# Improving Bayesian VAR density forecasts through autoregressive Wishart Stochastic Volatility*

Paul Karapanagiotidis†

*4th draft*
*March 10th, 2012*

**Abstract**

Dramatic changes in macroeconomic time series volatility pose a challenge to contemporary vector autoregressive (VAR) forecasting models. Traditionally, the conditional volatility of such models had been assumed constant over time or allowed for breaks across long time periods. More recent work, however, has improved forecasts by allowing the conditional volatility to be completely time variant by specifying the VAR innovation variance as a distinct discrete time process. For example, Clark (2011) specifies the volatility process as an independent log random walk for each time series in the VAR. Unfortunately, there is no empirical reason to believe that the VAR innovation volatility process of macroeconomic growth series follow log random walks, nor that the volatility of each series is independent of the others. This suggests that a more robust specification on the volatility process—one that both accounts for co-persistence in conditional volatility across time series and exhibits *mean reverting* behaviour—should improve density forecasts, especially over the long run forecasting horizon. In this respect, I employ a latent Inverse-Wishart autoregressive stochastic volatility specification on the conditional variance equation of a Bayesian VAR, with U.S. macroeconomic time series data, in evaluating Bayesian forecast efficiency against a competing log random walk specification by Clark (2011).

Keywords: Inverse Wishart distribution, stochastic volatility, predictive likelihoods, MCMC, macroeconomic time series, density forecasts, vector autoregression, steady state priors, Bayesian econometrics
JEL: C11, C32, C53, E17

# 1 Introduction

Forecasts of macroeconomic time series have become a ubiquitous component of any policymaker's toolkit. As such, central banks like the Federal Reserve typically publish density

*I'd like to thank John Maheu, Christian Gourieroux, Angelo Melino, Martin Burda, and Carolyn Pitchik for their helpful comments and suggestions.

†University of Toronto, Department of Economics, p.karapanagiotidis@utoronto.ca

forecasts for inflation, output, interest rates, or other major indicators. This information is important because it helps both industry and consumers make decisions consistent with economic fundamentals. However, forecasts themselves are not infallible. In fact, while major advances have been made in the area of statistical forecasting, there remains much room for improvement.

This paper resolves some of the relevant issues by proposing a key change in the volatility process of VARs (vector autoregressions) popular among macroeconomists. Instead of assuming cross-series independent log random walk processes on the volatilities of the VAR shocks, I employ an Inverse-Wishart process where the scale matrix is an autoregressive process of past covariance matrices (see the model section (3) for more details and a discussion of the pros and cons of this change). Furthermore, as data I employ four major U.S. macroeconomic time-series: GDP growth, the inflation rate, the interest rate, and unemployment rate.[1] A Bayesian approach is then taken in both estimation (MCMC estimation steps are provided) and in comparing forecasts between the benchmark model (Clark (2011)) and my competing Inverse-Wishart autoregressive volatility modification. Results suggest that incorporating the more sophisticated Inverse-Wishart autoregressive volatility process improves density forecasts in both the short and long run, with larger improvements seen as the horizon increases, despite a small sample size and increased parameterization of the model. With this in mind, the following discussion aims to provide a broader context surrounding the relevant forecasting issues precipitated this proposed modification to the typical VAR process volatility specification.

## 1.1 Background

A fundamental issue facing the production of good forecasts has been that of how to deal with the changing moments of the conditional forecast distributions. For example, dramatic changes in U.S. economic volatility have posed a modeling challenge to contemporary forecasters, specifically amongst macroeconomists where Gaussian vector autoregressive (VAR) models are popular. An analysis of major U.S. economic indicators, such as output growth over the past 100 years, illustrates that the economy goes through periods of changing volatility. For example, "The Great Moderation," which began in the 1980's, represented a period of unusually low volatility vis a vis both the lengthy prior period of erratic volatility and more recent instability we've experienced since 2007. In this respect, both Sims (2001) and Stock (2001) in separate discussions of Cogley and Sargent's (2001) paper, criticized their assumption of homoskedastic VAR variances, pointing to evidence as described above and analyzed by Bernanke and Mihov (1998a,1998b) (in the case of monetary policy shocks between 1979 and 1982), Kim and Nelson (1999), or McConnell and Perez Quiros (2000) (with respect to the growing stability of output around 1985). Finally, Clark (2011) also finds significant changes in conditional volatility across time when estimating the latent stochastic volatilities of his model.

It should not come as a surprise then that while, traditionally, the volatility of forecasting models was assumed constant over time—primarily for the sake of simplicity—it can be shown that this assumption leads to poor conditional forecasts. For example, Jore, Mitchell, and Vahey (2010) employ a model averaging approach to U.S. data, with both equal weights and recursively adapted weights based on log predictive density scores across a range of different specifications.

[1] Note that the data is taken from the RTDSM database – the same dataset, in fact, as Clark (2011), my benchmark comparison model (with the exception of the interest rate—see the data section below). Moreover, all data is at the aggregate U.S. level. Finally the interest rate employed in my paper is the 3-month Federal Treasury Bill rate.

Their results show strong support for a recursive weighting scheme across specifications. More interestingly, however, they find that during periods of changing macroeconomic volatility, for example when the US economy transitioned into "The Great Moderation", the weighting scheme tends to place more weight on specifications which dynamically account for structural breaks in volatility. Moreover, they find evidence of poor forecasting given a simple assumption of fixed volatility or equal weights across model specifications. However, it worth noting that the specifications which do respond to structural change within Jore et. al's (2010) framework are limited in that they are restricted to a finite set of possible volatility states and breaks.

Consequently, it is important to account for changing volatility in any forecasting specification. Furthermore, if such changes in volatility occurred relatively infrequently and could be extracted from the data with reasonable statistical significance, then employing a regime switching specification such as in Jore et. al. (2010) might prove sufficient in drawing good forecasts. However, the truth is that, given the complexity of the economy, changes in volatility probably occur much more frequently and take on many more values than can be effectively captured by a finite state model. For this reason forecasters have adopted a continuous state-space framework for estimating the conditional volatility of VAR models as opposed to the finite state regime switching type model applied to volatility, as popularized by Hamilton (1989), and employed by Jore et. al (2010). Moreover, the use of the so called continuous state "stochastic volatility" model has also grown in popularity given its usefulness in modeling a latent volatility process based on a filtration that includes more than just lagged VAR series shocks, as for example in the case of a GARCH model.[2]

Both Cogley & Sargent (2005) and Primiceri (2005) allow for time variation in the conditional covariance matrix across VAR series shocks according to a stochastic volatility law of motion, where the conditional volatility can take on any value in a continuous positive real set (and covariances can be any real number). Moreover, they also allow for time variation in the VAR parameters themselves, through another stochastic volatility law of motion on their state across time. Clark (2011), which will represent my benchmark model, also follows the same structure of the previous two studies, albeit without the time varying VAR parameters which are dropped in favour of tight Bayesian steady state priors on the deterministic trend parameters (which define the unconditional mean of the VAR process) and a rolling sample window which re-estimates the parameters across time. Villani (2009) showed that imposing Bayesian steady state prior distributions allow us to incorporate prior beliefs about macroeconomic variable steady states into our model. Furthermore, I believe that employing this information probably reduces the need for time varying VAR parameters since much of the time variation in the autoregressive parameters (which is not due to a lack of time variation in the shock covariance, as was the case with Cogley and Sargent (2001)), may in fact be due to a lack of a well defined deterministic trend (see Cogley and Sargent (2005) where they model their VAR intercepts[3] as stochastic random walks). Moreover, it should be noted that given the quarterly nature of most macroeconomic time series, small sample sizes are usually the norm. In this situation a tight prior also plays the role of constraining VAR parameters to aid in the identification of deterministic trends that might not otherwise be readily apparent.[4] In this respect, Villani (2009) also demonstrates that informative

[2]See also Sartore and Billio (2005) for a good, general, survey of Stochastic Volatility.

[3]Noting of course that given their formulation, the VAR long-run mean $\mu_t$ is both time varying, stochastic, and a function of the VAR intercepts, $\alpha_t$, as $\mu_t = (\mathrm{I} - \Phi_t)^{-1}\alpha_t$.

[4]In Clark (2011) for example, his rolling sample window is only of size $T = 80$.

steady state priors can greatly improve point forecasts, especially over the longer term horizon where correct specification of the unconditional mean of the series is important—see Clements and Hendry (1998). All of this of course assumes that our prior beliefs on the deterministic nature of the time series trends are correct. In fact, whether or not the trends in macroeconomic data are better modeled as stochastic (i.e. unit roots with drift) or deterministic is still an open question of debate.

However, most of these studies adopt certain features which could still be improved upon. For one, they all employ cross-series independent specifications of the individual volatility processes of each time-series. Therefore, the latent time-varying covariance matrices are constrained to be diagonal. This is done in order to reduce the parameterization of the model but I believe it severely limits the richness of the volatility dynamics since in order to generate conditional covariance matrices, some form of deterministic relationship must be drawn upon the diagonal covariance matrix, usually by pre-multiplying it by a lower triangular matrix. Therefore, the covariance between series is driven directly by some linear function of the variances and this relationship between variances and covariances does not depend on time.

Furthermore, all of the aforementioned studies employ a random walk formulation for the conditional volatility. The reason for this choice of specification is not entirely clear since an analysis of U.S. quarterly output growth suggests a great deal of autoregressive volatility persistence, as do other indicators such as inflation or the unemployment rate. Moreover, without explicitly parameterizing time $'t'$ covariances across series it is left unknown whether these series exhibit volatility spillovers across time as has been shown to be the case with financial time series from U.S. markets (see for example, Diebold and Yilmaz, (2007)). Furthermore, studies such as Cogley and Sargent (2005) seem to provide little explicit justification for this choice other than a brief comment that "the random walk specification is designed for permanent shifts in the innovation variance, such as those emphasized in the literature on the growing stability of the U.S. economy." Ultimately, if a random walk is mispecified and there does in fact exist co-persistence in the conditional volatility across macroeconomic time series, then a Wishart autoregressive process should capture this and allow us to extend our forecasts more accurately across longer horizons where the random walk specification would otherwise suggest an explosive volatility process.

Given this, the multivariate volatility process should be constructed to directly model the time varying covariance matrices without simply extending the univariate specification ad hoc to the multivariate case. Moroever, any autoregressive persistence in volatility should be captured and a finite unconditional mean should be specified. As described in Chib, Omori, and Asai (2009), one natural extension to the multivariate stochastic volatility framework involves modeling the time varying covariance matrices directly as a Wishart autoregressive process, since the scatter matrix[5] itself is a natural estimator of the covariance of multivariate Normal random variables.

[5]Given $n$ samples of $m$-dimensional data, represented as the $m \times n$ matrix $\mathbf{X} = [x_1, x_2, \ldots, x_n]$ the scatter matrix is the $nxn$ positive semi-definite matrix defined as:

$$\mathbf{S} = \sum_{j=1}^{n} (x_j - x_\mu)(x_j - x_\mu)^T$$

Or equivalently as:

$$\mathbf{S} = \mathbf{X}\mathbf{C_n}\mathbf{X}^T$$

where $\mathbf{C_n}$ is the $n \times n$ centering matrix and $x_\mu$ is the sample mean across the $n$ samples of dimension $m$.

In fact, Philipov and Glickman (2006) apply such an autoregressive Wishart process to analyze the conditional volatility of financial data and find that it improves volatility forecasts over simpler formulations, where a number of Bayesian and Frequentist measures are applied to compare forecast accuracy given a variety of competing specifications. It is worth noting, however, that there exist problems with the Philipov and Glickman (2006) implementation of the Wishart autoregressive volatility process as it stands—see Rinnergschwentner et. al (2011) for more details and quite a few corrections. Therefore, in light of Phillipov and Glickman's idea of direct Wishart modeling, I form my Inverse Wishart autoregressive volatility process by accounting for some of the weaknesses in their specification—see the model estimation section below for more details.

The rest of the paper is organized as follows. Section 2 discusses the data and deterministic trend methods employed in generating forecasts. Section 3 discusses both the benchmark model based on Clark (2011) and the proposed Inverse Wishart process modification based on Philipov and Glickman (2006). Moreover, this section discusses the Gibbs sampler Markov Chain Monte Carlo (MCMC) estimation technique employed to estimate both models. Section 4 details the steady state and other conjugate priors I employ within the Bayesian framework to sample from conditionally conjugate posterior densities of the model parameters. Section 5 discusses the method whereby I generate forecast densities for both the VAR data and covariance matrices across various horizons. Section 6 details the results of both the estimation process and the forecast comparisons based on Bayesian analysis of the predictive likelihoods. Finally, Section 7 summarizes and concludes.

# 2 Data

The VAR model will involve at most four macroeconomic time series generated from aggregate U.S. data: 1) output growth (i.e. real GDP) 2) the inflation rate 3) the unemployment rate and 4) some form of interest rate.

The data source considered is the same as in Clark (2011): the so-called "real-time" [6] data from the Federal Reserve Bank of Philadelphia's Real-Time Data Set for Macroeconomics (or "RTDSM"). The total sample size is quite small: only $T = 252$ data points extending from the $2^{nd}$ quarter of 1948 (hereon denoted as 1948:Q2) until the $1^{st}$ quarter of 2011. Output from the RTDSM database is quarterly real data and measured as either GDP or GNP depending on the data vintage.[7] Inflation from the RTDSM is also measured quarterly and as either a GDP or GNP deflator or a price index, depending on the vintage. I measure growth and inflation rates as annualized log changes.[8] The unemployment rate, however, is available on a monthly basis so I simply average across each quarter in matching the quarterly nature of output and inflation. Moreover, it should be noted that the unemployment rate tends to differ much less dramatically across vintages. Finally, while Clark (2011) employs the federal funds interest rate series, Primiceri (2005) recommends the nominal annualized yield on 3-month Federal Treasury Bills, since this series goes back much further. I therefore adopt the latter series, and again, average across

[6] That is data that is regenerated annually to conform to new changes in the way we measure macroeconomic indicators, or to take into account flaws in some previous set, observed ex-post. Each new issue is deemed a "vintage."

[7] The RTDSM generates entirely new time series each quarter (deemed "vintages") based on updated chain weighting techniques or other improvements. Thus newer vintages represent larger samples than older ones which were generated at previous dates.

[8] Since log differences are already continuously compounded, I simply multiply each quarterly value by 4.

quarters since the data is monthly.[9] Finally, output, inflation, and the unemployment rate are already seasonally adjusted by their source providers.

Clark and McCracken (2008,2010) also provide evidence that point forecasts of GDP growth, inflation, and interest rates are improved by specifying the latter two series as deviations from some form of deterministic trend simulating inflation expectations. Given this result Clark (2011) adopts the Blue Chip Consensus forecast produced from survey data and published by Aspen Publishers Ltd., as a form of long-term inflations expectations. Unfortunately, as Clark mentions in his online appendix, the data for this Blue Chip forecast of inflation expectations only extends back to the fourth quarter of 1979 (i.e. 1979:Q4). Therefore, Clark appends an exponentially smoothed trend from his inflation series to be beginning of the Blue Chip series in extending it back to 1964. Clark mentions that despite his attempts at keeping the data "as real time as possible" by employing every quarterly vintage of inflation data, in the end a trend based on his most recent vintage (2008:Q4) deviates little from the others. Moreover, as Clark notes, Kozicki and Tinsley (2001a,2001b) and Clark and McCracken (2008), both suggest that exponentially smoothed trends of the inflation rate match up reasonably well with survey-based measures of long-run expectations in the data since the early 1980's. Given both of these facts, I will simply employ an exponentially smoothed trend of the inflation rate through the most recent vintage currently available (2011:Q4) in generating a long-term inflations expectations series, skipping the Blue Chip survey data entirely and ignoring the previous vintages of inflation data.[10]

Citing studies such as Romer and Romer (2000), Sims (2002), and Croushore (2006), Clark (2011) also suggests that since the nature of how we measure macroeconomic aggregates has changed over time, GDP data available today, for example, is quite different than output as defined and measured in 1970. Unfortunately, this reality of data generation presents us with a difficult choice as to which vintage to use for forecasting comparisons. In this respect, Clark follows Romer and Romer (2000) in that when making out of sample forecast comparisons (from older vintages of smaller sample size), the forecast error should be based off of a newer vintage of data than that which generated the forecast. However, for simplicity, in my rolling sample forecast comparisons, I will simply employ the newest vintage of data since the focus of this paper is on a *relative* comparison across model specifications and not on absolute forecast efficiency (which could depend on data vintages employed). Moreover, I believe that given the small sample sizes available for the macroeconomic data in question, a trade off needs to be made between estimation sample size and out of sample forecast comparison. Therefore, the larger is the total sample size, the more flexibility is permitted in choosing our rolling window estimation size (to account for parameter drift) and out of sample comparisons given longer forecast horizons.

Finally, it should be mentioned that the unemployment rate series is also de-trended by an exponential smoother (in the same way the inflation rate was de-trended in order to generate the long-run inflation expectations (see footnote 10)). Therefore, to summarize, GDP growth is not de-trended (although I center it around a long run constant mean of $3.0\%$ through the

[9]The 3-month Federal Treasury bill rate series employed is a combination of two very similar series joined together at June 2000, since the first vintage was discontinued. "H15/discontinued/H1.RIFSGFPIM03_N.M" is the unique ID for the discontinued series and "H15/H15/RIFLGFCM03_N.M" is the newer series. Both series are available at the Federal Reserve website: http://www.federalreserve.gov/releases/h15/data.htm.

[10]The exponential smoother employed is as follows: $y_t^* = y_{t-1}^* + \alpha(y_t - y_{t-1}^*)$, where $y_t$ is the actual data series and $y_t^*$ is the exponentially smoothed trend. $\alpha$ is a parameter which can be adjusted depending on how "tight" we want the trend to follow the data series. For the inflation rate trend used as long-term inflation expectations, Clark suggests a value of $\alpha = 0.05$.

prior on $\mathbf{\Psi}$), unemployment is de-trended around its exponentially smoothed values lagged one period (with a smoothing parameter of $\alpha = 0.02$), the inflation rate is de-trended around its exponentially smoothed trend (with a smoothing parameter of $\alpha = 0.05$), and the interest rate (3-month Treasury bill) is de-trended around the same trend as inflation (which is supposed to simulate long-term inflation expectations), although I force a long run constant mean of $2.5\%$ above trend through the prior on $\mathbf{\Psi}$. See the model and estimation section for more details as to how these trends are implemented into the model through the $\mathbf{\Psi}$ and $\mathbf{d_t}$ terms.

# 3 Model specifications

The benchmark model is the BVAR-SSP-SV specification referred to in Clark (2011). This model employs a Bayesian $VAR(J)$ formulation for the multivariate mean equation and separate univariate log random walks with Normal shocks as the volatility processes for each macroeconomic series. Since the volatility shocks are uncorrelated, all of the covariance structure across macroeconomic series are driven by a deterministic lower-triangular matrix, $\mathbf{B}$. See the "benchmark model" below for the mean and volatility equations.

Moreover, the model employs informative "Steady-State priors" on the parameters of the steady state variables $\mathbf{d_t}$ as in Villani (2009). Note that the $VAR(J)$ is specified in what Villani (2009) refers to as "Steady-State" form—that is the $p \times 1$ macroeconomic time series vector $\mathbf{y_t}$ is differenced from its time-varying unconditional mean $\mu_\mathbf{t} = \mathbf{\Psi d_t}$, where $\mathbf{\Psi}$ is a $p \times q$ coefficient matrix and $\mathbf{d_t}$ is a $q \times 1$ vector of deterministic trends.[11] This will be important when we consider the role Bayesian steady-state priors play in identifying the deterministic trend coefficients of the VAR. See the section on priors for more details.

## 3.1 Benchmark model

In summary, I refer to this benchmark model as the $Clark$ specification:

$$\mathbf{\Pi}(L)(\mathbf{y_t} - \mathbf{\Psi d_t}) = \nu_\mathbf{t}$$

$$\nu_\mathbf{t} = \mathbf{B}^{-1}\mathbf{\Lambda}_t^{0.5}\varepsilon_\mathbf{t} \quad s.t. \quad \varepsilon_\mathbf{t} \sim MVN_p(\mathbf{0}, \mathbf{I_p})$$

$$\mathbf{\Lambda_t} = diag(\lambda_{1,t}, \lambda_{2,t}, \ldots, \lambda_{p,t})$$

$$\ln(\lambda_{i,t}) = \ln(\lambda_{i,t-1}) + \xi_{i,t} \ \forall i = 1, \ldots, p$$

$$\xi_{i,t} \sim i.i.d.N(0, \varphi_i)$$

$$and \quad \mathbf{\Pi}(L) = \mathbf{I_p} - \sum_{j=1}^{\mathbf{J}} \mathbf{\Pi_j} L^j \ \ where\ L\ is\ the\ lag\ operator.$$

[11] Note that the choice of values for $\mathbf{d_t}$ can have dramatic consequences. For example, if $\mathbf{d_t} = 1, \forall t$, i.e. takes on a single constant value for all time periods, then $\mathbf{\Psi}$ is a vector of regression constants, the values of which determine the time invariant long-run means of the autoregressive processes. However, if for example, $\mathbf{d_t} = t$, then the values in the vector $\mathbf{\Psi}$ measure individual responses to a linear time-trend relationship shared by $\mathbf{y_t}$. Moreover, if $\mathbf{d_t} = [t, f(t)]$, where $f(t)$ is perhaps some non-linear function of $t$, then $\mathbf{\Psi}$ becomes a matrix, the elements of which reflect how the time-varying long-run means of each process are expressed as a linear combination of both the linear-trend and the non-linear one simultaneously. Given this model flexibility, we can incorporate an exponential smoother as one possible method to de-trend the relevant macroeconomic series in a non-linear fashion. See the data section for more details.

## 3.2 Alternative volatility process specification

As an alternative to the $Clark$ log random walk volatility specification, I propose augmenting the volatility process to incorporate any persistence in (co)variation across time through the use of a multivariate Inverse Wishart specification (here on referred to as the $InvWishart(K)$). This formulation would eliminate the need for a $\mathbf{B}$ matrix since the covariance matrices across time are estimated directly in the hopes of capturing richer dynamics than the simple random walk can generate. While the random walk presumes that the series covariances are both constant across time and generated through the $\mathbf{B}$ matrix, the $InvWishart(K)$ estimates the latent covariance matrices directly for each time '$t$'. Moreover, the Inverse Wishart process is assumed to be autoregressive and includes a constant matrix $\mathbf{CC}^T$ which defines a finite long-run mean under suitable stationarity conditions (discussed below).

