Cholesky Stochastic Volatility

H. F. Lopes, R. E. McCulloch and R. S. Tsay*

September 11, 2011

Abstract

Multivariate volatility has many important applications in finance, including asset allocation and risk management. Estimating multivariate volatility, however, is not straightforward because of two major difficulties. The first difficulty is the curse of dimensionality. For p assets, there are $p(p+1)/2$ volatility and cross-correlation series. In addition, the commonly used volatility models often have many parameters, making them impractical for real application. The second difficulty is that the conditional covariance matrix must be positive definite for all time points. This is not easy to maintain when the dimension is high. In this paper, we develop a new approach to modeling multivariate volatility. We name our approach *Cholesky Stochastic Volatility (CSV)*. Our approach is Bayesian and we carefully derive the prior distributions with an appealing practical flavor that allows us to search for simplifying structure without placing hard restrictions on our model space. We illustrate our approach by a number of real and synthetic examples, including a real application based twenty of the S&P100 components.

KEY WORDS: Multivariate; Time series; Covariance, Time-Varying.

*Hedibert F. Lopes is Associate Professor of Econometrics and Statistics, Booth School of Business, University of Chicago, 5807 S. Woodlawn Ave, Chicago, IL 60637, hlopes@chicagobooth.edu. Robet E. McCulloch is Professor of Statistics, IROM Department, University of Texas at Austin, robert.mcculloch1@gmail.com. Ruey S. Tsay is H. G. B. Alexander Professor of Econometrics and Statistics, Booth School of Business, University of Chicago, 5807 S. Woodlawn Ave, Chicago, IL 60637, ruey.tsay@chicagobooth.edu.

Contents

1 Introduction

Since the pioneering works of Rosenberg (1972) and Engle (1982), models for the evolution of the variance have played a central role in time series analysis. Financial time series in particular, clearly exhibit differing levels of variability as time progresses. Stochastic volatility uses a state-space model approach. The latent state is the conditional variance of the observed series and the state equation describes how this conditional variance evolves over time. In this paper, we study multivariate stochastic volatility. We observe a multivariate time series and our latent state is the conditional covariance matrix which evolves over time. We wish to be able work with a large number of series without placing restrictions on the form of the matrices beyond the intrinsic constraint that covariance matrix at each time be positive definite.

Clearly, we face severe computational and modeling challenges. We focus on a strategy that allows for straightforward use of parallel computing. We use a Bayesian approach and develop priors that allows us to search for simplifying structure without placing hard restrictions on the model space. Our starting point is the commonly employed approach of reparameterizing the covariance as a series of regressions. That is, we model the joint distribution as a marginal and then a series of conditionals. Given the basic assumption of multivariate normality, each conditional is a linear regression. We then let all the parameters of each regression be time-varying. Posterior computations can then be done using the well-known Forward-Filter, Backward-Sampler (FFBS) algorithm of Carter & Kohn (1994) and Frühwirth-Schnatter (1994). Posteriors of time-varying residual variances are computed using the method proposed by Kim, Shephard & Chib (1998). While the basic components of our approach are well known, we describe how to optimally use parallel computations and document the gains in time. Crucially, we develop innovative priors which we have found to be essential to make the whole thing work.

Our Cholesky stochastic volatility model is introduced and discussed in Section 2. Section 2.1 reviews the current literature on multivariate stochastic volatility models. Section 2.2 details the time-varying regression approach and the relationships between the time-varying regression parameters, the covariance matrix, and the Cholesky decomposition. Section 2.3 illustrates the ideas and output of the approach by displaying the results obtained in a simple example with $p = 3$ series of asset returns. Section 3 details the Markov Chain Monte Carlo algorithm for posterior computation and the parallel computing strategy. Section 4 presents our new prior and illustrates its effects. Section 5 presents real and simulated examples with the number of series ranging from $p = 3$ to $p = 20$. Section 6 concludes the paper.

2 Cholesky stochastic volatility

Let Y_t denote our random vector of dimension p observed at time t, with

$$Y_t \sim N(0, \Sigma_t).$$

We assume that any mean structure 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. Two challenges arise in the multivariate context. Firstly, the number of distinct elements of Σ_t equals $p(p+1)/2$. 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 p. The vast majority of the papers available in the literature employed a small p or use highly parameterized models to simplify the computation. For instance, Engle (2002) and Tse & Tsui (2002) proposed dynamic conditional correlation (DCC) models that employ two parameters to govern the time evolution of correlations. The DCC models are often rejected in empirical volatility modeling. Secondly, the distinct elements of Σ_t can not be modeled independently since positive definiteness has to be satisfied. Section 2.1 briefly reviews the literature on multivariate stochastic volatility models, while Section 2.2 introduces our proposed Cholesky stochastic volatility model. A simple illustrative example is then presented in Section 2.3.

2.1 Brief literature review

There are at least three ways to decompose the covariance matrix Σ_t. In the first case,

$$\Sigma_t = D_t R_t D_t$$

where D_t is a diagonal matrix with the standard deviations, $D_t = \text{diag}(\sigma_{1t}, \ldots, \sigma_{pt})$, and R_t is the correlation matrix. The above two challenges remain in this parametrization, i.e. the number of parameters increases with p^2 and R_t has to be positive definite. In the second case, a standard factor analysis structure is used to produce

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

where β_t is the $p \times k$ 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, Ruiz & Shephard (1994), Pitt & Shephard (1999), Aguilar & West (2000), and, more recently, Lopes & Migon (2002), Chib, Nardari & Shephard (2006), Han (2006), Lopes & Carvalho (2007) and Philipov & Glickman (2006*a*), to name just a few[1].

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

$$\Sigma_t = A_t H_t A_t'$$

[1]Philipov & Glickman (2006*a*) extended the FSV model by allowing H_t to follow a Wishart random process and fit a 2-factor FSV to model the covariance of the returns of $p = 88$ S&P500 companies. Han (2006) fitted a similar FSV model to $p = 36$ CRSP stocks. Chib et al. (2006) analyzed $p = 10$ international weekly stock index returns (see also Nardari & Scruggs, 2007). Lopes & 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 $p = 5$ Latin America stock indexes.

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, its lower diagonal elements are unrestricted since positive definiteness is guaranteed. In the next section we show that there will be $p(p+1)/2$ dynamic linear models to be estimated and $3p(p+1)/2$ static parameters. When $p = 30$, for example, there are 465 latent states and 1395 static parameters.

The Cholesky decomposition approach has been studied elsewhere. Uhlig (1997) and Philipov & Glickman (2006*b*), for example, proposed models for the covariance matrix based on the temporal update of the parameters of a Wishart distribution (see also Asai & McAleer, 2009). Uhlig (1997) models $\Sigma_t^{-1} = B_{t-1}^{-1}\Theta_{t-1}(B_{t-1}^{-1})'\nu/(\nu+1)$, where $\Theta_{t-1} \sim Beta\left((\nu+pq)/2, 1/2\right)$, $B_t = A_t H_t^{1/2}$ and $Beta$ denotes the multivariate Beta distribution (Uhlig 1994). See also Triantafyllopoulos (2008) for a similar derivation in the context of multivariate dynamic linear models. Philipov & Glickman (2006*b*) 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 $E(\Sigma_t|\Sigma_{t-1},\theta) = \nu(C^{-1/2})(\Sigma_{t-1})^d(C^{-1/2})'/(\nu-p-1)$. The parameter d controls the persistence in the conditional variance process. A constant covariance model arises when $d = 0$, so $E(\Sigma_t) = \nu C^{-1}/(\nu-p-1)$ and C plays the role of a precision matrix. When $d = 1$ and $C = I_p$, it follows that $E(\Sigma_t) = \Sigma_{t-1}$ so generating random walk evolution for the conditional covariance. See Dellaportas & Pourahmadi (2011) for a similar model for time-invariant A and H_t following GARCH-type dynamics[2]. A thorough review of the multivariate stochastic volatility literature up to a few years is provided in Asai, McAleer & Yu (2006) and Lopes & Polson (2010).

2.2 Time-varying triangular regressions

In this section we lay out our basic parametrization of the time varying covariance structure. Let Y_t denote the mean-corrected vector of asset returns at time t, with $Y_t \sim N(0, \Sigma_t)$. We assume that any mean structure has been subtracted out as part of a larger MCMC algorithm. The dependence structure between components of Y_t is captured by a sequential regression structure. The regression structure represents a reparametrization of Σ_t. We allow all parameters of the regression structure to be time-varying.

Recall that $Y_t \sim N(0, \Sigma_t)$ and let $\Sigma_t = A_t H_t A_t'$ 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 $H_t = \text{diag}(\omega_{1t}^2, \ldots, \omega_{pt}^2)$. Therefore,

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

Let the (i,j) element of the lower triangular matrix A_t^{-1} be $-\phi_{ij}$ for $i < j$, while the diagonal element (i,i) is one. It follows that the joint normal distribution for Y_t, that is $N(0, \Sigma_t)$, can

[2]Uhlig (1997) models daily/current prices per tonne of aluminium, copper, lead and zinc, i.e. $p = 4$, exchanged in the London Metal Exchange. Philipov & Glickman (2006*b*) fit their model to returns data on $p = 5$ industry portfolios. Dellaportas & Pourahmadi (2011) model exchange rates of the US dollar against $p = 7$ other country/regions.

be rewritten as a set of p recursive conditional regressions where

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

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

$$Y_{it} \sim N(\phi_{i1t} Y_{1t} + \phi_{p2t} Y_{2t} + \cdots + \phi_{i(i-1)t} Y_{(i-1)t}, \omega_{it}^2). \tag{2}$$

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

$$\phi_{ijt} \sim N(\alpha_{ij} + \beta_{ij}\, \phi_{ij(t-1)}, \tau_{ij}^2) \tag{3}$$

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

$$d_{it} \sim N(\alpha_i + \beta_i\, d_{i(t-1)}, \tau_i^2) \tag{4}$$

for $d_{it} = \log(\omega_{it}^2)$ and $i = 1, \ldots, p$.

The actual parameters we work with are the ϕ_{ijt}s and d_{it}s. These parameters are our states variables in the state equations (3) and (4), while the recursive conditional regressions (or simply *triangular regressions*) are our observation in the observation equations (1) and (2). Our *Cholesky stochastic volatility* model comprises equations (1) to (4).

2.3 A simple example

In this section we present a simple example to illustrate the ideas in Section 2.2. Our example has $p = 3$ series. Each series consists of daily returns of a firm in the S&P100. In Section 5 we will consider $p = 20$ series.

The top panels of Figure 1 are time series plots of the three series in Y_t. The middle panels show the time-varying standard deviations and the bottom panels show the time varying correlations. The solid lines in the middle and bottom panels are plots of the posterior means of the standard deviations and correlations. The dashed lines in the middle and bottom panels are estimates obtained using a moving window of 50 observations. At each time t, we compute the sample covariance of 50 observations centered at t. We then plot the corresponding sample standard deviations and correlations.

The estimates in the middle panels clearly capture the volatility patterns evident in time-series plots of the data. Overall, the second series is more volatile. All series are less volatile in the later part of the data. There is a clear time segment with increased volatility for the second series around $t = 1,000$. The posterior means are a "smoothed" version of the estimates obtained from the simple moving-window approach.

The estimates in the bottom three panels show that there is time variation in the level of correlation amongst the three series. As with the standard deviations in the middle panels, the posterior mean appears to be a nice smooth of the sample quantities obtained from the moving-window. Unlike the standard deviations, the time-varying correlation cannot be seen easily in the time-series plots of the data.

Notice that the estimates of the time varying-correlations obtained from the moving-window are much more variable than those of the standard-deviations. Apparently, the correlations are more difficult to estimate. Of course, we could make the moving-window estimates smoother by increasing the size of the window. We control the smoothness of the posterior means through the choice of prior as discussed in Section 4. Choosing the size of window or "band-width" is always a difficult and important issue in practice. These issues become acute for large p. A basic assertion of this paper (and others like it) is that the Bayesian approach, with careful prior choice, is more effective in practice than any moving-window or simple smoothing approach, especially for large p.

Figure 1 about here.

