Bayesian Analysis (2014) **9**, Number 4, pp. 759–792

A Tractable State-Space Model for Symmetric Positive-Definite Matrices

Jesse Windle * and Carlos M. Carvalho †

Abstract. The Bayesian analysis of a state-space model includes computing the posterior distribution of the system's parameters as well as its latent states. When the latent states wander around $\mathbb{R}^n$ there are several well-known modeling components and computational tools that may be profitably combined to achieve this task. When the latent states are constrained to a strict subset of $\mathbb{R}^n$ these models and tools are either impaired or break down completely. State-space models whose latent states are covariance matrices arise in finance and exemplify the challenge of devising tractable models in the constrained setting. To that end, we present a state-space model whose observations and latent states take values on the manifold of symmetric positive-definite matrices and for which one may easily compute the posterior distribution of the latent states and the system's parameters as well as filtered distributions and one-step ahead predictions. Employing the model within the context of finance, we show how one can use realized covariance matrices as data to predict latent time-varying covariance matrices. This approach out-performs factor stochastic volatility.

Keywords: backward sample, forward filter, realized covariance, stochastic volatility

1 Introduction

A state-space model is often characterized by an observation density $f(y_t|x_t)$ for the responses $\{y_t\}_{t=1}^T$ and a transition density $g(x_t|x_{t-1})$ for the latent states $\{x_t\}_{t=1}^T$. Usually, the latent states can take on any value in $\mathbb{R}^n$; however, there are times when the states or the responses are constrained to a manifold embedded in $\mathbb{R}^n$. For instance, econometricians and statisticians have devised symmetric positive-definite matrix-valued statistics that can be interpreted as noisy observations of the conditional covariance matrix of a vector of daily asset returns. In that case, it is reasonable to consider a state-space model that has covariance matrix-valued responses (the statistics) characterized by $f(\mathbf{Y}_t|\mathbf{V}_t)$ and covariance matrix-valued latent quantities (the time-varying covariance matrices) characterized by $g(\mathbf{V}_t|\mathbf{V}_{t-1})$. (In general, we will write matrices as bold capital letters and vectors as bold lower case letters.)

Unfortunately, devising state-space models on curved spaces, like the set of covariance matrices, that lend themselves to Bayesian analysis is not easy. Just writing down the observation and transition densities can be difficult in this setting, since one must define sensible distributions on structured subsets of $\mathbb{R}^n$. Asking that these densities

*Duke University jesse.windle@stat.duke.edu
†The University of Texas at Austin carlos.carvalho@mccombs.utexas.edu

 DOI:10.1214/14-BA888

then lead to some recognizable posterior distribution for the latent states and the system's parameters compounds the problem. Filtering is slightly less daunting, since one can appeal to approximate or sequential methods. For instance, Tyagi and Davis (2008) develop a Kalman-like filter (Kalman 1960) for symmetric positive-definite matrices while Hauberg et al. (2013) develop an algorithm similar to the unscented Kalman filter (Julier and Uhlmann 1997) for the more general setting of geodesically complete manifolds. Sequential approaches to filtering for similar problems can be found in Srivastava and Klassen (2004), Tompkins and Wolfe (2007), and Choi and Christensen (2011) where the latent states take values on the Grassman manifold, the Steifel manifold, and the special Euclidean group respectively. (Filtering or forward filtering refers to iteratively deriving the distributions $p(x_t|\mathcal{D}_t, \boldsymbol{\theta})$; $\mathcal{D}_t$ is the data $\{y_s\}_{s=1}^t$ and $\boldsymbol{\theta}$ represents the system's parameters.)

However, forward filtering is just one component of the Bayesian analysis of state-space models. The complete Bayesian analysis of any state-space model requires that one be able to sample from the posterior distribution of the latent states and the system's parameters. The latter is important in practice, since one cannot even forward filter without sampling or at least estimating the system's unknown parameters.

We address these issues for a state-space model with symmetric positive-definite or positive semi-definite rank-k observations and symmetric positive-definite latent states. (Let $\mathcal{S}_{m,k}^+$ denote the set of order m, rank k, symmetric positive semi-definite matrices and let $\mathcal{S}_m^+$ denote the set of order m, symmetric positive-definite matrices.) The model builds on the work of Uhlig (1997), who showed how to construct a state-space model with $\mathcal{S}_{m,1}^+$ observations and $\mathcal{S}_m^+$ hidden states and how, using this model, one can forward filter in closed form. We extend his approach to observations of arbitrary rank and show how to forward filter, how to backward sample, and how to marginalize the hidden states to estimate the system's parameters, all without appealing to fanciful Markov chain Monte Carlo (MCMC) schemes. (Backward sampling refers to taking a joint sample of the posterior distribution of the latent states $p(\{x_t\}_{t=1}^T|\mathcal{D}_T, \boldsymbol{\theta})$ using the conditional distributions $p(x_t|x_{t+1}, \mathcal{D}_t, \boldsymbol{\theta})$.) The model's estimates and one-step ahead predictions are exponentially weighted moving averages (also called geometrically weighted moving averages). Exponentially weighted moving averages are known to provide simple and robust estimates and forecasts in many settings (Brown 1959). We find this to be the case within the context of multivariate volatility forecasting in finance. Specifically, we show that the one-step ahead predictions of the covariance matrix of daily asset returns generated by our Uhlig-like model, when using realized covariance matrices as data, out-performs factor stochastic volatility.

1.1 A Comment on the Original Motivation

Our interest in covariance-valued state-space models arose from studying the realized covariance statistic, which within the context of finance, roughly speaking, can be thought of as a good estimate of the conditional covariance matrix of a vector of daily asset returns. (The daily period is somewhat arbitrary; one may pick any reasonably "large" period.) We had been exploring the performance of factor stochastic volatility mod-

els, along the lines of Aguilar and West (2000), which use daily returns, versus exponentially weighted moving averages of realized covariance matrices and found that exponentially smoothing realized covariance matrices out-performed the more complicated factor stochastic volatility models. (Exponential smoothing refers to iteratively calculating a geometrically weighted average of observations and some initial value.) As Bayesians, we wanted to find a model-based approach that is capable of producing similar results and the following fits within that role. To that end, as shown in Section 5, this simple model, used in conjunction with realized covariances, provides better one-step ahead predictions of daily covariance matrices than factor stochastic volatility (which only uses daily returns).

2 Background

There are at least three prominent types of statistical models for modeling latent dynamic covariance matrices: (1) factor-like models, (2) GARCH-like models, and (3) stochastic volatility-like models. (GARCH refers to generalized autoregressive conditional heteroskedasticity.) In general, one observes either a vector, $\mathbf{r}_t$, or a covariance matrix, $\mathbf{Y}_t$, whose conditional distribution, $\mathbb{P}(\mathbf{V}_t)$, depends upon a sequence of covariance matrices, $\{\mathbf{V}_t\}_{t=1}^T$, that are correlated across time. While one can be agnostic about the specific setting in which these models are put to use, we find it helpful to think in terms of finance, in which case the observation is either a vector of asset returns, $\mathbf{r}_t$, or a realized covariance matrix, $\mathbf{Y}_t$. Thus, we use the vernacular of finance even though these models are applicable outside of that setting. (Realized covariance matrices are symmetric positive definite matrix-valued statistics of high-frequency asset price data.) It will also be helpful to think of the period over which one computes asset returns or realized covariance matrices to be a day or week—anything but high-frequency time scales. Both factor-like models and GARCH-like models differ significantly from the path followed in this paper, which is aligned with what we call stochastic volatility-like models, and hence we do not discuss them here. (Factor-like models are discussed further in Appendix 6; Bauwens et al. (2006) survey GARCH-like models.)

Univariate stochastic volatility models treat dynamic variances as stochastic processes; thus, the multivariate analog is to treat dynamic covariance matrices as stochastic processes as well. But, constructing multivariate stochastic volatility models is nontrivial. In particular, it is a challenge to construct a reasonable matrix-valued stochastic process that (1) respects positive definiteness and (2) couples nicely to the observation distribution. Often, ensuring positive definiteness is most easily accomplished by defining the process in a different, unconstrained coordinate system. However, such transformations tend to make it more difficult to do state-space inference, since the product of the observation and transition densities does not likely yield a recognizable and easily simulated posterior distribution for the latent states.

In the univariate case, reconciliation is possible. Within the context of finance, the observation distribution is often

$$r_t \sim N(0, v_t)$$

where $\{r_t\}_{t=1}^T$ are an asset's returns and $\{v_t\}_{t=1}^T$ is the variance process. One must pick a stochastic process for $\{v_t\}_{t=1}^T$ that couples to this observation equation in a way that yields a tractable posterior distribution $p(\{v_t\}_{t=1}^T, \boldsymbol{\theta} | \{r_t\}_{t=1}^T)$ for the hidden states $\{v_t\}_{t=1}^T$ and any system parameters $\boldsymbol{\theta}$. The most common approach is to model the log variances $h_t = \log v_t$ instead of the variances—that is, to use a more convenient coordinate system. (Taylor (1982) is often credited with initiating this path. Shephard (2005) provides an excellent review and anthology of stochastic volatility.) In the simplest case, h_t is assumed to be an AR(1) process. The key insight is that one may take the transformed observation equation,

$$\log r_t^2 = h_t + \nu_t, \quad \nu_t \sim \log \chi_1^2,$$

and introduce the auxiliary variables $\{\eta_t\}_{t=1}^T$, such that $(\nu_t | \eta_t) \sim N(0, \eta_t)$ marginalizes to a $\log \chi_1^2$ distribution or an approximation thereof, so that, conditional upon $\{\eta_t\}_{t=1}^T$, the observation and evolution equations are

$$\begin{cases} \log r_t^2 = h_t + \varepsilon_t, & \varepsilon_t \sim N(0, \eta_t), \\ h_t = \mu + \phi(h_{t-1} - \mu) + \omega_t, & \omega_t \sim N(0, W), \end{cases}$$

which is just a dynamic linear model. One can then sample from the posterior distribution of the latent states and system parameters using the usual MCMC techniques. In particular, one can forward filter and backward sample to efficiently generate a joint draw from the conditional posterior distribution $p(\{h_t\}_{t=1}^T | \{r_t\}_{t=1}^T, \boldsymbol{\theta})$. (Forward filtering and backward sampling was introduced by Carter and Kohn (1994) and Früwirth-Schnatter (1994).)

One can try this approach in the multivariate case. Generalizing the observation equation slightly, assume that the response is a matrix $\mathbf{Y}_t$ whose distribution, conditional upon the latent $m \times m$ covariances $\{\mathbf{V}_t\}_{t=1}^T$, is

$$\mathbf{Y}_t \sim W_m(k, \mathbf{V}_t / k). \tag{1}$$

(The observation equation $\mathbf{r}_t \sim N(0, \mathbf{V}_t)$ is just a special case of (1), since $\mathbf{r}_t \mathbf{r}_t' \sim W_1(1, \mathbf{V}_t)$.)

There are at least two ways to transform the coordinates of $\mathbf{V}_t$, or the inverse of $\mathbf{V}_t$, into an unconstrained coordinate system in $\mathbb{R}^n$, $n = m(m+1)/2$. One can transform the latent covariance matrices using the matrix logarithm and then vectorize the lower diagonal portion: $\mathbf{W}_t = \log \mathbf{V}_t$, $\mathbf{w}_t = \text{vech}(\mathbf{W}_t)$, where "vech" vectorizes the lower diagonal portion of a matrix. (The matrix logarithm is defined as $\log \mathbf{V} = \mathbf{U} \boldsymbol{\Delta} \mathbf{U}'$ where $\mathbf{U}\mathbf{D}\mathbf{U}'$ is the eigenvalue decomposition of $\mathbf{V}$ and $\boldsymbol{\Delta}$ is diagonal with $\boldsymbol{\Delta}_{ii} = \log \mathbf{D}_{ii}, i = 1, \ldots, m$. Chiu et al. (1996) describe properties of the matrix logarithm and show how it can be used to linearly model covariance matrices.) Or, one can model the covariance matrices in, essentially, the coordinates of its Cholesky decomposition: $\mathbf{W}_t = \text{lower chol } \mathbf{V}_t$, $\mathbf{w}_t' = (\log \text{vecd}(\mathbf{W}_t)', \text{vecl}(\mathbf{W}_t)')$, where "vecd" maps the diagonal portion of a matrix to a vector and "vecl" vectorizes the lower off-diagonal portion of a matrix. There are variations on the latter; for instance, one could use the $\mathbf{LDL}'$

factorization of $\mathbf{V}_t$ and arrive at a similar set of coordinates. In either case, the vector $\mathbf{w}_t$ is allowed to take any value in $\mathbb{R}^n$, since for any value in $\mathbb{R}^n$ one can invert $\mathbf{w}_t$ and recover a symmetric positive definite matrix.

Whatever stochastic process one chooses to put on $\{\mathbf{w}_t\}_{t=1}^T$ induces a stochastic process on $\{\mathbf{V}_t\}_{t=1}^T$. Bauer and Vorkink (2011) exploit the log-based coordinates while Chiriac and Voev (2010) and Loddo et al. (2011) take advantage of Cholesky-based coordinates. Both Bauer and Vorkink (2011) and Chiriac and Voev (2010) are interested in modeling realized covariance matrices directly, as opposed to treating them as noisy observations of some true latent covariance matrix; thus, they need not worry about how these stochastic processes interact with an observation equation, while Loddo et al. (2011) use the Cholesky-based coordinates to devise a factor-like model. Adapting any of these approaches to generate a tractable matrix-variate state-space model is difficult.

To mimic univariate stochastic volatility using any of these coordinate systems, one would need to find a transformation g and auxiliary variables so that, conditional upon those auxiliary variables, $g(\mathbf{Y}_t)$ is a linear model in $\mathbf{w}_t$. We are not aware of any successful attempts to do this. Thus, what works in the univariate case fails in the multivariate case. One could use a Metropolis-Hastings step to update the covariance matrices $\{\mathbf{V}_t\}_{t=1}^T$ one at a time, but as seen with other dynamic generalized linear models, this will make the samples of $\{\mathbf{V}_t\}_{t=1}^T$ much more correlated than if one were able to draw $\{\mathbf{V}_t\}_{t=1}^T$ jointly, like one can with forward filtering and backward sampling. (Windle et al. (2013) discuss this point further within the context of generalized dynamic linear models for binary or count data.)