In summary, the $InvWishart(K)$ as follows represents a modification of the autoregressive Wishart specification employed in Philipov and Glickman (2006):

$$\nu_{\mathbf{t}}|\mathbf{\Sigma_t} \sim MVN_p(\mathbf{0}, \mathbf{\Sigma_t})$$

$$\mathbf{\Sigma_t}|\mathbf{\Sigma_{t-1}}, \ldots, \mathbf{\Sigma_{t-K}} \sim IW_p(v, \mathbf{S_{t-1}})$$

$$\mathbf{S_{t-1}} = \left( \mathbf{CC}^T + \sum_{k=1}^{K} \mathbf{A}_k \mathbf{\Sigma_{t-k}} \mathbf{A}_k^T \right) (v - p - 1)$$

$$E\left[\mathbf{\Sigma_t} \mid \mathbf{\Sigma_{t-1}}, \ldots, \mathbf{\Sigma_{t-K}}\right] = \frac{\mathbf{S_{t-1}}}{v - p - 1} = \mathbf{CC}^T + \sum_{k=1}^{K} \mathbf{A}_k \mathbf{\Sigma_{t-k}} \mathbf{A}_k^T$$

$$Var\left[\sigma_{ij,t} \mid \mathbf{\Sigma_{t-1}}, \ldots, \mathbf{\Sigma_{t-K}}\right] = \frac{(v - p + 1)\, s_{ij,t-1}^2 + (v - p - 1)\, s_{ii,t-1} s_{ij,t-1}}{(v - p)(v - p - 1)^2 (v - p - 3)}$$

where $\sigma_{ij}$ is the $ij^{th}$ element of $\mathbf{\Sigma_t}$ and $s_{ij,t-1}$ is the $ij^{th}$ element of $\mathbf{S_{t-1}}$.[12]

Note that in this case $\mathbf{C}$ is lower triangular and the $\mathbf{A}_i$'s are not necessarily symmetric or positive-definite. However, given that $\mathbf{CC}^T$ is positive-definite, then the scale matrix $\mathbf{S}_t$ is symmetric and almost surely positive-definite. Moreover, the model is identified if the main diagonal of $\mathbf{C}$ and the top-left element of each $\mathbf{A}_i$ is strictly positive (see Engle and Kroner (1995)).

Note also that both volatility processes (i.e. either the $Clark$ or $InvWishart(K)$) are coupled to the $VAR(J)$ process for the conditional mean of the macroeconomic time series, and together constitute a state-space representation with the $VAR(J)$ as observation equation and the $InvWishart(K)$ or $Clark$ as the state equation, with the covariance as the "state" of the model.

In summary the improvements of such a change to the volatility process are as follows:

1. The direct estimation of the latent stochastic volatility covariance process precludes the need to estimate a $\mathbf{B}$ matrix and should capture more richly any existing co-persistence between macroeconomic series volatility across time.

2. The inclusion of autoregressive parameter matrices $\mathbf{A}_i$ avoids the curious assumption of a unit-root in the volatility processes of each series.

[12] See S.J. Press (1982).

3. The inclusion of the constant matrix $\mathbf{C}\mathbf{C}^T$ allows for a finite unconditional mean of the multivariate volatility process (the unconditional mean will be some symmetric positive-definite matrix).

4. Both (2) & (3) above should prove critical in resolving a major issue with the log random walk volatility specification—*that of poor long-term density forecasts*—since no longer will the forecasts "blow up" as the horizon increases, but rather should revert to a finite long-run mean given suitable stationarity conditions being satisfied.

However, there are some disadvantages as well:

1. The direct estimation of the latent stochastic volatility covariance process increases the number of latent parameters from $Tp$ to $Tp(p+1)/2$, raising the curse of dimensionality as an issue quickly as $p$ increases. Moreover, the number of regular parameters goes from $\frac{p(p-1)}{2} + p$ to $\frac{p(p+1)}{2} + Kp^2 + 1$, although in this latter case many possible reparameterizations are possible to reduce the number of regular parameters the $InvWishart(K)$ must estimate.

2. Since conditionally conjugate priors are unknown at this point for the conditional posterior densities of the $InvWishart(K)$ regular parameters, a Metropolis-Hastings random walk sampler is employed. This additional Metropolis-Within-Gibbs step requires some extra work to obtain reasonable draws.

Finally, both specifications suffer from the fact that they estimate the latent stochastic volatility processes sequentially rather than jointly. Clearly, if the volatility $\mathbf{\Sigma_t}$ is relatively correlated across time, joint sampling would reduce autocorrelation in the Markov process, since a conditional sampler will not propose draws sufficiently far from the previous value to be independent. In other words the sampler will fail to traverse the full parameter support in a reasonable amount of steps due to low variance of the conditional posterior. See Greenburg (2008) pg. 94 for a simple example illustrating the problem.

As in Philipov and Glickman (2006) the volatility autoregression of the $InvWishart(K)$ is defined directly on the scale matrix, $\mathbf{S}_t$, of the Inverse Wishart process. It should be noted, however, that this is not the only way to go about modeling the volatility in general. In fact, the literature suggests a number of different approaches.

One alternative is to place the autoregression on the underlying centrality parameter of the Wishart process such as in Gourieroux et. al. (2009). Recall that if $X_i = [x_{i,1}, x_{i,2}, \ldots, x_{i,p}]$ is an independent draw from a $p$-variate multivariate Normal distribution $MVN_p(0, \mathbf{\Sigma})$, then the matrix $\mathbf{A} = \mathbf{X}^T\mathbf{X} = \sum_{i=1}^{n} X_i^T X_i$, with $n$ rows made up of $X_i$'s, has the Wishart distribution denoted $\mathbf{A} \sim W_p(n, \mathbf{\Sigma})$. However, in this case the centrality parameter is implicitly zero and we have that $X_i = 0 + \mathbf{\Sigma}^{1/2}\varepsilon$ where $\varepsilon \sim MVN_p(0, \mathbf{I}_p)$. Gourieroux et. al. (2009) propose rather that if $X_i$ depends on time so that we have $X_{i,t} = \mathbf{M}X_{i,t-1} + \varepsilon_t$, then $\mathbf{A}_t = \mathbf{X}_t^T\mathbf{X}_t \sim W_p(n, \mathbf{M}, \mathbf{\Sigma})$ where $\mathbf{M}$ is the $p \times p$ centrality matrix. Therefore, in my model the emphasis is on updating the scale parameter through past covariance matrix estimates $\mathbf{\Sigma_t}$, since I assume that the time $'t'$ $VAR(j)$ shocks have zero mean and are uncorrelated (i.e. $\mathbf{M}$=0) but that their variance changes over time (i.e. $\nu_{i,t} \sim MVN_p(0, \mathbf{\Sigma_t})$ and $E[\mathbf{\Sigma_t} \mid v, \mathbf{S_{t-1}}] = \mathbf{C}\mathbf{C}^T + \sum_{k=1}^{K} \mathbf{A}_k \mathbf{\Sigma_{t-k}} \mathbf{A}_k^T$).

Fox and West (2011) also propose a novel class of stationary covariance matrix processes which exploit properties of Inverse Wishart partitioned matrix theory. Specifically, by augmenting the parameter state-space they show that we can easily obtain representations for the terms in a factorization of the joint density of covariance matrices across time, $f(\mathbf{\Sigma}_T, \ldots, \mathbf{\Sigma}_0) = \prod_{t=1}^{T} f(\mathbf{\Sigma}_t, \mathbf{\Sigma}_{t-1}) \prod_{t=2}^{T} f(\mathbf{\Sigma}_{t-1})$. Note that this expression defines a stationary first-order Markov process on the covariance matrices across time, with the marginal distribution given as $\mathbf{\Sigma}_t \sim IW_q(v+2, v\mathbf{S})$, since given the following augmented matrix:

$$\begin{pmatrix} \mathbf{\Sigma}_t & \phi_t^T \\ \phi_t & \mathbf{\Sigma}_{t-1} \end{pmatrix} \sim IW_{2q}(v+2, v \begin{pmatrix} \mathbf{S} & \mathbf{S}\mathbf{F}^T \\ \mathbf{F}\mathbf{S} & \mathbf{S} \end{pmatrix})$$

(such that $\phi_t = \mathbf{\Upsilon}_t \mathbf{\Sigma}_{t-1}$), we have that by Inverse Wishart partitioned matrix theory, the covariance process can be written as an AR(1): $\mathbf{\Sigma}_t = \mathbf{\Psi}_t + \mathbf{\Upsilon}_t \mathbf{\Sigma}_{t-1} \mathbf{\Upsilon}_t^T$, with $\mathbf{\Upsilon}_t$, representing a random coefficient matrix and $\mathbf{\Psi}_t$ representing an innovation (note both $\mathbf{\Upsilon}_t$ and $\mathbf{\Psi}_t$ are latent variables). Note that under this framework the conditional density $\mathbf{\Sigma}_t | \mathbf{\Sigma}_{t-1}$ is not of an analytical form but can nonetheless be explored theoretically. See Fox and West (2011) for more details.

At this point it is worth investigating the relevant stationarity conditions of the $Inv\text{-}Wishart(K)$ volatility process. As the $InvWishart(K)$ is just a special case of the $CAW(p,q)$ process of Golosnoy et. al. (2010) I adopt their **Proposition 1** directly:

"*The unconditional mean of the $InvWishart(K)$ process is finite iff all the eigenvalues of the matrix $\mathbf{\Upsilon} = \sum_{i=1}^{\mathrm{K}} \mathbf{\Xi_i}$ are less than 1 in modulus. In this case the unconditional mean is given by:*

$$E[\sigma_t] = (\mathbf{I}_g - \mathbf{\Upsilon})^{-1} c$$

where $g = \frac{p(p+1)}{2}$, $c = vech\left(\mathbf{C}\mathbf{C}^T\right)$, $\sigma_t = vech\left(\mathbf{\Sigma_t}\right)$, and $\mathbf{\Xi_i} = \mathbf{L}\left(\mathbf{A_i} \otimes \mathbf{A_i}\right)\mathbf{D}$.

Note that in this case $\mathbf{L}$ and $\mathbf{D}$ are the elimination and duplication matrices respectively so that $vec(\mathbf{X}) = \mathbf{D}vech(\mathbf{X})$ and $vech(\mathbf{X}) = \mathbf{L}vec(\mathbf{X})$. See Magnus and Neudecker (1980) for more details.

## 3.3 Model Estimation:

Both model specifications are estimated using Markov Chain Monte Carlo techniques. The benchmark specification, i.e. $Clark$, which employs a univariate random walk process for the stochastic volatility equations of each series is due to Clark (2011), given estimation steps I have derived myself based on those of Villani (2009) and Cogley and Sargent (2005).[13] The alternative multivariate Wishart specification for the stochastic volatility equation, i.e. $InvWishart(K)$, is built as a modified version of Philipov and Glickman (2006), and employs the same sequential sampling algorithm of the time-varying covariance matrices. Details for the estimation of each model specification are provided in the appendix as a series of Gibbs sampler steps (and

[13] Clark (2011) provides expressions for the conditional posterior distributions of the VAR parameters of his Gibbs sampler estimation process, obtained from Mattias Villani, who himself derived them based on the constant variance sampler employed in Villani (2009). However in my desire to be extremely clear, and precise, in explaining where all of my sampling steps come from, I have re-derived all of the relevant expressions for the posterior distributions employed in my Gibbs sampler and provide details in the appendix.

Metropolis-within-Gibbs steps where appropriate i.e. where the conditional posterior is not of a family from which we can easily draw samples).

The Gibbs sampler is a MCMC technique whereby we can indirectly sample from a distribution for which we have an expression for the p.d.f. but are unable to sample from it directly using traditional methods. In most cases, the distribution we wish to sample from is the joint posterior of the model parameters $p(\theta|y) \propto p(y \mid \theta)\,\pi(\theta)$ (by Bayes Theorem the left hand side is known as the "posterior", and the right hand side represents the likelihood times the "prior"). It turns out that a sequence of conditional draws from a Markov process[14] that satisfy certain ergodicity conditions will converge in unconditional distribution. Therefore, if we set the unconditional distribution of the Markov process draws to the target distribution we wish to sample from, we need only sample from the conditional distribution of some suitable Markov process to generate the desired unconditional.

Moreover, suppose we now partition the parameter set $\theta$ into $\theta_1$ and $\theta_2$. It can be shown that drawing the conditional Markov process in sequence from the conditional posteriors of ${\theta_1}'|\theta_1, \theta_2 \sim q_1(\theta_1'|\theta_1, \theta_2, y)$ and then ${\theta_2}'|\theta_1', \theta_2 \sim q_2(\theta_2'|\theta_1', \theta_2, y)$,[15] the repeated sequence of draws on $(\theta_1, \theta_2)$ will converge in unconditional distribution to the joint posterior $p(\theta_1, \theta_2|y)$, given a suitable "proposal" distribution $q(\cdot)$. This is known as the Gibbs sampler. See Greenburg (2008) for a much more detailed yet accessible treatment of Bayesian estimation and MCMC methods.

# 4 Priors

The Bayesian estimation framework employed requires of us to specify certain prior beliefs upon the parameter set before estimation and this is done through the specification of prior densities. In most cases the prior densities are chosen to be conditionally conjugate—that is, they are chosen of a known family such that the conditional posterior density works out to be of the same family as the prior. This facilitates estimation greatly since the need for a well fitting proposal density to the target (as in a Metropolis-Hastings or Accept-Reject algorithm) is obviated—in fact, in this case the proposal is always accepted since the conditional posterior *is* the target and we can draw directly from it. Moreover, this Gibbs sampling algorithm is designed such that if we draw from a series of conditionally conjugate posterior densities in sequence, each conditional on the last Markov process draw, this sequence will converge to the joint posterior density from which marginals are easily available as the individual processes themselves.

This section outlines and considers the specific forms and numerical values my prior densities will take. First, the prior for the VAR slope coefficients $\mathbf{\Pi}$, follow a modified *Minnesota specification* (see Litterman (1986)). In this case the prior mean assumes that the VAR follows an AR(1) process (i.e. means of additional lags beyond 1 are set to 0).[16] Since GDP growth displays less persistence in its levels, I set its autoregressive prior mean to $0.25$ and set the others to $0.8$. Cross

[14] Where the process is set to draw values from the parameter space $\theta \in \Theta$.

[15] Where the primed values denote the next draws in the Markov process, drawn from the proposed conditional distribution $q(\cdot)$. Note that $q(\cdot)$ is chosen by the analyst and may or may not be a good "fit" to the true conditional distribution of the Markov process. The better is the choice of the proposal distribution $q(\cdot)$, the better will be the convergence of the Markov process draws to the unconditional distribution we desire to sample from.

[16] However, in this case our prior beliefs do not assume unit-root levels processes–rather than they are autoregressive.

equation prior means are also set to 0. Moreover, Minnesota "own equation" variances shrink as a harmonic series for each additional lag (i.e. $\omega_{ii,j} = \frac{0.2}{j}\ s.t. j = 1, \ldots J$ and $\omega_{ii,j} \in \Omega_{\Pi}$ given the $i^{th}$ equation out of the $p$ series and where $j$ is the lag length of the autoregression). Also, "cross equation" variances are typically set to $\omega_{ik,j} = 0.5(\frac{0.2}{j} \times \frac{\sigma_i^*}{\sigma_k^*})$, where $\sigma_i^*$ is the estimated standard error of the residuals from a univariate autoregression on the $i^{th}$ equation with six lags, pre-fit to the data in advance. For simplicity, however, I will employ $\omega_{ik,j} = 0.5(\frac{0.2}{j})$ instead.

Priors on the deterministic parameters of $\mathbf{\Psi}$ are extremely important given the modest sample sizes employed and are chosen as to influence the series' steady-states toward certain reasonable values. In the case where we pre-difference the data series by their deterministic trends (specified in the data section above) $\mathbf{\Psi_{p \times q}}$ is such that $q = 1$ and $\mathbf{d_t} = 1, \forall t$. This dramatically reduces the number of parameters that need be estimated as the number of series increases, however, it places a prior constraint on the model by assuming that the trends chosen are correct. On the other hand, if we allow the trends to enter individually through the $\mathbf{d_t}$ term (where $\mathbf{d_t}$ then becomes a 3 element column vector for each $t$, such that $\mathbf{d_t}$=$[1, f(t-1), g(t)]^T$ where $f(t)$ and $g(t)$ are exponentially smoothed trends for the unemployment rate and inflation growth respectively) then $\mathbf{\Psi_{p \times q}}$ becomes a $p \times 3$ matrix from which we can statistically evaluate whether the relevant diagonal elements are indeed equal to 1 (which would imply the trends are in fact correct).

Either way, GDP growth is influenced to have a constant trend around $3.0\%$ through its prior mean, while unemployment and inflation are pushed to center around their trends (so their prior means would be 0 in the case of pre-differencing and 1 in the case where trends are added later through $\mathbf{d_t}$). Finally, the interest rate is centered around its trend as well; however, we also add to this the constant trend of $2.5\%$ to reflect the real long run rate. Therefore, in the case where the data is pre-differenced, the prior mean of $\mathbf{\Psi_{p \times q}}$ (i.e. $\mu_{\mathbf{\Psi}}$ in matrix form) will take the form:

$$\begin{bmatrix} 3.0 \\ 0 \\ 0 \\ 2.5 \end{bmatrix}$$

and when added to $\mathbf{d_t}$ it would be:

$$\begin{bmatrix} 3.0 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \\ 2.5 & 0 & 1 \end{bmatrix}$$

The prior variances of $\mathbf{\Psi_{p \times q}}$ (i.e. $\mathbf{\Omega_{\Psi}}$) are set as follows: GDP growth, 0.2 (0.3); unemployment, 0.2 (0.3); inflation, 0.2 (0.3); and the interest rate, 0.6 (0.75)—where these values have been adopted directly from Clark (2011). The first values, not in parenthesis, represent those employed in the recursive estimation scheme and are tighter since the gradually increasing sample size tends to limit the influence of the prior.

Priors on the volatility components of the model are as follows. For the components of the Clark (2011) independent random walk specification, I borrow numerical values directly from his paper (see Clark (2011) pg.331 for details). For my Wishart autoregressive specification, I employ multivariate Normal priors on both $\mathbf{A_k}\ \forall k$ and $\mathbf{C}$, and a Gamma prior on $(v - p)$. The Gamma prior is set with hyperparameters $\alpha = 30, \beta = 2$ (shape and rate) as to represent ignorance of its value while the multivariate Normal priors for the $\mathbf{C}$ and $\mathbf{A_k}$'s are set somewhat

loosely to let the data speak. In this respect, prior means are set at $diag\left(\mathbf{C}\right)$ =0.3 and $diag\left(\mathbf{A}_1\right) = 0.9$ (both with off-diagonals equal to 0, and all other $\mathbf{A}$ matrices set to 0). Variances are set equal to 0.002 (i.e. standard deviation of about 0.045).

# 5 Forecasts

Given the Bayesian model estimation framework employed, forecasts can be easily obtained with little extra computational overhead. Generally, the desired predictive density of some forecasted value $y_f$, given parameter set $\theta$ and data set $y$ is given within the Bayesian framework as:[17]

$$p\left(y_f \mid y\right) = \int p\left(y_f \mid \theta, y\right) \pi\left(\theta \mid y\right) d\theta$$

where $p(y_f|\theta, y)$ is the likelihood of $\theta$ given the data $y$, the forecast value $y_f$, and the particular model formulation chosen; and where $\pi(\theta)$ is the prior distribution assumed on $\theta$. Therefore after integrating out parameter uncertainty, the left hand side is known as the *predictive likelihood* of the given forecast value $y_f$.

Moreover, given our particular model specification this expression can be recast in terms of VAR forecasts of horizon length $h$, that is $\mathbf{y_{T+h}}$, where we have taken into account the stochastic/latent nature of the conditional variance matrices $\mathbf{\Sigma_t}$ :

$$p\left(\mathbf{y_{T+1}}, \ldots, \mathbf{y_{T+h}} \mid \mathbf{Y^T}\right) = \int \ldots \int p(\mathbf{y_{T+1}}, \ldots, \mathbf{y_{T+h}}, \mathbf{\Sigma}_{\mathrm{T+1}}, \ldots, \mathbf{\Sigma}_{\mathrm{T+h}}, \mathbf{\Omega} \left| \mathbf{Y^T}\right) d\mathbf{\Omega} \prod_{\mathbf{i=1}}^{\mathbf{h}} d\mathbf{\Sigma}_{\mathrm{T+i}}$$

$$s.t.\ \mathbf{Y^T} = \{\mathbf{y_T}, \mathbf{y_{T-1}}, \ldots, \mathbf{y_1}\}$$

where $\mathbf{\Omega}$ represents all the other the other parameters of the model, including past covariance matrices (i.e. $\mathbf{\Sigma}_{\mathrm{T-i}}, \forall i \geq 0$ in the case of the $InvWishart\left(K\right)$ model or $\mathbf{\Lambda}_{\mathrm{T-i}}, \forall i \geq 0$ for the $Clark$) and where $p\left(\cdot\right)$ is the joint density implied by the model formulation. In drawing forecasts then, it is simply a matter of factorizing the above joint density and then drawing simultaneously from each factor. That is we can rewrite the above as:

$$p\left(\mathbf{y_{T+1}}, \ldots, \mathbf{y_{T+h}} \mid \mathbf{Y^T}\right) =$$

$$\int \ldots \int p\left(\mathbf{y_{T+1}}, \ldots, \mathbf{y_{T+h}} | \mathbf{\Sigma}_{\mathrm{T+1}}, \ldots, \mathbf{\Sigma}_{\mathrm{T+h}}, \mathbf{\Omega}, \mathbf{Y^T}\right) p(\mathbf{\Sigma}_{\mathrm{T+1}}, \ldots, \mathbf{\Sigma}_{\mathrm{T+h}} | \mathbf{\Omega}, \mathbf{Y^T}) \cdot$$

$$p(\mathbf{\Omega} | \mathbf{Y^T}) d\mathbf{\Omega} \prod_{\mathbf{i=1}}^{\mathbf{h}} d\mathbf{\Sigma}_{\mathrm{T+i}}$$

Note that we already possess draws from $p(\mathbf{\Omega}|\mathbf{Y^T})$ since this is simply the target posterior distribution of the parameter set given the data (i.e. that which we employed the Gibbs sampler to generate and which we obtain estimates of the model parameters conditional on the data $\mathbf{Y^T}$). Therefore, each time we draw a value for the Markov process, we can simultaneously draw a $\{\mathbf{\Sigma}_{\mathrm{T+1}}, \ldots, \mathbf{\Sigma}_{\mathrm{T+h}}\}$ and $\{\mathbf{y_{T+1}}, \ldots, \mathbf{y_{T+h}}\}$, given the parameterization of the model and its

[17] See Greenburg (2008) section 3.2.3.

implied conditional Normality. That is, given a draw for $\mathbf{\Omega}$ from the Markov process, we can condition on these "fixed" parameter values (and the data) to draw both a covariance matrix at time $T+1$, $\mathbf{\Sigma_{T+1}}$ (given the relevant volatility process equation) and a $VAR(J)$ data vector $\mathbf{y_{T+1}}$. Consequently, these draws are conditional on both the current values of the latent covariances and the regular parameters of the VAR (all of which are embodied in $\mathbf{\Omega}$). We then simply repeat this process recursively up to period $T+h$. Finally, since what we have essentially done by this process is drawn from the joint density $p(\mathbf{y_{T+1}}, \ldots, \mathbf{y_{T+h}}, \mathbf{\Sigma}_{\mathrm{T+1}}, \ldots, \mathbf{\Sigma}_{\mathrm{T+h}}, \mathbf{\Omega}|\mathbf{Y^T})$, it is understood then that $\{\mathbf{y_{T+1}}, \ldots, \mathbf{y_{T+h}}\}$ is therefore a sample from the desired marginal distribution $p\left(\mathbf{y_{T+1}}, \ldots, \mathbf{y_{T+h}} \mid \mathbf{Y^T}\right)$. Moreover, the draws themselves for each time $T+i$ represent a finite sample distribution on the forecasted values themselves. Therefore, conditional moments can be estimated by their sample counterparts.[18]

For the purposes of forecasting, that which we wish to estimate is the conditional mean of the predictive density $p\left(\mathbf{y_{T+1}}, \ldots, \mathbf{y_{T+h}} \mid \mathbf{Y^T}\right) = p\left(\mathbf{y_{T+h}} \mid \mathbf{y_{T+h-1}}, \mathbf{Y^T}\right) \ldots p\left(\mathbf{y_{T+1}} \mid \mathbf{Y^T}\right)$. Therefore, for the first horizon out we have that a consistent estimator is the mean of $\mathbf{y_{T+1}}$ since:

$$plim\left(\frac{1}{N}\sum_{i=1}^{N}\mathbf{y_{T+1}}\right) = E\left[\mathbf{y_{T+1}} \mid \mathbf{Y^T}\right] \;\; by\ LLN$$

where $N$ is the number of Markov process draw iterations. Moreover, it can be shown by the Law of Iterated Expectations that the Rao-Blackwellization of the estimator is suggested as follows:

$$\begin{aligned} E\left[\mathbf{y_{T+1}} \mid \mathbf{Y^T}\right] &= \int^{y} \mathbf{y_{T+1}} \cdot p\left(\mathbf{y_{T+1}} \mid \mathbf{Y^T}\right) d\mathbf{y_{T+1}} = \int^{y} \mathbf{y_{T+1}} \int^{\Omega} p\left(\mathbf{y_{T+1}}, \mathbf{\Omega} \mid \mathbf{Y^T}\right) d\mathbf{\Omega} d\mathbf{y_{T+1}} \\ &= \int^{y} \mathbf{y_{T+1}} \int^{\Omega} p\left(\mathbf{y_{T+1}} \mid \mathbf{\Omega}, \mathbf{Y^T}\right) p\left(\mathbf{\Omega} \mid \mathbf{Y^T}\right) d\mathbf{\Omega} d\mathbf{y_{T+1}} \\ &= \int^{\Omega} \int^{y} \mathbf{y_{T+1}} p\left(\mathbf{y_{T+1}} \mid \mathbf{\Omega}, \mathbf{Y^T}\right) d\mathbf{y_{T+1}} p\left(\mathbf{\Omega} \mid \mathbf{Y^T}\right) d\mathbf{\Omega} \\ = \int^{\Omega} \mathrm{E}_y\left[\mathbf{y_{T+1}} \mid \mathbf{\Omega}, \mathbf{Y^T}\right] p\left(\mathbf{\Omega} \mid \mathbf{Y^T}\right) d\mathbf{\Omega} &= E_{\Omega}\left[\mathrm{E}_y\left[\mathbf{y_{T+1}} \mid \mathbf{\Omega}, \mathbf{Y^T}\right] \mid \mathbf{Y^T}\right] = E_y[\mathbf{y_{T+1}}|\mathbf{Y^T}] \end{aligned}$$

Therefore, an appropriate estimator is also,

$$\frac{1}{N}\sum_{i=1}^{N} E_y[\mathbf{y_{T+1}}|\mathbf{\Omega_i}, \mathbf{Y^T}]$$

so long as the conditional expectation has an analytical form we can draw directly from (which works in this case since the $VAR(J)$ formulation specifies a linear conditional mean).

