econometrics

MDPI

Article

The Stochastic Stationary Root Model

Andreas Hetland

Department of Economics, University of Copenhagen, 1353 Copenhagen K, Denmark; lxs601@ku.dk; Tel.: +45-20422606

Received: 31 March 2018; Accepted: 13 August 2018; Published: 21 August 2018

Abstract: We propose and study the stochastic stationary root model. The model resembles the cointegrated VAR model but is novel in that: (i) the stationary relations follow a random coefficient autoregressive process, i.e., exhibhits heavy-tailed dynamics, and (ii) the system is observed with measurement error. Unlike the cointegrated VAR model, estimation and inference for the SSR model is complicated by a lack of closed-form expressions for the likelihood function and its derivatives. To overcome this, we introduce particle filter-based approximations of the log-likelihood function, sample score, and observed Information matrix. These enable us to approximate the ML estimator via stochastic approximation and to conduct inference via the approximated observed Information matrix. We conjecture the asymptotic properties of the ML estimator and conduct a simulation study to investigate the validity of the conjecture. Model diagnostics to assess model fit are considered. Finally, we present an empirical application to the 10-year government bond rates in Germany and Greece during the period from January 1999 to February 2018.

Keywords: cointegration; particle filtering; random coefficient autoregressive model; state space model; stochastic approximation

JEL Classification: C15; C32; C51; C58

1. Introduction

In this paper, we introduce the multivariate stochastic stationary root (SSR) model. The SSR model is a nonlinear state space model, which resembles the Granger-Johansen representation of the cointegrated vector autoregressive (CVAR) model, see *inter alia* Johansen (1996) and Juselius (2007). The SSR model decomposes a p–dimensional observation vector into r stationary components and $p-r$ nonstationary components, which is similar to the CVAR model. However, the roots of the stationary components are allowed to be stochastic; hence the name 'stochastic stationary root'. The stationary and nonstationary dynamics of the model are observed with measurement error, which in this model prohibits close-form expressions for e.g., the log-likelihood, sample score and observed Information matrix. Likelihood-based estimation and inference therefore calls for non-standard methods.

Although the SSR model resembles the CVAR model, it is differentiated by its ability to characterize heavy-tailed dynamics in the stationary component. Heavy-tailed dynamics, and other types of nonlinear dependencies, are not amenable to analysis with the CVAR model, which has prompted work into nonlinear alternatives, see *inter alia* Bohn Nielsen and Rahbek (2014), Kristensen and Rahbek (2013), Kristensen and Rahbek (2010), and Bec et al. (2008). Similarly, cointegration in the state space setting has been considered in term of the common stochastic trend (CST) model by Chang et al. (2009) as well as the CVAR model with measurement errors by Bohn Nielsen (2016). Additionally, the SSR model is also related to the stochastic unit root literature, see *inter alia* Granger and Swanson (1997), Leybourne and McCabe (1996), Lieberman and Phillips (2014), Lieberman and Phillips (2017), McCabe and Tremayne (1995), and McCabe and Smith (1998). Relevant empirical applications where the SSR model could potentially provide a better fit than the CVAR model include, but are not limited to, (i) log-prices of assets

that exhibit random walk behavior in the levels and heavy-tailed error-correcting dynamics in the no-arbitrage relations, and (ii) interest rates for which the riskless rate exhibits random walk-type dynamics and the risk premia undergo periods of high levels and high volatility.

The stationary and nonstationary components of the SSR model are treated as unobserved processes, and consequently need to be integrated out in order to compute the log-likelihood function and its derivatives. Due to the nonlinearity of the model, this cannot be accomplished analytically. We appeal to the incomplete data framework and the simulation-based approach known as particle filtering to approximate the log-likelihood function, sample score and observed Information matrix. See *inter alia* Gordon et al. (1993), Doucet et al. (2001), Cappé et al. (2005), and Creal (2012) for an overview of the particle filtering literature. Moreover, we rely on stochastic approximation methods to obtain the maximum likelihood (ML) estimator, see Poyiadjis et al. (2011). Summarizing, the main contributions of this paper are to

i introduce and study the SSR model, and
ii propose a method for approximate frequentist estimation and inference.

It is beyond the scope of this paper to provide a complete proof of the asymptotic properties of the ML estimator. The study of the asymptotic properties of the ML estimator in general state space models, such as the SSR model, is an emerging area of research. Most existing results rely on compactness of the state space, which excludes the SSR model and is generally restrictive. For results in this direction, see e.g., Olsson and Rydén (2008) who derive consistency and asymptotic normality for the ML estimator by discretizing the parameter space. Douc et al. (2011) have shown consistency of the ML estimator without assuming compactness, but the regularity conditions are nonetheless too restrictive to encompass the SSR model. Instead of providing a complete proof of the asymptotic properties of the ML estimator, we conjecture the asymptotic properties of the derivatives of the log-likelihood function. We base the conjecture on known properties of models that are closely related to the SSR model, and corroborate it by a simulation study. Given the conjecture holds, it allows us to establish the asymptotic properties of the ML estimator. We leave proving the conjecture for future work, and focus in this paper on developing methods for approximate frequentist estimation and inference.

The rest of the paper is organized as follows. We introduce the SSR model in Section 2, and study some properties of the process in Section 3. In Section 4 we introduce likelihood-based estimation and inference for the unknown model parameter. In Section 5 we introduce the incomplete data framework. In Section 6 we introduce the particle filter-based approximations to the log-likelihood function, sample score and Information matrix. In Section 7 we propose how to approximate the ML estimator and classic standard errors. In Section 8 we consider model diagnostics. In Section 9 we conduct a simulation study of the asymptotic distribution of the ML estimator. In Section 10 we apply the SSR model to monthly observations of 10-year government bond rates in Germany and Greece from January 1999 to February 2018. We conclude in Section 11. All proofs have been relegated to Appendix B, while Appendix A contains various auxiliary results.

Notation-wise, we adopt the convention that the 'blackboard bold' typeface, e.g., $\mathbb{E}$, denotes operators, and the 'calligraphy' typeface, e.g., $\mathcal{X}$, denotes sets. We thus let $\mathcal{R}$ and $\mathcal{N}$ denote the real and natural numbers, respectively. For any matrix A, we denote by $|A|$ the determinant, by $\|A\| = \sqrt{\text{tr}(A'A)}$ the Euclidean norm, and by $\rho(A)$ the spectral radius. For some positive definite matrix A, we let $A^{1/2}$ denote the lower triangular Cholesky decomposition. For some function $f : \mathcal{R}^{d_z} \mapsto \mathcal{R}^{d_f}$, let $\partial f(z)/\partial z$ denote the derivative of $f(z)$ with respect to z. For some stochastic variable $z \in \mathcal{R}^{d_z}$ with Gaussian distribution with mean μ and covariance Σ, let $N(z; \mu, \Sigma)$ denote the Gaussian probability density function evaluated at z. We let $p(z)$ denote the probability density of stochastic variable $z \in \mathcal{R}^{d_z}$ with respect to the d_z–dimensional Lebesgue measure m, while $p(\mathrm{d}z) = p(z)\,\mathrm{d}m$ denotes the corresponding probability measure. Additionally, the letter 'p' is generic notation for probability density functions and measures induced by the model defined in

(1)–(3) below. The 'bold' typeface, e.g., $\boldsymbol{p}$, is generic notation for analytically intractable quantities, in the sense of having no closed-form expression. Finally, we denote a sequence of $n \in \mathcal{N}_+$ real d_z–dimensional vectors by $z_{1:n} := [\ z_1' \quad \dots \quad z_n'\ ]' \in \mathcal{R}^{n \times d_z}$.

2. The Model

The structure of the SSR model is similar to the Granger-Johansen representation of the CVAR model, cf. Johansen (1996, chp. 4), but departs from it in two respects. First, the stationary component is a random coefficient autoregressive process, cf. e.g., Feigin and Tweedie (1985), rather than an autoregressive process. Second, the stationary and nonstationary components are observed with measurement error. This makes the SSR model is a state space model, whereas the CVAR model is observation-driven. In addition to resembling the CVAR model, the SSR model constitutes an extension of the CST model, cf. Chang et al. (2009). However, while the CST model is a linear Gaussian state space model, the SSR model is a nonlinear Gaussian state space model as it allows the stationary component to be a random coefficient autoregressive process.

Formally, we consider the observable p-dimensional discrete time vector process y_t, for $t = 1, 2, \dots, T$ given by,

$$y_t = C(y_0) + B \sum_{i=1}^{t} \eta_i + A\xi_t + u_t \tag{1}$$

$$\xi_t = \mu + \Phi_t \xi_{t-1} + \nu_t\,, \tag{2}$$

for fixed initial values y_0 and ξ_0, and with u_t, Φ_t and $[\eta_t', \nu_t']'$ mutually independent. We define $\varepsilon_t := \sum_{i=1}^{t} \eta_i$ with $\varepsilon_0 = 0_{p-r}$. The sequences $\varepsilon_{1:T}$ and $\xi_{1:T}$ are unobserved and take values $\varepsilon_t \in \mathcal{R}^{p-r}$ and $\xi_t \in \mathcal{R}^r$ for $0 < r < p$. Additionally, the matrices are of dimensions $A \in \mathcal{R}^{p \times r}$ and $B \in \mathcal{R}^{p \times p-r}$, with $[A\ B] \in \mathcal{R}^{p \times p}$ and invertible. Let the random coefficient, Φ_t, be i.i.d. Gaussian,

$$\mathrm{vec}(\Phi_t) \sim N(\mathrm{vec}(\Phi),\, \Omega_\Phi)\,. \tag{3}$$

with Ω_Φ a positive definite covariance matrix. Let the observation error be i.i.d. Gaussian, such that $u_t \sim N(0,\, \Omega_u)$ with Ω_u a positive definite matrix, and let the innovations η_t and ν_t be jointly Gaussian such that $\eta_t \sim N(0,\, \Omega_\eta)$ and $\nu_t \sim N(0, \Omega_\nu)$ with cross-covariance $\mathbb{C}\mathrm{ov}\,[\eta_t,\, \nu_t] = \Omega_{\eta,\nu}$, such that the joint covariance matrix,

$$\Lambda := \begin{bmatrix} \Omega_\eta & \Omega_{\eta,\nu} \\ \Omega_{\eta,\nu}' & \Omega_\nu \end{bmatrix}, \tag{4}$$

is positive definite. Let all the introduced matrices be of appropriate dimensions and full rank. Furthermore, we introduce the orthogonal complements to A and B, which we denote $b \in \mathcal{R}^{p \times r}$ and $a \in \mathcal{R}^{p \times p-r}$, such that $b'B = 0$ and $a'A = 0$ with b and a of full column rank. Finally, we let $C(y_0) := B(a'B)^{-1}a'y_0$.

Define the parameter vectors,

$$\omega := \begin{bmatrix} \mathrm{vec}(B)' & \mathrm{vec}(A)' & \mathrm{vech}(\Omega_u)' \end{bmatrix}' \tag{5}$$

$$\lambda := \begin{bmatrix} \mu' & \mathrm{vec}(\Phi)' & \mathrm{vech}(\Omega_\Phi)' & \mathrm{vech}(\Lambda)' \end{bmatrix}', \tag{6}$$

which contain the parameters governing the observations y_t, and unobserved components ε_t and ξ_t, respectively. The parameter vectors take values in $\omega \in \Theta_\omega$ and $\lambda \in \Theta_\lambda$, respectively. Additionally, we define the full parameter vector as

$$\theta := \begin{bmatrix} \omega' & \lambda' \end{bmatrix}' \in \Theta_\omega \times \Theta_\lambda =: \Theta\,. \tag{7}$$

which indexes the model, and we refer to Θ as the parameter space. Note that ω and λ in θ are variation free in the sense of Engle et al. (1983). The parameter space is a subset of the d_θ-dimensional Euclidean space $\Theta \subseteq \mathcal{R}^{d_\theta}$, where d_θ denotes the number of elements in θ. In the case where no restrictions are imposed on θ, the dimension d_θ increases rapidly in r due to the $\frac{1}{2}(r^2+1)r^2$ parameters in Ω_Φ. We suggest restricting the off-diagonal elements of Ω_Φ to zero to avoid over-parameterization. The number of parameters is then $d_\theta = 2p^2 + p + 2r^2 + r$ when the model is otherwise unrestricted.

The log-likelihood function for any parameter vector $\theta \in \Theta$, fixed initial values $y_0 \in \mathcal{R}^p$, $\varepsilon_0 = 0_{p-r}$ and $\xi_0 \in \mathcal{R}^r$, and observation sequence $y_{1:T} \in \mathcal{R}^{p\times T}$ is given by,

$$\ell_T(\theta) := \log p_\theta(\varepsilon_0, \xi_0, y_{0:T}) . \tag{8}$$

The sample score is given by the first derivative of (8),

$$S_T(\theta) := \frac{\partial}{\partial\theta}\ell_T(\theta) , \tag{9}$$

and the observed Information matrix is given by minus the second derivative of (8),

$$I_T(\theta) := -\frac{\partial^2}{\partial\theta\partial\theta'}\ell_T(\theta) . \tag{10}$$

Due to the nonlinear dynamics of the unobserved process (2), the log-likelihood function (8) and its derivatives (9)–(10) do not have closed-form solutions. In the following, we suppress the dependence on the initial values ε_0, ξ_0 and y_0, but note they remain fixed.

3. Properties of the Process

In this section we consider some properties of the process defined by Equations (1)–(3) for a given parameter value $\theta \in \Theta$. Specifically, we study the nonstationary and stationary components, including conditions on the parameter θ that ensure strict stationarity of the stationary component. Additionally, we decompose the observation y_t into nonstationary and stationary directions.

3.1. The Unobserved Components

The first component of the model, ε_t, is a *random walk* (RW) in $p-r$ dimensions, equivalently expressed as an autoregressive process with a unit root. That is, for $t = 1, \ldots, T$,

$$\varepsilon_t = \varepsilon_{t-1} + \eta_t , \tag{11}$$

with $\varepsilon_0 = 0_{p-r}$. The process (11) admits the transition density $p_\lambda(\varepsilon_t \mid \varepsilon_{t-1})$ with respect to the $p-r$–dimensional Lebesgue measure; however, it does not have a stationary distribution. This type of process has been studied extensively, see e.g., Dickey and Fuller (1979). In summary, the RW process is linear and Gaussian, but nonstationary.

The second unobserved component of the model, ξ_t, is a *random coefficient autoregressive* (RCAR) process of lag order one in r dimensions. The RCAR process (2)–(3) is observationally equivalent to a double autoregressive (DAR) process with one lag, cf. Ling (2007), which we formalize in Lemma 1.

Lemma 1. *For $\theta \in \Theta$, the random coefficient autoregressive process (2)–(3) with $k = 1$ has the following double autoregressive process representation, $t = 1, 2, \ldots, T$*

$$\xi_t = \mu + \Phi\xi_{t-1} + \Omega_{\nu,t}^{1/2} z_t \tag{12}$$

$$\Omega_{\nu,t} = \Omega_\nu + \left(\xi'_{t-1} \otimes I_r\right) \Omega_\Phi \left(\xi'_{t-1} \otimes I_r\right)' , \tag{13}$$

for ξ_0 fixed, $z_t \sim N(0,\, I_r)$, cross-covariance $\mathbb{C}\mathrm{ov}\,[\eta_t,\, z_t] = \Omega_{\eta\,\nu}$, and with the joint innovation process $[\eta_t',\, z_t']'$ independent and identically distributed.

The DAR representation in Lemma 1 of the RCAR process in (2)–(3) characterizes the process dynamics in terms of the conditional mean and variance. The conditional mean $\mathbb{E}_\lambda\,[\xi_t \mid \xi_{t-1}]$ is autoregressive. However, the conditional variance $\mathbb{V}\mathrm{ar}_\lambda\,[\xi_t \mid \xi_{t-1}]$ depends positively on the lagged level 'squared'. The conditional variance is heteroskedastic, but not in the well-known ARCH sense of e.g., Engle (1982); rather, the lagged level of the process ξ_{t-1} enters the variance, not the lagged innovation ν_{t-1}. To illustrate the point, we consider for a moment the conditional variance in the univariate case $r = 1$, which is given by $\omega_{\nu,\,t}^2 = \omega_\nu^2 + \omega_\phi^2 \xi_{t-1}^2$. Here we see that a relatively large (in absolute terms) lagged level $|\xi_{t-1}|$ will result in a relatively large volatility $\omega_{\nu,\,t}$ in the present period, and vice versa.

We make the following assumption on the random coefficients (3) in order to ensure strict stationarity of the RCAR process (2)–(3).

