Australian Centre for Field Robotics
A Key Centre of Teaching and Research

The Rose Street Building J04
The University of Sydney 2006 NSW Australia

Robust Estimation in Non-linear State-space Models with State-dependent Noise

Gabriel Agamennoni and Eduardo M. Nebot

T: + 61 2 9351 8173
E: g.agamennoni@acfr.usyd.edu.au

Technical Report ACFR-TR-2013-002

31 May 2013
Released 13 June 2013

Contents

Abstract

In this paper we present a robust estimation algorithm for non-linear state-space models driven by state-dependent noise. We derive the algorithm step by step from first principles, from theory to implementation. The implementation is straightforward and consists mainly of two components: 1. a slightly modified version of the Rauch-Tung-Striebel recursions; and 2. a backtracking line search strategy. Since it preserves the underlying chain structure of the problem, its computational complexity grows linearly with the number of data. The algorithm is iterative and is guaranteed to converge, under mild assumptions, to a local optimum from any starting point. We validate our approach via experiments on synthetic data from a multi-variate stochastic volatility model.

1 Introduction

1.1 State-space models

State-space models (SSMs) are ubiquitous in many fields of engineering and applied sciences such as autonomous navigation [1, 2], target tracking [3, 4] and computational econometrics [5, 6]. An SSM is an abstract representation of a dynamic system with inputs and outputs, where its dynamics and input-output characteristics are represented in terms of state variables. Unlike inputs and outputs, the states are hidden and may not be measured directly (e.g. with a sensor).

Estimation is the problem of determining the system's states given pairs of input-output data. However, because the state and input-output processes are generally random (i.e. they are driven by noise), the states cannot be recovered exactly. The estimates are always afflicted with uncertainty.

1.2 Outliers and non-Gaussian noise

These random processes are often specified as conditional probability distributions. The most common distribution is the normal, or Gausian. It is justified by the central limit theorem[1] and favored for its convenient analytical properties, seldom motivated by the nature of the problem.

Since it appears relatively frequently, there is an unfortunate tendency to invoke the Gaussian in situations where it may not be applicable. Estimation is based on assumptions about the random processes driving the states. Therefore, if the Gaussian assumption does not hold then the estimates may be misleading and there is a significant risk of drawing incorrect conclusions about the system.

Outliers are a common type non-Gaussian phenomenon [7, 8]. Intuitively, outliers are measurements that do not agree with the bulk of the data. Although theoretically they may occur by chance in most distributions,[2] outliers often stem from effects that are either unknown or are deliberately excluded from the model as they are tedious or impractical to account for (e.g. external disturbances).

Systems that rely on high-quality data (e.g. mobile robotic platforms) are usually sensitive to outliers. In some cases even a few outliers that pass undetected may cause the system to fail to the point that a full recovery is impossible [4, 9, 10]. Hence the importance of outlier-robust estimators.

1.3 State-dependent noise

In addition to Gaussianity, another major, commonly made assumption is that the noise levels are constant (e.g. a sensor's noise characteristics can be condensed into a covariance matrix). In reality, however, the noise is a function of the system's state (e.g. a uniform acoustic linear array [11]).

We agree with the view of Saha and Gustaffson [12], namely that state-dependent noise might be more common in practice than what the literature acknowledges. Examples include:

- The discrete-time equivalent of a system of non-linear differential equations often results in a transition model with state-dependent variance;
- The noise properties of some transition models are inherently heterogeneous (e.g. a car's maneuverability depends on its speed);
- Some sensors (e.g. stereo cameras) typically exhibit constant *relative* noise characteristics — i.e. the *absolute* noise increases with the magnitude of the signal;

[1] The central limit theorem guarantees that the arithmetic mean of a large number of independent variates drawn from an arbitrary distribution —having finite mean and variance—is approximately Gaussian-distributed, provided the sample is large enough.

[2] For example, the Gaussian places over 99% of its probability mass within the interval $\mu \pm 3\sigma$. An outlier 5σ away from μ has less than *one in a million* chances of occurring. Even though the possibility exists, it is unrealistically small.

- Sometimes the dependency of the noise on the states is the only available cue for inferring the states (e.g. stochastic volatility models).

With a few exceptions [13, 14], there seems to be very little treatment of SSMs with state-dependent noise in the literature.

1.4 Related work

In this paper we address the problem of robust estimation in non-linear systems driven by state-dependent noise. To the best of the authors' knowledge, no previous work has addressed all three aspects simultaneously (i.e. robustness, non-linearity and state-noise interdependency) and deterministically.

The following work is closely related to ours, although the three aspects are only partially covered:

- Julier and Uhlmann [15] introduced a deterministic sampling method for non-linear systems called the unscented Kalman filter, which Särkkä [16] extended to the smoothing case;
- Särkkä and Nummenmaa [17] developed a filter based on variational approximations that tracks states and noise simultaneously, albeit for linear systems where the state and noise processes evolve independently;
- Agamennoni et al. [18, 19] extended this approach by accounting for cross-correlations between the output dimensions and deriving smoothing and parameter learning algorithms, though again in the context of linear systems;
- Aravkin et al. [20–22] formulated a number of robust smoothing methods for non-linear SSMs from an optimization viewpoint, although their treatment assumes that the noise is independent of the states;
- Piché et al. [23] presented an outlier-robust filter/smoother for non-linear systems in the assumed density filtering framework, again with state-independent noise;
- Spinello and Stilwell [13] generalized an iterated version of the extended Kalman filter [24] to cases where the observation noise is state-dependent and Gaussian, encoding the dependency via the covariance parameters.

1.5 List of major contributions

The major contributions in this paper are:

- A non-trivial generalization of Aravkin's work [21, 22] to models with state-dependent noise;
- A computationally efficient and provably convergent algorithm for solving the maximum a posteriori estimation problem which, unlike the approach of Aravkin et al., does *not* require approximating the Hessian matrix;
- A parameterization carefully designed in order to make equations readily interpretable and assert our algorithm's strong connection to the well-known Rauch-Tung-Striebel recursions.

We validate our approach via experiments on synthetic data from a challenging problem. The code for our implementation is available from the author's web page [25].

1.6 Outline of the paper

This paper is organized as follows. In section 2 we define the estimation problem that we aim to tackle and discuss a few interesting properties of heavy-tailed distributions. In section 3 we derive an upper bound on the objective function that we wish to minimize and show that this bound is a quadratic-composite function of the states. With this in mind, we develop an iterative optimization algorithm in section 4 to solve the estimation problem. In section 5 we test our algorithm on synthetic data and compare its performance against other methods. Finally, in section 6 we offer a brief summary and outline possibilities for future work.

2 The Estimation Problem

2.1 Notation and definitions

Let X be a sequence of n states and let Y be a corresponding sequence of n measurements:

$$X = (\mathbf{x}_1, \ldots, \mathbf{x}_n) \in \mathbb{R}^{d \times \ldots \times d} \tag{1a}$$

$$Y = (\mathbf{y}_1, \ldots, \mathbf{y}_n) \in \mathbb{R}^{d_1 \times \ldots \times d_n} \tag{1b}$$

where $\mathbb{R}^d$ stands for real d-dimensional Euclidean space.

Let $\mathbf{g}_k$ be the transition function from steps $k-1$ to k and let $\mathbf{h}_k$ be the observation function at step k. Let $\mathbf{Q}_k$ and $\mathbf{R}_k$ be, in that order, the noise matrices of the state and measurement processes. Namely,

$$\mathbf{g}_k : \mathbb{R}^d \to \mathbb{R}^d \qquad \mathbf{Q}_k : \mathbb{R}^d \to \mathbb{P}^d \tag{2a}$$

$$\mathbf{h}_k : \mathbb{R}^d \to \mathbb{R}^{d_k} \qquad \mathbf{R}_k : \mathbb{R}^d \to \mathbb{P}^{d_k} \tag{2b}$$

for $k \in \{1, \ldots, n\}$, where $\mathbb{P}^d$ denotes the cone of real $d \times d$ symmetric, positive-definite matrices.

