Skew-t Filter and Smoother with Improved Covariance Matrix Approximation

Skew-t Filter and Smoother with Improved Covariance Matrix Approximation

Henri Nurminen, Tohid Ardeshiri, Robert Piché, *Senior Member, IEEE*, and Fredrik Gustafsson, *Fellow, IEEE*

***Abstract*—Filtering and smoothing algorithms for linear discrete-time state-space models with skew-t-distributed measurement noise are proposed. The algorithms use a variational Bayes based posterior approximation with coupled location and skewness variables to reduce the error caused by the variational approximation. Although the variational update is done suboptimally using an expectation propagation algorithm, our simulations show that the proposed method gives a more accurate approximation of the posterior covariance matrix than an earlier proposed variational algorithm. Consequently, the novel filter and smoother outperform the earlier proposed robust filter and smoother and other existing low-complexity alternatives in accuracy and speed. We present both simulations and tests based on real-world navigation data, in particular GPS data in an urban area, to demonstrate the performance of the novel methods. Moreover, the extension of the proposed algorithms to cover the case where the distribution of the measurement noise is multivariate skew-t is outlined. Finally, the paper presents a study of theoretical performance bounds for the proposed algorithms.**

***Index Terms*— skew t, t-distribution, robust filtering, Kalman filter, RTS smoother, variational Bayes, expectation propagation, truncated normal distribution, Cramér–Rao lower bound**

I. Introduction

Asymmetric and heavy-tailed noise processes are present in many inference problems. In radio signal based distance estimation [1]–[3], for example, obstacles cause large positive errors that dominate over symmetrically distributed errors from other sources [4]. An example of this is the error histogram of time-of-flight in distance measurements collected in an indoor environment given in Fig. 1. The asymmetric distributions cannot be predicted by the normal or t-distributions that are equivalent in second order moments, because normal and t-distributions are symmetric distributions. The skew t-distribution [5]–[7] is a generalization of the t-distribution that has the modeling flexibility to capture both skewness and heavy-tailedness of such noise processes. To illustrate this, Fig. 2 shows the contours of the likelihood function for three range measurements where some of the measurements include large positive errors. In this example, skew-t, t, and normal measurement noise models are compared. Due to the additional modeling flexibility, the skew-t based likelihood provides a more apposite spread of the probability mass than the normal and t based likelihoods.

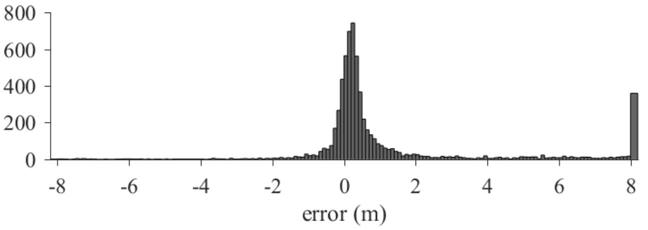


Fig. 1. The error histogram in an ultra-wideband (UWB) ranging experiment described in [8] shows positive skewness. The edge bars show the errors outside the figure limits.

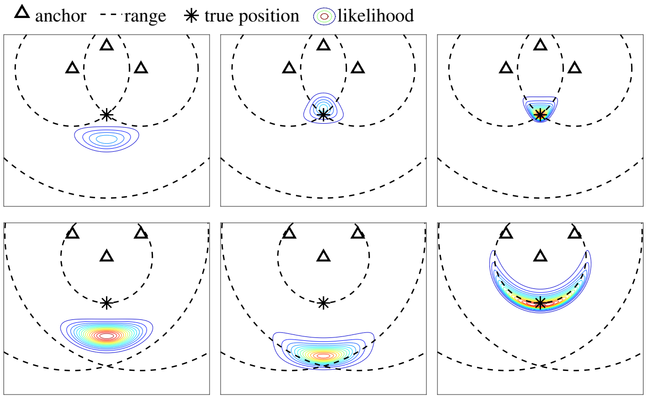


Fig. 2. The likelihood contours of distance measurements from three known anchors for the normal (left), t (middle) and skew-t (right) measurement noise models. The t and skew-t based likelihoods handle one large positive error (upper row), while only the skew-t model handles the two large positive errors (bottom row) due to its asymmetry. The measurement model parameters are selected such that the degrees-of-freedom values and the first two moments coincide.

The applications of the skew distributions are not limited to radio signal based localization. In biostatistics skewed distributions are used as a modeling tool for handling heterogeneous data involving asymmetric behaviors across subpopulations [9]. In psychiatric research skew normal distribution is used to model asymmetric data [10]. Further, in economics skew normal and skew t-distributions are used as models for describing claims in property-liability insurance [11]. More examples describing approaches for analysis and modeling using multivariate skew normal and skew t-distributions in econometrics and environmetrics are presented in [12].

There are various algorithms dedicated to statistical inference of time series when the data exhibit asymmetric distribution. Particle filters [13] can easily be adapted to skew noise distributions, but the computational complexity of these filters increases rapidly as the state dimension increases. A skew Kalman filter is proposed in [14], and in [15] this filter is extended to a robust scale-mixture filter using Monte Carlo integration. These solutions are based on state-space models where the measurement noise is a dependent process

H. Nurminen and R. Piché are with the Laboratory of Automation and Hydraulic Engineering, Tampere University of Technology (TUT), PO Box 692, 33101 Tampere, Finland (e-mails: henri.nurminen@here.com, robert.piche@tut.fi). H. Nurminen has received funding from TUT Graduate School, the Foundation of Nokia Corporation, Tekniikan edistämissäätiö, and Emil Aaltonen Foundation. Henri Nurminen is currently with HERE Technologies Inc.

T. Ardeshiri is with the Division of Automatic Control, Department of Electrical Engineering, Linköping University, 58183, Linköping, Sweden and has received funding from Swedish research council (VR), project scalable Kalman filters for this work. T. Ardeshiri is currently with the Department of Engineering, University of Cambridge, Trumpington Street, Cambridge, CB2 1PZ, UK, (e-mail: ta417@cam.ac.uk).

F. Gustafsson is with the Division of Automatic Control, Department of Electrical Engineering, Linköping University, 58183 Linköping, Sweden, (e-mail: fredrik@isy.liu.se).

with skewed marginals. The article [16] proposes filtering of independent skew measurement and process noises with the cost of increasing the filter state's dimension over time. In all the skew filters of [14]–[16], sequential processing requires numerical evaluation of multidimensional integrals. The inference problem with skew likelihood distributions can also be cast into an optimization problem; [3] proposes an approach to model the measurement noise in an ultra-wideband (UWB) based positioning problem using a tailored half-normal–half-Cauchy distribution. Skewness can also be modeled by a mixture of normal distributions (Gaussian mixtures, GM) [1]. There are many filtering algorithms for GM distributions such as Gaussian sum filter [17] and interactive multiple model (IMM) filter [18]. However, GMs have exponentially decaying tails and can thus be too sensitive to outlier measurements. Furthermore, in order to keep the computational cost of a Gaussian sum filter practicable, a mixture reduction algorithm (MRA) [19] is required, and these MRAs can be computationally expensive and involve approximations to the posterior density.

Variational Bayes (VB) method -based filtering and smoothing algorithms for linear discrete-time state-space models with skew-t measurement noise are proposed in [20]. The VB approach avoids the increasing filter state dimensionality and numerical integrations by finding an optimal approximation with the constraint that the state is independent of the non-dynamic latent variables; this makes analytical marginalisation straightforward. To our knowledge, VB approximations have been applied to the skew t-distribution only in our earlier works [8], [20], and by Wand et al. [21], and the latter use a VB factorization different from ours and do not consider time-series inference. In tests with real UWB indoor localization data [8], this filter is shown to be accurate and computationally inexpensive.

This paper proposes improvements to the robust filter and smoother proposed in [20]. Analogous to [20], the measurement noise is modeled by the skew t-distribution, and the proposed filter and smoother use VB approximations of the filtering and smoothing posteriors. However, the main contributions of this paper are (1) a novel VB factorization of the posterior and showing that at highly skewed models this factorization provides major improvement in both convergence speed of the VB iterations and accuracy of the estimate and covariance matrix, (2) an application of an existing expectation propagation (EP) based algorithm for approximating the statistics of a truncated multivariate normal distribution (TMND) that appears in the proposed VB algorithm, (3) a derivation of a greedy approach for a truncation ordering in the EP approximation of the TMND's moments, (4) a derivation of the Cramér–Rao lower bound (CRLB) for the proposed filter and smoother, and (5) the variational lower bound for the proposed VB factorization. A TMND is a multivariate normal distribution whose support is restricted (truncated) by linear constraints and that is re-normalized to integrate to unity. The aforementioned contributions improve the estimation performance of the skew-t filter and smoother by reducing the covariance matrix underestimation common to many VB inference algorithms [22, Chapter 10]. This is shown by evaluating the proposed algorithms with both simulations and real-data tests in positioning using GNSS (global navigation satellite system) based pseudorange measurements. The tests show clear improvement in estimation accuracy compared to state-of-the-art low-complexity algorithms. In both simulations and real-data tests the proposed algorithms also outperform the earlier VB-based methods of [20] in both estimation accuracy and speed of computations.

The rest of this paper is structured as follows. In Section II, the filtering and smoothing problem involving the univariate skew t-distribution is posed. In Section III a solution based on VB for the formulated problem is proposed. The proposed solution is evaluated in Sections IV and V. The essential expressions to extend the proposed filtering and smoothing algorithms to problems involving multivariate skew-t (MVST) distribution are given in Section VI. Performance bounds for time series data with MVST-distributed measurement noise are derived and evaluated in simulation in Section VII. The concluding remarks are given in Section VIII.

II. Inference Problem formulation

Consider the linear and Gaussian state evolution model

$$p(x_1) = \mathcal{N}(x_1; x_{1|0}, P_{1|0}), \tag{1a}$$

$$x_{k+1} = Ax_k + w_k, \qquad w_k \overset{\text{iid}}{\sim} \mathcal{N}(0, Q), \tag{1b}$$

where $\mathcal{N}(\cdot; \mu, \Sigma)$ denotes the probability density function (PDF) of the (multivariate) normal distribution with mean μ and covariance matrix Σ; $A \in \mathbb{R}^{n_x \times n_x}$ is the state transition matrix; $x_k \in \mathbb{R}^{n_x}$ indexed by $1 \le k \le K$ is the state to be estimated with initial prior distribution (1a), where the subscript "$a|b$" is read "at time a using measurements up to time b"; and $w_k \in \mathbb{R}^{n_x}$ is the process noise. Further, the measurements $y_k \in \mathbb{R}^{n_y}$ are assumed to be governed by the measurement equation

$$y_k = Cx_k + e_k, \tag{2}$$

where $C \in \mathbb{R}^{n_y \times n_x}$ is the measurement matrix, and the measurement noise vector e_k is independent of the process noise, and each component of e_k follows an independent univariate skew t-distribution

$$[e_k]_i \overset{\text{independent}}{\sim} \mathrm{ST}(0, R_{ii}, \Delta_{ii}, \nu_i), \tag{3}$$

where the operator $[\cdot]_i$ gives the ith entry of the argument vector, and $[\cdot]_{ij}$ gives the (i, j) entry of its argument matrix. The model parameters can also be time-varying, but for the sake of lighter notation the k subscripts on A, Q, C, R, Δ, and ν are omitted. The univariate skew t-distribution $\mathrm{ST}(\mu, \sigma^2, \delta, \nu)$ is parametrized by its location parameter $\mu \in \mathbb{R}$, spread parameter $\sigma \in \mathbb{R}_+$, shape parameter $\delta \in \mathbb{R}$ and degrees of freedom $\nu \in \mathbb{R}_+$, and has the PDF

$$\mathrm{ST}(z; \mu, \sigma^2, \delta, \nu) = 2\,\mathrm{t}(z; \mu, \sigma^2 + \delta^2, \nu)\,\mathrm{T}(\widetilde{z}; 0, 1, \nu + 1), \tag{4}$$

where

$$\mathrm{t}(z; \mu, \sigma^2, \nu) = \frac{\Gamma\left(\frac{\nu+1}{2}\right)}{\sigma\sqrt{\nu\pi}\,\Gamma\left(\frac{\nu}{2}\right)}\left(1 + \frac{(z-\mu)^2}{\nu\sigma^2}\right)^{-\frac{\nu+1}{2}} \tag{5}$$

is the PDF of Student's t-distribution, $\Gamma(\cdot)$ is the gamma function, and $\widetilde{z} = \frac{(z-\mu)\delta}{\sigma}\left(\frac{\nu+1}{\nu(\sigma^2+\delta^2)+(z-\mu)^2}\right)^{\frac{1}{2}}$. Also, $\mathrm{T}(\cdot; 0, 1, \nu)$ denotes the cumulative distribution function (CDF) of Student's t-distribution with degrees of freedom ν. Expressions for the first two moments of the univariate skew t-distribution can be found in [23].

The model (3) with independent univariate skew-t-distributed measurement noise components is justified when one-dimensional noises of different sensors can be assumed to be statistically independent [20]. Extension and comparison to multivariate skew-t-distributed noise will be discussed in Section VI.

The independent univariate skew-t noise model (3) induces the hierarchical representation of the measurement likelihood

$$y_k|x_k, u_k, \Lambda_k \sim \mathcal{N}(Cx_k + \Delta u_k, \Lambda_k^{-1}R), \tag{6a}$$

$$u_k|\Lambda_k \sim \mathcal{N}_+(0, \Lambda_k^{-1}), \tag{6b}$$

$$[\Lambda_k]_{ii} \sim \mathcal{G}(\tfrac{\nu_i}{2}, \tfrac{\nu_i}{2}), \tag{6c}$$

where $R \in \mathbb{R}^{n_y \times n_y}$ is a diagonal matrix whose diagonal elements' square roots $\sqrt{R_{ii}}$ are the spread parameters of the skew t-distribution in (3); $\Delta \in \mathbb{R}^{n_y \times n_y}$ is a diagonal matrix whose diagonal elements Δ_{ii} are the shape parameters; $\nu \in \mathbb{R}_+^{n_y}$ is a vector whose elements ν_i are the degrees of freedom. Λ_k is a diagonal matrix with a priori independent random diagonal elements $[\Lambda_k]_{ii}$. Also, $\mathcal{N}_+(\mu, \Sigma)$ is the TMND with closed positive orthant as support, location parameter μ, and squared-scale matrix Σ. Furthermore, $\mathcal{G}(\alpha, \beta)$ is the gamma distribution with shape parameter α and rate parameter β.