Assumption 1. *Assume that the top Lyapunov exponent is strictly negative,*

$$\gamma := \lim_{n\to\infty} \frac{1}{n}\, \mathbb{E}_\lambda \left[\log \left\| \prod_{t=1}^{n} \Phi_t \right\| \right] < 0\,. \tag{14}$$

Remark 1. *The top Lyapunov exponent (14) is intractable but can be approximated to arbitrary precision via simulation, cf. inter alia Ling (2007) and Francq and Zakoian (2010). The following approximation converges almost surely*

$$\hat{\gamma}_n := \frac{1}{n} \log \left\| \prod_{t=1}^{n} \Phi_t \right\| \quad \overset{a.s.}{\to} \quad \gamma\,, \tag{15}$$

as $n \to \infty$. In turn, $\hat{\gamma}_n$ can be computed efficiently via the QR-decomposition, cf. Dieci and Van Vleck (1995).

Assumption 1 ensures that the RCAR process can be characterized as a geometrically ergodic Markov chain, cf. Meyn and Tweedie (2005). This is formalized in the following theorem.

Theorem 1 (Feigin and Tweedie (1985), Theorem 3)**.** *Under Assumption 1, the process $\{\xi_t\}_{t=0,\,1,\,\ldots}$ is geometrically ergodic. In particular, the initial value ξ_0 can be given an initial distribution $\boldsymbol{p}_\theta(\xi_0)$ such that $\{\xi_t\}_{t=0,\,1,\,\ldots}$ is stationary and geometrically ergodic with some fractional moment.*

Remark 2. *The stationary component, ξ_t, exhibits heavy-tailed behavior since it satisfies a stochastic recurrence equation. Pedersen and Wintenberger (2018) have recently considered the tail properties of processes of the form (2) for a more general specification of the random coefficient, Φ_t, that includes BEKK-ARCH and DAR-type processes as special cases. It should be possible to show that the stationary distribution of ξ_t as defined in (2)–(3) also has power-law tails under suitable conditions.*

The RCAR process (2)–(3) admits the transition density $p_\lambda(\xi_t \mid \xi_{t-1})$ with respect to the r-dimensional Lebesgue measure. Moreover, the process has the stationary distribution $\boldsymbol{p}_\theta(\xi_t)$ under Assumption 1. In summary, the RCAR process is Gaussian and strictly stationary, but nonlinear.

3.2. The Observed Process

The observations $\{y_t\}_{t=1,\,2,\,\ldots}$ are conditionally independent given the sequence of unobserved components $\{\varepsilon_t,\, \xi_t\}_{t=1,\,2,\,\ldots}$. Thus, the dynamics of the observed process are determined by the dynamics of the unobserved components.

We use the orthogonal complements b' and a' of the loading matrices B and A, respectively, and the skew-projection identity of Johansen (1996) to decompose the observation vector y_t as follows,

$$y_t = B_a a' y_t + A_b b' y_t \,, \tag{16}$$

where we define $B_a := B(a'B)^{-1}$ and $A_b := A(b'A)^{-1}$. Here $a'B$ and $b'A$ are invertible thanks to our assumption that $[A \ B]$ is square and invertible. By premultiplying y_t by a' we eliminate the stationary directions, while leaving the nonstationary directions,

$$a' y_t = a' C(y_0) + a' B \varepsilon_t + a' u_t \,. \tag{17}$$

What is left after the linear transformation (17) is a random walk with Gaussian measurement error. Similarly, premultiplying y_t by b' eliminates the nonstationary directions while the stationary directions remain,

$$b' y_t = b' A \xi_t + b' u_t \,. \tag{18}$$

The process given by (18) is a stationary random coefficient autoregressive process with Gaussian measurement error.

The decomposition of the observation process (16) allows for a cointegration interpretation of the SSR model. The p observed variables in y_t share $p - r$ common stochastic trends (17) with loading matrix B_a, while the r linear combinations (18) are stationary and load into the levels with the matrix A_b. The observed process admits the conditional density $\boldsymbol{p}_\theta(y_t \mid y_{1:t-1})$ with respect to the p–dimensional Lebesgue measure; however, this density does not have a closed-form expression. Moreover, the observed process does not have a stationary distribution.

4. Likelihood-Based Estimation and Inference

In this section, we introduce the ML estimator and consider its asymptotic properties. We wish to conduct estimation and inference based on the true, but intractable, model likelihood. Due to the intractability of the likelihood, we can neither compute the ML estimator via numerical optimization of (8), nor compute classic standard errors via the observed Information matrix (10). We refer to the ML estimator as being 'doubly intractable', with reference to the concept from the literature in Bayesian statistics on models with intractable likelihoods, see e.g., Murray et al. (2006). It is beyond the scope of this paper to derive a full asymptotic theory for the SSR model. Instead, we conjecture the limiting properties of the likelihood function (8) and its derivatives (9)–(10). We obtain the asymptotic properties for the ML estimator based on the conjecture.

We recall preliminarily that the ML estimator is defined as the parameter vector $\theta \in \Theta$ that maximizes the log-likelihood function (8),

$$\hat{\boldsymbol{\theta}}_T := \underset{\theta \in \Theta}{\arg\sup} \, \boldsymbol{\ell}_T(\theta) \,, \tag{19}$$

noting that the ML estimator (19) is a function of the observation sequence $y_{1:T}$. We denote by $\theta^* \in \Theta$ the true parameter value for the data generating process (1)–(3). In the following, we make the below conjecture on the asymptotic properties of (8)–(10). Note that, having assumed that B^* is known, the score, information, and likelihood in the conjecture refer to the unknown parameters only; that is, all elements in θ excluding $\mathrm{vec}(B)$.

Conjecture 1. *If Assumption 1 holds, B^* is known, and $\theta^* \in \Theta \subseteq \mathcal{R}^{d_\theta}$, then the log-likelihood function $\boldsymbol{\ell}_T(\cdot) : \mathcal{R}^{d_\theta} \mapsto \mathcal{R}$ is three times continuously differentiable in θ, and*

1. $\frac{1}{\sqrt{T}} \boldsymbol{S}_T(\theta^*) \xrightarrow{D} N(0, \Omega_S)$ *as* $T \to \infty$, *with* $\Omega_S > 0$,

2. $\frac{1}{T}\boldsymbol{I}_T(\theta^*) \xrightarrow{P} \Omega_I$ *as* $T \to \infty$, *with* $\Omega_I > 0$, *and*
3. $\max_{h,i,j=1,\ldots,d_\theta} \sup_{\theta \in \mathcal{N}(\theta^*)} |\partial^3 \ell_T(\theta)/\partial\theta_h\partial\theta_i\partial\theta_j| \leq c_T$,

where $\mathcal{N}(\theta^*)$ *is a neighborhood of* θ^* *and* $0 \leq c_T \xrightarrow{P} c$, $0 < c < \infty$, *as* $T \to \infty$.

Remark 3. *Theorem 3 in Bohn Nielsen and Rahbek (2014) shows that Conjecture 1 holds in the case of the strictly stationary bivariate double autoregressive model with BEKK-type time-varying covariance. With* B^* *known, the SSR model corresponds closely to this model plus Gaussian measurement errors.*

It should be noted that we propose Conjecture 1 despite lack of finite moments of the RCAR process, cf. Theorem 1. This is in line with the results of *inter alia* Bohn Nielsen and Rahbek (2014) for the bivariate DAR model, and Ling (2004, 2007) for the univariate DAR model.

The result in Theorem 2 below states that if Conjecture 1 holds true, then the ML estimator (19) is unique, $\sqrt{T}$–consistent and asymptotically Gaussian. The result follows from applying Lemma 1 in Jensen and Rahbek (2004), the conditions of which correspond to (1.)–(3.) of Conjecture 1.

Theorem 2 (Jensen and Rahbek (2004), Lemma 1)**.** *If Conjecture 1 holds, then there exists a fixed open neighborhood* $\mathcal{U}(\theta^*) \subseteq \mathcal{N}(\theta^*)$ *of the true parameter* θ^*, *which is an interior point of* Θ, *such that with probability tending to one as* $T \to \infty$, *there exists a minimum point* $\hat{\boldsymbol{\theta}}_T$ *in* $\mathcal{U}(\theta^*)$ *and* $\boldsymbol{\ell}_T(\theta)$ *is convex in* $\mathcal{U}(\theta^*)$. *In particular,* $\hat{\boldsymbol{\theta}}_T$ *is unique and satisfies the score equation*

$$\boldsymbol{S}_T(\hat{\boldsymbol{\theta}}_T) = 0\,. \tag{20}$$

Additionally, the ML estimator is consistent $\hat{\boldsymbol{\theta}}_T \to \theta^*$, *and asymptotically Gaussian,*

$$\sqrt{T}(\hat{\boldsymbol{\theta}}_T - \theta^*) \quad \xrightarrow{D} \quad N(0, \Omega_I^{-1}\Omega_S\Omega_I^{-1})\,, \quad T \to \infty\,. \tag{21}$$

Proof. Conjecture 1 satisfies the Cramer-type conditions of Lemma 1 in Jensen and Rahbek (2004), which provides the result. □

We assume that the true value of B is known, because Chang et al. (2009) showed that the ML estimator of the loading matrix B exhibits T-convergence and is asymptotically mixed Gaussian in the CST model. The CST model corresponds to the SSR model with $p - r = 1$, but without the stationary components, i.e., $A = 0_{p\times r}$ for any p. We find it reasonable to believe that this result carries over to the SSR model. Moreover, fixing B is conceptually similar to classic cointegration analysis with known cointegrating vectors, which is an accepted starting point for new methodological developments, see e.g., Bec and Rahbek (2004). In applications we often have a predefined set of cointegrating vectors that we are interested in. In the context of the SSR model, the cointegrating vectors correspond to the rows of the orthogonal complement b'. As an example, for the empirical illustration in Section 10 we consider an interest rate spread in a bivariate system with one common stochastic trend, i.e., $p = 2$ and $p - r = 1$. The spread implies $b' = [\ 1 \quad -1\ ]$, which in turn corresponds to the loading matrix $B = [\ 1 \quad 1\ ]'$ when normalizing on the first element.

The Fisher Information matrix, Ω_I, is consistently estimated by the (scaled) observed Information matrix evaluated at $\hat{\boldsymbol{\theta}}_T$, cf. Conjecture 1.(3.). Moreover, the asymptotic variance of the score, Ω_S, is equal to the Fisher Information matrix when the model is well-specified; the information matrix equality holds, cf. e.g., Hamilton (1994, sct. 14.4). In this case, the asymptotic variance of the ML estimator (19) is simply the inverse Fisher Information matrix. Thus, we can use classic standard errors, that are based on the observed Information matrix (10), to conduct inference on the ML estimates.

5. The Incomplete Data Framework

In this section, we appeal to the incomplete data framework of Dempster et al. (1977) to deal with the unobserved components of the SSR model. We first formulate the state space representation of the model in (1)–(3) and its associated optimal filtering problem. Secondly, we formulate the intractable sample score (9) and observed information matrix (10) in terms of the *optimal filtering problem*. In Section 6 we introduce a particle filter algorithm with which we can approximate the optimal filtering problem. This enables approximation of the intractable sample score and observed information matrix via the particle filter algorithm.

5.1. The State Space Form and the Optimal Filtering Problem

Preliminarily, we collect the unobserved components in the vector $x_t := [\ \varepsilon_t' \quad \xi_t'\ ]'$, which we refer to as the *state vector*. The unobserved components are Markov, see (11)–(13), and the observation depends only on the contemporary values of the unobserved components. Thus, the SSR model in (1)–(3) has the dependency structure of a state space model. Formally, for $t = 1, \ldots, T$, the SSR model in (1)–(3) has the following state space representation,

$$y_t = C(y_0) + \Pi x_t + \Omega_u^{1/2} u_t \tag{22}$$

$$x_t = \alpha + \Gamma x_{t-1} + \Lambda_t^{1/2} v_t\,, \tag{23}$$

with y_0 and x_0 fixed, $u_t \sim N(0, I_p)$ and $v_t \sim N(0, I_p)$, and u_t and v_t mutually independent. We define accordingly,

$$\Pi := \begin{bmatrix} B' \\ A' \end{bmatrix}', \quad \alpha := \begin{bmatrix} 0 \\ \mu \end{bmatrix}, \quad \Gamma := \begin{bmatrix} I_{p-r} & 0 \\ 0 & \Phi \end{bmatrix} \quad \text{and} \quad \Lambda_t := \begin{bmatrix} \Omega_\eta & \Omega_{\eta,\nu} \\ \Omega_{\eta,\nu}' & \Omega_{\nu,t} \end{bmatrix}, \tag{24}$$

and recall that $\Omega_{\nu,t}$ is defined in Lemma (1). We refer to (22) as the *observation equation*, and to (23) as the *transition equation*. It is easy to verify that the state space representation in (22) and (23) is observationally equivalent to the SSR model as presented in (1)–(3). The observation and transition equations admit the densities with respect to the p-dimensional Lebesgue measure,

$$p_\omega\left(y_t \mid x_t\right) = N(y_t;\, C(y_0) + \Pi x_t,\, \Omega_u) \tag{25}$$

$$p_\lambda\left(x_t \mid x_{t-1}\right) = N(x_t;\, \alpha + \Gamma x_{t-1},\, \Lambda_t)\,, \tag{26}$$

respectively. We refer to (25) as the *observation density* and to (26) as the *transition density*. As mentioned previously, we suppress the dependence on the initial observation y_0.

One approach to conducting inference on the unobserved components, i.e., the state vector x_t, is the optimal filtering problem, cf. Anderson and Moore (1979). The optimal filtering problem refers to the general problem of computing the conditional expectation of some sequence of unobserved states given some sequence of observations. In the following, we consider the specific instance of the optimal filtering problem known as the *smoothing problem*. Formally, the smoothing problem is a conditional expectation of the form,

$$\mathbb{E}_\theta\left[\gamma_t(x_{1:t}) \mid y_{1:t}\right] = \int \gamma_t(x_{1:t}) \boldsymbol{p}_\theta\left(x_{1:t} \mid y_{1:t}\right) \mathrm{d}x_{1:t}\,, \tag{27}$$

for any function $\gamma_t(x_{1:t}) \in L^1\left[\mathcal{R}^{t\,p},\, \boldsymbol{p}_\theta\left(x_{1:t} \mid y_{1:t}\right)\right]$ and point in time $t \in \{1, \ldots, T\}$. We refer to the function $\gamma_t(x_{1:t})$ as the *test function* and to the density $\boldsymbol{p}_\theta\left(x_{1:t} \mid y_{1:t}\right)$ as the *smoothing density*. The test function may be time-varying, but of known form for a fixed observation sequence $y_{1:T}$. The smoothing density in (27) can be expressed as the recursion of the lagged smoothing density,

$$\boldsymbol{p}_\theta\left(x_{1:t} \mid y_{1:t}\right) = \frac{p_\omega\left(y_t \mid x_t\right) p_\lambda\left(x_t \mid x_{t-1}\right)}{\boldsymbol{p}_\theta\left(y_t \mid y_{1:t-1}\right)} \boldsymbol{p}_\theta\left(x_{1:t-1} \mid y_{1:t-1}\right), \tag{28}$$

initialized with $\boldsymbol{p}_\theta\left(x_1 \mid x_0, y_0, y_1\right)$. The normalizing constant in (28) is the likelihood contribution, which is given by the integral,

$$\boldsymbol{p}_\theta\left(y_t \mid y_{1:t-1}\right) = \int p_\omega\left(y_t \mid x_t\right) p_\lambda\left(x_t \mid x_{t-1}\right) \boldsymbol{p}_\theta\left(x_{1:t-1} \mid y_{1:t-1}\right) \mathrm{d}x_{1:t}\,. \tag{29}$$

We note the smoothing density recursion (28) is intractable due to the intractability of the likelihood contribution (29). In the following, we will use the smoothing problem (27) to address computation of the sample score (9) and observed Information matrix (10).

5.2. The Sample Score and Observed Information as Smoothing Problems

The incomplete data framework is closely associated with the classic expectation maximization (EM) algorithm, introduced in Dempster et al. (1977). The EM algorithm is a common approach to maximizing the log-likelihood function (8) to obtain the ML estimator (19) for models with unobserved variables. When the EM algorithm is applicable, it is also possible to evaluate the sample score (9) and observed Information matrix (10). For the SSR model, however, the EM algorithm does not apply directly, yet we may use the incomplete data framework to reformulate the sample score and observed Information in terms of intractable smoothing problems of the form (27).

A central concept of the EM algorithm is the auxiliary function called the *intermediate quantity*, which is defined as,

$$\begin{aligned} \boldsymbol{Q}_T\left(\theta \mid \vartheta\right) &:= \int \log p_\theta\left(y_{1:T}, x_{1:T}\right) \boldsymbol{p}_\vartheta\left(x_{1:T} \mid y_{1:T}\right) \mathrm{d}x_{1:T} \\ &= \boldsymbol{\ell}_T(\theta) - \boldsymbol{H}_T(\theta \mid \vartheta)\,, \end{aligned} \tag{30}$$

where

$$\boldsymbol{H}_T(\theta \mid \vartheta) := -\int \log p_\theta(x_{1:T} \mid y_{1:T}) \boldsymbol{p}_\vartheta(x_{1:T} \mid y_{1:T}) \,\mathrm{d}x_{1:T}\,, \tag{31}$$

for any parameter values $\theta, \vartheta \in \Theta$. We refer to $\log p_\theta\left(y_{1:T}, x_{1:T}\right)$ as the *complete data log-likelihood*. By the state space model structure (22)–(23) and variation freeness of θ defined in (7), we have that the complete data log-likelihood is given by,

$$\log p_\theta\left(y_{1:T}, x_{1:T}\right) = \sum_{t=1}^{T} \left[\log p_\omega\left(y_t \mid x_t\right) + \log p_\lambda\left(x_t \mid x_{t-1}\right)\right]. \tag{32}$$

The intermediate quantity (30) is sometimes also called the expected log-likelihood, since it is interpretable as the conditional expectation of the complete data log-likelihood (32) given the observations $y_{1:T}$. We note the term separating the log-likelihood (8) and the intermediate quantity (30) is the entropy of the smoothing density (28) with parameters ϑ and θ, defined in (31).

