Dealing with misspecification in structural macroeconometric models

Fabio Canova,
Norwegian Business School *

Christian Matthes,
Indiana University †

October 19, 2020

Abstract

We consider a set of potentially misspecified structural models, geometrically combine their likelihood functions, and estimate the parameters using composite methods. In a Monte Carlo study, composite estimators dominate likelihood-based estimators in mean squared error and composite models are superior to individual models in the Kullback-Leibler sense. We describe Bayesian quasi-posterior computations and compare our approach to Bayesian model averaging, finite mixture, and robust control procedures. We robustify inference using the composite posterior distribution of the parameters and the pool of models. We provide estimates of the marginal propensity to consume and evaluate the role of technology shocks for output fluctuations.

JEL Classification numbers: C13, C51, E17.

Keywords: Model misspecification, composite likelihood, Bayesian model averaging, finite mixture.

*Norwegian Business School, CAMP, and CEPR. Department of Economics, BI Norwegian Business School, Nydalsveien 37, 0484 Oslo, Norway; email: fabio.canova@bi.no

†Indiana University. Department of Economics, 100 S Woodlawn Avenue, Bloomington, IN 47405, USA. email: matthesc@iu.edu. We thank two anonymous referees, T. Zha (the editor in charge) M. Plagborg-Moller, G. Koop, B. Rossi, F. Schorfheide, J. Linde, the participants of the 2017 Minnesota alumni lecture; and of the conferences: Time varying uncertainty in macroeconomics, St. Andrews; the 16th EC^2, Amsterdam; Nonlinear models in Macroeconomics and Finance for an unstable world, Oslo; SBIES, St. Louis; GSE Summer Forum (Time Series), Barcelona; Time-Varying Parameter Models, Florence; St. Louis Fed Time Series Workshop, St. Louis; NBER Summer Institute 2018, Boston and of seminars at the Bank of Finland, University of Glasgow, University of Melbourne, University of Amsterdam for comments and suggestions. Canova acknowledges the financial support from the Spanish Ministerio de Economia y Competitividad through the grants ECO2015-68136-P and FEDER, UE.

1 Introduction

Over the last 20 years dynamic stochastic general equilibrium (DSGE) models have become more detailed and complex, and numerous features have been added to the original real business cycle core. Still, even the best practice DSGE model is likely to be misspecified either because features such as heterogeneities in expectations are missing, or because researchers leave out aspects deemed tangential to the analysis. While specifying an incomplete model is acceptable, for example, when qualitatively highlighting a mechanism which could be present in the data, misspecification becomes an issue when one wants to quantify the importance of certain shocks or estimate the magnitude of crucial policy trade-offs.

In theory, misspecification can be reduced by making structural models more comprehensive in their description of the economic relationships and of the interactions among agents. In practice, this is difficult because it is not clear which missing feature is relevant and jointly including several of them quickly makes the computations intractable and the interpretation difficult. Moreover, large scale models are hard to estimate with limited data and parameter identification problems are likely to be important (see e.g. Canova and Sala, 2009). The standard short cut to deal with misspecification is to use a structural model with ad-hoc reduced form features. However, in hybrid models it is often hard to distinguish the relative importance of structural vs. ad-hoc features in matching the data, making policy counterfactuals whimsical.

Structural vector autoregressive (VAR) models or limited information moment-based estimation approaches can deal with model incompleteness or partially specified dynamic relationships, when, e.g., characterizing the dynamics in response to shocks (see e.g. Kim, 2002; or Cogley and Sbordone, 2010). Full information likelihood-based methods, however, have a hard time dealing with misspecification other than that of the distribution of the error term, and are justified asymptotically only under the assumption that the estimated model correctly characterizes the data generating process (DGP) up to a set of serially and cross sectionally uncorrelated disturbances. To avoid this problem the recent econometric literature dealing with misspecification does not employ the likelihood in the estimation process (see e.g. Cheng and Liao, 2015; Thryphonides, 2016) and robustness approaches modify posterior inference to reduce the chance of incorrect decisions (see Hansen and Sargent, 2008; Giacomini and Kitigawa, 2017). The tension between theoretical developments and empirical practice becomes clear when one notices that the vast majority of the applied literature employs full information likelihood-based (classical or Bayesian) procedures to estimate structural parameters and policy prescriptions are often formulated on the basis of potentially misspecified models.

This paper proposes a new approach to reduce the inherent misspecification of DSGE models. Rather than enriching a particular model with structural or ad-hoc features, as it is common in the

literature, we jointly consider a finite set of potentially misspecified models, geometrically combine their likelihood functions, and estimate the parameters using the composite likelihood. With such an objective function, parameters common across models are estimated using the cross equation restrictions present in all specifications; model specific parameters are instead estimated using the cross-equation restrictions appearing only in that specification. When no parameter can be safely assumed to be common across models, composite and likelihood estimators coincide. Thus, the composite likelihood guards against misspecification by requiring estimates of the common parameters to be consistent with the structure of all models.

Although the composite likelihood approach is well established in the statistical literature (see e.g. Varin, et al. 2011), economic applications are limited to Engle et al. (2008), Qu (2018), and Chan et al. (2018). Nevertheless, in all the literature we are aware of the DGP is known; the composite likelihood combines marginals or conditionals of the DGP; and the composite weights are fixed. In our setup, instead, the DGP is unknown; the models entering the composite likelihood are assumed to be misspecified; and the composite weights are random variables. Whereas this paper focuses on misspecification, Canova and Matthes (2019) use the methodology to address a number of inferential and computational problems in structural estimation.

The Bayesian setup we work with is grounded in the Bayesian literature on misspecified models (see Walker, 2012, Bissiri et al. 2016) and related to the quasi-Bayesian estimation literature (see e.g. Kim, 2002, Marin et al. 2012, Scalone, 2018), to Bayesian shrinkage (see e.g. Del Negro and Schorfheide, 2004; Batthacharya et al. 2012) and to smoothness priors (see e.g. Barnichon and Brownlees, 2016). As in quasi-Bayesian approaches, we substitute the likelihood function with an alternative loss function and perform Bayesian inference with the resulting quasi-posterior; and as in the shrinkage and smoothness prior literature, we employ additional information to regularize parameter estimates. The posterior weight of a model plays a role in the inferential process, as in the Bayesian model averaging (BMA) literature (see Claeskens and Hjort, 2008). We differ in three aspects: BMA can be employed only when models share the same observables while our approach works even when models feature different observables. In BMA, each model is estimated separately and posterior weights are used to combine their predictions. Here estimates of the common parameters are jointly obtained and posterior weights can be used to combine models' predictions, if that is of interest. Our setup quantifies the uncertainty in the weight estimates. To the best of our knowledge, this can not be done in BMA exercises.

Our approach shares similarities with the methods of Del Negro and Schorfheide (2004) and Waggoner and Zha (2012), but three important differences need to be emphasized. We consider combinations of structural models; they combine a structural and a VAR model. Waggoner and Zha assume that the DGP is the mixture of the models; we leave open the possibility that the

composite model is still misspecified. Finally, while our approach allows for models with different observables, in the other approaches the models must share the same observable variables.

We describe a Monte Carlo Markov Chain (MCMC) approach to draw sequences from the quasi-posterior distribution of the parameters, show how to adjust the percentiles to ensure the right asymptotic coverage, discuss the computational costs of the approach, and explain how posterior weights inform us about the relative misspecification of the models entering the composite pool. We also show how to combine models and composite estimates for inference. While some researcher may use the posterior weights to select a model to conduct inference, we prefer to robustify the analysis using composite predictions.

Using a simple Monte Carlo design, we demonstrate that composite estimators are preferable to likelihood-based estimators when misspecification is present in a mean-squared error (MSE) sense, that composite models are closer to the true DGP in a Kullback-Leibner (KL) sense, and that BMA and the posterior mode of the weights provide similar information when the models share the same observables.

We apply the methodology to estimate the marginal propensity to consume (MPC) out of transitory income and to evaluate the role of technology shocks for output fluctuations. The MPC is generally low when models are separately estimated because transitory income has insufficient persistence, except when one allows for precautionary savings. When a composite estimate of the persistence parameter is used, the MPC generally increases. We show that problematic features of the basic specification such as quadratic preferences, separability of durable and non-durable consumption, exogenous real rate, lack of production, and consumers homogeneity are irrelevant for the estimation of the MPC and that composite and BMA estimates of the MPC are similar. We also show that a standard ad-hoc model is inferior to the composite model in a KL sense.

Consistent with the existing literature, we find that technology shocks account for about one-third of output fluctuations 20-30 quarters ahead in a standard medium scale New Keynesian model. We pair the model with a smaller scale New Keynesian model without capital, and jointly estimate the slope of the Phillips curve and the persistence of technology shocks. We find that the share of output fluctuations explained by technology shocks substantially increases because the smaller model receives a high posterior weight and forces estimation to move to a region of the parameter space where nominal rigidities are smaller, real rigidities are larger, and demand shocks are less autocorrelated, all of which make technology shocks more important.

The paper is organized as follows. The next section presents the problems one faces when a misspecified model is used for economic analyses and describes the approaches used to make the estimation results more credible. Section 3 presents our method. Section 4 describes a MCMC procedure to draw sequences from the quasi-posterior of the parameters and the weights; and

explains how to construct impulse responses, counterfactuals, and predictions using the pool of models. Section 5 applies the composite approach to two problems. Section 6 concludes. A number of on-line appendices contain relevant technical material.

2 Estimating misspecified structural models

Suppose a researcher is interested in measuring the MPC out of transitory income. Interest in the MPC may arise because the fiscal authority is planning to boost aggregate demand via a temporary tax cut, or because a researcher wants to design optimal policies to enhance aggregate savings and investments. Typically, one solves an off-the-shelf permanent-income, life-cycle model, and derives implications for the MPC. For example, in a representative agent model with quadratic preferences, constant real rate, when $\beta(1+r)=1$, and the exogenous labor income has permanent and transitory components, the decision rules are (see Inoue et al. 2017):

$$
\begin{aligned}
c_t &= \frac{r}{r+1}a_t + (y_t^P + \frac{r}{1-\rho+r}y_t^T) && (1)\\
a_{t+1} &= (1+r)(a_t + (y_t^T + y_t^P) - c_t) && (2)\\
y_t^T &= \rho y_{t-1}^T + e_{1t} && (3)\\
y_t^P &= y_{t-1}^P + e_{2t} && (4)
\end{aligned}
$$

where y_t^T is real transitory income, y_t^P is real permanent income, c_t is real non-durable consumption, a_t are real asset holdings, all in per-capita terms, $e_{it} \sim iidN(0,\sigma_i^2),\ i=1,2$, r is the constant real rate of interest, and ρ the persistence of the transitory income process.

(1)-(4) provide three important restrictions on the data. First, r and ρ are the only deep parameters mattering for the MPC; preference parameters are not identifiable from the decision rules. Second, the relationship between consumption and income is static. Third, the MPC out of transitory income, $MPC_{y^T} = \frac{r}{1-\rho+r}$, is intermediate between the MPC out of asset holdings, $MPC_a = \frac{r}{r+1}$, and the MPC out of permanent income, $MPC_{Y^P} = 1$.

Given this model, one could estimate MPC_{y^T} in a number of ways. If some unexpected temporary tax cut occurred in the past and individual consumer data is available, one can use this natural experiment to see how much of the transitory income the tax rebate has generated is spent. For example, in the US, Johnson et al. (2006) find that after the 2001 tax rebate, agents spent about 20-40 percent of the additional income in first quarter and about 60 percent of the cumulative income over two quarters. Parker et al. (2013) report that after the 2008 tax rebate, agents spent about 20 percent of the additional income on non-durable consumption goods and 30-40 percent on durable consumption goods.

Natural experiments are effective tools to understand how agents behave. However, they are not often available and, even if they were, individual consumer data is hard to get. One approach

to estimate MPC_y^T that uses theory as a guideline for the investigation but does not condition on the restrictions it provides in estimation, is to identify a permanent and a transitory shock in a VAR with (y_t, a_t, c_t) and then measure the effects on consumption of a transitory income shock, scaling the measurement by the income responses. Estimates obtained this way vary between 0.4 and 0.6, depending on the model specification and the sample employed.

To derive estimates of MPC_y^T, one could also partially condition on the restrictions of the model. For example, one could use moment conditions to estimate r and ρ. Since in industrialized countries the average real rate is about 1% per quarter and the persistence of the growth rate of aggregate income is around 0.5-0.7, MPC estimates obtained this way are in the range (0.05-0.10). Clearly, refinements are possible. One could group data according to characteristics of consumer i and report a (weighted) average of the resulting $MPC_{y_i^T}$. Estimates constructed this way are also low and in the range (0.10-0.15), see e.g. Carroll et al. (2017).

A final approach would be to take the implications of the model seriously, write down the likelihood function for (c_t, a_t, y_t) and impose the cross equation restrictions the decision rules imply (in particular, the fact that r and ρ appear in different equations) to estimate MPC_{y^T}. The evidence we present in section 5.1 suggests that likelihood-based estimates of MPC_{y^T} are in the range of 0.10-0.15 for the first quarter and 0.2-0.25 for the first year, roughly the same as when moments conditions are used.

In sum, MPC_{y^T} estimates obtained conditioning on the model's implications tend to be lower than estimates obtained otherwise. One reason for the difference is that the model employed in formal estimation is likely to be misspecified: the real rate is not constant; labor income is not exogenous; preferences may feature non-separable consumption-labor supply decisions. Moreover, the model leaves out aspects that could matter for understanding consumption decisions: income uncertainty does not play any role; home production and goods durability are disregarded; agents are homogeneous but, in the real world, some have zero assets; and others may be rich, but liquidity constrained. Finally, measurement errors in the real value of assets are probably important.

While moment-based and VAR-based estimates are robust to some form of misspecification (e.g. lack of dynamics in the decision rules) and to the omission of certain features from the model, likelihood-based estimates are not. Thus, if misspecification is suspected, estimates obtained relaxing the restrictions the model imposes may be preferable. However, if a researcher insists on using likelihood methods, how can she guard herself against misspecification?

