A Composite likelihood approach for dynamic structural models

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

January 30, 2017

Abstract

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

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

JEL Classification: C10, E27, E32.

[*] We thank Bruce Hansen, Mark Watson, Ulrich Mueller, Barbara Rossi, Refet Gurkenyak, Ferre de Grauve, Geert Dhaene and the participants at the ESEM 2016 invited session on 'New development in the analysis of Macroeconomic data', the 2016 ESOBE conference, and seminars at Singapore National University, Katholic University Leuven, the Riksbank for comments and suggestions. Canova acknowledges the financial support from the Spanish Ministerio de Economia y Competitividad through the grants ECO2012-33247, ECO2015-68136-P and FEDER, UE. The views presented in this paper do not reflect those of the Federal Reserve Bank of Richmond, or the Federal Reserve system.

1 Introduction

In macroeconomics it is now standard to construct dynamic stochastic general equilibrium (DSGE) models and analyze their properties. Until a decade ago, economic analyses, policy and forecasting exercises were conducted using parameters formally or informally calibrated. Nowadays, it is more common to conduct inference using classical or Bayesian full information likelihood approaches, see Andreasen et al. (2014) for an exception.

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

Moreover, inference in estimated DSGE models is also not as straightforward as one would like to think. For standard frequentist asymptotic theory to apply, one needs regularity conditions, which are often violated in practice. For example, while at the designing stage there is a general consensus that all available models are misspecified in, at least, some dimensions, at the estimation stage this fact is generally neglected. asymptotic approximations provide a poor characterization of the actual distribution of estimates. Bayesian methods may help when the sample size 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. While the recent literature has tried to address some of these estimation and inferential problems individually, to the best of our knowledge, no unified approach to deal with all these issues exists.

Because policymakers are aware of the fact that models are misspecified, there are structural breaks or instabilities in the data generating process, they are often interested in pooling estimates obtained from different models so as to robustify counterfactuals exercises and improve forecasting performance, see e.g. Aiolfi et al. (2010). However, rather than combining the outputs, one may want to combine the inputs of the estimation process so as to provide ex-ante rather than ex-post averaging,

This paper proposes an approach based on the *composite likelihood* which can potentially address the problems noted in this introduction. The procedure is simple to implement, helps with parameter identification, produces estimators with nice shrinkage properties and, in its Bayesian version, has a number of appealing features.

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

In our setup, the composite likelihood combines the likelihood of distinct structural or statistical models one may want to consider to answer a question of interest. Moreover, all models considered are assumed to be misspecified. Because of these features, standard asymptotic results do not necessarily apply. Still, as Lindsay (1980) pointed out, models may contain common features, which may be agreed upon by a number of researchers, and these may be all we care about for inference. We show how to conduct asymptotic inference with the composite likelihood constructed from misspecified models and show how exact small sample statements about the common parameters of interest can be obtained in a Bayesian setting, by treating the composite likelihood as a quasi-likelihood function, as long as the prior makes the composite posterior a proper density function. The idea that a prior can be used to regularize objective functions which are not likelihood functions is well established in the econometric literature, see e.g. Kim (2002) or Chernozukov and Hong (2003). 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.

We show that posterior inference is easy to perform and describe a block MCMC algorithm that can be used to draw posterior sequences of the parameters. We also show that, because of the data present in each model is not necessarily mutually exclusive, the asymptotic of Bayesian MCMC estimators are non-standard - under flat prior the asymptotic covariance matrix differ from the one produced by maximum likelihood. We describe a simple variation of the MCMC algorithm that achieves the correct asymptotic coverage, if that is of interest. We describe how to conduct forecasting exercises, impulse response analysis and counterfactual exercises using our composite likelihood setup and compare the structure we obtain with the one obtained with other pooling devices, such as mixture models or Bayesian model averaging (BMA).

We present examples indicating how the composite likelihood, constructed using the information present in distinct misspecified models, can be used to address the problems mentioned in this introduction, and show how the approach helps 1) to robustify parameter estimates and to reduce misspecification, 2) to ameliorate population and sample identification problems, 3) to solve singularity problems, 4) to combine information coming from different sources, frequencies and levels of aggregation, and 5) to produce more stable estimates of the parameters of large scale structural models. Finally, we apply the methodology to improve the quality of the inference a researcher is able to draw from the data about the slope of the Phillips curve.

The rest of the paper is organized as follows. The next section presents the traditional composite likelihood approach and describes the asymptotic properties of composite likelihood estimators. It also introduces our setup, highlight differences with the traditional setting and the implications for asymptotic inference and compare our approach to a finite mixture model. Section 3 discusses Bayesian estimation and some non-standard features of our estimation problem. Section 4 applies our approach to the problem of constructing density of forecasts, impulse responses and policy counterfactuals and discusses differences with BMA. Section 5 presents a number of examples which highlight the properties of the methodology in relevant settings. Section 6 present the results of estimating the slope of Phillips curve using a variety of structural models. Section 7 concludes. Three appendices provide formal statements of arguments discussed in the text.

2 The composite likelihood

The original composite likelihood formulation has been suggested to deal with hard computational problems. Two situations are of particular interest: when the full likelihood of the model is difficult to construct because of latent variables (the likelihood of the observables features multi-dimensional integrals); or when it is hard to manipulate because the model is of large scale and covariance matrix of the observables is nearly singular. In these situations, it might be preferable to use a weighted average of marginal or conditional distributions of submodels (or 'events' in the nomenclature of the composite likelihood literature) . In fact, this weighted average provides a limited information object that, under regularity conditions, allow consistent estimation of the parameters common to all submodels. As long as all limited objects entering the composite likelihood have similar properties, it is also relatively easy to derive asymptotic normality results for composite maximum likelihood estimators (e.g. Varin, et al., 2011). However, since the components entering the composite likelihood may be correlated, it is not necessarily true that adding components will improve the informational content of the composite likelihood. For this reason, component-selection criteria have been designed to optimally select the amount of "information" present in the composite likelihood.

As it will be clear later on, our submodels are not necessarily marginal or conditionals versions of the original model. Instead, they are potentially misspecified (or underspecified) representation of an unknown DGP. Since each component entering the composite likelihood is misspecified, the composite likelihood estimator is, in general, inconsistent. However, as we will show, it is possible to conduct frequentist inference with a misspecified composite likelihood; use a Bayesian framework to construct small sample distributions of the object of interests; and more importantly for our purposes, i) robustify inference and reduce submodel misspecification, and ii) address standard estimation problems.

2.1 The standard setup

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

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 versions of $F(y_t, \psi)$ depending on the specification of the problem. Each A_i defines a submodel, which has implications for a subvector y_{it} of length T_i of the observables.[1] In some applications, see Engle et al., 2008 or Varin, et al., 2011, $F(y_t, \psi)$ is conceptually tractable, but the dimensionality of ψ is such that handling marginal or conditional densities based on f is computationally appealing. Each A_i is associated with a vector of parameters $\psi_i = [\theta, \eta_i]'$, where η_i are (nuisance) event specific. We represent the information generated by A_i with the tuple (y_{it}, T_i, ψ_i).

Given a vector of weights ω_i, the composite likelihood is

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

$CL(\theta, \eta, y)$ is not a likelihood function and thus the properties of θ_{CL}, the maximum composite likelihood estimator, are unclear. If $y_{[1,t]}$ =$(y_1, \dots, y_t)$ is independent sample from $F(y_t)$, the weights ω_i are fixed, it can be shown that θ_{CL} is consistent and

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

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 A_i are constructed to get asymptotic results. For example, if they are independent, then (2) still holds. Note that a standard Newey-West correction to $J(\theta)$ can be used if $y_{[1,t]}$ is not an independent sample.

