Journal of Multivariate Analysis 123 (2014) 228–251

Contents lists available at ScienceDirect

Journal of Multivariate Analysis

journal homepage: www.elsevier.com/locate/jmva

Mixtures of skewed Kalman filters

Hyoung-Moon Kim [a], Duchwan Ryu [b], Bani K. Mallick [c], Marc G. Genton [d,*]

[a] Department of Applied Statistics, Konkuk University, Seoul, Republic of Korea
[b] Department of Biostatistics, Medical College of Georgia, Augusta, USA
[c] Department of Statistics, Texas A&M University, College Station, USA
[d] CEMSE Division, King Abdullah University of Science and Technology, Thuwal, Saudi Arabia

ARTICLE INFO

Article history:
Received 23 August 2012
Available online 20 September 2013

AMS 1991 subject classifications:
62F99
62H10

Keywords:
Closed skew-normal distribution
Discrete mixture
Kalman filter
Scale mixtures
Sequential importance sampling
Closed skew-*t* distribution

ABSTRACT

Normal state-space models are prevalent, but to increase the applicability of the Kalman filter, we propose mixtures of skewed, and extended skewed, Kalman filters. To do so, the closed skew-normal distribution is extended to a scale mixture class of closed skew-normal distributions. Some basic properties are derived and a class of closed skew-*t* distributions is obtained. Our suggested family of distributions is skewed and has heavy tails too, so it is appropriate for robust analysis. Our proposed special sequential Monte Carlo methods use a random mixture of the closed skew-normal distributions to approximate a target distribution. Hence it is possible to handle skewed and heavy tailed data simultaneously. These methods are illustrated with numerical experiments.

1. Introduction

State-space models have been widely investigated and used in applied fields such as computer vision, economics, engineering and statistics. The main idea of the state-space model is that the observation y_t at time t is generated by the observation and the state equations. Error terms are usually assumed to follow normal distributions independently.

The assumption of normality in the Kalman filter is not satisfied for a number of real applications. For example, the distributions in a state-space model can be skewed. Inspired by this idea, Naveau et al. [28] proposed a skewed Kalman filter based on the closed skew-normal distribution originally developed by González-Farías et al. [21,22]. We develop a new skewed Kalman filter and an extended skewed Kalman filter, and then extend these models to mixtures of skewed and extended skewed Kalman filters. They include the skewed Kalman filter as a special case when the mixing distribution is degenerated to 1. So we can handle skewed and heavy tailed data simultaneously. Furthermore, from a computational perspective, our extended skewed Kalman filter is faster than the model given by Naveau et al. [28] since there is no need to calculate some mean and covariance terms using numerical techniques.

To implement the skewed Kalman filter we extend the mixture Kalman filter [14] to the mixture of skewed Kalman filters in a direct way. These authors nicely defined partial conditional dynamic linear models and then developed the extended mixture Kalman filter (EMKF). The main idea of EMKF is to extract as many linear and Gaussian components from the system as possible, and then to integrate these components out using the Kalman filter before running a Monte Carlo filter on the

* Corresponding author.
E-mail address: marc.genton@kaust.edu.sa (M.G. Genton).

0047-259X/$ – see front matter
http://dx.doi.org/10.1016/j.jmva.2013.09.002

remaining components. Since, given nonlinear components, the extended skewed Kalman filter becomes linear and with Gaussian errors, the EMKF is directly employed to implement it.

The closed skew-normal distribution [21,22] is the basis distribution for the skewed Kalman filter [28]. It is a family of distributions which includes the (skew-)normal distributions as special cases. Furthermore it contains some skew-normal distributions suggested by Arnold and Beaver [6] and Liseo and Loperfido [25]. It preserves some important properties of the normal distribution, for instance, being closed under: marginalization, conditional distributions, linear transformation (full column or full row rank), sums of independent random variables from this family, and joint distribution of independent random variables in this family; see [17] for an overview.

We extend the closed skew-normal distribution to the scale mixtures of closed skew-normal distributions. Some basic properties are also obtained. As a special case, a class of closed skew-t distributions is derived in explicit form. The scale mixtures of closed skew-normal distributions contain the scale mixtures of skew-normal distributions and the scale mixtures of normal distributions [1,31] as special cases by the aforementioned relationships. The scale mixtures of skew-normal distributions appeared in [12], which include (skew-)normal distributions as special cases. One particular case of this distribution is the skew-normal distribution having a degenerate mixing density. The scale mixtures of normal distributions are widely used. For example, Choy et al. [15] applied those to the study of robust analysis of a normal location parameter, Chen et al. [13] used those for Bayesian modeling of correlated binary responses, and Bradley et al. [11] investigated non-Gaussian state-space models using those, to name a few.

This paper is organized as follows. In Section 2, a scale mixture class of closed skew-normal distributions is derived. Some basic properties are also obtained. As a special case, a class of closed skew-t distributions is derived in explicit form. Using the scale mixtures of closed skew-normal distributions, we develop two different versions of mixtures of skewed Kalman filters in Section 3. In Sections 3.3 and 3.5, we suggest simple generating methods for mixtures of (extended) skewed Kalman filters. We illustrate these methods with numerical experiments in Section 4.

2. A scale mixture class of closed skew-normal distributions

2.1. Definition and some examples

A random vector W has a multivariate closed skew-normal distribution according to González-Farías et al. [21,22] if it has the following probability density function (pdf):

$$C\phi_n(w;\mu,\Sigma)\Phi_m\{D(w-\mu);\nu,\Delta\},\quad w\in\mathbb{R}^n, \tag{1}$$

where $n\geq 1$, $m\geq 1$, $\mu\in\mathbb{R}^n$, $\nu\in\mathbb{R}^m$, $D\in\mathbb{R}^{m\times n}$, $\Sigma\in\mathbb{R}^{n\times n}$ and $\Delta\in\mathbb{R}^{m\times m}$ are both covariance matrices. Here $\phi_n(w;\mu,\Sigma)$ and $\Phi_m(w;\mu,\Sigma)$ are the normal pdf and cumulative distribution function (cdf) with mean μ and covariance matrix Σ. The normalizing constant C of the density function (1) is defined by

$$C^{-1}=\Phi_m(0;\nu,\Delta+D\Sigma D^T). \tag{2}$$

We shall then write $W\sim \mathrm{CSN}_{n,m}(\mu,\Sigma,D,\nu,\Delta)$. When $D=0$, it reduces to the normal distribution. Furthermore when $m=\Delta=1$ and $\nu=0$, it becomes the skew-normal distribution suggested by Azzalini and Dalla Valle [9] and Azzalini and Capitanio [7].

The closed skew-normal distribution will be extended to scale mixtures of closed skew-normal distributions similar to scale mixtures of (skew-)normal distributions. From now on all proofs can be found in the Appendix except for simple cases.

Lemma 1. *Let W and Z be defined as follows:*

$$W=\mu+K(\lambda)^{1/2}\epsilon_1,$$
$$Z=-\nu+D\epsilon_1+\epsilon_2,$$

for given λ where $\epsilon_1\sim N_n(0,\Sigma)$ and $\epsilon_2\sim N_m(0,\Delta)$ are independent random vectors. Here $n\geq 1$, $m\geq 1$, $\mu\in\mathbb{R}^n$, $\nu\in\mathbb{R}^m$, $\Sigma\in\mathbb{R}^{n\times n}$ and $\Delta\in\mathbb{R}^{m\times m}$ are both covariance matrices, $D\in\mathbb{R}^{m\times n}$ is an arbitrary matrix, λ is a mixing variable and $K(\lambda)$ is a weight function. Then the conditional distribution $(W|Z\geq 0,\lambda)\overset{d}{=}V\sim CSN_{n,m}(\mu,K(\lambda)\Sigma,K(\lambda)^{-1/2}D,\nu,\Delta)$.

Hence we define the pdf of scale mixtures of closed skew-normal distributions as follows:

$$f_W(w)=C\int_0^\infty \phi_n\{w;\mu,K(\lambda)\Sigma\}\Phi_m\{K(\lambda)^{-1/2}D(w-\mu);\nu,\Delta\}dH(\lambda),$$

where C is defined in (2) and $H(\lambda)$ is a cdf. We shall then write $W\sim \mathrm{SMCSN}_{n,m}(\mu,K(\lambda)\Sigma,K(\lambda)^{-1/2}D,\nu,\Delta)$. Thus the conditional distribution $W|\lambda\sim \mathrm{CSN}_{n,m}(\mu,K(\lambda)\Sigma,K(\lambda)^{-1/2}D,\nu,\Delta)$. Therefore there is a simple stochastic representation of the above class of distributions:

$$W=\mu+K(\lambda)^{1/2}Y, \tag{3}$$

where λ is independent of Y, and $Y \sim \text{CSN}_{n,m}(0, \Sigma, D, \nu, \Delta)$. This stochastic representation is useful for simulation and some theoretical purposes.

Our approach is similar to scale mixtures of the selection normal distribution [3] which have a mixing variable on the full covariance matrix of the multivariate normal distribution. So they have the density in two integration formulae appearing at both numerator and denominator. Hence statistical inferences could be harder than ours. Furthermore the integration should be done for their selection mechanism which cannot be calculated in closed form, whereas our pdf has one integration of mixing variable. In this sense we resolve the problem which they have for a mixing variable on the full covariance matrix. This approach is prevalent for most scale mixture approaches. Instead, we have a conditional class where the conditional mean and covariance of Z given λ do not depend on λ, a mixing variable (see the proof of Lemma 1 in the Appendix). Other approaches of scale mixtures of skewed distributions can be found in [5, Proposition 2.2] and [4, Section 5.3].

When $D = 0$, it reduces to scale mixtures of normal distributions. Furthermore when $m = \Delta = 1$ and $\nu = 0$, it turns out to be scale mixtures of skew-normal distributions. The scale mixtures of skew-normal distributions appeared in [12], and include (skew-)normal distributions as special cases. One particular case of this distribution is the skew-normal distribution, for which $K(\lambda) = 1$. Here are some special cases of scale mixtures of closed skew-normal distributions.

Example 1. When the distribution H is a discrete measure on $\{\lambda_1, \lambda_2, \ldots, \lambda_q\}$ with probabilities $p_1, p_2, \ldots, p_q$, respectively, then the density of finite mixtures of closed skew-normal distributions is

$$f_W(w) = C \sum_{j=1}^{q} p_j \phi_n\{w; \mu, K(\lambda_j)\Sigma\} \Phi_m\{K(\lambda_j)^{-1/2} D(w-\mu); \nu, \Delta\},$$

where $0 \le p_j \le 1$ and $\sum_{j=1}^{q} p_j = 1$. Furthermore when $K(\lambda) = 1/\lambda$ and the distribution H is a discrete measure on $\{\lambda_1 = \gamma, \lambda_2 = 1\}$ with probabilities p, $1-p$, respectively, then we have the contaminated closed skew-normal distribution with density

$$f_W(w) = C\left[p\phi_n(w; \mu, \gamma^{-1}\Sigma)\Phi_m\{\gamma^{1/2} D(w-\mu); \nu, \Delta\} + (1-p)\phi_n(w; \mu, \Sigma)\Phi_m\{D(w-\mu); \nu, \Delta\}\right],$$

where $0 < p < 1$ and $0 < \gamma \le 1$. This type of distributions, specifically the contaminated normal distribution, has been widely used in numerical studies of robustness requiring distributions that are elongated, that is, stretched relative to a Gaussian behavior [20].

Example 2. When $K(\lambda) = 4\lambda^2$ and λ follows an asymptotic Kolmogorov distribution with density

$$h(\lambda) = 8 \sum_{k=1}^{\infty} (-1)^{k+1} k^2 \lambda \exp(-2k^2\lambda^2), \quad \lambda > 0,$$

we have a closed skew-logistic distribution.

Example 3. Closed skew-stable distributions can be obtained by $K(\lambda) = 2\lambda$ and the mixture distribution $dH(\lambda) = S(\alpha/2, 1)$, where the pdf of the positive stable distribution $S(\alpha, 1)$ is

$$h_{PS}(\lambda|\alpha, 1) = \frac{\alpha}{1-\alpha} \lambda^{-\alpha/(1-\alpha)-1} \int_0^1 s(u) \exp\left\{-\frac{s(u)}{\lambda^{\alpha/(1-\alpha)}}\right\} du, \quad \lambda > 0, \tag{4}$$

for $0 < \alpha < 1$ with

$$s(u) = \left\{\frac{\sin(\alpha\pi u)}{\sin(\pi u)}\right\}^{\alpha/(1-\alpha)} \left[\frac{\sin\{(1-\alpha)\pi u\}}{\sin(\pi u)}\right].$$

When $\alpha = 1$, we get a closed skew-Cauchy distribution. The closed skew-normal distribution can also be obtained from the closed skew-stable distribution by taking $\alpha \to 1$.

Example 4. A closed skew-exponential power distribution can be obtained by choosing $K(\lambda) = 1/(2c_0\lambda)$ and $h(\lambda) = (1/\lambda)^{(k+1)/2} h_{SP}(\lambda|\alpha, 1)$, where $h_{PS}(\lambda|\alpha, 1)$ is given in (4), and c_0 is defined by $c_0 = \Gamma(3/2\alpha)/\Gamma(1/2\alpha)$ and $1/2 < \alpha < 1$. Closed skew-normal and closed skew-Laplace distributions can be obtained by taking $\alpha = 1$ and $\alpha = 1/2$, respectively.

Example 5. A closed skew-slash distribution can be derived by choosing $K(\lambda) = 1/\lambda^{2/q}$, $q > 0$ and $\lambda \sim U(0, 1)$. Wang and Genton [30] described multivariate and skewed multivariate extensions of the slash distribution.

2.2. Some basic properties of SMCSN

We develop some basic properties of scale mixtures of closed skew-normal distributions in this section. First, we derive a general expression for the moment generating function (mgf) of scale mixtures of closed skew-normal distributions.

Theorem 1. *Suppose $W \sim SMCSN_{n,m}(\mu, K(\lambda)\Sigma, K(\lambda)^{-1/2}D, \nu, \Delta)$. Then the mgf of W is given by*

$$M_W(t) = E\{\exp(t^T W)\} = \int_0^\infty M_{\mathrm{CSN}}(t)dH(\lambda),$$

where $\forall t \in \mathbb{R}^n$

$$M_{\mathrm{CSN}}(t) = C\Phi_m\{K(\lambda)^{1/2}D\Sigma t; \nu, \Delta + D\Sigma D^T\} \exp\left\{t^T\mu + \frac{K(\lambda)}{2}t^T\Sigma t\right\},$$

and C is defined in (2).

The moments of mixing distributions are defined as follows:

$$E\left\{K(\lambda)^{r/2}\right\} = c_r, \quad \text{where } r = 1, 2, \ldots.$$

We summarize those in Table 1 for special cases of scale mixtures of closed skew-normal distributions. Using the stochastic relationship (3), the mean and covariance matrix are derived. Their existence depends on the existence of c_1 and c_2. For example, there are no higher moments existing for closed skew-stable distributions. Since the mixing distribution is a positive stable distribution which has all moments of order less than $\alpha \in (0, 2]$, but none greater than α, the characteristic exponent α parameter defines the fatness of the tails (large α implies thin tails).

Theorem 2. *Suppose $W \sim SMCSN_{n,m}(\mu, K(\lambda)\Sigma, K(\lambda)^{-1/2}D, \nu, \Delta)$. Then the mean and covariance matrix of W are*

$$E(W) = \mu + c_1E(Y), \quad \text{if } c_1 < \infty,$$
$$Cov(W) = c_2\left\{\Sigma + C\Phi_m^{**}(0; \nu, \Delta + D\Sigma D^T)\right\} - c_1^2E(Y)E(Y)^T, \quad \text{if } c_2 < \infty,$$

where

$$E(Y) = C\Phi_m^{*}(0; \nu, \Delta + D\Sigma D^T),$$
$$\Phi_m^{*}(0; \nu, \Delta + D\Sigma D^T) = \left.\frac{\partial}{\partial t}\Phi_m(D\Sigma t; \nu, \Delta + D\Sigma D^T)\right|_{t=0},$$
$$\Phi_m^{**}(0; \nu, \Delta + D\Sigma D^T) = \left.\frac{\partial^2}{\partial t\partial t^T}\Phi_m(D\Sigma t; \nu, \Delta + D\Sigma D^T)\right|_{t=0},$$

and C is defined in (2).

In general, it is possible to calculate higher moments of scale mixtures of closed skew-normal distributions along the lines of Genton et al. [18].

Flecher et al. [16] derived their Proposition 2 to estimate the closed skew-normal distribution parameters using a weighted moments approach. We extend their Proposition 2 to the case of scale mixtures of closed skew-normal distributions as follows.

Theorem 3. *Suppose $W \sim SMCSN_{n,m}(\mu, K(\lambda)\Sigma, K(\lambda)^{-1/2}D, 0, \Delta)$, and $h(w) = h(w_1, \ldots, w_n)$ be any real valued function such that $E\{h(W)\}$ is finite. Then*

$$E\left\{h(W)\Phi_n^r(W; 0, I_n)\right\} = \int_0^\infty \frac{\Phi_{rn+m}\{0; \nu_+, \Delta_+ + K(\lambda)D_+\Sigma D_+^T\}}{\Phi_m(0; 0, \Delta + D\Sigma D^T)}E\{h(W_+)\}dH(\lambda),$$

where $W_+|\lambda \sim CSN_{n,rn+m}(\mu, K(\lambda)\Sigma, D_+, \nu_+, \Delta_+)$ with $\Delta_+ = \begin{pmatrix} I_{rn} & 0 \\ 0 & \Delta \end{pmatrix}$, $D_+ = \begin{pmatrix} D_* \\ K(\lambda)^{-1/2}D \end{pmatrix}$, D_* a $rn \times n$ matrix defined by $D_* = \begin{pmatrix} I_n \\ \vdots \\ I_n \end{pmatrix}$ with I_n the identity matrix of size n, and ν_+ a $rn + m$ vector defined by $\nu_+ = -(\mu^T \ldots \mu^T\ 0_m^T)^T$.

