PARSIMONY INDUCING PRIORS FOR LARGE SCALE STATE-SPACE MODELS[1]

Hedibert Freitas Lopes
Insper, Brazil

Robert McCulloch
Arizona State University, USA

Ruey Tsay
The University of Chicago Booth School of Business, USA

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. Our prior for the state-space component's standard deviation is able to accommodate different scenarios. 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.

[1]Previous versions of the manuscript has been presented in various institutions and conference meetings: Pennsylvania State University, Universidad Autonoma de Mexico, University of Missouri, University of Colorado-Boulder, Vienna University of Economics and Business, University of Washington, University of Waterloo, University of South Carolina, Columbia University, University of New Mexico, Rutgers Business School, University of São Paulo, Consiglio Nazionale delle Ricerche, George Mason University, University of California at Irvine, University of California at Santa Cruz, IMPA, Insper, Arizona State University, University of Texas at Austin. XX Brazilian Symposium of Probability and Statistics, VIII International Purdue Symposium on Statistics, Meetings of the Midwest Econometrics Group, Oxford-Man Institute Conference on Financial Econometrics and Vast Data, XXVI Brazilian Colloquium of Mathematics, V Workshop on Bayesian Inference in Stochastic Processes, Seminar on Bayesian Inference in Econometrics and Statistics, Joint Statistical Meeting, Minneapolis and XI School of Time Series and Econometrics. We are deeply thankful for all the discussants for their invariably useful and clarifying feedback. The also thank the editorial team, who has been fundamental to guide us to a much clearer and complete version of the manuscript. All inconsistencies and remaining typos are all on us.

Our model for this large system enables us to use parallel computing.

Keywords: Bayesian modeling, conditional heteroscedasticity, forward filtering and backward sampling, parallel computing, sparsity, shrinkage.

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, for instance, Migon et al. [2005] and Schmidt and Lopes [2019], for reviews of various kinds of dynamic models. In the modern and applied business and economics literature, 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]. See also Belmonte et al. [2014] and Bitto and Frühwirth-Schnatter [2019], amongst others, for regularisation inducing strategies in TVP models.

To be more specific, we will consider throughout the basic dynamics governing a given state-space component, which will be called s_t, and could potentially be a time-varying coefficient in the VAR model, a log-volatility in a SV model or a time-varying factor loading in a DFM. The most popular dynamics 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. He assumes $(\alpha, \beta) = (0, 1)$ when modeling the time-varying coefficients of the VAR model, giving random walk dynamics.

One of our key contributions is to avoid the conditionally conjugate normal-inverse gamma (N-IG) prior, commonly used in the Bayesian state-space literature to model (α, β, τ^2). The N-IG prior fails to properly account for common parsimonious and sparse cases. In other words, when a state-space component s_t mimics the behaviour of a time invariant, static parameter, the N-IG-

AR(1) model has a hard time identifying such behaviour. The main reason is that there are two two ways to represent such situation: i) $\beta = 0$ and α is close to the static parameter, possibly zero, or ii) $\alpha = 0$ and $\beta = 1$. In both cases, τ can be fairly small and hard to deal with. We propose a prior for τ that is able to accommodate these and other scenarios. For an illustrative example, see the top row of Figure 3, where the state-space component flat-lines at a nonzero constant. We argue, and show in our applications, that i) limiting the evolution of state-space components to a random walk, as in Primiceri [2005] and many others that followed, can be unrealistic and forces static states to artificially evolve over time, and consequently that ii) using the conditionally conjugate normal-inverse gamma prior for (α, β, τ^2) prevents sparsity in TVP models (see, for instance, Frühwirth-Schnatter [2004]). These limitations are particularly troubling when dealing with large-scale systems with several hundreds, or thousands, of state-space components flat-lining, rendering the random walk hypothesis meaningless. Section 3 carefully treats the Cholesky stochastic volatility model when modeling exchange rates and components of the S&P100 index. In both real data analysis contexts, we found that the data "recommends" that several of the state-space components be flat-lined.

The first of our two 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), $\alpha = \beta = 0$ (sparse component) , $\beta = 0$ (flat-line component), and $0 < \beta < 1$ (stationary component). Our mixture prior probability implicitly addresses the identifiability mentioned earlier. As it can be seen, two of these mixture components, namely the sparse component and the flat-line component, are sparse-inducing ones, therefore in line with the current literature above mentioned.

The second of our two main goals, also extensively discussed in Section 3, is 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 collapsing unnecessary coefficients at constant values, many of them potentially at zero. 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, with particular emphasis on the CSV model with 94 financial time series from components of the S&P100 index. Section 4 discusses our contributions and findings and suggests directions for further research.

2 Prior specification for the state equation

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

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

where s_t is the latent (hidden) state-space variable. The error terms η_t and ε_t are independent random shocks in the observation and state equations respectively, usually Gaussian, and we observe the pairs $(x_t, y_t),\ t = 1, 2, \ldots, T$. Two fairly common specifications for the observation equation are *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.

As one would expect, 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. As stated earlier in the paper, one of our goals 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, particularly promoting sparsity-inducing flat-lining structures. In addition, we will specify a prior for s_0, the initial state.

Initial state s_0. 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:

$$s_0 \sim \gamma\, N(0, (cw)^2) + (1 - \gamma)\, N(0, w^2)$$

where $\gamma \sim \text{Bernoulli}(p_*)$, 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.

2.1 A 4-component mixture prior

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): $(\alpha, \beta) = (0, 0)$ - (sparse component)

Case (3): $\alpha \in \Re$, $\beta = 0$ - (flat-line component)

Case (4): $\alpha \in \Re$, $\beta \in (0, 1)$ - (stationary component).

Our prior for (α, β, τ) mixes over these four cases. Without loss of generality, and keeping in mind the stochastic volatility dynamics, we place zero prior weight on $\beta < 0$. In addition, we use stock returns in our applications and it is well known 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 that the state is fixed near zero, which is often a possibility of particular interest. For example, if the state-space component is a time-varying coefficient in dynamic regression, then the corresponding regressor has no effect, provided that τ is negligible. Case (3) 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, or a *flat-line*. Case (4) allows the state to vary in a stationary fashion.

A near constant state can be achieved with Case (1) or Case (3), 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 for joint distributions, we have $[\alpha, \beta, \tau] = [\beta]\ [\alpha, \tau \mid \beta]$. All the specifications we consider in this paper make the additional simplifying assumption that τ and α are independent given β; or, more specifically, that $[\alpha, \tau \mid \beta] = [\alpha \mid \beta]\,[\tau \mid \beta]$.

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

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

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 α is unrestricted and $\beta = 0$, and p_{uu} is the probability that α is unrestricted and $\beta \in (0, 1)$.

The prior of β: 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 $(0,1)$:

$$p(\beta) \propto p_N(\beta \mid \bar{\beta}, \sigma_\beta^2)\, \delta_{\{\beta \in (0,1)\}}. \tag{2}$$

where $p_N(\cdot \mid \bar{\beta}, \sigma_\beta^2)$ denotes a normal density with mean $\bar{\beta}$ and standard deviation σ_β and, as before, δ_x denotes the Dirac measure.

The prior of $\alpha|\beta$: For the prior $p(\alpha \mid \beta)$ we use,

$$\alpha \mid \beta \sim N(0, \sigma_\alpha^2\,(1-\beta^2)). \tag{3}$$

When $\beta = 0$, we simply have $\alpha \sim N(0, \sigma_\alpha^2)$. As β increases toward one, we shrink our prior down towards the Case (1), the random walk component, 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. 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.

The prior for $\tau|\beta$: The commonly used prior for τ^2 is the inverse gamma $\tau^2 \sim IG(\nu/2, \nu\lambda/2)$. However, we have empirically found that it was very difficult to choose values for ν and λ that gave consistently good results, especially for dynamic models with hundreds or thousands of state-space components. 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. This is a crucial advantage of our prior set up in scenarios where the majority, but not all, of the state-space components are actually time invariant.

More specifically, we will set the prior of τ on a finite 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. We first choose the minimum and the maximum values of τ, namely 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 $t_{n_g} = \tau_{max}$. Then, we set $P(\tau = \tau_{min}) \equiv p_{min}$. For $i > 1$, $P(\tau = t_i) \propto \exp(-c_\tau |t_i - \tau_{min}|)$. 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, since a negligible τ will produce sparse and flat-line state-space as well. 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. 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.

Figure 1 illustrates how our discrete prior compares to the square root of the commonly used inverse gamma prior. In the left panel we see that our prior density pushes hard towards small values of 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. Finally, 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 $\beta = 0$. In what follows we study a few prior specifications with various values of c_τ.

2.2 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 are 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 computational time.

Smoother and rougher prior specifications. 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, 30% chance of a constant state and a 20% of a time-varying stationary state. We call this configuration the *rougher* prior specification. The name *rougher* refers to the fact that in our applications this prior allows for substantial state variation. A smoother version, named *smoother* prior, 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 sets $p_{01} = 0.85$, $p_{00} = 0.05$, $p_{u0} = 0.05$, $p_{uu} = 0.05$, so placing much more weight on the random walk.

Much smoother and mimicking prior specifications. In some of our examples we also entertain two additional prior specifications. A third specification, which we named *much smoother* prior, changes the *smoother* prior by setting $\tau_{max} = 0.02$, $c_\tau = 300$ and $c_\tau^0 = 600$, so inducing even more smoothness. Finally, we also entertain a *mimicking* prior specification, where we set the 4-component mixture prior to "mimic' the $G(1/2, 1/2)$ prior used to model τ^2 (see Kastner et al.

[2017]) coupled with independent normal priors for α and β:

$$\alpha \sim N(0, 100), \quad \beta \sim N(0.5, 100) \text{ and } \tau^2 \sim G(1/2, 1/2).$$

As we already highlighted in Section 2.1 and Figure 1, the (conditionally conjugate) normal-inverse gamma prior for (α, β, τ^2) bounds τ^2 away from zero a priori. Hence, we decided to try to "mimic" the gamma prior as an example of a more standard prior specification.

Table 1 lists the subset of tuning parameters that vary across these four different prior specifications, while Figure 2 shows the marginal prior densities under the *smoother*, *rougher* and *mimicking* priors. Our prior setup can give us shrinkage towards small τ or more spread out prior as in the the mimicking case which is particularly important in larger systems. 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.

Appendix D contains a summary of our `R` code `csv` that accommodates these two default prior specifications. The *smoother* prior is set when `defpri=-1` (this is the default), while the `defpri=0` sets the prior as the *rougher* prior. In addition, we defer the details about our customized Markov Chain Monte Carlo algorithm for approximate posterior inference to the Appendix A.

