DISCUSSION PAPER SERIES

DP13511

DEALING WITH MISSPECIFICATION IN STRUCTURAL MACROECONOMETRIC MODELS

Fabio Canova and Christian Matthes

MONETARY ECONOMICS AND FLUCTUATIONS

ISSN 0265-8003

DEALING WITH MISSPECIFICATION IN STRUCTURAL MACROECONOMETRIC MODELS

Fabio Canova and Christian Matthes

Discussion Paper DP13511
Published 05 February 2019
Submitted 28 January 2019

Centre for Economic Policy Research
33 Great Sutton Street, London EC1V 0DX, UK
Tel: +44 (0)20 7183 8801
www.cepr.org

DEALING WITH MISSPECIFICATION IN STRUCTURAL MACROECONOMETRIC MODELS

Abstract

We consider a set of potentially misspecified structural models, geometrically combine their likelihood functions, and estimate the parameters using composite methods. Composite estimators may be preferable to likelihood-based estimators in the mean squared error. Composite models may be superior to individual models in the Kullback-Leibler sense. We describe Bayesian quasi-posterior computations and compare the approach to Bayesian model averaging, finite mixture methods, and robustness 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: C13, C51, E17

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

Fabio Canova - fabio.canova@bi.no
Norwegian Business School and CEPR

Christian Matthes - christian.matthes@rich.frb.org
Federal Reserve Bank of Richmond

Dealing with misspecification in structural macroeconometric models

Fabio Canova,
Norwegian Business School *

Christian Matthes,
Federal Reserve Bank of Richmond †

January 20, 2019

Abstract

We consider a set of potentially misspecified structural models, geometrically combine their likelihood functions, and estimate the parameters using composite methods. Composite estimators may be preferable to likelihood-based estimators in the mean squared error. Composite models may be superior to individual models in the Kullback-Leibler sense. We describe Bayesian quasi-posterior computations and compare the approach to Bayesian model averaging, finite mixture methods, and robustness 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.

---

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

†We thank C. Michelacci, five anonymous referees, M. Plagborg-Moller, G. Koop, B. Rossi, T. F. Schorfheide, Zha, J. Linde and 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 section), 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 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. The views presented in this paper do not reflect those of the Federal Reserve Bank of Richmond, or the Federal Reserve system.

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 issues of interest. While specifying an incomplete model to explain the data is acceptable, for example, when qualitatively highlighting a mechanism which could be present in the data, misspecification becomes an issue when one want 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 model computationally intractable and difficult to interpret. 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 is to use a structural model with ad-hoc reduced form features. However, in hybrid models of this type it is often hard to distinguish the relative importance of structural vs. ad-hoc features in matching the data, making policy conclusions and 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. Perhaps because of this problem, the recent econometric literature dealing with misspecification (see e.g. Cheng and Liao, 2015; Thryphonides, 2016) does not employ the likelihood in the estimation process 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 decisions are often formulated on the basis of potentially misspecified models.

This paper proposes a new approach to deal with the inherent misspecification of DSGE models. Rather than enriching a model with structural or ad-hoc features, we jointly consider a finite set of potentially misspecified models, geometrically combine their likelihood functions, and estimate the parameters using the resulting composite likelihood. With such an objective function, parameters common across models are estimated using the cross equation restrictions present in all specifications. Thus, the composite likelihood guards against misspecification by requiring estimates of the common parameters to be consistent with the structures 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 (2016), Canova and Matthes (2017) and Chan et al. (2018). 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; however, in our setup the DGP is unknown, the models entering the composite likelihood are misspecified, and the weights are random variables.

The Bayesian setup we work with is related to the quasi-Bayesian estimation literature (see e.g. Kim, 2002, Marin et al. 2012, Bissiri et al., 2016, 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, as in the shrinkage and smoothness prior literature, we employ additional information to regularize parameter estimates. The posterior weight of a model plays an important 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 to share the same observables; 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. Finally, our setup quantifies the uncertainty in the weight estimates. To the best of our knowledge, this can not be done in BMA.

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 models may feature different observables, the models Del Negro and Schorfheide and Waggoner and Zha consider 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 and show how to adjust the percentiles to insure the right asymptotic coverage. We show with examples why composite estimators may be preferable to likelihood-based estimators and discuss how the posterior weights inform us about the relative misspecification of the models entering the composite pool. Finally, we demonstrate how to combine models and the composite posteriors for inference. While some researcher may want to use posterior weights to select a model, we prefer to robustify inference using the composite predictions of all available models.

We apply the methodology to the estimation of 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 and differences across models are significantly reduced. We show that problematic features of the basic specification such as quadratic preferences, separability, exogenous real rate, lack of production, consumer heterogeneities are irrelevant to characterize the MPC, and that composite and BMA estimates of the MPC are similar.

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 then pair such a model with a smaller scale New Keynesian model without capital, jointly estimate the slope of the Phillips curve and the persistence of technology shocks, and find that the share of output fluctuations explained by technology shocks substantially increases at all horizons. This change occurs because the smaller scale model receives high a posteriori 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, and these make technology shocks more important for output fluctuations.

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 approaches to make the estimation results more credible. Section 3 presents our method. Section 4 describes a MCMC procedure to draw sequences for the parameters and for the weights from their quasi-posterior distribution and explain 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 appendices contain relevant technical material.

2 Estimating misspecified structural models

Suppose a researcher is interested in measuring the MPC out of transitory income. Interest 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 starts with an off-the-shelf permanent-income, life-cycle model, solves it, and derives implications for the MPC. For example, in a representative agent model with quadratic preferences, constant interest rate, when $\beta(1+r)=1$, and the exogenous labor income has permanent and transitory components, the decision rules are:

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

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 consumption decision rule. 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 these predictions, 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.

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 consumer i characteristics 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) imposing the cross equation restrictions the decision rules imply (in particular, the facts 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.

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 interest 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; some agents may 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 does she guard herself against misspecification?

An obvious way is to estimate a more general model which includes potentially missing features, allows for general equilibrium effects on income and real interest rates, uses flexible functional forms for preferences and technologies, and allows for relevant heterogeneities. While feasible, it is computationally demanding to estimate large scale models, and identification issues are likely to lead to interpretation problems. 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 foundation of such a mechanism are still debatable (one exception is Ravn et al., 2006). With habit, the decision rules of our workhorse model are:

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

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

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

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

where h is the habit persistence parameter. Thus, habit helps to account for serial correlation in consumption and for the predictability of current consumption, given permanent wealth w_t; and makes the serial correlation properties of consumption disconnected from those of income. Nevertheless, ad-hoc features make the model less structurally interpretable and some ad-hoc additions may not lead to better models.

Adding non-structural 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 non-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 adds-on 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 take a step in acknowledging model misspecification, 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 they 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 adds-on are serially correlated and statistically important and no respecification of the structure is attempted.

3 A composite likelihood approach to misspecification

The approach we suggest is simple. 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 start from the assumption that a number of possibly misspecified structural models are available to investigate the question of interest. Models may differ in the assumptions they make, in the frictions they feature, or in the transmission mechanisms they emphasize, but they are theoretically relevant and have bearing to the phenomenon under investigation. We also assume they are sufficiently heterogeneous so that the information they provide does not entirely overlap. We construct the likelihood function of each model and geometrically combine them. The resulting composite likelihood is either maximized or used as an input for quasi-posterior analysis.

When the composite likelihood is used, common parameters are estimated using the cross-equation restrictions present in all the models. Model specific parameters are instead estimated using the cross-equation restrictions of that model, conditional on the estimate of the common parameters. Cross model restrictions robustify estimation in the presence of misspecification because they shrink single model estimates in a natural way. In addition, as long as the models are sufficiently different, the composite likelihood de-emphasizes idiosyncratic elements which are at odds with the data, making the composite model less misspecified than its components.

Let the DGP for y_t be represented by a density $F(y_t, \psi)$, where ψ is a parameter vector. The available models are indexed by $i = 1, \ldots, K$. Assume that each produces a density $f_i(y_{it}|\theta, \eta_i)$ for the observables y_{it} 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. Each model features a vector of parameters $\phi_i = [\theta, \eta_i]'$, where θ are common across specifications and η_i are model specific. We assume that there is no ϕ_i such that $f(y_{it}|\phi_i) = F(y_t, \psi),\ \forall i$. Investigators are typically free to choose what goes in θ and η_i: even though a parameter may appear in all K models, a researcher may decide to treat it as model specific because, for example, models are too incompatible with each other. Trivially, if $\theta = \emptyset$, and our approach produces likelihood-based estimates of $\phi_i = \eta_i$, model by model. Given a vector of weights, $0 < \omega_i < 1, \sum_i \omega_i = 1$, the composite likelihood is

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

In a traditional composite likelihood approach, ω_i are fixed quantities, chosen by the investigator. Here we work with random weights and the quasi-posterior of ω_i gives us a measure of the relative

misspecification of model i, given $(y_{1t}, \ldots, y_{Kt})$.

Before we discuss the details of the procedure, we present an example to show what kind of estimators the approach produces and to provide intuition for why estimation and inferential gains may emerge when misspecification is present.

3.1 Why are composite estimators preferable?

Suppose we have two misspecified structural models (A, B), with parameters $\phi_A = (\theta, \eta_A), \phi_B = (\theta, \eta_B)$ and implications for (y_{At}, y_{Bt}), where y_{At} may be different from y_{Bt} and T_A from T_B. Assume that $f(y_{it}|\theta, \eta_i),\ i = A, B$ are produced by the decision rules:

$$y_{it} \quad = \quad \rho_i y_{it-1} + \sigma_i e_t \tag{10}$$

where e_t and u_t are iid(0,I). For the sake of illustration, let $\rho_B = \delta\rho_A, \sigma_B = \gamma\sigma_A$, $\delta \neq 0, \gamma \neq 0$ and assume that y_{At} and y_{Bt} are scalars. Thus $\theta = (\rho_A, \sigma_A^2)$, $\eta_B = (\delta, \gamma^2)$ are (nuisance) parameters specific to model B, and $\eta_A = \emptyset$. The (normal) log-likelihood functions are:

$$\log L_i \quad \propto \quad -T_i \log \sigma_i - \frac{1}{2\sigma_i^2} \sum_{t=1}^{T_i} (y_{it} - \rho_i y_{it-1})^2 \tag{11}$$

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

$$\log CL = \omega \log L_A + (1 - \omega) \log L_B \tag{12}$$

Maximization of (12) with respect to θ leads to:

$$\rho_{A,CL} \quad = \quad (\sum_{t=1}^{T_A} y_{At-1}^2 + \zeta_2 \sum_{t=1}^{T_B} y_{Bt-1}^2)^{-1} (\sum_{t=1}^{T_A} y_{At} y_{At-1} + \zeta_1 \sum_{t=1}^{T_B} y_{Bt} y_{Bt-1}) \tag{13}$$

$$\sigma_{A,CL}^2 \quad = \quad \frac{1}{\xi} (\sum_{t=1}^{T_A} (y_{At} - \rho_A y_{At-1})^2 + \frac{1 - \omega}{\omega\gamma^2} \sum_{t=1}^{T_B} (y_{Bt} - \delta\rho_A y_{Bt-1})^2) \tag{14}$$

where $\zeta_1 = \frac{1-\omega}{\omega} \frac{\delta}{\gamma^2}$, $\zeta_2 = \frac{1-\omega}{\omega} \frac{\delta^2}{\gamma^2} = \zeta_1 \delta$, $\xi = (T_A + T_B \frac{1-\omega}{\omega} \log(\gamma))^{-1}$ is the effective sample size. Maximization of (12) with respect to η_B yields

$$\delta_{CL} = (\sum_{t=1}^{T_B} \rho_{A,CL}^2 y_{Bt-1}^2)^{-1} (\sum_{t=1}^{T_A} \rho_{A,CL} y_{Bt} y_{Bt-1}) \quad (15)$$

$$\gamma_{CL}^2 = \frac{1}{T_B \sigma_{A,CL}^2} \sum_{t=1}^{T_B} (y_{Bt} - \rho_{A,CL} y_{Bt-1})^2 \quad (16)$$

As (13)-(14) show, θ_{CL} combines information present in y_{At} and y_{Bt}. The formulas are similar to those i) obtained in least square problems with uncertain linear restrictions (Canova, 2007, Ch.10); ii) derived using a prior-likelihood approach, see e.g. Lee and Griffith (1979); and iii) implicitly produced by a DSGE-VAR setup (see Del Negro and Schorfheide, 2004), where T_B observations are added to the original T_A data points. As (15)-(16) indicate, model B parameters (δ, γ^2) are estimated using only model B information, conditional on $(\rho_{A,CL}, \sigma_{A,CL})$. In general, they will differ from likelihood estimates obtained with y_{Bt} only, because $\theta_{CL} \neq \theta_{B,ML}$.