Bayesian smoothing means finding the smoothing posterior $\pi(x_{1:K}, u_{1:K}, \Lambda_{1:K}) \triangleq p(x_{1:K}, u_{1:K}, \Lambda_{1:K}|y_{1:K})$. In [20], the smoothing posterior is approximated by a factorized distribution of the form $q_{\,[20]}(x_{1:K}, u_{1:K}, \Lambda_{1:K}) \triangleq q_x(x_{1:K})\, q_u(u_{1:K})\, q_\Lambda(\Lambda_{1:K})$. Subsequently, the approximate posterior distributions are computed using the VB approach. The VB approach minimizes the Kullback–Leibler divergence (KLD) $D_{\mathrm{KL}}(q||p) \triangleq \int q(x) \log \frac{q(x)}{p(x)} \mathrm{d}x$ [24] of the true posterior from the factorized approximation. That is, $D_{\mathrm{KL}}(q_{\,[20]}||\pi)$ is minimized in [20]. An approximate Bayesian filter update, i.e. an approximation the filtering posterior $p(x_k, u_k, \Lambda_k|y_{1:k})$ given a normal filtering prior for x_k, is then a smoother update with $K = 1$.

The numerical simulations in [20] manifest covariance matrix underestimation, which is a known weakness of the VB approach [22, Chapter 10]. One of the contributions of this paper is to reduce the covariance underestimation of the filter and smoother proposed in [20] by removing independence approximations from the VB factorization. The proposed filter and smoother are presented in Section III.

III. Proposed Filter and Smoother

A. VB factorization

Using Bayes' theorem, the state evolution model (1), and the likelihood (6), the joint smoothing posterior PDF is

$$\begin{aligned}
&\pi(x_{1:K}, u_{1:K}, \Lambda_{1:K}) \\
&\propto \mathcal{N}(x_1; x_{1|0}, P_{1|0}) \prod_{l=1}^{K-1} \mathcal{N}(x_{l+1}; Ax_l, Q) \\
&\times \prod_{k=1}^{K} \mathcal{N}(y_k; Cx_k + \Delta u_k, \Lambda_k^{-1}R)\, \mathcal{N}_+(u_k; 0, \Lambda_k^{-1}) \\
&\times \prod_{k=1}^{K} \prod_{i=1}^{n_y} \mathcal{G}\left([\Lambda_k]_{ii}; \frac{\nu_i}{2}, \frac{\nu_i}{2}\right).
\end{aligned} \tag{7}$$

The posterior is not analytically tractable. We propose to seek an approximation in the form

$$\pi(x_{1:K}, u_{1:K}, \Lambda_{1:K}) \approx \hat{q}_{xu}(x_{1:K}, u_{1:K})\, \hat{q}_\Lambda(\Lambda_{1:K}), \tag{8}$$

where the factors in (8) are specified by

$$\hat{q}_{xu}, \hat{q}_\Lambda = \underset{q_{xu}, q_\Lambda}{\operatorname{argmin}}\, D_{\mathrm{KL}}(q_{\mathrm{N}} \,||\, \pi), \tag{9}$$

where $q_{\mathrm{N}}(x_{1:K}, u_{1:K}, \Lambda_{1:K}) \triangleq q_{xu}(x_{1:K}, u_{1:K})\, q_\Lambda(\Lambda_{1:K})$ is the factorized approximation. Hence, $x_{1:K}$ and $u_{1:K}$ are not approximated as independent as in [20] because they can be highly correlated *a posteriori* [20]. The analytical solutions for $\hat{q}_{xu}$ and $\hat{q}_\Lambda$ are obtained by cyclic iteration of

$$\log q_{xu}(x_{1:K}, u_{1:K}) \leftarrow \underset{q_\Lambda}{\mathbb{E}}[\log p(y_{1:K}, x_{1:K}, u_{1:K}, \Lambda_{1:K})] + c_{xu} \tag{10a}$$

$$\log q_\Lambda(\Lambda_{1:K}) \leftarrow \underset{q_{xu}}{\mathbb{E}}[\log p(y_{1:K}, x_{1:K}, u_{1:K}, \Lambda_{1:K})] + c_\Lambda \tag{10b}$$

where $\leftarrow$ is the assignment or reassignment operator, and the expected values on the right hand sides are taken with respect to the current q_{xu} and q_Λ [22, Chapter 10] [25], [26]. Also, c_{xu} and c_Λ are constants with respect to the variables $(x_{1:K}, u_{1:K})$ and $\Lambda_{1:K}$, respectively.

The detailed derivation of the proposed smoother is given in Appendix A. The distribution $q_{xu}(x_{1:K}, u_{1:K})$ is a $K \times (n_x + n_y)$-dimensional TMND, where the underlying normal distribution can be obtained using the Rauch–Tung–Striebel smoother (RTSS) [27]. However, the first two moments of each x_k-marginal are required in the computation of the expectation in (10b), and a TMND's moments cannot be computed in closed form. This renders the smoother impractical, since there is no efficient algorithm for approximating the moments of a large TMND. To obtain a practical smoother algorithm, we replace the RTSS's forward filtering step with the assumed-normal filter where each joint filtering distribution of x_k and u_k, each of them being a TMND, is approximated by a normal distribution with the matched mean and covariance matrix. Because each of these filtering distributions is a low-dimensional ($(n_x + n_y)$-dimensional) TMND, their means and covariance matrices can be approximated efficiently using the computationally light algorithm discussed in Subsection III-B. The result of the assumed-normal filter is then fed into the standard RTSS's backward smoothing step. The obtained skew-t smoother (STS) algorithm is given in Algorithm 1.

In short, one iteration of the proposed smoother consists of a forward filtering step for the variables (x_k, u_k), of a standard RTSS backward smoothing step for the same variables, and of updating $q_\Lambda(\Lambda_k)$ based on the residuals and covariance matrices of each $q(x_k, u_k)$. The forward filtering step is done with a KF-type algorithm where each filtering distribution is modified with the approximative TMND's moments formula.

An approximative filtering update step can be derived as the smoother for a state-space model with just one time-instant. Because each $q_{xu}(x_k, u_k)$ distribution is again a low-dimensional TMND, the moments of each $q_{xu}(x_k, u_k)$ can be approximated quickly. By approximating the x_k-marginal $\int q_{xu}(x_k, u_k) \mathrm{d}u_k$ of the final VB iteration's TMND with a normal distribution, we obtain a recursive filtering algorithm, the skew-t filter (STF) of Algorithm 2. While the marginal $\int q_{xu}(x_k, u_k) \mathrm{d}u_k$ is not exactly normal but consists of non-truncated components of a TMND, it is unimodal and has $\mathbb{R}^{n_x}$ as support, so the normal distribution with the matching first and second moments is a standard approximation. This normality approximation does not affect the convergence of the filtering VB iterations, but there is no convergence proof for the VB iterations when the moments of the TMND are approximated. However, the approximative VB iterations show better accuracy and convergence speed in the numerical simulations presented in Sections IV than the VB iterations with the factorization $q_{[20]}$.

Algorithm 1 Smoothing for skew-t measurement noise

```
**Inputs:** $A$, $C$, $Q$, $R$, $\Delta$, $\nu$, $x_{1|0}$, $P_{1|0}$, $y_{1:K}$, APPROX_TMND
$\Lambda_{k|K} \leftarrow I_{n_y}$ for $k = 1 \cdots K$, $A_z \leftarrow \left[\begin{smallmatrix} A & 0 \\ 0 & 0 \end{smallmatrix}\right]$, $C_z \leftarrow \left[\begin{smallmatrix} C & \Delta \end{smallmatrix}\right]$
**repeat**
  *update* $q_{xu}(x_{1:K}, u_{1:K})$
    **for** $k$ = 1 to $K$ **do**
        $Z_{k|k-1} \leftarrow \text{blockdiagonal}(P_{k|k-1}, \Lambda_{k|K}^{-1})$
        $K_z \leftarrow Z_{k|k-1} C_z^{\mathrm{T}} (C P_{k|k-1} C^{\mathrm{T}} + \Delta \Lambda_{k|K}^{-1} \Delta^{\mathrm{T}} + \Lambda_{k|K}^{-1} R)^{-1}$
        $\widetilde{z}_{k|k} \leftarrow \left[\begin{smallmatrix} x_{k|k-1} \\ 0 \end{smallmatrix}\right] + K_z (y_k - C x_{k|k-1})$
        $\widetilde{Z}_{k|k} \leftarrow (I - K_z C_z) P_{k|k-1}$
        $[z_{k|k}, Z_{k|k}] \leftarrow$ APPROX_TMND$(\widetilde{z}_{k|k}, \widetilde{Z}_{k|k}, \{n_x+1 \cdots n_x+n_y\})$
        $x_{k|k} \leftarrow [z_{k|k}]_{1:n_x}, \ P_{k|k} \leftarrow [Z_{k|k}]_{1:n_x, 1:n_x}$
        $x_{k+1|k} \leftarrow A x_{k|k}$
        $P_{k+1|k} \leftarrow A P_{k|k} A^{\mathrm{T}} + Q$
    **end for**
    **for** $k$ = $K - 1$ down to 1 **do**
        $G_k \leftarrow Z_{k|k} A_z Z_{k+1|k}^{-1}$
        $z_{k|K} \leftarrow z_{k|k} + G_k (z_{k+1|K} - A_z z_{k|k})$
        $Z_{k|K} \leftarrow Z_{k|k} + G_k (Z_{k+1|K} - Z_{k+1|k}) G_k^{\mathrm{T}}$
        $x_{k|K} \leftarrow [z_{k|K}]_{1:n_x}, \ P_{k|K} \leftarrow [Z_{k|K}]_{1:n_x, 1:n_x}$
        $u_{k|K} \leftarrow [z_{k|K}]_{n_x+(1:n_y)}, U_{k|K} \leftarrow [Z_{k|K}]_{n_x+(1:n_y), n_x+(1:n_y)}$
    **end for**
  *update* $q_{\Lambda}(\Lambda_{1:K})$
    **for** $k$ = 1 to $K$ **do**
        $\Psi \leftarrow (y_k - C_z z_{k|K})(y_k - C_z z_{k|K})^{\mathrm{T}} R^{-1}$
              $+ C_z Z_{k|K} C_z^{\mathrm{T}} R^{-1} + u_{k|K} u_{k|K}^{\mathrm{T}} + U_{k|K}$
        **for** $i$ = 1 to $n_y$ **do** $[\Lambda_{k|K}]_{ii} \leftarrow \frac{\nu_i + 2}{\nu_i + \Psi_{ii}}$ **end for**
    **end for**
**until converged**
**Outputs:** $x_{k|K}$ **and** $P_{k|K}$ **for** $k = 1 \cdots K$
```

In short, one VB iteration in the proposed filter's measurement update step consists of updating $q(x_k, u_k)$ with a KF update, modifying its joint mean and covariance matrix with the approximative TMND's moments formulas, and finally updating $q_{\Lambda}(\Lambda_k)$ based on the residual and covariance matrix of $q(x_k, u_k)$.

We propose three stopping criteria for the VB iterations of the filter and smoother: small enough change in the estimate, small enough increase in the variational lower bound (practical only for the filter), and a fixed number of iterations. The computation of the variational lower bound is explained in Subsection III-C. In our tests we fix the number of VB iterations to five, because we found that the estimation accuracy does not improve after five iterations. Fixing the number of VB iterations is the most practical option in terms of predictability of the computation times, but the required number of iterations has to be verified for each model specifically.

B. TMND's moments

The mean and covariance matrix of a TMND can be computed using the formulas presented in [28]. They require evaluating the CDFs of general multivariate normal distributions. The MATLAB function `mvncdf` implements the numerical quadrature of [29] in 2 and 3 dimensional cases and the quasi-Monte Carlo method of [30] for the dimensionalities 4–25. However, these methods can be prohibitively slow. Therefore, we approximate the TMND's moments using a fast sequential algorithm that is based on the expectation propagation (EP) algorithm [31]. An EP algorithm for computing the mean, covariance matrix, and the truncated probability of a TMND is derived in [32]. The method is initialized with the original normal density whose parameters are then updated by applying one linear constraint at a time. For each constraint, the mean and covariance matrix of the once-truncated normal distribution are computed analytically, and the once-truncated distribution is approximated by a non-truncated normal with the updated moments. The EP is an iterative algorithm, so each truncation can be re-made when, roughly speaking, the effect of the previous iteration of the considered truncation is removed from the normal distribution's moments. One iteration of this method is illustrated in Fig. 3, where a bivariate normal distribution truncated into the positive quadrant is approximated with a non-truncated normal distribution.

Algorithm 2 Filtering for skew-t measurement noise

```
**Inputs:** $A$, $C$, $Q$, $R$, $\Delta$, $\nu$, $x_{1|0}$, $P_{1|0}$, $y_{1:K}$, APPROX_TMND
$\Lambda \leftarrow I_{n_y}, \quad C_z \leftarrow \left[\begin{smallmatrix} C & \Delta \end{smallmatrix}\right]$
**for** $k$ = 1 to $K$ **do**
    $[a_{k|k}]_i \leftarrow \frac{\nu_i + 2}{2}$, $[b_{k|k}]_i \leftarrow \frac{\nu_i + 2}{2}$ for $i = 1, \cdots, n_y$
    **repeat**
        $[\Lambda_{k|k}]_{ii} \leftarrow \frac{[a_{k|k}]_i}{[b_{k|k}]_i}$ for $i = 1, \cdots, n_y$
      *update* $q_{xu}(x_k, u_k)$
        $Z_{k|k-1} \leftarrow \text{blockdiagonal}(P_{k|k-1}, \Lambda_{k|k}^{-1})$
        $K_z \leftarrow Z_{k|k-1} C_z^{\mathrm{T}} (C P_{k|k-1} C^{\mathrm{T}} + \Delta \Lambda_{k|k}^{-1} \Delta^{\mathrm{T}} + \Lambda_{k|k}^{-1} R)^{-1}$
        $\widetilde{z}_{k|k} \leftarrow \left[\begin{smallmatrix} x_{k|k-1} \\ 0 \end{smallmatrix}\right] + K_z (y_k - C x_{k|k-1})$
        $\widetilde{Z}_{k|k} \leftarrow (I - K_z C_z) P_{k|k-1}$
        $[z_{k|k}, Z_{k|k}] \leftarrow$ APPROX_TMND$(\widetilde{z}_{k|k}, \widetilde{Z}_{k|k}, \{n_x+1 \cdots n_x+n_y\})$
        $x_{k|k} \leftarrow [z_{k|k}]_{1:n_x}, \ P_{k|k} \leftarrow [Z_{k|k}]_{1:n_x, 1:n_x}$
        $u_{k|k} \leftarrow [z_{k|k}]_{n_x+(1:n_y)}, U_{k|k} \leftarrow [Z_{k|k}]_{n_x+(1:n_y), n_x+(1:n_y)}$
      *update* $q_{\Lambda}(\Lambda_k)$
        $\Psi \leftarrow (y_k - C_z z_{k|k})(y_k - C_z z_{k|k})^{\mathrm{T}} R^{-1}$
              $+ C_z Z_{k|k} C_z^{\mathrm{T}} R^{-1} + u_{k|k} u_{k|k}^{\mathrm{T}} + U_{k|k}$
        **for** $i$ = 1 to $n_y$ **do** $[b_{k|k}]_i \leftarrow \frac{\nu_i + \Psi_{ii}}{2}$ **end for**
    **until converged**
    $x_{k+1|k} \leftarrow A x_{k|k}$
    $P_{k+1|k} \leftarrow A P_{k|k} A^{\mathrm{T}} + Q$
**end for**
**Outputs:** $x_{k|k}$ **and** $P_{k|k}$ **for** $k = 1 \cdots K$
```

The result of the EP algorithm depends on the order in which the constraints are applied. Finding the optimal order of applying the truncations is a problem that has combinatorial complexity. Hence, we adopt a greedy approach, whereby the constraint to be applied is chosen from among the remaining constraints so that the resulting once-truncated normal distribution is closest to the true TMND. By Lemma 1, the optimal constraint to select is the one that truncates the most probability. The optimality is with respect to a KLD as the measure. For example, in Fig. 3 the vertical constraint truncates more probability, so it is applied first.

