Vol. 1 (2018) 1–38
DOI: 0000

Parsimony inducing priors for large scale state-space models

Hedibert F. Lopes[1], **Robert E. McCulloch**[2] **and Ruey S. Tsay**[3]

[1] *INSPER Institute of Education and Research, Rua Quatá 300, São Paulo, SP 04546-042, Brazil. e-mail:* hedibertFL@insper.edu.br.

[2] *Arizona State University, 901 South Palm Walk, Tempe, AZ 85281-1804, USA. e-mail:* robert.mcculloch@asu.edu.

[3] *The University of Chicago Booth School of Business, 5807 S. Woodlawn Ave, Chicago, IL 60637, USA. e-mail:* ruey.tsay@chicagobooth.edu.

Abstract: State-space models are commonly used in the engineering, economic, and statistical literatures. They are flexible and encompass many well-known statistical models, including random coefficient autoregressive models and dynamic factor models. Bayesian analysis of state-space models has attracted much interest in recent years. However, for large scale models, prior specification becomes a challenging issue in Bayesian inference. In this paper, we propose a flexible prior for state-space models. The proposed prior is a mixture of four commonly entertained models, yet achieves parsimony in high-dimensional systems. Here "parsimony" is represented by the idea that in a large system, some states may not be time-varying. Simulation and simple examples are used throughout to demonstrate the performance of the proposed prior. As an application, we consider the time-varying conditional covariance matrices of daily log returns of the components of the S&P 100 index, leading to a state-space model with roughly five thousand time-varying states. Our model for this large system enables us to use parallel computing.

Keywords and phrases: Bayesian modeling, Conditional Heteroscedasticity, Forward Filtering and Backward Sampling, Parallel Computing, Prior, Random walk.

1. Introduction

State-space models, also known as *dynamic models*, are well established in many scientific areas ranging from signal processing to spatio-temporal modeling to marketing applications, to name only a few. See Migon et al. (2005) for a recent review of dynamic models. In the recent business and economics literature, these state-space structures have gained additional attention, particularly in macroeconomic and financial applications where they are used, respectively, when describing time-varying parameters (TVP) in vector autoregressive (VAR) models Primiceri (2005) or in large-scale dynamic factor models (DFM); and time-varying variances and covariances in stochastic volatility (SV) models Lopes and Polson (2010).

More specifically, the basic dynamics governing the state-space component, namely s_t, which can be a time-varying coefficient in the VAR model, a log-volatility in a SV model or a time-varying loading in a DFM, resembles a first

order autoregressive, AR(1), model,

$$s_t = \alpha + \beta\, s_{t-1} + \tau\varepsilon_t,$$

where the errors $\varepsilon_1, \ldots, \varepsilon_T$ are independent and identically distributed, usually standard normals. Primiceri (2005), for example, models the US economy with a trivariate TVP-VAR model containing inflation rate, unemployment rate and short-term interest rate. The paper assumes $(\alpha, \beta) = (0, 1)$ when modeling the time-varying coefficients of the VAR model, giving random walk dynamics. Similarly, the conditionally conjugate normal-inverse gamma prior, commonly used in the Bayesian state-space literature to model (α, β, τ^2), fails to properly account for common parsimonious/shrinkage cases. When a state s_t is close to time invariant, this AR(1) model is not easy to identify, because it can be represented in two ways. First, $\beta = 0$ and α is close to the constant state value. Second, $\alpha = 0$ and $\beta = 1$. In both cases, τ can be small and hard to deal with. For example, in a large system, some regression coefficients are likely to be close to zero for all time. Further discussion concerning identifiability is given below.

We argue, and show in our applications, that limiting the evolution of state-space components to a random walk and/or using conditionally conjugate normal-inverse gamma priors for (α, β, τ^2) are both unrealistic practices (see, for instance, Frühwirth-Schnatter (2004)). This is particularly troubling when dealing with large-scale systems where several hundreds, or thousands, of coefficient are essentially flat-line, rendering the random walk hypothesis meaningless. One of our main goals, extensively discussed in Section 2, is to propose a general mixture prior structure that allows us to entertain and investigate different kinds of state evolution within the simple AR(1) framework. More specifically, we will focus our attention on parsimonious/shrinkage cases, such as $(\alpha, \beta) = (0, 1)$ (random walk component), $\beta = 0$ (flat-lined component), $\alpha = \beta = 0$ (irrelevant state-space component) and $0 < \beta < 1$ (stationary component). Our mixture prior probability implicitly addresses the identifiability mentioned earlier.

Another major contribution of the paper is the modeling of time-varying covariance matrices in large-scale financial time series of log-returns, where the above-mentioned parsimonious prior structure will play a major regularizing role by shrinking unnecessary (or flat-line) coefficients toward zero (or toward constants). More specifically, we will rewrite the time-varying covariances Σ_t of the multivariate normal log-returns via a Cholesky transformation $\Sigma_t = A_t H_t A_t'$ and, in turn, model the recursive conditional regression coefficients in the lower-triangular matrix A_t and the log conditional variances from the diagonal matrix H_t, both with the above state-space AR(1) structure and mixture prior. Section 3 provides extensive details regarding this Cholesky stochastic volatility (CSV) structure along with a customized MCMC scheme for posterior Bayesian inference that takes advantage of the parallel nature of the CSV model.

We illustrate our approach by a number of real and synthetic examples, including a real application on the estimation of log-volatilities in a state-space model based on realized volatilities and a CSV model with 94 financial time series from components of the S&P100 index.

2. Prior Specification for the State Equation

To facilitate discussion, we begin with the univariate state-space model

$$\begin{array}{ll} \text{Observation equation:} & y_t = f(x_t, s_t, \eta_t) \\ \text{State equation:} & s_t = \alpha + \beta\, s_{t-1} + \tau\varepsilon_t, \end{array} \tag{2.1}$$

where s_t is the latent (hidden) state-space variable. η_t and ε_t are independent random shocks in the observation and state equations respectively, usually Gaussian, and we observe the pairs (x_t, y_t), $t = 1, 2, \ldots, T$. In our examples, we will consider two specifications for the observation equation: *i)* $y_t = x_t\, s_t + \eta_t$, a dynamic regression with a time-varying coefficient s_t, and *ii)* $y_t = \exp(s_t/2)\, \eta_t$, a standard stochastic volatility model.

The parameters (α, β, τ) strongly affect the posterior distribution of the state sequence $s = (s_1, s_2, \ldots, s_T)$. A basic observation is that if τ is small then the state sequence evolves smoothly. Consequently, the choice of prior for (α, β, τ) is influential. A basic goal of this paper is to specify a prior on (α, β, τ) that allows us to investigate different kinds of state evolutions within the simple AR(1) framework for the state equation. In addition, we will specify a prior for s_0, the initial state.

2.1. A Mixture Prior for AR Parameters

In this section we present a mixture prior for (α, β, τ). The basic notions our prior must be able to express are *i)* we may want τ small, and *ii)* the following four cases are of particular interest:

Case (1): $(\alpha, \beta) = (0, 1)$ - (random walk component)
Case (2): $\beta = 0$ - (flat-lined component)
Case (3): $(\alpha, \beta) = (0, 0)$ - (irrelevant state-space component)
Case (4): $0 < \beta < 1$ - (stationary component).

Our prior mixes over these four cases. We put zero prior weight on $\beta < 0$. In our applications, we use stock returns and the correlations between them tend to be positive. If we analyze returns of stocks and bond yields jointly, then we might have negative correlations. However, if negative correlations are to be expected, the sign of the data can be changed without affecting the analysis, yet we keep the correlations to be positive. Ultimately, this restriction can be relaxed without affecting the current structure of our mixture prior specification.

Case (1) corresponds to the classic "random-walk" prior. With τ small, this prior succinctly expresses the notion that the state evolves smoothly and may "wander". Many applications assume Case (1). Case (2) says the state simply varies about a fixed level α. With very small τ this is practically equivalent to a fixed value for the state. Case (3) says that the state is fixed near zero, which is often a possibility of particular interest. For example, if the state is a regression coefficient then the corresponding variable has no effect. Case (4) allows the state to vary in a stationary fashion.

A near constant state can be achieved with $(\alpha, \beta) = (0, 1)$ (Case (1)) or $(\alpha, \beta) = (\alpha, 0)$ (Case (2)), given τ small. Depending on the application, the user may choose to weight different mixture components. For example, if we are only extrapolating a few periods ahead, $\beta \approx 1$ may be fine. If, however, we wish to predict farther ahead, we may be more comfortable with $\beta < 1$, if the data allows it.

As usual, the prior allows us to push the inference in desired directions, without imposing it. In Section 3 we consider the problem of modeling high dimensional multivariate stochastic volatility. This large, complex model consists of thousands of univariate state-space models. In this application we found it essential to be able to flexibly consider the possibility that many of the states are constant over time. This leads to more parsimonious representations with time-invariant states greatly simplifying the model. Appropriately mixing over our four cases allows us to push our inference towards these parsimonious representations.

To specify our mixture we need prior probabilities for each of the cases and then a prior for (α, β, τ) given the case. As our four cases delineate, β is the key parameter for determining the state dynamics. Consequently, we specify the joint prior for (α, β, τ) by first choosing a marginal for β and then a conditional for (α, τ) given β. Using the Smith-Gelfand bracket notation we have

$$[\alpha, \beta, \tau] = [\beta]\,[\alpha, \tau \,|\, \beta].$$

All the specifications we consider in this paper make the additional simplifying assumption that τ and α are independent given β: $[\alpha, \tau \,|\, \beta] = [\alpha \,|\, \beta]\,[\tau \,|\, \beta]$.

Let δ_x denote the Dirac measure which assigns probability one to the value x. We use the Dirac measure to identify the special role that the values $\beta = 0$ and $\beta = 1$ play in our four cases. Our full mixture prior has the form

$$\begin{aligned} p(\alpha, \beta, \tau) &= p_{01}\,\delta_{\{\beta=1\}}\,\delta_{\{\alpha=0\}}\,p(\tau \,|\, \beta = 1) \\ &+ p_{00}\,\delta_{\{\beta=0\}}\,\delta_{\{\alpha=0\}}\,p(\tau \,|\, \beta = 0) \\ &+ p_{u0}\,\delta_{\{\beta=0\}}\,p(\alpha \,|\, \beta = 0)\,p(\tau \,|\, \beta = 0) \\ &+ p_{uu}\,p(\beta)\,p(\alpha \,|\, \beta)\,p(\tau \,|\, \beta) \end{aligned}$$

where p_{01}, p_{00}, p_{u0}, and p_{uu} are the mixture weights of our four components. p_{01} is the probability that $(\alpha, \beta) = (0, 1)$, p_{00} is the probability that $(\alpha, \beta) = (0, 0)$, p_{u0} is the probability that $\beta = 0$ and α is unrestricted, and p_{uu} is the probability that $\beta \in (0, 1)$ and α is unrestricted.

To specify the prior $p(\beta)$ for $\beta \in (0, 1)$ (used in our fourth "uu" component above) we use a normal distribution restricted to the interval $(0, 1)$:

$$p(\beta) \propto n(\beta \,|\, \bar{\beta}, \sigma_\beta^2)\,\chi_{(0,1)}(\beta), \tag{2.2}$$

where $n(\cdot \,|\, \bar{\beta}, \sigma_\beta^2)$ denotes a normal density with mean $\bar{\beta}$ and standard deviation σ_β and $\chi_{(0,1)}(\beta)$ is one for β in $(0, 1)$ and zero otherwise.

For the prior $p(\alpha \,|\, \beta)$ we use,

$$\alpha \,|\, \beta \sim N(0, \sigma_\alpha^2\,(1 - \beta^2)). \tag{2.3}$$

When $\beta = 0$ we simply have $\alpha \sim N(0, \sigma_\alpha^2)$. As β increases, we shrink our prior down towards the case where $\alpha = 0$ at $\beta = 1$.

Our mixture prior enables us to incorporate the special role of β in the AR(1) state equation. The parameter τ also plays a crucial role. We have developed a form of prior for τ that allows us to shrink towards small values but still have a right tail that allows for larger values. This prior is discussed in detail in Section 2.1.1.

Note that as soon as you think about what it might mean for τ to be small, you realize that it depends on β. In particular, the effect of τ depends on whether $\beta = 1$, $\beta = 0$, or $\beta \in (0, 1)$. These considerations make our mixture prior plausibly the minimally complex construction for serious prior thought.

2.1.1. The Prior for τ

We will specify our prior for τ by assuming a finite discrete set of possible values. While the basic idea of the prior could be expressed using a continuous (or mixture of discrete and continuous) distribution, we find it conceptually and computationally convenient to use the discrete construction.

To specify a prior for τ on a grid of values, we first choose minimum and maximum values τ_{min} and τ_{max}. Using n_g grid points, we have evenly spaced values $(t_1, t_2, \ldots, t_{n_g})$ with $t_1 = \tau_{min}$ and $t_{n_g} = \tau_{max}$. We let $P(\tau = \tau_{min}) \equiv p_{min}$. For $i > 1$, $P(\tau = t_i) \propto \exp(-c_\tau \, |t_i - \tau_{min}|)$. Thus, our τ prior has the four hyper-parameters $(\tau_{min}, \tau_{max}, p_{min}, c_\tau)$. Understanding and choosing the hyper-parameters of this prior is quite simple. We pick an interval, and then our choice of c_τ determines the degree to which we push τ towards smaller values.

We have chosen not to consider the case $\tau = 0$. There is no practical advantage in considering τ to be zero as opposed to small. A useful variation on the basic scheme above is to use a non-evenly spaced grid for τ. It might make sense to have the grid tighter for smaller τ. However, we have employed an evenly spaced grid in all our examples.

The commonly used prior for τ is the inverted chi-squared: $\tau^2 \sim \nu \, \lambda / \chi_\nu^2$. We found it very difficult to choose values for ν and λ that gave consistently good results. For small values of ν, the prior is not informative so that we cannot express a preference for smaller values. We can use an informative prior by using big values of ν. But, with large ν, if we choose λ to favor smaller τ, we find that we have too little prior probability attached to the possibility of larger τ. Our prior is designed to favor small τ but allow for large ones in the simplest possible way.

