Tronarp, Filip; Karvonen, Toni; Särkkä, Simo

**Student's t-Filters for Noise Scale Estimation**

*Published in:*
IEEE Signal Processing Letters

*DOI:*
10.1109/LSP.2018.2889440

*Published: 01/02/2019*

*Document Version*
Peer-reviewed accepted author manuscript, also known as Final accepted manuscript or Post-print

*Please cite the original version:*
Student’s $t$-filters for noise scale estimation

Filip Tronarp, Toni Karvonen, and Simo Särkkä

This article has been accepted for publication in IEEE Signal Processing Letters. When citing this article, use the following reference:


Copyright. ©2018 IEEE. Personal use of this material is permitted. Permission from IEEE must be obtained for all other uses, in any current or future media, including reprinting/republishing this material for advertising or promotional purposes, creating new collective works, for resale or redistribution to servers or lists, or reuse of any copyrighted component of this work in other works.
Student’s $t$-Filters for Noise Scale Estimation

Filip Tronarp, Student Member, IEEE, Toni Karvonen, and Simo Särkkä, Senior Member, IEEE

Abstract—We analyse certain Student’s $t$-filters for linear Gaussian systems with misspecified noise covariances. It is shown that the under appropriate conditions the filter both estimates the state and re-scales the noise covariance matrices in a Kullback–Leibler optimal fashion. If the noise covariances are misspecified, then the re-scaling is asymptotically exact. We also compare the Student’s $t$-filter scale estimates to the maximum likelihood estimates. Simulations demonstrating the results on the Wiener velocity model are provided.

Index Terms—Kalman filtering, Student’s $t$-filtering, model misspecification, noise covariance estimation

I. INTRODUCTION

Recently, Student’s $t$-filters, that assume the latent state and system noises have a joint Student’s $t$-distribution, for discrete-time systems have become popular in the signal processing community [1]–[10]. These filters were originally developed in the early 1990s [11] (see also [12], [13]) and later modified and popularised by Roth et al. [1]. The difference between the filters in [11], called here Student’s $t$-filter (ST; Alg. 2), and in [1, Sec. 3.1] (ST2; Rmk. 3) is that in the former (Thm. 6); and (iii) show that in the special case of covariance (ii) prove that, under certain assumptions on asymptotics of Student’s $t$-filter scale estimates to the maximum likelihood estimates. Simulations demonstrating the results on the Wiener velocity model are provided.

We (i) establish a relationship (Props. 4 and 5) between Student’s $t$-filter and classical misspecified Kalman filter (MKF); (ii) prove that, under certain assumptions on asymptotics of the scale estimator, Student’s $t$-filter estimates scaling of the noise covariance matrices in a Kullback–Leibler optimal way (Thm. 6); and (iii) show that in the special case of covariance matrices being misspecified by a common scalar factor the re-scaling due to Student’s $t$-filter is asymptotically correct (Cor. 7; in this setting, the scaling provided by the filter is compared to the maximum likelihood estimate in Thm. 8).

II. FILTERING FOR LINEAR SYSTEMS

We begin by introducing the classical Kalman filter [24], [25, Ch. 4] and Student’s $t$-filter as it appears in [11], [12]. The difference between ST and ST2 is discussed in Rmk. 3.

A. The System Description

We assume that the latent state $x_n \in \mathbb{R}^{d_x}$ and partial and corrupted measurements $y_n \in \mathbb{R}^{d_y}$ of it are, for $n \in \mathbb{N}$, generated by the linear Gaussian time-varying system

$$x_0 \sim \mathcal{N}(\mu_0, P_0),$$
$$x_n | x_{n-1} \sim \mathcal{N}(A_n x_{n-1}, Q),$$
$$y_n | x_n \sim \mathcal{N}(C_n x_n, R),$$

where $A_n \in \mathbb{R}^{d_x \times d_x}$ and $C_n \in \mathbb{R}^{d_y \times d_x}$ are model matrices, $Q$ and $R$ are the noise covariances, and $\mathcal{N}(\mu, \Sigma)$ is the Gaussian distribution with mean $\mu$ and variance $\Sigma$.

