Australian National University

Crawford School of Public Policy

CAMA

Centre for Applied Macroeconomic Analysis

Composite Likelihood Methods for Large Bayesian VARs with Stochastic Volatility

---

CAMA Working Paper 26/2018
May 2018

Joshua C.C. Chan
University of Technology Sydney
Centre for Applied Macroeconomic Analysis, ANU

Eric Eisenstat
The University of Queensland

Chenghan Hou
Hunan University

Gary Koop
University of Strathclyde

Abstract

Adding multivariate stochastic volatility of a flexible form to large Vector Autoregressions (VARs) involving over a hundred variables has proved challenging due to computational considerations and over-parameterization concerns. The existing literature either works with homoskedastic models or smaller models with restrictive forms for the stochastic volatility. In this paper, we develop composite likelihood methods for large VARs with multivariate stochastic volatility. These involve estimating large numbers of parsimonious models and then taking a weighted average across these models. We discuss various schemes for choosing the weights. In our empirical work involving VARs of up to 196 variables, we show that composite likelihood methods have similar properties to existing alternatives used with small data sets in that they estimate the multivariate stochastic volatility in a flexible and realistic manner and they forecast comparably. In very high dimensional VARs, they are computationally feasible where other approaches involving stochastic volatility are not and produce superior forecasts than natural conjugate prior homoscedastic VARs.

Keywords

Bayesian, large VAR, composite likelihood, prediction pools, stochastic volatility

JEL Classification

C11, C32, C53

Address for correspondence:

(E) cama.admin@anu.edu.au

ISSN 2206-0332

Composite Likelihood Methods for Large Bayesian VARs with Stochastic Volatility*

Joshua C.C. Chan
University of Technology Sydney and CAMA

Eric Eisenstat
The University of Queensland

Chenghan Hou
Hunan University

Gary Koop
University of Strathclyde

May 2018

Abstract: Adding multivariate stochastic volatility of a flexible form to large Vector Autoregressions (VARs) involving over a hundred variables has proved challenging due to computational considerations and over-parameterization concerns. The existing literature either works with homoskedastic models or smaller models with restrictive forms for the stochastic volatility. In this paper, we develop composite likelihood methods for large VARs with multivariate stochastic volatility. These involve estimating large numbers of parsimonious models and then taking a weighted average across these models. We discuss various schemes for choosing the weights. In our empirical work involving VARs of up to 196 variables, we show that composite likelihood methods have similar properties to existing alternatives used with small data sets in that they estimate the multivariate stochastic volatility in a flexible and realistic manner and they forecast comparably. In very high dimensional VARs, they are computationally feasible where other approaches involving stochastic volatility are not and produce superior forecasts than natural conjugate prior homoskedastic VARs.

Keywords: Bayesian, large VAR, composite likelihood, prediction pools, stochastic volatility

JEL Classifications: C11, C32, C53

*We would like to thank participants at the Workshop on Forecasting at the Deutsche Bundesbank (September 2017) for helpful comments. Gary Koop is a Senior Fellow at the Rimini Center for Economic Analysis. Joshua Chan and Eric Eisenstat would like to acknowledge financial support by the Australian Research Council via a Discovery Project (DP180102373).
Emails: joshuacc.chan@gmail.com, e.eisenstat@uq.edu.au, chenghan.hou@hotmail.com and gary.koop@strath.ac.uk.

1 Introduction

Empirical macroeconomists increasingly use Vector Autoregressions (VARs) with datasets involving a hundred or more variables. However, in many applications, there is strong evidence for stochastic volatility. Adding multivariate stochastic volatility to large VARs has proved challenging. With small data sets, popular specifications for VARs with stochastic volatility (VAR-SV) exist but Bayesian estimation and forecasting with these VAR-SV's is not computationally feasible with large data sets. This paper develops a computationally feasible approach with uses all the data available and allows for stochastic volatility. The idea of our approach is to work with many small VAR-SVs. Each of these contains only a few of the large number of variables available, but everyone one of the variables appears in one or more of the small models. Forecasts from the many small models are then combined to produce forecasts which reflect all the data available. We use composite likelihood methods to theoretically justify and implement such a strategy.

The fact that large VARs are being found increasingly useful in an era of Big Data needs little justification. The large VAR literature began with the US macroeconomic application of Banbura, Giannone and Reichlin (2010) but large VARs are now used with similar macroeconomic data sets for other countries (e.g. Bloor and Matheson, 2010). There are also applications where large VARs arise due to the need to build a model involving variables for many countries (e.g. Carriero, Kapetanios and Marcellino, 2010, and Koop and Korobilis, 2016). In addition, large VARs have arisen through having to deal with many related variants of a single variable (e.g. interest rates of different maturities or the different components that make up an inflation index), see Carriero, Kapetanios and Marcellino (2012) or Giannone, Lenza, Momferatou and Onorante (2014). They can also arise through the use of mixed frequency data (e.g. McCracken, Owyang and Sekhposyan, 2016). Large VARs have also been used for structural economic analysis or scenario forecasting in papers such as Bańbura, Giannone and Lenza (2015) and Jarociński and Maćkowiak (2016). In short, large VARs are increasingly used for a plethora of purposes and are promising to become one of the major tools of modern empirical macroeconomics.

Similarly, the facts that macroeconomic variables often exhibit structural instabilities and have variances that change over time is increasingly accepted. Papers such as Clark (2011) highlight the particular importance in macroeconomic applications of allowing for time-variation in the error covariance matrix. Hence, this is what we focus on in this paper (although the econometric methods we develop could also be used with the time-varying parameter VAR). Since the elements of this matrix enter impulse responses and have a large impact on predictive variances, use of mis-specified homoskedastic models can lead to invalid structural inference and poor forecasts.

The arguments in the preceding two paragraphs justify why there is a desire to work with large VAR-SVs. But Bayesian methods, requiring the use of Markov Chain Monte Carlo (MCMC) methods, quickly become computationally infeasible as the number of variables in the VAR increases. Bayesian methods are typically used with large VARs since they allow for prior shrinkage which is of great use with over-parameterized models. For instance, when working with a large VAR with $N = 100$ variables and a lag length of $p = 13$ (as might be required with monthly data), the researcher will have over $100,000$ VAR coefficients to estimate and $5,050$ free parameters in the error covariance matrix. Bayesian prior shrinkage, often using natural conjugate or Minnesota priors, is used to surmount the problems caused by a shortage of data information relative to the number of coefficients being estimated. Even with these priors, which imply that the posterior and one-step ahead predictive densities have analytical forms, the researcher can face a substantial computational burden. The main computational bottleneck is dealing with the huge posterior covariance matrix of the VAR coefficients (even in the absence of deterministic terms it is an $N^2p \times N^2p$ matrix).[1] Use of the natural conjugate prior in a standard homoskedastic VAR leads to a particular Kronecker product form for the posterior covariance matrix involving separate $N \times N$ and $Np \times Np$ matrices which can be manipulated independently of one another (see Chan, 2018 or Carriero, Clark and Marcellino, 2016a). This vastly simplifies computation. The problem is that small departures from the natural conjugate prior VAR destroy the Kronecker structure and, thus, lead to huge increases in the computational burden. With large VARs this makes many sensible alternative approaches untenable. This holds true for alternative approaches using less subjective priors that allow for automatic shrinkage of coefficients found to be unimportant (e.g. the variable selection prior of George, Sun and Ni, 2008, Koop, 2013 and Korobilis, 2013, or the Lasso prior of Gefang, 2014). It also holds true for specifications which allow for time-variation in parameters. It is the latter which is the focus of the present paper.

As noted, with large VARs standard approaches (e.g. Primiceri, 2005) which allow for multivariate stochastic volatility are not computationally feasible. But there are stochastic volatility specifications that can be used with larger VARs (e.g. Chan 2018 and Carriero, Clark and Marcellino, 2016a,b,c). However, these place restrictions on the form of time variation allowed for. And even these have a large computational burden which means they cannot be used for forecasting with the large VARs involving hundreds of dependent variables which are increasingly being used.[2]

These considerations motivate the present paper. Working with many small VAR-SVs is computationally feasible even with very high dimensional data sets

[1] It is worth stressing that the main computational hurdle does not relate to the error covariance matrix but the VAR coefficients. In finance, there are several methods (see, among many others, Creal and Tsay, 2015) for dealing with large-dimensional covariance matrices (e.g. involving asset returns for a huge number of assets) in models where the conditional means of the dependent variables are of low dimension (often zero). These are not relevant for our purposes.

[2] Perhaps the best of the current approaches is developed in Carriero, Clark and Marcellino (2016b). In this paper, impulse responses are presented using a 125 variable VAR, but when forecasting only a 20 variable VAR is used. Repeatedly forecasting with this model on an expanding window of data with the 196 variables used in this paper would take months or more of computer time on a good PC.

and composite likelihood methods can be used to combine forecasts from these many smaller models. So the methods we develop are practical and can be used with hundreds of variables. But we also address several other questions to further strengthen the case for our composite likelihood based methods. The first of these is whether there is a theoretically strong justification for use of composite likelihood methods in our context. We discuss relevant econometric theory in the next section of the paper. The second question is: How should the various small models that arise with composite likelihood methods be combined? This question we also address in the next section of the paper. In particular, we discuss various methods for doing so, drawing on the literature on opinion pools. The third question is: How well do these methods work in practice? We answer this using a large quarterly US macroeconomic data set involving 196 variables. We find our composite likelihood methods to forecast substantially better than the only computationally practical competitor: a homoskedastic VAR using a natural conjugate prior. We would like to compare our methods to other approaches which involve multivariate stochastic volatility using this large data set, but cannot do since the computational burden of popular Bayesian alternatives is too large. Instead, we compare our methods to a range of different Bayesian VARs with multivariate stochastic volatility using a small data set involving 7 variables. We find parameter estimates produced by our approach to be very similar to those produced by these alternatives. We also carry out a small Monte Carlo study which offers additional reassurance that the approximation inherent in the use of composite likelihood methods with VAR-SVs is an accurate one. We also find our (large data set) composite likelihood methods to forecast slightly better than the (small data set) Bayesian VAR-SV alternatives.

2 Composite Likelihood Methods for large VARs with Stochastic Volatility

2.1 Overview

A traditional likelihood function is based on the p.d.f. of the $N \times 1$ vector of dependent variables, y_t for $t = 1, .., T$. In many empirical cases, particularly if N is large, computation involving a likelihood function can be difficult or infeasible. In such cases, it may be possible to develop statistical methods for estimation of the parameters or forecasting using the composite likelihood instead of the full likelihood. The composite likelihood is built up as a weighted average of likelihoods for $y_{i,t}$ for $i = 1, .., M$ which are sub-vectors of y_t. The likelihoods for these sub-vectors are often called quasi-likelihoods and we will use this terminology. Bayesian methods can then be used by combining a prior with the composite likelihood in the standard way. Thus, if $y_{i,t}$ is of much lower dimension than y_t, a computationally difficult problem of working with a high dimensional likelihood can be turned into a much simpler one of working with many small quasi-likelihoods.

The statistical literature on composite likelihood methods (see, e.g., Varin, Reid and Firth, 2011, Ribatet, Cooley and Davison, 2012 and Roche 2016) provides theoretical and empirical justification for working with the composite likelihood function. For instance, depending on how the quasi-likelihoods are chosen, the composite likelihood can be shown to asymptotically converge to the likelihood function suggesting that, in such cases, the composite likelihood can provide a good approximation to the likelihood function. In addition, composite likelihood methods can be useful for reasons of robustness. That is, with high dimensional models, there are more ways to become mis-specified than with low dimensional densities and, thus, working with the latter can be more robust.

Composite likelihood methods can also have advantages in terms of parsimony. That is, high dimensional models like large VARs are hugely over-parameterized. The correct specification is likely a highly restricted version of the large VAR. The existing Bayesian large VAR literature tries to overcome this problem through the use of prior shrinkage. But this prior information often has to be very strong to obtain reasonable forecasts. For instance, the popular Minnesota prior involves a shrinkage parameter. Giannone, Lenza and Primiceri (2015) develop a method for estimating this shrinkage parameter and show how it becomes small (indicating stronger prior shrinkage) as the size of the data set increases. Using composite likelihood methods, we are only working with small VARs which can yield good forecasts even in the absence of strong prior information.

Another way of conceptualizing the over-parameterization issue is to note that there are various ways of treating this issue in large VARs. Some researchers use prior shrinkage to address it while others use parametric restrictions. These are useful approaches, but each has its possible drawbacks (i.e. sensitivity to prior and risk of mis-specification, respectively). Composite likelihood methods offer another avenue for inducing parsimony: working only with small models based on the chosen quasi-likelihoods. This approach, too, requires the researcher to make choices. But it may be easier to make (and justify) good choices for quasi-likelihoods than for a particular prior or parametric restriction. For instance, below we argue that the desire to forecast a core set of variables of interest in the context of a large data set which includes many other variables motivates our particular choice of quasi-likelihoods. Beyond the choice of quasi-likelihoods, we require no additional choices to be made about prior or parametric restrictions.

The preceding paragraphs provide the basic justifications and insights that underlie the methods we use in this paper and which we elaborate on in the remainder of this section. Composite likelihood methods have been exploited in several fields. For instance, Pakel, Shephard, Sheppard and Engle (2014) is a financial application involving a large number of stock returns. These methods have also been used in spatial statistics (e.g. Ribatet, Cooley and Davison, 2012). But they have been rarely used in macroeconomics. Two exceptions to this lie in the field of Dynamic Stochastic General Equilibrium (DSGE) modelling: Canova and Matthes (2017) and Qu (2016). To our knowledge, our paper is the first to use them in the large VAR field in order to add flexible and computationally feasible forms of multivariate stochastic volatility to large VARs.

2.2 The VAR-SV

We begin by defining the VAR-SVs that our quasi-likelihoods are based on. Specifications identical or similar to this have been used in a huge range of papers, including Primiceri (2005), Koop, Leon-Gonzalez and Strachan (2009), Clark (2011), D'Agostino, Gambetti and Giannone (2013) and Chan and Eisenstat (2018). The VAR-SV model can be written as:

$$A_{0t}y_t = c + A_1 y_{t-1} + \cdots + A_p y_{t-p} + \epsilon_t,$$

where c is an $N \times 1$ vector of intercepts, $A_1, \ldots, A_p$ are $N \times N$ matrices of VAR coefficients, $\Sigma_t = \text{diag}\left(e^{h_{1,t}}, \ldots, e^{h_{n,t}}\right)$ and A_{0t} is a time varying $N \times N$ lower triangular matrix with ones on the diagonal, to be specific,

$$A_{0t} = \begin{pmatrix} 1 & 0 & \cdots & 0 \\ a_{21,t} & 1 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ a_{n1,t} & a_{n2,t} & \cdots & 1 \end{pmatrix}.$$

It is convenient to re-write the VAR-SV as

$$y_t = X_t\beta + W_t a_t + \epsilon_t, \quad \epsilon_t \sim N(0, \Sigma_t), \tag{1}$$

where $X_t = I_n \otimes (1, y'_{t-1}, \ldots, y'_{t-p})$, a_t is an $\frac{N(N-1)}{2} \times 1$ vector consists of the free elements of A_{0t} stacked by rows, and W_t is an $N \times \frac{N(N-1)}{2}$ matrix,

$$W_t = \begin{pmatrix} 0 & 0 & 0 & \cdots & \cdots & \cdots & 0 \\ -y_{1,t} & 0 & 0 & \cdots & \cdots & \cdots & 0 \\ 0 & -y_{1,t} & -y_{2,t} & \cdots & \cdots & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots & \cdots & \cdots & 0 \\ 0 & \cdots & \cdots & -y_{1,t} & -y_{2,t} & \cdots & -y_{N-1,t} \end{pmatrix}.$$

The log-volatilities $h_t = (h_{1,t}, \ldots, h_{N,t})'$ and the time-varying parameters a_t are assumed to follow random walk processes:

$$h_t = h_{t-1} + \epsilon_t^h, \quad \epsilon_t^h \sim N(0, \Sigma_h), \tag{2}$$

$$a_t = a_{t-1} + \epsilon_t^a, \quad \epsilon_t^a \sim N(0, \Sigma_a), \tag{3}$$

where $\Sigma_h = \text{diag}(\sigma^2_{h,1}, \ldots, \sigma^2_{h,N})$ and $\Sigma_a = \text{diag}(\sigma^2_{a,1}, \ldots, \sigma^2_{a,\frac{N(N-1)}{2}})$.