The situation is illustrated in Figure 1. The solid black line is our prior with $\tau_{min} = 0.005$, $\tau_{max} = 0.15$, $p_{min} = 0.3$, $c_\tau = 200$. The other two lines are densities for the inverted chi-squared. The red, short-dashed line has $\nu = 5$ and the blue, long-dashed line has $\nu = 50$. Our discrete τ prior has been scaled to be comparable to the continuous distributions and λ is equal to the square of $E(\tau)$ under our prior. In the left panel we have the densities and in the right

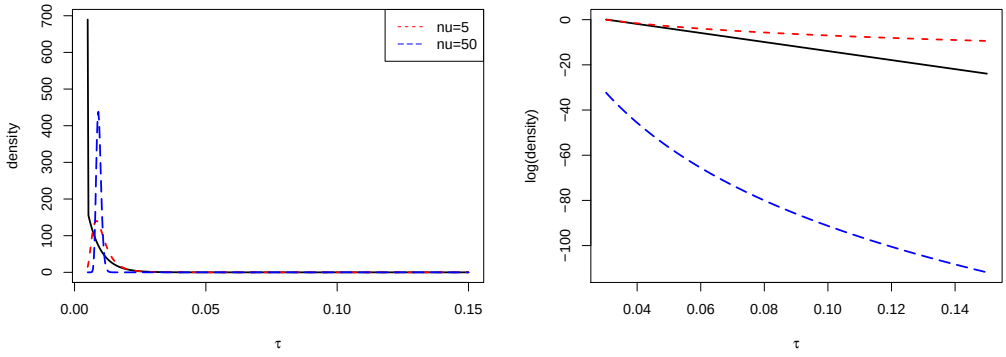

FIG 1. Prior for τ: *Here* $\tau_{min} = 0.005$, $\tau_{max} = 0.15$, $p_{min} = 0.3$, $c_\tau = 200$. *In each panel the black solid line is our* τ *prior and the other two correspond to inverted chi-squared densities with* λ *equal to the square of* $E(\tau)$ *under our prior and* ν *equal to 5 (red, short dash) or 50 (blue, long dash). In the left panel we have the three densities where our discrete distribution has been scaled to be comparable to the continuous distributions. In the right panel we have the log densities for large* τ.

panel we have the log densities for large τ. In the left panel we see that both our prior and the $\nu = 50$ density push hard towards small τ which is what is needed. In the right panel we see that the tail of our prior is more like the tail of the $\nu = 5$ density, so that, if the data demands it, larger values of τ are easily found.

To specify $p(\tau \mid \beta)$ in our general mixture prior we let the parameters τ_{min}, τ_{max}, p_{min} and c_τ depend on β. For example, in our applications we choose a value of c_τ to use for all $\beta > 0$ and then use twice that value when $\beta = 0$. The larger c_τ value allows us to express an even stronger desire for small τ when $\beta = 0$.

2.2. Markov Chain Monte Carlo Implementation

In this section we describe our implementation of a Markov chain Monte Carlo (MCMC) algorithm for drawing the state s and (α, β, τ) in the state-space model given by Equation 2.1. Let $y = (y_1, y_2, \ldots, y_T)$, $x = (x_1, x_2, \ldots, x_T)$, and $s = (s_1, s_2, \ldots, s_T)$. Let s_0 be the initial state. We start by employing the Gibbs sampler:

$$[(s_0, s) \mid (\alpha, \beta, \tau), y, x] \quad \text{and} \quad [(\alpha, \beta, \tau) \mid (s_0, s)]. \tag{2.4}$$

That is, we draw the states given the AR(1) parameters and the AR(1) parameters given the states. For draws of the states we use well known algorithms from the literature (see for example Frühwirth-Schnatter (2004) and Chib et al. (2006)).

Because the likelihood for (α, β, τ) given the states is that of a linear regression, the Gibbs sampler 2.4 allows us to develop a simple approach for the draw $[(\alpha, \beta, \tau) \mid (s_0, s)]$ using our non-conjugate mixture prior. However, this Gibbs sampler has the draw back that it may mix very slowly given the strong dependence between s and (α, β, τ). Our approach in this paper has been to use the Gibbs sampler 2.4 and then thin the draws to reduce dependence. In simple applications, thinning the draws is adequate. In our more complex examples (Section 3), we may simplify our use of the mixture prior by letting some components have zero prior probability. This strong prior information is appropriate in a high dimensional problem and simplifies the inferential complexity. We note however, that in some problems it may be worthwile to consider alternatives to 2.4. For example, in some cases it is possible to analytically or numerically integrate out the states making a direct draw of $[(\alpha, \beta, \tau) \mid y, x]$ possible.

We draw $[(\alpha, \beta, \tau) \mid (s_0, s)]$ jointly by drawing from $[(\beta, \tau) \mid (s_0, s)]$ and then $[\alpha \mid (\beta, \tau), (s_0, s)]$. Given (β, τ), α is either known to be zero or has the normal prior given by 2.3 depending on the mixture component. In the normal prior case, the prior is conditionally conjugate so it is a standard calculation to both integrate out α to obtain a marginal likelihood for the draw of $[(\beta, \tau) \mid (s_0, s)]$ and to draw $[\alpha \mid (\beta, \tau), (s_0, s)]$.

In order to make a joint draw of $[(\beta, \tau) \mid (s_0, s)]$ we must consider our four mixture components which we label 01, 00, $u0$ and uu as in the labeling of our mixture prior probabilities p_{01}, p_{00}, p_{u0}, and p_{uu}.

In the 01 component we know $\alpha = 0$ and $\beta = 1$ and we have a grid of n_g possible τ values with prior probabilities $p(\tau \mid \beta = 1)$. The prior probabilities $p(\tau \mid \beta = 1)$ will come from a choice of $(\tau_{min}, \tau_{max}, p_{min}, c_\tau)$ associated with $\beta = 1$. Each of the n_g grid points will have prior probability $p_{01}\, p(\tau \mid \beta = 1)$. Similarly, in the 00 component we have a set of n_g values of (α, β, τ) each having $\alpha = 0$ and $\beta = 1$ and prior probability $p_{00}\, p(\tau \mid \beta = 0)$. These two components gives us $2\, n_g$ values of (α, β, τ). At each of the values we can compute the simple linear regression likelihood resulting from the (s_0, s) state values.

In the $u0$ component, we know $\beta = 0$ and we again have a grid of τ values with prior probabilities $p(\tau \mid \beta = 0)$. In this case we have a $N(0, \sigma_\alpha^2)$ prior for α. Our likelihood for a $(\beta = 0, \tau)$ value is obtained by integrating out α in the regression likelihood.

Finally, we have the uu component in which $\beta \in (0, 1)$ rather than being zero or one. Again, given β and τ we can integrate out α to obtain an integrated likelihood. The integrated likelihood will depend on β in a non-conjugate manner because of the $N(0, \sigma_\alpha^2(1-\beta^2))$ prior in 2.3. We again look for a simple approach and discretize the prior 2.2 by picking n_b equally spaced grid points in $(0, 1)$. At each grid point β_i, $p(\beta_i) \propto n(\beta_i \mid \bar{\beta}, \sigma_\beta^2)$, $i = 1, 2, \ldots, n_b$. Thus in the uu compontent we have $n_g n_b$ possible (β, τ) pairs each having prior $p_{uu}\, p(\beta)\, p(\tau \mid \beta)$.

Combining the four components we have $3n_g + n_g n_b$ possible values of (β, τ). We draw from this discrete distribution. In the 01 and 00 components, α is known. In the other two components α is a draw from the normal given the states, the values of (β, τ), and a normal prior on α (2.3). In many problems this brute force grid approach is unappealing because of the time it takes to

evaluate the likelihood and prior at each grid point. However in our case the computation of likelihood and prior is so simple (given the states) that in our applications we do not incur a computational bottleneck relative to the other computations that are being made.

Note that given a (α, β, τ) value the mixture component can be identified by inspection. For example, if $\beta = 1$ you know you are in the 01 component. In some applications inferring the component is a major goal as it reveals the essential characteristics of the state evolution. Given draws of (α, β, τ), we can compute posterior probabilities of mixture components simply by counting the number of draws in each component. This solves an important and complex problem in a simple way. The drawback again is that the slow mixing of the basic Gibbs sampler 2.4 may necessitate a large number of runs. See Section 2.4.2 for an example of this kind of analysis.

We emphasize that the most crucial aspect of our prior is the prior on τ having the properites illustrated in Figure 1. This prior and the mixture elaboration, were developed in order to deal with the larger problem discussed in Section 3. Initially we tried using the standard inverted chi-squared prior for τ with parameters ν and λ: $\tau^2 \sim \frac{\nu\lambda}{\chi^2_\nu}$. If you run the MCMC with small ν and many states, the lack of prior information will give you signals that cannot be distinguished from noise. With big ν, the MCMC can be deceptive in that in short runs it appears to have converged but in longer runs the right tail of the prior is overcome, and large τ's are drawn. For additional discussion on the inverted chi-squared and its problematic use in state space models, see Frühwirth-Schnatter (2004) and Frühwirth-Schnatter and Wagner (2010).

A less important but still worth noting feature of our approach is the treatment of the initial state s_0. In many applications, zero is a value of particular importance for the state because it represents a model simplification. An important example is that of a time varying regression coefficient. To shrink the initial state s_0 towards zero we use a mixture prior along the lines of that used by George and McCulloch (1993) for variable selection:

$$\begin{aligned} s_0 &\sim \gamma\, N(0, (cw)^2) + (1-\gamma)\, N(0, w^2) \\ \gamma &\sim \text{Bernoulli}(p_*), \end{aligned}$$

where c is a large positive real number, w is small, and p_* is a hyper-parameter denoting the prior knowledge about the initial state. A small p_* favors zero initial state and $p_* = 0.5$ shows no preference. The variable γ is a latent variable. When $\gamma = 0$, the state is shrunk heavily towards zero and when $\gamma = 1$, the state may be large.

The basic Gibbs sampler 2.4 is modified by adding the draw of the latent $[\gamma \,|\, (s_0, s), (\alpha, \beta, \tau), y, x] = [\gamma \,|\, s_0]$ and drawing $[(s_0, s) \,|\, \gamma, (\alpha, \beta, \tau), y, x]$ with the pair of draws $[s \,|\, \gamma, (\alpha, \beta, \tau), y, x]$ and $[s_0 \,|\, s, \gamma, (\alpha, \beta, \tau)] = [s_0 \,|\, s_1, \gamma, (\alpha, \beta, \tau)]$.

Conditional on a draw of γ, we have a normal prior for the initial state with mean zero and standard deviation w in the case $\gamma = 0$ and standard deviation (cw) in the case $\gamma = 1$.

2.3. Issues in Prior Choice

In this section we review the hyperparameter choices associated with our mixture prior. We discuss some of the issues involved and simplifying choices we have used in application.

Perhaps the most basic issue in choosing the prior is that of scale. We have discussed the need to allow for small values of τ but the meaning of "small" depends of the scale (units) of the observed y and the relationship between y and the state s is the defined by the observation equation. While in any particular application there is no real substitute for careful thought about the prior, we have found it useful to simplify things in two ways.

First, we typically standardize y to have sample mean zero and sample stardard deviation one. This is a common practice in statistics (e.g the very popular `glmnet` R package defaults to standardize = TRUE). If y has outliers or extreme skewness, this can be inappropriate, but it typically put things in a reasonable "ballpark".

Second, to specify $p(\tau \mid \beta)$ we consider only the two case $\beta = 0$ and $\beta \in (0, 1]$. We choose values of $(\tau_{min}, \tau_{max}, p_{min}, c_\tau)$ to use for all $\beta \in (0, 1]$ and a different set to be used when $\beta = 0$. We keep τ_{max} and p_{min} the same in both cases, but use c_τ and τ_{min} when $\beta \in (0, 1]$ (the 01 and uu mixture components) and use c_τ^0 and τ_{min}^0 when $\beta = 0$ (the 00 and $u0$ components). The choice of this simplication was driven by our application in Section 3 where we wanted to push things towards a near constant state when $\beta = 0$. In some applications it might also make sense to pay particular attention to the $\beta = 1$ case and our general prior construction would facilitate this.

In summary, the hyperparameters we consider in our applications are *i)* p_{01}, p_{00}, p_{u0} and p_{uu} (for p), *ii)* τ_{min}, τ_{min}^0, τ_{max} p_{min}, c_τ and c_τ^0 (for τ), *iii)* σ_α (for α), *iv)* $\bar{\beta}$ and σ_β (for β), and *v)* p_*, w and c (for s_0). Additionally, one just chooses the grid sizes for both τ and β. We have use $n_g = n_b = 100$ throughout and these choices seem to give us a fine enough inference without taking too much time.

Specific hyper-parameter values we will use in some examples are given by *i)* $p_{01} = 0.5$, $p_{00} = 0.15$, $p_{u0} = 0.15$ and $p_{uu} = 0.2$ (for p), *ii)* $\tau_{min} = 0.005$, $\tau_{min}^0 = 0.001$, $\tau_{max} = 0.15$, $p_{min} = 0.5$, $c_\tau = 100$, and $c_\tau^0 = 200$ (for τ), *iii)* $\sigma_\alpha = 2.0$ (for α), *iv)* $\bar{\beta} = 1$ and $\sigma_\beta = 1$ (for β), and *v)* $p_* = 0.5$, $w = 0.1$ and $c = 10$ (for s_0). This prior suggests smaller τ when $\beta = 0$ which effectively leads to a constant state around α. We have a 50% prior probability of the random walk prior and 30% chance of a constant state. There is a 20% that we have a time-varying stationary state.

Appendix A contains a summary of our `R` routine `csv` and two default prior specifications. The above specification corresponds to a default *rougher* prior and is denoted by `defpri=0`. The name *rougher* refers to the fact that in our applications this prior allows for substantial state variation. We shall also use a default *smoother* prior which will result in inferring a smoother state by using larger c_τ ($c_\tau = 200, c_\tau^0 = 400$) and a smaller $\tau_{max} = 0.05$. In addition the smoother prior uses $p_{01} = 0.85$, $p_{00} = 0.05$, $p_{u0} = 0.05$, $p_{uu} = 0.05$, putting

much more weight on the random walk.

Of course, the hyperparameter choices are heavily influenced by our actual application. In other applications, other choices might be considered. Nevertheless, given that we have standardized the data, we hope that they might at least serve as useful starting points.

2.4. Examples

2.4.1. Study 1: DLM with Mixture Prior

In this section we illustrate our prior on (α, β, τ) in the simple normal dynamic linear model (NDLM),

$$\begin{aligned} y_t &= x_t\, s_t + \sigma\, \eta_t, \\ s_t &= \alpha + \beta\, s_{t-1} + \tau\, \varepsilon_t, \end{aligned}$$

where η_t and ε_t are independent and identically distributed $N(0,1)$. We simulate series of length $T = 200$ with $\sigma = 0.1$, $s_0 = 0$, and $x_t \sim N(0,9)$. The state is a time-varying regression coefficient.

