Robust and Trend-following Student's t-Kalman Smoothers ⋆

Aleksandr Aravkin [a] James V. Burke [b] Gianluigi Pillonetto [c]

[a] *Department of Earth and Ocean Sciences, University of British Columbia, Vancouver, Canada (e-mail: saravkin@eos.ubc.ca).*

[b] *Department of Mathematics, University of Washington, Seattle, USA (e-mail: burke@math.washington.edu)*

[c] *Department of Information Engineering, University of Padova, Padova, Italy (e-mail: giapi@dei.unipd.it)*

Abstract

Two nonlinear Kalman smoothers are proposed using the Student's t distribution. The first, which we call the *T-Robust smoother*, finds the maximum *a posteriori* (MAP) solution for Gaussian process noise and Student's t observation noise. It is extremely robust against outliers, outperforming the recently proposed ℓ_1-Laplace smoother in extreme situations with data containing 20% or more outliers. The second, which we call the *T-Trend smoother*, is a MAP solver for a model with Student's t-process noise and Gaussian observation noise. This smoother tracks sudden changes in the process model. A key ingredient of our approach is a novel technique to overcome the non-convexity of the Student's t loss function. By exploiting the structure of the underlying dynamics, the computational effort per iteration grows linearly with the length of the time series. Convergence analysis is provided for both smoothers. Numerical results for linear and nonlinear models illustrate the performance of the new smoothers for both robust and fast tracking applications, including an underwater tracking application with real data.

Key words: Robust estimation; non convex optimization; ℓ_1 loss functions; outliers; fast system dynamics; Student's t

1 Introduction

The Kalman filter is an efficient recursive algorithm for estimating the state of a dynamic system [14]. Traditional formulations are based on ℓ_2 penalties on model deviations, and are optimal under assumptions of linear dynamics and Gaussian noise. Kalman filters are used in a wide array of applications including navigation, medical technologies, and econometrics [8, 26, 22]. Many of these problems are nonlinear, and may require smoothing over past data in both online and offline applications to significantly improve estimation performance [11].
This paper focuses on two important areas in Kalman smoothing: robustness with respect to outliers in measurement data, and improved tracking of quickly changing system dynamics. Robust filters and smoothers have been a topic of significant interest since the 1970's, e.g. see [21]. A major body of work is devoted to designing filters robust to uncertainty in the *dynamical model*; see [18] and the numerous references within. For *known* dynamics, recent efforts have focused on building smoothers that are robust to outliers in the data [1, 2, 10], using convex loss functions such as ℓ_1, Huber or Vapnik, in place of the ℓ_2 penalty [13].
There have also been recent efforts to design smoothers able to better track fast system dynamics, e.g. jumps in the state values. A contribution can be found in [17] where the Laplace distribution, rather than the Gaussian, is used to model transition noise. This introduces an ℓ_1 penalty on the state evolution in time, resulting in an estimator interpretable as a dynamic version of the well known LASSO procedure [23].
For known dynamics, all of the smoothers mentioned above can be derived by modeling the process and the measurement noise using log-concave densities, taking the form

$$\mathbf{p}(\cdot) \propto \exp(-\rho(\cdot)), \quad \rho \text{ convex}. \tag{1.1}$$

Formulations exploiting (1.1) are nearly ubiquitous, in part because they correspond to convex optimization

⋆ This work was supported in part by the Natural Sciences and Engineering Research Council of Canada Discovery Grant (22R81254) and the Collaborative Research and Development Grant DNOISE II (375142-08). This research was carried out as part of the SINBAD II project with support from the following organizations: BG Group, BPG, BP, Chevron, Conoco Phillips, Petrobras, PGS, Total SA, and WesternGeco.

problems in the linear case. However, in order to model a regime with large outliers or sudden jumps in the state, we want to look beyond (1.1) and allow heavy-tailed densities, i.e. distributions whose tails are not exponentially bounded. All such distributions necessarily have non-convex loss functions [3, Theorem 2.1].

Several interesting candidates are possible, but in this contribution we focus on the Student's t-distribution for its convenient properties in the context of the applications we consider. The Student's t-distribution was successfully applied to a variety of robust inference applications in [15], and is closely related to re-descending influence functions [12].

In this work, we propose new nonlinear Student's t-based smoothers which we call T-Robust and T-Trend. The T-Robust smoother is derived from a dynamic system with *output noise* modeled by the Student's t-distribution. This is a further robustification of the estimator proposed in [1], which uses the Laplace density. The re-descending influence function of the Student's t guarantees that outliers in the measurements have less of an effect on the smoothed estimate than any convex loss function. In practice, the T-Robust smoother performs better than [1] for cases with a high proportion of outliers. The T-Trend smoother is similarly derived starting from a dynamic system with *transition noise* modeled by the Student's t-distribution. This allows T-Trend to better track sudden changes in the state.

In the context of Kalman filtering/smoothing, the idea of using Student's t-distributions to model the system noise for robust and tracking applications was first proposed in [9]. However, our work differs from that approach in some important aspects. First, our analysis includes nonlinear measurement and process models. Second, we provide a novel approach to overcome the non-convexity of the Student's t-loss function to solve both T-Robust and T-Trend smoothing problems. This approach differs significantly from the one proposed in [9]. [9] proposes using the random information matrix (i.e. full Hessian) when possible, or its expectation (Fisher information) when the Hessian is indefinite. Instead, we propose a modified Gauss-Newton method which builds information about the curvature of the Student's t-log likelihood into the Hessian approximation, and is guaranteed to be positive definite. As we show in Section 5, the new approach is provably convergent, and unlike the approach in [9] uses information about the relative sizes of the residuals in computing descent directions. These differences make it more stable than methods using random information, and more efficient than methods using Fisher information.

The major computational tradeoff in using non-convex penalties is that the loss function in the convex case is used directly, i.e. is not approximated, whereas in the nonconvex case, the loss function must be iteratively approximated with a local convex approximation. This introduces a fundamental difference in the convergence analysis.

The paper is organized as follows. In Section 2, we introduce the multivariate Student's t-distribution and describe the dynamic models underlying the T-Robust and T-Trend smoothers. In Section 3, we present the objective function for T-Robust, the quadratic approximation for this objective, and the convex quadratic program to solve this approximate subproblem. In Section 4, the T-Trend smoother is designed by modeling state transitions using Student's t, and the optimization approach is specified analogously to Section 3. In Section 5, we present the algorithm and a convergence theory for both smoothers. The T-Robust and T-Trend smoothers are tested using real and simulated data for linear and nonlinear models in Section 6. We end the paper with concluding remarks.

2 The T-Robust and T-Trend smoothing problems

For a vector $u \in \mathbb{R}^n$ and any positive definite matrix $M \in \mathbb{R}^{n\times n}$, let $\|u\|_M := \sqrt{u^{\mathrm{T}}Mu}$. We use the following generalization of the Student's t-distribution:

$$\mathbf{p}(v_k|\mu) = \frac{\Gamma(\frac{s+m}{2})}{\Gamma(\frac{s}{2})\det[\pi s R]^{1/2}}\left(1+\frac{\|v_k-\mu\|^2_{R^{-1}}}{s}\right)^{\frac{-(s+m)}{2}} \tag{2.1}$$

where μ is the mean, s is the degrees of freedom, m is the dimension of the vector v_k, and R is a positive definite matrix. A comparison of this distribution with the Gaussian and Laplacian distribution appears in Figure 2. Note that the Student's t-distribution has much heavier tails than the others, and that its influence function is re-descending, see [16] for a discussion of influence functions. This means that as we pull a measurement further and further away, its 'influence' decreases to 0, so it is eventually ignored by the model. Note also that the ℓ_1-Laplace is peaked at 0, while the Student's t-distribution is not, and so a Student's t-fit will not in general drive residuals to be exactly 0.

We use the following general model for the underlying dynamics: for $k = 1, \ldots, N$

$$\begin{aligned} x_k &= g_k(x_{k-1}) + w_k \\ z_k &= \quad h_k(z_k) + v_k \end{aligned} \tag{2.2}$$

with initial condition $g_1(x_0) = g_0 + w_1$, with g_0 a known constant, and where $g_k : \mathbb{R}^n \to \mathbb{R}^n$ are known smooth process functions, and $h_k : \mathbb{R}^n \to \mathbb{R}^{m(k)}$ are known smooth measurement functions.

For the T-Robust smoother, we assume that the vector $w_k \in \mathbb{R}^n$ is zero-mean Gaussian noise of known covariance $Q_k \in \mathbb{R}^{n\times n}$, and the vector $v_k \in \mathbb{R}^{m(k)}$ is zero-mean Student's t measurement noise (2.1) of known covariance $R_k \in \mathbb{R}^{m(k)\times m(k)}$ and degrees of freedom s.