Lemma 1. *Let $p(\mathbf{z})$ be a TMND with the support $\{\mathbf{z} \geq 0\}$ and $q(\mathbf{z}) = \mathcal{N}(\mathbf{z}; \mu, \Sigma)$. Then,*

$$\operatorname*{argmin}_i D_{\mathrm{KL}}\Big(p(\mathbf{z}) \,||\, \tfrac{1}{c_i} q(\mathbf{z}) [\![\mathbf{z}_i \geq 0]\!]\Big) = \operatorname*{argmin}_i \tfrac{\mu_i}{\sqrt{\Sigma_{ii}}}, \tag{11}$$

where μ_i is the ith element of μ, Σ_{ii} is the ith diagonal element of Σ, $[\![\cdot]\!]$ is the Iverson bracket, and $c_i = \int q(\mathbf{z}) [\![\mathbf{z}_i \geq 0]\!] \,\mathrm{d}\mathbf{z}$.

Proof: $D_{\mathrm{KL}}\Big(p(\mathbf{z}) \,||\, \tfrac{1}{c_i} q(\mathbf{z}) [\![\mathbf{z}_i \geq 0]\!]\Big)$

$$\overset{+}{=} -\int p(\mathbf{z}) \log(\tfrac{1}{c_i} q(\mathbf{z}) [\![\mathbf{z}_i \geq 0]\!]) \,\mathrm{d}\mathbf{z} \tag{12}$$

$$= \log c_i - \int p(\mathbf{z}) \log q(\mathbf{z}) \,\mathrm{d}\mathbf{z} \overset{+}{=} \log c_i, \tag{13}$$

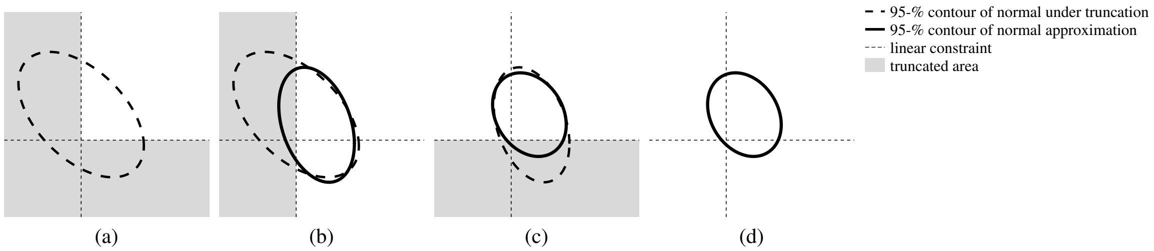

95-% contour of normal under truncation
95-% contour of normal approximation
linear constraint
truncated area
(a)
(b)
(c)
(d)

Fig. 3. An iteration of the EP algorithm for approximating a truncated normal distribution with a normal distribution: (a) the original normal distribution's contour ellipse that contains 95 % of the probability, and the truncated area in gray, (b) the first applied truncation in gray, and the 95-% contour of the resulting normal approximation, (c) the second applied truncation in gray, and the 95-% contour of the normal approximation, (d) the final normal approximation.

where $\overset{+}{=}$ means equality up to an additive constant. Since c_i is an increasing function of $\frac{\mu_i}{\sqrt{\Sigma_{ii}}}$, the proof follows. ■

The obtained EP algorithm with the greedy processing sequence for computing the mean and covariance matrix of a given multivariate normal distribution truncated to the positive orthant is given in Algorithm 3. The algorithm can also give the logarithm of the positive orthant's probability α, which is required in computing the variational lower bound. In many programming languages a numerically robust method to implement the line 12 of the algorithm in Algorithm 3 is using the scaled complementary error function erfcx through

$$\frac{\phi(\xi)}{\Phi(\xi)} = \frac{\sqrt{2/\pi}}{\mathrm{erfcx}(-\xi/\sqrt{2})}. \tag{14}$$

Unfortunately, the EP algorithm does not in general admit guaranteed convergence or error bounds. However, Cunningham et al. [32] present an extensive Monte Carlo study on the performance of the EP in approximating the truncated probability of a TMND, showing that this EP algorithm is reliable provided that there are no redundant truncating constraints and that the support of the distribution is hyperrectangular. Cunningham et al. also imply that the same result applies to approximating the moments of the TMND. These conditions are fulfilled in our case, as all the truncating hyperplanes are non-redundant and aligned with coordinate axes.

Algorithm 3 Greedy expectation propagation for the moments of normal distribution truncated to positive orthant (function $[\mu, \Sigma, \alpha] \leftarrow \texttt{APPROX\_TMND}(\mu, \Sigma, \mathcal{T})$)

```
 1: Inputs: $\mu$, $\Sigma$, and set of the truncated components' indices $\mathcal{T}$
 2: $\widetilde{\mu} \leftarrow \mu, \quad \widetilde{\Sigma} \leftarrow \Sigma$
 3: $\alpha \leftarrow -\frac{1}{2}\widetilde{\mu}^{\mathrm{T}}\widetilde{\Sigma}^{-1}\widetilde{\mu}, \quad M \leftarrow I_{n_\mu}$
 4: $\tau_k \leftarrow 0, \quad \eta_k \leftarrow 0$ for $k = 1, 2, \ldots, n_\mu$.
 5: repeat
 6:     $\mathcal{T}' \leftarrow \mathcal{T}$
 7:     while $\mathcal{T}' \neq \emptyset$ do
 8:         $k \leftarrow \mathrm{argmin}_i\{\mu_i/\sqrt{\Sigma_{ii}} \mid i \in \mathcal{T}'\}$
 9:         $s^2 \leftarrow 1/(1/\Sigma_{kk} - \tau_k)$
10:         $m \leftarrow s^2(\mu_k/\Sigma_{kk} - \eta_k)$
11:         $\xi \leftarrow m/s$
12:         $\epsilon \leftarrow \phi(\xi)/\Phi(\xi)$  ▷ $\phi$ is the PDF of $\mathcal{N}(0,1)$, $\Phi$ its CDF
13:         $\overline{m} \leftarrow m + \epsilon s$
14:         $\overline{s}^2 \leftarrow (1 - \xi\epsilon - \epsilon^2)s^2$
15:         $\overline{\tau}_k \leftarrow 1/\overline{s}^2 - 1/s^2 - \tau_k, \quad \tau_k \leftarrow \tau_k + \overline{\tau}_k$
16:         $\overline{\eta}_k \leftarrow \overline{m}/\overline{s}^2 - m/s^2 - \eta_k, \quad \eta_k \leftarrow \eta_k + \overline{\eta}_k$
17:         $\mu \leftarrow \mu + \frac{\overline{\eta}_k - \overline{\tau}_k\mu_k}{1+\overline{\tau}_k\Sigma_{kk}} \cdot \Sigma_{:,k}$  ▷ mean update
18:         $\Sigma \leftarrow \Sigma - \frac{\overline{\tau}_k}{1+\overline{\tau}_k\Sigma_{jj}} \cdot \Sigma_{:,k}\Sigma_{k,:}$  ▷ covariance update
19:         $M \leftarrow M + \tau_k L_{k,:}^{\mathrm{T}} L_{k,:}$  ▷ $LL^{\mathrm{T}} = \widetilde{\Sigma}$
20:         $\alpha \leftarrow \alpha + \log\big(\Phi(\xi)\big) + \frac{1}{2}\log(1 + \tau_k s^2) + \frac{1}{2}\tau_k\mu_k^2$
21:             $+\frac{1}{2}\frac{m^2\tau_k - 2m\eta_k - s^2\eta_k^2}{1+\tau_k s^2}$  ▷ log-probability update
22:         $\mathcal{T}' \leftarrow \mathcal{T}' \backslash \{k\}$
23:     end while
24:     $\alpha \leftarrow \alpha - \frac{1}{2}\log(\det(M)) + \frac{1}{2}\mu^{\mathrm{T}}\widetilde{\Sigma}^{-1}\mu$
25: until converged
26: Outputs: moments $\mu$, $\Sigma$, and the logarithm of the positive
    orthant's probability $\alpha$
```

C. Variational lower bound

When the PDF $p(x|y)$ is approximated with the PDF $q(x)$, the variational lower bound is

$$\mathcal{L}(q) = \int q(x) \log \frac{p(y,x)}{q(x)} \mathrm{d}x. \tag{15}$$

Minimizing the KLD is equivalent to maximizing the variational lower bound [33, Ch. 21]. Therefore, the variational lower bound can be used as a debugging means and convergence criterion for the VB iterations because the lower bound should increase at each iteration. Furthermore, because the logarithmic marginal likelihood $\log p(y)$ is the sum of the variational lower bound and the KLD, the maximal variational lower bound can be used as an approximation for $\log p(y)$. The model evidence in Bayesian comparison can thus be approximated with $\exp(\mathcal{L}(q))$ [33, Ch. 21.5.1.6].

When evaluated immediately after the VB filter update of $q_{xu}(x_k, u_k)$, the variational lower bound for the skew-t filter is

$$\begin{aligned}\mathcal{L}_{\mathrm{f}}(q) = &\log \mathcal{N}\big(y; Cx_{k|k-1}, CP_{k|k-1}C^{\mathrm{T}} + \Delta\Lambda_{k|k}^{-1}\Delta^{\mathrm{T}} + \Lambda_{k|k}^{-1}R\big)\\ &+ \sum_{j=1}^{n_y}\left[[a_{k|k}]_j\left(1 + \log\left(\tfrac{[a_{k|k}]_j - 1}{[b_{k|k}]_j}\right) - \tfrac{[a_{k|k}]_j - 1}{[b_{k|k}]_j}\right)\right.\\ &\left.- \log\left(\tfrac{[a_{k|k}]_j}{[b_{k|k}]_j}\right)\right] + n_y\log(2) + \log \alpha_{k|k},\end{aligned} \tag{16}$$

where the notations follow those in Algorithm 2, and $\alpha_{k|k}$ is the probability of the positive orthant for the distribution $\mathcal{N}([\widetilde{z}_{k|k}]_{n_x+(1:n_y)}, [\widetilde{Z}_{k|k}]_{n_x+(1:n_y), n_x+(1:n_y)})$. The probability $\alpha_{k|k}$ can be computed using the EP algorithm in Algorithm 3. The derivation of the lower bound (16) is straightforward but tedious and omitted here. Unfortunately, evaluation of the variational lower bound for the smoother is impractical because its expression includes a probability of the positive orthant given a high-dimensional normal distribution.

IV. Simulations

Our numerical simulations use satellite navigation pseudorange measurements and the model

$$[y_k]_i = \|s_i - [x_k]_{1:3}\| + [x_k]_4 + [e_k]_i,\ [e_k]_i \overset{\text{iid}}{\sim} \mathrm{ST}(0, 1\,\mathrm{m}, \delta\,\mathrm{m}, 4) \tag{17}$$

where $s_i \in \mathbb{R}^3$ is the ith satellite's position, $[x_k]_4 \in \mathbb{R}$ is bias with prior $\mathcal{N}(0, (0.75\,\mathrm{m})^2)$, and $\delta \in \mathbb{R}$ is a parameter. The model is linearized using the first order Taylor polynomial approximation, and the linearization error is negligible because the satellites are far relative to the magnitude of uncertainty in the prior. The satellite constellation of the Global Positioning System (GPS) from the first second of the year 2015 provided by the International GNSS Service [34] is used with 8 visible satellites. The root-mean-square error (RMSE) is computed for the position $[x_k]_{1:3}$ as

$$\mathrm{RMSE} = \sqrt{\frac{1}{K}\sum_{k=1}^{K} \left\| [x_{k|k}]_{1:3} - [x_k]_{1:3} \right\|^2}, \tag{18}$$

where $x_{k|k}$ is the filter estimate and x_k is the true state. The computations are made with Matlab.

A. Computation of TMND statistics

In this subsection we study the computation of the moments of the untruncated components of a TMND. For each Monte Carlo replication, one state value is generated from the prior $x \sim \mathcal{N}(0, \mathrm{diag}(20^2, 20^2, 0.22^2, 0.1^2)\,\mathrm{m}^2)$, and one measurement vector is generated from the model (17) with $\nu = \infty$ degrees of freedom (corresponding to skew-normal likelihood). 10 000 Monte Carlo replications are used. The compared methods are expectation propagation (EP) with the greedy truncation order and one, two, three, four, and five EP iterations (GEP1, GEP2, GEP3, GEP4, GEP5), the variational Bayes (VB), and the analytical formulas of [28] using Matlab function mvncdf (MVNCDF). VB is an update of the skew *t* VB filter (STVBF) [20] where the heavy-tailedness variable Λ_1 is fixed to identity I_{n_y} and the VB iteration is terminated when the position estimate changes less than 0.005 m or at the 1000th iteration. The reference solution for the expectation value is an importance sampling (IS) update with 50 000 samples and the prior as the importance distribution.

Fig. 4 shows the distributions of the estimates' differences from the IS estimate. The errors are given per cent of the IS's estimation error. The box levels are 5 %, 25 %, 50 %, 75 %, and 95 % quantiles and the asterisks show minimum and maximum values. The results indicate that the accuracy of the EP approximation of the mean does not improve after two EP iterations. MVNCDF is slightly more accurate than GEP2 in the cases with high skewness, but MVNCDF's computational load is roughly 40 000 times that of the GEP2. This justifies the use of the EP approximation.