Figure 1 presents point estimates. Our Bayesian approach also allows us to assess the uncertainty. Figure 2 displays our inference for $\{\rho_{12t}\}$, the time-varying correlation between Y_1 and Y_2. The solid line in the middle is the posterior mean (this is the same as in the bottom left panel of Figure 1). The smooth lines above and below the center lines give pointwise 90% posterior intervals for ρ_{12t} at each t. The more jumbly line is the estimate from the moving window (again, just as in Figure 1). The interval at time $t = 900$ is $(-0.32, -0.011)$ while the interval at time $t = 1500$ is $(0.12, 0.43)$. The intervals indicate that, while there is substantial uncertainty, the dependence structure at time $t = 900$ is quite different from that at time $t = 1500$.

Figure 2 about here.

Figure 3 displays the posteriors means of the actual states underlying our state-space representation of the time-varying Σ_t. The reader may wish to refer back to the first three lines of the first equation of Section 2.2. The diagonal panels plot the time series of posterior means of d_{1t}, d_{2t} and d_{3t}, respectively. Time series of the posterior means of ϕ_{21t}, ϕ_{31t} and ϕ_{32t} are shown in the middle left, bottom left and bottom middle panels, respectively. The top left panel of Figure 3 is quite easy to understand. For the first series, we are simply estimating univariate stochastic volatility, so that $\sigma_{1t} = \omega_{1t}$, $\omega_{1t} = \exp(d_{1t}/2)$, and $\sigma_{1t} = \exp(d_{1t}/2)$. Hence, the pattern is very similar to that of the top left panel of Figure 1. The patterns in bottom left panel of Figure 1 and middle left panel of Figure 3 also correspond fairly closely since ρ_{12t} is strongly related to ϕ_{12t}. After that, things are less transparent as detailed in Section 2.2.

Figure 3 about here.

For larger p it is quite difficult to display the fit of the model. In order to illustrate the fit and use of the model, we consider a simple application in the case where all of our series are asset returns. Let w_t denote the portfolio weights of the *global minimum variance portfolio*. That is, w minimizes $w' \Sigma w$ subject to the constraint that $\sum w_i = 1$, or

$$w(\Sigma) = \frac{\Sigma^{-1} 1_p}{1_p' \Sigma^{-1} 1_p},$$

where 1_p is a vector of ones. Figure 4 plots the posterior means of the components of $w(\Sigma_t)$. The dashed vertical lines at the ends of the time series plots indicates 80% posterior intervals for the weights at the final time. While the global minimum variance problem is admittebly a very stylized problem, we can still infer that the substantial time variation in both the volatility and dependence structure could have considerable implications for portfolio choice.

Figure 4 about here.

3 Posterior inference

We detail here the Markov Chain Monte Carlo algorithm for posterior computation of our CSV model introduced in Section 2.2. We also discuss how parallel computing can ease the heavy computational burden for moderate and large systems.

3.1 MCMC algorithm

Let p denote the number of series and T denote the number of time series observations on each series. Let $Y_i = \{Y_{it}\}_{t=1}^T$ and $d_i = \{d_{it}\}_{t=1}^T$, $i = 1, 2, \ldots, p$. Let $\phi_{ij} = \{\phi_{ijt}\}_{t=1}^T$, $i = 1, 2, \ldots, p$, $j = 1, 2, \ldots, (i-1)$. 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}. 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:

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

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

The terms $p(\alpha, \beta, \tau)$ and $p(s_0)$ (with i and j dropped and s standing for d or ϕ) denotes 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 is discussed in Section 4.

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"):

- d states: $(d_{i0}, d_i) \mid \circ$,
- ϕ states: $(\phi_{ij0}, \phi_{ij}) \mid \circ$,
- d AR parameters: $(\alpha_i, \beta_i, \tau_i) \mid \circ$,
- 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 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 FFBS and all of the d_i draws reduce to univariate stochastic volatility (we use the method of Kim, Shephard and Chib (1998), again based on the FFBS).

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

Thus, the whole thing boils down to repeated applications of the basic Gibbs sampler that cycles through $(s_0, s) \,|\, (\alpha, \beta, \tau)$ and $(\alpha, \beta, \tau) \,|\, (s_0, s)$, where s denotes a state series and s_0 the intial 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.

3.2 Parallel processing

One of the strengths of our CSV framework is that the triangular representation of the model naturally leads to parallelization in the MCMC scheme. More specifically, the Ti-dimensional state-space vector $(d_i, \phi_{i1}, \ldots, \phi_{i,i-1})$ and the $3i$-dimensional parameter vector $(\alpha_i, \beta_i, \tau_i, \alpha_{i1}, \beta_{i1}, \tau_{i1}, \ldots, \alpha_{i,i-1}, \beta_{i,i-1}, \tau_{i,i-1})$ 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 $(\alpha_i, \beta_i, \tau_i)$, for any i and j (see full conditional distributions in Section 3.1). 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

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

where $\kappa_1(p) = p$, $\kappa_2(p) = p(p-1)/2$ and $\kappa_3(p) = p(p+1)/2$. 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 in 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 $c_{1:i_1} = c_{(i_1+1):p} = c/2$. 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:

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

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

Figure 5 provides an illustration when there are $p = 100$ time series and up to $m = 20$ processors. The costs $(c_d, c_\phi, c_\theta) = (30, 2, 0)$ are based on actual run times (in seconds) for $T = 2516$ time points and 10,000 MCMC draws. It takes 15 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 215 minutes. For $m = 2$ processors, $i_1 = 67$ and the cost per processor is 107 minutes. For $m = 3$ processors, $(i_1, i_2) = (52, 79)$ and the cost per processor is 71 minutes. For $m = 4$ processors, $(i_1, i_2, i_3) = (44, 67, 84)$ and cost per processor is 53 minutes. For $m = 20$ processors, cost per processor is about 11 minutes.

Figure 5 about here.

4 Prior specification

For each $i = 1, 2, \ldots, p$ we need to specify priors for the initial condition d_{i0} and the AR parameters $(\alpha_i, \beta_i, \tau_i)$. For each $i = 2, 3, \ldots, p$ and $j = 1, 2, \ldots, (i-1)$ (a total of $p(p-1)/2$ instances) we need to specify priors for the initial state ϕ_{ij0} and the AR parameters $(\alpha_{ij}, \beta_{ij}, \tau_{ij})$. The choice of these priors is key to any approach that hopes to work for large p since it is only through these priors that we can smooth or "regularize" our high-dimensional model. The choice of these priors is, inevitably, influential.

We start by assuming that we are able to standardize our observations so each variable may be thought of as being on the same scale. To center ideas at a familiar location, we think of the base case as $\Sigma_t = I$, where I is the $p \times p$ identity matrix. This leads to the simple practical expedient of initially dividing each series by its sample standard deviation (we have already assumed the data is centered). We emphasize however, that is not necessary to use this data driven approach and a more subjective approach to using a common scale may be used.

Even with common-scale data we may want to choose different priors for different states. There are many more ϕ series than d series, so a stronger prior may be useful. If we had the prior belief that k factors drive the p series then we might want to tighten up our priors for $k < j < i$. In this way we can shrink towards a factor model without imposing it.

For the rest of this section we focus on the specification of a prior for s_0 and (α, β, τ) in the univariate state-space model

$$Y_t = f(s_t, x_t, Z_t), \quad s_t = \alpha + \beta\, s_{t-1} + \epsilon_t, \quad \epsilon_t \sim N(0, \tau^2),$$

where both Z_t and ϵ_t are the random shocks in the observation and state equations respectively. As discussed in Section 3, all of our state draws reduce to this case. In our application, s_t is either ϕ_{ijt} (in which case f is linear), or s_t is d_{it}.

4.1 The initial state

Thinking of our prior as roughly centered at $\Sigma_t = I$, it we can shrink s_0 towards zero in both the case where s stands for a d state or a ϕ state. Shrinking ϕ towards zero shrinks towards independence and shinking d towards zero shrinks towards a variance of $\exp(0) = 1$.

To shrink towards zero we use a mixture prior along the lines of that used by George & McCulloch (1992) for variable selection:

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

The variable γ is a latent variable which we introduce for each d and ϕ state. The basic Gibbs sampler discussed in Section 3 is augmented with draws (for each d and ϕ)

$$\gamma \,|\, \circ \equiv \gamma \,|\, s_0.$$

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

In all applications in this paper we use $p = 0.5$, $w = 0.1$ and $c = 10$. This is a fairly weak prior. In most applications, interest is not focused on the initial time periods. This prior gives us a very simple way to stabilize results in those applications where the initial parts of the series are of interest.

4.2 A mixture prior for AR parameters

In this section we present a mixture prior for (α, β, τ). We found the standard normal and inverted chi-squared priors to be inadequate. With $\tau^2 \sim \nu\lambda/\chi^2_\nu$, we could not find choices of (ν, λ) that worked well in a variety of applications. The basic notions our prior must be able to express are (i) we may want τ small, and (ii) the cases

Case (1): $(\alpha, \beta) = (0, 1)$

Case (2): $\beta = 0$

Case (3): $(\alpha, \beta) = (0, 0)$

are of particular interest. Our prior mixes over these three cases and the residual case $\beta \in (0,1)$. We put zero prior weight on $\beta < 0$.

Case (1) corresponds to the classic "random-walk" prior. With τ small, this prior succinctly expresses the notion that the state evolves smoothly. Just using $(\alpha, \beta) = (0,1)$ and a version of our τ prior given below (which pushes τ towards small values) is not a bad option, paricularly for large p. However, we found it useful to develop a more flexible prior by including the other components to give the investigator additional options whose attractiveness may depend on the application. Case (2) says the state simply varies about a fixed level α. With small τ this is practically equivalent to a fixed value for the state. Case (3) says that the state is fixed near zero, which, given our standardization, is a value of interest for both d states and ϕ states.

Note that in some cases the posterior is largely identified by the prior. A near constant state can be achieved with $(\alpha, \beta) = (0,1)$ (Case (1)) or $(\alpha, \beta) = (\alpha, 0)$ (Case (2)), given τ small, and the data does not care how you do it. 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 are foolish enough to predict farther ahead, we may be more comfortable going with $\beta \approx 0$, if the data allows it. As usual, the mixture representation allows us to push the inference in desired directions, without imposing it.

We have chosen not to consider the case $\tau = 0$. There is no great practical advantage in considering τ to be zero as opposed to small. Our experience suggests that the key is to be able to weight the cases enumerated above in a flexible manner and draw $(\alpha, \beta, \tau) \,|\, (s_0, s)$ jointly.

In order to make a joint draw and have a minimally restricted choice of prior specification, we put β and τ on a bivariate grid. We then restrict ourselves to a normal density for $p(\alpha \,|\, \beta, \tau)$. Given this normal prior we can integrate out α analytically to obtain $p(\beta, \tau \mid s_0, s)$ which is used to draw from the bivariate grid. Given a draw of (β, τ), we are in a conditionally conjugate situation so that the draw of $\alpha \,|\, \beta, \tau, s_0, s$ is just a normal. The likelihood is that of a univariate regression, so that all of these computations are easy and fast and based on a small number of sufficient statistics.

To specify a prior for τ on a grid of values, we first choose minimum and maximum values τ_{min} and τ_{max}. Using n grid points, we have evenly spaced values $(t_1, t_2, \ldots, t_n)$ with $t_1 = \tau_{min}$ and $t_n = \tau_{max}$. We let $P(\tau = \tau_{min}) \equiv p_{min}$. For $i > 1$, $P(\tau = t_i) \propto \exp(-c_\tau \, |t_i - \tau_{min}|)$. Thus, our τ prior has the four hyper-paramters $(\tau_{min}, \tau_{max}, p_{min}, c_\tau)$. This prior is very simple. We pick an interval, and then our choice of c_τ determines the degree to which we push τ towards smaller values. In principle, this could be done with the commonly used prior, $\tau^2 \sim \nu \, \lambda / \chi^2_\nu$. We found it very difficult to choose values for ν and λ that gave consistently good results.

To specify a prior for $\beta \in (0,1)$, on a grid of points $(b_1, b_2, \ldots, b_n)$, we let $p(\beta = b_i) \propto n(b_i \,|\, \bar{\beta}, \sigma^2_\beta)$, where $n(\cdot \,|\, \bar{\beta}, \sigma^2_\beta)$ denotes a normal density with mean $\bar{\beta}$ and standard deviation σ_β. Values we use in application are $\bar{\beta} = 1.0$ and $\sigma_\beta = 1.0$.