The distribution function of the scale mixtures of closed skew-normal distributions is as follows.

Theorem 4. *Suppose $W \sim SMCSN_{n,m}(\mu, K(\lambda)\Sigma, K(\lambda)^{-1/2}D, \nu, \Delta)$. Then the distribution function of W is given by*

$$F_W(w) = \int_0^\infty F_{\mathrm{CSN}}(w)dH(\lambda),$$

Table 1
Moments of mixing distributions for special cases of scale mixtures of closed skew-normal distributions.

Distribution	c_r
Closed skew-t	$\frac{(\nu/2)^{r/2}\Gamma\{(\nu-r)/2\}}{\Gamma(\nu/2)}$, $\nu > r$
Closed skew-logistic	$2^{1+r/2}\Gamma(r/2+1)\sum_{k=1}^{\infty}(-1)^{k+1}/k^r$
Closed skew-slash	$\frac{q}{q-r}$, $q > r$
Finite mixture of closed skew-normal	$\sum_{j=1}^{q} K(\lambda_j)^{r/2}p_j$
Contaminated closed skew-normal	$\frac{p}{\gamma^{r/2}} + 1 - p$

where

$$F_{\text{CSN}}(w) = C\Phi_{n+m}\left(\begin{pmatrix} w \\ 0 \end{pmatrix}; \begin{pmatrix} \mu \\ \nu \end{pmatrix}, \begin{pmatrix} K(\lambda)\Sigma & -K(\lambda)^{1/2}\Sigma D^T \\ -K(\lambda)^{1/2}D\Sigma & \Delta + D\Sigma D^T \end{pmatrix}\right),$$

and C is defined in (2).

2.3. A class of closed skew-t distributions

When $K(\lambda) = 1/\lambda$ and $\lambda \sim \text{Gamma}(\gamma/2, \gamma/2)$, we can derive a class of closed skew-t distributions which is equal to the skew-t distribution of Azzalini and Capitanio [8] and Branco and Dey [12] when $m = \Delta = 1$ and $\nu = 0$. In our parameterizations, $\text{Gamma}(\alpha, \beta)$ refers to a Gamma random variable with mean α/β and variance α/β^2. A preliminary result on Gamma distributions is required which is an extension of Lemma 11 of Azzalini and Capitanio [8].

Lemma 2. *If $\Lambda \sim \text{Gamma}(\alpha, \beta)$, then for any $a, b \in \mathbb{R}^n$*

$$E\{\Phi_n(\sqrt{\lambda}a + b)\} = P(T \le a\sqrt{\alpha/\beta}),$$

where T denotes a non-central multivariate t variate with 2α degrees of freedom and non-centrality parameter $-b$.

We now define a closed skew-t distribution using Lemma 2 on a Gamma variable with parameters $(\gamma/2, \gamma/2)$. Some simple algebra leads to the density of W

$$f_W(w) = Ct_n(w;\gamma)T_m\left\{\sigma R^{1/2}\Delta^{-1/2}D(w-\mu)\left(\frac{\gamma+n}{Q_w+\gamma}\right)^{1/2}\right\}, \tag{5}$$

where C is defined in (2), and

$$Q_w = (w-\mu)^T\Sigma^{-1}(w-\mu),$$
$$t_n(w;\gamma) = |\Sigma|^{-1/2}g_n(Q_w;\gamma) = \frac{\Gamma\{(\gamma+n)/2\}}{|\Sigma|^{1/2}(\pi\gamma)^{n/2}\Gamma(\gamma/2)}(1+Q_w/\gamma)^{-(\gamma+n)/2},$$

the density function of an n-dimensional t variate with γ degrees of freedom, and $T_m(\cdot)$ denotes an m-dimensional non-central multivariate t distribution function with $\gamma + n$ degrees of freedom and non-centrality parameter $\sigma R^{1/2}\Delta^{-1/2}\nu$. Here $\sigma > 0$ and R is a correlation matrix corresponding to a covariance matrix Δ, that is, $R = \rho^{-1}\Delta\rho^{-1}$, where $\rho = \text{diag}\{\delta_{11}, \ldots, \delta_{mm}\}$, and $\Delta = \{\delta_{ij}\}$ is an $m \times m$ covariance matrix. We obtain a closed skew-Cauchy distribution when $\gamma = 1$. When $\nu = 0$, $T_m(\cdot)$ appearing in (5) becomes an m-dimensional central multivariate t distribution function with $\gamma + n$ degrees of freedom.

Furthermore the closed skew-t distribution becomes the skew-t distribution of Azzalini and Capitanio [8] when $m = \Delta = 1$ and $\nu = 0$. So the closed skew-t distribution contains usual skew-t distributions obtained by Azzalini and Capitanio [8] and Branco and Dey [12]; see also [10,5,2].

3. Model and method

Consider the general state-space model of the following form: state equation

$$x_t = G_t x_{t-1} + \eta_t,$$

and observation equation

$$y_t = F_t x_t + \epsilon_t,$$

where η_t and ϵ_t are usually assumed to follow normal distributions independently. Here G_t and F_t are known matrices with dimensions $h \times h$ and $d \times h$, respectively. The x_t are unobserved state variables and the y_t are observations. Let

$\mathbf{y}_t = (y_1, \ldots, y_t)$ be the information available up to time t. There is a large literature on the estimation of the parameters for such models. For example, the Kalman filter provides an optimal way of estimating the model parameters. We use the term "Kalman filter" as a recursive procedure for inference about the state vector.

The usual Gaussian assumption is limited to express many different types of real data. It has been extended to non-Gaussian state-space models in the past. Smith and Miller [29], Bradley et al. [11], Meinhold and Singpurwalla [27], and Naveau et al. [28] proposed alternative methods to the classical Kalman filter. Naveau et al. [28] devised a skewed Kalman filter. Meinhold and Singpurwalla [27] used a multivariate distribution with Student t marginals. Bradley et al. [11] addressed the nonnormal situation with scale mixtures of normal distributions. In this sense, the latter approach is somewhat similar to ours, but our model contains their model since scale mixtures of closed skew-normal distributions contain scale mixtures of normal distributions as a special case. Smith and Miller [29] worked with exponential variables. Our approach is based on Naveau et al. [28] and we extend it to mixtures of skewed Kalman filters which also contain discrete mixtures of skewed Kalman filters.

3.1. Some preliminary results

In this section, we summarize some preliminary results required for developing mixtures of skewed Kalman filters. The first lemma studies closure under full row rank linear transformation of closed skew-normal distributions.

Lemma 3. *Let $y \sim CSN_{n,m}(\mu, \Sigma, D, \nu, \Delta)$ and let A be a $p \times n$ $(p \le n)$ matrix of rank p. Then*

$$Ay \sim CSN_{p,m}(\mu_A, \Sigma_A, D_A, \nu, \Delta_A),$$

where $\mu_A = A\mu$, $\Sigma_A = A\Sigma A^T$, $D_A = D\Sigma A^T \Sigma_A^{-1}$, and $\Delta_A = \Delta + (D - D_A A)\Sigma D^T$.

The proof is in [22]. Lemma 3 is useful for deriving marginal and conditional densities of closed skew-normal distributions as follows.

Lemma 4. *Let $y \sim CSN_{n,m}(\mu, \Sigma, D, \nu, \Delta)$ be partitioned into two components, y_1 and y_2, of dimensions h and $n-h$, respectively, and with a corresponding partition for μ, Σ, and D. Then the marginal distribution of y_1 is:*

$$y_1 \sim CSN_{h,m}(\mu_1, \Sigma_{11}, D^*, \nu, \Delta^*),$$

where $D^* = D_1 + D_2\Sigma_{21}\Sigma_{11}^{-1}$, $\Delta^* = \Delta + D_2\Sigma_{22.1}D_2^T$, $\Sigma_{22.1} = \Sigma_{22} - \Sigma_{21}\Sigma_{11}^{-1}\Sigma_{12}$,

$$\mu = \begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix}, \qquad \Sigma = \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix}, \quad \text{and} \quad D = \begin{pmatrix} D_1 \; D_2 \end{pmatrix}.$$

Furthermore the conditional distribution of y_2 given y_1 is:

$$y_2|y_1 \sim CSN_{n-h,m}(\mu_2 + \Sigma_{21}\Sigma_{11}^{-1}(y_1 - \mu_1), \Sigma_{22.1}, D_2, \nu - D^*(y_1 - \mu_1), \Delta).$$

The converse is also true.

The following lemma given by Naveau et al. [28] states that adding a normal noise to a closed skew-normal distribution does not change the distribution class. Note that a closed skew-normal distribution contains a normal distribution as a special case when $D = 0$.

Lemma 5. *Let $y \sim CSN_{n,m}(\mu, \Sigma, D, \nu, \Delta)$ and $z \sim N_n(\psi, \Omega)$ independently of y, then $y + z \sim CSN_{n,m}(\mu_{y+z}, \Sigma_{y+z}, D_{y+z}, \nu, \Delta_{y+z})$, where $\mu_{y+z} = \mu + \psi$, $\Sigma_{y+z} = \Sigma + \Omega$, $D_{y+z} = D\Sigma\Sigma_{y+z}^{-1}$, and $\Delta_{y+z} = \Delta + (D - D_{y+z})\Sigma D^T$.*

Since the normal distribution is a special case of a closed skew-normal distribution we have the following lemma which is useful for getting the joint distribution of independent normal and closed skew-normal distributions.

Lemma 6. *Let $y \sim CSN_{n,m}(\mu, \Sigma, D, \nu, \Delta)$ and $z \sim N_l(\psi, \Omega)$ independently of y, then $x = (y^T \; z^T)^T \sim CSN_{n+l,m}(\mu_x, \Sigma_x, D_x, \nu, \Delta)$, where*

$$\mu_x = \begin{pmatrix} \mu \\ \psi \end{pmatrix}, \qquad \Sigma_x = \begin{pmatrix} \Sigma & 0 \\ 0 & \Omega \end{pmatrix}, \qquad D_x = \begin{pmatrix} D & 0 \end{pmatrix}.$$

Remark. The distribution of x in Lemma 6 can also be written as, for any $p = 0, 1, 2, \ldots$, $\mathrm{CSN}_{n+l,p+m}(\mu_x, \Sigma_x, D_x, \nu_x, \Delta_x)$, where

$$\mu_x = \begin{pmatrix} \mu \\ \psi \end{pmatrix}, \qquad \Sigma_x = \begin{pmatrix} \Sigma & 0 \\ 0 & \Omega \end{pmatrix}, \qquad D_x = \begin{pmatrix} 0 & 0 \\ D & 0 \end{pmatrix}, \qquad \nu_x = \begin{pmatrix} 0 \\ \nu \end{pmatrix}, \qquad \Delta_x = \begin{pmatrix} I_p & 0 \\ 0 & \Delta \end{pmatrix}.$$

Except for μ_x and Σ_x, the parameters are in different forms even though we are using the same notation. Since

$$\Phi_{p+m}\left\{D_x(x-\mu_x);\, \nu_x,\, \Delta_x\right\} = \Phi_{p+m}\left\{\begin{pmatrix} 0 \\ D(y-\mu)\end{pmatrix};\, \nu_x,\, \Delta_x\right\}$$

and, by independence, it becomes $\Phi_p(0;0,I_p)\times\Phi_m\{D(y-\mu);\,\nu,\,\Delta\} = \left(\frac{1}{2}\right)^p \times \Phi_m\{D(y-\mu);\,\nu,\,\Delta\}$. By a similar argument we have that

$$\Phi_{p+m}\left(0;\, \nu_x,\, \Delta_x + D_x\Sigma_x D_x^T\right) = \left(\frac{1}{2}\right)^p \times \Phi_m(0;\,\nu,\,\Delta + D\Sigma D^T).$$

The remaining part of the proof is the same as for Lemma 6. Hence, by the definition of a closed skew-normal distribution, the result follows. Lemma 6 is the simplest case when $p = 0$.

3.2. Skewed state-space model

A skewed state-space model can be defined as follows:

$$\begin{aligned} x_t &= G_{\lambda_t}x_{t-1} + W_{\lambda_t}\eta_t, \\ y_t &= F_{\lambda_t}x_t + V_{\lambda_t}\epsilon_t, \end{aligned} \tag{6}$$

if $\Lambda_t = \lambda_t$, where $\eta_t \sim N_h(0,I)$ independent of $\epsilon_t \sim N_d(0,I)$ and all coefficient matrices are known given λ_t. Here G_{λ_t} and W_{λ_t} have the dimensions $h\times h$, F_{λ_t} and V_{λ_t} are $d\times h$ and $d\times d$, respectively. Define $\mathbf{\Lambda}_t = (\Lambda_1,\ldots,\Lambda_t)$. Let $\boldsymbol{\lambda}_t$ and λ_s be realizations of $\mathbf{\Lambda}_t$ and Λ_s, respectively. The Λ_t is a latent indicator process with a certain probabilistic structure. It can be either continuous or discrete. With discrete indicator variables, the model can be used to handle outliers, sudden jumps, environmental changes, etc. In this sense, our skewed state-space model does contain a linear state space model with regime switches when the state vector is finite and discrete [24]. Using continuous indicator variables, a skewed state-space model can deal with nonnormal innovations having both skewness and heavy tails.

So far there is no difference between a normal state-space model and a skewed state-space model. However introducing an initial distribution of x_0 as a closed skew-normal distribution, we have a difference. In this way we introduce the skewness. Using Lemmas 3 and 5, we derive the distributions of the state vector and the observation vector which are closed skew-normal distributions.

Theorem 5. *Let $x_0 \sim CSN_{h,m}(\psi_0, \Omega_0, D_0, \nu_0, \Delta_0)$ with a model (6). Then*

(i) $x_t|\boldsymbol{\lambda}_t \sim CSN_{h,m}(\psi_t, \Omega_t, D_t, \nu_t, \Delta_t)$, *where* $\psi_t = G_{\lambda_t}\psi_{t-1}$, $\Omega_t = G_{\lambda_t}\Omega_{t-1}G_{\lambda_t}^T + W_{\lambda_t}W_{\lambda_t}^T$, $D_t = D_{t-1}\Omega_{t-1}G_{\lambda_t}^T\Omega_t^{-1}$, $\nu_t = \nu_{t-1}$, *and* $\Delta_t = \Delta_{t-1} + (D_{t-1} - D_tG_{\lambda_t})\Omega_{t-1}D_{t-1}^T$.

(ii) $y_t|\boldsymbol{\lambda}_t \sim CSN_{d,m}(\mu_t, \Gamma_t, E_t, \gamma_t, \Theta_t)$, *where* $\mu_t = F_{\lambda_t}\psi_t$, $\Gamma_t = F_{\lambda_t}\Omega_tF_{\lambda_t}^T + V_{\lambda_t}V_{\lambda_t}^T$, $E_t = D_t\Omega_tF_{\lambda_t}^T\Gamma_t^{-1}$, $\gamma_t = \nu_t$, *and* $\Theta_t = \Delta_t + (D_t - E_tF_{\lambda_t})\Omega_tD_t^T$.

We use a Bayesian formulation [26] to derive the different Kalman filtering steps necessary to sequentially update the state of this model. Naveau et al. [28] also derived similar steps using their own lemmas with given trajectory $\boldsymbol{\lambda}_t$. We define $\mathbf{y}_t = (y_1,\ldots,y_t)$ as the information available up to time t.

Theorem 6. *Under model (6), let $x_0 \sim CSN_{h,m}(\psi_0, \Omega_0, D_0, \nu_0, \Delta_0)$, and $(x_{t-1}|\mathbf{y}_{t-1}, \boldsymbol{\lambda}_{t-1}) \sim CSN_{h,m}(\psi_{t-1}, \Omega_{t-1}, D_{t-1}, \nu_{t-1}, \Delta_{t-1})$, where*

$$\begin{aligned} &\psi_{t-1} = \psi_{t-1}(\boldsymbol{\lambda}_{t-1}), \qquad \Omega_{t-1} = \Omega_{t-1}(\boldsymbol{\lambda}_{t-1}), \qquad D_{t-1} = D_{t-1}(\boldsymbol{\lambda}_{t-1}), \\ &\nu_{t-1} = \nu_{t-1}(\boldsymbol{\lambda}_{t-1}), \qquad \Delta_{t-1} = \Delta_{t-1}(\boldsymbol{\lambda}_{t-1}). \end{aligned} \tag{7}$$

Then $(x_t|\mathbf{y}_t, \boldsymbol{\lambda}_t) \sim CSN_{h,m}(\psi_t, \Omega_t, D_t, \nu_t, \Delta_t)$, where

$$\begin{aligned} &\psi_t = G_{\lambda_t}\psi_{t-1} + \tilde{\Omega}_tF_{\lambda_t}^T\hat{\Omega}_t^{-1}e_t, \qquad \Omega_t = \tilde{\Omega}_t - \tilde{\Omega}_tF_{\lambda_t}^T\hat{\Omega}_t^{-1}F_{\lambda_t}\tilde{\Omega}_t, \\ &D_t = D_{t-1}\Omega_{t-1}G_{\lambda_t}^T\tilde{\Omega}_t^{-1}, \qquad \nu_t = \nu_{t-1} - D_t\tilde{\Omega}_tF_{\lambda_t}^T\hat{\Omega}_t^{-1}e_t, \\ &\Delta_t = \Delta_{t-1} + (D_{t-1} - D_tG_{\lambda_t})\Omega_{t-1}D_{t-1}^T, \\ &\tilde{\Omega}_t = W_{\lambda_t}W_{\lambda_t}^T + G_{\lambda_t}\Omega_{t-1}G_{\lambda_t}^T, \qquad \hat{\Omega}_t = V_{\lambda_t}V_{\lambda_t}^T + F_{\lambda_t}\tilde{\Omega}_tF_{\lambda_t}^T, \quad \textit{and} \\ &e_t = y_t - F_{\lambda_t}G_{\lambda_t}\psi_{t-1}. \end{aligned} \tag{8}$$

To emphasize the fact that all parameters depend on the mixing parameter $\boldsymbol{\lambda}_{t-1}$, we use the notation in (7). Obviously the parameters in (8) also depend on the mixing parameter $\boldsymbol{\lambda}_t$. We call the updating scheme of Theorem 6 as "the skewed Kalman filter" hereafter. By Theorem 2, we know that $G_{\lambda_t}\psi_{t-1}$ is not the mean of $(x_t|\mathbf{y}_{t-1}, \boldsymbol{\lambda}_t)$, it is a location parameter. So we define an error in predicting y_t from the previous time point $t-1$ using a location parameter, that is, $e_t = y_t - F_{\lambda_t}G_{\lambda_t}\psi_{t-1}$ in the proof. Instead of using a location parameter, we may use the mean of $(x_t|\mathbf{y}_{t-1}, \boldsymbol{\lambda}_t)$ as usual to define an error as follows:

$$e_t^* = y_t - F_{\lambda_t}E(x_t|\mathbf{y}_{t-1}, \boldsymbol{\lambda}_t), \tag{9}$$

where

$$E(x_t|\mathbf{y}_{t-1}, \boldsymbol{\lambda}_t) = G_{\lambda_t}\psi_{t-1} + C^*, \quad C^* = \frac{\Phi_m^*(0; \nu_t, \Delta_t + D_t\tilde{\Omega}_t D_t^T)}{\Phi_m(0; \nu_t, \Delta_t + D_t\tilde{\Omega}_t D_t^T)}, \text{ and}$$

$$\Phi_m^*(0; \nu_t, \Delta_t + D_t\tilde{\Omega}_t D_t^T) = \left.\frac{\partial}{\partial t}\Phi_m(D_t\tilde{\Omega}_t t; \nu_t, \Delta_t + D_t\tilde{\Omega}_t D_t^T)\right|_{t=0},$$

see formula (4) in [19].