The approximation of the posterior covariance matrix is tested by studying the normalized estimation error squared (NEES) values [35, Ch. 5.4.2]

$$\mathrm{NEES}_k = (x_{k|k} - x_k)^{\mathrm{T}} P_{k|k}^{-1} (x_{k|k} - x_k), \tag{19}$$

where $x_{k|k}$ and $P_{k|k}$ are the filter's output mean and covariance matrix, and x_k is the true state. The algorithms' NEES_1 values averaged over the Monte Carlo replications are given in Table I. If the covariance matrix is correct, the NEES_1 is χ^2-distributed with 3 degrees of freedom because the position is 3-dimensional, so the nominal expected value is 3 [35, Ch. 5.4.2]. VB shows large average NEES_1 values when δ is large, which indicates that VB underestimates the covariance matrix. Apart from MVNCDF, the GEP algorithms show average NEES_1 values closest to 3, so the EP provides a more accurate covariance matrix approximation than VB. Indicated by average NEES_1 being slightly smaller than 3, GEP1 in fact overestimates the covariance matrix when δ is large, but this issue is mostly fixed by the second EP iteration.

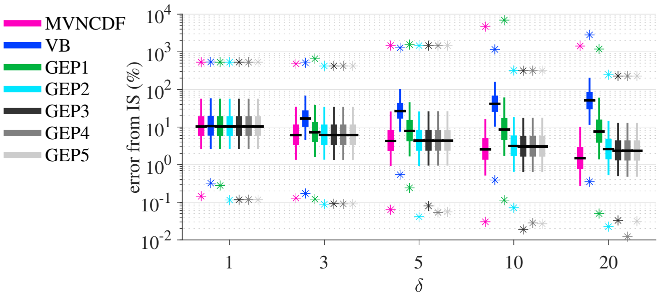


Fig. 4. Two EP iterations suffice. MVNCDF is slightly more accurate than the proposed GEP but computationally heavy.

Table I
The average NEES_1 values. GEP1's average NEES_1 is closer to the optimal value 3 than that of VB, so EP gives a more accurate posterior covariance matrix.

δ	1	3	5	10	20
MVNCDF	3.0	3.0	3.0	3.0	3.0
VB	3.8	9.1	19.1	65.6	229.2
GEP1	3.0	2.9	2.8	2.7	2.7
GEP2	3.0	3.0	3.0	2.9	2.9
GEP3	3.0	3.0	3.0	2.9	2.9
GEP4	3.0	3.0	3.0	2.9	2.9
GEP5	3.0	3.0	3.0	2.9	2.9

The order of the truncations in the EP algorithm affects the performance only when there are clear differences in the amounts of probability mass under each truncation. We compare GEP1 with the EP iteration with a random truncation order (REP1). In REP1 any of the non-optimal constraints is chosen randomly at each truncation. Fig. 5 presents an example where $\delta = 20$, and the measurement noise realization e has been generated from the skew normal distribution and then modified by

$$e_j = \min\{\min\{e_{1:n_y}\}, 0\} - c\sqrt{1 + 20^2}, \tag{20}$$

where j is a random index, and c is a parameter. A large c generates one negative outlier to each measurement vector, which results in one truncation with significantly larger truncated probability mass than the rest of the truncations. Fig. 5 shows the percentual difference of REP1 error from GEP1 error; i.e. a positive difference means that GEP1 is more accurate. The errors here refer to distance from the IS estimate. The figure shows that with large c GEP1 is more accurate than REP1. Thus, the effect of the truncation ordering on the accuracy of the EP approximation is more pronounced when there is one truncation that truncates much more than the rest. This justifies our greedy approach and the result of Lemma 1 for ordering the truncations.

The skew-*t* measurement model essentially implies that given the scaling variable Λ, we are observing the sum $Cx + \Delta u$ plus normally distributed noise. Fig. 6 compares the EP approximation and the 30-iteration VB approximation

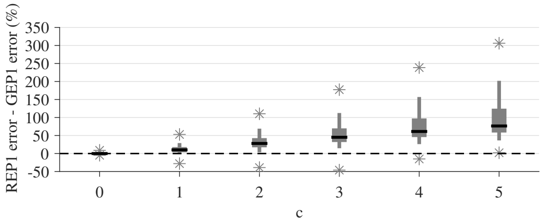
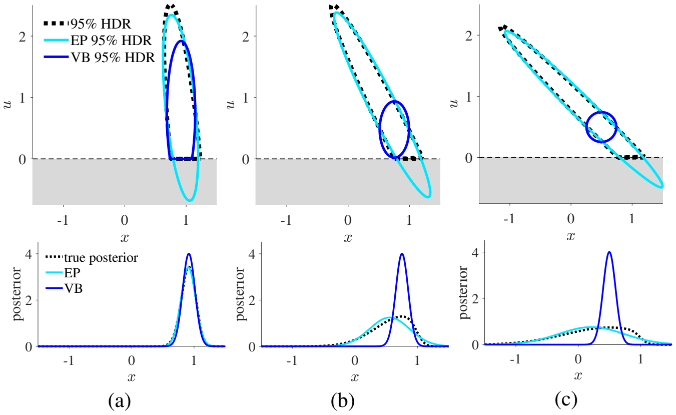


Fig. 5. The proposed GEP1 outperforms REP1 when one negative outlier is added to the measurement noise vector because there is one truncation that truncates much more probability than the rest.



(a)

(b)

(c)

Fig. 6. EP gives a better approximation than VB for a bivariate normal distribution of (x, u), where u is truncated to be positive. The figures show the 95 % high-density regions (HDR) of the posteriors $p(x, u|y = 1)$ (upper row) and the marginal posteriors $p(x|y = 1)$ (lower row) of the model (21). (a) $\delta = 0.1$, (b) $\delta = 0.5$, (c) $\delta = 1$.

of the posterior distribution for the model

$$p(x, u) = \mathcal{N}(x; 0, 1) \cdot \mathcal{N}_{+}(u; 0, 1) \qquad (21a)$$

$$p(y|x, u) = \mathcal{N}(y; x + \delta u, 0.1^2) \qquad (21b)$$

with the measurement value $y = 1$ and with δ values 0.1, 0.5, and 1. Fig. 6 illustrates that when δ is large, x and u are highly correlated. This makes VB seriously underestimate the covariance matrix, and EP provides a better approximation of the joint posterior and the marginal posterior of x.

B. Skew-*t* inference

In this section, the proposed skew-t filter (STF) is compared with state-of-the-art filters using numerical simulations of a 100-step trajectory. The tested STF uses two EP iterations. The measurement model is given in (17), and the state evolution model is a random walk with process noise covariance $Q = \operatorname{diag}(q^2, q^2, 0.2^2, 0)$ m, where q is a parameter. The compared methods are a bootstrap-type PF, STVBF [20], t variational Bayes filter (TVBF) [36], and Kalman filter (KF) with measurement validation gating [35, Ch. 5.7.2] that discards the measurement components whose normalized innovation squared is larger than the χ^2_1-distribution's 99 % quantile. The used KF parameters are the mean and variance of the used skew t-distribution, and the TVBF parameters are obtained by matching the degrees of freedom with that of the skew t-distribution and computing the maximum likelihood location and scale parameters for a set of pseudo-random numbers generated from the skew t-distribution. The results are based on 10 000 Monte Carlo replications.

Fig. 7 illustrates the filter iterations' convergence when the measurement noise components $[e_k]_i$ in (17) are generated independently from the univariate skew t-distribution. The figure shows that the proposed STF's median RMSE does not improve after five VB iterations, and STF outperforms the other filters in RMSE already with two VB iterations, except for PF that is the minimum-RMSE solution. Furthermore, Fig. 7 shows that STF's converged state is close to the PF's converged state in RMSE, and PF can require as many as 10 000 particles to outperform STF. In our implementation, the PF with 10 000 particles is computationally roughly 15 times heavier that the STF with five VB iterations. STF also converges faster than STVBF when the process noise variance parameter q is large.

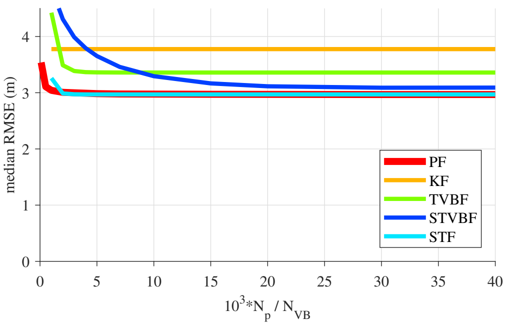


Fig. 7. The proposed STF's median RMSE does not improve after $N_{\mathrm{VB}} = 5$ VB iterations per time instant. The required number of PF particles N_{p} can be more than 10 000, and STVBF [20] can require 30 VB iterations. The x-axis is $10^3 \cdot N_{\mathrm{p}}$ for PF and N_{VB} for the rest of the filters. $q = 5$, $\delta = 5$.

Fig. 8 shows the distributions of the RMSE differences from the STF's RMSE as percentages of the STF's RMSE. STF1 is the skew-t filter with just one EP iteration per a VB iteration. STF, STF1, and TVBF use five VB iterations, and STVBF uses 30 VB iterations. STF clearly has the smallest RMSE when $\delta \geq 3$, i.e. when the skewness is high. STF1 and STF (with 2 EP iterations) have similar accuracies, so one EP iteration may be sufficient in practice. Unlike STVBF, the new STF improves accuracy even with small q and large δ, which can be explained by the improved covariance matrix approximation.

The proposed smoother is also tested with the measurement model (17) and the random-walk state model. The compared smoothers are the proposed skew-t smoother with two EP iterations (STS), skew-t variational Bayes smoother (STVBS) [20], t variational Bayes smoother (TVBS) [36], and the RTSS with 99 % measurement validation gating [27]. Fig. 9 shows that STS has lower RMSE than the smoothers based on symmetric distributions. Furthermore, STS's VB iteration converges in five iterations or less, so it is faster than STVBS.

V. Tests with real data

Two GNSS positioning data sets were collected in central London (UK) to test the filters' performance in a challenging real-world satellite positioning scenario with numerous non-line-of-sight measurements. The data include time-of-flight based pseudorange measurements from GPS satellites. Each set contains a trajectory that was collected by car using a u-blox M8 GNSS receiver. The lengths of the tracks are about 8 km and 10 km, the durations are about an hour for each, and measurements are received at about one-second intervals. The first track is used for fitting the filter parameters, while the second track is used for studying the filters' positioning errors. A ground truth was measured using an Applanix POS-LV220 system that improves the GNSS solution with tactical

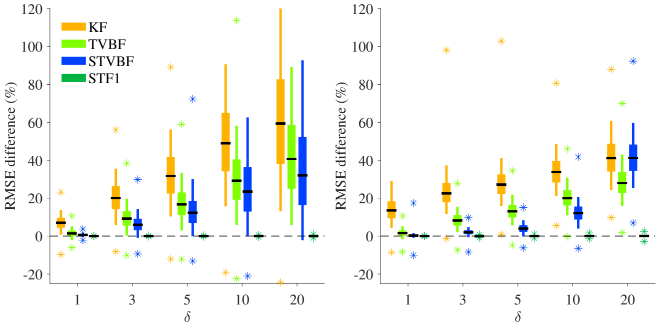


Fig. 8. The proposed STF outperforms the comparison methods with skew-t-distributed noise. RMSE differences from STF's RMSE per cent of the STF's RMSE. The difference to STVBF [20] increases as skewness increases or when process noise variance reduces. (left) $q=0.5$, (right) $q=5$.

grade inertial measurement units. The GPS satellites' locations were obtained from the broadcast ephemerides provided by the International GNSS Service [34]. The algorithms are computed with MATLAB.

In this test, both the user position $l_k \in \mathbb{R}^3$ and the receiver clock error $b_k \in \mathbb{R}$ follow the almost-constant velocity model used in [37, Section IV]. Thus, the filter state being $x_k = [\, l_k^{\mathrm T}\ \dot{l}_k^{\mathrm T}\ b_k\ \dot{b}_k \,]^{\mathrm T} \in \mathbb{R}^8$, the state evolution model is

$$x_{k+1} = \begin{bmatrix} I_3 & d_k I_3 & \mathrm{O}_{3\times2} \\ \mathrm{O}_3 & I_3 & \mathrm{O}_{3\times2} \\ \mathrm{O}_{2\times3} & \mathrm{O}_{2\times3} & \left[\begin{smallmatrix} 1 & d_k \\ 0 & 1 \end{smallmatrix}\right] \end{bmatrix} x_k + w_k, \tag{22}$$

where

$$w_k \sim \mathcal{N}\left(0, \begin{bmatrix} \frac{q^2 d_k^3}{3} I_3 & \frac{q^2 d_k^2}{2} I_3 & \mathrm{O}_{3\times2} \\ \frac{q^2 d_k^2}{2} I_3 & q^2 d_k I_3 & \mathrm{O}_{3\times2} \\ \mathrm{O}_{2\times3} & \mathrm{O}_{2\times3} & \begin{bmatrix} s_b d_k + \frac{s_f d_k^3}{3} & \frac{s_f d_k^2}{2} \\ \frac{s_f d_k^2}{2} & s_f d_k \end{bmatrix} \end{bmatrix}\right),$$

and d_k is the time difference of the measurements in seconds. The used parameter values are $q = 0.5\,\mathrm{m/s^{\frac{3}{2}}}$, $s_b = 70\,\frac{\mathrm{m}^2}{\mathrm{s}}$, and $s_f = 0.6\,\frac{\mathrm{m}^2}{\mathrm{s}^3}$. The initial prior is a normal distribution with mean given by the Gauss–Newton method with the first measurement and a large covariance matrix.

The measurement model is the same pseudorange model that is used in the simulations of Section IV, i.e.

$$[y_k]_i = \|s_{i,k} - [x_k]_{1:3}\| + [x_k]_7 + [e_k]_i, \tag{23}$$

where $s_{i,k}$ is the 3-dimensional position of the ith satellite at the time of transmission. The measurement model is linearized with respect to x_k at each prior mean using the first order Taylor series approximation. The compared filters are based on three different models for the measurement noise e_k where

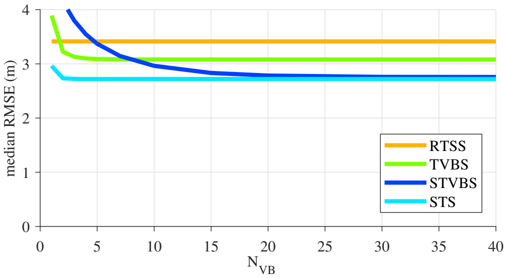


Fig. 9. Five STS iterations give the converged state's RMSE, while STVBS [20] can require 30 iterations. $q=5$, $\delta=5$.