It can be seen that the VAR-SV can have an enormous number of parameters when N is large. This has led large VAR researchers to work with restricted versions of the stochastic volatility process. An influential recent model is the common drifting volatility specification of Carriero, Clark and Marcellino (2016a) which we denote by VAR-CCM1 and use in our empirical work. This is the same as the VAR-SV except that $a_t = 0$ and $\Sigma_t = e^{h_t}\Sigma$, where the Σ is an $N \times N$ positive definite matrix and h_t is a scalar stochastic volatility process:

$$h_t = \rho h_{t-1} + \epsilon_t^h, \quad \epsilon_t^h \sim \mathcal{N}(0, \sigma_h^2).$$

This, much more parsimonious, specification has been successfully used with large VARs. But it does severely restrict the form that the time variation in the error covariance matrix can take. In our empirical work, we compare our new approach to the VAR-CCM1. We also use another specification proposed in Carriero, Clark and Marcellino (2016b) which we label VAR-CCM2. This amounts to the VAR-SV with a_t restricted to be time-invariant.

2.3 The Theory of Composite Likelihood Methods

2.3.1 Preliminaries

Assuming serially independent errors, the likelihood function for $y = (y_1', .., y_T')'$ can be written as:

$$L(y;\theta) = \prod_{t=1}^{T} L(y_t;\theta), \tag{4}$$

where $L(y_t;\theta) = p(y_t|\theta)$. The composite likelihood is defined as

$$L^C(y;\theta) = \prod_{t=1}^{T}\prod_{i=1}^{M} L^C(y_{i,t};\theta)^{w_i}, \tag{5}$$

where $L^C(y_{i,t};\theta) = p(y_{i,t}|\theta)$ is the quasi-likelihood and w_i is the weight attached to each quasi-likelihood with $\sum_{i=1}^{M} w_i = 1$. The weights will be discussed in sub-section 2.3.3.

The maximum composite likelihood estimator (MCLE) involves taking the maximum of $L^C(y;\theta)$. Bayesian estimation proceeds using a posterior based on the composite likelihood (i.e. the Bayesian composite posterior is $p^C(\theta|y) \propto L^C(y;\theta)\,p(\theta)$ where $p(\theta)$ is the prior).

In theory, the likelihood components used to build a composite likelihood can be anything. That is, $y_{i,t}$ for $i = 1, .., M$ can be any sub-sets of y_t and, indeed, $y_{i,t}$ and $y_{j,t}$ can overlap. For computational purposes, the key issue is that $y_{i,t}$ and M should be small enough to lead to fast estimation. For instance, Pakel, Shephard, Sheppard and Engle (2014), in an application involving stock returns for 129 companies, achieve these goals by considering all bivariate distributions involving each distinct pair of assets. Thus, they work with $M = \frac{N(N-1)}{2} = 8,256$ bivariate Dynamic Conditional Correlation (DCC) models which is much easier than trying to work with a 129 dimensional DCC model.

With large VARs, it is common to have a few core variables of interest either for impulse response analysis (e.g. as in the FAVAR approach of Bernanke, Boivin and Eliasz, 2005, where the interest rate is isolated in order to identify a monetary policy shock) or forecasting. In this spirit, we propose partitioning $y_t = \begin{pmatrix} y_t^* \\ z_t \end{pmatrix}$ where y_t^* is N_*-dimensional and contains the core variables of interest and z_t (with elements denoted by $z_{i,t}$) is the $N_{other} = N - N_*$ vector which contains the remaining variables. Then we can let $y_{i,t} = \begin{pmatrix} y_t^* \\ z_{i,t} \end{pmatrix}$ for

$i = 1,..,N_{other}$ and, thus, $M = N_{other}$. Our composite likelihood VAR-SV (VAR-CL-SV) application will involve quasi-likelihoods which are all $N_* + 1$ dimensional VAR-SVs.

2.3.2 Asymptotic Results

The standard frequentist way of investigating the theoretical properties of composite likelihoods is to assume that $L(y;\theta)$ is the true data generating process involving a true parameter value $\theta = \theta^0$ and derive the behavior of the MCLE. Results exist in the literature noting that the MCLE should converge asymptotically to θ^0 under certain assumptions (see, e.g., Varin, Reid and Firth, 2011 or Ribatet, Cooley and Davison, 2012). But such results are limited and model dependent. In a recent survey, Varin, Reid and Furth (2011, page 34) conclude: "Using the most general definition of composite likelihood, it may be difficult to derive very many specific properties beyond perhaps consistency of the point estimator." Ribatet, Cooley and Davison (2012, section 2.3.1) derive asymptotic Bayesian results using $p^C(\theta|y)$ and show that this posterior will also converge to θ^0 under certain assumptions. We take these results as offering general support for the idea that, in finite samples, the composite likelihood is often a reasonable approximation to $L(y_t;\theta)$.

However, it is important to dig a bit deeper into the assumptions that underlie both frequentist and Bayesian asymptotic theories discussed above. In (5), we have written the likelihood components as $L(y_{i,t};\theta)$ which all depend upon a common parameter vector θ. In the VAR-CL-SV this will not be the case. Some parameters will not appear in any of the likelihood components. For instance, consider the equations for $z_{i,t}$ and $z_{j,t}$ for $i \neq j$. A large VAR-SV will contain a time-varying error covariance between these two equations. However, this error covariance will not appear in the composite likelihood function and so it will be impossible to obtain consistent estimates of it using $L^C(y;\theta)$. In other words, our choice of quasi-likelihoods means that we can never aim for asymptotic convergence to an unrestricted large VAR-SV. However, it is interesting to investigate what our methods do converge to. In this sub-section, we prove asymptotic convergence to a particular restricted VAR-SV. We also highlight the connections between this restricted VAR-SV and the Minnesota prior, but emphasize that our approach allows for stochastic volatility while the conventional Minnesota prior does not.

Pakel, Shephard, Sheppard and Engle (2014) set up the composite likelihood function somewhat differently, involving likelihood components $L(y_{i,t};\theta,\eta_i)$ where η_i are nuisance parameters specific to sub-model i and θ are the parameters of interest which are common to all models. This set-up is more appropriate for our case since we are interested in the time-varying error covariance matrix corresponding to the upper left-hand $N_* \times N_*$ block of the error covariance matrix (which is common to all quasi-likelihoods). In our case, the time-varying error covariances of the other variables are of subsidiary interest. Pakel, Shephard, Sheppard and Engle (2014) show that, under a set of stronger assumptions,[3]

[3]Standard assumptions relating to asymptotic mixing either involve the dependence be-

θ is consistent (although they do not provide a central limit theorem). Under these assumptions they show that the incidental parameter bias present in many related approaches vanishes asymptotically. We rely on this theory to justify including a set of y_t^* variables in each quasi-likelihood and using the remaining z_t variables as only being useful insofar as they improve estimation of the error covariance matrix for the y_t^* variables.

For the choice of quasi-likelihoods made in the preceding sub-section, we have been able to prove asymptotic convergence of the composite likelihood to that of a restricted VAR-SV of the following form:

$$\begin{pmatrix} A_{y,t} & 0 & \cdots & 0 \\ -\alpha'_{z,1,t} & 1 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ -\alpha'_{z,M,t} & 0 & \cdots & 1 \end{pmatrix} \begin{pmatrix} y_t^* \\ z_{1,t} \\ \vdots \\ z_{M,t} \end{pmatrix} = \begin{pmatrix} c_y \\ c_{z,1} \\ \vdots \\ c_{z,M} \end{pmatrix} +$$

$$\sum_{j=1}^{p} \begin{pmatrix} B_{yy,j} & \frac{w_1}{g(M)}\beta_{yz,1,j} & \cdots & \frac{w_M}{g(M)}\beta_{yz,M,j} \\ \beta'_{zy,1,j} & \beta_{zz,1,j} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ \beta'_{zy,M,j} & 0 & \cdots & \beta_{zz,M,j} \end{pmatrix} \begin{pmatrix} y_{t-j}^* \\ z_{1,t-j} \\ \vdots \\ z_{M,t-j} \end{pmatrix} + \begin{pmatrix} \epsilon_{y,t} \\ \epsilon_{z,1,t} \\ \vdots \\ \epsilon_{z,M,t} \end{pmatrix},$$

with $\epsilon_{y,t} \sim N(0, \Sigma_{y,t})$, $\epsilon_{z,i,t} \overset{iid}{\sim} N(0, e^{h_{N^*+i,t} - \ln w_i})$ independent of each other and

$$A_{y,t} = \begin{pmatrix} 1 & 0 & \cdots & 0 \\ -\alpha_{21,t} & 1 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ -\alpha_{N_*1,t} & -\alpha_{N_*2,t} & \cdots & 1 \end{pmatrix}, \Sigma_{y,t} = \begin{pmatrix} e^{h_{1,t}} & & & \\ & e^{h_{2,t}} & & \\ & & \ddots & \\ & & & e^{h_{N_*,t}} \end{pmatrix}.$$

Observe that this is a VAR-SV of the form

$$\tilde{A}_t y_t = c + \sum_{j=1}^{p} \tilde{B}_j y_{t-j} + \epsilon_t, \tag{6}$$

with some elements of $\tilde{A}_t$ and $\tilde{B}_j$ restricted to zero and some elements of $\tilde{B}_j$ shrunk towards zero by factors $\frac{w_1}{g(M)}, \ldots, \frac{w_M}{g(M)}$ where $g(M)$ is a function of M. We stress that the target model is a restricted VAR-SV with random walk laws of motion as in equations (2) and (3).

A word of explanation is in order about $g(M)$. A sufficient condition for the proof of the following proposition requires $\frac{\sqrt{M}}{g(M)}$ to be bounded for all M (e.g. if $g(M) = \sqrt{M}$ our proof follows standard law of large numbers results). But this condition is exactly what prior shrinkage in VARs usually does. That

tween the same variable at different points in time or different variables at the same point in time. They add to these standard assumptions, additional mixing assumptions relating to the dependence between different variables at different points in time.

is, in our approach as M increases N also increases and the VAR dimension increases. It is standard for Bayesians working with large VARs to increase prior shrinkage (e.g. using the Minnesota prior) when VAR dimension increases (see, e.g., Table 1 of Banbura Giannone and Reichlin, 2010). Hence, the presence (and interpretation) of $g(M)$ is justified as being comparable to the types of prior shrinkage commonly used in large Bayesian VARs. Note too that $g(M)$ only applies to other lags in the equations for the core variables, so the convergence of the composite likelihood to a restricted VAR-SV only depends on the presence of shrinkage on these coefficients.

It is important to emphasize that $L^C(y;\theta)$ is not a true likelihood in the sense that it is not a density in the data (conditional on parameters) that integrates to one. To compare it to a conventional likelihood for the restricted VAR-SV given in (6), $L(y;\theta)$, we consider the normalized composite likelihood

$$\tilde{L}^C(y;\theta) = \frac{L^C(y;\theta)}{\int_y L^C(y;\theta)\,dy}.$$

A useful measure of the approximation error associated with using $L^C(y;\theta)$ instead of $L(y;\theta)$ is the Kulback-Liebler divergence of $L(y;\theta)$ from $\tilde{L}^C(y;\theta)$, denoted $D_{\mathrm{KL}}(L\|\tilde{L}^C)$, which is summarized in the following proposition.

Proposition 1 *Assume* $\max\{w_i\}$ *is decreasing in* M *and* $\frac{\sqrt{M}}{g(M)} < \infty$ *for all* $M \geq 1$. *Then*

$$\lim_{M\to\infty} D_{KL}(L\|\tilde{L}^C) = 0.$$

The proof of this proposition is in the Technical Appendix. The assumption that $\max\{w_i\}$ is decreasing in M is innocuous as it implies only that when we add a new sub-model it has non-zero weight which will leave less weight for the other models, including the model with maximum weight. Thus, our composite likelihood using small VAR-SVs as quasi-likelihoods asymptotically converges to the likelihood of particular large VAR-SV under sensible assumptions.

Of course, given the way we have defined our quasi-likelihoods, it is not possible to asymptotically converge to an unrestricted large VAR-SV since (as noted previously) some of the unrestricted model's parameters appear in none of the quasi-likelihoods. If interest lies in using composite likelihood methods to provide estimates of all the parameters in a large VAR-SV, then other quasi-likelihoods should be chosen to build a composite likelihood function (e.g. building a set of quasi-likelihoods involving all possible bivariate or tri-variate combinations of the variables). Our choice of quasi-likelihoods is based on our choice of empirical problem. We are interested in forecasting a small number of variables, using the other variables only to improve these forecasts. For this, our choice of quasi-likelihoods is a sensible one.

2.3.3 Composite Likelihoods as Opinion Pools

An alternative way of theorizing about composite likelihoods, popular among Bayesians (see, e.g., Roche, 2016) is to begin by assuming there is some feature of interest, θ (in our case, the error covariance matrix relating to the core variables). There are many "agents" each of which uses a (possibly agent-specific) information set to produce an "opinion" (i.e. a posterior) about θ. The opinions going into the pool can be obtained from any source. The question arises as to how to pool these opinions? There is a literature on such opinion or prediction pools. Hall and Mitchell (2007) and Geweke and Amisano (2011) are influential approaches in econometrics. Genest, Weerahandi, Zidek (1984) and Genest, McConway and Schervish (1986) are influential early references which establish or review many theoretical properties of opinion pools.

If, in our case, we interpret each quasi-likelihood, $L^C(y_{i,t};\theta)$, as arising from an agent, we can draw on this literature to obtain a theoretical justification for our approach. In sub-section 2.2.1, we defined the Bayesian composite posterior $p^C(\theta|y)$ based on the composite likelihood (5). Papers such Roche (2016) show that Bayesian inference using the composite likelihood can be interpreted as arising from a generalized logarithmic opinion pool. This offers strong theoretical justification for our approach. Genest et al (1984) show that such opinion pools have attractive properties including external Bayesianity. External Bayesianity implies that, if all agents agree on the same prior, then it does not matter whether the prior is added before or after the opinions are pooled. Generalized logarithmic opinion pools are the only class of opinion pools that have this property.

An alternative approach is to use linear opinion pools (e.g. Hall and Mitchell, 2007, and Geweke and Amisano, 2011). The use of linear opinion pools means this approach does not satisfy external Bayesianity nor lead to Bayesian inference based on $p^C(\theta|y)$. However, as discussed in Geweke and Amisano (2011), linear pools sometimes give results that are different from logarithmic opinion pools. Hence, even though they are not a composite likelihood approach, they are closely related and we include them in our set of empirical results.

The advantage of drawing on the opinion pool literature is that it offers insights into how the weights, w_i for $i = 1,..,M$, can be chosen. In our empirical work, we consider a range of approaches. Setting the weights to be equal ($w_i = \frac{1}{M}$) is simple and commonly done. However, this often leads to a problem known as "information overload". Adding more and more agents can lead to less precise inference as the agents with good opinions will find their signal swamped. In the linear opinion pool formulation, Geweke and Amisano (2011) derive a set of weights which are optimal for the linear pool and provide a method for calculating them.