For the T-Trend smoother, the roles are interchanged, i.e. w_k is Student's t noise while v_k is Gaussian. In both cases, we assume that the vectors $\{w_k\} \cup \{v_k\}$ are all

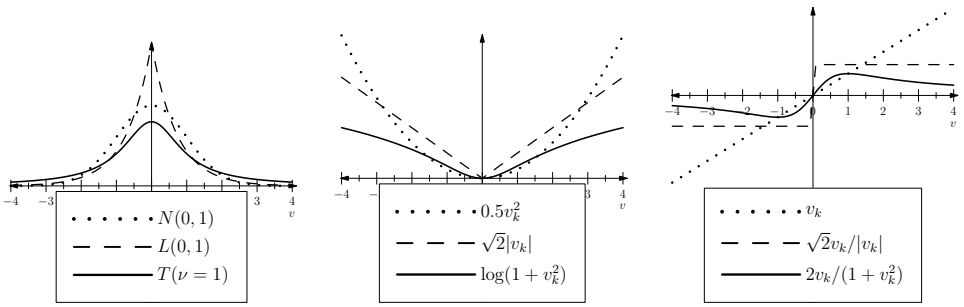


Fig. 1. Gaussian, Laplace, and Student's t Densities, Corresponding Negative Log Likelihoods, and Influence Functions (for scalar v_k).

mutually independent.
In the next sections, we design methods to find the MAP estimates of $\{x_k\}$ for both formulations.

3 T-Robust smoother

Given a sequence of column vectors $\{u_k\}$ and matrices $\{T_k\}$ we use the notation

$$\mathrm{vec}(\{u_k\}) = \begin{bmatrix} u_1 \\ u_2 \\ \vdots \\ u_N \end{bmatrix}, \ \mathrm{diag}(\{T_k\}) = \begin{bmatrix} T_1 & 0 & \cdots & 0 \\ 0 & T_2 & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0 \\ 0 & \cdots & 0 & T_N \end{bmatrix}.$$

We also make the following definitions:

$$\begin{aligned} R &= \mathrm{diag}(\{R_k\}) \\ Q &= \mathrm{diag}(\{Q_k\}) \\ x &= \mathrm{vec}(\{x_k\}) \end{aligned}, \quad \begin{aligned} w(x) &= \mathrm{vec}(\{x_k - g_k(x_{k-1})\}) \\ v(x) &= \mathrm{vec}(\{z_k - h_k(x_k)\}). \end{aligned}$$

Maximizing the likelihood for the model (2.2) is equivalent to minimizing the associated negative log likelihood

$$-\log \mathbf{p}(\{\nu_k\}, \{w_k\}) = -\log \mathbf{p}(\{\nu_k\}) - \log \mathbf{p}(\{w_k\})$$

Dropping the terms that do not depend on $\{x_k\}$, the objective corresponding to T-Robust is

$$\frac{1}{2}\sum_{k=1}^{N}(s_k + m_k)\log\left[1 + \frac{\|v_k\|^2_{R_k^{-1}}}{s_k}\right] + \|w_k\|^2_{Q_k^{-1}}, \qquad (3.1)$$

where s_k's are degrees of freedom parameters associated with measurement noise, and m_k are the dimensions of the kth observation.
A first-order accurate affine approximation to our model with respect to direction $d = \mathrm{vec}\{d_k\}$ near a fixed state sequence x is given by

$$\begin{aligned} \tilde{w}(x;d) &= \mathrm{vec}(\{x_k - g_k(x_{k-1}) - g_k^{(1)}(x_{k-1})d_k\}), \\ \tilde{v}(x;d) &= \mathrm{vec}(\{z_k - h_k(x_k) - h_k^{(1)}(x_k)d_k\}). \end{aligned}$$

Set $Q_{N+1} = I_n$ and $g_{N+1}(x_N) = 0$ (where I_n is the $n \times n$ identity matrix) so that the formulas are also valid for $k = N+1$.
We minimize the nonlinear nonconvex objective in (3.1) by iteratively solving quadratic programming (QP) subproblems of the form:

$$\min \ \tfrac{1}{2}d^{\mathrm{T}}Cd + a^{\mathrm{T}}d \quad \text{w.r.t } d \in \mathbb{R}^{nN}, \qquad (3.2)$$

where a is the gradient of objective (3.1) with respect to x and C has the form

$$C = \begin{bmatrix} C_1 & A_2^{\mathrm{T}} & 0 & \\ A_2 & C_2 & A_3^{\mathrm{T}} & 0 \\ 0 & \ddots & \ddots & \ddots \\ & 0 & A_N & C_N \end{bmatrix}, \qquad (3.3)$$

with $A_k \in \mathbb{R}^{n\times n}$ and $C_k \in \mathbb{R}^{n\times n}$ defined as follows:

$$\begin{aligned} A_k &= -Q_k^{-1}g_k^{(1)}, \\ C_k &= Q_k^{-1} + (g_{k+1}^{(1)})^{\mathrm{T}}Q_{k+1}^{-1}g_{k+1}^{(1)} + H_k, \\ H_k &= \frac{(h_k^{(1)})^{\mathrm{T}}R_k^{-1}h_k^{(1)}}{(s_k + \|v_k\|^2_{R_k^{-1}})/(s_k + m_k)}. \end{aligned} \qquad (3.4)$$

The Hessian approximation terms H_k in (3.4) are motivated in Section 5, and are crucial to both practical performance and theoretical convergence analysis. The solutions to the subproblem (3.2) have the form $d = -C^{-1}a$, and can be found in an efficient and numerically stable manner in $O(n^3N)$ steps, since C is tridiagonal and positive definite (see [4]).

4 T-Trend Smoother

The objective corresponding to T-Trend is

$$\frac{1}{2}\sum_{k=1}^{N}(r_k+n)\log\left[1+\frac{\|w_k\|^2_{Q_k^{-1}}}{r_k}\right]+\|v_k\|^2_{R_k^{-1}}, \quad (4.1)$$

where r_k are degrees of freedom parameters associated with process noise, and n is the dimension of each state x_k. A first-order accurate affine approximation to our model with respect to direction $d=\mathrm{vec}\{d_k\}$ near a fixed state sequence x is as follows:

$$\begin{aligned}\tilde{w}(x;d) &= \mathrm{vec}(\{x_k-g_k(x_{-1})-g_k^{(1)}(x_{k-1})d_k\}),\\ \tilde{v}(x;d) &= \mathrm{vec}(\{z_k-h_k(x_k)-h_k^{(1)}(x_k)d_k\}).\end{aligned}$$

As before, we set $Q_{N+1}=I_n$ and $g_{N+1}(x_N)=0$ (where I_n is the $n\times n$ identity matrix) so that the formula is also valid for $k=N+1$.
We minimize the nonlinear objective in (4.1) by iteratively solving quadratic programming (QP) subproblems of the form

$$\min\ \tfrac{1}{2}d^{\mathrm{T}}Cd+a^{\mathrm{T}}d \quad \text{w.r.t } d\in\mathbb{R}^{nN}, \quad (4.2)$$

where a is the gradient of objective (4) with respect to x and C again has form (3.3), but now with $A_k\in\mathbb{R}^{n\times n}$ and $C_k\in\mathbb{R}^{n\times n}$ defined as follows:

$$\begin{aligned}A_k &= -\frac{(r_k+n)Q_k^{-1}g_k^{(1)}}{r_k+\|w_k\|^2_{Q_k^{-1}}},\\ C_k &= Q_k^{-1}+(h_k^{(1)})^{\mathrm{T}}R_k^{-1}h_k^{(1)}+H_{k+1},\\ H_k &= \frac{(g_k^{(1)})^{\mathrm{T}}Q_k^{-1}g_k^{(1)}}{(r_k+\|w_k\|^2_{Q_k^{-1}})/(r_k+n)}.\end{aligned} \quad (4.3)$$

As for T-Robust, the Hessian approximation terms H_k in (4.3) are motivated in Section 5. One can show the resulting matrix C is positive definite by comparing to the Hessian of the standard Kalman smoothing objective [4, Eq. 12]. The solutions to the subproblem (4.2) again have the form $d=-C^{-1}a$, where C is tridiagonal and positive definite, so that they still can be found in an efficient and numerically stable manner in $O(n^3N)$ steps, see [4].