Consistency obtains because each component entering the composite likelihood is an unbiased estimating function converging, as T increases, to the true parameter

[1] Marginal or conditional here refers to densities that either integrate out all elements of y_t that are not in y_{it} or that condition on the elements of y_t that are not in y_{it}. For ease of reading, we generally do not make explicit the conditioning on past values of y_{it} if the subdensities assume that y_{it} is persistent.

vector. Asymptotic normality holds as the sampling distribution of the maximum likelihood estimator can be approximated quadratically around the mode. However, since the components entering the composite likelihood are limited information objects, the asymptotic covariance matrix of the composite likelihood estimator is different from the Fisher information matrix and thus θ_{CL} is not fully efficient - the Godambe matrix, $G(\theta)$ equals to the Fisher information matrix, $I(\theta)$ only if the composite likelihood is the likelihood of the true model. Carefully choosing ω_i may improve the efficiency of θ_{CL} and optimal weights can be obtained by either minimizing the distance between $G(\theta)$ and $I(\theta)$ or by making sure that the composite likelihood ratio statistics has a χ^2 asymptotic distribution (see Pauli et al., 2011).

If consistency is all that one cares (or efficiency issues are not of interest), a standard choice is $\omega_i = \frac{1}{K}, \forall i$. Alternatively, one could use a data-based approach to select ω_i. For example, one could 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 $\gamma_i = f_i(Y_{1,[1:\tau]}, ..., Y_{K,[1:\tau]})$. If these statistics are updated over time, ω_i could also be made time varying. Later on we describe a procedure to choose ω_i which miminizes a measure of distance between $F(y_t)$ and the composite density. There a large forecasting literature (see e.g. Aiolfi et al., 2010) which could be used refine training sample-based estimates of ω_i. As will be clear below, our approach is to treat ω_i as a random variable with a prior distribution (to be interpreted as the investigator prior assessment of the likelihood of submodel i) and jointly estimate the posterior distribution of the common parameters θ, of the nuisance parameters ψ_i, and of the weights ω_i).

When K or the number of nuisance parameters η_i is large, joint estimation of the vector of parameters may be demanding and a two-step estimation approach is possible. For example, if ω_i is treated as fixed, η_i could be estimated separately from each $\log f(y_{it} \in A_i, \theta, \eta_i)$ and plugged in the composite likelihood, which is then optimized with respect to θ, see e.g. Pakel et al. (2011). Consistency of θ_{CL} is unaffected as long as η_i are consistently estimated.

Given that the events A_i are not necessarily independent, there is a trade-off between the number of submodels K included in the composite likelihood and the efficiency gains obtained with the composite likelihood estimator. For this reason, one can design information criteria to optimally select K. Thus:

$$AIC_{i,CL} = -2CL(\theta_{CL}, \eta_{i,CL}, y) + 2dim(\theta) \tag{7}$$

$$BIC_{i.CL} = -2CL(\theta_{CL}, \eta_{i,CL}, y) + 2dim(\theta)\log K \tag{8}$$

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 with standard shrinkage methods(see Gao and Song, 2011).

2.2 Our setup

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

Second, $f(y_{it} \in A_i, \theta, \eta_i)$ are neither marginal nor conditional densities. Instead, they represent either misspecified approximations (simplifications) or alternative incompletely specified statistical descriptions of an unknown DGP.

To be concrete, in one leading example we have in mind, A_i are different structural models, e.g. a basic RBC model, a RBC model with financial frictions, a New Keynesian model with sticky price, a new Keynesian model with labor market frictions, etc., y_{it} is the data generated by these models, and $f(y_{it} \in A_i, \theta, \eta_i)$ the associated densities. Different structural models are treated as misspecified subdensities because they disregard aspects of the DGP - a closed economy model is used even if trade interdependencies are non-negligible; take short cuts to modelling the complexities of the DGP - a class of submodels may feature habit in consumption to mimic higher order serial correlation present in the DGP; or explicitly condition on certain features which may be present or absent from the DGP - certain markets are treated as competitive or non-competitive. Formally, there is no value of $\psi_i \in \Psi_i$ such that $F(y_t) = f(y_{it}, \theta, \eta_i)$, for all i. Here, θ are the parameters common to all submodels, e.g. the risk aversion coefficient, the Frisch elasticity, or the labor share, while η_i could be structural parameters, with a DGP counterpart, e.g. a LTV ratio or a Calvo parameter, or reduced form parameters used to approximate features of the DGP.

In another leading example we have in mind, $F(y_t)$ is a large scale structural model, for example, a multi-country model of trade and financial 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.

A third case of interest is one where $f(y_{it} \in A_i, \theta, \eta_i)$ are the densities generated by different approximate (perturbed or projected) solutions of a model or the densities of linear solutions, where only the k-th component of ψ is allowed to be time varying. Here A_i represent either the order of the approximation employed, or an indicator function describing the component of the parameter vector which is allowed to change.

A final case of interest is one where $f(y_{it} \in A_i, \theta, \eta_i)$ represent different *statistical* models. We term models 'statistical' if they are obtained from the same theoretical model but feature different observables. For instance, a standard three-equations New-Keynesian model could be estimated using inflation, the nominal interest rate, and a measure of output; or inflation, the nominal interest rate, and a measure of consumption - in the model consumption and output are equal. These two set of observables constitute what we call different statistical models. By extension, $F(y_t)$ could also be the density of an aggregate model and $f(y_{it} \in A_i, \theta, \eta_i)$ the densities obtained when i) data from cross sectional unit i is used; ii) data from a particular aggregation level (e.g. firm, industry, regional, etc.) is employed; or iii) data for different samples (say, pre-WWI, interwar, post-WWI, etc.) is used. As it is shown in the next few sections, in these situations a composite likelihood approach produces shrinkage estimators for the common parameters θ, which combine unit specific and average information contained in the remaining statistical models.

In all the examples we considered, $f(y_{it} \in A_i, \theta, \eta_i)$ ignores the potential dependence of the A_i. In addition, while in the traditional setup, submodels are compatible, in the sense that asymptotically likelihood estimates of θ converge to the same true value, this need to be the case here. Furthermore, since the subdensities may feature nuisance parameters η_i, inference for θ is generally affected. Finally, the elements of y_{it} need not be mutually exclusive across i; for example, the inflation rate could appear in each y_{it}. This fact has implications for the standard errors of the estimates we obtain.

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 models, researchers might prefer not to estimate a common value. In the case A_i represents different statistical models, one could estimate an independent 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 may be event specific. Clearly, one can think of situations where the parameters θ are common to a subset of the K events one wishes to consider.

Because of misspecification, the maximum likelihood estimators we obtain from each A_i will not be consistent and, as a consequence, the composite maximum likelihood estimator will not be consistent. Following seminar work by White (1982), Domowitz and White (1982) and others, it is possible to show that the estimator of θ_i obtained under regularity conditions and from properly scaled version of the likelihood function of each submodel will converge, as $T \to \infty$ to the pseudo-parameter value, say θ_0, which minimizes the Kullback-Liebler (KL) distance from the true DGP. Moreover, $\sqrt{T}(\theta_{i,ML} - \theta_0)$ has normal distribution with zero mean and variance equal to the Godambe information matrix for that submodel.

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

Mimicking the argument used for each submodel A_i, one can show that $\sqrt{T}(\theta_{CL} - \theta_{0,CL})$ has a normal distribution with zero mean and asymptotic covariance matrix equal to the Godambe information computed using the composite likelihood. Thus, the same results obtained in the standard framework holds also here with two qualifications: the pivot of the asymptotic distribution is the minimizer of the KL distance, rather than the true parameter vector; the Godambe information matrix is evaluated at the minimizer of the KL distance rather than the true parameter vector.

Appendix A provides formal support for these arguments.

2.3 A comparison with finite mixture models

The composite likelihood weights the likelihood of different submodels. An alternative way to pool the information contained in different submodels comes from ideas of Geweke and Amisano (2011), Billio et al. (2013), or Del Negro et al. (2016). The work of Waggoner and Zha (2012) is also relevant 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

$$\log L = \sum_{t=1}^{\tau} \log L(\theta, \eta_1 .... \eta_k | y_{1t} ... y_{kt}) \tag{9}$$

where $\tau = min(T_i)$.

