_1 \tag{47}$$

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

Similarly, using (33) and (36), for t=1 and t=2 we have

$$\varsigma_1 = \sigma^2(1+\zeta^2) \tag{48}$$

$$\varsigma_2 = \sigma^2 \frac{1+\zeta^2+\zeta^4}{1+\zeta^2} \tag{49}$$

$$\lambda_1 = \sigma^2((1-\beta\zeta)^2+\zeta^2) \tag{50}$$

$$\lambda_2 = \sigma^2(1-\beta\zeta)^2 \frac{1+\frac{\zeta^2}{(1-\beta\zeta)^2}+\frac{\zeta^4}{(1-\beta\zeta)^4}}{1+\frac{\zeta^2}{(1-\beta\zeta)^2}} \tag{51}$$

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