In the logarithmic opinion pool formulation, a logical thing to do (see Canova and Matthes, 2017) is to base the weights based on some measure of the fit of each quasi-likelihood. In our application, where each quasi-likelihood is a VAR-SV involving a set of core variables (y_t^*) and one other variable, it makes sense to use the marginal likelihood or an approximation to it to calculate the weights. Hence, we consider weighting schemes based on the Bayesian information criterion (BIC), the Deviance Information criterion (DIC) and the marginal likelihood. Letting BIC_i be the BIC for sub-model i, we have

$$\text{BIC}_i = -2\log L\left(y^*; \widehat{\theta}_i\right) + d\log(T),$$

where $\widehat{\theta}$ is the maximum likelihood estimate using sub-model i, $y^* = (y_1^{*\prime}, .., y_T^{*\prime})'$ and d is the number of free parameters. We stress that, in each quasi-likelihood, we are only using the core variables (which are common to all quasi-likelihoods) to define the BIC. The maximum likelihood estimate is computed using the integrated likelihood as in Chan and Eisenstat (2018). The weight for each sub-model is computed as

$$w_i^{BIC} = \frac{e^{-\frac{1}{2}\text{BIC}_i}}{\sum_{j=1}^M e^{-\frac{1}{2}\text{BIC}_j}}, \quad \text{for } i = 1, .., M.$$

Our second set of weights follows the same strategy, but using DIC instead of BIC. DIC is calculated based on the integrated likelihood for the core variables of interest (see Chan and Grant, 2016, for details).

The third weighting scheme is based on the marginal likelihood. We use the following marginal likelihood for sub-model i:

$$\text{ML}_i = \int p_i(y^*|\theta)p(\theta)d\theta,$$

where $p_i(y|\theta) = \prod_{t=1}^{T} L^C\left(y_{i,t};\theta\right)$ and $p_i(y^*|\theta)$ implies evaluating the marginal likelihood only using the core variables. The weight for each sub-model is computed as

$$w_i^{ML} = \frac{\text{ML}_i}{\sum_{j=1}^M \text{ML}_j}.$$

We calculate the marginal likelihood using the methods of Chan and Eisenstat (2018). We use the abbreviations, VAR-CL-BIC, VAR-CL-DIC and VAR-CL-ML for composite likelihood methods involving these three different weights.

In the linear opinion pool approach calculating the optimal weights involves the following steps. Let $p_i(y_t^*|y_{1:t-1})$ be the one-step-ahead predictive density for the core variables for the i^{th} sub-model and $w = (w_1, w_2, .., w_M)'$. The predictive log score function is given by

$$f(w) = \sum_{t=1}^{T}\log\left(\sum_{i=1}^{M} w_i p_i(y_t^*|y_{1:t-1}))\right).$$

The optimal weight is obtained by solving the optimization problem $\widehat{w} = \text{argmax}_w f(w)$. We use VAR-LIN as the abbreviation for this approach. Even though these weights are calculated to be optimal in the linear opinion pool case, we can use them as weights in the composite likelihood. We refer to such an approach as VAR-CL-LIN.

The main research question of interest in this paper is whether composite likelihood methods involving many small models can forecast well in the presence of large data sets. A subsidiary question though, is whether the general idea of combining many small models for forecasting is a good one. By including linear opinion pooling methods we can address the second question. To preview our empirical findings, we find that all approaches which combine many small models forecast well. That is, it seems that the empirical success of our approach is largely due to the choice of quasi-likelihoods as opposed to the way they are combined. However, it is worth noting that (as we shall see in Section 6) the linear opinion pool has significant computational drawbacks relative to the composite likelihood approach.

3 Bayesian Analysis Using the Composite Posterior

Our goal is to carry out Bayesian analysis on the composite posterior, $p^C(\theta|y)$, using MCMC draws from each of the quasi-posterior distributions. This section develops an algorithm for doing so.

3.1 Quasi-Posterior Distributions

We first extend our earlier notation to define the quasi-likelihoods. Remember that each of these is a VAR-SV that combines core variables of interest, y_t^*, with an additional variable, $z_{i,t}$. Thus, quasi-likelihood i (for $i = 1, \dots, M$) can be expressed in the form:

$$A_{y,t} y_t^* = X_{y,t}\beta_y + X_{z_i,t}\beta_{yz_i} + \epsilon_{y,t}, \qquad \epsilon_{y,t} \sim N(0, \Sigma_{y,t}), \tag{7}$$

$$z_{i,t} = y_t^* \alpha_{z_i,t} + X_t \beta_{z_i} + \epsilon_{z_i,t}, \qquad \epsilon_{z_i,t} \sim N\left(0, e^{h_{N_*+i,t}}\right), \tag{8}$$

$$A_{y,t} = \begin{pmatrix} 1 & 0 & \cdots & 0 \\ \alpha_{21,t} & 1 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ \alpha_{N_*1,t} & \alpha_{N_*2,t} & \cdots & 1 \end{pmatrix}, \quad \Sigma_{y,t} = \begin{pmatrix} e^{h_{1,t}} & & & \\ & e^{h_{2,t}} & & \\ & & \ddots & \\ & & & e^{h_{N_*,t}} \end{pmatrix}.$$

In (7), the matrix $X_{y,t}$ contains lags of y_t^*, and the matrix $X_{z_i,t}$ contains lags of $z_{i,t}$. The log-volatilities $h_{i,t}$ and the time-varying parameters $\alpha_{z_i,t}$ and $\alpha_{jk,t}$, $i = 1, \dots, M, j = 2, \dots, N_*, k = 1, \dots, j-1$ are assumed to follow random walk processes:

$$h_{i,t} = h_{i,t-1} + \epsilon_{i,t}^h, \quad \epsilon_{i,t}^h \sim N(0, \sigma_{h,i}^2), \tag{9}$$

$$\alpha_{jk,t} = \alpha_{jk,t-1} + \epsilon_{jk,t}^\alpha, \quad \epsilon_{jk,t}^\alpha \sim N(0, \sigma_{\alpha,jk}^2), \tag{10}$$

$$\alpha_{z_i,t} = \alpha_{z_i,t-1} + \epsilon_{i,t}^\alpha, \quad \epsilon_{i,t}^\alpha \sim N(0, \Sigma_{\alpha,i}), \tag{11}$$

where $\Sigma_{\alpha,i}$ is a diagonal matrix.

Let $\theta = \{\beta_y, A_{y,1}, \ldots, A_{y,T}, \Sigma_{y,1}, \ldots, \Sigma_{y,T}\}$ be the set of parameters that are common in all quasi-likelihoods, and denote by

$$\eta_i = \{\beta_{yz_i}, \beta_{z_i}, \alpha_{z_i,0}, \ldots, \alpha_{z_i,T}, h_{N_*+i,0}, \ldots, h_{N_*+i,T}, \Sigma_{\alpha,i}\}$$

the parameters that appear only in quasi-likelihood i. Each quasi-posterior i is given by

$$p_i(\theta, \eta_i \,|\, y^*, z_i) = p(\theta, \eta_i)p(y^*, z_i \,|\, \theta, \eta_i)/p(y^*, z_i),$$

where $z_i = (z_{i,1}, .., z_{i,T})'$.

A key feature of our set-up is that the density that defines each quasi-likelihood can be conveniently decomposed as:

$$\begin{aligned}
p(y^*, z_i \,|\, \theta, \eta_i) &= \prod_{t=1}^{T} p(y_t^* \,|\, y_{t-1}^*, \ldots, y_{t-p}^*, z_{i,t-1}, \ldots, z_{i,t-p}, \beta_y, \beta_{yz_i}, A_{y,t}, \Sigma_{y,t}) \\
&\quad \times p(z_{i,t} \,|\, y_t^*, y_{t-1}^*, \ldots, y_{t-p}^*, z_{i,t-1}, \ldots, z_{i,t-p}, \beta_{z_i}, \alpha_{z_i,t}, h_{N_*+i,t}), \\
&= \left(\prod_{t=1}^{T} p(y_t^* \,|\, \cdot\,)\right)\left(\prod_{t=1}^{T} p(z_{i,t} \,|\, y_t^*, \,\cdot\,)\right), \\
&= p(y^* \,|\, \tilde{z}_i, \theta, \beta_{yz_i})p(z_i \,|\, y^*, \tilde{\eta}_i),
\end{aligned}$$

where $\tilde{z}_i = \{z_{i,1}, \ldots, z_{i,T-1}\}$ and $\tilde{\eta}_i = \{\beta_{z_i}, \alpha_{z_i,0}, \ldots, \alpha_{z_i,T}, h_{N_*+i,0}, \ldots, h_{N_*+i,T}, \Sigma_{\alpha,i}\}$.

In this decomposition, $p(y^* \,|\, \tilde{z}_i, \theta, \beta_{yz_i})$ is the density of a multivariate normal distribution that can be regarded as the likelihood for the model in (7), with $z_{i,1}, \ldots, z_{i,T-1}$ treated as exogenous regressors. Moreover, this density can be integrated analytically with respect to a prior on β_{yz_i} to obtain a density that only contains common parameters θ, i.e., $p(y^* \,|\, \tilde{z}_i, \theta) = \int_{\beta_{yz_i}} p(\beta_{yz_i})p(y^* \,|\, \tilde{z}_i, \theta, \beta_{yz_i})d\beta_{yz_i}$. Similarly, $p(z_i \,|\, y^*, \tilde{\eta}_i)$ can be viewed as the multivariate normal likelihood for a time-varying parameter autoregressive distributed lag model (TVP-ARDL) with exogenous y_t^* defined by (8), with the important feature that it contains only nuisance parameters.

Consequently, if θ and $\tilde{\eta}_i$ are independent in the prior (as we assume in this paper), then they are also independent in the i-th quasi-posterior. Moreover, this independence carries over to the composite posterior defined as

$$\begin{aligned}
p^C(\theta, \tilde{\eta}_1, \ldots, \tilde{\eta}_M \,|\, y^*, z_1, \ldots, z_M) &\propto p(\theta)\prod_{i=1}^{M} p(\tilde{\eta}_i)p(y^*, z_i \,|\, \theta, \tilde{\eta}_i)^{w_i}, \\
&= p^C(\theta \,|\, y^*, \tilde{z}_1, \ldots, \tilde{z}_M)\prod_{i=1}^{M} p^C(\tilde{\eta}_i \,|\, y^*, z_i),
\end{aligned} \tag{12}$$

where

$$p^C(\theta \,|\, y^*, \tilde{z}_1, \ldots, \tilde{z}_M) \propto p(\theta)\prod_{i=1}^{M} p(y^* \,|\, \tilde{z}_i, \theta)^{w_i},$$
$$p^C(\tilde{\eta}_i \,|\, y^*, z_i) \propto p(\tilde{\eta}_i)p(z_i \,|\, y^*, \tilde{\eta}_i)^{w_i}.$$

Note that we have used the following identity to construct the composite likelihood in the definition of the composite posterior:

$$\begin{aligned} p(y^*, z_i \mid \theta, \tilde{\eta}_i) &= \int_{\beta_{yz_i}} p(\beta_{yz_i}) p(y^*, z_i \mid \theta, \eta_i) d\beta_{yz_i} \\ &= \int_{\beta_{yz_i}} p(\beta_{yz_i}) p(y^* | \tilde{z}_i, \theta, \beta_{yz_i}) p(z_i \mid y^*, \tilde{\eta}_i) d\beta_{yz_i} \\ &= p(y^* | \tilde{z}_i, \theta) p(z_i \mid y^*, \tilde{\eta}_i). \end{aligned}$$

The decomposition in (12) is crucial as it allows us to sample the common parameters θ and each $\tilde{\eta}_i, i = 1, \ldots, M$ independently. Consequently, we can parallelize the computations and vastly reduce the computational time. This is taken up in the next sub-section.

3.2 Simulation from the Composite Posterior

This section describes our computational algorithm to simulate from the composite posterior $p^C(\theta, \tilde{\eta}_1, \ldots, \tilde{\eta}_M \mid y^*, z_1, \ldots, z_M)$. Instead of designing an MCMC algorithm to directly sample from this composite posterior, we develop an accept-reject algorithm using MCMC draws from the individual quasi-posterior distributions as proposals. The key advantage of this approach is that sampling from each of the quasi-posterior distributions can be done in parallel and using standard MCMC methods for small VAR-SV models. This allows us to work with hundreds of variables where other approaches which involve use of MCMC methods with large VAR-SV are not feasible.

Using the decomposition of the composite posterior in (12), we can generate samples from $p^C(\theta \mid y^*, \tilde{z}_1, \ldots, \tilde{z}_M)$, $p^C(\tilde{\eta}_1 \mid y^*, z_1), \ldots, p^C(\tilde{\eta}_M \mid y^*, z_M)$ independently. We start with simulating the common parameters θ from $p^C(\theta \mid y^*, \tilde{z}_1, \ldots, \tilde{z}_M)$ by appropriately pooling draws of θ from the quasi-posteriors. We develop an accept-reject algorithm for this purpose.

Consider the proposal density $q(\theta)$ that is a mixture of the M quasi-posteriors, i.e.,

$$q(\theta) = \sum_{i=1}^{M} w_i p_i(\theta \mid y^*, z_i) = p(\theta) \sum_{i=1}^{M} \frac{w_i p(y^* \mid \tilde{z}_i, \theta)}{p(y^* \mid \tilde{z}_i)},$$

where $p(y^* | \tilde{z}_i) = \int_\theta p(\theta) p(y^* \mid \tilde{z}_i, \theta) d\theta$ can be regarded as the marginal likelihood of the VAR-SV with exogenous variables defined in (7).

Given draws from the M quasi-posteriors $p_i(\theta \mid y^*, z_i)$ for $i = 1, .., M$ and a set of weights w_i for $i = 1, .., M$—which can be any of those described in section 2.3.3—it is easy to obtain a set of draws from $q(\theta)$. Moreover, $q(\theta)$ can be readily evaluated: $p(y^* \mid \tilde{z}_i)$ can be computed using the algorithm of Chan and Eisenstat (2018) that we use to obtain the marginal likelihood in a VAR-SV (see Section 2.3.3) and $p(y^* \mid \tilde{z}_i, \theta)$ is a multivariate normal density.

To show the latter claim, let $\beta_{yz_i} \sim N(\underline{\beta}_{yz}, \underline{V}_{\beta,z})$ denote the prior for β_{yz_i}. Let $\alpha_{y,t}$ represent the free elements in $A_{y,t}$ stacked by row, and let $W_{y,t}$ denote the associated covariate matrix (see the discussion in Section 2.2). Then,

$p(y^* \mid \tilde{z}_i, \theta)$ has the following multivariate normal form:

$$(y^* \mid z_i, \theta) \sim N\left(W_y\alpha_y + X_y\beta_y + X_{z_i}\underline{\beta}_{yz}, X_{z_i}\underline{V}_{\beta,z}X_{z_i}' + \Sigma_y\right),$$

where Σ_y is a block diagonal matrix with diagonal blocks $\Sigma_{y,t}, t = 1, \ldots, T$. Finally, X_y, X_{z_i}, W_y and α_y respectively stack $X_{y,t}$, $X_{z_i,t}$ $W_{y,t}$ and $\alpha_{y,t}$ over $t = 1, \ldots, T$.