Moreover, again by the Iterated Law of Expectations we can employ the same estimators for any horizon $h = 1, 2, \ldots, H$ since we have that:

$$E_y[E_y\left[\mathbf{y_{T+h}} \mid \mathbf{y_{T+h-1}}, \mathbf{Y^T}\right] \left|\mathbf{y_{T+h-2}}, \mathbf{Y^T}\right] \ldots] = E_y[\mathbf{y_{T+h}}|\mathbf{Y^T}]$$

[18] There will therefore be one draw of $\{\mathbf{y_{T+1}}, \ldots, \mathbf{y_{T+h}}, \mathbf{\Sigma}_{\mathrm{T+1}}, \ldots, \mathbf{\Sigma}_{\mathrm{T+h}}\}$ for each iteration of the Markov process (i.e. for each draw of $\mathbf{\Omega}|\mathbf{Y^T}$). Thus taking the sample average of the former draws is a consistent estimator of $E[\mathbf{y_{T+i}}|\mathbf{Y^T}]$ or $E[\mathbf{\Sigma_{T+i}}|\mathbf{Y^T}]$ for each horizon $i = 1, \ldots, h$.

Furthermore, in comparing point forecasts we can employ the root mean squared error (RMSE) estimator for any forecast horizon $h$ :

$$RMSE_h = \sqrt{\frac{1}{T - T_0} \sum_{t=T_0}^{T} \mathbf{u_{t+h}} \mathbf{u_{t+h}^T}}$$

where the expected forecast error $\mathbf{u_{t+i}} = E\left[\mathbf{y_{t+i}} \mid \mathbf{Y^t}\right] - \mathbf{y^*_{t+i}}$ can be computed for out of sample forecasts[19] and $T - T_0$ represents the number of "runs" or forecasting attempts employing different out of sample sections (via a rolling or recursively growing sample window scheme—see below for a diagramatic explanation). Note that since we are taking the outer product of the forecast errors, the RMSE will be a Cholesky decomposition of a symmetric positive-definite matrix and there will exist $p(p+1)/2$ unique elements to compare. The off-diagonals in this case represent the squared forecast errors *across* macro series squared forecast errors and the main diagonal elements are the squared forecast errors for each individual series themselves.

Finally, the main measure of overall model fit can be computed as the sample estimator of the *predictive likelihood* described above:

$$p\left(\mathbf{y^*}_{\mathbf{t+h}} \mid \mathbf{Y^t}\right) = E_\Omega[p\left(\mathbf{y^*_{t+h}} \mid \mathbf{Y^t}, \mathbf{\Omega}\right)] \cong \frac{1}{\mathrm{N}} \sum_{i=1}^{N} p(\mathbf{y^*_{t+h}} | \mathbf{Y^t}, \mathbf{\Omega_i})$$

where $N$ is again the number of Markov process iterations, but in this case $\mathbf{y^*_{t+h}}$ is the true out of sample data point we had previously set aside for comparison. Therefore, this estimator gives us an idea of how well the model and parameter estimates "fit" the true outcomes that actually took place. That is, given that $\mathbf{y^*_{t+h}}$ is actually the value that occurred in reality, what is the probability of its occurrence given our model and estimated parameters? If our model is more congruent with the actual future outcomes than another, its *predictive likelihood* should be greater.

Note we typically take the log version of the above estimator and can then sum across the $T - T_0$ runs given different out of sample windows so as to gauge the model's performance across a number of independent forecasting attempts:

$$LPL_h = \sum_{t=T_0}^{T} \log p\left(\mathbf{y^*}_{\mathbf{t+h}} \mid \mathbf{Y^t}\right)$$

Finally, we can then compare the $Clark$ and $InvWishart(K)$ specifications by looking at the ratio or difference of these $LPL_h$ scores across different forecast horizons. More specifically, we can view the log-difference of two different model $LPL_h$'s as a means of comparing the "term structure" of competing forecast performance across all possible horizon choices $h = 1, 2, \ldots, H$. Furthermore, the log-difference can be "decomposed" into a sum of log-ratios which suggest at which points (i.e. which "runs") of the sample either model did better or worse at "fitting" the true VAR process at horizon $h$:

$$LPL_h^A - LPL_h^B = \sum_{t=T_0}^{T} \log p_A\left(\mathbf{y^*}_{\mathbf{t+h}} \mid \mathbf{Y^t}\right) - \sum_{t=T_0}^{T} \log p_B\left(\mathbf{y^*}_{\mathbf{t+h}} \mid \mathbf{Y^t}\right) =$$

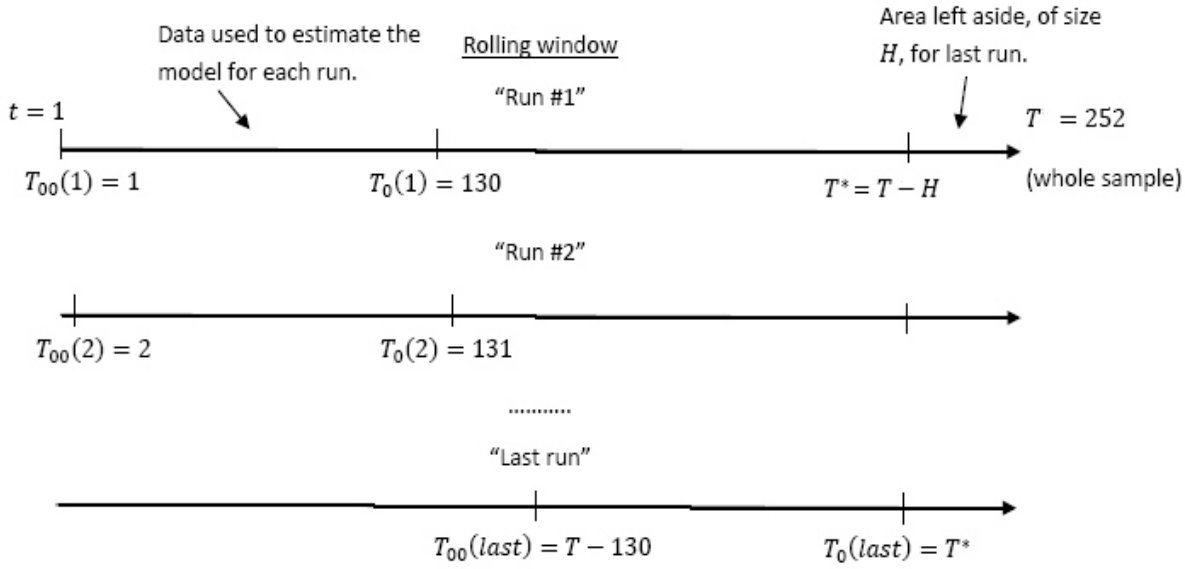
[19] That is, I set aside some portion of the sample from $T_0$ to $T$, s.t. $T_0 < T$. Therefore, forecast errors can be computed where $\mathbf{y^*_{t+h}}$ is the "real" out of sample data at forecast horizon $h$ and $E\left[\mathbf{y_{t+h}} \mid \mathbf{Y^t}\right]$ is estimated as described above and each time is conditioned on the sample data set given at $t$. This data set will depend on whether we are employing the rolling window or recursive scheme (see below).

$$\sum_{t=T_0}^{T} \log\left(\frac{p_A\left(\mathbf{y^*}_{\mathbf{t+h}} \mid \mathbf{Y^t}\right)}{p_B\left(\mathbf{y^*}_{\mathbf{t+h}} \mid \mathbf{Y^t}\right)}\right)$$

At this point it is worth mentioning that the *predictive likelihood* method represents an inherently Bayesian approach to forecast comparison and as such we do not require p-values, since we obtain the finite sample distribution directly. For more details on Bayesian versus Frequentist approaches to forecast generation and analysis see Geweke and Amisano (2010).

Again, forecasts comparisons are made by initially setting aside a portion of the data set to use as the "true" outcome of the macroeconomic time series (i.e. from $T_0$ to $T$, and leaving at least $H$ after $T$ available for comparison). As in Clark (2011) I employ both a recursive and rolling window estimation scheme. In the recursive process, for each iteration (or "run") one forward data point is added to the sample size each time so that the sample grows larger for later runs. Under the rolling window scheme, however, a set sample size is employed for each run but the sample employed moves forward in time. The idea here is that a rolling window should take into account changes in the underlying parameters across time.

To be absolutely clear, the following diagram illustrates both the rolling window and recursive schemes where $T_{00}(i)$ represents the beginning datum of the sample window under the rolling window scheme at "run" number $i$ and $T_0(i)$ represents the end of the window in both the rolling and recursive schemes under the same run number, in the case where we start with an initial window size of 130, chosen arbitrarily:




Note that the recursive scheme diagram is exactly the same except that $T_{00}(i) = 1, \forall i$; that is the window size grows from 130, by 1 each run, until it reaches $T^* = T - H$ for the last run.

After forecasting each iteration within the extended sample we've set aside, we eventually exhaust the entire dataset. Consequentially, given the estimated model parameters across each iteration, we can compute forecast RMSE's and LPL's for comparison. It is important to note that any deterministic trend components $\mathbf{d_{T+i}}$ must be constructed in real time for each forecast draw, as required according to their various specifications (e.g. exponential smoothed values), if we do not pre-difference the data by them beforehand.

The next section will detail not only the outcome of the parameter estimation process but the

results of Bayesian forecast comparisons between the $Clark$ benchmark forecasting model and the $InvWishart(K)$ modification.

# 6 Results

## 6.1 Simulated Data

The model is estimated separately for each run given the data set particular to that run's sample window. For each run, a Markov process is generated on the parameters of the model, of length 100,000 with a burn-in of 10,000. Sample means of the Markov processes draws are then interpreted as the estimated parameters given sufficiently well-behaved mixing.

To confirm that the estimation process is working correctly, I first simulate $1500$ data points according to the $VAR(3)\&InvWishart(3)$ combined processes given the following parameters:

| | |
|---|---|
| $Diag\left(\mathbf{C}\right) = 0.3$ | (all other elements are zero) |
| $Diag\left(\mathbf{A_1}\right) = 0.5,\ 0.75,\ 0.85,\ 0.98$ | (all other elements zero, including all other $\mathbf{A_2}, \mathbf{A_3}$ elements) |
| $Diag\left(\mathbf{\Pi_1}\right) = \{0.25,\ 0.8,\ 0.8,\ 0.8\}$ | (all other elements zero, including all other $\mathbf{\Pi_2}, \mathbf{\Pi_3}$ elements) |
| $v = 30$ | |

I then estimate the two competing specifications (i.e. $Clark$ and $InvWishart(K)$) according to a rolling window scheme, a fixed window size of $1000$, $100$ runs, and a forecast horizon extent of $100$.

The estimated parameters turn out to be quite close to their true values—see table A.1 in the appendix. The only parameter that seems systematically biased is the degree of freedom parameter of the $InvWishart(3)$ process, $v$. It tends to typically estimate lower than the true value for small sample sizes, and converges on the true value as $T$ gets large.

Moreover, under the $InvWishart(3)$ process, the tracking of the latent stochastic volatility estimated posterior means is also quite good, but tends to be worse when tracking series with low autoregressive coefficients on their lagged values (see figures 1 & 2 below).[20] It appears as though there exists a lower bound on the tracking of the posterior mean estimates across time of the latent covariance top-left element (i.e. that which corresponds to the series with a $0.5$ autoregressive coefficient—see figure 2). Multiple tests have revealed that the smaller is the autoregressive term, the more pronounced is this lower bound, and conversely, the closer is the autoregressive coefficient to $1$ the better the S.V. tracking. More generally, it is probably safe to assume that the tracking improves as the eigenvalues of the stability matrix $\mathbf{\Upsilon} = \sum_{i=1}^{\mathrm{K}} \mathbf{\Xi_i}$ approach $1$. More investigation is needed, however, to establish definitively the theoretical properties of this phenomenon.

[20]Note that all plots below that do not involve all runs simultaneously employ the $1^{st}$ run in each series, since subsequent runs are quite similar in their performance.

Figure 1: *InvWishart(3) - 6.1*

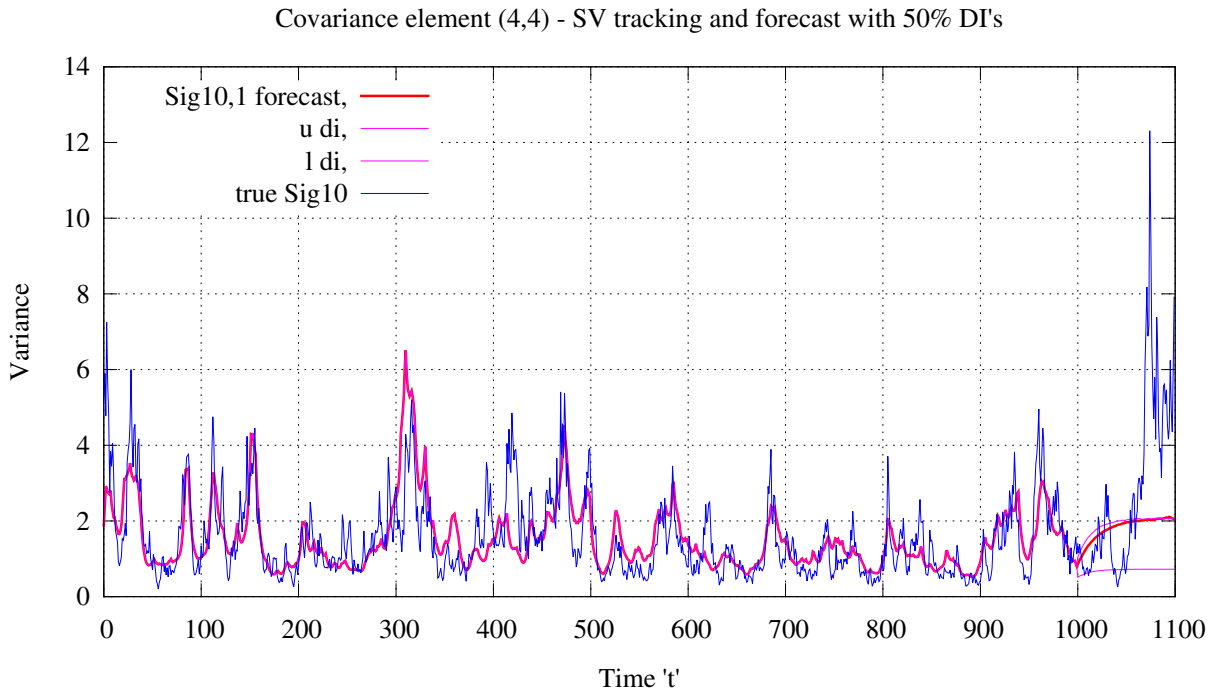



Figure 2: *InvWishart(3) - 6.1*

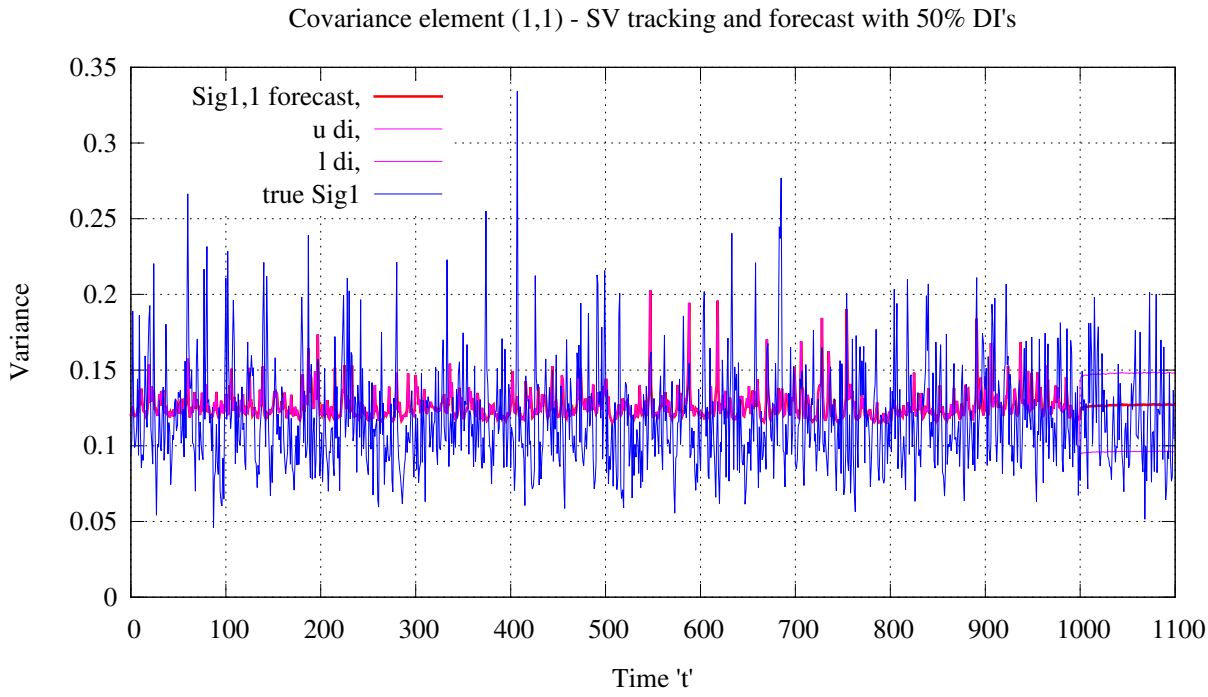
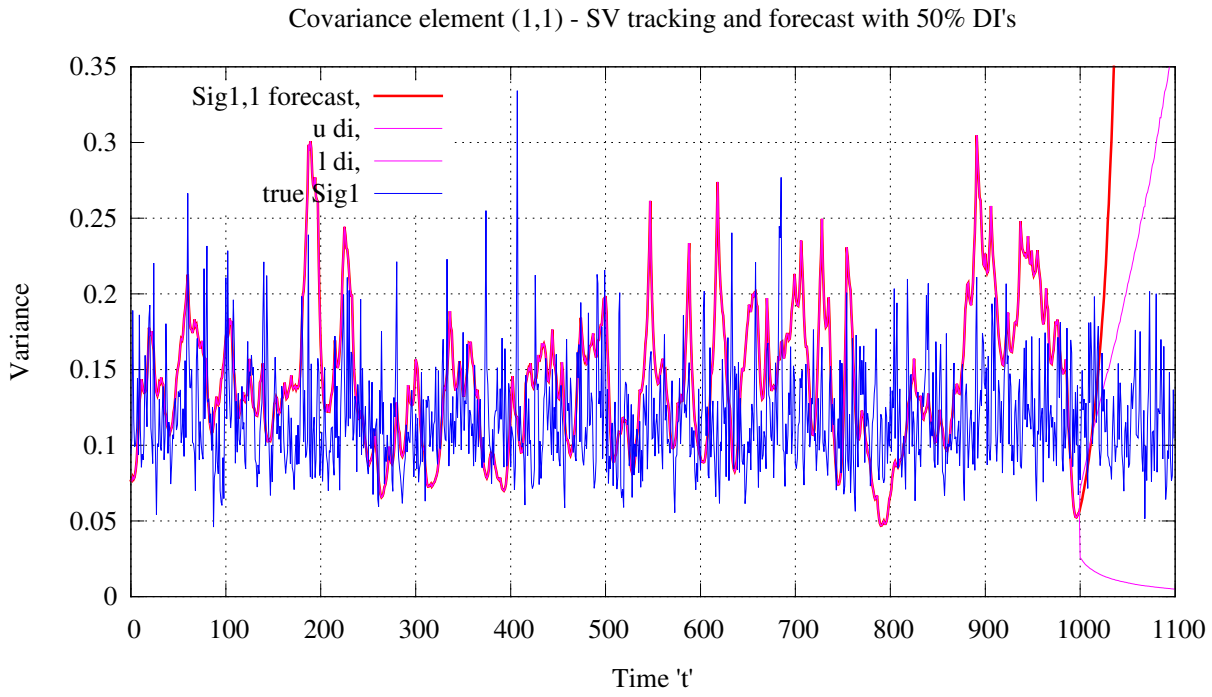


Figure 3: *Clark - 6.1*




Interestingly, the $Clark$ specification also suffers from poor performance in this case—see above figure 3. However, here the effect isn't so much a "sandwiching" or lower bound of the tracking (as in figure 2) but of a general meandering across the true values with no inherent pattern. Notice also the explosive forecast in red, appended to the S.V. tracking series (in pink) on the right-hand-side at $t = 1000$.