5 Algorithm and Global Convergence

When models g_k and h_k are linear, we can compare the algorithmic scheme proposed in the previous sections with the method in [9]. The latter uses the random information matrix (random Hessian) in place of the matrix C defined above, and recommends using the expected (Fisher) information when the full Hessian is indefinite. When the densities for w_k and v_k are Gaussian, this is equivalent to using Newton's method when possible, and using Gauss-Newton when the Hessian is indefinite. In general, using the expected information is known as the method of Fisher's scoring. In the Student's t-case, the scalar Fisher information matrix is computed in [15] to be

$$\frac{s+1}{s+3}\sigma^{-2}\,, \quad (5.1)$$

where σ^2 is the variance and s is the degrees of freedom parameter. [9] propose to use (5.1) as the Hessian approximation when the full Hessian is indefinite. Implementing this approach would effectively replace the terms $\|w_k\|_2^2$ or $\|v_k\|_2^2$, present in the denominators of H_k and A_k (see 3.4 and 4.3), with terms that depend only on s_k and r_k, the degrees of freedom. So while the random information (Hessian) matrix can become indefinite, the Fisher information is insensitive to outliers, and fails to down-weigh their contributions to the Hessian approximation.
To overcome these drawbacks, and find a middle ground between using the full Hessian and using a very rough approximation, we propose a Gauss-Newton method that is able to incorporate the relative size information of the residuals into the Hessian approximation. In the rest of this section we provide the details for the application of this method and a proof of convergence.
As in [1], the convergence theory is based upon the versatile convex-composite techniques developed in [6]. We begin by choosing the convex-composite structure for objectives (3.1) and (4.1). Note that one can write both in the convex-composite form $K=\rho\circ F$, with smooth F and convex ρ.
For T-Robust, we have

$$\rho_1\begin{pmatrix}c\\u\end{pmatrix}=c+\frac{1}{2}\|u\|^2_{Q^{-1}} \quad (5.2)$$

$$F_1(x)=\begin{pmatrix}f_1(x)\\w(x)\end{pmatrix} \quad (5.3)$$

$$f_1(x)=\frac{1}{2}\sum_{k=1}^{N}(b_k+l_k)\log\left[1+\frac{\|v_k(x)\|^2_{R_k^{-1}}}{b_k}\right], \quad (5.4)$$

where b_k and l_k correspond to the dimension and degrees of freedom parameters for the k-th measurement. For

T-Trend, we have

$$\rho_2 \begin{pmatrix} c \\ u \end{pmatrix} = c + \frac{1}{2}\|u\|^2_{R^{-1}} \tag{5.5}$$

$$F_2(x) = \begin{pmatrix} f_2(x) \\ v(x) \end{pmatrix} \tag{5.6}$$

$$f_2(x) = \frac{1}{2}\sum_{k=1}^{N}(n+a_k)\log\left[1 + \frac{\|w_k(x)\|^2_{Q_k^{-1}}}{n}\right], \tag{5.7}$$

where a_k corresponds to the degrees of freedom parameter for the k-th state. Note that the ranges of f_1 and f_2 are $\mathbf{R}_+$, and ρ_1 and ρ_2 are coercive on $\mathbb{R}_+ \times \mathbb{R}^{nN}$ and $\mathbb{R}_+ \times \mathbb{R}^M$, respectively, where $M = \sum m(k)$. The terms H_k in (3.4) and (4.3) are positive definite approximations to the Hessians of f_1 and f_2, respectively. To see this, consider the scalar function

$$\kappa(x) := \frac{1}{2}\log(1 + x^2/r)\ .$$

The second derivative of this function in x is given by

$$\frac{(r+x^2) - 2x^2}{(r+x^2)^2} = \frac{r - x^2}{(r+x^2)^2} \tag{5.8}$$

and is only positive on $(-\sqrt{r}, \sqrt{r})$. There are two reasonable globally positive approximations to take. The first,

$$\frac{r}{(r+x^2)^2},$$

simply ignores the subtracted term $-x^2$. In practice, we found this approximation to be too aggressive. Instead, we drop the $2x^2$ from the left of (5.8) to obtain the approximation

$$\frac{(r+x^2)}{(r+x^2)^2} = \frac{1}{(r+x^2)}\ . \tag{5.9}$$

Similarly, the terms H_k in (3.4) and (4.3) provide globally positive definite approximations to the Hessians of f_1 and f_2, using the strategy in (5.9). This strategy offers a significant computational advantage—the Hessian approximation that is built up down-weights the contributions of outliers, helping the algorithm proceed faster to the solution. As we shall see, these terms are also essential for the convergence theory in the T-Trend case.
Our approach exploits the objective structure by iteratively linearizing F about the iterates x^k and solving the *direction finding subproblem*

$$\min_{d\in\mathbb{R}^{nN}}\ \rho(F(x^k) + F^{(1)}(x^k)d) + \tfrac{1}{2}d^T H d. \tag{5.10}$$

For T-Robust and T-Trend, problem (5.10) can be solved with a single block-tridiagonal solve of the systems (3.2) and (4.2), respectively, yielding descent directions d for the objective $K(x)$.
We now develop a general convergence theory for convex-composite methods to establish the overall convergence to a stationary point of $K(x)$. This theory generalizes that of [1] to allow the inclusion of the correction term $\frac{1}{2}d^T H d$ in (5.10). This term was not necessary in [1], but is crucial here for both T-Robust and T-Trend formulations. Note that the theory does not rely at all on the technique used to solve the direction finding subproblem, and so the theory in this paper applies to the algorithm in [1] by taking $H_k = 0$.
Recall from [6] that the first-order necessary condition for optimality in the convex composite problem involving $K(x)$ is

$$0 \in \partial K(x) = \partial\rho\left(F(x)\right) F^{(1)}(x)$$

where $\partial K(x)$ is the generalized subdifferential of K at x [20] and $\partial\rho\left(F(x)\right)$ is the convex subdifferential of ρ at $F(x)$ [19]. Elementary convex analysis gives us the equivalence

$$0 \in \partial K(x) \quad \Leftrightarrow \quad K(x) = \inf_d \rho\Big(\ F(x) + F^{(1)}(x)d\ \Big)\ .$$

For both T-Robust and T-Trend, it is desirable to modify this objective by including curvature information, yielding the problem (5.10). We define the difference function

$$\Delta(x, H; d) = \rho\Big(\ F(x) + F^{(1)}(x)d\ \Big) + \frac{1}{2}d^{\mathrm{T}} H d - K(x)\ , \tag{5.11}$$

where $H = H(x)$ is positive semidefinite and varies continuously with x, and the minimum of $\Delta(x, H; d)$ with respect to direction d

$$\Delta^*(x, H) = \inf_d \Delta(x, H;\ d)\ . \tag{5.12}$$

Since $\frac{1}{2}d^{\mathrm{T}} H d$ is differentiable at the origin for any H, we have $\Delta^*(x, H) = 0$ if and only if $0 \in \partial K(x)$ [6].
Given $\eta \in (0, 1)$, we define a set of search directions at x by

$$D(x, H, \eta) = \{d \mid \Delta(x, H; d) \le \eta\Delta^*(x, H)\}\ . \tag{5.13}$$

Note that if there is a $d \in D(x, H, \eta)$ such that $\Delta(x, H;\ d) \ge -\eta\varepsilon$, then $\Delta^*(x, H) \ge -\varepsilon$. These ideas motivate the following algorithm [6].

Algorithm 5.1 *Gauss-Newton Algorithm.*

The inputs to this algorithm are

- $x^0 \in \mathbb{R}^{Nn}$*: initial estimate of state sequence*

- $H_0 \in \mathbb{S}^n_+$*: the initial curvature information*
- $\varepsilon \geq 0$*: overall termination criterion*
- $\eta \in (0,1)$*: termination criterion for subproblem*
- $\beta \in (0,1)$*: line search rejection criterion*
- $\gamma \in (0,1)$*: line search step size factor*

The steps are as follows:

(1) Set the iteration counter $\nu = 0$.
(2) (Gauss-Newton Step) Find d^ν *in the set* $D(x^\nu, H^\nu, \eta)$ *in 5.13. Set* $\Delta_\nu = \Delta(x^\nu, H^\nu; d^\nu)$ *in 5.11 and Terminate if* $\Delta_\nu \geq -\varepsilon$.
(3) (Line Search) Set

$$\begin{aligned} t_\nu = \quad & \max \gamma^i \\ & s.t.\ i \in \{0,1,2,\cdots\} \ \textit{and} \\ & s.t.\ \rho\left(F(x^\nu + \gamma^i d^\nu)\right) \leq \rho\left(F(x^\nu)\right) + \beta\gamma^i \Delta_\nu. \end{aligned}$$

(4) (Iterate) Set $x^{\nu+1} = x^\nu + t_\nu d^\nu$, *select* $H^{\nu+1} \in \mathbb{S}^n_+$ *and goto Step 2.*

We now present a general global convergence theorem that covers both T-robust and T-trend smoothers as special cases. This theorem generalizes [1, Theorem 5.1] to include positive definite correction terms in the Gauss-Newton framework.

Theorem 5.2 *If* $\{x^\nu\}$ *is a sequence generated by the Gauss-Newton Algorithm 5.1 with initial point* x^0 *and* $\varepsilon = 0$, *then one of the following must occur:*

(i) The algorithm terminates finitely at a point x^ν *with* $0 \in \partial K(x^\nu)$.
(ii) $\lim_{\nu \in I} \Delta_\nu = 0$ *for every subsequence* I *for which the set* $\{d^\nu, H^\nu \mid \nu \in I\}$ *is bounded.*
(iii) The sequence $\|d^\nu\|$ *diverges to* $+\infty$.

Moreover, if $\bar{x}$ *is any cluster point of a subsequence* $I \subset \mathbf{Z}_+$ *such that the subsequence* $\{d^\nu | \nu \in I\}$ *is bounded, then* $0 \in \partial K(\bar{x})$.