2.3 Local level and dynamic regression models

In this section we illustrate the implementation of our 4-component mixture prior of (α, β, τ) for two normal dynamic linear models (NDLMs): the first order model and a multiple dynamic linear regression model. We chose these two models as representative ones of the vast and recent applied business and economics literature where 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 (see Primiceri [2005]) or in large-scale dynamic factor models; and time-varying variances and covariances in stochastic volatility (SV) models (see, Lopes and Carvalho [2007], Kastner et al. [2017]). Also, regularisation inducing strategies in TVP models recently appeared in Belmonte et al. [2014], Kalli and Griffin [2014], Bitto and Frühwirth-Schnatter [2019], Rocková and McAlinn [2018], Kowal

et al. [2018] and Uribe and Lopes [2018], amongst several others.

Local level model. We consider a standard simple local level model, also commonly known as first order DLM:

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

where η_t and ε_t are independent and identically distributed $N(0,1)$. We simulate series of length $T = 300$ with $\sigma = 0.1$ and $s_0 = 0$. We chose three combinations of (α, β, τ):

$$\begin{aligned} \text{Flat-line state} &: (\alpha,\beta,\tau) = (0.0, 0.0, 0.01) \\ \text{AR(1) state} &: (\alpha,\beta,\tau) = (0.0, 0.99, 0.02) \\ \text{Random walk state} &: (\alpha,\beta,\tau) = (0.0, 1.0, 0.01) \end{aligned}$$

For each one of the above three simulated datasets, we employ the three prior specifications, namely the *smoother* prior, the *rougher* prior and the *mimicking* prior, which were detailed in Section 2.2 and Table 1 and visualized in Figure 2. We run our `csv` code for 5000 iterations as burn-in and keep the next 2000 for posterior inference. Our findings are summarized in Figure 3. The right column shows summaries of the posterior draws of τ under each one of the above four prior specifications. As it can be argued that the *much smoother* and, to some extent, the *smoother* priors always lead to more concentrated posteriors near zero, more so when the actual state-space component is a flat-line. The rougher leads to a more flat-lined state when the actual state-space component is a flat-line, while the *mimicking* prior behaves as a much noisier version of the *rougher* prior. We also argue that both *smoother* and *rougher* priors (the default ones in our `csv` code) are sensible compromises between over fitting and over smoothing the state-space component. Finally, the *mimicking* prior grossly overestimates the size of the standard deviations τ, as we anticipated earlier when discussing the various priors in Figure 1.

Dynamic regression model. The model is written as:

$$\begin{aligned} y_t &= s_{t1}x_{t1} + s_{t2}x_{t2} + s_{t3}x_{t3} + \sigma\,\eta_t, \\ s_{ti} &= \alpha_i + \beta_i\, s_{t-1,i} + \tau_i\,\varepsilon_{ti}, \qquad i = 1,2,3, \end{aligned}$$

where η_t and ε_{ti} are independent and identically distributed $N(0,1)$. We fix $\sigma = 0.1$, $(\alpha_1, \beta_1, \tau_1) = (0, 0, 0.1)$, $(\alpha_2, \beta_2, \tau_2) = (0, 1, 0.1)$ and $(\alpha_3, \beta_3, \tau_3) = (0.1, 0.9, 0.2)$. Therefore, the three state-space components are, respectively, a flat-line at zero (x_1 is irrelevant), a random walk component and a stationary AR(1) component. The predictors x_1 and x_2 are standard normal, while x_3 is $N(5,1)$. Figure 4 show posterior summaries for each one of the four prior specifications we entertained in the previous example. On the one hand, our mixture prior specifications, the *much smoother*, the *smoother* and the *rougher* priors, accurately captures the flat-lined state-space component and smooths out the random walk component. On the other hand, the *mimicking* prior adapts well to the AR(1) state-space component and overestimates the variability of the random walk component. Once again, both *smoother* and *rougher* priors seem to be reasonable compromises that avoids over smoothing and over noisy cases.

One can argue that these differences can become irrelevant from a practical point of view when the number of state-space components is small, or even moderate, say less than a dozen. However, we show in the next section that when the total number of state-space components is much larger, say in the hundreds or thousands, such minor differences accumulate and might make estimation and forecasting less efficient. Below we illustrate these issues by considering a multivariate stochastic volatility model for about one hundred time series of asset returns. In this case, the final number of state-space components is around five thousand and leads to fifteen thousand (α, β, τ) parameters and, therefore, flat-lining can lead to substantial decrease in parameter estimation.

3 Time-varying covariances

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:

$$y_t|F_{t-1} \sim N(0, \Sigma_t), \tag{4}$$

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 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. In what follows, we briefly review the literature on time-varying covariance models, while Section 3.1 introduces our proposed Cholesky stochastic volatility (CSV) model.

Brief literature review. There are at least three ways to decompose the covariance matrix Σ_t. 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. This is the parametrization used in dynamic conditional correlation models (see, Tse and Tsui [2002] and Engle [2002], amongst many others).

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 [2006b], to name

just a few. Philipov and Glickman [2006b] 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. See Kastner [2019], for instance, who model daily log-returns of 300 S&P500 members via a sparse FSV model, where sparsity is achieved through the factor loadings matrix.

We take a third alternative, one that parametrizes the full-rank time-varying covariance matrix Σ_t via a Cholesky decomposition (Wu and Pourahmadi [2003], Huang et al. [2006] and Pourahmadi [2013]):

$$\Sigma_t = A_t H_t A_t',$$

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. Next we show that there will be $q(q+1)/2$ state-space components to be estimated and, consequently, $3q(q+1)/2$ static parameters, the (α, β, τ) of Section 2. As an example of these magnitudes, modeling $q = 30$ time series of (log) returns with the CSV structure, for example, leads to around 500 state-space components and about 1500 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. Our CSV model also is sensitive to pre-defined order of the variables in the Cholesky transformation. Nonetheless, we argue that its impact is less pronounced due to the full Cholesky transformation and not a truncated, factor-based transformation. We argue that in practice the number of factors is likely to be time varying. For instance, the number of common factors might be 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 3.1, particularly equations 5 – 8, 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 Section 2, coupled with the computational approach developed below, enables us to search for simplifying structure in a large system without imposing it.

3.1 Cholesky stochastic volatility model

Below 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 $i > j$. 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

$$y_{1t} \sim N(0, \omega_{1t}^2) \tag{5}$$

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

$$y_{it} \sim N(\phi_{i1t} y_{1t} + \phi_{i2t} y_{2t} + \cdots + \phi_{i(i-1)t} y_{(i-1)t}, \omega_{it}^2). \tag{6}$$

Once ϕ_{ijt}s and ω_{it}^2s are available, so are A_t^{-1}, 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,

$$\phi_{ijt} \sim N(\alpha_{ij} + \beta_{ij}\, \phi_{ij(t-1)}, \tau_{ij}^2) \tag{7}$$

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

$$d_{it} \sim N(\alpha_i + \beta_i\, d_{i(t-1)}, \tau_i^2) \tag{8}$$

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 (7) and (8), while the recursive conditional regressions (or simply *triangular regressions*) are our observation in the observation equations (5) and (6). Our *Cholesky stochastic volatility* (CSV) model comprises equations (5) to (8). The connection between the state-space components ϕ_{ijt}s and d_{it}s and the parameters $(\alpha_i, \beta_i, \tau_i)$ and $(\alpha_{ij}, \beta_{ij}, \tau_{ij})$ and the set up of Section 2 is straightforward.

Other Cholesky-based models. The Cholesky decomposition approach has been studied elsewhere. On the one hand, various lasso-type regularization strategies for sparse modeling of covariance matrices appear, amongst others, in Levina et al. [2008], Rothman et al. [2010] and Leng and Li [2011]. For time varying covariance matrices, on the other hand, Uhlig [1997] and Philipov and Glickman [2006a], 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 DLMs. Philipov and Glickman [2006a] 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 [2006a] 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 [2019] 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].

3.2 Low-dimensional illustrations

In this section, we illustrate the performance of our CSV approach on i) simulated data with $q = 3$ and ii) real data with $q = 9$. In case i), we can assess the performance by comparing the CSV fit to the known true Σ_t. We make the true process rather smooth in order to see how adaptive our 4-component mixture prior handles such an extreme scenario. In case ii) we compare the results from our parsimony-inducing CSV model to the well-known factor stochastic volatility (Lopes and Carvalho [2007], Kastner et al. [2017] and Kastner [2019]). We also empirically illustrate how standard deviations and correlations are affected by the order in which the time series appear in our CSV model. 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 3.3

3.2.1 Smooth or piecewise constant covariance dynamics

To gain insight into the proposed analysis, in this section we present results from two simple simulated exercises where the covariance matrix Σ_t evolves smoothly over time (first case) or in a piece-wise constant fashion (second case).

Smooth case: We let $q = 3$, $\Sigma_1 = I_q$, Σ_T with standard deviations $\sigma_1 = 1.0$, $\sigma_2 = 0.81$, $\sigma_3 = 0.25$, and correlations $\rho_{12} = \rho_{23} = 0.95$, $\rho_{13} = 0.9025$ and $\Sigma_t = (1 - w_t)\,\Sigma_1 + w_t\,\Sigma_T$ 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 $T = 2500$ observations. We run our `R` package `csv` under three prior specifications. The first two ones are the default rougher prior and smoother prior specifications of Section 2.3. We have added a third specification, namely *much smoother* prior, where $\tau_{max} = 0.02$, $c_\tau = 300$ and $c_\tau^0 = 600$, which induce even more smoothness. Results appear in Figure 5. Naturally, the posterior medians from the *much smoother* prior are the ones that more closely follows the truth. Both smoother and rougher priors produce reasonable estimates, but both more noisy than the smoother prior.

Piecewise constant case: Here, we let $T = 1000$ and $q = 5$, while the covariance matrix Σ_t assumes one of four possible configurations. More specifically, $\Sigma_t = \Sigma_1^0$ for $t = 1, \ldots, 250$, $\Sigma_t = \Sigma_2^0$ for $t = 251, \ldots, 500$, $\Sigma_t = \Sigma_3^0$ for $t = 501, \ldots, 750$ and $\Sigma_t = \Sigma_4^0$ for $t = 751, \ldots, T$. $\Sigma_{l,ij}^0 = \rho_l^{|i-j|}$, for $l = 1, 2, 3, 4$ and $i, j = 1, \ldots, q$. Basic correlations are $\rho_1 = 0.3$, $\rho_2 = 0.5$, $\rho_3 = 0.7$ and $\rho_4 = 0.9$. Figure 6 shows posterior summaries for correlations, standard deviations as well as ϕ coefficients from the Cholesky parametrisation. For illustration, we compare to a factor stochastic volatility (FSV) model with $k = 2$ common factors. As expected, the posterior estimates of the ϕ coefficients collapse at zero when the true coefficients are zero.