Since the aforementioned approach breaks down, it appears that one should at least attempt to define a stochastic process on covariance matrices, in the coordinates of covariance matrices, that plays nicely with the multivariate observation equation (1). To that end, let us again consider the univariate case. Suppose we want to construct a Markov process $\{x_t\}_{t=1}^T$. When x_t can take on any real value, it is natural to use an additive model where $x_t = f(x_{t-1}) + \varepsilon_t$ and the innovations $\{\varepsilon_t\}_{t=1}^T$ can take on any real value. However, if x_t is constrained to be positive this does not make sense, since a sequence of negative innovations might force x_t to be negative as well. One remedy in the constrained case is to require ε_t to be positive and to pick f so that it shrinks x_t towards zero. Another option is to consider a multiplicative model in which $x_t = f(x_{t-1})\psi_t$ and the innovations $\{\psi_t\}_{t=1}^T$ are positive. The latter path is easier in the multivariate setting.

The analogous multiplicative process for covariance matrices is to let $\{\mathbf{\Psi}_t\}_{t=1}^T$ be a sequence of independent and identically distributed symmetric positive definite innovations and then to define

$$\mathbf{X}_t = \mathbf{S}_t \mathbf{\Psi}_t \mathbf{S}_t', \quad \mathbf{S}_t \mathbf{S}_t' = f(\mathbf{X}_{t-1}).$$

So long as the square root $\mathbf{S}_t$ is not singular, $\mathbf{X}_t$ will be symmetric and positive definite. One must be somewhat careful when choosing the distribution of $\mathbf{\Psi}_t$. In particular, one should pay attention to whether the distribution is invariant to transformations of the form $\mathbf{O}\mathbf{\Psi}_t\mathbf{O}'$ where $\mathbf{O}$ is an orthogonal matrix. If the distribution does not possess this

Source	$f(\mathbf{X}_{t-1})$	$\mathbf{\Psi}_t$
Philipov and Glickman (2006)	$\mathbf{A}^{1/2}\mathbf{X}_{t-1}^{d}\mathbf{A}^{1/2'}$	$W_m(\rho, \mathbf{I}_m/\rho)$
Asai and McAleer (2009)	$\mathbf{X}_{t-1}^{d/2}\mathbf{A}\mathbf{X}_{t-1}^{d/2}$	$W_m(\rho, \mathbf{I}_m/\rho)$
Uhlig (1997)	$\lambda^{-1}\mathbf{X}_{t-1}$	$\beta_m\left(\frac{n}{2}, \frac{1}{2}\right)$.

Table 1: Transformations and innovation distributions for multiplicative matrix variate processes.

invariance, then it matters how one computes the square root $\mathbf{S}_t$. Table 1 lists a few possible choices[1,2] for the transformation f and the innovation distribution. We list Uhlig (1997), since this is our primary motivation; however, Triantafyllopoulos (2008) has since extended Uhlig's approach using the transformation $f(\mathbf{X}_{t-1}) = \mathbf{A}^{1/2}\mathbf{X}_{t-1}\mathbf{A}^{1/2}$. Jin and Maheu (2012) suggest various extensions to the transformations of Philipov and Glickman (2006) and Asai and McAleer (2009) by letting f depend on not only $\mathbf{X}_{t-1}$, but also on $\mathbf{X}_{t-k}, k = 2, \ldots, p$. (We have strayed slightly from the exact scenarios examined by Asai and McAleer (2009) and Jin and Maheu (2012), but these deviations are not important for the present discussion.) Our contribution can be traced to Uhlig's initial proposal; in particular, we study the case that $\mathbf{\Psi}_t \sim \beta_m(n/2, k/2)$ for $k \neq 1$. Prado and West (2010) do this as well, but as we show below, our approach is more flexible.

Each of the proposed transformations and innovation distributions characterize a stochastic process $\{\mathbf{X}_t\}_{t=1}^T$. We may couple this stochastic process to the observation distribution (1) by setting $\mathbf{V}_t = \mathbf{X}_t^{-1}$ to get a state-space model:

$$\begin{cases} \mathbf{Y}_t \sim W_m(k, (k\mathbf{X}_t)^{-1}) \\ \mathbf{X}_t = \mathbf{S}_t\mathbf{\Psi}_t\mathbf{S}_t', \;\; \mathbf{S}_t\mathbf{S}_t' = f(\mathbf{X}_{t-1}), \end{cases}$$

$\{\mathbf{\Psi}_t\}_{t=1}^T$ are independent and identically distributed.

As mentioned above, an essential feature of univariate stochastic volatility is that it is amenable to forward filtering and backward sampling, which lets one take a joint draw of the latent states conditional upon the data and the system parameters. But it is a challenge to replicate this property is the multivariate case. Both Philipov and Glickman (2006) (p. 326) and Asai and McAleer (2009) (p. 191) resort to sampling the latent states one at a time. That is, to simulate from the distribution $p(\{\mathbf{X}_t\}_{t=1}^T | \{\mathbf{Y}_t\}_{t=1}^T)$, they suggest using Gibbs sampling whereby one iteratively draws samples from $p(\mathbf{X}_t | \mathbf{X}_{-t}, \{\mathbf{Y}_s\}_{s=1}^T)$ for $t = 1, \ldots, T$. (We implicitly condition on the system's parameters; $\mathbf{X}_{-t}$ denotes the sequence $\{\mathbf{X}_s\}_{s=1}^T$ with the tth element removed.) Further, both cases lack conjugacy so that $p(\mathbf{X}_t | \mathbf{X}_{-t}, \{\mathbf{Y}_s\}_{s=1}^T)$ is an unknown distribution and must be sampled via Metropolis-Hastings. Thus, the posterior samples are

[1] Raising a matrix to a non-integer power is defined like the matrix logarithm.

[2] By the Schur product theorem, the composition by Hadamard product, $\mathbf{B} \odot f$, of any symmetric positive definite matrix $\mathbf{B}$ and f is also a valid transformation.

likely to display much more autocorrelation than if one were able to jointly sample the latent states using known distributions. The Metropolis-Hastings step likely compounds this problem: presumably, as the dimension of the covariance matrices grows the acceptance rate of the Metropolis-Hastings step will diminish. In contrast, our approach, based upon Uhlig's work, can be forward filtered and backward sampled with known and easily simulated distributions. Thus, it should have better effective sampling rates and it should scale more easily to large problems.

While all of these stochastic processes lead to tractable state-space models, they are somewhat unsatisfying as descriptions of dynamic covariance matrices since they may not be stationary. Following an argument from Philipov and Glickman (2006) (p. 316), suppose the innovations $\{\mathbf{\Psi}_t\}_{t=1}^T$ are positive definite and that, for instance,

$$f(\mathbf{X}_{t-1}) = \mathbf{A}^{1/2}\mathbf{X}_{t-1}^d\mathbf{A}^{1/2}.$$

If the stochastic process $\{\mathbf{X}_t\}_{t=1}^T$ is stationary, then so is the log determinant process, $z_t = \log|\mathbf{X}_t|$, which evolves as

$$z_t = \log|\mathbf{A}| + dz_{t-1} + \log|\mathbf{\Psi}_t|.$$

When $d = 1$, as is the case for the model we study, z_t is a random walk, and a random walk with drift unless $\mathbf{A}$ is chosen very carefully. In either case, $\{z_t\}_{t=1}^T$ is not stationary, and hence neither is $\{\mathbf{X}_t\}_{t=1}^T$. Philipov and Glickman (2006) go on to say that their stochastic process $\{\mathbf{X}_t\}_{t=1}^T$ defined with $|d| < 1$ may be stationary, but that they cannot prove this to be the case. A similar remark holds for the work of Asai and McAleer (2009). It is unfortunate that these processes are non-stationary or that one is unable to show that they are stationary, since we believe that should be the case for latent covariances matrices within the context of finance.

Other covariance matrix-valued stochastic processes one might consider include the Wishart autoregressive process (Gourieroux et al. 2009) and the inverse Wishart autoregressive process (Fox and West 2011). Gourieroux et al. (2009) use their process to model realized covariance matrices directly. But one encounters difficulties when using this process to model the latent states in a covariance matrix-valued state-space model. In particular, the posterior distribution of the latent states is complicated, which dashes the hopes of easily sampling the latent states or the system's parameters. Fox and West (2011) study their inverse Wishart autoregressive process, which is inspired in part by the univariate processes studied by Pitt and Walker (2005), within the context of state-space modeling. However, a similar issue arises: one cannot forward filter in closed form (Fox and West 2011, p. 12), let alone backwards sample; nor can one easily draw the system's parameters.

Recapitulating, there are a variety of symmetric positive definite matrix-valued state-space models, each with a different evolution equation for the hidden states; but for all of those models it is difficult to simulate from the posterior distribution of the hidden states and the system's parameters. We can avoid this problem by building upon the work of Uhlig (1997).

3 A Covariance Matrix-Valued State-Space Model

The model herein is closely related to several models found in the Bayesian literature, all of which have their origin in variance discounting techniques (Quintana and West 1987; West and Harrison 1997). Uhlig (1997) provided a rigorous justification for variance discounting, showing that it is a form of Bayesian filtering for covariance matrices, and our model can be seen as a direct extension of Uhlig's work. (Shephard (1994) constructs a similar model for the univariate case.) The model of Prado and West (2010) (p. 273) is similar to ours, though less flexible.

Uhlig (1997) considers observations, $\mathbf{r}_t \in \mathbb{R}^m$, $t = 1, \ldots, T$, that are conditionally Gaussian given the hidden states $\{\mathbf{X}_t\}_{t=1}^T$, which take values in $\mathcal{S}_m^+$. In particular, assuming $\mathbb{E}[\mathbf{r}_t] = 0$, his model is

$$\begin{cases} \mathbf{r}_t \sim N(0, \mathbf{X}_t^{-1}), \\ \mathbf{X}_t = \mathbf{T}_{t-1}' \mathbf{\Psi}_t \mathbf{T}_{t-1} / \lambda, & \mathbf{\Psi}_t \sim \beta_m\left(\frac{n}{2}, \frac{1}{2}\right), \\ \mathbf{T}_{t-1} = \text{upper chol } \mathbf{X}_{t-1}, \end{cases}$$

where $n > m - 1$ is an integer and β_m is the multivariate beta distribution, which is defined in Appendix 6. This model possesses closed form formulas for forward filtering that only require knowing the outer product $\mathbf{r}_t \mathbf{r}_t'$; thus, one may arrive at equivalent estimates of the latent states by letting $\mathbf{Y}_t = \mathbf{r}_t \mathbf{r}_t'$ and using the observation distribution

$$\mathbf{Y}_t \sim W_m(1, \mathbf{X}_t^{-1})$$

where $W_m(1, \mathbf{X}_t^{-1})$ is the order m Wishart distribution with 1 degree of freedom and scale matrix $\mathbf{X}_t^{-1}$ as defined in Appendix 6. We show that one can extend this model for $\mathbf{Y}_t$ of any rank via

$$\begin{cases} \mathbf{Y}_t \sim W_m(k, (k\mathbf{X}_t)^{-1}), \\ \mathbf{X}_t \sim \mathbf{T}_{t-1}' \mathbf{\Psi}_t \mathbf{T}_{t-1} / \lambda, & \mathbf{\Psi}_t \sim \beta_m\left(\frac{n}{2}, \frac{k}{2}\right), \\ \mathbf{T}_{t-1} = \text{upper chol } \mathbf{X}_{t-1}, \end{cases} \tag{UE}$$

where $n > m - 1$ and k is a positive integer less than m or is a real number greater than $m - 1$. (When k is a positive integer less than m, $\mathbf{Y}_t$ has rank k.) Many of the mathematical ideas needed to motivate model (UE) (for Uhlig extension) can be found in a sister paper (Uhlig 1994) to the Uhlig (1997) paper, and Uhlig could have written down the above model given those results; though, he was focused specifically on the rank-deficient case, and the rank-1 case in particular, as his 1997 work shows. We contribute to this discourse by (1) constructing the model in a fashion that makes sense for observations of all ranks, by (2) showing that one may backward sample to generate a joint draw of the hidden states, and by (3) demonstrating that one may marginalize the hidden states to estimate the system's parameters n, k, and λ.

Model (UE) has a slightly different form and significantly more flexibility than the

model of Prado and West (2010) (see p. 273), which is essentially

$$\begin{cases} \mathbf{Y}_t \sim W_m(\eta, \mathbf{X}_{t-1}^{-1}), & \eta \in \{1, \dots, m-1\} \cup (m-1, \infty), \\ \mathbf{X}_t = \mathbf{T}_{t-1}' \boldsymbol{\Psi}_t \mathbf{T}_{t-1}/\lambda, & \boldsymbol{\Psi}_t \sim \beta_m\left(\frac{\lambda h_{t-1}}{2}, \frac{(1-\lambda)h_{t-1}}{2}\right), \\ \mathbf{T}_{t-1} = \text{upper chol } \mathbf{X}_{t-1}, & \\ h_{t-1} = \lambda h_{t-2} + 1. & \end{cases}$$

In this case, as noted by Prado and West, λ is constrained "to maintain a valid model, since we require either $h_t > m-1$ or h_t be integral [and less than or equal to m]. The former constraint implies that λ cannot be too small, $\lambda > (m-2)/(m-1)$ defined by the limiting value [of h_t as t grows]." That is, when $\boldsymbol{\Psi}_t$ has full rank, $\lambda > (m-2)/(m-1)$ so λ must be close to unity for even moderately sized matrices (moderately large m); and when $\boldsymbol{\Psi}_t$ is rank deficient, $h_t = h$ must be constant and an integer and λ must be equal to $(h-1)/h$. Thus, in either case, the choice of λ is restricted. (We have replaced Prado and West's β by λ and their q by m.) The parameter λ is important since it controls how much the model smooths observations when forming estimates and one-step ahead predictions; thus the constraints on λ are highly undesirable. In contrast, our model lets λ take on any value.

Given (UE), we can derive several useful propositions. The proofs of these propositions, which synthesize and add to results from Uhlig (1994), Muirhead (1982), and Díaz-García and Jáimez (1997), are technical, and hence we defer their presentation to Appendix 6. Presently, we focus on the *closed form* formulas that one may use when forward filtering, backward sampling, predicting one step into the future, and estimating n, k, and λ.

First, some notation: inductively define the collection of data $\mathcal{D}_t = \{\mathbf{Y}_t\} \cup \mathcal{D}_{t-1}$ for $t = 1, \dots, T$ with $\mathcal{D}_0 = \{\boldsymbol{\Sigma}_0\}$ where $\boldsymbol{\Sigma}_0$ is some covariance matrix. Let the prior for $(\mathbf{X}_1|\mathcal{D}_0)$ be $W_m(n, (k\boldsymbol{\Sigma}_0)^{-1}/\lambda)$. In the following, we implicitly condition on the parameters n, k, and λ.