Our full mixture prior has the form

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

where p_{01}, p_{00}, p_{u0}, and p_{uu} are the mixture weights of our four components. The notation δ_x represents the distribution such that x happens for sure. p_{01} is the probability that $(\alpha,\beta) = (0,1)$, p_{00} is the probability that $(\alpha,\beta) = (0,0)$, p_{u0} is the probability that $\beta = 0$ and α is unrestricted, and p_{uu} is the probability that $\beta \in (0,1)$ and α is unrestricted. $p(\tau \mid \beta)$ denotes a discrete distribution on a grid as dicussed above. We allow for the possibility that the choice of the parameters $(\tau_{min}, \tau_{max}, p_{min}, c)$ could depend on β. We may want to suggest smaller values of τ when $\beta = 0$. For example, in all our applications, the choice of c given $\beta = 0$ is twice the value used for non-zero β. $p(\beta)$ is the discrete distribution described above.

As discussed previously, our computational approach constrains us to choose a normal distribution for $\alpha \mid \beta, \tau$. However, we are free to let the paramters of normal depend on β and τ in any way. We use,

$$\alpha \mid \beta, \tau \sim N(0, \sigma_\alpha^2\,(1-\beta^2)).$$

When $\beta = 0$ we simply have $\alpha \sim N(0, \sigma_\alpha^2)$. As β increases, we shrink our prior down towards the case where $\alpha = 0$ at $\beta = 1$. A choice used in application is $\sigma_\alpha = 2.0$.

4.3 Mixture prior examples

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

$$Y_t = x_t\, s_t + Z_t, \quad s_t = \alpha + \beta\, s_{t-1} + \epsilon_t, \quad \epsilon_t \sim N(0, \tau^2).$$

We simulate series of length $T = 200$ with $\mathrm{Var}(Z_t) = 0.1$, $s_0 = 0$, and $x_t \sim N(0,9)$. As discussed in Sections (3) and (4.1), the posterior is computed using the Gibbs sampler with full conditionals given by i) $(s_0, s) \mid (\alpha,\beta,\tau), \gamma, y, x$, ii) $(\alpha,\beta,\tau) \mid (s_0, s), \gamma$, and iii) $\gamma \mid s_0$.

In all three examples, we have $(p, w, c) = (0.5, 0.1, 10)$ for the γ prior, $(\bar{\beta}, \sigma_\beta) = (1.0, 1.0)$ for the β prior, $\sigma_\alpha = 2.0$ for the α prior, and $\tau_{min} = 0.005$ (when $\beta \neq 0$) or $\tau_{min} = 0.001$ (when $\beta = 0$) and $p_{min} = 0.5$ for the τ prior. Also, c_τ is twice as big when $\beta = 0$ as it is when $\beta \neq 0$. We consider three different settings for the remaining hyper-parameters p_{01}, p_{00}, p_{u0}, p_{uu}, τ_{max}, and c_τ:

$$\begin{aligned} \text{Prior 0} &: p_{01} = 0.50,\ p_{00} = 0.15,\ p_{u0} = 0.15,\ p_{uu} = 0.20,\ \tau_{max} = 0.15,\ c_\tau = 100. \\ \text{Prior 1} &: p_{01} = 0.85,\ p_{00} = 0.05,\ p_{u0} = 0.05,\ p_{uu} = 0.05,\ \tau_{max} = 0.05,\ c_\tau = 200. \\ \text{Prior 2} &: p_{01} = 0.85,\ p_{00} = 0.05,\ p_{u0} = 0.05,\ p_{uu} = 0.05,\ \tau_{max} = 0.02,\ c_\tau = 300. \end{aligned}$$

As we go from prior 0 to prior 2, we tighten up our prior on smaller values of τ by decreasing τ_{max} and increasing c_τ. Priors 1 and 2 put more weight on the random-walk mixture component than prior 0.

Figure 6 displays prior 0. The top two panels are density smooths of prior 0 draws of β and τ respectively. The density smooths naturally "jitter" the draws (add a bit of normal noise) so that our mixture of discrete and continuous distributions can be displayed as a single continuous distribution. The marginal for β displays our preference for expressing a smooth state with either $\beta \approx 0$ or $\beta \approx 1$ with more weight being given to the vicinity of 1. The prior for τ expresses our desire for small values. Again, this is driven by our desire for a smooth state. The two τ modes reflect the choice of a smaller τ_{min} when $\beta = 0$. In this case the two modes are not very separated so this aspect of our prior has little practical effect. If we separated these modes more dramatically, we could use this aspect for our prior to identify $\beta \approx 0$ versus $\beta \approx 1$ be saying you can only have a really small τ if $\beta \approx 0$. The long right tail of our τ prior allows the data to push the inference towards larger values if needed.

The bottom two panels of Figure 6 display the joint prior of (α, β). The bottom left panel displays contours from a bivariate smooth of prior 0 draws of (α, β). The bottom right panel is a scatterplot of jittered prior 0 draws of (α, β). In the bivariate distribution we can see our preference for $(\alpha, \beta) \approx (0, 1)$ or $(\alpha, \beta) \approx (0, 0)$ with more weight given to the first pair of values. As β decreases, the conditional prior for α becomes more spread out. The contours appear to tighten as β approaches 0 because the choice $(\bar{\beta}, \sigma_\beta) = (1, 1)$ puts weight on larger β.

Figure 6 about here.

First simulated example. We set $(\alpha, \beta, \tau) = (0, 1, 0.04)$, which is consistent with the random-walk component of our mixture prior, and use prior 0. The results are depicted in Figure 7). The top left panel plots the time series of y and the top middle panel plots x against y. The top right panel plots the simulated $\{s_t\}$ state sequence and the posterior mean of the draws of s from our Gibbs sampler. The posterior mean does a good job of tracking the true state sequence. The bottom panels show the MCMC draws of α, β, and τ, respectively. Virtually all the draws of α and β are exactly equal to the true values of 0 and 1, respectively. Of course, if we drew long enough, we would get some (α, β) draws not equal to (0,1), but our mixture prior, gives us a very sharp posterior. The draws of τ vary about the true value of 0.04.

Figure 7 about here.

Figure 8 plots the posterior means of the states (as in the top right panel of Figure 7) obtained using all three of our priors. The posterior means corresponding to priors 0 and 1

are virtually identical and plot almost right on top of each other. The posterior mean from prior 2 is the smoother fit. Prior 2 has a tighter prior on smaller values of τ resulting in a smoother inference for the states.

Figure 8 about here.

Second simulated example. We set $(\alpha, \beta, \tau) = (0, 0.8, 0.1)$ and use prior 0. Figure 9 has the same format as Figure 7. The posterior mean does a very good job of fitting the simulated states. The draws of α are very tight about the true value. The draws of τ cover the true value, but the prior shinks the posterior somewhat towards smaller values. The draws of β are more interesting. With a 50% prior weight on $\beta = 1$, we see that the posterior draws split their time between $\beta = 1$ and variation about the true value of $\beta = 0.8$.

Figure 9 about here.

Third simulated example. We set $(\alpha, \beta, \tau) = (0.5, 0, 0.01)$ and use prior 0. Figure 10 reports results. In this case the state varies very slightly about the fixed level of 0.5. The state is fit very well and is essentially a flat line. The posteriors of α and β show that our posterior is concentrated on the mixture component where β is zero with α unrestricted and the component where β is in $(0, 1)$. 50% of the draws have $\beta = 0$, so the posterior probability of that mixture component is about 0.5 compared with the prior probability of 0.15.

The posterior for τ covers the true value. However, the draws exhibit substantial dependence. In the case where the state is essentially constant, we can fit the data with any (α, β) combination as long as τ is small. This, combined with the dependence built into the sampler which draws the state given the parameters and the parameters given the state, can lead to chains with strong dependence. The dependence can always be "remedied" by choosing a strong prior and how this prior is chosen will depend on the goals of the application. For example, setting $p_{01} = 1$, will certainly simplify things and give reasonable inference for the states. However, we feel that the full mixture prior provides the investigator with additional options.

Figure 10 about here.

5 Illustrations

5.1 A simulated example

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

Results obtained using priors 0 and 1 from Section 4.3 are shown in Figure 11 and Figure 12 respectively. The same (α, β, τ) prior was used for each of the six state series (three d state series and three ϕ state series). The time-varying standard deviations are plotted in the diagonal panels and the (i, j) correlation is plotted in the (i, j) lower panel below the diagonal. The very smooth line is the true value, the lighter smooth line is the posterior mean of the parmameter (σ_{it} or ρ_{ijt}) and the dashed line is the estimate obtained by computing the standard sample covariance of a moving window of observations. The moving window for estimation of Σ_t, includes all available observations within 50 time periods of t.

In both cases the posterior mean seems to nicely smooth the evidence from the data, with the tighter prior 1 giving smoother results.

Figures 11 and 12 about here.

5.2 Return predictors

Pástor & Stambaugh (2009) develop a *predictive systems* approach for investigating market return predictability. In Pástor & Stambaugh (2011) the predictive systems approach is used to assess the widely held belief that it is wise to invest in stocks when investing "for the long run" with the rather startling suggestion that this may not be the case.

The predictive systems approach is an vector autogression (VAR) for the observed market return, the observed variables thought to be predictive of market returns, and the unobserved expected value of the market return at time $(t+1)$ given information at time t. Three predictor variables are used in Pástor and Stambaugh's empirical results giving a five-dimenional VAR.

In the predictive systems model the covariance matrix of the innovations in the VAR is assumed to be constant over time. Pástor & Stambaugh (2011) discuss the possibility that the error covariance matrix may be time varying and argue that the basic results of the paper would not change if this was the case. However, no attempt is made to see what the data has to say about time variation. In this section we use the methods of this paper to

show that there is evidence of time variation. Carvalho, Lopes & McCulloch (2011) assess long run investment in stocks in the presence of multivariate stochastic volatility.

We focus on the three predictor variables (and thank Pástor and Stambaugh for providing us with the data). Allowing time variation for the covariance of market returns and expectation involves complex issues of prior specification discussed in Carvalho et al. (2011). The three predictor variables are dividend yield, consumption-wealth ratio (CAY), and bond-yield. We have quarterly data from 1952-Q1 to 2008-Q4 giving 228 quarters of observations. To get a simple look at the error distribution and demean the series, we used least-squares to fit first order autoregressive models, or AR(1) models, to each series individually and work with the residuals from these fits.

The data and inference are displayed in Figure 13. We used Prior 1 of Section 4.3. The top three panels plot the times series of AR(1) residuals (labelled $y1$, $y2$, and $y3$, respectively). The lines in the data plots are at $\pm 2\,\hat{\sigma}_{it}$ where $\hat{\sigma}_{it}$ is the posterior mean as estimated by our MCMC. The third series clearly exhibits time variation in the volatility. The first series suggests time variation, but not as strongly as the third. There is no clear indication of time variation for the second series.

The middle three panels report inference for the σ_{it}, $i = 1, 2, 3$ and the bottom three report inference for ρ_{21t}, ρ_{31t}, and ρ_{32t}. In each plot the solid line in the middle is the posterior mean, the dotted inner lines are the pointwise (at each t) 50% posterior intervals and the outer dashed lines are 95% intervals. While pointwise intervals are somewhat unsatisfactory, the middle panels do tend to confirm our impression from the data that the third series has time varying volatility. The bottom three panels suggest that there is strong, but not overwhelming evidence for time variation in ρ_{31t} and ρ_{32t}.

Figure 13 about here.

Figure 14 investigates the senstivity of our results to the choice of prior. Posterior means are plotted for three different prior specifications. The top three panels are for σ_{it} and the bottom three panels are for ρ_{21t}, ρ_{31t}, and ρ_{32t} (as in the middle and bottom rows of Figure 13). The vertical range for each σ_{it} is $(0.2\,\hat{\sigma}_i, 1.8\,\hat{\sigma}_i)$ where $\hat{\sigma}_i$ is the sample standard deviation of the i^{th} series. The solid line in each plot corresponds to Prior 1 from Section 4.3 (thus repeating information from Figure 13). The dashed lines correspond to Prior 0 from Section 4.3. Compared to Prior 1, Prior 0 puts more weight on larger τ and more weight on mixture components with $\beta = 0$. For the ρ_{ijt} and σ_{2t} there is negligible difference between the Prior 0 results and the Prior 1 results. For σ_{1t} and, particularly, σ_{3t}, the results from Prior 0 indicate more substantial variation in the volatility.