Using (9) in Theorem 6, we still have the same updating scheme, so we call it the "invariance property" of the skewed Kalman filter summarized in the following corollary.

Corollary 1. *Under the same assumptions as Theorem 6 with (9), we have the same updating scheme as (8).*

3.3. The method of mixtures of skewed Kalman filters

For clarity we reintroduce some notations defined earlier. Let $\mathbf{y}_t = (y_1, \ldots, y_t)$, $\boldsymbol{\Lambda}_t = (\Lambda_1, \ldots, \Lambda_t)$, and let $\boldsymbol{\lambda}_t$ and λ_s be realizations of $\boldsymbol{\Lambda}_t$ and Λ_s, respectively. We use a Bayesian formulation [26] to develop the different steps of the mixtures of skewed Kalman filters. For the skewed Kalman filter, we observe that

$$P(x_{t-1}|\mathbf{y}_{t-1}) = \int P(x_{t-1}|\boldsymbol{\lambda}_{t-1}, \mathbf{y}_{t-1})P(\boldsymbol{\lambda}_{t-1}|\mathbf{y}_{t-1})d\boldsymbol{\lambda}_{t-1}, \tag{10}$$

where $(x_{t-1}|\mathbf{y}_{t-1}, \boldsymbol{\lambda}_{t-1}) \sim \text{CSN}_{h,m}(\psi_{t-1}, \Omega_{t-1}, D_{t-1}, \nu_{t-1}, \Delta_{t-1})$ with parameters in (7).

Then the parameters can be obtained by running the skewed Kalman filter with given trajectory $\boldsymbol{\lambda}_{t-1}$. The idea of the mixtures of skewed Kalman filters is to use a weighted sample of the indicators,

$$S_{t-1} = \{(\boldsymbol{\lambda}_{t-1}^{(1)}, w_{t-1}^{(1)}), \ldots, (\boldsymbol{\lambda}_{t-1}^{(r)}, w_{t-1}^{(r)})\},$$

to represent the distribution of $\boldsymbol{\Lambda}_{t-1}|\mathbf{y}_{t-1}$, and then to use a random mixture of closed skew-normal distributions,

$$\frac{1}{W_{t-1}}\sum_{j=1}^{r} w_{t-1}^{(j)}\text{CSN}_{h,m}(\psi_{t-1}, \Omega_{t-1}, D_{t-1}, \nu_{t-1}, \Delta_{t-1}),$$

where $W_{t-1} = \sum_{j=1}^{r} w_{t-1}^{(j)}$ and every parameter depends on $\boldsymbol{\lambda}_{t-1}^{(j)}$, to approximate the distribution of $x_{t-1}|\mathbf{y}_{t-1}$ for the skewed Kalman filter.

Inspired by Chen and Liu [14], we use the following method of mixtures of the skewed Kalman filter. They used the method of mixture Kalman filter (MKF) when the posterior distribution follows a mixture of normal distributions. We extend their method to the posterior distribution which follows a mixture of closed skew-normal distributions. Let $\text{SKF}_{t-1}^{(j)}$ record the posterior parameters of x_{t-1}, conditional on $\mathbf{y}_{t-1}$ and a given trajectory $\boldsymbol{\lambda}_{t-1}^{(j)}$. This can be obtained by the skewed Kalman filter. For the skewed Kalman filter, suppose that the initial state x_0 of the system defined by (6) follows a closed skew-normal distribution, $\text{CSN}_{h,m}(\psi_0, \Omega_0, D_0, \nu_0, \Delta_0)$. Then the mixtures of skewed Kalman filters updating scheme consist of recursive applications of the following steps.

For $j = 1, \ldots, r$:

(a) generate $\lambda_t^{(j)}$ from a trial distribution $q_t(\lambda_t|\boldsymbol{\lambda}_{t-1}^{(j)}, \text{SKF}_{t-1}^{(j)})$;

(b) obtain $\text{SKF}_t^{(j)}$ by the skewed Kalman filter, conditional on

$$\{\text{SKF}_{t-1}^{(j)}, y_t, \lambda_t^{(j)}\},$$

that is, $(x_t|\mathbf{y}_t, \boldsymbol{\lambda}_t) \sim \text{CSN}_{h,m}(\psi_t, \Omega_t, D_t, \nu_t, \Delta_t)$, where the parameters are given in (8);

(c) update the new weights as $w_t^{(j)} = w_{t-1}^{(j)} \times u_t^{(j)}$, where

$$u_t^{(j)} = \frac{p(\boldsymbol{\lambda}_{t-1}^{(j)}, \lambda_t^{(j)}|\mathbf{y}_t)}{p(\boldsymbol{\lambda}_{t-1}^{(j)}|\mathbf{y}_{t-1})q_t(\lambda_t^{(j)}|\boldsymbol{\lambda}_{t-1}^{(j)}, \text{SKF}_{t-1}^{(j)})};$$

(d) (resampling–rejuvenation) if the coefficient of variation of the w_t exceeds a threshold value, resample a new set of SKF_t from $\{\mathrm{SKF}_t^{(1)}, \ldots, \mathrm{SKF}_t^{(r)}\}$ with probability proportional to the weights $w_t^{(j)}$.

Remark that part (b) of the above updating scheme is derived from Theorem 6 for the skewed Kalman filter. We first study Λ_{t-1} in the discrete case and then the continuous case follows. First, when Λ_{t-1} has values in a finite discrete set I, then

(i) for each $\Lambda_t = i$, $i \in I$, run the skewed Kalman filter to obtain

$$v_i^{(j)} \propto p(y_t|\Lambda_t = i, \mathrm{SKF}_{t-1}^{(j)})p(\Lambda_t = i|\boldsymbol{\lambda}_{t-1}^{(j)}),$$

where $p(\Lambda_t = i|\boldsymbol{\lambda}_{t-1}^{(j)})$ is the prior transition probability for the indicator. For the skewed Kalman filter, $p(y_t|\Lambda_t = i, \mathrm{SKF}_{t-1}^{(j)})$ can be obtained using (23) and Lemma 3 with $A = (0_{d\times h}\ I_{d\times d})$ as follows:

$$p(y_t|\mathbf{y}_{t-1}, \boldsymbol{\lambda}_t) \sim \mathrm{CSN}_{d,m}(0, \hat{\Omega}_t, D_A, \nu_t, \Delta_A), \tag{11}$$

where

$$D_A = D_t\tilde{\Omega}_t F_{\lambda_t}^T \hat{\Omega}_t^{-1}, \qquad \Delta_A = \Delta_t + (D_t - D_A F_{\lambda_t})\tilde{\Omega}_t D_t^T,$$

and other related parameters are given in (8);

(ii) sample a $\lambda_t^{(j)}$ from the set I, with probability proportional to $v_i^{(j)}$;

(iii) let $\mathrm{SKF}_t^{(j)}$ be the one with $\Lambda_t = \lambda_t^{(j)}$;

(iv) the new weight is $w_t^{(j)} = w_{t-1}^{(j)} \sum_{i\in I} v_i^{(j)}$.

When Λ_t is a continuous random variable a simple algorithm is:

(i) sample a $\lambda_t^{(j)}$ from $p(\Lambda_t|\boldsymbol{\lambda}_{t-1}^{(j)})$, the prior structure of the indicator variable;

(ii) run one step of the skewed and extended Kalman filters on

$$\{\lambda_t^{(j)}, \mathrm{SKF}_{t-1}^{(j)}, y_t\}$$

to obtain $\mathrm{SKF}_t^{(j)}$, using Eqs. (8) for the skewed Kalman filter;

(iii) the new weight is $w_t^{(j)} = w_{t-1}^{(j)}\, p(y_t|\lambda_t^{(j)}, \mathrm{SKF}_{t-1}^{(j)})$ using Eqs. (11) for the skewed Kalman filter.

Chen and Liu [14] explained some more algorithmic approaches about their mixture Kalman filtering of conditional dynamic linear models for discrete and continuous Λ_{t-1}.

For applying mixtures of skewed Kalman filters to the case of scale mixtures of skewed Kalman filters, we only need to replace some parameters of $P(x_{t-1}|\boldsymbol{\lambda}_{t-1}, \mathbf{y}_{t-1})$ in (10). That is, in (7),

$$\Omega_{t-1} = K(\boldsymbol{\lambda}_{t-1})\Omega_{t-1}, D_{t-1} = K(\boldsymbol{\lambda}_{t-1})^{-1/2}D_{t-1},$$

for the skewed Kalman filter. Then we change all the relevant parameters of the updating schemes. Note that mixtures of skewed Kalman filters contain scale mixtures of skewed Kalman filters as a special case.

3.4. Extended state-space model

The updating scheme appearing in Section 3.2 introduces skewness via the initial state x_0, but it would be better to include skewness at each time step. Let

$$y_t = F_{\lambda_t}x_t + V_{\lambda_t}\epsilon_t = Q_{\lambda_t}u_t + P_{\lambda_t}s_t + V_{\lambda_t}\epsilon_t, \tag{12}$$

for $\Lambda_t = \lambda_t$, where $F_{\lambda_t} = (Q_{\lambda_t}\ P_{\lambda_t})$, $x_t = (u_t^T\ s_t)^T$, and $\epsilon_t \sim N_d(0, I_d)$. The random vector u_t of length k and the scalar matrix Q_{λ_t} of dimension $d \times k$ represent the linear part of the observation equation. The random s_t of length 1 and the scalar matrix P_{λ_t} of dimension $d \times 1$ correspond to the additional skewness. All coefficient matrices are known given λ_t. Assume that

$$\begin{aligned} u_t &= K_{\lambda_t}u_{t-1} + H_{\lambda_t}\eta_t^*, \\ v_t &= -l_{\lambda_t}v_{t-1} + \sigma_{\lambda_t}\eta_t^+, \end{aligned} \tag{13}$$

for $\Lambda_t = \lambda_t$, where the normal error $\eta_t^* \sim N_k(0, I_k)$ is independent of $\eta_t^+ \sim N(0, 1)$. Here K_{λ_t} and H_{λ_t} have the dimension $k \times k$, whereas l_{λ_t} and σ_{λ_t} are constant. We also assume that ϵ_t is independent of η_t^* and η_t^+. Hence the joint distribution of $(u_t^T\ v_t)^T$ given λ_t, can be expressed as

$$\left(\begin{matrix} u_t \\ v_t \end{matrix}\middle| \lambda_t\right) \sim N_{k+1}\left(\begin{pmatrix} \psi_t^* \\ \psi_t^+ \end{pmatrix}\begin{pmatrix} \Omega_t^* & 0 \\ 0 & \sigma_t^{2+} \end{pmatrix}\right). \tag{14}$$

The following lemma is a cornerstone to develop another version of the skewed state-space model, namely the extended state-space model.

Lemma 7. *For $\boldsymbol{\Lambda}_t = \boldsymbol{\lambda}_t$, let $D_t^+ = \sigma_{t-1}^{2+} l_{\lambda_t}/\sigma_t^{2+}$, $\psi_t^+ = -l_{\lambda_t}\psi_{t-1}^+$, and $\Phi_t(\cdot) = \Phi(\cdot;\psi_t^+,\sigma_t^{2+})$ be the univariate normal cdf with mean ψ_t^+ and variance σ_t^{2+}. We define the skewness part s_t as*

$$s_t = \sigma_{\lambda_t}\eta_t^+ - l_{\lambda_t} w_{t-1}, \tag{15}$$

where w_{t-1} is as follows.
If $D_t^+\psi_t^+ \le \psi_{t-1}^+$, then

$$w_{t-1} = \begin{cases} v_{t-1} & \text{if } v_{t-1} \le D_t^+\psi_t^+, \\ 2\psi_{t-1}^+ - v_{t-1} & \text{if } v_{t-1} \ge 2\psi_{t-1}^+ - D_t^+\psi_t^+, \\ \Phi_{t-1}^{-1}\left[\dfrac{\Phi_{t-1}(D_t^+\psi_t^+)\left\{\Phi_{t-1}(v_{t-1}) - \Phi_{t-1}(D_t^+\psi_t^+)\right\}}{\Phi_{t-1}(2\psi_{t-1}^+ - D_t^+\psi_t^+) - \Phi_{t-1}(D_t^+\psi_t^+)}\right] & \text{otherwise,} \end{cases} \tag{16}$$

$$w_{t-1} = \begin{cases} v_{t-1} & \text{if } v_{t-1} \le D_t^+\psi_t^+, \\ \Phi_{t-1}^{-1}\left[\dfrac{\Phi_{t-1}(D_t^+\psi_t^+)\left\{\Phi_{t-1}(v_{t-1}) - \Phi_{t-1}(D_t^+\psi_t^+)\right\}}{1 - \Phi_{t-1}(D_t^+\psi_t^+)}\right] & \text{otherwise.} \end{cases}$$

Then $s_t \sim CSN_{1,1}(\psi_t^+, \sigma_t^{2+}, D_t^+, \nu_t^+, \Delta_t^+)$, where $\sigma_t^{2+} = l_{\lambda_t}^2\sigma_{t-1}^{2+} + \sigma_{\lambda_t}^2$, $\nu_t^+ = \psi_{t-1}^+ - D_t^+\psi_t^+$, and $\Delta_t^+ = \sigma_{t-1}^{2+} - (D_t^+)^2\sigma_t^{2+}$.

From the proof of the above lemma we find that $w_{t-1} \stackrel{d}{=} (v_{t-1}|v_{t-1} \le c)$ and $s_t \stackrel{d}{=} (v_t|v_{t-1} \le c)$ for $\Lambda_t = \lambda_t$, where $c = D_t^+\psi_t^+$ and $\stackrel{d}{=}$ denotes the equality of distributions hereafter. Instead of a simple stochastic relationship, $(v_{t-1}|v_{t-1} \le c)$, we define w_{t-1} as in Lemma 7 since simulating from $(v_t|v_{t-1} \le c)$ depends on the set $\{v_{t-1} \le c\}$, but we do not need the rejection algorithm for newly defined w_{t-1}. Hence, the variable s_t through a constant l_{λ_t} introduces at each time step a different skewness in the state vector. The price for this gain in skewness flexibility is that this state vector does not have anymore a linear structure. If $P_{\lambda_t} = 0$, $l_{\lambda_t} = \sigma_{\lambda_t} = 0$, then this extended state-space model becomes a skewed state-space model or standard state-space model depending on the initial distribution. We now derive the distributions of $x_t = (u_t^T\ s_t)^T$ and y_t which result in closed skew-normal distributions.