Proof: Assertions (i), (ii), and (iii) are a restatement of [6, Theorem 2.4] in our context, where the sets D_ν in [6, Theorem 2.4] are given by $D_\nu = D(x^\nu, \eta)$. The requirement that ρ be Lipschitz continuous on the set $\left\{(u) \,\middle|\, \rho(u) \leq K(x^0)\right\}$ is an immediate consequence of the fact that ρ is coercive on its domain, so this set is compact. This completes the proof of (i), (ii), and (iii). By compactness, the matrices H are uniformly continuous in x on this set. Suppose that $\bar{x}$ is a cluster point of a sequence $I \subset \mathbf{Z}_+$ for which $\{d^\nu, H^\nu\}$ is bounded. Since $\bar{x}$ is a cluster point of $\{x^\nu\}$, we can take a convergent subsequence along which $\{d^\nu\}$ are still bounded. By Bolzano-Weierstrass, we can then find a subsequence $J \subset I$ and $\bar{d}, \bar{H} \in \mathbb{R}^{Nn} \times \mathbb{S}^n_+$ such that $(x^\nu, d^\nu, H^\nu) \to_J (\bar{x},\ \bar{d},\ \bar{H})$. Fix any other point $\hat{d} \in \mathbb{R}^{Nn}$. By construction, we have

$$\begin{aligned} \Delta_\nu &= \rho\left(F(x^\nu) + F^{(1)}(x^\nu)d^\nu\right) + \tfrac{1}{2}\|d^\nu\|^2_{H^\nu} - \rho\left(F(x^\nu)\right) \\ &\leq \eta\Delta^*(x^\nu, H^\nu) \\ &\leq \eta\left(\rho\left(F(x^\nu) + F^{(1)}(x^\nu)\hat{d}\right) + \tfrac{1}{2}\|\hat{d}\|^2_{H^\nu} - \rho\left(F(x^\nu)\right)\right) . \end{aligned}$$

Taking the limit over J gives

$$\begin{aligned} 0 &= \rho\left(F(\bar{x}) + F^{(1)}(\bar{x})\bar{d}\right) + \tfrac{1}{2}\|\bar{d}\|^2_{\bar{H}} - \rho\left(F(\bar{x})\right) \\ &\leq \eta\left(\rho\left(F(\bar{x}) + F^{(1)}(\bar{x})\hat{d}\right) + \tfrac{1}{2}\|\hat{d}\|^2_{\bar{H}} - \rho\left(F(\bar{x})\right)\right) . \end{aligned}$$

But $\hat{d}$ was an arbitrary point in $\mathbb{R}^{Nn}$, so in particular we must have $\Delta^*(\bar{x}, \bar{H}) = 0$, which by [6, Theorem 3.6] implies that $0 \in \partial K$. ■

A stronger convergence result is possible under stronger assumptions on F and $F^{(1)}$, or on the sequence H^ν. Fix $x^0 \in \mathbb{R}^{Nn}$, and define

$$\Lambda = \left\{u \,\middle|\, \rho(u) \leq K(x^0)\right\} . \tag{5.14}$$

Corollary 5.3 *Suppose that* $F^{-1}(\Lambda) = \{x \mid F(x) \in \Lambda\}$ *is bounded, and take* H^ν *is as in step 2 of Algorithm 5.1, and assume further that* $\{H^\nu\}$ *are selected from a compact set in* $\mathbb{S}^n_+$. *Suppose either of the following assumptions hold:*

$$0 < \lambda_{\min} \leq \mathrm{eig}(H^\nu) \tag{5.15}$$

$$\mathrm{Null}\left(F^{(1)}(x)\right) = \{0\} \quad \forall\, x \in F^{-1}(\Lambda) , \tag{5.16}$$

If $\{x^\nu\}$ *is a sequence generated by Algorithm 5.1 with initial point* x^0 *and* $\varepsilon = 0$, *then* $\{x^\nu\}$ *and* $\{d^\nu\}$ *are bounded and either the algorithm terminates finitely at a point* x^ν *with* $0 \in \partial K(x^\nu)$, *or* $\Delta_\nu \to 0$ *as* $\nu \to \infty$, *and every cluster point* $\bar{x}$ *of the sequence* $\{x^\nu\}$ *satisfies* $0 \in \partial K(\bar{x})$.

Proof: First note that $F^{-1}(\Lambda)$ is closed since F is continuous, and therefore $F^{-1}(\Lambda)$ is compact, since by assumption it is bounded. Hence $\overline{\mathrm{co}}\left(F^{-1}(\Lambda)\right)$ is also compact. Therefore, $F^{(1)}$ is uniformly continuous on $\overline{\mathrm{co}}\left(F^{-1}(\Lambda)\right)$ which implies that the hypotheses of Theorem 5.2 are satisfied, and so one of (i)-(iii) must hold. If (i) holds we are done, so we will assume that the sequence $\{x^\nu\}$ is infinite. Since $\{x^\nu\} \subset F^{-1}(\Lambda)$, this sequence is bounded. We now show that the sequence $\{d^\nu\}$ of search directions is also bounded.
Suppose that (5.15) holds. For any direction d^ν, note that d^ν solves

$$\min_d\ \rho\left(F(x) + F^{(1)}(x)d\right) + \tfrac{1}{2}\|d\|^2_{H^\nu} .$$

Therefore we have

$$\rho\left(F(x) + F^{(1)}(x)d^\nu\right) + \frac{1}{2}\|d^\nu\|^2_{H^\nu} \leq \rho\left(F(x)\right) \tag{5.17}$$

since we can achieve $\rho(F(x))$ with $d = 0$. Since $\rho \geq 0$, we must have $\{\frac{1}{2}\|d^\nu\|^2_{H^\nu}\} \leq \rho(F(x))$, hence $\{\|d^\nu\|_{H^\nu}\}$ are bounded, and d^ν are bounded by (5.15).
Suppose instead that (5.16) holds. We claim that there exists $\kappa > 0$ such that

$$\kappa\|d\| \leq \|F^{(1)}(x)d\| \quad \forall\, d \in \mathbb{R}^{Nn} \text{ and } x \in F^{-1}(\Lambda)\ .$$

Indeed, if this were not the case, then there would exist sequences $\{y^i\} \subset F^{-1}(\Lambda)$ and $\{d^i\} \subset \mathbb{R}^{Nn} \setminus \{0\}$ such that

$$\|d^i\|/i > \|F^{(1)}(y^i)d^i\| \quad \forall\, i = 1, 2, \dots\ .$$

The set $F^{-1}(\Lambda)$ is compact, hence there exists a subsequence $J \subset \mathbf{Z}_+$, vector $\bar{x} \in F^{-1}(\Lambda)$, and vector $\bar{d} \in \mathbb{R}^{Nn}$ with $\|\bar{d}\| = 1$, such that $x^\nu \to_J \bar{x}$ and $d^i/\|d^i\| \to_J \bar{d} \neq 0$. It follows from the inequality above that

$$\frac{1}{i} \geq \left\| F^{(1)}(x^i)\frac{d^i}{\|d^i\|} \right\| .$$

Taking the limit with respect to the subsequence J, we obtain $0 \geq \|F^{(1)}(\bar{x})\bar{d}\|$. Thus $\bar{d}$ is in the kernel of $F^{(1)}(\bar{x})$ and $\bar{d} \neq 0$. This contradicts (5.16) and thereby proves the claim. For any direction d^ν, (5.17) holds and $F(x) + F^{(1)}(x)d^\nu \in \Lambda$ since $\frac{1}{2}\|d\|^2_H \geq 0$. Since Λ is compact, and $\{F(x^\nu),\ F(x^\nu) + F^{(1)}(x^\nu)d^\nu\} \subset \Lambda$ by construction, there is an $\alpha > 0$ such that $\|u\| \leq \alpha$ for all $u \in \Lambda$ and for $\nu = 1, 2, \dots,$

$$\begin{aligned}\kappa\|d^\nu\| &\leq \| F^{(1)}(x^\nu)d^\nu \| \\ &\leq \| F(x^\nu) + F^{(1)}(x^\nu)d^\nu \| + \|F(x^\nu)\| \ \leq\ 2\alpha\ .\end{aligned}$$

Hence the sequence $\{d^\nu\}$ of search directions is bounded. In both cases, Theorem 5.2 tells us that $\Delta_\nu \to 0$ as $\nu \to \infty$. The final statement of the corollary follows immediately from the final statement of Theorem 5.2. ■

Remark 5.4 *Both T-Robust and T-Trend can be solved by Algorithm 5.1, with the objective function $K(x)$ as above. In both cases, $F_i^{-1}(\Lambda)$ is compact because ρ_i is coercive on the range of f_i, since f_i are non-negative functions. We can then analyze conditions under which the remaining assumptions of Corollary 5.3 are satisfied.*

Convergence for T-Robust: In this case, (5.16) always holds. To show this, we compute compute $F_1^{(1)}$ for F_1 in (5.3):

$$F_1^{(1)}(x) = \begin{bmatrix} \nabla f_1(x)^T \\ G(x) \end{bmatrix} \tag{5.18}$$

where

$$G(x) := \begin{bmatrix} \mathrm{I} & 0 & & \\ -g_2^{(1)}(x_1) & \mathrm{I} & \ddots & \\ & \ddots & \ddots & 0 \\ & & -g_N^{(1)}(x_{N-1}) & \mathrm{I} \end{bmatrix} . \tag{5.19}$$

Since the matrix $G(x)$ is lower block bi-diagonal, with identities on the diagonal, we immediately have that F is nonsingular, and (5.16) holds. The matrices H^ν are given in (3.4), and are clearly bounded because they are continuous functions of x^ν, which remain in the compact set $F_1^{-1}(\Lambda)$. The conclusions of Corollary 5.3 therefore apply.