Simple manipulations reveal that (9) and the log of (1) differ by a Jensen's inequality term: in the composite likelihood the objective function is a convex combination of log-likelihoods; in the finite mixture, the objective function is the log-likelihood of a mixture [2].

While a-priori both specifications are appealing, for the problems we are interested in, the composite likelihood has superior properties and added flexibility. From a computational point of view, when the decision rules have an autoregressive structure, estimators of θ have a closed form expression in the composite log-likelihood case, but not in the log-likelihood of the finite mixture. 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 submodels represent alternatives that could have generated the observable data. These restrictions are unnecessary in the composite likelihood formulation and this gives flexibility to the methodology - see the next section for examples. In addition, in finite mixture setups misspecification of one or all models is not allowed. Third, when they are treated as fixed, the interpretation of ω_i differs: in the composite likelihood ω_i represents the proportion of observations coming from submodel i. In the mixture model ω_i represents the probability that one observation comes from model

[2] The difference between the two 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 (9) 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.

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

3 Bayesian estimation

Because we are interested in the exact small sample distribution of the common parameter vector θ, rather than its large sample approximation, we estimate θ using Bayesian methods. We treat ω as an additional vector of parameters whose posterior needs to be estimated. We combine the composite likelihood (1) with the prior for $\phi= (\theta, \eta_1, ...\eta_K, \omega_1, \ldots, \omega_K)$, compute the posterior for ϕ which we then integrate with respect to (η, ω) to obtain the posterior for θ. Because there is no a closed form expression for this posterior, we describe a Metropolis-within-Gibbs approach to numerically compute sequences for θ from $p(\theta|y_{1t}, \ldots y_{Kt}) = \int \ldots \int p(\phi|y_{1t}, \ldots, y_{Kt}) d\eta_1, \ldots d\eta_K, d\omega_1, \ldots, d\omega_K$.

Given (y_{it}, T_i), we assume the $\mathcal{L}(\psi_i|y_{i,T_i})$ can be constructed for each submodel i and that the composite likelihood $\mathcal{L}(\phi|y_{1,T_i}, \ldots, y_{K,T_k})$ can be computed for $0 < \omega_i < 1$, $\sum_i \omega_i = 1$. We assume that the priors for the parameters are of the form:

$$p(\phi_i) = p(\omega_i)p(\theta)p(\eta_i|\theta) \tag{10}$$

Note that in (10) we allow for an endogenous prior selection for η_i, in the spirit of Del Negro and Schorfheide (2008), which is advisable if we think of the composite likelihood components as distinct structural models and we want to put them on the same ground as far as matching certain statistics of the data.

The composite posterior kernel is:

$$\begin{aligned} \check{p}(\psi_1, \ldots, \psi_k|Y_{1,t_1}, \ldots, Y_{k,T_k}) &= \\ \mathcal{L}(\psi_1|Y_{1,T_1})^{\omega_1} p(\psi_1)^{\omega_1} p(\omega_i) \ldots \mathcal{L}(\psi_K|Y_{K,T_K})^{\omega_K} p(\psi_K)^{\omega_K} p(\omega_k) &= \\ \Pi_i \mathcal{L}(\psi_i|Y_{i,T_i})^{\omega_i} p(\psi_i)^{\omega_i} p(\theta)p(\omega_i) & \end{aligned} \tag{11}$$

which can be used to estimate the parameters as described in Kim (2002) or Chernozukov and Hong (2003). To apply standard MCMC techniques, we need the composite posterior kernel to be bounded, which occurs when $\sup_{\psi_i} f(y_{it} \in A_i, \theta, \eta_i) < b_i$, where b_i is finite, a condition generally satisfied for DSGE or VAR models.

For computational and efficiency reasons, rather than using a brute force Metropolis approach to compute the posterior, it may be advisable to employ a $2K + 1$ block Metropolis-within-Gibbs algorithm. Chib and Ramamurthy (2010) and Herbst and Schorfheide (2015) have also suggested drawing DSGE parameters in blocks. However, while they randomly split up the parameter vector in different blocks at each iteration, the blocks here are predetermined by the K submodels of interest. We characterize the uncertainty surrounding the posterior for the parameters using percentiles constructed from the MCMC draws. Chernozukov and Hong (2003) give conditions under which these percentiles asymptotically approximate valid frequentist confidence intervals in correctly specified models.

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

3.1 Estimation Algorithm

The algorithm consists of four steps:

1. Start with some $\Phi^0 = [\eta_1^0 \dots \eta_K^0, \theta^0, \omega_1^0 \dots \omega_K^0]$.
   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

$$\alpha_i = \min\left(1, \frac{\mathcal{L}(Y_{i,T_i}|\left[\eta_i^*, \theta^{iter-1}\right])^{\omega_i} p(\eta_i^*|\theta^{iter-1})^{\omega_i}}{\mathcal{L}(Y_{i,T_i}|\left[\eta_i^{iter-1}, \theta^{iter-1}\right])^{\omega_i} p(\eta_i^{iter-1}|\theta^{iter-1})^{\omega_i}}\right) \tag{12}$$

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

$$\beta = \min\left(1, \frac{\mathcal{L}(Y_{1,T_1}|\left[\eta_1^{iter}, \theta^*\right])^{\omega_1} \dots \mathcal{L}(Y_{K,T_K}|\left[\eta_K^{iter} \theta^*\right])^{\omega_K} p(\theta^*)}{\mathcal{L}(Y_{1,T_1}|\left[\eta_1^{iter}, \theta^{iter-1}\right])^{\omega_1} \dots \mathcal{L}(Y_{K,T_K}|\left[\eta_i^{iter}, \theta^{iter-1}\right])^{\omega_K} p(\theta^{iter-1})}\right) \tag{13}$$

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

$$\delta_i = \min\left(1, \frac{\mathcal{L}(Y_{1,T_1}|\left[\eta_1^{iter}, \theta^{iter}\right])^{\omega_1^*} \dots \mathcal{L}(Y_{K,T_K}|\left[\eta_K^{iter} \theta^{iter}\right])^{\omega_K^*} p(\omega^*)}{\mathcal{L}(Y_{1,T_1}|\left[\eta_1^{iter}, \theta^{iter}\right])^{\omega_1^{iter-1}} \dots \mathcal{L}(Y_{K,T_K}|\left[\eta_i^{iter}, \theta^{iter}\right])^{\omega_K^{iter-1}} p(\omega^{iter-1})}\right) \tag{14}$$

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

3.2 Adjusting percentiles of the posterior distribution

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

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

$$\tilde{\theta}^j = \hat{\theta} + V^{-1}(\theta^j - \hat{\theta}) \tag{15}$$

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

Note that the adjustment works well only when θ is well identified from the composite posterior and if the composite posterior has a unique maximum. As Canova and Sala (2009) have shown, such properties may not hold in a number of DSGE models. In general, when it is difficult to verify whether the composite posterior ha unique maximum or if it is known that the posterior is multimodal, it may be reasonable to report both standard and adjusted MCMC percentiles.

3.3 A sequential learning interpretation

For the sake of illustration, suppose that ω_i are fixed. It is easy to give a sequential, adaptive learning interpretation to the composite posterior kernel (11) and to the

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

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

Bayesian estimators we construct. When K=2, the composite posterior kernel $\check{p}$ is

$$\begin{aligned}\check{p}(\psi_{1....}\psi_2|Y_{1,T_1},Y_{2,T_2}) &= \\ \mathcal{L}(Y_{1,T_1}|\theta,\eta_1)^{\omega_1}p(\eta_1|\theta)^{\omega_1}p(\eta_2|Y_{2,T_2,}\theta)^{\omega_2}\{[p(\theta|Y_{2,T_2})ML(Y_{2,T_2})]^{\omega_2}p(\theta)^{\omega_1}\}\end{aligned} \tag{16}$$

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 $K = 2$.