To be a valid accept-reject algorithm with proposal density $q(\theta)$, we need to show that the ratio $p^C(\theta \mid y^*, \tilde{z}_1, \ldots, \tilde{z}_M)/q(\theta)$ is bounded for all θ in its support. To that end, observe that

$$r(\theta) = \frac{\prod_{i=1}^M [p(y^* \mid \tilde{z}_i, \theta)/p(y^* \mid \tilde{z}_i)]^{w_i}}{\sum_{i=1}^M w_i p(y^* \mid \tilde{z}_i, \theta)/p(y^* \mid \tilde{z}_i)} \leq 1.$$

This inequality follows from the fact that a geometric average is always less than or equal to the corresponding arithmetic average. Now, write the target density as

$$p^C(\theta \mid y^*, \tilde{z}_1, \ldots, \tilde{z}_M) = \frac{p(\theta)\prod_{i=1}^M p(y^* \mid \tilde{z}_i, \theta)^{w_i}}{K_i},$$

where $K_i = \int p(\theta)\prod_{i=1}^M p(y^* \mid \tilde{z}_i, \theta)^{w_i} d\theta$ is the normalizing constant. If we let $K = \prod_{i=1}^M p(y^* \mid \tilde{z}_i)^{w_i}/K_i$, then we can show that $p^C(\theta \mid y^*, \tilde{z}_1, \ldots, \tilde{z}_M) \leq Kq(\theta)$ for all θ:

$$\frac{p^C(\theta \mid y^*, \tilde{z}_1, \ldots, \tilde{z}_M)}{Kq(\theta)} = \frac{p(\theta)\prod_{i=1}^M [p(y^* \mid \tilde{z}_i, \theta)p(y^* \mid \tilde{z}_i)]^{w_i}}{p(\theta)\sum_{i=1}^M w_i p(y^* \mid \tilde{z}_i, \theta)/p(y^* \mid \tilde{z}_i)} = r(\theta) \leq 1.$$

This suggests an accept-reject sampling approach to pool draws of common parameters obtained from individual quasi-posteriors.[4] We summarize the algorithm as follows:

1. obtain a proposal draw $\theta^* \sim q(\theta)$;
2. accept θ^* with probability $r(\theta^*)$.

Next, we consider obtaining draws from $p^C(\tilde{\eta}_i \mid y^*, z_i)$ for each $i = 1, \ldots, M$.[5] To that end we focus on the TVP-ARDL model defined by (8). Note that each $\tilde{\eta}_i$ is relatively low-dimensional and is independent of θ. Consequently, sampling $\tilde{\eta}_1, \ldots, \tilde{\eta}_M$ can be done in parallel and is fast in practice. More specifically, when $w_i = 1$, sampling from $p^C(\tilde{\eta}_i \mid y^*, z_i)$ is equivalent to sampling from the TVP-ARDL posterior, which is standard. For the more general case with $w_i < 1$, the standard MCMC method needs only minor modifications.

[4] For a general discussion of the accept-reject method, see, e.g., Section 3.1.5 in Kroese, Taimre and Botev (2011). Since the proposal draws are obtained from the quasi-posteriors using MCMC, they are correlated by construction. Consequently, the sample obtained from this accept-reject algorithm would also be correlated.

[5] Note that the draws of η_i are only needed to compute MLs, DICs, and BICs, which are only used to compute the weights $w_1, \ldots, w_M$. If the weights are known (e.g. as in the equal weights case), then there is no need to obtain η_i.

In particular, the Gibbs steps for $\sigma^2_{h,i}$, $\Sigma_{\alpha,i}$, $\alpha_{z_i,0}$ and $h_{N_*+i,0}$, conditional on draws of $\alpha_{z_i,1},\ldots,\alpha_{z_i,T}$ and $h_{N_*+i,1},\ldots,h_{N_*+i,T}$ are identical to the standard case. The Gibbs steps to sample $\alpha_{z_i,1},\ldots,\alpha_{z_i,T}$ and β_{z_i} are also very similar—the only modification is to replace $h_{N_*+i,t}$ by $\bar{h}_{N_*+i,t} = h_{N_*+i,t} - \ln w_i$ for all $t = 1,\ldots,T$ and $i = 1,\ldots,M$ in the conditional distributions. Finally, we sample $h_{N_*+i,1},\ldots,h_{N_*+i,T}$ from its conditional distribution

$$\begin{aligned}
p(h_{N_*+i,1},\ldots,h_{N_*+i,T} \,|\, h_{N_*+i,0}, \sigma^2_{h,N_*+1}, \alpha_{z_i,0}, \alpha_{z_i,1},\ldots,\alpha_{z_i,T}, \beta_{z_i}, z_i, y^*) \\
\propto p(h_{N_*+i,1},\ldots,h_{N_*+i,T} \,|\, h_{N_*+i,0}, \sigma^2_{h,N_*+1}) \\
p(z_i \,|\, y^*, h_{N_*+i,1},\ldots,h_{N_*+i,T}, \alpha_{z_i,1},\ldots,\alpha_{z_i,T}, \beta_{z_i})^{w_i}. \quad (13)
\end{aligned}$$

This is done by considering the following auxiliary state space model

$$\begin{aligned}
z_{i,t} &= y_t^{*\prime}\alpha_{z_i,t} + x_t'\beta_{z_i} + \epsilon_{z_i,t}, & \epsilon_{z_i,t} &\sim \mathcal{N}\left(0, e^{\tilde{h}_{N_*+i,t}}\right), \\
\tilde{h}_{N_*+i,t} &= \frac{T-t+1}{2}(1-w_i)\sigma^2_{h,N_*+i} + \tilde{h}_{N_*+i,t-1} + \epsilon^h_{N_*+i,t}, & \epsilon^h_{N_*+i,t} &\sim \mathcal{N}\left(0, \sigma^2_{h,N_*+i}\right).
\end{aligned}$$

Clearly, we can sample $\tilde{h}_{N_*+i,1},\ldots,\tilde{h}_{N_*+i,T}$ using standard methods for stochastic volatility models. Given these draws, we set $h_{N_*+i,t} = \tilde{h}_{N_*+i,t} + \ln w_i$. It can be shown that the draws thus obtained follow the same conditional distribution given in (13).

4 Forecasting

In this section we describe how one can compute the joint predictive density of the core variables using simulation. To start we first introduce some notation. For a time series $x_1,\ldots,x_T$, we use $x_{s:t}$ to denote the observations from time s to time t, i.e., $x_{s:t} = \{x_s, x_{s+1},\ldots,x_{t-1},x_t\}$. For example, $\theta_{1:t}$ represents the set of common parameters from time 1 to time t, i.e., $\theta_{1:t} = \{\beta_y, A_{y,1},\ldots,A_{y,t},\Sigma_{y,1},\ldots,\Sigma_{y,t}\}$. Furthermore, let $z_{t-p:t-1}$ denote the set of non-core variables: $\{z_{1,t-p:t-1},\ldots,z_{M,t-p:t-1}\}$.

The one-step-ahead composite predictive density, conditional on the parameters up to time t, is given by:

$$\begin{aligned}
p^C(y_t^*, z_{1,t},\ldots,z_{M,t} \,|\, y_{t-p:t-1}, z_{t-p:t-1}, \theta_{1:t}, \tilde{\eta}_{1,1:t},\ldots,\tilde{\eta}_{M,1:t}) = \\
p^C(y_t^* \,|\, y^*_{t-p:t-1}, z_{t-p:t-1}, \theta_{1:t}) \prod_{i=1}^M p^C(z_{i,t} \,|\, y^*_{t-p:t}, z_{i,t-p:t-1}, \tilde{\eta}_{i,1:t}),
\end{aligned}$$

where

$$\begin{aligned}
p^C(y_t^* \,|\, y^*_{t-p:t-1}, z_{t-p:t-1}, \theta_{1:t}) &\propto \prod_{i=1}^M p(y_t^* \,|\, y^*_{t-p:t-1}, z_{i,t-p:t-1}, \theta_{1:t})^{w_i}, \\
p^C(z_{i,t} \,|\, y^*_{t-p:t}, z_{i,t-p:t-1}, \tilde{\eta}_{i,1:t}) &\propto p(z_{i,t} \,|\, y^*_{t-p:t}, z_{i,t-p:t-1}, \tilde{\eta}_{i,1:t})^{w_i}.
\end{aligned}$$

The density $p^C(y_t^* \,|\, y_{t-p:t-1}^*, z_{t-p:t-1}, \theta_{1:t})$ is multivariate normal and has the form

$$
\begin{aligned}
(y_t^* \,|\, y_{t-p:t-1}^*, z_{t-p:t-1}, \theta_{1:t}) &\sim N(\hat{y}_t, V_{y,t}), \\
\hat{y}_t &= W_{y,t}\alpha_{y,t} + X_{y,t}\beta_y + V_{y,t}\left(\sum_{i=1}^{M} w_i V_{y,i,t}^{-1} X_{z_i} \underline{\beta}_{yz}\right), \\
V_{y,t} &= \left(\sum_{i=1}^{M} w_i V_{y,i,t}^{-1}\right)^{-1}, \\
V_{y,i,t} &= X_{z_i,t} \underline{V}_{\beta,z} X_{z_i,t}' + \Sigma_{y,t}.
\end{aligned}
$$

The density $p^C(z_{i,t} \,|\, y_{t-p:t}^*, z_{i,t-p:t-1}, \tilde{\eta}_{i,1:t})$ is also normal and has the form

$$
(z_{i,t} \,|\, y_{t-p:t}^*, z_{i,t-p:t-1}, \tilde{\eta}_{i,1:t}) \sim N\left(y_t^{*\prime}\alpha_{z_i,t} + X_t\beta_{z_i}, e^{h_{N_*+i,t} - \ln w_i}\right).
$$

Accordingly, the one-step ahead predictive density is given by

$$
\begin{aligned}
p^C(y_{t+1}^* \,|\, y_{1:t}^*, z_{1,1:t}, \ldots, z_{M,1:t}) = \int_{\theta_{1:t+1}} & p^C(y_{t+1}^* \,|\, y_{t-p+1:t}^*, z_{t-p+1:t}, \theta_{1:t+1}) \\
& p^C(\theta_{1:t} \,|\, y_{1:t}^*, \tilde{z}_{1,1:t}, \ldots, \tilde{z}_{M,1:t}) p(\theta_{t+1}|\theta_t) d\theta_{1:t+1},
\end{aligned}
$$

where $p(\theta_{t+1}|\theta_t)$ is a product of normal densities implied by the state equations (9)–(10). Hence, we can obtain the one-step ahead predictive density as follows: given a posterior draw of $\theta_{1:t}$, we use the state equations (9)–(10) to obtain θ_{t+1}. Conditional on these draws, $p^C(y_{t+1}^* \,|\, y_{t-p+1:t}^*, z_{t-p+1:t}, \theta_{1:t+1})$ is a normal density given above. Finally, we average these densities over the posterior simulator output.

This predictive simulation method can be applied to generate forecasts for longer horizons. Specifically, the same procedure can be applied, once we generate future core and auxiliary variables using the model. Furthermore, observe that sampling from the one-step-ahead predictive density $p^C(y_{t+1}^* \,|\, y_{1:t}^*, z_{1,1:t}, \ldots, z_{M,1:t})$ does not require draws of $\tilde{\eta}_{i,1:t}$, and therefore, the extra steps involved in sampling $\tilde{\eta}_{i,t}$ can be omitted if the researcher is interested only in one-step ahead forecasting or uses the direct method of forecasting. The empirical results in the following section use the direct method of forecasting.

5 Empirical Results

5.1 Overview

We carry out an empirical investigation of our composite likelihood methods using a small data set of quarterly US data for 7 variables and a large quarterly data set involving 196 variables. The data is taken from the Federal Reserve

Bank of St. Louis' FRED-QD data set and runs from 1959Q1- 2015Q3.[6] All data are transformed to stationarity following the recommendations in the FRED-QD data base and then standardized to have mean zero and standard deviation one. We focus on empirical results relating to three core variables: CPI inflation, GDP growth and the Federal Funds rate. The 4 other variables in the small data set are the Civilian Unemployment Rate, Industrial Production Index, Real M2 Money Stock and S&P's Common Stock Price Index. A lag length of four is used for all models.

In our empirical application involving a high-dimensional data set, there are two issues that need to be kept conceptually clear. The first issue is whether composite likelihood methods forecast well when working with such high dimensional data as compared to other high dimensional approaches. The second is a general issue which is related to any modelling approach. This is whether working with large data sets improves forecasts or whether working with smaller data sets is adequate. To address the first issue the key comparison is with the homoskedastic large VAR with natural conjugate prior as this is the only approach that is computationally feasible using the large data set. Accordingly, this is the main model we use in our forecast comparison. It is labelled Large VAR in the tables.

With regards to the second issue, other papers working with similar US quarterly data sets and alternative modelling approaches have tended to find that working with large VARs does improve forecast performance relative to small VARs. However, the evidence is often not that strong. For instance, Koop (2013) finds that, compared to small VARs, moving towards larger VARs does improve forecast performance, but there comes a point where adding extra variables into the VAR offers only modest improvements in forecast performance. To shed light on this issue, we include the small data set. With this data set it is computationally feasible to estimate a wide range of VARs with stochastic volatility.

Our composite likelihood approach differs from some or all of the competitor models in three aspects: i) it uses all the variables in the large data set (whereas some of the other approaches do not), ii) it has stochastic volatility (whereas some of the other approaches do not) and iii) it involves combining results from many models (whereas other approaches work with single models). It is only through an extensive comparison involving both small and large data sets can we examine the separate roles of these three aspects.

The VAR-SV is the most flexible model we consider and, with 7 variables should not be over-parameterized. Thus, it should provide us reasonable benchmark estimates to compare the alternative approaches to. We begin by presenting a small Monte Carlo study where we use parameter estimates from the VAR-SV using the small data set to construct a data generating process (DGP). We artificially generate 100 artificial data sets from this DGP and investigate the performance of our composite likelihood methods using them. Subsequently,

[6]The data is available through https://research.stlouisfed.org/econ/mccracken/fred-databases/. See also McCracken and Ng (2015). Complete details of all the variables in the data set are provided there.

we present estimates using the small data set for our composite likelihood approaches, the VAR-CCM1 and VAR-CCM2 (see sub-section 2.2 for a definition of these models). We also present results from a homoskedastic VAR using the small data set (labelled VAR-HM in the tables).

Further details about the specification of all models, including prior choice, are given in the Technical Appendix. For the VAR coefficients in all models we make standard Minnesota prior choices. Where possible, we make identical specification and prior hyperparameter choices across models. It is worth stressing that, in conventional large VAR approaches where the number of parameters being estimated exceeds the number of observations, prior elicitation is crucial. Priors must be very informative and results can be sensitive to prior choice. An advantage of composite likelihood approaches is that, since all sub-models used are small, prior elicitation is a less important issue. It is possible to use less informative priors and prior sensitivity concerns are mitigated.

A summary of all the models used in the paper, including their acronyms, is given in Table 1.

Table 1: Models used in Forecasting Exercise

VAR-HM	7-variable Homoskedastic VAR
VAR-SV	7-variable VAR with stochastic volatility
VAR-CCM1	7-variable model of CCM (2016a)
VAR-CCM2	7-variable model of CCM (2016b)
Large VAR	Large Homoskedastic VAR
VAR-CL-BIC	VAR-CL-SV with BIC based weights
VAR-CL-DIC	VAR-CL-SV with DIC based weights
VAR-CL-EQ	VAR-CL-SV with equal weights
VAR-CL-ML	VAR–SV with ML weights
VAR-CL-LIN	VAR–CL-SV with linear pool weights
VAR-LIN	VAR-SV with linear pool weights
small VAR-HM	3-variable Homoskedastic VAR
small VAR-SV	3-variable VAR with stochastic volatility
VAR-SV-R	3-variable VAR-SV with the other 4 variables included on the RHS
VAR-SV-R1	3-variable VAR-SV with one lag of all variables included on the RHS

We discuss the empirical performance of each model in terms of their forecasting performance and the reasonableness of the estimates of features of interest they produce. Our features of interest focus on the error variances and covariances involving the three core variables.