Thus, when the decision rules feature an autoregressive structure, the composite likelihood shrinks the information in y_{At} by the information in y_{Bt} and the amount of shrinkage depends on (γ, δ, ω). The higher ω and γ are, the less important y_{Bt} information is. Similarly, the smaller is δ, the lower will be the shrinkage toward model B information. Thus, when estimating common parameters, the composite likelihood gives larger importance to data generated by a model with higher persistence and lower standard deviation because higher serial correlation implies important low frequency information; and lower standard deviation lower noise.

When an array of models is available, composite likelihood estimates of θ will be constrained by the structure present in all models. For example, equation (13) now becomes

$$\rho_A = (\sum_{t=1}^{T_A} y_{At-1}^2 + \sum_{i=1}^{K-1} \zeta_{i2} \sum_{t=1}^{T_i} y_{it-1}^2)^{-1} (\sum_{t=1}^{T_A} y_{At} y_{At-1} + \sum_{i=1}^{K-1} \zeta_{i1} \sum_{t=1}^{T_i} y_{it} y_{it-1}) \quad (17)$$

where $\zeta_{i1} = \frac{\omega_i}{\omega_A} \frac{\delta_i}{\gamma_i^2}$, $\zeta_{i2} = \zeta_{i1} \delta_i$. Thus, the composite likelihood robustifies estimation, because θ estimates are required to be consistent with the cross-equation restrictions of all models.

Two further aspects are worth some discussion. Since y_{At} and y_{Bt} could be different series, the procedure can be used to estimate parameters appearing in models featuring different observables (see section 5.2, for an example). y_{At} and y_{Bt} may also be the same series but with different levels of aggregation (say, aggregate vs. individual consumption). Furthermore, since T_A and T_B 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_A and T_B may also

represent two samples for the same vector of observables (e.g., before and after a financial crisis). Baumeister and Hamilton (2015) propose to downweight older information when conducting posterior inference. Their procedure mimics a composite estimator where data for the earlier part of the sample, say y_{At}, is weighted less than data in the later part of the sample, say y_{Bt}.

Because of the shrinkage nature of composite estimators, we expect them to do well in mean square error (MSE) relative to maximum likelihood estimators. We highlight this feature for the composite estimator of ρ_A. Algebraic manipulations give $\rho_{A,CL} = \chi\rho_{A,ML} + \frac{1-\chi}{\delta}\rho_{B,ML}$ where $\rho_{i,ML} = \frac{\sum_{t=1}^{T_i} y_{it}y_{it-1}}{\sum_{t=1}^{T_i} y_{it-1}^2}$; $i = A, B$ $\chi = \frac{\sum_{t=1}^{T_A} y_{At-1}^2}{\sum_{t=1}^{T_A} y_{At-1}^2 + \frac{\omega_B\delta^2}{\omega_A\gamma^2}\sum_{t=1}^{T_B} y_{Bt-1}^2} = \frac{1}{1+\frac{\omega_B\delta^2 var(\rho_{A,ML})}{\omega_A var(\rho_{B,ML})}}$ and $var(\rho_{A,ML}) = \sigma_A^2(\sum_t y_{At-1}^2)^{-1}; var(\rho_{B,ML}) = \gamma^2\sigma_A^2(\sum_t y_{Bt-1}^2)^{-1}$. Using (10)-(**??**) we have:

$$\rho_{A,CL} \quad = \quad \rho_A + \chi B_A + \frac{1-\chi}{\delta}B_B \tag{18}$$

where $B_i = \frac{\sum_{t=1}^{T_i} e_{it}y_{it-1}}{\sum_{t=1}^{T_i} y_{it-1}^2}$ is the bias in the ML estimator using y_{it}. To insure, e.g., $MSE_{CL} < MSE_{A,ML}$ we need $(1-\chi^2)EB_A^2 - \frac{(1-\chi)^2}{\delta^2}EB_B^2 - \frac{\chi(1-\chi)}{\delta}EB_AB_B > 0$. Suppose the biases in $\rho_{A,ML}, \rho_{B,ML}$ are independent. Then, the composite estimator is preferable if

$$\delta^2 \quad > \quad \frac{EB_B^2}{EB_A^2} - 2\frac{\omega_A}{\omega_B}\frac{var(\rho_{B,ML})}{var(\rho_{A,ML})} \tag{19}$$

(19) links the persistence of y_{Bt} to the relative weights, the relative biases, and the relative variances of the maximum likelihood estimators of two models. The higher is the bias of the maximum likelihood estimator obtained with y_{Bt}, the higher should δ be for the CL estimator to be MSE superior. Similarly, the higher is the variability of the ML estimator constructed with y_{At}, for a given ratio of ω weights, the lower needs to be δ for the CL estimator to dominate. When $\omega_B = \omega_A$ and the ML estimators have similar biases, $\delta^2 > 1 - 2\frac{var(\rho_{B,ML})}{var(\rho_{A,ML})}$ is sufficient for the CL estimator to be MSE superior, a condition easy to check in practice.

Another interesting case is when the biases of models A an B are negatively correlated, as in the experimental design of section 3.3. Here MSE improvements can be obtained under milder restrictions. For example, when y_{tB} is a noisy measure of y_{tA}, i.e. $\delta = 1$, a CL estimator improves the MSE as long as the bias in the ML estimator computed with y_{tB} is not too large:

$$EB_B^2 \quad < \quad \frac{1-\chi}{1+\chi}EB_A^2 - \frac{\chi}{1+\chi}EB_AB_B \tag{20}$$

In general, whenever the models we combine have biases which are negatively correlated we expect MSE improvements, even when the persistence properties of y_{At} and y_{Bt} are similar.

A common device to measure the degree of misspecification of a model is the Kullback-Leibler (KL) divergence. If y_t has been generated by a density $F(y_t, \psi)$ and the researcher uses the density $f_i(y_t, \phi_i), i = 1, \ldots, K$ for the analysis, the KL divergence is

$$KL_i(y, \psi, \phi_i) = \sum_{j=1}^{N} F(y_j, \psi) * \log(\frac{f_i(y_j, \phi_i)}{F(y_j, \psi)}) \tag{21}$$

which it is interpreted as the bits of information lost in characterizing y using f_i rather than F. The KL divergence can be used to rank misspecified models. In fact, if f_A and f_B are available and $KL_A(y, \phi_1, \psi) > KL_B(y, \phi_2, \psi)$, then f_B is less misspecified than f_A. Because the composite model averages different misspecified structural models, we expect it to reduce the misspecification of the original models. To check if this is the case in our Bayesian setup, 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 posterior 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 posterior of $\chi = (\phi_1, \ldots, \phi_K, \omega_1, \ldots, \omega_K)$. Section 3.3 provides evidence on the performance of composite estimators and composite models for some interesting DGPs.

When ω_i is a random variable, its posterior mode inform us about the relative misspecification of the models entering the composite likelihood. Let $y_{At} = y_{Bt}$, let $p(\omega) \propto \omega^{\alpha_A - 1}(1 - \omega)^{\alpha_B - 1}$, where α_A, α_B are known, and let the prior for $(\rho_A, \sigma_A^2, \delta, \gamma^2)$ be diffuse. The composite posterior kernel of ω, conditional on $(\rho_A, \sigma_A^2, \delta, \gamma^2)$ is $CP(\omega|\rho_A, \sigma_A^2, \delta, \gamma^2) = (L_A^{\omega} L_B^{1-\omega})\omega^{\alpha_A - 1}(1 - \omega)^{\alpha_B - 1}$. Taking logs and maximizing we have

$$\log L_A - \log L_B + \frac{(\alpha_A - 1)}{\omega} - \frac{(\alpha_B - 1)}{1 - \omega} = 0 \tag{22}$$

This is a quadratic equation in ω and the relevant solution is $0 < \omega_1 < 1$. Total differentiating (22) one finds that ω_1 is increasing in $\log L_A - \log L_B$. Completing the square terms of the likelihoods, and conditioning on the mode estimators of $(\rho_A, \sigma_A^2, \delta, \gamma^2)$, one obtains

$$\log L_A - \log L_B \propto -\frac{1}{2\sigma_A^2} \sum_{t=1}^{T_A} (y_{At|t-1} - \rho_A y_{At-1})^2 + \frac{1}{2\gamma^2 \sigma_A^2} \sum_{t=1}^{T_B} (y_{Bt|t-1} - \rho_A \delta y_{Bt-1})^2| \tag{23}$$

where $y_{it|t-1}$ is the predictor of y_{it} based on the correct model. Thus, $\log L_A - \log L_B$ reflects relative misspecification and the mode of ω is higher when model A is less misspecified [1].

[1] When y_{At} and y_{Bt} are vectors the equations should be adjusted accordingly. When y_{At} is a $m \times 1$ vector and y_{Bt} is, e.g., a scalar or when y_{At} is different from y_{Bt}, $\log L_A - \log L_B$ reflects, apart from differences in the variances, the average misspecification in all the equations of model A relative to the misspecification of the single

A popular method to rank models (and combine their predictions) is Bayesian model averaging (BMA). Asymptotically BMA puts all the weight on the model which is closest to the DGP in a KL sense. Because the posterior mode of ω measures the relative misspecification of the available models, and because, as $\min T_i \to \infty, i = 1, \dots K$, it will converge to one for the model which is closest to the DGP in a KL sense, one expects the two measures to provide similar ranking information. However, a BMA weight can only be computed when $y_{At} = y_{Bt}$ and $T_A = T_B$; 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, only BMA gives ex-post combination of individual model estimates. Some experimental evidence on the performance of the two ranking devices is in section 3.3.

3.2 Discussion

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

$$\log L = \sum_{t=1}^{\min\{T_A,T_B\}} \log(w_t L(\rho_A, \sigma_A | y_{At}) + (1 - w_t) L(\rho_A, \Sigma_A, \delta, \gamma | y_{Bt})) \tag{24}$$

Simple manipulations reveal that (24) and the log of (9) differ by Jensen's inequality terms [2].

While a-priori both composite and finite mixture devices are appealing to guard against misspecification, a composite likelihood has three advantages. From a computational point of view, when the decision rules have an autoregressive structure, estimators for θ have a closed form expression in the composite likelihood case, but not in the finite mixture case. In addition,

---

equation of model B. Thus, if model A has some very poorly specified equations, it may have low a-posteriori ω, even though certain equations are correctly specified (and θ appears in those equations).

[2] If $T_1 = T_2 = 2$ the composite log-likelihood is

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

while the log-likelihood in the mixture model is

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

Suppose $\omega_T = \omega = 1 - \omega$. Then, (9) and (24) differ because $\log \sum_{t=1}^{T} x_t \equiv \log x_1 + \log(1 + \sum_{t=2}^{T} \frac{x_t}{x_1})$, one has $\log \sum_{t=1}^{2} x_t = \log x_1 + \log(1 + \frac{x_2}{x_1})$ and this differs from $\sum_{t=1}^{2} \log x_t = \log x_1 + \log x_2$, since $\log(1 + \frac{x_2}{x_1}) \approx \frac{x_2}{x_1}$ if $\frac{x_2}{x_1}$ is small. When ω is time varying additional differences will be recorded.

in a finite mixture it must be the case that $y_{At} = y_{Bt}$ and $T_A = T_B$, 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.

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 SVAR 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 difference 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 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 unconditional orthogonality conditions are a weighted average of the scores of each structural model. The larger is the set of models considered, the more over-identified are the resulting estimators [3].