B. The (Misspecified) Kalman Filter

The classical Kalman filter [24], [25, Ch. 4] computes the conditional mean $\mathbb{E}[x_n | y_1, \ldots, y_n] = \mathbb{E}[x_n | y_{1:n}]$ and variance $\mathbb{V}[x_n | y_{1:n}]$ using simple linear algebraic recursion equations. However, the true noise covariances $Q$ and $R$ and the initial error covariance $P_0$ are often not available, forcing one to employ some other positive-definite matrices $Q^0$, $R^0$, and $P_0^0$ in the filter. The resulting MKF computes the state estimates $X_{n|n}^{\text{MKF}}$ and associated error covariances $P_{n|n}^{\text{MKF}}$ as follows.

Algorithm 1 (The misspecified Kalman filter; MKF). The misspecified Kalman filter consists of the prediction

$$X_{n|n-1}^{\text{MKF}} = A_n X_{n-1|n-1}^{\text{MKF}},$$
$$P_{n|n-1}^{\text{MKF}} = A_n P_{n-1|n-1}^{\text{MKF}} A_n^T + Q^0,$$

where $X_{0|0}^{\text{MKF}} = \mu_0$ and $P_{0|0}^{\text{MKF}} = P_0^0$, and the update

$$Y_{0|n}^{\text{MKF}} = C_n X_{n|n-1}^{\text{MKF}},$$
$$S_n^{\text{MKF}} = C_n P_{n|n-1}^{\text{MKF}} C_n^T + R^0,$$
$$K_n^{\text{MKF}} = P_{n|n-1}^{\text{MKF}} C_n^T S_n^{\text{MKF}}^{-1},$$
$$X_{n|n}^{\text{MKF}} = X_{n|n-1}^{\text{MKF}} + K_n^{\text{MKF}} (Y_n - Y_{n|n-1}^{\text{MKF}}),$$
$$P_{n|n}^{\text{MKF}} = P_{n|n-1}^{\text{MKF}} - K_n^{\text{MKF}} S_n^{\text{MKF}} K_n^{\text{MKF}}^T.$$

Due to the use of incorrect noise covariance matrices, $X_{n|n}^{\text{MKF}} \neq \mathbb{E}[x_n | y_{1:n}]$ and $P_{n|n}^{\text{MKF}} \neq \mathbb{V}[x_n | y_{1:n}]$ unless $Q^0 = Q$, $R^0 = R$, and $P_0^0 = P_0$. If there is no covariance misspecification, the above filter is the optimal Kalman filter and we will use superscript $\text{KF}$ instead of $\text{MKF}$. That is, $X_{n|n}^{\text{KF}} = \mathbb{E}[x_n | y_{1:n}]$ and $P_{n|n}^{\text{KF}} = \mathbb{V}[x_n | y_{1:n}]$.

Before presenting Student’s $t$-filter, we note that the effect of (partial) model misspecification as above to behaviour and stability of the Kalman filter has been studied for discrete-time systems in [26]–[30] and for continuous-time systems in [31].
C. Student’s t-Filter

Let \( N(x; \mu, \Sigma) \) stand for the Gaussian density with mean \( \mu \) and variance \( \Sigma \) and \( \Gamma^{-1}(s; \nu/2, \nu/2) \) for the reciprocal gamma density with shape and scale parameters \( \nu/2 \). Girón and Rojano [11] consider the Bayesian model

\[
p(x_0; y_{1:N}, s) = \frac{1}{\Gamma(\nu/2)} \left( \frac{\nu}{\nu + s} \right)^{\nu/2} (1 + s^{1/2} \nu)^{-\nu/2} \int_{\text{det} A_n > 0} \prod_{n=1}^{N} N(x_n; A_n x_{n-1}, s Q^0) \frac{1}{\sqrt{|A_n|}} \frac{1}{\sqrt{s}} \text{det} A_n^{\nu/2} dA_n
\]

for the full latent state and measurement sequences and noise models. Let \( S_n = \gamma_n^{-1} s_n - 1 \) and \( s_0 := s \). This results in the filter in Alg. 2.