As (16) makes it clear, the posterior kernel can be obtained in two stages. In the first stage the prior for ψ_2 and the likelihood for submodel $K = 2$ are used to construct the conditional posterior p($\theta|Y_{2,T_2}$). This conditional posterior, weighted by the marginal likelihood of the submodel $K = 2$, is geometrically combined with the prior $p(\theta)$ for the next estimation stage of θ. Suppose that $ML(Y_{2,T_2})$ is high, i.e. submodel $K = 2$ fits Y_{2,T_2} well. If $\omega_1 = \omega_2$, the prior for $K = 1$ will more heavily reflect the posterior of θ obtained from $K = 2$ relative to the initial prior $p(\theta)$. Suppose instead that the specification used for $K = 2$ fits Y_{2,T_2} poorly. In this case, the posterior for θ obtained from $K = 2$ will have a low weight relative to $p(\theta)$ when setting up the prior $K = 1$. In other words, the approach implicitly discounts information contained in submodels whose density poorly explain observable data. In general, the prior for θ in each stage depends on the relative weights assigned to the current and to the previous submodels and on the fit of all previous submodels. Thus, a composite Bayesian approach to estimation can be interpreted as an adaptive sequential learning process.

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

4 Predictions

It is relatively easy to use our Bayesian framework to construct prediction of future observations. Let $\tilde{y}_{t+l}$ be future values of the variables appearing in all the submodels, $l = 1, 2, \ldots$. Let $f(\tilde{y}_{t+l}|y_{it},\theta,\eta_i)$ be the prediction of $\tilde{y}_{t+k}$ made by submodel i and let $f^{cl}(\tilde{y}_{t+l}|y_{1t},\ldots,y_{kt},\theta,\eta_1,\ldots,\eta_K,\omega_1,\ldots\omega_K) = \prod f(\tilde{y}_{t+l}|y_{it},\theta,\eta_i)^{\omega_i}$. f^{cl} is a geometric pool (with weights ω_i) of K misspecified predictions of $\tilde{y}_{it+l}$, given y_{it}, and the parameters of the submodels. Then

$$\begin{aligned}p(\tilde{y}_{t+l}|y_{1t},\ldots,y_{Kt},\omega_1,\ldots\omega_K) = \int f^{cl}(\tilde{y}_{t+l}|y_{1t},\ldots,y_{Kt},\theta,\eta_1,\ldots,\eta_K,\omega_1,\ldots,\omega_K) \\ p(\theta,\eta_1,\ldots,\eta_K|\omega_1,\ldots,\omega_K,y_{1t},\ldots,y_{Kt})d\theta d\eta_1\ldots d\eta_K\end{aligned} \tag{17}$$

is the kernel of the composite predictive distribution of $\tilde{y}_{t+l}$, given the data and the weights.

When K=2, (17) reads

$$\begin{aligned} p(\tilde{y}_{t+l}|y_{1t}, y_{2t}, \omega, 1-\omega) &= \\ \int f(\tilde{y}_{t+l}|y_{1t}, \theta, \eta_1)^{\omega} f(\tilde{y}_{t+l}|y_{2t}, \theta, \eta_2)^{1-\omega} p(\theta, \eta_1, \eta_2|\omega, 1-\omega, y_{1t}, y_{2t}) d\theta d\eta_1 d\eta_K &= \\ \int [p(\tilde{y}_{t+l}, \theta, \eta_1|y_{1t})]^{\omega} [p(\tilde{y}_{t+l}, \theta, \eta_2|y_{2t})]^{1-\omega} d\theta d\eta_1 d\eta_2 &= \end{aligned} \tag{18}$$

where $p(\tilde{y}_{t+l}, \theta, \eta_i|y_{it})^{\omega_i} = f(\tilde{y}_{t+l}|y_{2i}, \theta, \eta_i)^{\omega_i} p(\theta, \eta_1, \eta_2|\omega, 1-\omega, y_{1t}, y_{2t})^{\omega_i}$. Thus, the composite predictions density can be obtained by taking the joint density of future observations, of the parameters, and of the available data for each model, geometrically weighting them, and integrating the resulting expression with respect to the parameters.

As it is clear from (17), the composite predictive distribution differs from the true predictive distribution in two respects. First, the prediction function is based on the composite prediction pool rather than the true prediction density. Second, the composite prediction pool is integrated with respect to the composite posterior density rather than their true posterior density of the parameters. Similarly, relative to the predictive distribution obtained with each submodel, the composite predictive distribution uses the information in all models for estimation and geometrically weighs their predictions.

To compute (17), we need an estimator for ω_i. One could use the mode or the posterior mean of the marginal of ω_i, depending on the relevant loss function employed. As an alternative, one could integrate the composite predictive distribution with respect to the posterior of ω. Given that in many applications it makes sense to condition the prediction on estimates of the weights (which will represent the posterior probability associated with each model) we believe (17) is more appropriate for standard applications. The prediction problem offers an alternative setup to choose ω_i if they are treated as fixed. In fact, one can setup a Kullback-Leibler (KL) risk measure between the misspecified predictive density and the true predictive density and choose the weights to minimize it (see Vidoni, 2012 and the next subsection).
In terms of computation, in many applications $f(\tilde{y}_{t+l}|y_{2i}, \theta, \eta_i)$ can be computed in a straightforward way (one example are linear Gaussian state space models). Then (17) can be approximated by first generating draws from the composite likelihood-based posterior, computing each predictive density for each parameter draw,[5] multiplying the weighted densities and finally averaging those results across all draws.

4.1 Comparison with other pooling devices

The problem of combining forecast densities is well established in the literature (see e.g. Geweke and Amisano, 2011). Two typical approaches are suggested: linear pooling,

[5] This procedure has to be carried out for each value of $\tilde{y}_{t+l}$ that we want to consider. As long as $\tilde{y}_{t+l}$ is not too high dimensional, this procedure is feasible.

which lead to finite mixtures predictive densities such as BMA or static pools, and logarithmic pooling, which is what a composite predictive density approach produces. Logarithmic pooling gives predictive densities which are generally unimodal and less dispersed than linear pooling and possess the property of being invariant to the arrival of new information (updating the components of the composite likelihood commutes with the pooling operator).

It is also worth noting that, the logarithmic combination formula we present can be obtained as the solution to a well known constrained optimization problem in information theory (see Cover and Thomas, 2006). Here we look for the predictive density solving the following problem

$$\hat{p} = arg\min KL(p(z|y), f(z)) \tag{19}$$

where z is any future sequence of y subject to

$$E_p\{\log \frac{f(z|y_i)}{f(z)}\} = E_{Z|Y=y}\{\ln \frac{f(z|y_i)}{f(z)}\} \quad i = 1, \ldots K \tag{20}$$

and the normalization condition $E_p(1) = 1$, where E_p is the expectation with respect to the density $p(z|y)$ and $f(z)$ is any preliminary density of z, for example its marginal. In words, we seek for the predictive density which is closest in the KL sense to any preliminary marginal density of z and reproduces the same conditional expectation as the true density $f(z|y)$ on $log\frac{f(z|y_i)}{f(z)}$ [6]. The solution to this problem is $\hat{p}(z|y) = f(z)\exp\{\lambda_i \log \frac{f(z|y_i)}{f(z)}\} - Z(y_i, \lambda)\}$ and $Z(y_i, \lambda)$ is a normalizing constant, λ_i are the Lagrange multipliers on the constraints (20). It is easy to recognize that $\hat{p}(z|y)$ has an exponential tilting format: we tilt $f(z)$ in the directions spanned by $\log \frac{f(z|y_i)}{f(z)}$. If $\lambda_i \geq 0, \sum_i \lambda_i \leq 1$, then $\hat{p}(z|y)$ is the composite predictive density previously derived with $\omega_i = \lambda_i,\ i = 1, \ldots, K$ and $\omega_0 = 1 - \sum_i \lambda_i$ where ω_0 is the weight on the marginal of $f(z)$. Finally, note that in this setup, the weights ω_i satisfy the following (score) equations:

$$\frac{\partial E_{z|Y=y} \log f_p(Z|y, \omega)}{\partial \omega_i} = 0, \quad i = 1, \ldots K \tag{21}$$

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

4.2 Computing standard statistics and counterfactuals

In analogy with the prediction problem, one can compute impulse responses and other summary statistics of interest geometrically weighting the outcomes obtained with each of the models and the composite posterior for the parameters. Take for example, impulse responses. Impulse responses to shock j for each model and given values of parameters can be obtained as the difference between a conditional and an unconditional predictive density $f(\tilde{y}_{t+l}|y_{it}, \epsilon_{it}^j = 1, \theta, \eta_i) - f(\tilde{y}_{t+l}|y_{it}, \epsilon_{it}^j = 0, \theta, \eta_i)$ $j = 1, 2, \ldots,$

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

where ϵ_{it} are the shocks of submodel A_i.The (kernel of) the density of the composite likelihood impulse responses can then be computed analogously to (17), with the differences between predictive densities defined above replacing the predictive densities in (17). Again, the composite expression defines a logarithmic pool of impulse responses for each value of the vector $(\omega_1, \ldots, \omega_K)$.