An obvious way is to estimate a more complex model which includes potentially missing features, allows for general equilibrium effects on income and the real rate, uses flexible functional forms for preferences and technologies, and permits relevant heterogeneities. While feasible, it is generally computationally demanding to estimate large scale models, identification issues linger in

the background, and it is often difficult to interpret the dynamics one obtains. Alternatively, one could enrich the model with ad-hoc features. For example, it is nowadays popular to use models with external habit in consumption, even if the micro foundations of such a mechanism are still debatable (one exception is Ravn et al., 2006). With habit, the decision rules of our workhorse model are (see Alessie and Lusardi, 1997):

$$c_t = \frac{h}{1+h}c_{t-1} + (1 - \frac{h}{1+h})w_t \tag{5}$$

$$w_t = \frac{r}{1+r}((1+r)a_{t-1} + \sum_{t=\tau}^{\infty}(1+r)^{t-\tau}E_t(y_\tau^P + y_\tau^T) \tag{6}$$

$$y_t^T = \rho y_{t-1}^T + e_{1t} \tag{7}$$

$$y_t^P = y_{t-1}^P + e_{2t} \tag{8}$$

where h is the habit parameter. Thus, habit helps to account for serial correlation in consumption and for the predictability of current consumption, given permanent wealth w_t; it also makes the serial correlation properties of consumption and income disconnected. Adding ad-hoc features is convenient but makes the model less structurally interpretable and may produce overfitting. In addition, some ad-hoc additions may not lead to better models. For example, adding a preference shock (to capture demand driven changes) to the baseline model would not alter MPC_{y^T}.

Adding these types of features may not be appealing to certain researchers. For this reason, a portion of the literature has instead preferred to alter the statistical properties of shocks, making the stochastic processes more flexible (see e.g. Del Negro and Schorfheide, 2009; Smets and Wouters, 2007) or allowing cross-shock correlation (Curdia and Reis, 2010).

A final approach has been to complete the probability space of the model by adding measurement errors to the decision rules (Ireland, 2004), wedges to optimality conditions (Chari et al., 2007), margins to preferences and technologies (Inoue et al., 2017), or agnostic structural shocks to the decision rules (Den Haan and Drechsel, 2018). Rather than tinkering with the inputs or the specification of the model, all these approaches take the structure as given and add non-structural features for estimation purposes only. Typically, the relevance of the add-ons is measured by the marginal likelihood. Kocherlakota (2007) has examples where using fit to select a model among potentially misspecified candidates may lead researchers astray.

While all these approaches acknowledge model misspecification and may be useful in specific situations of interest, they have at least three drawbacks. First, they condition on one model but there are many potential models a researcher could entertain - specifications could be indexed, e.g., by the economic frictions models impose. Second, they neglect the fact that different models may be more or less misspecified in different periods (see e.g. Del Negro et al., 2016). Third, the interpretation of the model's internal dynamics becomes difficult if the add-ons are serially

correlated and statistically important and no respecification of the structure is attempted.

3 A composite likelihood approach to misspecification

Rather than taking an off-the-shelf model and enriching it with non-structural features or shocks, or completing its probability space with measurement errors, wedges, or margins we take an alternative viewpoint because even with additions, the enlarged models may be far from the DGP. Our basic assumption is that, to investigate a question of interest, a researcher may employ a number of misspecified structural models. These models may differ in the assumptions they make, in the frictions they feature, in the aspects they leave out, or in the transmission mechanism they emphasize. We assume they are theoretically relevant, in the sense that they have implications for the phenomenon under investigation, that are sufficiently heterogeneous so that the information they provide does not entirely overlap, and that share some common parameters. We construct the likelihood function of each model and geometrically combine them. The resulting composite likelihood is either maximized with respect to the unknown parameters or used as an input for quasi-posterior analysis.

Our approach is not designed to *eliminate* misspecification. This is a titanic task, given our focus on structural models and can be achieved only if the set of models spans the DGP, a very strong requirement given the structures available in macroeconomics, or if we complement the set of misspecified structural models with an unrestricted VAR as in Waggoner and Zha (2012). More modestly, we propose an approach that has the potential to *reduce* misspecification, has useful economic interpretations, and sound econometric foundations.

Why would noticing that there are common parameters across models help to reduce misspecification when measuring the MPC? When likelihood methods are used, estimated parameters adjust to reduce the misspecification in the direction that it is largest. If different models are estimated separately, biases will tend to be heterogeneous and likely to reflect the worst misspecification ''direction" each model displays. When models are jointly estimated, however, common parameters are not as free to adjust, because they are constrained by the cross equations restrictions present in all models. Thus, if models which are misspecified in different directions are combined in estimation, biases in the common parameters may be reduced and the quality of inference may improve. We show in section 3.4 that this intuition works in a Monte Carlo setting.

Let the DGP for a vector of variables y_t be represented by a density $F(y_t|\psi)$, where ψ is a parameter vector. The available models are indexed by $i = 1, \ldots, K$ and each produces a density $f_i(y_{it}|\phi_i)$ for the observables y_{it}, which we assume it is of length T_i. y_{it} need not be the same for each i: there may be common and model specific variables. The sample size T_i could also be different and the frequency of the observations may vary with i. Let $\phi_i = [\theta', \eta_i']'$, where θ are

common across specifications and η_i are model specific. Investigators are typically free to choose what goes in θ and η_i. Even though a parameter may appear in all models, a researcher may decide to treat it as model specific because, for example, models are too incompatible with each other. We assume that the K models are misspecified, i.e. there is no ϕ_i such that $f(y_{it}|\phi_i) = F(y_t, \psi),\ \forall i$. Given a vector of weights, $0 < \omega_i < 1, \sum_i \omega_i = 1$, the composite likelihood is

$$CL(\theta, \eta_1, \ldots, \eta_K, y_{1t}, \ldots, y_{KT}) = \Pi_{i=1}^{K}\ \ f(y_{it}|\theta, \eta_i)^{\omega_i} \equiv \Pi_{i=1}^{K} \mathcal{L}(\theta, \eta_i|y_{it})^{\omega_i} \tag{9}$$

3.1 A Taxonomy of Misspecified DSGE Models

Let the data be generated by a (linear) Gaussian state space model:

$$x_t = A(\psi)x_{t-1} + B(\psi)e_t \tag{10}$$

$$z_t = C(\psi)x_{t-1} + D(\psi)e_t \tag{11}$$

where x_t is a $k \times 1$ vector of endogenous and exogenous states, z_t is a $m \times 1$ vector of endogenous controls, $e_t \sim N(0, \Sigma(\psi))$ is a $q \times 1$ vector of disturbances, $\Sigma(\psi)$ a diagonal matrix and ψ a vector of structural parameters; $A(\psi)$ is $k \times k$, $B(\psi)$ is $k \times q$, $C(\psi)$ is $m \times k$, $D(\psi)$ is $m \times q$. For convenience, let the eigenvalues of $A(\psi)$ all be less than one in absolute value. We assume that a researcher observes $y_t = [x_t',\ z_t']'$. If there are latent variables and only a subset of variables $y_{1t} \subset y_t$ is observed, the equations below apply replacing y_t with y_{1t}. There are three possible types of misspecification a DSGE model may display: it may feature the wrong disturbances, the wrong structure, or the wrong observable variables.

Misspecifying the disturbances. Assume that a researcher has the correct $A(\psi), B(\psi), C(\psi), D(\psi)$ matrices and the correct y_t but specifies only a subset of the disturbances present in the DGP, say e_{1t}. Thus, the researcher uses

$$x_t = A(\psi)x_{t-1} + \bar{B}_1(\psi)e_{1t} \tag{12}$$

$$z_t = C(\psi)x_{t-1} + \bar{D}_1(\psi)e_{1t} \tag{13}$$

to estimate the parameters ψ. The log-likelihood of (10)-(11) is proportional to $(y_t - M(\psi)y_{t-1})N(\psi)\Sigma_e N(\psi)'(y_t - M(\psi)y_{t-1})'$, where $M(\psi) = \begin{bmatrix} A(\psi) & 0 \\ C(\psi) & 0 \end{bmatrix}$, $N(\psi) = [B(\psi),\ D(\psi)]'$. The log-likelihood of (12)-(13) is proportional to $(y_t - M(\psi)y_{t-1})\bar{N}(\theta, \eta)\Sigma_{e_1}\bar{N}(\theta, \eta)'(y_t - M(\psi)y_{t-1})'$, where $\bar{N}(\psi) = [\bar{B}_1(\psi), \bar{D}_1(\psi)]'$ and $\Sigma_{e_1} = \Sigma(\theta, \eta)$, where θ and η are parameter vectors such that θ belongs to ψ, while η may not. While $M(\psi)$ could be consistently estimated as long as the omitted shocks are uncorrelated with y_{t-1}, the fact that $\bar{N}(\psi)$ is forced to capture the effect of omitted disturbances implies that ψ can not be consistently estimated from (12)-(13).

Misspecifying the structure. Assume that the researcher has the correct endogenous variables y_t, the correct number and the right sources of disturbances e_t, i.e. if there is a monetary disturbances in the data-generating process the misspecified structure also features a monetary shock, but employs the wrong model for the analysis, meaning either that the mapping between $A(\psi), B(\psi), C(\psi), D(\psi)$ and ψ is incorrect or that (θ, η) are used in place of ψ as structural parameters. Suppose, the researcher uses:

$$x_t = \tilde{A}(\theta,\eta)x_{t-1} + \tilde{B}(\theta,\eta)e_t \tag{14}$$

$$z_t = \tilde{C}(\theta,\eta)x_{t-1} + \tilde{D}(\theta,\eta)e_t \tag{15}$$

The log-likelihood of the estimated model is proportional to $(y_t - \tilde{M}(\theta,\eta)y_{t-1})\tilde{N}(\theta,\eta)\Sigma_e$ $\tilde{N}(\theta,\eta)'(y_t - \tilde{M}(\theta,\eta)y_{t-1})'$ where $\tilde{M}(\theta,\eta)$ and $\tilde{N}(\theta,\eta)$ have the same format as $M(\psi)$ and $N(\psi)$. Estimates of $\tilde{M}(\theta,\eta)$ and $\tilde{N}(\theta,\eta)$ will not asymptotically converge to $M(\psi)$ and $N(\psi)$, making it impossible to consistently estimate the structural parameters. Note that the first type of misspecification could be nested in the second type if shocks are specific to the structure used, but we keep them separated for the sake of clarity.

Misspecifying the observable variables. Here a researcher has the correct model, and the correct disturbances e_t, but uses a subvector of the endogenous variables y_t for estimation. Partition $x_t = [x_{1t}, x_{2t}], z_t = [z_{1t}, z_{2t}]$; partition $A(\psi), B(\psi), C(\psi), D(\psi)$ accordingly and let $w_t = [x_{1t}, z_{1t}]$ be the observables. In terms of w_t, the DGP is [1]:

$$\begin{aligned} x_{1t} &= (A_{11}(\psi) + A_{22}(\psi))x_{1t-1} + (A_{11}(\psi)A_{22}(\psi) - A_{12}(\psi)A_{21}(\psi))x_{1t-2} \\ &+ B_1(\psi)e_t - (A_{22}(\psi)B_1(\psi) - A_{21}(\psi)B_2(\psi))e_{t-1} \end{aligned} \tag{16}$$

$$\begin{aligned} z_{1t} &= A_{22}(\psi)z_{1t-1} + C_{11}(\psi)x_{1t-1} + (A_{22}(\psi)C_{11}(\psi) + C_{12}(\psi)A_{21}(\psi))x_{1t-2} \\ &+ D_1(\psi)e_t + A_{22}D_1(\psi)e_{t-1} \end{aligned} \tag{17}$$

or

$$\grave{x}_t = G(\psi)\grave{x}_{t-1} + F(\psi)\grave{e}_t \tag{18}$$

$$\grave{z}_t = H(\psi)\grave{x}_{t-1} + L(\psi)\grave{e}_t \tag{19}$$

where $\grave{x}_t = [x_{1t}, x_{1t-1}]', \grave{e}_t = [e_t, e_{t-1}]', \grave{z}_t = [z_{1t}, z'_{1t-1}]$. Letting $\grave{w}_t = [\grave{x}_t, \grave{z}_t]$, the log likelihood of the correct model is proportional to $(\grave{w}_t - R(\psi)\grave{w}_{t-1})S(\psi)\Sigma_{\grave{e}}S(\psi)'(\grave{w}_t - R(\psi)\grave{w}_{t-1})'$ where $R(\psi) = \begin{bmatrix} G(\psi) & 0 \\ H(\psi) & 0 \end{bmatrix}, S(\psi) = [F(\psi, L(\psi)]'$.

[1] To derive this expression we assume that x_{1t} and x_{2t} have the same dimension. If not the formulas are more complicated but the essence of the argument holds.

Misspecification may appear, for example, because the model used in the analysis features an insufficient number of lags of w_t to be able to capture the AR and the MA components present in (18)-(19). Let a researcher erroneously use

$$x_{1t} = A_1(\theta,\eta)x_{1t-1} + B_1(\theta,\eta)e_t \tag{20}$$

$$z_{1t} = C_1(\theta,\eta)x_{1t-1} + D_1(\theta,\eta)e_t \tag{21}$$

The log likelihood of the estimated model is proportional to $(w_t - R_1(\theta,\eta)w_{t-1})S_1(\theta,\eta)\Sigma_e S_1(\theta,\eta)'$ $(w_t - R_1(\theta,\eta)w_{t-1})'$ where $R_1(\theta,\eta) = \begin{bmatrix} A_1(\theta,\eta) & 0 \\ C_1(\theta,\eta) & 0 \end{bmatrix}, S_1(\theta,\eta) = [B_1(\theta,\eta), D_1(\theta,\eta)]'$. Clearly, $R_1(\theta,\eta)$, $S_1(\theta,\eta)$ are generally inconsistent estimators of the relevant elements of $R(\psi), S(\psi)$ and thus ψ can not be consistently estimated even when $\psi = (\theta,\eta)$.