3.3 Some experimental evidence

To understand what 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

[3] One could use this fact to provide an alternative definition of misspecification: a model is misspecified if the set of measures $\{Q \in L^1 | \int \frac{\partial L(\phi_i|y)}{\partial \phi_i} dQ| \neq 0\}$ does not contain $F(y, \psi)$ when Q is evaluated at ψ.

models used in estimation are an AR(1): $\log y_t = \rho_1 \log y_{t-1} + \log u_t$ and an MA(1): $\log y_t = \log \epsilon_t + \beta_1 \log \epsilon_{t-1}$. 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. Because both models disregard part of the serial correlation of the DGP, $\sigma_u, \sigma_\epsilon$ will be upward biased. Would geometrically combining the two 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 does the posterior mode of ω relates to a BMA weight?

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.

To measure the performance of composite specifications 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. Table 2 has the posterior mode of ω (our estimated weight on the AR(1)), the posterior standard deviation of ω, and the BMA weight on the AR(1). Given that the two models share the same observable, the comparison between BMA and posterior mode of ω is valid.

Composite specifications produce better estimates of σ and at least one of the composite model 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

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	0.166	0.176	0.253
		KL	15.04	4.07	9.67	10.58	6.96
$\sigma^2 = 0.5, \rho = 0.6, \beta = 0.5$	T=50	MSE	0.061	0.075	0.058	0.066	0.106
		KL	13.91	7.89	13.22	13.77	8.06
$\sigma^2 = 0.8, \rho = 0.6, \beta = 0.5$	T=50	MSE	0.021	0.027	0.019	0.026	0.050
		KL	12.55	5.87	11.45	12.16	5.96
$\sigma^2 = 1.0, \rho = 0.6, \beta = 0.5$	T=50	MSE	0.008	0.011	0.007	0.012	0.029
		KL	11.83	5.32	10.62	11.68	7.76
$\sigma^2 = 1.2, \rho = 0.6, \beta = 0.5$	T=50	MSE	0.006	0.007	0.005	0.007	0.017
		KL	9.34	4.48	8.02	9.07	7.92
$\sigma^2 = 0.5, \rho = 0.6, \beta = 0.8$	T=50	MSE	0.149	0.168	0.204	0.204	0.292
		KL	10.88	5.02	10.41	10.78	5.03
$\sigma^2 = 1.0, \rho = 0.6, \beta = 0.8$	T=50	MSE	0.009	0.011	0.036	0.035	0.060
		KL	8.90	5.35	9.54	10.40	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	11.82
$\sigma^2 = 1.0, \rho = 0.9, \beta = 0.2$	T=50	MSE	0.008	0.077	0.005	0.008	0.368
		KL	9.90	19.26	11.32	10.93	12.93
$\sigma^2 = 1.0, \rho = 0.9, \beta = 0.2$	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	12.33
$\sigma^2 = 0.5, \rho = 0.3, \beta = 0.8$	T=50	MSE	0.131	0.166	0.166	0.189	0.179
		KL	4.70	6.61	10.65	10.91	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	5.74
$\sigma^2 = 1.0, \rho = 0.3, \beta = 0.8$	T=100	MSE	0.007	0.011	0.023	0.033	0.011
		KL	4.45	4.74	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	6.66

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.

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 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	0.967 (0.03)	1.00
$\sigma^2 = 0.5, \rho = 0.6, \beta = 0.5$	T=50	0.967 (0.03)	1.00
$\sigma^2 = 0.8, \rho = 0.6, \beta = 0.5$	T=50	0.967 (0.03)	1.00
$\sigma^2 = 1.0, \rho = 0.6, \beta = 0.5$	T=50	0.967 (0.03)	1.00
$\sigma^2 = 1.2, \rho = 0.6, \beta = 0.5$	T=50	0.967 (0.03)	1.00
$\sigma^2 = 0.5, \rho = 0.6, \beta = 0.8$	T=50	0.967 (0.03)	1.00
$\sigma^2 = 1.0, \rho = 0.6, \beta = 0.8$	T=50	0.970 (0.03)	1.00
$\sigma^2 = 0.5, \rho = 0.9, \beta = 0.2$	T=50	0.995 (0.004)	1.00
$\sigma^2 = 1.0, \rho = 0.9, \beta = 0.2$	T=50	0.993 (0.004)	1.00
$\sigma^2 = 1.0, \rho = 0.9, \beta = 0.2$	T=100	0.993 (0.004)	1.00
$\sigma^2 = 1.0, \rho = 0.9, \beta = 0.2$	T=250	0.995 (0.002)	1.00
$\sigma^2 = 0.5, \rho = 0.3, \beta = 0.8$	T=50	0.116 (0.13)	0.994
$\sigma^2 = 1.0, \rho = 0.3, \beta = 0.8$	T=50	0.050 (0.05)	0.946
$\sigma^2 = 1.0, \rho = 0.3, \beta = 0.8$	T=100	0.041 (0.04)	0.105
$\sigma^2 = 1.0, \rho = 0.3, \beta = 0.8$	T=250	0.021 (0.02)	0.000

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

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 short, the ω and the BMA measures disagree regarding the likelihood of the AR(1). This divergence disappears when $T \geq 100$ and both models put smaller weight on the AR(1). Notice that the posterior of ω is updated in the direction of the basic model with smaller KL divergence even when $T = 50$.

What would happen to the composite posterior of ω when one of the estimated models is the DGP? Would it concentrate around 1 for the correct model as sample size increases? Table 3 reports evidence for two DGPs: an AR(1) and an MA(1). Clearly, the posterior of ω asymptotically concentrates at the corner solution corresponding to the correct model but at a somewhat slower rate than a BMA weight.

Table 3: Posterior estimates of ω

	Mode	Mean	Median	Std deviation	BMA weight
$y_t = 0.8y_{t-1} + e_t,\ e_t \sim N(0, \sigma^2)$, T=50					
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
$y_t = 0.7e_{t-1} + e_t,\ e_t \sim N(0, \sigma^2)$, T=50					
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

In sum, our simulations show that estimation outcomes can be improved and misspecification reduced using composite methods. Furthermore, the posterior mode of ω gives a consistent ranking device for misspecified models which has useful properties: its modal value agrees with a BMA weight in many specifications and it is superior when T is small and MA components dominate.

4 Estimation and inference

In a traditional setting, where the models entering the composite likelihood are marginal or conditional versions of the true DGP (see e.g. Varin, 2011), composite likelihood estimators are consistent and asymptotically normal. However, they are inefficient and one can select ω_i to minimize this inefficiency (see Appendix A).

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, $f(y_{it} \in A_i, \theta, \eta_i)$ are neither marginal nor conditional densities, but misspecified approximations of the unknown DGP. Thus, for all (θ, η_i), the KL divergence between $F(y_t, \psi)$ and $f(y_{it} \in A_i, \theta, \eta_i)$ is positive, $\forall i$. Third, $f(y_{it} \in A_i, \theta, \eta_i)$ need not be independent (models may share equations) nor compatible, in the sense that the likelihood estimator $\theta_{i,ML}$ asymptotically converges to the same value. Finally, we treat ω_i as a random variable, rather than a fixed number and wish to construct a quasi-posterior 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 $f(y_{it} \in A_i, \theta, \eta_i)$ are inconsistent and, as a consequence, the composite likelihood esti-

mator is also inconsistent. Following earlier work by White (1982) and Domowitz and White (1982), one can show, as sample size grows and under regularity conditions, that $\theta_{i,ML}$ converges to θ_0, the pseudo-parameter vector minimizing the KL divergence from the DGP. Moreover, $\sqrt{T}(\theta_{i,ML} - \theta_0) \sim N(0, G_i^{-1})$, where $G_i = H_i J_i^{-1} H_i$ is the Godambe information matrix for model i, J_i is the variability matrix and H_i the sensitivity matrix (see Appendix A). Thus, when misspecification is present the pivot of the asymptotic distribution is the minimizer of the KL divergence, rather than the true parameter vector; and the Godambe 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, θ_{CL} asymptotically approaches the pseudo-parameter value, say $\theta_{0,CL}$, minimizing the KL divergence between the density of the composite pool and the DGP. $\theta_{0,CL}$ is not, in general, a weighted average of $\theta_{0,i}$ because models are not necessarily independent. Furthermore, $\sqrt{T}(\theta_{CL} - \theta_{0,CL}) \sim N(0, G^{-1})$, where $G = HJ^{-1}H$ and H and J evaluated at the composite likelihood estimator (see Appendix A for details).

4.1 Bayesian quasi-posteriors

We use a Bayesian approach to estimation and inference, in part because of the special role the quasi-posterior of ω plays in our setup. 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. Lacking a closed form expression, we employ a multiple block Metropolis-Hastings approach to numerically compute sequences from these marginal.

Given (y_{it}, T_i), we assume that $\sup_{\{\theta,\eta_i\}} f(y_{it} \in A_i, \theta, \eta_i) < b_i \leq B < \infty$, 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 $CL(\chi | y_{1,T_i}, \ldots, y_{K,T_k})$ exists for $0 < \omega_i < 1$, $\sum_i \omega_i = 1$. Let the priors for model i parameters be of the form:

$$p(\theta, \eta_i) = p(\theta) p(\eta_i | \theta, y_{i0}) \tag{25}$$

where y_{i0} is a training sample. In (25) 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. The composite posterior kernel is:

$$\check{p}(\chi|y_{1,t_1},\ldots,y_{k,T_k}) \quad = \quad \Pi_i \mathcal{L}(\theta,\eta_i|y_{i,T_i})^{\omega_i} p(\eta_i)^{\omega_i} p(\theta) p(\omega_i) \tag{26}$$

which can be used to estimate χ as described, e.g. in Chernozukov and Hong (2003). Because the composite likelihood is an adequate loss function, the insights of the quasi-Bayesian computation literature (see Bisseri, et al. 2016) also apply. For computational and efficiency reasons, we employ a $K+1$ block Metropolis-Hastings algorithm. Herbst and Schorfheide (2015) also suggest 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.

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

$$\min\left(1, \frac{\mathcal{L}(\left[\eta_i^*, \theta^{iter-1}\right]|y_{i,T_i})^{\omega_i^{iter-1}} p(\eta_i^*|\theta^{iter-1})^{\omega_i^{iter-1}}}{\mathcal{L}(\left[\eta_i^{iter-1}, \theta^{iter-1}\right]|y_{i,T_i})^{\omega_i^{iter-1}} p(\eta_i^{iter-1}|\theta^{iter-1})^{\omega_i^{iter-1}}}\right) \tag{27}$$

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

$$\min\left(1, \frac{\mathcal{L}(\left[\eta_1^{iter}, \theta^*\right]|y_{1,T_1})^{\omega_1^{iter-1}} \ldots \mathcal{L}(\left[\eta_K^{iter}\theta^*\right]|y_{K,T_K})^{\omega_K^{iter-1}} p(\theta^*)}{\mathcal{L}([\eta_1^{iter}, \theta^{iter-1}]|y_{1,T_1})^{\omega_1^{iter-1}} \ldots \mathcal{L}([\eta_K^{iter}, \theta^{iter-1}]|y_{K,T_K})^{\omega_K^{iter-1}} p(\theta^{iter-1})}\right) \tag{28}$$

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

$$\min\left(1, \frac{\mathcal{L}(\left[\eta_1^{iter}, \theta^{iter}\right]|y_{1,T_1})^{\omega_1^*} \ldots \mathcal{L}(\left[\eta_K^{iter}\theta^{iter}\right]|y_{K,T_K})^{\omega_K^*} p(\omega^*)}{\mathcal{L}([\eta_1^{iter}, \theta^{iter}]\, y_{1,T_1})^{\omega_1^{iter-1}} \ldots \mathcal{L}([\eta_i^{iter}, \theta^{iter}]|y_{K,T_K})^{\omega_K^{iter-1}} p(\omega^{iter-1})}\right) \tag{29}$$

Note that in (27) only the likelihood of model i matters and when the proposals are asymmetric, the acceptance probability should be appropriately adjusted. A few interesting special cases are nested in the algorithm. For example, when the K models feature no nuisance parameters, steps

2.-3. can be combined in a single step. On the other hand, if ω_i's are fixed, step 4 disappears. 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, the "random walk Dirichlet" proposal seems appropriate (see Appendix B).