The dot-dash line in Figure 14 corresponds to a third prior having $p_{00} = 0.1$, $p_{u0} = 0.4$, $p_{01} = 0.1$, and $p_{uu} = 0.4$. This prior puts more weight on $\beta = 0$ than the other two priors. Given $\beta = 0$, the prior on τ is specified by the choices $\tau_{min} = 0.001$, and $c_\tau = 400$. Given $\beta = 1$, the prior on τ is specified by the choices $\tau_{min} = 0.01$, and $c_\tau = 200$. In both cases

$p_{min} = 0.5$ and $\tau_{max} = 0.1$. This prior distinguishes the cases $\beta = 0$ and $\beta = 1$ more sharply than the other two priors by preferring substantially smaller τ when $\beta = 0$. When $\beta = 1$ this prior supports τ which are larger than Prior 1 but smaller than Prior 0. Overall, this prior allows for more time-variation than Prior 1 (bigger τ) but is more focused on identifying $\beta = 0$. For most states, we obtain similar fit from the third prior. The fit for σ_{3t} is more like that of Prior 0 than Prior 1 because larger τ are allowed. The most markedly different posterior mean is that of ρ_{31t}. The third prior's emphasis on $\beta = 0$ has substantially reduced the time variation.

Figure 14 about here.

Figure 15 looks at some of the marginal posteriors of the state AR coefficients. We display density smooths of the draws of (β_i, τ_i) from the state equations $d_{it} \sim N(\alpha_i + \beta_i d_{i(t-1)}, \tau_i^2)$, $i = 1, 2, 3$. The top panels depict the posteriors of the three β_i and the bottom panels depict the posteriors of the three τ_i. In each plot the solid line is the posterior obtained using Prior 1 while the dashed line corresponds to the third prior described above. For the first and third series we see that the posterior for β is tight around $\beta \approx 1$ for both priors. For both the first and third series, the posterior for τ supports larger values under the third prior. Prior 1 may be "over smoothing" the states. The most striking difference is in the posteriors for β_2. The posterior from Prior 1 is concentrated at $\beta \approx 1$ while the posterior from the third prior allows for the possibility that the near constant variance of the second series may be obtained with either $\beta \approx 0$ or $\beta \approx 1$. The posterior of τ_2 under the third prior is concentrated on small τ values but exhibits two modes corresponding to the modes for β_2.

Figure 15 about here.

5.3 Many assets from the S&P100

Section 2.3 presents an example with $p = 3$ series of asset returns from the S&P100. We can now tell the reader that the results given there were obtained using prior 1 from Section 4.3. In this section we choose more assets (firms) from the S&P500 in order to illustrate the procedure with larger p.

We first consider a selection of returns on $p = 20$ of the firms. Again, we use prior 1. Figure 16 plots the posterior means of the σ_{it} and ρ_{ijt} series. The top panel show the 20 standard deviations series and the bottom panel shows the $20(19)/2 = 190$ correlations series. There is no simple way to plot so much information, but even with the many series, we can see that there is substantial time variation in both the standard deviations and the correlations. From the $\{\sigma_{it}\}$ series we see that there is an overall pattern to their behavior over time. For example, the volatility is generally lower at the end of the time period. However, there is substantion variation across assets (across i) both in the overall level of

volatility and the amount of time variation. Similarly, there are time periods where ρ_{ijt} is relatively large for most (i, j) pairs but some pairs behave quite differently from the rest.

Figure 16 about here.

Figure 17 plots the posterior means of the states with the d states in the top panel and the ϕ states in the bottom panel. The top panel shows the time variation in the residual variances. The bottom panel shows that most of the ϕ series have relatively little time variation and are centered near zero. This figure shows how our Bayesian model, with our particular prior choice, seeks a parsimonious representation of our very high dimensional problem.

Of course, the amount of "parsimony", "smoothness", or "regularization" inevitably is heavily influenced by our choice of prior. Figure 18 shows the posterior means of the states obtained when we use prior 2. This figure looks like a smoothed version of Figure 17. The "flat-line" appearance of many of the ϕ states is striking. The corresponding standard-deviation-correlation plot is, again, a smoothed version of Figure 16.

Figures 17 and 18 about here.

Figure 19 plots the posterior means of the portfolio weights for the global minimum variance portfolio using prior 1. 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.

Figure 20 plots two different estimates of a time-varying standard deviation. Again we used the $p = 20$ series but in in the second run we reversed the order of the series. We are plotting the posterior means of σ_{1t}, the standard deviation of the first series in the original order, and $\tilde{\sigma}_{pt}$, the standard deviation of the *same series* in the model where it is last in our Cholesky decomposition. These two fits result from different priors, so there is no reason that they be identical. However, their similarity is striking.

Figures 19 and 20 about here.

Figure 21 reports results for $p = 94$ assets using prior 2. In this case there are 94 standard deviation series (σ_{it}) and $94(93)/2 = 4{,}371$ correlation pairs (ρ_{ijt}) so it becomes quite difficult to present the results. The top panel displays results for the σ_{it} while the bottom panel displays the ρ_{ijt}. The two panels have the same format. The solid grey band gives pointwise quartiles for the posterior means. Thus, in the top panel, the grey band is the middle 50% of the 94 standard deviation posterior means $\hat{\sigma}_{it}$ for fixed t and in the bottom panel it is the middle 50% of the 4,371 correlation estimates. The thick solid (black) lines give 95% intervals. We can see that with 94 series we observe the same overall patterns we saw with $p = 20$.

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).

Figure 21 about here.

6 Conclusion

Clearly, modeling time-varying covariance matrices is very difficult, particularly when the dimension p is large. In most approaches, the set of possible Σ_t is highly restricted.

In this paper we present a method which works for large p without restricting the support of the prior for the Σ_t. The key aspects of our approach are (i) restricting our model and prior so the computations may be done in parallel, and, (ii) a very general and flexible specification for the prior on the parameters of the autoregressive state equations. The prior allows us the "softly-restrict" or "regularize" our inference. Although our prior has positive support on any $\{\Sigma_t\}$ sequence, we can put in the prior belief that they do not change too quickly. A key insight of this paper is that we may also want to put in the prior belief that many of the states describing how one series relates to another (the ϕ) are essentially constant over time. Even for relatively small p, we find the prior a useful tool for exploring the possible dynamic structure.

References

Aguilar, O. & West, M. (2000), 'Bayesian dynamic factor models and variance matrix discounting for portfolio allocation', *Journal of Business and Economic Statistics* **18**, 338–357.

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

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

Carter, C. K. & Kohn, R. (1994), 'On gibbs sampling for state space models', *Biometrika* **81**, 541–553.

Carvalho, C. M., Lopes, H. F. & McCulloch, R. E. (2011), Time-varying predictive systems, Technical report, The University of Chicago Booth School of Business.

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

Dellaportas, P. & Pourahmadi, M. (2011), 'Cholesky-GARCH models with applications to finance (to appear)', *Statistics and Computing* **19**, 237–271.

Engle, R. F. (1982), 'Autoregressive conditional heteroscedasticity with estimates of variance of united kingdom inflation', *Econometrica* **50**, 987–1008.

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

Frühwirth-Schnatter, S. (1994), 'Data augmentation and dynamic linear models', *Journal of Time Series Analysis* **15**, 183–802.

George, E. I. & McCulloch, R. E. (1992), 'Variable selection via gibbs sampling', *Journal of the American Statistical Association* **79**, 677–83.

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

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

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

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

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

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

Nardari, F. & Scruggs, J. T. (2007), '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.

Pástor, L. & Stambaugh, R. F. (2009), 'Predictive systems: living with imperfect predictors', *The Journal of Finance* **64**, 361–393.

Pástor, L. & Stambaugh, R. F. (2011), 'Are stocks really less volatile in the long run? (forthcoming)', *The Journal of Finance*.

Philipov, A. & Glickman, M. E. (2006*a*), 'Factor multivariate stochastic volatility via wishart processes', *Econometric Reviews* **25**, 311–334.

Philipov, A. & Glickman, M. E. (2006*b*), 'Multivariate stochastic volatility via wishart processes', *Journal of Business and Economic Statistics* **24**, 313–328.

Pitt, M. & Shephard, N. (1999), Time varying covariances: A factor stochastic volatility approach, *in* J. B. *et al*, ed., 'Bayesian statistics 6', London: Oxford University Press.

Rosenberg, B. (1972), The behaviour of random variables with nonstationary variance and the distribution of security prices, Technical report.

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

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

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

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

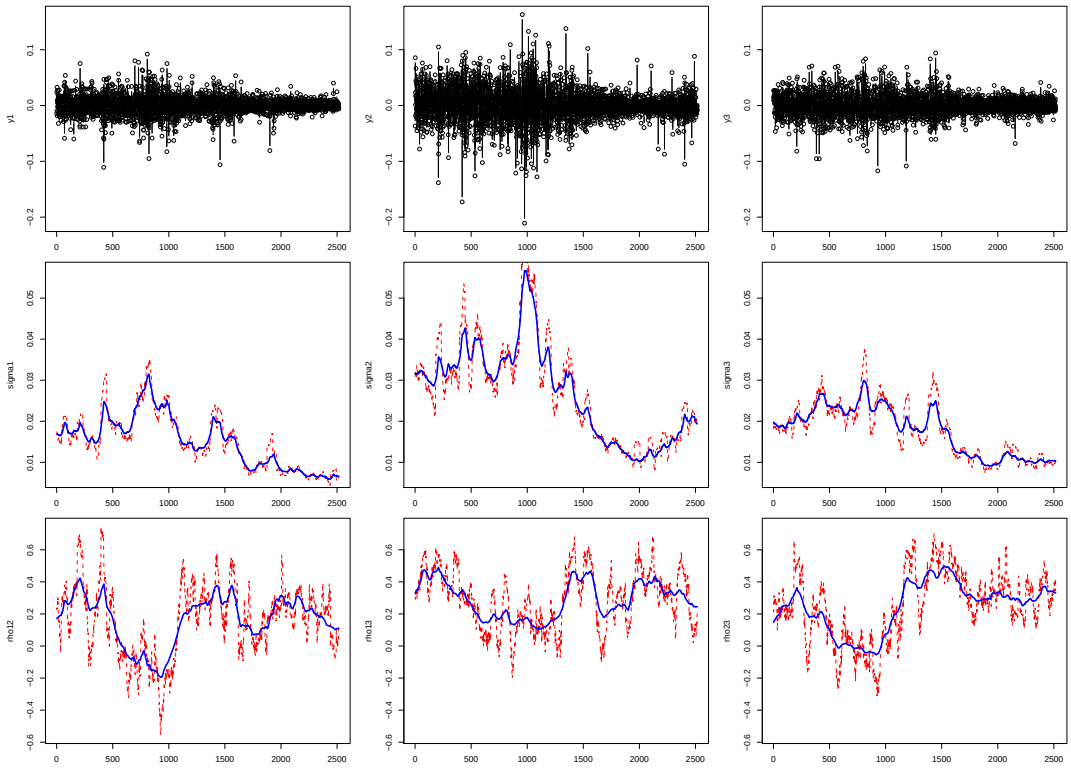

Figure 1: *Trivariate example.* The top three plots are time series of the three time series of daily returns. The middle three show the time-varying standard deviations and the bottom three show the time-varying correlations. The solid lines are posterior means while the dashed lines are estimates obtained from a moving window.

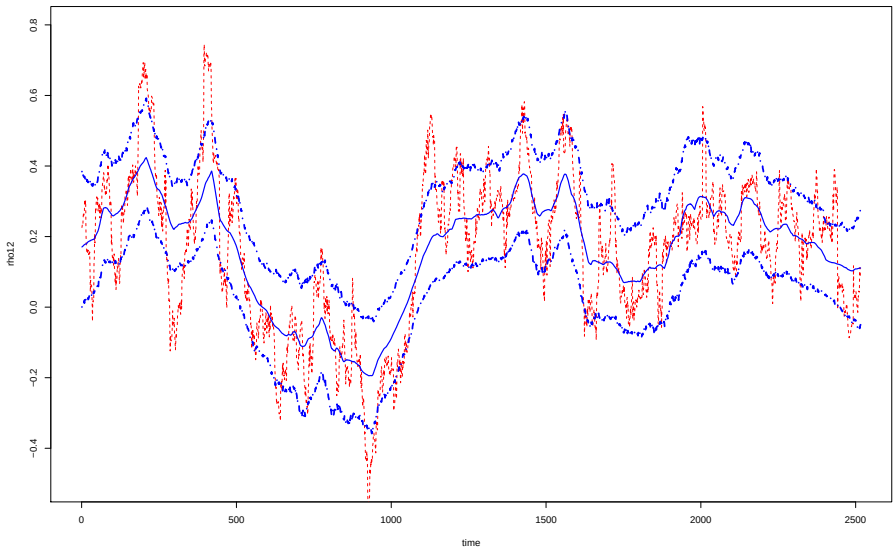

Figure 2: *Trivariate example.* Posterior mean and intervals for the time-varying correlation between Y_1 and Y_2.

