Inequality Constrained State Space Models

Hang Qian

Abstract

The standard Kalman filter cannot handle inequality constraints imposed on the state variables, as state truncation induces a non-linear and non-Gaussian model. We propose a Rao-Blackwellised particle filter with the optimal importance function for forward filtering and the likelihood function evaluation. The particle filter effectively enforces the state constraints when the Kalman filter violates them. Monte Carlo experiments demonstrate excellent performance of the proposed particle filter with Rao-Blackwellisation, in which the Gaussian linear sub-structure is exploited at both the cross-sectional and temporal levels.

Keywords: Kalman filter, Particle filter, Rao-Blackwellisation, Sequential Monte Carlo
JEL Classification: C32, C53

Address for correspondence: 55 Centre Street, Natick, MA 01760, USA.
E-mail: matlabist@gmail.com

1. Introduction

For economic applications of the state space models, the state variables often represent latent economic processes, some of which are inherently nonnegative or bounded. A leading example is the Gaussian short rate models such as Vasicek (1977) and Hull and White (1990). The conventional wisdom is that the nominal interest rate should be nonnegative (Black, 1995). In an era of low interest rates, the standard Kalman filter results are likely to violate the constraints.

Figure 1 demonstrates the binding inequality constraints in the Vasicek model, in which the instantaneous interest rate is the state variable and the entire term structure is a linear function of the state. Refer to Example 1 for the state and observation equations. The middle panel of Figure 1 shows that the Kalman filter produces negative state estimation after the year 2009. The upper panel plots five posterior samples of the state series by the standard simulation smoothing algorithm (Durbin and Koopman, 2002). All of them contain negative values. We generated millions of posterior series, but could not obtain one that satisfies the nonnegative constraints.

State space applications subject to inequality constraints are common. In the local level model using the Nile river data (see Commandeur et al., 2011), the river flow volume is necessarily a nonnegative state variable. In Stock and Watson (2007), the latent inflation rate is a bounded sequence if the central bank sets inflation targets. In Schwartz (1997) stochastic convenience yield models for pricing futures, the convenience yield represents the benefit of holding the underlying physical commodities rather than derivatives and thus it should be positive. In the Diebold et al. (2006) yield curve applications, the sign of the level, slope and curvature factors might be known if there is prior information on the shape of the yield curve. In the time-varying parameter (TVP) vector autoregressions (VAR) (see Cogley and Sargent, 2005), it is advisable to restrict the eigenvalues of the VAR process within the unit circle.

State constraints can be in the form of equalities and inequalities. Equality constraints can be reformulated as perfect measurement equations. Doran (1992) shows that equality constraints can be incorporated in the state estimation by augmenting measurement equations. Pizzinga (2012) provides a proof of the restricted Kalman filtering based on the Hilbert space geometry and demonstrates statistical efficiency of restricted filtering. Koop et al. (2010) consider a Bayesian application in which the states are subject to time-varying equality constraints.

Imposing inequality constraints on simulation smoothing has received some attention in the literature. Cogley and Sargent (2005) simulate the unrestricted posterior draws and rule out outcomes that violate the constraints by rejection sampling. The acceptance rate of the multi-move algorithm could be low. In our simulation of the Vasicek model, it is difficult to obtain a nonnegative path. Koop and Potter (2011) develop a single-move algorithm, which works well in their application. The single-move algorithm might produce highly correlated draws, as demonstrated in Carter and Kohn (1994).

Imposing inequality constraints on forward filtering receives little attention. To the best of our knowledge, there are no rigorous approaches addressing inequality constrained filtering in econometrics literature. However, there are reasonable ways to add constraints to the Kalman filter. In engineering applications, Simon and Simon (2005) and Gupta and Hauser (2008) adapt the Kalman filter by treating an active set of inequality constraints as equality constraints. Simon and Simon (2010) truncates univariate normal densities for an adapted Kalman filter.

This paper provides a rigorous treatment of the inequality constrained state filtering. Our main contribution is a Rao-Blackwellised particle filter with the optimal importance function, which effectively enforces the inequality constraints when the Kalman filter violates them. Our algorithm departs from the Kalman filter, but analytic integration by the Kalman filter is utilized by Rao-Blackwellisation at both cross-sectional and temporal levels. Our algorithm is based on the particle filter, but not as computationally intensive, since marginalization reduces the state dimensions for particle filtering and muffles Monte Carlo noises.

The reminder of the paper is organized as follows. Section 2 specifies the transition and observation distributions of the inequality constrained model, based on which a particle filter is proposed in Section 3. Section 4 - 6 discuss the cross-sectional and temporal Rao-Blackwellisation, as well as automatic constraint detection for temporal Rao-Blackwellisation. Section 7 extends the model by a more flexible specification on posterior constraints. Section 8 is devoted to numeric comparison of alternative constrained filtering algorithms. Section 9 applies the Rao-Blackwellised particle filter to a nonnegative term structural model. Section 10 concludes the paper by suggesting directions for future research.

2. The Model

Let $x_t, t = 1, \dots, T$ be a $m\times1$ state vector, and y_t be a $n\times1$ observation vector. We define a probabilistic model by the joint density $p(x_{1:T}, y_{1:T})$, where $x_{1:T} = (x_1', \dots, x_T')'$ and $y_{1:T} = (y_1', \dots, y_T')'$. The joint density, decomposed as $\prod_{t=1}^{T} p(x_t|x_{1:t-1}, y_{1:t-1})p(y_t|x_{1:t}, y_{1:t-1})$, is said to be an inequality constrained state space model (ICSSM) if

$$p(x_t|x_{1:t-1}, y_{1:t-1}) = \frac{\phi(x_t; A_t x_{t-1}, Q_t)}{F(A_t x_{t-1}, Q_t, \mathcal{X}_t)} \cdot 1(x_t \in \mathcal{X}_t), \tag{1}$$

$$p(y_t|x_{1:t}, y_{1:t-1}) = \phi(y_t; C_t x_t, R_t), \tag{2}$$

where the matrices A_t, C_t, Q_t, R_t are time-varying coefficients, which could be functions of past observations $y_{1:t-1}$ in economic applications (e.g., autoregressive terms in C_t). The set $\mathcal{X}_t \subset \mathbb{R}^m$ represents the state constraints and the function $1(x_t \in \mathcal{X}_t)$ is a binary indicator for the event $\{x_t|x_t \in \mathcal{X}_t\}$. Also, the density $\phi(x_t; A_t x_{t-1}, Q_t)$ denotes the multivariate normal $N(A_t x_{t-1}, Q_t)$ density evaluated at x_t, and the normalisation term $F(A_t x_{t-1}, Q_t, \mathcal{X}_t)$ denotes the probability of $N(A_t x_{t-1}, Q_t)$ in the region $\mathcal{X}_t$. Note that the normalisation term is a function of the past state x_{t-1}, hence a non-linear model. We assume that $F(\cdot) > 0$, as we address inequality constraints. Equality constraints can be cast as perfect measurement equations and put in Eq (2) instead. As an example of inequality constraints, nonnegative states are represented by $\mathcal{X}_t = \{x_t|x_t \geq 0\}$ with $F(\cdot)$ as the upper cumulative distribution function (c.d.f.). Inequality constraints can be a non-linear function of the states, say $\mathcal{X}_t = \left\{(x_{1t}, x_{2t}, x_{3t}, x_{4t}) \middle| \text{eigenvalues for} \begin{pmatrix} x_{1t} & x_{2t} \\ x_{3t} & x_{4t} \end{pmatrix} \text{in unit circle}\right\}$.

ICSSM follows the state space tradition. First, Markovian transition: $p(x_t|x_{1:t-1}, y_{1:t-1}) = p(x_t|x_{t-1}, y_{1:t-1})$, which is a truncated normal distribution denoted by $TN(A_t x_{t-1}, Q_t, \mathcal{X}_t)$. Second, contemporaneous observations: $p(y_t|x_{1:t}, y_{1:t-1}) = p(y_t|x_t, y_{1:t-1})$. If $\mathcal{X}_t = \mathbb{R}^m$, then Eq (1) and (2) reduce to a Gaussian linear state space form:

$$x_t = A_t x_{t-1} + \varepsilon_t, \tag{3}$$

$$y_t = C_t x_t + v_t, \tag{4}$$

where $\varepsilon_t \sim N(0, Q_t)$, $v_t \sim N(0, R_t)$.

We assume that the initial state vector x_0 is deterministic, without loss of generality because the time-varying coefficient matrices can replicate a non-deterministic initial state distribution. For example, suppose that we require $x_0 \sim TN(\mu_0, \Sigma_0, \mathcal{X}_0)$. Then we may put $x_{-1} = \mu_0$ with $A_0 = I$, $Q_0 = \Sigma_0$, $C_0 = 0$, $R_0 = 0$, $y_0 = 0$. Forward-shifting the time script for all variables in the model by one period (i.e., rewrite x_{-1} as x_0, A_0 as A_1, etc.), we obtain an equivalent state space model with deterministic initial states.

The posterior state distribution takes the form

$$p(x_{1:t}|y_{1:t}) \propto \prod_{\tau=1}^{t} \left[\frac{\phi(x_\tau; A_\tau x_{\tau-1}, Q_\tau)\phi(y_\tau; C_\tau x_\tau, R_\tau)}{F(A_\tau x_{\tau-1}, Q_\tau, \mathcal{X}_\tau)} \cdot 1(x_\tau \in \mathcal{X}_\tau)\right].$$

Due to the normalisation term $F(A_\tau x_{\tau-1}, Q_\tau, \mathcal{X}_\tau)$ in the denominator, the posterior state distribution does not have a closed form for $t > 1$.

3. The Particle Filter

Introduced by Gordon et al. (1993), the particle filter is a powerful tool for characterizing a series of target distributions of increasing dimensions: $p(x_{1:t}|y_{1:t}), t = 1, \ldots, T$. The target density is proportional to $p(x_{1:t}, y_{1:t})$, which can be evaluated pointwise. The proportionality constant equals the likelihood function $p(y_{1:t})$.

Particle filtering is developed in the importance sampling framework. Particles are generated from a well-chosen proposal density $f_t(x_{1:t})$, and assigned the unnormalised importance weights $w_t(x_{1:t}) = \frac{p(x_{1:t}, y_{1:t})}{f_t(x_{1:t})}$. The weighted particles approximate the target distribution, and the sample average of the unnormalised weights approximates the likelihood function. Refer to Liu and Chen (1998), Chopin (2004), Doucet and Johansen (2009), and others.