Proposition 1 (Forward Filtering)**.** *Suppose* $(\mathbf{X}_t|\mathcal{D}_{t-1}) \sim W_m(n, (k\boldsymbol{\Sigma}_{t-1})^{-1}/\lambda)$. *After observing* $\mathbf{Y}_t$*, the updated distribution is*

$$(\mathbf{X}_t|\mathcal{D}_t) \sim W_m(k+n, (k\boldsymbol{\Sigma}_t)^{-1})$$

where

$$\boldsymbol{\Sigma}_t = \lambda\boldsymbol{\Sigma}_{t-1} + \mathbf{Y}_t.$$

Evolving $\mathbf{X}_t$ *one step leads to*

$$(\mathbf{X}_{t+1}|\mathcal{D}_t) \sim W_m(n, (k\boldsymbol{\Sigma}_t)^{-1}/\lambda).$$

Proposition 2 (Backward Sampling)**.** *The joint density of* $(\{\mathbf{X}_t\}_{t=1}^T|\mathcal{D}_T)$ *can be decomposed as*

$$p(\mathbf{X}_T|\mathcal{D}_T) \prod_{t=1}^{T-1} p(\mathbf{X}_t|\mathbf{X}_{t+1}, \mathcal{D}_t)$$

(with respect to the T-fold product of $\mathcal{S}_m^+$ embedded in $\mathbb{R}^{m(m+1)/2}$ with Lebesgue measure) where the distribution of $(\mathbf{X}_t|\mathbf{X}_{t+1}, \mathcal{D}_t)$ is a shifted Wishart distribution

$$(\mathbf{X}_t|\mathbf{X}_{t+1}, \mathcal{D}_t) = \lambda\mathbf{X}_{t+1} + \mathbf{Z}_{t+1},\ \mathbf{Z}_{t+1} \sim W_m(k, (k\boldsymbol{\Sigma}_t)^{-1}).$$

Proposition 3 (Marginalization)**.** *The joint density of of $\{\mathbf{Y}_t\}_{t=1}^T$ is given by*

$$p(\{\mathbf{Y}_t\}_{t=1}^T|\mathcal{D}_0) = \prod_{t=1}^T p(\mathbf{Y}_t|\mathcal{D}_{t-1})$$

with respect to the differential form $\bigwedge_{t=1}^T (d\mathbf{Y}_t)$ where $(d\mathbf{Y}_t)$ is as found in Definition 4 for either the rank-deficient or full-rank cases, depending on the rank of $\mathbf{Y}_t$. (Differential forms, otherwise known as K-forms, are vector fields that may be used to simplify multivariate analysis. In particular, one may define densities with respect to differential forms. Mikusiński and Taylor (2002) provide a good introduction to differential forms while Muirhead (1982) shows how to use them in statistics.) The density $p(\mathbf{Y}_t|\mathcal{D}_{t-1})$ is

$$\pi^{-(mk-k^2)/2} \frac{\Gamma_m(\frac{\nu}{2})}{\Gamma_m(\frac{n}{2})\Gamma_m(\frac{k}{2})} \frac{|\mathbf{L}_t|^{(k-m-1)/2}|\mathbf{C}_t|^{n/2}}{|\mathbf{C}_t + \mathbf{Y}_t|^{\nu/2}}$$

with respect to $(d\mathbf{Y}_t)$ in the rank-deficient case and is

$$\frac{\Gamma_m(\frac{\nu}{2})}{\Gamma_m(\frac{n}{2})\Gamma_m(\frac{k}{2})} \frac{|\mathbf{Y}_t|^{(k-m-1)/2}|\mathbf{C}_t|^{n/2}}{|\mathbf{C}_t + \mathbf{Y}_t|^{\nu/2}}$$

with respect to $(d\mathbf{Y}_t)$ in the full-rank case, where $\nu = n + k$, and $\mathbf{C}_t = \lambda\boldsymbol{\Sigma}_{t-1}$ with $\boldsymbol{\Sigma}_t = \lambda\boldsymbol{\Sigma}_{t-1} + \mathbf{Y}_t$ like above.

Examining the one-step ahead forecasts of $\mathbf{Y}_t$ elucidates how the model smooths. Invoking the law of iterated expectations, one finds that $\mathbb{E}[\mathbf{Y}_{t+1}|\mathcal{D}_t] = \mathbb{E}[\mathbf{X}_{t+1}^{-1}|\mathcal{D}_t]$. Since $(\mathbf{X}_{t+1}^{-1}|\mathcal{D}_t)$ is an inverse Wishart distribution, its expectation is proportional to $\boldsymbol{\Sigma}_t$. Solving the recursion for $\boldsymbol{\Sigma}_t$ from Proposition 1 shows that

$$\boldsymbol{\Sigma}_t = \sum_{i=0}^{t-1} \lambda^i \mathbf{Y}_{t-i} + \lambda^t \boldsymbol{\Sigma}_0. \tag{2}$$

Thus, the forecast of $\mathbf{Y}_{t+1}$ will be a scaled, geometrically weighted sum of the previous observations. If, further, one enforces the constraint

$$\frac{1}{\lambda} = 1 + \frac{k}{n-m-1} \tag{3}$$

then taking a step from $\mathbf{X}_t$ to $\mathbf{X}_{t+1}$ does not change the latent state's harmonic mean, that is $\mathbb{E}[\mathbf{X}_t^{-1}|\mathcal{D}_t] = \mathbb{E}[\mathbf{X}_{t+1}^{-1}|\mathcal{D}_t]$. It also implies that the one-step ahead point forecast of $(\mathbf{X}_{t+1}^{-1}|\mathcal{D}_t)$ is

$$\mathbb{E}[\mathbf{X}_{t+1}^{-1}|\mathcal{D}_t] = (1-\lambda)\boldsymbol{\Sigma}_t = (1-\lambda)\sum_{i=0}^{t-1} \lambda^i \mathbf{Y}_{t-i} + (1-\lambda)\lambda^t \boldsymbol{\Sigma}_0. \tag{4}$$

Hence in the constrained case, the one-step ahead forecast is essentially the geometrically weighted average of past observations. For a geometrically weighted average, the most recent observations are given more weight as λ decreases. It has been known for some time that such averages provide decent one-step ahead forecasts (Brown 1959).

4 Multiple Smoothing Parameters

As mentioned earlier, our initial goal was to find a tractable state-space model whose one-step ahead forecasts were like those generated by exponential smoothing, and as shown above, this is exactly what we have done. However, we can introduce a more sophisticated smoothing procedure via

$$\begin{cases} \mathbf{Y}_t \sim W_m(k, (k\mathbf{X}_t)^{-1}), \\ \mathbf{X}_t \sim \mathbf{R}^{-1'}\mathbf{T}'_{t-1}\boldsymbol{\Psi}_t\mathbf{T}_{t-1}\mathbf{R}^{-1}, \quad \boldsymbol{\Psi}_t \sim \beta_m\left(\frac{n}{2}, \frac{k}{2}\right), \\ \mathbf{T}_{t-1} = \text{upper chol } \mathbf{X}_{t-1}, \end{cases} \tag{5}$$

where $\mathbf{R}$ is a square matrix. When $\mathbf{R} = \sqrt{\lambda}\mathbf{I}_m$ we recover the initial model. Under model (5), Propositions 1 through 3 above are essentially unchanged, except that now $\boldsymbol{\Sigma}_t = \mathbf{R}\boldsymbol{\Sigma}_{t-1}\mathbf{R}' + \mathbf{Y}_t$ so that

$$\boldsymbol{\Sigma}_t = \sum_{i=0}^{t-1} \mathbf{R}^i \mathbf{Y}_{t-i} \mathbf{R}^{i'} + \mathbf{R}^t \boldsymbol{\Sigma}_0 \mathbf{R}^{t'}.$$

It is still the case that $\mathbb{E}[\mathbf{Y}_t|\mathcal{D}_{t-1}] = \mathbb{E}[\mathbf{X}_t^{-1}|\mathcal{D}_{t-1}] = k\mathbf{C}_t/(n-m-1)$, where $\mathbf{C}_t = \mathbf{R}\boldsymbol{\Sigma}_{t-1}\mathbf{R}'$; however, when $\mathbf{R}$ is not $\sqrt{\lambda}\mathbf{I}_m$ we no longer can constrain $\mathbf{R}$, n, and k so to ensure that $\mathbb{E}[\mathbf{X}_{t+1}^{-1}|\mathcal{D}_t] = \mathbb{E}[\mathbf{X}_t^{-1}|\mathcal{D}_t]$ nor can we arrive at an exponential smoothing interpretation like above. We will show that, for our financial example at least, introducing this more complicated structure does not improve predictive performance.

5 Example: Covariance Forecasting

As noted initially, (UE) is an extension of the model proposed by Uhlig (1997). For the original model, when $k = 1$, one might consider observing a vector of heteroskedastic asset returns $\mathbf{r}_t \sim N(0, \mathbf{X}_t^{-1})$ where the precision matrix $\mathbf{X}_t$ changes at each time step. The extended model allows the precision matrix to change less often than the frequency with which the returns are observed. For instance, one may be interested in estimating the variance of the daily returns, assuming that the variance only changes from day to day, using *multiple* observations taken from *within* the day.

To that end, suppose the vector of intraday stock prices evolves as geometric Brownian motion so that on day t the m-vector of log prices is

$$\mathbf{p}_{t,s} = \mathbf{p}_{t,0} + \boldsymbol{\mu}s + \mathbf{V}_t^{1/2}(\mathbf{w}_{t+s} - \mathbf{w}_t)$$

at time s, where s the fraction of the trading day that has elapsed, $\{\mathbf{w}_s\}_{s\geq 0}$ is an m-dimensional Brownian motion, and $\mathbf{V}_t^{1/2}\mathbf{V}_t^{1/2'} = \mathbf{X}_t^{-1}$. In practice, $\boldsymbol{\mu}$ is essentially zero, so we will ignore that term. Further, suppose one has viewed the vector of prices at $k+1$ equispaced times throughout the day so that $\mathbf{r}_{t,i} = \mathbf{p}_{t,i} - \mathbf{p}_{t,i-1/k},\ i = 1/k, \ldots, 1$. Then $\mathbf{r}_{t,i} \sim N(0, \mathbf{X}_t^{-1}/k)$ and $\mathbf{Y}_t = \sum_{i=1}^{k} \mathbf{r}_{t,i}\mathbf{r}_{t,i}'$ is distributed as $W_m(k, \mathbf{X}_t^{-1}/k)$. Letting $\mathbf{X}_t = \mathcal{U}(\mathbf{X}_{t-1})'\boldsymbol{\Psi}_t\mathcal{U}(\mathbf{X}_{t-1})/\lambda$ where $\mathcal{U}(\cdot)$ computes the upper Cholesky decomposition and $\boldsymbol{\Psi}_t \sim \beta_m(n/2, k/2)$, we recover model (UE) exactly. Of course, in reality, returns are not normally distributed; they are heavy tailed and there are diurnal patterns within the day. Nonetheless, the realized covariance literature, which we discuss in more detail in Appendix 6, suggests that taking $\mathbf{Y}_t$ to be an estimate of the daily variance $\mathbf{X}_t^{-1}$ is a reasonable thing to do; though to suppose that the error is Wishart is a strong assumption. More dubious is the choice of the evolution equation for $\mathbf{X}_t$ since the subsequent stochastic process is not stationary. But the evolution equation for $\{\mathbf{X}_t\}_{t=1}^{T}$ does accommodate closed form forward filtering and backward sampling formulas and possesses only a few parameters, which makes it a relatively cheap model to employ.

The one mild challenge when applying the model is estimating $\boldsymbol{\Sigma}_0$. However, it is possible to "cheat" and not actually estimate $\boldsymbol{\Sigma}_0$ at all. Consider (2) and ponder the following two observations. First, $\boldsymbol{\Sigma}_t$ is a geometrically weighted sum in $\{\boldsymbol{\Sigma}_0, \mathbf{Y}_1, \ldots, \mathbf{Y}_t\}$. Second, the least important term in the sum is $\boldsymbol{\Sigma}_0$. Thus, one can reasonably ignore $\boldsymbol{\Sigma}_0$ if t is large enough. To that end, we suggest setting aside the first τ_1 observations and using $\{\boldsymbol{\Sigma}_{\tau_1}, \mathbf{Y}_{\tau_1+1}, \ldots, \mathbf{Y}_{\tau_2}\}$ where $\boldsymbol{\Sigma}_{\tau_1} = \sum_{i=0}^{\tau_1} \lambda^i \mathbf{Y}_{\tau_1-i}$ to learn n, k, and λ using Proposition 3 and the prior $p(\mathbf{X}_{\tau_1+1}|\mathcal{D}_{\tau_1}) \sim W_m(n, (k\boldsymbol{\Sigma}_{\tau_1})^{-1}/\lambda)$. It may seem costly to disregard the first τ_1 observations, but since there are so few parameters to estimate this is unlikely to be a problem—the remaining data will suffice.

This is the process used to generate Figure 1 (with $\tau_1 = 50$ and $\tau_2 = 100$). The data set follows the $m = 30$ stocks that comprised the Dow Jones Industrial Average in October, 2010. Eleven intraday observations were taken every trading day from February 27, 2007 to October 29, 2010 to produce 927 daily, rank-10 observations $\{\mathbf{Y}_t\}_{t=1}^{927}$. Since the observations are rank-deficient, we know that $k = 10$. (In the full-rank case, we will estimate k.) We constrain λ using (3) so that the only unknown is n. Given an improper flat prior for $n > 29$, the posterior mode is $n = 215$, implying that $\lambda = 0.95$, a not unusual value for exponential smoothing. Once n is set, one can forward filter, backward sample, and generate one-step ahead predictions in closed form. The right side of Figure 1 shows the filtered covariance between Alcoa Aluminum and American Express on the correlation scale.