T-Trend: We compute $F_2^{(1)}$ for F_2 in (5.6).

$$F_2^{(1)}(x) = \begin{bmatrix} \nabla f_2(x)^T \\ h^{(1)}(x) \end{bmatrix} , \tag{5.20}$$

where

$$h^{(1)} := \begin{bmatrix} h_1^{(1)}(x) & & & \\ & h_2^{(1)}(x) & & \\ & & \ddots & \\ & & & h_n^{(1)}(x) \end{bmatrix} . \tag{5.21}$$

For T-Trend, (5.16) holds if all $h_k^{(1)}(x)$ are nonsingular for all $x \in F_2^{(-1)}(\Lambda)$. This condition may not hold in general, but note that (5.15) holds if all $g_k^{(1)}(x_k)$ are nonsingular on $F_2^{-1}(\Lambda)$. This implies that

$$(g_{k+1}^{(1)}(x))^{\mathrm{T}} Q_{k+1}^{-1} g_{k+1}^{(1)}(x)$$

are uniformly positive definite, i.e. there exists $\lambda_{min} > 0$ such that

$$\lambda_{min}\|d\|^2 \leq d^T (g_{k+1}^{(1)}(x))^{\mathrm{T}} Q_{k+1}^{-1} g_{k+1}^{(1)}(x) d$$

for all $d \in \mathbb{R}^n$, $k = 1, \dots, N$. Let w_m be the maximum residual, i.e. $\sup_{x \in F_2^{-1}(\Lambda)} \|w_k(x)\|^2_{Q_k}$, which if finite because $w_k(x)$ is continuous and $F_2^{-1}(\Lambda)$ is compact. Observe that using H^ν in (4.3),

$$u^T H^\nu u = \frac{u^T (g_{k+1}^{(1)})^{\mathrm{T}} Q_{k+1}^{-1} g_{k+1}^{(1)} u}{(r_k + \|w_k\|^2_{Q_k^{-1}})/(r_k + n)} \geq \|u\|^2 \frac{C\lambda_{min}}{(1 + w_m)} ,$$

for any $x \in F^{-1}(\Lambda)$, where C is some positive constant that depends on degrees of freedom parameters $\{r_k\}$ and

state dimension n. Finally, H^ν are clearly bounded because they are continuous functions of x^ν, which remain in the compact set $F_2^{-1}(\Lambda)$. The conclusions of Corollary 5.3 therefore apply.

6 Numerical Experiments

6.1 T-Robust Smoother: function reconstruction using splines

In this section we compare the new T-robust smoother with the ℓ_2-Kalman smoother [4] and with the ℓ_1-Laplace robust smoother [1], both implemented in [5]. The *ground truth* for this simulated example is

$$x(t) = \begin{bmatrix} -\cos(t) & -\sin(t) \end{bmatrix}^{\mathrm{T}} .$$

The time between measurements is a constant Δt. We model the two components of the state as the first and second integrals of white noise, so that

$$g_k(x_{k-1}) = \begin{bmatrix} 1 & 0 \\ \Delta t & 1 \end{bmatrix} x_{k-1} , \qquad Q_k = \begin{bmatrix} \Delta t & \Delta t^2/2 \\ \Delta t^2/2 & \Delta t^3/3 \end{bmatrix} .$$

This stochastic model for function reconstruction underlies the Bayesian interpretation of cubic smoothing splines, see [25] for details.
The measurement model for the conditional mean of measurement z_k given state x_k is defined by

$$h_k(x_k) = \begin{bmatrix} 0 & 1 \end{bmatrix} x_k = x_{2,k} , \qquad R_k = \sigma^2 ,$$

where $x_{2,k}$ denotes the second component of x_k, $\sigma^2 = 0.25$ for all experiments, and the degrees of freedom parameter k was set to 4 for the Student's t methods.
The measurements $\{z_k\}$ were generated as a sample from

$$z_k = x_2(t_k) + v_k, \quad t_k = 0.04\pi \times k$$

where $k = 1, 2, \ldots, 100$. The measurement noise v_k was generated according to the following schemes.

(1) **Nominal**: $v_k \sim \mathbf{N}(0, 0.25)$.

(2) **Gaussian contamination**

$$v_k \sim (1-p)\mathbf{N}(0, 0.25) + p\mathbf{N}(0, \phi),$$

for $p \in \{0.1, 0.2, 0.5\}$ and $\phi \in \{1, 4, 10, 100\}$.

(3) **Uniform contamination**

$$v_k \sim (1-p)\mathbf{N}(0, 0.25) + p\mathbf{U}(-10, 10),$$

for $p \in \{0.1, 0.2, 0.5\}$.

Each experiment was performed 1000 times. Table 1 presents the results for our simulated fitting showing the median Mean Squared Error (MSE) value and a quantile interval containing 95% of the results. The MSE is defined by

$$\frac{1}{N}\sum_{k=1}^{N}[x_1(t_k) - \hat{x}_{1,k}]^2 + [x_2(t_k) - \hat{x}_{2,k}]^2, \qquad (6.1)$$

where $\{\hat{x}_k\}$ is the corresponding estimating sequence.
From Table 1 one can see that T-Robust and the ℓ_1-smoother perform as well as the (optimal) ℓ_2-smoother at nominal conditions, and that both continue to perform at that same level for a variety of outlier generating scenarios. T-Robust always performs at least as well as the ℓ_1-smoother, and it gains an advantage when either the probability of contamination is high, or the contamination is uniform. This is likely due to the re-descending influence function of the Student's t-distribution — the smoother effectively throws out bad points rather than simply decreasing their impact to a certain threshold, as is the case for the ℓ_1-smoother. As an example, results coming from a single run for the case where 50% of measurements are contaminated with the uniform distribution on $[-10, 10]$ are displayed in Figure 2. Notice that T-Robust has an advantage over the ℓ_1-smoother.

6.2 T-Robust Smoother: Van Der Pol oscillator

In this section, we present results for the Van Der Pol oscillator (VDP), described in detail in [1]. The VDP oscillator is a coupled nonlinear ODE defined by

$$\dot{x}_1(t) = x_2(t)$$
$$\dot{x}_2(t) = \mu[1 - x_1(t)^2]x_2(t) - x_1(t)$$

The process model here is the Euler approximation for $X(t_k)$ given $X(t_{k-1})$:

$$g_k(x_{k-1}) = \begin{pmatrix} x_{1,k-1} + x_{2,k-1}\Delta t \\ x_{2,k-1} + \{\mu[1 - x_{1,k}^2]x_{2,k} - x_{1,k}\}\Delta t \end{pmatrix} .$$

For this simulation, the *ground truth* is obtained from a stochastic Euler approximation of the VDP. To be specific, with $\mu = 2$, $N = 164$ and $\Delta t = 16/N$, the ground truth state vector x_k at time $t_k = k\Delta t$ is given by $x_0 = (0, -0.5)^{\mathrm{T}}$ and for $k = 1, \ldots, N$, $x_k = g_k(x_{k-1}) + w_k$, where $\{w_k\}$ is a realization of independent Gaussian noise with variance 0.01.
In [1], the ℓ_1-Laplace smoother was shown to have superior performance to the ℓ_2-smoother, both implemented in [5]. We compared the performance of the nonlinear T-robust and nonlinear ℓ_1-Laplace smoothers, and found that T-robust gains an advantage in the extreme cases of 70% outliers. Figure 3 illustrates results coming from a single representative run. For 40% or fewer outliers,

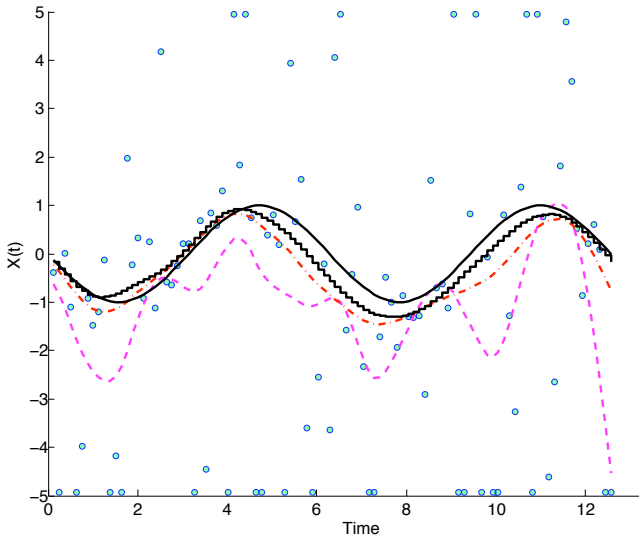


Fig. 2. Function reconstruction via spline: performance of ℓ_2 Kalman smoother (dash), ℓ_1-Laplace Robust smoother (dash-dot), and T-Robust (stair-case solid) on contaminated normal model with 50% outliers distributed uniformly on [-10,10]. True state x(t) is drawn as solid line. Measurements appear as 'o' symbols, and all measurements visible off of the true state are outliers in this case. Values outside [-5,5] are plotted on the axis limits.