In sequential importance sampling, the proposal is formulated recursively such that $f_t(x_{1:t}) = f_{t-1}(x_{1:t-1}) \cdot g(x_t|x_{1:t-1})$, where $g(x_t|x_{1:t-1})$ is a well-chosen transition kernel. Weights have a minimum conditional variance if the transition kernel equals $p(x_t|x_{t-1}, y_{1:t})$, which is termed as the *optimal importance function* (See Doucet et al., 2000, p. 199). Under that optimal choice, the weights can be recursively computed by $w_t(x_{1:t}) = w_{t-1}(x_{1:t-1}) \cdot p(y_t|x_{t-1}, y_{1:t-1})$, where $p(y_t|x_{t-1}, y_{1:t-1})$ is termed as the *incremental importance weights*.

Proposition 1: *The optimal importance function for ICSSM particle filtering is given by:*

$$p(x_t|x_{t-1}, y_{1:t}) = \frac{\phi(x_t;\mu_t,\Sigma_t)}{F(\mu_t,\Sigma_t,\mathcal{X}_t)} \cdot 1(x_t \in \mathcal{X}_t), \tag{5}$$

where

$$\mu_t = A_t x_{t-1} + Q_t C_t' (C_t Q_t C_t' + R_t)^{-1} (y_t - C_t A_t x_{t-1}),$$
$$\Sigma_t = Q_t - Q_t C_t' (C_t Q_t C_t' + R_t)^{-1} C_t Q_t.$$

The incremental importance weights are

$$p(y_t|x_{t-1}, y_{1:t-1}) = \phi(y_t; C_t A_t x_{t-1}, C_t Q_t C_t' + R_t) \cdot \frac{F(\mu_t,\Sigma_t,\mathcal{X}_t)}{F(A_t x_{t-1},Q_t,\mathcal{X}_t)}. \tag{6}$$

A proof of Proposition 1 is in the online supplementary appendix. Eq (5) indicates that $p(x_t|x_{t-1}, y_{1:t})$ follows a truncated normal distribution $TN(\mu_t, \Sigma_t, \mathcal{X}_t)$, where μ_t and Σ_t are the single-period Kalman filter outputs. Meanwhile, y_t is subject to incidental truncation (refer to sample selection econometric models; Greene, 2008, p. 883) and $p(y_t|x_{t-1}, y_{1:t-1})$ follows an extended skewed normal distribution, the density of which is given by Eq (6).

To implement the particle filter, we generate period-t particles by Eq (5), and assign them weights by multiplying the previous weights by Eq (6). To evaluate the likelihood function, we take the sample average of the unnormalised weights. In practice, it is necessary to resample the particles when the weights are dispersed. Under the optimal importance function, the weights are not functions of the period-t particles. It is legitimate to reverse the order of sampling and resampling so as to preserve the diversity of the particles.

Example 1*: The Vasicek model*

The Vasicek (1977) model is a one-factor affine term structure model. Let r_t be the latent instantaneous rate and $y_t(\tau)$ be the zero-coupon yield with the maturity τ. Refer to Hull (2003, p. 539) for the model specification. The discretized state and observation equations are

$$r_{t+\Delta t} = e^{-k\Delta t} r_t + (1 - e^{-k\Delta t})\theta + \sqrt{\frac{\sigma^2}{2k}(1 - e^{-2k\Delta t})}\, \varepsilon_t,$$

$$y_t(\tau) = -\frac{1}{\tau}\alpha(\tau) + \frac{1}{\tau}\beta(\tau) r_t + v_t,$$

where $\Delta t = \frac{1}{12}$, $\beta(\tau) = \frac{1-e^{-k\tau}}{k}$, $\alpha(\tau) = [\beta(\tau) - \tau]\left(\theta - \frac{\sigma^2}{2k^2}\right) - \frac{\sigma^2[\beta(\tau)]^2}{4k}$, and ε_t, v_t are independent Gaussian noises. We impose the state constraint $r_t \geq 0, \forall t$.

The estimation data are monthly U.S. treasury rates of maturities from three months to thirty years, 1982-2016. The bottom panel of Figure 1 plots the estimated state series by the particle filter. In contrast with the Kalman filter results (middle panel), the constraints are honored for all outcomes of the particle-filtered distributions. Both the posterior means (the solid line) and the 95% intervals (the dotted lines) of the short rate series are positive.

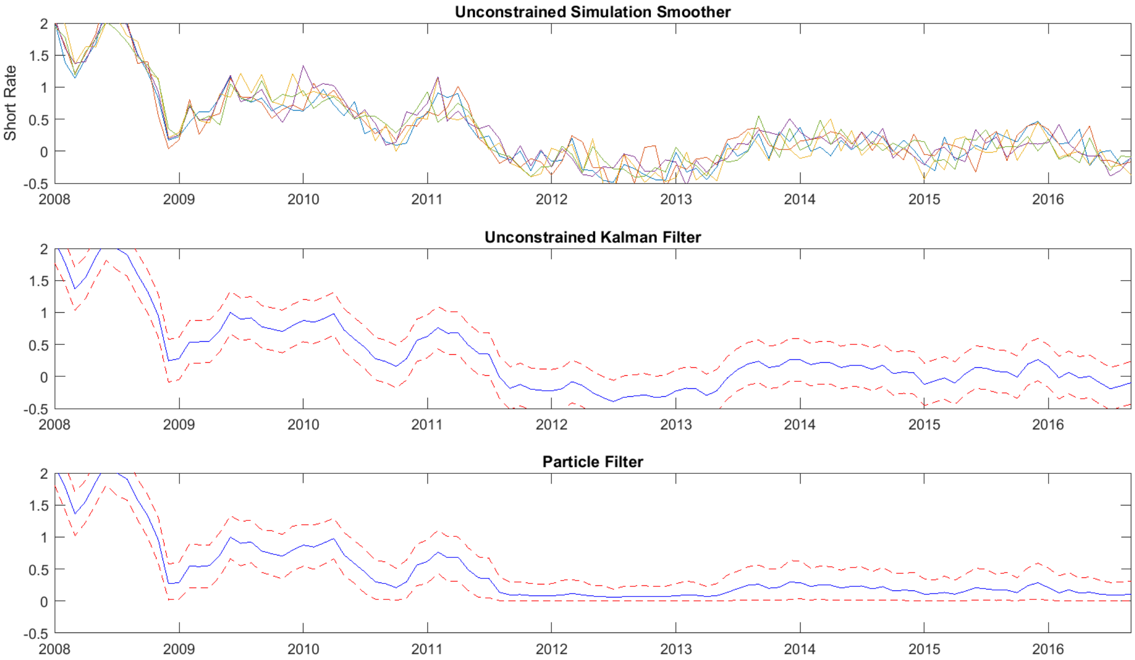

Figure 1 Short rate estimation for the Vasicek model

The upper panel illustrates five paths of the simulated posterior state series using the unconstrained simulation smoother. The middle panel plots the unconstrained Kalman filter estimation of the short rates. The solid line represents the means of the filtered state series and the two dotted lines are the 95% intervals of the filtering distributions. The bottom panel shows the particle filtering results with nonnegative constraints on the short rate series. The estimation sample is 1982 – 2016. To highlight the low interest rate era, the curves are plotted from 2008 to 2016.

4. Cross-sectional Rao-Blackwellisation

In some ICSSM applications, not all the state variables are subject to inequality constraints; some states might be free. It is desirable to decompose the filtering distribution into the analytically tractable and intractable components. The former has a conditionally linear sub-structure and thus can be marginalized by the Kalman filter. We only apply the particle filter to the latter so as to reduce Monte Carlo errors. That technique is known as Rao-Blackwellisation (see Doucet et al., 2001), or termed as mixture Kalman filters (Chen and Liu, 2000) or marginalized particle filtering (Schon et al., 2005).

Partition the state variables and let $x_t = (\xi_t', \eta_t')'$, where the $m_1 \times 1$ constrained states ξ_t must fall into the set $\Xi_t \subset \mathbb{R}^{m_1}$, while $m_2 \times 1$ states η_t are unconstrained. For notational convenience, we consider a diagonal model in which the state transition of ξ_t and η_t has no interactions (this assumption can be relaxed, see below), so that the transition matrix takes a block diagonal form $A_t = diag(A_{1t}, A_{2t})$, $Q_t = diag(Q_{1t}, Q_{2t})$, and $C_t = (C_{1t}, C_{2t})$. The transition and observation densities for the state space model can be written as

$$p(\xi_t|\xi_{t-1}) = \frac{\phi(\xi_t; A_{1t}\xi_{t-1}, Q_{1t})}{F(A_{1t}\xi_{t-1}, Q_{1t}, \Xi_t)} \cdot 1(\xi_t \in \Xi_t),$$

$$p(\eta_t|\eta_{t-1}) = \phi(\eta_t; A_{2t}\eta_{t-1}, Q_{2t}),$$

$$p(y_t|\xi_t, \eta_t) = \phi(y_t; C_{1t}\xi_t + C_{2t}\eta_t, R_t).$$

The target distributions for particle filtering are $p(\xi_{1:t}, \eta_{1:t}|y_{1:t}), t = 1, \ldots, T$, which can be decomposed as

$$p(\xi_{1:t}, \eta_{1:t}|y_{1:t}) = p(\xi_{1:t}|y_{1:t}) \cdot p(\eta_{1:t}|\xi_{1:t}, y_{1:t}).$$

On the one hand, $p(\eta_{1:t}|\xi_{1:t}, y_{1:t})$ is analytically tractable. Conditional on $\xi_{1:t}$, the system reduces to a Gaussian linear sub-model (GLSM), in which η_t is the state vector:

$$\eta_t = A_{2t}\eta_{t-1} + \varepsilon_{2t},$$

$$\tilde{y}_t = C_{2t}\eta_t + v_t,$$

where $\tilde{y}_t = y_t - C_{1t}\xi_t$, and $\varepsilon_{2t} \sim N(0, Q_{2t})$, $v_t \sim N(0, R_t)$.

On the other hand, the intractable component $p(\xi_{1:t}|y_{1:t})$ requires particle filtering. The optimal importance function $p(\xi_t|\xi_{1:t-1}, y_{1:t})$ and the incremental importance weights $p(y_t|\xi_{1:t-1}, y_{1:t-1})$ are summarized by the following proposition.