Counterfactuals can be similarly computed. Let $\bar{y}_{kt+l}$ a selected path for future in the k-th element of $\tilde{y}_{t+l}$. Then using $f(\bar{y}_{kt+l}|y_{it}, \epsilon^j_{it+l}, \theta, \eta_i)$ for model A_i, one can find the path of ϵ^j_{it+l} which is consistent with the assumed $\bar{y}_{kt+l}$. With this path one can then compute $f(\bar{y}_{k't+l}|y_{it}, \epsilon^j_{it+l}, \theta, \eta_i)$, for $k' \neq k$. The composite counterfactual can then again be computed analogously to (17).

5 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 and shows how small sample identification problems can be resolved. The next example show how the approach can ameliorate population identification problems; the third example deals with singularity issues; the fourth 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.

5.1 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}), for some subvectors y_{At} and y_{Bt} of the observale variables y_t and may feature common parameters (such as utility function parameters or policy rule coefficients) and model specific parameters. Both models as misspecified in the sense that there is no parameter vector that makes the density of the data $F(y_t)$ equal to the density of the model $(F(y_t) \neq f(y_{At}, \psi_A), F(y_t) \neq f(y_{Bt}, \psi_B))$.

Assume that f_A and f_B are produced by the decision rules:

$$y_{At} = \rho_A y_{At-1} + \sigma_A e_t \tag{22}$$

$$y_{Bt} = \rho_B y_{Bt-1} + \sigma_B u_t \tag{23}$$

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

Thus $\theta = (\rho_A, \sigma_A)$, $\eta_1 = \varnothing, \eta_2 = (\delta, \gamma)$. The (normal) log-likelihood functions associated with each model are:

$$\log L_A \propto -T_A \log \sigma_A - \frac{1}{2\sigma_A^2} \sum_{t=1}^{T_A} (y_{At} - \rho_A y_{At-1})^2 \tag{24}$$

$$\log L_B \propto -T_B \log \sigma_B - \frac{1}{2\sigma_B^2} \sum_{t=1}^{T_B} (y_{Bt} - \rho_B y_{Bt-1})^2 \tag{25}$$

and the composite likelihood is

$$\log C = \omega \log L_A + (1-\omega) \log L_B \tag{26}$$

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

Maximization of (26) with respect to θ leads to:

$$\rho_A = (\sum_{t=1}^{T_A} y_{At-1}^2 + \phi_2 \sum_{t=1}^{T_B} y_{Bt-1}^2)^{-1} (\sum_{t=1}^{T_A} y_{At} y_{At-1} + \phi_1 \sum_{t=1}^{T_B} y_{Bt} y_{Bt-1}) \tag{27}$$

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

$$\sigma_A^2 = \frac{1}{\xi} (\sum_{t=1}^{T_A} (y_{At} - \rho_A y_{At-1})^2 + \frac{1-\omega}{\omega\gamma^2} \sum_{t=1}^{T_B} (y_{Bt} - \delta \rho_A y_{Bt-1})^2) \tag{28}$$

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 model B likelihood are

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

and

$$\sigma_{AA}^2 = \frac{1}{T_A} \sum_{t=1}^{T_A} (y_{At} - \rho_{AA} y_{At-1})^2; \quad \sigma_{AB}^2 = \frac{1}{T_B} \sum_{t=1}^{T_B} (y_{Bt} - \delta \rho_{AB} y_{Bt-1})^2) \tag{30}$$

By comparing, e.g. ((27))-((28)) and ((29))-((30)) one can see that the estimator for θ obtained from a composite likelihood approach combine the information coming from y_{At} and y_{Bt} Furthermore, when estimating θ, model B plays the role of a prior for model A. The formulas in (27) and (28) are also similar to those i) obtained in least square problems with uncertain linear restrictions (Canova, 2007, Ch.10), ii) derived using a prior-likelihood approach, see e.g. Lee and Griffith (1979) or Edwards (1969) and iii) implicitly produced by a DSGE-VAR setup (see Del Negro and Schorfheide, 2004), where T_B are the additional observations added to the original T_A available to estimate θ. Note that the parameters specific to model B $\eta_B = (delta, \gamma)$ are estimated using only the information present in y_{tB}.

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, when estimating common parameters, the composite likelihood gives more importance to data assumed to be generated by a model with higher persistence and lower standard deviation and which is a-priori more likely. The reason is straightforward: higher serial correlation implies important low frequency information; lower standard deviation implies lower noise.

When an array of models are available, estimates of θ will be constrained by the structure present in all models. For example, equation (27) now becomes

$$\rho_A = (\sum_{t=1}^{T_A} y_{At-1}^2 + \sum_{i=1}^{K-1} \zeta_{i2} \sum_{t=1}^{T_i} y_{it-1}^2)^{-1} (\sum_{t=1}^{T_A} y_{At} y_{At-1} + \sum_{i=1}^{K-1} \zeta_{i1} \sum_{t=1}^{T_i} y_{it} y_{it-1}) \tag{31}$$

where $\zeta_{i1} = \frac{\omega_i}{\omega_A} \frac{\delta_i}{\gamma_i^2}$, $\zeta_{i2} = \zeta_{i1}\delta_i$. Thus, the composite likelihood robustifies inference, in the sense that estimates of the common parameters are shrunk to be consistent with the data generated by all available models. 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 may 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 parameters appearing in models featuring different observables. Alternatively, y_{At} and y_{Bt} may represent the same series with different levels of aggregation (say, aggregate vs. individual consumption). In general, y_{At} and y_{Bt} may have common components (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 can be used 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 time periods, for example, consumption and the real rate before and after a financial crisis. Baumeister and Hamilton (2015) suggested a procedure to downweigh the information contained in the first part of a sample that mimics a composite likelihood estimator in this situation.

The setup of this example also allows us to discuss how a composite likelihood approach may help to reduce small sample identification problems. Suppose the density of model A is well behaved, but because T_A is short, the likelihood we can construct from it is flat. Thus, in part of the domain and θ may be poorly identified using y_{At}. It is easy to show θ could become better identified if (y_{At}, y_{Bt}) are jointly used in estimation. This is because the curvature of the composite likelihood depends on the effective sample size is ξ which is a function ofT_A and $T_B \frac{1-\omega}{\omega\gamma^2}$. Thus, for example, if γ 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 around the mode than the likelihood constructed using only y_{At}.