Last of all, let $\{\mathbf{u}_k\}$ and $\{\mathbf{v}_k\}$ be, respectively, the state and measurement noise processes:

$$\mathbf{u}_k \sim \mathcal{N}_d(\mathbf{0}, \mathbf{I}) \tag{3a}$$

$$\mathbf{v}_k \sim \mathcal{T}_{d_k}(\mathbf{0}, \mathbf{I}, s_k) \tag{3b}$$

for all k, $1 \leq k \leq n$, where $\mathcal{N}_d(\mu, \mathbf{\Sigma})$ is a d-variate normal, or Gaussian distribution with mean vector μ and variance-covariance matrix $\mathbf{\Sigma} \succ \mathbf{0}$, and $\mathcal{T}_d(\mu, \mathbf{\Sigma}, \nu)$ is a d-variate t distribution with location vector μ, scale matrix $\mathbf{\Sigma} \succ \mathbf{0}$ and $\nu > 0$ degrees of freedom [26].

2.2 Robust non-linear estimation

Assume that the state and observation sequences are generated by the following processes:

$$\mathbf{x}_k = \mathbf{g}_k(\mathbf{x}_{k-1}) + \mathbf{Q}_k^{\frac{1}{2}}(\mathbf{x}_{k-1})\, \mathbf{u}_k \tag{4a}$$

$$\mathbf{y}_k = \mathbf{h}_k(\mathbf{x}_k) + \mathbf{R}_k^{\frac{1}{2}}(\mathbf{x}_k)\, \mathbf{v}_k \tag{4b}$$

for $k = 1$ to n,[3,4] Or equivalently,

$$\mathbf{x}_k | \mathbf{x}_{k-1} \sim \mathcal{N}_d(\mathbf{g}_k(\mathbf{x}_{k-1}), \mathbf{Q}_k(\mathbf{x}_{k-1})) \tag{5a}$$

$$\mathbf{y}_k | \mathbf{x}_k \sim \mathcal{T}_{d_k}(\mathbf{h}_k(\mathbf{x}_k), \mathbf{R}_k(\mathbf{x}_k), s_k) \tag{5b}$$

Then, the probability density function of the joint distribution over state and measurement sequences is[5]

$$p(X, Y) = \prod_{k=1}^{n} p(\mathbf{x}_k | \mathbf{x}_{k-1})\, p(\mathbf{y}_k | \mathbf{x}_k) \tag{6}$$

Note that, even though states are *conditionally* Gaussian- and observations are *conditionally* t-distributed, in general they are not *jointly* Gaussian nor t.[6]

2.3 The t distribution

The t is a sub-exponential distribution —its tails decay to zero at a less-than-exponential rate. It has much heavier tails than the Gaussian, which is super-exponential. The weight of the tails depends on the number of degrees of freedom: in the limit $\nu \to \infty$ the tails flatten and the t reduces to a Gaussian. As ν becomes smaller the t spreads its probability mass more and more evenly across its sample space and further away from the mode, assigning outliers a non-negligible probability.

[3] The notation $\mathbf{A}^{1/2}$ stands for the lower-triangular Cholesky factor of a symmetric, positive-definite matrix $\mathbf{A}$.

[4] With the convention $\mathbf{f}_1(\mathbf{x}_0) = \mu$ and $\mathbf{Q}_1(\mathbf{x}_0) = \mathbf{\Sigma}$.

[5] Throughout this manuscript we will use the same symbols to denote both the random *variable* and its realization, the random *variate*. Although this is a slight abuse of notation, it is in the interest of clarity and should cause no confusion.

[6] The t is not closed under convolution. Therefore, even if $\mathbf{g}_k$ and $\mathbf{h}_k$ are affine and $\mathbf{Q}_k$ and $\mathbf{R}_k$ are constant, (X, Y) is still not t-distributed.

Placing a non-negligible probability on outliers is not a disadvantage; it simply reflects reality. The Gaussian concentrates most of its probability mass in a small region around the mean, essentially ruling out the possibility that a measurement is ever wrong. The t makes no such mistake.

Our observation model in (5b) acknowledges the fact that, occasionally, the measurement process may produce inconsistent readings. By imparting this information directly into the model we enable it to deal with outliers natively within the estimation framework. As a result, there is no need for us to explicitly pre-process outliers or treat them separately because the model is now capable of explaining them.

2.4 Characteristics of the t family

The family of t distributions has a few unique and attractive features, which sets it apart from the Gaussian:

2.4.1 The influence function

The influence function [27] quantifies the sensitivity of a distribution with respect to small changes in the data. It is directly related to the derivative of the negative log-density function. Fig. 1 shows the t's influence function for different numbers of degrees of freedom. The Gaussian (i.e. the limit $\nu \to \infty$) is augmented with a 95% confidence χ^2 test, a common practice in Kalman filtering for rejecting outliers [28].

Fig. 1 illustrates the difference in the way the t and the Gaussian react to measurement outliers. Data lying far away from the origin will exert an increasingly large influence on the Gaussian until reaching the $\chi^2_{95\%}$ threshold. Beyond this threshold, the data are discarded and their influence is null. The t is not so categorical. Its influence gradually tends to zero, down-weighting the data in a continuous fashion as they are pulled further and further away from the origin. Eventually, the data are ignored. In the meantime, however, the information gain —albeit small—remains positive.

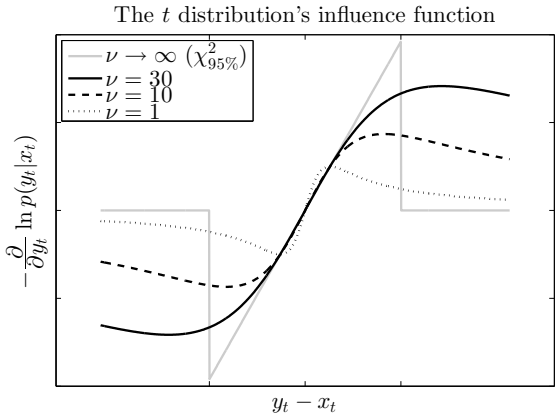

Figure 1: The influence function captures how a distribution changes with the data. The t's influence function is re-descending: it decreases to zero away from the origin. In contrast, the Gaussian's influence function ($\nu \to \infty$) is linear and thus unbounded, unless a χ^2 validation threshold is in place.

2.4.2 The moment of indecision

The "moment of indecision," coined by O'Hagan [29] and addressed in [30] and more recently in [4], is an interesting phenomenon characteristic of heavy-tailed distributions. At times it is hard to tell a true outlier from the outcome of a large but natural variance. For instance, when a measurement conflicts with the prior but is not far enough so that the distinction is not obvious. In this gray area the t behaves in a peculiar way that, although logical, is not fully intuitive.

The term "moment of indecision" refers to the fact that a product of t density functions can have multiple maxima. In other words, Bayes' rule can produce a multi-modal posterior. When the prior and the data strongly disagree, the posterior splits into two distinct peaks that capture two hypotheses: one where the measurement is wrong (i.e. an outlier) and another where the prior is wrong. The ambiguity is resolved later on when more data become available, reinforcing one of the modes and suppressing the other.