The paper focuses on the second type of misspecification, which is the most severe and the most common when estimating structural models. However, other types of misspecification can be analyzed with composite methods; Qu (2018), for example, focuses on the first form of misspecification; and Canova and Matthes (2019) on situations where different types of misspecification may be simultaneously present.

3.2 Why are composite estimators preferable under misspecification? An example

In this example we employ partial equilibrium dynamic models as it is possible to derive a simple, closed form representation for the optimality conditions that allows the reader to understand the properties of our composite estimator. The intuition we discuss also applies, although with considerable complications, to numerical solutions obtained from general equilibrium models.

The first model is an asset pricing model which gives the following Euler equation:

$$1 + R_{t,t+1} = \beta E_t \left(\frac{c_{t+1}}{c_t} \right)^{\gamma} \tag{22}$$

where $R_{t,t+1}$ is the exogenous, and known at t, one period real rate on safe bonds, β is the discount factor and γ is the risk aversion coefficient of the investor' utility. The second is a labor market model which gives the following labor supply equation:

$$N_t^{\eta} = c_t^{-\gamma} \frac{Y_t}{N_t}(1-\alpha)v_t \tag{23}$$

where η is the inverse of the Frisch elasticity, $1-\alpha$ the labor share, $(1-\alpha)\frac{Y_t}{N_t}v_t = w_t$ is the competitive real wage and v_t is a log-normal iid shock to the real wage, which we assume is realized at each t after production and hiring decisions are made. We assume that output and hours are exogenous with respect to the consumption process.

Suppose one want to estimate the risk aversion γ. Log linearizing (22)-(23) we have

$$0 = -\ln\beta + \ln(1+R_{t-1,t}) - \gamma\Delta c_t + u_{1t} \tag{24}$$

$$0 = \ln(1-\alpha) - \gamma c_t - (1+\eta)\ln N_t + \ln Y_t + u_{2t} \tag{25}$$

where u_{1t} captures the expectational consumption growth error and $u_{2t} \equiv v_t$.

Equation (24)-(25) can be compactly written as

$$y_{1t} = A + \rho x_{1t} + v_{1t} \tag{26}$$

$$y_{2t} = B + \rho x_{2t} + \delta x_{3t} + v_{2t} \tag{27}$$

where $y_{1t} = \Delta \ln c_t$ $A = -\frac{\ln\beta}{\gamma}$ $B = \ln(1-\alpha)$ $x_{1t} = \ln(1+R_{t-1,t})$, $\rho = \frac{1}{\gamma}$, $y_{2t} = c_t$, $x_{2t} = \ln Y_t, x_{3t} = -\ln N_t$, $\delta = \frac{1+\eta}{\gamma}$ $v_{1t} = \frac{u_{1t}}{\gamma}$, $v_{2t} = \frac{u_{2t}}{\gamma}$ Here $\theta = \rho = \frac{1}{\gamma}$ is common to the two models; while $\eta_1 = (A, \sigma_1^2)$, $\eta_2 = (B, \delta, \sigma_2^2)$ are (nuisance) parameters specific to each model.

Suppose we have T_1 observations pertaining to the first model and T_2 observations to the second model. The (normal) log-likelihood functions, conditional on x_t, are:

$$\log L_1 \propto -T_1 \log\sigma_1 - \frac{1}{2\sigma_1^2}\sum_{t=1}^{T_1}(y_{1t} - A - \rho x_{1t})^2 \tag{28}$$

$$\log L_2 \propto -T_2 \log\sigma_2 - \frac{1}{2\sigma_2^2}\sum_{t=1}^{T_2}(y_{2t} - \rho x_{1t} - \delta x_{2t})^2 \tag{29}$$

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

$$\log CL = \omega \log L_A + (1-\omega)\log L_B \tag{30}$$

For simplicity, let $\beta = 1$. The maximizers of (30) are:

$$\rho_{CL} = (\sum_{t=1}^{T_1} x_{1t}^2 + \zeta_{1,CL}\sum_{t=1}^{T_1} x_{2t}^2)^{-1}(\sum_{t=1}^{T_1} y_{1t}x_{1t} + \zeta_{1,CL}\sum_{t=1}^{T_1}(y_{2t} - \delta_{CL}x_{3t})x_{2t}) \tag{31}$$

$$\sigma_{1,CL}^2 = \frac{1}{T_1}(\sum_{t=1}^{T_1}(y_{1t} - \rho_{CL}x_{1t})^2 \tag{32}$$

$$\sigma_{2,CL}^2 = \frac{1}{T_2}(\sum_{t=1}^{T_2}(y_{2t} - \rho_{CL}x_{2t} - \delta_{CL}x_{3t})^2 \tag{33}$$

$$\delta_{CL} = = (\sum_{t=1}^{T_2} x_{3t}^2)^{-1}(\sum_{t=1}^{T_2}(y_{2t} - \rho_{CL}x_{2t})x_{3t}) \tag{34}$$

where $\zeta_{1,CL} = \frac{1-\omega}{\omega}\frac{\sigma_{1,CL}^2}{\sigma_{2,CL}^2}$ measures the relative importance of the two types of information for

composite estimation. Instead, the Maximum Likelihood (ML) estimates are:

$$\rho_{1,ML} = (\sum_{t=1}^{T_1} x_{1t}^2)^{-1}(\sum_{t=1}^{T_1} y_{1t}x_{1t}) \tag{35}$$

$$\rho_{2,ML} = (\sum_{t=1}^{T_2} x_{2t}^2)^{-1}(\sum_{t=1}^{T_2} (y_{2t} - \delta_{ML}x_{3t})x_{2t}) \tag{36}$$

$$\sigma_{1,ML}^2 = \frac{1}{T_1}(\sum_{t=1}^{T_1} (y_{1t} - \rho_{1,ML}x_{1t})^2 \tag{37}$$

$$\sigma_{2,CL}^2 = \frac{1}{T_2}(\sum_{t=1}^{T_2} (y_{2t} - \rho_{2,ML}x_{2t} - \delta_{ML}x_{3t})^2 \tag{38}$$

$$\delta_{ML} = = (\sum_{t=1}^{T_2} x_{3t}^2)^{-1}(\sum_{t=1}^{T_2} (y_{2t} - \rho_{2,ML}x_{2t})x_{3t}) \tag{39}$$

The formula in (31) is 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); and iii) implicitly produced by a DSGE-VAR setup (see Del Negro and Schorfheide, 2004), where T_2 observations are added to the original T_1 data points. As (31) indicates, the composite estimator shrinks the information present in (y_{1t}, x_{1t}) towards the information present in (y_{2t}, x_{2t}, x_{3t}) and the amount of shrinkage depends on $(\sigma_1^2, \sigma_2^2, \omega)$, all of which enter ζ_1. The higher ω and σ_2^2 are, the less important (y_{2t}, x_{2t}, x_{3t}) information is. Thus, when estimating common parameters, the composite likelihood gives more importance to data generated by a model with a larger weight and lower relative standard deviation. As (32)-(33)-(34) indicate, model specific parameters are estimated using the information that only that model provides. Although the formulas are similar, these estimates differ from those computed with the likelihood function of each model, see equations (37)-(39), because $\rho_{CL} \neq \rho_{i,ML}, i = 1, 2$. When $\theta = \emptyset$, that is, there are no common parameters, composite estimates are simply likelihood estimates, model by model.

When an array of models is available, composite likelihood estimates of ρ will be constrained by the structure present in all models. For example, when an additional K-1 models have two regressors, equation (31) becomes

$$\rho_{CL} = (\sum_{t=1}^{T_1} x_{1t}^2 + \sum_{i=2}^{K}\zeta_{i,CL}\sum_{t=1}^{T_i} x_{i2t}^2)^{-1}(\sum_{t=1}^{T_1} y_{1t}x_{1t} + \sum_{i=2}^{K}\zeta_{i,CL}\sum_{t=1}^{T_i}((y_{it} - \delta_{i,CL}x_{i3t})x_{i2t}) \tag{40}$$

where $\zeta_{i,CL} = \frac{\omega_i}{\omega_1}\frac{\sigma_{1,CL}^2}{\sigma_{i,CL}^2}$. Hence, the composite likelihood robustifies estimation, because $\rho = \frac{1}{\gamma}$ estimates are required to be consistent with the cross-equation restrictions present in all models.

As the example indicates, y_{1t} and y_{2t} could be different series. The setup we use is also consistent with the possibility that there a single model and (y_{1t}, x_{1t}) (y_{2t}, x_{2t}) are the same series

but with different levels of aggregation (say, aggregate vs. individual consumption). Furthermore, since T_1 and T_2 may be different, the procedure can be used to combine data of various length or the information available at different frequencies (e.g., a quarterly and an annual model). T_1 and T_2 may also represent two samples for the same vector of observables (e.g., before and after a financial crisis). Baumeister and Hamilton (2019) downweight older information when conducting posterior inference. Their procedure mimics a composite estimator where data for the earlier part of the sample, say (y_{1t}, x_{1t}), is more noisy and thus given less weight than more recent data.

Given the shrinkage nature of composite estimators, we expect them to do well in mean square error (MSE) relative to maximum likelihood estimators. Algebraic manipulations of (31) gives $\rho_{A,CL} = \chi\rho_{1,ML} + (1-\chi)\rho_{2,ML} = \chi\rho_1 + (1-\chi)\rho_2 + \chi B_1 + (1-\chi)B_2$ where $\chi = \frac{1}{1+\frac{\omega_2 var(\rho_{1,ML})}{\omega_1 var(\rho_{2,ML})}}$, $var(\rho_{i,ML})$ are the variances of the ML estimators, i=1,2; $\rho_1 = E(\rho_{1,ML})$ and $\rho_2 = E(\rho_{2,ML})$; $B_1 = \frac{\sum_t x_{1t}(y_{1t}-\rho_{1,ML}x_{1t})}{\sum_t x_{1t}^2}$ and $B_2 = \frac{\sum_t x_{2t}[(y_{2t}-\delta_{ML}x_{3t}-\rho_{2,ML}x_{2t})}{\sum_t x_{2t}^2}$.

Let ρ^* be the expected value of the CL estimator and assume that $\rho^* = \chi\rho_1 + (1-\chi)\rho_2$ [2]. To insure that MSE_{CL} is less, say, of $MSE_{1,ML}$, we need $(1-\chi^2)E(B_1^2) - (1-\chi)^2E(B_2^2) - 2\chi(1-\chi)EB_1B_2 > 0$, where E denotes the expectation operator. Suppose $EB_1B_2 = 0$, i.e. the biases in $\rho_{1,ML}, \rho_{2,ML}$ are independent. Then, the composite estimator is preferable if

$$1 \quad > \quad \frac{EB_2^2}{EB_1^2} - 2\frac{\omega}{(1-\omega)}\frac{var(\rho_{2,ML})}{var(\rho_{1,ML})} \tag{41}$$

(41) links the relative weights, the relative biases, and the relative variances of the maximum likelihood estimators of two models. Other things being equal, the higher is the bias of the maximum likelihood estimator obtained with (y_{2t}, x_{2t}, x_{3t}), the higher should ω be for the CL estimator to be MSE superior. Similarly, the higher is the variability of the ML estimator constructed with (y_{1t}, x_{1t}), the lower needs to be $1-\omega$ for the CL estimator to dominate. When the ML estimators have similar biases, $\frac{1-\omega}{\omega} > 1 - 2\frac{var(\rho_{2,ML})}{var(\rho_{1,ML})}$ is sufficient for the CL estimator to be MSE superior, a condition easy to check in practice.

When the biases are negatively correlated, as in the experimental design of section 3.4, MSE improvements can be obtained under milder restrictions. For example, a CL estimator is preferable as long as the bias of the second ML estimator is not too large:

$$EB_2^2 \quad < \quad \frac{1-\chi}{1+\chi}EB_1^2 - \frac{\chi}{1+\chi}EB_1B_2 \tag{42}$$

Thus, as intuition would suggest, whenever individual estimators have negatively correlated biases, we expect the CL estimator to produce MSE improvements.

When y_t has been generated by a density $F(y_t, \psi)$ but a researcher uses the density $f_i(y_t, \phi_i), i =$

[2]This is a valid assumption in our setup because the models have no common equations.

$1, \ldots, K$ for the analysis, one can define the Kullback-Leibler (KL) divergence as:

$$KL_i(y, \psi, \phi_i) = \sum_{j=1}^{N} F(y_j, \psi) * \log(\frac{f_i(y_j, \phi_i)}{F(y_j, \psi)}) \tag{43}$$

which it is interpreted as the bits of information lost in characterizing y_t using f_i rather than F. The KL divergence has appealing decision theory foundations and can be used to rank misspecified models. In fact, if f_1 and f_2 are available and $KL_1(y, \phi_1, \psi) > KL_2(y, \phi_2, \psi)$, then f_2 is less misspecified than f_1. Because the composite model averages different misspecified structural models, we expect it to reduce the misspecification of the original models. To examine if this is the case, one could compute $\tilde{KL}_i = \int KL_i(y, \phi_i, \psi)p(\phi_i|y)d\phi_i$ where $KL_i(y, \phi_i, \psi)$ is the KL divergence of model i and $p(\phi_i|y)$ is the (asymptotic or posterior) distribution of ϕ_i computed in model i and compare it with $\tilde{KL}_g = \int KL_g(y, \chi, \psi)p(\chi|y)d\chi$, where $g(y, \chi, \psi) = \sum_i f_i(y, \phi_i)^{\omega_i}$ is the density of the composite model, and $p(\chi|y)$ the composite (asymptotic or posterior) distribution of $\chi = (\phi_1, \ldots, \phi_K, \omega_1, \ldots, \omega_K)$. Section 3.4 provides evidence on the performance of composite estimators and composite models for some DGPs. To approximate $F(y_t, \psi)$ one can use the histogram of the data or a VAR as long as standard regularity conditions are met.