We are interested in the intermediate quantity (30) because it provides a convenient way to derive the sample score and observed Information matrix in terms of the derivatives of the complete data log-likelihood (32). The first and second derivatives of the complete data log-likelihood function in (32) are the sum of the first and second order derivatives of the observation and transition log-densities

with respect to ω and λ, respectively. These can be computed by either analytical or numerical differentiation of (32). For $\vartheta \in \Theta$, we define the derivatives of (32) in terms of the functions,

$$U_T\left(x_{1:T};\, \vartheta\right) \coloneqq \frac{\partial}{\partial \theta} \log p_\theta\left(y_{1:T},\, x_{1:T}\right)\big|_{\theta=\vartheta} = \sum_{t=1}^{T} u_t\left(x_t,\, x_{t-1};\, \vartheta\right) \tag{33}$$

$$V_T\left(x_{1:T};\, \vartheta\right) \coloneqq \frac{\partial^2}{\partial \theta \partial \theta'} \log p_\theta\left(y_{1:T},\, x_{1:T}\right)\big|_{\theta=\vartheta} = \sum_{t=1}^{T} v_t\left(x_t,\, x_{t-1};\, \vartheta\right), \tag{34}$$

where, taking advantage of the variation freeness of the model parameter, θ, we define the summands of (33) and (34), respectively, as

$$u_t\left(x_t,\, x_{t-1};\, \vartheta\right) \coloneqq \left[\begin{array}{c} \frac{\partial}{\partial \omega} \log p_\omega\left(y_t \mid x_t\right) \\ \frac{\partial}{\partial \lambda} \log p_\lambda\left(x_t \mid x_{t-1}\right) \end{array}\right]\Bigg|_{\theta=\vartheta}, \tag{35}$$

and

$$v_t\left(x_t,\, x_{t-1};\, \vartheta\right) \coloneqq \left[\begin{array}{cc} \frac{\partial^2}{\partial \omega \partial \omega'} \log p_\omega\left(y_t \mid x_t\right) & 0_{d_\omega \times d_\lambda} \\ 0_{d_\lambda \times d_\omega} & \frac{\partial^2}{\partial \lambda \partial \lambda'} \log p_\lambda\left(x_t \mid x_{t-1}\right) \end{array}\right]\Bigg|_{\theta=\vartheta}, \tag{36}$$

We note that the functions (35) and (36) should not be confused with the measurement error in (22) and innovations in (23), respectively.

If the first and second order derivatives of the complete data log-likelihood in (33) and (34), respectively, are integrable with respect to the smoothing density (28), then we may appeal to Fisher's and Louis' identities (defined below) to express the sample score (9) and observed Information matrix (10) in terms of smoothing problems of the form (27).

Conjecture 2. *For any $\theta \in \Theta$ and observation sequence $y_{1:T} \in \mathcal{R}^{p \times T}$, it holds that $U_T(x_{1:T};\, \theta) \in L^2[\mathcal{R}^{p \times T},\, \boldsymbol{p}_\theta(x_{1:T} \mid y_{1:T})]$ and $V_T(x_{1:T};\, \theta) \in L^1[\mathcal{R}^{p \times T},\, \boldsymbol{p}_\theta(x_{1:T} \mid y_{1:T})]$.*

For the same reasons we conjectured the asymptotic properties of the true log-likelihood function, sample score, observed information matrix, we conjecture integrability of the derivatives of the complete data log-likelihood (33) and (34).

Fisher's identity, cf. Dempster et al. (1977), states the first derivative of the intermediate quantity (30) is equivalent to the sample score (9). Similarly, Louis' identity of Louis (1982) establishes a relation between the first and second derivatives of the intermediate quantity (30) and the observed Information matrix (10).

Lemma 2 (Fisher's and Louis' identities, cf. Cappé et al. (2005), Proposition 10.1.6)**.** *If Conjecture 2 holds and $\theta \in \Theta$, then the sample score (9) is equivalently given by*

$$\boldsymbol{S}_T\left(\theta\right) = \int U_T(x_{1:T};\, \theta) \boldsymbol{p}_\theta\left(x_{1:T} \mid y_{1:T}\right)\, \mathrm{d}x_{1:T}, \tag{37}$$

and the observed Information (10) is equivalently given by

$$\boldsymbol{I}_T\left(\theta\right) = \boldsymbol{S}_T\left(\theta\right) \boldsymbol{S}_T\left(\theta\right)' - \boldsymbol{G}_T(\theta) - \boldsymbol{K}_T(\theta), \tag{38}$$

where

$$\boldsymbol{G}_T(\theta) \coloneqq \int V_T(x_{1:T};\, \theta) \boldsymbol{p}_\theta\left(x_{1:T} \mid y_{1:T}\right)\, \mathrm{d}x_{1:T} \tag{39}$$

$$\boldsymbol{K}_T(\theta) \coloneqq \int U_T(x_{1:T};\, \theta) U_T(x_{1:T};\, \theta)' \boldsymbol{p}_\theta\left(x_{1:T} \mid y_{1:T}\right)\, \mathrm{d}x_{1:T}, \tag{40}$$

and the functions $U_T(x_{1:T};\theta)$ and $V_T(x_{1:T};\theta)$ are defined in (33) and (34), respectively.

Although Lemma 2 shows the sample score (9) and observed Information (10) can be restated as smoothing problems of the form (27), we still cannot obtain closed-form expressions due to the intractability of the optimal filtering problem, cf. Section 5.1. In the next section, we introduce a particle filter algorithm that can approximate smoothing problems for appropriately chosen test functions, such as the functions $U_T(x_{1:T};\theta)$ and $V_T(x_{1:T};\theta)$ under Conjecture 2.

6. Particle Filter-Based Approximations

In this section, we introduce a particle filter algorithm that produces pointwise approximations to the true but intractable log-likelihood function (8), sample score (9), and observed Information matrix (10) for any parameter $\theta \in \Theta$ and fixed observation sequence $y_{1:T} \in \mathcal{R}^{p\times T}$. In Section 7, we show how to apply the particle filter-based approximations introduced in this section to approximate the true, intractable ML estimator and classic standard errors, which we introduced in Section 4.

6.1. Particle Filtering

A particle filter is a simulation-based algorithm that produces approximations to smoothing problems of the form (27) for state space models. We introduce here a standard particle filter, which produces empirical measures that recursively approximate the smoothing density (28) for each time point in the observed sample $t \in \{1,\ldots,T\}$. The empirical measures consist of point masses, which we refer to as *particles*, and we use these for Monte Carlo integration in order to approximate the smoothing problem (27). Additionally, the particle filter produces a point-wise approximation of the log-likelihood function as a by-product. For an introduction to particle filtering in the context of economics and finance see Creal (2012).

The particle filter algorithm relies on an *importance density*, denoted $q_\theta(x_{1:t} \mid y_{1:t})$, that has the same support and recursive structure as the smoothing density (28). Formally, for $t = 1,\ldots,T$, we define the importance density as,

$$q_\theta(x_{1:t} \mid y_{1:t}) \coloneqq q_\theta(x_t \mid x_{t-1}, y_t)\, q_\theta(x_{1:t-1} \mid y_{1:t-1}) , \tag{41}$$

initialized by $q_\theta(x_1 \mid x_0, y_0, y_1)$. We note the importance density (41) is defined recursively by $q_\theta(x_t \mid x_{t-1}, y_t)$, which we refer to as the *importance transition density*.

Assuming the smoothing density (28) is absolutely continuous with respect to the importance density (41), we can write the former as a the product of the importance density and a weight function,

$$\boldsymbol{p}_\theta(x_{1:t} \mid y_{1:t}) = \bar{\boldsymbol{w}}_t(x_{1:t})\, q_\theta(x_{1:t} \mid y_{1:t}) , \quad \bar{\boldsymbol{w}}_t(x_{1:t}) \coloneqq \frac{\boldsymbol{p}_\theta(x_{1:t} \mid y_{1:t})}{q_\theta(x_{1:t} \mid y_{1:t})} . \tag{42}$$

We refer to the weight function $\bar{\boldsymbol{w}}_t(x_{1:t})$ as the *normalized importance weight*. We note that (42) constitutes a change of measure from the smoothing density to the importance density, and the normalized importance weight is a Radon-Nikodym derivative between the two densities.

Substituting the recursive expressions for the smoothing density (28) and importance density (41) into the expression for the normalized importance weight in (42), we obtain a recursive expression for the normalized importance weight,

$$\bar{\boldsymbol{w}}_t(x_{1:t}) = \frac{\tilde{w}_t(x_{t-1:t})}{\boldsymbol{p}_\theta(y_t \mid y_{1:t-1})} \bar{\boldsymbol{w}}_{t-1}(x_{1:t-1}) , \tag{43}$$

where we define

$$\tilde{w}_t(x_{t-1:t}) \coloneqq \frac{p_\omega(y_t \mid x_t)\, p_\lambda(x_t \mid x_{t-1})}{q_\theta(x_t \mid x_{t-1}, y_t)} . \tag{44}$$

We refer to (44) as the *incremental importance weights*. The recursion for the normalized importance weight (43) is normalized by the likelihood contribution (29) and is therefore also intractable.

For particle filtering in general, the importance transition density is subject to choice under mild regularity conditions, cf. e.g., Assumption 9.4.1 in Cappé et al. (2005). We let the importance transition density be the corresponding model density; formally,

$$q_\theta\left(x_t \mid x_{t-1}, y_t\right) \coloneqq p_\theta\left(x_t \mid x_{t-1}, y_t\right) . \tag{45}$$

We refer to (45) as the *locally optimal transition density*. This choice of importance transition density is optimal in the sense that it is conditional on the the contemporary observation y_t, cf. Doucet et al. (2000). This is sometimes also referred to as 'fully adapted', cf. e.g., Pitt and Shephard (1999b). If we instead let the importance transition density be the model transition density (26), we omit the information about x_t that is contained in y_t. The locally optimal transition density is not necessarily available in closed-form for nonlinear state space models. It is, however, available for the SSR model and we present it in Lemma 3.

Lemma 3. *For $\theta \in \Theta$, the locally optimal transition density has the closed-form expression*

$$p_\theta\left(x_t \mid x_{t-1}, y_t\right) = N(x_t; \mu^x_{t|t}, \Sigma^x_{t|t}) , \tag{46}$$

where the conditional mean and variance are given by,

$$\mu^x_{t|t} = \mu^x_{t|t-1} + \Sigma^x_{t|t-1}\Pi' \left[\Sigma^y_{t|t-1}\right]^{-1} \left(y_t - \mu^y_{t|t-1}\right) \tag{47}$$

$$\Sigma^x_{t|t} = \Sigma^x_{t|t-1} - \Sigma^x_{t|t-1}\Pi' \left[\Sigma^y_{t|t-1}\right]^{-1} \Pi\Sigma^x_{t|t-1} , \tag{48}$$

with

$$\mu^y_{t|t-1} = C(y_0) + \Pi\mu^x_{t|t-1} \tag{49}$$

$$\Sigma^y_{t|t-1} = \Pi\Sigma^x_{t|t-1}\Pi' + \Omega_u \tag{50}$$

$$\mu^x_{t|t-1} = \alpha + \Gamma x_{t-1} \tag{51}$$

$$\Sigma^x_{t|t-1} = \Lambda_t , \tag{52}$$

and the state space form definitions given in (24).

Remark 4. *The locally optimal transition density (46) is related to the Kalman (1960) filter, which solves the optimal filtering problem analytically for linear and Gausian models. Equations (49)–(52) correspond the Kalman filter for a known value of x_{t-1}. Related methods for efficient particle filtering include the mixture Kalman filter and Rao-Blackwellisation, cf. Chen and Liu (2000) and Andrieu and Doucet (2002).*

It is straightforward to use the general expression for the incremental importance weight in (44) to show that letting the importance transition density be the locally optimal transition density, i.e., (45), results in the following specific expression for incremental importance weights,

$$\tilde{w}_t(x_{t-1}) = p_\theta\left(y_t \mid x_{t-1}\right) . \tag{53}$$

We refer to the density in (53) as the *predictive observation density*. It has a closed-form expression that follows from the closed-form expression of the locally optimal transition density in Lemma 3.

Corollary 1. *For $\theta \in \Theta$, the predictive observation density has the closed-form expression*

$$p_\theta(y_t \mid x_{t-1}) = N(y_t;\, \mu^y_{t|t-1},\, \Sigma^y_{t|t-1}), \tag{54}$$

recalling the definitions in (49)–(52).

Proof. Contained in the proof of Lemma 3. □

Remark 5. *The choice of importance transition density (45) is locally optimal in the sense that the conditional variance of the incremental importance weights (53) given x_{t-1} is zero, cf. Doucet et al. (2000).*

The particle filter, presented in Algorithm 1 below, produces weighted particle samples approximately distributed as the smoothing density (28) at each point in time $t = 1, \ldots, T$. The algorithm consists of iterating over three steps. At point t in time, the first step is to sample N particles, denoted $\{\tilde{x}^{(i)}_{1:t}\}_{i=1}^N$, from the importance density (41) given the particle sample from $t-1$. This is called the *propagation step*. Step two consists of computing self-normalized importance weights, denoted $\{\bar{w}^{(i)}_t\}_{i=1}^N$, that approximate the normalized importance weights (43). This is the *weighting step*. The third step is to sample N particle indices, denoted $\{I^{(i)}\}_{i=1}^N$, with replacement. We sample index j with probability $\bar{w}^{(j)}_t$ for $j \in \{1, \ldots, N\}$. We retain the number of particles indicated by the resulting sample of particle indices, denoted $\{x^{(i)}_{1:t}\}_{i=1}^N$, and let the importance weights be uniform. This is the *resampling step*. After resampling, we store the particle samples and proceed to $t+1$.

For a fixed parameter value $\theta \in \Theta$ and observation sequence $y_{1:T} \in \mathcal{R}^{p \times T}$, we run the locally optimal particle filter for the SSR model as specified in Algorithm 1 below.

Algorithm 1: Locally Optimal Particle Filter.

Given a parameter $\theta \in \Theta$, initialize by setting $x^{(i)}_0 := x_0$ and $\bar{w}^{(i)}_0 := 1/N$ for $i = 1, \ldots, N$. For $t = 0, 1, \ldots, T$:

1. Sample particles $\{\tilde{x}^{(i)}_t\}_{i=1}^N$ with distribution

$$\tilde{x}^{(i)}_t \sim p_\theta(x_t \mid x^{(i)}_{t-1}, y_t), \tag{55}$$

and set $\tilde{x}^{(i)}_{1:t} := [\; x^{(i)}_{1:t-1} \quad \tilde{x}^{(i)}_t \;]$ for $i = 1, 2, \ldots, N$.

2. Calculate the unnormalized importance weights, $\{w^{(i)}_t\}_{i=1}^N$,

$$w^{(i)}_t = p_\theta(y_t \mid \tilde{x}^{(i)}_{t-1,t}) \bar{w}^{(i)}_{t-1}, \tag{56}$$

for $i = 1, \ldots, N$. Then compute the normalized importance weights

$$\bar{w}^{(i)}_t = \frac{w^{(i)}_t}{W^N_t}, \quad W^N_t := \sum_{i=1}^N w^{(i)}_t, \tag{57}$$

for $i = 1, \ldots, N$.

3. Sample N particle indices $\{I^{(i)}\}_{i=1}^N$, $I^{(i)} \in \{1, \ldots, N\}$, with probabilities

$$\Pr(I^{(i)} = j \mid \tilde{\mathcal{F}}_t, y_{1:t}) = \bar{w}^{(j)}_t, \quad j \in \{1, \ldots, N\} \tag{58}$$

for $i = 1, \ldots, N$. Set the resampled particles $x^{(i)}_{1:t} := \tilde{x}^{(I^{(i)})}_{1:t}$, and the normalized importance weights $\bar{w}^{(i)}_t := 1/N$ for $i = 1, \ldots, N$.