Proposition 2: *The optimal importance function for particle filtering* $p(\xi_{1:t}|y_{1:t})$*,* $t = 1, \dots, T$*, takes the form:*

$$p(\xi_t|\xi_{1:t-1}, y_{1:t}) = \frac{\phi(\xi_t; \mu_{\xi t}, \Sigma_{\xi t})}{F(\mu_{\xi t}, \Sigma_{\xi t}, \Xi_t)} \cdot 1(\xi_t \in \Xi_t), \tag{7}$$

where

$$\mu_{\xi t} = A_{1t}\xi_{t-1} + Q_{1t}C_{1t}'\Sigma_{yt}^{-1}(y_t - \mu_{yt}),$$

$$\Sigma_{\xi t} = Q_{1t} - Q_{1t}C_{1t}'\Sigma_{yt}^{-1}C_{1t}Q_{1t},$$

$$\mu_{yt} = C_{1t}A_{1t}\xi_{t-1} + C_{2t}\underline{\mu}_{\eta t},$$

$$\Sigma_{yt} = C_{1t}Q_{1t}C_{1t}' + C_{2t}\underline{\Sigma}_{\eta t}C_{2t}' + R_t.$$

The predictive moments $\underline{\mu}_{\eta t}$ *and* $\underline{\Sigma}_{\eta t}$ *are functions of* $\xi_{1:t-1}$*, and can be recursively computed by the Kalman filter using the GLSM. To be specific,*

$$\underline{\mu}_{\eta t} = A_{2,t}\bar{\mu}_{\eta,t-1},$$

$$\underline{\Sigma}_{\eta t} = A_{2,t}\bar{\Sigma}_{\eta,t-1}A_{2,t}' + Q_{2,t},$$

$$\bar{\mu}_{\eta,t} = \underline{\mu}_{\eta t} + \underline{\Sigma}_{\eta t}C_{2t}'\left(C_{2t}\underline{\Sigma}_{\eta t}C_{2t}' + R_t\right)^{-1}\left(y_t - C_{1t}\xi_t - C_{2t}\underline{\mu}_{\eta t}\right),$$

$$\bar{\Sigma}_{\eta,t} = \underline{\Sigma}_{\eta t} - \underline{\Sigma}_{\eta t}C_{2t}'\left(C_{2t}\underline{\Sigma}_{\eta t}C_{2t}' + R_t\right)^{-1}C_{2t}\underline{\Sigma}_{\eta t}.$$

The incremental importance weights under the optimal importance function are given by

$$p(y_t|\xi_{1:t-1}, y_{1:t-1}) = \phi(y_t; \mu_{yt}, \Sigma_{yt}) \cdot \frac{F(\mu_{\xi t}, \Sigma_{\xi t}, \Xi_t)}{F(A_{1t}\xi_{t-1}, Q_{1t}, \Xi_t)}. \tag{8}$$

A proof is in the online appendix. In the Rao-Blackwellised filter, each particle has a Kalman filter, which has contemporaneous interactions with importance sampling, as the Kalman filter "waits for" the particle realizations before it updates the state distributions. Specifically, given the particles ξ_{t-1}, the Kalman filter calculates the filtering distribution $(\bar{\mu}_{\eta,t-1}, \bar{\Sigma}_{\eta,t-1})$ and the predictive distribution $\left(\underline{\mu}_{\eta t}, \underline{\Sigma}_{\eta t}\right)$ based on the GLSM. Then the Kalman filter pauses. The particle filter generates new particles ξ_t from $TN(\mu_{\xi t}, \Sigma_{\xi t}, \Xi_t)$ and assigns importance weights. Taking the particles for ξ_t as given, the Kalman filter updates $(\bar{\mu}_{\eta,t}, \bar{\Sigma}_{\eta,t})$ and proceeds to period $t + 1$ for $\left(\underline{\mu}_{\eta,t+1}, \underline{\Sigma}_{\eta,t+1}\right)$, and so on.

The assumption on the block diagonal A_t and Q_t can be relaxed. Cross-sectional Rao-Blackwellisation is applicable provided that there is an embedded conditional Gaussian linear sub-structure, which exists when the normalisation term $F(A_t x_{t-1}, Q_t, \mathcal{X}_t)$ in Eq (1) is not a function of the past unconstrained states. For example, when A_t is a block lower-triangular matrix and Q_t is a full matrix, the normalisation term only depends on the past constrained states, and thus can be treated as a constant term conditional on $\xi_{1:t-1}$. It follows that $p(x_{1:t-1}|\xi_{1:t-1}, y_{1:t-1})$ is a Gaussian density whose mean and variance are outputs of the Kalman filter using an expanded linear state space model for $\tau = 1, \dots, t-1$:

$$x_\tau = A_\tau x_{\tau-1} + \varepsilon_\tau, \tag{9}$$

$$y_\tau = C_\tau x_\tau + v_\tau, \tag{10}$$

$$\xi_\tau = \left(I_{m_1 \times m_1}, 0_{m_1 \times m_2}\right) \cdot x_\tau, \tag{11}$$

where $x_\tau = (\xi_\tau', \eta_\tau')'$, and Eq (11) is a perfect measurement as the state itself is observed.

Since $p(x_{1:t-1}|\xi_{1:t-1}, y_{1:t-1})$ is a Gaussian density, the state constraints take effects only in period t. The optimal important function $p(\xi_t|\xi_{1:t-1}, y_{1:t})$ remains a low-dimensional truncated normal distribution, and $p(y_t|\xi_{1:t-1}, y_{1:t-1})$ is still an extended skewed normal distribution.

We may interpret the cross-sectional Rao-Blackwellised particle filter as a two-step Kalman filter for each particle. Denote $x_t|\xi_{1:t-1}, y_{1:t} \sim TN\left(\underline{\mu}_{x,t}, \underline{\Sigma}_{x,t}, \mathcal{X}_t\right)$ and $x_t|\xi_{1:t}, y_{1:t} \sim N\left(\bar{\mu}_{x,t}, \bar{\Sigma}_{x,t}\right)$, where $\underline{\mu}_{x,t}, \underline{\Sigma}_{x,t}, \bar{\mu}_{x,t}, \bar{\Sigma}_{x,t}$ can be recursively computed by a two-step process. In the first step, given $\left(\bar{\mu}_{x,t-1}, \bar{\Sigma}_{x,t-1}\right)$, we employ a single-period Kalman filter based on Eq (9) and (10) to calculate $\left(\underline{\mu}_{x,t}, \underline{\Sigma}_{x,t}\right)$, which will be used for generating period-t particles. In the second step, given $\left(\bar{\mu}_{x,t-1}, \bar{\Sigma}_{x,t-1}\right)$ and the new particles, we use a single-period Kalman filter based on Eq (9), (10) and (11) to compute $\left(\bar{\mu}_{x,t}, \bar{\Sigma}_{x,t}\right)$.

Example 2*: Time-Varying Parameter Autoregression*

To illustrate cross-sectional Rao-Blackwellisation, we consider a time-varying AR(2) model:

$$y_t = \phi_0 + \phi_{1t} y_{t-1} + \phi_{2t} y_{t-2} + \varepsilon_t,$$

where $\phi_{it} = \phi_{i,t-1} + u_{it}$ with $u_{it} \sim N(0, \sigma_i^2)$, $i = 1,2$. To ensure a non-explosive series, we impose the triangular constraints: $\phi_{2t} + \phi_{1t} \leq 1$, $\phi_{2t} - \phi_{1t} \leq 1$, $\phi_{2t} \geq -1$.

Our data are quarterly U.S. civilian unemployment rates 1969 – 2015. The unconstrained Kalman filter results, shown in Figure 2, violate the constraint $\phi_{2t} + \phi_{1t} \leq 1$ in 21 quarters, mostly during economic recessions. The upper panel of Figure 3 demonstrates that the particle filter with the triangular constraints suppresses all the spikes that violate the constraints.

The middle and bottom panels of Figure 2 indicate that the other two constraints are not tightly binding. No state point estimator violates constraints. Even the 95% intervals are away from the bounds. Removing those two constraints and reparametrizing the state vector as $(\phi_{2t} + \phi_{1t}, -\phi_{2t})'$, we apply Rao-Blackwellisation to marginalize the second state variable. Figure 3 demonstrates that the filtered states are nearly identical, with the maximum discrepancy 0.0059. The example shows the principle of parsimony, as exclusion of unnecessary constraints will not change the filtering results, but will accelerate the filter (the computing time drops from 11 seconds to 0.2 seconds on a laptop computer).

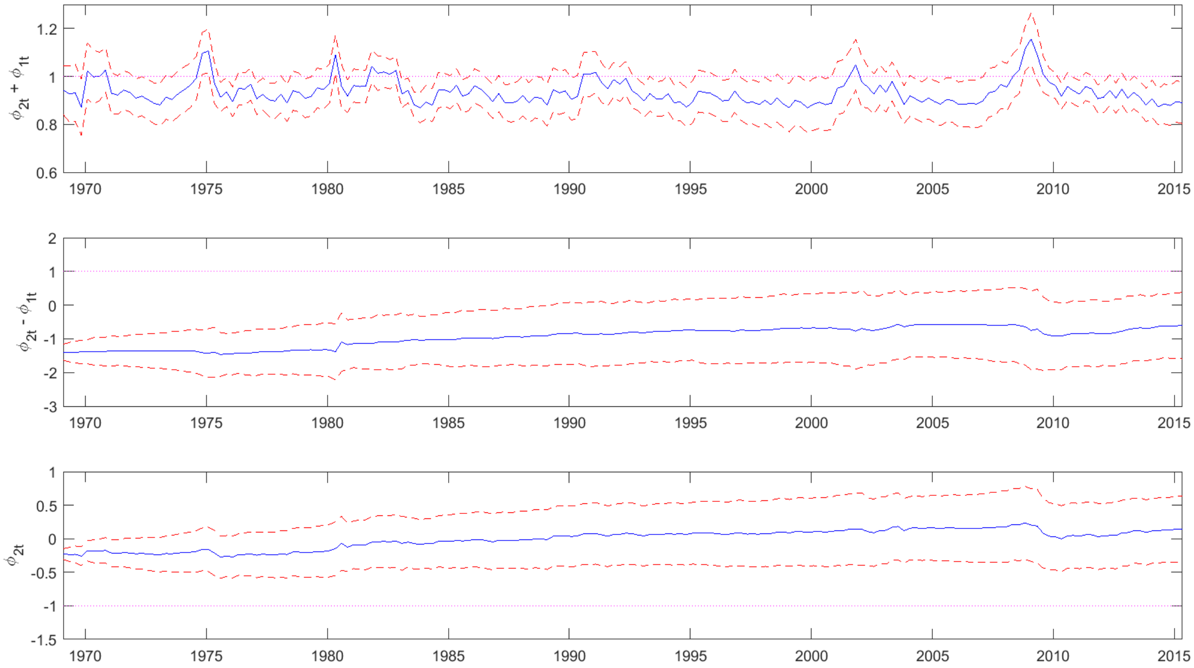

Figure 2 Kalman filter for time-varying AR(2) regression

The filtered state series for $\phi_{2t} + \phi_{1t}$ (upper panel), $\phi_{2t} - \phi_{1t}$ (middle panel) and ϕ_{2t} (lower panel) are obtained from the unconstrained Kalman filter. Solid lines plot the means of the filtering distributions and dashed lines characterize the 95% intervals. The dotted horizontal lines are the stationary AR(2) triangular constraints.