Finally, the log-predictive-likelihood ($LPL_h$) term structure of the $InvWishart(3)$ estimation is much better than the $Clark$ estimation, as is expected–see figure 4.

## 6.2 Real data (larger data window size used in each estimation but less runs)

Next, I turn to the real Clark (2011) data set. I first estimate the model with an initial window size of 204, 30 runs, and a horizon extent of 16, under the assumption of a $VAR\,(3)$ process on the mean equation and either an $InvWishart(3)$ or $Clark$ on the volatility. Both rolling window and recursive schemes are estimated. Note various lag values of $J$ and $K$ were tried and the general result was the more lags the better (as measured by $LPL_h$ term structure graphs)—three lags were ultimately chosen as a balance between parameterization and improvement in fit.

The following figure 5 suggests that the $InvWishart(3)$ formulation outperforms the $Clark$ volatility in both the short-term and long-term horizon forecasts, with the improvement becoming more dramatic as the horizon increases. Moreover, note that the recursive schemes tend to do better at longer horizons than their rolling window scheme counterparts, and vice versa at shorter horizons:

Figure 4: *Simulated data - 6.1*

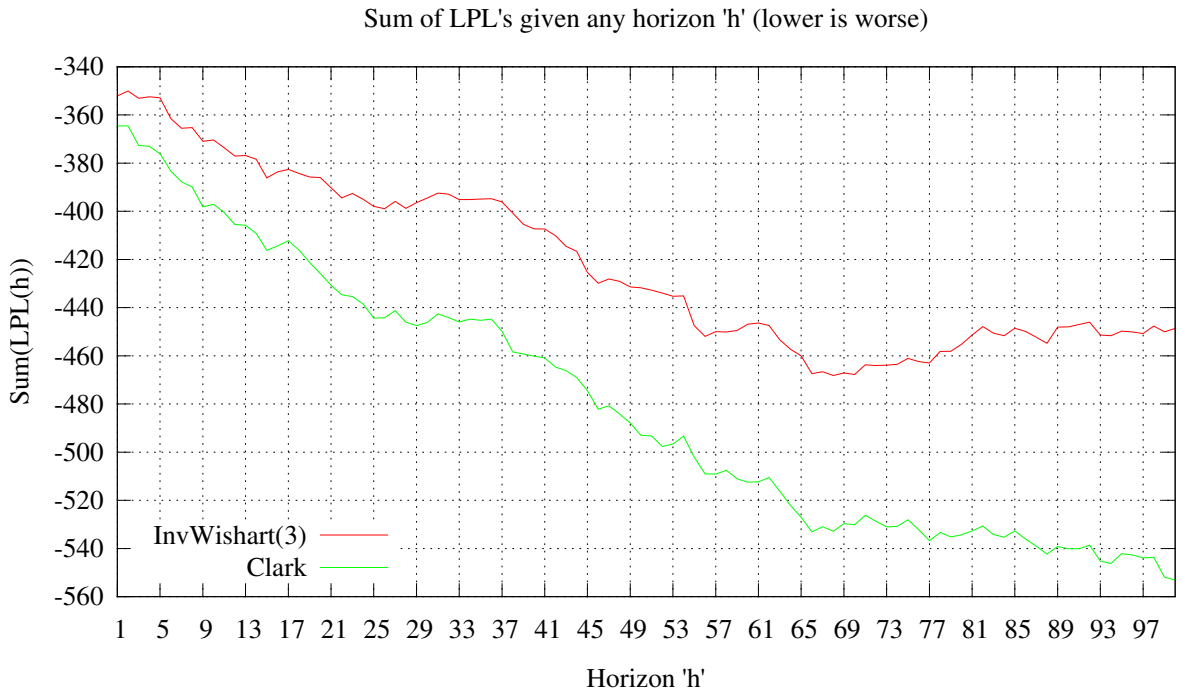
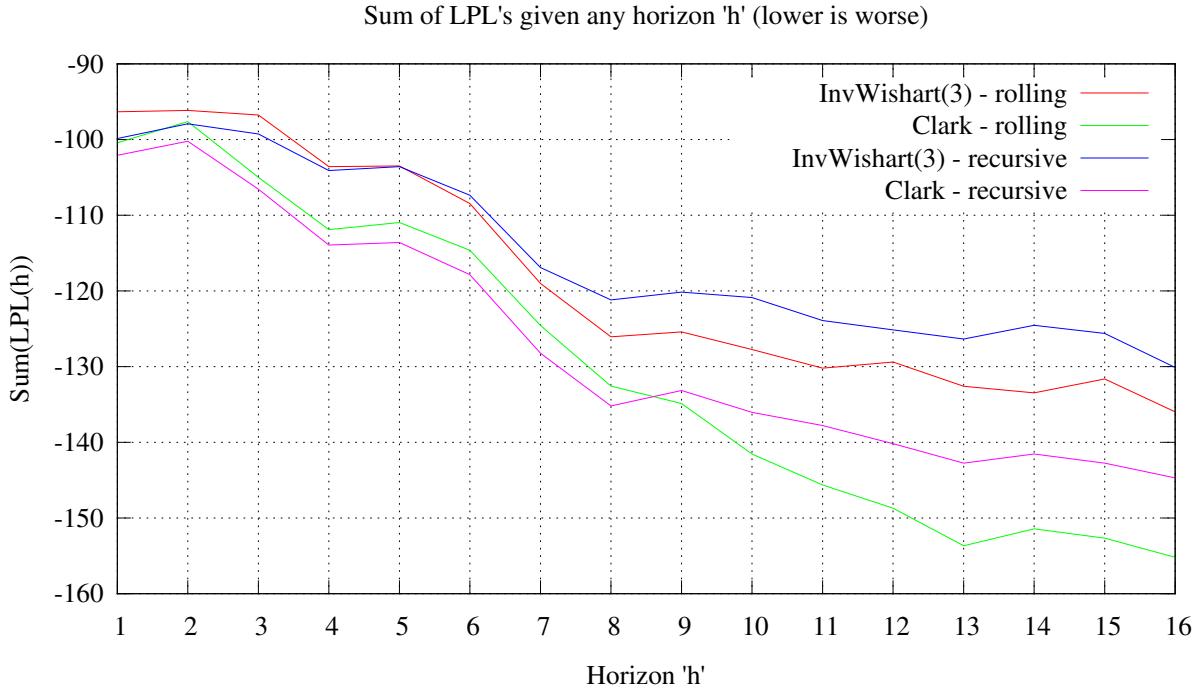


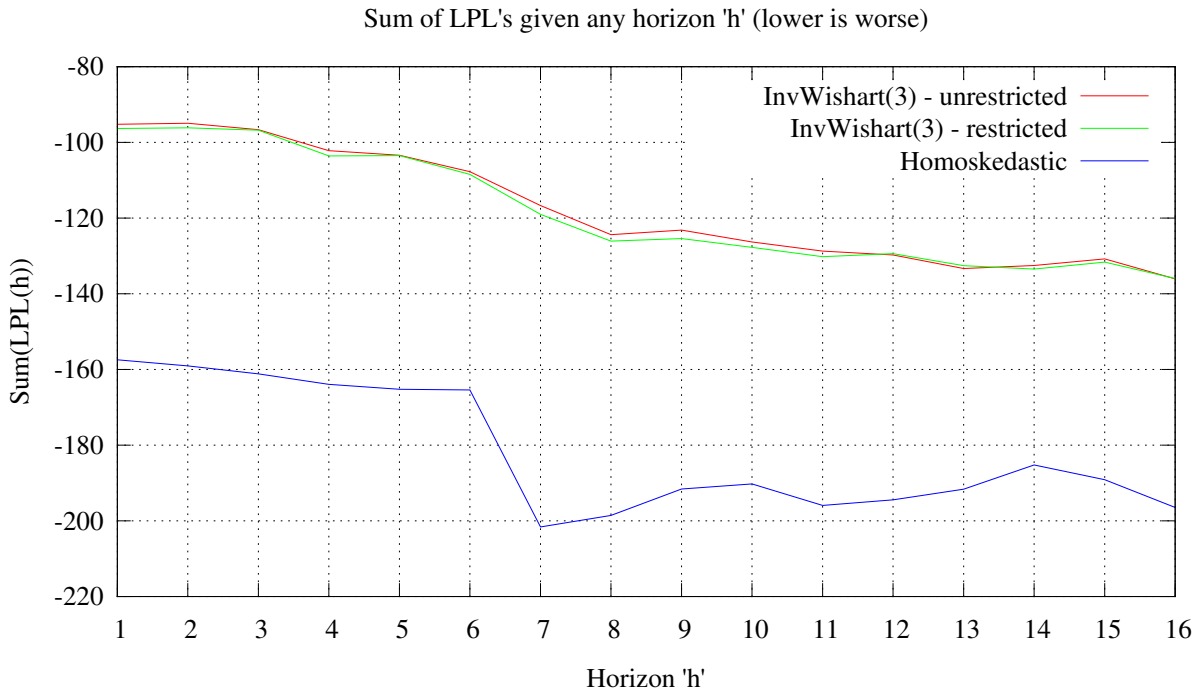

Figure 5: *Real data - 6.2*




Unsurprisingly, if I do not constrain the joint draws on the $\mathbf{C}$, $\mathbf{A_i}$ and $v$ parameters to remain in the area of the parameter space within which the unconditional mean of the $InvWishart(3)$ pro-

cess exists and is finite, the results are slightly improved (although the improvement is marginal at best). This is illustrated in figure 6. Note I also include results for the case where the covariance process is estimated as homoskedastic for reference (see footnote in section 6.4 below).

Figure 6: *Real data - 6.2*

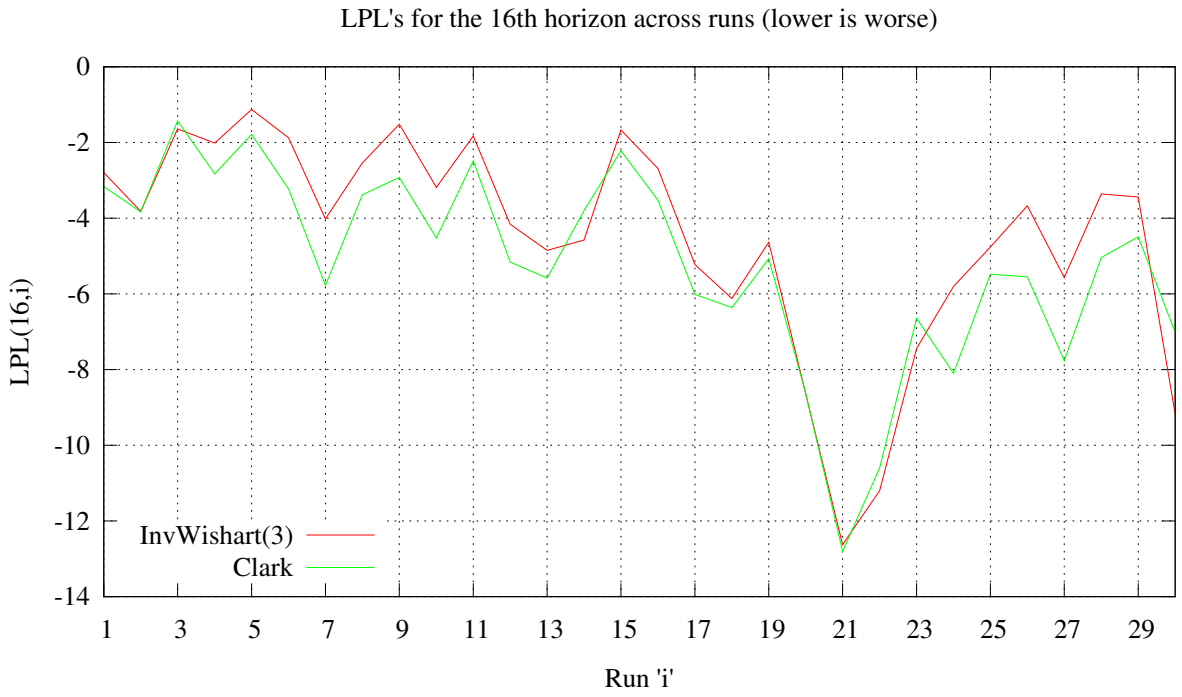



Moreover, the superiority of the $InvWishart(3)$ in fitting the data appears to be relatively uniform across the runs (and thus across the data set) at the $16^{th}$ horizon, although both specifications had problems at $t = 243 = 252 - 9$ (see figure 7, $run = 21$). Note that this datum corresponds with the 4th quarter of 2009 and represents an outlier given our small sample size. It will be shown later that this specifically represents a failure on the part of *both* models to predict the volatility correctly during this period.

Considering the earlier discussion from section 6, tracking of the latent stochastic volatility estimated posterior means of each element of the covariance matrices is assumed satisfactory given that the estimated volatilities of all the series are highly autoregressive (i.e. the eigenvalues of the stability matrix are close to 1, especially the one associated with GDP growth). Of course, we do not know the "true" path of the volatility (and no model is ever 100% correctly specified), so it is impossible to speak of how "well" the estimation tracks the truth. Nonetheless, figure 8 & 9 present the time-paths of the estimates of the covariance posterior means for both the volatility specifications, where the elements are interpreted as the $vech(\mathbf{\Sigma_t})$ (recall that since $\mathbf{\Sigma_t}$ is a $4 \times 4$ symmetric matrix, we have that 6 elements are redundant). Moreover, the posterior mean tracking is extended beyond $t = 204$ by means of the estimated forecasted means (which in the constrained $InvWishart(K)$ case exist and are finite – recall the earlier discussion in section 3).

Note first that the variance of the first series (i.e. GDP growth) is much larger than that of the other series in both cases. Secondly, the covariances across series tend to be largest on those between GDP growth and the inflation rate, and GDP growth and the interest rate. Third, the

Figure 7: *Real data - 6.2*




$InvWishart(3)$ model tends to highlight covariance changes more than the $Clark$. Perhaps most important to consider, however, is the bizarre nature of the covariance forecasts under the $Clark$ specification. Since the specification assumes a log random walk on the volatility processes of each series, the end result are forecasts that become explosive within a few periods out.[21]

[21] Interestingly, the result is the same whether we consider a volatility process on the covariance or precision matrix; both explode since the random walk sample paths of the variances (or precisions) under $Clark$ are bounded by 0.

Figure 8: *Real data - 6.2*

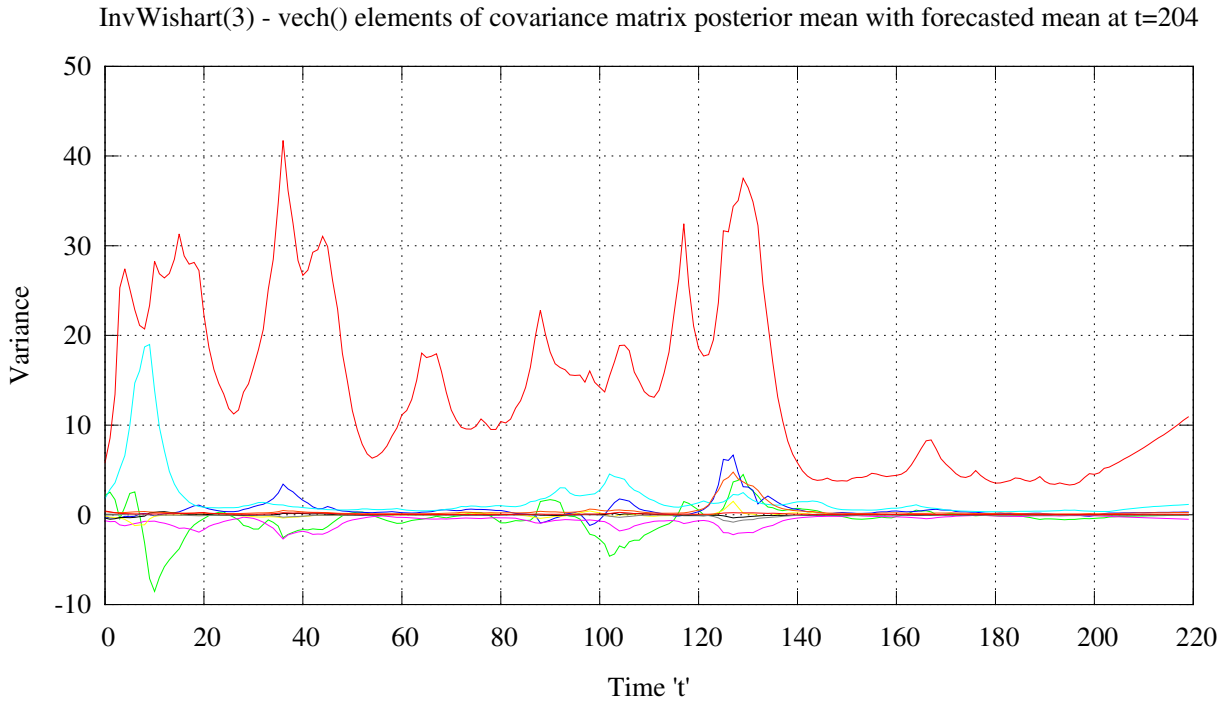
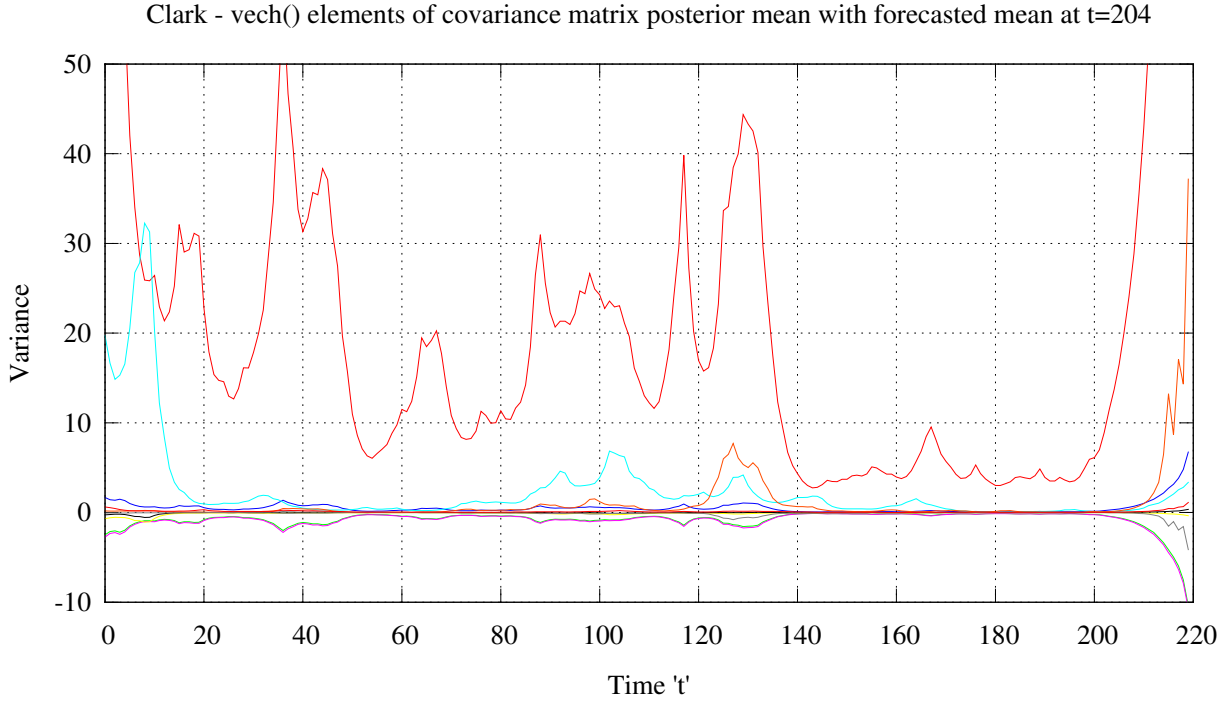
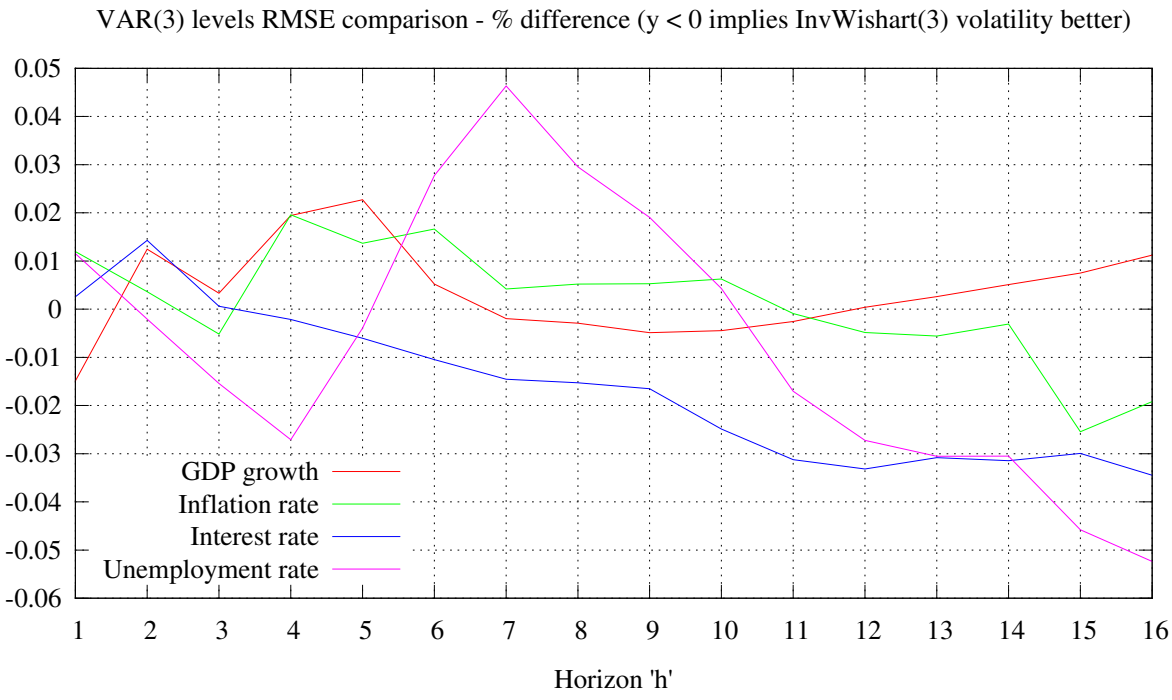



Figure 9: *Real data - 6.2*



The performance of the $VAR(3)$ forecasts of the series' means can be analyzed through the $RMSE_h$ term structure plots across horizons (recall the formula provided in section 5). In this case I take the percentage difference between the $RMSE_h$ values between the $InvWishart(3)$ and $Clark$ versions (given a rolling window scheme)–see figure 10.