Remark 6. *The resampling method applied in step (3.) of Algorithm 1 is known as* multinomial resampling. *Alternative methods that are guaranteed to produce lower Monte Carlo variance exists, cf. Douc et al. (2005). We consider multinomial resampling for its analytical tractability, and recommend applying one of the more efficient alternatives in practice.*

Remark 7. *The notation $x_{1:t}^{(i)}$ is ambiguous due to the resampling step of Algorithm 1, since the elements of the ith particle path at time $t-1$, denoted $x_{1:t-1}^{(i)}$, are not necessarily the same as the first $t-1$ elements of the ith particle path at time t, denoted $x_{1:t}^{(i)}$. By convention, $x_{1:t}^{(i)}$ always refers to the particle chain after resampling at time t (similarly $\tilde{x}_{1:t}^{(i)}$ refers to the chain before resampling). We refer to elements k to l of the ith particle chain after resampling at time t as $x_{l:k,t}^{(i)}$.*

The particle filter in Algorithm 1 produces two particle samples at each point in time, t. The first set, $\{\tilde{x}_{1:t}^{(i)}\}_{i=1}^{N}$, is produced at the propagation step (1.) and is associated with importance weights in the weighting step (2.), $\{\bar{w}_t^{(i)}\}_{i=1}^{N}$. The second set, $\{x_{1:t}^{(i)}\}_{i=1}^{N}$, is produced at the resampling step (3.). Both sets are approximately drawn from the smoothing density (28). We note the resampling step introduces additional sampling error, cf. Chopin (2004), so we calculate approximations using the weighted sample unless otherwise specified.

The particle filter iterates over over the propagation, weighting and resampling steps throughout the sequence, $t = 1, \ldots, T$, after which the algorithm terminates. We note the two sets of particles produced during each iteration are themselves random variables measurable with respect to the sub-σ-algebras $\tilde{\mathcal{F}}_t$ and $\mathcal{F}_t$, defined next.

Definition 1. *Define the sub-σ-algebras $\tilde{\mathcal{F}}_t := \mathcal{F}_{t-1} \cup \sigma(\tilde{x}_t^{(1)}, \ldots, \tilde{x}_t^{(N)})$, $\mathcal{F}_t := \tilde{\mathcal{F}}_t \cup \sigma(x_t^{(1)}, \ldots, x_t^{(N)})$ for $t = 1, \ldots, T$, initialized by $\mathcal{F}_0 := \emptyset$.*

At each point in time, we associate an empirical measure with the weighted particle sample generated by the propagation (1.) and reweighting (2.) steps in Algorithm 1. Formally, for $t = 1, 2, \ldots, T$, we define the empirical measure,

$$\tilde{p}_\theta^N(\mathrm{d}x_{1:t} \mid y_{1:t}) := \sum_{i=1}^{N} \bar{w}_t^{(i)} \delta_{\tilde{x}_{1:t}^{(i)}}(\mathrm{d}x_{1:t}), \tag{59}$$

where $\delta_{x'}(\mathrm{d}x)$ denotes the point measure at $x' \in \mathcal{R}^p$ with respect to $\mathrm{d}x$. The weighted particles that constitute the empirical measure (59) are approximately distributed according to the smoothing density (28). We emphasize the weighted particles are not independent draws from (28), because the resampling step introduces dependence between the particles at each iteration of the algorithm. We use the empirical measure (59) to define a particle filter-based approximation of the intractable smoothing problem in (27),

$$\tilde{\mathbb{E}}_\theta^N[\gamma_t(x_{1:t}) \mid y_{1:t}] := \int \gamma_t(x_{1:t}) \tilde{p}_\theta^N(\mathrm{d}x_{1:t} \mid y_{1:t}) = \sum_{i=1}^{N} \bar{w}_t^{(i)} \gamma_t(\tilde{x}_{1:t}^{(i)}), \tag{60}$$

for any point in time $t \in \{1, \ldots, T\}$. Due to dependence between the weighted particles, we cannot establish the asymptotic properties of the approximation (60) based on the law of large numbers and central limit theorem for independent random variables. For appropriately chosen test functions $\gamma_t(x_{1:t})$, the approximation (60) is both consistent and asymptotically Gaussian as the number of particles tends to infinity, $N \to \infty$, cf. Theorem 9.4.5 in Cappé et al. (2005).

The particle filter in Algorithm 1 also produces an approximation of the log-likelihood function (8) evaluated at the parameter value θ and the observation sequence $y_{1:T}$,

$$\tilde{\ell}_T^N(\theta) := \sum_{t=1}^{T} \log W_t^N \,. \tag{61}$$

We note that the approximate log-likelihood function (61) consists of the logarithm of the product of normalizing constants produced by Algorithm 1. The approximate log-likelihood (61) is consistent in the sense that it converges in probability to the true log-likelihood function, as the number of particles tends to infinity, see Lemma 4.

Lemma 4. *For the model (1)–(3) and $\theta \in \Theta$, the approximate log-likelihood function (61) produced by Algorithm 1 is a consistent estimator of the true log-likelihood (8),*

$$\tilde{\ell}_T^N(\theta) \quad \xrightarrow{P} \quad \ell_T(\theta) \,, \tag{62}$$

as $N \to \infty$.

In addition to producing an approximation of the intractable log-likelihood function (8), we apply the approximation (60) of the intractable smoothing problem in (27) to produce approximations of the sample score and observed Information matrix via Fisher's and Louis' identities in Lemma 2.

6.2. The Approximate Sample Score and Observed Information Matrix

We showed in Section 5 that the sample score and observed Information matrix can be expressed in terms of smoothing problems of the form (27). Appealing to Fisher's identity (37) in Lemma 2, and to the approximation of the smoothing problem (60), we define the particle filter-based approximate sample score as,

$$\tilde{S}_T^N(\theta) := \sum_{i=1}^{N} U_T(\tilde{x}_{1:T}^{(i)}; \theta) \bar{w}_T^{(i)} \,, \tag{63}$$

for any parameter $\theta \in \Theta$, with the function $U_T(x_{1:T}; \theta)$ as defined in (33). If Conjecture 2 holds, then the approximate sample score in (63) is both consistent and asymptotically normal.

Lemma 5. *If Conjecture 2 holds and $\theta \in \Theta$, then the approximate sample score (63) is asymptotically normal,*

$$\sqrt{N}\left\{\tilde{S}_T^N(\theta) - S_T(\theta)\right\} \quad \xrightarrow{D} \quad N(0, \tilde{\mathbb{S}}_T[U_T(x_{1:T}; \theta)]) \,, \tag{64}$$

as $N \to \infty$. An intractable expression for the asymptotic covariance matrix $\tilde{\mathbb{S}}_T[U_T(x_{1:T}; \theta)]$ is given in Lemma A.5 by setting $t = T$ and $\gamma_T(x_{1:T}) = U_T(x_{1:T}; \theta)$.

Similarly, by appealing to Louis' identity (38) in Lemma 2, and to the approximation of the smoothing problem (60), we define the particle filter-based approximate observed Information matrix as,

$$\tilde{I}_T^N(\theta) := \tilde{S}_T^N(\theta)\tilde{S}_T^N(\theta)' - \tilde{G}_T^N(\theta) - \tilde{K}_T^N(\theta) \,, \tag{65}$$

for any parameter $\theta \in \Theta$, where we define the approximations to (39) and (40) as

$$\tilde{G}_T^N(\theta) := \sum_{i=1}^{N} V_T(\tilde{x}_{1:T}^{(i)}; \theta)\bar{w}_T^{(i)} \tag{66}$$

$$\tilde{K}_T^N(\theta) := \sum_{i=1}^{N} U_T(\tilde{x}_{1:T}^{(i)}; \theta)U_T(\tilde{x}_{1:T}^{(i)}; \theta)'\bar{w}_T^{(i)}\,, \tag{67}$$

and the functions $U_T(x_{1:T}; \theta)$ and $V_T(x_{1:T}; \theta)$ are defined in (33) and (34), respectively. If Conjecture 2 holds, then the approximate observed Information in (65) is consistent, stated in the following lemma.

Lemma 6. *If Conjecture 2 holds and $\theta \in \Theta$, then the approximate observed Information matrix (65) is consistent,*

$$\tilde{I}_T^N(\theta) \quad \xrightarrow{P} \quad \boldsymbol{I}_T(\theta) \tag{68}$$

as $N \to \infty$.

Both the approximate sample score (63) and observed Information matrix (65) are biased for finite N. This is a general issue related to the particle filter-based approximation of the smoothing problem (60). At each iteration, the particle filter in Algorithm 1 relies on an approximation of the normalized constant, i.e., likelihood contribution. This induces a finite-sample bias in (60) that gradually disappears as the number of particles N tends to infinity and is negligible for large enough N, cf. e.g., Robert and Casella (2010, sct. 3.3.2).

The particle filter-based approximation of the sample score (63) and observed Information matrix (65) correspond to a batch version of Algorithm A in Poyiadjis et al. (2011), which is of computational cost $O(N)$, but exhibits quadratically increasing variance of the approximate sample score as a function of the sample size T. We note that Poyiadjis et al. (2011) also suggest an alternative algorithm, that exhibits linearly increasing variance as a function of T, but at the computational cost $O(N^2)$. For smaller sample sizes, such as monthly observations as usually encountered in economics, we have found that the $O(N)$ algorithm is adequate.

7. Particle Filter-Based Estimation and Inference

In this section, we show how the approximate sample score (63) and observed Information matrix (65) can be used to perform parameter estimation and inference. We apply a stochastic approximation method based on the approximate sample score to approximate the ML estimator (19). This has recently been suggested in Poyiadjis et al. (2011). We then use the approximate observed Information matrix to obtain approximate standard errors for the approximate ML estimates. Although these quantities are 'approximate', we note that they can be made arbitrarily precise by increasing the number of particles, N, at the expense of increased computational effort.

Recall from Section 4 that the ML estimator (19) is doubly intractable. Consequently, we cannot apply gradient-based optimization algorithms to maximize the log-likelihood function (8). Originally proposed in Robbins and Monro (1951), stochastic approximation methods are conceptually similar to gradient-based optimization methods, but rely on noisy rather than exact evaluations of the sample score to optimize the objective function. The basic idea is that appropriately decreasing the step sizes provides an averaging of the random errors induced by the noisy evaluations of the sample score. For a book-length treatment of stochastic approximation, we refer to Kushner and Yin (2003).

The stochastic approximation algorithm proposed in Poyiadjis et al. (2011, sct. 3.1) consists of a recursion that is conceptually similar to the steepest descent method, cf. e.g., Nocedal and Wright (2006, chp. 3). Prior to executing the algorithm, we choose a fixed initial parameter value $\theta_0 \in \Theta$, a sequence of particle counts $\{N_j\}_{j=1}^{\infty}$, a sequence of step sizes $\{\gamma_j\}_{j=1}^{\infty}$,

and a sequence of weight matrices $\{B_j\}_{j=1}^{\infty}$. The particle counts must be monotonically increasing positive integers, the step sizes must be strictly positive, non-summable but square summable,

$$\sum_{j=1}^{\infty} \gamma_j = \infty \quad \text{and} \quad \sum_{j=1}^{\infty} \gamma_i^2 < \infty\,, \tag{69}$$

and the weight matrices must be positive definite. Having chosen the initial parameter, particle counts, step sizes, and weight matrices, we run the recursion,

$$\theta_{j+1} = \theta_j + \gamma_j B_j \tilde{S}_T^{N_j}(\theta_j)\,, \tag{70}$$

for $j = 0, 1, \ldots, K$. Here K has to be sufficiently large in the sense that the sequence of parameter values generated by the recursion (70) has stabilized in a neighborhood of the true ML estimate. Additionally, if the particle count N_j is large enough, the approximation error affecting the stochastic approximation recursion (70) will be approximately normal, cf. Lemma 5. In this case large disturbances will be rare, such that the parameter sequence $\{\theta_j\}_{j=1}^{K}$ is likely to stabilize without exhibiting large jumps.

We denote by $\{\tilde{x}_{1:T,j}^{(i)}, \bar{w}_{t,j}^{(i)}\}_{i=1}^{N_j}$ the particle paths produced by the particle filter in Algorithm 1 at iteration j of the stochastic approximation recursion (70). The iteration index j is notationally identical to time index of the particle path, cf. Remark 7. Although this is abuse of notation, it is clear from the context whether we refer to the parameter iteration or particle path time index. The parameter θ_{j+1} produced by iteration j of (70) is a random variable that is measurable with respect to the sub-σ-algebra $\mathcal{G}_j$, defined next.

Definition 2. *Let $\mathcal{F}_{T,j} := \sigma(x_{1:T,j}^{(1)}, \ldots, x_{1:T,j}^{(N_j)})$ denote the sub-σ-algebra in Definition 1 generated with the parameter value θ_j, and define the sub-σ-algebras $\mathcal{G}_j := \mathcal{G}_{j-1} \cup \mathcal{F}_{T,j}$ for $j = 1, \ldots$, initialized by $\mathcal{G}_0 := \mathcal{F}_{T,0}$.*

One of the main benefits of the stochastic approximation method is that the method is known to stabilize for a wide variety of initial values, sample counts, step sizes, and weight matrices. In practice, all of these choices affect the number of iterations needed to bring the parameter sequence into the neighborhood of the true ML estimator. The choice of step sizes is particularly important, since large step sizes generally speed up the convergence, but fail to dampen the approximation-induced noise. Small step sizes reduce the noise, but cause slow convergence. The particle count has a similar effect, since a low number of particles will result in a computationally cheap but noisy approximation of the sample score, while a large number of particles reduces the noise but increases the computational cost. Heuristically, it is appropriate to use a combination of large step sizes and small particle counts until the parameter sequence has reached a neighborhood of the ML estimator, and then switch to a combination of smaller step sizes and larger particle counts to reduce the noise. The intuition is that, while far away from the ML estimator, a relatively noisy approximation of the sample score will still on average lead the algorithm in the right direction.

The presence of noise in the sample score is not an impediment when applying stochastic approximation, since the use of decreasing step sizes provides an averaging of the errors. However, the finite sample bias of the particle filter-based approximate sample score, cf. Section 6.2, poses a problem since its effect is not mitigated by decreasing the step sizes. Bias reduction is possible by increasing the particle count N_j together with the iteration number j.

The stochastic approximation method is presented in Algorithm 2 below.[1]

[1] We use the Choleski factorization to ensure positive definiteness of the covariance matrices Ω_u, Λ and Ω_Φ. Thus, we estimate the parameters B, A, $\Omega_u = C_u C_u'$, μ, Φ, $\Omega_\Phi = C_\Phi C_\Phi'$ and $\Lambda = C_\Lambda C_\Lambda'$ using Algorithm 2 and transform the covariances to the original parametrization. We obtain standard errors via the δ-method.

Algorithm 2: Stochastic Approximation.

Choose the initial parameter $\theta_0 \in \Theta$, the particle counts $\{N_j\}_{j=1}^{\infty}$, the step sizes $\{\gamma_j\}_{j=1}^{\infty}$ and weighting matrices $\{B_j\}_{j=1}^{\infty}$. For $j = 0, 1, \ldots, K$:

1. Run Algorithm 1 for θ_j to generate N_j weighted particle paths, denoted $\{x_{1:T,j}^{(i)}, \bar{w}_{t,j}^{(i)}\}_{i=1}^{N_j}$.
2. Compute the approximate sample score (63), denoted

$$\tilde{S}_T^{N_j}(\theta_j) = \sum_{i=1}^{N_j} U_T(x_{1:T,j}^{(i)}; \theta_j)\bar{w}_{t,j}^{(i)} . \tag{71}$$

3. With step size γ_j, ascend along the direction B_j,

$$\theta_{j+1} = \theta_j + \gamma_j B_j \tilde{S}_T^{N_j}(\theta_j) . \tag{72}$$

Polyak (1990) and Polyak and Juditsky (1992) showed that if the step sizes $\{\gamma_j\}_{j=1}^{\infty}$ satisfy the summability conditions (69) and tend to zero slower than j^{-1}, then the average of the last $j - K_0$ iterations converges at an optimal rate. Here $K_0 < K$ denotes the iteration number at which the averaging begins; implicitly, we discard the initial K_0 iterations. We define the approximate ML estimator as,

$$\tilde{\theta}_T := \frac{1}{K - K_0} \sum_{j=K_0}^{K} \theta_j , \tag{73}$$

suppressing the dependence on the particle count. Establishing convergence of the approximate ML estimator (73) to the true ML estimator (19) is outside the scope of this paper. However, if (73) converges in probability to (19) for any fixed T, then (73) inherits the consistency property, cf. Theorem 2, of the true ML estimator.

Convergence of the particle filter-based stochastic approximation method proposed in Poyiadjis et al. (2011) has, to the author's knowledge, not been studied yet. The finite-sample bias of the approximate sample score (63) presents the primary obstacle to establishing convergence results. Intuition suggests that increasing the number of particles N_j with the iteration number j solves the problem. However, convergence of such schemes has not been carefully established, cf. Douc et al. (2014, sct. 12.1.2). Poyiadjis et al. (2011) report stabilization of the particle filter-based stochastic approximation method with constant particle count. In Section 10, we report similar stabilization with increasing particle counts.

If the model is correctly specified, we would conduct inference on the ML estimator via the observed Information matrix, cf. Section 4. Analogously, since the approximate observed Information matrix (65) converges in probability to the true observed Information matrix (10), we can conduct inference for the approximate ML estimator (73) via the approximate observed Information matrix (65), the same way we would conduct inference given the true observed Information matrix (10).