However, one need not take such a literal interpretation of the model. Instead of trying to justify its use on first principles, one may simply treat it as a symmetric positive definite matrix-valued state-space model, which we do presently. As noted in the introduction and elaborated on in Appendix 6, the realized covariance matrix is a good estimate of the daily covariance matrix of a vector of financial asset returns. Since realized covariance matrices are good estimates it is natural to try to use them for prediction. The statistics themselves place very few restrictions on the distribution of asset prices and their construction is non-parametric. In other words, the construction

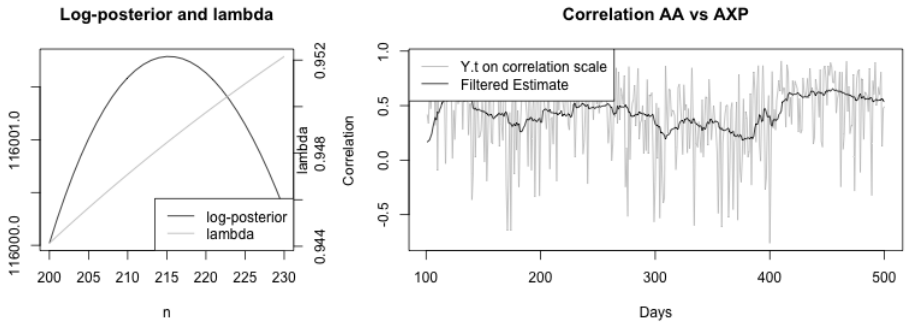


Figure 1: Level sets of the log posterior and filtered estimates on the correlation scale. On the left is the log posterior of n calculated using $\{\mathbf{\Sigma}_{50}, \mathbf{Y}_{51}, \ldots, \mathbf{Y}_{100}\}$ and constraint (3). The black line is the log posterior in n, which has a mode at $n = 215$ corresponding to $\lambda = 0.95$; $k = 10$ is fixed. The gray line is λ as a function of n. On the right are the values of $\mathbf{Y}_t$ and the estimate $\mathbb{E}[\mathbf{X}_t^{-1}|\mathcal{D}_t, n]$, $t = 101, \ldots, 500$, on the correlation scale, for Alcoa by American Express. A truncated time series was used to provide a clear picture.

of a realized covariance matrix (at least the construction we use) says little about the evolution of the latent daily covariance matrices.

But we do not need to know the exact evolution of the latent daily covariance matrices to employ model (UE) to make short-term predictions. To that end, we may treat realized covariance matrices $\{\mathbf{Y}_t\}_{t=1}^T$ as $\mathcal{S}_m^+$-valued data that track the latent daily covariances $\{\mathbf{X}_t^{-1}\}_{t=1}^T$. We construct the realized covariance matrices using the same $m = 30$ stocks over the same time period as above, but using all of the intraday data, which results in full-rank observations (see Appendix 6 for details). We follow the procedure outlined above to estimate k and n, and implicitly λ by constraint (3). Selecting an improper flat prior for $n > 29$ and $k > 29$ yields the log-posterior found in Figure 2. The posterior mode is at $(67, 396)$ implying $\lambda = 0.85$. The gray lines in Figure 2 correspond to level sets of λ in k and n. As seen in the figure, the uncertainty in (k, n) is primarily in the direction of the steepest ascent of λ. One can use Proposition 3 and the previously described method of generating Σ_{τ_1} to construct a random walk Metropolis sampler as well. Doing that we find the posterior mean to be $(67, 399)$, which implies an essentially identical estimate of λ. A histogram of the posterior of λ is in Figure 3, showing that, though the the direction of greatest variation in (k, n) corresponds to changes in λ, the subsequent posterior standard deviation of λ is small.

Recall, our original motivation for studying $\mathcal{S}_m^+$-valued state-space models was the observation that exponentially smoothing realized covariance matrices generates better one-step ahead predictions than factor stochastic volatility (FSV). In those initial experiments, we used cross-validation to pick the smoothing parameter λ. Figure 3 shows

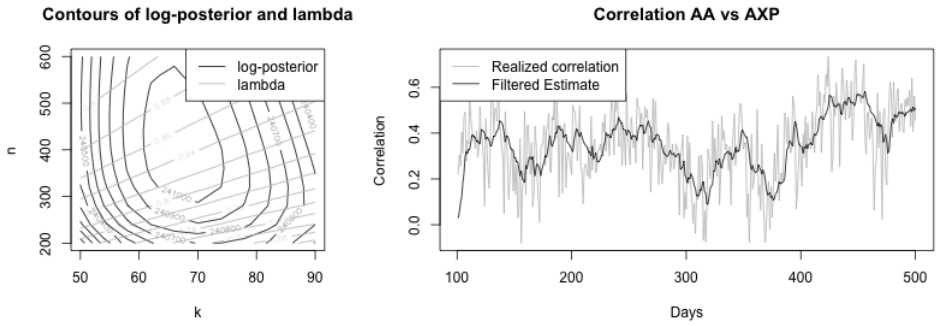


Figure 2: Level sets of the log posterior and filtered estimates on the correlation scale. On the left is the log posterior of (k, n) calculated using $\{\mathbf{\Sigma}_{50}, \mathbf{Y}_{51}, \ldots, \mathbf{Y}_{100}\}$ and constraint (3). The black line is the level set of the log posterior as a function of (k, n), which has a mode at $(67, 396)$ corresponding to $\lambda = 0.85$. The gray line is the level set of λ as a function of (k, n). On the right are the values of $\mathbf{Y}_t$ and the estimate $\mathbb{E}[\mathbf{X}_t^{-1}|\mathcal{D}_t, n, k]$, $t = 101, \ldots, 500$, on the correlation scale, for Alcoa by American Express.

that one arrives at the same conclusion for two different out-of-sample exercises using model UE. (The measures of performance are defined in the caption to Figure 3.)

To summarize the results: it is better to use our simple $\mathcal{S}_m^+$-valued state-space model with realized covariance matrices to make short term predictions than to use factor stochastic volatility with only daily returns. One could argue that this is an unfair comparison: factor stochastic volatility only uses opening and closing prices while model (UE) uses all of the prices observed throughout the day; hence, model (UE) has an inherent advantage. To address this claim, we also generate one-step ahead forecasts of the daily covariance matrices using an extended version of factor stochastic volatility that incorporates some information from the realized covariance matrices. The exact model can be found in Appendix 6. However, even in this case, model UE does better. Figure 4 plots the cumulative squared returns of the various out-of-sample minimum variance portfolios. Model (UE) appears to gain most of its advantage during the recent financial crisis.

Repeating this out-of-sample exercise for the Uhlig-like model with more smoothing parameters does not improve the predictive performance. In particular, we fit model (5) when $\mathbf{R} = \mathbf{D}$ is a diagonal matrix using the same method as described above with $\tau_1 = 50$ and $\tau_2 = 100$. An improper flat prior is placed on $n > 29$, $k > 29$ and $\log(D_{ii}), i = 1, \ldots, 30$. Using an independence Metropolis sampler, the posterior mean is at $(68, 334)$ for (k, n) and in $[0.82, 0.84]$ for D_{ii}^2, $i \in 1, \ldots, 30$. Since k and n are similar to the estimates for the simpler model and since $D'D$ is close to $0.85I_{30}$, it follows that adding the extra smoothing parameters will not alter the predictions much; indeed, the

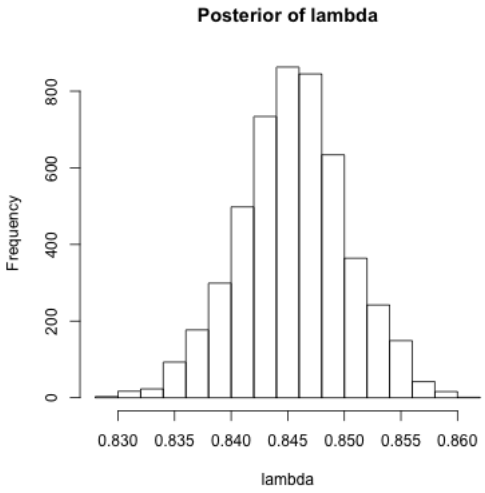
out-of-sample predictive performance is essentially the same as seen on the right hand side of Figure 3 and in Figure 4.

In all of the examples above, we did not account for uncertainty in the values of n, k, and the smoothing parameter (λ or $\mathbf{R}$). But, as seen in Lence and Hayes (1994a,b), it is sometimes important to include parameter uncertainty when making decisions, like when picking a one-step ahead portfolio. One simple way to incorporate such uncertainty into the exercises above is to sample from, for instance, $(n, k|\mathcal{D}_{\tau_1})$; compute the implied value of λ; generate one-step ahead forecasts, $\mathbb{E}[\mathbf{X}_t^{-1}|\mathcal{D}_{t-1}, n, k]$, using those sampled values; and then average over those forecasts. Doing this we find that the average forecasts produce essentially identical out-of-sample results for both the minimum variance portfolios (0.00929) and the predictive log-likelihood (96779). A more rigorous approach would average the forecasts $\mathbb{E}[\mathbf{X}_t^{-1}|\mathcal{D}_{t-1}, n, k]$ using values of n and k sampled from $(n, k|\mathcal{D}_{t-1})$, in effect computing $\mathbb{E}[\mathbf{X}_t^{-1}|\mathcal{D}_{t-1}]$. However, an examination of the modes of $(n, k|\mathcal{D}_{t-1})$ as t changes suggests that time variation in the distribution of $(n, k|\mathcal{D}_{t-1})$ is more important than the dispersion of the distribution $(n, k|\mathcal{D}_{t-1})$ for fixed t. To see how this time variation affects the results, we repeat the out-of-sample exercise above using the forecasts $\mathbb{E}[\mathbf{X}_t^{-1}|\mathcal{D}_{t-1}, n, k]$ where n and k are set to the mode of $(n, k|\mathcal{D}_{t-1})$. This procedure still performs better than the factor stochastic volatility-like models, though slightly worse than when the parameters n and k are chosen from $(n, k|\mathcal{D}_{\tau_1})$.

6 Discussion

Employing exponentially weighted moving averages to generate short-term forecasts is not new. These methods were popular at least as far back as the first half of the 20th century (Brown 1959). In light of this, it may seem that model (UE) is rather unglamorous. But this is only because we have explicitly identified how the model uses past observations to make predictions. In fact, many models of time-varying variances behave similarly. For instance, GARCH (Bollerslev 1986) does exponential smoothing with mean reversion to predict daily variances using squared returns. Stochastic volatility (Taylor 1982) does exponential smoothing with mean reversion to predict log variances using log square returns. Models that include a leverage effect do exponential smoothing so that the amount of smoothing depends on the direction of the returns. Thus, it should not be surprising or uninteresting when a state-space model generates predictions with exponential smoothing or some variation thereof.

This helps explains why a simple model (UE) with high-quality observations can generate better short-term predictions than a complicated model (factor stochastic volatility) with low-quality data. First, both models, in one way or another, are doing something similar to exponential smoothing. Second, the true covariance process seems to revert quite slowly. Thus, there will not be much difference between a one-step ahead forecast that lacks mean reversion (the Uhlig extension) and a one-step ahead forecast that includes mean reversion (factor stochastic volatility). Since the prediction mechanisms are similar, the model that uses a "higher resolution" snapshot of the latent



Model	MVP	PLLH
FSV 1	0.01021	94910
FSV 2	0.00992	95155
FSVE 1	0.00996	94220
FSVE 2	0.00977	94719
UE 1	0.00929	96781
UE 30	0.00930	96788

MVP: lower is better
PLLH: higher is better

Figure 3: The posterior of λ and the predictive performance of the Uhlig-like model using 1 and 30 smoothing parameters. On the left: the posterior of λ for model (UE) calculated using $\{\mathbf{\Sigma}_{50}, \mathbf{Y}_{51}, \ldots, \mathbf{Y}_{100}\}$, constraint (3), and the posterior samples of (k, n). On the right: the performance of model (UE) (UE 1) and its extension (UE 30) found in (5) when $\mathbf{R}$ is a diagonal matrix versus factor stochastic volatility (FSV) with one and two factors and an extension to factor stochastic volatility (FSVE) with one and two factors that uses information from realized covariance matrices (see Appendix 6). "MVP" stands for minimum variance portfolios and "PLLH" stands for predictive log-likelihood. For all of the models, a sequence of one-step ahead predictions of the latent covariance matrices $\{\hat{\mathbf{X}}_t^{-1}\}_{t=101}^{920}$ was generated. For model (UE), we set λ to be the posterior mode found from the data $\{\mathbf{\Sigma}_{50}, \mathbf{Y}_{51}, \ldots, \mathbf{Y}_{100}\}$, as described in Section 5, to generate the one-step ahead predictions. A similar procedure was followed for model (5). For the factor stochastic volatility-like models, we picked the point estimate $\hat{\mathbf{X}}_t^{-1}$ to be an approximation of the mean of $(\mathbf{X}_t^{-1}|\mathcal{F}_{t-1})$ where $\mathbf{r}_t$ is the vector of open to close log-returns on day t and $\mathcal{F}_t = \{\mathbf{r}_1, \ldots, \mathbf{r}_t\}$. For the MVP column, the one-step ahead predictions were used to generate minimum variance portfolios for $t = 101, \ldots, 920$; the column reports the empirical standard deviation of the subsequent portfolios. A lower empirical standard deviation is better. For the PLLH column, the one-step ahead predictions were used to calculate the predictive log-likelihood $\sum_{i=101}^{920} \log \phi(\mathbf{r}_t; 0, \hat{\mathbf{X}}_t^{-1})$ where ϕ is a multivariate Gaussian kernel. A higher predictive log-likelihood is better. The Uhlig-like models do better on both counts. For this data set, one does not gain much by incorporating multiple smoothing parameters.