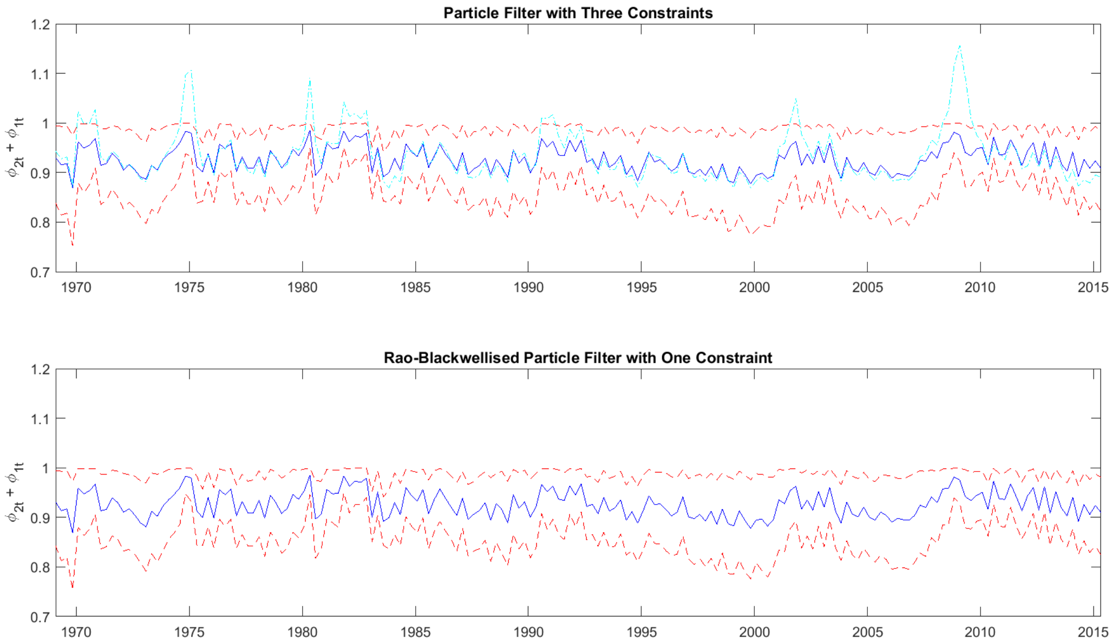

Figure 3 Particle filter for time-varying AR(2) regression

The filtered state series for $\phi_{2t} + \phi_{1t}$ are obtained from the particle filter with and without cross-sectional Rao-Blackwellisation. In the upper panel, all the three inequality constraints are imposed. In the lower panel, only one constraint, namely $\phi_{2t} + \phi_{1t} \leq 1$, is imposed to apply Rao-Blackwellisation. Solid lines plot the means of filtering distributions and dashed lines characterize the 95% intervals. The dash-dot line in the upper panel represents the Kalman filter results.

5. Temporal Rao-Blackwellisation

In the era of high interest rates, few practitioners concerned about the negative rates. Not until recent years when the interest rates plummeted did the concern loom large. Though an inequality constraint always binds the posterior state distribution, the restriction can be tight or loose, depending on the probability that the unrestricted state distribution violates the constraint. It is sensible to impose a constraint only if there is a substantial probability that the constraint is violated. We design a Rao-Blackwellised particle filter that exploits the Gaussian linear sub-structure whenever the constraints are absent. In contrast with cross-sectional marginalization that employs the Kalman filter on a subset of the states, temporal Rao-Blackwellisation resorts to the Kalman filter in a subsample.

Consider Eq (1) – (3) with time-varying constraints: $\mathcal{X}_t = \mathbb{R}^m$ for $t = S+1, \dots, V$, where $1 < S < V < T$. That is, ICSSM reduces to a linear system Eq (3) and (4) in that subsample. Suppose that we have employed the particle filter in the first S periods and the filtering distribution $p(x_{1:S}|y_{1:S})$ is represented by K particles $x_{1:S}^{(i)}$ with the unnormalised weights $\underline{w}_S^{(i)}, i = 1, \dots, K$. In practice, we may only store $x_S^{(i)}$ instead of the entire series.

The question is how to switch to the Kalman filter. It is tempting to initialize the Kalman filter by computing $E(x_S|y_{1:S})$ and $Var(x_S|y_{1:S})$ using the weighted particles. The Kalman filter can produce the best linear state estimator, but cannot characterize the non-Gaussian filtering distribution and cannot represent the likelihood function for ICSSM. It is also tempting to apply the Kalman filter under each of the deterministic initial state $x_S^{(i)}$, and then use the weight $\underline{w}_S^{(i)}$ to average the Kalman filter outputs. As is shown in the following proposition, that method is flawed because the correct weights should incorporate the information contents of $y_{S+1:V}$.

Proposition 3: *Assume that* $p(x_S|y_{1:S})$ *is represented by the* K *particles* $x_S^{(i)}$ *with the unnormalised weights* $\underline{w}_S^{(i)}, i = 1, \dots, K$. *For the unconstrained periods* $t = S+1, \dots, V$,

$$E(x_t|y_{1:t}) = X_{t|t} \cdot \begin{bmatrix} 1 \\ E(x_S|y_{1:t}) \end{bmatrix},$$

$$Var(x_t|y_{1:t}) = P_{t|t} + X_{t|t} \cdot \begin{bmatrix} 0 & 0_{1\times m} \\ 0_{m\times 1} & Var(x_S|y_{1:t}) \end{bmatrix} \cdot X_{t|t}',$$

where $E(x_S|y_{1:t})$ and $Var(x_S|y_{1:t})$ *are the mean and variance for the smoothed distribution defined by the same particles* $x_S^{(i)}$ *with the updated weights* $\bar{w}_S^{(i)}, i = 1, \dots, K$:

$$\bar{w}_S^{(i)} \propto \underline{w}_S^{(i)} \cdot \prod_{\tau=S+1}^{t} \phi\left[V_\tau \cdot \begin{pmatrix} 1 \\ x_S^{(i)} \end{pmatrix}; 0, O_{\tau|\tau-1}\right]. \tag{12}$$

To evaluate the likelihood function,

$$\hat{p}(y_{1:t}) = \frac{1}{K}\sum_{i=1}^{K}\left\{\underline{w}_S^{(i)} \cdot \prod_{\tau=S+1}^{t} \phi\left[V_\tau \cdot \begin{pmatrix} 1 \\ x_S^{(i)} \end{pmatrix}; 0, O_{\tau|\tau-1}\right]\right\} \tag{13}$$

is a consistent estimator for the likelihood value $p(y_{1:t})$.

The matrices $X_{t|t}, P_{t|t}, V_t, O_{t|t-1}$ *are obtained from the augmented Kalman filter (see Durbin and Koopman, 2012, p. 141). The forward recursion starts from* $X_{S|S} = (0_{m\times 1}, I_{m\times m})$, $P_{S|S} = 0_{m\times m}$. *For periods* $t = S+1, \dots, V$, *we sequentially compute the following variables:*

$$X_{t|t-1} = A_t X_{t-1|t-1},$$

$$P_{t|t-1} = A_t P_{t-1|t-1} A_t' + Q_t,$$

$$Y_{t|t-1} = C_t X_{t|t-1},$$

$$O_{t|t-1} = C_t P_{t|t-1} C_t' + R_t,$$

$$V_t = [y_t, 0_{n\times m}] - Y_{t|t-1},$$

$$X_{t|t} = X_{t|t-1} + P_{t|t-1} C_t' \left(O_{t|t-1}\right)^{-1} V_t,$$

$$P_{t|t} = P_{t|t-1} - P_{t|t-1} C_t' \left(O_{t|t-1}\right)^{-1} C_t P_{t|t-1}.$$

A proof is in the online appendix. Proposition 3 shows that $E(x_t|y_{1:t})$ can be computed by the law of iterated expectations. Given a deterministic initial state x_S, the conditional mean $E(x_t|x_S, y_{1:t})$ is a Kalman filter output. Since each particle represents a different initial state, it is legitimate to take the weighted average of the Kalman filter outputs. However, the correct weights come from the smoothing distribution $p(x_S|y_{1:t})$.

Random samples from $p(x_t|y_{1:t})$ can be generated using the following identity:

$$p(x_s, x_t|y_{1:t}) = p(x_s|y_{1:t}) \cdot p(x_t|x_s, y_{1:t}). \qquad (14)$$

We can first generate a draw from the smoothing distribution $p(x_S|y_{1:t})$, which is essentially a resample of the particles with weights given by Eq (12). Conditional on that draw, we sample from $p(x_t|x_S, y_{S+1:t})$ using the Kalman filter. Those equally-weighted samples fully characterize the filtering distribution $p(x_t|y_{1:t})$. Alternatively, weighted draws can also represent that distribution. Note that

$$p(x_t|y_{1:t}) = \int p(x_s, x_t|y_{1:t}) dx_s = \sum_{i=1}^{K} \left\{ \bar{w}_S^{(i)} \cdot p\left(x_t \middle| x_S^{(i)}, y_{1:t}\right) \right\}.$$

We may use the Kalman filter to generate a draw from $p\left(x_t \middle| x_S^{(i)}, y_{1:t}\right)$ based on the original particles, then assign it with the smoothing weight $\bar{w}_S^{(i)}$.

In practice, we only need to generate random samples or weighted samples in period V, as the state constraints will be in effect again and we switch to the particle filter for $t = V + 1, \ldots, T$. If we treat the random samples generated from $p(x_V|y_{1:V})$ as the period-V particles, we assign them the unnormalised weights $p(y_{1:V})$, an estimator of which is given by Eq (13). If we treat the weighted samples as the period-V particles, we assign them the unnormalised weights given by the right hand side of Eq (12), namely the unnormalised version of $\overline{w}_S^{(i)}$. Both methods ensure that the average unnormalised weights approximate the likelihood function.