4.3 Adjusting percentiles of the MCMC distribution

Our estimation problem is non-standard since models are misspecified and y_{it} are not necessarily mutually exclusive across i (see Mueller, 2013). 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 they were independent across i. 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, as a referee suggested, a normal posterior with sandwich covariance matrix. Here we follow Ribatet et al. (2012)and Qu (2016), and directly add two steps to the MCMC algorithm. 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

$$\tilde{\chi}^j = \hat{\chi} + V^{-1}(\chi^j - \hat{\chi}) \tag{30}$$

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 [4].

The adjustment works well when χ is well identified from the composite posterior and if the composite posterior has a unique maximum. As Canova and Sala (2009) have shown, such properties may not hold in a number of structural models. Thus, we recommend users to report both standard and adjusted percentiles.

4.4 Intepretations and time varations

One can think of composite posterior analysis in different ways. One is the sequential learning interpretation provided in Canova and Matthes (2017): the composite posterior kernel can be

[4]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). In this way MCMC draws are recursively centered, which insures faster convergence, but a numerical optimization is needed at each step of the procedure.

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 downweighted. Here, 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.

Our composite posterior estimators are special quasi-Bayesian estimators. In this literature (see e.g., Marin, 2012, Bissiri et al., 2016, and Scalone, 2018), one updates prior beliefs using a loss function which downplays some undesirable features of the likelihood. In particular, a moment-based loss functions provide estimators which reduce the inconsistencies of likelihood-based methods when misspecification is present. Seen through this lens, the composite likelihood is a moment-based loss function which uses a weighted average of each model's scores.

Since the composite likelihood can be interpreted as an "opinion" pool of agents/models using different pieces of information, the composite quasi-posterior statistics we compute in the next subsection are a Bayesian pool 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 information, see Roche (2016). Thus, a composite likelihood can be associated with a probability distribution on hypotheses, extending Bayesian analysis to problems where the likelihood function is unknown.

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, for example, assume a random walk 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), cross equation restrictions present in different models could receive different weights in different portions of the sample.

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. 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.

Let $\tilde{y}_{t+l}$ be future values of the variables appearing in all models. Let $f(\tilde{y}_{t+l}|y_{it}, \theta, \eta_i)$ be the prediction of $\tilde{y}_{t+l}, l = 1, 2, \ldots$ made by model i, given (θ, η_i) and let $f^{cl}(\tilde{y}_{t+l}|y_{1t}, \ldots, y_{Kt}, \chi) =$

$\prod_{i=1}^{K} f(\tilde{y}_{t+l}|y_{it}, \theta, \eta_i)^{\omega_i}$ be a geometric pool of predictions of $\tilde{y}_{t+l}$, given data up to t, the models, and the parameters. Then

$$\begin{aligned} p(\tilde{y}_{t+l}|y_{1t}, \ldots, y_{Kt}, \omega_1, \ldots \omega_K) &\propto \int \ldots \int f^{cl}(\tilde{y}_{t+l}|y_{1t}, \ldots, y_{Kt}, \chi) \\ & p(\theta, \eta_1, \ldots, \eta_K|y_{1t}, \ldots, y_{Kt}, \omega_1, \ldots, \omega_K) d\theta d\eta_1 \ldots d\eta_K \\ &= \int \ldots \int \prod_i p(\tilde{y}_{t+l}, \theta, \eta_i|y_{it})^{\omega_i} d\theta d\eta_1 \ldots d\eta_K \end{aligned} \tag{31}$$

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 (see Roche, 2016)

Depending on the investigator's loss function, one could compute (31) using the mode or the posterior mean of ω_i. One could also integrate (31) with respect to the marginal of ω, but given that in many applications it makes sense to condition on the estimated ω's (which represents the posterior probability associated with each model), we believe (31) has larger appeal.

$f(\tilde{y}_{t+l}|y_{it}, \theta, \eta_i)$ is straightforward to compute since the models we consider have linear (Gaussian) state space representation. Thus, (31) 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 lead to finite mixtures predictive densities such as BMA or static pools, and logarithmic pooling, which is what a composite predictive density approach produces. Logarithmic pooling 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 naive ex-post pools of predictive densities, the composite predictive density uses the information in all models for estimation and to compute weights [5]. This may lead to large differences, especially when models are misspecified in different ways. 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 preferrable if nonlinearities are suspected to exist over time, the logarithmic pooling implicit in (31) generally suffices for the purposes of guarding against misspecification of linear macroeconometric models.

[5]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.

In analogy with the prediction problem, one can compute statistics of interest by geometrically weighting the densities of outcomes obtained with each model and the composite posterior for the parameters. Take, for example, the computation of the responses for the subset of variables present in all models to a shock also present in all models. Given (θ, η_i), responses to shock j for model i can be computed setting all other structural shocks to 0 - which is reasonable given that the models considered are linear and shocks are uncorrelated. The density of outcome paths, computed randomizing (θ, η_i) from their posterior, is the impulse response of interest. The kernel of the composite posterior responses can then be computed analogously to (31), with the density of outcome paths replacing predictive densities.

Counterfactuals can be similarly computed. Let $\bar{y}_{kt+l}$ be a selected path for the future values $t+l$ in the k-th element of $\tilde{y}_{t+l}$. Using $f(\bar{y}_{kt+l}|y_{it}, \epsilon^j_{it+l}, \theta, \eta_i)$ 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 $f(\bar{y}_{k't+l}|y_{it}, \epsilon^j_{it+l}, \theta, \eta_i)$, for $k' \neq k$. Composite counterfactuals can be computed as in (31).

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, how to shed light on the role of technology shocks as drivers of output fluctuations.

5.1 Measuring the marginal propensity to consume

We consider five models suggested in the literature to explain the dynamics of the MPC in the data: the first is a standard permanent income model; the others add aspects of the consumption-income relationship left out of the workhorse model. In the baseline model there is a representative agent with quadratic preferences, constant interest rate, $(1+r)\beta = 1$, and exogenous permanent and transitory income components. The second model has similar features but preferences are exponential (in the spirit of Caballero, 1990). Because the variance of income shocks affect consumption decisions, precautionary saving matter and consumption is no longer a random walk. To make the model empirically interesting we allow the volatility of both income components to be time dependent and assume a simple AR(1) specification for the log of the variance. In the third model we make the real rate endogenous. We consider a real business cycle (RBC) structure featuring consumption-leisure choices, production requiring capital and labor, and a technological disturbance with transitory and permanent components. Here preferences have a separable CRRA format. The fourth specification introduces agents' heterogeneity: a

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
Composite	0.85	0.90	0.96
Composite (without RBC)	0.80	0.85	0.91

fourth of the agents consume all their current income, as in Gali et al., 2004. Preferences and constraints are the same 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, constant interest 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 (which depreciate 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. The log-linearized conditions are in appendix E.

Although models feature different endogenous variables, we use aggregate real per-capita non-durable consumption, real per-capita income, and real per-capita value of assets as observables for 1980:1-2017:2 for all specifications - in the RBC model we equate real per-capita assets with the per-capita capital of the representative agent. This choice of observables allows us to compare composite and BMA ranking of models and predictions. All variables are quadratically detrended. Estimation is performed with MCMC techniques using the likelihood of each model or the composite likelihood, restricting 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 0.20. The priors for all other parameters are proper but loose and truncated, when needed, to the region with economic interpretation.

Table 4 presents a summary of the posterior of ρ. The first five rows display single model percentiles; the sixth row the composite percentiles. Although Cogley and Nason (1995) have shown that income persistence in a RBC model is largely driven by the persistence of TFP, one may argue that TFP persistence and exogenous income persistence are parameters with different economic interpretations. Thus, the last row of table 4 presents composite percentiles when ρ

is restricted to be common only across models with exogenous labor 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 very persistent and 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 composite posterior percentiles of ρ practically unchanged.

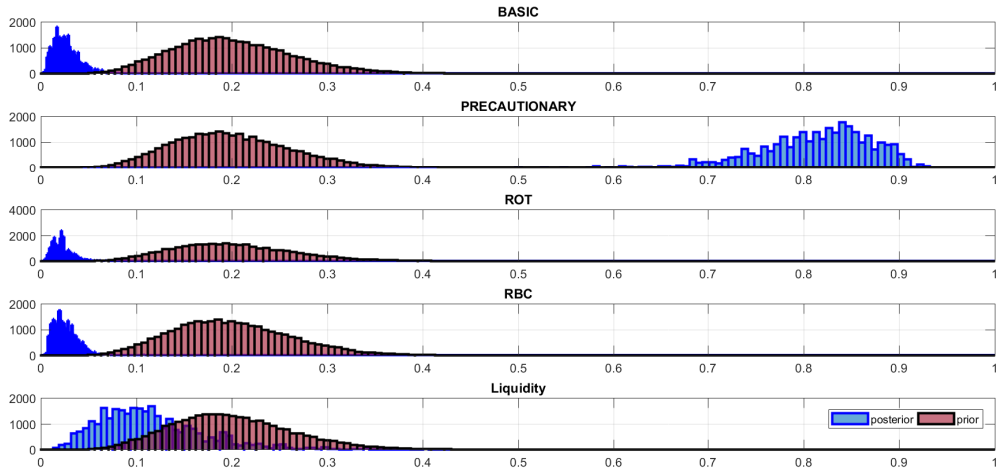

Figure 1: Prior and posterior for ω

Why is the composite posterior median of ρ high? Figure 1, which presents the prior and the posterior of ω for the five models, shows that the precautionary model receives the highest a-posteriori weight. Thus, the fact that interest rate is constant, that labor supply decision and heterogeneities are disregarded are less of a problem when characterizing the MPC than leaving precautionary motives out of the basic model. Note that the weights are very stable over time (estimates available on request). Thus, income uncertainty is not a dominant factor only in the last 10 years of data.

Figure 2 present dynamic estimates of MPC_{y^T}, computed as $MPC_y^T(l) = \frac{\sum_{j=1}^{l} c_{t+j}|e_t^T}{\sum_{j=1}^{l} y_{t+j}|e_t^T}$, $l = 1, 2, ...$, 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. When ρ is estimated to be low, MPC_{y^T} is also low. Consistent with the discussion in section 2, instantaneous posterior estimates

of MPC_{y^T} obtained with BASIC, RBC, and LIQUIDITY models are around 0.05. Estimates increase at longer horizons but after two years the 68 percent range is still below 0.10. The instantaneous MPC slightly higher in the ROT model (the median value is now 0.25). Still, after two years the economy 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, only 15 percent of transitory income is spent the first quarter and less than 25 percent after two years.

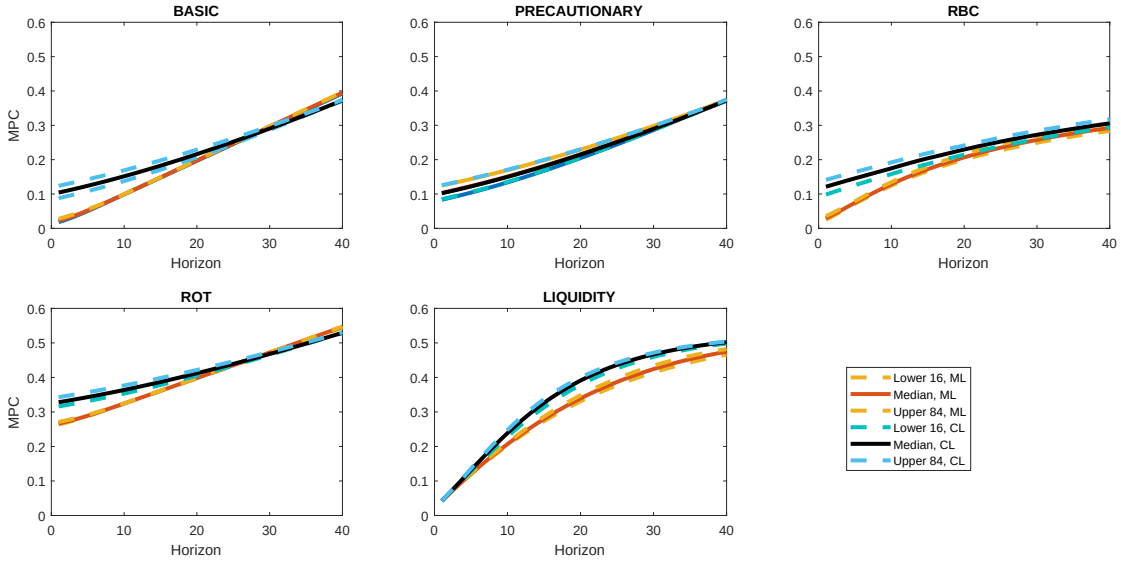

Figure 2: Likelihood and composite likelihood estimates of MPC_{y^T}

With composite posterior estimates ρ, 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 for the first two years. Thus, even when income is relatively persistent, rational consumers save the majority of their transitory income. Interestingly, with composite estimates of ρ, differences in MPC_{y^T} estimates across models are smaller.