5.2 Ameliorating population identification problems

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

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

$$
\begin{aligned}
R_{At} &= \tau E_t \pi_{At+1} + e_{1t} && (32)\\
y_{At} &= \delta E_t y_{At+1} - \sigma(R_{At} - E_t \pi_{At+1}) + e_{2t} && (33)\\
\pi_{At} &= \beta E_t \pi_{At+1} + \gamma y_{At} + e_{3t} && (34)
\end{aligned}
$$

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

$$
\begin{bmatrix} R_{At} \\ y_{At} \\ \pi_{At} \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 \\ \sigma & 1 & 0 \\ \sigma\gamma & \sigma & 1 \end{bmatrix} \begin{bmatrix} e_{1t} \\ e_{2t} \\ e_{3t} \end{bmatrix} \equiv A e_t \qquad (35)
$$

Clearly, β is underidentified (it disappears from the solution) and 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 (35) 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 with exogenous marginal costs:

$$
\begin{aligned}
\pi_{Bt} &= \beta E_t \pi_{Bt+1} + \gamma y_{Bt} + u_{2t} && (36)\\
y_{Bt} &= \rho y_{Bt-1} + u_{1t} && (37)
\end{aligned}
$$

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

$$\pi_{Bt} = \frac{\gamma}{1-\beta\rho} y_{Bt} + u_{2t} \tag{38}$$

$$y_{Bt} = \rho y_{Bt-1} + u_{1t} \tag{39}$$

We can rewrite (38) and (39) in terms of $x_t \equiv (1-\rho\ell)y_{Bt}$, $w_t \equiv (1-\rho\ell)\pi_{Bt}$ as

$$\left[\begin{array}{c} x_t \\ w_t \end{array}\right] = \left[\begin{array}{cc} 1 & 0 \\ \frac{\gamma}{1-\beta\rho} & 1-\rho\ell \end{array}\right] \left[\begin{array}{c} u_{1t} \\ u_{2t} \end{array}\right] \tag{40}$$

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) it 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 (40) and thus in the composite likelihood, even when $1-\omega$ is small. In this particular example, shrinking model A toward model B which is more misspecified but has sharper information about the parameters of interest may be beneficial in terms of identification and estimation. Note that all the arguments made here 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}).

5.3 Solving singularity problems

DSGE models are typically singular. That is, since they generally feature more endogenous variables than shocks, the theoretical covariance matrix of the observables is singular and the likelihood function of the model can not be constructed and optimized. There are two types of responses to this problem in the literature : 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 use the the log-likelihood of this subvector for estimation; 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, 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, we use a stylized asset pricing example.

Suppose that the dividend process is $d_t = e_t - \zeta e_{t-1}$, 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) unless we add a measurement error, which is difficult to justify since neither dividends nor stock prices are subject to revisions, or make βa random variable [7]. When the composite likelihood is employed, information present in both series can be used to identify and estimate (ζ, σ^2) and β, if it is of interest. The optimization process makes dividends and stock prices contain different types of information; the composite likelihood combines them 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

$$\log L(\tilde{d}_t|\zeta, \sigma^2) = -0.5T\log(2\pi) - 0.5\sum_{t=1}^{T}\log\varsigma_t - 0.5\sum_{t=1}^{T}\frac{\tilde{d}_t^2}{\varsigma_t} \tag{41}$$

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

$$\tilde{d}_t \quad = \quad d_t - \zeta\frac{1+\zeta^2+\zeta^4+\ldots+\zeta^{2(t-2)}}{1+\zeta^2+\zeta^4+\ldots+\zeta^{2(t-1)}}\tilde{d}_{t-1} \tag{42}$$

$$\varsigma_t \quad = \quad \sigma^2\frac{1+\zeta^2+\zeta^4+\ldots+\zeta^{2t}}{1+\zeta^2+\zeta^4+\ldots+\zeta^{2(t-1)}} \tag{43}$$

and

$$\log L(\tilde{p}_t|\beta, \zeta, \sigma^2) = -0.5T\log(2\pi) - 0.5\sum_{t=1}^{T}\log\lambda_t - 0.5\sum_{t=1}^{T}\frac{\tilde{p}_t^2}{\lambda_t} \tag{44}$$

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

$$\tilde{p}_t \quad = \quad p_t - \frac{\gamma^2}{\zeta}\frac{1+\gamma^2+\gamma^4+\ldots+\gamma^{2(t-2)}}{1+\gamma^2+\gamma^4+\ldots+\gamma^{2(t-1)}}\tilde{p}_{t-1} \tag{45}$$

$$\lambda_t \quad = \quad \sigma^2(1-\beta\zeta)^2\frac{1+\gamma^2+\gamma^4+\ldots+\gamma^{2t}}{1+\gamma^2+\gamma^4+\ldots+\gamma^{2(t-1)}} \tag{46}$$

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

$$\frac{\partial\log L(\tilde{d}_t)}{\partial\zeta} \quad = \quad -0.5\sum_t\frac{\partial\log\varsigma_t}{\partial\zeta} - 0.5\sum_t\frac{\partial\log\tilde{d}_t}{\partial\zeta}\frac{1}{\varsigma_t} + 0.5\sum_t\frac{\partial\log\varsigma_t}{\partial\zeta}\frac{\tilde{d}_t}{\varsigma_t^2} \tag{47}$$

$$\frac{\partial\log L(\tilde{p}_t)}{\partial\gamma} \quad = \quad -0.5\sum_t\frac{\partial\log\lambda_t}{\partial\zeta} - 0.5\sum_t\frac{\partial\log\tilde{p}_t}{\partial\zeta}\frac{1}{\lambda_t} + 0.5\sum_t\frac{\partial\log\lambda_t}{\partial\zeta}\frac{\tilde{p}_t}{\lambda_t^2} \tag{48}$$

[7] Clearly, if the dividend process is used, β is underidentified.

For a given ω assigned to $\tilde{d}_t$, the composite likelihood is a weighted sum of (47) and (48). 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 (47) and (48) 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 (48) and (47) are generally different 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 ofω, either the serial correlation or the variance properties of (d_t, p_t) or both will be employed. Clearly, if the low frequency components of $\tilde{p}_t$ are poorly characterized because, for example, the sample is short or because 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 is likely to give a more stable estimate of ζ than the individual likelihood functions.

5.4 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 and 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 N(0,$\Sigma(\theta)$), θ is a vector of structural parameters, y_t is of large dimension and, generally, dim(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 A(θ) corresponding to $\tilde{y}_t$ - we assume that for a given vector of structural parameters the solution for $\tilde{y}_t$ still follows a VAR of order 1. The likelihood function is

$$L(\tilde{y}|A(\theta), \Sigma(\theta)) = (2\pi)^{-T/2}|\Sigma|^{T/2}\exp\{(\tilde{y}_t - \widetilde{A(\theta)}\tilde{y}_{t-1})\Sigma(\theta)^{-1}(\tilde{y}_t - \widetilde{A(\theta)}\tilde{y}_{t-1})'\} \quad (49)$$

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,

$$L(\tilde{y}_1|A, \Sigma) = \int L(\tilde{y}_1|\tilde{y}_2, \widetilde{A(\theta)}, \Sigma(\theta))g(\tilde{y}_2)d\tilde{y}_2 \quad (50)$$

which may be intractable.

Rather than computing the likelihood for $\tilde{y}_{1t}$, we can take a limited information point of view 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 are:

$$CL_1(\hat{y}_t | A_i(\theta), \Sigma(\theta)) = \sum_{i=1}^{K} \omega_i \log L(\hat{y}_{it} | A_i(\theta), \Sigma(\theta)) \tag{51}$$

$$CL_2(\hat{y}_{it}, A_i(\theta), \Sigma(\theta)) = \sum_{i=1}^{K} \omega_i \log L(\hat{y}_{it} | \hat{y}_{-it}, A_i(\theta), \Sigma(\theta)) \tag{52}$$

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 the variables of all other countries $\hat{y}_{-it}$. Which composite likelihood one would use depends on the problem and the tractability of conditional vs. marginal likelihoods.

5.5 Dealing with short T when a panel is available.

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

$$CL(\hat{y}_{1t}, \hat{y}_{2t}, ...\hat{y}_{Kt} | A(\theta), \Sigma(\theta)) = \sum_{i=1}^{K} \omega_i \log L(\hat{y}_{it} | A(\theta), \Sigma(\theta)) \tag{53}$$

In this case (27) 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 combine individual and a (weighted) average of the information present in the cross section of data. This is clear when looking at (31), 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). Note that 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, stochastically exploits commonalities in cross section, if they exist, and may lead to improved estimates of the vector of common parameters θ.