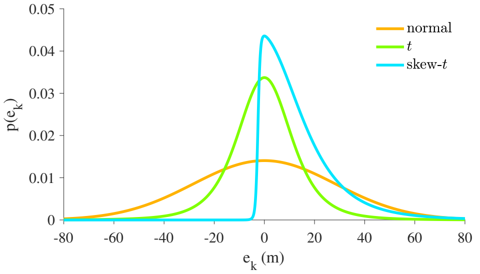


Fig. 10. Measurement error distributions fitted to the real GNSS data for normal, t, and skew-t error models. The modes are fixed to zero.

Table II
FILTER PARAMETERS FOR REAL GNSS DATA

Skew-t, $\nu=4$			t, $\nu_{\mathrm t}=4$		Normal	
μ (m)	σ (m)	δ (m)	$\mu_{\mathrm t}$ (m)	$\sigma_{\mathrm t}$ (m)	$\mu_{\mathrm n}$ (m)	$\sigma_{\mathrm n}$ (m)
-2.5	0.8	16.8	0	11.1	0	28.4

$$[e_k]_i \sim \mathrm{ST}(\mu, \sigma^2, \delta, \nu); \tag{24}$$
$$[e_k]_i \sim \mathcal{T}(\mu_{\mathrm t}, \sigma_{\mathrm t}^2, \nu_{\mathrm t}); \tag{25}$$
$$[e_k]_i \sim \mathcal{N}(\mu_{\mathrm n}, \sigma_{\mathrm n}^2). \tag{26}$$

The skew-t model (24) is the basis for STF and STVBF, the t model (25) is the basis for TVBF, and the normal model (26) is the basis for the extended KF (EKF) with 99 % measurement validation gating. The pseudoranges are unbiased in the line-of-sight case, so the location parameters are fixed to $\mu_{\mathrm n} = \mu_{\mathrm t} = 0$. Furthermore, the degrees of freedom are fixed to $\nu = \nu_{\mathrm t} = 4$, which according to our experience is in general a good compromise between outlier robustness and performance based on inlier measurements, provides infinite kurtosis but finite skewness and variance, and is recommended in [38]. The deviation parameter $\sigma_{\mathrm n}$ of the normal model was then fitted to the data using the expectation–maximization algorithm [39, Ch. 12.3.3] and the parameter $\sigma_{\mathrm t}$ of the t model as well as the parameters σ and δ of the skew-t model were fitted with the particle–Metropolis algorithm [39, Ch. 12.3.4]. The location parameter μ was obtained by numerically finding the point that sets the mode of the skew-t noise distribution to zero. Furthermore, we added a heuristic method for mitigating the STVBF's covariance underestimation, namely a posterior covariance scaling factor that scales each STVBF posterior covariance matrix with the number 3.25^2. This scaling was found to provide the lowest RMSE for our data set, which ensures that we do not favor the proposed STF over STVBF. These three error distributions' parameters are given in Table II, and the PDFs are plotted in Fig. 10.

Fig. 11 shows the filter RMSEs as a function of the number of VB iterations. Both STF and TVBF converge within five VB iterations, while the STVBF does not converge within 30 iterations but requires about 150 iterations. The empirical CDF graphs of the user position errors with five VB iterations for STF and TVBF and with 150 iterations for STVBF are shown in Fig. 12, and the RMSEs as well as the relative running times are given in Table III. The results show that modelling the skewness improves the positioning accuracy and is important especially for the accuracy in vertical direction. This can be explained by the sensitivity of the vertical direction to large measurement errors; due to bad measurement geometry the accuracy in the vertical direction is low even with line-of-

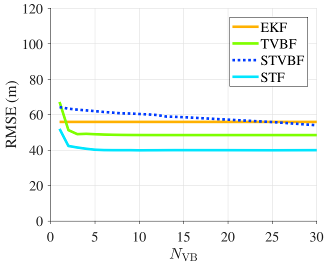


Fig. 11. RMSE of horizontal (left) and vertical (right) position for real GNSS data as a function of the number of VB iterations

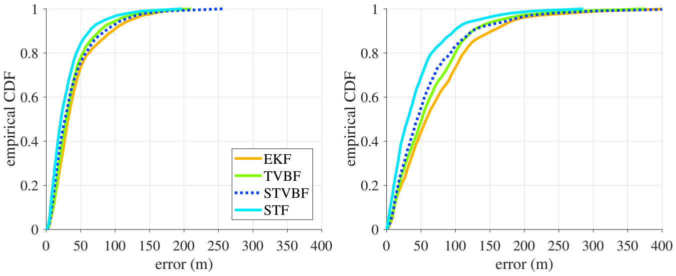


Fig. 12. Empirical error CDFs for the real GNSS data for the horizontal error (left) and the vertical error (right)

sight measurements, so correct downweighting of erroneous altitude information requires careful modelling of the noise distribution's tails. The computational burden of our STF implementation with five VB iterations is almost four times that of TVBF, but Fig. 11 shows that two STF iterations would already be enough to match TVBF's average RMSE.

Fig. 12 and Table III also show that the proposed STF is more accurate than STVBF despite STVBF being considerable heavier computationally due to STVBF's 150 VB iterations. Furthermore, achieving this STVBF performance required awkward and data-dependent tuning to reduce the underestimation of posterior covariance matrix. The issues shown by STVBF are probably due to the highly skewed measurement noise distribution.

VI. Extension to MVST

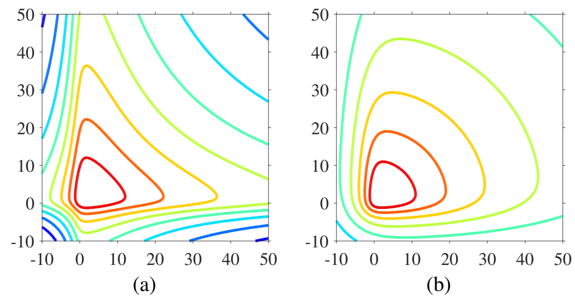
The skew t-distribution has several multivariate versions. In [5]–[7] the PDF of the multivariate skew t-distribution (MVST) involves the CDF of a univariate t-distribution, while the definition of skew t-distribution given in [40] involves the CDF of a multivariate t-distribution. These versions of MVST are special cases of more general multivariate skew-t-type distribution families, which include the multivariate canonical fundamental skew t-distribution (CFUST) [41] and the multivariate unified skew t-distribution [42]. A comprehensive review on the different variants of the MVST is given in [23].

The MVST variant used in this article is based on the CFUST discussed in [23], and it is the most general variant of the MVST. In this variant the parameter matrix $R \in \mathbb{R}^{n_z \times n_z}$ is a square positive-definite matrix, and $\Delta \in \mathbb{R}^{n_z \times n_z}$ is an arbitrary matrix. The PDF is

$$\mathrm{MVST}(z; \mu, R, \Delta, \nu) = 2^{n_z} \mathrm{t}(z; \mu, \Omega, \nu)\, \mathrm{T}(\overline{z}; 0, L, \nu + n_z), \tag{27}$$

Table III
The RMSEs and relative running times for real GNSS data

	EKF	TVBF	STVBF	STF
RMSE$_{\text{horizontal}}$ (m)	56	49	52	40
RMSE$_{\text{vertical}}$ (m)	95	84	84	61
Running time	1	1.3	18.7	3.8



(a)

(b)

Fig. 13. PDF of bivariate measurement noise from (a) independent univariate skew-t components model (3) with $\Delta = 5I_2$, $R = I_2$, $\nu = \left[\begin{smallmatrix}4\\4\end{smallmatrix}\right]$ and (b) MVST model (30) with $\Delta = 5I_2$, $R = I_2$, $\nu = 4$.

where $L = I_{n_z} - \Delta^{\mathrm{T}} \Omega^{-1} \Delta$, $\Omega = R + \Delta\Delta^{\mathrm{T}}$,

$$\begin{aligned}&\mathrm{t}(z; \mu, \Sigma, \nu)\\ &= \frac{\Gamma\left(\frac{\nu+n_z}{2}\right)}{(\nu\pi)^{\frac{n_z}{2}} \det(\Sigma)^{\frac{1}{2}} \Gamma\left(\frac{\nu}{2}\right)} \left(1 + \tfrac{1}{\nu}(z-\mu)^{\mathrm{T}} \Sigma^{-1} (z-\mu)\right)^{-\frac{\nu+n_z}{2}}\end{aligned} \tag{28}$$

is the PDF of the n_z-variate t-distribution and $\mathrm{T}(z; \mu, \Sigma, \nu)$ its CDF, and

$$\overline{z} = \Delta^{\mathrm{T}} \Omega^{-1} (z - \mu) \sqrt{\tfrac{\nu + n_z}{\nu + (z-\mu)^{\mathrm{T}} \Omega^{-1} (z-\mu)}}. \tag{29}$$

The inference algorithms proposed in this paper can be extended to cover the case where the elements of the measurement noise vector are not statistically independent but jointly multivariate skew-t-distributed. When the measurement noise follows a MVST, i.e.

$$e_k \sim \mathrm{MVST}(0, R, \Delta, \nu), \tag{30}$$

the smoothing and filtering algorithms presented in Algorithms 1 and 2 apply with slight modifications. At the core of this convenient extension is the fact that the MVST can be represented by a similar hierarchical model as in (6). However, the shape matrices Δ and R are not required to be diagonal, and the matrix Λ_k has the form $\lambda_k \cdot I_{n_y}$, where λ_k is a scalar with the prior

$$\lambda_k \sim \mathcal{G}(\tfrac{\nu}{2}, \tfrac{\nu}{2}). \tag{31}$$

Notice that when λ_k admits a small value, all the measurement components can potentially be outliers simultaneously unlike with the independent univariate skew-t components model. A univariate skew-t is also a MVST, but a vector of univariate independently skew-t distributed components is not a special case of MVST. This difference is illustrated by the PDF contour plots in Fig. 13. See also further discussion in [23].

The specific modification required by MVST measurement noise to the STS algorithm in Algorithm 1 is replacing line 23 by

$$\Lambda_{k|K} \leftarrow \frac{\nu + 2n_y}{\nu + \mathrm{tr}\{\Psi_k\}} \cdot I_{n_y} \tag{32}$$

Similarly, the specific modification required by MVST measurement noise to the STF algorithm in Algorithm 2 is replacing line 15 by

$$\Lambda_{k|k} \leftarrow \frac{\nu + 2n_y}{\nu + \operatorname{tr}\{\Psi_k\}} \cdot I_{n_y}. \tag{33}$$

VII. Performance Bound

A. Cramér–Rao lower bound

The Bayesian Cramér–Rao lower bound (CRLB) B is a lower bound for the mean-square-error (MSE) matrix of the state estimator $\hat{x}$ of the random variable x using the observations y

$$M = \mathop{\mathbb{E}}_{p(x,y)} [(x - \hat{x})(x - \hat{x})^{\mathrm{T}}] \tag{34}$$

in the sense that the matrix difference $M - B$ is positive semidefinite for any state estimator [43, Ch. 2.4]. The regularity conditions sufficient for the positive-semidefiniteness to hold [43, Ch. 2.4] are the integrability of the first two partial derivatives of the joint density $p(x_{1:k}, y_{1:k})$ for an asymptotically unbiased estimator. These conditions are satisfied by the skew-t likelihood and the normal prior distribution, even though they do not hold for $p(x_{1:k}, u_{1:k}, \Lambda_{1:k}, y_{1:k})$ of the hierarchical model used in the proposed variational estimator due to restriction of $u_{1:k}$ to the positive orthant. This is sufficient, since we only seek the CRLB for the actual state x, not for the artificial variables u and Λ.

The filtering CRLB $B_{k|k}$ for the state-space model (1)–(2) follows the recursion [44]

$$B_{1|0} = P_{1|0} \tag{35a}$$

$$B_{k+1|k+1} = \left((AB_{k|k}A^{\mathrm{T}} + Q)^{-1} + \mathop{\mathbb{E}}_{p(x_k|y_{1:k-1})}[\mathcal{I}(x_k)]\right)^{-1}, \tag{35b}$$

where $\mathcal{I}(e_k)$ is the Fisher information matrix of the measurement noise distribution. Furthermore, the smoothing CRLB for the state-space model (1)–(2) follows the recursion [44]

$$B_{k|K} = B_{k|k} + G_k(B_{k+1|K} - B_{k+1|k})G_k^{\mathrm{T}}, \tag{36}$$

where

$$G_k = B_{k|k}A^{\mathrm{T}}B_{k+1|k}^{-1}, \tag{37}$$

$$B_{k+1|k} = AB_{k|k}A^{\mathrm{T}} + Q. \tag{38}$$

This coincides with the covariance matrix update of Rauch–Tung–Striebel smoother's backward recursion [27].

The Fisher information matrix for the multivariate skew-t-distributed measurement noise of (30) is

$$\mathcal{I}(x) = C^{\mathrm{T}}(R + \Delta\Delta^{\mathrm{T}})^{-\frac{\mathrm{T}}{2}} E (R + \Delta\Delta^{\mathrm{T}})^{-\frac{1}{2}} C, \tag{39}$$

where

$$E = \mathop{\mathbb{E}}_{p(r)} \left[\tfrac{\nu + n_y}{\nu + r^{\mathrm{T}} r} \left(I_{n_y} - \tfrac{2}{\nu + r^{\mathrm{T}} r} r r^{\mathrm{T}} + \widetilde{R}_r \widetilde{R}_r^{\mathrm{T}} \right) \right] \tag{40}$$

with $r \sim \mathrm{MVST}(0, I_{n_y} - \Theta\Theta^{\mathrm{T}}, \Theta, \nu)$, $\Theta = (R + \Delta\Delta^{\mathrm{T}})^{-\frac{1}{2}}\Delta$, $A^{\frac{1}{2}}$ is a square-root matrix such that $A^{\frac{1}{2}}(A^{\frac{1}{2}})^{\mathrm{T}} = A$, $A^{-\frac{1}{2}} \triangleq (A^{\frac{1}{2}})^{-1}$, $A^{-\frac{\mathrm{T}}{2}} \triangleq ((A^{\frac{1}{2}})^{-1})^{\mathrm{T}}$, and

$$\begin{aligned} \widetilde{R}_r = {} & \left(\mathrm{T}\left(\Theta^{\mathrm{T}} r \sqrt{\tfrac{\nu + n_y}{\nu + r^{\mathrm{T}} r}}; 0, I_{n_y} - \Theta^{\mathrm{T}}\Theta, \nu + n_y \right) \right)^{-1} \\ & \times (I_{n_y} - \tfrac{1}{\nu + r^{\mathrm{T}} r} r r^{\mathrm{T}}) \Theta \\ & \times \nabla_u \mathrm{T}(u; 0, I_{n_y} - \Theta^{\mathrm{T}}\Theta, \nu + n_y) \Big|_{u = \Theta^{\mathrm{T}} r \sqrt{\frac{\nu + n_y}{\nu + r^{\mathrm{T}} r}}}, \end{aligned} \tag{41}$$

where ∇_u is the gradient with respect to u. The derivation is given in Appendix B. The evaluation of the expectation in (40) is challenging with high-dimensional measurements due to the requirement to evaluate the CDF of the multivariate t-distribution and its partial derivatives. By the Woodbury matrix identity, the recursion (35) is equivalent to the covariance matrix update of the Kalman filter with the measurement noise covariance $(R + \Delta\Delta^{\mathrm{T}})^{\frac{1}{2}} E^{-1} ((R + \Delta\Delta^{\mathrm{T}})^{\frac{1}{2}})^{\mathrm{T}}$.