8. Model Diagnostics

In this section, we introduce a method to conduct model diagnostics, such that we may assess whether the SSR model is well-specified for a given parameter θ and observation sequence $y_{1:T}$. Recall that the disturbances u_t, η_t and ν_t are normally distributed and serially independent with mean zero and unit variances. Because the components ε_t and ξ_t are hidden to us, we cannot directly compute the residuals corresponding to the disturbances. Instead, we introduce the normalized one-step prediction errors, cf. Durbin and Koopman (2012, sct. 2.12), that can be approximated via particle filtering.

This approach to model diagnostics for state space models has also previously been considered in Pitt and Shephard (1999a).

We define the normalized one-step prediction errors as,

$$\boldsymbol{e}_t := \mathbb{V}\mathrm{ar}_\theta \left[y_t \mid y_{1:t-1} \right]^{-1/2} \left(y_t - \mathbb{E}_\theta \left[y_t \mid y_{1:t-1} \right] \right) , \tag{74}$$

for $t = 1, \ldots, T$. For a well-specified model, the sequence of normalized one-step prediction errors should be serially independent with mean zero with unit variance. Any deviation from these characteristics are indicative of model misspecification.

The conditional mean and variance in (74) can be stated in terms of smoothing problems, where the test functions are the conditional mean and variance of the predictive observation density,

$$\mathbb{E}_\theta \left[y_t \mid y_{1:t-1} \right] = \mathbb{E}_\theta \left[\mathbb{E}_\theta \left[y_t \mid x_{t-1} \right] \mid y_{1:t-1} \right] , \tag{75}$$

$$\mathbb{V}\mathrm{ar}_\theta \left[y_t \mid y_{1:t-1} \right] = \mathbb{E}_\theta \left[\mathbb{V}\mathrm{ar}_\theta \left[y_t \mid x_{t-1} \right] \mid y_{1:t-1} \right] + \mathbb{V}\mathrm{ar}_\theta \left[\mathbb{E}_\theta \left[y_t \mid x_{t-1} \right] \mid y_{1:t-1} \right] . \tag{76}$$

We note that the conditional mean and variance of the predictive observation density are given in Lemma 3. Using the locally optimal particle filter in Algorithm 1, we define approximations to (75) and (76) as

$$\tilde{\mathbb{E}}_\theta^N \left[y_t \mid y_{1:t-1} \right] := \sum_{i=1}^N \tilde{\mu}_{t|t-1}^{y,(i)} \bar{w}_{t-1}^{(i)} \tag{77}$$

$$\widetilde{\mathbb{V}\mathrm{ar}}_\theta^N \left[y_t \mid y_{1:t-1} \right] := \sum_{i=1}^N \tilde{\Sigma}_{t|t-1}^{y,(i)} \bar{w}_{t-1}^{(i)} + \sum_{i=1}^N (\tilde{\mu}_{t|t-1}^{y,(i)})(\tilde{\mu}_{t|t-1}^{y,(i)})' \bar{w}_{t-1}^{(i)} - \tilde{\mathbb{E}}_\theta^N \left[y_t \mid y_{1:t-1} \right] \tilde{\mathbb{E}}_\theta^N \left[y_t \mid y_{1:t-1} \right]' , \tag{78}$$

respectively, where we have defined the conditional moments given each individual particle as,

$$\tilde{\mu}_{t|t-1}^{y,(i)} := \mathbb{E}_\theta [y_t \mid \tilde{x}_{t-1}^{(i)}] \tag{79}$$

$$\tilde{\Sigma}_{t|t-1}^{y,(i)} := \mathbb{V}\mathrm{ar}_\theta [y_t \mid \tilde{x}_{t-1}^{(i)}] , \tag{80}$$

for $i = 1, \ldots, N$. Finally, we use the approximations (77) and (78) to define the approximate normalized likelihood contributions as follows,

$$\tilde{e}_t^N := \widetilde{\mathbb{V}\mathrm{ar}}_\theta^N \left[y_t \mid y_{1:t-1} \right]^{-1/2} \left(y_t - \tilde{\mathbb{E}}_\theta^N \left[y_t \mid y_{1:t-1} \right] \right) , \tag{81}$$

for $t = 1, \ldots, T$. Thus, by applying the particle filter in Algorithm 1, we obtain the sequence of approximate normalized one-step prediction errors $\tilde{e}_{1:T}$ via (77)–(81). For N sufficiently large, we can use the sequence $\tilde{e}_{1:T}$ to test whether the true sequence of normalized one-step prediction errors $\boldsymbol{e}_{1:T}$ is serially independent with mean zero and unit variance. For common tests for serial dependence and ARCH effects see e.g., Doornik and Hendry (2013, sct. 11.9.2–3).

9. Simulation Study

In this section, we conduct a simulation study of the asymptotic properties of the ML estimator, stated in Theorem 2. We limit our treatment to B, A, Φ and Ω_Φ, leaving aside the remaining parameters Ω_u, μ, Ω_η, Ω_ν and $\Omega_{\eta,\nu}$. Recall, the loading matrix for the stationary components A is conjectured to be asymptotically normal, while the loading matrix of the nonstationary components B is kept fixed. Due to the results of Chang et al. (2009), we expect the asymptotic distribution of B to be mixed normal, and we tentatively investigate this. Moreover, we consider the case where Φ_t is a stochastic unit root. A deterministic unit root is associated with the Dickey-Fuller distribution, cf. Dickey and Fuller (1979), while a stochastic unit root has been shown to be asymptotically normal, see e.g., Ling (2007) and Bohn Nielsen and Rahbek (2014).

Recall, Theorem 2 is based on the conjectured properties of the true, intractable log-likelihood function and its derivatives, cf. Conjecture 1. The aim is to substantiate this conjecture by obtaining the distribution of the approximate ML estimator based on simulated data sets. Usually, the number of realizations in a simulation study of this type is in excess of 1000 and the sample length in excess of 2500 observations. Due to the computational intensity of the particle filter-based stochastic approximation method in Algorithm 2, we limit ourselves to 250 realizations and 500 observations.

We let each of the simulated data sets be a bivariate $p = 2$ series of length $T = 500$ observations with $r = 1$ stationary component and $p - r = 1$ nonstationary component. We use the parameter

$$B = \begin{bmatrix} 1 \\ 1 \end{bmatrix}, \quad A = \begin{bmatrix} 0 \\ 1 \end{bmatrix}, \quad \Omega_u = \begin{bmatrix} 2.5^2 & 0 \\ 0 & 2.5^2 \end{bmatrix}, \tag{82}$$

$$\mu = 0\,, \quad \phi = 1\,, \quad \omega_\phi^2 = 0.25^2\,, \quad \omega_\eta^2 = 15^2\,, \quad \omega_{\eta,\nu} = 0\,, \quad \text{and} \quad \omega_\nu^2 = 2.5^2\,, \tag{83}$$

to generate the simulated data sets. We note the parameter values (83) result in a top Lyapunov coefficient of $\gamma_n = -0.035$, computed via (15) with $n = 10^6$, such that the RCAR process $\{\xi_t\}_{t=0,1,\ldots}$ is strictly stationary.

Having simulated 250 series with the data generating process given by (1)–(3) and (83), we apply Algorithm 2 with $K = 600$ iterations to obtain the approximate ML estimate for the parameter in question, e.g., ϕ, keeping all other parameters fixed at the true values in (83). We initialize the algorithm at the true parameter value, and initiate Polyak averaging at iteration $K_0 = 100$.[2] Moreover, we let the particle count increase as

$$N_j = 50 + \lfloor 1/20j \rfloor\,, \tag{84}$$

where $\lfloor \cdot \rfloor$ denotes the largest integer that is smaller than the argument. We let the step size sequence to decrease as

$$\gamma_j = 100(j + 500)^{-2/3}\,, \tag{85}$$

and set the weight matrix to

$$B_j = T^{-1}\text{diag}\left(\begin{bmatrix} 10^{-5} & 1 & 1 & 1 & 1 & 10^{-2} & 1 & 1 & 1 & 10^{-3} \end{bmatrix}\right), \tag{86}$$

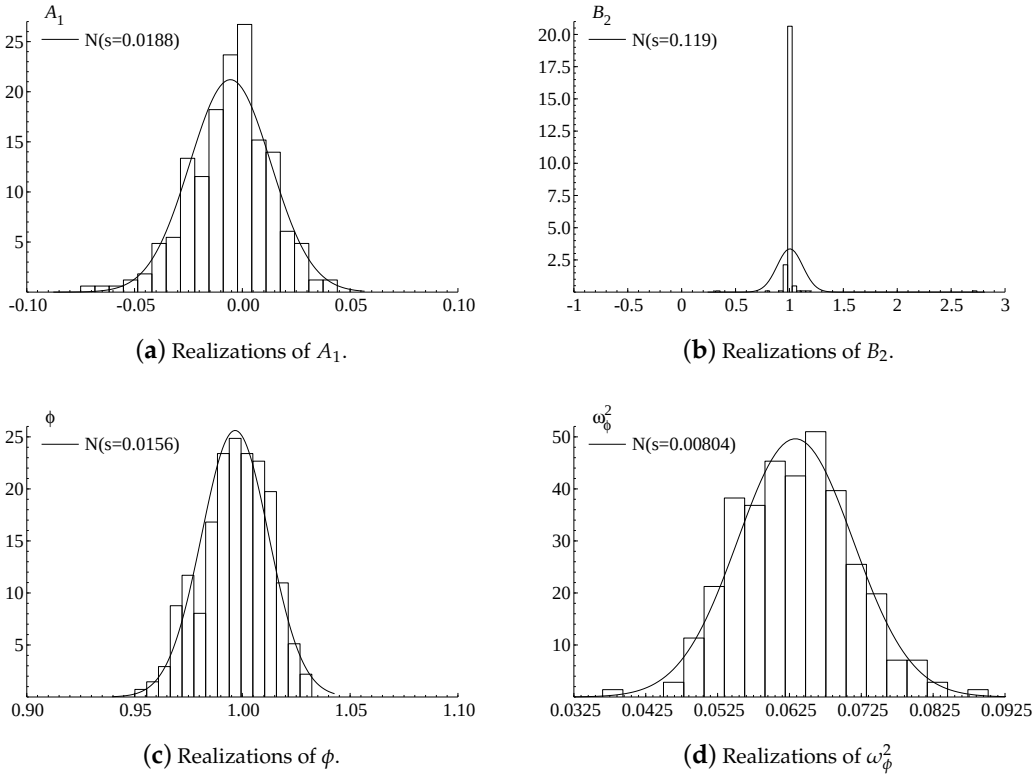
for $j = 1, 2, \ldots, K$. Note the particle count (84) tends to infinity as $j \to \infty$, eliminating the finite-sample bias of (63)–(65), the step sizes satisfy (69), and the weight matrix is constant.[3]

The results from the simulation experiment are presented in Figure 1. Despite the relatively low number of realizations and observations, Figure 1 is instructive of the asymptotic distributions of A_1, ϕ and ω_ϕ^2, cf. Panels (a), (c) and (d). These all appear to be normal. Recall, Theorem 2 does not state the asymptotic distribution of the ML estimator for B_2, and from Panel (b) it does not appear to be normal. Rather, the realizations in Panel (b) are consistent with mixed normality, as we would expect from the closely-related CST model, cf. Chang et al. (2009). To investigate further, one could to simulate the t-ratios of B_2, which should be standard normal. This involves the approximation of the observed Information matrix for each realization, which further increases the computational cost. For this reason, and because we consider B fixed, we do not pursue this further here.

---

[2] Because we initialize at the true parameter value, the parameter sequences stabilize within the first 100 realizations. Using $K = 600$ iterations is sufficient to reduce the impact of the approximation error.

[3] The choice of weight matrix is based on hand-tuning the convergence speed of Algorithm 2 by running a small number of trial-and-error runs with $N = 50$ particles and constant step size $\gamma = 1$.



(**a**) Realizations of A_1. (**b**) Realizations of B_2.

(**c**) Realizations of ϕ. (**d**) Realizations of ω_ϕ^2

Figure 1. Simulation study with 250 realizations of the approximate MLE for A_1, B_2, ϕ, and ω_ϕ^2.

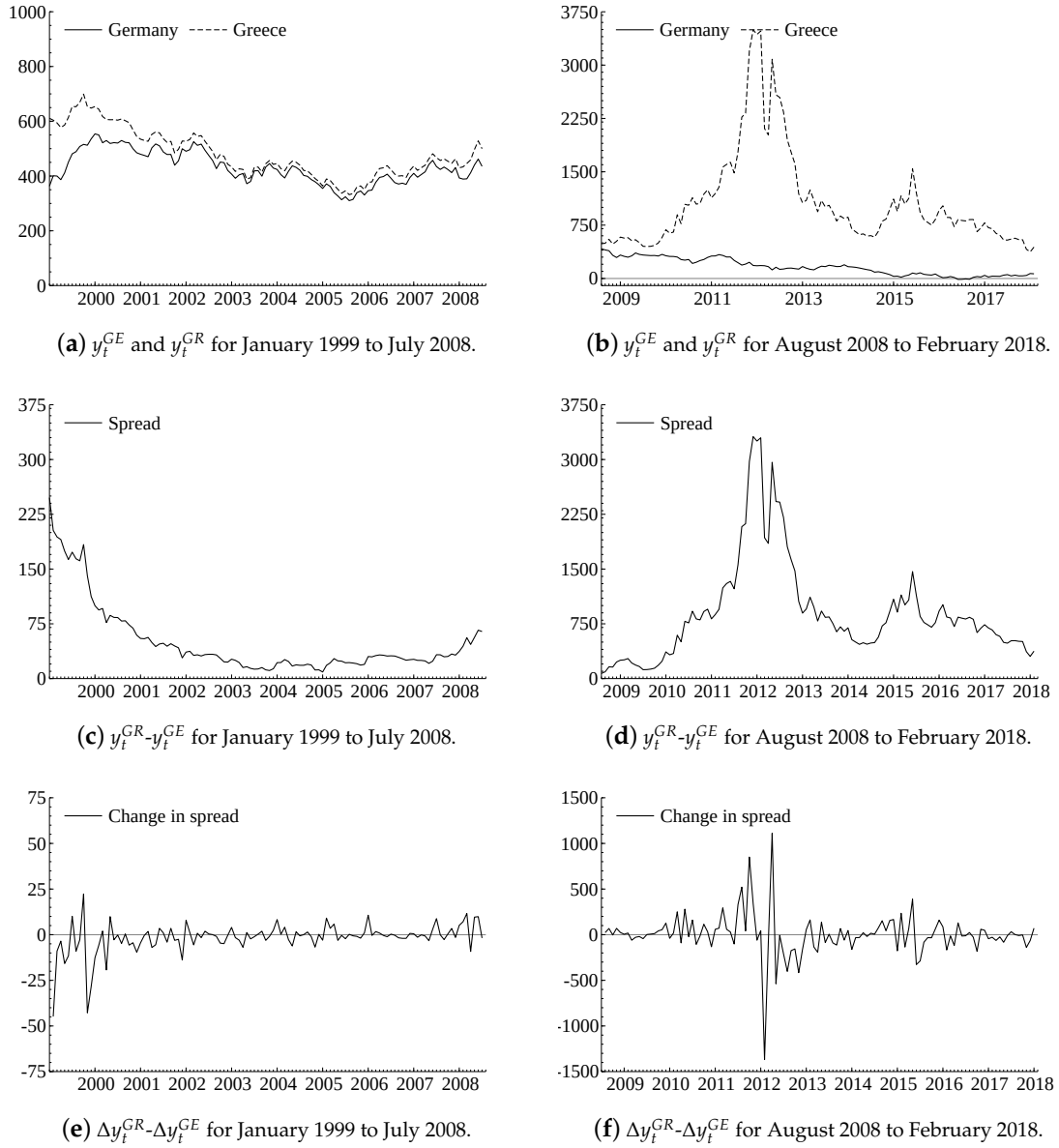
In summary, the findings of the simulation study tentatively support the conjecture made in Section 4. Namely, the ML estimator for A, Φ and Ω_Φ is asymptotically normal. The ML estimator for B_2 appears to be consistent with mixed normality. We have not investigated the remaining parameters.

10. An Illustration

In this section, we illustrate the use of the SSR model by applying it to the monthly 10-year government bond rates for Germany and Greece from January 1999 to February 2018.[4] We denote the German and Greek bond rates y^{GE} and y^{GR}, respectively, and measure these in basis points per year. The sample begins at the introduction of the euro area and ends at present day. During this period, the rates initially exhibit convergence towards a common 'euro area rate', until interrupted by the euro area crisis beginning in 2009 and culminating in 2011. The rates, the spread and the changes in the spread are illustrated in Figure 2 below. Because the spread is up to 75 times larger during the second half of the sample than during the first half, we split the display of the sample into the first and second half, respectively.