3.2.2 Order of the time series

The two main goals of the following empirical applications are: i) Compare the posterior summaries for time-varying standard deviations and correlations based on the various prior specifications we have outlined in the paper for the CSV model, as well as those from the factor stochastic volatility models implemented in the `R` packages `factorstochvol` (see Kastner et al. [2017] and Kastner [2019]).

S&P100 data. We randomly selected $q = 9$ time series of returns (companies) out of the 100 components of the S&P100 index. Posterior summaries are presented in Figures 7, 8 and 9. 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 practically different covariance

estimates. More specifically, the $q = 9$ S&P100 time series, $y_1, \ldots, y_9$, was fit under two different orders: the original order, $y_1, \ldots, y_9$, and the full reverse order, $y_9, \ldots, y_1$. Figures 8 and 9 present posterior medians for correlations and Cholesky time-varying coefficients. We argue that the correlations are quite similar regardless of the type of prior smoothing strategy is implemented. Nonetheless, when comparing ϕ_{ijt} coefficients, is becomes apparent, as expected, that the order of the time series matters as many coefficients in on order becomes flat-lined in the reserve order. This is more pronounced when we use the smoother prior and seen in the bottom row of Figure 9. 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. Figures 8 and 9 suggest 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 at least, 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.

Figure 10 exhibits the posterior densities for the standard deviations τ corresponding to all ϕ coefficients. As expected, the rougher priors lead to similar rougher posteriors. Nonetheless, even in the rougher prior set up, many ϕ coefficient flat-line and the corresponding τ shrinks toward zero. Finally, Figure 11 plot through time the posterior means of the weights for the global minimum variance portfolio based on the much *smoother* prior discussed above in Section 3.2.1. 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. For instance, some of weights are always zero or positive, while others start at zero and half way through become positive. Both CSV and FSV models produce similar weights, with the FSV ones more jittered for some of the financial assets.

Exchange rate returns data. Figure 12 compares our CSV model with the FSV with $k = 4$ common factors for a subset of $q = 9$ exchange times series randomly selected from the 23 ones

from the database `exrates`, included in the `R` package `stochvol` (see, Kastner et al. [2017]). More specifically, the exchange rates that were selected correspond to the Canadian dollar (CAD), UK pound (GBP), Hong Kong dollar (HKD), Mexican Peso (MXN), New Zealand dollar (NZD), Poland zloty (PLN), Sweden krona (SEK), Singapore dollar (SGD) and the US dollar (USD). The data spans from January 3rd, 2000 to April, 4th 2012, with a total of $T = 3140$ observations. We compare cumulative portfolio performance and cumulative likelihood ratios relative to the FSV(4) model. The comparisons are based on posterior point estimates of Σ_t, say $\widehat{\Sigma}_t$. More specifically, we compute $\sum_{l=1}^{t} \log p(y_l|y_{l-1}, \widehat{\Sigma}_l)$ for $t = t_0, \ldots, T$ and some $t_0 \geq 1$. Similarly, cumulative portfolio returns are computed from point estimates of portfolio weights, i.e. $\widehat{\omega}_t = \widehat{\Sigma}_t^{-1} 1_q / 1_q' \widehat{\Sigma}_t^{-1} 1_q$. The CSV model under the *much smoother* and *smoother* prior specifications lead to superior or similar portfolio performances when compared to the FSV(4) model. Also, when comparing cumulative likelihoods, the *much smoother* and the *mimicking* priors are the winners, but all CSV prior specifications lead to superior performance when compared to the FSV(4) model. Similar results were obtained when comparing to FSV(3), while FSV(1) and FSV(2) have much worse performance overall.

3.3 High-dimensional illustration

In this final illustration, we use asset returns from firms making up the S&P100 index in order to show that our CSV framework is particularly suited to deal with moderate to large settings. More specifically, we stress that one of the crucial strengths of our proposed Cholesky SV framework is that the triangular representation of the model naturally leads to parallelization in the MCMC scheme. We refer to Appendix C 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, the entire system is need, i.e. we need to compute the full covariance matrix Σ_t.

First case - $q = 20$: We first consider a selection of returns on $q = 20$ of the firms and use both the *smoother* prior and the *much smoother* prior discussed in Section 3.2.1. We use these priors since more smoothing may be desirable for larger q. Figure 13 plots the posterior means of the d and ϕ states that represent the Cholesky SV framework. As expected, the time variation in the residual variances visible, while the ϕ series have relatively little time variation and are centered

near zero. However, under the *smoother* prior, a few of the ϕ 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. The amount of "parsimony", "smoothness", or "regularization" inevitably is heavily influenced by our choice of prior. Under the *much smoother* prior, the "flat-line" appearance of many of the ϕ states is striking. The results are quite similar to those shown by Figure 9, when $q = 9$ and two orders of the time series were investigated. 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 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.

Second case - $q = 94$: Finally, Figure 14 reports results for $q = 94$ assets using the *much smoother* prior specification. 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 $q = 94$ latent standard deviation processes (σ_{it}) and $q(q-1)/2 = 4,371$ latent correlation processes (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. 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 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.

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 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. 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. We empirically showed that our approach is comparable to the popular factor stochastic volatility (FSV) methodology and gives stable intuitively plausible MCMC results.

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. 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 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], for static factor models, and Kastner et al. [2017], for factor SV models.

Both shrinkage and time-evolution of factor loadings in moderate and large scale factor stochastic volatility and related models has emerged over the last decade. See, amongst others, Lopes and Carvalho [2007], Belmonte et al. [2014], Zhao et al. [2016], Kastner et al. [2017], Bitto and Frühwirth-Schnatter [2019], Rocková and McAlinn [2018], Kowal et al. [2018] and Uribe and Lopes [2018] and Kastner [2019]. 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 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 $\beta = 1$ 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 (Equa-

tion 9) 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 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

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

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

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

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

M. A. Belmonte, G. Koop, and D. Korobilis. Hierarchical shrinkage in time-varying parameter models. *Journal of Forecasting*, 33(1):80–94, 2014.

A. Bitto and S. Frühwirth-Schnatter. Achieving shrinkage in a time-varying parameter model framework. *Journal of Econometrics*, 210(1):75–97, 2019.

C. M. Carvalho, H. F. Lopes, and R. E. McCulloch. On the long run volatility of stocks. *Journal of the American Statistical Association (in press)*, 2018.

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

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

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

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

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

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

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

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

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

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

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

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

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

J. Z. Huang, N. Liu, M. Pourahmadi, and L. Liu. Covariance matrix selection and estimation via penalised normal likelihood. *Biometrika*, 93(1):85–98, 2006.

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

G. Kastner. Sparse Bayesian time-varying covariance estimation in many dimensions. *Journal of Econometrics*, 210(1):98–115, 2019.

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

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

D. R. Kowal, D. S. Matteson, and D. Ruppert. Dynamic shrinkage processes. Technical report, 2018.

C. Leng and B. Li. Forward adaptive banding for estimating large covariance matrices. *Biometrika*, 98(4):821–830, 2011.

E. Levina, A. Rothman, and J. Zhu. Sparse estimation of large covariance matrices via a nested lasso penalty. *The Annals of Applied Statistics*, 2(1):245–263, 2008.

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

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

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

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

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

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

F. Nardari and J. T. Scruggs. Bayesian analysis of linear factor models with latent factors, multivariate stochastic volatility, and apt pricing restrictions. *Journal of Financial and Quantitative Analysis*, 42:857–892, 2007.

A. Philipov and M. E. Glickman. Multivariate stochastic volatility via wishart processes. *Journal of Business and Economic Statistics*, 24:313–328, 2006a.

A. Philipov and M. E. Glickman. Factor multivariate stochastic volatility via wishart processes. *Econometric Reviews*, 25:311–334, 2006b.

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

M. Pourahmadi. *High-Dimensional Covariance Estimation*. John Wiley & Sons, 2013.

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

V. Rocková and K. McAlinn. Dynamic variable selection with spike-and-slab process priors. Technical report, Booth School of Business, University of Chicago, 2018.

A. Rothman, E. Levina, and J. Zhu. A new approach to cholesky-based covariance regularization in high dimensions. *Biometrika*, 97(3):539–550, 2010.

A. M. Schmidt and H. F. Lopes. Dynamic models. In A. E. Gelfand, M. Fuentes, J. A. Hoeting, and R. L. Smith, editors, *Handbook of Environmental and Ecological Statistics*, pages 57–80. Chapman & Hall/CRC, Boca Raton, FL, USA, 2019.

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

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

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

H. Uhlig. Bayesian vector autoregressions with stochastic volatility. *Econometrica*, 65:59–73, 1997.

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

W. B. Wu and M. Pourahmadi. Nonparametric estimation of large covariance matrices of longitudinal data. *Biometrika*, 90(4):831–844, 2003.

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

A State equation: MCMC 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 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. Posterior inference is obtained by cycling through the two Gibbs full conditional distributions

$$[(s_0, s) \mid (\alpha, \beta, \tau), y, x] \quad \text{and} \quad [(\alpha, \beta, \tau) \mid (s_0, s)]. \tag{9}$$

That is, we draw the state-space components conditional on the AR(1) parameters and then draw the AR(1) parameters conditional on the state-space components. To draws the whole vector state-space components s, we use the well known *forward filtering, backward sampling* (FFBS) algorithm (see, for instance, 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 (9) 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 9 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 (9). 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 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 3. We again look for a simple approach and discretize the prior 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 (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 (9) may necessitate a large number of runs.

We emphasize that the most crucial aspect of our prior is the prior on τ having the properties 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 inverse gamma prior for τ^2, i.e. $\tau^2 \sim IG(\nu/2, \nu\lambda/2)$. 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 inverse gamma and its problematic use in state space models, see Frühwirth-Schnatter [2004] and Frühwirth-Schnatter and Wagner [2010].

Initial state s_0. The basic Gibbs sampler (9) is modified by adding the draw of the latent variable

$$[\gamma \mid (s_0, s), (\alpha, \beta, \tau), y, x] = [\gamma \mid s_0],$$

then a draw from $[(s_0, s) \mid \gamma, (\alpha, \beta, \tau), y, x]$ from $[s_0 \mid s, \gamma, (\alpha, \beta, \tau)] = [s_0 \mid s_1, \gamma, (\alpha, \beta, \tau)]$ and, finally, a draw form $[s \mid \gamma, (\alpha, \beta, \tau), y, x]$. Conditional on a draw of γ, we have a normal prior for the initial state with mean zero and standard deviation w (when $\gamma = 0$) or cw (when $\gamma = 1$).