We employ the *rougher* prior of Section 2.3. Figure 2 displays the prior. The top two panels are kernel density approximations of the marginal priors of β and τ based on prior draws from our full mixture prior. The density smooths naturally "jitter" the draws (add a bit of normal noise) so that the marginals from our mixture prior of discrete and continuous distributions can be displayed as a single continuous distribution. The marginal for β displays our preference for expressing a smooth state with either $\beta \approx 0$ or $\beta \approx 1$ with more weight being given to the vicinity of 1. The prior for τ expresses our desire for small values. Again, this is driven by our desire for a smooth state. The two τ modes reflect the choice of a smaller τ_{min} when $\beta = 0$. In this case the two modes are not very separated so this aspect of our prior has little practical effect. If we separated these modes more dramatically, we could use this aspect for our prior to help identify $\beta \approx 0$ versus $\beta \approx 1$ by saying you can only have a really small τ if $\beta \approx 0$. The long right tail of our τ prior allows the data to push the inference towards larger values if needed.

The bottom two panels of Figure 2 display the joint prior of (α, β). The bottom left panel displays contours from a bivariate smooth of draws of (α, β). The bottom right panel is a scatterplot of jittered draws of (α, β). In the bivariate distribution we can see our preference for $(\alpha, \beta) \approx (0, 1)$ or $(\alpha, \beta) \approx (0, 0)$ with more weight given to the first pair of values. As β decreases, the conditional prior for α becomes more spread out. The contours appear to tighten as β approaches 0 because the choice $(\bar{\beta}, \sigma_\beta) = (1.0, 1.0)$ puts weight on larger β.

Figure 3 displays the results from three different data simulations. Each row corresponds to a different simulation scenario. In the first row results are for data simulated with $(\alpha, \beta, \tau) = (0, 1, 0.04)$, in the second row we used $(\alpha, \beta, \tau) = (0, 0.8, 0.1)$, and in the third row we used $(\alpha, \beta, \tau) = (0.5, 0, 0.01)$. Thus, in the first row our state follows a random walk, in the second row the state is time

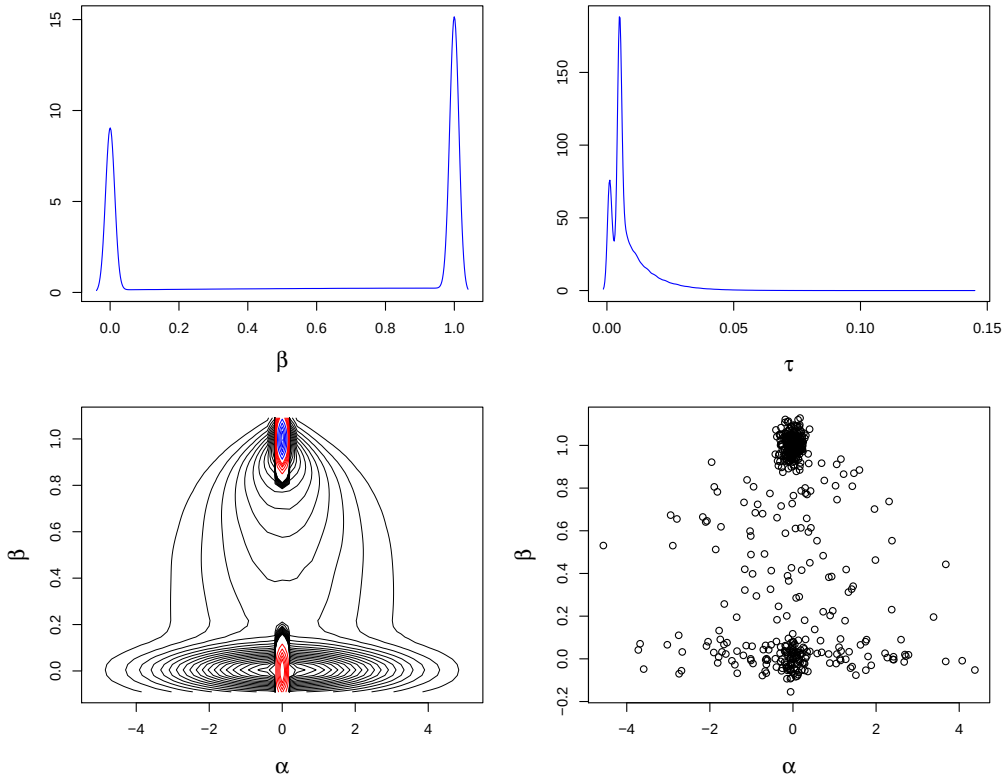

FIG 2. Mixture prior for AR parameters – *This is the* rougher *prior specification of Section 2.3. The top two panels are density smooths of prior draws of β and τ. The bottom left panel displays contours from a bivariate smooth of draws of (α, β). The bottom right panel are jittered draws of (α, β).*

varying but stationary, while in the third row the state is essentially constant at 0.5.

The first column of plots in Figure 3 shows the simulated states (small circles) and the posterior mean of the state draws (with a solid line). In each row we see that the posterior mean nicely smooths the true states and that the essential nature of the state is quite different reflecting our three scenarios. The second, third, and fourth columns of Figure 3 display time series plots of MCMC draws of α, β, and τ respectively. In each plot a dashed horizontal line indicates the true value of the parameter.

In each case the posterior nicely captures the true value. In some cases, our mixture prior has some interesting shrinkage effects. In the case of the first scenario (row 1), virtually all of our posterior draws have $\alpha = 0$ and $\beta = 1$. Thus, while our prior probability of the random walk case was 0.5, our posterior

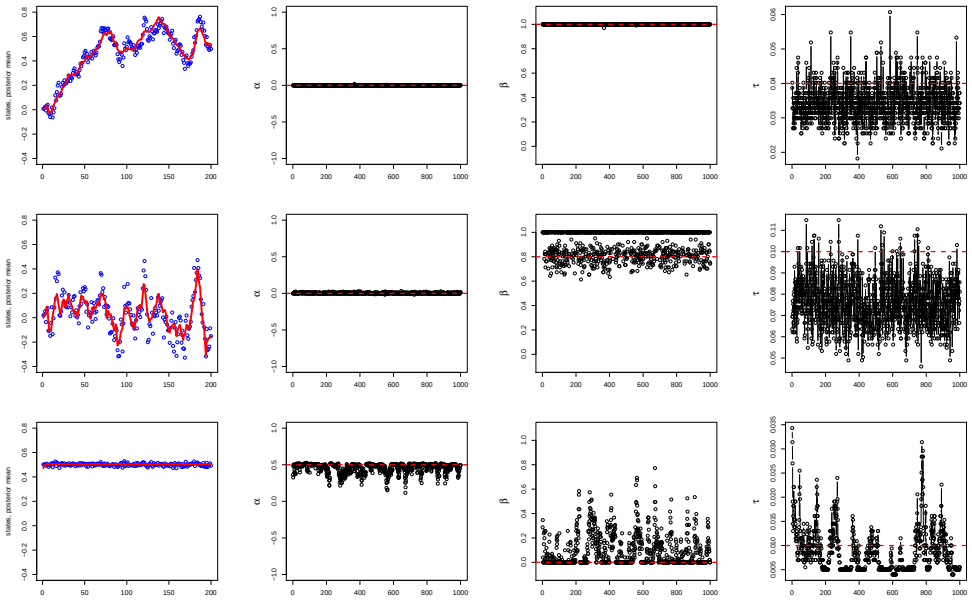

FIG 3. Simulation study of Section 2.4.1 – *Each row shows the results for a different simulation. The plot in the first column shows the simulated states (the plotted points) and the posterior mean (the solid curve). The second, third, and fourth columns of plots show time-series of MCMC draws from the posteriors of α, β, and τ respectively. Dashed horizontal lines represent the true values of (α, β, τ).*

probability is very close to one. For the second data set our posterior for β mixes between the two components $\beta = 1$ and $\beta \in (0, 1)$. The posterior probabilities of the components are 0.55 and 0.45 compared to prior probabilities 0.5 and 0.2. Given the data we are not sure whether β is one or not, but we know it is not zero. In the third scenario our posterior mixes between $\beta = 0, \alpha \approx 0.5$ and $\beta \in (0, 1)$. The posterior probability that $\beta = 0$ and $\alpha \neq 0$ is 0.52 compared with the prior probability of 0.15. The posterior for τ also reflects our prior. In both the second and third scenarios our prior pushes the posterior towards small τ but still covers the true values.

Note that the inference for τ in the third simulation (the (3,4) frame of Figure 3) shows the difficulties our MCMC algorithm may have with mixing. Our inference strongly suggests that τ is small and in practice that is typically all that is needed. However, a more accurate inference would call for a longer thinned run.

2.4.2. Study 2: Realized Volatility with Mixture Prior

In this section we apply our mixture prior to an analysis of the log series of daily realized volatility of Alcoa stock from January 2, 2003 to March 7, 2004 for 340

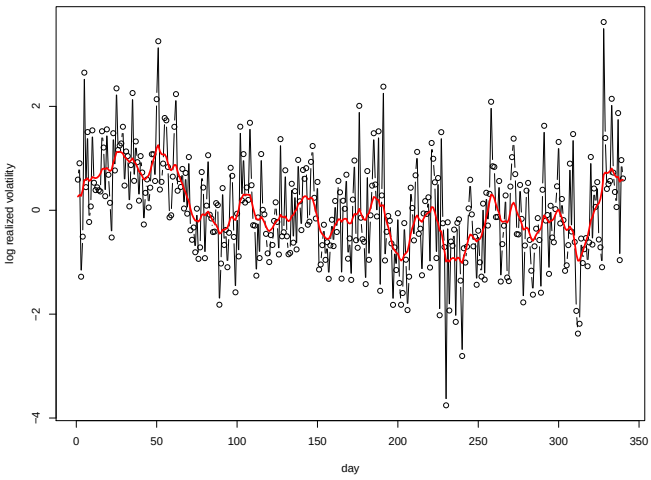

FIG 4. Realized volatility – *Log realized volatility for Alcoa stock (standardized), with the posterior mean of the state.*

observations. The realized volatility is calculated using the intraday 10-minute log returns. It is well known that realized volatility is subject to the impact of market micro-structure noises. See, for instance, Bandi and Russell (2008) and Zhang et al. (2005). Consequently, we entertain the model

$$y_t = s_t + \sigma\eta_t \quad \text{and} \quad s_t = \alpha + \beta\, s_{t-1} + \tau\varepsilon_t,$$

for the log series of Alcoa realized volatility. Here the shock $\sigma\eta_t$ denotes the impact of market micro-structure noises and the state s_t is simply the level of the log volatility. The AR(1) state equation with error term $\tau\varepsilon_t$ allows for volatility to be time-varying.

Our first step in the analysis is to standardize the data to have zero mean and standard deviation one. Given this standardization, we can use the same prior as in Section 2.4.1. Figure 4 is the time series plot of the standardized data and the posterior mean of the state. In this particular case the posterior probability of the mixture component with $\beta = 1$ is 0.99. The prior probability was 0.5. Thus, we have strong support for the random walk state specification, which is in good agreement with empirical characteristics of asset volatility. For example, consider the VIX index of the Chicago Board Options Exchange (CBOE), which is the most widely used daily volatility index in the U.S. The log VIX series fails to reject the null hypothesis of a unit root in hypothesis testing. This analysis of realized volatility demonstrates that the mixture prior of Section 2 can easily produce reasonable results.

3. Cholesky Stochastic Volatility

In this section we show the impact of our mixture prior in a much larger set up where thousands of state variables might evolve over time according to an AR(1) process. More specifically, we are interested in the case where $y_t = (y_{1t}, \ldots, y_{qt})'$ denotes a q-dimensional vector of financial time series observed at time t and consider posterior inference regarding the (possibly large) covariance matrices Σ_t driven by the observation equation:

$$y_t | F_{t-1} \sim N(0, \Sigma_t), \tag{3.1}$$

where F_{t-1} denotes the information available at time $t-1$. Without loss of generality, we assume that any mean structure of y_t has been subtracted out as part of a larger MCMC algorithm.

The main focus is on modeling the dynamic behavior of the conditional covariance matrix Σ_t, which is known as the volatility matrix in finance. Two challenges arise in the multivariate context. Firstly, the number of distinct elements of Σ_t equals $q(q+1)/2$. This quadratic growth has made the modeling Σ_t computationally very expensive and, consequently has created, up to a few years ago, a practical upper bound for q. The vast majority of the papers available in the literature employed a small q or use highly parametrized models to simplify the computation. For instance, Engle (2002) and Tse and Tsui (2002) proposed dynamic conditional correlation (DCC) models where the time evolution of correlations is essentially driven by a pair of parameters. We argue that that such models unrealistically over-simplify the complexity of the covariance dynamics. Secondly, the distinct elements of Σ_t cannot be modeled independently since positive definiteness has to be satisfied. Section 3.1 briefly reviews the literature on multivariate stochastic volatility models, while Section 3.2 introduces our proposed Cholesky stochastic volatility (CSV) model.

3.1. Brief Literature Review

There are at least three ways to decompose the covariance matrix Σ_t.

Correlations and Standard Deviations. In the first case, the covariance matrix is decomposed as

$$\Sigma_t = D_t R_t D_t$$

where D_t is a diagonal matrix with standard deviations, i.e. $D_t = \text{diag}(\sigma_{1t}, \ldots, \sigma_{qt})$ with σ_{it} being the volatility of y_{it}, and R_t is the correlation matrix. The above two challenges remain in this parametrization, i.e. the number of parameters increases with q^2 and R_t has to be positive definite.

Factor Analysis. In the second case, a standard factor model is used to produce

$$\Sigma_t = \beta_t H_t \beta_t' + \Psi_t$$

where β_t is the $q \times k$ matrix of factor loadings and is block lower triangular with diagonal elements equal to one. Ψ_t and H_t are the diagonal covariance matrices of the specific factors and common factors, respectively. This is the *factor stochastic volatility* (FSV) model of Harvey et al. (1994), Pitt and Shephard (1999), Aguilar and West (2000), and, more recently, Lopes and Migon (2002), Chib et al. (2006), Han (2006), Lopes and Carvalho (2007) and Philipov and Glickman (2006a), to name just a few. Philipov and Glickman (2006a) extended the FSV model by allowing H_t to follow a Wishart random process and fit a 2-factor FSV model to the covariance of the returns of $q = 88$ S&P500 companies. Han (2006) fitted a similar FSV model to $q = 36$ CRSP stocks. Chib et al. (2006) analyzed $q = 10$ international weekly stock index returns (see also Nardari and Scruggs (2007)). Lopes and Carvalho (2007) extended the FSV model to allow for Markovian regime shifts in the dynamic of the variance of the common factors and apply their model to study $q = 5$ Latin America stock indexes.

A Cholesky Approach. In this paper we take a third alternative that decomposes Σ_t via a Cholesky decomposition as

$$\Sigma_t = A_t H_t A_t'$$

where $A_t H_t^{1/2}$ is the lower triangular Cholesky decomposition of Σ_t. H_t is a diagonal matrix, the diagonal elements of A_t are all equal to one and, more importantly, the lower diagonal elements of A_t are unrestricted since positive definiteness is guaranteed. In the next section we show that there will be $q(q+1)/2$ dynamic linear models to be estimated and $3q(q+1)/2$ static parameters. When $q = 30$, for example, there are 465 latent states at each time and 1395 static parameters.