Fig. 2 illustrates this phenomenon. The prior (dotted) and the likelihood (dashed) are one-dimensional t densities with $\nu = 2$ and 1 degrees of freedom, respectively. The posterior (solid) is a poly-t density [31]. Initially, the prior and the likelihood are concentric ($y_t = x_t$) and the posterior is uni-modal (upper-left panel). As the prior and the likelihood diverge (i.e. $y_t - x_t$ increases) the posterior widens and shifts to the right (upper-right). When the difference is large enough, the single mode breaks into two modes that move in opposite directions (lower-left). Eventually, the probability mass under the right mode becomes negligible and dies out, and the posterior reverts to the prior.[7]

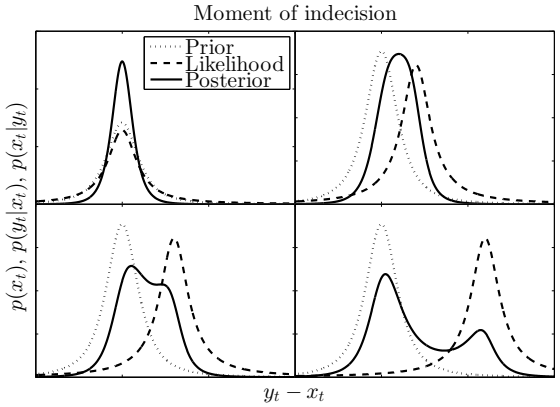

Figure 2: The "moment of indecision" is a phenomenon typical of heavy-tailed distributions. As the prior and likelihood diverge, the posterior becomes multi-modal in order to account for two possibilities: the first where the measurement is an outlier, and the second where the prior is off.

3 Robust Non-linear Estimation

3.1 Smoothing vs. maximum a posteriori estimation

Given Y, smoothing consists of finding and manipulating $X|Y$, the posterior distribution over state sequences. This distribution is analytically intractable since its probability density function, $p(X|Y)$, has no closed-form expression. Hence the smoothing problem is extremely challenging in the robust non-linear model.

In contrast to smoothing, maximum *a posteriori* estimation seeks to find a maximum of $p(X|Y)$ — or, equivalently, a maximum of $p(X, Y)$—rather than the density function itself. Maximization provides a point estimate of the state sequence and is analytically tractable. Still, the density function (6) is not log-concave and the problem involves a large number of variables, meaning that naïve approaches are prone to get stuck in poor local optima.

3.2 The objective function

Maximizing (6) is equivalent to minimizing

$$-\ln p(X, Y) = \sum_{k=1}^{n} \left(-\ln p(\mathbf{x}_k|\mathbf{x}_{k-1}) - \ln p(\mathbf{y}_k|\mathbf{x}_k)\right) \tag{7}$$

The main difficulties of this minimization problem are:

1. the conditional location parameters are non-linear and the conditional scale parameters non-constant; and
2. even if they were, the t's probability density function is not log-concave.

Consequently, the objective function (7) is non-linear and non-convex.

[7]The number of degrees of freedom determines which mode survives. In this case, ν is larger for the prior than for the likelihood. If the opposite were true, the left mode would ultimately become extinct and the posterior would follow the likelihood instead of returning to the prior.

3.3 Gauss-Gamma decomposition

A t-distributed random variable can be expressed as a scale mixture of Gaussian variables. Namely, (5b) is equivalent to

$$\begin{cases} \mathbf{y}_k|\mathbf{x}_k, w_k \sim \mathcal{N}_{d_k}(\mathbf{h}_k(\mathbf{x}_k)\,,\, \mathbf{R}_k(\mathbf{x}_k)/w_k\,) \\ w_k \sim \mathcal{G}\,(\,s_k/2\,,\,s_k/2\,) \end{cases}$$

where $\mathcal{G}\,(\alpha, \beta)$ stands for a Gamma distribution with shape $\alpha > 0$ and rate $\beta > 0$ [26].

The weight w_k is an auxiliary random variable. It is particularly convenient because it renders the measurement process conditionally Gaussian. Therefore, if w_k was known for all k, minimizing eq. (7) would be a quadratic-composite problem.[8] With this in mind, we shall now derive a quadratic-composite upper bound on the objective function.

3.4 A quadratic-composite upper bound

Consider the second term inside the summation in the right-hand side of eq. (7). Let $p\,(w_k)$ be w_k's probability density function and $q\,(w_k)$ be an arbitrary probability density function over the positive real line.

If $q\,(w_k)$ is fixed, this term is bounded from above. Invoking the Gauss-Gamma decomposition, multiplying and dividing by $q\,(w_k)$ and applying Jensen's inequality [33] produces

$$\begin{aligned} -\ln p\,(\mathbf{y}_k|\mathbf{x}_k) &= -\ln \int p\,(\mathbf{y}_k|\mathbf{x}_k, w_k)\, p\,(w_k)\, dw_k \\ &= -\ln \int p\,(\mathbf{y}_k|\mathbf{x}_k, w_k)\, \frac{p\,(w_k)}{q\,(w_k)} q\,(w_k)\, dw_k \\ &\leq -\int \ln \left[p\,(\mathbf{y}_k|\mathbf{x}_k, w_k)\, \frac{p\,(w_k)}{q\,(w_k)} \right] q\,(w_k)\, dw_k \end{aligned}$$

After expanding the logarithm and rearranging we arrive at

$$-\ln p\,(\mathbf{y}_k|\mathbf{x}_k) \leq \underbrace{-\int \ln p\,(\mathbf{y}_k|\mathbf{x}_k, w_k)\, q\,(w_k)\, dw_k}_{\text{Quadratic-composite function of } \mathbf{x}_k} + \underbrace{\int \ln \frac{q\,(w_k)}{p\,(w_k)} q\,(w_k)\, dw_k}_{\text{Independent of } \mathbf{x}_k} \tag{8}$$

Up to an additive constant independent of $\mathbf{x}_k$, the right-hand side of this inequality is

$$\frac{\omega_k}{2} \left\| \mathbf{v}_k(\mathbf{x}_k) \right\|^2 + \frac{1}{2} r_k(\mathbf{x}_k)$$

where we have defined

$$\omega_k \triangleq \int w_k\, q\,(w_k)\, dw_k \tag{9a}$$

$$\mathbf{v}_k(\mathbf{x}_k) \triangleq \mathbf{R}_k^{\frac{1}{2}}(\mathbf{x}_k)^{-1} \left(\mathbf{y}_k - \mathbf{h}_k(\mathbf{x}_k) \right) \tag{9b}$$

$$r_k(\mathbf{x}_k) \triangleq \ln \det \mathbf{R}_k(\mathbf{x}_k) \tag{9c}$$

Hence the upper bound in ineq. (8) is a quadratic-composite function of $\mathbf{x}_k$, i.e. it is quadratic in $\mathbf{v}_k$ and r_k even though it is not necessarily quadratic in $\mathbf{x}_k$.

For a fixed $\mathbf{x}_k$, the bound attains its minimum if and only if $q\,(w_k) = p\,(w_k|\mathbf{x}_k, \mathbf{y}_k)$ for all $w_k > 0$[9] (the reader may verify this via substitution and applying Bayes' rule). In this case the bound is tight and the inequality becomes an equality.

[8] A convex-composite problem [32] is of the form

$$\min_{\mathbf{x} \in S} f\,(\mathbf{g}\,(\mathbf{x}))$$

where $f : \mathbb{R}^m \to \mathbb{R}$ is convex, $\mathbf{g} : \mathbb{R}^n \to \mathbb{R}^m$ is continuously differentiable and $S \subseteq \mathbb{R}^n$ is affine. In particular, if f is a quadratic function then the problem is quadratic-composite.

[9] Except, perhaps, on a set of measure zero.