In a traditional composite likelihood approach, ω_i are fixed quantities, chosen by the investigator. When ω_i is a random variable, its quasi-posterior mode informs us about the relative misspecification of the models entering the composite likelihood. To illustrate this property, let $p(\omega) \propto \omega^{\alpha_1 - 1}(1-\omega)^{\alpha_2 - 1}$, where α_1, α_2 are known, and let the prior for $(\rho, \sigma_1^2, \delta, \sigma^2)$ be diffuse. The composite posterior kernel of ω, conditional on $(\rho, \sigma_1^2, \delta, \sigma^2)$ is $CP(\omega|\rho, \sigma_1^2, \delta, \sigma^2) = (L_1^{\omega} L_2^{1-\omega})\omega^{\alpha_1 - 1}(1-\omega)^{\alpha_2 - 1}$. Taking logs and maximizing we have

$$\log L_1 - \log L_2 + \frac{(\alpha_1 - 1)}{\omega} - \frac{(\alpha_1 - 1)}{1-\omega} = 0 \tag{44}$$

This is a quadratic equation in ω and the relevant solution is $0 < \omega_1 < 1$. Totally differentiating (44) one finds that ω_1 is increasing in $\log L_1 - \log L_2$. Completing the square terms of the likelihoods, and conditioning on the mode estimators of $(\rho, \sigma_1^2, \delta, \sigma_2^2)$, one obtains

$$\log L_1 - \log L_2 \propto -\frac{1}{2\sigma_1^2}\sum_{t=1}^{T_1}(y_{1t|t-1} - \rho x_{1t})^2 + \frac{1}{2\sigma_2^2}\sum_{t=1}^{T_2}(y_{2t|t-1} - \rho x_{2t} - \delta x_{3t})^2| \tag{45}$$

where $y_{it|t-1}$ is the optimal predictor of y_{it}. Thus, $\log L_1 - \log L_2$ reflects relative misspecification (how far the predictions of each model are from the optimal predictor for each y_{it}) and the mode of ω is higher when model 1 is less misspecified [3]. In finite samples, $0 < \omega < 1$. The same will hold in large samples, if $y_{1t} \neq y_{2t}$, and the models are equally poor in characterizing y_{1t} and y_{2t}.

[3]When (y_{1t}, X_{1t}) and (y_{2t}, X_{2t}) are vectors the equations should be adjusted accordingly. When y_{1t} is a $m \times 1$ vector and y_{2t} is, e.g., a scalar or when y_{1t} is different from y_{2t}, $\log L_1 - \log L_2$ reflects, apart from differences in the variances, the average misspecification in all the equations of model 1 relative to the misspecification of the single equation of model 2. Thus, if model 1 has some very poorly specified equations, it may have low a-posteriori ω, even though certain equations are correctly specified (and ρ appears in those equations).

While we work under the assumption that all models are misspecified, one may like to know what happens to our estimation approach when one model is close to best in a KL sense. When $y_{1t} \neq y_{2t}$, no general conclusions can be drawn, even though we expect the model close to the best to receive larger weight. When $y_{1t} = y_{2t}$ and as sample size grows to infinity, $\omega_i \to 1$ for the model closest to the best in a KL sense. Thus, our composite estimates will be close to those obtained by minimizing the KL distance. We provide some evidence on these issues in samples of moderate size in subsection 3.4.

3.3 Relationship with the literature

Researchers often use Bayes factors to rank models and Bayesian model averaging (BMA) to combine their predictions. Asymptotically, when models are misspecified the Bayes factor selects the model closest to the data in a KL sense, regardless of the prior, and BMA puts all weight on that model (see e.g. Fernandez Villaverde and Rubio Ramirez, 2004). Because the quasi-posterior mode of ω measures the relative misspecification of the available models, we expect it to provide similar ranking information when the data used by each model is the same. However, Bayes factors and BMA weights can only be computed when $y_{1t} = y_{2t}$ and $T_1 = T_2$; the posterior of ω can be computed even without these restrictions. Also, our analysis provides a measure of uncertainty for ω. No such measure is generally available for BMA weights. Finally, BMA only gives an ex-post combination of individual model estimates. Some experimental evidence on the performance of the two ranking devices is in section 3.4.

It is useful to highlight how a composite setup relates to the mixture procedure of Waggoner and Zha (2012) and to robustness approaches (Hansen and Sargent, 2008, Giacomini and Kitagawa, 2017). In Waggoner and Zha, the estimated model linearly (rather than geometrically) combines the likelihoods of a structural model and a VAR (rather than K structural models), but the weights have a Markov switching structure. Their objective function is:

$$\log L = \sum_{t=1}^{\min\{T_1,T_2\}} \log(w_t L(\rho, \sigma_1^2 | y_{1t}, x_{1t}) + (1 - w_t) L(\rho, \sigma_2^2, \delta | y_{2t}, x_{2t}, x_{3t})) \tag{46}$$

Simple manipulations reveal that (46) and the log of (9) differ by Jensen's inequality terms.

While a-priori both composite and finite mixture devices are appealing, a composite likelihood has three advantages. From a computational point of view, when the model' decision rules have a linear structure, estimators for θ have a closed form expression in the composite likelihood case, but not in the finite mixture case. In addition, in a finite mixture it must be the case that $y_{1t} = y_{2t}$, and $T_1 = T_2$, since the models represent alternatives that could have generated the same data. These restrictions are unnecessary in the composite likelihood formulation. Finally,

in Waggoner and Zha the composite model is the DGP; here the composite model could still be misspecified, hopefully less than the individual models.

Hansen and Sargent (2008) robustify decisions and counterfactuals using a density for the parameters which is a tilted version of the posterior distribution. Let $p(\phi_i) \equiv p(\phi_i|y_t)$ be the posterior of ϕ_i, computed using the information in y_t. Hansen and Sargent's density is $\pi(\phi_i) = \frac{\exp\{\lambda L(\phi_i)\}p(\phi_i)}{\int \exp\{\lambda L(\phi_i)\}p(\phi_i)d\phi_i}$, where $L(\phi_i)$ is a loss function and λ is the ray of a ball around $p(\phi_i)$ in which we seek robustness. Two differences between Hansen and Sargent's and our approach are immediately evident. In the latter, robustness is sought for all parameters within a model; we seek robust estimators of a subset of the parameters across models. Moreover, Hansen and Sargent's approach protects a researcher from the worst possible outcome but it is not suited to deal with instabilities or time variations in the DGP, if the ball is small. In our approach, the weights are endogenously adaptable to the features of the sample.

Giacomini and Kitagawa (2017) propose a method to conduct posterior inference on the impulse responses of partially identified SVARs that is robust to prior choices for the rotation matrices. They summarize the class of posteriors generated by alternative priors by reporting a posterior mean bounds interval, interpreted as an estimator of the identified set, and a robustified credible region, measuring the uncertainty about the identified set. Once again, two differences with our approach are evident. First, they seek robustness with respect to prior rotations; we are looking for estimators which are robust across structural models. Second, they care about impulse responses in SVARs; we care about (common) parameters in structural models.

It is also useful to relate composite and GMM estimators. A composite likelihood estimator with fixed model weights solves moment conditions of the form $\sum_i \omega_i \frac{\partial L(\phi_i|y)}{\partial \phi_i} = 0$. Thus, composite likelihood estimators are over-identified GMM estimators, where the orthogonality conditions are a weighted average of the scores of each structural model with fixed weights. The larger is the set of models considered, the more over-identified the estimators are. When ω_i are optimized, the moment conditions are similar to those of generalized empirical likelihood (GEL) methods (see Newey and Smith, 2004) and of minimum distance estimators (see Ragusa, 2011).

3.4 Some experimental evidence

To understand the kind of gains one should expect from composite estimators and the situations when these are more likely to materialize, we perform an experiment where the DGP is a univariate ARMA(1,1): $\log y_t = \rho \log y_{t-1} + \theta \log e_{t-1} + \log e_t,\ \log e_t \sim (0, \sigma^2)$, and the models used in estimation are an AR(1): $\log y_t = \rho_1 \log y_{t-1} + \log u_t$ and an MA(1): $\log y_t = \log \epsilon_t + \beta_1 \log \epsilon_{t-1}$. Thus, the example fits case 2 of section 3.1: both models use the incorrect decision rules. We present results for four different combinations of (ρ, θ): two generating proper ARMA processes

(DGP1:$\beta = 0.6, \theta = 0.5$ and DPG2: $\beta = 0.6, \theta = 0.8$, which produces larger first order autocorrelation in $\log y_t$); one close to an AR(1) (DGP3: $\beta = 0.9, \theta = 0.2$); and one close to an MA(1) (DGP4:, $\beta = 0.3, \theta$=0.8). For DGP1 we present results varying $\sigma = 0.2, 0.5, 0.8, 1.0, 1.5$ and for DGP3 and DGP4 results varying $T = 50, 100, 250$. Since DGP3 and DGP4 are close to one of the estimated models, one should expect the sample size to be more important for the conclusions one draws about composite estimators in these cases.

We focus attention on the relationship between the true and the estimated σ, which is common across models [4]. Because both models disregard part of the serial correlation of the DGP, $\sigma_u, \sigma_\epsilon$ are upward biased. Would geometrically combining the likelihoods give a better estimate of σ? Would a composite model be less misspecified than both the AR(1) and the MA(1)? Do the conclusions depend on the DGP or the sample size? How do the posterior mode of ω and a BMA weight relate?

We set $\omega_2 = 1 - \omega_1$ and treat $\omega = \omega_1$ either as fixed or as random. When it is fixed, we construct composite estimates equally weighting the two models ($\omega = 0.50$) or using weights that reflect the relative mean square error (MSE) in a training sample with 100 observations. In the baseline specifications T=50. Since there are only two parameters in the AR(1) and MA(1), and three in the composite models, this is actually a medium sized sample.

We estimate the three composite specifications, the AR(1), and the MA(1) models with Bayesian methods. The prior for the AR (MA) parameter is truncated normal with mean zero and variance 0.2 and the prior for σ is flat in the positive orthant. The prior for ω is Beta(1,1). We draw sequences with 50000 elements and keep 1 out of every 5 of the last 25000 draws for inference. The scale parameter of the Metropolis random walk is optimized using an adaptive scheme and the Hessian at the mode is used for the proposal density.

Table 1 presents the mean square error of σ, computed using posterior (composite posterior) draws (MSE_j) and the KL divergence (KL_j), computed averaging over posterior (composite posterior) draws of the parameters, j=1,. . . ,5.

Composite specifications produce better estimates of σ and at least one of the composite models has lower MSE than both the AR(1) and the MA(1). The magnitude of the gains depends on the DGP and the persistence of the data, but not on the true σ or the sample size T. Furthermore, there is a composite model which reduces the misspecification of both the AR(1) and the MA(1) models - the equally weighted specification for DGP1 and DGP2 and the random ω specification for DGP3 and DGP4 - and for many of the cases examined more than one composite model has smaller KL divergence. The superiority of composite models is unaffected by T. The random

[4] σ may not be the most natural parameter one would focus attention on to perform joint estimation. We have decided to measure the improvements of composite approaches looking just at σ to keep the design as simple as possible.

Table 1: Monte Carlo results

$\log y_t = \rho \log y_{t-1} + \beta \log e_{t-1} + \log e_t,\ \log e_t \sim N(0, \sigma^2)$							
DGP	Sample Size	Statistic	CL, random weigths	CL, equal weights	CL, MSE weights	AR(1)	MA(1)
$\sigma^2 = 0.2, \rho = 0.6, \beta = 0.5$	T=50	MSE	0.173	0.202	0.167	0.176	0.253
		KL	14.99	8.13	13.26	13.94	4.70
$\sigma^2 = 0.5, \rho = 0.6, \beta = 0.5$	T=50	MSE	0.061	0.075	0.058	0.066	0.107
		KL	13.91	7.89	13.22	13.77	6.06
$\sigma^2 = 0.8, \rho = 0.6, \beta = 0.5$	T=50	MSE	0.021	0.027	0.019	0.026	0.050
		KL	12.55	5.87	11.46	12.17	5.98
$\sigma^2 = 1.0, \rho = 0.6, \beta = 0.5$	T=50	MSE	0.008	0.011	0.007	0.012	0.030
		KL	11.83	5.32	10.63	11.70	7.77
$\sigma^2 = 1.2, \rho = 0.6, \beta = 0.5$	T=50	MSE	0.006	0.007	0.005	0.007	0.017
		KL	9.34	4.49	8.03	9.07	9.10
$\sigma^2 = 0.5, \rho = 0.6, \beta = 0.8$	T=50	MSE	0.148	0.168	0.205	0.204	0.292
		KL	11.00	5.02	10.53	11.41	4.93
$\sigma^2 = 1.0, \rho = 0.6, \beta = 0.8$	T=50	MSE	0.009	0.011	0.036	0.035	0.060
		KL	8.90	5.33	9.54	10.41	9.07
$\sigma^2 = 0.5, \rho = 0.9, \beta = 0.2$	T=50	MSE	0.028	0.169	0.020	0.021	0.429
		KL	11.25	16.93	13.21	12.40	7.78
$\sigma^2 = 1.0, \rho = 0.9, \beta = 0.2$	T=50	MSE	0.008	0.077	0.005	0.008	0.368
		KL	9.90	19.27	11.32	10.93	14.61
$\sigma^2 = 1.0, \rho = 0.9, \beta = 0.2$	T=100	MSE	0.006	0.152	0.005	0.007	0.173
		KL	17.07	29.60	22.83	20.91	36.75
$\sigma^2 = 1.0, \rho = 0.9, \beta = 0.2$	T= 250	MSE	0.002	0.136	0.002	0.002	0.414
		KL	5.93	16.66	9.48	9.07	21.33
$\sigma^2 = 0.5, \rho = 0.3, \beta = 0.8$	T=50	MSE	0.131	0.152	0.171	0.189	0.179
		KL	4.73	5.91	7.11	11.74	3.74
$\sigma^2 = 1.0, \rho = 0.3, \beta = 0.8$	T=50	MSE	0.006	0.009	0.017	0.027	0.009
		KL	4.88	5.32	6.11	9.62	5.94
$\sigma^2 = 1.0, \rho = 0.3, \beta = 0.8$	T=100	MSE	0.007	0.011	0.023	0.033	0.011
		KL	4.45	4.14	7.02	7.73	5.06
$\sigma^2 = 1.0, \rho = 0.3, \beta = 0.8$	T= 250	MSE	0.003	0.012	0.024	0.032	0.004
		KL	6.20	8.11	9.25	10.89	6.06

The MSE weights for the AR(1) and the MA(1) are computed in a pre-sample with T=100. MSE is the mean square error of the estimated σ; KL measures the divergence with respect to the DGP on average using the posterior (composite posterior) distribution of the parameters.