To evaluate forecast performance, we use two point forecast metrics and two density forecast metrics for the core variables. Let $y_t^* = \left(y_{t,1}^*, y_{t,2}^*, y_{t,3}^*\right)'$ denote the random variables being forecast and $y_t^R = \left(y_{t,1}^R, y_{t,2}^R, y_{t,3}^R\right)'$ be their realizations. For the point forecast, we report the root mean squared forecast error (RMSFE) and the mean absolute forecast error (MAFE),

$$\text{RMSFE}_i = \sqrt{\frac{\sum_{t=t_0}^{T-h}\left(y^R_{t+h,i} - E(y^*_{t+h,i}|y^R_{1:t})\right)^2}{T-h-t_0+1}}.$$

$$\text{MAFE}_i = \frac{\sum_{t=t_0}^{T-h}\left|y^R_{t+h,i} - \hat{y}_{t+h,i}\right|}{T-h-t_0+1},$$

for $i = 1, 2, 3$ where $E(y_{t+h}|y^R_{1:t})$ is the mean of the predictive density and $\hat{y}^M_{t+h}$ is the median of the predictive density. For the density forecasts, we report the average log-predictive likelihoods (ALPL) and the average continuous rank probability score (ACRPS),

$$\text{ALPL}_i = \frac{\sum_{t=t_0}^{T-h}\log p_{t+h}(y^*_{t+h,i} = y^R_{t+h,i}|y^R_{1:t})}{T-h-t_0+1},$$

$$\text{ACRPS}_i = \frac{1}{T-h-t_0+1}\sum_{t=t_0}^{T-h}\text{CRPS}_{t,i},$$

for $i = 1, 2, 3$ where $\text{CRPS}_{t,i} = \int_{-\infty}^{\infty}\left(F_{t+h}(z) - 1(y^R_{t+h} < z)\right)^2 dz = E_{p_{t+h}}|y^*_{t+h,i} - y^R_{t+h,i}| - 0.5E_{p_{t+h}}|y^*_{t+h,i} - y^R_{t+h,i}|$ and $F_{t+h}(\bullet)$ is the c.d.f. of the predictive density. A small value of the ACRPS_i indicates a better forecasting performance.

We also present a joint ALPL for the three core variables of interest:

$$\text{ALPL} = \frac{\sum_{t=t_0}^{T-h}\log p_{t+h}(y^*_{t+h} = y^R_{t+h}|y^R_{1:t})}{T-h-t_0+1}.$$

5.2 Estimating Variances and Covariances

5.2.1 Monte Carlo Study

The DGP is obtained by first estimating the VAR-SV in (1), (2) and (3) using the small data set so as to obtain estimates (posterior means) of a_t, h_t and β. We then generate 100 artificial datasets (with same sample size as the actual data) from the VAR-SV with parameters and states set to these estimates. For each dataset, we use various VAR-CL approaches to estimate σ_{ijt} for $i, j = 1, 2, 3$ where σ_{ijt} denotes the $(i, j)^{th}$ element of the error covariance matrix at time t. The results are in Figures 1 through 4. All lines in these figures (except the one for the true parameter path) represents an average over the 100 datasets. It can be seen that the average of the point estimates for all approaches tracks the true parameter path fairly well and that the coverage of the intervals is excellent for all the choices of weights used with the composite likelihood approaches. Even the use of equal weights leads to good coverage properties.

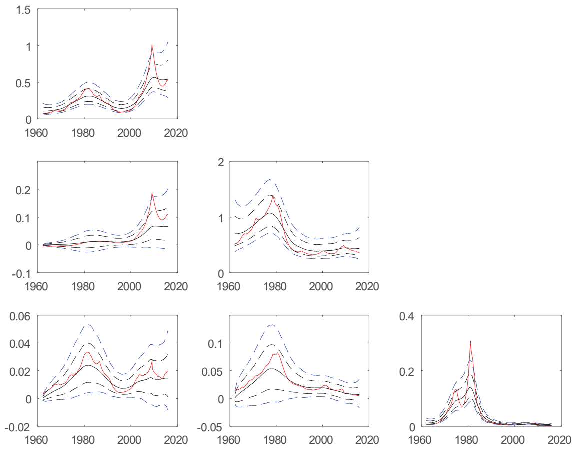

Figure 1: Monte Carlo Results for VAR-CL-ML. Solid red line: true parameter path. Black line: posterior median. Dotted lines: 16th/84th and 5th/95th percentiles.

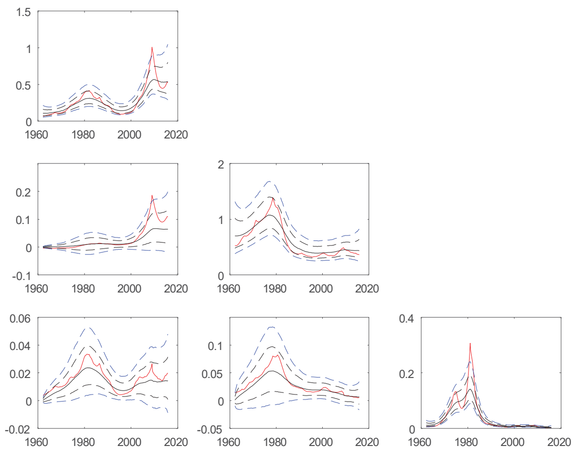

Figure 2: Monte Carlo Results for VAR-CL-BIC. Solid red line: true parameter path. Black line: posterior median. Dotted lines: 16th/84th and 5th/95th percentiles.

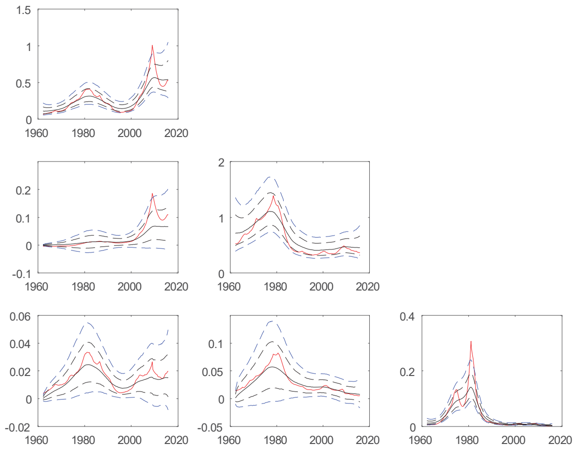

Figure 3: Monte Carlo Results for VAR-CL-DIC. Solid red line: true parameter path. Black line: posterior median. Dotted lines: 16th/84th and 5th/95th percentiles.

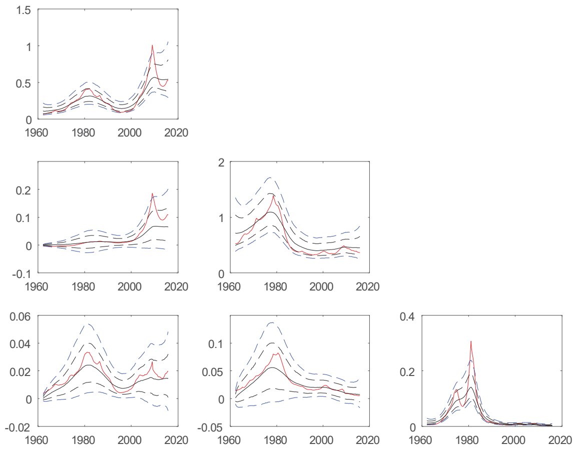

Figure 4: Monte Carlo Results for VAR-CL-EQ. Solid red line: true parameter path. Black line: posterior median. Dotted lines: 16th/84th and 5th/95th percentiles.

5.2.2 Estimation Results Using the Small Data Set

In this sub-section, we present results for a variety of approaches using the actual data. For the sake of brevity, the figures only presents results for a few main approaches. The credible intervals in this sub-section cover the 16th to 84th percentiles.

Figure 5 provides point estimates from two of the main composite likelihood approaches as well as VAR-SV and VAR-CCM2 (as we shall see below, VAR-CCM2 is in many cases the best alternative approach). It can be seen that all of the approaches track the VAR-SV fairly well, although VAR-CCM2 tracks it slightly more closely than our composite likelihood approaches for σ_{31t} and σ_{32t}.

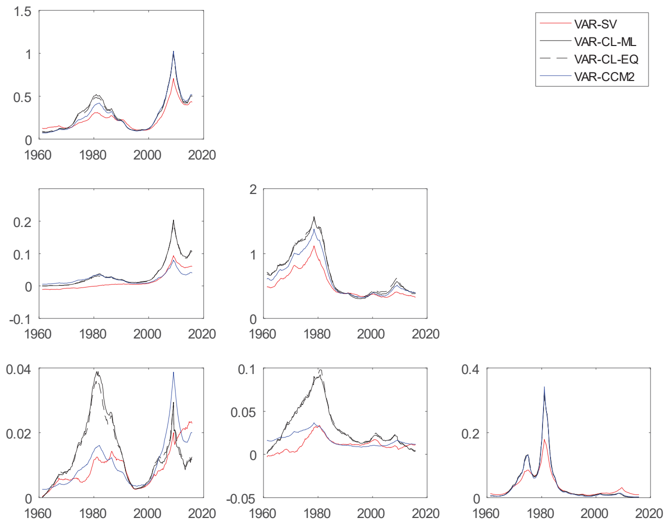

Figure 5: Point estimates of $\sigma_{ij,t}$ for $i, j = 1, 2, 3$

Figure 6 offers a more detailed comparison of one of our major composite likelihood approaches (VAR-CL-ML) to the unrestricted VAR-SV (the other composite likelihood and linear pooling approaches reveal similar patterns). It can be seen that, even for σ_{31t} and σ_{32t}, where the point estimates differ somewhat, the credible intervals always overlap. We take this as evidence that our composite likelihood approaches are doing a good job of matching the VAR-SV. The VAR-CCM2 produces similarly accurate estimates. However, it is worth noting that the VAR-CCM1 and VAR-HM do not. This is revealed in Figures 7 and 8 which present detailed results for these two models. From the former, we can see that the common drifting volatility assumption in VAR-CCM1 is too restrictive, with high volatility in $\sigma_{11,t}$ spilling over inappropriately into some of the other variances and covariances. From Figure 8 we can see the homoskedastic model is failing to pick up changes in volatility that are clearly present in the data.

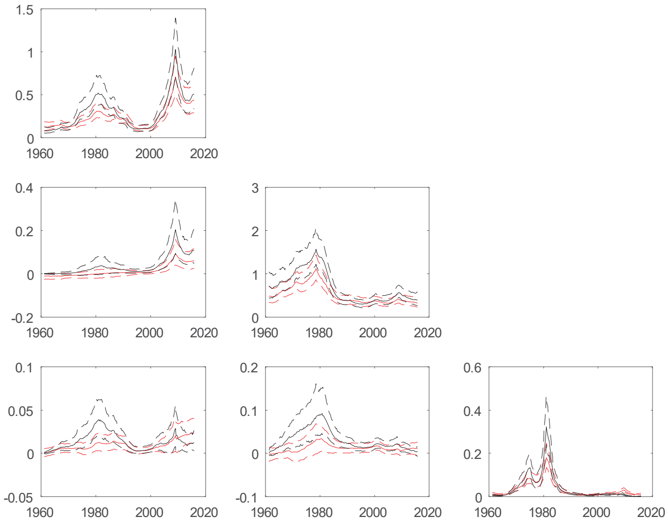

Figure 6: Comparison of VAR-CL-ML to VAR-SV (Point estimates of $\sigma_{ij,t}$ with 16%-84th percentiles, VAR-SV in red)

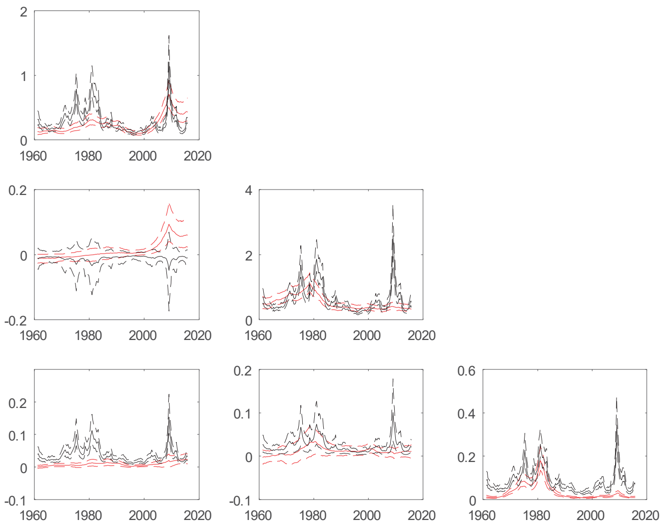

Figure 7: Comparison of VAR-CCM1 to VAR-SV (Point estimates of $\sigma_{ij,t}$ with 16%-84th percentiles, VAR-SV in red)

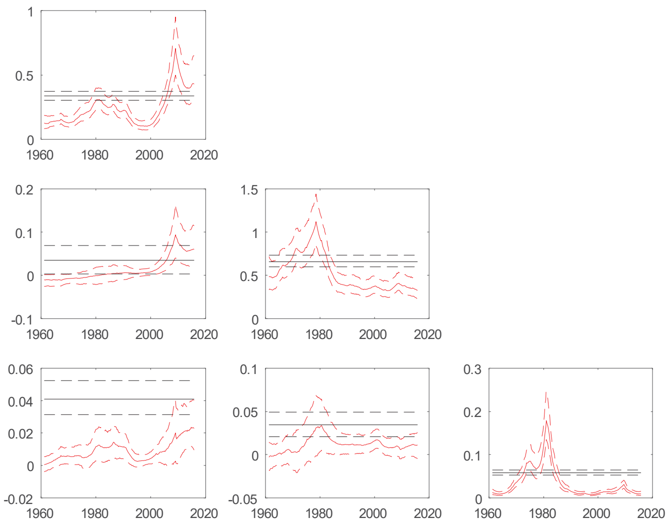

Figure 8: Comparison of VAR-HM to VAR-SV (Point estimates of $\sigma_{ij,t}$ with 16%-84th percentiles, VAR-SV in red)

In this sub-section, we have compared our composite likelihood approaches to a range of alternatives using a small data set where such a comparison is feasible. Of course, with such a small data set, the researcher would probably want to work with a VAR-SV (or similar model) since it is the more flexible approach and, thus, more able to capture empirically-relevant features of the data. But it is re-assuring to see that even with the small data set, composite likelihood methods are producing results which are very similar to the VAR-SV.

5.3 Forecasting

In this sub-section, we investigate how well composite likelihood methods forecast using the large data set involving 196 variables. We remind the reader that, with this many variables, the only other feasible Bayesian VAR approach is the one with acronym Large VAR which is homoskedastic and uses a natural conjugate prior. We also include all the models of the preceding sub-section, but for these other models we are using the small data set to produce forecasts. We present results for a long forecast evaluation period (beginning in 1970) and a short forecast evaluation period that begins in 2008Q1 so as to take in the financial crisis and subsequent period. In both cases the forecast evaluation period runs to the end of the sample. We provide forecasts of quarterly variables ($h = 1$) and quarterly variables one year in the future ($h = 4$). To aid in interpretation, note that all variables are standardized to have zero mean and unit standard deviation and that our forecast metrics are not benchmarked against any model. We carry out the sign test of equal predictive accuracy of Diebold and Mariano (1995). All tests compare a specific model to the homoskedastic large VAR. In the tables, ***, ** and * denote rejection of the null hypothesis of equal predictive accuracy of a model and the benchmark at the 1%, 5% and 10% level of significance, respectively.