Example 3*: The Vasicek model (continued)*

To illustrate temporal Rao-Blackwellisation in the Vasicek model, we remove the state constraint $r_t \geq 0$ when the interest rates were high, and only impose that constraint after December 2008. The Rao-Blackwellised filter starts from the Kalman filter and then switches to the particle filter. Figure 4 compares the particle filtering results with and without temporal Rao-Blackwellisation. The upper and middle panel are almost identical, with the maximum discrepancy 2.8 basis points. The example demonstrates the principle of parsimony, for the constraint before 2008 has no significant impact on the filtering results. However, imposing the constraint after 2008 effectively removes the anomaly produced by the Kalman filter.

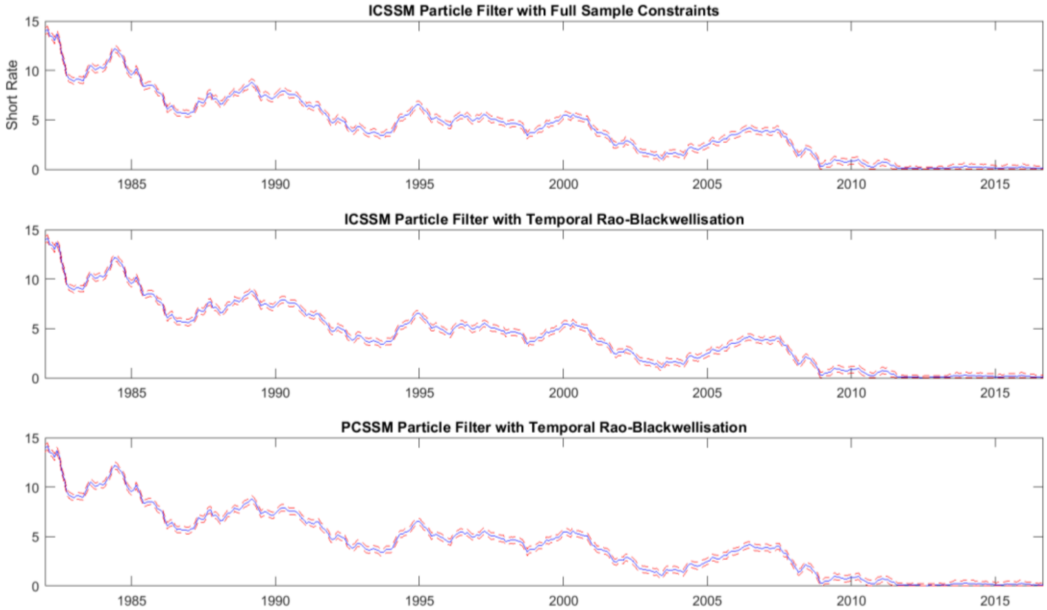

Figure 4 Temporal Rao-Blackwellised particle filter for the Vasicek model

The upper panel plots the short rates estimated by the particle filter under the full-sample nonnegative constraints. The curves are identical to the bottom panel of Figure 1, except that the x-axis range is 1982 - 2016. The solid line represents the means of the filtered states and the two dotted lines are the 95% intervals. The middle panel illustrates the ICSSM particle filter with temporal Rao-Blackwellisation, with the constraint imposed after December 2008. The bottom panel corresponds to the Rao-Blackwellised PCSSM results.

6. Automatic Constraint Selection

For simple applications, the unconstrained subsample for temporal Rao-Blackwellisation can be found by a priori knowledge or visual inspection on the Kalman filter results. In more complex settings, we wish automatic constraint selection and real-time switching between the particle and Kalman filters. Our idea is to let the particles decide whether the constraints can be lifted in the next period by evaluating the bias of ignoring constraints.

Suppose that we run the particle filter with resampling every period and have obtained K equally weighted particles $x_{t-1}^{(i)}, i = 1, \dots, K$ that characterize $p(x_{t-1}|y_{1:t-1})$. Those particles shed light on the bias on the state estimator $E(x_t|y_{1:t})$ induced by switching to the unconstrained Kalman filter. We have two scenarios:

First, if we continue the particle filter, by Proposition 1, we generate new particles $x_t^{(i)}$ from $TN\left(\mu_t^{(i)}, \Sigma_t, \mathcal{X}_t\right)$, with the unnormalised weight $w_t^{(i)}$, where

$$\mu_t^{(i)} = A_t x_{t-1}^{(i)} + Q_t C_t' (C_t Q_t C_t' + R_t)^{-1} \left(y_t - C_t A_t x_{t-1}^{(i)}\right),$$

$$w_t^{(i)} = \phi\left(y_t; C_t A_t x_{t-1}^{(i)}, C_t Q_t C_t' + R_t\right) \cdot \frac{F\left(\mu_t^{(i)}, \Sigma_t, \mathcal{X}_t\right)}{F\left(A_t x_{t-1}^{(i)}, Q_t, \mathcal{X}_t\right)}.$$

Thus the particle filter estimates $E(x_t|y_{1:t})$ by $G_1 = \sum_{i=1}^{K} \frac{w_t^{(i)}}{\sum_{j=1}^{K} w_t^{(j)}} x_t^{(i)}$.

Second, if we switch to the unconstrained filter with temporal Rao-Blackwellisation (loosely speaking, switch to the Kalman filter), by Proposition 3, we average the Kalman filter outputs, namely $\mu_t^{(i)}$, by the updated weights proportional to $\bar{w}_t^{(i)}$. Refer to Eq (12). The estimated $E(x_t|y_{1:t})$ is given by $G_0 = \sum_{i=1}^{K} \frac{\bar{w}_t^{(i)}}{\sum_{j=1}^{K} \bar{w}_t^{(j)}} \mu_t^{(i)}$, where $\bar{w}_t^{(i)} = \phi\left(y_t; C_t A_t x_{t-1}^{(i)}, C_t Q_t C_t' + R_t\right)$.

The bias is defined as $G_1 - G_0$, which is random because we are given the particles $x_{t-1}^{(i)}$ but have not generated $x_t^{(i)}$. The following proposition calculates the expected bias and variance.

Proposition 4: *For ICSSM particle filtering with the optimal important function, conditional on the period $t-1$ particles $x_{t-1}^{(i)}, i = 1, \dots, K$ and observations y_t, the bias induced by switching to the Kalman filter in period t has the mean and variance*

$$E(G_1 - G_0) = \sum_{i=1}^{K} \left[\frac{w_t^{(i)}}{\sum_{j=1}^{K} w_t^{(j)}} M_t^{(i)} - \frac{\bar{w}_t^{(i)}}{\sum_{j=1}^{K} \bar{w}_t^{(j)}} \mu_t^{(i)} \right],$$

$$Var(G_1 - G_0) = \sum_{i=1}^{K} \left[\frac{w_t^{(i)}}{\sum_{j=1}^{K} w_t^{(j)}} \right]^2 H_t^{(i)},$$

where $M_t^{(i)}$ and $H_t^{(i)}$ are the mean and variance of $TN\left(\mu_t^{(i)}, \Sigma_t, \mathcal{X}_t\right)$, respectively.

Proposition 4 provides guidance on real-time constraint selection. Given the mean and variance of the bias, we may apply the Chebyshev inequality to estimate the probability that the bias falls within a tolerance level. Alternatively, a simple switching rule can be the expected relative bias, defined as $\left|\frac{E(G_1 - G_0)}{G_0}\right|$, falling below some threshold α.

Proposition 4 requires the optimal importance function, because the importance weights are predetermined one period in advance. For other particle filtering schemes such as the bootstrap filter, the weights are functions of the random particles $x_t^{(i)}$ and the expected bias could not be easily computed.

Example 4*: Time-Varying Parameter Autoregression (continued)*

To illustrate automatic constraint selection, we continue the time-varying AR(2) example, in which we showed that only one constraint, namely $\phi_{2t} + \phi_{1t} \leq 1$, tightly binds the states. Now we make the constraint even more parsimonious by automatic switching to the unconstrained filter if the expected relative bias drops below the level $\alpha = 1\%, 2\%, 5\%$.

In Figure 5, automatically selected dates are marked on the horizontal axis. When $\alpha = 1\%$, 39 out of 186 periods are subject to the constraint. They cover all the periods when the Kalman filter produces large spikes. As α rises to 2% and 5%, the selected periods drop to 21 and 9, respectively. They correspond to the largest-spike dates. In general, automatic detection is as accurate as visual inspection. The choice of the tolerance level α is a trade-off between precision and speed, both of which are critical for online filtering.

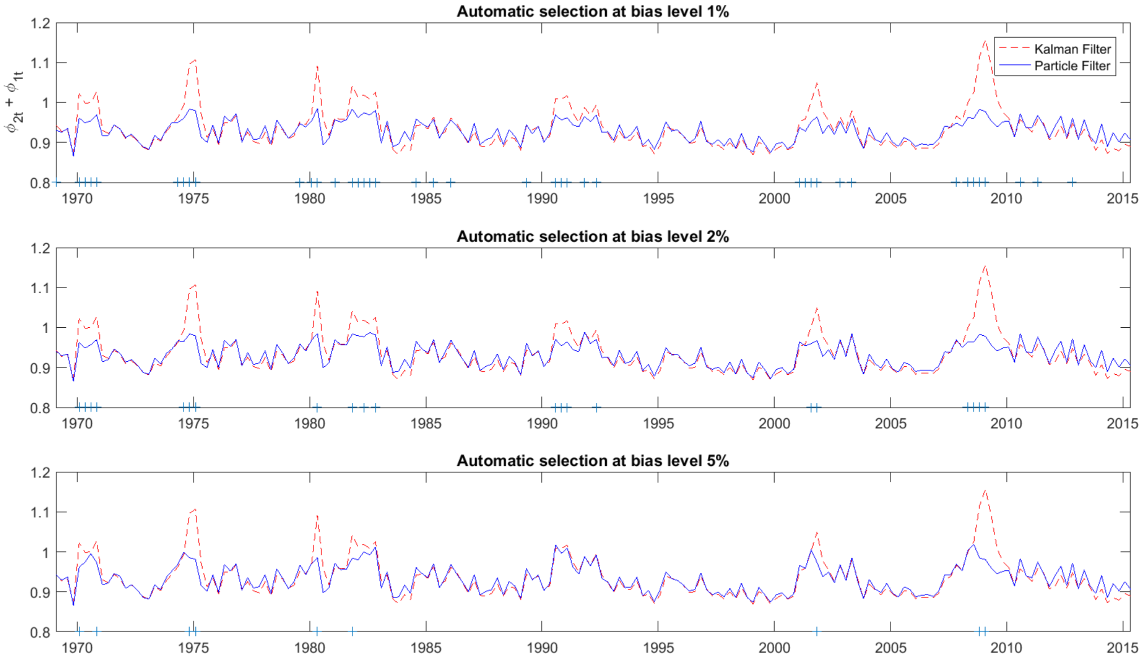

Figure 5 Automatic constraint selection for time-varying AR(2) regression

The three panels plot the temporal Rao-Blackwellised filtering results (solid lines) for $\phi_{2t} + \phi_{1t}$ with a subsample of constraints automatically selected by the expected bias criterion at levels $\alpha = 1\%, 2\%, 5\%$, respectively. The symbol "+" marked on the horizontal axis indicates the dates when constraints are imposed. For comparison, the unconstrained Kalman filter results (dashed lines) are shown in the same figure.