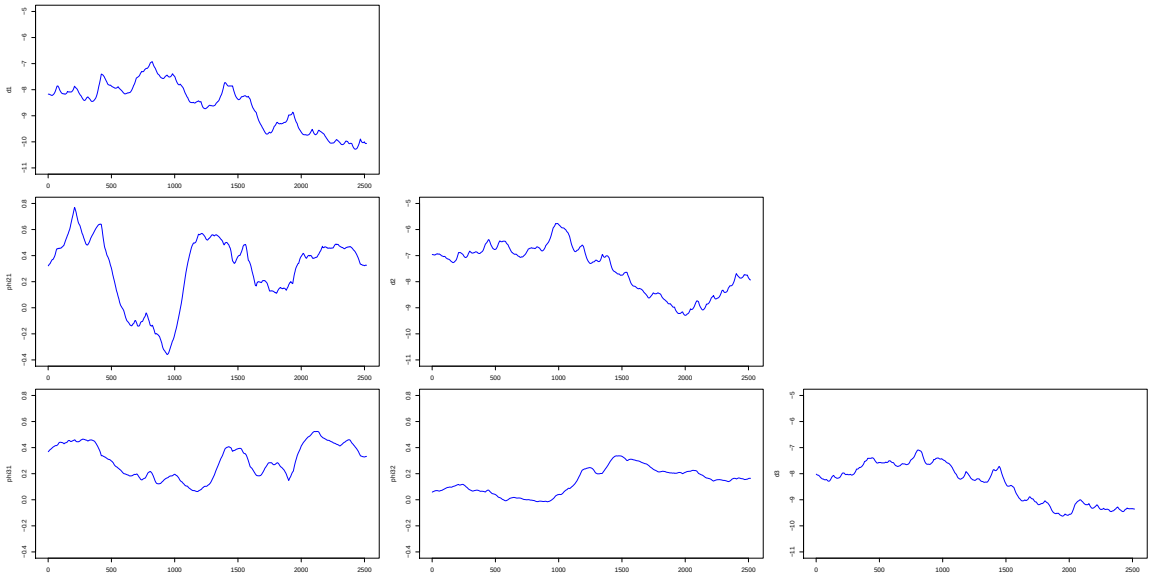

Figure 3: *Trivariate example.* Posterior means of time-varying states. The i^{th} diagonal panels plot the posterior mean of $\{d_{it}\}$ and the (i, j) diagonal panel plots the posterior mean of $\{\phi_{ijt}\}$ for $i = 2, 3,\ j < i$.

Figure 4: *Trivariate example.* Portolio weights for the global minimum variance portfolio.

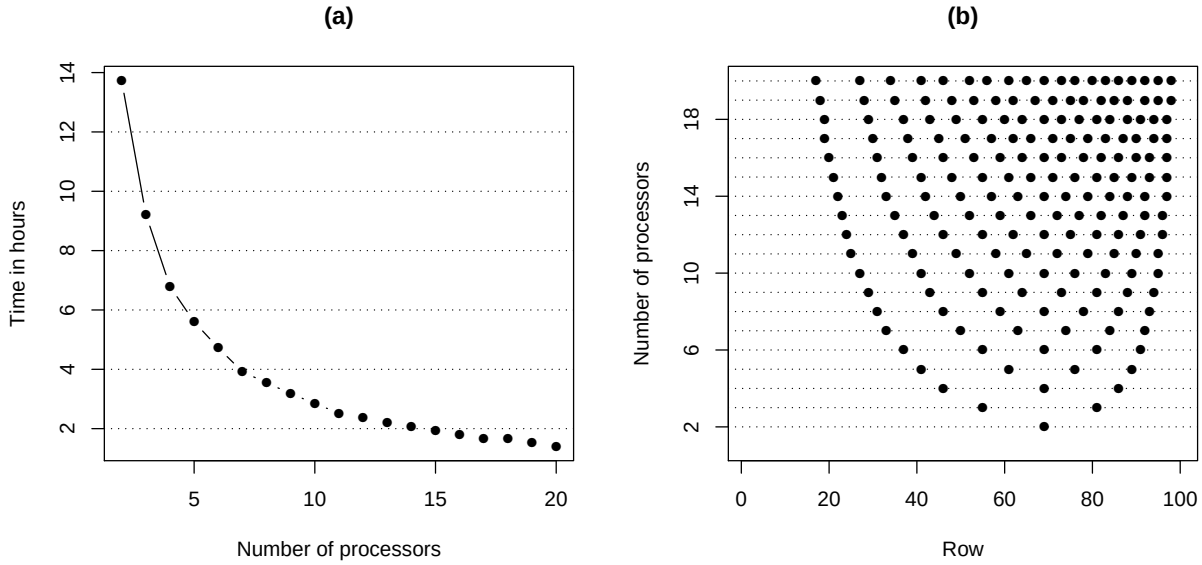

Figure 5: *Multiple processors.* In panel (a) we plot the number of processors vs. the total time in minutes to run 50,000 iterations for a 100×100 ($p = 100$) time varying covariance matrix. It takes 15 times longer to draw a d state than it does to draw a ϕ state. Code was run on a 2.93 GHz Intel Core 2 Duo processor. With 20 processors, the time is about 11 minutes. In panel (b) we have the number of processssors 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.

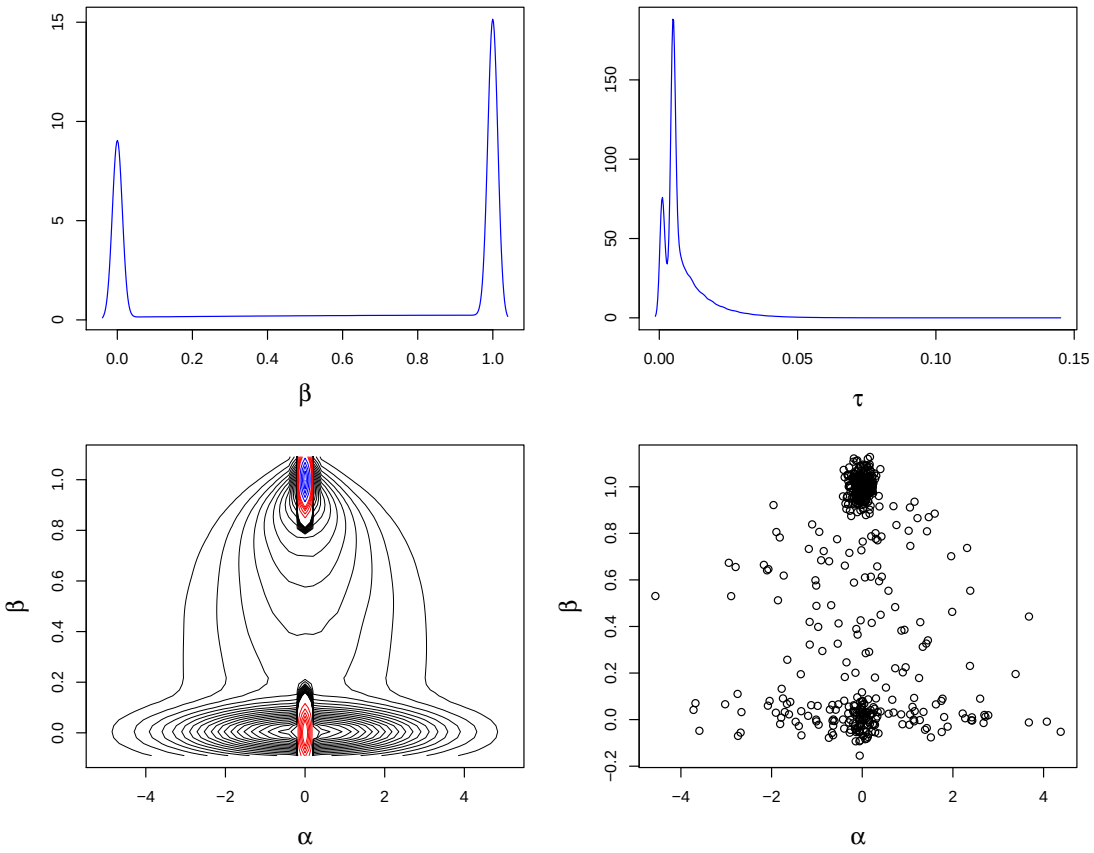

Figure 6: *Prior 0.* The top two panels are density smooths of draws of β and τ. The bottom left panel displays contours from a bivariate smooth of draws of (α, β). The bottom right panel are jittered draws of (α, β).

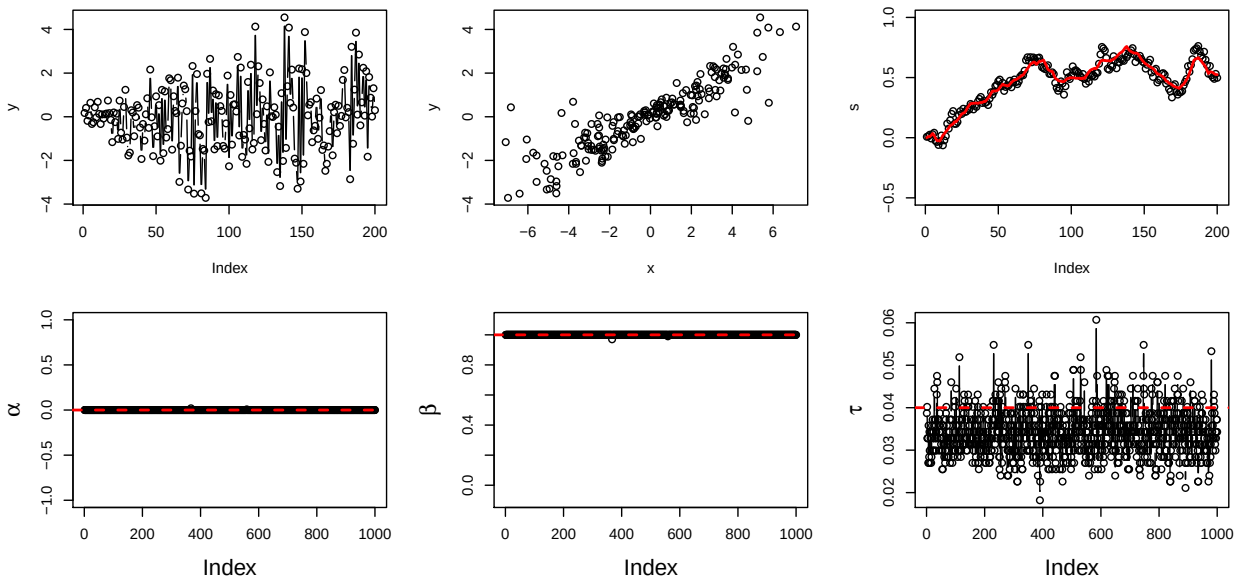

Figure 7: *Prior 0.* Inference for a DLM with a random walk state, $\beta = 1.0$.

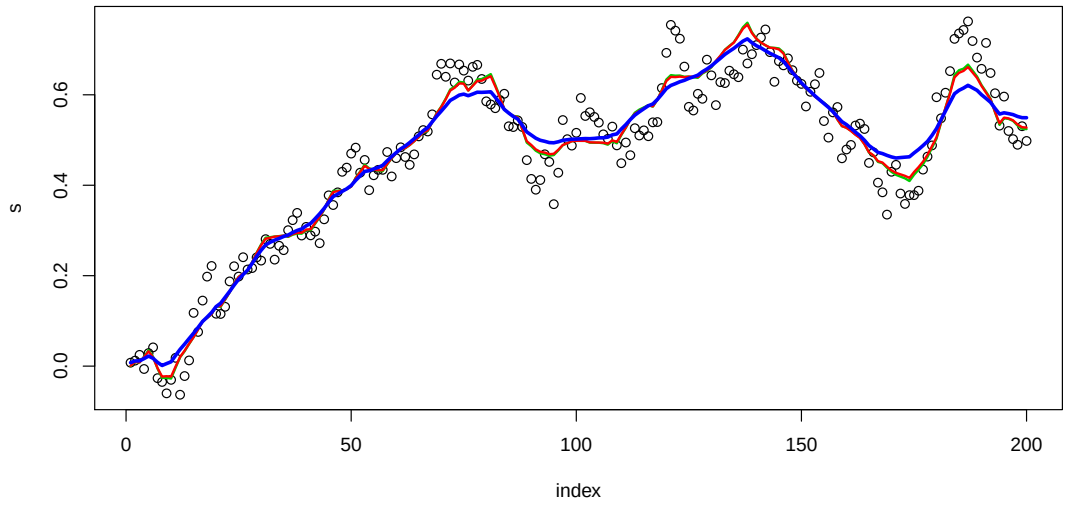

Figure 8: *Priors 0, 1 and 2.* Simulated states and posterior mean estimates from the three priors. The smoother fit corresponds to prior 2 (blue line), while priors 0 and 1 give very similar results.