B CSV: MCMC implementation

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 (6). 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: $(d_{i0}, d_i) \mid \circ$,

ii. ϕ states: $(\phi_{ij0}, \phi_{ij}) \mid \circ$,

iii. d AR parameters: $(\alpha_i, \beta_i, \tau_i) \mid \circ$, and

iv. ϕ AR parameters: $(\alpha_{ij}, \beta_{ij}, \tau_{ij}) \mid \circ$.

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,

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

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

Thus, all the ϕ_{ij} draws reduce to simple applications of the forward filtering backward sampling (FFBS) algorithm 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 $(s_0, s) \mid (\alpha, \beta, \tau)$ and $(\alpha, \beta, \tau) \mid (s_0, s)$, 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.

C 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 Appendix B).

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.

D Prior setup in R package csv

Recalling the set up of Section 2.3, In the univariate state-space model with observation equation $y_t = f(x_t, s_t, \eta_t)$ and state equation $s_t = \alpha + \beta s_{t-1} + \tau\varepsilon_t$, 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}=tauminp, 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.

Finally, the prior on the initial state 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.

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)
```

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.

List of Tables

		Prior set up			
Parameter	`csv` name	*much smoother*	*smoother*	*rougher*	*mimicking*
τ_{min}	`taumin`	0.005	0.005	0.001	0.001
τ^0_{min}	`taumin0`	0.001	0.001	0.01	0.001
τ_{max}	`taumax`	0.02	0.05	0.15	6
p_{min}	`tauminp`	0.50	0.50	0.50	0.001
c_τ	`tauc`	300	200	100	1.25
c^0_τ	`tauc0`	600	400	200	1.25
σ_α	`sa`	2.0	2.0	2.0	10
$\bar{\beta}$	`bbar`	1.0	1.0	1.0	0.5
σ_β	`sb`	1.0	1.0	1.0	10
p_{00}	`p00`	0.05	0.05	0.15	0.01
p_{u0}	`pu0`	0.05	0.05	0.15	0.01
p_{01}	`p01`	0.85	0.85	0.5	0.01

Table 1: *4-component mixture prior for AR parameters* – Hyper-parameters of the *much smoother*, *smoother*, *rougher* and *mimicking* prior specifications presented in Section 2.2. The densities are depicted in Figure 2.

List of Figures

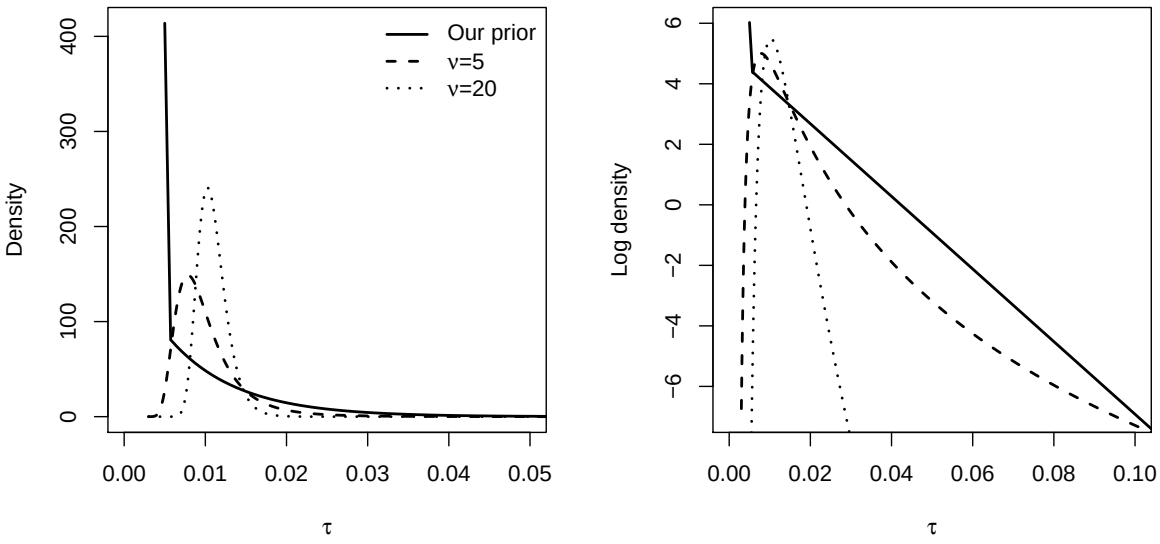

Figure 1: *Prior for* τ: Here $\tau_{min} = 0.005$, $\tau_{max} = 0.15$, $p_{min} = 0.3$, $c_\tau = 120$. In each panel the solid line is our τ prior and the other two correspond to the densities for τ derived from inverse gamma densities for τ^2 with λ equal to the square of $E(\tau)$ under our prior and ν equal to 5 (dashed lines) or 20 (dotted lines). These are, respectively, $IG(2.5, 0.0002)$ and $IG(10, 0.001)$. When $\tau^2 \sim IG(\nu/2, \nu\lambda/2)$, it follows that the density of τ is $p(\tau) = [(\nu\lambda/2)^{\nu/2}/\Gamma(\nu/2)]\tau^{-(\nu+1)}\exp\{-0.5\nu\lambda/\tau^2\}$. 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.

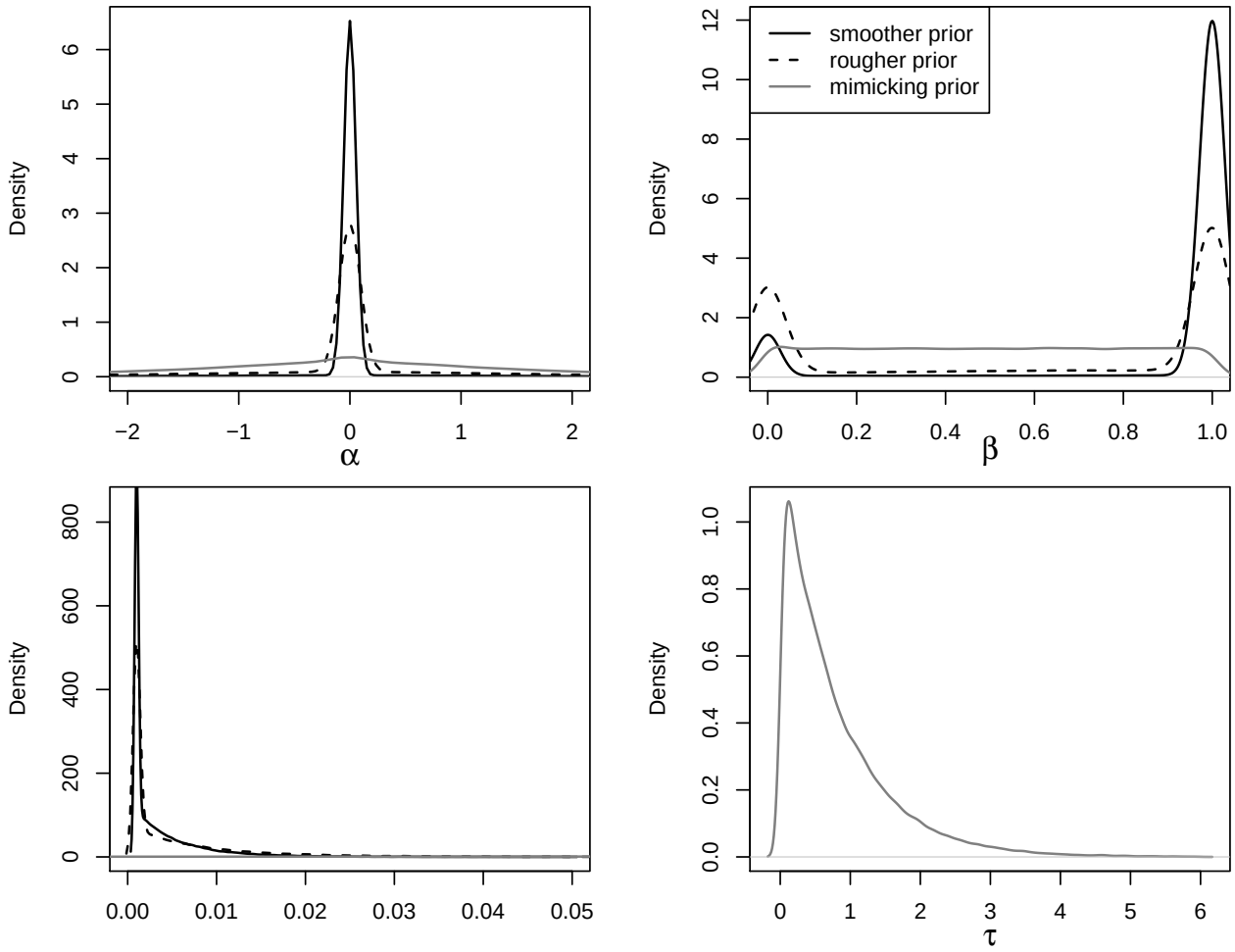