3.5 Putting the pieces together

Applying (8) to each of the terms inside the summation in eq. (7) leads to a quadratic-composite upper bound on the objective function. Let us define

$$\mathbf{u}_k(\mathbf{x}_{k-1}, \mathbf{x}_k) \triangleq \mathbf{Q}_k^{\frac{1}{2}}(\mathbf{x}_{k-1})^{-1}(\mathbf{x}_k - \mathbf{g}_k(\mathbf{x}_{k-1})) \tag{10a}$$

$$q_k(\mathbf{x}_{k-1}) \triangleq \ln\det \mathbf{Q}_k(\mathbf{x}_{k-1}) \tag{10b}$$

in analogy with eqs. (9b) and (9c) and let $W = (w_1, \ldots, w_n)$ be the sequence of weights. Then,

$$-\ln p(X, Y) \leq b(X, q(W)) \tag{11a}$$

where

$$b(X, q(W)) = \frac{1}{2}\sum_{k=1}^{n}\left[\|\mathbf{u}_k(\mathbf{x}_{k-1}, \mathbf{x}_k)\|^2 + \omega_k \|\mathbf{v}_k(\mathbf{x}_k)\|^2 + q_k(\mathbf{x}_{k-1}) + r_k(\mathbf{x}_k) + \ldots\right] \tag{11b}$$

with the dots denoting terms independent of the states.

Minimizing the bound with respect to $q(W)$ tightens it and turns ineq. (11a) into an equation. Thus we can write

$$\arg\min_X -\ln p(X|Y) = \arg\min_{X,\,q} b(X, q(W)) \tag{12}$$

This equation is remarkable as it decouples the original non-linear, non-convex minimization problem into a problem that is quadratic-composite in X —which allows for a much more efficient solution— and trivial to solve in $q(W)$. In addition, it suggests a coordinate descent algorithm for minimizing the objective function by iteratively tightening the bound.

4 Algorithm and Implementation

4.1 The maximum a posteriori estimation algorithm

We now propose an algorithm for carrying out maximum a posteriori estimation on robust non-linear state-space models. Starting from an initial guess of the state sequence, repeat the following steps until convergence:

1. While keeping the states fixed, minimize the bound with respect to the weights; and
2. While keeping the weights fixed, minimize the bound with respect to the states.

Assuming that the objective is well-behaved,[10] convergence to a local minimum is guaranteed even if neither 1) nor 2) minimize the bound, provided they do not increase it.

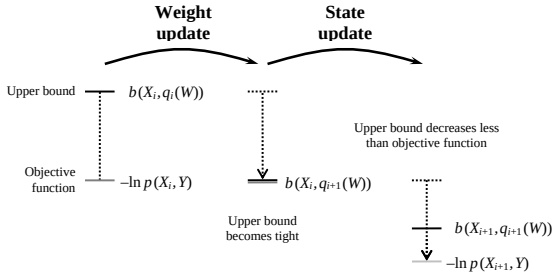

Figure 3: Illustration of the coordinate descent algorithm (ith iteration). First, step 1) updates the weights' densities. At this point the upper bound is tight. Then, step 2) updates the state sequence and decreases the bound further. The objective function must decrease by at least as much as the bound since it is no longer tight. As iterations progress, the gap becomes smaller and smaller and eventually falls below the relative tolerance.

Fig. 3 illustrates steps 1) and 2) of the algorithm at the ith iteration. Let us go over both steps in more detail:

[10] That is, assuming that a minimum exists.

4.1.1 Updating the weights

For a fixed state sequence, minimization with respect to the weights' density is trivial. From 3.4 we know that the bound is tightened when $q(W)$ is equal to the density of $W|X,Y$. Since the weights are conditionally independent given the states, the density factorizes as $\prod_{k=1}^{n} q(w_k)$. Applying Bayes' rule to the Gauss-Gamma decomposition of the kth factor yields

$$w_k|\mathbf{x}_k, \mathbf{y}_k \sim \mathcal{G}(\alpha_k, \beta_k)$$

$$\alpha_k = \frac{s_k + d_k}{2} \quad \beta_k = \frac{s_k + \|\mathbf{v}_k(\mathbf{x}_k)\|^2}{2}$$

and hence the optimal $q(w_k)$ is given by

$$q(w_k) = \frac{\beta_k^{\alpha_k}}{\Gamma(\alpha_k)} w_k^{\alpha_k - 1} \exp(-\beta_k w_k)$$

for $w_k > 0$, where Γ is the gamma function [34].

Now that we know w_k's density we can calculate all of the moments that eq. (11b) requires. In particular, evaluating ω_k as defined in eq. (9a) produces[11]

$$\omega_k = \frac{\alpha_k}{\beta_k} = \frac{s_k + d_k}{s_k + \|\mathbf{v}_k(\mathbf{x}_k)\|^2} \tag{13}$$

Given $\mathbf{x}_k$, the expected value of w_k decreases according to the square of the normalized residual. Intuitively, measurements are down-weighted by their disagreement with the states.

4.1.2 Updating the states

Minimization with respect to X, for a fixed $q(W)$, is achieved numerically. We apply a variant of the Gauss-Newton method developed by Burke and Ferris [32] in the context of convex-composite optimization. At each iteration of the method, we compute a sequence of search directions by solving a quadratic sub-problem and then perform a line search along these directions.

The objective function for the quadratic sub-problem results from replacing $\mathbf{u}_k$, $\mathbf{v}_k$, q_k and r_k in eq. (11b) by their first-order Taylor expansions around $\mathbf{x}_{k-1}$ and $\mathbf{x}_k$. Specifically, if ΔX is the sequence of linear variables,

$$\Delta X = (\Delta\mathbf{x}_1, \ldots, \Delta\mathbf{x}_n)$$

then the search directions are computed by solving

$$\min_{\Delta X} \sum_{k=1}^{n} f_k(\Delta\mathbf{x}_{k-1}, \Delta\mathbf{x}_k) \tag{14}$$

where f_k is given by eqs. (15) and (16). (Deriving these equations is lengthy but straightforward, involving little more than differentiation and algebraic manipulation.)

If we take a close look at (15) we realize that the mathematical structure of the sub-problem is almost the same as that of a linear, time-varying Kalman smoother. The objective function is quadratic in ΔX and consists of a sum of pairwise terms (involving $\Delta\mathbf{x}_{k-1}$ and $\mathbf{x}_k$) and singleton terms (involving only $\mathbf{x}_k$). Thus we can solve prob. (14) via a modified form of the Rauch-Tung-Striebel (RTS) recursions [37].[12]

Line search may be performed in a number of ways [38]. We chose a backtracking strategy, which is fast and effective. Let h be the step size along the search direction ΔX. Starting from $h = 1$, we either accept the step if it satisfies the Wolfe conditions, or shrink it by a constant factor and try again.

[11] The dots in (11b) represent terms independent of $\mathbf{x}_k$. Evaluating these terms, although tedious, is straightforward. They are given by

$$\frac{d_k}{2} \mathrm{E}[\ln(2\pi w_k)] + \mathrm{KL}[q(w_k) \| p(w_k)]$$

where E denotes expectation with respect to $q(w_k)$ and KL is the Kullback-Leibler divergence [35].