ω specification performs well in the KL metric for several parameter configurations and seems preferable for highly persistent data or when the DGP is "close" to one of the two basic models.

Table 2: Posterior of ω and BMA weight

$\log y_t = \rho \log y_{t-1} + \beta \log e_{t-1} + \log e_t,\ \log e_t \sim N(0, \sigma^2)$			
DGP	Sample size	ω estimate (s.d)	BMA weight
$\sigma^2 = 0.2, \rho = 0.6, \beta = 0.5$	T=50	0.984 (0.03)	1.00
$\sigma^2 = 0.5, \rho = 0.6, \beta = 0.5$	T=50	0.984 (0.03)	1.00
$\sigma^2 = 0.8, \rho = 0.6, \beta = 0.5$	T=50	0.992 (0.03)	1.00
$\sigma^2 = 1.0, \rho = 0.6, \beta = 0.5$	T=50	0.992 (0.03)	1.00
$\sigma^2 = 1.2, \rho = 0.6, \beta = 0.5$	T=50	0.994 (0.03)	1.00
$\sigma^2 = 0.5, \rho = 0.6, \beta = 0.8$	T=50	0.984 (0.03)	1.00
$\sigma^2 = 1.0, \rho = 0.6, \beta = 0.8$	T=50	0.990 (0.03)	1.00
$\sigma^2 = 0.5, \rho = 0.9, \beta = 0.2$	T=50	0.999 (0.004)	1.00
$\sigma^2 = 1.0, \rho = 0.9, \beta = 0.2$	T=50	0.999 (0.008)	1.00
$\sigma^2 = 1.0, \rho = 0.9, \beta = 0.2$	T=100	1.000 (0.007)	1.00
$\sigma^2 = 1.0, \rho = 0.9, \beta = 0.2$	T=250	0.999 (0.004)	1.00
$\sigma^2 = 0.5, \rho = 0.3, \beta = 0.8$	T=50	0.014 (0.103)	0.994
$\sigma^2 = 1.0, \rho = 0.3, \beta = 0.8$	T=50	0.012 (0.057)	0.946
$\sigma^2 = 1.0, \rho = 0.3, \beta = 0.8$	T=100	0.008 (0.044)	0.105
$\sigma^2 = 1.0, \rho = 0.3, \beta = 0.8$	T=250	0.002 (0.02)	0.002

The table reports the posterior mode and the standard deviation of ω and the BMA weight on the AR(1).

Table 2 has the posterior mode of ω (which is our estimated weight on the AR(1) model), the posterior standard deviation of ω, and the BMA weight on the AR(1) model. Because the two models share the same observable, a comparison between BMA and the posterior mode of ω is possible. The mode of ω and a BMA weight have similar information in the majority of cases we consider. However, when the DGP is close to an MA(1) and T is short, the two measures disagree regarding the likelihood of the AR(1) model. This divergence disappears when $T \geq 100$ and both models put smaller weight on such a model. Note that the posterior of ω is updated in the direction of the model with smaller KL divergence, even when $T = 50$.

Although our approach is not designed for situations where one of the models in the pool is the DGP, it works well also in these cases. Table 3, which presents the evolution of the posterior of the weights as sample size increases, shows that the posterior of ω asymptotically concentrates at the corner solution corresponding to the correct model, although at a somewhat slower rate than a BMA weight. Furthermore, when T is small our approach gives more conservative estimates of the weights than BMA.

In sum, our simulations show that estimation outcomes can be improved and misspecification reduced with composite methods. Furthermore, the posterior mode of ω gives a model ranking device with useful properties: its modal value agrees with a BMA weight in many specifications

Table 3: Posterior estimates of ω

	Mode	Mean	Median	Std deviation	BMA weight
DGP=$y_t = 0.8y_{t-1} + e_t,\ e_t \sim N(0,1)$					
Prior		0.5	0.5	0.288	
T=50	0.994	0.978	0.985	0.023	0.991
T=100	0.997	0.983	0.986	0.018	1.000
T=250	0.998	0.990	0.993	0.010	1.000
T=500	0.999	0.993	0.995	0.006	1.000
DGP=$y_t = 0.7e_{t-1} + e_t,\ e_t \sim N(0,1)$					
Prior		0.5	0.5	0.288	
T=50	0.356	0.468	0.432	0.187	0.024
T=100	0.007	0.220	0.147	0.177	0.015
T=250	0.003	0.048	0.030	0.050	0.006
T=500	0.002	0.034	0.021	0.030	0.002

and it is superior when T is small and MA components dominate. Finally, the quasi-posterior standard deviation of ω gives us a way measure the credibility of the rankings - no uncertainty can be generally attached to a BMA weight.

4 Estimation and inference

In a traditional setting, where the models entering the composite likelihood are marginal or conditional versions of the true DGP, composite likelihood estimators are consistent and asymptotically normal (see e.g. Varin, 2011) but are inefficient, because information about the DGP is disregarded, and ω_i can be selected to minimize their inefficiency.

Our setup differs from the traditional one in four respects. First, $F(y_t, \psi)$ is unavailable - the process generating the data is unknown. Second, $f(y_{it} \in A_i, \phi_i)$ are neither marginal nor conditional densities, but misspecified approximations of the unknown DGP. Thus, for all (ϕ_i), the KL divergence between $F(y_t, \psi)$ and $f(y_{it} \in A_i, \phi_i)$ is positive, $\forall i$. Third, $f(y_{it} \in A_i, \phi_i)$ need not be independent (models may share equations) nor compatible, in the sense that the likelihood estimator $\phi_{i,ML}$ asymptotically converges to the same value. Finally, we treat ω_i as a random variable and wish to construct estimators for the common parameters θ, the nuisance parameters η_i, and the weights $\omega_i, i = 1, 2, \ldots, K$.

Because all available models are misspecified, maximum likelihood estimators obtained from each $f(y_{it} \in A_i, \phi_i)$ are inconsistent and, as a consequence, the composite likelihood estimator obtained for given ω_i is also inconsistent. As earlier work by White (1982) and Domowitz and White (1982) shows, as the sample size grows and under regularity conditions, $\phi_{i,ML}$ converges to ϕ_0, the pseudo-parameter vector minimizing the KL divergence from the DGP. Moreover, $\sqrt{T}(\phi_{i,ML} - \phi_0) \sim N(0, G_i^{-1})$, where $G_i = H_i J_i^{-1} H_i$ is the Godambe information matrix for model

i, J_i the variability matrix and H_i the sensitivity matrix. Thus, with model misspecification the pivot of the asymptotic distribution is the minimizer of the KL divergence, rather than the true parameter vector; and the Godambe (sandwich) information matrix is evaluated at the minimizer of the KL divergence, rather than the true parameter vector.

The composite pool defines a density for a different misspecified model (a weighted average of the K models). When w_i are fixed, ϕ_{CL} asymptotically approaches the pseudo-parameter value, say $\phi_{0,CL}$, minimizing the KL divergence between the density of the composite pool and the DGP. $\phi_{0,CL}$ is not, in general, a weighted average of $\phi_{0,i}$ because models are not necessarily independent. Furthermore, $\sqrt{T}(\phi_{CL} - \phi_{0,CL}) \sim N(0, G^{-1})$, where $G = HJ^{-1}H$ and H and J evaluated at the composite likelihood estimator (see Appendix A for details).

We work in a Bayesian framework rather than a classical likelihood setup. There is a growing literature examining the properties of Bayesian estimator under model misspecification. For example, Fernandez Villaverde and Rubio Ramirez (2004) show, that under mild regularity conditions - the most important ones being that the support of the prior includes the KL optimizer and that the likelihood function can be computed - the prior asymptotically vanishes; the posterior mode convergences in probability to the KL optimizer; and that Bayes factor of any model over the best model under KL distance approaches zero asymptotically. These results have been refined in a number of papers using weaker or alternative assumptions (see e.g. Clydec and Iversen, 2013). Furthermore, Klein and Van der Vaart (2012) have shown that the Bernstein-Von Mises theorem holds under misspecification; and Bissiri et al. (2016) provide a general framework for updating prior beliefs when the data is represented with a general loss function. Thus, valid posterior inference can be performed, even when the model is misspecified.

Our analysis treats ω as a random variable and thus seeks to construct the quasi-posterior distributions for the structural parameters and for the ω vector. As long as $0 < \omega < 1$, standard asymptotic results derived in the literature hold. When this is not the case, we conjecture that results similar to those of Andrews (1998) could be established.

4.1 Bayesian quasi-posteriors

In this paper, we do not rely on asymptotic results. We combine the composite likelihood (9) with a prior for $\chi = (\theta, \eta_1, ...\eta_K, \omega_1, \ldots, \omega_K)$, compute the joint quasi-posterior, which we then integrate with respect to the nuisance parameters to obtain the marginals of θ and ω. We employ a multiple block Metropolis-Hastings approach to numerically compute sequences from this joint quasi-posterior distribution.

Given (y_{it}, T_i), we assume that $\sup_{\{\phi_i\}} f(y_{it} \in A_i, \phi_i) < b_i \leq B < \infty$, a condition generally satisfied for structural macroeconometric models, that $\mathcal{L}(\theta, \eta_i | y_{i,T_i})$ can be constructed for each i,

and that the composite likelihood $CL(\chi|y_{1,T_1},\ldots,y_{K,T_k})$ exists for $0<\omega_i<1$, $\sum_i \omega_i = 1$. Let the priors for ϕ_i be of the form:

$$p(\theta,\eta_i) = p(\theta)p(\eta_i|\theta,y_{i0}) \tag{47}$$

where y_{i0} is a training sample. In (47) we allow for a data-based prior specification for η_i, as in Del Negro and Schorfheide (2008), which is advisable to put models on the same ground as far as matching certain statistics of the data. Making the prior of η_i data-based also helps to avoid identification problems when ω_i is close to zero and to make it more likely that the minimizer of the KL divergence belongs to the support of the prior , see e.g. Walker (2012).

The composite posterior kernel is:

$$\check{p}(\chi|y_{1,T_1},\ldots,y_{k,T_k}) \quad = \quad \Pi_i \mathcal{L}(\theta,\eta_i|y_{i,T_i})^{\omega_i} p(\eta_i|\theta,y_{i0})^{\omega_i} p(\theta)p(\omega_i) \tag{48}$$

which can be used to estimate χ as described, e.g. in Chernozukov and Hong (2003). For computational and efficiency reasons, we employ a $K+1$ block Metropolis-Hastings algorithm. Herbst and Schorfheide (2015) also suggested drawing parameters in blocks. While they randomly split the parameter vector in blocks at each iteration, the blocks here are predetermined by the K models of interest. In the applications of section 5, the prior for ω we employ is subjective. However, one can also consider using a training sample to calibrate it, i.e. use $p(\omega|y_{io})$. This could help to obtain faster convergence of the algorithm described below under standard stationarity assumptions.

When K is large, the parameter space will also be large and computations may be demanding. Hence, one may want to preliminarily obtain the posterior of η_i using (y_i,T_i), condition on these posterior distributions when estimating (θ,ω), and iterate. Since only the information contained in model i is used to estimate η_i, the approach seems sensible and practical.

4.2 MCMC Algorithm

The algorithm consists of four steps:

1. Start with some $\chi_0 = [\eta_1^0 \ldots \eta_K^0, \theta^0, \omega_1^0 \ldots \omega_K^0]$. For $iter = 1 : draws$ do steps 2.-4.

2. For $i = 1:K$, draw η_i^* from a symmetric proposal P^{η_i}. Set $\eta^{iter} = \eta_i^*$ with probability

$$\min\left(1, \frac{\mathcal{L}([\eta_i^*,\theta^{iter-1}]\,|y_{i,T_i})^{\omega_i^{iter-1}} p(\eta_i^*|\theta^{iter-1},y_{i0})^{\omega_i^{iter-1}}}{\mathcal{L}(\left[\eta_i^{iter-1},\theta^{iter-1}\right]\,|y_{i,T_i})^{\omega_i^{iter-1}} p(\eta_i^{iter-1}|\theta^{iter-1},y_{i0})^{\omega_i^{iter-1}}}\right) \tag{49}$$

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

$$\min\left(1, \frac{\mathcal{L}([\eta_1^{iter},\theta^*]\,|y_{1,T_1})^{\omega_1^{iter-1}} \ldots \mathcal{L}([\eta_K^{iter}\theta^*]\,|y_{K,T_K})^{\omega_K^{iter-1}} p(\theta^*)}{\mathcal{L}([\eta_1^{iter},\theta^{iter-1}]\,|y_{1,T_1})^{\omega_1^{iter-1}} \ldots \mathcal{L}([\eta_K^{iter},\theta^{iter-1}]\,|y_{K,T_K})^{\omega_K^{iter-1}} p(\theta^{iter-1})}\right) \tag{50}$$

4. Draw ω_i^* from a symmetric proposal P^ω. Set $\omega^{iter} = \omega^* = (\omega_1^* ... \ \omega_k^*)$ with probability

$$\min\left(1, \frac{\mathcal{L}([\eta_1^{iter}, \theta^{iter}] \ |y_{1,T_1})^{\omega_1^*} \dots \mathcal{L}([\eta_K^{iter} \theta^{iter}] \ |y_{K,T_K})^{\omega_K^*} p(\omega^*)}{\mathcal{L}([\eta_1^{iter}, \theta^{iter}] \ y_{1,T_1})^{\omega_1^{iter-1}} \dots \mathcal{L}([\eta_i^{iter}, \theta^{iter}] \ |y_{K,T_K})^{\omega_K^{iter-1}} p(\omega^{iter-1})}\right) \tag{51}$$

When the proposals are asymmetric, the acceptance probability should be adjusted appropriately. Note that in (49) only the likelihood of model i matters. When the K models feature no nuisance parameters, steps 2.-3. can be combined in a single step. Similarly, when $\theta = \emptyset$ steps 3 and 4 can be eliminated. Also, when ω_i's are fixed, step 4 disappears. Finally, when $\omega_i = 0, i \neq k$, $\omega_k = 1$, the algorithm collapses into a standard Block Metropolis MCMC. A random walk proposal for (θ, η_i) works well in practice; a multivariate logistic proposal or an independent Dirichlet proposal are natural choices for ω_i if K is small. For large K, a "random walk Dirichlet" proposal seems appropriate (see Appendix B).