Panels (a) and (b) in Figure 2 show the bond rates, Panels (c) and (d) show the spread, and Panels (e) and (f) show the changes in the spread in the two periods. We note two features of the observations. First, Panel (a) suggests the rates can be characterized by a shared common stochastic trend, since these tend to move in tandem. Second, Panels (d) and (f) suggest the spread can be characterized by a RCAR process, since the changes in the spread, cf. Panel (f), are clearly positively associated with the level of the spread itself, cf. Panel (d).

[4] Obtained via a Bloomberg LP Terminal using the ticker codes 'GDBR10 Index' and 'GGGB10YR Index'.



(**a**) y_t^{GE} and y_t^{GR} for January 1999 to July 2008.

(**b**) y_t^{GE} and y_t^{GR} for August 2008 to February 2018.

(**c**) y_t^{GR}-y_t^{GE} for January 1999 to July 2008.

(**d**) y_t^{GR}-y_t^{GE} for August 2008 to February 2018.

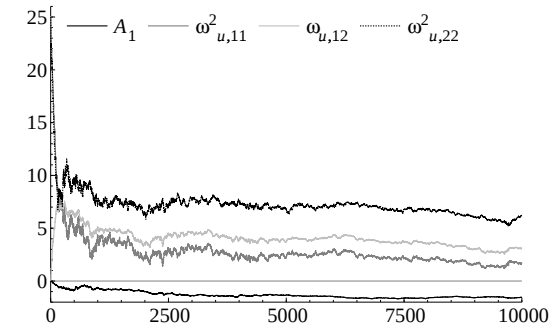
(**e**) Δy_t^{GR}-Δy_t^{GE} for January 1999 to July 2008.

(**f**) Δy_t^{GR}-Δy_t^{GE} for August 2008 to February 2018.

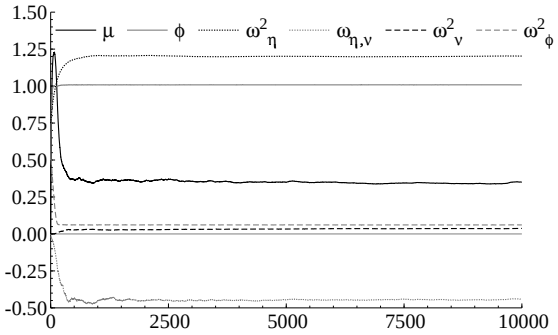
Figure 2. German and Greek 10-year government bond rates, spread and changes in the spread. Monthly observations in basis points from January 1999 to February 2018.

We define the observation vector as $y_t := \begin{bmatrix} y_t^{GE} & y_t^{GR} \end{bmatrix}'$. We condition on the observation for January 1999, which we denote y_0, such that the effective sample spans $t = 1, \ldots, 229$. From visual inspection of Figure 2, our working assumption is that the spread $y_t^{GR} - y_t^{GE}$ is strictly stationary, while the rates y_t share a common stochastic trend. With a $p = 2$ dimensional system, we thus have $r = 1$ stationary component and $p - r = 1$ nonstationary component. Moreover, we fix $B = \begin{bmatrix} 1 & 1 \end{bmatrix}'$, such that the orthogonal complement $b = \begin{bmatrix} -1 & 1 \end{bmatrix}'$ produces the spread. To ensure the model is just-identified, we normalize on the second element of A, such that $A_2 = 1$.

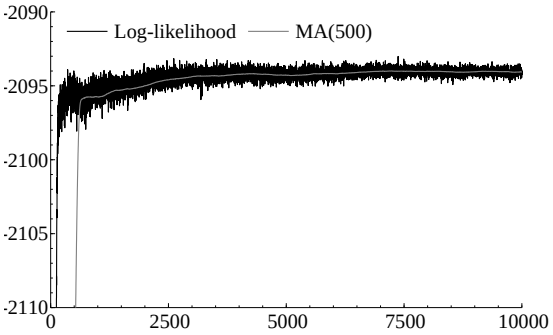
We apply the particle filter-based stochastic approximation method in Algorithm 2 to obtain the approximate ML estimate of the model parameter θ. For this illustration, we run the algorithm for $K = 10,000$ iterations. We let the particle count increase as (84), the step size sequence decrease as (85), and the weighting matrix as (86). We initiate Polyak averaging at iteration $K_0 = 5000$.



(**a**) Parameter sequence, $\{\omega_j\}_{j=1}^K$.



(**b**) Parameter sequence, $\{\lambda_j\}_{j=1}^K$.



(**c**) Simulated log-likelihood sequence, $\{\tilde{\ell}_j\}_{j=1}^K$.

Figure 3. Parameter and log-likelihood sequences from stochastic approximation with $K = 10,000$ iterations. We also show a moving average of lag order 500 for the log-likelihood sequence. To avoid large differences in the scales of the displayed sequences, we have scaled the sequences for A_1, ω^2_η, $\omega_{\eta,\nu}$, ω_ν, and ω_ϕ by 100, 1/300, 1/50, 1/50, and 2 respectively.

Figure 3 shows the results of running the particle filter-based stochastic approximation method. Panel (a) displays the iterations for the parameters in the observation Equation (22), Panel (b) displays the iterations for the parameters in the transition Equation (23), and Panel (c) displays the sequence of realized approximate log-likelihoods together with a moving average of lag order 500. The algorithm has been implemented in the Ox 7 programming language, cf. Doornik (2012), using analytical derivatives of the complete data log-likelihood (32) for the evaluation of the function (33). The elements of the parameter sequence shown in Panels (a) and (b) have stabilized after the initial 7500 iterations. At the 10,000th iteration, the particle count has increased to 550, the step size decreased to 0.2085, and the sequences have stabilized. By inspection of the sequence of the approximate log-likelihood in Panel (c), we see that the value has also stabilized after approximately 7500 iterations.

The estimation results are presented in Table 1, together with approximate classic standard errors.[5] Before considering inference, we assess the model fit. We compute the normalized one-step prediction errors $\tilde{e}^N_{1:T}$ via (81) using $N = 1000$ particles. Table 2 presents univariate tests for autocorrelation (AR) of order one and two, autoregressive conditional heteroskedasticity (ARCH) of order one, and a multivariate test for AR of order one and two, cf. Doornik and Hendry (2013, sct. 11.9.2–3). We cannot reject the null hypothesis of no-AR of order one and two in the univariate as well as multivariate tests at a 5% critical level. Nor can we reject the null hypothesis of no-ARCH for the residuals at a 5% critical level. However, we note the test for the German rate is close to, but below, our chosen critical level. This could suggest unmodeled heteroskedasticity in the German bond rate. In conclusion, the overall specification of the model is acceptable. Moreover, computing the top Lyapunov coefficient via (15) with $n = 10^5$ produces a coefficient of $\hat{\gamma}_n = -0.007$, which indicates the stationary direction is strictly stationary for $\tilde{\theta}_T$.

Table 1. Approximate ML estimate, $\tilde{\theta}_T$.

Parameter	Estimate	Std.err.	Parameter	Estimate	Std.err.
B_1	1.0000	–	μ	0.3449	0.5526
B_2	1.0000	–	ϕ	1.0085	0.0152
A_1	-0.0154	3.4×10^{-5}	ω^2_ϕ	0.0306	0.0031
A_2	1.0000	–	ω^2_η	360.1600	33.6250
$\omega^2_{u\,11}$	2.1063	0.1974	$\omega_{\eta,\nu}$	-22.2400	0.8119
$\omega_{u\,12}$	3.5924	0.3435	ω^2_ν	1.7880	2.0728
$\omega^2_{u\,22}$	6.6327	0.6214			

Note: The approximate log-likelihood is $\tilde{\ell}_T = -2094.1$. The approximate ML estimate has been obtained by running Algorithm 2 for $K = 10,000$ iterations with the particle count increasing to $N = 550$ particles, as described in the main text. The standard errors are based on the inverse of the approximate observed Information matrix computed with $N = 1000$ particles.

Table 2. Model diagnostics.

Univariate tests for AR 1-2:	$\tilde{e}_{t,1}$	$F(2,227) = 1.4523$	$p = 0.2362$
	$\tilde{e}_{t,2}$	$F(2,227) = 1.2086$	$p = 0.3005$
Multivariate test for AR 1-2:		$F(8,448) = 1.6084$	$p = 0.1200$
Univariate tests for ARCH:	$\tilde{e}_{t,1}$	$F(1,227) = 4.7008$	$p = 0.0312$
	$\tilde{e}_{t,2}$	$F(1,227) = 0.58861$	$p = 0.4438$

Note: The approximate normalized one-step prediction errors $\tilde{e}_{1:T}$ have been computed with $N = 1000$ particles for the approximate ML estimate $\tilde{\theta}_T$, cf. Table 1.

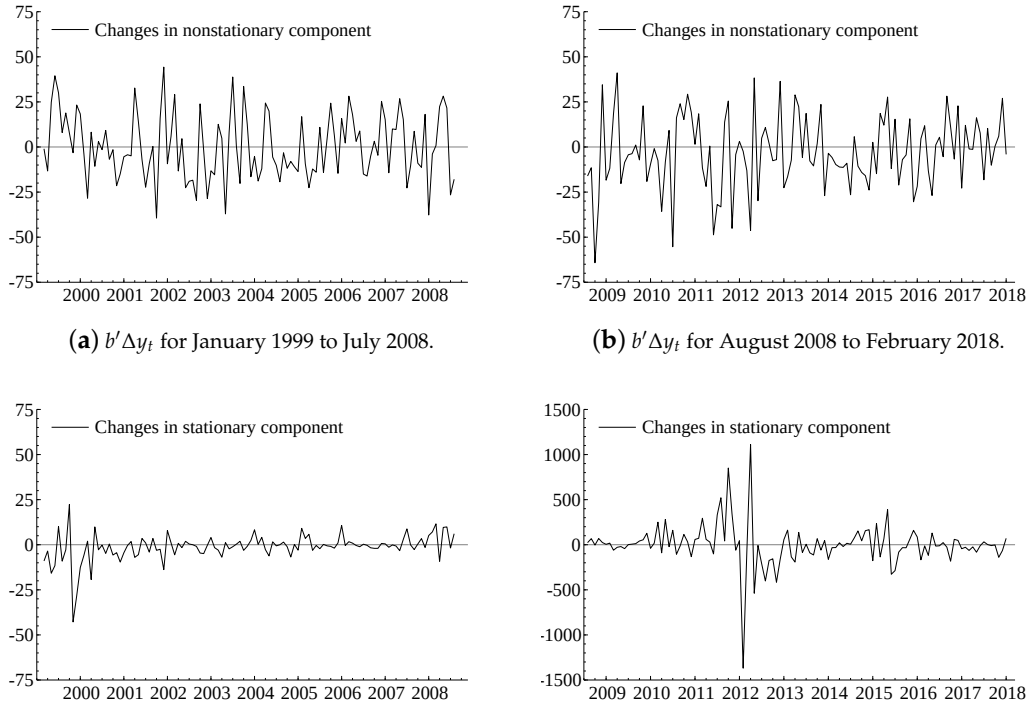
The model is reasonably well-specified, and we therefore proceed to use the approximate classic standard errors to conduct inference on the approximate ML estimates. First, we note the standard error of the estimate of A_1 is extremely small. Since the test for no-ARCH for the residuals associated with the German rate is rejected at the 5% critical level, this could affect the approximate classic standard errors.[6] Nevertheless, it is economically plausible that the stationary component also loads into the German rate, given that a large increase in the Greek rate would in this case coincide with a small drop in the German rate, which is consistent with risk-averse investors seeking safer assets in times of uncertainty, such as the euro area crisis. Second, we cannot reject the null hypothesis that $H_0 : \phi = 1$ at a 5% critical level with $p = 0.577$. Third, the estimate of ω^2_ϕ is significantly different from

[5] The difference between computing the classic standard errors with $N = 1000$ and $N = 10,000$ particles is negligible.

[6] Particle filter-based approximate robust standard errors have been suggested in Doucet and Shephard (2012), but we do not pursue this idea further in the present context.

zero at any commonly used critical level. However, the constant term μ is not significantly different from zero with $p = 0.533$. Fourth, the measurement errors are highly positively correlated with coefficient 0.961, and the innovations of the unobserved components are highly negatively correlated with coefficient -0.876. The results in Table 1 suggest the level of the stationary direction is a stochastic unit root process without a constant term. An approximate likelihood ratio test for the joint null hypothesis $H_0 : \phi = 1$, $\mu = 0$ fails to reject the null at a 5% critical level with $p = 0.374$.

Based on the estimates in Table 1, we use the orthogonal complements b and a to compute the changes of the nonstationary and stationary components, given by $b'\Delta y_t$ and $a'\Delta y_t$, respectively. These are illustrated in Figure 4. First, we note the magnitude of the changes in Panels (a) and (b) of Figure 4 are slightly larger during the second half of the sample than during the first (standard deviations 18.01 and 20.16, respectively). Otherwise, the series in Panels (a) and (b) in Figure 4 are consistent with a homoskedastic random walk plus measurement error, cf. (17). The magnitude of the changes in Panels (c) and (d) of Figure 4 is positively associated with the level, just as observed in Figure 2. This is consistent with a random coefficient autoregressive process plus measurement error, cf. (18).



(a) $b'\Delta y_t$ for January 1999 to July 2008.

(b) $b'\Delta y_t$ for August 2008 to February 2018.

(c) $a'\Delta y_t$ for January 1999 to July 2008.

(d) $a'\Delta y_t$ for August 2008 to February 2018.

Figure 4. Changes in the nonstationary $b'y_t$ and stationary $a'y_t$ components.

Summarizing, the empirical illustration suggests that the SSR model successfully characterizes the 10-year government bond rates for Germany and Greece during the period from January 1999 to February 2018. During this sample, the spread exhibits bubble-like behavior, which is captured by the random coefficient autoregressive dynamics of the stationary component. Additionally, the levels exhibit a shared common stochastic trend, which is captured by the random walk dynamics of the nonstationary component.

11. Conclusions

In this paper, we have proposed and studied the stochastic stationary root model, which is a multivariate nonlinear state space model. We introduced particle filter-based approximations of the intractable log-likelihood function, sample score and observed Information matrix. In turn, we used

these to approximate the ML estimator via stochastic approximation, and showed how to perform inference via the approximate observed Information matrix. We considered model diagnostics to assess the model fit. Additionally, we conducted a simulation study to investigate the asymptotic properties of the ML estimator. Finally, we presented an empirical application to the 10-year government bond rates in Germany and Greece in the period from January 1999 to February 2018 to illustrate the usefulness of the SSR model.

Acknowledgments: The author gratefully acknowledges comments by two anonymous referees that have led to substantial improvements of the paper. The author also thanks the editors, Rocco Mosconi and Paolo Paruolo, for constructive feedback, and the assistant editor, Lu Liao, for assisting in the publication process. Finally, the author would like to thank Anders Rahbek, Michael Pitt, Siem Jan Koopman, Heino Bohn Nielsen, Katarina Juselius, Søren Johansen, Simon Hetland, Gareth Roberts, Adam Johansen, Axel Finke, and Anthony Lee for helpful comments and dicsussions. Part of the work was undertaken while the author was a PhD student at the Department of Economics at the University of Copenhagen and part of the work was undertaken while the author was a CRiSM Research Fellow at the Department of Statistics at the University of Warwick. While at the University of Warwick, funding from the 36 Engineering and Physical Sciences Research Council (EPSRC) is gratefully acknowledged (Grant EP /D002060/1). All errors and omissions are the sole responsibility of the author.

Conflicts of Interest: The author declares no conflict of interest.

Appendix A. Auxiliary Results

Lemma A.1. *For the SSR model (1)–(3) with $\theta \in \Theta$, it holds that*

i $\int p_\omega(y_t \mid x_t) p_\lambda(x_t \mid x_{t-1}) \, \mathrm{d}x_t > 0$ *for all* $x_{t-1} \in \mathcal{R}^p$, *and*
ii $\sup_{x_t \in \mathcal{R}^p} p_\omega(y_t \mid x_t) < \infty$,

for any $t \in \{1, \ldots, T\}$.