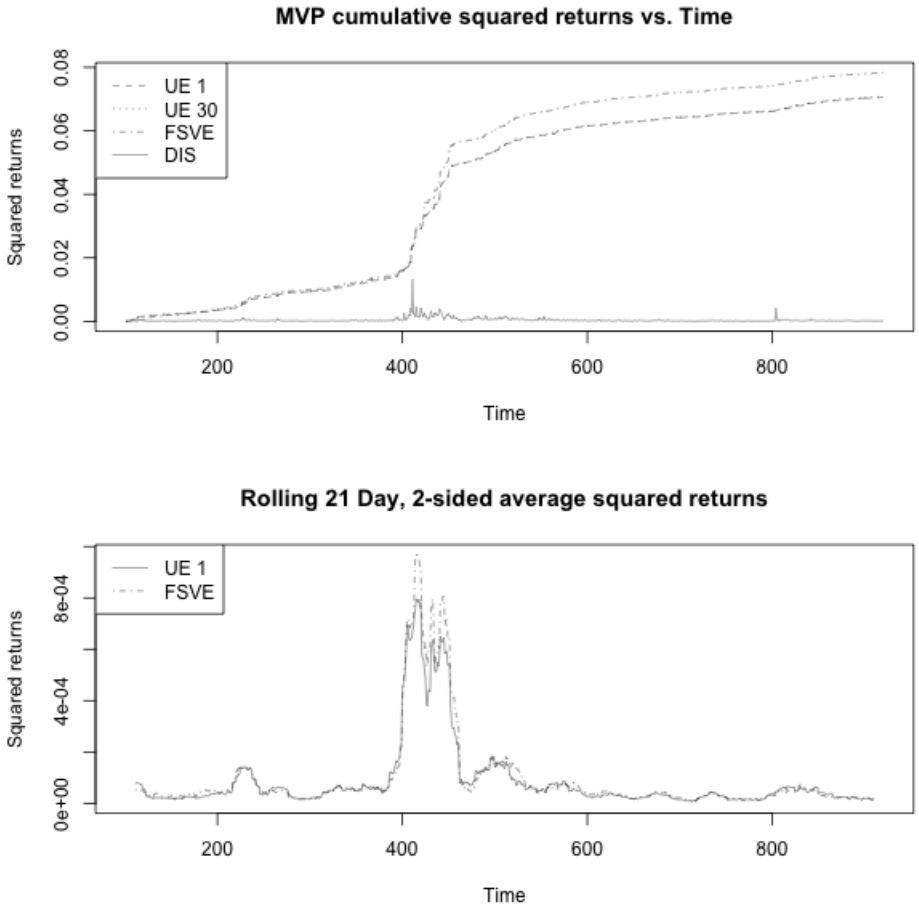


Figure 4: The cumulative squared returns and a rolling average of squared returns for several minimum variance portfolios. Minimum variance portfolios were generated using one-step ahead predictions for model (UE) (UE 1 in plot), for model (5) when $\mathbf{R}$ is a diagonal matrix (UE 30), and for the extension to factor stochastic volatility described in Appendix 6 (FSVE). The procedure for generating the one-step ahead portfolios is described in Figure 3. The cumulative squared returns of these portfolios are plotted in the top graphic; a lower cumulative squared return is better. The Uhlig-like models perform almost identically. The factor stochastic volatility-like model performs about the same as the Uhlig-like models initially, worse during the recent financial crisis, and then about the same after the crisis. The curve on the lower portion of the top graphic is the realized variance of Disney, which shows when the crisis occurred and the relative magnitude of market volatility. A 21 day, 2-sided rolling average of the squared returns for UE 1 and FSVE are plotted on the bottom graphic. Both graphics suggests that model (UE) does best when the latent covariances are changing rapidly.

covariance matrices has the advantage. Of course, these observations only apply when using factor stochastic volatility with daily returns. It may be the case that one can use intraday information along with some specialized knowledge about the structure of market fluctuations (like factor stochastic volatility) to generate better estimates and predictions.

Despite Model UE's short-term forecasting success, it does have some faults. In particular, the evolution of $\{\mathbf{X}_t\}_{t=1}^T$ is rather degenerate, since $\{\mathbf{X}_t\}_{t=1}^T$ is not stationary. This becomes apparent if you try to simulate data from the model as the hidden states will quickly become numerically singular. The lack of stationarity means that the model's k-step ahead predictions do not revert to some mean, which is what one would expect when modeling latent covariance matrices in finance, or most applications for that matter. If the true latent covariance matrices do indeed mean revert, then the model's predictions will becomes worse as the time horizon increases. However, as discussed in Section 2, other models of time varying covariance matrices can suffer from the same problem, and, further, are much less amenable to Bayesian analysis. Thus, though our model may not capture all of the features one would like to find in a covariance-matrix valued state-space model, at least it is tractable.

While it is difficult to remedy the lack of stationarity of the hidden states and maintain tractability of the posterior inference, one may feasibly relax the distributional assumptions for the observation equation. In particular, one may consider observations of the form

$$\mathbf{Y}_t \sim W_m(k, (k\mathbf{X}_t)^{-1}/\phi_t)$$

where the ϕ_t are independently and identically distributed and centered at unity; for instance, $\phi_t \sim \mathrm{Ga}(\nu/2, \nu/2)$, a gamma distribution. Introducing the auxiliary variables lets $\mathbf{Y}_t$ have fatter tails, which will reduce the impact of relatively large deviations in $\mathbf{Y}_t$ from $\mathbf{X}_t^{-1}$. To see this, consider the conditional density for $\mathbf{Y}_t$ above, and in particular the term in the exponent, which is proportional to $-\phi_t \operatorname{tr} \mathbf{X}_t\mathbf{Y}_t$, where tr denotes the trace functional. When $\operatorname{tr} \mathbf{X}_t\mathbf{Y}_t$ is relatively large, a smaller than average value of ϕ_t may be chosen to increase the likelihood of having observed $\mathbf{Y}_t$. (Remember that the distribution of ϕ_t is chosen so that its average value is unity.) Conditional upon $\{\phi_t\}_{t=1}^T$, the one step ahead prediction $(\mathbf{X}_t^{-1}|\mathcal{D}_{t-1})$ is still $k\mathbf{C}_t/(n-m-1)$ where $\mathbf{C}_t = \lambda\mathbf{\Sigma}_{t-1}$, but now $\{\mathbf{\Sigma}_t\}_{t=1}^T$ is recursively defined by

$$\mathbf{\Sigma}_t = \lambda\mathbf{\Sigma}_{t-1} + \phi_t\mathbf{Y}_t.$$

Thus, a smaller than average ϕ_t, that is $\phi_t < 1$, has the effect of down-weighting the contribution of $\mathbf{Y}_t$ to $\mathbf{\Sigma}_t$. Hence, relatively large deviations play a less important role in smoothing. However, this flourish comes at a computational cost. Now, instead of simply sampling the posterior distribution of the system's parameters $\boldsymbol{\theta}$ using Proposition 3 and a random walk Metropolis sampler, one must employ a Metropolis within Gibbs sampler and draw all of the unknown quantities, not just $\boldsymbol{\theta}$. To do this, one can sample $(\boldsymbol{\theta}|\{\phi_t\}_{t=1}^T)$ using Proposition 3 and a random walk Metropolis step followed by forward filtering and backward sampling $(\{\mathbf{X}_t\}_{t=1}^T|\boldsymbol{\theta}, \{\phi_t\}_{t=1}^T)$ to generate a joint draw of $(\boldsymbol{\theta}, \{\mathbf{X}_t\}_{t=1}^T|\{\phi_t\}_{t=1}^T)$. One can then sample $(\{\phi_t\}_{t=1}^T|\{\mathbf{X}_t\}_{t=1}^T, \boldsymbol{\theta})$, which factorizes

into the independent components $(\phi_t|\mathbf{X}_t, \boldsymbol{\theta})$, using conjugate updating. Iterating this procedure yields the Gibbs sampler. The practical impact of this more flexible form of smoothing remains future work.

Acknowledgments

The referees provided several useful comments and criticisms. In particular, they helped us improve the background section, prompted us to consider the case of multiple smoothing parameters, and suggested including a more flexible observation equation via data augmentation. Thank you to whomever you are. All subsequent errors and failures of exposition are our own.

References

Aguilar, O. and West, M. (2000). "Bayesian Dynamic Factor Models and Portfolio Allocation." *Journal of Business and Economic Statistics*, 18(3): 338–357. 761

Andersen, T. G., Bollerslev, T., Diebold, F. X., and Ebens, H. (2001). "The Distribution of Realized Stock Return Volatility." *Journal of Financial Econometrics*, 61: 43–76. 789

Asai, M. and McAleer, M. (2009). "The Structure of Dynamic Correlations in Multivariate Stochastic Volatility Models." *Journal of Econometrics*, 150: 182–192. 764, 765

Barndorff-Nielsen, O. E., Hansen, P. R., Lunde, A., and Shephard, N. (2009). "Realized Kernels in Practice: Trades and Quotes." *Econometrics Journal*, 12(3): C1–C32. 789, 791

— (2011). "Multivariate Realized Kernels: Consistent Positive Semi-Definite Estimators of the Covariation of Equity Prices with Noise and Non-Synchronous Trading." *Journal of Econometrics*, 162: 149–169. 789, 790

Barndorff-Nielsen, O. E. and Shephard, N. (2002). "Econometric Analysis of Realized Volatility and Its Use in Estimating Stochastic Volatility." *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 64(2): 253–280. 788

Bauer, G. H. and Vorkink, K. (2011). "Forecasting Multivariate Realized Stock Market Volatility." *Journal of Econometrics*, 160: 93–101. 763

Bauwens, L., Laurent, S., and Rombouts, J. V. K. (2006). "Multivariate GARCH Models: a Survey." *Journal of Applied Econometrics*, 21: 7109. 761

Bollerslev, T. (1986). "Generalized Autoregressive Conditional Heteroskedasticity." *Journal of Econometrics*, 31: 307–327. 773

Brown, R. G. (1959). *Statistical Forecasting for Inventory Control*. McGraw-Hill. 760, 769, 773

Carter, C. K. and Kohn, R. (1994). "On Gibbs Sampling for State Space Models." *Biometrika*, 81: 541–533. 762

Carvalho, C. M., Lopes, H. F., and Aguilar, O. (2011). "Dynamic Stock Selection Strategies: A Structured Factor Model Framework." In *Bayesian Statistics 9*. Oxford University Press. 787

Chib, S., Nardari, F., and Shephard, N. (2002). "Markov Chain Monte Carlo Methods for Stochastic Volatility Models." *Journal of Econometrics*, 108: 281–316. 787

Chiriac, R. and Voev, V. (2010). "Modelling and Forecasting Multivariate Realized Volatility." *Journal of Applied Econometrics*, 26: 922–947. 763

Chiu, T. Y. M., Leonard, T., and Tsui, K.-W. (1996). "The Matrix-Logarithmic Covariance Model." *Journal of the American Statistical Association*, 91: 198–210. 762

Choi, C. and Christensen, H. I. (2011). "Robust 3D Visual Tracking Using Particle Filters on the SE(3) Group." In *IEEE International Conference on Robotics and Automation*, 4384–4390. 760

Cochrane, J. H. (2005). *Asset Pricing*. Princeton University Press. 786

Díaz-García, J. A. and Jáimez, R. G. (1997). "Proof of the Conjectures of H. Uhlig on the Singular Multivariate Beta and the Jacobian of a Certain Matrix Transformation." *The Annals of Statistics*, 25: 2018–2023. 767, 782

Edelman, A. (2005). "The Mathematics and Applications of (Finite) Random Matrices." See handouts 1-4.
URL http://web.mit.edu/18.325/www/handouts.html 781

Fama, E. F. and French, K. R. (1993). "Common Risk Factors in the Returns on Stocks and Bonds." *Journal of Financial Economics*, 33: 3–56. 786

Fox, E. B. and West, M. (2011). "Autoregressive Models for Variance Matrices: Stationary Inverse Wishart Processes." Technical report, Duke University. 765

Früwirth-Schnatter, S. (1994). "Data Augmentation and Dynamic Linear Models." *Journal of Time Series Analysis*, 15: 183–202. 762

Gourieroux, C., Jasiak, J., and Sufana, R. (2009). "The Wishart Autoregressive Process of Multivariate Stochastic Volatility." *Journal of Econometrics*, 150: 167–181. 765

Harvey, A., Ruiz, E., and Shephard, N. (1994). "Multivariate Stochastic Volatility Models." *The Review of Economic Studies*, 61: 247–264. 786

Hauberg, S., Lauze, F., and Pedersen, K. S. (2013). "Unscented Kalman Filtering on Riemannian Manifolds." *Journal of Mathematical Imaging and Vision*, 46: 103–120. 760

Jacod, J. and Shiryaev, A. N. (2003). *Limit Theorems For Stochastic Processes*. Springer. 789

Jacquier, E., Polson, N. G., and Rossi, P. E. (2004). "Bayesian Analysis of Stochastic Volatility Models with Fat-Tails and Correlated Errors." *Journal of Econometrics*, 122: 185–212. 786

Jin, X. and Maheu, J. M. (2012). "Modelling Realized Covariances and Returns." URL http://homes.chass.utoronto.ca/~jmaheu/jin-maheu.pdf 764

Julier, S. J. and Uhlmann, J. K. (1997). "New Extensions of the Kalman Filter to Nonlinear Systems." In *Signal Processing, Sensor Fusion, and Target Recognition VI*, volume 3068. 760

Kalman, R. E. (1960). "A New Approach to Linear Filtering and Prediction Problems." *Journal of Basic Engineering*, 82 (Series D): 35–45. 760

Karatzas, I. and Shreve, S. E. (1991). *Brownian Motion and Stochastic Calculus*. Springer. 788

Koopman, S. J., Jungbacker, B., and Hol, E. (2005). "Forecasting Daily Variability of the S&P 100 Stock Index using Historical, Realised and Implied Volatility Measurements." *Journal of Empirical Finance*, 12: 445–475. 789

Lence, S. H. and Hayes, D. J. (1994a). "The Empirical Minimum-Variance Hedge." *American Journal of Agricultural Economics*, 76: 94–104. 773

— (1994b). "Parameter-Based Decision Making Under Estimation Risk: An Application to Futures Trading." *The Journal of Finance*, 49: 345–357. 773

Liu, Q. (2009). "On Portfolio Optimization: How and When Do We Benefit from High-Frequency Data?" *Journal of Applied Econometrics*, 24: 560–582. 789

Loddo, A., Ni, S., and Sun, D. (2011). "Selection of Multivariate Stochastic Volatility Models via Bayesian Stochastic Search." *Journal of Business and Economic Statistics*, 29: 342–355. 763

Mikusiński, P. and Taylor, M. D. (2002). *An Introduction to Multivariate Analysis*. Birkhäuser. 768, 781

Muirhead, R. J. (1982). *Aspects of Multivariate Statistical Theory*. Wiley. 767, 768, 780, 781, 782

Philipov, A. and Glickman, M. E. (2006). "Multivariate Stochastic Volatility via Wishart Processes." *Journal of Business and Economic Statistics*, 24: 313–328. 764, 765

Pitt, M. K. and Walker, S. G. (2005). "Constructing Stationary Time Series Models Using Auxiliary Variables with Applications." *Journal of the American Statistical Association*, 100(470): 554–564. 765

Prado, R. and West, M. (2010). *Time Series: Modeling, Computation, and Inference*, chapter Multivariate DLMs and Covariance Models, 263–319. Chapman & Hall/CRC. 764, 766, 767

Quintana, J. M. and West, M. (1987). "An Analysis of International Exchange Rates Using Multivariate DLMs." *The Statistician*, 36: 275–281. 766

Shephard, N. (1994). "Local Scale Models: State Space Alternative to Integrated GARCH Processes." *Journal of Econometrics*, 60: 181–202. 766

— (2005). *Stochastic Volatility: Selected Readings*. Oxford University Press. 762

Srivastava, A. and Klassen, E. (2004). "Bayesian and Geometric Subspace Tracking." *Advances in Applied Probability*, 36(1): 43–56. 760

Taylor, S. J. (1982). *Financial Returns Modelled by the Product of Two Stochastic Processes–a Study of Daily Sugar Prices 1961-1979*, 203–226. Amersterdam: North-Holland. 762, 773

Tompkins, F. and Wolfe, P. J. (2007). "Bayesian Filtering on the Stiefel Manifold." In *Computational Advances in Multi-Sensor Adaptive Processing*, 261 – 264. 760

Triantafyllopoulos, K. (2008). "Multivariate stochastic volatility with Bayesian dynamic linear models." *Journal of Statistical Planning and Inference*, 138: 1021–1037. 764

Tyagi, A. and Davis, J. W. (2008). "A Recursive Filter For Linear Systems on Riemannian Manifolds." In *IEEE Conference on Computer Vision and Pattern Recognition*. 760

Uhlig, H. (1994). "On Singular Wishart and Singular Multivariate Beta Distributions." *The Annals of Statistics*, 22(1): 395–495. 766, 767, 780, 781, 782

— (1997). "Bayesian Vector Autoregressions with Stochastic Volatility." *Econometrica*, 65(1): 59–73. 760, 764, 765, 766, 769

West, M. and Harrison, J. (1997). *Bayesian Forecasting and Dynamic Models*. Springer Verlag. 766

Windle, J. (2013). "Forecasting High-Dimensional Variance-Covariance Matrices with High-Frequency Data and Sampling Pólya-Gamma Random Variates for Posterior Distributions Derived from Logistic Likelihoods." Ph.D. thesis, The University of Texas at Austin. 787

Windle, J., Carvalho, C. M., Scott, J. G., and Sun, L. (2013). "Efficient Data Augmentation in Dynamic Models for Binary and Count Data."
URL http://arxiv.org/abs/1308.0774 763

Appendix A: Technical Details

Much of the calculus one needs can be found Uhlig (1994) or Muirhead (1982). We synthesize those results here. We are not aware of results in either regarding backward sampling or marginalization.