The best overall summary of forecast performance involves the entire joint predictive density for the three core variables. This is presented in Table 2 for $h = 1$ and $h = 4$. The most important comparison is between the Large VAR and the methods which pool results from many small models, since these are the only feasible approaches with large data sets. In this comparison, it can be seen that the composite likelihood approaches are clearly winning for both forecast horizons, particularly for the forecast evaluation period which begins in 2008Q1. For the longer forecast evaluation period with $h = 1$, the linear pooling method actually forecasts slightly better than logarithmic pooling used with the composite likelihood approaches. In general, any method which involves homoskedasticity or highly restrictive forms for the error covariance matrix (i.e. VAR-CCM1) forecast poorly when evaluated using the ALPL for the 3 core variables. The less restrictive VAR-CCM2 forecasts well over the longer forecast evaluation period but is beaten by composite likelihood methods for the shorter evaluation period. These statements hold true for both $h = 1$ and $h = 4$. The forecast improvements relative to the Large VAR are statistically significant in almost every case. The only exceptions are for $h = 4$ for the forecast evaluation period which begins in 2008Q1.

Table 2: Forecasting Evaluation Using Joint ALPL for 3 Core Variables

Horizon	$h = 1$		$h = 4$	
Evaluation begins:	1970Q1	2008Q1	1970Q1	2008Q1
VAR-HM	0.33^{***}	-0.58^{***}	-1.04^{***}	-1.60
VAR-SV	0.65^{***}	0.44^{***}	-1.04^{***}	-1.61
VAR-CCM1	0.06^{***}	-0.51^{**}	-0.98^{***}	-1.85
VAR-CCM2	0.90^{***}	0.52^{***}	-0.84^{***}	-1.58
Large VAR	-0.47	-1.69	-1.41	-2.02
VAR-CL-ML	0.90^{***}	1.27^{***}	-0.99^{***}	-1.49
VAR-CL-DIC	0.85^{***}	0.67^{***}	-0.72^{***}	-0.92^{***}
VAR-CL-BIC	0.90^{***}	1.15^{***}	-0.88^{***}	-1.51
VAR-CL-EQ	0.88^{***}	0.89^{***}	-0.71^{***}	-0.84^{***}
VAR-CL-LIN	0.89^{***}	0.92^{***}	-0.71^{***}	-0.79^{***}
VAR-LIN	0.91^{***}	1.01^{***}	-0.75^{***}	-0.83^{***}
small VAR-HM	0.27^{***}	-0.36^{***}	-1.03^{***}	-0.69
small VAR-SV	0.63^{***}	0.77^{***}	-0.98^{***}	-0.48^{**}
VAR-SV-R	0.89^{***}	0.59^{***}	-0.76^{***}	-0.75
VAR-SV-R1	-0.01^{***}	-0.14^{**}	-1.19^{**}	-0.99

The following tables present detailed results for the individual variables using the full range of forecast metrics. The good forecast performance of composite likelihood methods and relatively poor forecasting performance of the large homoskedastic VAR noted in Table 2 are also found in these tables, but there are some differences across variables and forecast metrics worth noting.

The general pattern is that composite likelihood and linear pooling methods forecast particularly well for inflation and the interest rate, for the post-2008 period and using metrics that involve the entire predictive density (i.e. ACRPS and ALPL). The last point is not that surprising in that incorporation of stochastic volatility is usually found to be more important in getting the shape of the entire predictive density correct as opposed to just getting a reasonable point forecast. For example, for inflation over the longer forecast evaluation period with $h = 1$, the homoskedastic VAR-HM model is actually forecasting quite well if we look at RMSFE and MAE. However its ALPL is not as high as other methods for this case. It is worth noting that this pattern does not hold for $h = 4$. Also, for the interest rate, the small homoskedastic model produces poor point forecasts, especially after 2008. And the homoskedastic large VAR often produces high RMSFEs. So the reader should not take away the message that, if point forecasts are all that matter, then working with homoskedastic models is adequate.

For GDP growth, Tables 7 and 8 indicate that the small VAR-SV forecasts best for $h = 1$ and, in general, small models such as VAR-CCM2 tend to forecast well. But even here, the forecast performance of composite likelihood methods is only slightly worse than these models. For $h = 4$, composite likelihood methods tend to produce superior forecasts.

In general, of the alternative models, the VAR-CCM2 tends to forecast almost as well as our methods (and forecasts much better than VAR-CCM1). However, we stress that VAR-CCM2 is not computationally feasible in the really large VARs macroeconomists are increasingly interested in.

These tables also reinforce the finding that, among the various composite likelihood approaches, the alternative ways of doing the weighting typically do not make a great deal of difference for forecasting. There is no consistent pattern where one weighting method dominates and it is always possible to find case where a particular set of weights forecasts best. There are also cases where a linear pool of sub-models forecasts best. Indeed, even using equal weights produces forecasts which are only slightly inferior to other methods which estimate weights in a data-based fashion.

Table 3: Evaluation of Inflation Forecasts Beginning in 1970

	$h = 1$				$h = 4$			
	RMSFE	MAE	ACRPS	ALPL	RMSFE	MAE	ACRPS	ALPL
VAR-HM	0.66	0.45	0.36^{***}	-0.15^{***}	0.88	0.64	0.50^{***}	-0.53
VAR-SV	0.67	0.46	0.36^{***}	-0.06^{***}	0.88^{**}	0.65^{**}	0.51^{***}	-0.49^{**}
VAR-CCM1	0.71	0.51	0.39^{***}	-0.12	0.90	0.65	0.50^{***}	-0.43^{***}
VAR-CCM2	0.67	0.46	0.36^{***}	-0.00^{***}	0.87^{***}	0.64^{***}	0.50^{***}	-0.39^{***}
Large VAR	0.73	0.52	0.56	-0.14	1.03	0.79	0.82	-0.64^{***}
VAR-CL-ML	0.69	0.47	0.36^{***}	-0.01^{**}	0.68^{***}	0.48^{***}	0.48^{***}	-0.39^{***}
VAR-CL-DIC	0.68	0.47	0.36^{***}	-0.01	0.67^{***}	0.47^{***}	0.49^{***}	-0.36^{***}
VAR-CL-BIC	0.69	0.46	0.36^{***}	-0.01^{**}	0.66^{***}	0.47^{***}	0.48^{***}	-0.38^{***}
VAR-CL-EQ	0.68	0.47	0.36^{***}	-0.01	0.66^{***}	0.47^{***}	0.48^{***}	-0.35^{***}
VAR-CL-LIN	0.68	0.47	0.36^{***}	0.00	0.66^{***}	0.47^{***}	0.48^{***}	-0.34^{***}
VAR-LIN	0.68	0.47	0.38^{***}	-0.00	0.67^{***}	0.48^{***}	0.50^{***}	-0.36^{***}
small VAR-HM	0.67^{*}	0.46	0.35^{***}	-0.16^{**}	0.88	0.64	0.49^{***}	-0.52
small VAR-SV	0.68	0.46	0.36^{***}	-0.08^{***}	0.87^{**}	0.63^{**}	0.48^{***}	-0.47^{***}
VAR-SV-R	0.67	0.46	0.36^{***}	0.00^{***}	0.86^{***}	0.62^{***}	0.49^{***}	-0.37^{***}
VAR-SV-R1	0.81	0.58	0.66^{***}	-0.18^{***}	1.04	0.79	0.85^{***}	-0.53^{***}

Table 4: Evaluation of Inflation Forecasts Beginning in 2008

	$h = 1$				$h = 4$			
	RMSFE	MAE	ACRPS	ALPL	RMSFE	MAE	ACRPS	ALPL
VAR-HM	1.04	0.66	0.52^{***}	-1.16	1.13	0.79	0.61^{***}	-0.94
VAR-SV	1.06	0.68	0.54^{***}	-0.68	1.11	0.75	0.60^{***}	-0.80
VAR-CCM1	1.04	0.66	0.52^{***}	-0.71	1.06	0.71	0.56^{***}	-0.78
VAR-CCM2	1.05	0.68	0.53^{***}	-0.57	1.08	0.72	0.58^{***}	-0.65
Large VAR	1.03	0.65	0.69	-0.71	1.25	0.88	0.94	-1.00
VAR-CL-ML	1.04	0.65	0.51^{***}	-0.54	0.97	0.60	0.54^{***}	-0.59
VAR-CL-DIC	1.04	0.66	0.52^{**}	-0.57	0.95	0.59	0.54^{***}	-0.57^{***}
VAR-CL-BIC	1.02	0.63	0.50^{***}	-0.50	0.97	0.60	0.54^{***}	-0.60
VAR-CL-EQ	1.04	0.66	0.52^{***}	-0.57	0.96	0.61	0.54^{***}	-0.57^{***}
VAR-CL-LIN	1.04	0.66	0.52^{**}	-0.50	0.95	0.58	0.54^{***}	-0.55^{***}
VAR-LIN	1.03	0.66	0.54^{**}	-0.48	0.96	0.61	0.54^{***}	-0.55^{***}
small VAR-HM	1.02	0.63	0.49^{***}	-1.11	1.10	0.67	0.53^{***}	-0.77
small VAR-SV	1.04	0.66	0.51^{**}	-0.67	1.02	0.58	0.49^{***}	-0.59
VAR-SV-R	1.05	0.68	0.53^{***}	-0.56^{**}	1.05	0.64	0.53^{***}	-0.52
VAR-SV-R1	1.28^{**}	0.88^{**}	0.77	-0.63^{**}	1.21	0.77	0.78	-0.68^{***}

Table 5: Evaluation of Interest Rate Forecasts Beginning in 1970

	$h = 1$				$h = 4$			
	RMSFE	MAE	ACRPS	ALPL	RMSFE	MAE	ACRPS	ALPL
VAR-HM	0.29***	0.18***	0.15***	0.81***	0.62**	0.47**	0.36***	−0.06***
VAR-SV	0.28***	0.17***	0.14***	1.03***	0.59**	0.45**	0.35***	−0.01***
VAR-CCM1	0.51***	0.33***	0.25***	0.53***	0.68	0.51	0.39***	−0.10**
VAR-CCM2	0.28***	0.17***	0.14***	1.19***	0.59***	0.44***	0.34***	0.02***
Large VAR	0.56	0.42	0.44	0.17	0.75***	0.55***	0.60	−0.15
VAR-CL-ML	0.28***	0.17***	0.13***	1.18***	0.46***	0.34***	0.38***	−0.20
VAR-CL-DIC	0.28***	0.16***	0.13***	1.17***	0.36***	0.26***	0.34***	0.04***
VAR-CL-BIC	0.28***	0.17***	0.13***	1.20***	0.41***	0.32***	0.36***	−0.10
VAR-CL-EQ	0.27***	0.16***	0.13***	1.19***	0.37***	0.26***	0.34***	0.02***
VAR-CL-LIN	0.27***	0.16***	0.13***	1.20***	0.36***	0.26***	0.34***	0.01***
VAR-LIN	0.27***	0.16***	0.13***	1.21***	0.37***	0.26***	0.35***	−0.01***
small VAR-HM	0.28***	0.17***	0.15***	0.82***	0.62	0.46	0.35***	−0.06
small VAR-SV	0.28***	0.17***	0.14***	1.06***	0.60	0.44	0.34***	−0.03***
VAR-SV-R	0.28***	0.17***	0.14***	1.18***	0.58**	0.43**	0.34***	0.04***
VAR-SV-R1	0.40***	0.26***	0.37***	0.64***	0.76	0.55	0.68***	−0.11**

Table 6: Evaluation of Interest Rate Forecasts Beginning in 2008

	$h = 1$				$h = 4$			
	RMSFE	MAE	ACRPS	ALPL	RMSFE	MAE	ACRPS	ALPL
VAR-HM	0.25***	0.18***	0.14***	0.97***	0.65	0.56	0.39***	−0.12
VAR-SV	0.18***	0.12***	0.10***	1.50***	0.62	0.56	0.39***	−0.16
VAR-CCM1	0.36***	0.30***	0.20***	0.66***	0.74	0.70	0.48***	−0.49
VAR-CCM2	0.20***	0.12***	0.10***	1.45***	0.64	0.57	0.40***	−0.15
Large VAR	0.51	0.45	0.46	0.09	0.72	0.62	0.62	−0.21
VAR-CL-ML	0.13***	0.07***	0.06***	2.00***	0.46**	0.43	0.39***	−0.32
VAR-CL-DIC	0.13***	0.08***	0.07***	1.67***	0.27***	0.25***	0.26***	0.27***
VAR-CL-BIC	0.13***	0.07***	0.06***	1.88***	0.48	0.45	0.39***	−0.28
VAR-CL-EQ	0.12***	0.08***	0.07***	1.79***	0.28***	0.26***	0.26***	0.27***
VAR-CL-LIN	0.12***	0.08***	0.07***	1.78***	0.29***	0.26***	0.27***	0.24***
VAR-LIN	0.12***	0.08***	0.07***	1.83***	0.29***	0.26***	0.30***	0.25***
small VAR-HM	0.16***	0.11***	0.10***	1.17***	0.45*	0.36*	0.25***	0.29**
small VAR-SV	0.12***	0.08***	0.07***	1.83***	0.42	0.33	0.23***	0.36**
VAR-SV-R	0.17***	0.11***	0.09***	1.54***	0.56	0.49	0.34***	0.04
VAR-SV-R1	0.52***	0.27***	0.32***	0.90***	0.67	0.48	0.52	0.08

Table 7: Evaluation of GDP Growth Forecasts Beginning in 1970

	$h = 1$				$h = 4$			
	RMSFE	MAE	ACRPS	ALPL	RMSFE	MAE	ACRPS	ALPL
VAR-HM	0.89	0.68	0.51***	−0.38	1.01**	0.76**	0.58***	−0.51**
VAR-SV	0.86	0.65	0.50***	−0.32**	1.00***	0.74***	0.57***	−0.51***
VAR-CCM1	0.87	0.67	0.51***	−0.36	1.00**	0.76**	0.58***	−0.52***
VAR-CCM2	0.86	0.66	0.50***	−0.31**	1.00***	0.74***	0.58***	−0.50***
Large VAR	0.93	0.70	0.77	−0.39	1.14	0.89	0.98	−0.62
VAR-CL-ML	0.92	0.67	0.51***	−0.35	0.98**	0.72***	0.56***	−0.49***
VAR-CL-DIC	0.91	0.67	0.51***	−0.36	0.97**	0.72**	0.56***	−0.48***
VAR-CL-BIC	0.93	0.68	0.52***	−0.35***	0.99**	0.73**	0.57***	−0.49***
VAR-CL-EQ	0.92	0.68	0.51***	−0.35	0.97**	0.71**	0.56***	−0.47***
VAR-CL-LIN	0.92	0.68	0.51***	−0.35	0.97**	0.71**	0.55***	−0.47***
VAR-LIN	0.92	0.68	0.54***	−0.36	0.98**	0.72**	0.57***	−0.47***
small VAR-HM	0.96	0.71	0.54***	−0.45***	1.01	0.74	0.56***	−0.49**
small VAR-SV	0.92	0.68	0.52***	−0.39	0.98**	0.71**	0.55***	−0.47***
VAR-SV-R	0.86	0.65	0.49***	−0.30*	0.98**	0.72**	0.56***	−0.46***
VAR-SV-R1	1.04	0.78	0.89***	−0.46***	1.17	0.91	0.98	−0.61***

Table 8: Evaluation of GDP Growth Forecasts Beginning in 2008

	$h = 1$				$h = 4$			
	RMSFE	MAE	ACRPS	ALPL	RMSFE	MAE	ACRPS	ALPL
VAR-HM	0.96	0.72	0.56***	−0.48	1.14	0.85	0.64***	−0.70
VAR-SV	0.86	0.63	0.50***	−0.42	1.07	0.77	0.62***	−0.83
VAR-CCM1	0.94	0.73	0.57***	−0.57	1.16	0.87	0.67***	−0.88
VAR-CCM2	0.88	0.65	0.52***	−0.46	1.11	0.82	0.65***	−0.85
Large VAR	0.96	0.77	0.80	−0.47	1.14	0.87	0.99	−0.69
VAR-CL-ML	0.95	0.65	0.52***	−0.46	1.20	0.81	0.63***	−0.83
VAR-CL-DIC	0.95	0.66	0.53***	−0.50	1.11	0.76	0.61***	−0.79
VAR-CL-BIC	0.96	0.67	0.52***	−0.47	1.20	0.83	0.65***	−0.86
VAR-CL-EQ	0.95	0.66	0.52***	−0.47	1.11	0.75	0.60***	−0.76
VAR-CL-LIN	0.95	0.66	0.52***	−0.45	1.10	0.75	0.60***	−0.77
VAR-LIN	0.96	0.68	0.56**	−0.46	1.12	0.76	0.62***	−0.77
small VAR-HM	1.07	0.76	0.58**	−0.57	1.04	0.69	0.53***	−0.48
small VAR-SV	0.94	0.66	0.52**	−0.50	0.95	0.61	0.49***	−0.52
VAR-SV-R	0.86	0.63	0.51***	−0.42	0.98	0.66	0.53***	−0.52
VAR-SV-R1	1.11	0.84	0.87	−0.54	1.05*	0.75*	0.85	−0.47

6 The Computational Advantages of Composite Likelihood Methods

We have argued in this paper that the main advantage of our composite likelihood approach is computational. It is computationally feasible in a Big Data context where other approaches which incorporate stochastic volatility are not. To reinforce this point, in this section we present some results showing the computational properties of the composite likelihood approaches relative to others.