6 Reducing misspecification

The shrinkage estimators that the composite likelihood generates may 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 (misspecified) models restrict the range of values that the common parameters can take. 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 the relationship between estimates of σ^2, the variance of the estimated error term, which is common across models, and 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 information in 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 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 four 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 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,

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 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 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,

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 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 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 parametrization dependent.

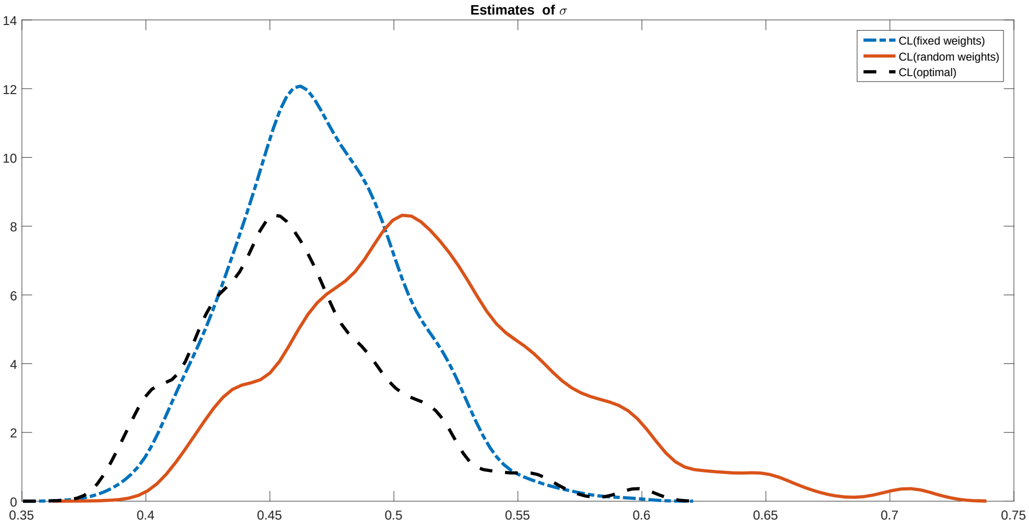

Figure 1: Composite posterior density and two posterior mixtures

7 An example

Schorfheide (2008) surveyed estimates of the slope of the Phillips curve present in the literature. He documented large cross sectional variations and found that differences seem to depend on i) the specification of the model used for estimation, ii) the observability of marginal costs, iii) the number and the type of variables used in estimation. Given this background, and given that policymakers are typically interested in the posterior distribution of this parameter for forecasting and counterfactual exercises, we want to examine how the composite posterior distribution of the slope of the Phillips curve obtained with a variety of models featuring different numbers of observables looks like relative to i) the posterior distribution obtained with each single model; ii) the posterior distribution obtained ex-post averaging the posterior distributions of each single model.

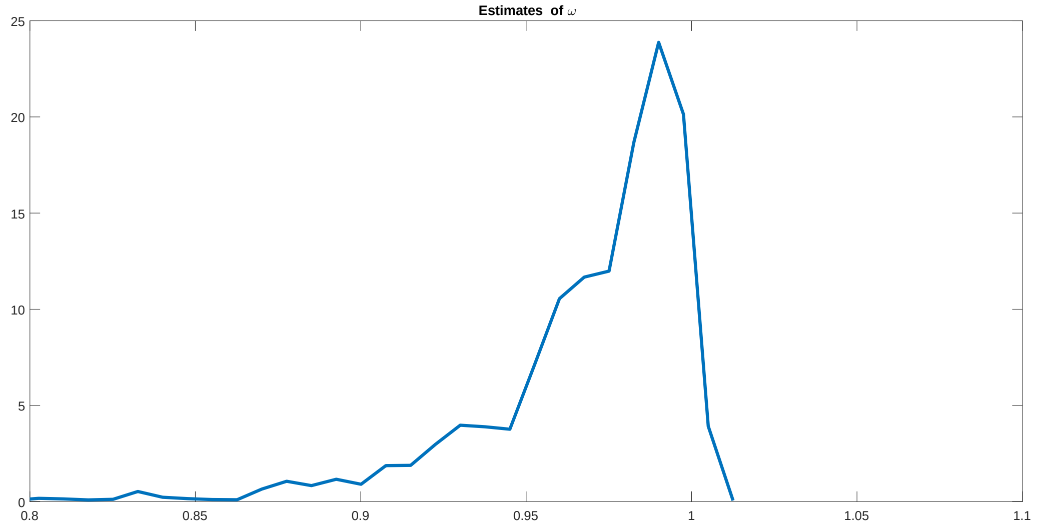

Figure 2: Posterior density of ω

We consider five models in the exercise: 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, detrended consumption, detrendend investment, demeaned π, demeaned R, detrended hours, and detrended nominal wage W; a New Keynesian model with search and matching labour market frictions, as in Christoffel and Kuester (2008) where the variables used in estimation are detrended Y, demeaned π, demeaned R and detrended real wage w; and the Bernanke, Gertler, and Gilchrist (1999) model, estimated with detrended output Y, demeaned inflation π, and demeaned nominal rate R [8]. For all models, a quadratic trend is used to detrend the data, and the series used are from Smets and Wouter (2007). Furthermore, for comparability, the estimation sample is 1960:1-2005:4 for all models. The priors for the structural parameters of various models are standard. To make sure that all models approximately reproduce the cross correlation function of output, inflation and the nominal rate, the priors for the nuisance parameters are

[8] The parameters governing financial frictions are mainly calibrated, along the lines of Cogley et al (2011). This helps us to avoid the tedious choice of exactly what data series to compare to the model-implied spread series.

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

	5%	50%	95%
Prior	0.01	0.80	1.40
Basic NK	0.06	0.18	0.49
Basic NK with nominal wages	0.05	0.06	0.07
SW with capital and adj.costs	0.04	0.05	0.07
Search	0.44	0.62	0.86
BGG	0.0175	0.02	0.024
CL	0.15	0.19	0.25

Reported are relevant percentiles of the posterior of the slope of the Phillips curve 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 7 observables (SW with capital and adj. costs), the four variable search and matching model (Search) and the three variable financial friction model (BGG). The row with CL reports composite posterior percentiles using equal weights on lal models. Estimation sample is 1960:1-2005:4.

endogenously selected as in Del Negro and Schorfheide (2008).

In the baseline exercise we report, we fix $\omega = (0.2, 0.2, 0.2, 0.2, 0.2)$; choosing a higher weight ($\omega_3 = 2/6, \omega_i = 1/6$, i =1,2,4,5) or a lower weight ($\omega_3 = 1/9, \omega_i = 2/9$, i =1,2,4,5) on the larger model produces a composite posterior with similar features.

Table 2 displays some percentiles of the posterior distribution for the slope of the Phillips curve (κ_p) obtained with each of the five models and when the composite likelihood is used to aggregate their information. The posteriors for the first three models have medians close to zero and, as Schorfheide suggested, having non-observables marginal costs tend to increase the location of the mode. The median value of the posterior obtained in financial friction model is also low and very precisely estimated. The search and matching model instead has a much higher median and the posterior does not overlap with the posteriors obtained with the other four models. Again, in agreement with Schorfheide, estimation results depend on the model used and the observable variables used in estimation. Notice that, in many cases, the spread of the posterior is relatively large and in a few cases the overlap in the posterior distributions is also large.

The composite posterior is centered around 0.15, and has a credible 90 percentile smaller than most individual models. Thus, sharper inference about the effect of, say, labour market or product market reforms can be made using composite posterior estimates in most of the models we consider. More importantly, inference is likely to be more robust because estimates are obtained using information from a number of non-necessarily nested models.