First, some notation: Assume $k, m \in \mathbb{N}$, $k \leq m$. Let $\mathcal{S}^+_{m,k}$ denote the set of positive semi-definite symmetric matrices of rank k and order m. When $k = m$, we drop k from the notation so that $\mathcal{S}^+_m$ denotes the set of positive-definite symmetric matrices of order m. For symmetric matrices $\mathbf{A}$ and $\mathbf{B}$, let $\mathbf{A} < \mathbf{B}$ denote $\mathbf{B} - \mathbf{A} \in \mathcal{S}^+_m$. For $\mathbf{A} \in \mathcal{S}^+_m$ let

$$\{\mathcal{S}^+_{m,k} < \mathbf{A}\} = \{\mathbf{C} \in \mathcal{S}^+_{m,k} : \mathbf{C} < \mathbf{A}\}.$$

If $k > m - 1$ is real and we write $\mathcal{S}^+_{m,k}$ then we implicitly mean $\mathcal{S}^+_m$. We will use $|\cdot|$ to denote the determinant of a matrix and $\mathbf{I}$ to denote the identity. We at times follow Muirhead (1982) and define densities with respect to differential forms (also known as K-forms or differential K-forms). Mikusiński and Taylor (2002) is a good introduction to calculus on manifolds. The handouts of Edelman (2005) provide a more succinct introduction.

Definition 4 (Wishart distribution)**.** *A positive semi-definite symmetric matrix-valued random variable* $\mathbf{Y}$ *has Wishart distribution* $W_m(k, \mathbf{V})$ *for* $k \in \mathbb{N}$ *and* $\mathbf{V} \in \mathcal{S}^+_m$ *if*

$$\mathbf{Y} \sim \sum_{i=1}^{k} \mathbf{r}_i \mathbf{r}_i', \quad \mathbf{r}_i \overset{iid}{\sim} N(0, \mathbf{V}), \quad i = 1, \ldots, k.$$

When $k > m - 1$*, the density for the Wishart distribution is*

$$\frac{|\mathbf{Y}|^{(k-m-1)/2}}{2^{mk/2}\Gamma_m\left(\frac{k}{2}\right)|\mathbf{V}|^{k/2}} \exp\left(\mathsf{tr} - \frac{1}{2}\mathbf{V}^{-1}\mathbf{Y}\right)$$

(Muirhead 1982) with respect to the volume element

$$(d\mathbf{Y}) = \bigwedge_{1 \leq i \leq j \leq m} d\mathbf{Y}_{ij}.$$

When $k \leq m - 1$ *and* $\mathbf{Y}$ *is rank deficient, the density is*

$$\frac{\pi^{-(mk-k^2)/2}|\mathbf{L}|^{(k-m-1)/2}}{2^{mk/2}\Gamma_k\left(\frac{k}{2}\right)|\mathbf{V}|^{k/2}} \exp\left(\mathsf{tr} - \frac{1}{2}\mathbf{V}^{-1}\mathbf{Y}\right)$$

with respect to the volume element

$$(d\mathbf{Y}) = 2^{-k} \prod_{i=1}^{k} l_i^{m-k} \prod_{i<j}^{k} (l_i - l_j)(\mathbf{H}_1' d\mathbf{H}_1) \wedge \bigwedge_{i=1}^{k} dl_i$$

where $\mathbf{Y} = \mathbf{H}_1 \mathbf{L} \mathbf{H}_1'$*,* $\mathbf{H}_1$ *is a matrix of orthonormal columns of order* $m \times k$*, and* $\mathbf{L} = diag(l_1, \ldots, l_k)$ *with decreasing positive entries (Uhlig 1994, Thm. 6). The notation* $(\mathbf{H}_1' d\mathbf{H}_1)$ *is shorthand for a differential* K*-form from the Steifel manifold* $V_{m,k}$ *embedded in* $\mathbb{R}^{m \times k}$ *where* $K = mk - k(k+1)/2$ *(Muirhead 1982, p. 63). One can extend the definition of the Wishart distribution to real values of* $k > m - 1$ *for* $\mathcal{S}^+_m$*-valued random variables by defining* $\mathbf{Y} \sim W_m(k, \mathbf{V})$ *to have the full rank density defined above.*

Definition 5 (the bijection τ)**.** *Assume $m \in \mathbb{N}$ and $k \in \{1, \ldots, m\}$. A single bijection provides the key to both the evolution of $\mathbf{X}_t$ in model (UE) and to the definition of the beta distribution. In particular, let $\tau : \mathcal{S}^+_{m,k} \times \mathcal{S}^+_m \to \mathcal{S}^+_m \times \{\mathcal{S}^+_{m,k} < \mathbf{I}\}$ take $(\mathbf{A}, \mathbf{B})$ to $(\mathbf{S}, \mathbf{U})$ by letting $\mathbf{T}'\mathbf{T} = \mathbf{A} + \mathbf{B}$ be the Cholesky factorization of $\mathbf{A} + \mathbf{B}$ and letting*

$$\begin{cases} \mathbf{S} = \mathbf{A} + \mathbf{B}, \\ \mathbf{U} = \mathbf{T}^{-1\prime}\mathbf{A}\mathbf{T}^{-1}. \end{cases}$$

Conversely, let $g : \mathcal{S}^+_m \times \{\mathcal{S}^+_{m,k} < \mathbf{I}\} \to \mathcal{S}^+_{m,k} \times \mathcal{S}^+_m$ take $(\mathbf{S}, \mathbf{U})$ to $(\mathbf{A}, \mathbf{B})$ by letting $\mathbf{T}'\mathbf{T} = \mathbf{S}$ be the Cholesky decomposition of $\mathbf{S}$ and

$$\begin{cases} \mathbf{A} = \mathbf{T}'\mathbf{U}\mathbf{T}, \\ \mathbf{B} = \mathbf{T}'(\mathbf{I} - \mathbf{U})\mathbf{T}. \end{cases}$$

One can see that g is the inverse of τ since $\tau(g(\mathbf{S}, \mathbf{U})) = (\mathbf{S}, \mathbf{U})$ and $g(\tau(\mathbf{A}, \mathbf{B})) = (\mathbf{A}, \mathbf{B})$.

Definition 6 (beta distribution)**.** *Let $\mathbf{A} \sim W_m(k, \mathbf{\Sigma}^{-1})$ and $\mathbf{B} \sim W_m(n, \mathbf{\Sigma}^{-1})$ be independent where $n > m - 1$ and either $k < m$ is an integer or $k > m - 1$ is real-valued. Let $(\mathbf{S}, \mathbf{U}) = \tau(\mathbf{A}, \mathbf{B})$. The beta distribution, $\beta_m(k/2, n/2)$, is the distribution of $\mathbf{U}$. When $k < m$ is an integer, the beta distribution $\beta_m(n/2, k/2)$ is the distribution of $\mathbf{I} - \mathbf{U}$ where $\mathbf{U} \sim \beta_m(k/2, n/2)$. (See Definition 1 from Uhlig (1994) and p. 109 in Muirhead (1982).)*

The following theorem synthesizes results from Uhlig (1994), Muirhead (1982), and Díaz-García and Jáimez (1997).

Theorem 7. *Based on Muirhead (1982, Thm. 3.3.1), Uhlig (1994, Thm. 7), and Díaz-García and Jáimez (1997, Thm. 2) . Let $n > m - 1$ and let either $k < m$ be an integer or $k > m - 1$ be real-valued. The bijection $\tau : \mathcal{S}^+_{m,k} \times \mathcal{S}^+_m \to \mathcal{S}^+_m \times \{\mathcal{S}^+_{m,k} < \mathbf{I}\}$ from Definition 5 changes*

$$\mathbf{A} \sim W_m(k, \mathbf{\Sigma}^{-1}) \perp \mathbf{B} \sim W_m(n, \mathbf{\Sigma}^{-1}) \tag{6}$$

to

$$\mathbf{S} \sim W_m(n + k, \mathbf{\Sigma}^{-1}) \perp \mathbf{U} \sim \beta_m(k/2, n/2). \tag{7}$$

Proof. Thm. 3.3.1 in Muirhead (1982) proves this in the full rank case. Thm. 7 in Uhlig (1994) proves this in the rank 1 case. Thm. 2 in Díaz-García and Jáimez (1997) proves it in the general rank deficient case. □

Theorem 7 justifies forward filtering in models (UE) and (5) as follows.

Proof of Proposition 1, Forward Filtering. Suppose we start at time $t - 1$ with data $\mathcal{D}_{t-1}$, so that the joint distributions of $\mathbf{X}_{t-1}$ and $\mathbf{\Psi}_t$ is characterized by

$$\mathbf{X}_{t-1} \sim W_m(n + k, (k\mathbf{\Sigma}_{t-1})^{-1}) \perp (\mathbf{I} - \mathbf{\Psi}_t) \sim \beta_m(k/2, n/2),$$

which looks like (7). Theorem 7 shows that the bijection τ^{-1} takes $(\mathbf{X}_{t-1}, \mathbf{I} - \mathbf{\Psi}_t)$ to

$$\mathbf{Z}_t \sim W_m(k, (k\mathbf{\Sigma}_{t-1})^{-1}) \perp \mathbf{R}'\mathbf{X}_t\mathbf{R} \sim W_m(n, (k\mathbf{\Sigma}_{t-1})^{-1}),$$

which is (6) after applying the transformation summarized by

$$\begin{cases} \mathbf{Z}_t = \mathbf{T}'_{t-1}(\mathbf{I} - \mathbf{\Psi}_t)\mathbf{T}_{t-1}, \\ \mathbf{R}'\mathbf{X}_t\mathbf{R} = \mathbf{T}'_{t-1}\mathbf{\Psi}_t\mathbf{T}_{t-1}, \\ \mathbf{X}_{t-1} = \mathbf{Z}_t + \mathbf{R}'\mathbf{X}_t\mathbf{R}, \\ \mathbf{T}_{t-1} = \text{upper chol } \mathbf{X}_{t-1}. \end{cases} \tag{8}$$

The transformation includes the evolution equation in (UE) since

$$\mathbf{X}_t = \mathbf{R}^{-1'}\mathbf{T}'_{t-1}\mathbf{\Psi}_t\mathbf{T}_{t-1}\mathbf{R}^{-1}.$$

It also yields $(\mathbf{X}_t|\mathcal{D}_{t-1}) \sim W_m(n, (k\mathbf{R}\mathbf{\Sigma}_{t-1}\mathbf{R}')^{-1})$. Conjugate updating then yields $(\mathbf{X}_t|\mathcal{D}_t) \sim W_m(n+k, (k\mathbf{\Sigma}_t)^{-1})$ where $\mathbf{\Sigma}_t = \mathbf{R}\mathbf{\Sigma}_{t-1}\mathbf{R}' + \mathbf{Y}_t$. □

The reader may notice that the choice of distribution for $\mathbf{\Psi}_t$ is precisely the one that facilitates forward filtering. In particular, assuming that $(\mathbf{X}_{t-1}|\mathcal{D}_{t-1})$ has an acceptable distribution to start, then $(\mathbf{X}_t|\mathcal{D}_{t-1})$ will have an acceptable distribution to update, so that $(\mathbf{X}_t|\mathcal{D}_t)$ will have a distribution that lets us repeat the process. However, we cannot easily write down the distribution of $(\mathbf{X}_{t+k}|\mathcal{D}_t)$ for anything but $k = 0$ or 1. To see why, assume that we start at time $t-1$ with data $\mathcal{D}_{t-1}$ and evolve to $(\mathbf{X}_t|\mathcal{D}_{t-1}) \sim W_m(n, (k\mathbf{R}\mathbf{\Sigma}_{t-1}\mathbf{R}')^{-1})$, just like above. Now consider moving from $\mathbf{X}_t$ to $\mathbf{X}_{t+1}$ without updating:

$$\begin{cases} \mathbf{T}_t = \text{upper chol } \mathbf{X}_t \\ \mathbf{R}'\mathbf{X}_{t+1}\mathbf{R} = \mathbf{T}'_t\mathbf{\Psi}_t\mathbf{T}_t, \quad \mathbf{\Psi}_t \sim \beta_m(n/2, k/2). \end{cases}$$

The distribution of $\mathbf{I}-\mathbf{\Psi}_t$ is $\beta_m(k/2, n/2)$ but the distribution of $(\mathbf{X}_t|\mathcal{D}_{t-1})$ is $W_m(n, \ldots)$. We cannot apply Theorem 7 at this point because there is a mismatch between the parameters of $\mathbf{I}-\mathbf{\Psi}_t$ and the degrees of freedom of $(\mathbf{X}_t|\mathcal{D}_{t-1})$—we need $n+k$ not n degrees of freedom! Thus, the distribution of $(\mathbf{X}_{t+1}|\mathcal{D}_{t-1})$ is unknown. Despite not knowing its distribution, one can show that the evolution of $\{\mathbf{X}_t\}_{t=1}^T$ is rather degenerate as seen in Section 2.