In the model (3) the measurement noise components are independently univariate skew-t-distributed. In this case the Fisher information is obtained by applying (39) to each conditionally independent measurement component and summing. The resulting formula matches with (39), the matrix E now being a diagonal matrix with the diagonal entries

$$\begin{aligned} E_{ii} = \mathop{\mathbb{E}}_{p(r_i)} \Bigg[& \tfrac{\nu_i - r_i^2}{(\nu_i + r_i^2)^2} \\ & + \tfrac{\theta_i^2}{1 - \theta_i^2} \tfrac{\nu_i^2}{(\nu_i + r_i^2)^3} \left(\tau_{\nu_i + 1}\left(\tfrac{\theta_i}{\sqrt{1 - \theta_i^2}} r_i \sqrt{\tfrac{\nu_i + 1}{\nu_i + r_i^2}} \right) \right)^2 \Bigg], \end{aligned} \tag{42}$$

where $r_i \sim \mathrm{ST}(0, 1 - \theta_i^2, \theta_i, \nu_i)$ is a univariate skew-t-distributed random variable, $\theta_i = \Delta_{ii}/\sqrt{R_{ii} + \Delta_{ii}^2}$ and $\tau_\nu(x) = \mathrm{t}(x; 0, 1, \nu)/\mathrm{T}(x; 0, 1, \nu)$. Substituted into (39), this formula matches the Fisher information formula obtained for the univariate skew t-distribution in [45]. In this case only integrals with respect to one scalar variable are to be evaluated numerically.

B. Simulation

We study the CRLB in (35) of a linear state-space model with skew-t-distributed measurement noise by generating realizations of the model

$$x_{k+1} = \left[\begin{smallmatrix} 1 & 1 \\ 0 & 1 \end{smallmatrix}\right] x_k + w_k, \;\; w_k \sim \mathcal{N}(0, Q) \tag{43a}$$

$$y_k = \left[\begin{smallmatrix} 1 & 0 \end{smallmatrix}\right] x_k + e_k, \;\; e_k \sim \mathrm{ST}(\mu, \sigma^2, \delta, \nu), \tag{43b}$$

where $x \in \mathbb{R}^2$ is the state, $Q = \left[\begin{smallmatrix} 1/3 & 1/2 \\ 1/2 & 1 \end{smallmatrix}\right]$ is the process noise covariance matrix, $y_k \in \mathbb{R}$ is the measurement, and ν and δ_c are parameters that determine other parameters by the formulas

$$\mu = -\gamma\delta_c\sigma, \tag{44a}$$

$$\sigma^2 = \tfrac{\omega^2}{\frac{\nu}{\nu - 2}(1 + \delta_c^2) - \gamma^2\delta_c^2}, \tag{44b}$$

$$\delta = \delta_c\sigma, \tag{44c}$$

$$\gamma = \sqrt{\tfrac{\nu}{\pi}} \tfrac{\Gamma((\nu - 1)/2)}{\Gamma(\nu/2)}. \tag{44d}$$

Thus, the measurement noise distribution is zero-mean and has the variance $\omega^2 = 5^2$. We generate 10 000 realizations of a 50-step process, and compute the CRLB and mean-square-errors (MSE) of the bootstrap PF with 2000 particles and the STF. The CRLB and the MSEs were computed for the first component of the state at the last time instant $[x_{50}]_1$.

Fig. 14 shows the CRLB of the model (43). The figure shows that increase in the skewness as well as heavy-tailedness can decrease the CRLB significantly, which suggests that a nonlinear filter can be significantly better than the KF, which gives MSE 11.8 for all δ_c and ν. Fig. 15 shows the MSEs of PF and STF. As expected, when $\nu \to \infty$ and $\delta_c \to 0$, the PF's MSE approaches the CRLB. STF is only slightly worse than PF. The figures also show that although the CRLB becomes looser when the distribution becomes more skewed and/or heavy-tailed, it correctly indicates that modeling the skewness still improves the filtering performance.

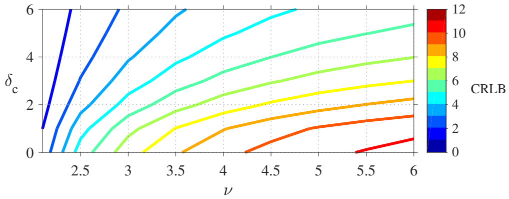


Fig. 14. The CRLB of the 50th time instant for the model (43) with a fixed measurement noise variance. Skewness and heavy-tailedness decreases the CRLB significantly.

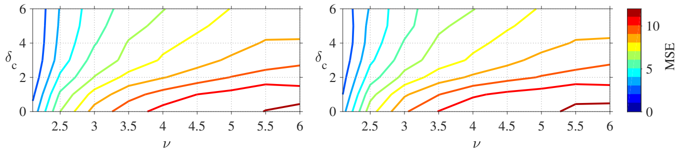


Fig. 15. The MSEs of PF (left) and STF's (right) are close to each other.

VIII. Conclusions

We have proposed a novel approximate filter and smoother for linear state-space models with heavy-tailed and skewed measurement noise distribution, and derived the Cramér–Rao lower bounds for the filtering and smoothing estimators. The algorithms are based on the variational Bayes approximation, where some posterior independence approximations are removed from the earlier versions of the algorithms to avoid significant underestimation of the posterior covariance matrix. Removal of independence approximations is enabled by the expectation propagation (EP) algorithm for approximating the mean and covariance matrix of truncated multivariate normal distribution. A greedy processing sequence is given for the EP. Simulations and real-data tests with GNSS positioning data show that the proposed algorithms outperform the state-of-the-art low-complexity methods, including the earlier skew-t VB filter, in a real-world estimation problem.

References

[1] F. Gustafsson and F. Gunnarsson, "Mobile positioning using wireless networks: possibilities and fundamental limitations based on available wireless network measurements," *IEEE Signal Processing Magazine*, vol. 22, no. 4, pp. 41–53, July 2005.

[2] B.-S. Chen, C.-Y. Yang, F.-K. Liao, and J.-F. Liao, "Mobile location estimator in a rough wireless environment using Extended Kalman-based IMM and data fusion," *IEEE Transactions on Vehicular Technology*, vol. 58, no. 3, pp. 1157–1169, March 2009.

[3] M. Kok, J. D. Hol, and T. B. Schön, "Indoor positioning using ultra-wideband and inertial measurements," *IEEE Transactions on Vehicular Technology*, vol. 64, no. 4, 2015.

[4] K. Kaemarungsi and P. Krishnamurthy, "Analysis of WLAN's received signal strength indication for indoor location fingerprinting," *Pervasive and Mobile Computing*, vol. 8, no. 2, pp. 292–316, 2012, special Issue: Wide-Scale Vehicular Sensor Networks and Mobile Sensing.

[5] M. D. Branco and D. K. Dey, "A general class of multivariate skew-elliptical distributions," *Journal of Multivariate Analysis*, vol. 79, no. 1, pp. 99–113, October 2001.

[6] A. Azzalini and A. Capitanio, "Distributions generated by perturbation of symmetry with emphasis on a multivariate skew t-distribution," *Journal of the Royal Statistical Society. Series B (Statistical Methodology)*, vol. 65, no. 2, pp. 367–389, 2003.

[7] A. K. Gupta, "Multivariate skew t-distribution," *Statistics*, vol. 37, no. 4, pp. 359–363, 2003.

[8] H. Nurminen, T. Ardeshiri, R. Piché, and F. Gustafsson, "A NLOS-robust TOA positioning filter based on a skew-t measurement noise model," in *International Conference on Indoor Positioning and Indoor Navigation (IPIN)*, October 2015, pp. 1–7.

[9] S. Frühwirth-Schnatter and S. Pyne, "Bayesian inference for finite mixtures of univariate and multivariate skew-normal and skew-t distributions," *Biostatistics*, vol. 11, no. 2, pp. 317–336, 2010.

[10] N. Counsell, M. Cortina-Borja, A. Lehtonen, and A. Stein, "Modelling psychiatric measures using skew-normal distributions," *European Psychiatry*, vol. 26, no. 2, pp. 112–114, 2010.

[11] M. Eling, "Fitting insurance claims to skewed distributions: Are the skew-normal and skew-student good models?" *Insurance: Mathematics and Economics*, vol. 51, no. 2, pp. 239–248, 2012.

[12] Y. V. Marchenko, "Multivariate skew-t distributions in econometrics and environmetrics," Ph.D. dissertation, Texas A&M University, December 2010.

[13] A. Doucet, S. Godsill, and C. Andrieu, "On sequential Monte Carlo sampling methods for Bayesian filtering," *Statistics and Computing*, vol. 10, no. 3, pp. 197–208, July 2000.

[14] P. Naveau, M. G. Genton, and X. Shen, "A skewed Kalman filter," *Journal of Multivariate Analysis*, vol. 94, pp. 382–400, 2005.

[15] H.-M. Kim, D. Ryu, B. K. Mallick, and M. G. Genton, "Mixtures of skewed Kalman filters," *Journal of Multivariate Analysis*, vol. 123, pp. 228–251, 2014.

[16] J. Rezaie and J. Eidsvik, "Kalman filter variants in the closed skew normal setting," *Computational Statistics and Data Analysis*, vol. 75, pp. 1–14, 2014.

[17] D. L. Alspach and H. W. Sorenson, "Nonlinear Bayesian estimation using Gaussian sum approximations," *IEEE Transactions on Automatic Control*, vol. 17, no. 4, pp. 439–448, Aug. 1972.

[18] Y. Bar-Shalom and T. Fortmann, *Tracking and Data Association*, ser. Mathematics in Science and Engineering Series. Academic Press, 1988.

[19] J. L. Williams and P. S. Maybeck, "Cost-function-based hypothesis control techniques for multiple hypothesis tracking," *Mathematical and Computer Modelling*, vol. 43, no. 9–10, pp. 976–989, May 2006.

[20] H. Nurminen, T. Ardeshiri, R. Piche, and F. Gustafsson, "Robust inference for state-space models with skewed measurement noise," *IEEE Signal Processing Letters*, vol. 22, no. 11, pp. 1898–1902, 2015.

[21] M. P. Wand, J. T. Ormerod, S. A. Padoan, and R. Frühwirth, "Mean field variational Bayes for elaborate distributions," *Bayesian Analysis*, vol. 6, no. 4, pp. 847–900, 2011.

[22] C. M. Bishop, *Pattern Recognition and Machine Learning*. Springer, 2007.

[23] S. X. Lee and G. J. McLachlan, "Finite mixtures of canonical fundamental skew t-distributions – the unification of the restricted and unrestricted skew t-mixture models," *Statistics and Computing*, no. 26, pp. 573–589, 2016.

[24] T. M. Cover and J. Thomas, *Elements of Information Theory*. John Wiley and Sons, 2006.

[25] D. G. Tzikas, A. C. Likas, and N. P. Galatsanos, "The variational approximation for Bayesian inference," *IEEE Signal Processing Magazine*, vol. 25, no. 6, pp. 131–146, Nov. 2008.

[26] M. J. Beal, "Variational algorithms for approximate Bayesian inference," Ph.D. dissertation, Gatsby Computational Neuroscience Unit, University College London, 2003.

[27] H. E. Rauch, C. T. Striebel, and F. Tung, "Maximum Likelihood Estimates of Linear Dynamic Systems," *Journal of the American Institute of Aeronautics and Astronautics*, vol. 3, no. 8, pp. 1445–1450, 1965.

[28] G. Tallis, "The moment generating function of the truncated multi-normal distribution," *Journal of the Royal Statistical Society. Series B (Methodological)*, vol. 23, no. 1, pp. 223–119, 1961.

[29] A. Genz, "Numerical computation of rectangular bivariate and trivariate normal and t probabilities," *Statistics and Computing*, vol. 14, pp. 251–260, 2004.

[30] A. Genz and F. Bretz, "Comparison of methods for the computation of multivariate t probabilities," *Journal of Computational and Graphical Statistics*, vol. 11, no. 4, pp. 950–971, 2002.

[31] T. P. Minka, "Expectation propagation for approximate Bayesian inference," in *17th Annual Conference on Uncertainty in Artificial Intelligence (UAI)*, 2001, pp. 362–369.

[32] J. P. Cunningham, P. Hennig, and S. Lacoste-Julien, "Gaussian probabilities and expectation propagation," Arxiv, November 2013. [Online]. Available: arxiv.org/abs/1111.6832

[33] K. P. Murphy, *Machine Learning: A Probabilistic Perspective*. Cambridge, MA: The MIT Press, 2012.

[34] J. M. Dow, R. Neilan, and C. Rizos, "The international GNSS service in a changing landscape of global navigation satellite systems," *Journal of Geodesy*, vol. 83, no. 7, p. 689, February 2009.

[35] Y. Bar-Shalom, R. X. Li, and T. Kirubarajan, *Estimation with Applications to Tracking and Navigation, Theory Algorithms and Software*. John Wiley & Sons, 2001.

[36] R. Piché, S. Särkkä, and J. Hartikainen, "Recursive outlier-robust filtering and smoothing for nonlinear systems using the multivariate Student-t distribution," in *IEEE International Workshop on Machine Learning for Signal Processing (MLSP)*, September 2012.

[37] P. Axelrad and R. Brown, "GPS navigation algorithms," in *Global Positioning System: Theory and Applications I*, B. W. Parkinson and J. J. Spilker Jr., Eds. Washington D.C.: AIAA, 1996, ch. 9.

[38] K. L. Lange, R. J. Little, and J. M. Taylor, "Robust statistical modeling using the t distribution," *Journal of the Americal Statistical Association*, vol. 84, no. 408, pp. 881–896, December 1989.

[39] S. Särkkä and J. Hartikainen, "On Gaussian optimal smoothing of non-linear state space models," *IEEE Transactions on Automatic Control*, vol. 55, no. 8, pp. 1938–1941, August 2010.