Even though the factor stochastic volatility structure might, at first, seem more parsimonious than our cholesky stochastic volatility model, it suffers from well known, unresolved problems, such as the selection of the order of the variables and, perhaps more importantly, it relies on the selection of a suitable, time invariant number of common factors. We argue that practice the number of factors is likely to be time varying. For instance, the number of common factors is lower during financial crises. Our approach avoid the need to determine time-varying number for factors, which can be endogenously accommodated by the ϕ-states dynamics, where the ϕ-states are the non-zero components of the inverse of the matrix A_t (see Section 3.2, particularly equations 3.2-3.5, for details). Put differently, the factor model limitations became, under our CSV structure, modeling tools that might suggest more parsimonious (zero columns in the lower triangular Cholesky matrix) and/or more sparse (zeros in the lower triangular Cholesky matrix).

The prior developed in above Section 2 coupled with the compuational approach developed below enables us to search for simplifying structure in a large system without imposing it.

3.2. Time-Varying Triangular Regressions

In this section we lay out our basic parametrization of the time-varying covariance structure in terms of linear regressions. Recall that $y_t \sim N(0, \Sigma_t)$ and $\Sigma_t = A_t H_t A_t'$ where $A_t H_t^{1/2}$ is the lower triangular Cholesky decomposition of Σ_t. The matrix A_t is lower triangular with ones in the main diagonal and $H_t = \text{diag}(\omega_{1t}^2, \ldots, \omega_{qt}^2)$. Therefore,

$$A_t^{-1} y_t \sim N(0, H_t).$$

Let the (i,j)th element of the lower triangular matrix A_t^{-1} be $-\phi_{ij}$ for $i > j$, while the diagonal (i,i) element is one. It follows that the joint normal distribution for y_t given F_{t-1}, that is $N(0, \Sigma_t)$, can be rewritten as a set of q recursive conditional regressions where

$$y_{1t} \sim N(0, \omega_{1t}^2) \tag{3.2}$$

and, for $i = 2, \ldots, q$,

$$y_{it} \sim N(\phi_{i1t} y_{1t} + \phi_{i2t} y_{2t} + \cdots + \phi_{i(i-1)t} y_{(i-1)t}, \omega_{it}^2). \tag{3.3}$$

Once ϕ_{ijt}s and ω_{it}^2s are available, so are A_t^{-1} (and A_t), H_t and, consequently, $\Sigma_t = A_t H_t A_t'$. To make Σ_t fully time-varying without any restrictions, we simply make each parameter in the regression representation time-varying. More precisely,

$$\phi_{ijt} \sim N(\alpha_{ij} + \beta_{ij}\, \phi_{ij(t-1)}, \tau_{ij}^2) \tag{3.4}$$

for $i = 2, \ldots, q$ and $j = 1, \ldots, i-1$, and

$$d_{it} \sim N(\alpha_i + \beta_i\, d_{i(t-1)}, \tau_i^2) \tag{3.5}$$

for $d_{it} = \log(\omega_{it}^2)$ and $i = 1, \ldots, q$, where τ_{ij}^2 and τ_i^2 are hyper-parameters. It is understood that the aforementioned distributions are all conditional on the available information F_{t-1}.

The actual parameters we work with are the ϕ_{ijt}s and d_{it}s. These parameters are our state variables in the state equations (3.4) and (3.5), while the recursive conditional regressions (or simply *triangular regressions*) are our observation in the observation equations (3.2) and (3.3). Our *Cholesky stochastic volatility* (CSV) model comprises equations (3.2) to (3.5).

Order of the time series. The Cholesky approach requires us to pick an order for the time series, which can be mistakenly seen as a weakness of the proposed modeling framework. We argue that the order can be important, but it does not necessarily implies that different orders will lead to different covariance estimates. More specifically, in one of our smaller studies (Study 4: S&P500, Small q), a model with $q = 9$ stocks was entertained with two different orders for the time series. Figure 9 plots two different estimates of time-varying

standard deviations (Microsoft and IBM) and one time-varying correlation (between Microsoft and IBM). In the second run we reversed the order of 9 time series, so the last becomes the first, the second last becomes the second and so on. The two CSV fits result from different priors on the d-states and ϕ-states under both orderings, so there is no reason that they be identical. However, their similarity is striking, regardless of the time of prior specification and regardless of that fact we are comparing standard deviations or correlations. In some cases, this may be a convenient way to express prior information. For example, if one series represented returns on the market (or some "factor") we may want to put it early in the list. In some applications, a natural ordering may not be apparent. Figure 9 suggests that when the data is reasonably informative, we do not have to worry too much about getting the "right" order. It also shows that, in this example, our MCMC is remarkably stable. In addition, one possible solution to overcome the ordering issue is randomly selecting a few orders and then averaging them out in order to obtain a more precise estimate of the covariance matrix.

Finally, since the set of recursive regressions are simply a standard decomposition of the q-variate normal distribution into the product of one univariate marginal normal distribution and $q-1$ univariate conditional normal distributions, it follows that the order would only matter as a function of the prior distribution for the parameters of such decomposition. Since the likelihoods are exactly the same, their maximizations lead to exactly the same estimates regardless of the order. The same can be said about the prior distribution if it is invariant to the permutation of the series. This is the case in our parameterization, since we do not treat differently, *a priori*, parameters from different equations. More precisely, if and when there is additional information regarding the order of the variables in the system, then the prior distribution plays an important role in shrinking and/or zeroing out irrelevant coefficients. Otherwise, our invariant prior seems to be working just as fine, at least in the empirical exercise we have performed.

Cholesky decomposition. The Cholesky decomposition approach has been studied elsewhere. Uhlig (1997) and Philipov and Glickman (2006b), for example, proposed models for the covariance matrix based on the temporal update of the parameters of a Wishart distribution (see also Asai and McAleer (2009)). Uhlig (1997) models $\Sigma_t^{-1} = B_{t-1}^{-1}\Theta_{t-1}(B_{t-1}^{-1})'\nu/(\nu+1)$, where $B_t = A_t H_t^{1/2}$ and $\Theta_{t-1} \sim Beta\left((\nu+pq)/2, 1/2\right)$ is a multivariate Beta distribution Uhlig (1994). See also Triantafyllopoulos (2008) for a similar derivation in the context of multivariate dynamic linear models. Philipov and Glickman (2006b) model $\Sigma_t^{-1} \sim W(\nu, S_{t-1}^{-1})$, where $S_{t-1}^{-1} = \frac{1}{\nu}(C^{1/2})(\Sigma_{t-1}^{-1})^d(C^{1/2})'$, such that $E(\Sigma_t|\Sigma_{t-1},\theta) = \nu(C^{-1/2})(\Sigma_{t-1})^d(C^{-1/2})'/(\nu-q-1)$. The parameter d controls the persistence in the conditional variance process. A constant covariance model arises when $d=0$, so $E(\Sigma_t) = \nu C^{-1}/(\nu-q-1)$ and C plays the role of a precision matrix. When $d=1$ and $C=I_q$, it follows that $E(\Sigma_t) = \Sigma_{t-1}$ so generating random walk evolution for the conditional covariance. See Dellaportas

and Pourahmadi (2012) for a similar model for time-invariant A and H_t following GARCH-type dynamics. Uhlig (1997) models daily/current prices per ton of aluminum, copper, lead and zinc, i.e. $q = 4$, exchanged in the London Metal Exchange. Philipov and Glickman (2006b) fit their model to returns data on $p = 5$ industry portfolios. Dellaportas and Pourahmadi (2012) model exchange rates of the US dollar against $q = 7$ other country/regions. A thorough review of the multivariate stochastic volatility literature up to a few years is provided in Asai et al. (2006) and Lopes and Polson (2010). See also Bauwens et al. (2012).

Shrinkage prior for large scale state-space models. Our parsimony inducing priors, when applied to the Cholesky SV problem, falls into the emerging literature on shrinkage priors for large scale state-space models. Frühwirth-Schnatter and Wagner (2010), for instance, uses spike-and-slab priors for shrinking states towards zero or nonzero constant in dynamic models, while Belmonte et al. (2014) and Bitto and Frühwirth-Schnatter (2016) implement hierarchical shrinkage to large dynamic systems. More recently, there has a several contributions to tackle sparsity dynamically, i.e. when a state variable (s_t in our generic notation) goes on and off throughout time. A few prominent contributions are Nakajima and West (2013), who proposes a thresholding scheme for dynamic sparsity, and Kalli and Griffin (2014) who extends the Normal-Gamma prior of Griffin and Brown (2010) with a stationary gamma autoregressive process. Additional related contributions are Rocková and McAlinn (2018), Kowal et al. (2018) and Uribe and Lopes (2018). Finance and economics applications appeared in Dangl and Halling (2012), Zhao et al. (2016), Eisenstat et al. (2016) and Carvalho et al. (2018).

3.3. Posterior Inference

We detail here the Markov chain Monte Carlo algorithm for posterior computation of our CSV model introduced above. Before we proceed and to make the prior specification less sensitive to scale, we recommend the standardization of the time series upfront, as it is commonly done in virtually all statistics and econometrics applications of finance, economics and related datasets.

Let q denote the number of series and T denote the number of observations on each time series. Let $Y_i = \{y_{it}\}_{t=1}^T$ and $d_i = \{d_{it}\}_{t=1}^T$, $i = 1, 2, \ldots, q$. Let $\phi_{ij} = \{\phi_{ijt}\}_{t=1}^T$, $i = 2, 3, \ldots, q$, $j = 1, 2, \ldots, (i-1)$. That is, Y_i is the time series of observations on the i^{th} variable, d_i is the time-varying state corresponding to the residual variance of the regression of y_{it} on y_{jt}, $j < i$, and ϕ_{ij} is the time-varying state corresponding to the regression coefficient of y_{it} on y_{jt}. See Equation (3.3). Let d_{i0} and ϕ_{ij0} denote initial states.

With $p(\cdot)$ denoting a generic probability density function, the full joint distribution of everything we need to think about is then given by the product of the following four hierarchical terms:

i. Likelihood function: $\prod_{i=2}^q p(Y_i \mid Y_1, \ldots, Y_{i-1}, d_i, \phi_{i1}, \ldots, \phi_{i(i-1)}) \times p(Y_1|d_1)$,
ii. (d, ϕ) *states:* $\prod_{i=1}^q p(d_i \mid \alpha_i, \beta_i, \tau_i, d_{i0}) \prod_{j<i} p(\phi_{ij} \mid \alpha_{ij}, \beta_{ij}, \tau_{ij}, \phi_{ij0})$,

iii. AR parameters: $\prod_{i=1}^{q} p(\alpha_i, \beta_i, \tau_i) \prod_{j<i} p(\alpha_{ij}, \beta_{ij}, \tau_{ij})$, and
iv. Initial states: $\prod_{i=1}^{q} p(d_{i0}) \prod_{j<i} p(\phi_{ij0})$,

where $\prod_{j<i} = 1$ when $i = 1$. The joint densities in *iii.* and in *iv.* denote our prior on the parameters of the autoregressive specification of the state evolution and our prior on the initial state, respectively. The choice of this prior is a key component of our approach and was extensively discussed in Section 2.

Our Markov chain Monte Carlo is a (large-scale) Gibbs sampler where we (efficiently) draw from the following full conditional distributions (with $\circ$ denoting "everything else"):

i. d states: $(d_{i0}, d_i) \,|\, \circ$,
ii. ϕ *states:* $(\phi_{ij0}, \phi_{ij}) \,|\, \circ$,
iii. d AR parameters: $(\alpha_i, \beta_i, \tau_i) \,|\, \circ$, and
iv. ϕ *AR parameters:* $(\alpha_{ij}, \beta_{ij}, \tau_{ij}) \,|\, \circ$.

The key property in this potentially large system is that, in the conditionals above, the states and parameters for a given equation are independent of the states and parameters of the other equations. This is readily seen in the structure of the full joint distributions given above. Thus, to draw d_i, we simply compute $\widetilde{y}_{it} = y_{it} - \sum_{j<i} \phi_{ijt}\, y_{jt}$ and use standard methods developed for univariate stochastic volatility given the model:

$$\begin{aligned} \widetilde{y}_{it} &\sim N(0, \exp\{d_{it}/2\}), \\ d_{it} &\sim N(\alpha_i + \beta_i\, d_{i(t-1)}, \tau_i^2). \end{aligned}$$

Similarly, the draw of ϕ_{ij} reduces to the analysis of a basic dynamic linear model (DLM) for $\widetilde{y}_{ijt} = y_{it} - \sum_{k<i, k\neq j} \phi_{ikt}\, y_{kt}$:

$$\begin{aligned} \widetilde{y}_{ijt} &\sim N(\phi_{ijt}\, y_{jt}, \exp\{d_{it}/2\}), \\ \phi_{ijt} &\sim N(\alpha_{ij} + \beta_{ij}\, \phi_{ij(t-1)}, \tau_{ij}^2). \end{aligned}$$

The draws of the AR parameter also reduce to consideration of a single state,

$$(\alpha_i, \beta_i, \tau_i) \,|\, \circ \equiv (\alpha_i, \beta_i, \tau_i) \,|\, (d_{i0}, d_i),$$

$$(\alpha_{ij}, \beta_{ij}, \tau_{ij}) \,|\, \circ \equiv (\alpha_{ij}, \beta_{ij}, \tau_{ij}) \,|\, (\phi_{ij0}, \phi_{ij}).$$

Thus, all the ϕ_{ij} draws reduce to simple applications of FFBS and all of the d_i draws reduce to those of the univariate stochastic volatility model. We use the method of Kim et al. (1998), again based on FFBS, for the univariate stochastic volatility model.

In order to keep the entire system manageable for large q, we use a univariate DLM for each ϕ in each equation rather than running a multivariate FFBS to jointly draw all the ϕ series for a given equation. This approach avoids a great many high-dimensional matrix operations. Potentially, this could put dependence into our chain depending upon the application. This does not seem to be a severe problem in our examples.

Thus, the whole thing boils down to repeated applications of the basic Gibbs sampler that cycles through $(s_0, s)\,|(\alpha, \beta, \tau)$ and $(\alpha, \beta, \tau)\,|(s_0, s)$, where s denotes a state series and s_0 the initial state. Since we need to put a strong prior on (α, β, τ) there is unavoidable dependence in the basic chain. Because of this dependence, we have found it useful to draw (α, β, τ) jointly as discussed in Section 2.1.