Proposition 8. *Assume* $\mathbf{S}$ *and* $\mathbf{U}$ *are as in Theorem 7 and let* $(\mathbf{A}, \mathbf{B}) = \tau^{-1}(\mathbf{S}, \mathbf{U})$. *Then the conditional distribution of* $(\mathbf{S}|\mathbf{B})$ *is*

$$(\mathbf{S}|\mathbf{B}) = \mathbf{B} + \mathbf{Z}, \ \mathbf{Z} \sim W_m(k, \mathbf{\Sigma}^{-1}). \tag{9}$$

Proof. Let $\mathbf{S}$ and $\mathbf{U}$ be as in Theorem 7 and let $(\mathbf{A}, \mathbf{B}) = \tau^{-1}(\mathbf{S}, \mathbf{U})$. Let p be the rank of $\mathbf{A}$. Fix $\mathbf{B}$ and define a change of variables g by $\mathbf{A} = \mathbf{S} - \mathbf{B}$. Jointly, $(\mathbf{A}, \mathbf{B})$ has

a density with respect to the differential form $(d\mathbf{A}) \wedge (d\mathbf{B})$ where $\mathbf{A}$ is a K-form with $K = np - p(p-1)/2$:

$$(d\mathbf{A}) = \sum_{i_1<\ldots<i_K} f_{i_1<\ldots<i_K}(\mathbf{A})\; d\mathbf{A}_{i_1} \wedge \cdots \wedge d\mathbf{A}_{i_K}$$

where the index of $d\mathbf{A}$ corresponds to the vectorized (by column) upper triangular portion of $\mathbf{A}$. Under g, the pull back of $d\mathbf{A}_i$ is

$$g^*(d\mathbf{A}_i) = d\mathbf{S}_i;$$

thus,

$$(d\mathbf{S}) = \sum_{i_1<\ldots<i_K} f_{i_1<\ldots<i_K}(\mathbf{S}-\mathbf{B})\; d\mathbf{S}_{i_1} \wedge \cdots \wedge d\mathbf{S}_{i_K},$$

where, again, the index corresponds to the vectorized upper triangular portion. Let $f_A(\mathbf{A})f_B(\mathbf{B})$ be the density of $(\mathbf{A}, \mathbf{B})$ with respect to the differential form $(d\mathbf{A}) \wedge (d\mathbf{B})$. Under g, the differential form corresponding to the density of $(\mathbf{A}, \mathbf{B})$,

$$f_A(\mathbf{A})f_B(\mathbf{B})\; (d\mathbf{A}) \wedge (d\mathbf{B}),$$

becomes

$$f_A(\mathbf{S}-\mathbf{B})f_B(\mathbf{B})\; (d\mathbf{S}) \wedge (d\mathbf{B})$$

on the manifold

$$\{(\mathbf{S}, \mathbf{B}) : \mathbf{S} \in \mathcal{S}_m^+,\ \mathbf{B} \in \mathcal{S}_m^+,\ \mathbf{S}-\mathbf{B} \in \mathcal{S}_{m,k}^+\}.$$

We know that $f_B(\mathbf{B})(d\mathbf{B})$ is the differential form corresponding to the distribution of $\mathbf{B}$, hence $f_A(\mathbf{S}-\mathbf{B})(d\mathbf{S})$ describes the conditional distribution of $(\mathbf{S}|\mathbf{B})$. Doing another change of variables shows that $(\mathbf{S}|\mathbf{B})$ is a shifted Wishart distribution, that is

$$(\mathbf{S}|\mathbf{B}) = \mathbf{B} + \mathbf{Z},\ \mathbf{Z} \sim W_m(k, \mathbf{\Sigma}^{-1}).$$

□

Proof of Proposition 2, Backward Sampling. The Markovian structure of the model ensures that we can decompose the joint density of the latent states given $\mathcal{D}_T$ (and n, k, $\mathbf{R}$) as

$$p(\mathbf{X}_T|\mathcal{D}_T) \prod_{i=1}^{T-1} p(\mathbf{X}_t|\mathbf{X}_{t+1}, \mathcal{D}_t).$$

(The density is taken with respect to the product measure on the T-fold product of $\mathcal{S}_m^+$ embedded in $\mathbb{R}^{m(m+1)/2}$ with Lebesgue measure). Applying Proposition 8 with $(\mathbf{X}_{t-1}|\mathcal{D}_{t-1})$ as $\mathbf{S}$, $\mathbf{I} - \mathbf{\Psi}_t$ as $\mathbf{U}$, and $(\mathbf{R}'\mathbf{X}_t\mathbf{R}|\mathcal{D}_{t-1})$ as $\mathbf{B}$, we find that the distribution of $(\mathbf{X}_{t-1} \,|\, \mathbf{X}_t, \mathcal{D}_{t-1})$ is

$$(\mathbf{X}_{t-1}|\mathbf{X}_t, \mathcal{D}_{t-1}) = \mathbf{R}'\mathbf{X}_t\mathbf{R} + \mathbf{Z}_t,\ \mathbf{Z}_t \sim W_m(k, (k\mathbf{\Sigma}_{t-1})^{-1}).$$

□

Proof of Proposition 3, Marginalization. First, by conditioning we can express the density $p(\{\mathbf{Y}_t\}_{t=1}^T|\mathcal{D}_0)$ as

$$\prod_{t=1}^{T} p(\mathbf{Y}_t|\mathcal{D}_{t-1})$$

with respect to the differential form $\bigwedge_{t=1}^T(d\mathbf{Y}_t)$ where $(d\mathbf{Y}_t)$ is as in Definition 4.

Thus, we just need to derive the distribution of $(\mathbf{Y}_t|\mathcal{D}_{t-1})$. Assume that $n > m-1$ and that either $k < m$ is a positive integer or $k > m-1$ is real-valued. Suppose that $(\mathbf{Y}_t|\mathbf{X}_t) \sim W_m(k, (k\mathbf{X}_t)^{-1})$ and $(\mathbf{X}_t|\mathcal{D}_{t-1}) \sim W_m(n, (k\mathbf{C}_t)^{-1})$ where $\mathbf{C}_t = \mathbf{R}\boldsymbol{\Sigma}_{t-1}\mathbf{R}'$. Then the density for $(\mathbf{Y}_t|\mathcal{D}_{t-1})$ is

$$\pi^{-(mk-k^2)/2}\frac{\Gamma_m(\frac{\nu}{2})}{\Gamma_m(\frac{n}{2})\Gamma_k(\frac{k}{2})}\frac{|\mathbf{L}_t|^{(k-m-1)/2}|\mathbf{C}_t|^{n/2}}{|\mathbf{C}_t+\mathbf{Y}_t|^{\nu/2}}$$

in the rank-deficient case and is

$$\frac{\Gamma_m(\frac{\nu}{2})}{\Gamma_m(\frac{n}{2})\Gamma_m(\frac{k}{2})}\frac{|\mathbf{Y}_t|^{(k-m-1)/2}|\mathbf{C}_t|^{n/2}}{|\mathbf{C}_t+\mathbf{Y}_t|^{\nu/2}}$$

in the full-rank case, with respect to the differential form $(d\mathbf{Y}_t)$ as found in Definition 4 for either the rank-deficient or full-rank cases respectively.

We will only prove the rank-deficient case, since the full-rank case is essentially identical. Consider the joint density $p(\mathbf{Y}_t|\mathbf{X}_t)p(\mathbf{X}_t|\mathcal{D}_{t-1})$:

$$\begin{aligned}&\frac{\pi^{-(mk-k^2)/2}|k\mathbf{X}_t|^{k/2}}{2^{mk/2}\Gamma_k\left(\frac{k}{2}\right)}|\mathbf{L}_t|^{(k-m-1)/2}\exp\left(\frac{-1}{2}\mathsf{tr}k\mathbf{X}_t\mathbf{Y}_t\right)\\&\cdot\frac{|k\mathbf{C}_t|^{n/2}}{2^{nm/2}\Gamma_m(\frac{n}{2})}|\mathbf{X}_t|^{(n-m-1)/2}\exp\left(\frac{-1}{2}\mathsf{tr}k\mathbf{C}_t\mathbf{X}_t\right)\end{aligned}$$

(where $\mathbf{Y}_t = \mathbf{H}_t\mathbf{L}_t\mathbf{H}_t$, $\mathbf{L}_t$ is a $k \times k$ diagonal matrix with decreasing entries, and $\mathbf{H}_t$ is in the Steifel manifold $V_{m,k}$) with respect to $(d\mathbf{Y}_t) \wedge (d\mathbf{X}_t)$, which is

$$\pi^{-(mk-k^2)/2}\frac{|\mathbf{L}_t|^{(k-m-1)/2}}{2^{km/2}\Gamma_k(\frac{k}{2})}\frac{|k\mathbf{C}_t|^{n/2}}{2^{nm/2}\Gamma_m(\frac{n}{2})}\,k^{km/2}|\mathbf{X}_t|^{(\nu-m-1)/2}\exp\left(\frac{-1}{2}\mathsf{tr}k\big(\mathbf{C}_t+\mathbf{Y}_t\big)\mathbf{X}_t\Big),\right.$$

$\nu = n+k$. The latter terms are the kernel for a Wishart distribution in $\mathbf{X}_t$. Integrating the kernel with respect to $\mathbf{X}_t$ yields

$$\frac{2^{\nu m/2}\Gamma_m(\frac{\nu}{2})}{|k(\mathbf{C}_t+\mathbf{Y}_t)|^{\nu/2}}.$$

Hence the density of $(\mathbf{Y}_t|\mathcal{D}_{t-1})$ is

$$\pi^{-(mk-k^2)/2}\frac{\Gamma_m(\frac{\nu}{2})k^{\nu m/2}}{\Gamma_m(\frac{n}{2})\Gamma_k(\frac{k}{2})}\frac{|\mathbf{L}_t|^{(k-m-1)/2}|\mathbf{C}_t|^{n/2}}{|k(\mathbf{C}_t+\mathbf{Y}_t)|^{\nu/2}}$$

with respect to $(d\mathbf{Y}_t)$. Factoring the k in the denominator gives us

$$\pi^{-(mk-k^2)/2}\frac{\Gamma_m(\frac{\nu}{2})}{\Gamma_m(\frac{n}{2})\Gamma_k(\frac{k}{2})}\frac{|\mathbf{L}_t|^{(k-m-1)/2}|\mathbf{C}_t|^{n/2}}{|\mathbf{C}_t+\mathbf{Y}_t|^{\nu/2}}.$$

□

Appendix B: Factor-like Models

Factor-like models capture variation in asset returns using multiple linear regression. In particular, the conditional returns are modeled linearly on some covariates; integrating over the covariates yields the marginal covariation of the returns. To see how this works, suppose the vector of asset returns $\mathbf{r}_t$ depends linearly on the covariate $\mathbf{x}_t$ so that

$$\mathbf{r}_t = \boldsymbol{\beta}\mathbf{x}_t + \boldsymbol{\varepsilon}_t, \;\; \boldsymbol{\varepsilon}_t \sim N(\boldsymbol{\mu}, \mathbf{D}_t).$$

Given the fixed regression coefficient $\boldsymbol{\beta}$ and assuming that the covariates are independent of the error terms, the marginal variance of $\mathbf{r}_t$ is

$$\text{var}(\mathbf{r}_t) = \boldsymbol{\beta}\text{var}(\mathbf{x}_t)\boldsymbol{\beta}' + \mathbf{D}_t.$$

Thus, conditionally, the individual elements of the vector $\mathbf{r}_t$ are independent, but marginally they are correlated. From this perspective, there are two ways to proceed. First, one may pick the covariates $\mathbf{x}_t$ so that they are known and so that they capture as much of the predictable marginal variation in $\mathbf{r}_t$ as possible. This is essentially the route followed by the well-known work of Fama and French (1993); though their objective is to find common factors that contribute to an asset's returns, which is slightly different than modeling covariance matrices. Second, instead of cleverly choosing some known covariates, one may use the data to infer a set of latent covariates $\{\mathbf{x}_t\}_{t=1}^T$. This is the path taken by Harvey et al. (1994) and the one traditionally followed by Bayesian statisticians. In that case, one places a prior on the covariates $\{\mathbf{x}_t\}_{t=1}^T$ and the errors' variances $\{\mathbf{D}_t\}_{t=1}^T$. Usually, this is done so that $\text{var}(\mathbf{x}_t)$ and $\mathbf{D}_t$ are diagonal and change slowly. The latter model is called factor stochastic volatility; however, stochastic volatility only enters the model through $\text{var}(\mathbf{x}_t)$ and through $\mathbf{D}_t$, and usually as multiple univariate processes at that. Since the important idea is really that there are a few factors that determine the correlation between elements of $\mathbf{r}_t$ we classify this model as factor-like.

There are many flourishes on these two basic approaches. In finance, for the former, one is interested in finding the covariates that reflect the non-diversifiable sources of risk and return. Chapter 20 of Cochrane (2005) provides a good discussion of the major work in this direction. For the latter, there have been a variety of suggestions to capture more features of asset returns or to improve predictive performance. For instance, one may incorporate a leverage effect, heavy tails, or jumps; one may impose sparsity on the regression coefficients; or one may let the regression coefficients change in time. (Jacquier et al. (2004) incorporate heavy tails and a leverage effect into univariate

stochastic volatility and Chib et al. (2002) examine heavy tails and jumps in univariate stochastic volatility. Carvalho et al. (2011) show how one may use dynamic regression coefficients and sparsity in factor stochastic volatility models.)