Theorem 7. *Let the initial vector $(u_0^T\ v_0)^T$ defined by (13) follow*

$$N_{k+1}\left(\begin{pmatrix}\psi_0^*\\ \psi_0^+\end{pmatrix}\begin{pmatrix}\Omega_0^* & 0\\ 0 & \sigma_0^{2+}\end{pmatrix}\right).$$

Then $x_t = (u_t^T\ s_t)^T$ and y_t defined by (13)–(15) and (12) follow $x_t|\boldsymbol{\lambda}_t \sim CSN_{k+1,1}(\psi_t, \Omega_t, D_t, \nu_t, \Delta_t)$ and $y_t|\boldsymbol{\lambda}_t \sim CSN_{d,1}(\mu_t, \Gamma_t, E_t, \gamma_t, \Theta_t)$ for $t \ge 1$, respectively. The parameters are related as follows:

$$\psi_t = \begin{pmatrix}\psi_t^*\\ \psi_t^+\end{pmatrix}, \qquad \Omega_t = \begin{pmatrix}\Omega_t^* & 0\\ 0 & \sigma_t^{2+}\end{pmatrix}, \qquad D_t = (0\ D_t^+),\ \nu_t = \nu_t^+, \qquad \Delta_t = \Delta_t^+,$$

$$\psi_t^* = K_{\lambda_t}\psi_{t-1}^*, \qquad \psi_t^+ = -l_{\lambda_t}\psi_{t-1}^+, \qquad \Omega_t^* = K_{\lambda_t}\Omega_{t-1}^*K_{\lambda_t}^T + H_{\lambda_t}H_{\lambda_t}^T,$$

$$\sigma_t^{2+} = l_{\lambda_t}^2\sigma_{t-1}^{2+} + \sigma_{\lambda_t}^2, \qquad D_t^+ = l_{\lambda_t}\sigma_{t-1}^{2+}/\sigma_t^{2+}, \qquad \nu_t^+ = \psi_{t-1}^+ - D_t^+\psi_t^+,$$

$$\Delta_t^+ = \sigma_{t-1}^{2+} - (D_t^+)^2\sigma_t^{2t},$$

$$\mu_t = F_{\lambda_t}\psi_t, \qquad \Gamma_t = F_{\lambda_t}\Omega_t F_{\lambda_t}^T + V_{\lambda_t}V_{\lambda_t}^T, \qquad E_t = D_t\Omega_t F_{\lambda_t}^T\Gamma_t^{-1}, \qquad \gamma_t = \nu_t^+,$$

$$\text{and} \quad \Theta_t = \Delta_t^+ + (D_t - E_tF_{\lambda_t})\Omega_t D_t^T.$$

For a Bayesian sequential formulation [26], we assume that

$$\left(\begin{matrix}u_{t-1}\\ v_{t-1}\end{matrix}\middle|\,\mathbf{y}_{t-1}, \boldsymbol{\lambda}_{t-1}\right) \sim N_{k+1}\left(\begin{pmatrix}\psi_{t-1}^*\\ \psi_{t-1}^+\end{pmatrix}\begin{pmatrix}\Omega_{t-1}^* & 0\\ 0 & \sigma_{t-1}^{2+}\end{pmatrix}\right), \quad t = 1, 2, \ldots. \tag{17}$$

Note that (17) holds for $t = 1$ which is an initial time. So, similar to Section 3.2, we develop the different steps of the Kalman filtering. We use an independent multivariate normal distributional assumption in (17) whereas Naveau et al. [28] used a dependent multivariate normal distribution.

Theorem 8. *Let the extended state-space model defined by (12), (13) and (15) satisfy (17). Then the posterior distribution of $(x_t|\mathbf{y}_t, \boldsymbol{\lambda}_t)$ follows a closed skew-normal distribution, $CSN_{k+1,1}(\psi_t, \Omega_t, D_t, \nu_t, \Delta_t)$ with*

$$\psi_t = \begin{pmatrix}\psi_t^*\\ \psi_t^+\end{pmatrix}, \qquad \Omega_t = \begin{pmatrix}\Omega_t^* & \Omega_t^{*+}\\ (\Omega_t^{*+})^T & \sigma_t^{2+}\end{pmatrix}, \qquad D_t = \begin{pmatrix}0 & \tilde{D}_t^+\end{pmatrix},$$

$$\nu_t = \tilde{\nu}_t^+ - \tilde{\sigma}_t^{2+}\tilde{D}_t^+P_{\lambda_t}^T\Sigma_t^{-1}(y_t - Q_{\lambda_t}K_{\lambda_t}\psi_{t-1}^* - P_{\lambda_t}\tilde{\psi}_t^+), \quad \text{and}$$

$$\Delta_t = \tilde{\Delta}_t^+. \tag{18}$$

The parameters of the posterior distribution are updated as follows:

$$\begin{pmatrix}\psi_t^*\\ \psi_t^+\end{pmatrix} = \begin{pmatrix} K_{\lambda_t}\psi_{t-1}^* + \tilde{\Omega}_t^* Q_{\lambda_t}^T \Sigma_t^{-1}(y_t - Q_{\lambda_t}K_{\lambda_t}\psi_{t-1}^* - P_{\lambda_t}\tilde{\psi}_t^+)\\ -l_{\lambda_t}\psi_{t-1}^+ + \tilde{\sigma}_t^{2+}P_{\lambda_t}^T\Sigma_t^{-1}(y_t - Q_{\lambda_t}K_{\lambda_t}\psi_{t-1}^* - P_{\lambda_t}\tilde{\psi}_t^+)\end{pmatrix},$$

$$\begin{pmatrix}\Omega_t^*\\ \sigma_t^{2+}\\ \Omega_t^{*+}\end{pmatrix} = \begin{pmatrix}\tilde{\Omega}_t^* - \tilde{\Omega}_t^* Q_{\lambda_t}^T\Sigma_t^{-1}Q_{\lambda_t}\tilde{\Omega}_t^*\\ \tilde{\sigma}_t^{2+}(1 - \tilde{\sigma}_t^{2+}P_{\lambda_t}^T\Sigma_t^{-1}P_{\lambda_t})\\ -\tilde{\sigma}_t^{2+}\tilde{\Omega}_t^* Q_{\lambda_t}^T\Sigma_t^{-1}P_{\lambda_t}\end{pmatrix},$$

$\tilde{\Omega}_t^* = K_{\lambda_t}\Omega_{t-1}^* K_{\lambda_t}^T + H_{\lambda_t}H_{\lambda_t}^T$, *and* $\Sigma_t = Q_{\lambda_t}\tilde{\Omega}_t^* Q_{\lambda_t}^T + \tilde{\sigma}_t^{2+}P_{\lambda_t}P_{\lambda_t}^T + V_{\lambda_t}V_{\lambda_t}^T$. *Intermediate parameters are*

$$\tilde{\psi}_t^+ = -l_{\lambda_t}\psi_{t-1}^+, \qquad \tilde{\sigma}_t^{2+} = l_{\lambda_t}^2\sigma_{t-1}^{2+} + \sigma_{\lambda_t}^2, \qquad \tilde{D}_t^+ = l_{\lambda_t}\sigma_{t-1}^{2+}/\tilde{\sigma}_t^{2+},$$
$$\tilde{\nu}_t^+ = \psi_{t-1}^+ - \tilde{D}_t^+\tilde{\psi}_t^+, \quad \text{and} \quad \tilde{\Delta}_t^+ = \sigma_{t-1}^{2+} - (\tilde{D}_t^+)^2\tilde{\sigma}_t^{2+}.$$

We call the updating scheme of Theorem 8 "the extended skewed Kalman filter" hereafter. In Theorem 8, we used the expectations $E(u_t|\mathbf{y}_{t-1}, \boldsymbol{\lambda}_t)$ and $E(s_t|\mathbf{y}_{t-1}, \boldsymbol{\lambda}_t)$ to define e_t which denotes the error in predicting y_t from the time point $t-1$. Instead of the expectations, we may use the location parameters of $(u_t|\mathbf{y}_{t-1}, \boldsymbol{\lambda}_t)$ and $(s_t|\mathbf{y}_{t-1}, \boldsymbol{\lambda}_t)$ which are $K_{\lambda_t}\psi_{t-1}^*$ and $\tilde{\psi}_t^+ = -l_{\lambda_t}\psi_{t-1}^+$. Remark that $E(u_t|\mathbf{y}_{t-1}, \boldsymbol{\lambda}_t)$ is equal to the location parameter, but $E(s_t|\mathbf{y}_{t-1}, \boldsymbol{\lambda}_t)$ is not equal to the location parameter. Define a new error as $e_t^* = y_t - Q_{\lambda_t}K_{\lambda_t}\psi_{t-1}^* - P_{\lambda_t}\tilde{\psi}_t^+$. Although we use a different error, we still have the same updating scheme (18) as follows.

Corollary 2. *Under the same assumptions as Theorem 8 except for a new error* $e_t^* = y_t - Q_{\lambda_t}K_{\lambda_t}\psi_{t-1}^* - P_{\lambda_t}\tilde{\psi}_t^+$, *we have the same updating scheme as (18).*

Similar to Corollary 1, we call this property "invariance property" of the extended skewed Kalman filter.

If we compare the skewness parameters of Theorems 6 and 8, it is straightforward to find the difference. For the former model, $D_t = D_{t-1}\Omega_{t-1}G_{\lambda_t}^T\tilde{\Omega}_t^{-1}$ so D_t becomes 0 as soon as $D_{t-1} = 0$. However for the latter model, D_t is not directly related to the previous skewness parameter D_{t-1} so it can be very different from the previous skewness. In this way the skewness is introduced via l_{λ_t} at each time point. One good point of the independence assumption (17) is that we do not need to calculate some means and covariances using any numerical techniques, for example in their notation C_t, $\tau_t^{(i)}$, $i = 1, 2$ [28].

Using the assumptions of Theorem 8 and Lemma 7, it can be proved that the distribution of $x_{t-1}|\mathbf{y}_{t-1}, \boldsymbol{\lambda}_{t-1}$ is:

$$\mathrm{CSN}_{k+1,1}(\psi_{t-1}, \Omega_{t-1}, D_{t-1}, \nu_{t-1}, \Delta_{t-1}) \tag{19}$$

with

$$\psi_{t-1} = \begin{pmatrix}\psi_{t-1}^*\\ \tilde{\psi}_{t-1}^+\end{pmatrix}, \qquad \Omega_{t-1} = \begin{pmatrix}\Omega_{t-1}^* & 0\\ 0 & \tilde{\sigma}_{t-1}^{2+}\end{pmatrix}, \qquad D_{t-1} = \begin{pmatrix}0 & \tilde{D}_{t-1}^+\end{pmatrix},$$
$$\nu_{t-1} = \tilde{\nu}_{t-1}^+, \quad \text{and} \quad \Delta_{t-1} = \tilde{\Delta}_{t-1}^+.$$

As a by-product the distribution of $(x_{t-1}|\mathbf{y}_{t-1}, \boldsymbol{\lambda}_{t-1})$ is derived and this distribution is not used directly to derive the distribution of $(x_t|\mathbf{y}_t, \boldsymbol{\lambda}_t)$ since the state vector x_t does not have a linear structure like the skewed Kalman filter and the normal Kalman filter. So we cannot use the EMKF algorithm of Chen and Liu [14] who suggested extended mixture Kalman filters which can be used for implementing the extended skewed Kalman filter. However, given a trajectory $\boldsymbol{\lambda}_t$, our suggested filter still works well similar to Naveau et al. [28].

3.5. The method of mixtures of extended skewed Kalman filters

To implement mixtures of the extended skewed Kalman filter, we introduce the extended mixture Kalman filters suggested by Chen and Liu [14]. They introduced the partial conditional dynamic linear models (PCDLM). The concept of PCDLM is such that the state and observation equations given the non-linear component are linear in state and observed variables and have (complex) Gaussian errors. They then devised the extended mixture Kalman filter (EMKF) to implement PCDLM.

Chen and Liu [14] also gave some examples of PCDLM. For example, the Rayleigh flat fading channel has the following form: state equations

$$\mathbf{x}_t = F\mathbf{x}_{t-1} + Ww_t,$$
$$\alpha_t = G\mathbf{x}_t,$$
$$s_t \sim p(\cdot|s_{t-1});$$

and the observation equation

$$y_t = \alpha_t s_t + Vv_t,$$

where s_t are the input digital signals (symbols), y_t are the received complex signals and α_t are the unobserved (changing) fading coefficients. Both w_t and v_t are complex Gaussian with identity covariance matrices. This system is clearly PCDLM since the system is linear in $\mathbf{x}_t$ and y_t given the input signals s_t.

The extended state-space model defined by (12), (13) and (15) is a special case of PCDLM. The system is linear in u_t and y_t^* and has Gaussian errors since, given the nonlinear component s_t which is equally distributed as $(v_t|v_{t-1} \le c)$, it has the following form: state equation

$$u_t = K_{\lambda_t} u_{t-1} + H_{\lambda_t} \eta_t^*,$$

and the observation equation

$$y_t^* = y_t - P_{\lambda_t} s_t = Q_{\lambda_t} u_t + V_{\lambda_t} \epsilon_t.$$

Thus, in the EMKF we approximate the joint distribution of $p(u_t, s_t|\mathbf{y}_t)$ as a Monte Carlo approximation of the marginal distribution $p(s_t|\mathbf{y}_t)$ and an exact Gaussian conditional distribution $p(u_t|s_t, \mathbf{y}_t)$. We introduce the algorithm that appeared in [14] for completeness. Let $\mathbf{s}_{t-1} = (s_1, s_2, \dots, s_{t-1})$. Suppose that at time $t-1$ there is a sample

$$(\boldsymbol{\lambda}_{t-1}^{(j)}, \mathbf{s}_{t-1}^{(j)}, KF_{t-1}^{(j)}, w_{t-1}^{(j)}), \quad j = 1, 2, \dots, r,$$

where $KF_{t-1}^{(j)}$ denotes the mean and the covariance matrix of $p(u_{t-1}|\boldsymbol{\lambda}_{t-1}^{(j)},\ s_{t-1}^{(j)}, \mathbf{y}_{t-1})$ obtained by the Kalman filter. By the assumption of the extended skewed Kalman filter, this distribution is expressed as follows:

$$(u_{t-1}|\boldsymbol{\lambda}_{t-1}^{(j)}, s_{t-1}^{(j)}, \mathbf{y}_{t-1}) \sim N_k(\psi_{t-1}^*, \Omega_{t-1}^*). \tag{20}$$

The EMKF updating algorithm recursively applies the following steps.
For $j = 1, \dots, r$:

(a) generate $(\lambda_t^{(j)}, s_t^{(j)})$ from a trial distribution $q_t(\lambda_t, s_t|\boldsymbol{\lambda}_{t-1}^{(j)}, s_{t-1}^{(j)}, KF_{t-1}^{(j)})$;
(b) obtain $KF_t^{(j)}$ by a one-step Kalman filter, conditional on $\{\lambda_t^{(j)}, s_t^{(j)},\ KF_{t-1}^{(j)}, y_t\}$,

$$\begin{aligned} R_t &= K_{\lambda_t} \Omega_{t-1}^* K_{\lambda_t}^T + H_{\lambda_t} H_{\lambda_t}^T, \\ S_t &= Q_{\lambda_t} R_t Q_{\lambda_t}^T + V_{\lambda_t} V_{\lambda_t}^T, \\ \psi_t^* &= K_{\lambda_t} \psi_{t-1}^* + R_t Q_{\lambda_t}^T S_t^{-1} (y_t^* - Q_{\lambda_t} K_{\lambda_t} \psi_{t-1}^*), \\ \Omega_t^* &= R_t - R_t Q_{\lambda_t}^T S_t^{-1} Q_{\lambda_t} R_t; \end{aligned} \tag{21}$$

(c) evaluate the incremental weight

$$u_t^{(j)} = \frac{p(\boldsymbol{\lambda}_{t-1}^{(j)}, s_t^{(j)}, \lambda_t^{(j)}|\mathbf{y}_t)}{p(\boldsymbol{\lambda}_{t-1}^{(j)}, s_{t-1}^{(j)}|\mathbf{y}_{t-1}) q_t(\lambda_t^{(j)}, s_t^{(j)}|\boldsymbol{\lambda}_{t-1}^{(j)}, s_{t-1}^{(j)}, KF_{t-1}^{(j)})} \tag{22}$$

and update the new weight as $w_t^{(j)} = w_{t-1}^{(j)} u_t^{(j)}$;
(d) (resampling–rejuvenation) if the coefficient of variation of the w_t exceeds a threshold value, re-sample a new set of KF_t from $\{KF_{t-1}^{(1)}, \dots,\ KF_{t-1}^{(r)}\}$ with probability proportional to the weights $w_t^{(j)}$.

For applying mixtures of extended skewed Kalman filters to the case of the scale mixtures of extended skewed Kalman filters, we only need to replace the covariance matrix of $P(u_{t-1}|\boldsymbol{\lambda}_{t-1}, s_{t-1}, \mathbf{y}_{t-1})$ in (20). That is,

$$\Omega_{t-1}^* = K(\boldsymbol{\lambda}_{t-1}) \Omega_{t-1}^*.$$

Then we change all relevant parameters of the updating schemes in (21). Mixtures of extended skewed Kalman filters contain scale mixtures of extended skewed Kalman filters as a special case. To calculate the incremental weight in (22) it is straightforward to find some useful facts. Because

$$p(\boldsymbol{\lambda}_{t-1}, s_t, \lambda_t|\mathbf{y}_t) = p(s_t|\boldsymbol{\lambda}_t, \mathbf{y}_t) \times p(\boldsymbol{\lambda}_t|\mathbf{y}_t),$$

we can use

$$(s_t|\mathbf{y}_t, \boldsymbol{\lambda}_t) \sim \mathrm{CSN}_{1,1}(\psi_t^+, \sigma_t^{2+}, \tilde{D}_t^+, \nu_t, \Delta_t)$$

by Theorem 8 and Lemma 4. Similarly

$$p(\boldsymbol{\lambda}_{t-1}, s_{t-1}|\mathbf{y}_{t-1}) = p(s_{t-1}|\boldsymbol{\lambda}_{t-1}, \mathbf{y}_{t-1}) \times p(\boldsymbol{\lambda}_{t-1}|\mathbf{y}_{t-1}),$$

where $(s_{t-1}|\mathbf{y}_{t-1}, \boldsymbol{\lambda}_{t-1})$ follows $\mathrm{CSN}_{1,1}(\tilde{\psi}_{t-1}^+, \tilde{\sigma}_{t-1}^{2+}, \tilde{D}_{t-1}^+, \tilde{\nu}_{t-1}^+, \tilde{\Delta}_{t-1}^+)$ by (19). To finish the implementation we only need a good trial distribution.