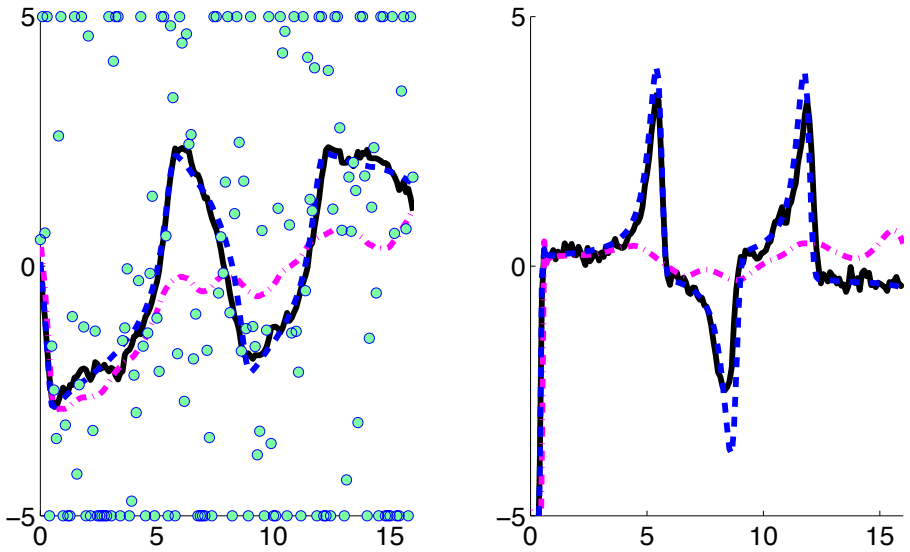


Fig. 3. Van Der Pol oscillator: smoother fits for X-component (left) and Y-component (right), with **70% outliers** $N(0, 100)$. Black solid line is truth, magenta dash-dot is the ℓ_1 smoother result, and blue dashed line is T-robust. Measurements on X-component are shown as dots, with outliers outside the range $[-5, 5]$ plotted on top and bottom axes.

Table 1
Function reconstruction via spline: median MSE over 1000 runs and intervals containing 95% of MSE results.

Outlier	p	ℓ_2 **MSE**	ℓ_1 **MSE**	**Student's t MSE**
Nominal	—	.04(.02, .1)	.04(.01, .1)	.04(.01, .09)
$\mathbf{N}(0,1)$	.1	.06(.02, .12)	.04(.02, .10)	.04(.02, .10)
$\mathbf{N}(0,4)$	.1	.09(.04, .29)	.05(.02, .12)	.04(.02, .11)
$\mathbf{N}(0,10)$	.1	.17(.05, .55)	.05(.02, .13)	.04(.02, .11)
$\mathbf{N}(0,100)$	.1	1.3(.30, 5.0)	.05(.02, .14)	.04(.02, .11)
$\mathbf{U}(-10,10)$	.1	.47(.12, 1.5)	.05(.02, .13)	.04(.02, .10)
$\mathbf{N}(0,10)$	.2	.32(.11, .95)	.06(.02, .19)	.05(.02, .16)
$\mathbf{N}(0,100)$	.2	2.9(.94, 8.5)	.07(.02, .22)	.05(.02, .14)
$\mathbf{U}(-10,10)$	.2	1.1(.36, 3.0)	.07(.03, .26)	.05(.02, .13)
$\mathbf{N}(0,10)$	.5	.74(.29, 1.9)	.13(.05, .49)	.10(.04, .45)
$\mathbf{N}(0,100)$	.5	7.7(2.9, 18)	.21(.06, 1.6)	.09(.03, .44)
$\mathbf{U}(-10,10)$	.5	2.6(1.0, 5.8)	.20(.06, 1.4)	.10(.03, .44)

it is hard to differentiate the performance of the two smoothers for this nonlinear example.

6.3 T-Robust Smoother: underwater Tracking Application

This application is described in detail in [1], so we just give a brief overview here. In [1] we used the application to test the ℓ_1-Laplace smoother. Here we use it for a qualitative comparison between the T-Robust smoother, the ℓ_1-Laplace smoother, and the ℓ_2 smoother with outlier removal.

In this experiment, a tracking target was hung on a steel cable approximately 200 meters below a ship. The pilot was attempting to keep the ship in place (hold station) at specific coordinates, but the ship was pitching and rolling due to wave action. The measurements for the smoother were sound travel times between the tracking target and four bottom mounted transponders at known locations, and pressure readings from a gauge that was placed on the target. Tracking data was independently verified using a GPS antenna mounted on a ship, and the GPS system provided sub-meter accuracy in position.

Pressure measurements in absolute bars were converted to depth in meters by the formula

$$\text{depth} = 9.9184(\text{pressure} - 1).$$

We use N to denote the total number of time points at which we have tracking data. For $k = 1, \ldots, N$, the state vector at time t_k is defined by $x_k = (e_k, n_k, d_k, \dot{e}_k, \dot{n}_k, \dot{d}_k)^{\mathrm{T}}$ where (e_k, n_k, d_k) is the (east, north, depth) location of the object (in meters from the origin), and $(\dot{e}_k, \dot{n}_k, \dot{d}_k)$ is the time derivative of this location.

The measurement vector at time t_k is denoted by z_k. The first four components of z_k are the range measurements to the corresponding bottom mounted transponders and the last component is the depth corresponding to the pressure measurement. For $j = 1, \ldots, 4$, the model for the mean of the corresponding range measurements was

$$h_{j,k}(x_k) = \|(e_k, n_k, d_k) - b_j\|_2 - \Delta r_j \ .$$

These measurements were assumed independent with standard deviation 3 meters. These depth measurements were assumed to have standard deviation of 0.05 meters. We use Δt_k to denote $t_{k+1} - t_k$. The model for the mean of x_{k+1} given x_k was

$$g_{k+1}(x_k) = \\ (e_k + \dot{e}_k \Delta t_k,\ n_k + \dot{n}_k \Delta t_k,\ d_k + \dot{d}_k \Delta t_k,\ \dot{e}_k,\ \dot{n}_k,\ \dot{d}_k)^{\mathrm{T}} \ .$$

The process noise corresponding to east, north, and depth components of the conditional distribution of x_{k+1} given x_k was assumed to be Gaussian, with mean zero and standard deviation $.01\Delta t_k$. The process noise corresponding to the derivative vector of east, north, and depth components of the conditional mean x_{k+1} given x_k was also assumed Gaussian with mean zero and standard deviation $.2\Delta t_k$.

ℓ_2-smoother results without outlier removal are shown in Figure 4. There are three large peaks (two in the east component and one in the north component of the state) that are due to measurement outliers, and require either an outlier removal strategy or robust smoothing.

Three fits are shown in Figure 5: ℓ_1-Laplace, T-Robust, and ℓ_2-smoother with outlier removal. The darker curves appearing below the track are independent verifications using the GPS tracking near the top of the cable. A depth of 198 meters was added to the depth location of the GPS antenna so that the depth comparison can use the same axis for both the GPS data and the tracking

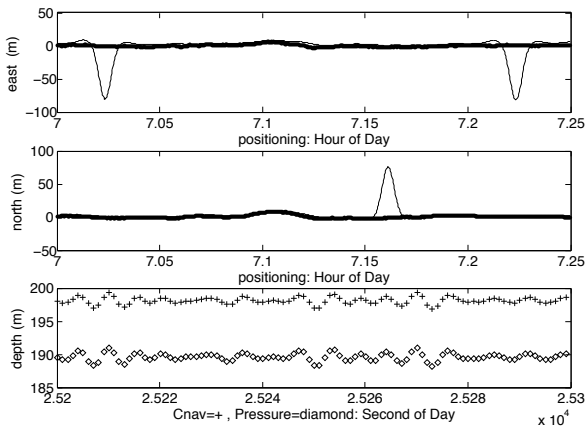


Fig. 4. Track: Independent GPS verification (thick line and +), ℓ_2-smoother estimate (thin line). Note the large outliers in the data.

results. Note that the time scale for the depth plots different (much finer) than the north, east, down plots, and demonstrates the accuracy of the GPS tracking as validated by the pressure sensor.

T-Robust, like the ℓ_1-Laplace smoother, was able to use the whole data sequence, despite large outliers in the data. The fits look very similar, and it is clear that T-Robust can also be used for smoothing in the presence of outliers. Note that the T-Robust track (b) is smoother than the ℓ_1-Laplace track (a) but has more detail than the ℓ_2-smoother track with outlier removal (c). This is easiest to see by comparing the east coordinates in (a), (b), and (c) of Figure 5, between 7.2 and 7.25 hours.