[12] The only difference between the objective function of the quadratic sub-problem and that of the Kalman smoother is the presence of the additional terms

$$\mathbf{q}_k(\mathbf{x}_{k-1})^\top \Delta\mathbf{x}_{k-1} \quad \mathbf{r}_k(\mathbf{x}_k)^\top \Delta\mathbf{x}_k$$

outside of the square in eq. (15). These are caused by state-dependent noise. However, since they are linear in $\Delta\mathbf{x}_{k-1}$ and $\Delta\mathbf{x}_k$, they can be accounted for during the forward pass by executing an extra correction step with "fictitious" zero-valued measurements.

$$
\begin{aligned}
f_k(\Delta\mathbf{x}_{k-1}, \Delta\mathbf{x}_k) = \frac{1}{2} &\left\| \mathbf{Q}_k^{\frac{1}{2}}(\mathbf{x}_{k-1})^{-1} \left(\Delta\mathbf{x}_k - \mathbf{G}_k(\mathbf{x}_{k-1}, \mathbf{x}_k)\, \Delta\mathbf{x}_{k-1} - \mathbf{g}_k(\mathbf{x}_{k-1}) + \mathbf{x}_k \right) \right\|^2 \\
&+ \frac{\omega_k}{2} \left\| \mathbf{R}_k^{\frac{1}{2}}(\mathbf{x}_k)^{-1} \left(\mathbf{y}_k - \mathbf{H}_k(\mathbf{x}_k)\, \Delta\mathbf{x}_k - \mathbf{h}_k(\mathbf{x}_k) \right) \right\|^2 \\
&+ \mathbf{q}_k(\mathbf{x}_{k-1})^\top \Delta\mathbf{x}_{k-1} + \mathbf{r}_k(\mathbf{x}_k)^\top \Delta\mathbf{x}_k + \dots
\end{aligned}
\tag{15}
$$

$$
\mathbf{G}_k(\mathbf{x}_{k-1}, \mathbf{x}_k) = \begin{bmatrix} \dots & \dfrac{\partial \mathbf{g}_k}{\partial x_i}(\mathbf{x}_{k-1}) + \dfrac{\partial \mathbf{Q}_k^{\frac{1}{2}}}{\partial x_i}(\mathbf{x}_{k-1})\ \mathbf{u}_k(\mathbf{x}_{k-1}, \mathbf{x}_k) & \dots \end{bmatrix} \tag{16a}
$$

$$
\mathbf{H}_k(\mathbf{x}_k) = \begin{bmatrix} \dots & \dfrac{\partial \mathbf{h}_k}{\partial x_i}(\mathbf{x}_k) + \dfrac{\partial \mathbf{R}_k^{\frac{1}{2}}}{\partial x_i}(\mathbf{x}_k)\ \mathbf{v}_k(\mathbf{x}_k) & \dots \end{bmatrix} \tag{16b}
$$

$$
\mathbf{q}_k(\mathbf{x}_{k-1}) = \begin{bmatrix} \dots & \operatorname{tr}\left(\mathbf{Q}_k^{\frac{1}{2}}(\mathbf{x}_{k-1})^{-1} \dfrac{\partial \mathbf{Q}_k^{\frac{1}{2}}}{\partial x_i}(\mathbf{x}_{k-1}) \right) & \dots \end{bmatrix}^\top \tag{16c}
$$

$$
\mathbf{r}_k(\mathbf{x}_k) = \begin{bmatrix} \dots & \operatorname{tr}\left(\mathbf{R}_k^{\frac{1}{2}}(\mathbf{x}_k)^{-1} \dfrac{\partial \mathbf{R}_k^{\frac{1}{2}}}{\partial x_i}(\mathbf{x}_k) \right) & \dots \end{bmatrix}^\top \tag{16d}
$$

Figure 4: The kth term of the objective function of the quadratic sub-problem (14). The ith columns of matrices $\mathbf{G}_k$ and $\mathbf{H}_k$ are shown in eqs. (16a) and (16b), and the ith elements of vectors $\mathbf{q}_k$ and $\mathbf{r}_k$ appear in eqs. (16c) and (16d), respectively. Calculating Cholesky factors and evaluating their derivatives can be done simultaneously and requires the same order of complexity [36].

After each successful step, we check the following termination criterion to assess convergence:

$$
-n^{-1} \operatorname{tr}\left(\frac{\partial b}{\partial X} \left(X, q\left(W\right) \right)^\top \Delta X \right) \leq \epsilon \tag{17a}
$$

where $\epsilon > 0$. (Refer to appendix A for an expression of the derivatives of the upper bound.) When this criterion is met we consider that the Gauss-Newton method converged to a local minimum and return to step 1).

4.2 Assessing convergence

To assess convergence of the overall algorithm we monitor the objective function between two consecutive iterations. If

$$
\ln p\left(X_{i+1}, Y\right) - \ln p\left(X_i, Y\right) \leq \delta \tag{17b}
$$

for $\delta > 0$ then we deem that a local minimum has been found and terminate the algorithm.

4.3 Pseudo-code

Alg. 1 shows the pseudo-code implementing the robust non-linear estimator. Line 2 implements step 1), while lines 3 to 7 implement step 2). In line 4 the search directions are computed via modified RTS recursions, and in line 5 a backtracking line search strategy selects the step size.

Algorithm 1 The robust non-linear estimator.

Require: an initial guess X

1: **repeat**
2: for all k, $\omega_k \leftarrow$ (13) *# step 1)*
3: **repeat**
4: $\Delta X \leftarrow$ the solution to prob. (14)
5: calculate h via backtracking
6: $X \leftarrow X + h\,\Delta X$
7: **until** (17a) is **true** *# end of step 2)*
8: **until** (17b) is **true**

Our implementation of alg. 1 is in the MatLab programming language. The source code is available on-line and includes a test function for a quick demonstration. Interested readers must download the files from [25] into their working directories, run the initialization script (`init.m`)[13] and type `TestRNLE()` in the command prompt. Documentation and details of our implementation may be found by typing `help RNLE` and/or by examining the comments in the source code.

5 Experimental Results

5.1 Multi-variate stochastic volatility models

Multi-variate Stochastic Volatility (MSV) models track the price variations of a group of financial assets. The variation is treated as a random process governed by state variables, and thus an MSV model can be expressed as an SSM.

One such MSV model advocated by Tsay [39] is written in state-space form as

$$\mathbf{x}_k = \phi \circ (\mathbf{x}_{k-1} - \mu) + \mu + \mathbf{\Lambda} \operatorname{diag}(\sigma/2)\, \mathbf{u}_k \tag{18a}$$

$$\mathbf{y}_k = \mathbf{\Gamma} \operatorname{diag}(\exp(\mathbf{x}_k/2))\, \mathbf{v}_k \tag{18b}$$

where $\circ$ denotes the Hadamard, or element-wise product and $\exp$ the element-wise exponential function. Vector $\mathbf{x}_k$ contains the latent log-volatilities, and vector $\mathbf{y}_k$ contains the percentage rate of return for each asset for the kth trading day. Vectors $\mathbf{u}_k$ and $\mathbf{v}_k$ are noise processes as defined in eqs. (3a) and (3b), respectively.

The transition model in eq. (18a) captures the tendency of the states (log-volatilities) to revert to a stationary value (μ) in the long run. The observation model in eq. (18b) asserts that the variation of the measurements (returns) is a non-linear function of the states. Note that eq. (18b) implies

$$\mathbf{h}_k(\mathbf{x}_k) = \mathbf{0} \qquad \mathbf{R}_k(\mathbf{x}_k) = \mathbf{\Gamma} \operatorname{diag}(\exp(\mathbf{x}_k))\, \mathbf{\Gamma}^\top$$

that is, the only link between states and measurements is the state-dependent measurement noise.

Table 1 summarizes the parameters of this MSV model. The last column lists the distributions used to generate random parameters during the experiment. The notation $\mathcal{U}(a, b)$ denotes a uniform distribution in the interval $[a, b]$. Matrices $\mathbf{\Lambda}$ and $\mathbf{\Gamma}$ are lower-triangular with unit diagonal. The number of states is equal to the number of measurements, i.e. $d = d_k$ for all $k = 1, \ldots, n$.

Table 1: Summary of model parameters

Symbol	Shape	Sampling distribution
μ	$d \times 1$	$\mu_i \sim \mathcal{U}(-2, -1)$
ϕ	$d \times 1$	$\phi_i \sim \mathcal{U}(0.95, 1)$
σ	$d \times 1$	$\sigma_i \sim \mathcal{U}(0.05, 0.15)$
$\mathbf{\Lambda}$	$d \times d$	$\lambda_{ij} \sim \mathcal{U}(-0.10, 0.95),\ i > j$
$\mathbf{\Gamma}$	$d \times d$	$\gamma_{ij} \sim \mathcal{U}(-0.15, 0.95),\ i > j$

The percentage rate of return is usually assumed conditionally Gaussian [6]. In this paper $\mathbf{y}_k$ is t-distributed with $s_k = \nu$ degrees of freedom. This accounts for the higher kurtosis [40] typical of daily price variations.