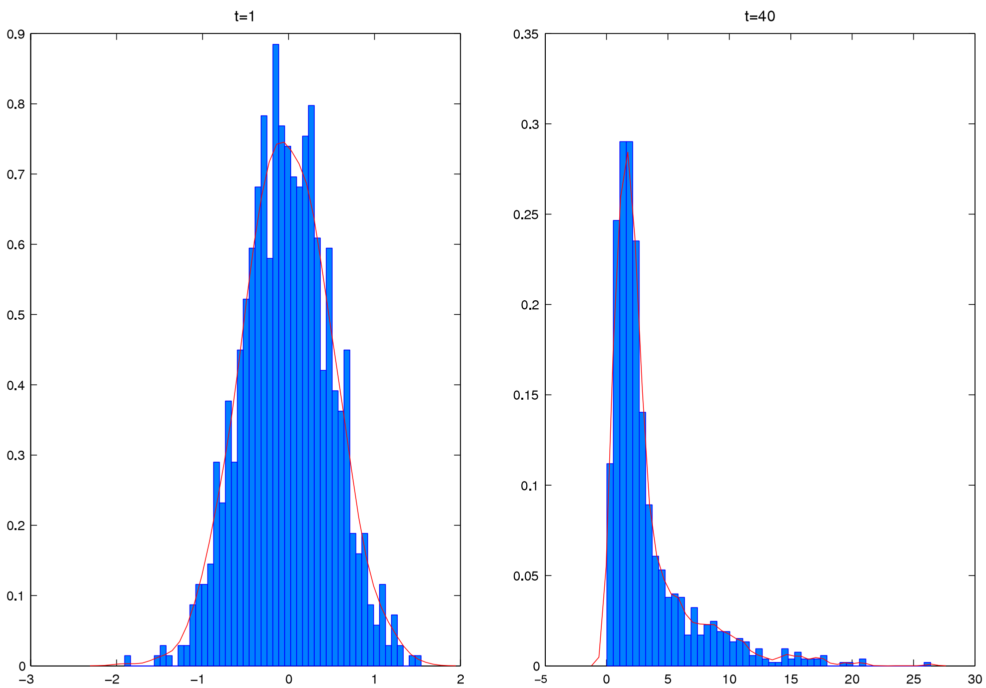


Fig. 1. Density of y_t with histograms from simulated values. The left panel corresponds to the time $t = 1$ and the right panel to the time $t = 40$ for the parameters described in Fig. 2. The red lines are estimated density curves through kernel density estimation. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

4. Numerical experiments

4.1. Simulation study for mixtures of skewed Kalman filters

To illustrate the distribution driven by the state-space model (6), two histograms of y_t were plotted in Fig. 1, where the distribution is slightly skewed at time $t = 1$ and the distribution is heavily skewed at $t = 40$. The simulation data have been collected by assuming $\eta_t \sim N(0, 0.1^2)$ and $\epsilon_t \sim G(3, 3)$. The latent indicator $\Lambda_t \in \{0, 1\}$ at time t has been assumed from a discrete distribution such that $P(\Lambda_t = 0|\Lambda_{t-1} = 0) = 0.99$, $P(\Lambda_t = 1|\Lambda_{t-1} = 0) = 0.01$, $P(\Lambda_t = 0|\Lambda_{t-1} = 1) = 0.01$, and $P(\Lambda_t = 1|\Lambda_{t-1} = 1) = 0.99$. For $\Lambda_t = \lambda_t$, the coefficients in (6) have been set by $F_{\lambda_t} = G_{\lambda_t} = \lambda_t'$, $V_{\lambda_t} = 1 - \lambda_t'$, and $W_{\lambda_t} = 1 + \lambda_t'$, where $\lambda_t' = 0.2 + 0.6\lambda_t$. Fig. 2 describes the temporal evolution of the coefficients.

In the procedure of mixture of skewed Kalman filters, a flexible prior transition probability has been considered $P(\Lambda_t = 0|\Lambda_{t-1} = 0) = P(\Lambda_t = 1|\Lambda_{t-1} = 0) = P(\Lambda_t = 0|\Lambda_{t-1} = 1) = P(\Lambda_t = 1|\Lambda_{t-1} = 1) = 0.5$. In Fig. 3, the solid line represents the observed path for x_t, the circles denote the estimated $\hat{x}_t$ from the mixture of Kalman filters as a reference, and the dots indicate the mixture of skewed Kalman filters, respectively. At the beginning, the two estimators show similar performance, but as the time passes ($t > 30$), the mixture of skewed Kalman filters captures the true pattern better than the mixture of Kalman filters. Furthermore, as shown in Fig. 4, the mixture of skewed Kalman filters captures y_t correctly and quickly, while the mixture of Kalman filters is biased. We set $\nu_t = 0$ for every time point since the closed skew-normal distribution is not identifiable [2]. The same approach was used by Flecher et al. [16] and González-Farías et al. [21].

4.2. Simulation study for extended skewed Kalman filters and mixtures of those

The variable s_t, defined at (15), through l_{λ_t} introduces a different skewness at each time point. To illustrate the distribution of the skewness, s_t, two histograms of s_t are plotted in Fig. 5 at two different time points: $t = 1$ (slight skewness, see left panel) and $t = 40$ (large skewness, see right panel). These simulated data were generated by setting $F_{\lambda_t} = P_{\lambda_t} = (-1)^t/2$, $V_{\lambda_t} = 1$, $Q_{\lambda_t} = K_{\lambda_t} = H_{\lambda_t} = 0$, and $\sigma_{\lambda_t} = 1$. The other parameters were set according to Fig. 6 which explain the temporal evolution of l_{λ_t}, σ_t^{2+}, D_t^+, and Δ_t^+. The two remaining parameters are set to $\psi_t^+ = \nu_t^+ = 0$. To explain the difference between the classical Gaussian Kalman filter and the extended skewed Kalman filter, two filters were used to estimate the temporal evolution of the state vector s_t from simulated observations y_t. In Fig. 7, the extended skewed Kalman filters captures s_t correctly and quickly, while the Kalman filter is severely biased. The root mean-squared errors at a given time point, $\sqrt{\sum_{i=1}^{1000}\{s(i)_t - \hat{s}(i)_t\}^2/1000}$, are plotted in Fig. 8 for both filters. This plot obviously indicates that the classical Kalman filter lost some efficiency when skewness was introduced. For the simulation we used the result of Theorem 8. We

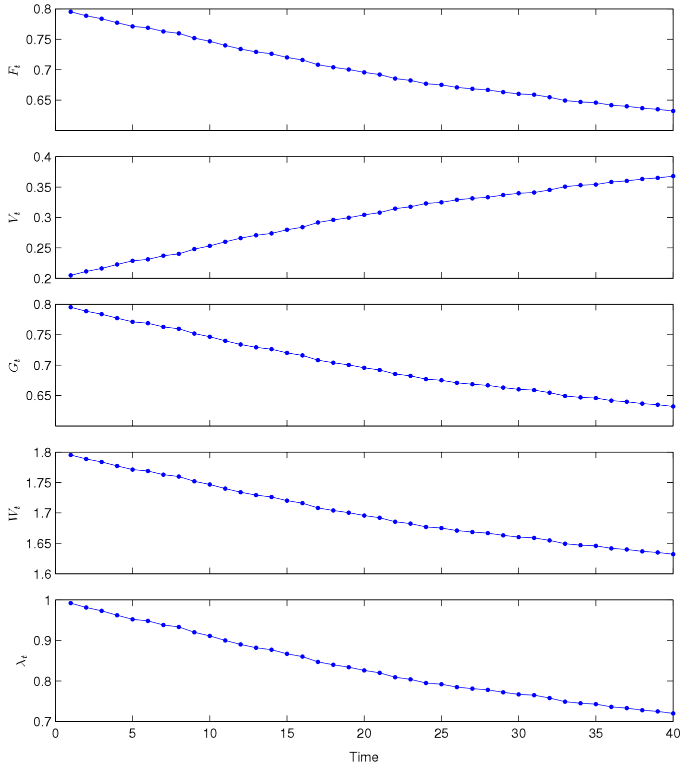


Fig. 2. Temporal evolution of the parameters used to simulate y_t in Fig. 1. We set $F_t = G_t = \lambda'_t$, $V_{\lambda_t} = 1 - \lambda'_t$, $W_{\lambda_t} = 1 + \lambda'_t$, and $\lambda'_t = 0.6\lambda_t + 0.2$.

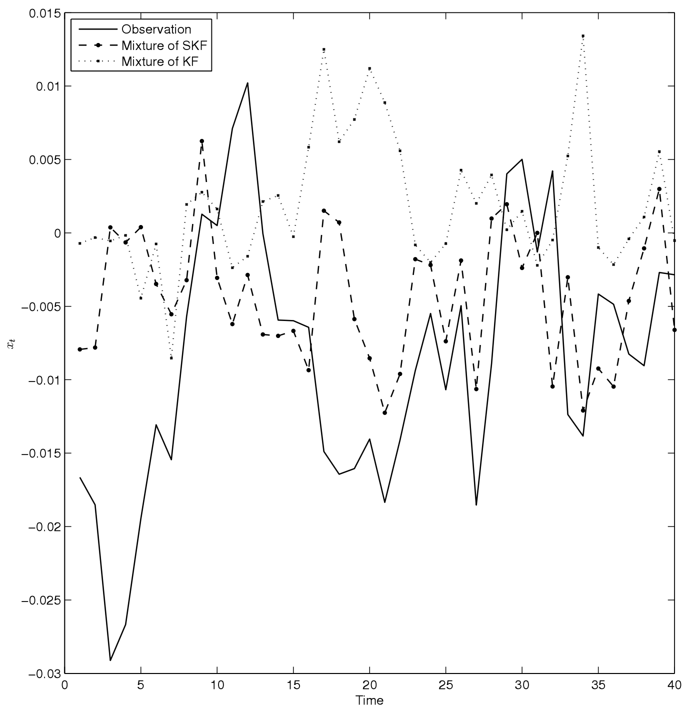


Fig. 3. Estimation of the temporal evolution of the state-space variable x_t by using the mixture of Kalman filters (dotted) and the mixture of skewed Kalman filters (dashed). The plot shows average values over 1000 observations at each time. The solid line represents the simulated values of x_t and the circles are the estimated values of x_t. The skewness is introduced through the time evolution of λ_t in Fig. 2.

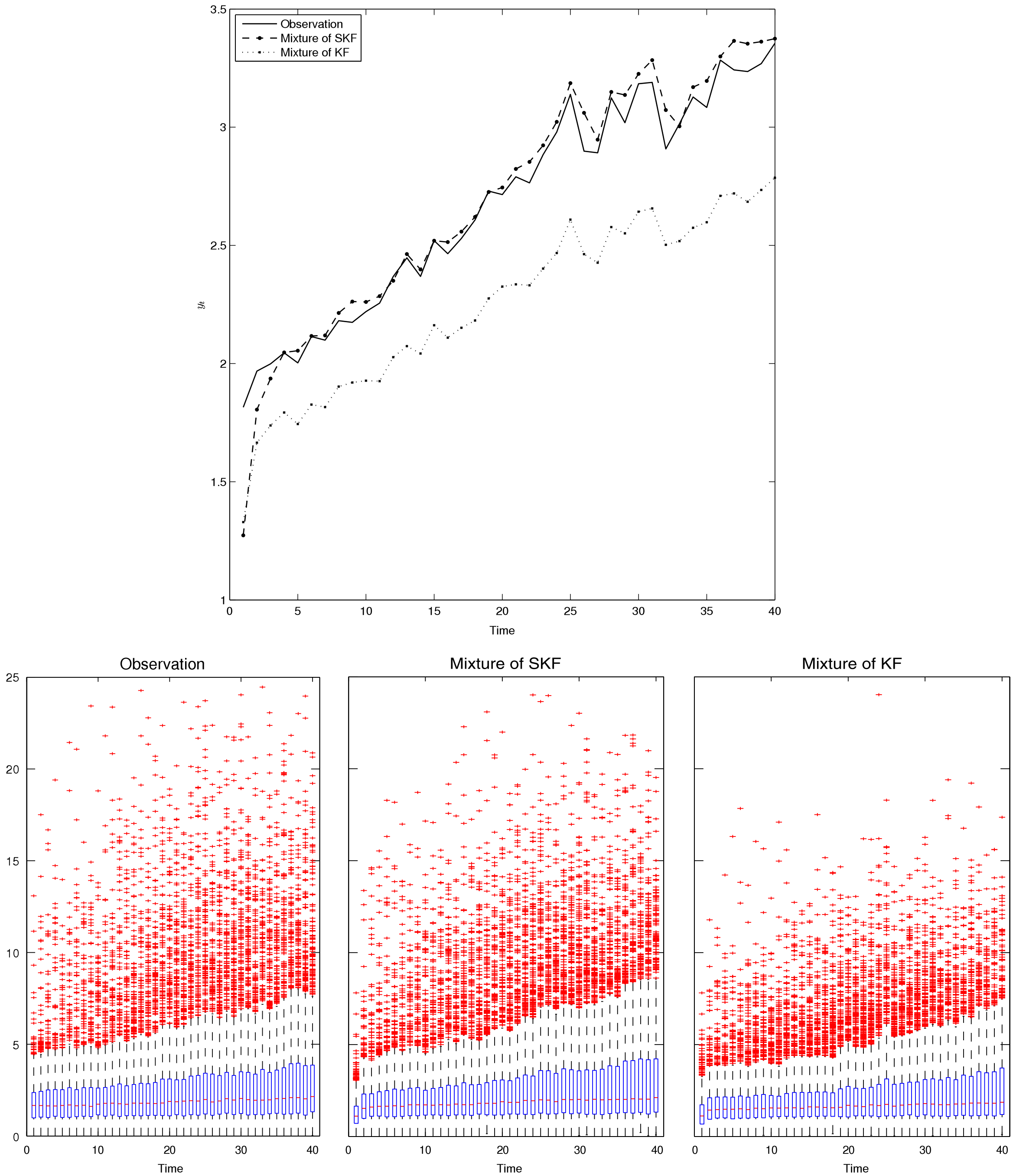


Fig. 4. Estimation of the response y_t by using the mixture of Kalman filters (dotted) and by the mixture of skewed Kalman filters (dashed). The plot on top shows average values over 1000 observations at each time. The solid line represents the simulated values of y_t and the circles are the estimated values of y_t. The three plots in the bottom show the distributions of estimated y_t.

also used $\nu_t = 0$ for every time point for the aforementioned reason. So far, by Lemma 7, we generated samples from a closed skew-normal distribution.

By a simple stochastic representation of scale mixtures of closed skew-normal distributions, (3), it is direct to simulate samples from a closed skew-t distribution taking $K(\lambda) = 1/\lambda$, $\lambda \sim \text{Gamma}(\gamma/2, \gamma/2)$, and $\gamma = 5$. The distribution of s_t is given in Fig. 9 where the same parameters' setup as Fig. 6 is used, but $Q_{\lambda_t} = (-1)^t/2$ and $K_{\lambda_t} = H_{\lambda_t} = 1$. For the left panel ($t = 1$) there is slight skewness and kurtosis, whereas large skewness and kurtosis at the right panel ($t = 40$). Similar to the extended skewed Kalman filter, mixtures of extended skewed Kalman filters best capture the trend of s_t (see Fig. 10) whereas the others still lack efficiency. The Kalman filter misses the trend completely. The root mean-squared errors at a

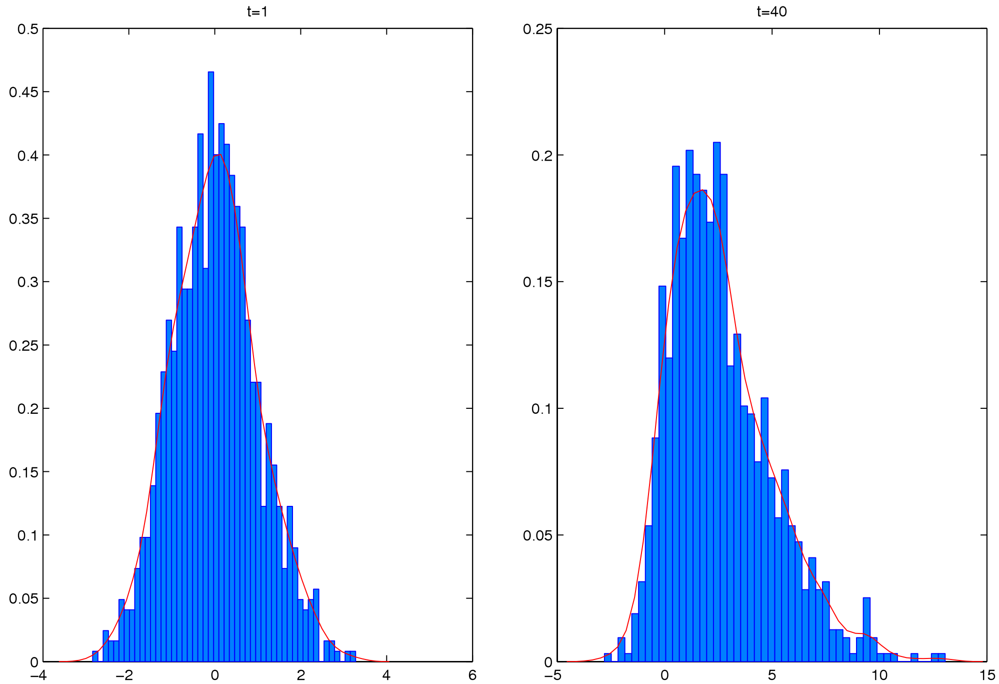


Fig. 5. Density of s_t with histograms from simulated values. The left panel corresponds to the time $t = 1$ and the right panel to the time $t = 40$ for the parameters described in Fig. 6. The red lines are estimated density curves through kernel density estimation. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

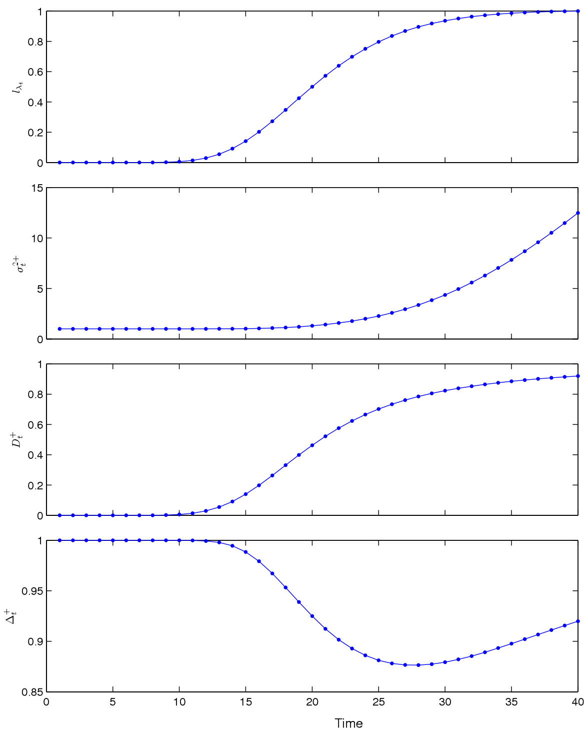


Fig. 6. Temporal evolution of the parameters used to simulate s_t in Fig. 5.

given time point are plotted in Fig. 11 for three filters. This plot obviously indicates that the classical Kalman filter and the extended skewed Kalman filter lost some efficiency when skewness and kurtosis were introduced. For this simulation, we also used $\nu_t = 0$ for every time point for the aforementioned reason.