Before diving into a few illustrative examples, it is worth emphasizing that one of the strengths of the proposed CSV framework is that the triangular representation of the model naturally leads to parallelization in the MCMC scheme. See Appendix A for a simple account of the processing times when various processors are used in parallel to estimate a standard mid-size CSV model. Nonetheless, when it comes to recomputing full correlations matrices from ρ_{ijt}, the entire system is need, i.e. we need to compute the full covariance matrix Σ_t.

3.4. Small q Illustrations

In this section, we illustrate the performance of our CSV approach on simulated data with $q = 3$, and real data with $q = 3$ and $q = 9$. In the simulated example we can assess the performance by comparing the CSV fit to the known true Σ_t. In the real examples we compare our fit to that obtained from the well-known dynamic conditional correlation (DCC) model of Engle (2002) and Tse and Tsui (2002).

The DCC approach uses a two-step modeling procedure. In the first step, univariate GARCH models are built for individual return series. This step provides estimates of the volatility series σ_{it}. In the second step, a DCC model is applied jointly to the standardized return series, y_{it}/σ_{it}. Both the correlations and volatility series σ_{it} are time-varying so that, like the proposed CSV approach, a DCC model also delivers time-varying covariance matrices Σ_t.

In all three examples we also compare fits to those obtained from a simple moving window. In the moving window approach we estimate Σ_t with the sample covariance obtained from a subset of the data with times close to t.

We chose to compare to DCC/GARCH since we view it as the leading methodology available in the literature for small to moderate q. We chose to compare to the moving window since it is the only simple way to get fairly direct "look" at the data without making modeling assumptions.

These examples are meant to illustrate CSV. Our more ambitious goal is the analysis of large scale systems using an unrestricted model coupled with prior information to "regularize" the fit and we know of no competing methodology with these features. This is illustrated in Section 3.5

Note that we fit the DCC approach using the functions `dccPre` and `dccFit` available in the `R` package `MTS`.

3.4.1. Study 3: Simulated Example, Smooth Covariance Dynamics

To gain insight into the proposed analysis, in this section we present results from a simple simulated example. We let $q = 3$, Σ_0 be the identity and Σ_1 be the covariance matrix corresponding to standard deviations of $\sigma_1 = 4$, $\sigma_2 = 1$, $\sigma_3 = 0.25$, and correlations $\rho_{12} = \rho_{23} = 0.9$, $\rho_{13} = 0.81$. We then let $\Sigma_t = (1 - w_t)\,\Sigma_0 + w_t\,\Sigma_1$ where w_t increases from 0 to 1 as t goes from 1 to T. At each t we draw $y_t \sim N(0, \Sigma_t)$. We simulate $T = 500$ tri-variate observations.

We run our `R` package `csv` under two prior specifications. The first one is the default rougher prior specification (`defpri=0`), which is the *rougher* prior of Section 2.4.1:

$$p_{01} = 0.50, \quad p_{00} = 0.15, \quad p_{u0} = 0.15, \quad \tau_{max} = 0.15, \quad c_\tau = \{100, 200\},$$

where $c_\tau = 100$ when $\beta > 0$ and $c_\tau = 200$ when $\beta = 0$. We consider a second prior specification that changes hyperparameters of the *rougher* prior in the following way:

$$p_{01} = 0.85, \quad p_{00} = 0.05, \quad p_{u0} = 0.05, \quad \tau_{max} = 0.05, \quad c_\tau = \{200, 400\}.$$

This is the default *smoother* prior of our `R` package `csv` (`defpri=1`).

Results appear in Figure 5, corresponding to the *smoother* prior. The *smoother* prior puts more weight on the random walk component and favors smaller τ (and hence smoother states) by decreasing τ_{max} and increasing c_τ. The same (α, β, τ) prior was used for each of the six state series (three d-state series and three ϕ-state series). The moving window for estimation of Σ_t, includes all available observations within 50 time periods of t. The posterior median seems to nicely smooth the evidence from the data, with the tighter prior giving smoother results. Quite similar results (not shown) are obtained when using the *rougher* prior.

3.4.2. Study 4: S&P100, Small q

Our real data consists of daily stock returns. In what follows, the first example looks at returns for three (randomly chosen) companies ($q = 3$). The second example considers nine companies ($q = 9$). In both cases, the companies were randomly chosen out of the 94 companies we consider in the large q study of Section 3.5.

Three time series. The three financial time series are returns on Coke, Dell and DuPont. We have 2,516 daily returns from the beginning of 1997 to the end of 2006. We ran our `R` routine `csv` with the *rougher* default prior of Section 2.4.1 (`defpri=0`) for 15,000 draws, discarding the first 5,000.

Fits from the CSV and DCC models are checked for adequacy using the four tests implemented in the `R` package `MTS`. Appendix B provides a brief description of the four model checking statistics used. The results summarized in Table 1

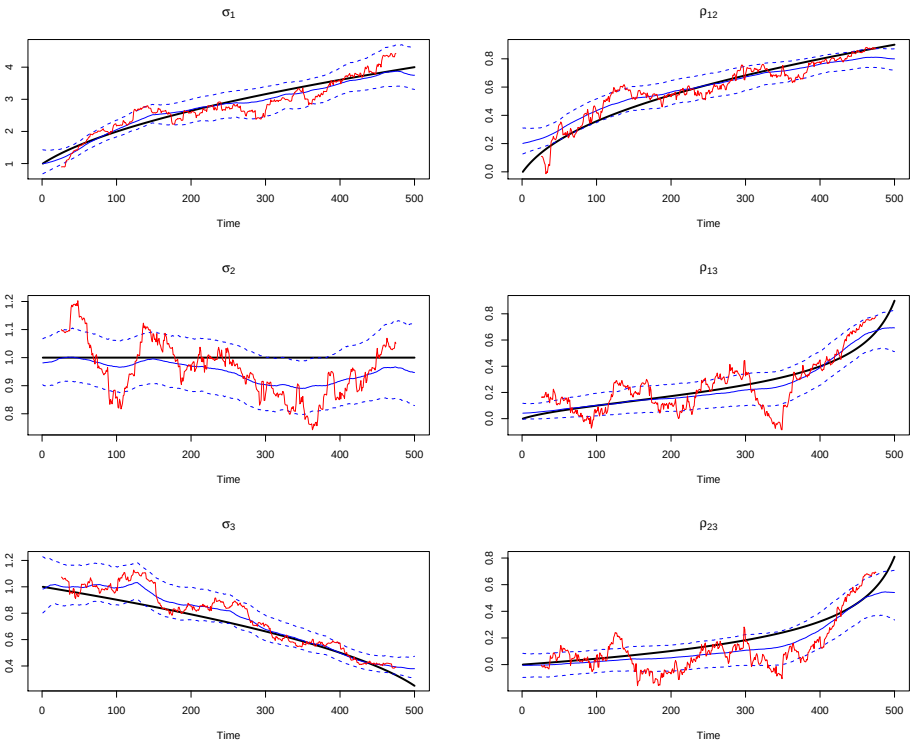

FIG 5. Simulated example based on the *smoother* prior – *Standard deviations are on the left column and the correlations are on the right column.* True values: *very smooth black curves.* Posterior median and 95% credible intervals: *lighter smooth blue (solide and dashed) curves.* Moving window estimates: *red dashed curves.*

show that none of the tests reject the null hypthesis that CSV model fits the data while the two robust tests reject the DCC fit.

Time-varying standard deviations and correlations estimates from both CSV and DCC models are compared to moving window (each window uses 5% of the data). Broadly, all three approaches agree as similar time paths are obtained DCC fits appear to be even more "jumbly" than those obtained from our *rougher* prior (figure not shown).

Figure 6 provides more detail on the fit of the $\{\rho_{32t}\}$ sequence. The top frame compares the CSV fit to the moving window fit and the bottom from compares the DCC fit to the moving window fit. In both frames the solid black line indicates the moving window fit and the dotted magenta lines give pointwise 90% posterior intervals for ρ_{32t} at each t (obtained from the CSV MCMC). In the top frame the CSV fit (posterior mean) is indicated by the dot-dash blue line and in the bottom frame the DCC fit is indicated by the dashed red line. Again, all three approches give similar fits. The CSV fit seems to track the

	CSV		DCC	
Test	Statistic	p-value	Statistic	p-value
$Q(m)$ of e_t	14.62	0.147	10.50	0.398
Rank-based $Q(m)$	7.93	0.636	100.81	0.000
$Q(m)$ of ϵ_t	82.17	0.709	61.49	0.991
Robust $Q(m)$	90.50	0.466	152.36	0.000

TABLE 1

Test validity of fitted model using `MCHdiag` *from* `R` *package* `MTS`. *Small p-values is "evidence" against the model. The residuals* e_t *and* ϵ_t *are, respectively,* $a_t'\Sigma_t^{-1}a_t$ *and* $\Sigma_t^{-1/2}a_t$, *where* a_t *and* Σ_t *are residuals and fitted covariance matrix. See Appendix B for more details on the four tests.*

moving window fit better. The correlation between the CSV fits and the moving window fits is 0.95 and the DCC - moving window correlation is 0.86. While we cannot argue that the moving window is the gold standard, we do find this reassuring. The uncertainty intervals appear to be quite reasonable and cover all three fitting approaches.

Despite relatively vague prior specification for the τ parameters (error standard deviations in the state equations), we observed that the posterior distributions of τ's corresponding to the stochastic volatilities, d_{it}, are larger than the ones corresponding to time-varying regression coefficients, ϕ_{ijt}, as expected (figure not shown). Sensitivity to prior hyperparameters, particularly related to the standard deviations of the state variables, namely τ_i and τ_{ij}, are presented in Figure 7. Notice how the right tail of our tail prior allows τ to get big for the d states and the tighter prior (`defpri=1`) shrinks τ's to slightly smaller values. In large q problems, where the data is less powerful relative to the complexity of the problem, this shrinkage plays a bigger role.

Nine time series. The 9 stocks used are Microsoft, IBM, Apple, Intel Corp, Cisco Systems, Qualcomm, Wal-Mart Stores, Home Depot and Costco Wholesale Corp. The time span is from 2004 to 2014. Figures 8 to 9 illustrates several aspects of implementing CSV. Here we present results based on the default *smoother* prior (`defpri=1`) of Section 3.4.1.

Both CSV and DCC fitted models are rejected by all four tests previously used for the case $q = 3$ (see Table 1). We argue that, even at a moderately size problem with $q = 9$, the tests may not be useful as "all models are false".

Similar to the smaller $q = 3$ case, DCC becomes more erratic when estimating standard deviations. In addition, DCC seems unable to follow the more abrupt changes in time-varying correlations captured by both CSV and the moving window. Figure 8 summarizes these findings for all 9 standard deviations and 36 correlations. It plots sample correlations of CSV point estimates of time-varying standard deviations and time-varying correlations against moving window counterparts; and it does similarly for DCC against moving window. In all cases CSV is "closer" to the moving window estimates. Again, while we cannot argue that the moving window results are the "correct" we find this reassuring.

Figure 9 plots two different estimates of time-varying standard deviations

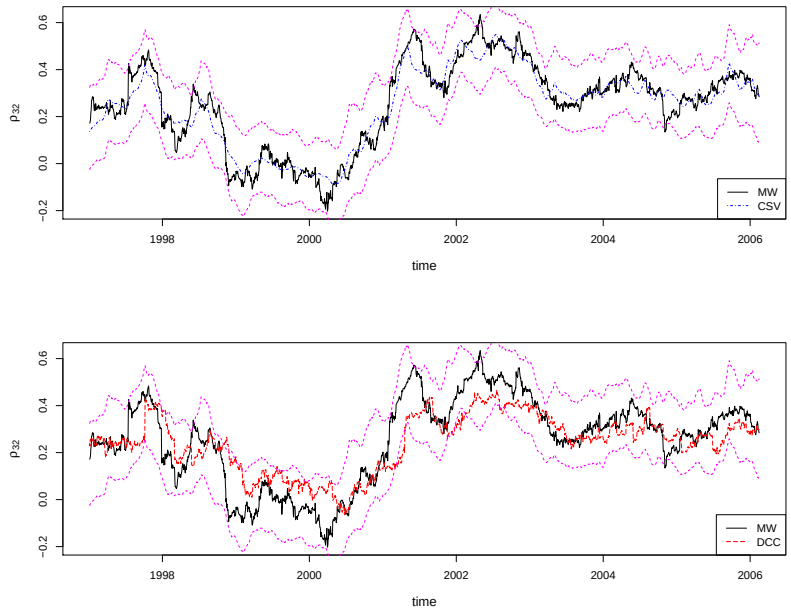

FIG 6. S&P100 data, $q = 3$, comparing CSV, DCC and moving window for ρ_{32t} – *CSV posterior medians (blue, top frame) and 95% credibility intervals (magenta, both frames) for correlation coefficients* ρ_{32t}. *Comparing to DCC estimates (red, bottom frame) and moving window (black, both frames).*

(Microsoft and IBM) and one time-varying correlation (between Microsoft and IBM). In the second run we reversed the order of 9 time series, so the last becomes the first, the second last becomes the second and so on. The two CSV fits result from different priors on the d-states and ϕ-states under both orderings, so there is no reason that they be identical. However, their similarity is striking, regardless of the time of prior specification (*rougher*, *smoother* or *much smoother*) and regardless of that fact we are comparing standard deviations or correlations. The *rougher* and *smoother* priors were specified in Section 3.4.1, while the *much smoother* prior is specified later in Section 3.5. The Cholesky approach requires us to pick an order for the time series. In some cases, this may be a convenient way to express prior information. For example, if one series represented returns on the market (or some "factor") we may want to put it early in the list. In some applications, a natural ordering may not be apparent. Figure 9 suggests that when the data is reasonably informative, we do not have to worry too much about getting the "right" order. It also shows that, in this example, our MCMC is remarkably stable.