Figure 2: *4-component mixture prior for AR parameters* – Marginal prior distributions for α (top left panel), β (top right panel) and τ (bottom left panel) for the three prior specifications: *smoother*, *rougher* and *mimicking* priors presented in the text of Section 2.2 and in Table 1. Given the extreme differences in variation, the bottom right panel is again the marginal prior of τ under the *mimicking* prior.

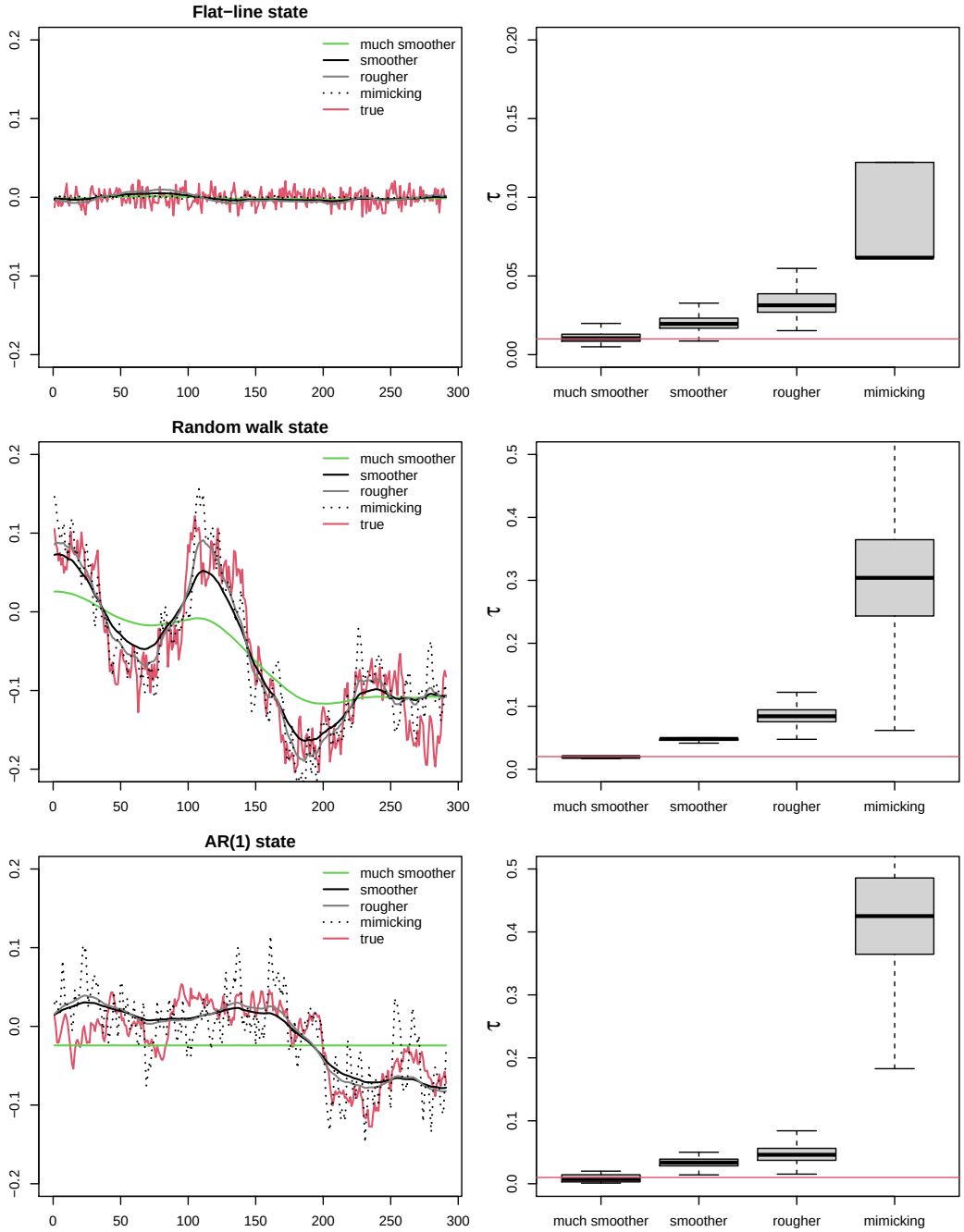

Figure 3: *Local level model* – Each row shows the results for a different configuration of the local level model: flat-line state component (top row), AR(1) state component (middle row) and random walk state component (bottom row). Posterior medians of the state-space components are presented in the left column, while posterior summaries for the state-space standard deviations, τ, appear in the right column. For each configuration (flat-line, AR(1) and random walk), we fit the local level models based on the four priors discussed in the text: *much smoother*, *smoother*, *rougher* and *mimicking* priors (see Table 1).

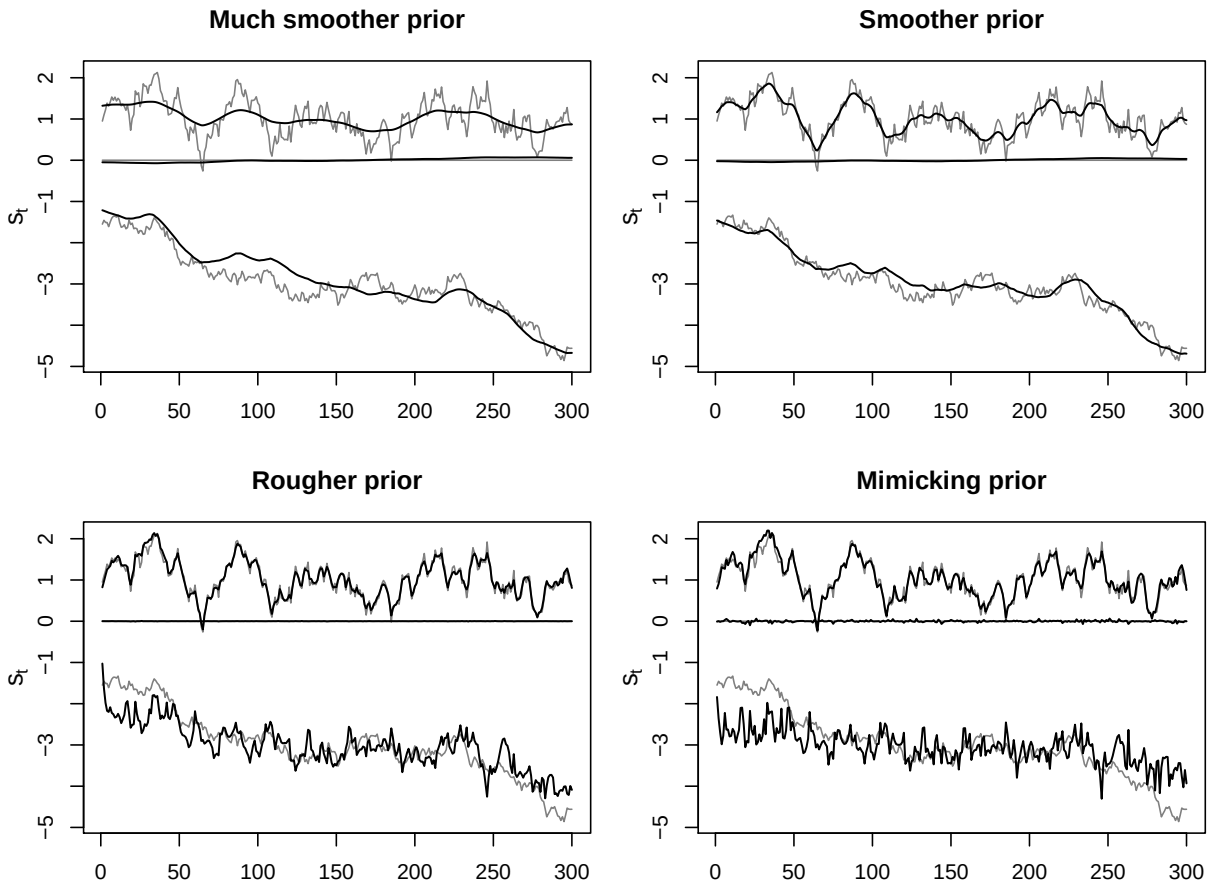

Figure 4: *Dynamic regression model* – Each plot shows posterior medians of the three state-space components, s_{t1}, s_{t2} and s_{t3} based on each one of the four prior specifications outlined in Section 2.2: *much smoother* prior (top left), *smoother* prior (top right), *rougher* prior (bottom left) and *mimicking* prior (bottom right).

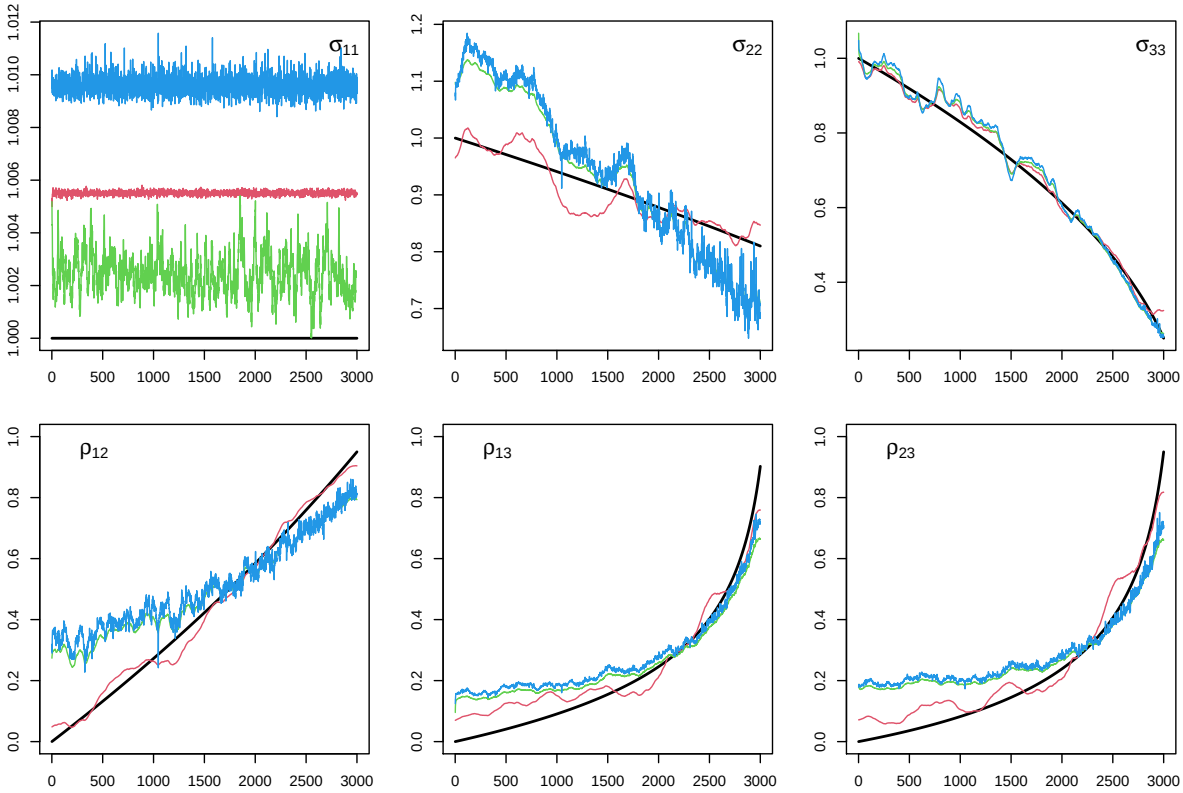

Figure 5: *Smooth covariance dynamics* – Standard deviations are on the top row and correlations are on the bottom row. True values (thicker black lines), *much smoother* prior (red lines), *smoother* prior (green lines) and *rougher* prior (blue lines).