Proof of Lemma A.1. By Corollary 1 we have that $\int p_\omega(y_t \mid x_t) p_\lambda(x_t \mid x_{t-1}) \, \mathrm{d}x_t = p_\theta(y_t \mid x_{t-1})$ is Gaussian, and therefore strictly positive for all $x_{t-1} \in \mathcal{R}^p$ and $\theta \in \Theta$, which yields part (*i*). Moreover, because the observation density (25) is Gaussian with constant and non-singular covariance matrix, we obtain part (*ii*). □

Lemma A.2. *For the SSR model (1)–(3) with $\theta \in \Theta$, the model likelihood $\boldsymbol{p}_\theta(y_{1:T})$ is strictly positive and finite,*

$$0 < \boldsymbol{p}_\theta(y_{1:T}) < \infty \,. \tag{A1}$$

Proof of Lemma A.2. Preliminarily, we observe the likelihood in (A1) can equivalently be written in terms of the complete data likelihood $p_\theta(y_{1:T}, x_{1:T})$,

$$\boldsymbol{p}_\theta(y_{1:T}) = \int p_\theta(y_{1:T}, x_{1:T}) \, \mathrm{d}x_{1:T} \,, \tag{A2}$$

which, by the state space structure of the model, cf. (25)–(26), is equivalently

$$\boldsymbol{p}_\theta(y_{1:T}) = \int \prod_{t=1}^{T} p_\omega(y_t \mid x_t) p_\lambda(x_t \mid x_{t-1}) \, \mathrm{d}x_{1:T} \,. \tag{A3}$$

By Lemma A.1.(*i*) and (A3), we have that the likelihood in (A1) is strictly positive, since

$$\int \prod_{t=1}^{T} p_\omega(y_t \mid x_t) p_\lambda(x_t \mid x_{t-1}) \, \mathrm{d}x_{1:T} > 0 \,. \tag{A4}$$

Moreover, by Lemma A.1.(*ii*), the likelihood in (A1) is also finite, since

$$\begin{aligned}
&\int \prod_{t=1}^{T} p_\omega(y_t \mid x_t) p_\lambda(x_t \mid x_{t-1}) \, \mathrm{d}x_{1:T} \\
&\quad \leq \prod_{t=1}^{T} \sup_{x_t \in \mathcal{R}^p} p_\omega(y_t \mid x_t) \int \prod_{t=1}^{T} p_\lambda(x_t \mid x_{t-1}) \, \mathrm{d}x_{1:T} \\
&\quad = \prod_{t=1}^{T} \sup_{x_t \in \mathcal{R}^p} p_\omega(y_t \mid x_t) < \infty ,
\end{aligned} \tag{A5}$$

which completes the proof of Lemma A.2. □

Lemma A.3. *For the model (1)–(3) with $\theta \in \Theta$, it holds that*

i $\quad p_\omega(y_t \mid x_t) p_\lambda(x_t \mid x_{t-1}) \ll p_\theta(x_t \mid x_{t-1}, y_t)$ *for all* $x_{t-1} \in \mathcal{R}^p$,

ii $\quad \sup_{x_{t-1}, x_t \in \mathcal{R}^p \times \mathcal{R}^p} \frac{p_\omega(y_t|x_t) p_\lambda(x_t|x_{t-1})}{p_\theta(x_t|x_{t-1}, y_t)} > 0$, *and*

iii $\quad p_\theta(x_t \mid x_{t-1}, y_t) > 0$ *for all* $x_{t-1} \in \mathcal{R}^p$,

for $t \in \{1, \ldots, T\}$

Proof of Lemma A.3. We preliminarily note that the locally optimal transition density (46) can be written as

$$p_\theta(x_t \mid x_{t-1}, y_t) = \frac{p_\omega(y_t \mid x_t) p_\lambda(x_t \mid x_{t-1})}{p_\theta(y_t \mid x_{t-1})} , \tag{A6}$$

where the predictive observation density is given by the integral,

$$p_\theta(y_t \mid x_{t-1}) = \int p_\omega(y_t \mid x_t) p_\lambda(x_t \mid x_{t-1}) \, \mathrm{d}x_t . \tag{A7}$$

By (A6) and the definition of absolute continuity, part (*i*) states that for every Borel-measurable set $\mathcal{A} \in \mathcal{B}(\mathcal{R}^p)$, it holds that

$$\int_{\mathcal{A}} \frac{p_\omega(y_t \mid x_t) p_\lambda(x_t \mid x_{t-1})}{p_\theta(y_t \mid x_{t-1})} \, \mathrm{d}x_t = 0 \quad \Longrightarrow \quad \int_{\mathcal{A}} p_\omega(y_t \mid x_t) p_\lambda(x_t \mid x_{t-1}) \, \mathrm{d}x_t = 0 . \tag{A8}$$

By (A7) and Lemma A.1.(*i*), we know the predictive observation density is strictly positive $p_\theta(y_t \mid x_{t-1}) > 0$ for all $x_{t-1} \in \mathcal{R}^p$ and $\theta \in \Theta$. Therefore (A8) is true for all $x_{t-1} \in \mathcal{R}^p$ and $\theta \in \Theta$, and part (*i*) holds.

To show part (*ii*), we first use (A6) to write

$$\frac{p_\omega(y_t \mid x_t) p_\lambda(x_t \mid x_{t-1})}{p_\theta(x_t \mid x_{t-1}, y_t)} = p_\theta(y_t \mid x_{t-1}) , \tag{A9}$$

where, by Corollary 1, we have that $p_\theta(y_t \mid x_{t-1})$ is Gaussian and therefore strictly positive for all x_t, $x_{t-1} \in \mathcal{R}^p \times \mathcal{R}^p$ and $\theta \in \Theta$, and part (*ii*) holds.

Part (*iii*) follows from $p_\theta(x_t \mid x_{t-1}, y_t)$ being Gaussian, cf. Lemma 3, and therefore strictly positive for all $x_{t-1} \in \mathcal{R}^p$. Thus, part (*iii*) holds. □

Lemma A.4. *If $\theta \in \Theta$ and $\gamma_t(x_{1:t}) \in L^1[\mathcal{R}^{tp}, \boldsymbol{p}_\theta(x_{1:t} \mid y_{1:t})]$, then it holds that the approximation (60) is consistent,*

$$\tilde{\mathbb{E}}_\theta^N [\gamma_t(x_{1:t}) \mid y_{1:t}] \quad \xrightarrow{P} \quad \mathbb{E}_\theta [\gamma_t(x_{1:t}) \mid y_{1:t}] , \tag{A10}$$

for any $t \in \{1, \ldots, T\}$, *as* $N \to \infty$.

Proof of Lemma A.4. We apply Theorem 9.4.5.(i) in Cappé et al. (2005) by verifying its conditions, i.e., Assumptions 9.4.1–3. We note the theorem is stated for scalar test functions, but generalizes to higher-dimensional test functions. Assumptions 9.4.1–2 is hold by Lemma A.1, while Assumption 9.4.3 holds by Lemma A.3. Thus, the conditions for Theorem 9.4.5.(i) in Cappé et al. (2005) are satisfied, which completes the proof of Lemma A.4. □

Lemma A.5. *If* $\theta \in \Theta$ *and* $\gamma_t(x_{1:t}) \in L^2[\mathcal{R}^{tp}, \boldsymbol{p}_\theta(x_{1:t} \mid y_{1:t})]$, *then it holds that the approximation (60) is consistent and asymptotically normal,*

$$\sqrt{N}\left\{\tilde{\mathbb{E}}_\theta^N[\gamma_t(x_{1:t}) \mid y_{1:t}] - \mathbb{E}_\theta[\gamma_t(x_{1:t}) \mid y_{1:t}]\right\} \xrightarrow{D} N(0, \tilde{\mathbb{S}}_t[\gamma_t(x_{1:t})]), \tag{A11}$$

for any $t \in \{1, \ldots, T\}$, *as* $N \to \infty$. *Initialized by* $\tilde{\mathbb{S}}_0 := 0$, *the asymptotic covariance matrix* $\tilde{\mathbb{S}}_t[\gamma_t(x_{1:t})]$ *is given by*

$$\begin{aligned}\tilde{\mathbb{S}}_t[\gamma_t(x_{1:t})] = \tilde{\mathbb{S}}_{t-1}\Bigg[&\mathbb{E}_{q,t}\left[(\gamma_t(x_{1:t}) - \mathbb{E}_\theta[\gamma_t(x_{1:t}) \mid y_{1:t}])\frac{\tilde{w}_t(x_{t-1:t})}{\boldsymbol{p}_\theta(y_t \mid y_{1:t-1})} \mid x_{1:t-1}\right]\Bigg] \\ &+ \mathbb{V}\mathrm{ar}_\theta\left[\mathbb{E}_{q,t}\left[(\gamma_t(x_{1:t}) - \mathbb{E}_\theta[\gamma_t(x_{1:t}) \mid y_{1:t}])\frac{\tilde{w}_t(x_{1:t})}{\boldsymbol{p}_\theta(y_t \mid y_{1:t-1})} \mid x_{1:t-1}\right] \mid y_{1:t-1}\right] \\ &+ \mathbb{E}_\theta\left[\mathbb{V}\mathrm{ar}_{q,t}\left[(\gamma_t(x_{1:t}) - \mathbb{E}_\theta[\gamma_t(x_{1:t}) \mid y_{1:t}])\frac{\tilde{w}_t(x_{1:t})}{\boldsymbol{p}_\theta(y_t \mid y_{1:t-1})} \mid x_{1:t-1}\right] \mid y_{1:t-1}\right],\end{aligned} \tag{A12}$$

where, for any appropriately integrable function $\gamma(x_{1:t})$, *we define the operators*

$$\mathbb{E}_{q,t}[\gamma(x_{1:t}) \mid x_{1:t-1}] := \int \gamma(x_{1:t})\, q_\theta(x_t \mid f_{1:x-1}, y_{1:t-1})\, \mathrm{d}x_{1:t} \tag{A13}$$

$$\mathbb{V}\mathrm{ar}_{q,t}[\gamma(x_{1:t}) \mid x_{1:t-1}] := \mathbb{E}_{q,t}\left[\gamma(x_{1:t})\,\gamma(x_{1:t})' \mid x_{1:t-1}\right] - \mathbb{E}_{q,t}[\gamma(x_{1:t}) \mid x_{1:t-1}]\,\mathbb{E}_{q,t}[\gamma(x_{1:t}) \mid x_{1:t-1}]', \tag{A14}$$

omitting dependence on θ.

Proof of Lemma A.5. We apply Theorem 9.4.5.(ii) in Cappé et al. (2005) by verifying its conditions, i.e., Assumptions 9.4.1–3. Similar to the proof of Lemma A.4, we note the theorem is stated for scalar test functions, but generalizes to higher-dimensional test functions. Assumptions 9.4.1–2 is hold by Lemma A.1, while Assumption 9.4.3 holds by Lemma A.3. Thus, the conditions for Theorem 9.4.5.(ii) in Cappé et al. (2005) are satisfied, which completes the proof of Lemma A.5. □

Appendix B. Main Results

Proof of Lemma 1. We compute conditional mean and variance of ξ_t in Equation (2). First the mean

$$\begin{aligned}\mathbb{E}_\lambda[\xi_t \mid \xi_{t-1}] &= \mathbb{E}_\lambda[\mu + \Phi_t\xi_{t-1} + \nu_t \mid \xi_{t-1}] \\ &= \mu + \Phi\xi_{t-1},\end{aligned} \tag{A15}$$

and then the variance

$$\begin{aligned}\mathbb{V}\mathrm{ar}_\lambda[\xi_t \mid \xi_{t-1}] &= \mathbb{V}\mathrm{ar}_\lambda[\mu + \Phi_t\xi_{t-1} + \nu_t \mid \xi_{t-1}] \\ &= \mathbb{V}\mathrm{ar}_\lambda\left[\mu + (\xi_{t-1}' \otimes I_r)\,\mathrm{vec}(\Phi_t) + \nu_t \mid \xi_{t-1}\right] \\ &= (\xi_{t-1}' \otimes I_r)\,\mathbb{V}\mathrm{ar}_\lambda[\mathrm{vec}(\Phi_t)]\,(\xi_{t-1}' \otimes I_r)' + \mathbb{V}\mathrm{ar}_\lambda[\nu_t] \\ &= (\xi_{t-1}' \otimes I_r)\,\Omega_\Phi\,(\xi_{t-1}' \otimes I_r)' + \Omega_\nu.\end{aligned} \tag{A16}$$

Since the conditional distribution of ξ_t given ξ_{t-1} is Gaussian, it is completely characterized by its first and second conditional moments. Thus, we obtain equations (12)–(13), which completes the proof of Lemma 1. □

Proof of Lemma 2. The result is an application of the Fisher's and Louis' identities to the SSR model. We use Proposition 10.1.6 in Cappé et al. (2005), by verifying the conditions.

First, we verify that Assumption 10.1.3 in Cappé et al. (2005) holds. We have that Θ is an open subset of $\mathcal{R}^{d_\theta}$, which satisfies Assumption 10.1.3.(i). Assumption 10.1.3.(ii) is satisfied via Lemma A.2. Assumption 10.1.3.(iii) is encompassed by condition (b) of Proposition 10.1.6 in Cappé et al. (2005), shown below. Thus, Assumption 10.1.3 in Cappé et al. (2005) holds.

Second, we verify conditions (a) and (b) of Proposition 10.1.6 in Cappé et al. (2005). Condition (a) holds by Conjecture 1. For condition (b), we begin with the third and last part, which states that

$$\frac{\partial}{\partial\theta}\int \log p_\theta(y_{1:t},\, x_{1:T})\boldsymbol{p}_\vartheta(x_{1:T} \mid y_{1:T})\,\mathrm{d}x_{1:T} = \int \frac{\partial}{\partial\theta}\log p_\theta(y_{1:t},\, x_{1:T})\boldsymbol{p}_\vartheta(x_{1:T} \mid y_{1:T})\,\mathrm{d}x_{1:T}\,. \tag{A17}$$

For θ, $\vartheta \in \Theta$, the complete data log-likelihood (32) is log-Gaussian and therefore continuous with respect to θ, and (A17) holds.

The second part of condition (b) states that for $\theta \in \Theta$,

$$\int \|U_T(x_{1:T};\,\theta)\|\,\boldsymbol{p}_\theta(x_{1:T} \mid y_{1:T})\,\mathrm{d}x_{1:T} < \infty \tag{A18}$$

$$\int \|V_T(x_{1:T};\,\theta)\|\,\boldsymbol{p}_\theta(x_{1:T} \mid y_{1:T})\,\mathrm{d}x_{1:T} < \infty\,, \tag{A19}$$

which is holds by Conjecture 2.

The first part of condition (b) states that for θ, $\vartheta \in \Theta$, the entropy function in (31) is twice-differentiable with respect to θ for fixed ϑ and $y_{1:T}$. Using (A17) and that the complete data log-likelihood (32) is twice-differentiable with respect to θ, we have that (31) is also twice-differentiable with respect to θ. Thus, Proposition 10.1.6 in Cappé et al. (2005) applies for the SSR model, which completes the proof of Lemma 2. □

Proof of Lemma 3. Define the conditional moments of the locally optimal transition density (46),

$$\mu^x_{t|t} \coloneqq \mathbb{E}_\theta\left[x_t \mid x_{t-1},\, y_t\right] \quad \text{and} \quad \Sigma^x_{t|t} \coloneqq \mathbb{V}\mathrm{ar}_\theta\left[x_t \mid x_{t-1},\, y_t\right]\,. \tag{A20}$$

Applying the Gaussian projection, we can write these as

$$\begin{aligned}\mu^x_{t|t} &= \mathbb{E}_\lambda\left[x_t \mid x_{t-1}\right] + \mathbb{C}\mathrm{ov}_\theta\left[x_t,\, y_t \mid x_{t-1}\right]\mathbb{V}\mathrm{ar}_\theta\left[y_t \mid x_{t-1}\right]^{-1}\left(y_{t-1} - \mathbb{E}_\theta\left[y_t \mid x_{t-1}\right]\right)\\ &= \mu^x_{t|t-1} + \Sigma^x_{t|t-1}\Pi'\left[\Sigma^y_{t|t-1}\right]^{-1}\left(y_t - \mu^y_{t|t-1}\right)\end{aligned} \tag{A21}$$

$$\begin{aligned}\Sigma^x_{t|t} &= \mathbb{V}\mathrm{ar}_\lambda\left[x_t \mid x_{t-1}\right] + \mathbb{C}\mathrm{ov}_\theta\left[x_t,\, y_t \mid x_{t-1}\right]\Pi'\,\mathbb{V}\mathrm{ar}_\theta\left[y_t \mid x_{t-1}\right]^{-1}\Pi\,\mathbb{C}\mathrm{ov}_\theta\left[y_t,\, x_t \mid x_{t-1}\right]\\ &= \Sigma^x_{t|t-1} - \Sigma^x_{t|t-1}\Pi'\left[\Sigma^y_{t|t-1}\right]^{-1}\Pi\Sigma^x_{t|t-1}\,,\end{aligned} \tag{A22}$$

where we have used that,

$$\mathbb{C}\mathrm{ov}_\theta\left[x_t,\, y_t \mid x_{t-1}\right] = \mathbb{C}\mathrm{ov}_\theta\left[x_t,\, C(y_0) + \Pi x_t \mid x_{t-1} \mid x_{t-1}\right] = \Sigma^x_{t|t-1}\Pi'\,. \tag{A23}$$