**Algorithm 2** (Student’s t-filter; ST). Student’s t-filter consists of the prediction

\[
X_{st}^{n|n-1} = A_n X_{st}^{n-1|n-1}, \quad P_{st}^{n|n-1} = A_n P_{st}^{n-1|n-1} A_n^T + Q_{st}^{n-1},
\]

where \( X_{st}^{0|0} = \mu_0, P_{st}^{0|0} = P_0^0, \) and \( Q_0 = Q^0 \), and the update

\[
Y_{st}^{n|n-1} = C_n X_{st}^{n|n-1}, \quad S_n = C_n P_{st}^{n|n-1} C_n^T + R_{st}^{n-1},
\]

\[
\gamma_n = \frac{\nu_0 + (n - 1)d_y + \|Y_n - Y_{st}^{n|n-1}\|^2_{S_n^{-1}}}{\nu_0 + nd_y},
\]

\[
K_n = P_{st}^{n|n-1} C_n^T S_n^{-1}, \quad X_{st}^{n|n} = X_{st}^{n|n-1} + K_n (Y_n - Y_{st}^{n|n-1}),
\]

\[
P_{st}^{n|n} = \gamma_n (P_{st}^{n|n-1} - K_n S_n[K_n]^T), \quad Q_n = \gamma_n Q_{st}^{n-1}, \quad R_n = \gamma_n R_{st}^{n-1},
\]

**Remark 3.** The difference between ST of Alg. 2 and ST2 proposed in [1, Sec. 3.1] under the name exact filter is that

\[
Q_n = Q_{st}^{n-1} = \cdots = Q^0 \quad \text{and} \quad R_n = R_{st}^{n-1} = \cdots = R^0.
\]

This is because ST2 is based on a sequence of probability models \( p_{n+1}(y_{n+1}, x_{n+1}, x_n, s) \) of the form (2) such that \( p_{n+1}(x_n | s) \) matches \( p_n(x_n | y_n, s) \) of the previous time-step. That is, future noise realisations are not included in these models.

III. PROPERTIES OF STUDENT’S T-FILTER

In this section we prove a number of properties of Student’s t-filter of Alg. 2.

A. Basic Properties of Student’s t-FILTER

We begin by establishing some important connections between Student’s t-filter, misspecified Kalman filter, and the optimal Kalman filter. For the analysis it is useful to define

\[
\xi_n := \prod_{i=1}^{n} \gamma_i, \quad \xi_0 := 1.
\]

**Proposition 4.** The misspecified Kalman filter and Student’s t-filter of Algs. 1 and 2 admit the following relations:

\[
X_{st}^{n|n} = X_{st}^{MKF}, \quad P_{st}^{n|n} = \xi_n P_{st}^{MKF}, \quad S_n = \xi_n S_n^{MKF}, \quad Q_n = \xi_n Q^0, \quad R_n = \xi_n R^0.
\]

**Proof.** From the definition of \( \xi_n \) we immediately observe that

\[
S_{st}^{n|n} = \xi_n S_n^{MKF} \quad \text{and} \quad P_{st}^{n|n} = \xi_n P_{st}^{MKF}.
\]

Because of cancellation of \( K_n = K_{st}^{MKF} \), it follows that

\[
P_{st}^{n|n} = \xi_n P_{st}^{MKF}.
\]

This establishes that \( S_{st}^{n|n} = \xi_n S_n^{MKF} \) and, due to cancellation of \( \xi_n \), \( K_n = K_{st}^{MKF} \).

Hence Student’s t-filter can be interpreted as computing estimates \( \xi_n \) of the common scaling of the noise covariance matrices while producing the same state estimates as the MKF. More can be said when \( Q^0, R^0, \) and \( P_0^0 \) are merely misscaled.

**Proposition 5.** Suppose that the model (1) takes the form

\[
X_0 \sim N(\mu_0, P_0^0), \quad X_n \mid X_{n-1} \sim N(A_n X_{n-1}, \lambda Q^0), \quad Y_n \mid X_n \sim N(C_n X_n, \lambda R^0)
\]

for some \( \lambda > 0 \). Then, in addition to relations in Prop. 4, we have

\[
E[X_n \mid Y_{1:n}] = X_{st}^{n|n}, \quad \text{X_{st}^{n|n} = X_{st}^{MKF}}.
\]

**Proof.** In a manner similar to the proof of Prop. 4, it can be proved that \( P_{st}^{MKF} = \lambda P_{st}^{MKF} \) and \( S_n^{MKF} = \lambda S_n^{MKF} \). From these equations it follows that \( X_{st}^{n|n} = X_{st}^{MKF} \).