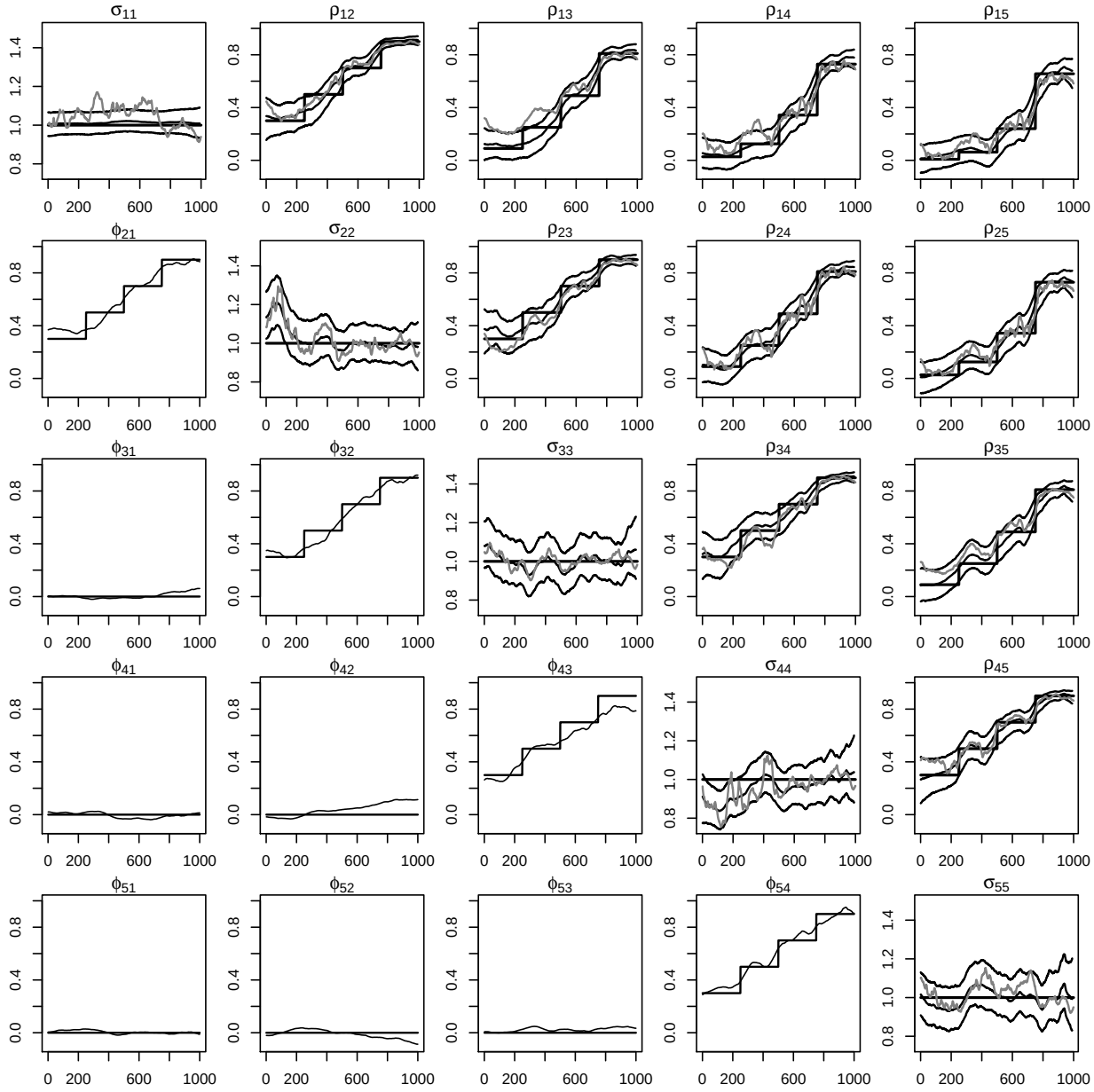

Figure 6: $q = 5$ *example* – Structural break. $n = 1000$ with $\Sigma_t = \Sigma_1^0$ constant for $t = 1, \ldots, 250$, $\Sigma_t = \Sigma_2^0$ constant for $t = 251, \ldots, 500$, $\Sigma_t = \Sigma_3^0$ constant for $t = 501, \ldots, 750$ and $\Sigma_t = \Sigma_4^0$ constant for $t = 750, \ldots, n$. $\Sigma_{l,ij}^0 = \rho_l^{|i-j|}$, for $l = 1, 2, 3, 4$ and $i, j = 1, \ldots, q$. Basic correlations are $\rho_1 = 0.3$, $\rho_2 = 0.5$, $\rho_3 = 0.7$ and $\rho_4 = 0.9$. The flat-lines and the step-lines are true values of ϕ_{tij} and ρ_{tij}. CSV are the thicker lines and FSV with $k = 2$ are the grey lines.

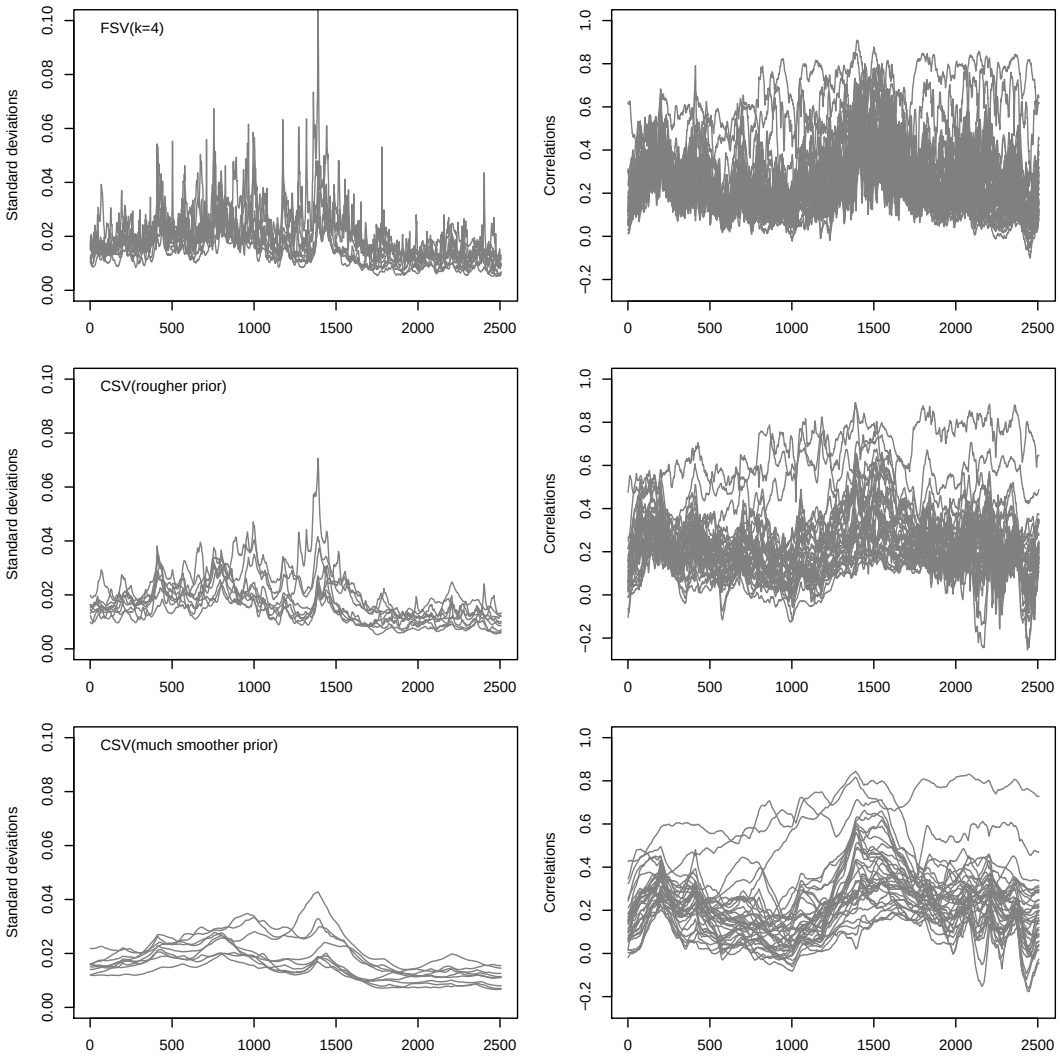

Figure 7: $q = 9$ *example* – Comparing the factor stochastic volatility model with $k = 4$ common factors (top row), with the Cholesky stochastic volatility models based on two priors: the *rougher* prior (middle row) and the *much smoother* prior (bottom row). Standard deviations are in the left column and correlations are in the right column.

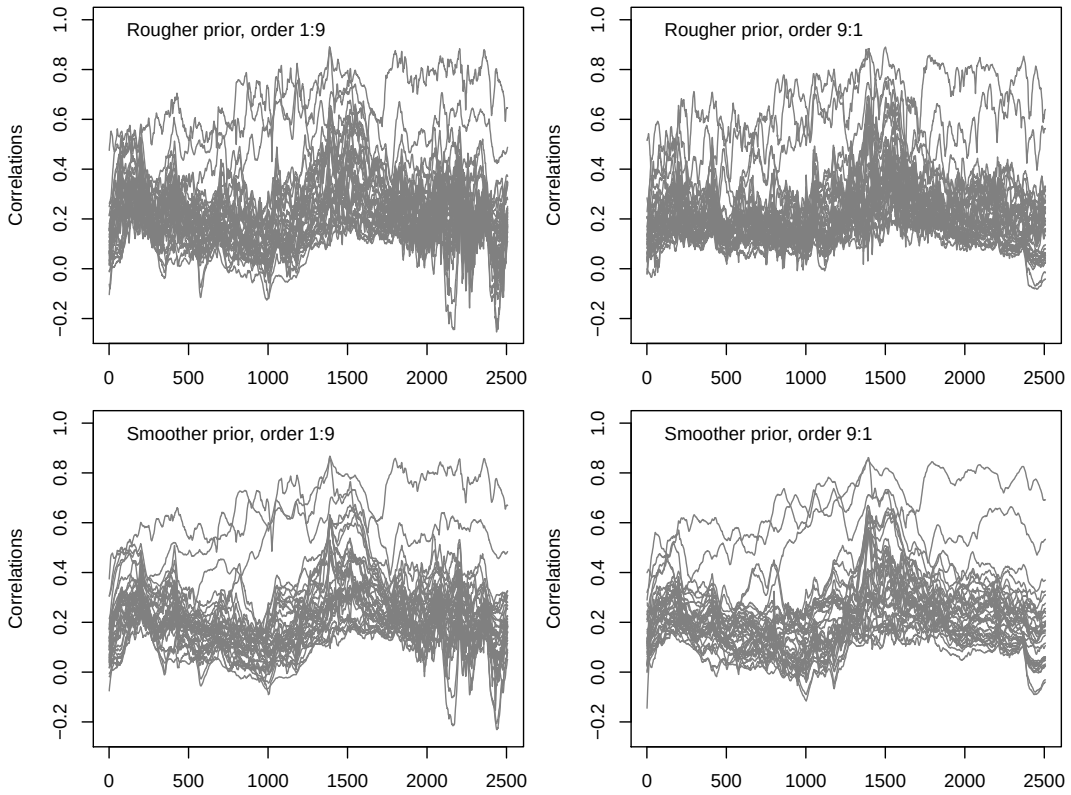

Figure 8: $q = 9$ *example* – Posterior medians of correlations. *Rougher* prior (top row) and *smoother prior* (bottom row). Two orders for the time series: $(1, 2, \ldots, 9)$ (left column) and $(9, 8, \ldots, 1)$ (right column).

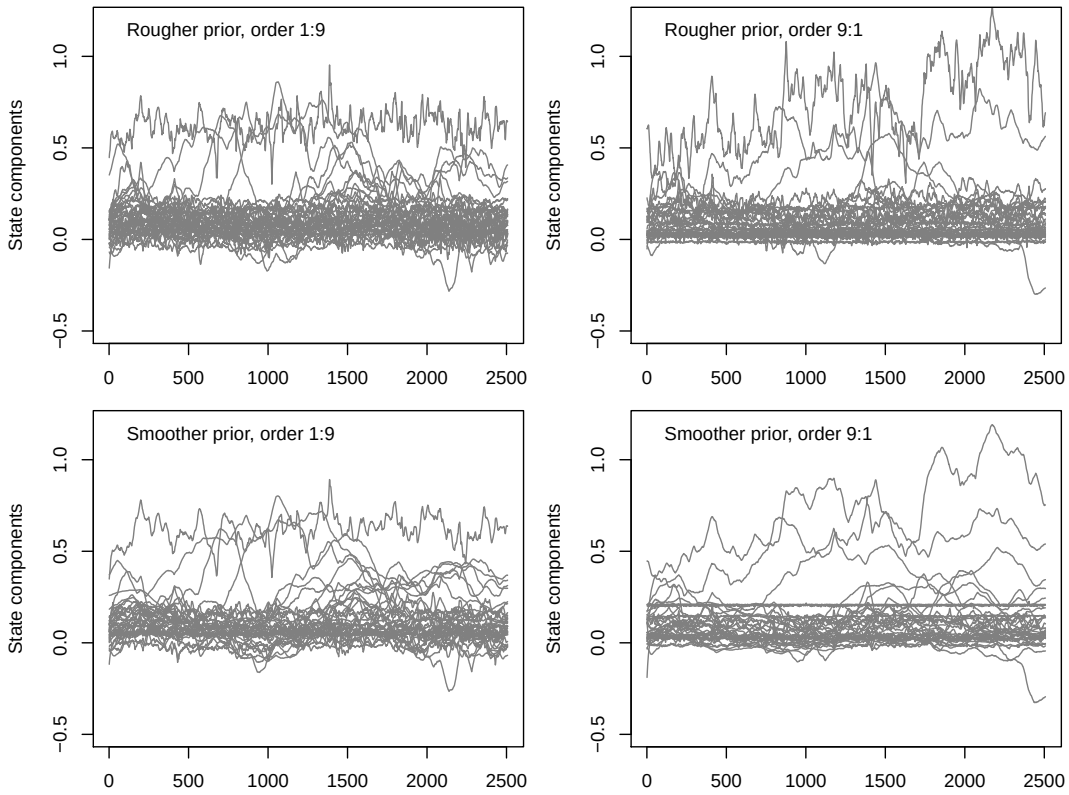

Figure 9: $q = 9$ *example* – Posterior medians of ϕ-states. *Rougher* prior (top row) and *smoother prior* (bottom row). Two orders for the time series: $(1, 2, \ldots, 9)$ (left column) and $(9, 8, \ldots, 1)$ (right column).