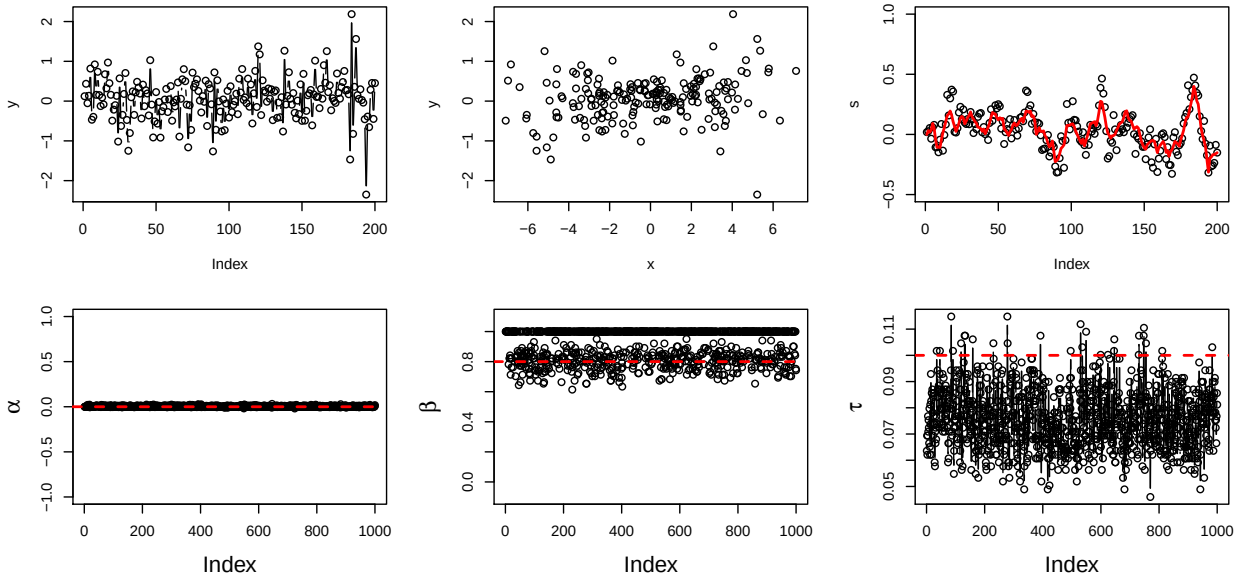

Figure 9: *Prior 0.* Inference for a DLM with a stationary state, $\beta = 0.8$.

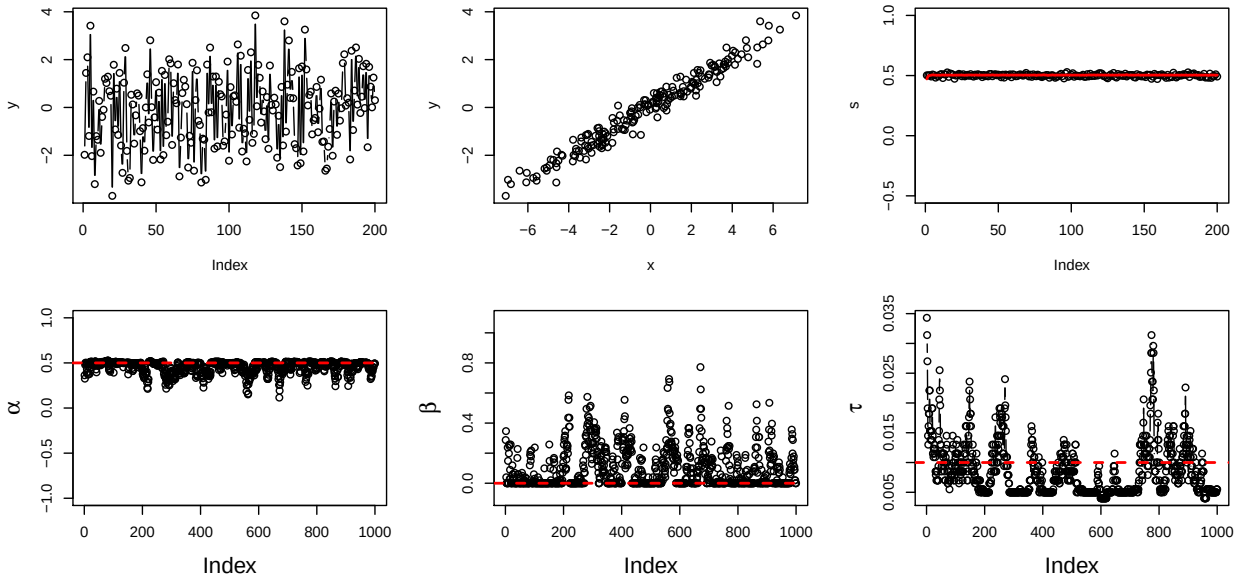

Figure 10: *Prior 0.* Simulation and inference for a DLM with a flat-line state, $\beta = 0$.

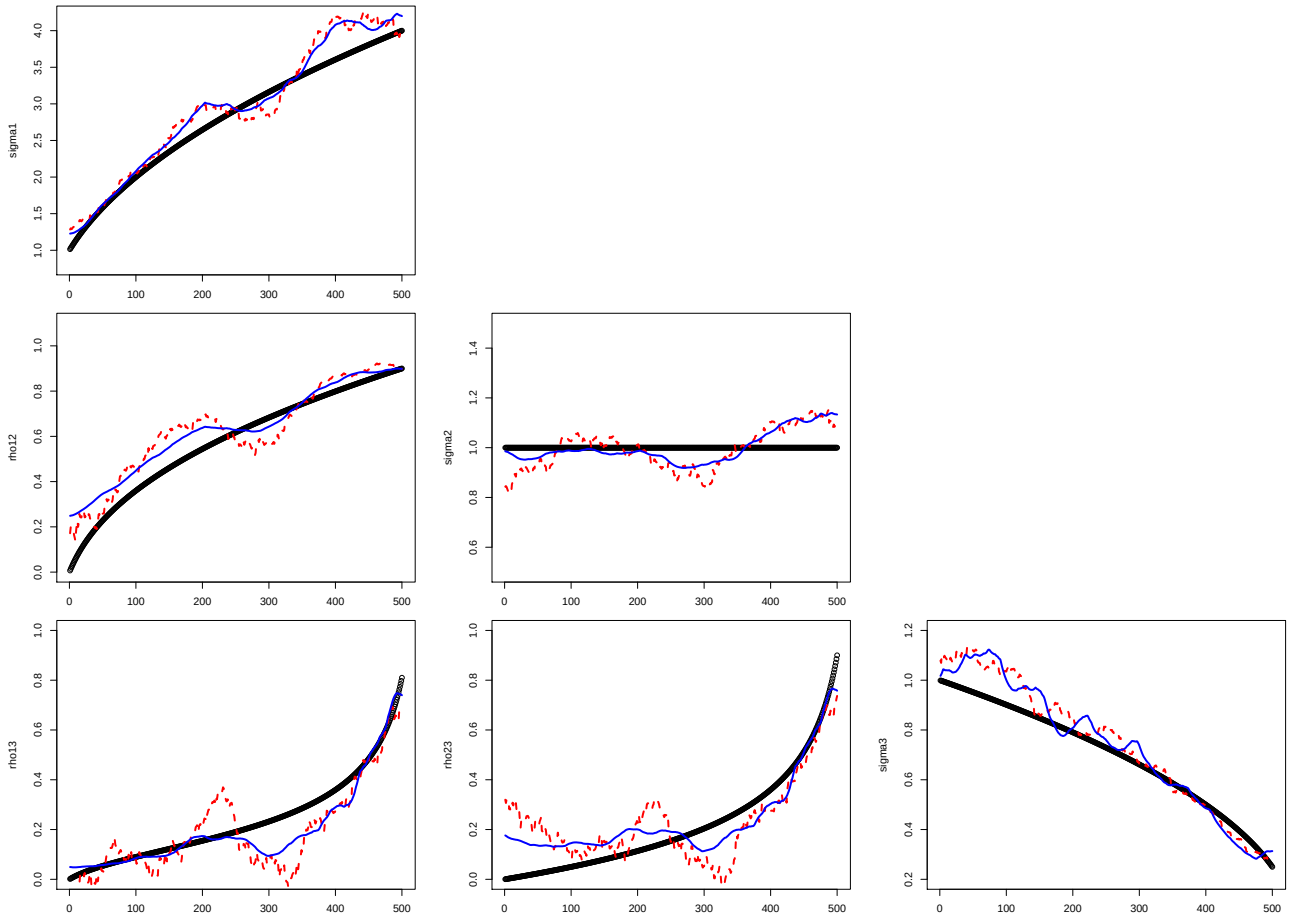

Figure 11: *Simulated example - prior 0.* Time-varying standard deviations are plotted in the diagonal panels and correlations are plotted in the lower panel below the diagonal. *True values:* very smooth black lines. *Posterior means:* lighter smooth blue lines. *Moving window estimates:* red dashed lines.

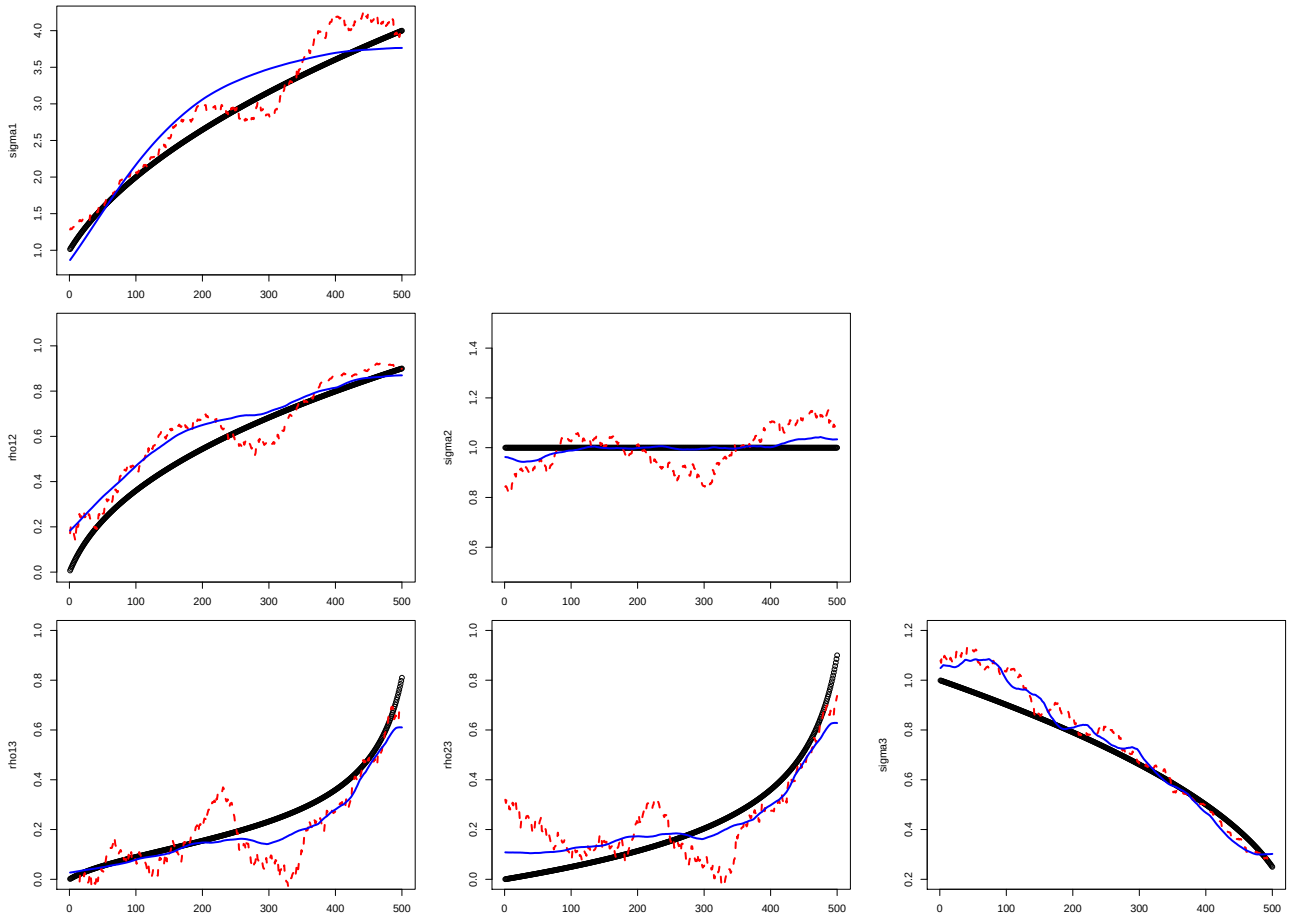

Figure 12: *Simulated example - prior 1.* Time-varying standard deviations are plotted in the diagonal panels and correlations are plotted in the lower panel below the diagonal. *True values:* very smooth black lines. *Posterior means:* lighter smooth blue lines. *Moving window estimates:* red dashed lines.