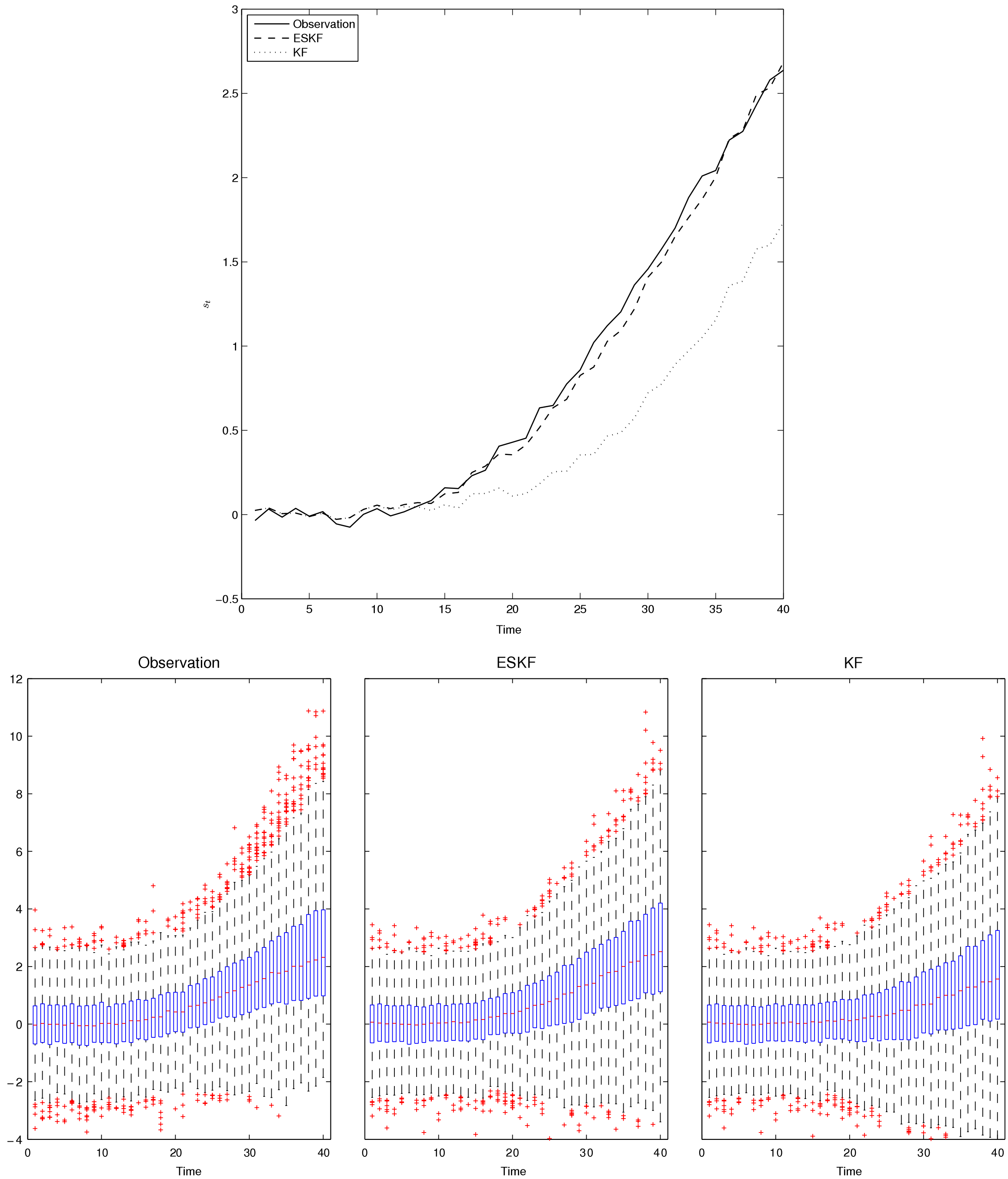


Fig. 7. Estimation of s_t by using the Kalman filter (dotted) and by the extended skewed Kalman filter (dashed). The plot on top shows average values over 1000 observations at each time. The solid line represents the simulated values of s_t and the two other lines are the estimated values of s_t. The three plots in the bottom show the distributions of estimated s_t.

Acknowledgments

This publication is based in part on work supported by Award No. KUS-C1-016-04 made by King Abdullah University of Science and Technology (KAUST), and by NSF grant DMS-1007504. The first author's research was supported by Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education, Science and Technology (2013R1A1A2005995).

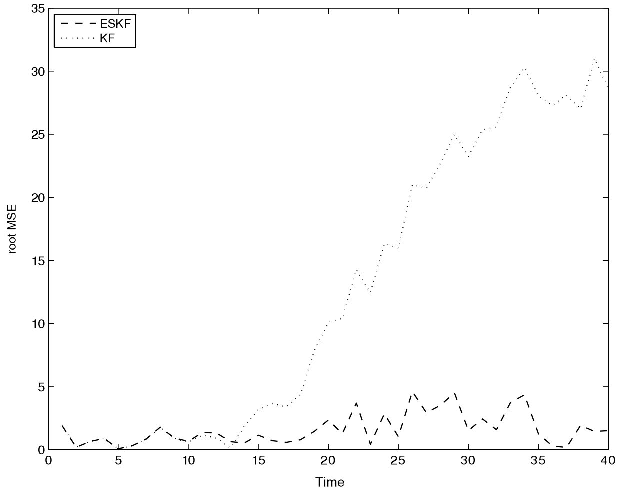


Fig. 8. Root MSE for each time point. ESKF denotes the extended skewed Kalman filter and KF denotes the classical Kalman filter.

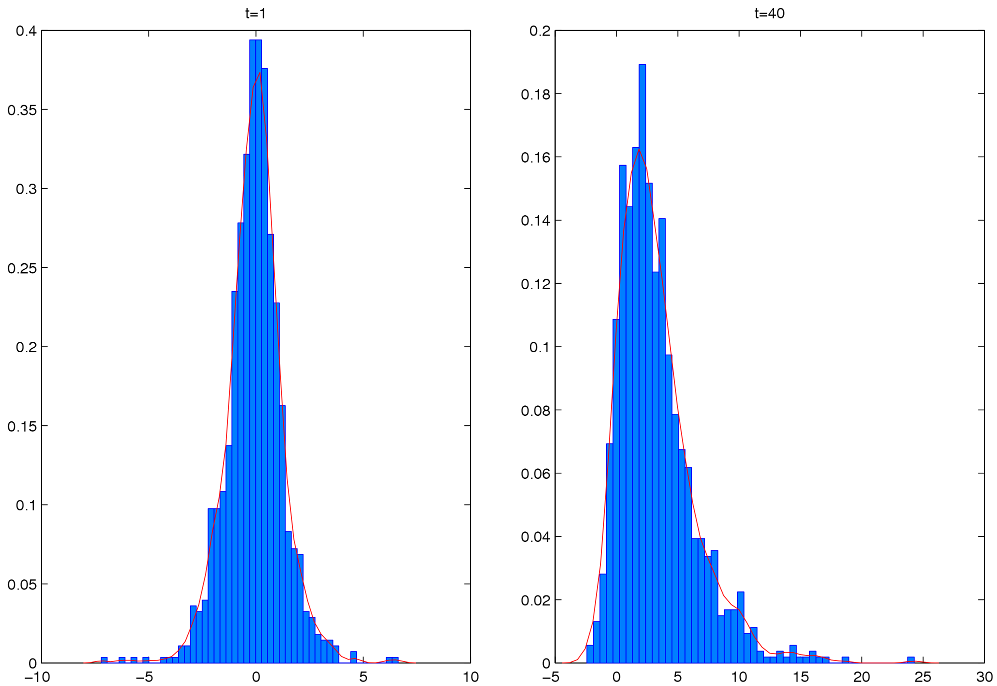


Fig. 9. Density of s_t with histograms from simulated values. The left panel corresponds to the time $t = 1$ and the right panel to the time $t = 40$ for the parameters described in Fig. 6. The red lines are estimated density curves through kernel density estimation. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

Appendix

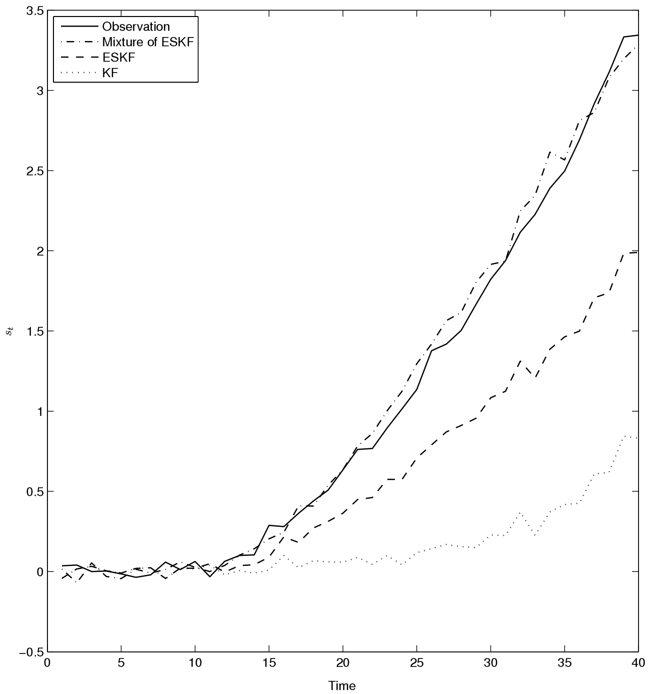
Proof of Lemma 1. The joint distribution of W and Z is

$$\left(\begin{matrix} W \\ Z \end{matrix}\middle| \lambda\right) \sim N_{n+m}\left(\left(\begin{matrix} \mu \\ -\nu \end{matrix}\right), \left(\begin{matrix} K(\lambda)\Sigma & K(\lambda)^{1/2}\Sigma D^T \\ K(\lambda)^{1/2}D\Sigma & D\Sigma D^T + \Delta \end{matrix}\right)\right)$$

and since

$$f_{W|Z\geq 0,\lambda}(w) = \frac{f_{W|\lambda}(w)}{P(Z \geq 0|\lambda)} P(Z \geq 0|W = w, \lambda),$$

we obtain the result using standard properties of normal distributions. □



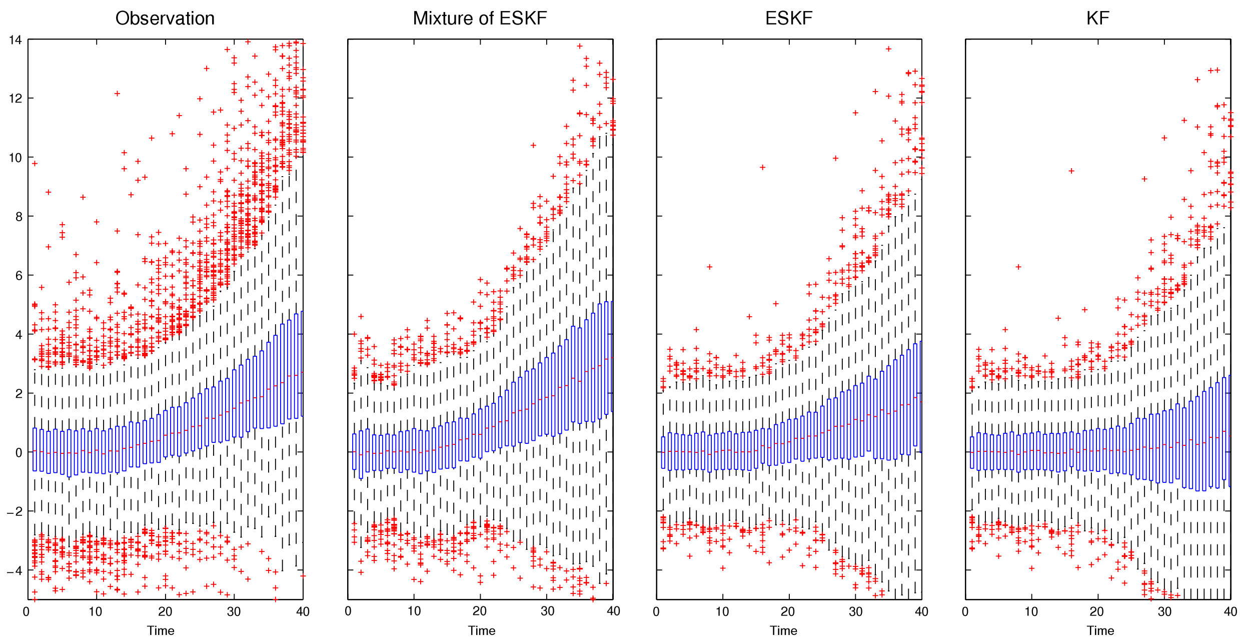


Fig. 10. Estimation of s_t by using the Kalman filters (dotted), by the extended skewed Kalman filters (dashed) and by the mixtures of extended skewed Kalman filters (dashed–dotted). The plot on top shows average values over 1000 observations at each time. The solid line represents the simulated values of s_t and the three other lines are the estimated values of s_t. The four plots in the bottom show the distributions of estimated s_t.

Proof of Theorem 1.

$$M_W(t) = E\{\exp(t^T W)\} = \int_{\mathbb{R}^n} e^{t^T w} f_W(w)dw$$
$$= \int_0^\infty \int_{\mathbb{R}^n} e^{t^T w} f_{W|\lambda}(w)dwdH(\lambda) = \int_0^\infty M_{\text{CSN}}(t)dH(\lambda),$$

where $M_{\text{CSN}}(t)$ is given by Lemma 1 of González-Farías et al. [21]. □

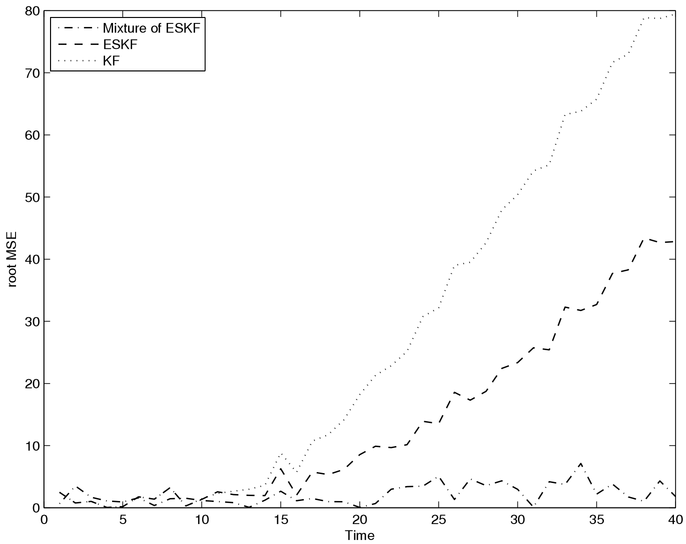


Fig. 11. Root MSE for each time point. Mixture of ESKF denotes mixture of extended skewed Kalman filters, ESKF denotes the extended skewed Kalman filter and KF denotes the classical Kalman filter.