5.2 Benchmarks

For the purpose of validation, we compared our Robust Non-Linear Estimator (RNLE) to two other estimators:

- The Sampling-Importance-Resampling Particle Smoother (SIR-PS) [41], a sequential sampling algorithm that splits the state-space into a finite number of particles; and
- A block component-wise Metropolis-Hastings Sampler (MHS) [42], a Markov chain Monte Carlo Method that exploits the sequential nature of the data.

[13] The initialization script compiles two C files and creates corresponding MatLab executable (MEx) files.

Both of these estimators are asymptotically optimal, i.e. they produce increasingly better estimates as the number of particles grows infinitely large. Hence they serve as a benchmark for the method we propose.

We point out that the methods mentioned in sub-section 1.4 are not applicable to this problem. Linearization algorithms such as the Extended Kalman Filter/Smoother (EKF/S) [43] and their iterated cousins [24] only linearize the observation function, not the noise. Deterministic sampling algorithms such as the Unscented Kalman Filter (UKF) [15] and its retrospective counterpart, the Unscented Rauch-Tung-Striebel Smoother (URTSS) [16], use too few samples to reason about variance-covariance parameters.[14] Perhaps the method of Spinello and Stilwell [13] would be a viable candidate, were it not for the fact that their algorithm is designed for filtering and that they assume Gaussian noise.

5.3 Performance metrics

In order to evaluate and compare the performance of different estimation algorithms we selected a set of error metrics. Namely, we chose the Root Mean Squared (RMS), Maximum (Max), Mean Absolute (MAbs) and Maximum Absolute (MaxAbs) errors, given by:

$$\begin{aligned}
\text{RMS} &= \sqrt{\frac{1}{n}\sum_{k=1}^{n} \|\mathbf{x}_k - \hat{\mathbf{x}}_k\|^2} \\
\text{Max} &= \max_{k=1,\ldots,n} \|\mathbf{x}_k - \hat{\mathbf{x}}_k\| \\
\text{MAbs} &= \frac{1}{n}\sum_{k=1}^{n} |\mathbf{x}_k - \hat{\mathbf{x}}_k| \\
\text{MaxAbs} &= \max_{k=1,\ldots,n} |\mathbf{x}_k - \hat{\mathbf{x}}_k|
\end{aligned}$$

where $\hat{\mathbf{x}}_k$ is an estimate of the true state $\mathbf{x}_k$ at step k and $|\cdot|$ is the Manhattan norm.

For the SIR-PS, $\hat{\mathbf{x}}_k$ is computed as the weighted average of the particles for the kth day. For the MHS, it is the average of the kth element over samples.

5.4 Experimental setup

Our experiment consisted of repeating the following steps a total of 50 times:

1. Sample a set of parameters according to table 1;
2. Generate a sequence of 200 data from this model;
3. Run the SIR-PS, the RNLE (algorithm 1) and the MHS on this data with full knowledge of the model; and
4. Compute the error metrics for each estimator.

The SIR-PS was run with 100 particles. Its proposal distribution was the same as the state transition model.

The initial guess of the state sequence for the RNLE was obtained with a windowed filter of width $2m + 1$. Specifically, we initialized each component of the state vector as follows:

$$\hat{x}_k = \ln \frac{1}{2m+1} \sum_{i=k-m}^{k+m} \left[y_i - \frac{1}{2m+1} \sum_{j=k-m}^{k+m} y_j \right]^2$$

for each k, $m < k \leq n - m$. In our experiment $m = 3$.

The MHS was initialized with the same initial guess. We ran it for 1000 burn-in iterations and then 20000 more iterations, thinning by a factor of 20. This produced a total of 100 samples from the posterior distribution over state sequences. We hand-tuned the parameters of the proposal distribution to achieve an average acceptance rate of 20%–40%.

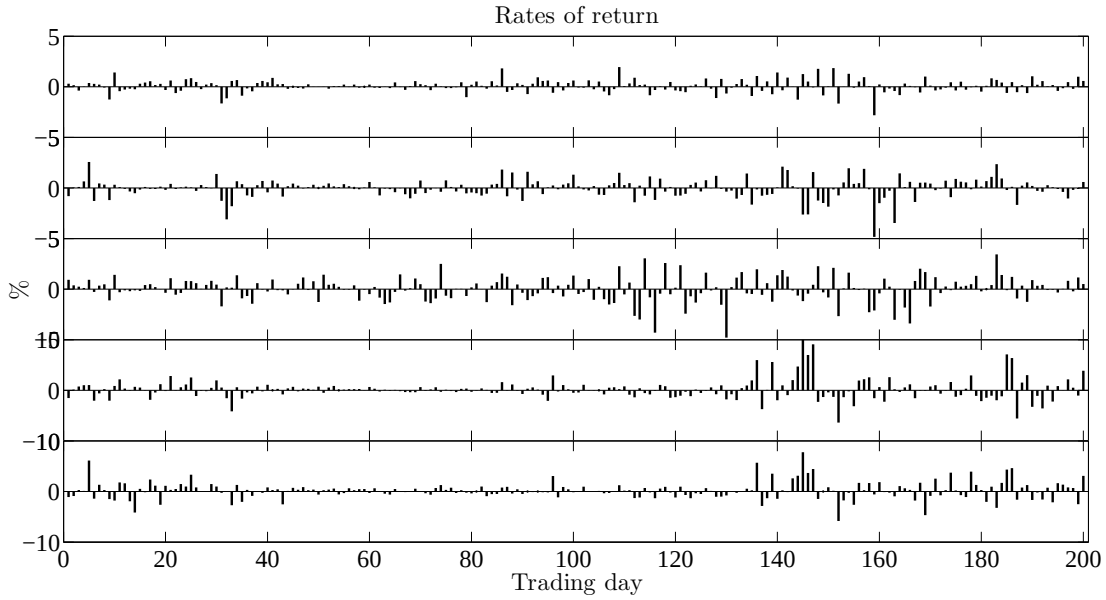

Figure 5: Typical sequences of data generated by the multi-variate stochastic volatility model in eqs. (18)**. The vertical bars denote the percentage rate of return of a given asset, i.e. the percent change in its price between consecutive trading days.**

5.5 Results

Fig. 5 shows a typical sequence of measurements generated by the MSV model in eqs. (18). Note that this is not a standard signal processing problem where we wish to recover a signal buried in noise. What we wish to estimate is the instantaneous variation in the signal (its envelope, so to speak).

Fig. 6 shows the sequence of log-volatilities of the 3rd asset (3rd row in fig. 5), plus the estimates. For the SIR-PS (top), the estimates are depicted as particle clouds, where each particle's weight is proportionally mapped to a different tone of gray. For the RNLE (middle), the estimates are 99% confidence intervals derived from the variance-covariance parameters computed during the RTS recursions. For the MHS, the estimates are samples in the ensemble.