[40] S. K. Sahu, D. K. Dey, and M. D. Branco, "A new class of multivariate skew distributions with applications to Bayesian regression models," *Canadian Journal of Statistics*, vol. 31, no. 2, pp. 129–150, 2003.

[41] R. B. Arellano-Valle and M. G. Genton, "On fundamental skew distributions," *Journal of Multivariate Analysis*, no. 96, pp. 93–116, 2005.

[42] ——, "Multivariate extended skew-*t* distributions and related families," *METRON - International Journal of Statistics*, vol. 68, no. 3, pp. 201–234, 2010.

[43] H. L. Van Trees, *Detection, Estimation, and Modulation Theory, Part I: Detection, Estimation, and Linear Modulation Theory*. New York: John Wiley & Sons, Inc., 1968.

[44] M. Šimandl, J. Královec, and P. Tichavský, "Filtering, predictive, and smoothing Cramér–Rao bounds for discrete-time nonlinear dynamic systems," *Automatica*, vol. 37, pp. 1703–1716, 2001.

[45] T. J. Di Ciccio and A. C. Monti, "Inferential aspects of the skew *t*-distribution," *Quaderni di Statistica*, vol. 13, pp. 1–21, 2011.

[46] S. Särkkä, *Bayesian Filtering and Smoothing*. Cambridge, UK: Cambridge University Press, 2013.

[47] R. Piché, "Cramér-Rao lower bound for linear filtering with t-distributed measurements," in *19th International Conference on Information Fusion (FUSION)*, July 2016, pp. 536–540.

Appendix A
Derivations for the skew-t smoother

A. Derivations for q_{xu}

Eq. (10a) gives

$$
\begin{aligned}
\log q_{xu}(x_{1:K}, u_{1:K}) &= \log \mathcal{N}(x_1; x_{1|0}, P_{1|0}) \\
&+ \sum_{l=1}^{K-1} \log \mathcal{N}(x_{l+1}; Ax_l, Q) \\
&+ \sum_{k=1}^{K} \mathop{\mathbb{E}}_{q_\Lambda}[\log \mathcal{N}(y_k; Cx_k + \Delta u_k, \Lambda_k^{-1} R) \\
&+ \log \mathcal{N}_+(u_k; 0, \Lambda_k^{-1})] + c \qquad (45)\\
= \log \mathcal{N}(x_1; x_{1|0}, P_{1|0}) &+ \sum_{l=1}^{K-1} \log \mathcal{N}(x_{l+1}; Ax_l, Q) \\
&- \frac{1}{2} \sum_{k=1}^{K} \mathop{\mathbb{E}}_{q_\Lambda}[(y_k - Cx_k - \Delta u_k)^{\mathrm{T}} R^{-1} \Lambda_k (y_k - Cx_k - \Delta u_k) \\
&+ u_k^{\mathrm{T}} \Lambda_k u_k] + c \qquad (46)\\
= \log \mathcal{N}(x_1; x_{1|0}, P_{1|0}) &+ \sum_{l=1}^{K-1} \log \mathcal{N}(x_{l+1}; Ax_l, Q) \\
&- \frac{1}{2} \sum_{k=1}^{K} \{(y_k - Cx_k - \Delta u_k)^{\mathrm{T}} R^{-1} \Lambda_{k|K} (y_k - Cx_k - \Delta u_k) \\
&+ u_k^{\mathrm{T}} \Lambda_{k|K} u_k\} + c \qquad (47)\\
= \log \mathcal{N}(x_1; x_{1|0}, P_{1|0}) &+ \sum_{l=1}^{K-1} \log \mathcal{N}(x_{l+1}; Ax_l, Q) \\
&+ \sum_{k=1}^{K} \{\log \mathcal{N}(y_k; Ax_k + \Delta u_k, \Lambda_{k|K}^{-1} R) \\
&+ \log \mathcal{N}(u_k; 0, \Lambda_{k|K}^{-1})\} + c \qquad (48)\\
= \log \mathcal{N}\left(\left[\begin{smallmatrix} x_1 \\ u_1 \end{smallmatrix}\right]; \left[\begin{smallmatrix} x_{1|0} \\ 0 \end{smallmatrix}\right], \left[\begin{smallmatrix} P_{1|0} & \mathrm{O} \\ \mathrm{O} & \Lambda_{1|K}^{-1} \end{smallmatrix}\right]\right) \\
&+ \sum_{l=1}^{K-1} \log \mathcal{N}\left(\left[\begin{smallmatrix} x_{l+1} \\ u_{l+1} \end{smallmatrix}\right]; \left[\begin{smallmatrix} A & \mathrm{O} \\ \mathrm{O} & \mathrm{O} \end{smallmatrix}\right] \left[\begin{smallmatrix} x_l \\ u_l \end{smallmatrix}\right], \left[\begin{smallmatrix} Q & \mathrm{O} \\ \mathrm{O} & \Lambda_{l+1|K}^{-1} \end{smallmatrix}\right]\right) \\
&+ \log \mathcal{N}\left(y_k; \left[\begin{smallmatrix} C & \Delta \end{smallmatrix}\right] \left[\begin{smallmatrix} x_k \\ u_k \end{smallmatrix}\right], \Lambda_{k|K}^{-1} R\right) + c, u_{1:K} \geq 0, \qquad (49)
\end{aligned}
$$

where c is a term that is constant with respect to $(x_{1:K}, u_{1:K})$ but admits different values in different equations, $\Lambda_{k|K} \triangleq \mathbb{E}_{q_\Lambda}[\Lambda_k]$ is derived in Appendix A, Subsection B, and $u_{1:K} \geq 0$ means that all the components of all u_k are required to be nonnegative for each $k = 1 \cdots K$. Up to the truncation of the u components, $q_{xu}(x_{1:K}, u_{1:K})$ has thus the same form as the joint smoothing posterior of a linear state-space model with the state transition matrix $\widetilde{A} \triangleq \left[\begin{smallmatrix} A & \mathrm{O} \\ \mathrm{O} & \mathrm{O} \end{smallmatrix}\right]$, process noise covariance matrix $\widetilde{Q}_k \triangleq \left[\begin{smallmatrix} Q & \mathrm{O} \\ \mathrm{O} & \Lambda_{k+1|K}^{-1} \end{smallmatrix}\right]$, measurement model matrix $\widetilde{C} \triangleq \left[\begin{smallmatrix} C & \Delta \end{smallmatrix}\right]$, and measurement noise covariance matrix $\widetilde{R} \triangleq \Lambda_{k|K}^{-1} R$. We denote the PDFs related to this state-space model with $\widetilde{p}$.

It would be possible to compute the truncated multivariate normal posterior of the joint smoothing distribution $\widetilde{p}(\left[\begin{smallmatrix} x_{1:K} \\ u_{1:K} \end{smallmatrix}\right] | y_{1:K})$, and account for the truncation of $u_{1:K}$ to the positive orthant using the sequential truncation. However, this would be impractical with large K due to the large dimensionality $K \times (n_x + n_y)$. A feasible solution is to approximate each filtering distribution in the Rauch–Tung–Striebel smoother's (RTSS [27]) forward filtering step with a multivariate normal distribution by

$$
\begin{aligned}
\widetilde{p}(x_k, u_k | y_{1:k}) &= \tfrac{1}{C} \mathcal{N}\left(\left[\begin{smallmatrix} x_k \\ u_k \end{smallmatrix}\right]; z'_{k|k}, Z'_{k|k}\right) \cdot [\![u_k \geq 0]\!] \qquad (50)\\
&\approx \mathcal{N}\left(\left[\begin{smallmatrix} x_k \\ u_k \end{smallmatrix}\right]; z_{k|k}, Z_{k|k}\right) \qquad (51)
\end{aligned}
$$

for each $k = 1 \cdots K$, where $[\![u_k \geq 0]\!]$ is the Iverson bracket notation, C is the normalization factor, and $z_{k|k} \triangleq \mathbb{E}_{\widetilde{p}}[\left[\begin{smallmatrix} x_k \\ u_k \end{smallmatrix}\right] | y_{1:k}]$ and $Z_{k|k} \triangleq \mathrm{Var}_{\widetilde{p}}[\left[\begin{smallmatrix} x_k \\ u_k \end{smallmatrix}\right] | y_{1:k}]$ are approximated using the sequential truncation. Given the multivariate normal approximations of the filtering posteriors $\widetilde{p}(x_k, u_k | y_{1:k})$, by Lemma 2 the backward recursion of the RTSS gives multivariate normal approximations of the smoothing posteriors $\widetilde{p}(x_k, u_k | y_{1:K})$. The quantities required in the derivations of Subsection B are the expectations of the smoother posteriors $x_{k|K} \triangleq \mathbb{E}_{q_{xu}}[x_k]$, $u_{k|K} \triangleq \mathbb{E}_{q_{xu}}[u_k]$, and the covariance matrices $Z_{k|K} \triangleq \mathrm{Var}_{q_{xu}}\left[\begin{smallmatrix} x_k \\ u_k \end{smallmatrix}\right]$ and $U_{k|K} \triangleq \mathrm{Var}_{q_{xu}}[u_k]$.

Lemma 2. *Let $\{z_k\}_{k=1}^K$ be a linear–Gaussian process, and $\{y_k\}_{k=1}^K$ a measurement process such that*

$$
\begin{aligned}
z_1 &\sim \mathcal{N}(z_{1|0}, Z_{1|0}) \qquad (52a)\\
z_k | z_{k-1} &\sim \mathcal{N}(Az_{k-1}, Q) \qquad (52b)\\
y_k | z_k &\sim p(y_k | z_k), \qquad (52c)
\end{aligned}
$$

where $p(y_k|z_k)$ is a known distribution, and the standard Markovianity assumptions hold. Then, if the filtering posterior $p(z_k|y_{1:k})$ is a multivariate normal distribution for each k, then for each $k < K$ holds $z_k|y_{1:K} \sim \mathcal{N}(z_{k|K}, Z_{k|K})$, where

$$
\begin{aligned}
z_{k|K} &= z_{k|k} + G_k(z_{k+1|K} - Az_{k|k}), \qquad (53)\\
Z_{k|K} &= Z_{k|k} + G_k(Z_{k+1|K} - AZ_{k|k}A^{\mathrm{T}} - Q)G_k^{\mathrm{T}}, \qquad (54)\\
G_k &= Z_{k|k}A^{\mathrm{T}}(AZ_{k|k}A^{\mathrm{T}} + Q)^{-1}, \qquad (55)
\end{aligned}
$$

and $z_{k|k}$ and $Z_{k|k}$ are the mean and covariance matrix of the filtering posterior $p(z_k|y_{1:k})$.

Proof: The details are omitted here because the proof is mostly similar to that of [46, Theorem 8.2]. ■

B. Derivations for q_Λ

Eq. (10b) gives

$$
\log q_\Lambda(\Lambda_{1:K}) = \sum_{k=1}^{K} \Big\{ \mathop{\mathbb{E}}_{q_{xu}} [\log p(y_k | x_k, u_k, \Lambda_k) + \log p(u_k | \Lambda_k)]
$$

$$+\log p(\Lambda_k)\big\}+c. \tag{56}$$

Thus, $q_\Lambda(\Lambda_{1:K})=\prod_{k=1}^{K}q_\Lambda(\Lambda_k)$.

In the model with independent univariate skew-t-distributed measurement noise components (3), the diagonal entries of Λ_k are separate random variables, as given in (6c). Therefore,

$$\begin{aligned}&\log q_\Lambda(\Lambda_k)\\ &=-\frac{1}{2}\mathop{\mathbb{E}}_{q_{xu}}\big[\mathrm{tr}\{(y_k-Cx_k-\Delta u_k)(y_k-Cx_k-\Delta u_k)^{\mathrm{T}}R^{-1}\Lambda_k\}\\ &\quad+\mathrm{tr}\{u_ku_k^{\mathrm{T}}\Lambda_k\}\big]+\sum_{i=1}^{n_y}\big(\tfrac{\nu_i}{2}\log[\Lambda_k]_{ii}-\tfrac{\nu_i}{2}[\Lambda_k]_{ii}\big)+c \qquad (57)\\ &=-\frac{1}{2}\mathrm{tr}\,\Big\{\Big[\Big((y_k-Cx_{k|K}-\Delta u_{k|K})(y_k-Cx_{k|K}-\Delta u_{k|K})^{\mathrm{T}}\\ &\quad+\left[\,C\;\Delta\,\right]Z_{k|K}\begin{bmatrix}C^{\mathrm{T}}\\ \Delta^{\mathrm{T}}\end{bmatrix}\Big)R^{-1}+(u_{k|K}u_{k|K}^{\mathrm{T}}+U_{k|K})\Big]\Lambda_k\Big\}\\ &\quad+\sum_{i=1}^{n_y}\big(\tfrac{\nu_i}{2}\log[\Lambda_k]_{ii}-\tfrac{\nu_i}{2}[\Lambda_k]_{ii}\big)+c \qquad (58)\\ &=\sum_{i=1}^{n_y}\big(\tfrac{\nu_i}{2}\log[\Lambda_k]_{ii}-\tfrac{\nu_i+\Psi_{ii}}{2}[\Lambda_k]_{ii}\big)+c, \qquad (59)\end{aligned}$$

where

$$\begin{aligned}\Psi=&(y_k-Cx_{k|K}-\Delta u_{k|K})(y_k-Cx_{k|K}-\Delta u_{k|K})^{\mathrm{T}}R^{-1}\\ &+\left[\,C\;\Delta\,\right]Z_{k|K}\begin{bmatrix}C^{\mathrm{T}}\\ \Delta^{\mathrm{T}}\end{bmatrix}R^{-1}+u_{k|K}u_{k|K}^{\mathrm{T}}+U_{k|K}. \qquad (60)\end{aligned}$$

Therefore,

$$q_\Lambda(\Lambda_k)=\prod_{i=1}^{n_y}\mathcal{G}\left([\Lambda_k]_{ii};\tfrac{\nu_i}{2}+1,\tfrac{\nu_i+\Psi_{ii}}{2}\right). \tag{61}$$