Table 9 presents results relating to the composite likelihood approach when using the large data set and doing one run of our simulation algorithm using the full sample.[7] Note that our algorithm involves two steps: (i) estimating all the quasi-posteriors using MCMC (labelled "Estimation" in Table 9) and (ii) using an accept-reject algorithm to pool draws (labelled "Pooling" in Table 9). Note also that, within step (i), we run things in parallel across different quasi-posteriors and some variants require marginal likelihood or information criteria estimation. The results in the table are based on taking 22,000 draws from each quasi-posterior which, we have found, is the minimum necessary to obtain reasonable effective sample sizes for all our MCMC algorithms. The first 2,000 draws from each quasi-posterior are burn-in draws which are dropped. The remaining 20,000 are then thinned to 1,000 to reduce correlation between draws. This leaves us with 1,000×193=193,000 draws which are used in step (ii). The time to do step (ii) is calculated as the time taken to obtain 1,000 retained draws from these 193,000 draws. Note that these final 1,000 may contain some repeating draws and, if such repeats are too high, this will make the effective sample size of the algorithm low. To show that this is not a substantive problem with our composite likelihood approaches, the table also contains a column labelled "Unique" which is the percentage of draws which are unique and do not repeat.

It can be seen that, even with our very large data set, computation can easily be done by a good PC with running time being roughly an hour. The equally-weighted composite likelihood approach is faster due to the fact that it does not require the calculation of marginal likelihoods or an information criterion.

Table 9: Computational Time in Minutes of Composite Likelihood Approaches

	Estimation	Pooling	Total	Unique
VAR-CL-ML	62.4	5.5	67.9	46.7%
VAR-CL-DIC	60.6	0.4	61.0	65.0%
VAR-CL-BIC	60.6	2.7	63.3	53.2%
VAR-CL-EQ	34.3	11.2	45.5	99.9%

It is worth noting that the linear opinion pool (VAR-LIN) is much more computationally demanding since it involves recursive estimation and numerical optimization (see Geweke and Amisano, 2011). The computational time comparable to those reported in Table 9 is 62.4 hours.

Table 10 presents computational time for VARs of different dimensions for the alternative approaches which allow for stochastic volatility and for one of our composite likelihood approaches. For the composite likelihood approach, the time reported is to carry out the same exercise as was used to produce the numbers in Table 9. For the other approaches, it is the time to produce 22,000 MCMC draws.

Table 10 shows that the composite likelihood approach and VAR-CCM1 are

[7] All computation was done on a Dell Precision Tower 7910 with 2 Intel Xeon 3.10Ghz processors (total of 20 cores) and 256GB of memory.

the only approaches likely to be computationally feasible in truly large VARs. But, as we have seen, VAR-CCM1 is likely to be too restrictive in many empirical contexts. Computational times, of course, increase with VAR dimension. But with the composite likelihood approach this increase is approximately linear in N. With VAR-SV the increase is much more rapid (not quadratic, but close to it). Even for $N = 100$, computation time with the VAR-SV is more than a week on a good PC for a single run of the algorithm. VAR-CCM2 is not this bad (for $N = 100$ its running time is a few hours), but running time is much more than for our composite likelihood approaches and it is increasing at a more than linear rate in N. The latter fact is likely to preclude its use in very large VARs.

Table 10: Computation Time in Minutes for Different VAR dimensions

N	VAR-CL-EQ	VAR-SV	VAR-CCM1	VAR-CCM2
3	n.a.	0.74	0.27	0.44
7	2.67	6.34	0.40	1.34
20	4.01	112.67	0.82	5.42
50	10.95	1602.61	1.79	37.89
100	21.14	13071.09	7.72	160.24

7 Summary and Conclusions

Large VARs are emerging as a popular tool in modern macroeconomics. Adding multivariate stochastic volatility to them has emerged as one of the unresolved challenges in the field. It arises since it is not computationally practical to carry out Bayesian estimation in large VARs with multivariate stochastic volatility. Even if computation were possible, conventional approaches can be over-parameterized when working with large data sets leading to problems with over-fitting, imprecise estimation and the need for strong prior information. In this paper, we propose the use of composite likelihood methods for meeting this challenge. These involve averaging over many smaller models. In our context, we use many small VAR-SVs thus enabling computation to be feasible even in data sets involving hundreds of variables. By working with smaller models, concerns over over-parameterization and the need for careful prior elicitation are lessened. We explore these themes in the paper. In addition, we discuss the econometric theory of composite likelihood methods drawing on conventional asymptotic results as well as the literature on prediction pools. All in all, there are strong theoretical reasons for thinking composite likelihood methods may be an attractive way of adding stochastic volatility to large VARs.

The issue of how well composite likelihood methods work in practice is explored in our empirical work. Working with a large US quarterly macroeconomic data set involving 196 variables, we find encouraging results. When we use all 196 variables and compare the forecast performance of our composite likelihood methods against the only practical alternative (a large homoskedastic VAR with natural conjugate prior), we find strong evidence of the superiority of our methods. Clearly, stochastic volatility is an important feature of this data set and our VAR-CL-SV methods allow for this.

When we compare our methods to a range of existing methods which include stochastic volatility we must restrict ourselves to smaller data sets. Using these, we find our composite likelihood methods are producing parameter estimates which are similar to those produced by state-of-the-art approaches. We also find that composite likelihood methods using the large data set forecast well relative to these other methods which use the small data set. Overall, we conclude that the strategy of combining forecasts from many small models is computationally feasible even with large VARs and leads to forecast performance that is better than other computationally feasible approaches.

References

Banbura, M., Giannone, D. and Reichlin, L. (2010). Large Bayesian vector autoregressions. Journal of Applied Econometrics, 25, 71-92.

Bernanke, B., Boivin, J. and Eliasz, P. (2005). Measuring monetary policy: A Factor augmented vector autoregressive (FAVAR) approach. Quarterly Journal of Economics, 120, 387-422.

Bloor, C. and Matheson, T. (2010). Analysing shock transmission in a data-rich environment: a large BVAR for New Zealand. Empirical Economics 39, 537-558.

Canova, F. and Matthes, C. (2017). A composite likelihood approach for dynamic structural models. Manuscript available at
http://apps.olin.wustl.edu/conf/SBIES/Files/pdf/2017/228.pdf.

Carriero, A., Clark, T. and Marcellino, M. (2016a). Common drifting volatility in large Bayesian VARs. Journal of Business & Economic Statistics, 34, 375-390.

Carriero, A., Clark, T. and Marcellino, M. (2016b). Large Vector Autoregressions with stochastic volatility and flexible priors. Working paper 1617, Federal Reserve Bank of Cleveland.

Carriero, A., Clark, T. and Marcellino, M. (2016c). Measuring uncertainty and its impact on the economy. Working paper 1622, Federal Reserve Bank of Cleveland.

Carriero, A., Kapetanios, G. and Marcellino, M. (2010). Forecasting exchange rates with a large Bayesian VAR. International. Journal of Forecasting, 25, 400-417.

Chan, J. (2018). Large Bayesian VARs: A flexible Kronecker error covariance structure. Journal of Business & Economic Statistics, forthcoming.

Chan, J. and Eisenstat, E. (2018). Bayesian model comparison for time-varying parameter VARs with stochastic volatility. Journal of Applied Econometrics, forthcoming.

Chan, J. and Grant, A. (2016). On the observed-data Deviance Information Criterion for volatility modeling. Journal of Financial Econometrics, 14, 772-802.

Clark, T. (2011). Real-time density forecasts from BVARs with stochastic volatility. Journal of Business and Economic Statistics 29, 327-341.

Creal, D. and Tsay, R. (2015). High dimensional dynamic stochastic copula models. Journal of Econometrics 189, 335-345.

D'Agostino, A., Gambetti, L. and Giannone, D. (2013). Macroeconomic forecasting and structural change. Journal of Applied Econometrics 28, 82-101.

Diebold, F. and Mariano, R. (1995). Comparing predictive accuracy. Journal of Business and Economic Statistics 13, 253-263.

Durrett, R. (2010). Probability: Theory and Examples, Fourth Edition. Cambridge University Press, Cambridge.

Gefang, D. (2014). Bayesian doubly adaptive elastic-net Lasso for VAR shrinkage. International Journal of Forecasting, 30, 1-11.

Genest, C., Weerahandi, S., Zidek, J. (1984). Aggregating opinions through logarithmic pooling. Theory and Decision 17, 61–70.

Genest, C., McConway, K. and Schervish, M. (1986). Characterization of externally Bayesian pooling operators. Annals of Statistics, 14, 487-501.

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

Giannone, D., Lenza, M. and Primiceri, G. (2015). Prior selection for Vector Autoregressions 97, 436-451.

Hall, S. and Mitchell, J. (2007). Combining density forecasts. International Journal of Forecasting 23, 1-13.

Horn, R. and Johnson, C. (1991). Topics in Matrix Analysis. Cambridge University Press, Cambridge.

Koop, G. (2013). Forecasting with medium and large Bayesian VARs. Journal of Applied Econometrics 28, 177-203.

Koop, G. and Korobilis, D. (2016). Model uncertainty in panel vector autoregressive models. European Economic Review 81, 115-131.

Koop, G., Leon-Gonzalez, R. and Strachan, R. (2009). On the evolution of the monetary policy transmission mechanism. Journal of Economic Dynamics and Control, 33, 997–1017.

Korobilis, D. (2013). VAR forecasting using Bayesian variable selection. Journal of Applied Econometrics 28, 204-230.

Kroese, D., Taimre, T. and Botev, Z. (2011). Handbook of Monte Carlo Methods. John Wiley and Sons, New York.

McCracken, M. and Ng, S. (2015). FRED-MD: A monthly database for macroeconomic research. Federal Reserve Bank of St. Louis, working paper 2015-012A.

McCracken, M., Owyang, M. and Sekhposyan, T. (2016). Real-time forecasting with a large, mixed frequency Bayesian VAR, manuscript available at http://www.tateviksekhposyan.org/.

Pakel, C., Shephard, N., Sheppard, K. and Engle, R. (2014). Fitting vast dimensional time-varying covariance models,
available at http://staff.bilkent.edu.tr/cavit/research/.

Primiceri. G. (2005). Time varying structural vector autoregressions and monetary policy. Review of Economic Studies 72, 821-852.

Qu, Z. (2016). A composite likelihood approach to analyze singular DSGE models, Boston University manuscript.

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

Roche, A. (2016). Composite Bayesian inference, working paper available at https://arxiv.org/abs/1512.07678.

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

Technical Appendix

Priors and Specification Choices

For the VAR-SV model, we assume normal priors for the initial condition $a_0 \sim N(0, V_a)$ and $h_0 \sim N(0, V_h)$. Moreover, we assume an independent prior for parameters in Σ_h and Σ_a which are distributed as

$$\sigma^2_{h,i} \sim IG(\nu_{h,i}, S_{h,i}), \quad \sigma^2_{a,j} \sim IG(\nu_{a,j}, S_{a,2}),$$

for $i = 1, \ldots, N$ and $j = 1, \ldots, \frac{N(N-1)}{2}$. We set $\nu_{h,i} = 10$, $S_{h,i} = 0.1^2(\nu_{h,i} - 1)$, $\nu_{a,j} = 10$ and $S_{h,j} = 0.01^2(\nu_{h,j} - 1)$. For the initial states, we set $V_h = 10 \times I_N$ and $V_a = 10 \times I_{\frac{N(N-1)}{2}}$.

For the VAR coefficients $\beta = \text{vec}\left((c, A_1, , \ldots, A_p)'\right)$, we use a Minnesota prior and assume $\beta \sim N(\beta_0, V_\beta)$. For the prior mean, we set $\beta_0 = 0$. The prior covariance matrix V_β is set to be diagonal and its corresponding values are set as follows:

$$\text{Var}(c) = 10 \times I_N,$$
$$\text{Var}(A_l^{ij}) = \begin{cases} \frac{\lambda_1^2 \lambda_2}{l^{\lambda_3}} \frac{\sigma_i}{\sigma_j} & \text{for } l = 1, \ldots, p \text{ and } i \neq j, \\ \frac{\lambda_1^2}{l^{\lambda_3}} & \text{for } l = 1, \ldots, p \text{ and } i = j. \end{cases}$$

where A_l^{ij} denotes the (i, j) th element of the matrix A_l and σ_r is set equal to the standard deviation of the residual from an AR(p) model for the variable r. For the hyperparameters, we set $\lambda_1 = 0.2$, $\lambda_2 = 0.5$, $\lambda_3 = 2$, $p = 4$.

The VAR-CCM2 is the same as the VAR-SV except that the a_t is restricted to be time-invariant, i.e. $a_t = a$. We assume a normal prior $a \sim N(0, \Omega_a)$ with $\Omega_a = 10 \times I_{\frac{N(N-1)}{2}}$. The priors for other parameters are set the same as those in the VAR-SV.

For the VAR-HM

$$y + X\beta + \epsilon, \quad \epsilon_t \sim N(0, I_N \otimes \Sigma),$$

we assume an independent prior for the model parameters. The prior for the VAR coefficients is set equal to that in the VAR-SV. For the covariance matrix, we set $\Sigma \sim IW(\Sigma_0, \nu_0)$ with $\nu_0 = N + 2$ and $\Sigma_0 = (\nu_0 - N - 1)I_N$, where $IW(\cdot, \cdot)$ denotes the inverse Wishart distribution. This implies that the prior mean $E(\Sigma) = I_N$. We also include a natural conjugate prior version of the homoskedastic VAR for use with the large data set. For this we choose the same prior with the exception that the prior covariance matrix for β is the same as for VAR-CCM1 (see below).