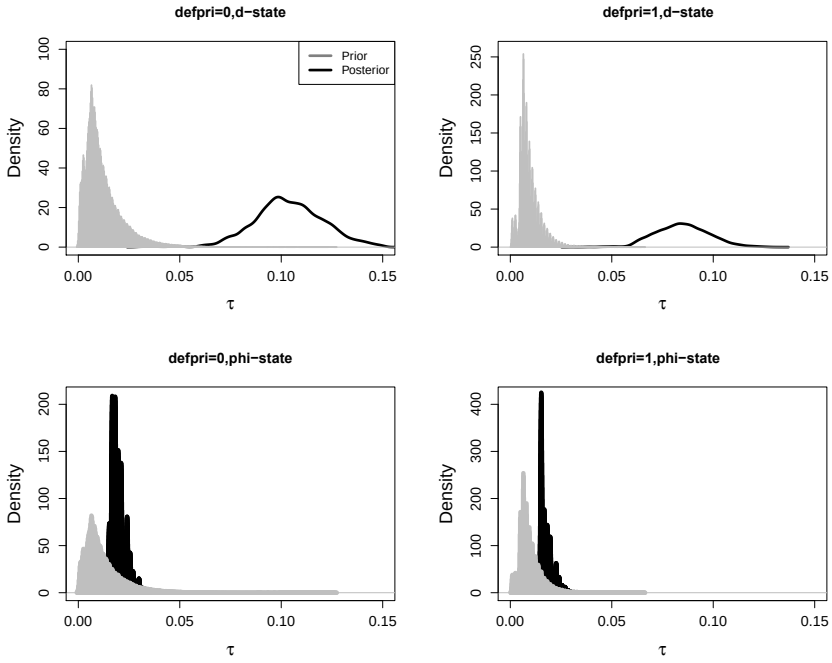

FIG 7. S&P100 data, $q = 3$, prior sensitivity analysis – *Prior distribution (gray) and posterior (black) distribution of standard deviations τ_1 (top row, d-state) and τ_{21} (bottom row, ϕ-state).* Rougher *prior,* `defpri=0` *(left column) and* smoother *prior,* `defpri=1` *(right column).*

3.5. Study 5: S&P100, Large q

In this section we use asset returns from firms making up the S&P100 index in order to illustrate the procedure with large q. We first consider a selection of returns on $q = 20$ of the firms and use the *smoother* prior discussed in Section 3.4.1. We use this prior because with larger q, more smoothing may be desirable.

Figure 10 plots the posterior means of the σ_{it} and ρ_{ijt} series. The top panel shows the 20 standard deviations series and the bottom panel shows the $20(19)/2 = 190$ correlations series. There is no simple way to plot so much information, but even with the many series, we can see that there is substantial time variation in both the standard deviations and the correlations. From the $\{\sigma_{it}\}$ series we see that there is an overall pattern to their behavior over time. For example, the volatility is generally lower at the end of the time period. However, there is substantial variation across assets (across i) both in the overall level of volatility and the amount of time variation. Similarly, there are time periods where ρ_{ijt} is relatively large for most (i, j) pairs but some pairs behave quite differently from the rest.

Figure 11 plots the posterior means of the states with the d states in the

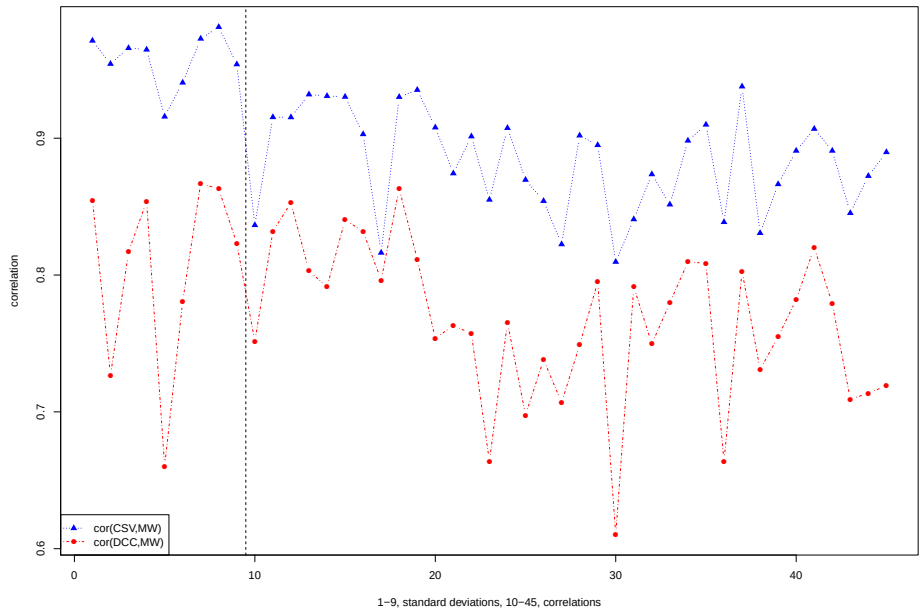

FIG 8. US stocks, $q = 9$ – *Sample correlations of CSV point estimates of 9 time-varying standard deviations (and 36 time-varying correlations) against moving window counterparts (blue lines). Similar for DCC against moving window (red line). CSV is based on our default* smoother *prior (*`defpri=1`*). The vertical dashed line separates standard deviations (1 to 9) from correlations (10 to 45).*

top panel and the ϕ states in the bottom panel. The top panel shows the time variation in the residual variances which vary markedly over time. The bottom panel shows that most of the ϕ series have relatively little time variation and are centered near zero. However, a few of the ϕ_{ijt} series do vary substantially over time. This figure shows how our Bayesian model, with our particular prior choice, seeks a parsimonious representation of a very high-dimensional problem.

Of course, the amount of "parsimony", "smoothness", or "regularization" inevitably is heavily influenced by our choice of prior. Figure 12 shows the posterior means of the states obtained when we use yet another prior given by

$$p_{01} = 0.85, \quad p_{00} = 0.05, \quad p_{u0} = 0.05, \quad \tau_{max} = 0.02, \quad c_\tau = \{300, 600\} \tag{3.6}$$

where again the two values for c_τ correspond to $\beta > 0$ and $\beta = 0$. This figure looks like a smoothed version of Figure 11. The "flat-line" appearance of many of the ϕ states is striking. The corresponding standard-deviation-correlation plot (not shown here) is, again, a smoothed version of Figure 10. One could argue that different levels of smoothness of the prior should be applied to state variables, perhaps with smoother specifications to the ϕ coefficients and less smooth ones

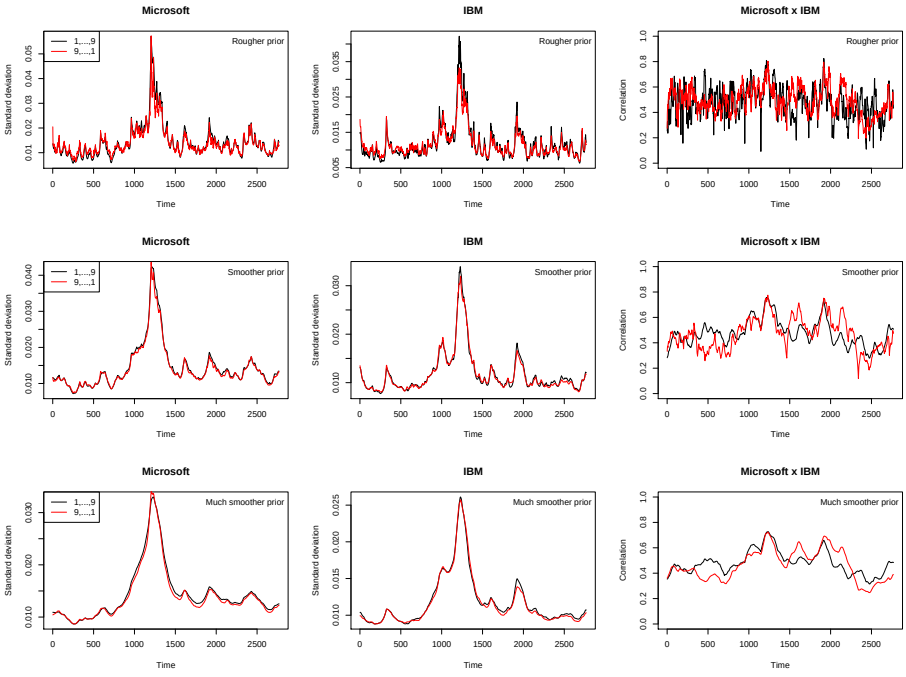

FIG 9. US stocks, $q = 9$, ordering of the data – $\{1, \ldots, 9\}$ *and* $\{9, \ldots, 1\}$ *indicate the orders of the time series when estimating CSV. The specifications of the* rougher *prior (*`defpri=0`*) and the* smoother *prior (*`defpri=1`*) appear in Section 3.4.1, while the specification of the* much smoother *prior appears Section 3.5.*

to the log-volatilities. In addition, the level of roughness or smoothness might vary according to the relation between q, number of time series, and T, number of time points.

Figure 13 plot through time the posterior means of the weights for the global minimum variance portfolio based on the *smoother* prior discussed in Section 3.4.1. The time t global minimum variance portfolio weights are computed as $\omega_t = \Sigma_t^{-1} 1_q / 1_q' \Sigma_t^{-1} 1_q$, where 1_q is a column vector of ones of length q. We see that the time variation in the standard deviations and correlations may be of real practical importance in that the corresponding portfolio weights change over time substantially. At time T, most of the weights are negligible, while being positive for about half a dozen assets (figure not shown).

Figure 14 reports results for $q = 94$ assets using the prior in Equation (3.6). Six of the return series from the 100 companies have missing values, due to inclusion and exclusion to the index, leaving us with 94. In this case there are 94 standard deviation series (σ_{it}) and $94(93)/2 = 4{,}371$ correlation pairs (ρ_{ijt}) so it becomes quite difficult to present the results. The top panel displays results for the σ_{it} while the bottom panel displays the ρ_{ijt}. The two panels have the same format. The solid gray band gives pointwise quartiles for the posterior means.

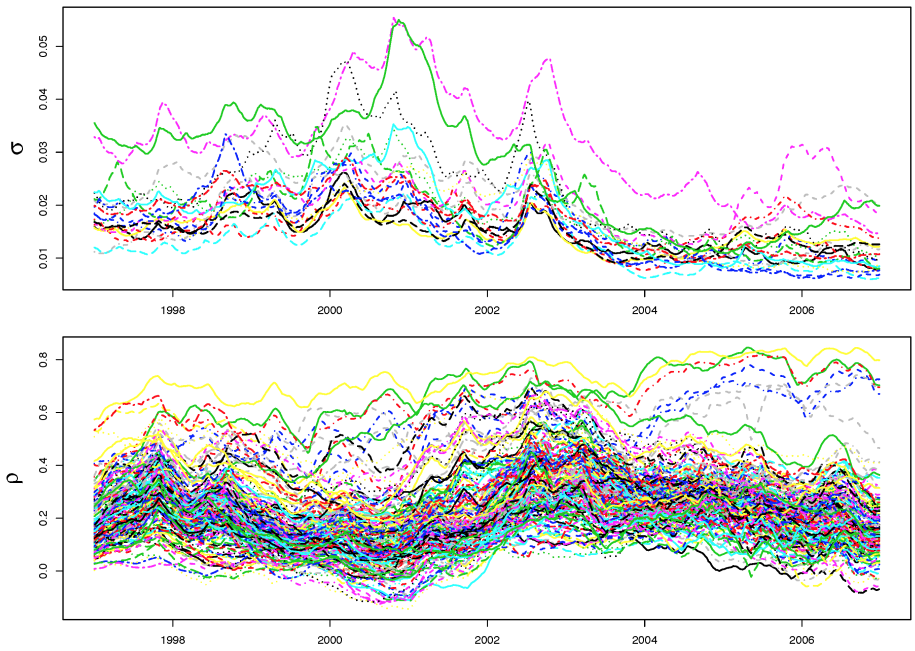

FIG 10. S&P100 data, $q = 20$, *smoother* prior of Section 3.4.1 – *Posterior means of time-varying standard deviations (top frame) and correlations (bottom frame).*

Thus, in the top panel, the gray band is the middle 50% of the 94 standard deviation posterior means $\hat{\sigma}_{it}$ for each fixed t and in the bottom panel it is the middle 50% of the 4,371 correlation estimates for each fixed t. The thick solid (black) lines give 95% intervals. We can see that with 94 series we observe the same overall patterns we saw with $q = 20$.

We also randomly picked 20 of the $\{\hat{\sigma}_{it}\}$ series to plot in the top panel and 20 of the $\{\hat{\rho}_{ijt}\}$ series to plot in the bottom panel. These plots, along with the size of the 95% intervals, indicate the while there is an overall pattern over time, there are substantial differences amongst the $\{\hat{\sigma}_{it}\}$ across i (assets) and the $\{\hat{\rho}_{ijt}\}$ across (i, j) (pairs of assets).

4. Final discussion and remarks

In this paper we develop a new prior specification for the parameters of the state equation in a state-space model. We then develop an approach for modeling high dimensional time varying covariance matrices in which the covariance at each time is a high dimensional state. We are able to compute the posterior of the states using parallel computation and shrinkage based on our new prior. In high dimensions, some form of shrinkage is essential given the large number of parameters and that we do not want to impose restrictions on the set of possible

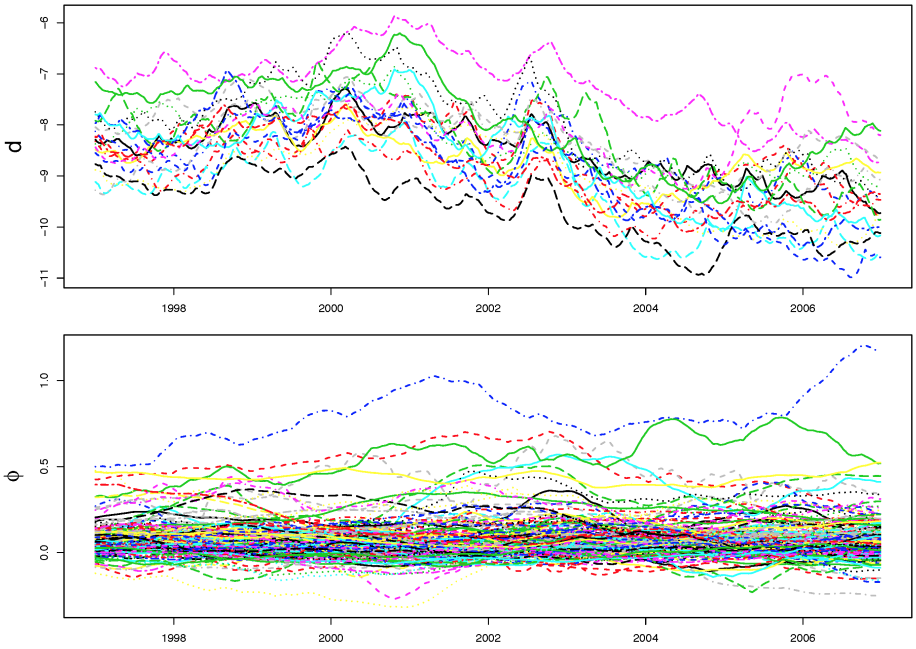

FIG 11. S&P100 data, $q = 20$, *smoother* prior of Section 3.4.1 – *Posterior means of the d-states (top frame) and the ϕ-states (bottom frame).*

covariance matrices. For the important example of vectors of asset returns, our prior allows us to uncover a novel form of shrinkage in which some state elements remain essentially constant over time while others vary.