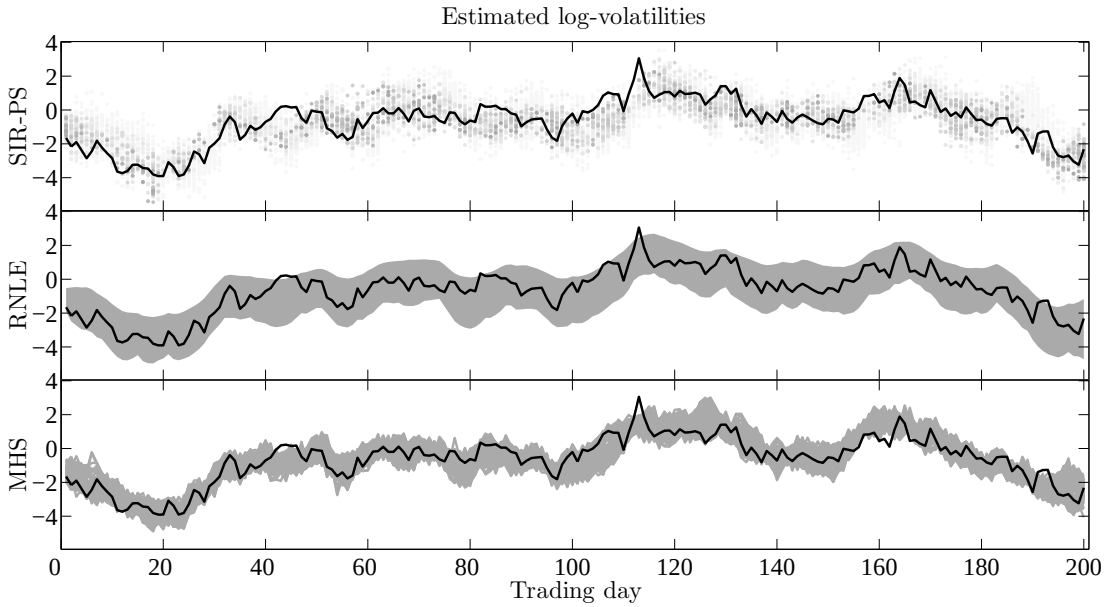

Figure 6: Typical sequence of state estimates returned by the three estimation algorithms. The estimated log-volatilities of a given asset, as estimated by the SIR-PS (top), RNLE (middle) and MHS (bottom), are plotted in gray. The black line depicts the true log-volatilities.

Table 2 and fig. 7 summarize the error metrics that each algorithm scored in our experiment. The last

[14]The unscented transform propagates $2d + 1$ points, while a covariance matrix contains $d(d-1)/2$ free parameters.

row in the table lists the execution time per sequence in seconds. Statistics were collected over the 50 runs. All differences are statistically significant at the 0.5% level.

Table 2: Summary of performance metrics

Metric	SIR-PS		RNLE		MHS	
	Mean	Std. dev.	Mean	Std. dev.	Mean	Std. dev.
RMS	1.47	0.0928	1.22	0.0855	1.05	0.0697
Max	3.04	0.443	2.57	0.346	2.20	0.200
MAbs	2.60	0.152	2.15	0.143	1.88	0.125
MaxAbs	6.08	0.956	5.21	0.764	4.34	0.467
Exec. time	77.6	1.49	1.05	0.0823	144	1.56

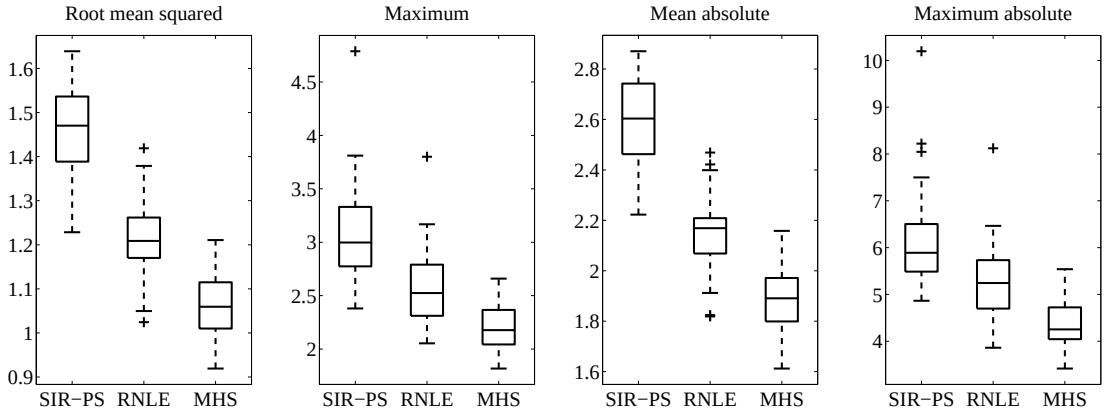

Figure 7: Box plots of the error metrics achieved by the SIR-PS, RNLE and MHS. The sample statistics are computed over the 50 runs of the experiment.

5.6 Comparison and discussion

In terms of estimation accuracy, the RNLE is almost on par with the MHS (e.g. its RMS error is 14% larger on average). Assuming that the sample size is sufficiently large (100 in our experiment), we can regard the MHS as the optimal estimator and conclude that, for this experiment, the RNLE performed close to the best possible estimator.

The advantage of the RNLE is its running time. With an average of 1.05 ± 0.135 sec. per sequence, it is more than 120 times faster than the MHS and over 70 times faster than the SIR-PS, which took an average of 144 ± 2.56 and 77.6 ± 2.45 seconds per sequence, respectively.

The RNLE outperformed the SIR-PS, although this is due to the relatively small particle cloud. Increasing the number of particles beyond 100 tends to decrease the errors, albeit at a computational cost that increases quadratically and thus renders it impractical compared to the MHS.

6 Summary and Conclusions

In this paper we proposed a robust estimation algorithm for non-linear state-space models driven by state-dependent noise. We derived the algorithm from first principles as an iterative solver for a quadratic-composite minimization problem. We tested it on simulated data and showed that its performance is comparable to that of the optimal Bayes estimator.

We believe that the RNLE has great potential for complex tasks with outliers and missing data. To the best of the authors' knowledge, no other deterministic algorithm in the literature has these capabilities.

It would be interesting to extend our approach to allow for more general distributional assumptions. For instance, elliptical distributions [44] are attractive because of their generality and their compact parametrization.[15] One could imagine a model parameterized by a density generator function that determines the shape of the observation density.

[15] In fact, the t is an elliptical distribution with density generator function $g : x \mapsto (1 + x/\nu)^{-(\nu+d)/2}$ for $x \geq 0$.

In our experiment we relied on a filter to provide us with an initial guess of the state sequence. There are many other, more general initialization schemes, e.g. the EKF/S, the UKF etc. Studying and assessing their relative merits would be an interesting direction to pursue in the future.

Acknlowledgements

This work is supported in part by the Australian Research Council (ARC) under the linkage grant LP120100700.

A Derivatives of the Quadratic-composite Function

The derivatives of b —defined in eq. (11b)—with respect to $\mathbf{x}_k$, evaluated at X, are given by

$$\begin{aligned}\frac{\partial}{\partial \mathbf{x}_k} b\left(X, q\left(W\right)\right) = \mathbf{Q}_k^{\frac{1}{2}}\left(\mathbf{x}_{k-1}\right)^{-\top} \mathbf{u}_k(\mathbf{x}_{k-1}, \mathbf{x}_k) \\ - \mathbf{G}_{k+1}(\mathbf{x}_k, \mathbf{x}_{k+1})^\top \mathbf{Q}_{k+1}^{\frac{1}{2}}(\mathbf{x}_k)^{-\top} \mathbf{u}_{k+1}(\mathbf{x}_k, \mathbf{x}_{k+1}) \\ - \omega_k \mathbf{H}_k(\mathbf{x}_k)^\top \mathbf{R}_k^{\frac{1}{2}}(\mathbf{x}_k)^{-\top} \mathbf{v}_k(\mathbf{x}_k) + \mathbf{q}_{k+1}(\mathbf{x}_k) + \mathbf{r}_k(\mathbf{x}_k)\end{aligned}$$

The gradient, ∇b, of b with respect to X is a $d \times n$ matrix formed by concatenating these partial derivatives.

References

[1] S. Thrun, Y. Liu, D. Koller, A. Ng, Z. Ghahramani, and H. Durrant-Whyte, "Simultaneous localization and mapping with sparse extended information filters," *The International Jounal of Robotics Research*, vol. 23, no. 7–8, pp. 693–716, July–August 2004.