7. An Alternative View on State Constraints

State space models use observations to update the prior state distributions. Constraints can be viewed as additional observations, conditional on which the posterior state distributions satisfy the inequality constraints. A model is said to be a posterior constrained state space model (PCSSM) if

$$x_t = A_t x_{t-1} + \varepsilon_t, \tag{15}$$

$$y_t = C_t x_t + v_t, \tag{16}$$

$$z_t = 1(x_t \in \mathcal{X}_t), \tag{17}$$

where $\varepsilon_t \sim N(0, Q_t)$, $v_t \sim N(0, R_t)$. We introduce an auxiliary measurement variable z_t for the state constraints. In addition to the regular observations y_t, we also observe $z_t = 1, \forall t$.

ICSSM imposes constraints at the prior stage, while the constraints in PCSSM are honored in the posterior distributions. The TVP-VAR model in Cogley and Sargent (2005) can be interpreted as a PCSSM, while that in Koop and Potter (2011) can be viewed as an ICSSM. The key difference is that the ICSSM normalisation term is a function of past states, while that of PCSSM is truly a constant. Consequently, PCSSM has analytic properties that ICSSM does not have.

We rewrite Eq (15) and (16) in the matrix form: $Ax_{1:t} = c_0 + \varepsilon_{1:t}$, $y_{1:t} = Cx_{1:t} + v_{1:t}$, where $A = \begin{pmatrix} I & 0 & \cdots & 0 & 0 \\ -A_2 & I & \cdots & 0 & 0 \\ \cdots & \cdots & \ddots & \cdots & \cdots \\ 0 & 0 & \cdots & -A_t & I \end{pmatrix}$, $c_0 = \begin{pmatrix} A_1 x_0 \\ 0 \\ \cdots \\ 0 \end{pmatrix}$, and $C = diag(C_1, \dots, C_t)$.

Proposition 5: *Conditional on $z_{1:t} = 1$, the state filtering distribution for PCSSM is truncated multivariate normal with the density:*

$$p(x_{1:t}|z_{1:t}) = \frac{\phi\left(x_{1:t}; \underline{\mu}_{1:t}, \underline{\Sigma}_{1:t}\right)}{F\left(\underline{\mu}_{1:t}, \underline{\Sigma}_{1:t}, \mathcal{X}_{1:t}\right)} \cdot 1(x_{1:t} \in \mathcal{X}_{1:t}),$$

$$p(x_{1:t}|y_{1:t}, z_{1:t}) = \frac{\phi(x_{1:t}; \overline{\mu}_{1:t}, \overline{\Sigma}_{1:t})}{F(\overline{\mu}_{1:t}, \overline{\Sigma}_{1:t}, \mathcal{X}_{1:t})} \cdot 1(x_{1:t} \in \mathcal{X}_{1:t}),$$

where $\underline{\mu}_{1:t} = A^{-1}c_0$, $\underline{\Sigma}_{1:t} = A^{-1}QA'^{-1}$,

$$\overline{\mu}_{1:t} = \underline{\mu}_{1:t} + \underline{\Sigma}_{1:t}C'\left(C\underline{\Sigma}_{1:t}C' + R\right)^{-1}\left(y_{1:t} - C\underline{\mu}_{1:t}\right),$$

$$\overline{\Sigma}_{1:t} = \underline{\Sigma}_{1:t} - \underline{\Sigma}_{1:t}C'\left(C\underline{\Sigma}_{1:t}C' + R\right)^{-1}C\underline{\Sigma}_{1:t},$$

$$Q = diag(Q_1, \dots, Q_t), R = diag(R_1, \dots, R_t), \mathcal{X}_{1:t} = \{x_{1:t}|x_\tau \in \mathcal{X}_\tau, \forall\tau = 1, \dots, t\}.$$

The PCSSM particle filter with the optimal importance function is still tractable. The following result is the PCSSM version of Proposition 1.

Proposition 6: *Given $z_t = 1, \forall t$, the optimal importance function for PCSSM particle filtering $p(x_{1:t}|y_{1:t}, z_{1:t})$, $t = 1, \dots, T$ takes the form:*

$$p(x_t|x_{t-1}, y_{1:t}, z_{1:t}) = \frac{\phi(x_t; \mu_t, \Sigma_t)}{F(\mu_t, \Sigma_t, \mathcal{X}_t)} \cdot 1(x_t \in \mathcal{X}_t).$$

The incremental importance weights are

$$p(y_t, z_t|x_{t-1}, y_{1:t-1}, z_{1:t-1}) = \phi(y_t; C_tA_tx_{t-1}, C_tQ_tC_t' + R_t) \cdot F(\mu_t, \Sigma_t, \mathcal{X}_t),$$

where μ_t, Σ_t are the same as those defined below Eq (5).

PCSSM has computational advantages. First, PCSSM weights involve a multivariate normal probability, while ICSSM requires the ratio of two probabilities. Second, for an outlier particle, the probability ratio could be vulnerable to numerical errors and some particles with unreasonably large weights may propagate to the next-period filtering. However, PCSSM has a self-stabilizing mechanism: the incremental weights assigned to outlier particles are small. They are likely to be discarded by resampling, and numerical errors do not propagate.

Rao-Blackwellisation is still applicable. Proposition 3 on temporal Rao-Blackwellisation holds for PCCSM without modification. For cross-sectional Rao-Blackwellisation, the PCSSM counterpart of Proposition 2 is stated below.

Proposition 7: *Given $z_t = 1, \forall t$, the optimal importance function for PCSSM particle filtering $p(\xi_{1:t}|y_{1:t}, z_{1:t})$, $t = 1, \dots, T$ in the diagonal model takes the form:*

$$p(\xi_t|\xi_{1:t-1}, y_{1:t}, z_{1:t}) = \frac{\phi(\xi_t; \mu_{\xi t}, \Sigma_{\xi t})}{F(\mu_{\xi t}, \Sigma_{\xi t}, \Xi_t)} \cdot 1(\xi_t \in \Xi_t).$$

The incremental importance weights are

$$p(y_t, z_t|\xi_{1:t-1}, y_{1:t-1}) = \phi(y_t; \mu_{yt}, \Sigma_{yt}) \cdot F(\mu_{\xi t}, \Sigma_{\xi t}, \Xi_t),$$

where $\mu_{\xi t}, \Sigma_{\xi t}, \mu_{yt}, \Sigma_{yt}$ are the same as those defined in Proposition 2.

For cross-sectional Rao-Blackwellisation, PCSSM allows more general dependencies between the constrained and unconstrained state components. We provide some intuition: Rao-Blackwellisation is applicable when a nonlinear model has an embedded Gaussian linear sub-structure. If conditioning on a subset of variables removes all the nonlinear factors and reveals the Gaussian linear sub-structure, then we can perform Rao-Blackwellisation. Nonlinearity in PCSSM is induced by a single factor: the binary function indicating truncation. Conditioning on the truncated variable removes nonlinearity, and Rao-Blackwellisation is applicable no matter how state components interact with each other in transition. However, nonlinearity in ICSSM is caused by two factors: the truncation indicator and a normalisation term that depends on both truncated and unconstrained states. Conditioning on the truncated variable removes nonlinearity from the first factor, but not the second. In the special case that the second factor is absent (i.e., the past unconstrained states have no impact on the normalisation term), we have the Gaussian linear sub-structure and can perform Rao-Blackwellisation.

Despite various virtues, PCSSM is awkward in its data generating process, as it is difficult to draw the triple $(x_{1:T}, y_{1:T}, z_{1:T})$ by Eq (15) – (17). We may first generate a candidate $x_{1:T}$ by Eq (15), but have to reject it if any state violates the constraints. The observations $y_{1:T}$, as seen in a real-world data set, are the result of a lucky state sequence that survives all the constraints.

Example 5*: The Vasicek model (continued)*

The bottom panel of Figure 4 illustrates temporal Rao-Blackwellised particle filter under PCSSM. The filtering results are similar to the ICSSM counterparts, with the maximum discrepancy 2.1 basis points.

8. Monte Carlo Simulations

In this section, we perform numerical exercises to answer two questions: How does our approach compare with the constrained Kalman filtering proposed in the literature? Can our approach handle large state space models?

The constrained Kalman filtering (Simon and Simon, 2005 and Gupta and Hauser, 2008) is a reasonable method that incorporates inequality constraints in the Kalman filter by solving quadratic programming problems. In each period, an active set of binding inequality constraints translate to equality constraints, which are handled by the equality-constrained Kalman filter.

In our first Monte Carlo experiment, we consider an unobserved component state space model. The linear version is

$$x_{1t} = x_{1,t-1} + \varepsilon_{1t},$$

$$x_{2t} = \phi x_{2,t-1} + \varepsilon_{2t},$$

$$y_t = x_{1t} + x_{2t} + \varepsilon_{3t},$$

where $x_{2t} \geq 0$ by the state equation truncation and $\varepsilon_{it} \sim N(0, \sigma_i^2), i = 1, 2, 3$. The parameter values are $\phi = 0.6$, $\sigma_1 = 0.1$, $\sigma_2 = 0.2$, $\sigma_3 = 0.3$, and states are initialized at zeros.

The Monte Carlo experiment involves 500 replications. In each replication, we simulate 100 periods of states and observations, and then run the unconstrained Kalman filter, active-set constrained Kalman filter, ICSSM and PCSSM Rao-Blackwellised particle filters with 1000 particles. Since the true states are simulated (but not used by any filter), we can calculate the root mean squared error (RMSE): $\sqrt{\frac{1}{T}\sum_{t=1}^{T}(x_{it}-\hat{x}_{it})^2}$, $i = 1{,}2$, where x_{it} is the true state and $\hat{x}_{it}$ is the state estimator by the Kalman/active-set/particle filters.