Although ω's are time independent, adjusting the MCMC algorithm to allow for time varying ω's is easy. For example, one can accommodate time-varying weights non-parametrically, repeating the computations using a rolling window of fixed-size data. Alternatively, one could consider a parametric specification for the time variations and add a MCMC step which draws the innovations from a Dirichlet distribution. With time varying weights, one could look at their evolution to understand how the data is filtered. Thus, as in Waggoner and Zha (2012), the cross equation restrictions of different models could receive different weights in different portions of the sample.

4.3 Computational costs

It might be useful to highlight the computational costs of our approach. Given the structure of our algorithm, we can derive some bounds on the computation time needed in each loop. Suppose that in a standard MCMC setup, generating a draw from the proposal for the parameters, evaluating the associated priors, solving the model, and evaluating the likelihood takes x seconds for the "slowest" model. Then, if we study K models, $K * x$ seconds is the upper bound for the time it takes to go through one loop of the MCMC for the K models. How does this number compare with the time needed to go through one loop in our MCMC algorithm?

In our sampler, we first need to generate a draw for the proposals for the common parameters, evaluate the associated priors, solve each model and compute the model-specific likelihood functions.This will also take roughly $K * x$ seconds. [5] We also need to draw model-specific parameters,

[5]Since we only need to generate the common parameters, the proposal might be slightly faster than in the case of generating a draw for each model separately. In practice this difference is negligible. In this calculation we also do not explicitly consider the cost of computing the acceptance probability in each Metropolis step which is very fast.

for which we have to generate a draw from the proposal, evaluate the associated priors, solve each model and compute the likelihood functions. In a brute force implementation this would take $K * x$ seconds as well. But, conditional on the common parameters, these steps can be carried out in parallel. If we have access to K cores, this block of commands takes approximately x seconds. The final step of our algorithm is the updating of the weights. Here we do not need to solve the models or compute likelihood values because neither model-specific nor common parameters are updated, which are the main costs in terms of time. Because the cost of this final step is negligible, the computational cost of one loop in our algorithm is roughly $(K + 1) * x$ seconds.

4.4 Adjusting percentiles of the MCMC distribution

Our estimation problem is non-standard since y_{it} are not necessarily mutually exclusive across i. Thus, for example, if all models feature a nominal interest rate, that series may be used K times. Naive implementations of a MCMC approach produce marginal posterior percentiles for θ which are too concentrated, because the procedure treats y_{it} as if it was independent across i (see Mueller, 2013). In Appendix B we show that, under regularity conditions, the composite posterior has an asymptotically normal shape, but the covariance matrix is the sensitivity matrix H, rather than the Godambe matrix G.

To obtain the correct asymptotic coverage one could use a normal posterior with sandwich covariance matrix. Following Ribatet et al. (2012) and Qu (2018), we directly add two steps to the MCMC algorithm to take care of the problem. In the first we compute the "sandwich" matrix, $H(\chi)J(\chi)^{-1}H(\chi)$, where $H(\chi) = -E(\nabla_2 p(\chi|Y))$ and $J(\chi) = Var[\nabla p(\chi|Y)]$ are obtained maximizing the composite posterior $p(\chi|Y)$. In the second, we adjust draws as

$$\tilde{\chi}^j = \hat{\chi} + V^{-1}(\chi^j - \hat{\chi}) \tag{52}$$

where $\hat{\chi}$ is the posterior mode, $V = C^T HC$ and $C = M^{-1}M_A$ is a semi-definite square matrix; $M_A^T M_A = HJ^{-1}H, M^T M = H$; M_A and M are obtained via a singular value decomposition [6].

The adjustment works well when the composite posterior has a unique maximizer and χ is well identified from the composite likelihood. As Canova and Sala (2009) have shown, uniqueness and identifiability may fail in a number of structural models. Although identification problems may be eased with a composite approach, see e.g. Canova and Matthes, 2019, multiple composite posterior modes can not be ruled out. Thus, we recommend users to report both standard and adjusted percentiles.

[6] Rather than finding H and J once, prior to running the algorithm, one could perform the adjustment adaptively, using $C(\phi^j|\phi^{j-1}, y)C(\phi|y)$ (see Ribatet et al, 2012, p. 826). Because MCMC draws are recursively centered, faster convergence is likely to occur, but at the costs of needing a numerical optimization at each iteration.

4.5 Composite posterior statistics

Once composite estimates of the common parameters are available, one can proceed with standard analysis using the "best" model as selected by the posterior of ω. Since ω_i measures the relative misspecification of model i and since the experimental evidence suggests that ω_i has properties similar to BMA when $y_{it} = y_{jt}$, for all i,j, such an approach is equivalent to comparing the marginal data densities, when one of the models is the minimizer KL divergence.

Because of the instabilities present in economic data and our Bayesian philosophy, we prefer to average the information contained in various models using posterior estimates and the posterior weights. Thus, rather than choosing one model, we pool them for inference. However, instead of using the posterior estimates based on each model being estimated individually, we use composite posterior estimates in the exercise.

Let $\tilde{y}_{t+l}$ be future values of the variables appearing in all models. Let $f(\tilde{y}_{t+l}|y_{it}, \phi_i)$ be the prediction of $\tilde{y}_{t+l}, l = 1, 2, \ldots$ in model i, given ϕ_i and let $f^{cl}(\tilde{y}_{t+l}|y_{1t}, \ldots, y_{Kt}, \chi) = \prod_{i=1}^{K} f(\tilde{y}_{t+l}|y_{it}, \phi_i)^{\omega_i}$ be a geometric pool of predictions, given y_t, the K models, and the parameters ϕ_i. Then

$$\begin{aligned} p(\tilde{y}_{t+l}|y_{1t}, \ldots, y_{Kt}, \omega_1, \ldots \omega_K) \propto & \int \ldots \int f^{cl}(\tilde{y}_{t+l}|y_{1t}, \ldots, y_{Kt}, \chi) \\ & p(\theta, \eta_1, \ldots, \eta_K|y_{1t}, \ldots, y_{Kt}, \omega_1, \ldots, \omega_K) d\theta d\eta_1 \ldots d\eta_K \\ = & \int \ldots \int \prod_i p(\tilde{y}_{t+l}, \phi_i|y_{it})^{\omega_i} d\theta d\eta_1 \ldots d\eta_K \end{aligned} \tag{53}$$

is the composite predictive density of $\tilde{y}_{t+l}$, given the data and the weights, and $p(\tilde{y}_{t+l}, \theta, \eta_i|y_{it})^{\omega_i} \equiv (f(\tilde{y}_{t+l}|y_{it}, \theta, \eta_i)p(\theta, \eta_1, \ldots \eta_K|\omega, y_{1t}, \ldots, y_{Kt}))^{\omega_i}$ is an "opinion" pool. In words, the composite prediction density is obtained by taking the joint density of future observations and of the parameters for each model, geometrically weighting them, and integrating the resulting expression with respect to the nuisance parameters' composite posterior. Note that the composite predictive density is not the true predictive density because the prediction function uses the composite prediction pool density rather than the true prediction density; and because the composite prediction density is integrated with respect to the composite posterior rather than the true posterior.

Depending on the investigator's loss function, one could compute (53) using the mode or the posterior mean of ω_i. One could also integrate (53) with respect to the marginal of ω, but given that in many applications it makes sense to condition on estimated ω's (which represent the posterior probability associated with each model), we believe (53) has stronger appeal.

$f(\tilde{y}_{t+l}|y_{it}, \phi_i)$ is straightforward to compute for each i since the models we consider have a linear (Gaussian) state space representation. Thus, (53) can be approximated by first generating draws from the composite posterior, computing the predictive density for each draw in each i, geometrically combining the predictions and, finally, averaging across draws of $(\theta, \eta_1, \ldots, \eta_K)$. The

problem of combining prediction densities is well studied in the literature (see e.g. Geweke and Amisano, 2011 or Del Negro et al., 2016). Two approaches are typically suggested: linear pooling, which leads to finite mixtures predictive densities such as BMA or static pools, and logarithmic pooling, which is what a composite predictive density produces. Logarithmic pooling generates predictive densities which are generally unimodal and less dispersed than linear pooling; and satisfy external Bayesianity, the property of being invariant to the arrival of new information (updating the components of the composite likelihood commutes with the pooling operator). Relative to standard pools of predictive densities, the composite predictive density uses the information in all models for estimation and to compute weights [7]. This may lead to differences, especially when models are misspecified in different ways and when the models feature different observables. There is an expanding literature dealing with nonlinear model combinations (see e.g. Gneiting and Rajan (2010) or Billio et al. (2013)). While such an approach is preferable if nonlinearities are suspected to exist, the logarithmic pooling implicit in (53) generally suffices for the purposes of reducing the misspecification of linear macroeconometric models.

In analogy with the prediction problem, one can compute statistics of interest by geometrically weighting the densities of outcomes and the composite posterior for the parameters. Take, for illustration, the computation of the responses for the subset of variables present in all models to a shock also present in all models. Given ϕ_i, responses to shock j for model i can be computed setting all other structural shocks to zero - which is reasonable given that the models considered are linear and shocks are uncorrelated. The density of outcome paths, computed randomizing ϕ_i from their posterior, is the impulse response of interest. The kernel of the composite posterior responses can then be computed analogously to (53), with the density of outcome paths replacing the predictive densities.

Counterfactuals are also easy to compute. Let $\bar{y}_{kt+l}$ be a selected path for the future values in the k-th element of $\tilde{y}_{t+l}$. Using $f(\bar{y}_{kt+l}|y_{it}, \epsilon^j_{it+l}, \phi_i)$ for submodel i, one can find the path of ϵ^j_{it+l} consistent with the assumed $\bar{y}_{kt+l}$. With this path one can then compute $f(\bar{y}_{k't+l}|y_{it}, \epsilon^j_{it+l}, \phi_i)$, for $k' \neq k$. Composite counterfactuals can be computed as in (53).

4.6 Interpretations

One can think of composite posterior analysis in at least three different ways. One is the sequential learning interpretation provided in Canova and Matthes (2019): the composite posterior kernel can be obtained in K stages via an adaptive sequential learning process, where the information contained in models whose density poorly relates to the observables is appropriately

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

downweighted.The prior for θ at each stage of the learning process depends on the relative weights assigned to the current and to all previous models and on their relative fit for θ. Thus, by examining the posterior distribution at each stage, one can identify the inferential contribution of each model for the posterior of the common parameters, reduce the set of relevant models, if that is of interest. Furthermore, one robustifies estimates and inference since, at each stage, a change in estimates reflect the contribution the cross-equations restrictions present in that model.

An alternative interpretation comes from noting that since the composite likelihood describes an "opinion" pool, where agents/models construct their likelihood using different pieces of information and different structures. Hence, the composite quasi-posterior distribution we construct and the composite statistics we consider can be interpreted as Bayesian pools of opinions, where each agent/model acts as a local Bayesian statistician expressing an opinion in the form of a posterior distribution on the unknown parameters, given a specific piece of information. The Bayesian pool weighs the posterior of each agents/models, based on their posterior weights. One can also show that the composite posterior is a "message" approximator, that is, it minimizes the KL divergence to the probabilistic opinions: $p_{CL} = argmin_p \sum_{i=1}^{K} \omega_i D(p||p_i)$ where $p_i \propto \pi(\psi_i)L(y_i|\psi_i)$ is the posterior of the parameters of model i. In words, it provides the best possible way to extract consensus among differing agents/models, see Roche (2016).

A final interpretation of our composite posterior estimators comes from noticing that they are special cases of quasi-Bayesian estimators. In this literature (see e.g., Marin, 2012; Bissiri et al., 2016; Scalone, 2018), one updates prior beliefs using a loss function which downplays some undesirable features of the likelihood. Different loss functions can be used for different purposes. A moment-based or a zero-one loss function are typical, because they provide estimators which reduce the inconsistencies of likelihood-based methods when misspecification is present. Seen through these lenses, the composite likelihood is a moment-based loss function, weighting the average of each model's scores. As Grunwald and van Ommen, 2017 or Baumeister and Hamilton, 2019 have noticed, a similar outcome can also be obtained by properly weighting different observations entering the likelihood. Rather than downweighting the likelihood of certain observations, our approach downweights the likelihood of models, while maintaining convexity of the composite objective function.

5 Two applications

We evaluate our framework of analysis in two applications. In the first we show how to robustify inference about the marginal propensity to consume (MPC) out of transitory income. In the second, we show how to shed light on the role of technology shocks as drivers of output fluctuations.