[2] F. Dellaert and M. Kaess, "Square root SAM: Simultaneous localization and mapping via square root information smoothing," *The International Journal of Robotics Research*, vol. 25, no. 12, pp. 1181–1203, December 2006.

[3] X. Li and V. Jilkov, "Survey of maneuvering target tracking. part i: Dynamic models," *IEEE Transactions on Aerospace and Electronic Systems*, vol. 39, no. 4, pp. 1333–1364, October 2003.

[4] J. Loxam and T. Drummond, "Student t mixture filter for robust, real-time visual tracking," in *Proceedings of the 10th European Conference on Computer Vision: Part III*, 2008.

[5] J. Stroud, P. Müller, and N. Polson, "Non-linear state-space models with state-dependent variances," *Journal of the American Statistical Association*, vol. 98, no. 462, pp. 377–386, June 2003.

[6] M. Asai, M. McAleer, and J. Yu, "Multivariate stochastic volatility: A review," *Econometric Reviews*, vol. 25, no. 2–3, pp. 145–175, 2006.

[7] D. Moore and G. McCabe, *Introduction to the Practice of Statistics*. W.H. Freeman, 1993.

[8] V. Barnett and T. Lewis, *Outliers in Statistical Data*. John Wiley & Sons, 1994.

[9] T. Bailey and H. Durrant-Whyte, "Simultaneous localization and mapping (SLAM): Part II," *IEEE Robotics and Automation Magazine*, vol. 13, no. 3, pp. 108–117, September 2006.

[10] J. Ting, A. D'Souza, and S. Schaal, "Automatic outlier detection: A Bayesian approach," in *Proceedings of the IEEE International Conference on Robotics and Automation*, 2007.

[11] A. Gadre, M. Roan, and D. Stilwell, "Sensor error model for a uniform linear array," Virginia Center for Autonomous Systems, Tech. Rep. VaCAS-2008-01, May 2008.

[12] S. Saha and F. Gustaffson, "Particle filtering with dependent noise processes," *IEEE Transactions on Signal Processing*, vol. 60, no. 9, pp. 4497–4508, September 2012.

[13] D. Spinello and D. Stilwell, "Non-linear estimation with state-dependent gaussian observation noise," *IEEE Transactions on Automatic Control*, vol. 55, no. 6, pp. 1358–1366, June 2010.

[14] V. Peltola and A. Honkela, "Variational inference and learning for non-linear state-space models with state-dependent observation noise," in *IEEE International Workshop on Machine Learning for Signal Processing*, 2010, pp. 190–195.

[15] S. Julier and J. Uhlmann, "A new extension of the Kalman filter to non-linear systems," in *International Symposium on Aerospace and Defense Sensing, Simulation and Control*, 1997.

[16] S. Särkkä, "Unscented Rauch–Tung–Striebel smoother," *IEEE Transactions on Automatic Control*, vol. 53, no. 3, pp. 845–849, April 2008.

[17] S. Särkkä and A. Nummenmaa, "Recursive noise adaptive Kalman filtering by variational Bayesian approximations," *IEEE Transactions on Automatic Control*, vol. 54, no. 3, pp. 596–600, March 2009.

[18] G. Agamennoni, J. Nieto, and E. Nebot, "An outlier-robust Kalman filter," in *Proceedings of the 2011 IEEE International Conference on Robotics and Automation*, 2011.

[19] G. Agamennoni, J. I. Nieto, and E. Nebot, "Approximate inference in state-space models with heavy-tailed noise," *IEEE Transactions on Signal Processing*, vol. 60, no. 10, pp. 5024–5037, October 2012.

[20] A. Aravkin, B. Bell, J. Burke, and G. Pillonetto, "An ℓ_1-Laplace robust Kalman smoother," *IEEE Transactions on Automatic Control*, vol. 56, no. 12, pp. 2898–2911, December 2011.

[21] A. Aravkin, J. Burke, and G. Pillonetto, "Optimization viewpoint on Kalman smoothing, with applications to robust and sparse estimation," March 2013, pre-print available from arXiv:1303.1993.

[22] ——, "Robust and trend-following Student's t Kalman smoothers," March 2013, pre-print available from arXiv:1303.5588.

[23] R. Piché, S. Särkkä, and J. Hartikainen, "Recursive outlier-robust filtering and smoothing for non-linear systems using the multi-variate student-t distribution," in *Proceedings of the IEEE International Conference on Machine Learning for Signal Processing*, 2012.

[24] B. Bell and F. Cathey, "The iterated Kalman filter update as a Gauss-Newton method," *IEEE Transactions on Automatic Control*, vol. 38, no. 2, pp. 294–297, February 1993.

[25] G. Agamennoni. Community profile at MathWorks. [Online]. Available: www.mathworks.com/matlabcentral/fileexchange/authors/52834

[26] S. Kotz and S. Nadarajah, *Multivariate t Distributions and their Applications*. Cambridge University Press, 2004.

[27] F. Hampel, E. Ronchetti, P. Rousseeuw, and W. Stahel, *Robust Statistics: The Approach Based on Influence Functions*. John Wiley & Sons, March 1986.

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

[29] A. O'Hagan, "A moment of indecision," *Biometrika*, vol. 68, no. 1, pp. 329–330, 1981.

[30] R. Meinhold and N. Singpurwalla, "Robustification of Kalman filter models," *Journal of the American Statistical Association*, vol. 84, no. 406, pp. 479–486, June 1989.

[31] J. Drèze, "Bayesian regression analysis using poly-t densities," *Journal of Econometrics*, vol. 6, no. 3, pp. 329–354, November 1977.

[32] J. Burke and M. Ferris, "A Gauss-Newton method for convex composite optimization," Department of Mathematics, University of Washington and Department of Computer Sciences, University of Wisconsin, Tech. Rep., 1993.

[33] J. Jensen, "Sur les fonctions convexes et les inégalités entre les valeurs moyennes," *Acta Mathematica*, vol. 30, no. 1, pp. 175–193, December 1906.

[34] M. Abramowitz and I. Stegun, *Handbook of Mathematical Functions, with Formulas, Graphs and Mathematical Tables*. Dover Publications, Incorporated, 1974.

[35] S. Kullback and R. Leibler, "On information and sufficiency," *Annals of Mathematical Statistics*, vol. 22, no. 1, pp. 79–86, 1951.

[36] S. Smith, "Differentiation of the cholesky algorithm," *Journal of Computational and Graphical Statistics*, vol. 4, no. 2, pp. 134–147, June 1995.

[37] H. E. Rauch, C. T. Striebel, and F. Tung, "Maximum likelihood estimates of linear dynamic systems," *American Institute of Aeronautics and Astronautics Journal*, vol. 3, no. 8, pp. 1445–1450, 1965.

[38] J. Nocedal and S. Wright, *Numerical Optimization*, P. Glynn and S. Robinson, Eds. Springer, 1999.

[39] R. Tsay, *Analysis of Financial Time Series: Financial Econometrics*. John Wiley & Sons, 2002.

[40] A. Harvey, E. Ruiz, and N. Shephard, "Multivariate stochastic variance models," *The Review of Economic Studies*, vol. 61, no. 2, pp. 247–264, April 1994.

[41] A. Doucet and A. Johansen, "A tutorial on particle filtering and smoothing: Fifteen years later," in *Oxford Handbook of Non-linear Filtering*. Oxford University Press, 2009.

[42] R. Levine, Z. Yu, W. Hanley, and J. Nitao, "Implementing component-wise Hastings algorithms," *Computational Statistics & Data Analysis*, vol. 48, no. 2, pp. 363–389, February 2005.

[43] D. Simon, *Optimal State Estimation*. John Wiley & Sons, 2006.

[44] K.-T. Fang, S. Kotz, and K. Ng, *Symmetric Multi-variate and Related Distributions*. Chapman & Hall, 1987.