Table 1 reports the simulation results. The unconstrained Kalman filter violates the constraint in nearly half of the sample periods, hence the large RMSE of 0.274 and 0.265 for the two states respectively. The constrained Kalman filter censors the second state estimator to the lower bound when the constraint is active, which reduces RMSE to 0.222 and 0.207. The Rao-Blackwellised particle filter marginalizes the first state variable and enforces the constraint for all outcomes of the filtering distribution of the second state. The ICSSM particle filter has the lowest RMSE: 0.175 and 0.144. We also tested the PCSSM particle filter, and the results are very close: 0.180 and 0.146.

	x1	x2
Kalman	0.274	0.265
	(0.002)	(0.001)
Active-set	0.222	0.207
	(0.002)	(0.001)
ICSSM	0.175	0.144
	(0.001)	(0.001)
PCSSM	0.180	0.146
	(0.001)	(0.001)

Table 1 RMSE of alternative filtering methods

Two states (x1 and x2) are simulated from an unobserved component state space model. RMSE of the filtered states are reported under the unconstrained Kalman filter, active-set constrained Kalman filter, ICSSM and PCSSM particle filters. The numeric standard errors are in parentheses, which reflect Monto Carlo variations.

Our second Monte Carlo experiment evaluates the feasibility of filtering a larger model with 30 state variables, the first 10 of which are bounded above by zero. The sample size is 500, with 20 repeated measurements every period. The linear version of the model is

$$x_{it} = \frac{i}{30}x_{i,t-1} + u_{it}, i = 1, \dots, 30,$$

$y_{jt} = \sum_{i=1}^{30} \frac{i}{30} x_{it} + \varepsilon_{jt}, j = 1, \dots, 20,$

where $u_{it} \sim N(0, 1)$ with $cov(u_{it}, u_{kt}) = 0.3, \forall i, k$. Also, $\varepsilon_{jt} \sim N(0,1)$ without correlation.

The computational challenge of particle filtering comes from the importance weights, which involve a 10-dimension multivariate normal probability for each of the 1000 particles, even after Rao-Blackwellisation. We first resort to the popular Geweke-Hajivassiliou-Keane (GHK) simulator, under which the particle filter works but is slow (about 15 minutes). We also try the simple Monte Carlo, which estimates probabilities by frequencies that random draws fall into the constraint set. The conventional wisdom is that the simple Monte Carlo is crude, but runs fast, does not suffer from the curse of dimensionality, and can handle nonlinear constraints.

We calculate the RMSE between the true states and the estimated states by the Kalman/particle filters. In this example, the state dimension is high but the repeated measurements provide inadequate information, so the average RMSE of the Kalman filter amounts to 5.675. Knowledge on the state upper bound is valuable, and the RMSE of the particle filter under the GHK simulator decreases to 1.326. In particular, for the 10 constrained states, the RMSE is 0.708. A surprising finding is that the filtering results under the simple Monte Carlo simulator (with 500 draws) are highly similar to those under the GHK simulator, with the average discrepancy of 0.026. The RMSE under the simple Monte Carlo is 1.329 and the 10 constrained states have the average RMSE 0.708, which is marginally larger than the GHK results. However, the particle filter under the simple Monte Carlo method is faster and it only costs 1/5 of the computing time compared to the GHK simulator.

The results suggest that our particle filter can handle larger state space models, possibly with nonlinear constraints, as the obstacle of evaluating multivariate normal probabilities can be overcome by the simple Monte Carlo approach. We provide some intuition: suppose that the true probability of the $F(\cdot)$ term is 0.6. A good simulator produces 0.601 and a coarse simulator yields 0.61. However, the unconstrained Kalman filter assumes that the $F(\cdot)$ term always equals to 1. As a result, a coarse particle filter can still rectify most of the bias induced by the Kalman filter.

9. An Application

We consider estimating the zero-lower-bound (ZLB) yields based on the arbitrage-free Nelson-Siegel (AFNS) model proposed by Christensen et al. (2011). The dynamic term structure model involves the level, slope and curvature factors (L_t, S_t, C_t) that follow the Ornstein-Uhlenbeck processes. The instantaneous rate r_t is the sum of L_t and S_t. The zero-coupon yield with τ years to maturity, $y_t(\tau)$, is a linear function of the latent factors. An independent-factor AFNS model under the P-measure reads

$$\begin{pmatrix} dL_t \\ dS_t \\ dC_t \end{pmatrix} = \begin{pmatrix} \kappa_1 & 0 & 0 \\ 0 & \kappa_2 & 0 \\ 0 & 0 & \kappa_3 \end{pmatrix} \left[\begin{pmatrix} \theta_1 \\ \theta_2 \\ \theta_3 \end{pmatrix} - \begin{pmatrix} L_t \\ S_t \\ C_t \end{pmatrix} \right] dt + \begin{pmatrix} \sigma_1 & 0 & 0 \\ 0 & \sigma_2 & 0 \\ 0 & 0 & \sigma_3 \end{pmatrix} \begin{pmatrix} dW_{1t} \\ dW_{2t} \\ dW_{3t} \end{pmatrix}, \tag{18}$$

$$y_t(\tau) = L_t + \frac{1-e^{-\lambda\tau}}{\lambda\tau} S_t + \left(\frac{1-e^{-\lambda\tau}}{\lambda\tau} - e^{-\lambda\tau} \right) C_t - \frac{A(\tau)}{\tau} + \varepsilon_t, \tag{19}$$

$$r_t = L_t + S_t, \tag{20}$$

where $\frac{A(\tau)}{\tau}$ is the yield-adjustment term, which is a deterministic function of τ and the model parameters $\sigma_1, \sigma_2, \sigma_3$. Refer to Christensen et al. (2011) for the analytical formula for that term. We follow Christensen and Rudebusch (2013) and let κ_1 be infinitesimal and $\theta_1 = 0$, which leads to a persistent level factor.

The continuous-time AFNS model can be discretized as a Gaussian linear state space model. Meanwhile, it is desirable to impose the ZLB constraint on the instantaneous rate $r_t \geq 0$, which translates to an ICSSM or PCSSM. The state filtering distributions are determined by 1) the asset pricing model dynamics, 2) observations on the yields, and 3) knowledge of ZLB.

Christensen and Rudebusch (2015) assume that r_t is censored at zero when the shadow rate is negative. Using an option-based approach, they approximate the arbitrage-free pricing relationships and provide the bond yield formula, which is a nonlinear function of the latent factors estimated by the extended Kalman filter. In contrast, PCSSM views ZLB from a new perspective: the affine term structure model (i.e., Eq (18) and (19)) predicts the theoretic bond yields. Knowledge of ZLB is not part of the theoretic model, but an auxiliary observation. Such knowledge can further update the state distributions.

An unconstrained AFNS model can be fitted by maximum likelihood estimation (MLE) via the standard Kalman filter. See Christensen et al. (2011). Under the ZLB constraint, MLE is challenging due to Monte Carlo errors in likelihood evaluation via the particle filter. We find that the Bayesian approach works better in this application. We adopt diffuse priors on the model parameters, and use a random-walk Metropolis sampler with the multivariate t proposal distribution. The posterior density is proportional to the prior times the likelihood, which is evaluated by the Rao-Blackwellised particle filter. Refer to the appendix for details on the estimation and filtering procedures.

Our estimates are based on monthly US treasury constant maturity rates of 1/4, 1/2, 1, 2, 3, 5, 7,10, 20, 30 years, from January 1982 to September 2016. We impose the ZLB constraint after December 2008 when the federal funds target range is lowered to 0 to 1/4 percent. Table 2 compares the unrestricted MLE to the Bayesian estimator with and without the ZLB constraint. The estimated parameters are similar and significant, with an exception κ_2 not significantly different from zero, which “accords with the usual finding that one or more of the interest rate factors are close to being nonstationary process under the P-measure” (Christensen et al., 2011, p.7). The one-month conditional mean-reversion matrix, estimated by the Bayesian approach with the ZLB constraint, is given by $diag(1, 0.99, 0.97)$. The estimated volatility terms $\sigma_1, \sigma_2, \sigma_3$ are significant, and the associated one-month conditional covariance matrix is $diag(1.25, 7.90, 39.5) \cdot 10^{-6}$. The model fits the data well, as evidenced by the small RMSE of the fitted yield curve: 12.9, 6.1, 10.2, 10.6, 4.8, 7.7, 8.2, 11.8, 12.2, 10.2 basis points for the above-mentioned maturities, respectively.

	MLE	Bayesian	Particle
σ_1	0.0039	0.0039	0.0039
	(0.0000)	(0.0001)	(0.0001)
σ_2	0.0098	0.0098	0.0097
	(0.0002)	(0.0004)	(0.0004)
σ_3	0.0217	0.0218	0.0218
	(0.0007)	(0.0009)	(0.0009)
σ_{obs}	0.0037	0.0038	0.0038
	(0.0000)	(0.0001)	(0.0001)
λ	0.4853	0.4857	0.5013
	(0.0031)	(0.0060)	(0.0057)
κ_2	0.0056	0.1292	0.1223
	(0.1627)	(0.1089)	(0.1072)
κ_3	0.4099	0.4002	0.3849
	(0.1535)	(0.1730)	(0.1744)

Table 2 Parameter estimation of AFNS model

The second and third columns show the maximum likelihood and Bayesian estimators for the unconstrained model. The last column corresponds to the Bayesian estimator subject to the ZLB constraint. The point estimators (posterior means) are reported, with standard errors (posterior standard deviations) in parentheses.

The AFNS model is known for its superior predictive performance, but ignoring ZLB induces anomalies in the forecasted yield curve. Figure 6 highlights the problem when yields at historic lows. We filter the states by observations from January 1982 to January 2015, and then make a 6-month-ahead forecast. The predicted curves with and without ZLB largely overlap for maturities between 1 to 10 years, but the unconstrained Kalman filter predicts negative yields for maturities $\tau < 0.25$ years. The anomaly in yield curve prediction is caused by the flawed state estimation. The instantaneous rates estimated by the Kalman filter falls below zero at various points in time after 2008, but the particle filter always observes the ZLB constraint and the predicted yield curve is positive for all maturities.

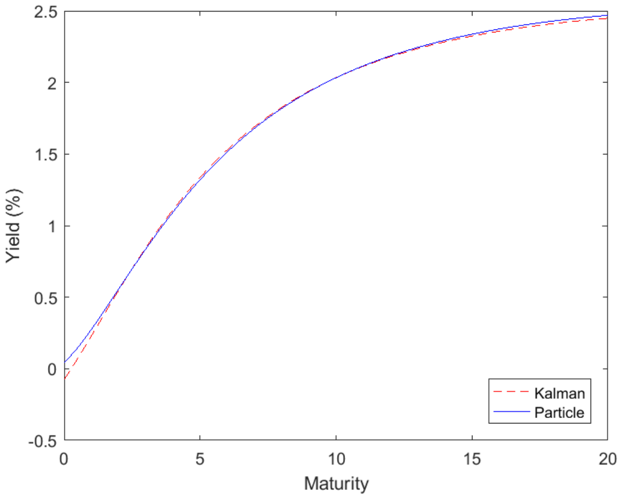

Figure 6 Six-month-ahead yield curve forecast

The solid line represents the yield curve predicted by the Rao-Blackwellised particle filter with the ZLB constraint, and the dashed line corresponds to the Kalman filter prediction.