In the derivations of Subection A, $\Lambda_{k|K}\triangleq\mathbb{E}_{q_\Lambda}[\Lambda_k]$ is required. $\Lambda_{k|K}$ is a diagonal matrix with the diagonal elements

$$[\Lambda_{k|K}]_{ii}=\tfrac{\nu_i+2}{\nu_i+\Psi_{ii}}. \tag{62}$$

In the model (30) with multivariate skew-t-distributed measurement noise Λ_k is of the form $\lambda_k\cdot I_{n_y}$. There, λ_k is a scalar random variable, and there is just one degrees-of-freedom parameter ν, as given in (31). Therefore,

$$\begin{aligned}&\log q_\Lambda(\lambda_k)\\ &=-\frac{1}{2}\mathop{\mathbb{E}}_{q_{xu}}[\mathrm{tr}\{(y_k-Cx_k-\Delta u_k)(y_k-Cx_k-\Delta u_k)^{\mathrm{T}}R^{-1}\lambda_k\}]\\ &\quad-\frac{1}{2}\mathop{\mathbb{E}}_{q_{xu}}[\mathrm{tr}\{u_ku_k^{\mathrm{T}}\lambda_k\}]+\tfrac{\nu+2n_y-1}{2}\log\lambda_k-\frac{\nu}{2}\lambda_k+c \qquad (63)\\ &=\tfrac{\nu+2n_y-1}{2}\log\lambda_k-\tfrac{\nu+\mathrm{tr}\{\Psi\}}{2}\lambda_k, \qquad (64)\end{aligned}$$

where Ψ is given in (60). Thus,

$$q_\Lambda(\lambda_k)=\mathcal{G}\left(\lambda_k;\tfrac{\nu+2n_y}{2},\tfrac{\nu+\mathrm{tr}\{\Psi\}}{2}\right). \tag{65}$$

so the required expectation is

$$\Lambda_{k|K}=\tfrac{\nu+2n_y}{\nu+\mathrm{tr}\{\Psi\}}\cdot I_{n_y}. \tag{66}$$

Appendix B
Derivation for the Fisher information of MVST

Consider the multivariate skew-t measurement model $y|x\sim\mathrm{MVST}(Cx,R,\Delta,\nu)$, where $C\in\mathbb{R}^{n_y\times n_x}$, $R\in\mathbb{R}^{n_y\times n_y}$, $\Delta\in\mathbb{R}^{n_y\times n_y}$, and $\nu\in\mathbb{R}^+$. The logarithm of the PDF of $y|x$ is

$$\begin{aligned}\log p(y|x)=&\log(2^{n_y}/\det(\Omega)^{\frac{1}{2}})+\log\mathrm{t}(r;0,I_{n_y},\nu)\\ &+\log\mathrm{T}(\Delta^{\mathrm{T}}\Omega^{-\frac{\mathrm{T}}{2}}r\sqrt{\tfrac{\nu+n_y}{\nu+r^{\mathrm{T}}r}};0,L,\nu+n_y), \qquad (67)\end{aligned}$$

where $r=\Omega^{-\frac{1}{2}}(y-Cx)$ is a function of x and y, $\Omega=R+\Delta\Delta^{\mathrm{T}}$, $L=I_{n_y}-\Delta^{\mathrm{T}}\Omega^{-1}\Delta$, and $\mathrm{t}(\cdot;\mu,\Sigma,\nu)$ and $\mathrm{T}(\cdot;\mu,\Sigma,\nu)$ denote the PDF and CDF of the scaled non-central multivariate t-distribution with ν degrees of freedom. $A^{\frac{1}{2}}$ is a square-root matrix such that $A^{\frac{1}{2}}(A^{\frac{1}{2}})^{\mathrm{T}}=A$, $A^{-\frac{1}{2}}\triangleq(A^{\frac{1}{2}})^{-1}$, and $A^{-\frac{\mathrm{T}}{2}}\triangleq((A^{\frac{1}{2}})^{-1})^{\mathrm{T}}$.

The Hessian matrix of the term $\log\mathrm{t}(r;0,I_{n_y},\nu)$ is derived in [47], and it is

$$\begin{aligned}&\tfrac{\mathrm{d}^2}{\mathrm{d}x^2}\log\mathrm{t}(r;0,I_{n_y},\nu)\\ &=\tfrac{\nu+n_y}{\nu}C^{\mathrm{T}}\Omega^{-\frac{\mathrm{T}}{2}}\left(-\tfrac{1}{1+\frac{1}{\nu}r^{\mathrm{T}}r}I_{n_y}+\tfrac{2/\nu}{(1+\frac{1}{\nu}r^{\mathrm{T}}r)^2}rr^{\mathrm{T}}\right)\Omega^{-\frac{1}{2}}C \qquad (68)\\ &=\tfrac{\nu+n_y}{\nu+r^{\mathrm{T}}r}C^{\mathrm{T}}\Omega^{-\frac{\mathrm{T}}{2}}\left(-I_{n_y}+\tfrac{2}{\nu+r^{\mathrm{T}}r}rr^{\mathrm{T}}\right)\Omega^{-\frac{1}{2}}C \qquad (69)\end{aligned}$$

The second term in (67) can be differentiated twice using the chain rule $\frac{\mathrm{d}^2\log(f)}{\mathrm{d}x^2}=\frac{1}{f}\frac{\mathrm{d}^2f}{\mathrm{d}x^2}-\frac{1}{f^2}\left(\frac{\mathrm{d}f}{\mathrm{d}x}\right)^{\mathrm{T}}\frac{\mathrm{d}f}{\mathrm{d}x}$, which gives

$$\begin{aligned}&\tfrac{\mathrm{d}^2}{\mathrm{d}x^2}\log\mathrm{T}(\Delta^{\mathrm{T}}\Omega^{-\frac{\mathrm{T}}{2}}r\sqrt{\tfrac{\nu+n_y}{\nu+r^{\mathrm{T}}r}};0,L,\nu+n_y)\\ &=\big(\mathrm{T}(\Delta^{\mathrm{T}}\Omega^{-\frac{\mathrm{T}}{2}}r\sqrt{\tfrac{\nu+n_y}{\nu+r^{\mathrm{T}}r}};0,L,\nu+n_y)\big)^{-1}g(r)\\ &\quad-\big(\mathrm{T}(\Delta^{\mathrm{T}}\Omega^{-\frac{\mathrm{T}}{2}}r\sqrt{\tfrac{\nu+n_y}{\nu+r^{\mathrm{T}}r}};0,L,\nu+n_y)\big)^{-2}D_r^{\mathrm{T}}P_r^{\mathrm{T}}P_rD_r, \qquad (70)\end{aligned}$$

where the function g is antisymmetric because it is the second derivative of a function that is antisymmetric up to an additive constant,

$$P_r=\tfrac{\mathrm{d}}{\mathrm{d}u}\mathrm{T}(u;0,L,\nu+n_y)\Big|_{u=\Delta^{\mathrm{T}}\Omega^{-\frac{\mathrm{T}}{2}}r\sqrt{\frac{\nu+n_y}{\nu+r^{\mathrm{T}}r}}}, \tag{71}$$

and

$$\begin{aligned}D_r=&\tfrac{\mathrm{d}}{\mathrm{d}x}\Delta^{\mathrm{T}}\Omega^{-\frac{\mathrm{T}}{2}}r\sqrt{\tfrac{\nu+n_y}{\nu+r^{\mathrm{T}}r}} \qquad (72)\\ =&\sqrt{\tfrac{\nu+n_y}{\nu+r^{\mathrm{T}}r}}\Delta^{\mathrm{T}}\Omega^{-\frac{\mathrm{T}}{2}}(\tfrac{1}{\nu+r^{\mathrm{T}}r}rr^{\mathrm{T}}-I_{n_y})\Omega^{-\frac{1}{2}}C. \qquad (73)\end{aligned}$$

Because the function g is antisymmetric, $\int g(r)p(r)\,\mathrm{d}y=0$ for any symmetric function p for which the integral exists.

We now outline the proof of integrability of certain functions to show that the CRLB exists and fulfils the regularity conditions given in [43, Ch. 2.4]. The integral $\int g(r)\,\mathrm{t}(r;0,1,\nu)\,\mathrm{d}y$ exists because the terms of g are products of positive powers of rational expressions where the denominator is of a higher degree than the nominator and derivatives of $\mathrm{T}(u;0,1,\nu+n_y)$ evaluated at $\Delta^{\mathrm{T}}\Omega^{-\frac{\mathrm{T}}{2}}r\sqrt{\frac{\nu+n_y}{\nu+r^{\mathrm{T}}r}}$, which is a bounded continuous function of y. The integral

$$\begin{aligned}&\int\big(\mathrm{T}(\Delta^{\mathrm{T}}\Omega^{-\frac{\mathrm{T}}{2}}r\sqrt{\tfrac{\nu+n_y}{\nu+r^{\mathrm{T}}r}};0,L,\nu+n_y)\big)^{-1}D_r^{\mathrm{T}}P_r^{\mathrm{T}}P_rD_r\\ &\quad\times\tfrac{2}{\det(\Omega)^{\frac{1}{2}}}\mathrm{t}(r;0,I_{n_y},\nu)\,\mathrm{d}y\end{aligned}$$

also exists because $\big(\mathrm{T}(\Delta^{\mathrm{T}}\Omega^{-\frac{\mathrm{T}}{2}}r\sqrt{\frac{\nu+n_y}{\nu+r^{\mathrm{T}}r}};0,L,\nu+n_y)\big)^{-1}$ and P_r are bounded and continuous and D_r is a positive power of a rational expression where the denominator is of a higher degree than the nominator. Similar arguments show the integrability of the first and second derivative of the likelihood $p(y|x)$, which guarantees that the regularity conditions of the CRLB are satisfied.

Thus, the expectation of (70) is

$$\begin{aligned}
&\mathop{\mathbb{E}}_{p(y|x)}\left[\tfrac{\mathrm{d}^2}{\mathrm{d}x^2}\log\mathrm{T}(\Delta^{\mathrm{T}}\Omega^{-\frac{\mathrm{T}}{2}}r\sqrt{\tfrac{\nu+n_y}{\nu+r^{\mathrm{T}}r}};0,L,\nu+n_y)\right]\\
&=\int g(r)\,\tfrac{2}{\det(\Omega)^{\frac{1}{2}}}\mathrm{t}(r;0,I_{n_y},\nu)\,\mathrm{d}y-\int\tfrac{2}{\det(\Omega)^{\frac{1}{2}}}\mathrm{t}(r;0,I_{n_y},\nu)\\
&\quad\times\big(\mathrm{T}(\Delta^{\mathrm{T}}\Omega^{-\frac{\mathrm{T}}{2}}r\sqrt{\tfrac{\nu+n_y}{\nu+r^{\mathrm{T}}r}};0,L,\nu+n_y)\big)^{-1}D_r^{\mathrm{T}}P_r^{\mathrm{T}}P_rD_r\mathrm{d}y \qquad (74)\\
&=\int 2g(r)\mathrm{t}(r;0,I_{n_y},\nu)\,\mathrm{d}r-\int 2\,\mathrm{t}(r;0,I_{n_y},\nu)\\
&\quad\times\big(\mathrm{T}(\Delta^{\mathrm{T}}\Omega^{-\frac{\mathrm{T}}{2}}r\sqrt{\tfrac{\nu+n_y}{\nu+r^{\mathrm{T}}r}};0,L,\nu+n_y)\big)^{-1}D_r^{\mathrm{T}}P_r^{\mathrm{T}}P_rD_r\,\mathrm{d}r \qquad (75)\\
&=-\mathop{\mathbb{E}}_{p(r|x)}\left[\big(\mathrm{T}(\Theta^{\mathrm{T}}r\sqrt{\tfrac{\nu+n_y}{\nu+r^{\mathrm{T}}r}};0,L,\nu+n_y)\big)^{-2}D_r^{\mathrm{T}}P_r^{\mathrm{T}}P_rD_r\right], \qquad (76)
\end{aligned}$$

where $\Theta=\Omega^{-\frac{1}{2}}\Delta$, and $r|x\sim\mathrm{MVST}(0,I_{n_y}-\Theta\Theta^{\mathrm{T}},\Theta,\nu)$ because $z\sim\mathrm{MVST}(\mu,R,\Delta,\nu)$ implies $Az\sim\mathrm{MVST}(A\mu,ARA^{\mathrm{T}},A\Delta,\nu)$. This gives

$$\begin{aligned}
&\mathop{\mathbb{E}}_{p(y|x)}\left[\tfrac{\mathrm{d}^2}{\mathrm{d}x^2}\log\mathrm{T}(\Delta^{\mathrm{T}}\Omega^{-\frac{\mathrm{T}}{2}}r\sqrt{\tfrac{\nu+n_y}{\nu+r^{\mathrm{T}}r}};0,L,\nu+n_y)\right]\\
&\qquad=-\,C^{\mathrm{T}}\Omega^{-\frac{\mathrm{T}}{2}}\mathop{\mathbb{E}}_{p(r|x)}\left[\tfrac{\nu+n_y}{\nu+r^{\mathrm{T}}r}\widetilde{R}_r\widetilde{R}_r^{\mathrm{T}}\right]\Omega^{-\frac{1}{2}}C, \qquad (77)
\end{aligned}$$

where

$$\begin{aligned}
\widetilde{R}_r=&\big(\mathrm{T}(\Theta^{\mathrm{T}}r\sqrt{\tfrac{\nu+n_y}{\nu+r^{\mathrm{T}}r}};0,L,\nu+n_y)\big)^{-1}(I_{n_y}-\tfrac{1}{\nu+r^{\mathrm{T}}r}rr^{\mathrm{T}})\Theta\\
&\times\left(\tfrac{\mathrm{d}}{\mathrm{d}u}\mathrm{T}(u;0,L,\nu+n_y)\Big|_{u=\Theta^{\mathrm{T}}r\sqrt{\frac{\nu+n_y}{\nu+r^{\mathrm{T}}r}}}\right)^{\mathrm{T}}, \qquad (78)
\end{aligned}$$

where $L=I_{n_y}-\Theta^{\mathrm{T}}\Theta$. Thus, the Fisher information for the measurement model $y|x\sim\mathrm{MVST}(Cx,R,\Delta,\nu)$ is

$$\begin{aligned}
\mathcal{I}(x)=&\mathop{\mathbb{E}}_{p(y|x)}\left[-\tfrac{\mathrm{d}^2}{\mathrm{d}x^2}\log p(y|x)\right] \qquad (79)\\
=&C^{\mathrm{T}}\Omega^{-\frac{\mathrm{T}}{2}}\mathop{\mathbb{E}}_{p(r|x)}\left[\tfrac{\nu+n_y}{\nu+r^{\mathrm{T}}r}(I_{n_y}-\tfrac{2}{(\nu+r^{\mathrm{T}}r)^2}rr^{\mathrm{T}}+\widetilde{R}_r\widetilde{R}_r^{\mathrm{T}})\right]\Omega^{-\frac{1}{2}}C, \qquad (80)
\end{aligned}$$

where $r|x\sim\mathrm{MVST}(0,I_{n_y}-\Theta\Theta^{\mathrm{T}},\Theta,\nu)$, $\Theta=\Omega^{-\frac{1}{2}}\Delta$, $\Omega=R+\Delta\Delta^{\mathrm{T}}$, and $\widetilde{R}_r$ is defined in (78).