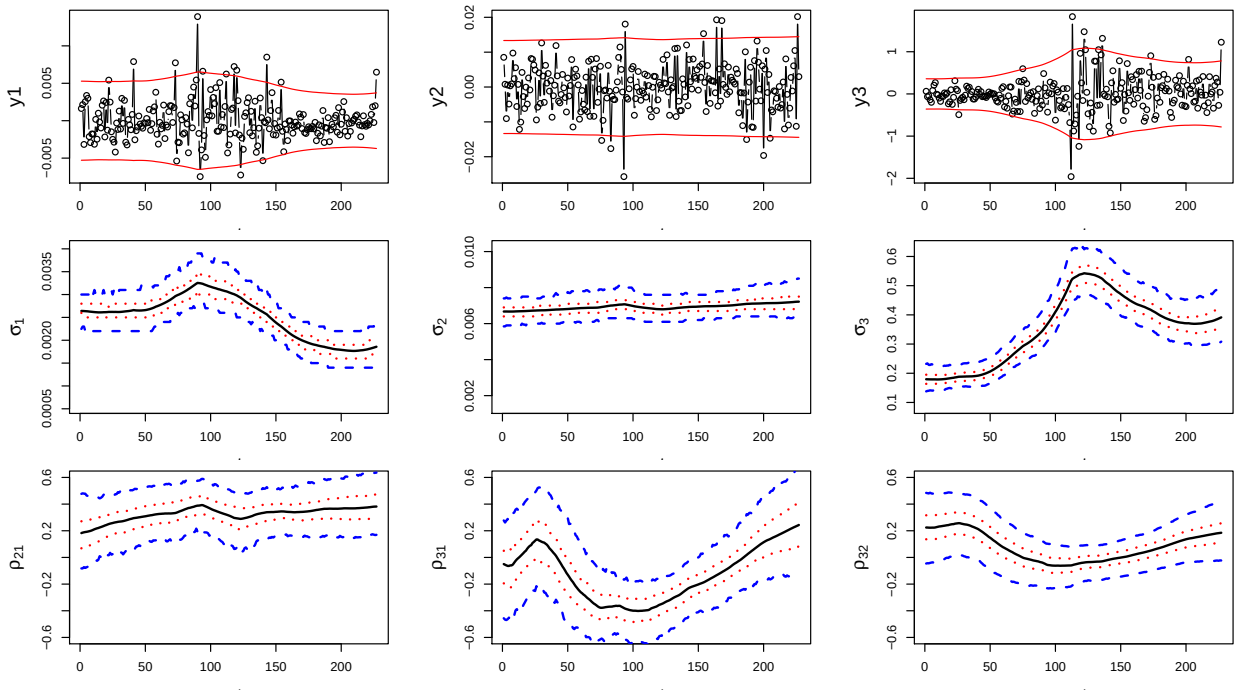

Figure 13: *Return predictors.* The top three panels are time series plots of the data with $\pm 2\,\hat{\sigma}_{it}$. The middle three panels display inference for σ_{it}, $i = 1, 2, 3$, and the bottom three panels display inference for ρ_{21t} ρ_{31t} and ρ_{32t}. In each "inference", the solid line is the posterior mean, the inner dotted lines are pointwise 50% intervals and the outer dashed lines are pointwise 95% intervals.

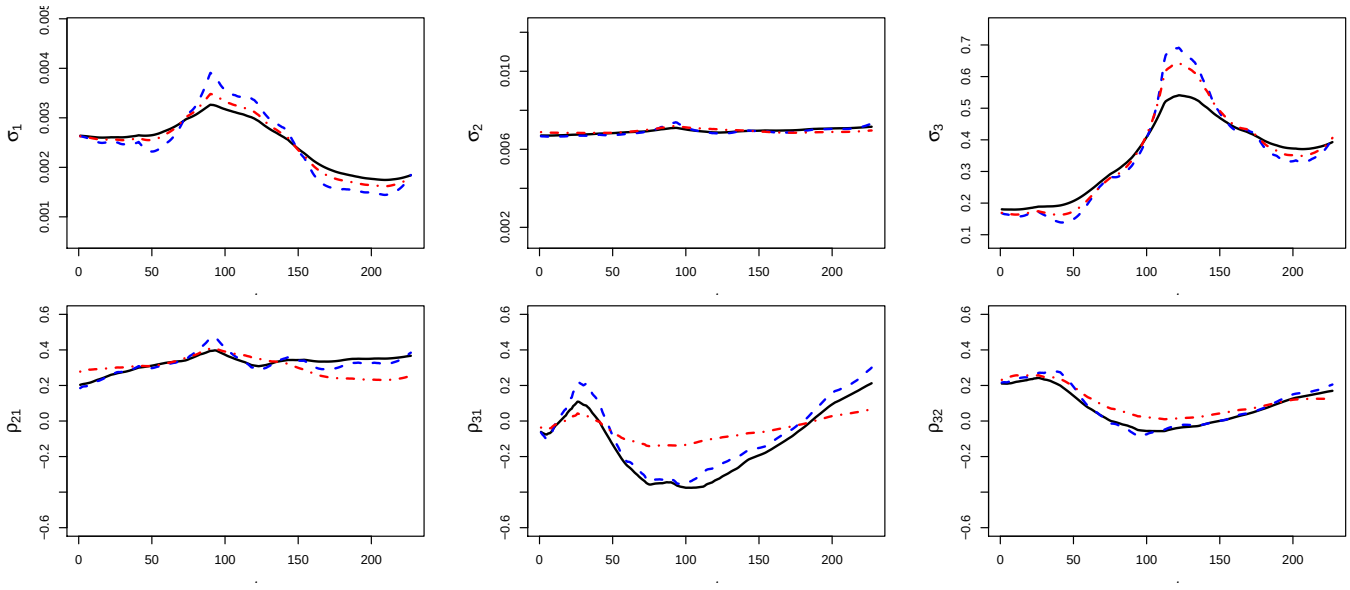

Figure 14: *Return predictors.* Posterior means from three different priors. The top three panels display inference for σ_{it}, $i = 1, 2, 3$, and the bottom three panels display inference for ρ_{21t} ρ_{31t} and ρ_{32t}. In each panel the solid line is the posterior obtained using prior 1, the dashed line is the posterior obtained using prior 0, and the dot-dash line is the posterior obtained using the prior which places more weight on $\beta = 0$.

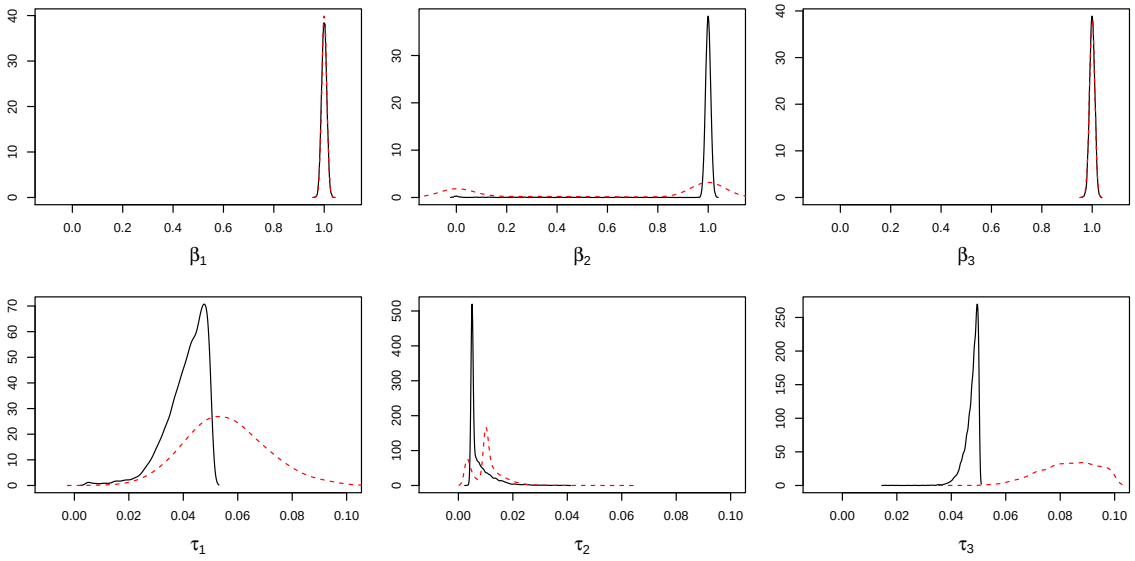

Figure 15: *Return predictors.* Inference for (β_i, τ_i) from the d_{it} state equations $d_{it} \sim N(\alpha_i + \beta_i\, d_{i(t-1)}, \tau_i^2)$. In each panel the solid curve is the posterior density obtained using prior 1 and the dashed curve is the posterior obtained using the prior which places more weight on $\beta = 0$. The top row give densities for β_i and the bottom row gives densities for τ_i.

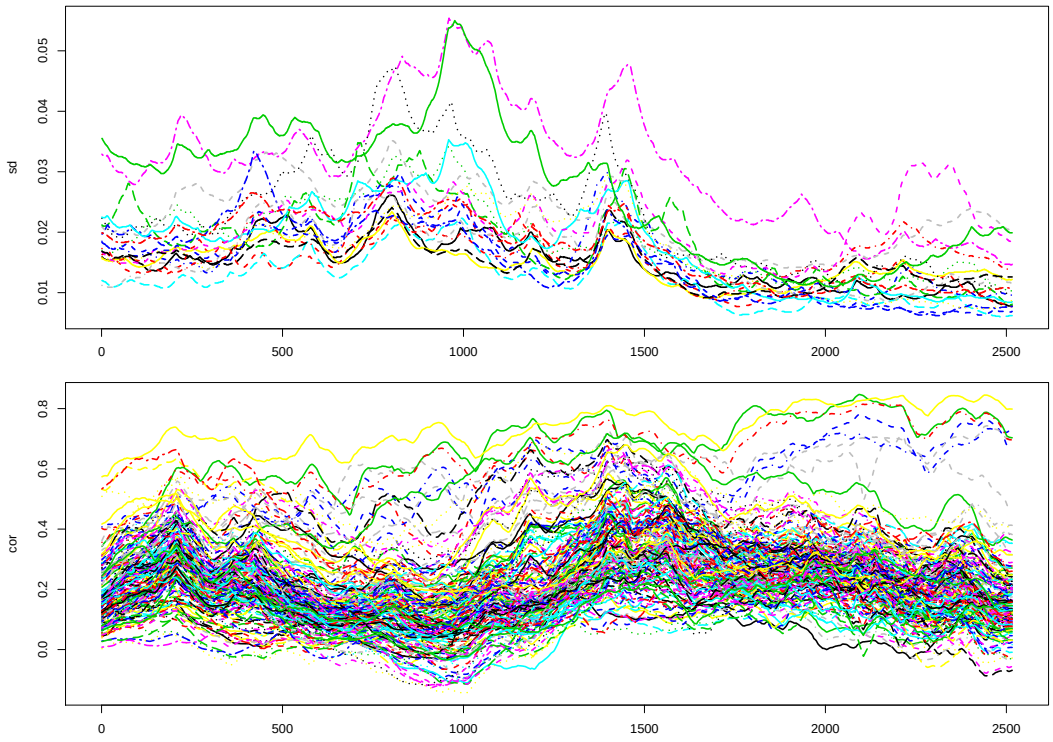

Figure 16: *S&P100 data.* Posterior means of time-varying standard deviations and correlations.