Rather than plugging composite posterior estimates in a model, one may want 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 standard combinations: one constructed using BMA weights and one using naive equal weights.

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

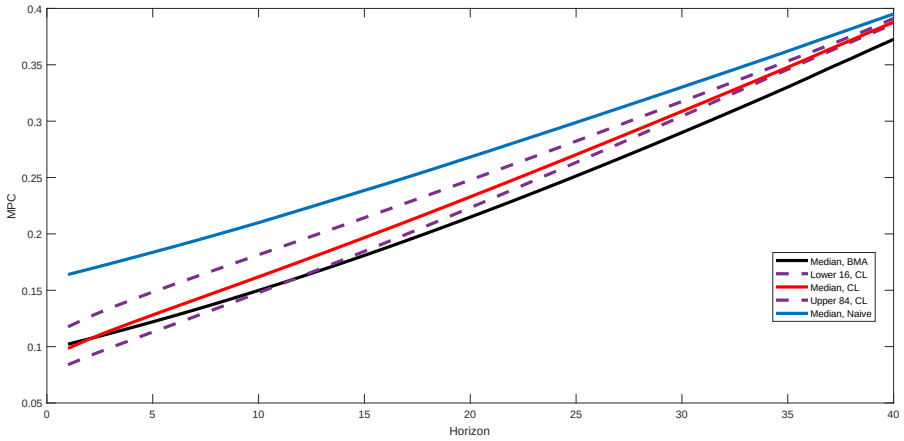

Figure 3: Composite, BMA and naive posterior estimates of the MPC

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.

Finally, we examine whether the composite model is less misspecified by computing the average KL divergence for detrended real per-capita consumption for each of the five models and for the composite one. Unsurprisingly, given ω estimates, the PRECAUTIONARY model is closest to the DGP (KL=117.80) but the composite model is significantly better (KL=76.60).

In sum, our analysis indicates that over a two-years horizon, US consumers spend at most one-third of their cumulative transitory income on non-durable goods. Whether the rest is used to repay debt, to purchase non-durable goods or to make intergenerational transfers is an important question we leave for future research.

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 results are due, in part, to specification choices and, in part, to the sample used in the computations. 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 issue 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, one could restrict other parameters (e.g. Taylor rule coefficients). We constrain only a few parameters to be common to highlight the stark differences obtained when estimating the JPT model separately or jointly with the HS model. The optimality conditions are in appendix F. Note that both models feature permanent and transitory technological disturbances; that the HS model is not nested in the JPT model via parametric restrictions; and that we can approximate a RBC framework through prior parameter restrictions. Thus, one can also think of our exercise as combining NK and RBC frameworks without having to worry about the typical poor fit of RBC models for nominal variables.

We estimate the weights assuming that the two models are a-priori equally likely. Since we use different observables 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 here.

When the JPT model is estimated in isolation, estimates of κ and ρ_z are low (means 0.02 and 0.14, standard deviations 0.0001 and 0.0041, respectively). The mean estimates are similar to the point estimates reported by Justiniano et al. (0.10 and 0.24), despite a different estimation sample [6]. They imply that technology shocks explain 30-40 percent of output fluctuations at typical business cycle horizons. 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 (see figure 4).

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 JPT model is estimated in isolation, technology are used to fit a number of equations, but when it is paired with the HS model, they are restricted to fit well the Euler equation. This constraint moves posterior estimation to a region of the parameter space where nominal rigidities are smaller (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 (mean value of the persistence of preference shocks drops from 0.76 to 0.23). The combined effect of higher persistence of the stationary component of technology shock, of higher real and lower price rigidities, and of

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

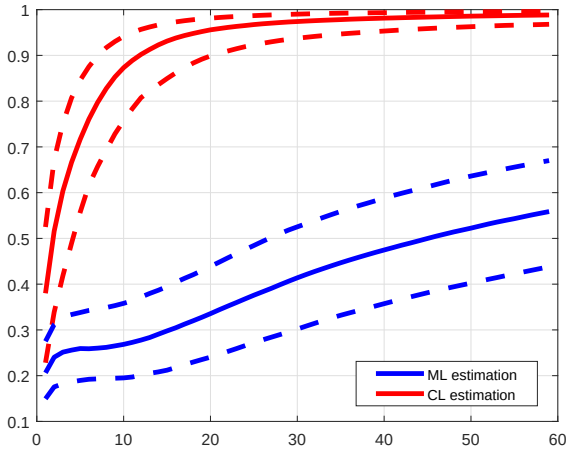

Figure 4: Fraction of output fluctuations due to TFP shocks, JPT model

lower persistence of demand shocks make technology disturbances more important for output fluctuations. These conclusions remain when we restrict the HS model 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. While misspecification is roughly the same (average KL is 0.025 for the composite model and 0.021 for the JPT model), our results indicate that the JPT model possesses modes featuring mechanics of transmission of structural disturbances different from the usual ones. Clearly, more work is needed 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 current DSGE models. We consider a set of potentially misspecified models, geometrically combine their likelihood functions, and estimate the parameters with the resulting composite likelihood. The composite likelihood shrinks individual likelihood estimates toward a weighted average of all other models' estimates. Thus, composite estimation guards against misspecification by requiring estimates of the common parameters consistent with the structure present in all models. We highlight the properties of our approach and relate our methodology to existing ones.

We describe a MCMC approach to draw sequences from the composite posterior distributions, and show how to adjust the MCMC percentiles to produce posterior credible sets with the

right asymptotic coverage. We also discuss how posterior weights inform us about the relative misspecification of different models and show how to construct composite posterior statistics.

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 they 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. Technology shocks explain about one-third of output fluctuations in a standard medium scale NK model at business cycle horizons but their importance increases when such a model is paired with a smaller scale model without capital and the persistence of technology shocks jointly estimated.

We conclude with some practical suggestions to potential users. First, to make the approach meaningful the models entering the composite likelihood should capture different aspects left out (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, the composite likelihood can still be employed since if $\theta = \emptyset$, the approach produces likelihood estimates, model by model. Third, while the methodology robustifies estimation and 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, the approach has a number of benefits relative to likelihood-based estimation of the structural parameters (see Canova and Matthes, 2017). For example, when a large scale model is available, the composite likelihood constructed using model blocks has shape and properties which are similar to those of the likelihood of the full model, without the numerical difficulties. Thus, our approach is not only useful to examine in which direction a model should be improved, but also to estimate the larger scale models one is likely to build after the initial experimentation. Fifth, although we focus on linearized models, one can combine the likelihoods of models perturbed at higher order and we expect the gains to remain. Finally, the approach is suited to deal with structural time varying coefficients models, which are complicated to estimate and interpret with standard likelihood-based technology.

References

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

Barnichon, R. and C. Browlees (2016). Impulse response estimation by smooth local projection. Forthcoming, *Review of Economics and Statistics*.

Baumeister, C. and J. D. Hamilton (2015). Structural interpretation of vector autoregressions with incomplete identification: revisiting the role of oil supply and demand shocks. Forthcoming, *American economic Review*.

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 frameowrk for updating belief distributions. *Journal of the royal Statistical Society*, Ser B, 78, 1103-1130.

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

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

Canova, F. and C. Matthes (2017). Solving computational and estimation problems in dynamic structural models: a composite likelihood approach, manuscript.

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. ECB working paper 1655. *Quantitative Economics*, 8, 9771020.

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.

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. Dreschel (2018). Agnostic structrual 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.

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 Paritally Identified VARs.UCL manuscript.

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

Hansen, L. and T. Sargent (2008). Robustness. Princeton University Press, Priceton, 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. Universitat Pompeu Fabra, manuscript.

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, 11671180.

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

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. (2015). A composite likelihood approach to analyze singular DSGE models. Forthcoming, *Review of Economics and Statistics*.

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.

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

On-line Appendices

Appendix A: Classical composite likelihood estimators

Asymptotic properties of composite likelihood estimators In a standard approach one has a known DGP which produces a parametric density $F(y_t, \psi)$ for an $m \times 1$ vector of observables y_t, given a $q \times 1$ vector of parameters $\psi = (\theta, \eta)$, where θ is $q_1 \times 1$ and η is $q - q_1 \times 1$. When y_t is of high dimensions or contains latent variables, it may be difficult to use $F(y_t, \psi)$ for estimation purposes.

The key idea of composite methods is to construct arbitrary sets of low dimensional densities and to combine them for estimation purposes. This may be viewed as *divide-and-conquer* method of approximating the full likelihood.

Let $f(y_{it} \in A_i, \phi_i)$ be sub-densities of $F(y_t, \psi)$ obtained by marginalizing (or conditioning on portions of) $F(y_t, \psi)$, where A_i is a set and $i = 1, \ldots, K$. For ease of reading, the integrals and the conditioning sets are left implicit. Each sub-density defines a sub-model, has an associated vector of parameters $\phi_i = [\theta, \eta_i]'$, where η_i are (nuisance) sub-density specific, and has implications for a sub-vector y_{it} of length T_i. The elements of y_{it} need not be mutually exclusive across i and T_i may be different than T_j. Given a vector of fixed weights ω_i, the composite likelihood is

$$CL(\theta, \eta_1, \ldots, \eta_K, y_{1t}, \ldots, y_{KT}) = \Pi_{i=1}^{K} \; f(y_{it} \in A_i, \theta, \eta_i)^{\omega_i} \equiv \Pi_{i=1}^{K} \mathcal{L}(\theta, \eta_i | y_{it} \in A_i)^{\omega_i} \tag{32}$$

Although $CL(\phi, y) \equiv CL(\theta, \eta_1, \ldots, \eta_K, y_{1t}, \ldots, y_{KT})$ is not a likelihood function, if $y_{[1,t]}$ $=(y_1, \ldots, y_t)$ is an independent sample from $F(y_t, \psi)$ and ω_i are fixed, θ_{CL}, the maximum composite likelihood estimator satisfies $\theta_{CL} \xrightarrow{P} \theta$ and

$$\sqrt{T}(\theta_{CL} - \theta) \xrightarrow{D} N(0, G^{-1}) \tag{33}$$

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

$$\begin{aligned}
G &= HJ^{-1}H \text{ Godambe information} && (34)\\
J &\equiv var_\theta u(\phi, y_{[1,t]}|\omega) \quad \text{Variability matrix} && (35)\\
H &\equiv -E_\theta[\bigtriangledown_\theta u(\phi, y_{[1,t]}|\omega)] \quad \text{Sensitivity matrix} && (36)\\
u(\phi, y_{[1,t]}|\omega) &= \sum_i \omega_i \bigtriangledown_\theta l_i(\theta, \eta_i, y_{[1,t]}) \quad \text{Composite scores} && (37)
\end{aligned}$$

$\bigtriangledown_\theta l_i(\theta, \eta_i, y_{[1,t]})$ are the score associated with the log of $f(y_{it} \in A_i, \theta, \eta_i)$, and $H \neq J$ [7].

Consistency obtains because each sub-model i provides an unbiased estimating function for θ. Since the ML estimator of each sub-model converges to the true parameter vector as T increases, $\theta_{CL} \xrightarrow{P} \theta$. Asymptotic normality holds because, typically, the sampling distribution of the maximum likelihood estimator of each i can be approximated quadratically around the same mode. θ_{CL} is inefficient - G equals Fisher information matrix, I, only if the composite likelihood is the likelihood of the true model. Careful choices ω_i may improve efficiency and optimal weights can be designed by minimizing the distance between G and I, or by insuring that the composite likelihood ratio statistics has an asymptotic χ^2 distribution, see Pauli et al. (2011).