B. Asymptotic Properties of Student’s t-FILTER

This section adduces some asymptotic properties of Student’s t-filter. We prove that \( \xi_n \) tends to a value optimal in the sense of Kullback–Leibler divergence (Thm. 6) and that, under the model (4), \( \xi_n \) converges to \( \lambda \) (Cor. 7). The results are based on the recursion

\[
\xi_n = \frac{\nu_0 + (n - 1)d_y + \|Y_n - Y_{st}^{n|n-1}\|^2_{S_n^{-1}}}{\nu_0 + nd_y} \xi_{n-1}
\]

\[
= \frac{\nu_0 + (n - 1)d_y}{\nu_0 + nd_y} \xi_{n-1} + \frac{\|Y_n - Y_{st}^{n|n-1}\|^2_{S_n^{-1}}}{\nu_0 + nd_y}.
\]
from which it follows by induction and Prop. 4 that
\[ \xi_n = \frac{\nu_0 + \sum_{i=1}^{n} \| Y_i - Y_{\text{MKF}_{i|i-1}} \|^2_{\mathbb{S}_{i|1}^{\text{MKF}}}}{\nu_0 + nd_y}. \]  

(5)

Under certain assumptions, \( \xi_n \) converges to a value that is Kullback–Leibler optimal in the class of probability models that (possibly incorrectly) assume the noise covariances are merely misscaled; this is the class defined in (6). Note that it is not necessary to assume that (4) is the true model.

**Theorem 6.** Let the sequence \( Y_{1:n} \) be governed by the model (1) and denote its joint density by \( p^n(Y_{1:n}) \). Define the density
\[ q^n_\eta(y_{1:n}) = \frac{1}{n!} \prod_{i=1}^{n} N(y_i; y_{\text{MKF}_{i|i-1}}, \eta S_{i|1}^{\text{MKF}}) \]  

(6)

for \( \eta > 0 \) and the Kullback–Leibler divergence minimiser
\[ \eta_n = \arg \min_{\eta > 0} D_{KL}(p^n \parallel q^n_\eta). \]
Assume that \( \lim_{n \to \infty} \mathbb{E}[\xi_n^2/n^2] = \lim_{n \to \infty} \mathbb{V}[\xi_n] = 0 \). Then
\[ \lim_{n \to \infty} \mathbb{E}[(\xi_n - \eta_n)^2] = 0, \]
where the expectation is taken with respect to \( p^n \).

**Proof.** The Kullback–Leibler divergence is given by
\[ D_{KL}(p^n \parallel q^n_\eta) = \mathbb{E}[\log p^n(Y_{1:n})] + \frac{nd_y}{2} \log(2\pi\eta) + \frac{1}{2} \sum_{i=1}^{n} \log |S_{i|1}^{\text{MKF}}| \]
\[ + \frac{1}{2\eta^2} \sum_{i=1}^{n} \mathbb{E}\left[ \| Y_i - Y_{\text{MKF}_{i|i-1}} \|^2_{\mathbb{S}_{i|1}^{\text{MKF}}} \right]. \]

Differentiation gives
\[ \frac{dD_{KL}(p^n \parallel q^n_\eta)}{d\eta} = \frac{nd_y}{2\eta} - \frac{1}{2\eta^2} \sum_{i=1}^{n} \mathbb{E}\left[ \| Y_i - Y_{\text{MKF}_{i|i-1}} \|^2_{\mathbb{S}_{i|1}^{\text{MKF}}} \right]. \]

Therefore,
\[ \eta_n = \frac{1}{nd_y} \sum_{i=1}^{n} \mathbb{E}\left[ \| Y_i - Y_{\text{MKF}_{i|i-1}} \|^2_{\mathbb{S}_{i|1}^{\text{MKF}}} \right], \]

and from (5) it follows that
\[ \eta_n = \frac{\nu_0 + nd_y}{\nu_0 + nd_y} \mathbb{E}[[\xi_n] - \frac{\nu_0}{nd_y}] = \left( 1 + \frac{\nu_0}{nd_y} \right) \mathbb{E}[[\xi_n] - \frac{\nu_0}{nd_y}], \]

and
\[ \mathbb{E}[(\xi_n - \eta_n)^2] = \mathbb{V}[[\xi_n] + \left( \frac{\nu_0}{nd_y} \right)^2 (\mathbb{E}[[\xi_n]] + 1)^2, \]

which vanishes as \( n \to \infty \) by assumption.