We define the conditional moments of the predictive observation density,

$$\mu^y_{t|t-1} \coloneqq \mathbb{E}_\theta\left[y_t \mid x_{t-1}\right] = \mathbb{E}_\theta\left[C(y_0) + \Pi x_t \mid x_{t-1}\right] = C(y_0) + \Pi\mu^x_{t|t-1} \tag{A24}$$

$$\Sigma^y_{t|t-1} \coloneqq \mathbb{V}\mathrm{ar}_\theta\left[y_t \mid x_{t-1}\right] = \mathbb{V}\mathrm{ar}_\theta\left[C(y_0) + \Pi x_t \mid x_{t-1}\right] = \Pi\Sigma^x_{t|t-1}\Pi' + \Omega_u\,, \tag{A25}$$

where we have used (22). Similarly, we define the conditional moments of the transition density,

$$\mu^x_{t|t-1} := \mathbb{E}_\lambda \left[x_t \mid x_{t-1} \right] = \alpha + \Pi x_{t-1} \tag{A26}$$

$$\Sigma^x_{t|t-1} := \mathbb{V}\mathrm{ar}_\lambda \left[x_t \mid x_{t-1} \right] = \Lambda_t \,, \tag{A27}$$

where we have used (23), which concludes the proof of Lemma 3. □

Proof of Lemma 4. Lemma A.4 establishes that Theorem 9.4.5 in Cappé et al. (2005) holds. It is a corollary to Theorem 9.4.5 in Cappé et al. (2005) that

$$\tilde{L}^N_T(\theta) := \prod_{t=1}^{T} W^N_t \quad \xrightarrow{P} \quad \boldsymbol{p}_\theta(y_{1:T}) =: \boldsymbol{L}_T(\theta) \,, \tag{A28}$$

as $N \to \infty$. By continuity of the logarithm, the continuous mapping theorem and the definitions (8) and (61), we therefore have that,

$$\tilde{\ell}^N_T(\theta) = \log \tilde{L}^N_T(\theta) \quad \xrightarrow{P} \quad \log \boldsymbol{L}_T(\theta) = \boldsymbol{\ell}_T(\theta) \,, \tag{A29}$$

as $N \to \infty$, which completes the proof of Lemma A.4. □

Proof of Lemma 5. We apply Lemma A.5 for $t = T$ setting the test function to $\gamma_T(x_{1:T}) := U_T(x_{1:T}; \theta)$, cf. (33). By Conjecture 2 we have that $U_T(x_{1:T}; \theta) \in L^2[\mathcal{R}^{p \times T}, \boldsymbol{p}_\theta(x_{1:T} \mid y_{1:T})]$, which satisfies the condition, and Lemma A.5 applies. □

Proof of Lemma 6. We apply Lemma A.4 to the functions $U_T(x_{1:T}; \theta)$, $V_T(x_{1:T}; \theta)$ and the outer product $U_T(x_{1:T}; \theta)U_T(x_{1:T}; \theta)'$ for $\theta \in \Theta$. First, Conjecture 2 implies that $U_T(x_{1:T}; \theta) \in L^1[\mathcal{R}^{p \times T}, \boldsymbol{p}_\theta(x_{1:T} \mid y_{1:T})]$, such that by setting the test function to $\gamma_T(x_{1:T}) := U_T(x_{1:T}; \theta)$, Lemma A.4 gives us that,

$$\tilde{\mathbb{E}}^N_\theta \left[U_T(x_{1:T}; \theta) \mid y_{1:t} \right] \quad \xrightarrow{P} \quad \mathbb{E}_\theta \left[U_T(x_{1:T}; \theta) \mid y_{1:t} \right] , \tag{A30}$$

as $N \to \infty$. Second, Conjecture 2 states $V_T(x_{1:T}; \theta) \in L^1[\mathcal{R}^{p \times T}, \boldsymbol{p}_\theta(x_{1:T} \mid y_{1:T})]$, such that by setting the test function to $\gamma_T(x_{1:T}) := V_T(x_{1:T}; \theta)$, Lemma A.4 gives us that,

$$\tilde{\mathbb{E}}^N_\theta \left[V_T(x_{1:T}; \theta) \mid y_{1:t} \right] \quad \xrightarrow{P} \quad \mathbb{E}_\theta \left[V_T(x_{1:T}; \theta) \mid y_{1:t} \right] , \tag{A31}$$

as $N \to \infty$. Third, we note that by the Cauchy-Schwarz inequality it holds that,

$$\| U_T(x_{1:T}; \theta) U_T(x_{1:T}; \theta)' \| \leq \| U_T(x_{1:T}; \theta) \| \| U_T(x_{1:T}; \theta)' \| = \| U_T(x_{1:T}; \theta) \|^2 , \tag{A32}$$

such that, by Conjecture 2, we have that

$$\int \| U_T(x_{1:T}; \theta) U_T(x_{1:T}; \theta)' \| \boldsymbol{p}_\theta(x_{1:T} \mid y_{1:T}) \, \mathrm{d}x_{1:T} \leq \int \| U_T(x_{1:T}; \theta) \|^2 \boldsymbol{p}_\theta(x_{1:T} \mid y_{1:T}) \, \mathrm{d}x_{1:T} < \infty \,. \tag{A33}$$

Thus, by setting the test function to $\gamma_T := U_T(x_{1:T}; \theta)U_T(x_{1:T}; \theta)'$, Lemma A.4 gives us that,

$$\tilde{\mathbb{E}}^N_\theta \left[U_T(x_{1:T}; \theta) U_T(x_{1:T}; \theta)' \mid y_{1:t} \right] \quad \xrightarrow{P} \quad \mathbb{E}_\theta \left[U_T(x_{1:T}; \theta) U_T(x_{1:T}; \theta)' \mid y_{1:t} \right] , \tag{A34}$$

as $N \to \infty$. Now, by (37), (39), (40), (63), (66), and (67), we have that (A30)–(A34) correspond to,

$$\tilde{S}_T^N(\theta) \xrightarrow{P} \boldsymbol{S}_T(\theta) \tag{A35}$$

$$\tilde{G}_T^N(\theta) \xrightarrow{P} \boldsymbol{G}_T(\theta) \tag{A36}$$

$$\tilde{K}_T^N(\theta) \xrightarrow{P} \boldsymbol{K}_T(\theta)\,, \tag{A37}$$

as $N \to \infty$, respectively, such that we get by the continuous mapping theorem that,

$$\tilde{I}_T^N(\theta) = \tilde{S}_T^N(\theta)\tilde{S}_T^N(\theta)' - \tilde{G}_T^N(\theta) - \tilde{K}_T^N(\theta) \xrightarrow{P} \boldsymbol{S}_T(\theta)\boldsymbol{S}_T(\theta)' - \boldsymbol{K}_T(\theta) - \boldsymbol{G}_T(\theta) = \boldsymbol{I}_T(\theta)\,, \tag{A38}$$

as $N \to \infty$, which completes the proof of Lemma 6. □

References

Anderson, Brian D. O., and John B. Moore. 1979. *Optimal Filtering*. Upper Saddle River: Prentice-Hall, pp. 1–367. [CrossRef]

Andrieu, Christophe, and Arnaud Doucet. 2002. Particle filtering for partially observed Gaussian state space models. *Journal of the Royal Statistical Society B* 64: 827–36. [CrossRef]

Bec, Frederique, and Anders Rahbek. 2004. Vector Equilibrium Correction Models with Non-Linear Discontinuous Adjustments. *Econometrics Journal* 7: 628–51. [CrossRef]

Bec, Frédérique, Anders Rahbek, and Neil Shephard. 2008. The ACR Model: A Multivariate Dynamic Mixture Autoregression. *Oxford Bulletin of Economics and Statistics* 70: 583–618. [CrossRef]

Cappé, Olivier, Eric Moulines, and Tobias Rydén. 2005. *Inference in Hidden Markov Models*. New York: Springer.

Chang, Yoosoon, J. Isaac Miller, and Joon Y. Park. 2009. Extracting a Common Stochastic Trend: Theory with Some Applications. *Journal of Econometrics* 150: 231–47. [CrossRef]

Chen, Rong, and Jun S. Liu. 2000. Mixture Kalman Filters. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 62: 493–508. [CrossRef]

Chopin, Nicolas. 2004. Central Limit Theorem for Sequential Monte Carlo Methods and its Application to Bayesian Inference. *The Annals of Statistics* 32: 2385–411.

Creal, Drew. 2012. A Survey of Sequential Monte Carlo Methods for Economics and Finance. *Econometric Reviews* 31: 245–96.

Dempster, Arthur P., Nan M. Laird, and Donald B. Rubin. 1977. Maximum Likelihood from Incomplete Data via the EM Algorithm. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 37: 1–38.

Dickey, David A., and Wayne A. Fuller. 1979. Distribution of the Estimators for Autoregressive Time Series with a Unit Root. *Journal of the American Statistical Association* 74: 427–31. [CrossRef]

Dieci, Luca, and Erik S. Van Vleck. 1995. Computation of a Few Lyapunov Exponents for Continuous and Discrete Dynamical Systems. *Applied Numerical Mathematics* 17: 275–91.

Doornik, Jurgen A., and David F. Hendry. 2013. *Modelling Dyamic Systems—PcGive 14: Volume II*. London: Timberlake Consultants Ltd., pp. 1–284. [CrossRef]

Doornik, Jurgen A. 2012. *Developer's Manual for Ox 7*; London: Timberlake Consultants Ltd., pp. 1–182. [CrossRef]

Douc, Randal, Olivier Cappé, and Eric Moulines. 2005. Comparison of Resampling Schemes for Particle Filtering. Paper presented at the 4th International Symposium on Image and Signal Processing and Analysis, Zagreb, Croatia, September 15–17, pp. 64–69.

Douc, Randal, Eric Moulines, Jimmy Olsson, and Ramon Van Handel. 2011. Consistency of the Maximum Likelihood Estimator for General Hidden Markov Models. *The Annals of Statistics* 39: 474–513.

Douc, Randal, Eric Moulines, and David Stoffer. 2014. *Nonlinear Time-Series: Theory, Methods, and Applications with R Examples*. Boca Raton: CRC Press. [CrossRef]

Doucet, Arnaud, and Neil Shephard. 2012. *Robust Inference on Parameters via Particle Filters and Sandwich Covariance Matrices*. Economics Series Working Papers 606, University of Oxford, Oxford, UK.

Doucet, Arnaud, Nando De Freitas, Kevin Murphy, and Stuart Russell. 2000. Rao-Blackwellised Particle Filtering for Dynamic Bayesian Networks. In *Proceedings of the 16th Conference on Uncertainty in Artificial Intelligence*. San Francisco: Morgan Kaufmann Publishers Inc.

Doucet, Arnaud, Nando de Freitas, and Neil Gordon. 2001. *Sequential Monte Carlo Methods in Practice*. New York: Springer.

Durbin, James, and Siem Jan Koopman. 2012. *Time Series Analysis by State Space Methods*, 2nd ed. Oxford: Oxford University Press. [CrossRef]

Engle, Robert F., David F. Hendry, and Jean-Francois Richard. 1983. Exogeneity. *Econometrica* 51: 277–304. [CrossRef]

Engle, Robert F. 1982. Autoregressive Conditional Heteroscedasticity with Estimates of the Variance of United Kingdom Inflation. *Econometrica* 50: 987–1007. [CrossRef]

Feigin, Paul D., and Richard L. Tweedie. 1985. Random Coefficient Autoregressive Processes: A Markov Chain Analysis of Stationarity and Finiteness of Moments. *Journal of Time Series Analysis* 6: 1–14. [CrossRef]

Francq, Christian, and Jean-Michel Zakoian. 2010. *GARCH Models: Structure, Statistical Inference and Financial Applications*. Hoboken: Wiley, p. 489. [CrossRef]

Gordon, Neil J., David J. Salmond, and Adrian F. M. Smith. 1993. Novel Approach to Nonlinear/Non-Gaussian Bayesian State Estimation. *IEE Proceedings F, Radar and Signal Processing* 140: 107–13. [CrossRef]

Granger, Clive W. J., and Norman R. Swanson. 1997. An Introduction to Stochastic Unit-Root Processes. *Journal of Econometrics* 80: 35–62. [CrossRef]

Hamilton, James Douglas. 1994. *Time Series Analysis*. Princeton: Princeton University Press.

Jensen, Søren Tolver, and Anders Rahbek. 2004. Asymptotic Inference for Nonstationary GARCH. *Econometric Theory* 20: 1203–26. [CrossRef]

Johansen, Søren. 1996. *Likelihood-Based Inference in Cointegrated Vector Autoregressive Models (Advanced Texts in Econometrics)*. Oxford: Oxford University Press.

Juselius, Katarina. 2007. *The Cointegrated VAR Model: Methodology and Applications (Advanced Texts in Econometrics)*. Oxford: Oxford University Press. [CrossRef]

Kalman, Rudolph Emil. 1960. A New Approach to Linear Filtering and Prediction Problems. *Transactions of the ASME–Journal of Basic Engineering* 82: 35–45.

Kristensen, Dennis, and Anders Rahbek. 2010. Likelihood-based Inference for Cointegration with Nonlinear Error-Correction. *Journal of Econometrics* 158: 78–94.

Kristensen, Dennis, and Anders Rahbek. 2013. Testing and Inference in Nonlinear Cointegrating Vector Error Correction Models. *Econometric Theory* 29: 1238–88. [CrossRef]

Kushner, Harold, and G. George Yin. 2003. *Stochastic Approximation and Recursive Algorithms and Applications*, 2nd ed. New York: Springer.

Leybourne, Stephen J., Brendan P. M. McCabe, and Andrew R. Tremayne. 1996. Can Economic Time Series Be Differenced To Stationarity? *Journal of Business & Economic Statistics* 14: 435–46.

Lieberman, Offer, and Peter C. B. Phillips. 2014. Norming rates and limit theory for some time-varying coefficient autoregressions. *Journal of Time Series Analysis* 35: 592–623.

Lieberman, Offer, and Peter C. B. Phillips. 2017. A multivariate stochastic unit root model with an application to derivative pricing. *Journal of Econometrics* 196: 99–110. [CrossRef]

Ling, Shiqing. 2004. Estimation and Testing Stationarity for Double-Autoregressive Models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 66: 63–78. [CrossRef]

Ling, Shiqing. 2007. A Double AR(p) Model: Structure and Estimation. *Statistica Sinica* 17: 161–75.

Louis, Thomas A. 1982. Finding the Observed Information Matrix when Using the EM Algorithm. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 44: 226–33. [CrossRef]

McCabe, Brendan P. M., and Richard J. Smith. 1998. The power of some tests for difference stationarity under local heteroscedastic integration. *Journal of the American Statistical Association* 93: 751–61.

McCabe, Brendan P. M., and Andrew R. Tremayne. 1995. Testing a time series for difference stationarity. *Annals of Statistics* 23: 1015–28.

Meyn, Sean P., and Richard L. Tweedie. 2005. *Markov Chains and Stochastic Stability*; Cambridge: Cambridge University Press. [CrossRef]

Murray, Iain, Zoubin Ghahramani, and David MacKay. 2006. MCMC for Doubly-Intractable Distributions. Paper presented at the Twenty-Second Conference on Uncertainty in Artificial Intelligence (UAI2006), Cambridge, MA, USA, July 13–16.

Nocedal, Jorge, and Stephen Wright. 2006. *Numerical Optimization*, 2nd ed. New York: Springer. [CrossRef]

Olsson, Jimmy, and Tobias Rydén. 2008. Asymptotic Properties of Particle Filter-Based Maximum Likelihood Estimators for State Space Models. *Stochastic Processes and Their Applications* 118: 649–80.

Pedersen, Rasmus Søndergaard, and Olivier Wintenberger. 2018. On the tail behavior of a class of multivariate conditionally heteroskedastic processes. *Extremes* 21: 261–84.

Pitt, Mark, and Neil Shephard. 1999a. Time Varying Covariances: A Factor Stochastic Volatility Approach. In *Bayesian Statistics*. Edited by José M. Bernardo, James O. Berger, A. P. Dawid and Adrian F. M. Smith. Oxford: Oxford University Press, vol. 6, pp. 547–70.

Pitt, Michael K., and Neil Shephard. 1999b. Filtering via Simulation: Auxiliary Particle Filters. *Journal of the American Statistical Association* 94: 590–99. [CrossRef]

Polyak, Boris T., and Anatoli B. Juditsky. 1992. Acceleration of Stochastic Approximation by Averaging. *SIAM Journal on Control and Optimization* 30: 838–55.

Polyak, Boris Teodorovich. 1990. A New Method of Stochastic Approximation Type. *Automation and Remote Control* 51: 98–107. [CrossRef]

Poyiadjis, George, Arnaud Doucet, and Sumeetpal S. Singh. 2011. Particle Approximations of the Score and Observed Information Matrix in State Space Models with Application to Parameter Estimation. *Biometrika* 98: 65–80.

Robbins, Herbert, and Sutton Monro. 1951. A Stochastic Approximation Method. *The Annals of Mathematical Statistics* 22: 400–7. [CrossRef]

Robert, Christian P., and George Casella. 2010. *Introducing Monte Carlo Methods with R*, 1st ed. New York: Springer, p. 296. [CrossRef]

Nielsen, Heino Bohn, and Anders Rahbek. 2014. Unit Root Vector Autoregression with Volatility Induced Stationarity. *Journal of Empirical Finance* 29: 144–67. [CrossRef]

Nielsen, Heino Bohn. 2016. The Co-integrated Vector Autoregression with Errors-in-Variables. *Econometric Reviews* 35: 169–200. [CrossRef]