If consistency is all that one cares about, one could set $\omega_i = \frac{1}{K}, \forall i$ or use a data-based approach, e.g. select $\omega_i = \frac{\exp(\zeta_i)}{1+\sum_{i=1}^{K-1}\exp(\zeta_i)}$, where ζ_i are functions of some statistics of past data, $\zeta_i = \zeta(Y_{i,[-\tau:0]})$. If these statistics are updated over time, ω_i could also be made time varying. There is a large forecasting literature (see e.g. Aiolfi et al., 2010) which can be used to select training sample-based estimates of ω_i and to make them time varying.

The asymptotic properties of θ depend on $(\eta_1, \ldots, \eta_K)$. In standard exercises η_i are assumed to be known, so the dependence disappears. When η_i are unknown, but estimable a two-step approach is generally implemented: η_i are estimated separately from each $\log f(y_{it} \in A_i, \theta, \eta_i)$ and plugged in the composite likelihood, which is then optimized with respect to θ, see e.g. Pakel et al. (2011). Consistency of θ_{CL} is unaffected as long as η_i are consistently estimated, but standard errors need to be properly adjusted. A two-step approach is convenient when K or the number of nuisance parameters is large, since joint estimation of $(\theta, \eta_1, \ldots, \eta_K)$ may be demanding.

Asymptotic properties of composite estimators under misspecification When $f(y_{it} \in A, \theta, \eta_i)$ are not marginal or conditional representations of $F(y_t, \psi)$, the previous conclusions need to be modified. Let $y_{[1,t]}$ be a sample from $F(y_t, \psi)$ with respect to some σ-measure μ. Suppose model i with density $f_i(y_{[1,t]}, \phi_i)$, where $\phi_i \in \Phi \subset R^m$ is a vector of parameters, is used in the analysis and let its log-likelihood be $l_i(\phi_i) = \sum_t \log f_i(y_t, \phi_i)$. The model is misspecified because $F(y_{[1,t]}, \psi) \neq f_i(y_{[1,t]}, \phi_i), \ \forall \phi_i$. Let $\phi_{i,ML} = sup_{\phi_i} l_i(\phi_i)$. Since $T^{-1} l_i(\phi_i) \to E(\log f_i(y_{[1,t]}, \phi_i))$, by the uniform law of large numbers, $\phi_{i,ML}$ is consistent for $\phi_{i,0} = arg\max_{\phi_i} E \log f_i(y_{[1,t]}, \phi_i)$, where the expectations are taken with respect to F. If F is absolutely continuous with respect

[7] If T is fixed, but $K \to \infty$, and the sub-models are independent, the result still holds. On the other hand, when $\{y_t\}_{t=1}^T$ has correlated observations similar results can be proved, see Engle et al. (2008). Note also that a standard Newey-West correction to $J(\theta)$ can be used if $y_{[1,t]}$ is not an independent sample.

to f_i

$$E \log f_i(y_{[1,t]}, \phi_i) - E \log F(y_{[1,t]}, \psi) = - \int F(y_{[1,t]}, \psi) \log \frac{F(y_{[1,t]}, \psi)}{f_i(y_{[1,t]}, \phi_i)} d\mu(y_{[1,t]}) = -KL_i(\phi_i) \tag{38}$$

Hence, $\phi_{i,0}$ is also the minimizer of KL_i, the Kullback-Liebler divergence between F and f_i.

Let $s_t^i(\phi_i) = \bigtriangledown_{\phi_i} \ln f_i(y_t, \phi_i)$ be the score of observation t and let $h_t^i(\phi_i) = \bigtriangledown_{\phi_i} s_t^i(\phi_i)$. When the maximum is in the interior of Φ, $\sum_t s_t^i(\phi_i) = 0$, and taking a first order expansion we have

$$0 \approx T^{-0.5} \sum_t s_t^i(\phi_{i,0}) + T^{0.5} V_1^{-1} (\phi_{i,ML} - \phi_{i,0}) \tag{39}$$

where $V_1 = -E(h_t^i(\phi_{i,0})) = \bigtriangledown_\phi^2 KL_i(\phi_i)_{\phi_i = \phi_{i,0}}$. Using a central limit theorem for uncorrelated observations, we have $T^{-0.5}(\phi_{i,ML} - \phi_{i,0}) \sim N(0, V)$, where $V = V_1 V_2 V_1'$, $V_2 = E(s_t^i(\phi_i) s_t^i(\phi_i)')_{\phi_i = \phi_{i,0}}$, with the standard correction for V_2, if $y_{[1,t]}$ has correlated observations.

In typical applications $s_t^i(\phi_i)$ are computed with the Kalman filter and are function of martingale difference processes (the shocks of the model). Thus, $\sum_t s_t^i(\phi_i) = 0$ is likely to hold. Further regularity conditions need to be imposed for the arguments to hold precisely (see, e.g. Mueller, 2013).

The composite likelihood geometrically averages different $f_i(y_t, \phi_i)$, each of which is misspecified. Thus, the composite model is, in general, misspecified with density $g(y_{1t}, \ldots, y_{Kt}, \theta, \eta_1, \ldots \eta_K)$ $\equiv g(y_t, \phi) = \prod_i f_i(y_{it}, \phi_i)^{\omega_i}$. Repeating the argument of the previous paragraph, and under regularity conditions discussed in Xu and Reid (2011), when ω_i are fixed, ϕ_{CL}, the composite likelihood estimator, is consistent for $\phi_{0,CL}$, the minimizer of the KL divergence between the g and F. Furthermore, the scaled difference between ϕ_{CL} and $\phi_{CL,0}$ has an asymptotic normal distribution with zero mean and covariance matrix $V_{CL} = V_{CL,1} V_{CL,2} V_{CL,1}'$ where $V_{CL,2} = E(s_{CL,t}(\phi) s_{CL,t}(\phi)')$, $V_{CL,1} = -E[\bigtriangledown_\phi s_{CL,t}(\phi)]$ and $s_{CL,t}(\phi) = \bigtriangledown_\phi \ln g(y_t, \phi)$, all evaluated at $\phi = \phi_{CL}$. When the sub-models have different sample size, one needs to let $\min T_i \to \infty$.

When the weights are random, the asymptotic distribution of ϕ_i depends on $\omega_1, \ldots, \omega_K$. Under standard assumptions that ensure that the estimator of $\omega_1, \ldots \omega_K$ converges to the KL pseudo value $\omega_{10}, \ldots, \omega_{K0}$, and that no ω_{i0} is on the boundary of the parameter space, asymptotic normality still holds but the standard errors for ϕ_{CL} need to be adjusted for the randomness in ω_i. As long as the Godambe matrix is block diagonal in (ϕ, ω), one can ignore this extra uncertainty for inferential purposes.

Appendix B: Issues in quasi-posterior estimation

Drawing ω_i in MCMC algorithm There are various ways to draw candidate weights ω_i, $i = 1, ..., K$. If K is small, an independent Dirichlet proposal works well. When K is large, one could first logistically transform the weights and then use a random walk proposal for the transformed weights. This approach has the disadvantage that the proposal is no longer a multivariate random walk (in particular, it is no longer symmetric). Furthermore, one needs to compute the Jacobian of the mapping, which may be tedious to code and may lead to numerical instabilities because of non-linearities.

Our preferred approach is to use a proposal density which directly operates on the weights. We call it 'random-walk Dirichlet', since the expected value of the proposal is the last accepted draw. Denote by ω^a the last accepted vector of weights, by ω^p a proposal draw, and by $\lambda > 0$ a scalar regulating the variance of the proposal. The proposal density is Dirichlet, denoted by $p_D(\omega^p|\omega^a, \lambda)$, with parameter $\lambda\omega^a$. The mean of this proposal is independent of λ and equal to ω^a. The variance of any element of ω^p is a decreasing function of λ. In an initial adaptive phase, where draws are discarded before computing posterior quantities, we adjust λ so as to achieve a reasonable acceptance probability (20-30%). This proposal density is not symmetric, and thus the acceptance probability needs to be properly modified.

Asymptotic properties of MCMC estimators Let χ_{CL} be the maximum composite likelihood estimator of $\chi = (\theta, \eta_1, \ldots, \eta_K, \omega_1, \ldots, \omega_K)$ and let χ_p be the mode of the prior $p(\chi)$. Suppose both χ_{CL} and χ_p are in the interior of the parameter space. Let $h(\chi_{CL}) = -\nabla^2_\chi \log CL(\chi_{CL}|y_t)$ and $h(\chi_p) = -\nabla^2_\chi \log p(\chi_p)$. Expanding quadratically the composite posterior $p_{CL}(\chi|y_t)$ we have

$$\begin{aligned} &\propto \exp\{\log CL(\chi_{CL}|y_t) - 0.5(\chi - \chi_{CL})^T h(\chi_{CL})(\chi - \chi_{CL}) + \log p(\chi_p) - 0.5(\chi - \chi_p)^T h(\chi_p)(\chi - \chi_p)\} \\ &\approx N(\hat{\chi}, h(\chi_{CL}, \chi_p)^{-1}) \end{aligned} \tag{40}$$

where $\hat{\chi} = h(\chi_{CL}, \chi_p)^{-1}(h(\chi_{CL})\chi_{CL} + h(\chi_p)\chi_p)$ and $h(\chi_{CL}, \chi_p) = h(\chi_{CL}) + h(\chi_p)$.

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

$$T^{-1}h(\chi_{CL}, \chi_p) \to -E(\nabla^2 \log CL(\hat{\chi}_0|y_t)) \equiv H(\hat{\chi}_0) \tag{41}$$

$$\hat{\chi} = (T^{-1}h(\chi_{CL}, \chi_p))^{-1}(T^{-1}h(\chi_{CL})\chi_{CL} + T^{-1}h(\chi_p)\chi_p) \to \hat{\chi}_0 \tag{42}$$

. Thus as $T \to \infty$ $p_{CL}(\chi|y_t) \approx N(\hat{\chi}_0, T^{-1}H(\hat{\chi}_0)^{-1})$. Sufficient conditions that insure that is the case are, for example, in Deblasi and Walker (2013). Rubio and Villaverde (2004) provide conditions which are somewhat easier to verify in practice.

When χ_{CL} is not in the interior of the parameter space, for example, because $\omega_i \to 0$, for some i, η_i may become non-identifiable from the composite likelihood and the above result may not hold. If we let $p(\eta_i) = p(\eta_i|y_{0t})$, where y_{0t} is a training sample of size $\bar{T}$, letting both T an $\bar{T}$ go to infinity, we will have that (41)-(42) hold for identified parameters while for those η_i for which $\omega_i \to 0$, $p_{CL}(\eta_i|y_t, y_{0t}) \approx N(\hat{\eta}_{i0}, \bar{T}^{-1}H(\hat{\eta}_{i0})^{-1})$, where $\hat{\eta}_{i0}$ is the asymptotic pivot of, e.g., ML estimator for η_i in the training sample.

Note that when weak identification problems are present, the above results should be carefully evaluated. In particular, the properties of ω_i may deviate from the standard ones stated in the text.