Table 4: Posterior distribution of ρ

Model	16th	50th	84th
BASIC	0.44	0.57	0.66
PRECAUTIONARY	0.90	0.91	0.91
RBC	0.41	0.52	0.63
ROT	0.46	0.56	0.65
LIQUIDITY	0.70	0.77	0.84
Unadjusted Composite	0.85	0.90	0.96
Adjusted Composite	0.80	0.87	0.95
Composite (without RBC)	0.80	0.85	0.91

5.1 Measuring the marginal propensity to consume

We consider five models commonly used in the literature to explain the dynamics of the MPC: the first is a standard permanent income specification; the others add aspects left out of the workhorse model. In the baseline model there is a representative agent with quadratic preferences, the real rate of interest is assumed to be constant, $(1 + r)\beta = 1$, income is exogeneous and features permanent and transitory components. The second model has similar features but preferences are exponential (in the spirit of Caballero, 1990). Because the variance of income shocks affects consumption decisions, precautionary savings matter and consumption is no longer a random walk. To make the model empirically interesting we allow the volatility of the two income components to be time dependent and assume a simple AR(1) for the log of the variance. In the third model, we make the real rate endogenous by considering a real business cycle structure featuring consumption-leisure choices, production requiring capital and labor, and a technological disturbance with transitory and permanent components. The representative agent has separable CRRA preferences. The fourth specification introduces agent heterogeneity: a fourth of the agents consume all of their current income, as in Gali et al., (2004). Preferences and constraints for the optimizing agents are as in the basic specification. The last model also has two types of agents, but one is liquidity constrained (in the spirit of Chah et al., 2006). This model retains exogenous income, a constant real rate equal to the inverse of the rate of time preference of the non-liquidity constrained agent but features a non-separable utility in non-durable and durable consumption goods (depreciating at the rate δ). Furthermore, constrained agents must finance a fraction of non-durable expenditure with accumulated assets. We make the liquidity constraint binding in the steady state by assuming that constrained agents are more impatient. We name the models: BASIC, PRECAUTIONARY, RBC, ROT, LIQUIDITY, respectively; their log-linearized conditions are in appendix D.

Although models feature different endogenous variables, we use aggregate real per-capita non-durable consumption (FRED name: A796RX0Q048SBEA), real per-capita income, (FRED name:

A067RO1Q156NBEA) and real per-capita value of assets (Household and non-profit organization total financial assets, FRED name: TFAABSHNO) as observables in estimation for all specifications - in the RBC model we equate real per-capita assets with per-capita capital of the representative agent. This choice of observables allows us to compare composite and BMA ranking of models and predictions. The sample size is 1980:1-2017:2 and all variables are quadratically detrended. Estimation is performed with MCMC techniques using the likelihood of each model or the composite likelihood. In the latter case, we restrict the persistence of the transitory income process ρ, which as seen in section 2, matters for the MPC_{y^T}, to be common across specifications. The prior for ω_i, i=1...5, is Dirichlet with mean equal to 0.20. The priors for all other parameters are proper but loose and truncated, when needed, to the region with economic interpretation.

Table 4 summarizes of the posterior features of ρ. The first five rows display single model percentiles; the sixth and seventh rows composite percentiles (unadjusted and adjusted). Although Cogley and Nason (1995) showed that income persistence in a RBC model is largely driven by TFP persistence, one may argue that TFP and exogenous income persistence are parameters with different economic interpretations. Thus, the eight row of table 4 presents composite percentiles when ρ is restricted to be common only across models featuring exogenous income.

For BASIC, ROT and RBC models the median estimate is around 0.55 and the envelope of the 68 percent posterior ranges is [0.40-0.65]; for the model with liquidity constraint the median estimate is 0.77 and significantly different from those of the first three models. Finally, in a model with precautionary motive, transitory income is highly persistent and very precisely estimated. The composite posterior estimate is also high: its median value (0.90) is close to the one obtain in the precautionary model (0.91), but the posterior range is larger, reflecting the heterogeneity of single model estimates. Eliminating the RBC model from the composite estimation leaves the posterior percentiles of ρ practically unchanged.

Why is the composite posterior median of ρ high? Figure 1, which presents the prior and the posterior of ω_i, shows that the precautionary model receives the highest weight in the composite pool. Thus, the fact that real rate is constant, that labor supply decision and heterogeneities are disregarded are less crucial when characterizing the MPC than leaving precautionary motives out. Since the weights are stable over time (estimates available on request), income uncertainty is not a dominant factor only in the post 2008 part of the sample.

Figure 2 presents dynamic estimates of MPC_{y^T}, computed cumulating over horizons consumption and output responses to transitory shocks, i.e., $MPC_y^T(l) = \frac{\sum_{j=1}^{l} c_{t+j}|e_t^T}{\sum_{j=1}^{l} y_{t+j}|e_t^T}$, $l = 1, 2, ...$, where $c_{t+j}(y_{t+j})$ is the response of real per-capita consumption (transitory income) at $t+j$, e_t^T is a transitory income shock, and l the horizon. In individual models, when ρ is estimated to be low, the profile of MPC_{y^T} is also low and, consistent with the discussion of section 2, the in-

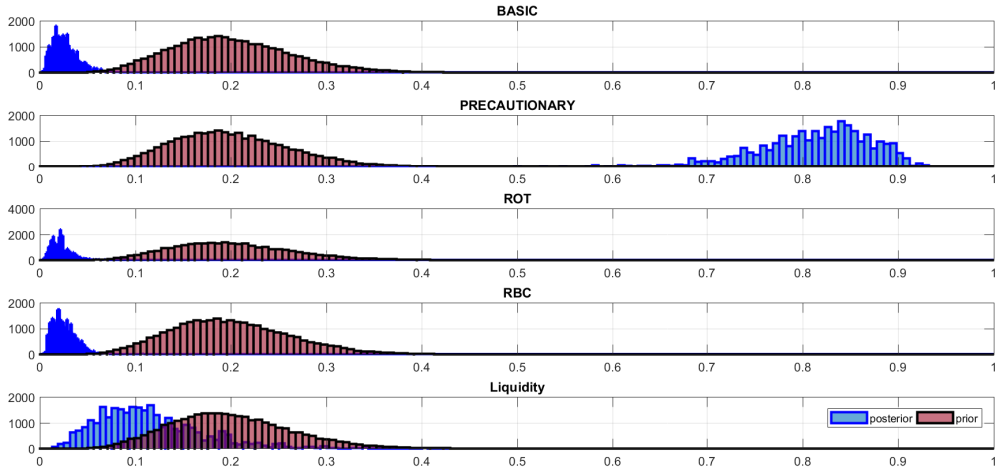

Figure 1: Prior and posterior for ω

stantaneous posterior estimates of MPC_{y^T} obtained with BASIC, RBC, and LIQUIDITY models are only around 0.05. Estimates increase somewhat at longer horizons but after two years the 68 percent range is still below 0.10. The instantaneous MPC is slightly higher in the ROT model (the median value is 0.25). Still, after two years the representative agent cumulatively consumes only 30 percent of the cumulative transitory income. With the PRECAUTIONARY model, the instantaneous posterior estimate of MPC_y^T is also higher. However, also with this specification, less than 25 percent of cumulatively transitory income is cumulatively consumed after two years. Hence, no matter what model one employs, MPC estimates suggest that at most 30 percent of cumulative income is cumulatively consumed at the two years horizon.

When the composite posterior estimate of ρ is employed, the instantaneous value of MPC_y^T generally increases but, with the exception of the ROT model, MPC_{y^T} estimates are still below 30 percent at the two years horizon. Thus, even when income is relatively persistent, rational consumers save the majority of their transitory income. Perhaps more interesting from our point of view is the fact that, when composite estimates of ρ are used, cross model differences in MPC_{y^T} estimates decrease considerably. For example, the time profile of MPC estimates in PRECAUTIONARY and RBC models (the models with the highest and the lowest median estimate of ω) are very similar and differences previously noted decrease substantially.

Rather than plugging composite posterior estimates in a model, one may choose to robustify inference by computing a composite MPC_y^T estimate, weighting the MPC_{y^T} of each model by the posterior ω_i. Figure 3 presents such a measure together with two other standard combinations: one constructed using BMA weights and one using naive, equal weights.

Composite and BMA estimates of MPC_{y^T} are similar, given that BMA puts all posterior weight on the PRECAUTIONARY model. Since posterior standard errors are also similar, the

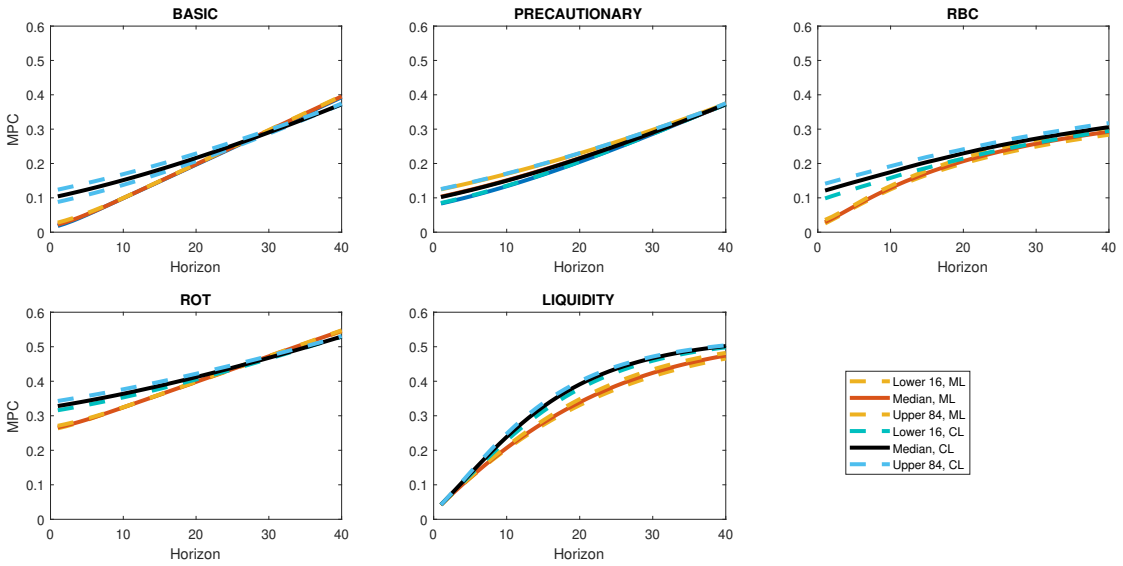

Figure 2: Likelihood and composite likelihood estimates of MPC_{y^T}

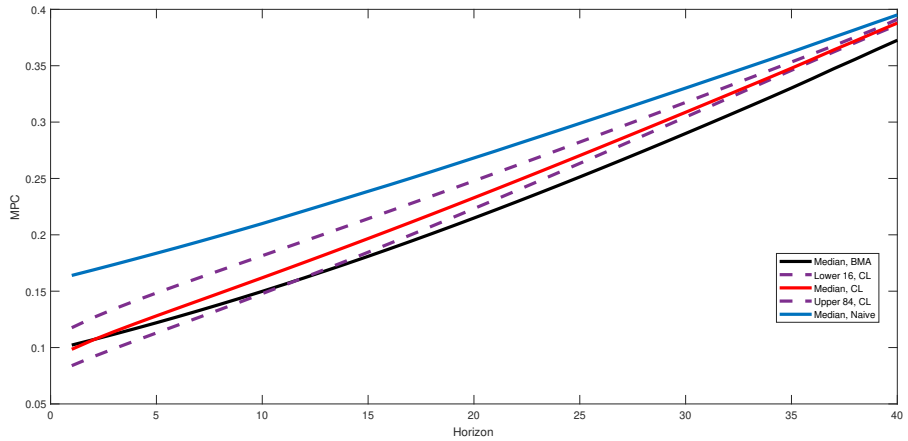

Figure 3: Composite, BMA and naive posterior estimates of the MPC

two measures give similar conclusions about the propensity to consume of US agents. The naive combination, instead, produces MPC_y^T estimates which are almost twice as large for the first two years, because the ROT model gets a much larger weight than in the other two combinations.

It is instructive to compute the average Kullback-Leibler (KL) divergence for detrended real per-capita output to have a further sense of the misspecification of the various models we entertain. Recall that while ω median estimates provide a small sample measure of relative misspecification, KL estimates are absolute measures and valid only in large samples. The PRECAUTIONARY model turns out to be the closest to the DGP (KL Divergence=0.0041) also according to this metric, and the other four models all feature KL divergence exceed 0.030. The composite model's KL divergence is larger than for the PRECAUTIONARY model (0.009), but substantially smaller than standard ad-hoc specifications. To illustrate we consider introduces habits in consumption

and a random disturbance to the budget constraint of consumers. In addition, measurement errors are added to all observables in estimation as this error drives a wedge between model implications and the dynamics of observable variables. The log-linearized conditions of this ad-hoc model are in appendix D. The KL divergence of the this habit persistence model is 0.323.

One may wonder whether the PRECAUTIONARY or the composite model should be employed for inference, given that the former is best in the KL metric. Our results indicate that once composite posterior estimates of ρ are used, MPC differences across models wash out. Thus, a good estimate of ρ is more important for thinking about MPC than the exact features a model displays. Nevertheless, one may worry about robustness to potential model switches and structural breaks. When this is a concern, composite posterior estimates of the MPC should be preferred.

In sum, our approach seems successful in many dimensions: it gives high posterior weight to the model with the lowest KL divergence; it reduces differences in MPC estimates across potentially misspecified models, making policy decisions less uncertain. Furthermore, composite inferences is close to BMA inference, despite the fact that the latter assumes that one of the models in the pool is the true one, and features lower KL divergence than an alternative ad-hoc specification.

5.2 The role of technology shocks for output fluctuations

The importance of technology shocks in accounting output fluctuations has been discussed for over 35 years with contrasting conclusions (see e.g. Kydland and Prescott, 1982 or Gali, 1999). Differences in the conclusions are due, in part, to specification choices and, in part, to the sample used. In general, larger models featuring dynamic evolution for the capital stock find a smaller role than smaller models featuring no or constant capital.

To show how a composite approach can shed light on the role of technology shocks we first estimate the medium scale New Keynesian (NK) model of Justiniano et al. (2010) (JPT henceforth) using post-1984 US data. We then pair it with the small NK model without capital of Herbst and Schorfheide (2015) (HS henceforth) and jointly estimate two models by composite methods, restricting the slope of the New Keynesian Phillips curve κ and the persistence of the stationary TFP shock ρ_z to be common. Clearly, there are other parameters which are common and could be restricted (e.g. Taylor rule coefficients). We chose to constrain only a few parameters to be common to highlight the stark differences obtained when estimating the JPT model in isolation or jointly with the HS model. The optimality conditions of the two models are in appendix E. Note that both models feature permanent and transitory technological disturbances; and we can approximate a RBC framework through prior parameter restrictions in the HS model. Thus, one can also think of our exercise as combining NK and RBC frameworks without having to worry about the poor fit that RBC models have for nominal variables.