State space models have become increasingly important in economic and financial applications as well as in problems in the physical sciences. Inevitably, the specification of the prior on the parameters of the state equation plays an important role in the overall model. Often, simple and possibly naive choices (such as imposing a random walk) are made. Our prior allows for consideration of the possibilities of interest to most researchers. Important cases such as the random walk model and the iid model become simple special cases whose presence may be inferred. In the example in Section 2.4.2, we find that the posterior probability of the random walk model is 0.99 given a prior probability of 0.5.

While our full prior specification provides the user with a lot of choice, the essential feature of state smoothness in easily controlled by choosing the τ_{min}, τ_{max}, and c_τ parameters. We show in the examples that a few simple choices give good results. As the dimension increases, we make the prior stronger (τ_{min} and τ_{max} smaller and c_τ bigger).

The problem of estimating time-varying covariance matrices Σ_t is important and difficult when the dimension q is large. Our approach was guided by the desire to enable parallel computation which is essential for large q and a desire to

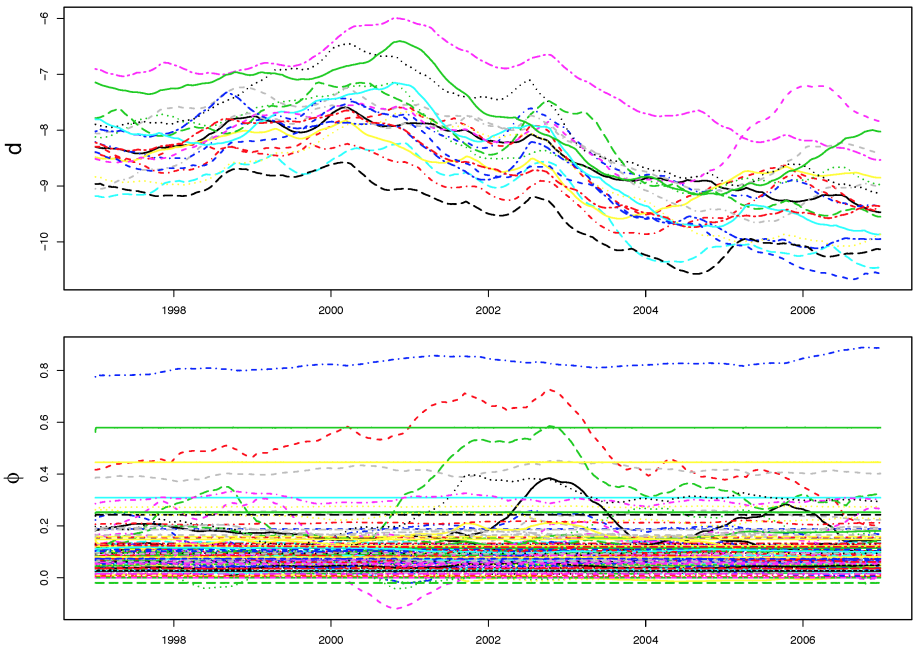

FIG 12. S&P100 data, $q = 20$, *much smoother* prior of Equation (3.6) – *Posterior means of the d-states (top frame) and the ϕ-states (bottom frame).*

keep the model as simple as possible without restricting the Σ_t. See Appendix A for an illustration of how parallel processing is a natural tool Bayesian inference in our time-varying covariance models. Approaches such as factor stochastic volatility achieve parsimony by making strong assumptions (the number of factors) which may not be time invariant.

However, without restrictions, some form of prior shrinkage (regularization) becomes essential to stop the model from overfitting. We show that our prior enables us to shrink towards smooth state evolution in a simple way and identify states which are essentially constant (see Figure 12). This is a novel form of shrinkage we feel is both an important empirical observation for the returns data as well as a useful general insight for high-dimensional state-space modeling.

For moderate dimensions we show that our approach is competitive with the popular GARCH-DCC methodology and gives stable intuitively plausible MCMC results (Table 1 and Figures 6 and 8).

While the examples in this paper show that a few simple prior choices work very well, it may also be of interest to go beyond these choices in practice. For example, Figure 7 and the simple fact that there are many more ϕ states than d states, suggest that a stronger prior might be used for the ϕ states rather than using the same prior for each state equation. A more exploratory modeling approach might reorder the components of the vector y from the most important

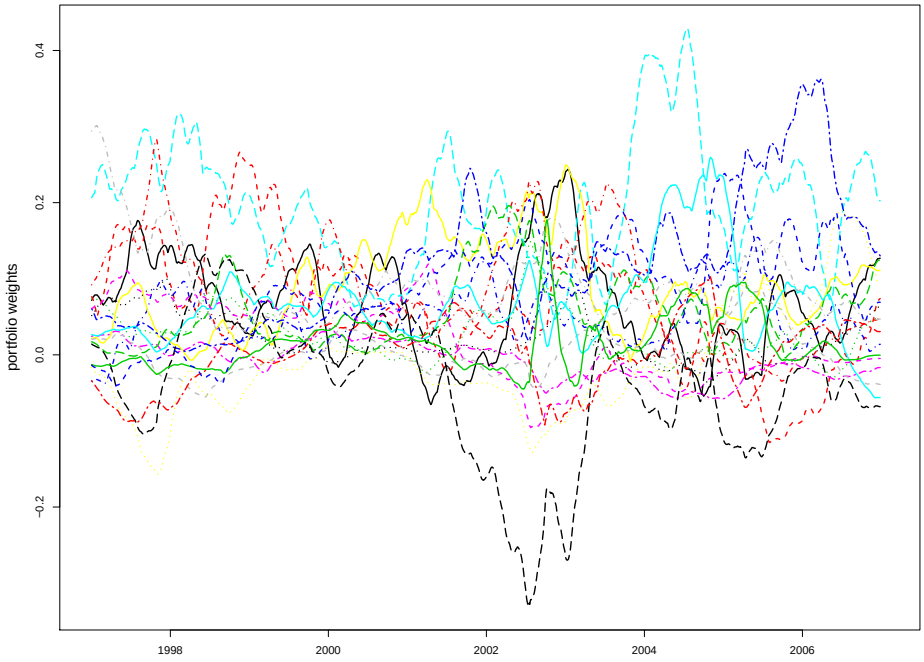

FIG 13. S&P100 data, $q = 20$, *smoother* prior of in Section 3.4.1 – *Posterior means of portfolio weights for the global minimum variance portfolio.*

to the least important relative to common latent factors. We argue that many of the ϕ-states would wonder around zero as the row of the Cholesky equation increases, mimicking the usual block, lower triangular factor loadings structure, as in Lopes and West (2004).

Both shrinkage and time-evolution of factor loadings in moderate and large scale factor stochastic volatility models has emerged over the last decade. See, amongst others, Lopes and Carvalho (2007), Zhao et al. (2016), Kastner et al. (2017). See also Frühwirth-Schnatter and Tüchler (2008) for a connection between rank reduction in the Cholesky decomposition and identification issues.

Of course, our approach has some key additional advantages stemming from its Bayesian formulation. It can be embedded in a large MCMC as a conditional model. A basic example is that any real example would have to have a model for the mean. In addition, the posterior uncertainty naturally qualifies our inference, something that is difficult to do in high dimensional models without the Bayesian machinery. An `R` package `csv` is being made available. A version (testing on Ubuntu and the Mac) will soon be available at `www.rob-mcculloch.org/csv`.

Final remark. A key contribution of our paper is the mixture prior on the AR(1) parameters (α, β, τ) of the state equation. In many applications, this state equation specification lies at the heart of the model. Our prior coherently

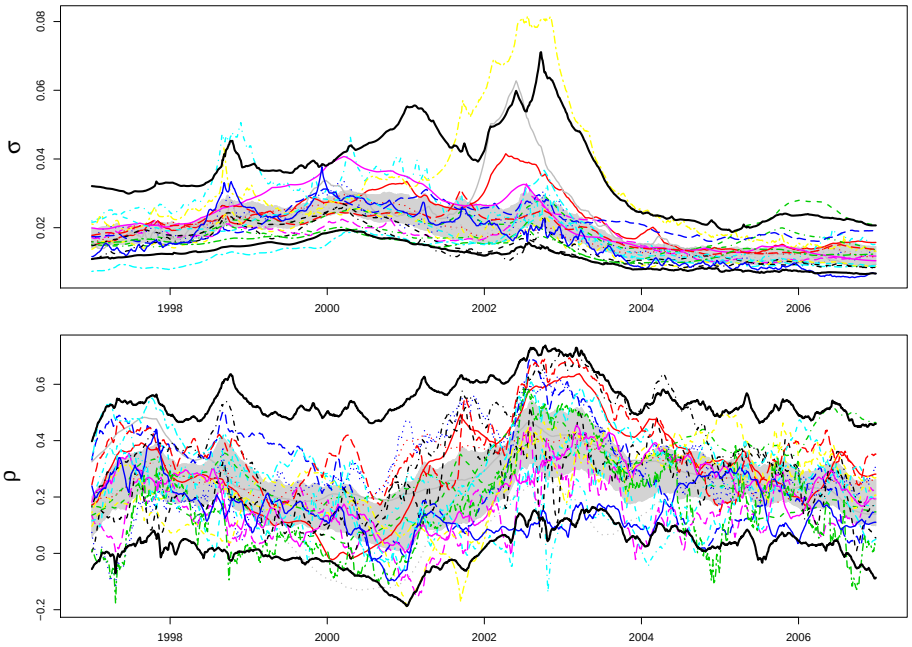

FIG 14. S&P100 data, $q = 94$, *much smoother* prior of Equation (3.6) – *Posterior means of time-varying standard deviations (top frame) and correlations (bottom frame).*

delineates the structural implications of prior choice. Our prior prior is the first to do so, and should be helpful both in inputting prior information about structure, and extracting posterior inferences about structure. A key element of our prior is the simple specification for the τ prior. This priors allows us to express prior beliefs appropriate for the larger context of the state space model and is superior to the commonly use conditionally conjugate prior.

Our MCMC approach to the posterior computation is simple and allows us to obtain posterior probabilities of key quantities like the probability $\beta = 1$ (Section 2.4.2) in a relatively straightforward manner. However, our MCMC algorithm was tailored to the applications in this paper and modifications of the algorithm could be of interest in other situations. In particular, the simple Gibbs sampler (Equation 2.4) mixes slowly and in some applications it might be worth computing a marginal likelihood by integrating out the state so the the parameters may be drawn directly. In this paper, inferential details our full mixture model prior were only of interest in low dimension problems (Section 2.4) so that the slow mixing was handled by using long runs.

Another contribution of our paper is inference for high-dimensional time varying covariance matrices (Section 3). Our approach builds upon our prior specification and much of the development of the prior was driven by this problem. Our MCMC for the for this problem draws each $\{\phi_{ijt}\}$ sequence for a given

i and j conditionally. In some applications, a multivariate approach may be preferable. In our high dimensional examples, the correlations were not extreme so that the univariate approach worked well.

References

Aguilar, O. and West, M. (2000). "Bayesian dynamic factor models and portfolio allocation." *Journal of Business and Economic Statistics*, 18: 338–357.

Asai, M. and McAleer, M. (2009). "The structure of dynamic correlations in multivariate stochastic volatility models." *Journal of Econometrics*, 150: 182–192.

Asai, M., McAleer, M., and Yu, J. (2006). "Multivariate stochastic volatility: a review." *Econometric Reviews*, 25: 145–175.

Bandi, F. M. and Russell, J. R. (2008). "Microstructure noise, realized Variance, and optimal sampling." *Review of Economic Studies*, 75: 339–369.

Bauwens, L., Hafner, C., and Laurent, S. (2012). "Volatility Models." In Bauwens, L., Hafner, C., and Laurent, S. (eds.), *Handbook of Volatility Models and Their Applications*, 1–45. John Wiley & Sons, Inc., Hoboken, NJ, USA.

Belmonte, M. A., Koop, G., and Korobilis, D. (2014). "Hierarchical Shrinkage in Time-Varying Parameter Models." *Journal of Forecasting*, 33(1): 80–94.

Bitto, A. and Frühwirth-Schnatter, S. (2016). "Achieving Shrinkage in a Time-Varying Parameter Model Framework." *arXiv preprint arXiv:1611.01310*.

Carvalho, C. M., Lopes, H. F., and McCulloch, R. E. (2018). "On the Long Run Volatility of Stocks." *Journal of the American Statistical Association (in press)*.

Chib, S., Nardari, F., and Shephard, N. (2006). "Analysis of high dimensional multivariate stochastic volatility models." *Journal of Econometrics*, 134: 341–371.

Dangl, T. and Halling, M. (2012). "Predictive regressions with time-varying coefficients." *Journal of Financial Economics*, 106: 157–181.

Dellaportas, P. and Pourahmadi, M. (2012). "Cholesky-GARCH models with applications to finance." *Statistics and Computing*, 22: 849–855.

Dufour, J. M. and Roy, R. (1986). "Generalized portmanteau statistics and tests of randomness." *Communications in Statistics: Theory and Methods*, 15: 2953–2972.

Eisenstat, E., Chan, J. C. C., and Strachan, R. (2016). "Stochastic model specification search for time-varying parameter VARs." *Econometric Reviews*, 35: 1638–1665.

Engle, R. F. (2002). "Dynamic conditional correlation: a simple class of multivariate generalized autoregressive conditional heteroskedasticity models." *Journal of Business and Economic Statistics*, 20: 339–350.

Frühwirth-Schnatter, S. (2004). "Efficient Bayesian parameter estimation." In Harvey, A., Koopman, S. J., and Shephard, N. (eds.), *State Space and Unobserved Component Models*, 1123–151. Cambridge University Press, CAmbridge.

Frühwirth-Schnatter, S. and Tüchler, R. (2008). "Bayesian parsimonious covariance estimation for hierarchical linear mixed models." *Statistics and Computing*, 18: 123–151.

Frühwirth-Schnatter, S. and Wagner, H. (2010). "Stochastic model specification search for Gaussian and partial non-Gaussian state space models." *Journal of Econometrics*, 154(1): 85–100.

Frühwirth-Schnatter, S. and Wagner, H. (2010). "Stochastic model specification search for Gaussian and partially-Gaussian state space models." *Journal of Econometrics*, 154: 85–100.

George, E. I. and McCulloch, R. E. (1993). "Variable selection via Gibbs sampling." *Journal of the American Statistical Association*, 79: 677–83.

Griffin, J. and Brown, P. (2010). "Inference with normal-gamma prior distributions in regression problems." *Bayesian Analysis*, 5(1): 171–188.

Han, Y. (2006). "Asset allocation with a high dimensional latent factor stochastic volatility model." *The Review of Financial Studies*, 19: 237–271.

Harvey, A. C., Ruiz, E., and Shephard, N. (1994). "Multivariate stochastic variance models." *Review of Economic Studies*, 61: 247–264.

Kalli, M. and Griffin, J. E. (2014). "Time-varying sparsity in dynamic regression models." *Journal of Econometrics*, 178(2): 779–793.