For the VAR-CCM1, we first let $x_t' = (1, y_{t-1}', \ldots, y_{t-p}')$. It is convenient to specify the model as

$$Y = XA + U, \quad \text{vec}(U) \sim N(0, \Sigma \otimes \Omega)$$

where $Y = (y_1, \ldots, y_T)'$, $X = (x_1, \ldots, x_T)'$, $A = (c, A_1, \ldots, A_p)'$ and $\Omega = \text{diag}(e^{h_1}, \ldots, e^{h_T})$. Recall that the log volatility follow an AR(1) process

$$h_t = \rho h_{t-1} + \epsilon_t^h, \quad \rho \sim \mathcal{N}(0, \sigma_h^2),$$

with $|\rho| < 1$. A standard normal-inverse-Wishart prior are set for model parameters (A, Σ) as

$$\Sigma \sim IW(\Sigma_0, \nu_0), \quad \text{vec}(A)|\Sigma \sim N(\text{vec}(A_0), \Sigma \otimes V_{\mathbf{A}}).$$

The hyperparameters Σ_0 and ν_0 are set equal to those in VAR-HM. We set $A_0 = 0$ for the prior mean of the VAR coefficients. For the covariance matrix, we assume it to be $V_A = \text{diag}(v_1, \ldots, v_k)$ and set $v_i = \frac{\lambda_1^2 \sigma_r}{l^{\lambda_3}}$ for coefficients associated to lag l of variable r for $i = 2, \ldots, k$ and $v_1 = 10$. The other hyperparameters are set equal to those in VAR-SV. For the AR coefficient and the variance of the log volatility process, we assume

$$\rho \sim \mathcal{N}(\rho_0, V_\rho) \text{ for } |\rho| < 1, \quad \sigma_h^2 \sim \mathcal{N}(\nu_h, S_h)$$

with $\rho_0 = 0.9$, $V_\rho = 0.2^2$, $\nu_h = 10$ and $S_h = 0.1^2(\nu_h - 1)$.

Proof of Proposition 1

Proof. Defining $\tilde{y}_t^* = A_{y,t}y_t^* - c_y - \sum_{j=1}^{p} B_{yy,j}y_{t-j}^*$ it is straightforward to show the form of the restricted VAR-SV implies:

$$
\begin{aligned}
p(y_t \mid \cdot\ ) \propto \exp &\left\{ -\frac{1}{2} \left(\tilde{y}_t^* - \sum_{i=1}^{M} w_i \sum_{j=1}^{p} \frac{\beta_{yz,i,j} z_{i,t-j}}{g(M)} \right)' \Sigma_{y,t}^{-1} \left(\tilde{y}_t^* - \sum_{i=1}^{M} w_i \sum_{j=1}^{p} \frac{\beta_{yz,i,j} z_{i,t-j}}{g(M)} \right) \right\} \\
&\times \prod_{i=1}^{M} \exp \left\{ -\frac{1}{2} \left[h_{N^*+i,t} - \ln w_i \right.\right. \\
&\qquad \left.\left. + \ e^{-h_{N^*+i,t} + \ln w_i} \left(z_{i,t} - \alpha'_{z,i,t} y_t^* - c_{z,i} - \sum_{j=1}^{p} \beta'_{zy,i,j} y_{t-j}^* - \sum_{j=1}^{p} \beta_{zz,i,j} z_{i,t-j} \right)^2 \right] \right\} \\
\propto \exp &\left\{ \sum_{i=1}^{M} -\frac{w_i}{2} \left(\tilde{y}_t^* - \sum_{j=1}^{p} \frac{\beta_{yz,i,j} z_{i,t-j}}{g(M)} \right)' \Sigma_{y,t}^{-1} \left(\tilde{y}_t^* - \sum_{j=1}^{p} \frac{\beta_{yz,i,j} z_{i,t-j}}{g(M)} \right) \right\} \\
&\times \exp \left\{ -\frac{1}{2g(M)^2} \left(\sum_{i=1}^{M} w_i \sum_{j=1}^{p} \beta_{yz,i,j} z_{i,t-j} \right)' \Sigma_{y,t}^{-1} \left(\sum_{i=1}^{M} w_i \sum_{j=1}^{p} \beta_{yz,i,j} z_{i,t-j} \right) \right\} \\
&\times \exp \left\{ \sum_{i=1}^{M} \frac{w_i}{2g(M)^2} \left(\sum_{j=1}^{p} \beta_{yz,i,j} z_{i,t-j} \right)' \Sigma_{y,t}^{-1} \left(\sum_{j=1}^{p} \beta_{yz,i,j} z_{i,t-j} \right) \right\} \\
&\times \prod_{i=1}^{M} \exp \left\{ -\frac{1}{2} \left[h_{N^*+i,t} - \ln w_i \right.\right. \\
&\qquad \left.\left. + \ e^{-h_{N^*+i,t} + \ln w_i} \left(z_{i,t} - \alpha'_{z,i,t} y_t^* - c_{z,i} - \sum_{j=1}^{p} \beta'_{zy,i,j} y_{t-j}^* - \sum_{j=1}^{p} \beta_{zz,i,j} z_{i,t-j} \right)^2 \right] \right\},
\end{aligned}
$$

where we used the fact that $(y_t^*)'\Sigma_{y,t}^{-1}(y_t^*) = \sum_{i=1}^{M} w_i (y_t^*)'\Sigma_{y,t}^{-1}(y_t^*)$. The likelihood of the restricted VAR-SV is

$$
L(y;\theta) = \prod_{t=1}^{T} p(y_t \mid \cdot\ ). \tag{14}
$$

Now, suppose that our composite likelihood is constructed from sub-models:

$$
A_{y,t}y_t = c_y + \sum_{j=1}^{p} B_{yy,j} y_{t-j}^* + \sum_{j=1}^{p} \frac{\beta_{yz,i,j} z_{i,t-j}}{g(M)} + \epsilon_{y,t}, \qquad \epsilon_{y,t} \sim N(0, \Sigma_{y,t}), \tag{15}
$$

$$
z_{i,t} - \alpha'_{z,i,t} y_t^* = c_{z,i} + \sum_{j=1}^{p} \beta'_{zy,j} y_{t-j}^* + \sum_{j=1}^{p} \beta'_{zz,i,j} z_{i,t-j} + \epsilon_{z,i,t}, \quad \epsilon_{z,i,t} \sim N\left(0, e^{h_{N_*+i,t}}\right), \tag{16}
$$

which leads to

$$p^C(y_t \mid \ \cdot\ ) \propto \exp\left\{\sum_{i=1}^{M} -\frac{w_i}{2}\left(\tilde{\mathbf{y}}_t^* - \sum_{j=1}^{p}\frac{\beta_{yz,i,j}z_{i,t-j}}{g(M)}\right)' \Sigma_{y,t}^{-1}\left(\tilde{\mathbf{y}}_t^* - \sum_{j=1}^{p}\frac{\beta_{yz,i,j}z_{i,t-j}}{g(M)}\right)\right\}$$
$$\times \prod_{i=1}^{M}\exp\left\{-\frac{1}{2}\left[w_i h_{N^*+i,t} + e^{-h_{N^*+i,t}+\ln w_i}\left(z_{i,t} - \alpha_{z,i,t}' y_t^* - c_{z,i} - \sum_{j=1}^{p}\beta_{zy,i,j}' y_{t-j}^* - \sum_{j=1}^{p}\beta_{zz,i,j}z_{i,t-j}\right)^2\right]\right\}$$

and the composite likelihood $L^C(y;\theta) = \prod_{t=1}^{T} p^C(y_t \mid \ \cdot\ )$.

Observe that

$$L^C(y;\theta) \propto L(y;\theta)$$
$$\times \exp\left\{-\frac{1}{2g(M)^2}\sum_{t=1}^{T}\left[\sum_{i=1}^{M} w_i\left(\sum_{j=1}^{p}\beta_{yz,i,j}z_{i,t-j}\right)'\Sigma_{y,t}^{-1}\left(\sum_{j=1}^{p}\beta_{yz,i,j}z_{i,t-j}\right)\right.\right.$$
$$\left.\left. - \left(\sum_{i=1}^{M} w_i\sum_{j=1}^{p}\beta_{yz,i,j}z_{i,t-j}\right)'\Sigma_{y,t}^{-1}\left(\sum_{i=1}^{M} w_i\sum_{j=1}^{p}\beta_{yz,i,j}z_{i,t-j}\right)\right]\right\}$$
$$\propto L(y;\theta)\exp\left\{-\frac{1}{2g(M)^2}\sum_{t=1}^{T}\tilde{z}_t'\Xi_t\tilde{z}_t\right\},$$

where $\tilde{z}_t = (z_{1,t-1},\ldots,z_{1,t-p},\ldots,z_{M,t-1},\ldots,z_{M,t-p})'$, $B_i = (\beta_{yz,i,1},\ldots,\beta_{yz,i,p})$, and Ξ_t is a $Mp \times Mp$ positive semi-definite matrix with the (i,k) block given by

$$\Xi_{ik,t} = \begin{cases} w_i(1-w_i)B_i'\Sigma_{y,t}^{-1}B_i & \text{if} \quad i = k, \\ -w_i w_k B_i'\Sigma_{y,t}^{-1}B_k & \text{if} \quad i \neq k. \end{cases}$$

Let $\tilde{z}_i = (z_{i,1},\ldots,z_{i,T-1})'$, $\tilde{z} = (z_1',\ldots,z_M')'$, $z_T = (z_{i,T},\ldots,z_{M,T})'$ and $y^* = ((y_1^*)',\ldots,(y_T^*)')'$. Then, we may write the likelihood $L(y;\theta)$ as the density $L(y;\theta) = p(y^*, z_T, \tilde{z} \mid \vartheta)$. Consequently,

$$\tilde{L}^C(y;\theta) = \frac{p(y^*, z_T, \tilde{z} \mid \theta)\exp\left\{-\frac{1}{2g(M)^2}\sum_{t=1}^{T}\tilde{z}_t'\Xi_t\tilde{z}_t\right\}}{\int_{\tilde{z}}\int_{\mathbf{y}^*,z_T} p(y^*, z_T, \tilde{z} \mid \theta)d(y^*, z_T)\exp\left\{-\frac{1}{2g(M)^2}\sum_{t=1}^{T}\tilde{z}_t'\Xi_t\tilde{z}_t\right\}d\tilde{z}}$$
$$= \frac{p(y^*, z_T, \tilde{z} \mid \theta)\exp\left\{-\frac{1}{2g(M)^2}\sum_{t=1}^{T}\tilde{z}_t'\Xi_t\tilde{z}_t\right\}}{\mathrm{E}_{\tilde{z}}\left(\exp\left\{-\frac{1}{2g(M)^2}\sum_{t=1}^{T}\tilde{z}_t'\Xi_t\tilde{z}_t\right\}\right)},$$

and

$$D_{\mathrm{KL}}(L\|\tilde{L}^C) = \ln \mathrm{E}_{\tilde{z}}\left(\exp\left\{-\frac{1}{2g(M)^2}\sum_{t=1}^{T}\tilde{z}_t'\Xi_t\tilde{z}_t\right\}\right) - \mathrm{E}_{\tilde{\mathbf{z}}}\left(-\frac{1}{2g(M)^2}\sum_{t=1}^{T}\tilde{z}_t'\Xi_t\tilde{z}_t\right).$$

To prove that $D_{\mathrm{KL}}(L\|\tilde{L}^C) \longrightarrow 0$ as $M \longrightarrow \infty$, note that Ξ_t can be represented by the Hadamard product $\tilde{\Xi}_t \odot (W \otimes \iota_p\iota_p')$, with the $M \times M$ matrix W defined by elements

$$W_{ik} = \begin{cases} w_i(1-w_i) & \text{if} \quad i = k \\ -w_i w_k & \text{if} \quad i \neq k \end{cases},$$

and $\iota_p = (1, \ldots, 1)'$ being the $p \times 1$ vector of ones. In particular, W is positive semi-definite and contains information regarding the weights, while $\tilde{\Xi}_{ik,t} = B_i'\Sigma_{y,t}^{-1}B_k$, for all i and k, depends only on the parameters.

Accordingly,

$$\frac{\tilde{z}_t'\Xi_t\tilde{z}_t}{g(M)^2} = \frac{\tilde{z}_t'\tilde{z}_t}{g(M)^2} \times \frac{\tilde{z}_t'\Xi_t\tilde{z}_t}{\tilde{z}_t'\tilde{z}_t} \leq \frac{\tilde{z}_t'\tilde{z}_t}{g(M)^2}\|\Xi_t\|,$$

where $\|\cdot\|$ denotes the spectral norm. Since $\tilde{\Xi}_t$ and $W \otimes \iota_p\iota_p'$ are positive semi-definite, Schur's inequality (Horn and Johnson, 1991, Theorem 5.5.1) implies $\|\Xi_t\| \leq p\|\tilde{\Xi}_t\|\|W\|$. Moreover, there exists a unit vector u (satisfying $u'u = 1$) such that

$$\|W\| = u'Wu = \sum_{i=1}^{M} w_i u_i^2 - \left(\sum_{i=1}^{M} u_i w_i\right)^2.$$

Since $\sum_{i=1}^{M} w_i u_i^2 \leq \max\{w_i\} \sum_{i=1}^{M} u_i^2 = \max\{w_i\}$ and $\left(\sum_{i=1}^{M} u_i w_i\right)^2 \geq 0$, we obtain $\|W\| \leq \max\{w_i\}$. Consequently, $\max\{w_i\} \longrightarrow 0$ implies $\|W\| \longrightarrow 0$ and $\|\Xi_t\| \longrightarrow 0$ follows from the fact that $\|\tilde{\Xi}_t\|$ is constant with respect to M.

It remains to show that $\frac{\tilde{z}_t'\tilde{z}_t}{g(M)^2} = \sum_{j=1}^{p} \frac{\sum_{i=1}^{M} z_{i,t-j}^2}{g(M)^2}$ does not diverge for fixed T and $M \longrightarrow \infty$. Since $z_{i,t-j}$ is normally distributed conditional on y^*, with conditional expectation $\mu_i(y^*) \equiv \mathrm{E}(z_{i,t-j} \,|\, y^*)$ and variance v_i^2, the quantity $\zeta_i = \frac{z_{i,t-j} - \mu_i(y^*)}{g(M)}$ is conditionally independently (though not identically) distributed, and has the following properties:

1. $\mathrm{E}(\zeta_i \,|\, y^*) = 0$,
2. $\mathrm{E}(\zeta_i^2 \,|\, y^*) = \frac{v_i^2}{g(M)^2}$,
3. $\sum_{i=1}^{M} \mathrm{Var}(\zeta_i \,|\, y^*) = \bar{v}\frac{M}{g(M)^2} < \infty$, where $\bar{v} = \frac{1}{M}\sum_{i=1}^{M} v_i^2$,
4. $\sum_{i=1}^{M} \mathrm{Var}(\zeta_i^2 \,|\, y^*) \leq 3\tilde{v}\frac{M}{g(M)^4} < \infty$, where $\tilde{v} = \frac{1}{M}\sum_{i=1}^{M} v_i^4$.

Hence $\sum_{i=1}^{M} \zeta_i$ and $\sum_{i=1}^{M} \zeta_i^2 - \bar{v}\frac{M}{g(M)^2}$ both converge in $\mathbb{R}$ almost surely (Durrett, 2010, Theorem 2.5.3), which implies $\frac{\sum_{i=1}^{M} z_{i,t-j}^2}{g(M)^2}$ converges in $\mathbb{R}$ almost surely. In this case, the product $\frac{\tilde{z}'\tilde{z}_t}{g(M)^2}\|\Xi_t\| \longrightarrow 0$ and $D_{\mathrm{KL}}(L\|\tilde{L}^C)$ vanishes in the limit. ■