We further investigate out-of-sample forecast accuracy by an expanding-sample exercise. The first sample for state filtering is January 1982 to January 2012, and we forecast the yield curve one year ahead. The second sample is January 1982 to February 2012, and so on. The RMSE of the yield curve forecast is reported in Table 3. The particle filter with ZLB has higher predictive power for short-term bonds. For example, to forecast the yields of 6 months to maturity, the RMSE of the particle filter is 19.02, while that of the Kalman filter is 20.74. For maturities longer than two years, results are similar between the Kalman and particle filters.

Maturity	0.25	0.5	1	2	3	5	7	10	20
Kalman	20.17	20.74	19.40	15.78	19.13	34.40	45.11	58.28	71.45
Particle	19.19	19.02	18.54	15.70	19.17	34.25	45.05	58.52	72.31

Table 3 One-year-ahead yield curve forecast RMSE (basis point)

Forecast RMSE are compared between the unconstrained Kalman filter and the Rao-Blackwellised particle filter with the ZLB constraint. The yield curve maturities range from 1/4 to 20 years.

10. Conclusion

When a state space model is subject to inequality constraints, analytic tractability offered by the Kalman filter is lost. We developed a Monte Carlo method that enforces the constraints for all outcomes of the filtering distributions. Our method is based on the particle filter, but we exploit the Gaussian linear sub-structure for analytic integration whenever possible. The major features of our particle filter include the optimal importance function, cross-sectional and temporal Rao-Blackwellisation. The optimal importance function is a single-period analytic filter. Cross-sectional Rao-Blackwellisation marginalizes the unconstrained state components by the Kalman filter. Temporal Rao-Blackwellisation skips particle filtering in the unconstrained subsample and bridges particles of two disjoint periods by the augmented Kalman filter.

There are several directions for model extension and future research.

First, generalization towards nonlinear systems. Suppose that Eq (1) is replaced by $p(x_t|x_{1:t-1}, y_{1:t-1}) = \frac{\phi(x_t; f(x_{t-1}), Q_t)}{F(f(x_{t-1}), Q_t, \mathcal{X}_t)} \cdot 1(x_t \in \mathcal{X}_t)$, where $f(\cdot)$ is a nonlinear function. The particle filter is still applicable by replacing the $A_t x_{t-1}$ term by $f(x_{t-1})$ in Proposition 1. In addition, cross-sectional Rao-Blackwellisation given in Proposition 2 may be generalized if $f(\cdot)$ is partly linear conditional on some constrained and nonlinear state components.

Second, initialisation of the filter. Section 2 discusses initial states that are either deterministic or properly distributed. Initialisation for the diffuse states can be an interesting extension. In the particle filtering framework, we want initial particles generated from $p(x_0, x_1|y_1)$, assuming it is a proper distribution (that is, an observation makes the posterior state distribution proper). Note that $p(x_0, x_1|y_1) \propto p(x_0) \cdot p(x_1|x_0) \cdot p(y_1|x_1)$, where $p(x_0)$ could be non-informative (proportional to one), $p(x_1|x_0)$ and $p(y_1|x_1)$ are given by Eq (1) and (2). We may resort to the Metropolis sampler to generate initial draws, since the density can be evaluated pointwise up to a proportionality constant. Furthermore, the posterior state distribution under PCSSM has an analytic form, if $x_0 \sim TN(0, \kappa I_m, \mathcal{X}_0)$, $\kappa \to \infty$. Proposition 5 suggests that $p(x_0, x_1|y_1)$ is multivariate normal with truncation. The normal mean and variance can be obtained by the diffuse Kalman filter (De Jong, 1991) or Koopman and Durbin (2003), or simply by the "large-kappa" Kalman filter/smoother. Then we may generate initial particles by rejection sampling. For $t \geq 2$, the particle filtering procedures are the same as those proposed in Proposition 1 and 2.

Third, enhancement on automatic Rao-Blackwellisation. Section 6 shows real-time subsample selection for temporal Rao-Blackwellisation. It would be interesting to study automatic subset selection for cross-sectional Rao-Blackwellisation. The expected bias criterion might still be useful for identifying the state components subject to tight inequality constraints.

SUPPLEMENTARY MATERIAL

MATLAB Toolbox for ICSSM/PCSSM: software for inequality constrained particle filtering. The package also contains all datasets and code used in Example 1 - 5 and Section 8 - 9.

Proofs: proof of Proposition 1 - 7.

Temporal and Cross-sectional Rao-Blackwellisation: exposition of temporal Rao-Blackwellisation as a special case of cross-sectional Rao-Blackwellisation.

Filtering and Estimating AFNS model with ZLB constraint: a more detailed exposition of the methodology for estimating the AFNS model discussed in Section 9.

References

Black, F. (1995). Interest rates as options. *The Journal of Finance 50* (5), 1371-1376.

Carter, C. K. and R. Kohn (1994). On Gibbs sampling for state space models. *Biometrika 81* (3), 541-553.

Chen, R. and J. S. Liu (2000). Mixture Kalman filters. *Journal of the Royal Statistical Society: Series B (Statistical Methodology) 62* (3), 493-508.

Chopin, N. (2004). Central limit theorem for sequential Monte Carlo methods and its application to Bayesian inference. *The Annals of Statistics 32* (6), 2385-2411.

Christensen, J. H., F. X. Diebold, and G. D. Rudebusch (2011). The affine arbitrage-free class of Nelson-Siegel term structure models. *Journal of Econometrics 164* (1), 4-20.

Christensen, J. H. and G. D. Rudebusch (2013). Modeling yields at the zero lower bound: Are shadow rates the solution? Working Paper Series 2013-39, Federal Reserve Bank of San Francisco.

Christensen, J. H. E. and G. D. Rudebusch (2015). Estimating shadow-rate term structure models with near-zero yields. *Journal of Financial Econometrics 13* (2), 226-259.

Cogley, T. and T. J. Sargent (2005). Drift and volatilities: Monetary policies and outcomes in the post WWII U.S. *Review of Economic Dynamics 8* (2), 262-302.

Commandeur, J. J. F., S. J. Koopman, and M. Ooms (2011). Statistical software for state space methods. *Journal of Statistical Software 41* (1), 1-18.

De Jong, P. (1991). The diffuse Kalman filter. *The Annals of Statistics 19* (2), 1073-1083.

Diebold, F. X., G. D. Rudebusch, and B. S. Aruoba (2006). The macroeconomy and the yield curve: a dynamic latent factor approach. *Journal of Econometrics 131* (1-2), 309-338.

Doran, H. E. (1992). Constraining Kalman filter and smoothing estimates to satisfy time-varying restrictions. *The Review of Economics and Statistics 74* (3), 568-72.

Doucet, A., S. Godsill, and C. Andrieu (2000). On sequential Monte Carlo sampling methods for Bayesian filtering. *Statistics and Computing 10* (3), 197-208.

Doucet, A., N. Gordon, and V. Krishnamurthy (2001). Particle filters for state estimation of jump Markov linear systems. *Signal Processing, IEEE Transactions on 49* (3), 613-624.

Doucet, A. and A. M. Johansen (2009). *A Tutorial on Particle Filtering and Smoothing: Fifteen Years Later.* Oxford: Oxford University Press.

Durbin, J. and S. J. Koopman (2002). A simple and efficient simulation smoother for state space time series analysis. *Biometrika 89* (3), 603-615.

Durbin, J. and S. J. Koopman (2012). *Time Series Analysis by State Space Methods: Second Edition.* Oxford: Oxford University Press.

Gordon, N., D. Salmond, and A. Smith (1993). Novel approach to nonlinear/non-Gaussian Bayesian state estimation. *Radar and Signal Processing, IEE Proceedings F 140* (2) ,107-113.

Greene, W. H. (2008). *Econometric Analysis, Sixth Edition*. New Jersey: Prentice Hall.

Gupta, N. and R. Hauser (2008). Kalman filtering with equality and inequality state constraints. Manuscript: http://arxiv.org/pdf/0709.2791.pdf.

Hull, J. (2003). *Options, Futures and Other Derivatives, Fifth Edition*. New Jersey: Prentice Hall.

Hull, J. and A. White (1990). Pricing interest-rate-derivative securities. *Review of Financial Studies 3* (4), 573-592.

Koop, G., R. Leon-Gonzalez, and R. W. Strachan (2010). Dynamic probabilities of restrictions in state space models: An application to the Phillips curve. *Journal of Business & Economic Statistics 28* (3), 370-379.

Koop, G. and S. M. Potter (2011). Time varying VARs with inequality restrictions. *Journal of Economic Dynamics and Control 35* (7), 1126-1138.

Koopman, S. J. and J. Durbin (2003). Filtering and smoothing of state vector for diffuse state-space models. *Journal of Time Series Analysis 24* (1), 85-98.

Liu, J. S. and R. Chen (1998). Sequential Monte Carlo methods for dynamic systems. *Journal of the American Statistical Association 93* (443), 1032-1044.

Pizzinga, A. (2012). *Restricted Kalman Filtering, Theory, Methods, and Application*. Springer: New York.

Schon, T., F. Gustafsson, and P. Nordlund (2005). Marginalized particle filters for mixed linear/nonlinear state-space models. *Signal Processing, IEEE Transactions on 53* (7), 2279-2289.

Schwartz, E. S. (1997). The stochastic behavior of commodity prices: Implications for valuation and hedging. *Journal of Finance 52* (3), 923-73.

Simon, D. J. and D. L. Simon (2005). Aircraft turbofan engine health estimation using constrained Kalman filtering. *Journal of Engineering for Gas Turbines and Power 127* (2), 323-328.

Simon, D. J. and D. L. Simon (2010). Constrained Kalman filtering via density function truncation for turbofan engine health estimation. *International Journal of Systems Science 41* (2), 159-171.

Stock, J. H. and M. W. Watson (2007). Why has U.S. inflation become harder to forecast? *Journal of Money, Credit and Banking 39* (s1), 3-33.

Vasicek, O. (1977). An equilibrium characterization of the term structure. *Journal of Financial Economics 5* (2), 177-188.