Figure 10: *Real data - 6.2*




Of interest here is the fact that none of the percentage differences are greater than 5%. This suggests that for the most part, the volatility specification bears little effect on the estimated forecasts of the $VAR(3)$ mean.[22] Also, notice that the $InvWishart(3)$ volatility process tends to improve the $VAR(3)$ forecasted means out near the further horizons (i.e. $h > 11$) on the latter three series, while the $Clark$ does better in this case on the first series, GDP growth.

For completeness, I also include plots of the actual Clark (2011) data series, along with forecasts, in the appendix—see figures 17 to 24.

## 6.3 Real data (smaller data window size used for each estimation but more runs)

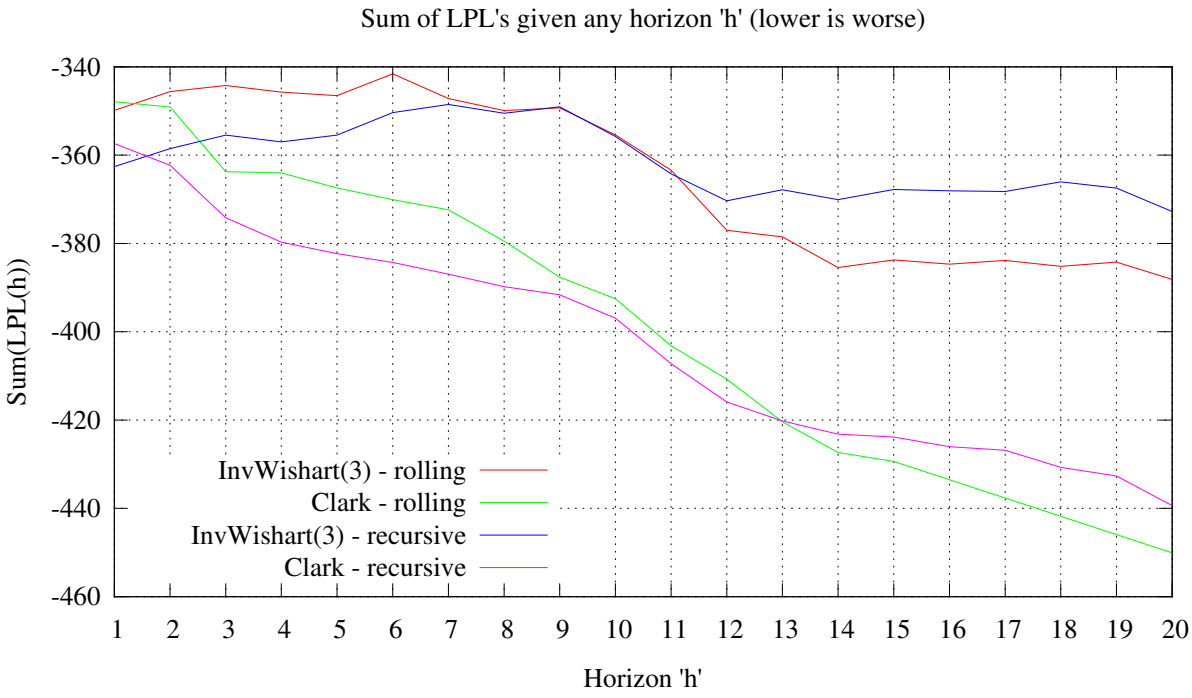
Now suppose that we wish to consider how the models compare when we employ a smaller initial window size, but more runs. That is, we estimate the model more often across the total sample so that our $LPL_h$ estimators are more representative (i.e. include a sum with more terms) of performance across more of the sample.

[22] To be more definitive on this issue, below I will demean the Clark (2011) data by estimating the $VAR(3)$ first with homoskedastic covariance assumed, and then re-estimate the volatility process directly on this demeaned data to isolate the volatility forecast quality.

In this case I employ an initial window size of 130, 100 runs, and a horizon extent of 20, again under the assumption of a $VAR(3)$ process on the mean equation and either an $InvWishart(3)$ or $Clark$ on the volatility. Both rolling window and recursive schemes are again estimated.

The corresponding $LPL_h$ term structure plot (i.e. the analog to figure 5 above) is given below in figure 11 and the analog to figure 7 is given in figure 12.

Figure 11: *Real data - 6.3*




These results serve to further reinforce that the $InvWishart(3)$ is preferred to the $Clark$ even when the sample size is relatively small. This is a surprising result considering the greater parameterization of the $InvWishart(3)$.

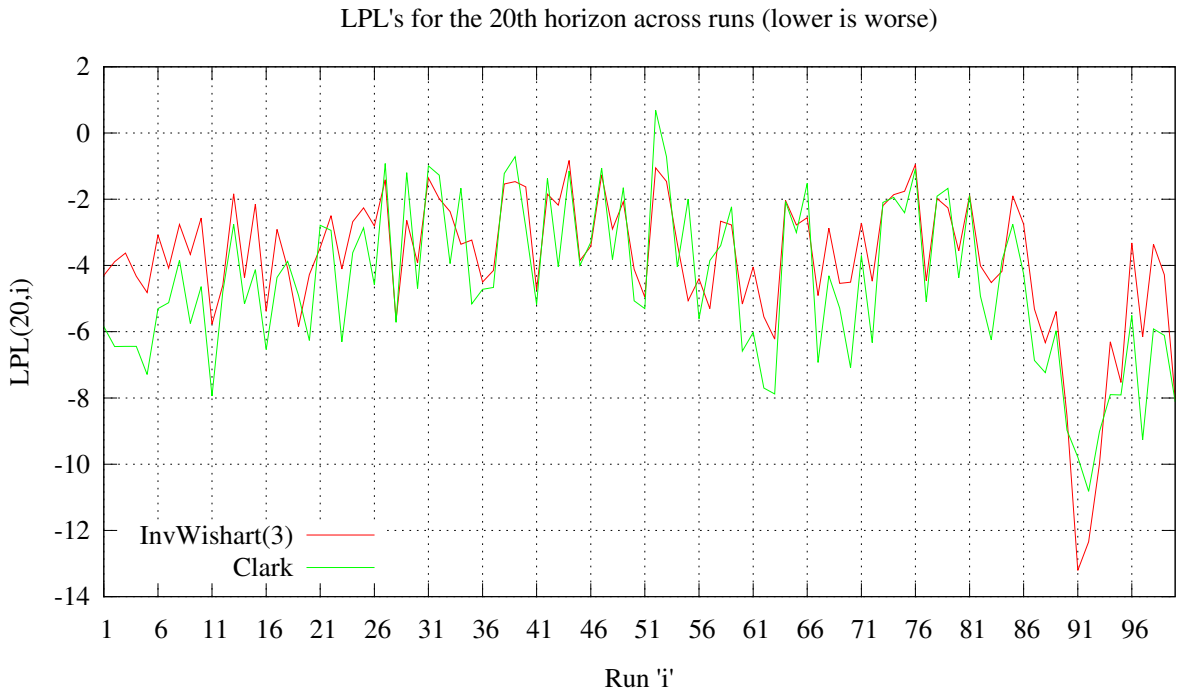
## 6.4 Real data (demeaned data)

In order to emphasize that the improvements in the $InvWishart(3)$ are due to the improved modeling on the time-varying covariance matrix forecasts and not due to any possible improvements in the $VAR(3)$ forecasted means, I now demean the Clark (2011) data prior to estimating the volatility process parameters.

Again, as in section 6.3, I employ an initial window size of 130, 100 runs, and a horizon extent of 20 in order to emphasize the forecasts across a majority of the sample. However, in this case, I first estimate a $VAR(3)$ on the whole data set with the assumption of homoskedastic covariance.[23]

[23]This simply involves replacing the Gibbs sampler steps for the volatility estimation with a single Gibbs step that employs an Inverse-Wishart prior density. Since the Inverse-Wishart prior is conditionally conjugate with the

Figure 12: *Real data - 6.3*




The corresponding $LPL_h$ term structure plot (i.e. the analog to figures 5 & 11 above) is given in figure 13.

What is impressive here is that after we demean the data series according to a $VAR(3)$ process with homoskedastic covariance matrix, and then re-estimate each of the two volatility processes (so as to focus solely on the volatility forecast quality) the superiority of the $InvWishart(3)$ over the $Clark$ is even more dramatic at $h = 20$ than was the case originally. In fact, it appears that the volatility forecasts do better at $h = 20$, than they do at $h = 1$, where the $Clark$ seems to have a small advantage. Given the small sample size employed in estimating the highly parameterized $InvWishart(3)$ the results appear even more impressive. Finally, notice that in the case where the $InvWishart(3)$ parameter draws are constrained to the stationary process region, we see that the volatility forecasts do better in the long-run than the short, and vice versa for the unconstrained version (as was in the case in sections 6.2 & 6.3).

Moreover, note that given the demeaned data we now find that the recursive scheme strictly dominates the rolling window given both volatility specifications. This suggests that for pure volatility forecasting a the recursive scheme is preferred.

Finally, I present the analog to figures 7 & 12 in figure 14.

---

multivariate Normal, the conditional posterior is also Inverse-Wishart. That is, if $\pi(\mathbf{\Sigma}) \sim IW(a_0, \mathbf{V}_0) \Rightarrow p(\mathbf{\Sigma} \mid \nu) \sim IW(a_1, \mathbf{V}_1)$ s.t $\mathbf{V}_1 = \sum_{t=1}^{T} \nu_t \nu_t{}^T + \mathbf{V}_0$ and $a_1 = T + a_0$, where $\nu_t$ is the $p \times 1$ vector of $VAR(J)$ residuals. In this case $\mathbf{V}_0$ is set to the unconditional sample covariance matrix of simulated VAR residuals (generated with reasonable guesses on the VAR parameters) and $a_0 = 15$.

Figure 13: *Real data - 6.4*

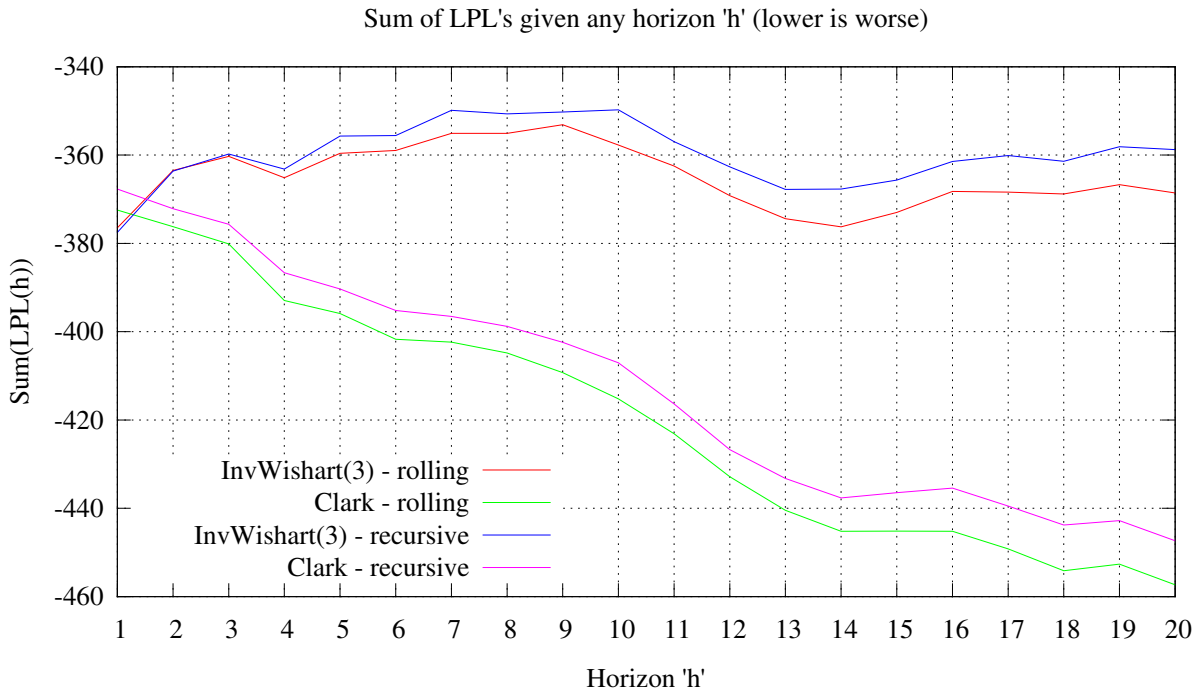
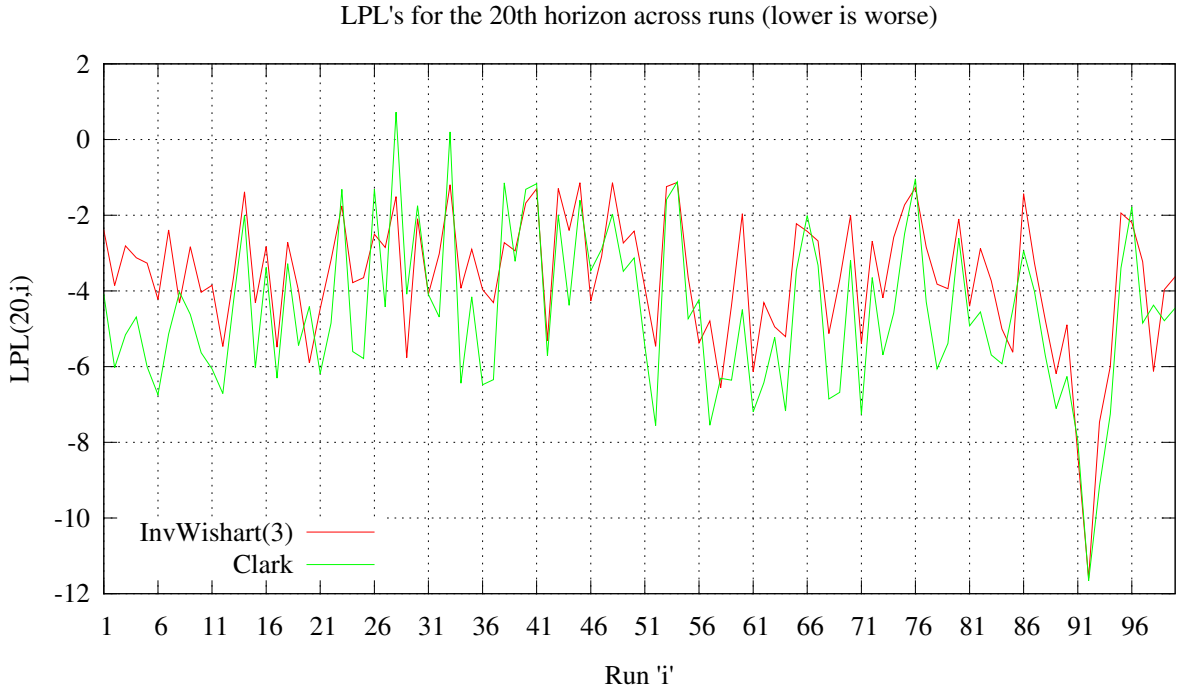



Figure 14: *Real data - 6.4*




From figure 14 it appears the forecasting improvements in the $InvWishart(3)$ are pretty uniform across runs, although there are a few cases where the $Clark$ actually outperforms the

$InvWishart(3)$ at the $20^{th}$ horizon (e.g. runs $i = 28$ and 33).

## 6.5 Real data ($LDL^T$ factorization)

Further evidence can also be drawn by considering the form of the constraint which the Clark model places on the time-varying structure of the covariance process. Recall from section 3.1 that the Clark model imposes the following parameterization on the $VAR(J)$ innovations:

$$\nu_{\mathbf{t}}=\mathbf{B}^{-1}\mathbf{\Lambda}_t^{0.5}\varepsilon_{\mathbf{t}} \quad s.t. \quad \varepsilon_{\mathbf{t}} \sim MVN_p\left(\mathbf{0},\mathbf{I_p}\right)$$

Which implies that:

$$\mathbf{\Gamma_t}=\mathbf{B}^{-1}\mathbf{\Lambda_t}\left(\mathbf{B}^{-1}\right)^T = var\left(\nu_{\mathbf{t}}\right)$$

The interesting point to note here is that this parameterization is equivalent to imposing an $LDL^T$ factorization on the covariance matrix of the $VAR(J)$ innovations, where $L$ is a lower triagular matrix with ones on the diagonal and $D$ is a diagonal matrix. Note that this $LDL^T$ factorization always exists for positive definite, real, symmetric matrices and is unique.

This result implies a method for testing whether Clark's parameteric assumption on the volatility process is correct given the Clark data set. Suppose that we estimate the model, given the entire sample, under an $InvWishart(K)$ specification. At each point in time that we draw a covariance matrix estimate, we then factorize this covariance matrix by $\mathbf{\Sigma}_t = L_t D_t L_t^T$. This provides us with finite sample distibutions of the $L_t$ matrices implied by the $InvWishart(K)$ specification for each time period 't'. Since these $L_t$ matrices are unique, their time-varying distributions must suggest something about whether or not it is appropriate to assume that the $\mathbf{B}^{-1}$ matrix in the Clark specification is constant across time.

Of course, while there are other ways to check the validity of this assumption (e.g. estimate the Clark with time-varying $\mathbf{B}$ matrices and compare results according to some metric), the aforementioned test proves the most immediately applicable.

Figures 15 and 16 illustrate the results of this test which seem to suggest that there does exist significant time variation in the elements of the $L$ matrix factor across time, especially with regards to the elements corresponding to the pairing of GDP growth with the inflation rate and the interest rate (i.e. $L_{2,1}$ and $L_{3,1}$).

Figure 15: *L matrix element densities*

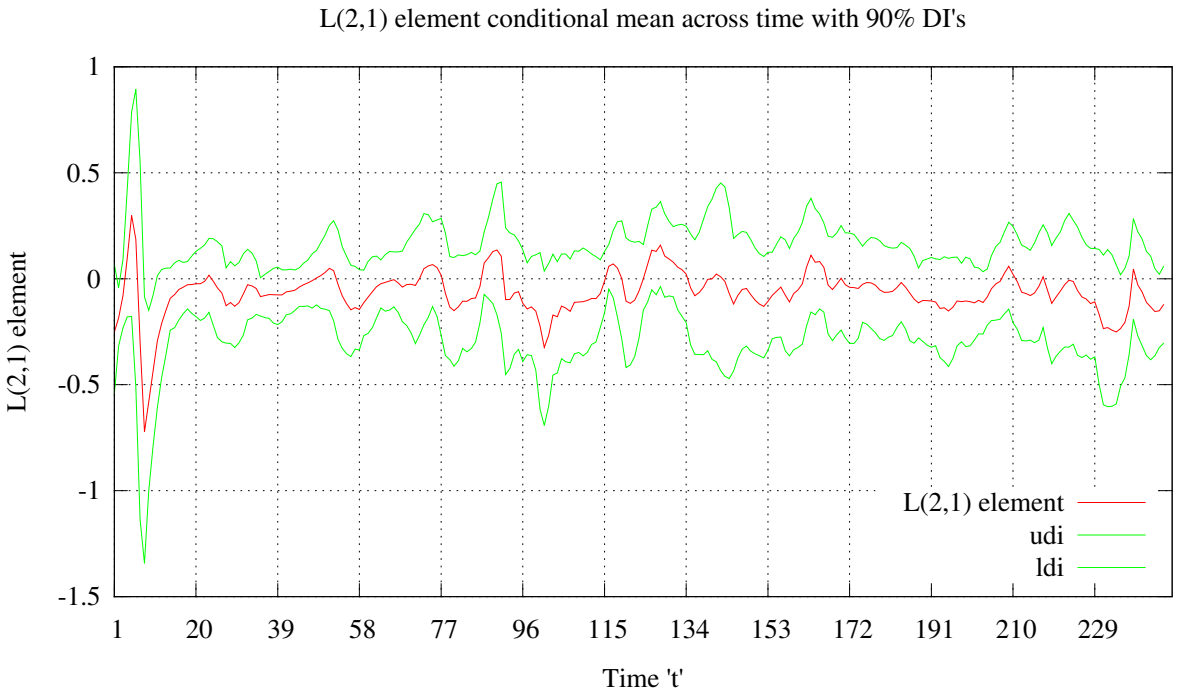
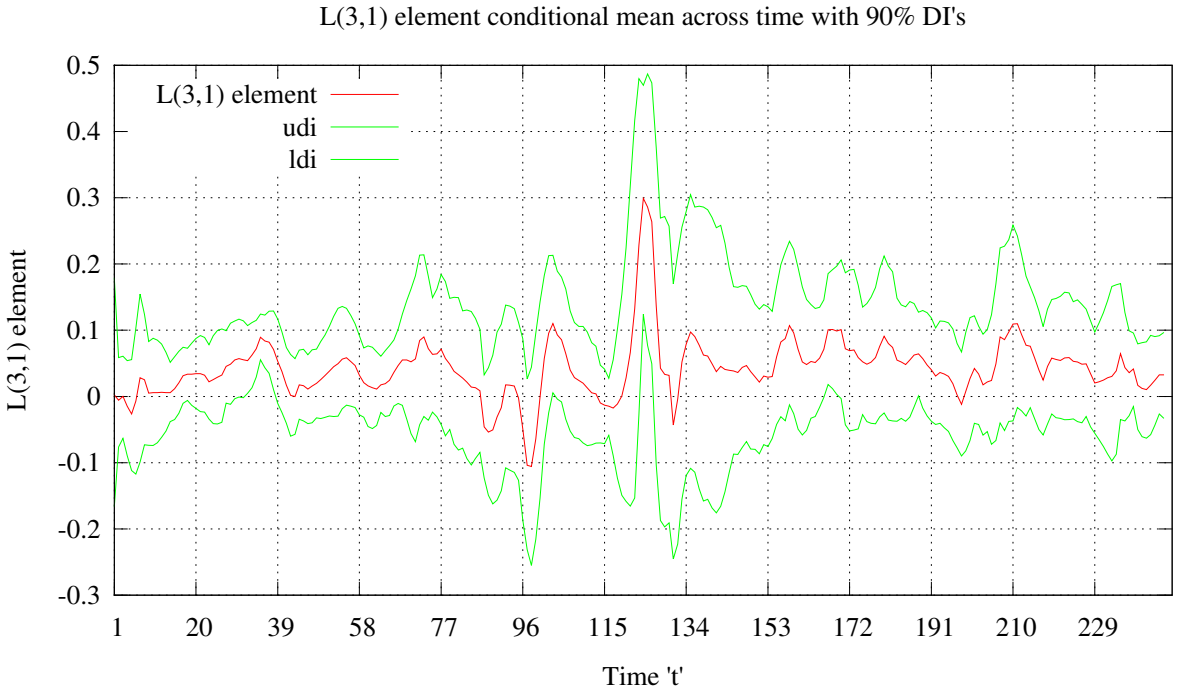



Figure 16: *L matrix element densities*



# 7 Conclusion

In conclusion, I have shown that given a competing volatility specification and accompanying data set from Clark (2011), direct modeling of the time-varying covariance process of VAR forecast shocks is superior to the assumption, common in the macroeconomic forecasting literature, of cross-series independent random walks in the variance. More specifically, the proposed $InvWishart(K)$ specification works to capture both cross and own series autoregressive persistence and mean reverting behaviour in the volatility process, and in doing so, dramatically improves long-run forecast fit as measured by Bayesian predictive likelihoods. Moreover, these improvements are achieved despite the small Clark (2011) sample size, the high degree of parameterization of the $InvWishart(K)$ specification, and are robust to varying sample window schemes (i.e. rolling window or recursive) and sizes.