Kastner, G., Frühwirth-Schnatter, S., and Lopes, H. F. (2017). "Efficient Bayesian inference for multivariate factor stochastic volatility models." *Journal of Computational and Graphical Statistics*, 26: 905–917.

Kim, S., Shephard, N., and Chib, S. (1998). "Stochastic volatility: likelihood inference and comparison with ARCH models." *Review of Economic Studies*, 65: 361–393.

Kowal, D. R., Matteson, D. S., and Ruppert, D. (2018). "Dynamic Shrinkage Processes." Technical report.

Lopes, H. F. and Carvalho, C. M. (2007). "Factor stochastic volatility with time varying loadings and Markov switching regimes." *Journal of Statistical Planning and Inference*, 137: 3082–3091.

Lopes, H. F. and Migon, H. S. (2002). "Comovements and contagion in emergent markets: stock indexes volatilities." *Case Studies in Bayesian Statistics*, 6: 285–300.

Lopes, H. F. and Polson, N. G. (2010). "Bayesian inference for stochastic volatility modeling." In Bocker, K. (ed.), *Rethinking Risk Measurement and Reporting: Uncertainty, Bayesian Analysis and Expert Judgement*, 515–551. RiskBooks.

Lopes, H. F. and West, M. (2004). "Bayesian model assessment in factor analysis." *Statistica Sinica*, 14: 41–67.

Migon, H. S., Gamerman, D., Lopes, H. F., and Ferreira, M. A. R. (2005). "Dynamic models." In Dey, D. and Rao, C. R. (eds.), *Handbook of Statistics: Bayesian Thinking, Modeling and Computation*, volume 25, 553–588. Elsevier.

Nakajima, J. and West, M. (2013). "Bayesian analysis of latent threshold dynamic models." *Journal of Business & Economic Statistics*, 31(2): 151–164.

Nardari, F. and Scruggs, J. T. (2007). "Bayesian analysis of linear factor mod-

els with latent factors, multivariate stochastic volatility, and APT pricing restrictions." *Journal of Financial and Quantitative Analysis*, 42: 857–892.

Philipov, A. and Glickman, M. E. (2006a). "Factor multivariate stochastic volatility via Wishart processes." *Econometric Reviews*, 25: 311–334.

— (2006b). "Multivariate stochastic volatility via Wishart processes." *Journal of Business and Economic Statistics*, 24: 313–328.

Pitt, M. and Shephard, N. (1999). "Time varying covariances: a factor stochastic volatility approach." In *et al*, J. B. (ed.), *Bayesian statistics 6*. London: Oxford University Press.

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

Rockova, V. and McAlinn, K. (2018). "Dynamic Variable Selection with Spike-and-Slab Process Priors." Technical report, Booth School of Business, University of Chicago.

Triantafyllopoulos, K. (2008). "Multivariate stochastic volatility with Bayesian dynamic linear models." *Journal of Statistical Planning and Inference*, 138: 1021–1037.

Tsay, R. S. (2014). *Multivariate Time Series Analysis: with R and Financial Applications*. Wiley.

Tse, Y. K. and Tsui, A. K. C. (2002). "A multivariate generalized autoregressive conditional heteroscedasticity model with time-varying correlations." *Journal of Business and Economic Statistics*, 20: 351–362.

Uhlig, H. (1994). "On singular Wishart and singular multivariate beta distributions." *The Annals of Statistics*, 22: 395–405.

— (1997). "Bayesian vector autoregressions with stochastic volatility." *Econometrica*, 65: 59–73.

Uribe, P. W. and Lopes, H. F. (2018). "Dynamic sparsity on dynamic regression models." Technical report.

Zhang, L., Mykland, P. A., and Ait-Sahalia, Y. (2005). "A tale of two time scales: determining integrated volatility with noisy high-frequency data." *Journal of the American Statistical Association*, 100: 1394–1411.

Zhao, Z. Y., Xie, M., and West, M. (2016). "Dynamic dependence networks: Financial time series forecasting and portfolio decisions." *Applied Stochastic Models in Business and Industry*, 32: 311–332.

Appendix A: Parallel processing

One of the strengths of the proposed CSV framework is that the triangular representation of the model naturally leads to parallelization in the MCMC scheme. More specifically, the $T \times i$-dimensional state-space matrix

$$(d_i, \phi_{i1}, \ldots, \phi_{i,i-1}),$$

and the $3 \times i$-dimensional parameter matrix

$$(\alpha_i, \beta_i, \tau_i, \alpha_{i1}, \beta_{i1}, \tau_{i1}, \ldots, \alpha_{i,i-1}, \beta_{i,i-1}, \tau_{i,i-1}),$$

corresponding to the i-th recursive conditional regression can be drawn independently from the other recursive conditional regressions.

However, it is well known that sampling d_i (log-volatilities) is more computationally expensive (more time consuming) than sampling ϕ_{ij}. In fact, for a small to moderate i, it is likely that the computational burden is due to d_i almost exclusively. Let c_d, c_ϕ and c_θ be the computational cost (in seconds, for instance) to draw the T-dimensional vectors d_i and ϕ_{ij} and the 3-dimensional vectors $\theta_i = (\alpha_i, \beta_i, \tau_i)$, for any i and j (see full conditional distributions in Section 3.3). Usually c_θ is negligible when compared to c_d and c_ϕ. The cost to draw the states from recursive conditional regression i is $c_i = c_d + (i-1)c_\phi + ic_\theta$, and the total cost is

$$c = \kappa_1(q)c_d + \kappa_2(q)c_\phi + \kappa_3(q)c_\theta$$

where $\kappa_1(q) = q$, $\kappa_2(q) = q(q-1)/2$ and $\kappa_3(q) = q(q+1)/2$. Similarly, the total cost of running regressions $i_a + 1$ to i_b ($i_b - i_a$ regressions) is

$$c_{i_a:i_b} = \Delta\kappa_1^{ab}c_d + \Delta\kappa_2^{ab}c_\phi + \Delta\kappa_3^{ab}c_\theta$$

where $\Delta\kappa_j^{ab} = \kappa_j(i_b) - \kappa_j(i_a)$, for $j = 1, 2, 3$. Assume that computation can be split between two parallel processors. Due to the imbalance between (mainly) c_d and c_ϕ (and c_θ), it is not immediately obvious which recursive conditional regression i_1 will make $c_{1:i_1} = c_{(i_1+1):q} = c/2$. Similarly, what are the optimal i_1 and i_2 when three processors are available? In general, for m processors, the goal is to find the cut-offs $(i_1, i_2, \ldots, i_{m-1})$ such that the cost within each group of recursive conditional regressions is the same:

$$c_{1:i_1} = c_{(i_1+1):i_2} = \cdots = c_{(i_{m-2}+1):i_{m-1}} = c_{(i_{m-1}+1):q} = c/m.$$

The search for the cut-offs is performed recursively with i_1 selected from $\{1, \ldots, q\}$ such that $c_{1:i_1} < c/m$ and $c_{1:(i_1+1)} > c/m$, i_2 selected from $\{i_1 + 1, \ldots, q\}$ such that $c_{1:i_2} < 2c/m$ and $c_{1:(i_2+1)} > 2c/m$, and so forth.

Figure 15 provides an illustration when there are $q = 100$ time series and up to $m = 20$ processors. The costs $(c_d, c_\phi, c_\theta) = (310, 23, 0)$ are based on actual run times (in seconds) for $T = 2,516$ time points and 50,000 MCMC draws. It takes 13.5 times longer to draw d_i than it does to draw ϕ_{ij}. These costs were based on our code running in a 2.93 GHz Intel Core 2 Duo processor. For $m = 1$ processor, the total cost is about 26 hours. For $m = 2$ processors, $i_1 = 67$ and the cost per processor is about 21 hours. For $m = 3$ processors, $(i_1, i_2) = (52, 79)$ and the cost per processor is about 14 hours. For $m = 4$ processors, $(i_1, i_2, i_3) = (44, 67, 84)$ and cost per processor is about 10.5 hours. For $m = 20$ processors, cost per processor is about 2 hours.

Appendix B: Prior setup in `R` package `csv`

Recalling the set up of Section 2.4.1, In the univariate state-space model with observation equation $y_t = f(x_t, s_t, \eta_t)$ and state equation $s_t = \alpha + \beta\, s_{t-1} + \tau\varepsilon_t$,

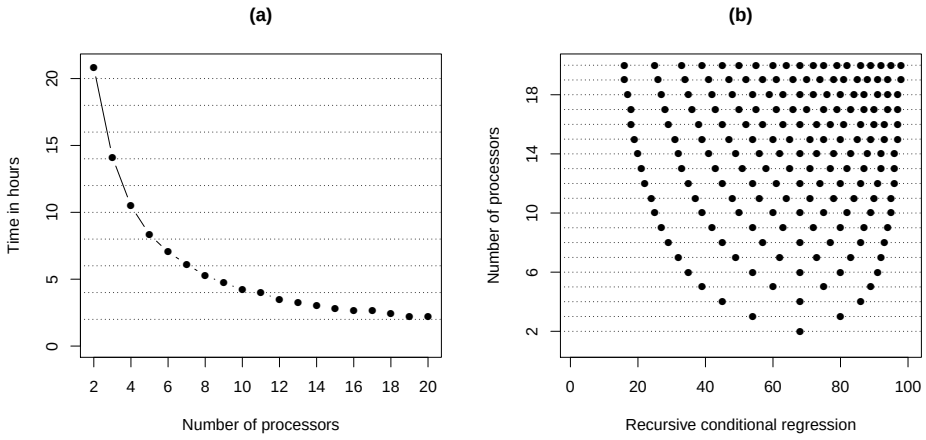

FIG 15. Multiple processors – *In panel (a) we plot the number of processors vs. the total time in hours to run 50,000 iterations for a 100 × 100 ($q = 100$) time varying covariance matrix with $T = 2,516$. It takes about 13.5 times longer to draw a d state than it does to draw a ϕ state. Code was run on a 2.93 GHz Intel Core 2 Duo processor. With 1 processor, the time is about 26 hours. With 20 processors, the time is about 2 hours. In panel (b) we have the number of processors on the vertical axis and each set of points along the dotted lines indicate how the 100 conditional regressions in the Cholesky decomposition are allocated to the different processors. For example, when $m = 2$ the cut-off is regression $i_1 = 67$, i.e. the first processor runs regressions 1 to 67 while the second processor runs regressions 68 to 100.*

the full mixture prior for the parameters (α, β, τ) of the state equation is

$$\begin{aligned} p(\alpha,\beta,\tau) &= p_{01}\, p(\tau|\beta=1)\, \delta_{\{\alpha=0,\beta=1\}} + p_{00}\, p(\tau|\beta=0)\, \delta_{\{\alpha=0,\beta=0\}} \\ &+ p_{u0}\, p(\tau|\beta=0)\, p(\alpha|\beta=0,\tau)\, \delta_{\{\beta=0\}} + p_{uu}\, p(\beta)\, p(\tau|\beta\neq 0)\, p(\alpha|\beta), \end{aligned}$$

where p_{01}=`p01`, p_{00}=`p00` and p_{u0}=`pu0`, and

- Prior on $\tau|\beta$: $Pr(\tau = \tau_i|\beta) \propto \exp\{-c_\tau|\tau_i - \tau_{min}|\}$, where $Pr(\tau = \tau_{min}|\beta) = p_{min}$, $\tau_i \in \{\tau_{min} + h_\tau, \ldots, \tau_{max}\}$, with h_τ is defined on a grid of length `ngt`, p_{min}=`taming`, τ_{max} =`taumax`. Additionally, when $\beta = 0$, τ_{min} =`taumin0` and c_τ =`tauc0`, and when $\beta \neq 0$, τ_{min} =`taumin` and c_τ =`tauc`.
- Prior on $\alpha|\beta$: $\alpha|\beta \sim N\{0, \sigma_\alpha^2(1-\beta^2)\}$, where σ_α =`sa`.
- Prior on β: $Pr(\beta = \beta_i) \propto p_N(\beta_i, \bar{\beta}, \sigma_\beta^2)$, where $\bar{\beta}$=`bbar`, σ_β=`sb`, and $\beta_i \in (0,1)$ on a grid of length `ngb`.

The prior on initial state, s_0, is $s_0 \sim \gamma N(0,(cw)^2) + (1-\gamma)N(0,w^2)$ and $\gamma \sim Ber(p^*)$, where p^* =`gamp`, w =`wgam`, and c =`cgam`.

B.1. Default smoother *prior* - `defpri=1`

This is the default prior we set-up for `csv`. In other words, running `csv(y)` is the same as running

```
csv(y,burn=500,nd=1000,thin=1,taumin=0.005,taumin0=0.001,taumax=0.05,
    tauminp=0.5,tauc=200,tauc0=400,p00=0.05,pu0=0.05,p01=0.85,sa=2.0,
    bbar=1.0,sb=1.0,gamp=0.5,wgam=0.1,cgam=10.0,ngb=100,ngt=100,defpri=1)
```

B.2. Default rougher *prior* - `defpri=0`

Running `csv(y,defpri = 0)` sets `p01=0.5`, `p00=0.15`, `pu0=0.15`, `taumax=0.15`, `tauc=100` and `tauc0=200`, while all other values are kept the same as in the case of the *smoother* prior.

Appendix C: Model checking statistics

The test statistics that appear on Table 1 of Section 3.4 are listed below. The quantities a_t and Σ_t are, at time t, the q-dimensional vector of residuals and fitted covariance matrix of the fitted model.

Ljung-Box $Q_e(m)$ test statistic: $Q_e(m)$ is the well-known Ljung-Box statistic, $Q(m)$, of the transformed residual series $e_t = a_t'\Sigma_t^{-1}a_t$. Under the assumption that the fitted model is the true model, the e_t are independent and identically distributed χ^2_q random variates so that $Q_e(m)$ follows asymptotically a χ^2_m distribution.

Rank-based $Q^r_e(m)$ test statistic: $Q^r_e(m)$ is $Q_e(m)$ for the rank series of e_t. Dufour and Roy (1986) show that $Q^r_e(m)$ follows asymptotically a χ^2_m distribution if e_t has no serial dependence.

Ljung-Box $Q_\epsilon(m)$ test statistic: $Q_\epsilon(m)$ is the multivariate $Q(m)$ statistic of the standardized residuals $\epsilon_t = \Sigma_t^{-1/2}a_t$.

Robust $Q^R_\epsilon(m)$ test statistic: $Q^R_\epsilon(m)$ is a robust version of $Q_\epsilon(m)$ by 5% trimming of the residual a_t based on the order statistics of e_t. As shown in Tsay (2014), $Q\epsilon^R(m)$ works well in detecting conditional heteroscedasticity of multivariate return series.