B.1 Extensions to Factor-like Models

In our original out-of-sample experiments, we benchmarked exponentially smoothed realized covariance matrices against factor stochastic volatility. However, factor stochastic volatility is at an inherent disadvantage because it does not use any data from the realized covariance statistic, which makes use of intraday data. In an attempt to level the playing field, one can incorporate some exogenous information from the realized covariance matrices, as seen in Section 3.1 of Windle (2013). For instance, one might have gleaned some exogenous information from the realized covariance matrices that concerns the factor loadings found in factor stochastic volatility. In that case, one may extend the factor stochastic volatility model to have dynamic loadings that track this exogenous information via

$$\begin{cases} \mathbf{r}_t = \boldsymbol{\beta}_t \mathbf{x}_t + \boldsymbol{\varepsilon}_t, \;\; \boldsymbol{\varepsilon}_t \sim N(0, \mathbf{D}_t) \\ \boldsymbol{\alpha}_t = \text{vecl}(\boldsymbol{\beta}_t) \\ \mathbf{z}_t = \boldsymbol{\alpha}_t + \boldsymbol{\omega}_t, \;\; \boldsymbol{\omega}_t \sim N(0, \boldsymbol{\Delta}) \\ \boldsymbol{\alpha}_t = \boldsymbol{\alpha}_{t-1} + \boldsymbol{\eta}_t, \;\; \boldsymbol{\eta}_t \;\; \sim N(0, \mathbf{W}). \end{cases}$$

where $\boldsymbol{\beta}_t$ is an $n \times p$ matrix of dynamic factor loadings; "vecl" vectorizes the lower diagonal portion of a matrix; $\mathbf{x}_t$ are the latent common factors and $\boldsymbol{\varepsilon}_t$ are the idiosyncratic factors such that the individual components of each are independent univariate stochastic volatility processes; $\boldsymbol{\Delta}$ and $\mathbf{W}$ are diagonal; and $\mathbf{z}_t$ is the information gleaned from the realized covariance matrices. One must be careful about identifying the factor loadings matrices, and we assume that they are lower triangular with ones along the diagonal. Thus, the obvious choice of exogenous information is the first p columns of $\mathbf{L}$ from the $\mathbf{LDL}'$ decomposition of the realized covariance matrices, that is $\mathbf{L}_t \boldsymbol{\Lambda}_t \mathbf{L}_t' = \mathbf{RC}_t$, where $\mathbf{L}_t$ is lower triangular and $\boldsymbol{\Lambda}_t$ is diagonal, and $\mathbf{z}_t$ is the vectorization of the lower triangular portion of the first p columns of $\mathbf{L}_t$. While this approach is ad hoc and does not jointly model returns and realized covariance matrices, it generates reasonable predictions.

Appendix C: Realized Covariance Matrices

Realized covariance matrices are symmetric positive-definite estimates of the daily quadratic variation of a multidimensional continuous-time stochastic process. Within the context of financial time series, there is both theoretical and empirical evidence to suggest that a realized covariance matrix can be interpreted as an estimate of the conditional covariance matrix of the open to close log returns.

Imagine that the market in which the assets are traded is open 24 hours a day and that we are interested in estimating the covariance matrix of daily log returns. Following

Barndorff-Nielsen and Shephard (2002), let $\mathbf{p}_s$ be the m-vector of log prices where s is measured in days and suppose that it is a Gaussian process of the form

$$\mathbf{p}_s = \int_0^s \mathbf{V}_u^{1/2} d\mathbf{w}_u$$

where $\{\mathbf{w}_s\}_{s\geq 0}$ is an m-dimensional Brownian motion and $\{\mathbf{V}_s^{1/2}\}_s$ is a continuous, deterministic, symmetric positive definite $m\times m$ process such that the square of $\{\mathbf{V}_s^{1/2}\}_s$ is integrable. Then the day t vector of log returns $\mathbf{r}_t = (\mathbf{p}_t - \mathbf{p}_{t-1})$ is distributed as

$$\mathbf{r}_t \sim N\Big(0, \int_{t-1}^t \mathbf{V}_u du\Big)$$

where $\mathbf{V}_u = \mathbf{V}_u^{1/2}\mathbf{V}_u^{1/2'}$. The quadratic covariation matrix (quadratic variation henceforth) measures the cumulative local (co)-fluctuations of the sample paths:

$$\langle \mathbf{p}\rangle_s = \operatorname*{plim}_{|\Delta_N|\to 0} \sum_{i=1}^{K_N} (\mathbf{p}_{u_i} - \mathbf{p}_{u_{i-1}})(\mathbf{p}_{u_i} - \mathbf{p}_{u_{i-1}})'$$

where the limit holds for any sequence of partitions of the form $\Delta_N = \{u_0 = 0 < \ldots < u_{K_N} = s\}$ and $|\Delta_N| = \max\{u_i - u_{i-1} : i \in 1, \ldots, K_N\}$. It is always the case, even when $\{\mathbf{V}_s^{1/2}\}_s$ is a stochastic process correlated with $\{\mathbf{w}_s\}_s$, that

$$\int_{t-1}^t \mathbf{V}_u du = \langle \mathbf{p}\rangle_t - \langle \mathbf{p}\rangle_{t-1}.$$

(See Proposition 2.10 in Karatzas and Shreve (1991).) Thus, in the Gaussian process case, the variance of $\mathbf{r}_t$ is related to the quadratic variation by

$$\operatorname{var}(\mathbf{r}_t) = \langle \mathbf{p}\rangle_t - \langle \mathbf{p}\rangle_{t-1}.$$

If the assets under consideration are traded frequently, then the day-t partition of trading times $\Delta_t^* = \{u_0 = t-1 < \ldots < u_{K_t} = t\}$ has $|\Delta_t^*|$ near zero so that

$$\mathbf{RC}_t = \sum_{i=1}^{K_t} (\mathbf{p}_{u_i} - \mathbf{p}_{u_{i-1}})(\mathbf{p}_{u_i} - \mathbf{p}_{u_{i-1}})',$$

where the summation is over Δ_t^*, is a good estimate of $\langle \mathbf{p}\rangle_t - \langle \mathbf{p}\rangle_{t-1}$. This is the realized covariance.

The same logic proceeds when $\{\mathbf{V}_s^{1/2}\}_s$ is a stochastic process that is independent of the Brownian motion. In that case, the only major change is

$$\Big(\mathbf{r}_t \Big| \int_{t-1}^t \mathbf{V}_u du\Big) = N\Big(0, \int_{t-1}^t \mathbf{V}_u du\Big),$$

that is the log returns are a mixture of normals, so that

$$\text{var}\Big(\mathbf{r}_t \Big| \int_{t-1}^{t} \mathbf{V}_u du\Big) = \langle \mathbf{p} \rangle_t - \langle \mathbf{p} \rangle_{t-1}.$$

Since $\mathbf{RC}_t$ is a good estimate of $\langle \mathbf{p} \rangle_t - \langle \mathbf{p} \rangle_{t-1}$ regardless of $\{\mathbf{V}_s\}_s$, so long as the assets are traded often enough, one still has a good estimate of the daily conditional variance despite the fact that $\{\mathbf{V}_s^{1/2}\}_s$ is random. The nice thing about quadratic variation is that it is well-defined for any process that is a semimartingale (Jacod and Shiryaev 2003, Thm. 4.47). In that sense, it is a completely non-parametric statistic; though the derivations above do not necessarily hold once $\{\mathbf{V}_s\}_s$ is correlated with the underlying Brownian motion. Empirical work has shown that $\{\mathbf{RC}\}_{t=1}^T$ can be used to estimate and forecast the conditional variance of the daily returns in the univariate case (Andersen et al. 2001; Koopman et al. 2005) and the conditional covariance matrix of the vector of daily returns in the multivariate case Liu (2009).

We treat the realized covariances $\{\mathbf{RC}_t\}_{t=1}^T$ (or rather a different, related approximation to $\langle \mathbf{p} \rangle_t - \langle \mathbf{p} \rangle_{t-1}$ called realized kernels) as the noisy observations $\{\mathbf{Y}_t\}_{t=1}^T$ in Section 5 and then infer n, k, and λ to generate filtered estimates and one-step ahead predictions of the latent covariance matrices $\{\mathbf{X}_t^{-1}\}_{t=1}^T$. Barndorff-Nielsen et al. (2011) describe how to construct the matrix valued data and we follow their general approach to produce symmetric positive-definite valued data $\{\mathbf{Y}_t\}_{t=1}^{927}$ for 927 trading days and 30 assets. Details of the construction and the data can be found in Appendix 6.

Appendix D: Construction of Realized Kernel and Data

The data set follows the thirty stocks found in Table 2, which comprised the Dow Jones Industrial Average as of October, 2010. The raw data consists of intraday tick-by-tick trading prices from 9:30 AM to 4:00 PM provided by the Trades and Quotes (TAQ) database through Wharton Research Data Services[1] . The data set runs from February 27, 2007 to October 29, 2010 providing a total of 927 trading days.

Our construction of the realized kernels is based upon Barndorff-Nielsen et al. (2009, 2011). Warning: we re-use the letters $\mathbf{x}$ and $\mathbf{y}$, but now they refer to vector-valued continuous-time processes! Barndorff-Nielsen et al.'s model, which takes into account market microstructure noise, is

$$\mathbf{x}_{t_i} = \mathbf{y}_{t_i} + \mathbf{u}_{t_i}$$

where $\{t_i\}_{i=1}^n$ are the times at which the m-dimensional vector of log stock prices, $\{\mathbf{x}_t\}_{t\geq 0}$, are observed, $\{\mathbf{y}_t\}_{t\geq 0}$ is the latent log stock price, and $\{\mathbf{u}_{t_i}\}_{t=1}^n$ are errors introduced by market microstructure. The challenge is to construct estimates of the quadratic variation of $\{\mathbf{y}_t\}$ with the noisy data $\{\mathbf{x}_{t_i}\}_{i=1}^n$. They do this using a kernel

[1]Wharton Research Data Services (WRDS) was used in preparing this paper. This service and the data available thereon constitute valuable intellectual property and trade secrets of WRDS and/or its third-party suppliers.

Alcoa (AA)	American Express (AXP)	Boeing (BA)
Bank of America (BAC)	Caterpillar (CAT)	Cisco (CSCO)*
Chevron (CVX)	Du Pont (DD)	Disney (DIS)
General Electric (GE)	Home Depot (HD)	Hewlett-Packard (HPQ)
IBM (IBM)	Intel (INTC)*	Johnson & Johnson (JNJ)
JP Morgan (JPM)	Kraft (KFT)	Coca-Cola (KO)
McDonald's (MCD)	3M (MMM)	Merk (MRK)
Microsoft (MSFT)*	Phizer (PFE)	Proctor & Gamble (PG)
AT&T (T)	Traveler's (TRV)	United Technologies (UTX)
Verizon (VZ)	Walmart (WMT)	Exxon Mobil (XOM)

Table 2: The thirty stocks that make up the data set. The asterisk denotes companies whose primary exchange is the NASDAQ. All other companies trade primarily on the NYSE.

approach,

$$K(\mathbf{x}_t) = \sum_{h=-H}^{H} k\Big(\frac{h}{H}\Big)\Gamma_h$$

where

$$\Gamma_h(\mathbf{x}_t) = \sum_{j=h+1}^{n} \mathbf{r}_j \mathbf{r}'_{j-h}, \text{ for } h \geq 0,$$

with $\mathbf{r}_j = \mathbf{x}_{s_j} - \mathbf{x}_{s_{j-1}}$ and $\Gamma_h = \Gamma'_{-h}$ for $h < 0$. The kernel k is a weight function and lives within a certain class of functions. While this provides a convenient formula for calculating realized kernels, the choice of weight function and proper bandwidth H requires some nuance. Barndorff-Nielsen et al. (2011) discuss both issues. We follow their suggestions, using the Parzen kernel for the weight function and picking H as the average of the collection of bandwidths $\{H_i\}_{i=1}^m$ one calculates for each asset individually. Before addressing either of those issues one must address the practical problem of cleansing and synchronizing the data.

Clean the data : The data was cleaned using the following rules.

- Retrieve prices from only one exchange. For most companies we used the NYSE, but for Cisco, Intel, and Microsoft we used FINRA's Alternative Display Facility.
- If there are several trades with the same time stamp, which is accurate up to seconds, then the median price across all such trades is taken to be the price at that time.
- Discard a trade when the price is zero.
- Discard a trade when the correction code is not zero.
- Discard a trade when the condition code is a letter other than 'E' or 'F'.

Synchronize Prices : Regarding synchronization, prices of different assets are not updated at the same instant in time. To make use of the statistical theory for constructing the realized measures one must decide how to "align" prices in time so that they appear to be updated simultaneously. Barndorff-Nielsen et al. suggest constructing a set of refresh times $\{\tau_j\}_{j=1}^{J}$ which corresponds to a "last most recently updated approach." The first refresh time τ_1 is the first time at which all asset prices have been updated. The subsequent refresh times are inductively defined so that τ_n is the first time at which all assets prices have been updated since τ_{n-1}. After cleansing and refreshing the data, one is left with the collection $\{\mathbf{x}_{\tau_j}\}_{j=1}^{J}$ from which the realized kernels will be calculated.

Jitter End Points : For their asymptotic results to hold Barndorff-Nielsen et al. suggest jittering the first and last observations $\{\mathbf{x}_{\tau_j}\}_{j=1}^{J}$. We do this by taking the average of the first two observations and relabeling the resulting quantity as the first observation and taking the average of the last two observations and labeling the resulting quantity as the last observation.

Calculate Bandwidths :

We follow Barndorff-Nielsen et al. (2009) when calculating each H_i individually using the time series $\{x_{t_j}^{(i)}\}_{j=1}^{n}$ before it has been synchronized or jittered. Fix i and suppress it from the notation—we are only considering a single asset. In particular, for asset i the bandwidth H is estimated as

$$\hat{H} = c^*(\hat{\xi}^2)^{2/5} n^{3/5}$$

where $c^* = 0.97$ for the Parzen kernel, n is the number of observations, and

$$\hat{\xi}^2 = \hat{\omega}^2 / \widehat{IV}.$$

$\widehat{IV}$ is the realized variance sampled on a 20 minute grid. $\hat{\omega}^2$ is an estimate of the variance of $\{u_{t_i}\}_{i=1}^{n}$ and is given by

$$\hat{\omega}^2 = \frac{1}{q}\sum_{k=1}^{q} \hat{\omega}_k^2 \quad \text{with} \quad \hat{\omega}_k^2 = \frac{RV_{dense}^{(k)}}{2n_k}.$$

The quantity $RV_{dense}^{(k)}$ is the sum of square increments taken at a high frequency,

$$RV_{dense}^{(k)} = \sum_{j=0}^{n_k - 1} r_j^{(k)2}, \; r_j^{(k)} = (x_{t_{qj+k}} - x_{t_{q(j-1)+k}}), k = 1, \ldots, q,$$

and n_k is the total number of well-defined differences, $x_{t_{qj+k}} - x_{t_{q(j-1)+k}}$, given the data, over $j \in \mathbb{N}$. For each time series we choose $q = \lfloor n/195 \rfloor$, which is the average number of ticks on that day per two minute period (Barndorff-Nielsen et al. 2009).