Figure 3 presents the composite posterior for the slope of the Phillips curve together with two alternative naive posterior combination estimators: one that equally weights the posteriors of κ_p obtained with the five models; and one which weights the

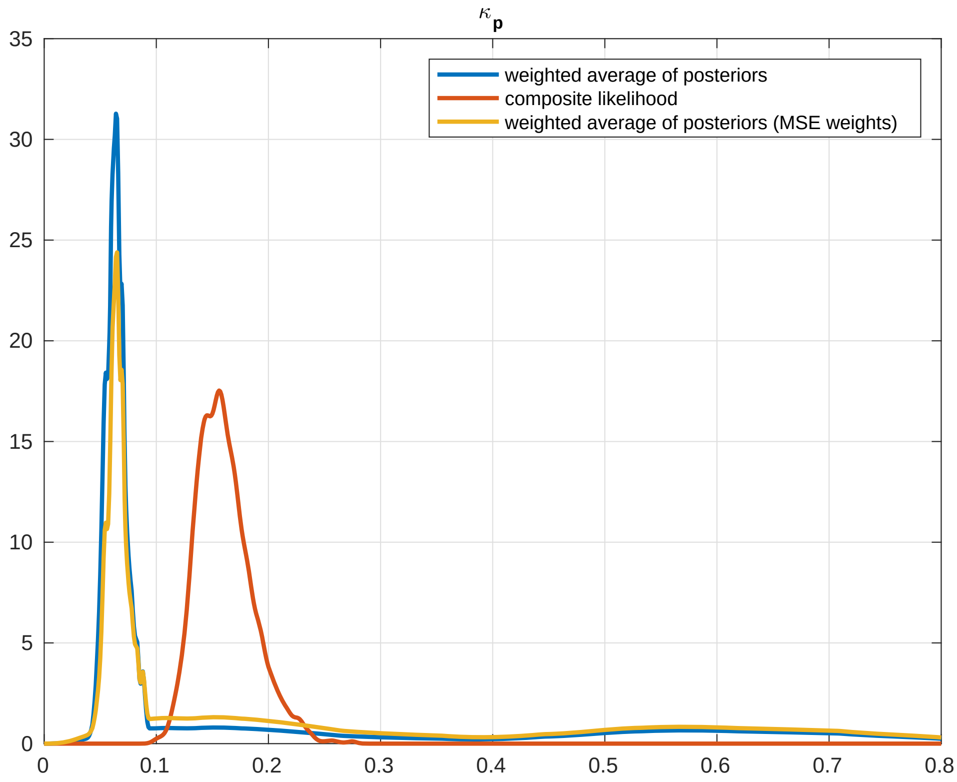

Figure 3: Composite posterior and two naive posterior mixtures

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, combining ex-post estimates of κ_p generate distributions which are different from those obtained with a composite posterior approach; the location of the distributions is different and the spread much larger when naive posterio combinations are computed. Second, it matters the way naive combinations estimators are constructed. For example, the combination posterior produced using MSE weights has two modes: one of which is indistinguishable from the mode obtained with equal weights, and one which is considerably lower.

8 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 implement, 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 .

9 References

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

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

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

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

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

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

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

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

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

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

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

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.

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.

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

Appendix A

Let y_t be a sample from the density $f(y_t)$ with respect to some σ-measure μ. Suppose a model with the density $g(y_t,\lambda),\ \lambda \in \Lambda \subset R^m$ is use and the log-likelihood is $L_g(\lambda) + \sum_t \log g(y_t,\lambda)$. the model is misspecified because $f(y_t) \neq g(y_t,\lambda),\ \forall\lambda$. Let $\hat{\lambda}$ be the maxmimum likelihood estimator i.e. $\hat{\lambda} = sup_\lambda L_g(\lambda)$. Since $T^{-1}L_g(\lambda) \to E(\ln g(y_t,\lambda))$ by a uniform law of large numbers, $\hat{\lambda}$ will be consistent for $\lambda_0 = arg\max_\lambda E\ln g(y_t,\lambda)$, where the expectations are taken with respect to the density f.

If f is absolutely continuous with respect to g then

$$E\ln g(y_t,\lambda) - Ef(y_t) = -\int f(y_t)\ln\frac{f(y_t)}{g(y_t,\lambda)}d\mu(y) - KL(\lambda) \tag{54}$$

where $KL(\lambda)$ is the Kullback-Liebler divergence between f and g. Hence λ_0 is also the minimizer of $KL(\lambda)$.

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

$$0 \approx T^{-0.5}\sum_t s_t(\theta_0) + T^{0.5}\Sigma_1^{-1}(\lambda_{ML} - \lambda_0) \tag{55}$$

where $\Sigma_1 = -E(h_t(\theta_0)) = \frac{\partial^2 KL(\lambda)}{\partial\lambda\partial\lambda^T}|_{\lambda=\lambda_0}$. Then using a central limit theorem for correlated observations we have that $T^{-0.5}(\theta_{ML} - \theta_0) \sim N(0,V)$ where $V = \Sigma_1\Sigma_2\Sigma_1$ and $\Sigma_2 = E(s_t(\theta)s_t(\theta)')$.

Note that in standard DSGE applications $s_t(\lambda)$ are typically computed via the Kalman filter and are function of martingale difference processes(the shocks of the model). Thus the condition $\sum_t s_t(\lambda) = 0$ is likely to hold.

Regularity conditions which need to be imposed for this to hold are in Mueller (2013).

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

Appendix B

NEED ASSUMPTIONS ON $p(\theta)$

Let θ_{CL} be the maximum composite likelihood estimator and let θ_p be the mode of the prior distribution for $p(\theta)$.Let $h(\theta_{CL}) = -\nabla^2_\theta CL(\theta|y)$ and $h(\theta_p) = -\nabla^2_\theta p(\theta)$. For

$T \rightarrow \infty$

$$\begin{aligned} p_{CL}(\theta|Y) &\propto \{CL(\theta_{CL}|y) - 0.5(\theta) - \theta_{CL})^T h_(\theta_{CL})(\theta - \theta_{CL}) + logp(\theta_p) - 0.5(\theta) - \theta_p)^T h_(\theta_p)(\theta - \theta_p)\} \\ &\approx N(\hat{\theta}, h(\theta_{CL}, \theta_p)) \end{aligned} \tag{56}$$

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

Given the assumptions made on $p(\theta)$ it will vanish at $T \rightarrow \infty$. Then, almost surely, the strong law of large number implies that

$$T^{-1}h(\theta_{CL}, \theta_p) \rightarrow -E(\nabla^2 CL(\theta_0|Y)) \equiv H(\theta_0) \tag{57}$$

$$th\hat{e}ta = (T^{-1}h(\theta_{CL}, \theta_p))^{-1}(T^{-1}h(\theta_{CL})\theta_{CL} + T^{-1}h(\theta_p)\theta_p) \rightarrow \theta_0 \tag{58}$$

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

Appendix C

Consider observations t=1 and t=2. From (42) and (45) we have

$$\tilde{d}_1 = d_1 \tag{59}$$

$$\tilde{d}_2 = d_2 - \frac{\zeta}{1+\zeta^2}d_1 \tag{60}$$

$$\tilde{p}_1 = p_1 \tag{61}$$

$$\tilde{p}_2 = p_2 - \frac{\zeta}{(1-\beta\zeta)^2+\zeta^2}p_1 \tag{62}$$

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 (43) and (46), for t=1 and t=2 we have

$$\varsigma_1 = \sigma^2(1+\zeta^2) \tag{63}$$

$$\varsigma_2 = \sigma^2\frac{1+\zeta^2+\zeta^4}{1+\zeta^2} \tag{64}$$

$$\lambda_1 = \sigma^2((1-\beta\zeta)^2+\zeta^2) \tag{65}$$

$$\lambda_2 = \sigma^2(1-\beta\zeta)^2\frac{1+\frac{\zeta^2}{(1-\beta\zeta)^2}+\frac{\zeta^4}{(1-\beta\zeta)^4}}{1+\frac{\zeta^2}{(1-\beta\zeta)^2}} \tag{66}$$

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.