Proof of Theorem 2. $E(W) = \mu + E\{K(\lambda)^{1/2}\}E(Y)$ by independence of λ and Y and the stochastic representation (3). Here c_r and $E(Y)$ are given by Table 1 and Gupta et al. [23], respectively. So the mean follows. By definition of the covariance matrix, $\mathrm{Cov}(W) = E[\{W - E(W)\}\{W - E(W)\}^T]$, we have $\mathrm{Cov}(W) = c_2 E(YY^T) - c_1^2 E(Y)E(Y)^T$ after simple algebra. So the covariance follows by $\mathrm{Cov}(Y)$ given in [23]. □

Proof of Theorem 3.

$$\begin{aligned} E\{h(W)\Phi_n^r(W; 0, I_n)\} &= \int_{\mathbb{R}^n}\int_0^\infty h(w)\Phi_n^r(w; 0, I_n) f_{W|\lambda}(w)dH(\lambda)dw \\ &= \int_0^\infty \int_{\mathbb{R}^n} h(w)\Phi_n^r(w; 0, I_n) f_{W|\lambda}(w)dwdH(\lambda) \\ &= \int_0^\infty \int_{\mathbb{R}^n} h(w)\Phi_{rn}\{D_*(w-\mu); -[\mu^T \dots \mu^T]^T, I_{rn}\} f_{W|\lambda}(w)dwdH(\lambda) \\ &= C\int_0^\infty \int_{R^n} h(w)\phi_n\{w; \mu, K(\lambda)\Sigma\}\Phi_{rn+m}\{D_+(w-\mu); \nu_+, \Delta_+\}dwdH(\lambda). \end{aligned}$$

We note that $W_+|\lambda \sim \mathrm{CSN}_{n,rn+m}(\nu, K(\lambda)\Sigma, D_+, \nu_+, \Delta_+)$ so the result follows. □

Proof of Theorem 4.

$$\begin{aligned} F_W(w) &= \int_{-\infty}^w f_W(t)dt = \int_{-\infty}^w \int_0^\infty f_{W|\lambda}(w)dH(\lambda)dt \\ &= \int_0^\infty \int_{-\infty}^w f_{W|\lambda}(w)dtdH(\lambda) = \int_0^\infty F_{\mathrm{CSN}}(w)dH(\lambda), \end{aligned}$$

where $F_{\mathrm{CSN}}(w)$ is given by Lemma 2.2.1 of González-Farías et al. [22]. □

Proof of Lemma 2. Let $Y \sim N_n(0, \sigma^2 R)$, where R is a correlation matrix and $\sigma > 0$. Then for any $a, b \in \mathbb{R}^n$

$$\begin{aligned} E\{\Phi_n(\sqrt{\lambda}a + b)\} &= E_\Lambda\{P(Y \le \sqrt{\lambda}a + b|\Lambda = \lambda)\} \\ &= E_\Lambda[P\{(Y-b)/\sqrt{\beta\lambda/\alpha} \le a\sqrt{\alpha/\beta}|\Lambda = \lambda\}] \\ &= P(T \le a\sqrt{\alpha/\beta}), \end{aligned}$$

where $T = (Y-b)/\sqrt{\beta\Lambda/\alpha}$ has the quoted multivariate t distribution. □

Proof of Lemma 4. See [22]. They used the full row rank linear transformation to prove this lemma. Another approach is to use the moment generating function method, i.e., $M_{y_1}(t_1) = M_{y_1,y_2}(t_1, 0)$. So the marginal distribution of y_1 can be obtained. By direct computations, we obtain the conditional distribution of y_2 given y_1 by the definition of the conditional distribution. The converse is obtained by direct multiplication of the conditional distribution and the marginal distribution which is an obvious reverse calculation used to derive the marginal and conditional distributions. □

Proof of Lemma 6. By independence of y and z, we have the joint pdf of $x = (y^T\ z^T)^T$ as follows:

$$f(x) = C_y \phi_n(y; \mu, \Sigma)\Phi_m\{D(y-\mu); \nu, \Delta\}\phi_l(z; \psi, \Omega),$$

where $C_y^{-1} = \Phi_m(0; \nu, \Delta + D\Sigma D^T)$. By the property of the multivariate normal distribution, we have that $\phi_n(y; \mu, \Sigma) \times \phi_l(z; \psi, \Omega) = \phi_{n+l}(x; \mu_x, \Sigma_x)$, where

$$\mu_x = \begin{pmatrix} \mu \\ \psi \end{pmatrix}, \qquad \Sigma_x = \begin{pmatrix} \Sigma & 0 \\ 0 & \Omega \end{pmatrix}.$$

Also $\Phi_m\{D(y-\mu); \nu, \Delta\} = \Phi_m\{D_x(x-\mu_x); \nu, \Delta\}$, where $D_x = \begin{pmatrix} D & 0 \end{pmatrix}$, and $\Phi_m(0; \nu, \Delta + D\Sigma D^T) = \Phi_m(0; \nu, \Delta + D_x\Sigma_x D_x^T)$. Hence the result follows by the definition of the closed skew-normal distribution. □

Proof of Theorem 5. Using Lemmas 3 and 5, and mathematical induction, it is straightforward. Naveau et al. [28] proved a similar theorem using their own lemma with given trajectory $\boldsymbol{\lambda}_t$. □

Proof of Theorem 6. Note that $(x_t|x_{t-1}, \mathbf{y}_{t-1}, \boldsymbol{\lambda}_t) \stackrel{d}{=} (G_{\lambda_t}x_{t-1}|x_{t-1}, \mathbf{y}_{t-1}, \boldsymbol{\lambda}_t) + W_{\lambda_t}\eta_t|\boldsymbol{\lambda}_t$. So $(x_t|x_{t-1}, \mathbf{y}_{t-1}, \boldsymbol{\lambda}_t) \sim N_h(G_{\lambda_t}x_{t-1}, W_{\lambda_t}W_{\lambda_t}^T)$ and we know that $(x_{t-1}|\mathbf{y}_{t-1}, \boldsymbol{\lambda}_{t-1})$ follows a $\text{CSN}_{h,m}(\psi_{t-1}, \Omega_{t-1}, D_{t-1}, \nu_{t-1}, \Delta_{t-1})$ by assumption. By Lemma 4, the joint distribution of $(x_{t-1}^T\ x_t^T)^T$ given $\mathbf{y}_{t-1}, \boldsymbol{\lambda}_t$ follows a closed skew-normal distribution, i.e.

$$\text{CSN}_{2h,m}\left(\begin{pmatrix} \psi_{t-1} \\ G_{\lambda_t}\psi_{t-1} \end{pmatrix}, \begin{pmatrix} \Omega_{t-1} & \Omega_{t-1}G_{\lambda_t}^T \\ G_{\lambda_t}\Omega_{t-1} & \tilde{\Omega}_t \end{pmatrix}, (D_{t-1}\ 0), \nu_{t-1}, \Delta_{t-1}\right),$$

where $\tilde{\Omega}_t = W_{\lambda_t}W_{\lambda_t}^T + G_{\lambda_t}\Omega_{t-1}G_{\lambda_t}^T$. Again using Lemma 4, the marginal distribution of x_t given $\mathbf{y}_{t-1}, \boldsymbol{\lambda}_t$ follows $\text{CSN}_{h,m}$ $(G_{\lambda_t}\psi_{t-1}, \tilde{\Omega}_t, D_t, \nu_t, \Delta_t)$, where

$$D_t = D_{t-1}\Omega_{t-1}G_{\lambda_t}^T\tilde{\Omega}_t^{-1}, \qquad \nu_t = \nu_{t-1}, \text{ and}$$
$$\Delta_t = \Delta_{t-1} + (D_{t-1} - D_tG_{\lambda_t})\Omega_{t-1}D_{t-1}^T.$$

Let $e_t = y_t - F_{\lambda_t}G_{\lambda_t}\psi_{t-1}$ be an error in predicting y_t from the time point $t-1$, then $(e_t|x_t, \mathbf{y}_{t-1}, \boldsymbol{\lambda}_t) \stackrel{d}{=} (F_{\lambda_t}(x_t - G_{\lambda_t}\psi_{t-1})|x_t, \mathbf{y}_{t-1}, \boldsymbol{\lambda}_t) + V_{\lambda_t}\epsilon_t|\boldsymbol{\lambda}_t$. So $(e_t|x_t, \mathbf{y}_{t-1}, \boldsymbol{\lambda}_t)$ follows $N_d(F_{\lambda_t}(x_t - G_{\lambda_t}\psi_{t-1}), V_{\lambda_t}V_{\lambda_t}^T)$. Applying Lemma 4 to $(e_t|x_t, \mathbf{y}_{t-1}, \boldsymbol{\lambda}_t)$ and $(x_t|\mathbf{y}_{t-1}, \boldsymbol{\lambda}_t)$, we have the joint distribution of $(x_t^T\ e_t^T)^T$ given $\mathbf{y}_{t-1}, \boldsymbol{\lambda}_t$ as follows:

$$\text{CSN}_{h+d,m}\left(\begin{pmatrix} G_{\lambda_t}\psi_{t-1} \\ 0 \end{pmatrix}, \begin{pmatrix} \tilde{\Omega}_t & \tilde{\Omega}_tF_{\lambda_t}^T \\ F_{\lambda_t}\tilde{\Omega}_t & \hat{\Omega}_t \end{pmatrix}, (D_t\ 0), \nu_t, \Delta_t\right), \tag{23}$$

where $\hat{\Omega}_t = V_{\lambda_t}V_{\lambda_t}^T + F_{\lambda_t}\tilde{\Omega}_tF_{\lambda_t}^T$. Since $(x_t|e_t, \mathbf{y}_{t-1}, \boldsymbol{\lambda}_t) \stackrel{d}{=} (x_t|\mathbf{y}_t, \boldsymbol{\lambda}_t)$, the result follows using Lemma 4. □

Proof of Corollary 1. The steps before defining e_t in the proof of Theorem 6 are still valid to prove the "invariance property". Now define e_t^* as in (9), then $(e_t^*|x_t, \mathbf{y}_{t-1}, \boldsymbol{\lambda}_t) \stackrel{d}{=} (F_{\lambda_t}(x_t - G_{\lambda_t}\psi_{t-1} - C^*)|x_t, \mathbf{y}_{t-1}, \boldsymbol{\lambda}_t) + (V_{\lambda_t}\epsilon_t|\boldsymbol{\lambda}_t)$. So $(e_t^*|x_t, \mathbf{y}_{t-1}, \boldsymbol{\lambda}_t)$ follows $N_d(F_{\lambda_t}(x_t - G_{\lambda_t}\psi_{t-1} - C^*), V_{\lambda_t}V_{\lambda_t}^T)$. Applying Lemma 4 to $(e_t^*|x_t, \mathbf{y}_{t-1}, \boldsymbol{\lambda}_t)$ and $(x_t|\mathbf{y}_{t-1}, \boldsymbol{\lambda}_t)$, we have the joint distribution of $(x_t^T, (e_t^*)^T)^T$ given $\mathbf{y}_{t-1}, \boldsymbol{\lambda}_t$ as follows:

$$\text{CSN}_{h+d,m}\left(\begin{pmatrix} G_{\lambda_t}\psi_{t-1} \\ -F_{\lambda_t}C^* \end{pmatrix}, \begin{pmatrix} \tilde{\Omega}_t & \tilde{\Omega}_tF_{\lambda_t}^T \\ F_{\lambda_t}\tilde{\Omega}_t & \hat{\Omega}_t \end{pmatrix}, (D_t\ 0), \nu_t, \Delta_t\right),$$

where $\hat{\Omega}_t = V_{\lambda_t}V_{\lambda_t}^T + F_{\lambda_t}\tilde{\Omega}_tF_{\lambda_t}^T$. Since $(x_t|e_t^*, \mathbf{y}_{t-1}, \boldsymbol{\lambda}_t) \stackrel{d}{=} (x_t|\mathbf{y}_t, \boldsymbol{\lambda}_t)$, the result follows using Lemma 4. Note that $e_t^* + F_{\lambda_t}C^* = y_t - F_{\lambda_t}(G_{\lambda_t}\psi_{t-1} + C^*) + F_{\lambda_t}C^*$ equal to e_t. □

Proof of Lemma 7. We assume that $\boldsymbol{\Lambda}_t = \boldsymbol{\lambda}_t$, so this lemma is proved for a given trajectory to avoid notational complexity. Let $c = D_t^+\psi_t^+$ and $c < \psi_{t-1}^+$. The case $c \geq \psi_{t-1}^+$ can be treated in a similar manner. For the first step it is proved that $w_{t-1} \stackrel{d}{=} (v_{t-1}|v_{t-1} \leq c)$. Suppose $v_{t-1}^* = 2\psi_{t-1}^+ - v_{t-1}$ and $\tilde{v}_{t-1} = \Phi_{t-1}^{-1}\{h_{t-1}(v_{t-1})\}$, where $h_{t-1}(x) = a\Phi_{t-1}(x) + b$ with $a = \frac{\Phi_{t-1}(c)}{\Phi_{t-1}(2\psi_{t-1}^+ - c) - \Phi_{t-1}(c)}$ and $b = -\Phi_{t-1}(c) \times a$. Then $\{x : c < x < 2\psi_{t-1}^+ - c\} = \{x : -\infty < \Phi_{t-1}^{-1}\{h_{t-1}(x)\} < c\}$ by straightforward algebra. So (16) is equal to

$$w_{t-1} = \begin{cases} v_{t-1} & \text{if } v_{t-1} \leq c, \\ v_{t-1}^* & \text{if } v_{t-1}^* \leq c, \\ \tilde{v}_{t-1} & \tilde{v}_{t-1} < c. \end{cases}$$

By the law of total probability,

$$\begin{aligned} P(w_{t-1} \leq x) &= P(v_{t-1} \leq x|v_{t-1} \leq c)P(v_{t-1} \leq c) + P(v_{t-1}^* \leq x|v_{t-1}^* \leq c)P(v_{t-1} \geq 2\psi_{t-1}^+ - c) \\ &\quad + P(\tilde{v}_{t-1} \leq x|\tilde{v}_{t-1} < c)P(c < v_{t-1} < 2\psi_{t-1}^+ - c). \end{aligned}$$

Since $v_{t-1} \sim N(\psi_{t-1}^{+}, \sigma_{t-1}^{2+})$, $v_{t-1}^{*} = 2\psi_{t-1}^{+} - v_{t-1} \sim N(\psi_{t-1}^{+}, \sigma_{t-1}^{2+})$, i.e.

$$v_{t-1} \overset{d}{=} v_{t-1}^{*}. \tag{24}$$

By the probability integral transform, $\Phi_{t-1}(v_{t-1})$ follows a $U(0, 1)$ hence the conditional distribution of $(\Phi_{t-1}(v_{t-1})|c < v_{t-1} < 2\psi_{t-1}^{+} - c)$ follows a similar uniform distribution, i.e. $U(\Phi_{t-1}(c), \Phi_{t-1}(2\psi_{t-1}^{+} - c))$ by simple algebra. Furthermore

$$\begin{aligned} P(\tilde{v}_{t-1} \le x|\tilde{v}_{t-1} < c) &= P[\Phi_{t-1}^{-1}\{h_{t-1}(v_{t-1})\} \le x|\Phi_{t-1}^{-1}\{h_{t-1}(v_{t-1})\} < c] \\ &= P[h_{t-1}(v_{t-1}) \le \Phi_{t-1}(x)|c < v_{t-1} < 2\psi_{t-1}^{+} - c] \\ &= \Phi_{t-1}(x)/\Phi_{t-1}(c) = P(v_{t-1} \le x|v_{t-1} \le c). \end{aligned} \tag{25}$$

By (24) and (25), we know that $P(w_{t-1} \le x) = P(v_{t-1} \le x|v_{t-1} \le c)$, i.e. $w_{t-1} \overset{d}{=} (v_{t-1}|v_{t-1} \le c)$. Hence $s_t = \sigma_{\lambda_t}\eta_t^{+} - l_{\lambda_t} w_{t-1} \overset{d}{=} (\sigma_{\lambda_t}\eta_t^{+} - l_{\lambda_t} v_{t-1}|v_{t-1} \le c) \overset{d}{=} (v_t|v_{t-1} \le c)$ by the definition of v_t. Second we prove that $s_t \overset{d}{=} (v_t|v_{t-1} \le c)$ follows a closed skew-normal distribution. By (13) and (14), we know that

$$\begin{pmatrix} v_t \\ v_{t-1} \end{pmatrix} \sim N_2\left(\begin{pmatrix} \psi_t^{+} \\ \psi_{t-1}^{+} \end{pmatrix}\begin{pmatrix} \sigma_t^{2+} & -l_{\lambda_t}\sigma_{t-1}^{2+} \\ -l_{\lambda_t}\sigma_{t-1}^{2+} & \sigma_{t-1}^{2+} \end{pmatrix}\right).$$

By the standard property of the multivariate normal distribution, $(v_{t-1}|v_t = y) \sim N(\mu_{2|1}, \Delta_t^{+})$, where $\mu_{2|1} = \psi_{t-1}^{+} - D_t^{+}(y - \psi_t^{+})$, $D_t^{+} = l_{\lambda_t}\sigma_{t-1}^{2+}/\sigma_t^{2+}$, and $\Delta_t^{+} = \sigma_{t-1}^{2+} - (D_t^{+})^2\sigma_t^{2+}$. Hence

$$\begin{aligned} f_{s_t}(y|v_{t-1} \le c) &= \frac{f_{v_t}(y)P(v_{t-1} \le c|v_t = y)}{P(v_{t-1} \le c)} \\ &= \phi(y; \psi_t^{+}, \sigma_t^{2+})\frac{\Phi(c; \mu_{2|1}, \Delta_t^{+})}{\Phi(c; \psi_{t-1}^{+}, \sigma_{t-1}^{2+})} \\ &= \phi(y; \psi_t^{+}, \sigma_t^{2+})\frac{\Phi\{D_t^{+}(y - \psi_t^{+}); \psi_{t-1}^{+} - D_t^{+}\psi_t^{+}, \Delta_t^{+}\}}{\Phi\{0; \psi_{t-1}^{+} - D_t^{+}\psi_t^{+}, \Delta_t^{+} + (D_t^{+})^2\sigma_t^{2+}\}}, \end{aligned}$$

that is, $s_t \sim \mathrm{CSN}_{1,1}(\psi_t^{+}, \sigma_t^{2+}, D_t^{+}, \nu_t^{+}, \Delta_t^{+})$ by comparing $f_{s_t}(y|v_{t-1} \le c)$ to the pdf of a closed skew-normal distribution. □

Proof of Theorem 7. To get the distribution of $x_t|\boldsymbol{\lambda}_t$, the proof is based on Lemmas 6 and 7. Using Lemmas 3 and 5, it is straightforward to get the distribution of $y_t|\boldsymbol{\lambda}_t$. □

Proof of Theorem 8. By assumption we know that

$$\left(\begin{matrix} u_{t-1} \\ v_{t-1} \end{matrix}\middle| \mathbf{y}_{t-1}, \boldsymbol{\lambda}_{t-1}\right) \sim N_{k+1}\left(\begin{pmatrix} \psi_{t-1}^{*} \\ \psi_{t-1}^{+} \end{pmatrix}, \begin{pmatrix} \Omega_{t-1}^{*} & 0 \\ 0 & \sigma_{t-1}^{2+} \end{pmatrix}\right), \quad t = 1, 2, \ldots. \tag{26}$$

Note that (26) holds for $t = 1$ which is an initial time. Hence from (13), we find that

$$\left(\begin{matrix} u_t \\ v_t \\ v_{t-1} \end{matrix}\middle| \mathbf{y}_{t-1}, \boldsymbol{\lambda}_t\right) = \left(\begin{matrix} K_{\lambda_t}u_{t-1} \\ -l_{\lambda_t}v_{t-1} \\ v_{t-1} \end{matrix}\middle| \mathbf{y}_{t-1}, \boldsymbol{\lambda}_t\right) + \left(\begin{matrix} H_{\lambda_t}\eta_t^{*} \\ \sigma_{\lambda_t}\eta_t^{+} \\ 0 \end{matrix}\middle| \boldsymbol{\lambda}_t\right)$$

follows a multivariate normal distribution

$$N_{k+2}\left(\begin{pmatrix} K_{\lambda_t}\psi_{t-1}^{*} \\ -l_{\lambda_t}\psi_{t-1}^{+} \\ \psi_{t-1}^{+} \end{pmatrix}, \begin{pmatrix} \tilde{\Omega}_t^{*} & 0 & 0 \\ 0 & \tilde{\sigma}_t^{2+} & -l_{\lambda_t}\sigma_{t-1}^{2+} \\ 0 & -l_{\lambda_t}\sigma_{t-1}^{2+} & \sigma_{t-1}^{2+} \end{pmatrix}\right), \tag{27}$$

where $\tilde{\Omega}_t^{*} = K_{\lambda_t}\Omega_{t-1}^{*}K_{\lambda_t}^{T} + H_{\lambda_t}H_{\lambda_t}^{T}$, and $\tilde{\sigma}_t^{2+} = l_{\lambda_t}^2\sigma_{t-1}^{2+} + \sigma_{\lambda_t}^2$.