If the true model is (4), it is a straightforward consequence of Thm. 6 that \( \xi_n \) asymptotically attains the correct scaling since, in this case, the true model belongs to the class (6) of approximating models.

**Corollary 7.** Under the model (4), \( \eta_n = \lambda \) and
\[ \lim_{n \to \infty} \mathbb{E}[(\xi_n - \lambda)^2] = 0. \]

**Proof.** By assumptions of model (4), \( p^n = q^n_\lambda \). Consequently, \( D_{KL}(p^n \parallel q^n_\lambda) = 0 \) and thus \( \eta_n = \lambda \) is the global KL minimiser. Furthermore, from Prop. 4 it follows that
\[ \| Y_i - Y_{\text{MKF}_{i|i-1}} \|^2_{\mathbb{S}_{i|1}^{\text{MKF}}} = \lambda \| Y_i - Y_{\text{MKF}_{i|i-1}} \|^2_{\mathbb{S}_{i|1}^{\text{MKF}}}, \]
and \( \| Y_i - Y_{\text{MKF}_{i|i-1}} \|^2_{\mathbb{S}_{i|1}^{\text{MKF}}} \) are i.i.d. with a \( \chi^2(d_y) \) distribution (see, e.g., [32]). From this it follows that
\[ \xi_n = \frac{\nu_0}{\nu_0 + nd_y} + \frac{\lambda}{\nu_0 + nd_y} \sum_{i=1}^{n} \mathbb{V}[Y_i - Y_{\text{MKF}_{i|i-1}}], \]

and
\[ \mathbb{E}[[\xi_n]] = \frac{\nu_0 + nd_y}{\nu_0 + nd_y}, \quad \mathbb{V}[[\xi_n]] = \lambda^2 \mathbb{E}[[\xi_n]] = \left( \frac{\lambda}{\nu_0 + nd_y} \right)^2 2nd_y. \]

(7)

The assumptions of Thm. 6 are thus satisfied and hence
\[ \lim_{n \to \infty} \mathbb{E}[(\xi_n - \eta_n)^2] = \lim_{n \to \infty} \mathbb{E}[(\xi_n - \lambda)^2] = 0. \]

(8)

\[]

Fig. 1 depicts variance and squared bias of \( \xi_n \) under the model (4).

**C. Comparison to Maximal Likelihood**

Here, \( \xi_n \) as an estimator of \( \lambda \) in (4) is compared to the maximum likelihood (ML) estimate. The derivative of the log-likelihood
\[ \ell(\lambda) = \sum_{i=1}^{n} \log p(Y_i \mid Y_{1:i-1}), \]
is given by
\[ \frac{d\ell}{d\lambda} = \frac{1}{2\lambda^2} \sum_{i=1}^{n} \| Y_i - Y_{\text{MKF}_{i|i-1}} \|^2_{\mathbb{S}_{i|1}^{\text{MKF}}} - \frac{nd_y}{2\lambda}, \]

(9)

the ML estimate \( \lambda_{\text{ML}} \) is given by
\[ \lambda_{\text{ML}} = \frac{\lambda}{nd_y} \sum_{i=1}^{n} \| Y_i - Y_{\text{MKF}_{i|i-1}} \|^2_{\mathbb{S}_{i|1}^{\text{MKF}}} . \]

(10)
By Prop. 4 \( \lambda_{n}^{\text{ML}} \) is thus a scaled sum of i.i.d. \( \chi^2(d_y) \) random variables. Therefore \( \mathbb{E}[\lambda_{n}^{\text{ML}}] = \lambda \) and \( \mathbb{V}[\lambda_{n}^{\text{ML}}] = 2\lambda/(nd_y) \). The relationship between \( \xi_{n} \) and \( \lambda_{n}^{\text{ML}} \) is elucidated in Thm. 8.

**Theorem 8.** Under the model (4), \( \xi_{n} \to \lambda_{n}^{\text{ML}} \) as \( \nu_{0} \to 0 \). Furthermore, \( \Gamma^{-1}(s; \nu_{0}/2, \nu_{0}/2) \) tends to Jeffrey’s prior [33].