The residual plots in Figure 6 are quite revealing. Outliers are defined as measurements corresponding to residuals with absolute value greater than three standard deviations from the mean. All outliers are shown as 'o' characters, and those that fall outside the axis limits are plotted on the vertical axis limit lines. Note that the ℓ_2-smoother with outlier removal detects outliers after the first fit that are not outliers after the second fit. The peaks in Figure 4 are large enough to influence the entire fit, and so some points which are actually 'good' measurements are removed by the 3-σ edit rule, resulting in 'over-smoothing' of the outlier removal track and more detail in both of the robust smoothers in Figure 5.

The ℓ_1-Laplace smoother pushes more of the residuals to zero, particularly those corresponding to depth measurements, which are the most reliable and frequent. The T-Robust smoother is somewhere in between — the residuals for the depth track are smaller in comparison to the residuals of the ℓ_2-smoother, but are not set to zero as by the ℓ_1-Laplace smoother. As discussed previously, these features are artifacts of the behavior of the distributions at zero, and the choice of smoother should be guided by particular applications.

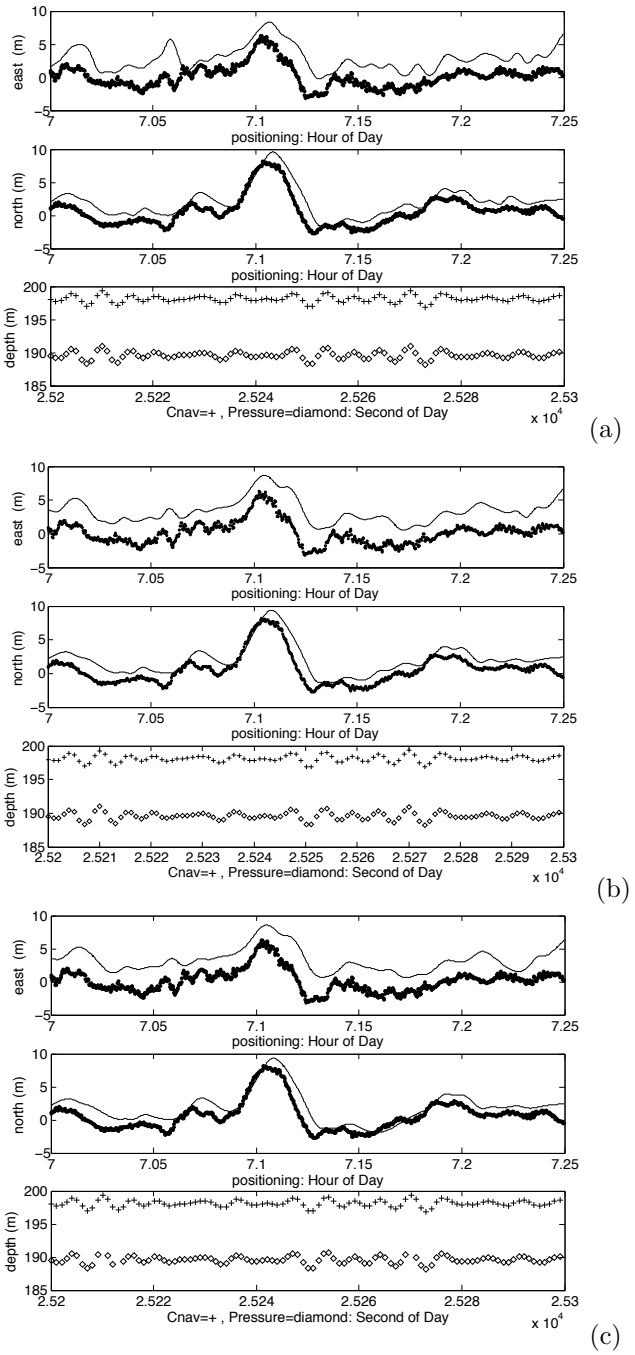
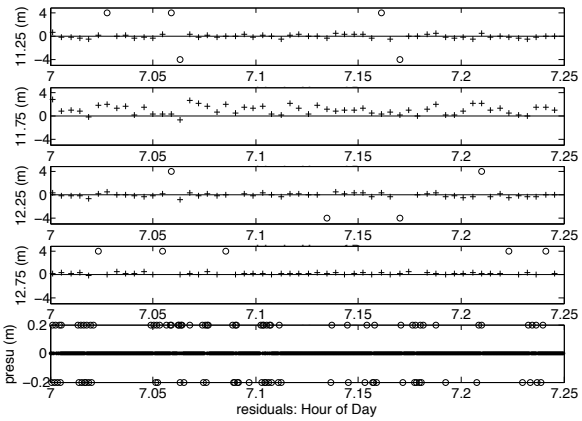


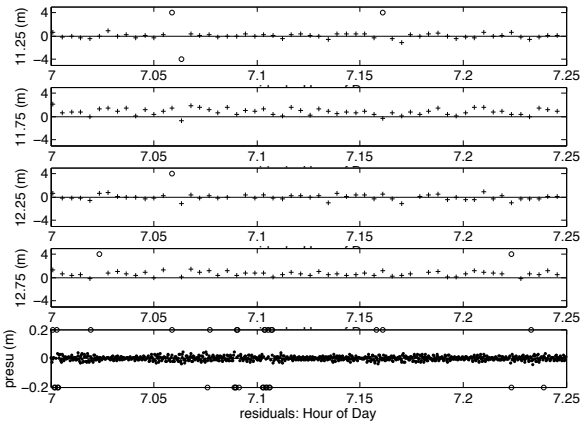
Fig. 5. Track: Independent GPS verification (thick line and +), (a): ℓ_1-Laplace smoother (thin line), (b): T-Robust smoother (thin line), (c): ℓ_2-smoother with outlier removal.

6.4 T-Trend Smoother: reconstruction of a sudden change in state

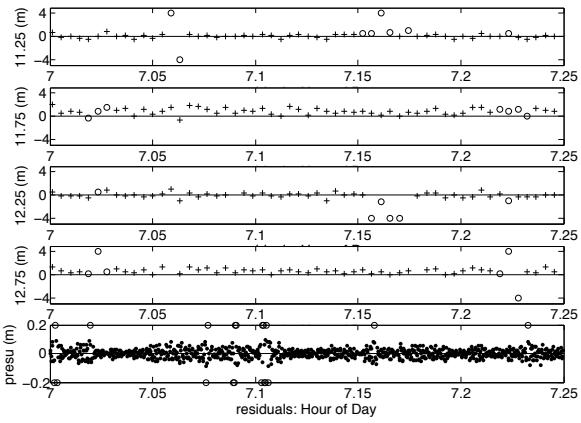
We present a proof of concept result for the T-Trend smoother, using two Monte Carlo studies of 200 runs. In the first study, the state vector, as well as the process and measurement models, are the same as in Sec. 6.1. At any run, x_2 has to be reconstructed from 20 measurements



(a)



(b)



(c)

Fig. 6. Residuals, (a): ℓ_1-Laplace smoother, (b): T-Robust smoother, (c): Gaussian smoother with outlier removal

corrupted by a white Gaussian noise of variance 0.05 and collected on $[0, 2\pi]$ using a uniform sampling grid. The top panel of Figure 7 reports the boxplot of the 200 root-MSE errors, with MSE defined by

$$\sqrt{\frac{1}{N}\sum_{k=1}^{N}[x_2(t_k) - \hat{x}_{2,k}]^2}$$

calculated using the ℓ_2-, ℓ_1-, and T-Trend smoothers, while the top panel of Figure 8 displays the estimate obtained in a single run. It is apparent that the performance of the three estimators is very similar.

The second experiment is identical to the first one except that we introduce a 'jump' at the middle of the sinusoidal wave. The bottom panel of Figure 7 reveals the superior performance of the T-Trend smoother under these perturbed conditions. The result depicted in the bottom panel of Figure 8 for a single run of the experiment is representative of the average performance of the estimators. The estimate achieved by the ℓ_2-smoother (dashed-line) does not follow the jump well (the true state is the solid line). The ℓ_1-smoother (dashdot) does a better job than the ℓ_2-smoother, and the T-trend smoother outperforms the ℓ_1-smoother, following the jump very closely while still providing a good solution along the rest of the path.

7 Conclusion

We have designed two new nonlinear smoothers, called T-Robust and T-Trend. These estimators efficiently solve for the MAP estimates of the states in a state-space model with Student's t measurement or Student's t process noise.

Similar to contributions in other applications, e.g. sparse system identification [27, 7, 24], our results underscore the significant advantages of using heavy tailed distributions in statistical modeling. Heavy tailed distributions force the use of non-convex loss functions to solve for the associated MAP estimates [3, Theorem 2]. The consequent challenge is to optimize a non-convex objective even when the system dynamics are linear. In contrast to the convex case, this requires an iterative smoother. The convergence analysis for these methods is still developed within the general framework of convex-composite optimization [6], although the details of the analysis differ.