# A Gibbs estimation steps

The following section describes in detail the steps required in generating the joint posterior distribution of the model parameters according to a Gibbs sampler approach. Note that it includes two sections: one for estimating the $VAR(J)$ coupled with $Clark$ volatility and another that employs instead $InverseWishart(K)$ volatility.

*BVAR-SSP-ClarkSV Gibbs Estimation Steps:*

1. Draw the slope coefficients $\mathbf{\Pi}^T_{\mathbf{p}\times\mathbf{pJ}}=[\mathbf{\Pi_1}, \mathbf{\Pi_2}, \ldots, \mathbf{\Pi_J}]$ of the VAR, conditional on $\mathbf{\Psi_{p\times q}}$, $\mathbf{\Lambda_t}$ $\forall t$, $\mathbf{B}$, and $\mathbf{\Phi}$=$diag(\varphi_1, \varphi_2, \ldots, \varphi_p)$ and given multivariate Normal prior, $\mathbf{\Pi} \sim N(\mu_{\mathbf{\Pi}}, \mathbf{\Omega_\Pi})$.

   For this step we rewrite the VAR as:

$$\mathbf{Y_t}=\mathbf{\Pi}^T\mathbf{X_t}+\nu_\mathbf{t}$$

$$where\ \mathbf{Y_t}=\mathbf{y_t}-\mathbf{\Psi d_t},\quad \nu_\mathbf{t}=\mathbf{B}^{-1}\mathbf{\Lambda}_t^{0.5}\varepsilon_\mathbf{t}$$

$$and\ \mathbf{X_t}=\left[\left(\mathbf{y_{t-1}}-\mathbf{\Psi d_{t-1}}\right)^\mathbf{T}, \left(\mathbf{y_{t-2}}-\mathbf{\Psi d_{t-2}}\right)^T, \ldots, \left(\mathbf{y_{t-J}}-\mathbf{\Psi d_{t-J}}\right)^T\right]^T$$

$$so\ \mathbf{X_t}\ is\ pJ\times 1$$

$$\Rightarrow\quad \mathbf{Y_t}=vec\left(\mathbf{\Pi}^T\mathbf{X_t}\right)+\nu_\mathbf{t}=vec\left(\left(\mathbf{\Pi}^T\mathbf{X_t}\right)^T\right)+\nu_\mathbf{t}=vec\left(\mathbf{X}_\mathbf{t}^T\mathbf{\Pi}\right)+\nu_\mathbf{t}$$

$$\Rightarrow\quad \mathbf{Y_t}=\left(\mathbf{I_p}\otimes\mathbf{X}_\mathbf{t}^T\right)\cdot vec\left(\mathbf{\Pi}\right)+\nu_\mathbf{t}$$

$$so\left(\mathbf{I_p}\otimes\mathbf{X}_\mathbf{t}^T\right)\ is\ p\times p^2J$$

   Eliminating the heteroskedasticity by pre-multiplication we have:

$$\mathbf{Y}_\mathbf{t}^*=\mathbf{\Gamma}_\mathbf{t}^{-0.5}\mathbf{Y_t}=\mathbf{\Gamma}_\mathbf{t}^{-0.5}\left(\mathbf{I_p}\otimes\mathbf{X}_\mathbf{t}^T\right)\cdot vec\left(\mathbf{\Pi}\right)+\varepsilon_\mathbf{t}$$

$$s.t.\ \varepsilon_\mathbf{t}\ \sim N\left(\mathbf{0}, \mathbf{I_p}\right)\ and\ \mathbf{\Gamma}_\mathbf{t}^{-0.5}=\mathbf{\Lambda}_\mathbf{t}^{-0.5}\mathbf{B}$$

$$Since\ \mathbf{\Gamma_t}=\mathbf{B}^{-1}\mathbf{\Lambda_t}\left(\mathbf{B}^{-1}\right)^T = var\left(\nu_\mathbf{t}\right)$$

$$\implies \mathbf{\Gamma_t}= \mathbf{B}^{-1}\mathbf{\Lambda_t}^{0.5}\left(\mathbf{\Lambda_t}^{0.5}\right)^T\left(\mathbf{B}^{-1}\right)^T = \mathbf{B}^{-1}\mathbf{\Lambda_t}^{0.5}\left(\mathbf{B}^{-1}\mathbf{\Lambda_t}^{0.5}\right)^T = \mathbf{\Gamma_t}^{0.5}\left(\mathbf{\Gamma_t}^{0.5}\right)^T$$

$$\implies \mathbf{\Gamma_t}^{-0.5}=\mathbf{\Lambda_t}^{-0.5}\mathbf{B}$$

Thus stacking the columns by time $'t'$, we have:

$$\mathbf{Y}^*_{Tp\times 1}=\left[\left(\mathbf{\Gamma_1}^{-0.5}\mathbf{Y_1}\right)^T,\ldots,\left(\mathbf{\Gamma}_T^{-0.5}\mathbf{Y}_T\right)^T\right]^T$$

$$\mathbf{X}^*_{Tp\times p^2J}=\left[\left(\mathbf{\Gamma_1}^{-0.5}\left(\mathbf{I}_p\otimes\mathbf{X}_\mathbf{1}^T\right)\right)^T,\ldots,\left(\mathbf{\Gamma}_T^{-0.5}\left(\mathbf{I}_p\otimes\mathbf{X}_T^T\right)\right)^T\right]^T$$

Finally, by standard results,[24] we have that the conditional posterior for $\mathbf{\Pi}$ is distributed Normal with mean vector $\mu^*_\mathbf{\Pi}$ and covariance matrix $\mathbf{\Omega}^*_\mathbf{\Pi}$ s.t:

$$\mathbf{\Omega}^*_\mathbf{\Pi} = \left[\mathbf{\Omega}^{-1}_\mathbf{\Pi} + \mathbf{X}^{*T}\mathbf{X}^*\right]^{-\mathbf{1}}$$

$$\mu^*_\mathbf{\Pi}=\mathbf{\Omega}^*_\mathbf{\Pi}[\mathbf{\Omega}^{-1}_\mathbf{\Pi}\mu_\mathbf{\Pi}+\mathbf{X}^{*T}\mathbf{Y}^*]$$

where $\mu_\mathbf{\Pi}$ and $\mathbf{\Omega}^{-1}_\mathbf{\Pi}$ are the mean vector and co-precision matrix of the Normal prior density for $\mathbf{\Pi}$. See the section on priors for more details on actual numerical values.

Note that it can also be shown that:

$$\mathbf{X}^{*T}\mathbf{X}^* = \sum_{t=1}^{T}\left[\mathbf{\Gamma_t}^{-0.5}\left(\mathbf{I}_p\otimes\mathbf{X_t}^T\right)\right]^T\left[\mathbf{\Gamma_t}^{-0.5}\left(\mathbf{I_p}\otimes\mathbf{X_t}^T\right)\right] = \sum_{t=1}^{T}\left(\mathbf{\Gamma_t}^{-0.5}\otimes\mathbf{X_t}^T\right)^T\left(\mathbf{\Gamma_t}^{-0.5}\otimes\mathbf{X_t}^T\right) =$$

$$\sum_{t=1}^{T}\left(\mathbf{\Gamma_t}^{-0.5^T}\otimes\mathbf{X_t}\right)\left(\mathbf{\Gamma_t}^{-0.5}\otimes\mathbf{X_t}^T\right) = \sum_{t=1}^{T}\left(\mathbf{\Gamma_t}^{-1}\otimes\mathbf{X_t}\mathbf{X_t}^T\right)$$

where the second equality holds only because $\mathbf{X}_\mathbf{t}^T$ has row order 1 (i.e is a row vector). This expression for $\mathbf{X}^{*\mathrm{T}}\mathbf{X}^*$ is desirable because it involves a sum of smaller matrices rather than the single product of an overly large matrix.

Moreover, we also have that:

$$\mathbf{X}^{*\mathrm{T}}\mathbf{Y}^*=\sum_{t=1}^{T}\left[\mathbf{\Gamma_t}^{-0.5}\left(\mathrm{I_p}\otimes\mathbf{X_t}^T\right)\right]^T\left(\mathbf{\Gamma_t}^{-0.5}\mathbf{Y_t}\right)=\sum_{t=1}^{T}\left(\mathbf{I_p}\otimes\mathbf{X_t}\right)\left(\mathbf{\Gamma_t}^{-0.5}\right)^T\mathbf{\Gamma_t}^{-0.5}\mathbf{Y_t}=$$

$$\sum_{t=1}^{T}\left(\mathbf{I_p}\otimes\mathbf{X_t}\right)\mathbf{\Gamma_t}^{-1}\mathbf{Y_t}=\sum_{t=1}^{T}\left(\mathbf{\Gamma_t}^{-1}\otimes\mathbf{X_t}\right)\mathbf{Y_t}=\sum_{t=1}^{T}\left(\mathbf{\Gamma_t}^{-1}\otimes\mathbf{X_t}\right)vec(\mathbf{Y_t}) =$$

$$\sum_{t=1}^{T}vec(\left[\mathbf{\Gamma_t}^{-1}\mathbf{Y_t}\mathbf{X_t}^T\right]^T)=vec(\sum_{t=1}^{T}\mathbf{X_t}\mathbf{Y_t}^T\mathbf{\Gamma_t}^{-1})$$

where again, the fourth equality holds because $\mathbf{X}_t$ has column order 1 (i.e. is a column vector).

[24] See Tsay (2005) section 12.3.2, or Box and Tiao (1973).

2. Draw the steady state coefficients $\mathbf{\Psi_{p\times q}}$ of the VAR, conditional on $\mathbf{\Pi}$', $\mathbf{\Lambda_t}$ $\forall t$, $\mathbf{B}$, and $\mathbf{\Phi}$=$diag(\varphi_1, \varphi_2, \ldots, \varphi_p)$ and given multivariate Normal prior, $\mathbf{\Psi} \sim N(\mu_{\mathbf{\Psi}}, \mathbf{\Omega_\Psi})$.

   This time, rewrite the VAR somewhat differently as:

$$\mathbf{Y_t}=\mathbf{\Pi}\left(L\right)\mathbf{y_t}=\mathbf{\Pi}\left(L\right)\mathbf{\Psi d_t}+\nu_\mathbf{t}$$

   This can be further simplified as:

$$\Rightarrow \mathbf{Y_t} = \mathbf{\Pi}(L)\mathbf{y_t} = \left(\mathbf{I_p} - \sum_{j=1}^{J} \mathbf{\Pi}_j L^j\right)\mathbf{\Psi d_t} + \nu_t$$

$$= \mathbf{I_p\Psi d_t} - \mathbf{\Pi}_1\mathbf{\Psi d_{t-1}} - \cdots - \mathbf{\Pi}_J\mathbf{\Psi d_{t-J}} + \nu_t$$

$$= \left(\mathbf{d_t}^T \otimes \mathbf{I_p}\right)\cdot vec(\mathbf{\Psi}) - \left(\mathbf{d_{t-1}}^T \otimes \mathbf{\Pi}_1\right)\cdot vec(\mathbf{\Psi}) - \cdots - \left(\mathbf{d_{t-J}}^T \otimes \mathbf{\Pi}_J\right)\cdot vec(\mathbf{\Psi}) + \nu_t$$

$$= \left(\left(\mathbf{d_t}^T \otimes \mathbf{I_p}\right) - \left(\mathbf{d_{t-1}}^T \otimes \mathbf{\Pi}_1\right) - \cdots - \left(\mathbf{d_{t-J}}^T \otimes \mathbf{\Pi}_J\right)\right)\cdot vec(\mathbf{\Psi}) + \nu_t$$

$$= \mathbf{X_t}\cdot vec(\mathbf{\Psi}) + \nu_t$$

$$s.t. \qquad \mathbf{X_t} = \left(\left(\mathbf{d_t}^T \otimes \mathbf{I_p}\right) - \left(\mathbf{d_{t-1}}^T \otimes \mathbf{\Pi}_1\right) - \cdots - \left(\mathbf{d_{t-J}}^T \otimes \mathbf{\Pi}_J\right)\right)$$
$$\text{is} \quad p \times pq$$

   Therefore, removing the heteroskedasticity again by pre-multiplication we have:

$$\mathbf{Y_t^*}=\mathbf{\Gamma_t^{-0.5}Y_t}=\mathbf{\Gamma_t^{-0.5}}(\mathbf{\Pi}\left(L\right)\mathbf{y_t}) =\mathbf{\Gamma_t^{-0.5}X_t}\cdot vec\left(\mathbf{\Psi}\right)+\varepsilon_\mathbf{t}$$

   And again, stacking by columns across $'t'$ as before, we have:

$$\mathbf{Y}^*_{Tp\times 1}=\left[\left(\mathbf{\Gamma_1^{-0.5}Y_1}\right)^T,\ldots,\left(\mathbf{\Gamma}_T^{-0.5}\mathbf{Y}_T\right)^T\right]^T$$

$$\mathbf{X}^*_{Tp\times pq}=\left[\left(\mathbf{\Gamma_1^{-0.5}X_1}\right)^T,\ldots,\left(\mathbf{\Gamma}_T^{-0.5}\mathbf{X}_T\right)^T\right]^T$$

   Finally, by standard results, we have that the conditional posterior for $\mathbf{\Psi}$ is distributed Normal with mean vector $\mu^*_{\mathbf{\Psi}}$ and covariance matrix $\mathbf{\Omega^*_\Psi}$ s.t:

$$\mathbf{\Omega^*_\Psi} = \left[\mathbf{\Omega_\Psi^{-1}} + \mathbf{X}^{*\mathrm{T}}\mathbf{X}^*\right]^{-\mathbf{1}}$$

$$\mu^*_{\mathbf{\Psi}}=\mathbf{\Omega^*_\Psi}[\mathbf{\Omega_\Psi^{-1}}\mu_{\mathbf{\Psi}}+\mathbf{X}^{*T}\mathbf{Y}^*]$$

   where $\mu_{\mathbf{\Psi}}$ and $\mathbf{\Omega_\Psi^{-1}}$ are the mean vector and co-precision matrix of the prior density for $\mathbf{\Psi}$. See the section on priors for more details on actual numerical values.

   Again, similar to step (1) above, we can rewrite the expressions for $\mathbf{X}^{*T}\mathbf{X}^*$ and $\mathbf{X}^{*T}\mathbf{Y}^*$ in more computationally tractable forms:

$$\mathbf{X}^{*\mathrm{T}}\mathbf{X}^*=\sum_{t=\mathbf{1}}^{T}\left(\mathbf{\Gamma_t^{-0.5}X_t}\right)^T\mathbf{\Gamma_t^{-0.5}X_t}=$$

$$\sum_{t=1}^{T} \mathbf{X_t}^T \mathbf{\Gamma_t}^{-0.5T} \mathbf{\Gamma_t}^{-0.5} \mathbf{X_t} = \sum_{t=1}^{T} \mathbf{X_t}^T \mathbf{\Gamma_t}^{-1} \mathbf{X_t}$$

$$\mathbf{X}^{*T}\mathbf{Y}^{*}=\sum_{t=\mathbf{1}}^{T} \left(\mathbf{\Gamma_t}^{-0.5} \mathbf{X_t}\right)^T \mathbf{\Gamma_t}^{-0.5} \mathbf{Y_t} = \sum_{t=1}^{T} \mathbf{X_t}^T \mathbf{\Gamma_t}^{-1} \mathbf{Y_t}$$

3. Draw the elements of $\mathbf{B}$ (lower triangular with ones in the diagonal) conditional on $\mathbf{\Pi}^{'}$, $\mathbf{\Psi}^{'}$, $\mathbf{\Lambda_t}$ $\forall t$, and $\mathbf{\Phi}$=$diag(\varphi_1, \varphi_2, \ldots, \varphi_p)$, given Normal, independent, priors on each of the elements of the $\mathbf{B}$ matrix.

   This time rewrite the VAR as:

$$\mathbf{B\Pi}\left(L\right)\left(\mathbf{y_t} - \mathbf{\Psi d_t}\right) = \mathbf{B}\widetilde{\mathbf{y_t}}=\mathbf{\Lambda_t}^{0.5}\varepsilon_{\mathbf{t}}$$

   Since $\mathbf{B}$ is lower triangular, this system of equations reduces to:

$$\widetilde{y_{1,t}} = \lambda_{1,t}^{0.5}\varepsilon_{1,t}$$

$$\widetilde{y_{2,t}} = -b_{21}\widetilde{y_{1,t}} + \lambda_{2,t}^{0.5}\varepsilon_{2,t}$$

$$\widetilde{y_{3,t}} = -b_{31}\widetilde{y_{1,t}} - b_{32}\widetilde{y_{2,t}} + \lambda_{3,t}^{0.5}\varepsilon_{3,t}$$

$$\widetilde{y_{4,t}} = -b_{41}\widetilde{y_{1,t}} - b_{42}\widetilde{y_{2,t}} - b_{43}\widetilde{y_{3,t}} + \lambda_{4,t}^{0.5}\varepsilon_{4,t}$$

$$\ldots.$$

$$\widetilde{y_{p,t}} = -b_{p1}\widetilde{y_{1,t}} - b_{p2}\widetilde{y_{2,t}} - b_{p3}\widetilde{y_{3,t}} - \ldots - b_{p,(p-1)}\widetilde{y_{(p-1),t}} + \lambda_{p,t}^{0.5}\varepsilon_{p,t}$$

   where $\widetilde{y_{i,t}}$ is the i$^{th}$ element of the $p \times 1$ column vector $\mathbf{\Pi}\left(L\right)\left(\mathbf{y_t} - \mathbf{\Psi d_t}\right) = \widetilde{\mathbf{y_t}}$.

   We can therefore treat each of the $i = 2, \ldots, p$ equations above as linear regressions. Again, pre-multiplication of each of the $i$ equations by $\lambda_{i,t}^{-0.5}$ $\forall t$ removes the heteroskedasticity. Furthermore, given the assumption of independent Normal prior densities, the conditional posterior for each row vector of $\mathbf{B}$ is also Normal. Consequently, we can easily draw the $b_{ij}$ elements for each row, equation by equation in sequence, according to $N\left(\beta_i{}^*, \mathbf{G}_i{}^*\right)$ $\forall i = 2, \ldots, p$ s.t:

$$\mathbf{G}_i^*=\left[\mathbf{G}_i{}^{-1} + \mathbf{X}_i^{*T}\mathbf{X}_i^*\right]^{-1}$$

$$\beta_i^*=\mathbf{G}_i^*[\mathbf{G}_i^{-1}\beta_i + \mathbf{X}_i^{*T}\mathbf{Y}_i^*]$$

   where $\mathbf{G}_i$=$\alpha \times I_{i-\mathbf{1}}$ (see priors section for the numerical value of $\alpha$), is the prior variance for equation $i$, and $\beta_i$ is the prior mean vector. Finally, $\mathbf{Y}_i^*=\left[\lambda_{i,1}^{-0.5}\widetilde{y_{i,1}}, \ldots, \lambda_{i,T}^{-0.5}\widetilde{y_{i,T}}\right]^T$ and:

$$\mathbf{X}_i^*=\begin{bmatrix} -\lambda_{i,1}^{-0.5}\widetilde{y_{1,1}} & -\lambda_{i,1}^{-0.5}\widetilde{y_{2,1}} & \ldots & -\lambda_{i,1}^{-0.5}\widetilde{y_{i-1,1}} \\ \ldots & \ldots & \ldots & \ldots \\ -\lambda_{i,T}^{-0.5}\widetilde{y_{1,T}} & -\lambda_{i,T}^{-0.5}\widetilde{y_{2,T}} & \ldots & -\lambda_{i,T}^{-0.5}\widetilde{y_{i-1,T}} \end{bmatrix}$$

4. Draw the elements of the time-varying covariance matrix $\mathbf{\Lambda_t}$for each time $'t'$ in sequence, each conditional on $\mathbf{\Pi}^{'}$, $\mathbf{\Psi}^{'}$, $\mathbf{B}'$, and $\mathbf{\Phi}$=$diag(\varphi_1, \varphi_2, \ldots, \varphi_p)$.

   First, note that since the stochastic volatility equations are independent of each other for all $i = 1, \ldots, p$, (i.e. $\mathbf{\Phi}$ is diagonal), we can estimate them each separately. In order to do so, however, what we need is an expression for the posterior density of each $\lambda_{i,t}$ conditional on everything else, including the entire macroeconomic series values for all $T$. This is also known as the "smoothed" density for $\lambda_{i,t}$.

   Since we have that the volatility equation is Markov of the first order we can write the conditional kernel for each $i = 1, \ldots, p$ as:[25]

$$g\left(\lambda_{i,t} \mid \lambda_{i,\backslash t}, \varphi_i, \mathbf{Y}_i^*\right) \propto f\left(\mathbf{Y}_i^* \mid \lambda_i\right) g\left(\lambda_{i,t} \mid \lambda_{i,\backslash t}, \varphi_i\right) \propto f\left(y_{i,t}^* \mid \lambda_{i,t}\right) g\left(\lambda_{i,t} \mid \lambda_{i,\backslash t}, \varphi_i\right) =$$

$$f\left(y_{i,t}^* \mid \lambda_{i,t}\right) g\left(\lambda_{i,t} \mid \lambda_{i,t-1}, \varphi_i\right) g\left(\lambda_{i,t+1} \mid \lambda_{i,t}, \varphi_i\right) = f\left(y_{i,t}^* \mid \lambda_{i,t}\right) g\left(\lambda_{i,t} \mid \lambda_{i,t-1}, \lambda_{i,t+1}, \varphi_i\right)$$

   s.t. $\mathbf{Y}_i^* = \left\{y_{i,1}^*, \ldots, y_{i,T}^*\right\}$ where $y_{i,t}^*$ is the $i^{th}$ element of $\mathbf{B\Pi}\left(L\right)\left(\mathbf{y_t} - \mathbf{\Psi d_t}\right)$. Furthermore, since $\lambda_{i,t}|\lambda_{i,t-1} \sim LN\left(e^{\ln(\lambda_{i,t-1}) + \frac{\varphi_i}{2}}, (e^{\varphi_i} - 1)e^{2\ln(\lambda_{i,t-1}) + \varphi_i}\right)$, we have that:

$$f\left(y_{i,t}^* \mid \lambda_{i,t}\right) g\left(\lambda_{i,t} \mid \lambda_{i,t-1}, \lambda_{i,t+1}, \varphi_i\right) \propto$$

$$\lambda_{i,t}^{-0.5} \exp\left(-\frac{\left(y_{i,t}^*\right)^2}{2\lambda_{i,t}}\right) \lambda_{i,t}^{-1} \exp\left(-\frac{\left(ln\left(\lambda_{i,t}\right) - \mu_{i,t}\right)^2}{2\sigma^2}\right)$$

   where we can solve for missing values according to section 12.6.1 of Tsay (2005) and find that:

$$\mu_{i,t} = \frac{1}{2}\left(\ln\left(\lambda_{i,t+1}\right) + \ln\left(\lambda_{i,t-1}\right)\right), \quad \sigma^2 = \frac{1}{2}\varphi_i$$

   Therefore, in implementing a Metropolis-within-Gibbs step we can draw a proposal for $\lambda_{i,t}^m$ from $\lambda_{i,t}^m \sim LN(e^{\mu_{i,t} + \frac{\sigma^2}{2}}, (e^{\mu_{i,t}} - 1)e^{2\mu_{i,t} + \sigma^2})$, and accept it with probability:

$$a\left(\lambda_{i,t}^{m-1}, \lambda_{i,t}^m\right) = \min\{1, \frac{f\left(y_{i,t}^* \mid \lambda_{i,t}^m\right)}{f\left(y_{i,t}^* \mid \lambda_{i,t}^{m-1}\right)}\}$$

   Since the proposal densities cancel out in the ratio.