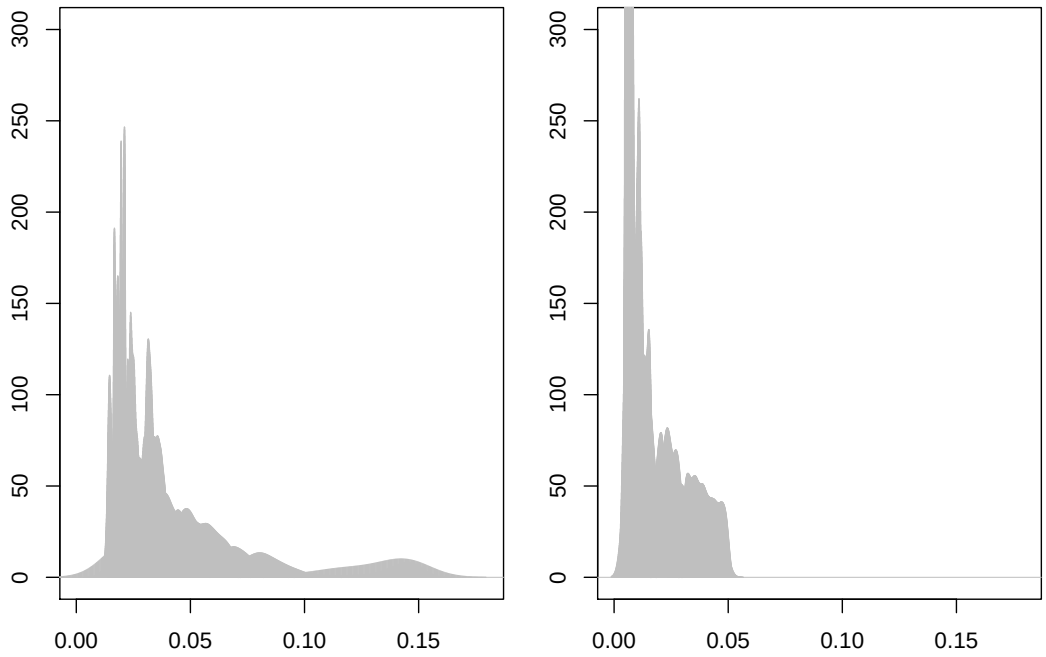

Figure 10: $q = 9$ *example* – Posterior densities of the standard deviations τ of the ϕ-states. *Rougher* prior (left frame) and *smoother* prior (right frame).

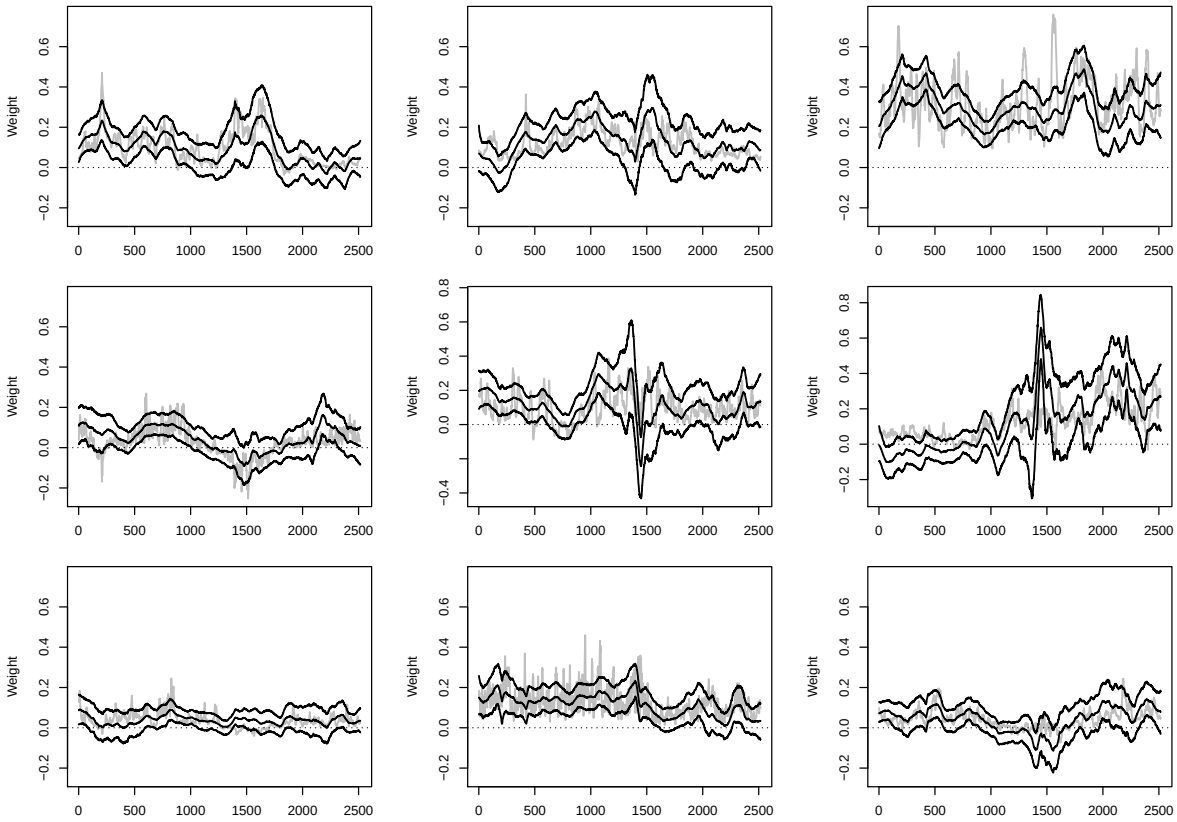

Figure 11: $q = 9$ *example* – Global minimum variance portfolio weights. Comparing FSV with $k = 4$ and CSV with the *much smoother* prior.

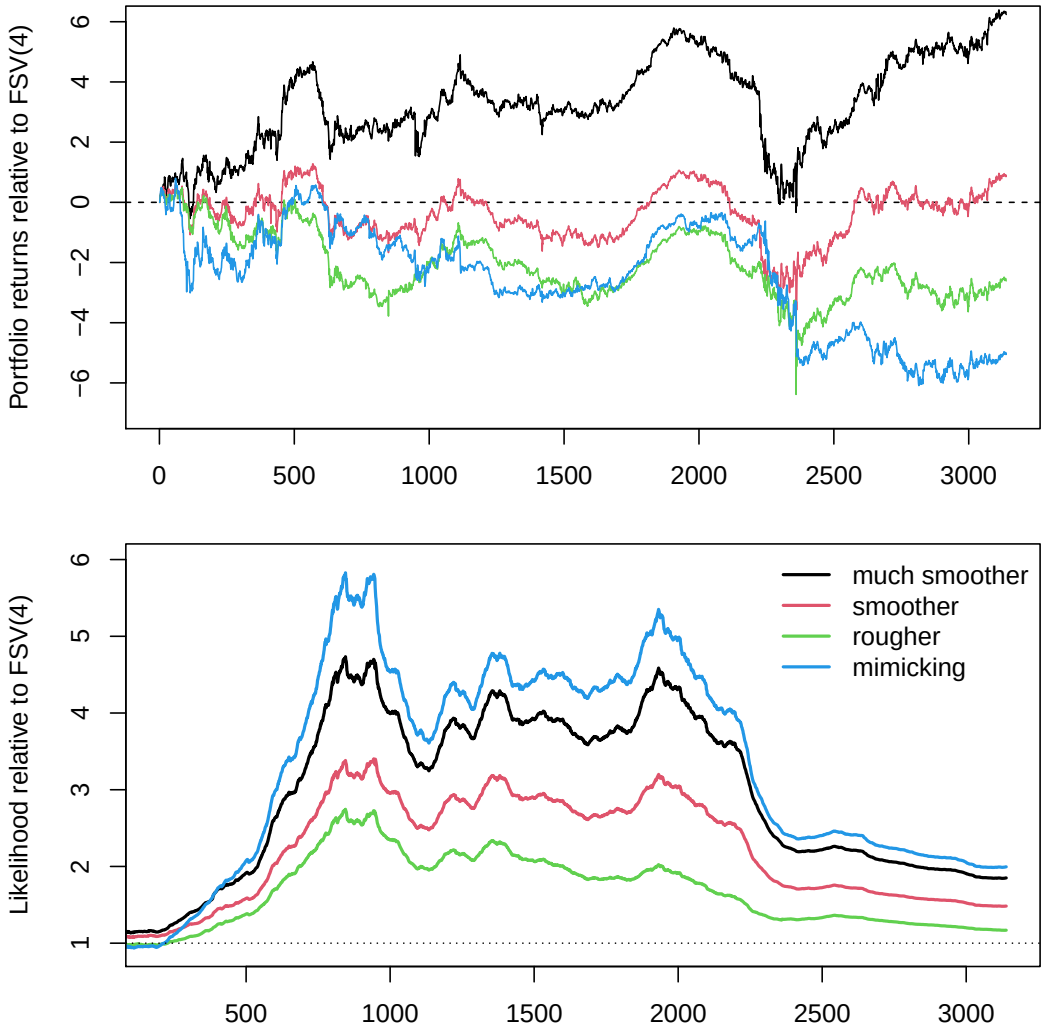

Figure 12: *Exchange rates example* – Global minimum variance portfolio cumulative performance and cumulative likelihood for CSV models under the four prior specifications detailed in the text and against a factor stochastic volatility model with $k = 4$ common factors.

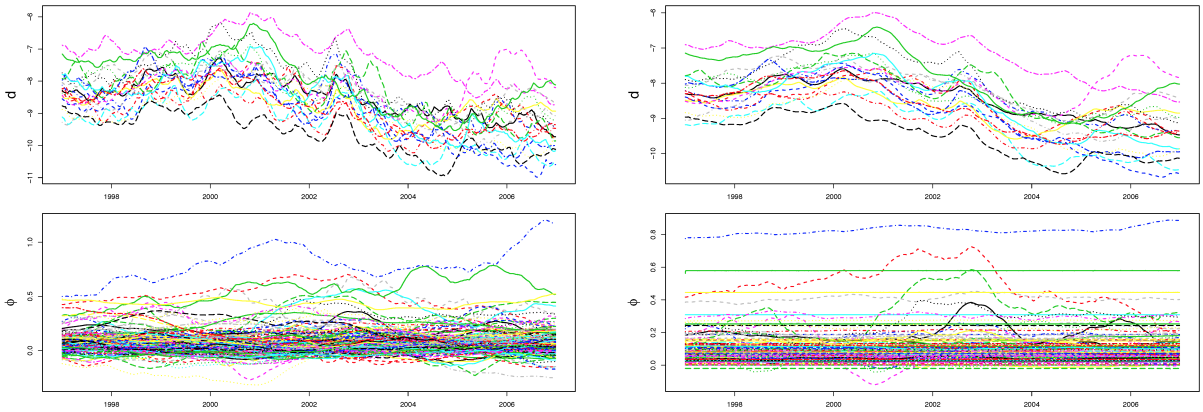

Figure 13: *S&P100 data,* $q = 20$ - Posterior means of the d-states (top row) and the ϕ-states (bottom row). Posteriors are based on the *smoother* prior (left column) and the *much smoother* prior (right column) specifications.

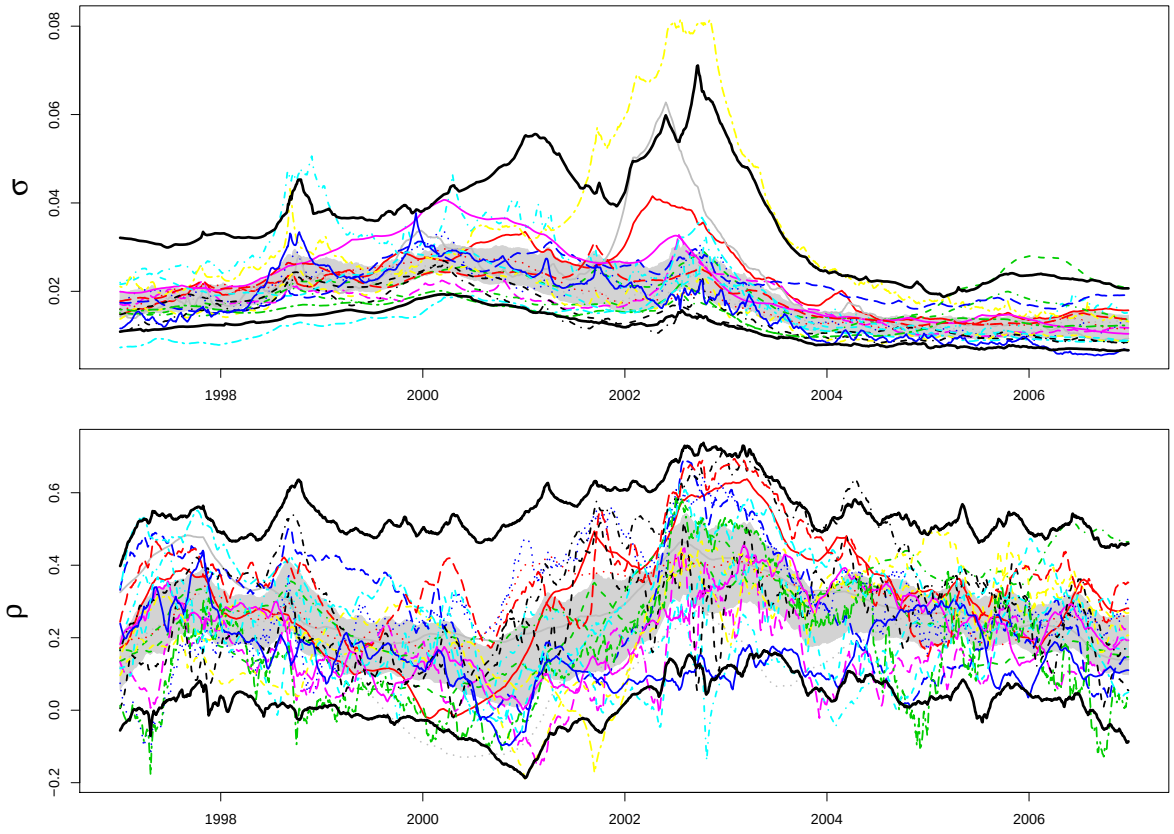

Figure 14: *S&P100 data,* $q = 94$ - Posterior means of time-varying standard deviations (top frame) and correlations (bottom frame), based on our *much smoother* prior specification.

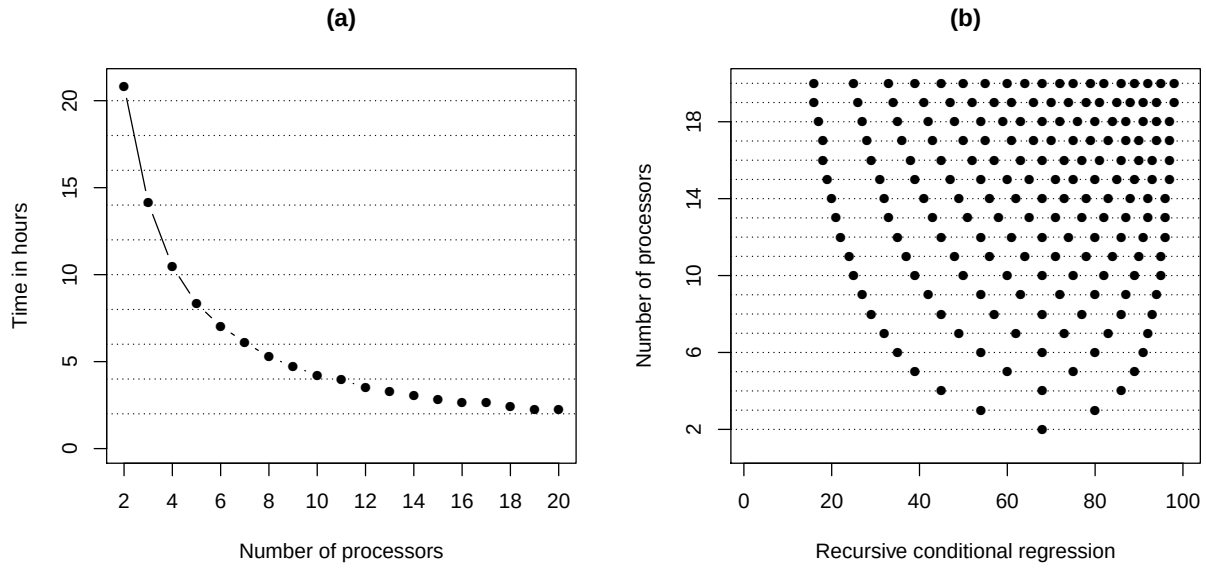

Figure 15: *Multiple processors* – In panel (a) we plot the number of processors vs. the total time in hours to run 50,000 iterations for a 100×100 ($q = 100$) time varying covariance matrix with $T = 2,516$. It takes about 13.5 times longer to draw a d state than it does to draw a ϕ state. Code was 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.