Appendix C: Tilting vs composite predictors

We look for a predictive density $p(z|y)$ solving:

$$\hat{p} = arg\min_{p} KL(p(z|y), f(z,\phi)) \tag{43}$$

where z is any future sequence of y and ϕ a vector of parameters, subject to the constraint

$$E_p\{\log\frac{f(z|y_t,\phi)}{f(z,\phi)}\} = E_{Z|Y=y}\{\log\frac{f(z|y_t,\phi)}{f(z,\phi)}\} \;\; t = 1,\ldots,T \tag{44}$$

and the normalization $E_p(1) = 1$, where E_p is the expectation with respect to the density $p(z|y)$, and $f(z,\phi)$ is any preliminary density of z, for example, its marginal. In words, we seek for the predictive density which is closest in the KL sense to any preliminary density $f(z,\phi)$ and reproduces the same conditional expectation as the true density $f(z|y,\phi)$ on functions $\log\frac{f(z|y_t,\psi)}{f(z,\phi)}$. Note that when $f(z)$ is disregarded, the problem becomes one of maximizing the entropy -$E_p[\log p(z|y)]$, subject to the constraints (44). The solution is $\hat{p}(z|y) = f(z,\phi)\exp\{\sum_t \xi_t \log\frac{f(z|y_t,\phi)}{f(z,\phi)}\} - \kappa(y_t,\phi,\xi)\}$ where $\kappa(y_t,\phi,\xi)$ is a normalizing constant, ξ_t are the Lagrange multipliers on the constraints (44). $\hat{p}(z|y)$ has an exponential tilting format: we tilt $f(z,\phi)$ in the directions spanned by $\log\frac{f(z|y_i,\phi)}{f(z,\phi)}$. If $\xi_t \geq 0, \sum_t \xi_t \leq 1$, then $\hat{p}(z|y)$ is the scaled version of the composite predictive density derived in section 4.5 with $\omega_t = \xi_t,\ t = 1,\ldots,T$ and $\omega_0 = 1 - \sum_t \xi_t$, where ω_0 is the weight on $f(z,\phi)$. Note that in this setup, ω_t satisfies the

following (score) equation:

$$\frac{\partial E_{z|Y=y} \log f_p(Z|y,\phi,\omega_t)}{\partial \omega_t} = 0, \quad t = 1,\ldots,T \tag{45}$$

Thus, it can be chosen to maximize the conditional expected logarithmic score (45).

Appendix D: Choosing the composite pool when K is large

In the paper we have assumed that K is given (and small) and that a researcher includes all models in the composite likelihood. Here we discuss how to choose the optimal combination of models entering the composite likelihood when K is large. That is, how to choose both the dimensionality of the composite pool and the models entering the pool. This problem may be relevant when the information provided by the models is not necessarily independent. In this case, there may be a trade-off between the number of models to be included and the estimation gains that can be obtained with a composite methods when the K models are all misspecified.

Let $S = \sum_{k=2}^{K-2} \frac{k!}{r!(k-r)!}$ be an index for the composite combination, where we allow a minimum of r=2 models to appear in the composite pool, let $y = y_1 = y_2 = \ldots = y_S$ and let ω_i be fixed. Let α_s be the weight on combination $s = 1,\ldots,S$. When the prior for α_s is proportional to the expected Kullback-Leibler (KL) divergence between that combination and the best fitting pool $D(s, s_0)$, and the prior for the parameters $p(\phi_s)$ satisfies standard regularity conditions that allow the composite marginal likelihood to be computed in the neighbor of the composite likelihood estimator, one can follow the steps of Lv and Liu (2014) and show that a Laplace expansions of the marginal composite likelihood leads to the generalized BIC criteria:

$$GBIC_{s,CL} = -2CL(\phi_{s,CL}, y_t) + 1 + 2dim(\phi_{s,CL}) \log T_s + 2I(H_s, J_s) \tag{46}$$

where $I(H_s, J_s) = \frac{1}{2}(tr(Q_s) - \ln|Q_s| - dim(\phi_s))$, $Q_s = J_s^{-1} H_s$. $I(H_s, J_s)$ is the log of the KL divergence between two $dim(\phi_s)$ vectors of normal random variables, one with zero mean and covariance J_s and one with zero mean and covariance H_s, where J_s and H_s are the variability and the sensitivity matrices of the composite combination s.

(46) features the two standard elements of a BIC criteria (a measure of fit, and a term penalizing model complexity) and an additional term reflecting composite model misspecification relative to the best fitting model (in a KL sense). When the composite model $\bar{s}$ is correctly specified, $J_{\bar{s}} \approx H_{\bar{s}}$, $I(J_{\bar{s}}, H_{\bar{s}}) \cong 0$, and GBIC=BIC. Thus, there are three dimensions that

matter when choosing composite pools: fit, dimensionality, and misspecification.

When models do not share the same observables, the expansion in (46) becomes non-comparable across composite pools - the measure of fit will always be weakly higher for a pool that includes larger scale models. To apply (46), the measure of fit must be restricted to the same observables. This can be done, solving out equations until the same observables are present in all the models. While ω informs us about the relative support of a model in the composite pool, $I(H_s, J_s)$ tells us about the relative misspecification of different estimation pools.

If we allow composite pools to include just one model and assume that the prior on the pool s has a Dirichlet format with $\alpha_1 = \alpha_2 = ... = \alpha_S$, the expression in (46) simplifies to

$$GBIC_{s,CL} = -2CL(\phi_{s,CL}, y_t) + 2dim(\phi_{s,CL}) \log T_s - \ln |Q_s| \tag{47}$$

(47) can be used to compare composite vs. maximum likelihood estimators of the parameters. Also in this case, there will be three terms that matter: the fit of each model as measured by the maximized value of the likelihood relative to the maximized value of the marginal likelihood; a penalty for the relative dimensionality of the parameter space; and a term reflecting misspecification. Note that $\ln |Q_s|$ provides an approximation to the KL divergence. Thus, if the $\bar{s}$-model is correctly specified, $Q_s = I_{dim(\phi_s)}$ and $\ln |Q_s| = 0$. Once again, (47) can be applied only to models sharing the same observables.

Appendix E: Models of section 5.1

1) Basic model with quadratic preferences, constant interest rate, exogenous permanent and transitory income process. Let $G = 1 + g$ be the growth rate of permanent income. Let $\tilde{c}_t = \frac{c_t}{y_t^P}; \tilde{a}_t = \frac{a_t}{y_t^P}, y_t = y_t^T y_t^P$. The log linearized conditions are

$$\hat{\tilde{c}}_t = \hat{e}_{2t+1} + \hat{\tilde{c}}_{t+1} \tag{48}$$

$$\hat{\tilde{a}}_t = \frac{1}{\bar{a}/G + \bar{y}^T - \bar{c}}(\bar{a}/G\hat{\tilde{a}}_{t-1} - \bar{a}/G\hat{e}_{2t} + \bar{y}\hat{y}_t^T - \bar{c}\hat{\tilde{c}}_t) \tag{49}$$

$$\hat{y}_t^P = \hat{y}_{t-1}^P + \hat{e}_{2t} \tag{50}$$

$$\hat{y}_t^T = \rho\hat{y}_{t-1}^T + \hat{e}_{1t} \tag{51}$$

$$\hat{c}_t = \hat{\tilde{c}}_t + \hat{y}_t^P \tag{52}$$

$$\hat{a}_t = \hat{\tilde{a}}_t + \hat{y}_t^P \tag{53}$$

where c_t is consumption, a_t are savings, y_t is income,ρ the persistence of transitory income, $(1+r)$ the gross real rate of interest $(1+r)\beta = 1$, $\sigma_i, i = 1,2$ the standard deviation of the transitory and permanent income, and variables with a bar indicate steady state quantities.

2) Model with exponential utility, constant interest rate, exogenous permanent and transitory income process. The instantaneous utility function is $u(c) = \frac{-1}{\theta}\exp(-\theta c_t)$, where $\theta > 0$ is the coefficient of risk aversion. The log linearized equations are:

$$
\begin{aligned}
-\hat{c}_t &= -\hat{c}_{t+1} + \frac{1}{\theta\bar{c}}(\hat{\sigma}_t + \hat{e}_{2t}) && (54)\\
\hat{\tilde{a}}_t &= \frac{1}{\frac{\bar{a}}{G*\bar{\sigma}} + \bar{y}^T - \bar{\tilde{c}}}\left(\frac{\bar{a}}{G\bar{\sigma}}\hat{\tilde{a}}_{t-1} - \frac{\bar{a}}{G\bar{\sigma}}\hat{e}_{2t} - \frac{\bar{a}}{G}\hat{\sigma}_t + \bar{y^T}\hat{y}_t^T - \bar{\tilde{c}}\hat{\tilde{c}}_t\right) && (55)\\
\hat{y}_t^P &= \hat{y}_{t-1}^P + \hat{\sigma}_t + \hat{e}_{2t} && (56)\\
\hat{y}_t^T &= \rho_1\hat{y}_{t-1}^T + \hat{\sigma}_t + \hat{e}_{1t} && (57)\\
\hat{\sigma}_t &= \rho_2\hat{\sigma}_{t-1} + \hat{e}_{3t} && (58)\\
\hat{c}_t &= \hat{\tilde{c}}_t + \hat{y}_t^P && (59)\\
\hat{a}_t &= \hat{\tilde{a}}_t + \hat{y}_t^P && (60)
\end{aligned}
$$

where σ_t is the standard deviation of the permanent and transitory income shock, and ρ_2 the persistence of the volatility process.

3) RBC model with separable CRRA preferences, labor supply decisions, capital accumulation, endogenous interest rate, permanent and transitory technology shocks. Letting α be the share of capital in production, γ the risk aversion coefficient, δ the capital depreciation rate, η the inverse of the Frish elasticity of labor supply. and assuming that $\log e_{2t}$

has zero mean, the log-linearized conditions are

$$\begin{aligned}
\gamma\hat{\tilde{c}}_t + \eta\hat{N}_t &= \hat{\tilde{Y}}_t - \hat{N}_t && (61)\\
-\gamma\hat{\tilde{c}}_t &= (1-\gamma)\hat{e}_{2t+1} - \gamma\hat{\tilde{c}}_{t+1} + \frac{r}{1+r}\hat{r}_{t+1} && (62)\\
\hat{r}_t &= \frac{\alpha}{1+r}(\hat{\tilde{Y}}_t - \hat{\tilde{K}}_{t-1}) && (63)\\
\hat{\tilde{Y}}_t &= \alpha(\hat{\tilde{K}}_{t-1}) + (1-\alpha)(\hat{N}_t + \hat{\zeta}_t^T) && (64)\\
\hat{\tilde{Y}}_t &= \frac{\bar{c}}{\bar{Y}}\hat{\tilde{c}}_t + \frac{\bar{K}}{\bar{Y}}\hat{\tilde{K}}_t + \frac{(1-\delta)}{G}\frac{\bar{K}}{\bar{Y}}\hat{\tilde{K}}_{t-1} - \frac{(1-\delta)}{G}\frac{\bar{K}}{\bar{Y}}\hat{e}_{2t+1} && (65)\\
\hat{\zeta}_t^P &= G + \hat{\zeta}_{t-1}^P + \hat{e}_{2t} && (66)\\
\hat{\zeta}_t^T &= \rho\hat{\zeta_{t-1}^T} + \hat{e}_{1t} && (67)\\
\hat{c}_t &= \hat{\tilde{c}}_t + \hat{y}_t^P && (68)\\
\hat{k}_t &= \hat{\tilde{k}}_t + \hat{y}_t^P && (69)\\
\hat{y}_t &= \hat{\tilde{y}}_t + \hat{y}_t^P && (70)
\end{aligned}$$

where k_t is the capital stock and N_t is hours, ζ_t the technology disturbance and ρ the persistence of its transitory component.

4) Model with two types of agents optimizers and Rule of thumb (ROT) consumers, constant interest rate, permanent and transitory income components. Let $1-\omega$ be the share of ROT consumers. The log linearized conditions are

$$\begin{aligned}
-\gamma\hat{\tilde{c}}_{1t} &= (1-\gamma)\hat{e}_{2t+1} - \gamma\hat{\tilde{c}}_{1t+1} && (71)\\
\hat{\tilde{c}}_t^{ROT} &= \hat{y}_t^T && (72)\\
\hat{\tilde{a}}_t &= \frac{1}{\bar{a}/G + \bar{y}^T - \bar{c}}(\bar{a}/G\hat{\tilde{a}}_{t-1} - \bar{a}/G\hat{e}_{2t} + \bar{y}\hat{y}_t^T - \bar{c}_1\hat{\tilde{c}}_{1t}) && (73)\\
\hat{y}_t^P &= G + \hat{y}_{t-1}^P + \hat{e}_{2t} && (74)\\
\hat{y}_t^T &= \rho\hat{y}_{t-1}^T + \hat{e}_{1t} && (75)\\
\hat{c}_{1t} &= \hat{\tilde{c}}_{1t} + \hat{y}_t^P && (76)\\
\hat{c}_t^{ROT} &= \hat{\tilde{c}}_t^{ROT} + \hat{y}_T^P && (77)\\
\hat{c}_t &= \omega\hat{c}_{1t} + (1-\omega)\hat{c}_t^{ROT} && (78)\\
\hat{a}_t &= \hat{\tilde{a}}_t + \hat{y}_t^P && (79)
\end{aligned}$$

where γ is the coefficient of relative risk aversion and the superscript ROT indicate the variables of the agents which do not save. We calibrate $\omega = 0.2, (1+r) = 1.01$.