5. Draw the diagonal elements of $\mathbf{\Phi}$ conditional on $\mathbf{\Pi}^{'}$, $\mathbf{\Psi}^{'}$, $\mathrm{B}'$, and $\mathbf{\Lambda}'_{\mathbf{t}}$ $\forall t$.

   Note that the Inverse-Gamma prior is conjugate to the variance of the Normal density. Therefore, it can be shown that the conditional posterior of $\varphi_i$ is also Inverse-Gamma:

$$f(\varphi_i|\lambda_i) \propto h\left(\lambda_i \mid \varphi_i\right) p(\varphi_i) \propto$$

$$\prod_{t=1}^{T} \frac{1}{\varphi_i^{0.5}} \exp\left\{-\frac{\left(ln\left(\lambda_{i,t}\right) - ln\left(\lambda_{i,t-1}\right)\right)^2}{2\varphi_i}\right\} \times \varphi_i^{-\left(\frac{\gamma}{2}+1\right)} e^{-\frac{\delta}{2\varphi_i}}$$

[25] Note that $\lambda_{i,\backslash t}$ denotes all elements of the $\lambda_i$ vector except for the $t^{th}$ element.

Furthermore, the right hand side above is equal to:

$$\varphi_i^{-\left(\frac{\gamma}{2}+1\right)-\frac{T}{2}} \exp\left\{-\frac{\delta}{2\varphi_i} - \frac{1}{2\varphi_i}\sum_{t=1}^{T} ln\left(\frac{\lambda_{i,t}}{\lambda_{i,t-1}}\right)^2\right\}$$

$$= \varphi_i^{-\left(\frac{\gamma+T}{2}+1\right)} \exp\left\{-\frac{\delta + \sum_{t=1}^{T} ln\left(\frac{\lambda_{i,t}}{\lambda_{i,t-1}}\right)^2}{2\varphi_i}\right\}$$

Consequently, assuming identical Inverse-Gamma priors on each $\varphi_i \sim IG(\frac{\gamma}{2}, \frac{\delta}{2})$, we have that the conditional posterior is also Inverse-Gamma, or $IG(\frac{\gamma^*}{2}, \frac{\delta^*}{2})$, s.t.:

$$\gamma^* = \gamma + \mathrm{T}, \;\; \delta^* = \delta + \sum_{t=1}^{T}\left(\ln\left(\frac{\lambda_{i,t}}{\lambda_{i,t-1}}\right)\right)^2$$

*BVAR-SSP-WishartSV Gibbs Estimation Steps:*

1. First we repeat the exact same step as step (1) above, except that this time we replace $\mathbf{\Gamma_t}$=$\mathbf{B}^{-1}\mathbf{\Lambda_t}(\mathbf{B}^{-1})^T = var(\nu_\mathbf{t})$, with $\mathbf{\Sigma_t}$. That is, we no longer condition on $\mathbf{B}$, $\mathbf{\Lambda_t}$ $\forall t$, and $\mathbf{\Phi}$, but rather on $\mathbf{A_k}$ $\forall k = \{1, \ldots, K\}$, $\mathbf{C}$, $v$, and $\mathbf{\Sigma_t}$ $\forall t$.

2. Again the same thing applies: we repeat step (2) above again, except that this time we replace $\mathbf{\Gamma_t}$=$\mathbf{B}^{-1}\mathbf{\Lambda_t}(\mathbf{B}^{-1})^T = var(\nu_\mathbf{t})$, with $\mathbf{\Sigma_t}$.

3. Next we draw the parameters $\mathbf{A_k}$ $\forall k$, $\mathbf{C}$, and $v$ jointly, conditional on $\mathbf{\Pi}$, $\mathbf{\Psi}$, and $\mathbf{\Sigma_t}$ $\forall t$.

   All of the individual elements of the parameter matrices $\mathbf{A_k}$ $\forall k$, $\mathbf{C}$ and $v$ are drawn jointly by a random-walk Metropolis-within-Gibbs step. The joint proposal is multivariate Normal, and I impose MVN priors on both $\mathbf{A_k}$ $\forall k$ and $\mathbf{C}$ and a Gamma prior on $(v - p)$. See the section on priors for more details.

   Further note that the random-walk multivariate Normal proposal is symmetric and conditioned on the last value in the process through its mean vector; therefore it drops out of the acceptance ratio. The variance of the proposal is initially set to the inverse of the observed negative Hessian matrix at the mode of the conditional posterior for a first run, and then a second run is employed using the covariance matrix of the initial Markov process draws themselves for improved mixing.

   Moreover, the likelihood of the $InvWishart(K)$ model is now:

$$f(\nu \mid \theta) = L(\theta) = \prod_{t=1}^{\mathrm{T}} f(\nu_\mathbf{t} \mid \mathbf{\Sigma_t})\, g(\mathbf{\Sigma_t} \mid \mathbf{\Sigma_{t-1}}, \ldots, \mathbf{\Sigma_{t-K}}; \theta)$$

$$= \prod_{t=1}^{T} \frac{1}{(2\pi)^{\frac{p}{2}}|\mathbf{\Sigma_t}|^{\frac{1}{2}}} \exp\left\{-\frac{1}{2}\nu_\mathbf{t}^T \mathbf{\Sigma_t}^{-1}\nu_\mathbf{t}\right\} \times$$

$$2^{-\left(\frac{vp}{2}\right)}\left|\mathbf{S_{t-1}}\right|^{\frac{v}{2}}\Gamma_p\left(\frac{v}{2}\right)^{-1}\left|\mathbf{\Sigma_t}\right|^{-(v+p+1)/2}\exp\left\{-\frac{1}{2}tr\left[\mathbf{S_{t-1}\Sigma_t^{-1}}\right]\right\}$$

Note that $\nu$=$\mathbf{\Pi}\left(L\right)\left(\mathbf{y_t}-\mathbf{\Psi d_t}\right)$, so it is a function of the data, $\mathbf{y}$. Therefore, by Bayes Theorem we can consider the conditional posterior of $\theta$ as proportional to the likelihood (which is really a function of the data) times the prior density for $\theta$ (where $\theta=\{\mathbf{A_1},\ldots,\mathbf{A_k},\mathbf{C},v\}$) as follows:[26]

$$p\left(\theta\mid\mathbf{y};\mathbf{\Pi},\mathbf{\Psi},\mathbf{\Sigma}\right)\propto L\left(\theta\right)\pi\left(\theta\right)=f\left(\mathbf{y},\mathbf{\Sigma}\mid\theta;\mathbf{\Pi},\mathbf{\Psi}\right)\pi(\theta)\propto f\left(\nu\mid\theta\right)\pi\left(\theta\right)$$

Therefore, the Metropolis acceptance probability of the random walk sampler can be expressed as:

$$\alpha\left(\theta,\theta'\right)=\min\left\{1,\frac{p\left(\theta'\mid\mathbf{y};\mathbf{\Pi},\mathbf{\Psi},\mathbf{\Sigma}\right)}{p\left(\theta\mid\mathbf{y};\mathbf{\Pi},\mathbf{\Psi},\mathbf{\Sigma}\right)}\right\}$$

4. Finally, in a similar fashion as was employed in step (4) above, we now draw $\mathbf{\Sigma_t}$ conditional on $\mathbf{\Sigma}_{\backslash\mathbf{t}}$, $\mathbf{A'}_{\mathbf{k}}$ $\forall k$, $\mathbf{C}'$, $v'$, $\mathbf{\Pi}'$, and $\mathbf{\Psi}'$ for all $'t'$ in sequence. Again we have that:

$$P\left(\mathbf{\Sigma_t}\;\middle|\;\mathbf{\Sigma}_{\backslash\mathbf{t}},\nu\right)\propto P\left(\nu_{\mathbf{t}}\mid\mathbf{\Sigma_t}\right)P\left(\mathbf{\Sigma_t}\mid\mathbf{\Sigma_{t-1}}\right)P\left(\mathbf{\Sigma_{t+1}}\mid\mathbf{\Sigma_t}\right)\propto$$

$$\left|\mathbf{\Sigma_t}\right|^{-\frac{1}{2}}\left|\mathbf{S_t}\right|^{\frac{v}{2}}\left|\mathbf{\Sigma_t}\right|^{-(v+p+1)/2}\exp\left\{-\frac{1}{2}tr\left[\left(\mathbf{S_{t-1}}+\nu_{\mathbf{t}}\nu_{\mathbf{t}}^T\right)\mathbf{\Sigma_t^{-1}}\right]\right\}\exp\left\{-\frac{1}{2}tr\left[\mathbf{S_t\Sigma_{t+1}^{-1}}\right]\right\}$$

$$s.t.\;\;\frac{\mathbf{S_{t-1}}}{v-p-1}=\mathbf{CC}^T+\sum_{k=1}^{K}\mathbf{A}_k\mathbf{\Sigma_{t-k}^{-1}}\mathbf{A}_k^T$$

Therefore, by letting the proposal be Inverse-Wishart $\mathbf{\Sigma_t}\sim IW_p\left(v,\mathbf{S_{t-1}^*}\right)\;\;s.t.\;\;\mathbf{S_{t-1}^*}=\mathbf{S_{t-1}}+\nu_{\mathbf{t}}\nu_{\mathbf{t}}^T$ the proposal drops out of the Metropolis-Hastings ratio. Therefore, we have that the probability of accepting a draw of $\mathbf{\Sigma_t}$ (sequentially for each $'t'$, starting with 1) is:[27]

$$\alpha\left(\mathbf{\Sigma_t},\mathbf{\Sigma_t}'\right)=\min\left\{1,\frac{\left|\mathbf{\Sigma_t}'\right|^{-\frac{1}{2}}\left|\mathbf{S_t}'\right|^{\frac{v}{2}}\exp\left\{-\frac{1}{2}tr\left[\mathbf{S_t}'\mathbf{\Sigma_{t+1}^{-1}}\right]\right\}}{\left|\mathbf{\Sigma_t}\right|^{-\frac{1}{2}}\left|\mathbf{S_t}\right|^{\frac{v}{2}}\exp\left\{-\frac{1}{2}tr\left[\mathbf{S_t\Sigma_{t+1}^{-1}}\right]\right\}}\right\}$$

[26]Note that the lack of a subscript on say, $\nu$ or $\mathbf{\Sigma}$, indicates the set of all $T$ elements.

[27]Note: to avoid numerical problems logs are taken of both the numerator and denominator, then differenced, before finally taking their exponential (i.e. the inverse function of the initial logs taken). This avoids issues when the non-logged function values grow either too large or too small to be machine comparable.

# A.1 Tables and Figures

Table A.1

| Parameter | Pop. values | Estimates | 25% DI | 75% DI | Parameter | Pop. values | Estimates | 25% DI | 75% DI |
|---|---|---|---|---|---|---|---|---|---|
| vech(**C**) | 0.3 | 0.3048067 | 0.2986989 | 0.3106097 | $\boldsymbol{\Psi}$ | 3 | 2.989978 | 2.9817319 | 2.9982527 |
| | 0 | -0.004543 | -0.01307 | 0.0041573 | | 0 | -0.00742 | -0.026974 | 0.0121962 |
| | 0 | 0.0040364 | -0.008375 | 0.0165825 | | 2.5 | 2.5000301 | 2.480028 | 2.5199119 |
| | 0 | -0.004546 | -0.017613 | 0.0091572 | | 0 | 0.00597 | -0.015043 | 0.0268138 |
| | 0.3 | 0.3080188 | 0.3000249 | 0.3160946 | | | | | |
| | 0 | -0.005497 | -0.017898 | 0.0071127 | | | | | |
| | 0 | 0.017993 | 0.0058068 | 0.0299014 | | | | | |
| | 0.3 | 0.3064622 | 0.2967589 | 0.3155485 | | | | | |
| | 0 | 0.0059086 | -0.009476 | 0.0205295 | | | | | |
| | 0.3 | 0.3144775 | 0.3007619 | 0.328237 | | | | | |
| vec(**A_1**) | 0.5 | 0.4958902 | 0.4816772 | 0.5102997 | vec($\boldsymbol{\Pi}$**_1**) | 0.2 | 0.203067 | 0.1818246 | 0.2242618 |
| | 0 | 0.0024868 | -0.012623 | 0.0173502 | | 0 | 0.0191343 | -0.008198 | 0.0465216 |
| | 0 | 0.0049993 | -0.009395 | 0.0191631 | | 0 | 0.0092204 | -0.024373 | 0.0426402 |
| | 0 | 0.0007086 | -0.014299 | 0.0160316 | | 0 | 0.0785083 | 0.0170608 | 0.1402327 |
| | 0 | 0.0051099 | -0.00988 | 0.0207817 | | 0 | 0.0224101 | 0.0059593 | 0.0389707 |
| | 0.75 | 0.7422374 | 0.7284686 | 0.7558432 | | 0.8 | 0.8377045 | 0.815873 | 0.8595974 |
| | 0 | 0.0027259 | -0.011197 | 0.0163227 | | 0 | 0.0310997 | 0.0049297 | 0.0574669 |
| | 0 | 0.0083849 | -0.005988 | 0.0228076 | | 0 | 0.0045642 | -0.044952 | 0.05423 |
| | 0 | -0.005162 | -0.017116 | 0.0062144 | | 0 | 0.0275359 | 0.0144587 | 0.0405824 |
| | 0 | 0.0029687 | -0.008882 | 0.0143381 | | 0 | 0.0052861 | -0.011677 | 0.0222193 |
| | 0.85 | 0.845122 | 0.8344947 | 0.8563897 | | 0.8 | 0.7477523 | 0.7260779 | 0.7696097 |
| | 0 | 0.0013346 | -0.011855 | 0.0149829 | | 0 | 0.0295836 | -0.009455 | 0.0683902 |
| | 0 | -0.004354 | -0.009064 | 0.0004438 | | 0 | -0.00732 | -0.013495 | -0.001152 |
| | 0 | -0.007785 | -0.012145 | -0.003305 | | 0 | 0.0120076 | 0.0039424 | 0.0200806 |
| | 0 | -0.001321 | -0.005601 | 0.0030164 | | 0 | -0.010948 | -0.02075 | -0.001181 |
| | 0.98 | 0.9731136 | 0.9682256 | 0.9786167 | | 0.8 | 0.8322424 | 0.8100259 | 0.8543505 |
| vec(**A_2**) | 0 | 0.0181686 | 0.0072719 | 0.0262129 | vec($\boldsymbol{\Pi}$**_2**) | 0 | -0.035788 | -0.057587 | -0.014035 |
| | 0 | -0.000187 | -0.015698 | 0.0154666 | | 0 | -0.048275 | -0.075841 | -0.020669 |
| | 0 | 0.0021431 | -0.012429 | 0.0168972 | | 0 | 0.0023807 | -0.032155 | 0.0368863 |
| | 0 | -0.000201 | -0.015658 | 0.0147217 | | 0 | -0.181691 | -0.245393 | -0.117468 |
| | 0 | 0.0001618 | -0.014322 | 0.0148431 | | 0 | 0.0039256 | -0.01715 | 0.0249879 |
| | 0 | -0.000836 | -0.014431 | 0.0144369 | | 0 | -0.084736 | -0.11276 | -0.05653 |
| | 0 | -0.000973 | -0.016262 | 0.0149579 | | 0 | 0.0102287 | -0.023407 | 0.0438769 |
| | 0 | -0.001339 | -0.01691 | 0.0145956 | | 0 | -0.097716 | -0.161552 | -0.033722 |
| | 0 | 0.0013751 | -0.012526 | 0.0148502 | | 0 | -0.021575 | -0.037903 | -0.005195 |
| | 0 | -0.004763 | -0.019613 | 0.0092085 | | 0 | 0.007884 | -0.013093 | 0.0288199 |
| | 0 | 0.0040955 | -0.010645 | 0.0199832 | | 0 | 0.0573499 | 0.0303681 | 0.0842961 |
| | 0 | 0.0031088 | -0.013321 | 0.0197369 | | 0 | -0.000885 | -0.049737 | 0.0482779 |
| | | | | | | | | | |
| | 0 | 0.0001272 | -0.013017 | 0.0125786 | | 0 | 0.015079 | 0.0071316 | 0.0230623 |
| | 0 | 0.0015171 | -0.015073 | 0.0177086 | | 0 | -0.001201 | -0.011513 | 0.0091777 |
| | 0 | 0.0028091 | -0.010796 | 0.0165486 | | 0 | 0.0145664 | 0.0020213 | 0.0271719 |
| | 0 | -0.000922 | -0.01551 | 0.0137548 | | 0 | -0.062941 | -0.09114 | -0.034784 |
| vec(**A_3**) | 0 | 0.0169788 | 0.0066137 | 0.0245011 | vec($\boldsymbol{\Pi}$**_3**) | 0 | -0.016282 | -0.037365 | 0.0048585 |
| | 0 | 0.0018876 | -0.013159 | 0.0170465 | | 0 | 0.003695 | -0.02342 | 0.0308181 |
| | 0 | 0.0032602 | -0.010407 | 0.017481 | | 0 | 0.0083428 | -0.025006 | 0.0416217 |
| | 0 | 0.0031293 | -0.012355 | 0.0180682 | | 0 | 0.194614 | 0.1328817 | 0.2564857 |
| | 0 | -0.002463 | -0.017236 | 0.0123474 | | 0 | -0.010037 | -0.026447 | 0.0062422 |
| | 0 | 7.779E-05 | -0.014382 | 0.015211 | | 0 | 0.0398893 | 0.0184721 | 0.0611843 |
| | 0 | 0.002959 | -0.013434 | 0.0184915 | | 0 | -0.051051 | -0.076812 | -0.025254 |
| | 0 | -0.000346 | -0.014503 | 0.0144719 | | 0 | 0.0764383 | 0.0268855 | 0.1258434 |
| | 0 | -0.005501 | -0.021799 | 0.0110332 | | 0 | 0.0069021 | -0.006159 | 0.0200284 |
| | 0 | -0.001555 | -0.015631 | 0.0126231 | | 0 | -0.011329 | -0.02822 | 0.0054462 |
| | 0 | 0.0010484 | -0.013776 | 0.0160192 | | 0 | -0.008135 | -0.02934 | 0.0130758 |
| | 0 | -0.004589 | -0.019862 | 0.0118073 | | 0 | -0.00101 | -0.040783 | 0.0384009 |
| | 0 | -0.003385 | -0.016385 | 0.0092182 | | 0 | -0.006783 | -0.012892 | -0.000671 |
| | 0 | 0.0016518 | -0.013823 | 0.016844 | | 0 | -0.004481 | -0.012473 | 0.0035003 |
| | 0 | 0.0005964 | -0.014179 | 0.0155372 | | 0 | -0.001215 | -0.010827 | 0.0083997 |
| | 0 | 0.0010657 | -0.014317 | 0.0164658 | | 0 | 0.0407181 | 0.0197003 | 0.0618022 |
| v | 30 | 25.486853 | 23.775212 | 27.257565 | | | | | |