Because the problems are non-convex, iterative methods may converge to local rather than global minima. This problem can be mitigated by an appropriate initialization procedure—for example, the ℓ_1-Laplace smoother can be used to obtain a starting point for the optimizer, in which case we can improve on the ℓ_1 solution when the data is highly contaminated with outliers. This approach was not taken in our numerical experiments, which used the same initial points. For all the linear experiments, the initial point was simply the null state sequence. For the Van Der Pol, the initial state x_0 was correctly specified in all experiments, and the remaining state estimates in the initial sequence were null.

The T-Robust smoother compares favourably to the ℓ_1-Laplace smoother described in [1], and outperforms it in our experiments when the data is heavily contaminated by outliers. The T-Trend smoother was designed for tracking signals that may exhibit sudden changes, and therefore has many potential applications in areas such as navigation and financial trend tracking. It was

demonstrated to follow a fast jump in the state better than a smoother with a convex penalty on model deviation. Both T-Robust and T-Trend smoothers can be implemented with minor modifications to an existing implementation of the ℓ_2 nonlinear smoother.

8 Acknowledgements

The authors thank the North Pacific Acoustic Laboratory (NPAL) investigators of the Applied Physics Laboratory, University of Washington for the underwater tracking data used in this paper (NPAL is sponsored by the Office of Naval Research code 321OA).

References

[1] A.Y. Aravkin, B.M. Bell, J.V. Burke, and G. Pillonetto. An ℓ_1-Laplace robust Kalman smoother. *IEEE Transactions on Automatic Control*, 2011.

[2] A.Y. Aravkin, B.M. Bell, J.V. Burke, and G. Pillonetto. Learning using state space kernel machines. In *Proc. IFAC World Congress 2011*, Milan, Italy, 2011.

[3] A.Y. Aravkin, M.P. Friedlander, F.J. Herrmann, and T. van Leeuwen. Robust inversion, dimensionality reduction, randomized sampling. *Submitted to Math. Prog.*, 2011.

[4] B. M. Bell, J. V. Burke, and G. Pillonetto. An inequality constrained nonlinear Kalman-Bucy smoother by interior point likelihood maximization. *Automatica*, 2008.

[5] B.M. Bell, G. Pillonetto, A.Y. Aravkin, and J.V. Burke. Matlab/Octave Package for constrained and robust Kalman smoothing, 2007-2011.

[6] J.V. Burke. Descent methods for composite nondifferentiable optimization problems. *Mathematical Programming*, 33:260–279, 1985.

[7] A. Chiuso and G. Pillonetto. Learning sparse dynamic linear systems using stable spline kernels and exponential hyperpriors. In *In Advances in Neural Information Processing Systems (NIPS*, 2010.

[8] Charles Chui and Guanrong Chen. *Kalman Filtering*. Springer, 2009.

[9] Ludwig Fahrmeir and Rita Kunstler. Penalized likelihood smoothing in robust state space models. *Metrika*, 49:173–191, 1998.

[10] S. Farahmand, G.B. Giannakis, and D. Angelosante. Doubly robust smoothing of dynamical processes via outlier sparsity constraints. *IEEE Transactions on Signal Processing*, 59:4529–4543, 2011.

[11] A. Gelb. *Applied Optimal Estimation*. The M.I.T. Press, Cambridge, MA, 1974.

[12] Frank R. Hampel, Elvezio M. Ronchetti, Peter J. Rousseeuw, and Werner A. Stahel. *Robust Statistics: The Approach Based on Influence Functions*. Wiley Series in Probability and Statistics, 1986.

[13] T. J. Hastie, R. J. Tibshirani, and J. Friedman. *The Elements of Statistical Learning. Data Mining, Inference and Prediction*. Springer, Canada, 2001.

[14] R. E. Kalman. A new approach to linear filtering and prediction problems. *Transactions of the AMSE - Journal of Basic Engineering*, 82(D):35–45, 1960.

[15] Kenneth L. Lange, Roderick J. A. Little, and Jeremy M. G. Taylor. Robust statistical modeling using the t distribution. *Journal of the American Statistical Association*, 84(408):881–896, 1989.

[16] Ricardo A. Maronna, Douglas Martin, and Yohai. *Robust Statistics*. Wiley Series in Probability and Statistics. Wiley, 2006.

[17] H. Ohlsson, F. Gustafsson, L. Ljung, and S. Boyd. State smoothing by sum-of-norms regularization. *Automatica (to appear)*, 2011.

[18] Ian R. Petersen and Andrey V. Savkin. *Robust Kalman Filtering for Signals and Systems with Large Uncertainties*. Control Engineering. Birkhauser, 1999.

[19] R. T. Rochafellar. *Convex Analysis*. Princeton University Press, 1970.

[20] R.T. Rockafellar and R.J.B. Wets. *Variational Analysis*, volume 317. Springer, 1998.

[21] I.C. Schick and S.K. Mitter. Robust recursive estimation in the presence of heavy-tailed observation noise. *The Annals of Statistics*, 22(2):1045–1080, June 1994.

[22] J.C. Spall. Estimation via Markov chain Monte Carlo. *Control Systems Magazine, IEEE*, 23(2):34 – 45, April 2003.

[23] R. Tibshirani. Regression shrinkage and selection via the LASSO. *Journal of the Royal Statistical Society, Series B.*, 58, 1996.

[24] M. Tipping. Sparse bayesian learning and the relevance vector machine. *Journal of Machine Learning Research*, 1:211–244, 2001.

[25] G. Wahba. *Spline models for observational data*. SIAM, Philadelphia, 1990.

[26] Mike West and Jeff Harrison. *Bayesian Forecasting and Dynamic Models*. Springer, second edition, 1999.

[27] D.P. Wipf and B.D. Rao. An empirical bayesian strategy for solving the simultaneous sparse approximation problem. *IEEE Transactions on Signal Processing*, 55(7):3704–3716, 2007.

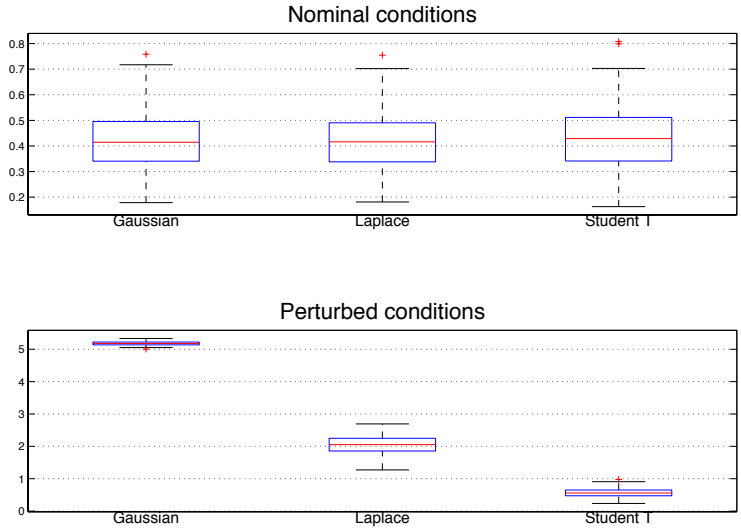


Fig. 7. Reconstruction of a sudden change in state: Monte Carlo simulation. Boxplot of errors obtained using Gaussian, Laplace and Student's T Kalman smoother under nominal (top) and perturbed (bottom) conditions.

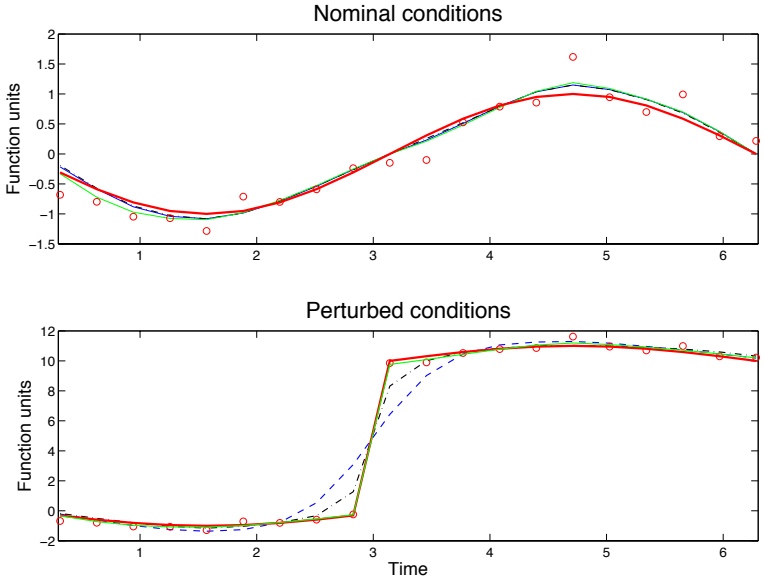


Fig. 8. Reconstruction of a sudden change in state: results from a Monte Carlo run under nominal (top) and perturbed (bottom) conditions using ℓ_2 (dashed), ℓ_1 (dashdot) and $T - Trend$ (thin line) smoother. The thick line is the true state.