**Proof.** From Prop. 4 and Eq. (5) it follows that

\[
\lim_{\nu_{0} \to 0} \xi_{n} = \frac{1}{nd_y} \sum_{i=1}^{n} \| Y_{i} - Y_{i|i-1}^{\text{ST}} \|_{[S_{n}^{\text{MKF}}]^{-1}}^{2}
= \frac{\lambda}{nd_y} \| Y_{i} - Y_{i|i-1}^{\text{MKF}} \|_{[S_{n}^{\text{MKF}}]^{-1}}^{2}
= \lambda_{n}^{\text{ML}}.
\]

Furthermore, the log-likelihood derivative (9) is

\[
\frac{d\ell}{d\lambda} = \frac{nd_y}{2\lambda} + \frac{1}{2\lambda} \sum_{i=1}^{n} \| Y_{i} - Y_{i|i-1}^{\text{MKF}} \|_{[S_{n}^{\text{MKF}}]^{-1}}^{2},
\]

and hence the Fisher information is

\[
\mathcal{I}(\lambda) = \mathbb{V}\left[\frac{d\ell}{d\lambda}\right] = \left(\frac{1}{2\lambda}\right)^{2} \sum_{i=1}^{n} 2d_{y},
\]

Jeffrey’s prior is then \( p_{2}(\lambda) \propto \sqrt{\mathcal{I}(\lambda)} \propto \lambda^{-1} \) which is proportional to \( \Gamma^{-1}(\lambda; \nu_{0}/2, \nu_{0}/2) \) as \( \nu_{0} \to 0 \).

**IV. A NUMERICAL EXAMPLE**

For numerical examples of the Student’s t-filter in Alg. 2 we use the Wiener velocity model with the model matrices

\[
A_{n} = \begin{bmatrix}
I_{d_{y} \times d_{y}} & \Delta t I_{d_{y} \times d_{y}} \\
0_{d_{y} \times d_{y}} & I_{d_{y} \times d_{y}}
\end{bmatrix}
\quad \text{and} \quad
C_{n} = \begin{bmatrix}
1 & 0 & 0 & 0 \\
0 & 1 & 0 & 0
\end{bmatrix}
\]

and true noise covariances

\[
Q = \lambda \begin{bmatrix}
\Delta t^{2} I_{d_{y} \times d_{y}} & \Delta t I_{d_{y} \times d_{y}} \\
\Delta t I_{d_{y} \times d_{y}} & I_{d_{y} \times d_{y}}
\end{bmatrix}
\quad \text{and} \quad
R = \lambda \mu_{d_{y} \times d_{y}}.
\]

where \( \lambda = 0.5 \) and \( \Delta t = 0.1 \). The system was initialised with \( P_{0} = \lambda I_{d_{y} \times d_{y}} \) and \( \mu_{0} = 0_{d_{y} \times 1} \). To contrast \( \xi_{n} \) from \( \lambda_{n}^{\text{ML}} \), similar when \( \nu_{0} \) is small by Thm. 8, we set \( \nu_{0} = 50 \).

**V. CONCLUSIONS**

We analysed some properties of the Student’s t-filter based on the joint model (2). Relations between the filter and misspecified Kalman filters were studied. Student’s t-filter was shown to converge to a certain Kullback–Leibler optimal Kalman filter under some assumption as well as recover the correct scale when the noise covariance matrices are misscaled. The scale estimator was compared to the maximum likelihood estimator. Simulations involving the Wiener velocity model demonstrated the theoretical results.

An interesting future research direction would be to see if any analysis is possible for ST2 (recall Rmk. 3). However, as the formulation of this filter does not correspond to a joint probability model for the entire state and measurement sequence, different approaches and results ought to be sought.

**Figure 2.** Student’s t-filter covariance scale estimates \( \xi_{n} \) and their theoretical mean \( \mathbb{E}[\xi_{n}] \) and maximum likelihood estimates \( \lambda_{n}^{\text{ML}} \) given in (10) for five trajectories of the Wiener velocity model.

**Figure 3.** Frobenius errors (compared to the optimal Kalman filter) of the error covariance and measurement covariance produced by the ST, ST2, and MKF for five trajectories of the Wiener velocity model under structurally misspecified filter covariances (11). Note that the MKF covariances are data-independent.