Figure 17: *Clark data*

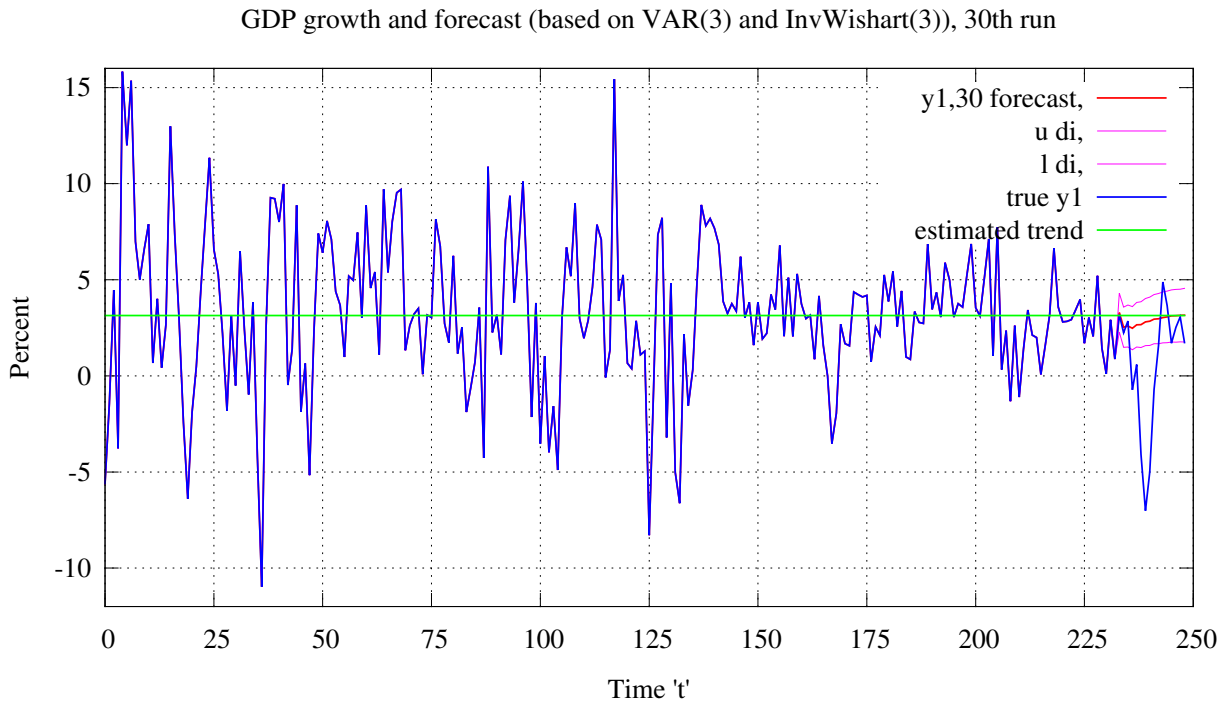



Figure 18: *Clark data*

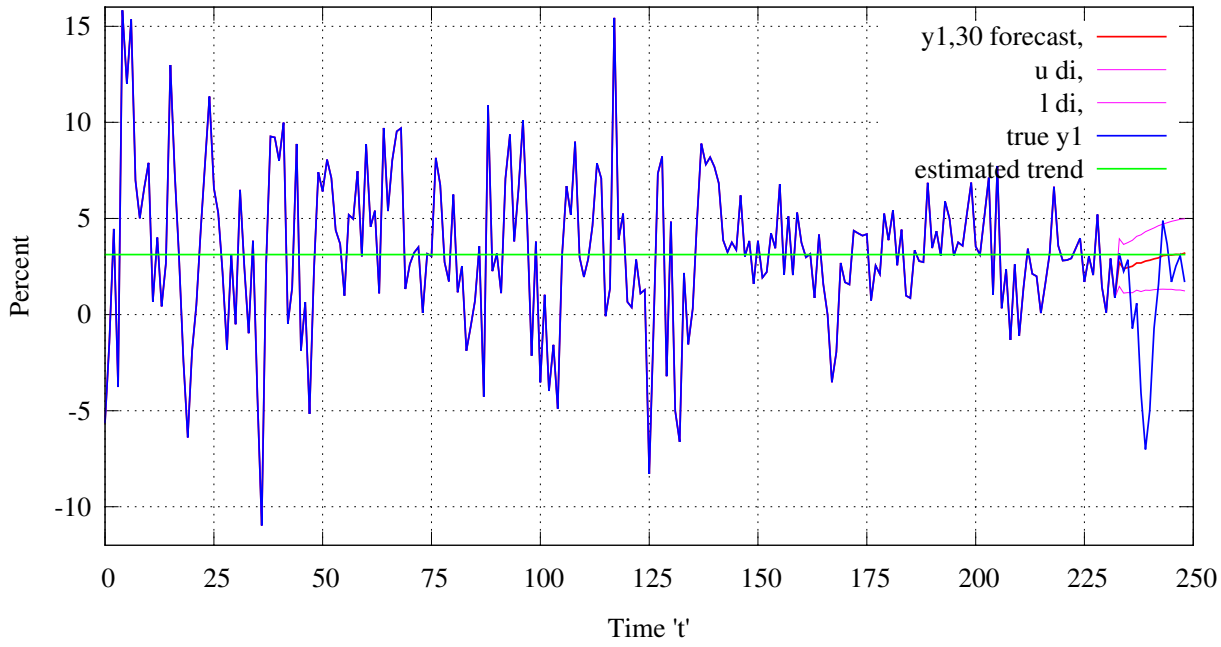


Figure 19: *Clark data*

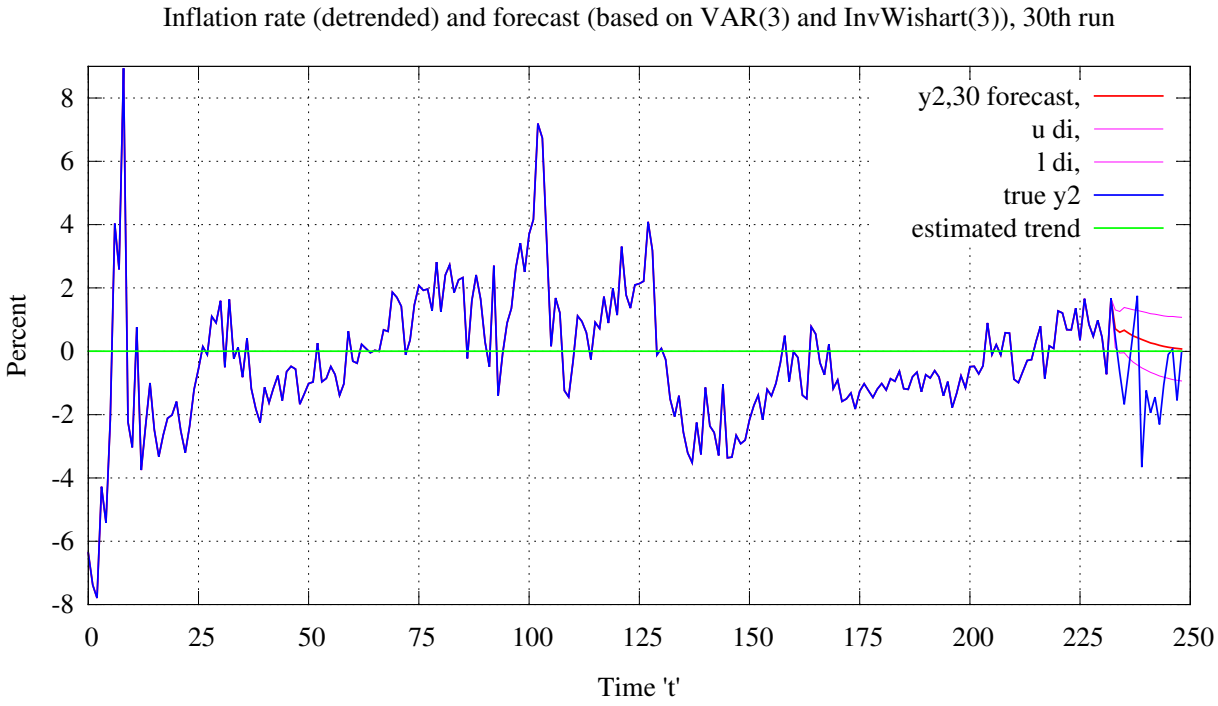



Figure 20: *Clark data*

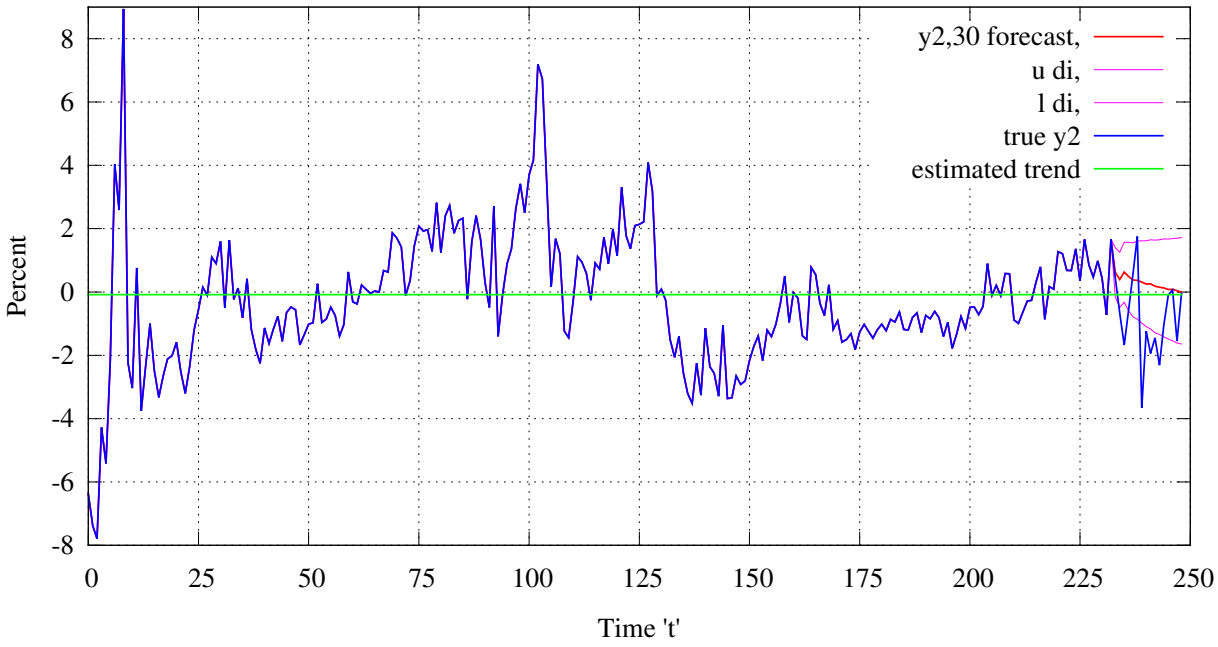


Figure 21: *Clark data*

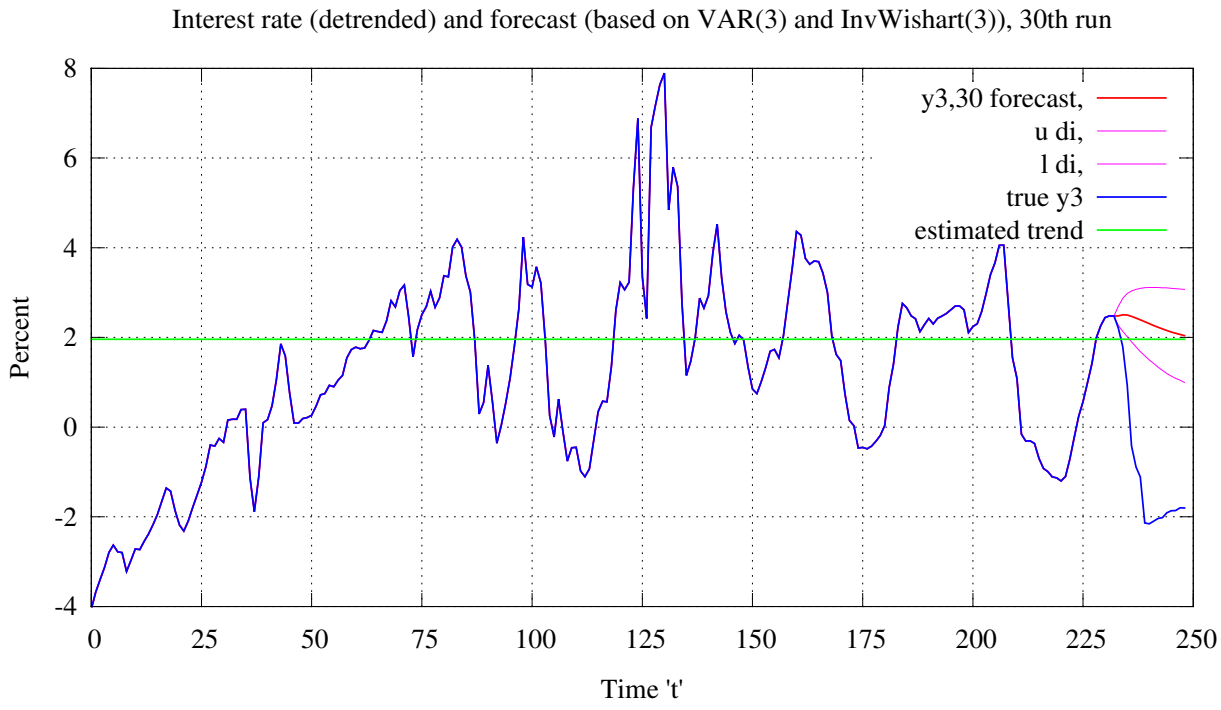



Figure 22: *Clark data*

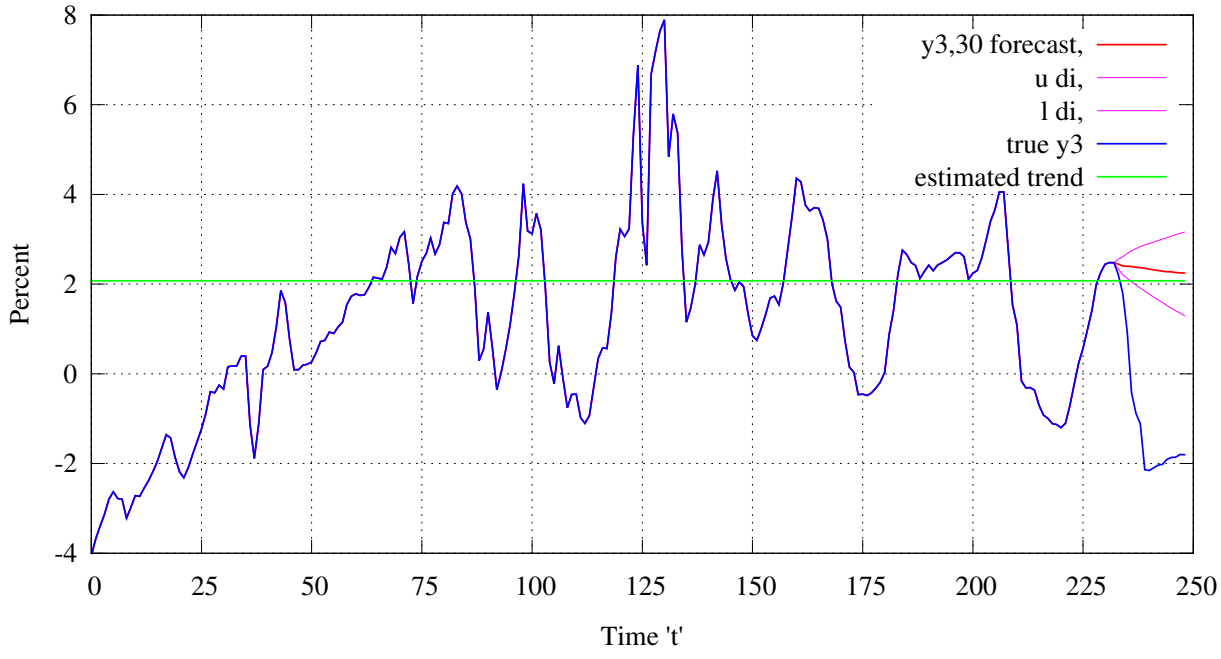


Figure 23: *Clark data*

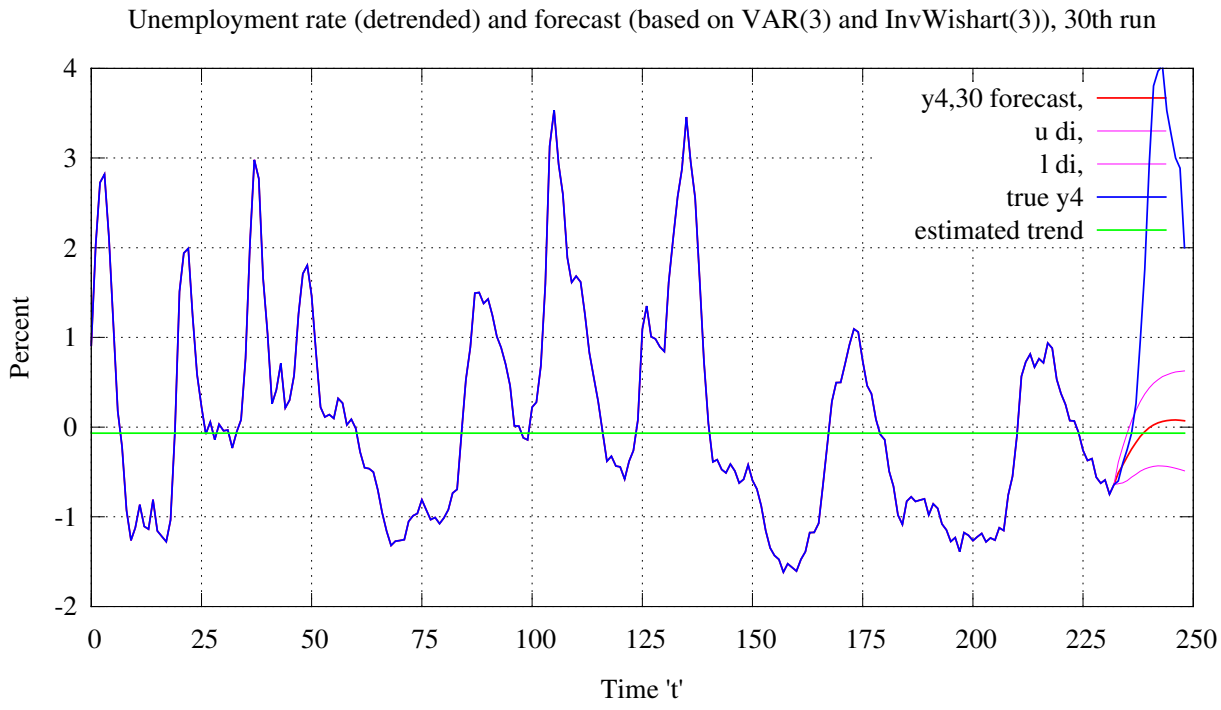
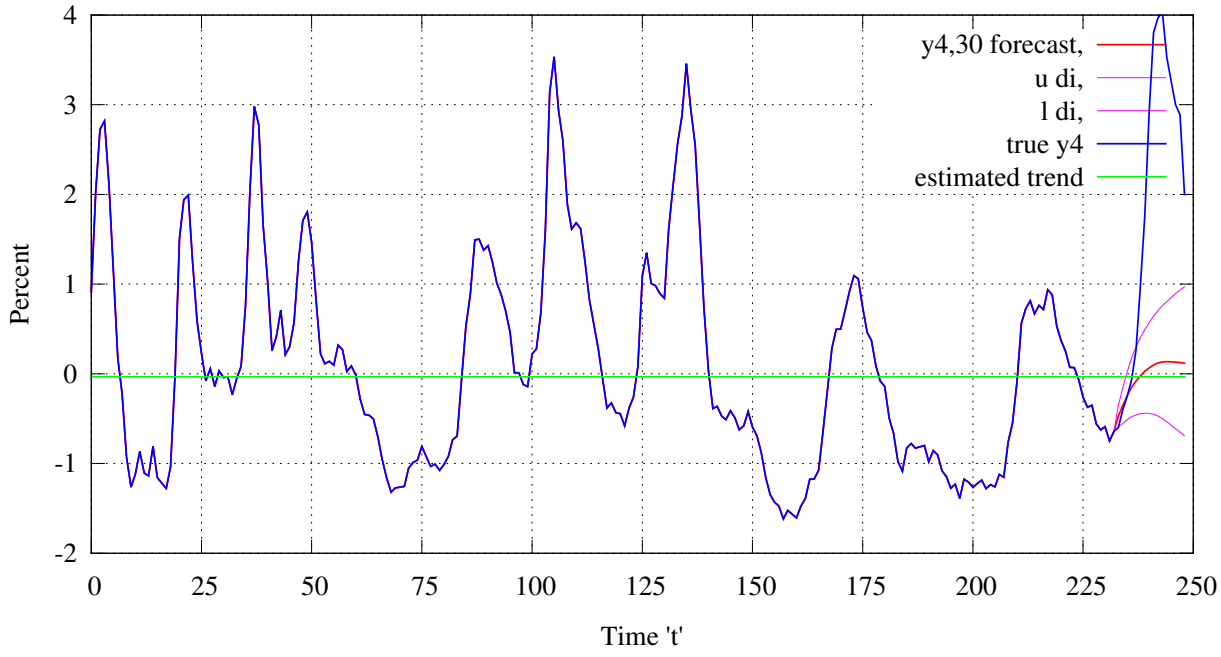



Figure 24: *Clark data*



## A.2 References


Bernanke, B., Mihov, I. (1998a) "The liquidity effect and long-run neutrality," *Carnegie-Rochester Conference Series on Public Policy*, 49, 149-194

————————- (1998b) "Measuring monetary policy." *Quarterly Journal of Economics,* 113, 869-902

Chib, S., Omori, Y., Asai, M., (2009) "Multivariate Stochastic Volatility," in *Handbook of Financial Time Series.,* ed. Anderson, T.G., et al., Springer-Verlag Publishing: Berlin

Clark, T.E. (2011) "Real-time density forecasts from Bayesian vector autoregressions with stochastic volatility," *Journal of Business & Economic Statistics,* 29, 3

Clark, T.E., McCracken, M.W. (2001) "Tests of Equal Forecast Accuracy and Encompassing for Nested Models," *Journal of Econometrics,* 105, 85-110

————————- (2008) "Forecasting with small macroeconomic VARs in the presence of instability", in *Forecasting in the Presence of Structural Breaks and Model Uncertainty,* eds. Rapach, D.E., Wohar, M.E., Bingley. Emerald Publishing: U.K.

Clements, M.P., Hendry, D.F., (1998) *Forecasting Economic Time Series*. Cambridge University Press: Cambridge, U.K.

Cogley, T., Sargent, T.J. (2001) "Evolving post-World War II U.S. inflation dynamics," *NBER Macroeconomics Annual*, 16, 331-373

————————- (2005) "Drifts and volatilities: monetary policies and outcomes in the post-World War II U.S.," *Review of Economic Dynamics*, 8, 262-302

Croushore, D., (2006) "Forecasting with real-time macroeconomic data," in *Handbook of Economic Forecasting*, eds. Elliott, G., Granger, C., Timmermann, A., North-Holland Publishers: Amsterdam

Diebold, F.X., Yilmaz, K., (2007) "Measuring financial asset return and volatility spillovers, with application to global equity markets," *NBER Working Paper*

Engle, R.F., Kroner, K.F., (1995) "Multivariate simultaneous generalized ARCH," *Economic Theory*, 11, 122-150

Fox, E.B., West, M., (2011) "Autoregressive Models for Variance Matrices: Stationary Inverse Wishart Processes," *arXiv:*1107.5239v1

Geweke, J., Amisano, G., (2010) "Comparing and evaluating Bayesian predictive distributions of asset returns," *International Journal of Forecasting*, 26, 216-230

Gourieroux, C., Jasiak, J., Sufana, R., (2009) "The Wishart autoregressive process of multivariate stochastic volatility," *Journal of Econometrics,* 150, 167-181

Golosnoy, V., Gribisch, B., Liesenfeld, R., (2010) "The conditional autoregressive Wishart model for multivariate stock market volatility," *Working Paper: Christian-Albrechts-Universitat zu Kiel,* 07

Greenburg, E., (2008) *Introduction to Bayesian Econometrics,* Cambridge University Press: Cambridge, U.K.

Hamilton, J.D., (1989) "A new approach to the economic analysis of nonstationary time series and the business cycle," *Econometric,* 57, 357-384

Jore, A.S., Mitchell, J., Vahey, S.P., (2010) "Combining forecast densities from VARs with uncertain instabilities," *Journal of Applied Econometrics,* 25, 621-634

Kim, C.J., Nelson, C.R., (1999) "Has the US economy become more stable? A Bayesian approach based on a Markov switching model of the business cycle," *Review of Economics and Statistics,* 81, 608-661

Kozicki, S., Tinsley, P.A., (2001a) "Shifting endpoints in the term structure of interest rates," *Journal of Monetary Economics,* 47, 613-652

———————- (2001b) "Term structure views of monetary policy under alternative models of agent expectations," *Journal of Economic Dynamics and Control,* 25, 149-184

Litterman, R.B., (1986) "Forecasting with Bayesian vector autoregressions: Five years of experience," *Journal of Business and Economic Statistics,* 4, 25-38

Magnus, J.R., Neudecker, H., (1980) "The elimination matrix: some lemmas and applications," *SIAM Journal on Algebraic and Discrete Methods,* 1, 422-449

McConnell, M., Perez Quiros, G., (2000) "Output fluctuations in the United States: What has changed since the early 1980s?," *American Economic Review*, 90, 1464-1476

Philipov, A., Glickman, M.E., (2006) "Multivariate stochastic volatility via Wishart processes," *Journal of Business and Economic Statistics,* 24, 3

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

Rinnergschwentner, W., Tappeiner, G., Walde, J. (2011) "Multivariate stochastic volatility via Wishart processes – A continuation," *Working Papers in Economics and Statistics: University of Innsbruck,* 19

Romer, C.D., Romer, D.H., (2000) "Federal reserve information and the behaviour of interest rates," *American Economic Review,* 90, 429-457

Sartore, D., Billio, M., (2005) "Stochastic volatility models: A survey with applications to option pricing and Value at Risk," in *Applied Quantitative Methods for Trading and Investment,* eds. Dunis, C.L., Laws, J., Naim, P., John Wiley & Sons Publishers

Sims, C.A., (2001) "Comment on Sargent and Cogley's 'Evolving post-World War II US inflation dynamics'," *NBER Macroeconomics Annual,* 16, 373-379

———————- (2002) "The role of models and probabilities in the monetary policy process," *Brookings Papers on Economic Activity,* 2, 1-40

Stock, J.H., (2001) "Discussion of Cogley and Sargent 'Evolving post-World War II US inflation dynamics'," *NBER Macroeconomics Annual*, 16, 379-387

Villani, M., (2009) "Steady-state priors for vector autoregressions," *Journal of Applied Econometrics*, 24, 630-650