We estimate the weights assuming that the two models are a-priori equally likely. Since we use different observable variables in estimation (output, inflation and the nominal rate for the HS model; output, inflation, the nominal rate, consumption, investment, hours and real wages for the JPT model), no comparison with BMA is possible. When the JPT model is estimated in isolation, estimates of κ and ρ_z are low (posterior means 0.02 and 0.14, standard deviations 0.0001 and 0.0041, respectively). The mean estimates are close to the point estimates reported by JPT (0.10 and 0.24)[8]. The quantitative differences are due to a different estimation sample. The posterior estimates obtained imply that technology shocks explain 30-40 percent of output fluctuations at typical business cycle horizons of 8-32 quarters (see figure 4). Mean estimates increase to $\kappa = 0.22$ and $\rho_z = 0.93$ when composite methods are used (standard deviations are 0.0023 and 0.0002, respectively). With composite posterior estimates technology shocks become the major source of output fluctuations at horizons greater than one year.

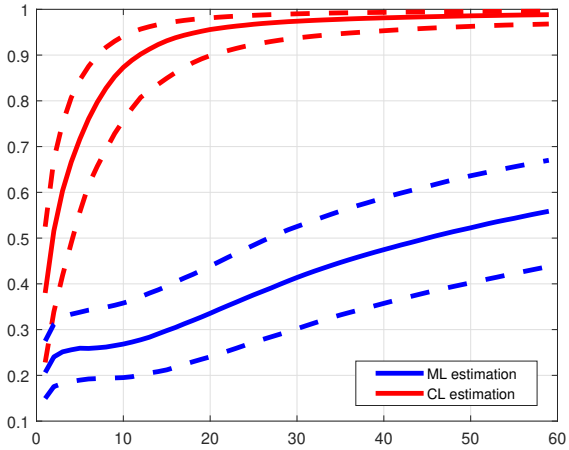

Figure 4: Fraction of output fluctuations due to TFP shocks, JPT model

How does one interpret these findings? First, notice that the HS model receives a-posteriori higher weight (mean estimate for ω is 0.63 and standard deviation 0.0003). Second, in the HS model technology shocks enter only the Euler equation, while in the JPT model they affect several equations. Thus, when the latter is estimated in isolation, technology shocks must propagate in a way that helps to fit well the dynamics of a number of endogenous variables. If the JPT model is misspecified, in particular, in equations other than the Euler equation, pairing it with the HS model relaxes incorrect cross equations restrictions the model imposes. Because the JPT model has been designed to give monetary shocks their best chance to explain output and inflation fluctuations, it is likely that the mechanics of transmission of other shocks are misspecified. The fact that the HS model has a higher ω estimate and that potentially incorrect cross equation

[8]The value of κ is obtained using estimates of the parameters they report.

restrictions are relaxed imply that posterior estimation in the JPT model moves to a region of the parameter space where nominal rigidities are smaller (the price stickiness mean estimate drops from 0.66 to 0.47), real rigidities are larger (the investment adjustment cost parameter mean estimate increases from 1.54 to 2.57) and demand shocks less persistent (the mean value of the persistence of preference shocks drops from 0.76 to 0.23), all of which make technology disturbances more important for output fluctuations. These conclusions remain valid when HS is restricted to mimic a RBC model.

To know whether composite inference should be trusted, we compute the KL divergence for output and inflation for the JPT model (using posterior estimates) and the composite pool. Because misspecification is roughly the same (average KL is 0.025 for the composite model and 0.021 for the JPT model), our results indicate the JPT model possesses multiple posterior modes featuring different mechanics of structural transmission but similar KL divergence.

Clearly, additional work is needed to more comprehensively explore the posterior of the JPT model but our evidence warns about dismissing technology shocks as major sources of output fluctuations in medium scale New Keynesian models.

6 Conclusions and implications for practice

This paper proposes a new approach to deal with the inherent misspecification of the current generation of DSGE models. We consider a set of potentially misspecified models, geometrically combine their likelihood functions, and perform posterior estimation using the composite likelihood. The composite likelihood shrinks individual likelihood estimates of the common parameters toward a weighted average of all other models' estimates, while leaving untouched estimates of idiosyncratic parameters. Thus, composite estimation guards against misspecification by requiring estimates of the common parameters to be consistent with the structure present in all models. We highlight the properties of our approach and relate it to existing methodologies.

We describe a MCMC approach to draw sequences from the composite posterior distribution, show how to adjust the MCMC percentiles to produce posterior credible sets with the right asymptotic coverage, highlight how to construct composite posterior statistics, such as impulse responses or counterfactuals, and discuss how posterior weights inform us about the relative misspecification of the models entering the pool.

We use the methodology to estimate the marginal propensity to consume out of transitory income, and to evaluate of the role of technology shocks for output fluctuations. MPC estimates are generally low when models are estimated separately but significantly increase when models are jointly estimated. Composite posterior and BMA MPC estimates are similar and lower than a naive combination of individual MPC estimates. Furthermore, the composite model is closer

to the process generating the data than a standard ad-hoc model with habit in consumption. Technology shocks explain about one-third of output fluctuations in a standard medium scale NK model but their importance increases when such a model is paired with a less restricted and smaller scale model without capital.

We conclude with some practical suggestions to potential users and highlight a few issues which need be developed in future research. First, to make the approach meaningful the models entering the composite likelihood should capture different aspects disregarded (or mis-represented) in the baseline specification. Gains from composite estimators depend on a careful selection of models entering the pool. Second, when a researcher perceives that the models are economically incompatible, making parameters with the same name different economic objects, the composite likelihood can still be employed since if $\theta = \emptyset$, the approach produces likelihood estimates, model by model. Third, while the methodology has the potential to reduce misspecification and to improve inference, given existing models, it is not a substitute for having better models. Section 5 shows how it can be used to gauge which missing features should be included in a benchmark model, and how conclusions could be altered when estimation is restricted in a meaningful way. Fourth, apart from misspecification issues, the approach has a number of other benefits relative to likelihood-based estimation of the structural parameters (see Canova and Matthes, 2019). For example, when a large scale model is available, the composite likelihood constructed using blocks of equations has shape and properties which are similar to those of the likelihood of the full model, without the numerical difficulties. Thus, the approach is not only useful to examine in which direction a model should be improved. It also provides a way to estimate the larger scale models one is likely to build after the initial experimentation. Fifth, although we focus on linearized models, one can also combine the likelihoods of models perturbed at higher order. We expect the gains to remain also in these more complicated frameworks. Finally, by treating data subsamples as different models that are combined for inference via the composite likelihood, the approach is suited to deal with structural time varying coefficients models, which are complicated to interpret with standard likelihood-based technology, see e.g. Canova et al. (2020).

One question that needs careful attention is one of overfitting. Standard models with ad-hoc additions may lead to overfitting, making their out-of-sample performance poor. One relevant question is whether our approach faces a similar problem. While we have not performed out-of-sample checks, the literature on model combination suggests that it is unlikely to be the case because shrinkage estimates give superior performance to standard estimates; and model combinations dominate single model forecasts in the presence of even mild instabilities in the data generating process, see e.g. Aiolfi et al. (2010). We plan to investigate the issue in future work.

References

Alessi, R. and A. Lusardi (1997). Consumption, Savings and Habit formation. *Economic Letters*, 55, 103-108.

Andrews, D. (1999). Estimation when a parameter is on the boundary. *Econometrica*, 67, 1341-1383.

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

Batthacharya, A., Pati, D., Pillai, N and D. Dunson (2012). Bayesian Shrinkage. https://arxiv.org/pdf/1212.6088.pdf

Barnichon, R. and C. Brownlees (2019). Impulse response estimation by smooth local projection. *Review of Economics and Statistics*, 101, 522-530.

Baumeister, C. and J. D. Hamilton (2019). Structural interpretation of vector autoregressions with incomplete identification: revisiting the role of oil supply and demand shocks. *American economic Review*, 109, 1873-1910.

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

Bissiri, P. G., Holmes, C.C. and S.G. Walker (2016) A general framework for updating belief distributions. *Journal of the royal Statistical Society*, Ser B, 78, 1103-1130.

Caballero, R. (1990) .Consumption puzzles and precautionary savings. *Journal of Monetary Economics*, 25, 113-136.

Canova, F. (2007). Methods for Applied Macroeconomic Research. Princeton University Press, Princeton, NJ.

Canova, F., F. Ferroni and C. Matthes (2020). Detecting and Analyzing the Effects of Time-Varying Parameters in DSGE Models, *International Economic Review*, Vol 61-1

Canova, F. and C. Matthes (2019). Solving computational and estimation problems in dynamic structural models: a composite likelihood approach, manuscript, https://sites.google.com/view/fabio-canova-homepage/home/current-research.

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

Carroll, C., Slacalek, J. and K. Tokouka (2017). The distribution of wealth and the marginal propensity to consume. *Quantitative Economics*, 8, 977-1020.

Chah, E., Ramey, V. and R. Starr (1995). Liquidity constraint and intertemporal consumption optimization: theory and evidence from durable goods. *Journal of Money, Credit and Banking*, 27, 272-287.

Chan, J., Eisenstat, E., Hou, C. and G. Koop (2018). Composite likelihood methods for large BVAR with stochastic volatility, *Journal of Applied Econometrics*, 33, 509-533.

Cheng, X and Z. Liao (2015). Select the valid and relevant moments: An information-based LASSO for GMM with many instruments. *Journal of Econometrics*, 186, 443-464.

Chari, V., Kehoe, P. and E. McGrattan (2007). Business cycle accounting. *Econometrica*, 75, 781-836.

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

Claeskens, G., and N. L. Hjort (2008). Model selection and model averaging. Cambridge University Press, Cambridge, UK.

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

Cogley, T. and Nason, J (1995). Output dynamics in RBC models. *American Economic Review*, 85, 492-515.

Cogley, T. and A. Sbordone (2008). Trend inflation, indexation, and inflation persistence in the New Keynesian Phillips curve. *American Economic Review*, 98, 2101-2126.

Cover, T. and J. Thomas (2006). Elements of information theory. Wiley, New York, NY.

Curdia, V. and R. Reis (2010). Correlated disturbances and US business cycles. Columbia University,manuscript.

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

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

Del Negro, M. and F. Schorfheide (2009). Monetary Policy analysis with potentially misspecified models. *American Economic Review*, 99, 1415-1450.

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.

Den Haan, W. and T. Drechsel (2018). Agnostic structural disturbances (ASDS): detecting and reducing misspecification in empirical macroeconomic models. CEPR working paper 13145.

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

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

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

Gali, J. (1999) Technology, employment, and the business cycle: Do technology shocks explain aggregate fluctuations? *American Economic Review*, 89, 249-271.

Gali, J., Lopez Salido, D. and J. Valles (2004). Rule of thumb consumers and the design of interest rate rules. *Journal of Money, Credit and Banking*, 36, 739-764.

Geweke, J. and G. Amisano (2011). Optimal prediction pools. *Journal of Econometrics*, 164, 130-141.

Giacomini, R. and T. Kitigawa (2017). Robust Inference in Partially Identified VARs. UCL manuscript.

Grunwald, P. and T. van Ommen (2017). Inconsistency of Bayesian inference for misspecified linear models, and a proposal for repairing it. *Bayesian Analysis*, 12: 1069-1103.

Gneiting, T. and R. Rajan (2010). Combining predictive distributions. GE research manuscript.

Hansen, L. and T. Sargent (2008). Robustness. Princeton University Press, Princeton, NJ.

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

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

Kocherlakota, N. (2007). Model fit and model selection. *Review, Federal Reserve Bank of St. Louis*, July, 349-360.

Kydland, F. and E. Prescott (1982). Time to build and aggregate fluctuations.*Econometrica*, 50, 1345-1370.

Inoue, A., Rossi, B. and C. Kuo (2017). Identifying sources of model misspecification, forthcoming, *Journal of Monetary Economics*.

Ireland, P. (2004). Taking a model to the data. *Journal of Economic Dynamics and Control*, 28, 1205-1226.

Johnson, D., Parker, J. and N. Souleles (2006). Household expenditure and the tax rebate of 2001. *American Economic Review*, 96, 1589-1610.

Justiniano, A., Primicieri, G. and A. Tambalotti (2010). Investment shocks and business cycles. *Journal of Monetary Economics*, 57, 132-145.

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

Marin, J., P., Pudlo, C. Robert and R. Ryder (2012). Approximate Bayesian computational methods. *Statistics and Computing*, 22, 1167?1180.

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

Newey, W. K., and R. J. Smith (2004). Higher Order Properties of GMM and Generalized Empirical Likelihood Estimators. *Econometrica*, 72, 219?255.

Parker, J. , Souleles, N, D. Johnson, and R. McClelland (2013). Consumer Spending and the Economic Stimulus of 2008. *American Economic Review*, 103, 2530-2553.

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

Ragusa, G. (2011). Properties of minimum divergence estimators. *Econometric Review*, 30, 406-456.

Ravn, M., Schmitt-Grohe, S. and M. Uribe (2006). Deep Habits. *Review of Economic Studies*, 73, 195-218.

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

Roche, A. (2016). Composite Bayesian inference. CHUV, Siemens Healthcare, EPFL manuscript.

Smets, F. and R. Wouters (2007). Shocks and frictions in US business cycles: a Bayesian DSGE approach. *American Economic Review*, 97, 586-606.

Scalone, V. (2018). Estimating nonlinear DSGEs with approximate Bayesian computations: an application to the zero lower bound, Banque de France, manuscript.

Thryphonides, A. (2016). Robust inference for dynamic economies with an application to financial frictions. Humboldt University manuscript.

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

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

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

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