Prior to observing y_t, we want to derive the prior distribution of

$$(x_t|\mathbf{y}_{t-1}, \boldsymbol{\lambda}_t) = ((u_t^T\ s_t)^T|\mathbf{y}_{t-1}, \boldsymbol{\lambda}_t).$$

To do so, apply Lemma 7 to the joint distribution of $((v_t\ v_{t-1})^T|\mathbf{y}_{t-1}, \boldsymbol{\lambda}_t)$, then we have $(s_t|\mathbf{y}_{t-1}, \boldsymbol{\lambda}_t) \sim \mathrm{CSN}_{1,1}(\tilde{\psi}_t^{+}, \tilde{\sigma}_t^{2+}, \tilde{D}_t^{+}, \tilde{\nu}_t^{+}, \tilde{\Delta}_t^{+})$, where

$$\tilde{\psi}_t^{+} = -l_{\lambda_t}\psi_{t-1}^{+}, \qquad \tilde{\sigma}_t^{2+} = l_{\lambda_t}^2\sigma_{t-1}^{2+} + \sigma_{\lambda_t}^2, \qquad \tilde{D}_t^{+} = l_{\lambda_t}\sigma_{t-1}^{2+}/\tilde{\sigma}_t^{2+}, \qquad \tilde{\nu}_t^{+} = \psi_{t-1}^{+} - \tilde{D}_t^{+}\tilde{\psi}_t^{+},$$
$$\text{and} \quad \tilde{\Delta}_t^{+} = \sigma_{t-1}^{2+} - (\tilde{D}_t^{+})^2\tilde{\sigma}_t^{2+}.$$

Hence $E(s_t|\mathbf{y}_{t-1}, \boldsymbol{\lambda}_t) = \tilde{\psi}_t^+ + \tau_t^{(1)}$, where

$$\tau_t^{(1)} = \frac{\tilde{D}_t^+ \tilde{\sigma}_t^{2+}}{\sqrt{\tilde{\Delta}_t^+ + (\tilde{D}_t^+)^2 \tilde{\sigma}_t^{2+}}} \lambda \left(\frac{\tilde{\nu}_t^+}{\sqrt{\tilde{\Delta}_t^+ + (\tilde{D}_t^+)^2 \tilde{\sigma}_t^{2+}}} \right), \quad \text{and} \quad \lambda(t) = \frac{s'(t)}{s(t)}.$$

Here $\lambda(t)$ is the hazard rate function of the standard normal distribution, where $s(t) = 1 - F(t) = \int_t^\infty \phi(u)du$. This expectation was obtained by using the cumulant generating function of the closed skew-normal distribution, $(s_t|\mathbf{y}_{t-1}, \boldsymbol{\lambda}_t)$.

From (27), we know that $(u_t|\mathbf{y}_{t-1}, \boldsymbol{\lambda}_t) \sim N_k(K_{\lambda_t}\psi_{t-1}^*, \tilde{\Omega}_t^*)$. Hence the distribution of $(x_t|\mathbf{y}_{t-1}, \boldsymbol{\lambda}_t) = ((u_t^T\ s_t)^T|\mathbf{y}_{t-1}, \boldsymbol{\lambda}_t)$ is

$$\mathrm{CSN}_{k+1,1}\left(\begin{pmatrix} K_{\lambda_t}\psi_{t-1}^* \\ \tilde{\psi}_t^+ \end{pmatrix}, \begin{pmatrix} \tilde{\Omega}_t^* & 0 \\ 0 & \tilde{\sigma}_t^{2+} \end{pmatrix}, (0\ \tilde{D}_t^+), \tilde{\nu}_t^+, \tilde{\Delta}_t^+ \right) \tag{28}$$

by applying Lemma 6 since u_t and s_t are independent given $\mathbf{y}_{t-1}$ and $\boldsymbol{\lambda}_t$.

On observing y_t, we need to calculate the posterior distribution of $(x_t|\mathbf{y}_t, \boldsymbol{\lambda}_t)$. To achieve this first define $e_t = y_t - Q_{\lambda_t}K_{\lambda_t}\psi_{t-1}^* - P_{\lambda_t}(\tilde{\psi}_t^+ + \tau_t^{(1)})$. Therefore $(e_t|u_t, s_t, \mathbf{y}_{t-1}, \boldsymbol{\lambda}_t) = (\{Q_{\lambda_t}(u_t - K_{\lambda_t}\psi_{t-1}^*) + P_{\lambda_t}(s_t - \tilde{\psi}_t^+ - \tau_t^{(1)})\}|x_t, \mathbf{y}_{t-1}, \boldsymbol{\lambda}_t) + (V_{\lambda_t}\epsilon_t|\boldsymbol{\lambda}_t)$ follows $N_d(Q_{\lambda_t}(u_t - K_{\lambda_t}\psi_{t-1}^*) + P_{\lambda_t}(s_t - \tilde{\psi}_t^+ - \tau_t^{(1)}), V_{\lambda_t}V_{\lambda_t}^T)$. Treating this distribution as the conditional distribution and the distribution of $(x_t|\mathbf{y}_{t-1}, \boldsymbol{\lambda}_t)$ as the marginal distribution, we obtain the joint distribution of $((x_t^T\ e_t^T)^T|\mathbf{y}_{t-1}, \boldsymbol{\lambda}_t)$ by Lemma 4 as follows:

$$\mathrm{CSN}_{k+1+d,1}\left(\begin{pmatrix} K_{\lambda_t}\psi_{t-1}^* \\ \tilde{\psi}_t^+ \\ -P_{\lambda_t}\tau_t^{(1)} \end{pmatrix}, \begin{pmatrix} \tilde{\Omega}_t^* & 0 & \tilde{\Omega}_t^* Q_{\lambda_t}^T \\ 0 & \tilde{\sigma}_t^{2+} & \tilde{\sigma}_t^{2+} P_{\lambda_t}^T \\ Q_{\lambda_t}\tilde{\Omega}_t^* & \tilde{\sigma}_t^{2+} P_{\lambda_t} & \Sigma_t \end{pmatrix}, (0\ \tilde{D}_t^+\ 0), \tilde{\nu}_t^+, \tilde{\Delta}_t^+ \right), \tag{29}$$

where $\Sigma_t = Q_{\lambda_t}\tilde{\Omega}_t^* Q_{\lambda_t}^T + \tilde{\sigma}_t^{2+} P_{\lambda_t} P_{\lambda_t}^T + V_{\lambda_t}V_{\lambda_t}^T$.

Again applying Lemma 4 we have that the distribution of $(x_t|e_t, \mathbf{y}_{t-1}, \boldsymbol{\lambda}_t) = (x_t|\mathbf{y}_t, \boldsymbol{\lambda}_t)$ follows

$$\mathrm{CSN}_{k+1,1}(\psi_t, \Omega_t, D_t, \nu_t, \Delta_t).$$

The parameters can be obtained by simple algebra. This is the required posterior distribution. □

Proof of Corollary 2. The proof steps before defining the error e_t are the same as in Theorem 8. Define a new error $e_t^* = y_t - Q_{\lambda_t}K_{\lambda_t}\psi_{t-1}^* - P_{\lambda_t}\tilde{\psi}_t^+$. Then $(e_t^*|u_t, s_t, \mathbf{y}_{t-1}, \boldsymbol{\lambda}_t)$ is equal to $([Q_{\lambda_t}(u_t - K_{\lambda_t}\psi_{t-1}^*) + P_{\lambda_t}(s_t - \tilde{\psi}_t^+)]|x_t, \mathbf{y}_{t-1}, \boldsymbol{\lambda}_t) + (V_{\lambda_t}\epsilon_t|\boldsymbol{\lambda}_t)$. So the distribution of a new error follows a normal distribution, i.e. $N_d(Q_{\lambda_t}(u_t - K_{\lambda_t}\psi_{t-1}^*) + P_{\lambda_t}(s_t - \tilde{\psi}_t^+), V_{\lambda_t}V_{\lambda_t}^T)$.

We know that the distribution of $(x_t|\mathbf{y}_{t-1}, \boldsymbol{\lambda}_t) = ((u_t^T\ s_t)^T|\mathbf{y}_{t-1}, \boldsymbol{\lambda}_t)$ follows

$$\mathrm{CSN}_{k+1,1}\left(\begin{pmatrix} K_{\lambda_t}\psi_{t-1}^* \\ \tilde{\psi}_t^+ \end{pmatrix}, \begin{pmatrix} \tilde{\Omega}_t^* & 0 \\ 0 & \tilde{\sigma}_t^{2+} \end{pmatrix}, (0\ \tilde{D}_t^+), \tilde{\nu}_t^+, \tilde{\Delta}_t^+ \right)$$

by (28). Treating this distribution as the marginal distribution and the distribution of $(e_t^*|x_t, \mathbf{y}_{t-1}, \boldsymbol{\lambda}_t)$ as the conditional distribution, we have the joint distribution of $((x_t^T\ (e_t^*)^T)^T|\mathbf{y}_{t-1}, \boldsymbol{\lambda}_t)$ by Lemma 4 as follows:

$$\mathrm{CSN}_{k+1+d,1}\left(\begin{pmatrix} K_{\lambda_t}\psi_{t-1}^* \\ \tilde{\psi}_t^+ \\ 0 \end{pmatrix}, \begin{pmatrix} \tilde{\Omega}_t^* & 0 & \tilde{\Omega}_t^* Q_{\lambda_t}^T \\ 0 & \tilde{\sigma}_t^{2+} & \tilde{\sigma}_t^{2+} P_{\lambda_t}^T \\ Q_{\lambda_t}\tilde{\Omega}_t^* & \tilde{\sigma}_t^{2+} P_{\lambda_t} & \Sigma_t \end{pmatrix}, (0\ \tilde{D}_t^+\ 0), \tilde{\nu}_t^+, \tilde{\Delta}_t^+ \right), \tag{30}$$

where $\Sigma_t = Q_{\lambda_t}\tilde{\Omega}_t^* Q_{\lambda_t}^T + \tilde{\sigma}_t^{2+} P_{\lambda_t} P_{\lambda_t}^T + V_{\lambda_t}V_{\lambda_t}^T$. The only difference between (30) and (29) is that the location parameter of $(e_t^*|\mathbf{y}_{t-1}, \boldsymbol{\lambda}_t)$ is 0 whereas the location parameter of $(e_t|\mathbf{y}_{t-1}, \boldsymbol{\lambda}_t)$ is $-P_{\lambda_t}\tau_t^{(1)}$.

Since $e_t = e_t^* - P_{\lambda_t}\tau_t^{(1)}$, all remaining proofs are similar to those of Theorem 8. So the result follows. □

References

[1] D.F. Andrews, C.L. Mallows, Scale mixtures of normal distribution, Journal of the Royal Statistical Society: Series B 36 (1974) 99–102.

[2] R.B. Arellano-Valle, A. Azzalini, On the unification of families of skew-normal distributions, Scandinavian Journal of Statistics 33 (2006) 561–574.

[3] R.B. Arellano-Valle, M.D. Branco, M.G. Genton, A unified view on skewed distributions arising from selections, The Canadian Journal of Statistics 34 (2006) 581–601.

[4] R.B. Arellano-Valle, M.G. Genton, An invariance property of quadratic forms in random vectors with a selection distribution, with application to sample variogram and covariogram estimators, Annals of the Institute of Statistical Mathematics 62 (2010) 363–381.

[5] R.B. Arellano-Valle, M.G. Genton, Multivariate extended skew-*t* distributions and related families, Metron 68 (2010) 201–234.

[6] B.C. Arnold, R.J. Beaver, Skewed multivariate models related to hidden truncation and/or selective reporting, TEST 11 (2002) 7–54.

[7] A. Azzalini, A. Capitanio, Statistical applications of the multivariate skew normal distribution, Journal of the Royal Statistical Society: Series B 61 (1999) 579–602.

[8] A. Azzalini, A. Capitanio, Distributions generated by perturbation of symmetry with emphasis on a multivariate skew *t* distribution, Journal of the Royal Statistical Society, Series B 65 (2003) 367–389.
[9] A. Azzalini, A. Dalla Valle, The multivariate skew-normal distribution, Biometrika 83 (1996) 715–726.
[10] A. Azzalini, M.G. Genton, Robust likelihood methods based on the skew-*t* and related distributions, International Statistical Review 76 (2008) 106–129.
[11] P.C. Bradley, N.G. Polson, D.S. Stoffer, A Monte Carlo approach to nonnormal and nonlinear state-space modeling, Journal of the American Statistical Association 87 (1992) 493–500.
[12] M.D. Branco, D.K. Dey, A general class of multivariate skew elliptical distributions, Journal of Multivariate Analysis 79 (2001) 99–113.
[13] M.-H. Chen, D.K. Dey, Bayesian modeling of correlated binary responses via scale mixture of multivariate normal link functions, Sankhyā 60 (1998) 322–343.
[14] R. Chen, J.S. Liu, Mixture Kalman filters, Journal of the Royal Statistical Society: Series B 62 (2000) 493–508.
[15] S.T.B. Choy, A.F.M. Smith, On robust analysis of a normal location parameter, Journal of the Royal Statistical Society: Series B 59 (1997) 463–474.
[16] C. Flecher, P. Naveau, D. Allard, Estimating the closed skew-normal distribution parameters using weighted moments, Statistics & Probability Letters 79 (2009) 1977–1984.
[17] M.G. Genton, Skew-Elliptical Distributions and their Applications: A Journey Beyond Normality, CRC/Chapman & Hall, London, 2004.
[18] M.G. Genton, L. He, X. Liu, Moments of skew-normal random vectors and their quadratic forms, Statistics & Probability Letters 51 (2001) 319–325.
[19] M.G. Genton, Y. Ma, H. Sang, On the likelihood function of Gaussian max-stable processes, Biometrika 98 (2011) 481–488.
[20] J.R. Gleason, Understanding elongation: the scale contaminated normal family, Journal of the American Statistical Association 88 (1993) 327–337.
[21] G. González-Farías, A. Domínguez-Molina, A.K. Gupta, Additive properties of skew normal random vectors, Journal of Statistical Planning and Inference 126 (2004) 521–534.
[22] G. González-Farías, A. Domínguez-Molina, A.K. Gupta, The closed skew-normal distribution, in: M.G. Genton (Ed.), Skew-Elliptical Distributions and their Applications: A Journey Beyond Normality, CRC/Chapman & Hall, London, 2004, pp. 25–42.
[23] A.K. Gupta, G. González-Farías, A. Domínguez-Molina, A multivariate skew-normal distribution, Journal of Multivariate Analysis 89 (2004) 181–190.
[24] James D. Hamilton, Regime switching models, in: Steven N. Durlauf, Lawrence E. Blume (Eds.), The New Palgrave Dictionary of Economics, second ed., Palgrave Macmillan Ltd., 2008.
[25] B. Liseo, N. Loperfido, A Bayesian interpretation of the multivariate skew-normal distribution, Statistics & Probability Letters 61 (2003) 395–401.
[26] R.J. Meinhold, N.D. Singpurwalla, Understanding the Kalman filter, The American Statistician 37 (1983) 123–127.
[27] R.J. Meinhold, N.D. Singpurwalla, Robustification of Kalman filter models, Journal of the American Statistical Association 84 (1989) 479–486.
[28] P. Naveau, M.G. Genton, X. Shen, A skewed Kalman filter, Journal of Multivariate Analysis 94 (2005) 382–400.
[29] R.L. Smith, J.E. Miller, A non-Gaussian state space model and application to prediction of records, Journal of the Royal Statistical Society: Series B 86 (1986) 763–769.
[30] J. Wang, M.G. Genton, The multivariate skew-slash distribution, Journal of Statistical Planning and Inference 136 (2006) 209–220.
[31] M. West, On scale mixtures of normal distributions, Biometrika 74 (1987) 646–648.