5) Model with two types of optimizing agents, liquidity and non-liquidity constrained, constant interest rate, permanent and transitory income components. Utility depends on durable and non-durable consumption, relative price of non-durable is exogenous. The log-linear conditions are

$$\begin{aligned}
\hat{\hat{c}}_{1t} - \hat{\hat{d}}_{1t} &= \\
\frac{1}{\zeta_1}(\hat{p}_t - \frac{1-\delta}{1+r}\hat{p}_{t+1}) & \qquad (80)\\
\hat{\hat{a}}_{1t} - \hat{\hat{d}}_{1t} - \hat{p}_t &= 0 \qquad (81)\\
\bar{p}\bar{d}_1\delta(\hat{p}_t + \hat{\hat{d}}_{1t}) + \bar{c}_1\hat{\hat{c}}_{1t} + \bar{a}_1\hat{\hat{a}}_{1t} &= \\
(1+r)\bar{a}_1\hat{\hat{a}}_{1t-1} + \hat{y}_t^T - [(1+r)\bar{a}_1 - (1-\delta)\bar{p}\bar{d}_1)]\hat{e}_{2t} & \qquad (82)\\
\hat{\hat{c}}_{2t} - \hat{\hat{d}}_{2t} &= \\
\frac{1}{\zeta_2}(\hat{p}_t(1+\psi(\beta_2(1+r)-1)) - \beta_2(1-\delta)\hat{p}_{t+1} &+ \\
(\gamma-1)\beta_2[\psi(1+r)-(1-\delta)](\bar{c}_2\hat{\hat{c}}_{2t+1} - \bar{c}_2\hat{\hat{c}}_{2t} - \bar{d}_2\hat{\hat{d}}_{2t+1} + \bar{d}_2\hat{\hat{d}}_{2t+1})) & \qquad (83)\\
\bar{p}\bar{d}_2\delta(\hat{p}_t + \hat{\hat{d}}_{2t}) + \bar{c}_2\hat{\hat{c}}_{2t} + \bar{a}_2\hat{\hat{a}}_{2t} &= \\
(1+r)\bar{a}_2\hat{\hat{a}}_{2t-1} + \hat{y}_t^T - [(1+r)\bar{a}_2 - (1-\delta)\bar{p}\bar{d}_2)]\hat{e}_{2t} & \qquad (84)\\
\frac{\bar{a}_2}{B}\hat{\hat{a}}_{2t} + \frac{1-\bar{a}_2}{B}(\hat{p}_t + \hat{\hat{d}}_{2t}) &= 0 \qquad (85)
\end{aligned}$$

These equations have six unknowns $(\hat{\hat{c}}_{1t}, \hat{\hat{c}}_{2t}, \hat{\hat{d}}_{1t}, \hat{\hat{d}}_{2t}, \hat{\hat{a}}_{1t}, \hat{\hat{a}}_{2t})$, given y_t^T, y_t^P, p_t. The remaining

equations are

$$
\begin{aligned}
(\gamma-1)(\bar{c}_2\hat{\tilde{c}}_{2t}-\bar{d}_2\hat{\tilde{d}}_{2t}) &+ \\
\frac{1}{\beta_2(1+r)-1}(\beta_2(1+r)(\gamma-1)(\bar{c}_2\hat{\tilde{c}}_{2t+1}-\bar{c}_2\hat{\tilde{c}}_{2t}-\bar{d}_2\hat{\tilde{d}}_{2t+1}+\bar{d}_2\hat{\tilde{d}}_{2t})) &= \hat{\mu}_t && (86)\\
\hat{\tilde{c}}_{it}+\hat{y}_t^P &= \hat{c}_{it} && (87)\\
\hat{\tilde{a}}_{it}+\hat{y}_t^P &= \hat{a}_{it} && (88)\\
\hat{\tilde{d}}_{it}+\hat{y}_t^P &= \hat{d}_{it} && (89)\\
\hat{y}_{t-1}^P+\hat{e}_{2t} &= \hat{y}_t^P && (90)\\
\rho_1\hat{y}_{t-1}^T+\hat{e}_{1t} &= \hat{y}_t^T && (91)\\
\rho_2\hat{p}_{t-1}+\hat{e}_{3t} &= \hat{p}_t && (92)\\
\hat{y}_t^P+\hat{y}_t^T &= \hat{y}_t && (93)\\
\omega\hat{c}_{1t}+(1-\omega)\hat{c}_{2t} &= \hat{c}_t && (94)\\
\omega\hat{a}_{1t}+(1-\omega)\hat{a}_{2t} &= \hat{a}_t && (95)\\
\omega\hat{d}_{1t}+(1-\omega)\hat{d}_{2t} &= \hat{d}_t && (96)
\end{aligned}
$$

where i=1,2, $1-\omega$ is the share of liquidity constrained consumers and γ the Cobb-Douglass share of non-durable good d_{it} in the utility. $\zeta_1=(1-\frac{1-\delta}{1+r}), \zeta_2=(1-\beta_2((1-\delta)-\psi(1+r))-1)$, ψ is the share of durable financiable with assets, $\beta_2>\beta_1$ and $\hat{\mu}$ is the Lagrange multiplier on the liquidity constraint (in percentage deviation from steady states). We calibrate $\omega=0.2, \psi=0.95, B=0.05, (1+r)=1.01$.

Appendix F: Models of section 5.2

1) Herbst and Schorfheide (2015) model

$$
\begin{aligned}
y_t &= E_t(y_{t+1})-\frac{1}{\tau}(R_t-E_t(\pi_{t+1})-E_t(z_{t+1}))+g_t-E_t(g_{t+1}) && (97)\\
\pi_t &= \beta E_t(\pi_{t+1})+\kappa(y_t-g_t) && (98)\\
R_t &= \rho_R R_{t-1}+(1-\rho_R)(\phi_1\pi_t+\phi_2(y_t-g_t))+\varepsilon_{R,t} && (99)\\
z_t &= \rho_z z_{t-1}+\varepsilon_{z,t} && (100)\\
g_t &= \rho_g g_{t-1}+\varepsilon_{g,t} && (101)
\end{aligned}
$$

where y_t is output, π_t inflation, R_t the nominal rate, z_t a technology shock, g_t a demand shock

and $e_{R,t}$ a monetary policy shock.

2) Justiniano, Primiceri and Tambalotti (2010) model

$$\hat{y}_t = \frac{y+F}{y}\left[\alpha\hat{k}_t + (1-\alpha)\,\hat{L}_t\right] \tag{102}$$

$$\hat{\rho}_t = \hat{w}_t + \hat{L}_t - \hat{k}_t \tag{103}$$

$$\hat{s}_t = \alpha\hat{\rho}_t + (1-\alpha)\,\hat{w}_t \tag{104}$$

$$\hat{\pi}_t = \gamma_f E_t\hat{\pi}_{t+1} + \gamma_b\hat{\pi}_{t-1} + \kappa\hat{s}_t + \kappa\hat{\lambda}_{p,t} \tag{105}$$

$$\hat{\lambda}_t = \frac{h\beta e^{\gamma}}{(e^{\gamma}-h\beta)(e^{\gamma}-h)}E_t\hat{c}_{t+1} - \frac{e^{2\gamma}+h^2\beta}{(e^{\gamma}-h\beta)(e^{\gamma}-h)}\hat{c}_t + \frac{he^{\gamma}}{(e^{\gamma}-h\beta)(e^{\gamma}-h)}\hat{c}_{t-1} \tag{106}$$

$$+ \frac{h\beta e^{\gamma}\rho_z - he^{\gamma}}{(e^{\gamma}-h\beta)(e^{\gamma}-h)}\hat{z}_t + \frac{e^{\gamma}-h\beta\rho_b}{e^{\gamma}-h\beta}\hat{b}_t \tag{107}$$

$$\hat{\lambda}_t = \hat{R}_t + E_t\left(\hat{\lambda}_{t+1} - \hat{z}_{t+1} - \hat{\pi}_{t+1}\right) \tag{108}$$

$$\hat{\rho}_t = \chi\hat{u}_t \tag{109}$$

$$\hat{\phi}_t = (1-\delta)\,\beta e^{-\gamma}E_t\left(\hat{\phi}_{t+1} - \hat{z}_{t+1}\right) + \left(1-(1-\delta)\,\beta e^{-\gamma}\right)E_t\left[\hat{\lambda}_{t+1} - \hat{z}_{t+1} + \hat{\rho}_{t+1}\right] \tag{110}$$

$$\hat{\lambda}_t = \hat{\phi}_t + \hat{u}_t - e^{2\gamma}S''\left(\hat{\imath}_t - \hat{\imath}_{t-1} + \hat{z}_t\right) + \beta e^{2\gamma}S''E_t\left[\hat{\imath}_{t+1} - \hat{\imath}_t + \hat{z}_{t+1}\right] \tag{111}$$

$$\hat{k}_t = \hat{u}_t + \hat{\bar{k}}_{t-1} - \hat{z}_t \tag{112}$$

$$\hat{\bar{k}}_t = (1-\delta)\,e^{-\gamma}\left(\hat{\bar{k}}_{t-1} - \hat{z}_t\right) + \left(1-(1-\delta)\,e^{-\gamma}\right)(\hat{u}_t + \hat{\imath}_t) \tag{113}$$

$$\hat{w}_t = \frac{1}{1+\beta}\hat{w}_{t-1} + \frac{\beta}{1+\beta}E_t\hat{w}_{t+1} - \kappa_w\hat{g}_{w,t} + \tag{114}$$

$$+ \frac{\iota_w}{1+\beta}\hat{\pi}_{t-1} + \frac{1+\beta\iota_w}{1+\beta}\pi_t + \frac{\beta}{1+\beta}E_t\hat{\pi}_{t+1} + \tag{115}$$

$$+ \frac{\iota_w}{1+\beta}z_{t-1} - \frac{1+\beta\iota_w-\rho_z\beta}{1+\beta}z_t + \kappa_w\hat{\lambda}_{w,t} \tag{116}$$

$$\hat{g}_{w,t} = \hat{w}_t - \left(\nu\hat{L}_t + \hat{b}_t - \hat{\lambda}_t\right) \tag{117}$$

$$\hat{R}_t = \rho_R\hat{R}_{t-1} + (1-\rho_R)\left[\phi_\pi\hat{\pi}_t + \phi_X\left(\hat{x}_t - \hat{x}_t^*\right)\right] + \phi_{dX}\left[\left(\hat{x}_t - \hat{x}_{t-1}\right) - \left(\hat{x}_t^* - \hat{x}_{t-1}^*\right)\right] + \hat{\eta}_{mp,t} \tag{118}$$

$$\hat{x}_t = \hat{y}_t - \frac{\rho k}{y}\hat{u}_t \tag{119}$$

$$\frac{1}{g}\hat{y}_t = \frac{1}{g}\hat{g}_t + \frac{c}{y}\hat{c}_t + \frac{i}{y}\hat{\imath}_t + \frac{\rho k}{y}\hat{u}_t \tag{120}$$

Additional References

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

Deblasi, P. and S. Walker (2013) Bayesian asymptotics with misspecified models. *Statistica Sinica*, 23, 169-187.

Lv, J. and J. Liu (2014) Model selection principles in misspecified models. *Journal of the Royal Statistical Society*, 76, part 1, 141-167

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

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

J. Fernandez Vilaverde and J. Rubio Ramirez (2004). Comparing dynamic general equilibrium models to data: A bayesian approach.*Journal of Econometrics*, 123, 128-157.

Xu, X. and N. Reid (2011) On the robustness of the maximum composite likelihood estimator. *Journal of Statistical Planning and Inference*, 141, 3047-3054.