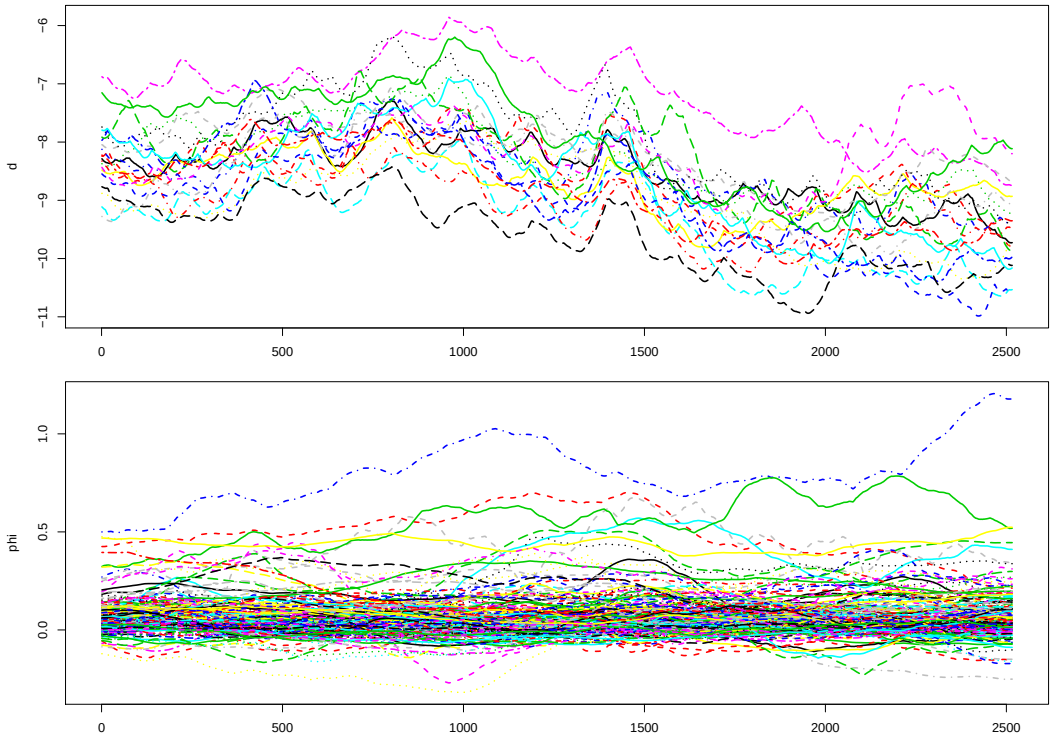

Figure 17: *S&P100 data - prior 1.* Posterior means of d and ϕ states.

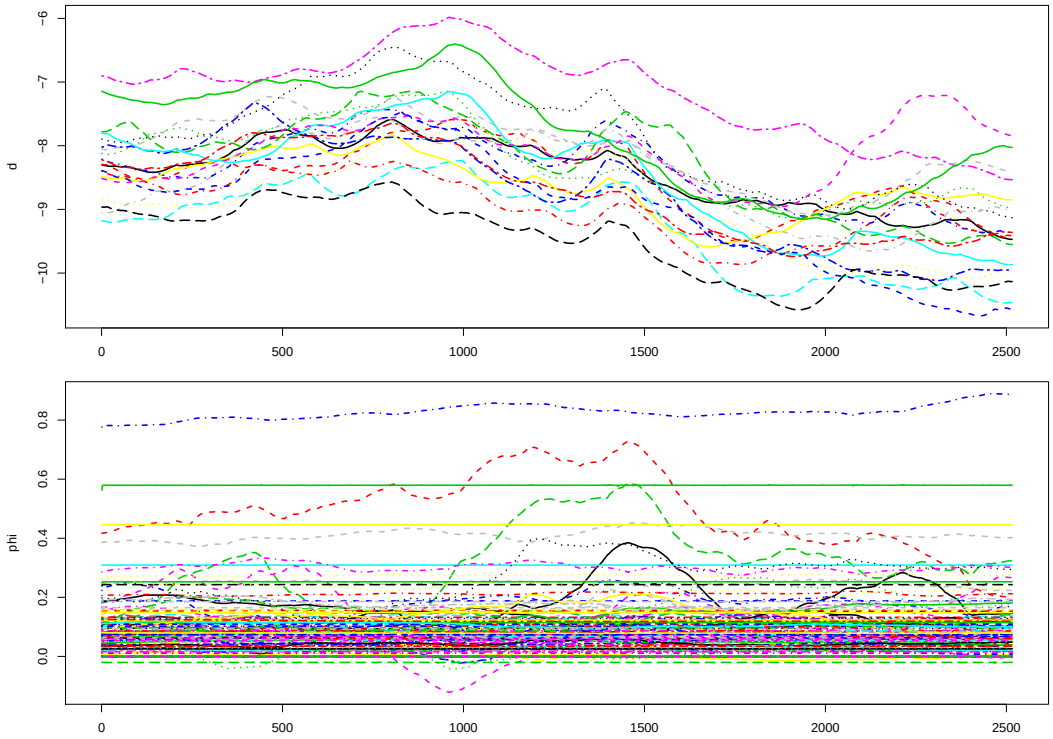

Figure 18: *S&P100 data - prior 2.* Posterior means of d and ϕ states.

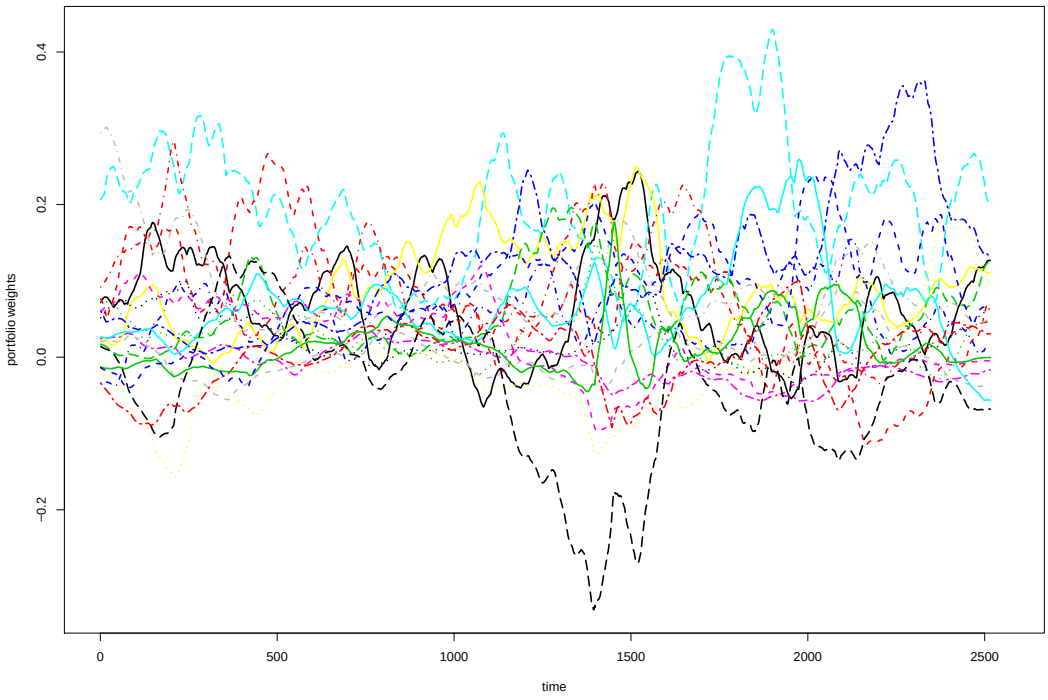

Figure 19: *S&P100 data - prior 1.* Posterior means of portolio weights for the global minimum variance portfolio.

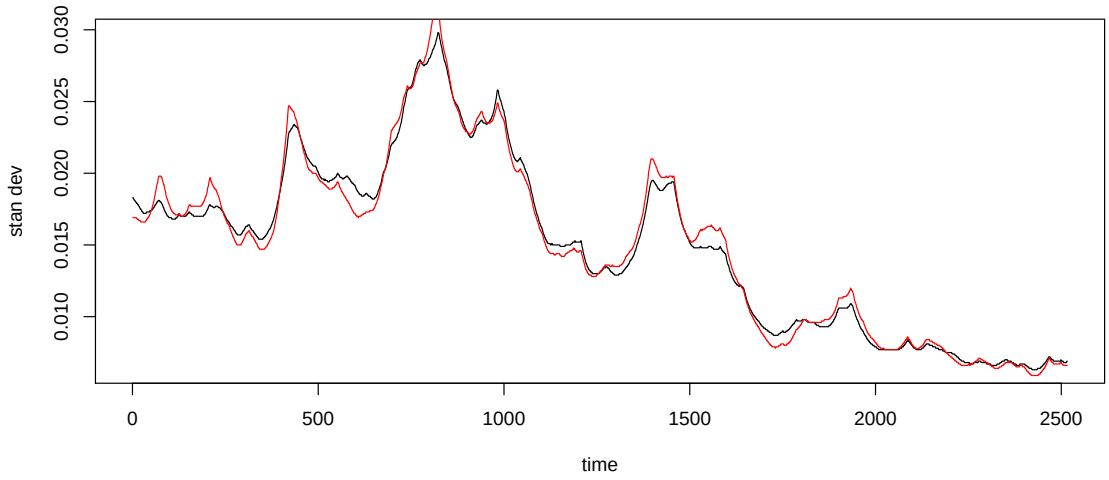

Figure 20: *S&P100 data.* Comparison of the posterior means of the time-varying standard deviation of a series with the original order of the series and the order reversed.

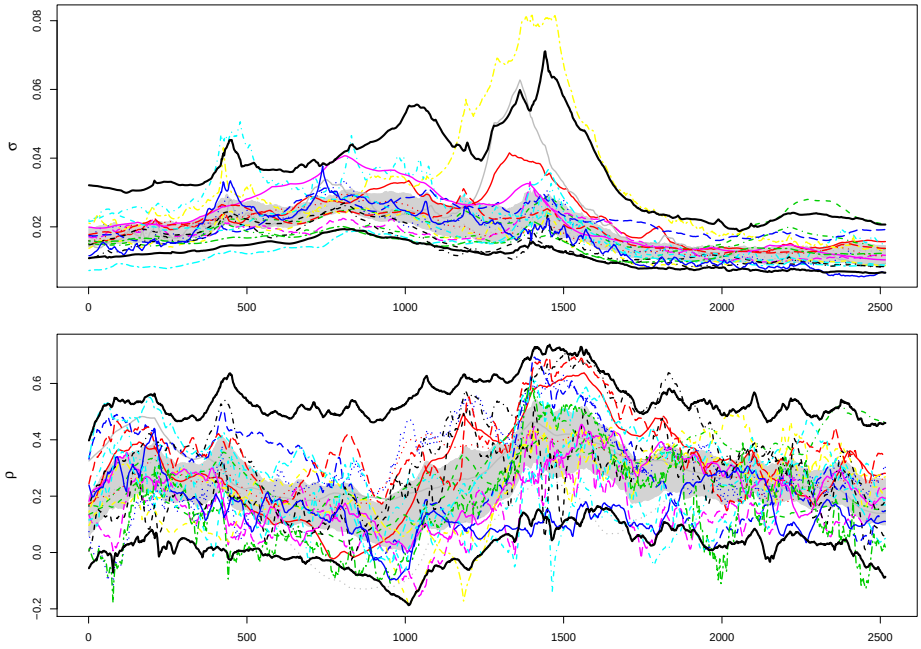

Figure 21: *S&P100 data, $p = 94$, prior 2.* Top panel displays $\hat{\sigma}_{it}$. Bottom panel displays $\hat{\rho}_{ijt}$.