2011 IEEE International Conference on Robotics and Automation
Shanghai International Conference Center
May 9-13, 2011, Shanghai, China

An Outlier-Robust Kalman Filter

Gabriel Agamennoni, Juan I. Nieto and Eduardo M. Nebot

Abstract— **We introduce a novel approach for processing sequential data in the presence of outliers. The outlier-robust Kalman filter we propose is a discrete-time model for sequential data corrupted with non-Gaussian and heavy-tailed noise. We present efficient filtering and smoothing algorithms which are straightforward modifications of the standard Kalman filter Rauch-Tung-Striebel recursions and yet are much more robust to outliers and anomalous observations. Additionally, we present an algorithm for learning all of the parameters of our outlier-robust Kalman filter in a completely unsupervised manner. The potential of our approach is borne out in experiments with synthetic and real data.**

I. Introduction

Sequential data arises in many domains of robotics. It arises in sensor fusion and navigation problems as a sequence of GPS and inertial measurements, in point tracking applications as a stream of radar or laser returns or in autonomous mapping scenarios as a sequence of noisy relative position measurements. Stochastic models for filtering sequential data are therefore of considerable interest.

The Kalman filter is often regarded as the predecessor of modern filtering systems [1], [2]. It is the optimal estimator for linear-Gaussian dynamical systems [3] in the sense that it yields the smallest expected mean-squared error. However, its performance breaks down in the presence of non-Gaussian noise. The squared error criterion is very sensitive to outlying measurements [4] and yields very poor estimates for systems with transient disturbances. The problem lies in the lightweight tails of the Gaussian distribution, due to which the Kalman filter effectively rules out the idea that any measurement is ever an outlier.

An outlier is an observation that lies outside an overall pattern of distribution [5]. An outlying observation is numerically distant from other members of the sample in which it occurs. Although it may occur by chance in any distribution, it often stems from unmodeled factors or anomalous causes. For instance, outliers may originate from transient environmental disturbances, temporary sensor failure, erroneous measurements or from noise characteristics that are intrinsic to the sensor.

There are many sensors which have well-defined noise characteristics. For example, potentiometers, inertial measurement units and optical encoders can be suitably modeled with additive white Gaussian noise. However, other sensors such as visual tracking systems, GPS receivers, sonars and radars are not as easily interpretable. The track returned by a computer vision system is often infested with outliers. The position estimated by a GPS receiver frequently exhibits large drifts caused by multi-path reflections. Sonar and radar data are strongly affected by speckle and phase noise, which is usually far from Gaussian.

Australian Centre for Field Robotics, The University of Sydney, NSW (2006), Australia, e-mail: g.agamennoni, j.nieto, e.nebot @acfr.usyd.edu.au.

II. Prior work

There have been a number of approaches aimed at improving the robustness of the Kalman filter in the face of outliers and non-Gaussian noise. Researchers have proposed many alternative noise models, usually in the form of heavy-tailed distributions or ad-hoc cost functions. Unfortunately, any distribution other than the Gaussian renders the filter analytically intractable and requires some form of approximation. A few authors have proposed the use of heavy-tailed non-Gaussian [6] or t-distributed [7], [8], [9] noise models, although their treatment is purely one-dimensional and, due to their methodology, it is not extendable into higher-dimensional space. Others have developed ad-hoc cost functions [10], [11], [12] for the update step, which are fairly involved and require hand-tuning of one or more parameters. Furthermore, all these approaches are substantially more difficult to implement than the original Kalman filter.

Other approaches, which are more readily implemented, draw on numerical techniques. Some authors have proposed statistical sampling [13], [14] or numerical integration methods [15] in order to perform robust filtering. However, their high computational demand often makes them unsuitable for real-time applications and may even be completely prohibitive for systems with a high-dimensional state space. In addition, assessing the convergence of such methods is usually nontrivial. Other nonparametric methods [16], [17], [18], [19] bypass the constraints of analytical tractability by representing the hidden state distribution as a finite set of samples, or particles. These methods also suffer from the curse of dimensionality, as the number of particles scales exponentially with the dimension of the state space.

Some recent work in the literature has shown promising results. The authors of [20], [21] introduced a modified Kalman filter capable of performing robust real-time tracking in the presence of measurement outliers. Their model does not require any manual tuning and allows for fully unsupervised parameter learning. The authors successfully apply their method to the task of filtering noisy motion capture data from a mobile robotic platform. A similar approach was reported in [22]. Here, the authors additionally define a heuristic transition model for the noise parameters in order to simultaneously track the system state as well as the level of sensor noise. Simulation results are given but no results were shown with real-world data.

978-1-61284-385-8/11/$26.00 ©2011 IEEE

These recent proposals suffer from two important limitations. First, they are both aimed at real-time applications and forsake problems such as fixed-interval smoothing. State smoothing has many relevant applications in robotics and thus is an interesting problem in its own right. The second limitation relates to implicit assumptions made by the authors about the noise distribution. In [20], [21] it is assumed that noise characteristics are homogeneous across all measurements. In practice, sequential data is often comprised of multiple physically different sensors. The data from these sensors is concatenated together and fed to the filter as a single stream of observations. Since the sensors have distinct properties, their noise characteristics are often considerably different and may shift independently from one another. In [22] a more flexible measurement noise model is assumed. However, it still fails to model correlations amongst observations and hence it is also discarding valuable information that may be beneficial for the filtering problem.

In this manuscript we generalize and extend prior work by [20], [21] and [22]. In Section III we define an outlier-robust Kalman filter that inherits the advantages of both approaches and encompasses them as special cases. In Section IV we derive filtering and smoothing algorithms which are much more robust yet only slightly more involved than the standard inference algorithms for the Kalman filter. We present an algorithm for learning the parameters of the model in Section V. Finally, in Section VI we demonstrate the potential of our approach with synthetic data and with data collected by a mobile sensing platform. Conclusions and further research directions are outlined in Section VII.

III. Model

A. The outlier-robust Kalman filter

The outlier-robust Kalman filter (ORKF) we propose is similar to the well-known Kalman filter (KF) in the sense that it too is a generative model for sequential data. A sequence of observations is assumed to be generated by a corresponding sequence of hidden states. The hidden state sequence is a stochastic process that follows first-order autoregressive dynamics, whereas the observations are conditionally independent from each other given all hidden variables.

The novelty of the ORKF lies in its noise model. The classical KF assumes that observations are corrupted with additive white Gaussian noise. The level of noise is constant and is encoded via the sensor covariance matrix. Our version also assumes additive noise. However, we do not restrict the level of noise to be neither constant nor Gaussian. Instead, it is allowed to vary over time and may have heavier tails than the Gaussian distribution. This endows the model with much more flexibility when it comes to explaining outliers in the data, at the cost of only one additional parameter.

B. A generative model

The ORKF model can be regarded as a probabilistic generative model. It is shown as a dynamic Bayesian network [23] (DBN) in Fig. 1. Round nodes represent random variables and directed edges encode stochastic relationships between them. White nodes depict hidden or unobserved variables while grey nodes correspond to observed data points. Directed edges represent causal links between pairs of nodes by means of conditional probability distributions.

The set of all conditional probability distributions associated with the edges provides a complete specification of the model. Let $\mathbf{y}_1, \ldots, \mathbf{y}_n$ be a finite sequence of observations generated by a corresponding sequence $\mathbf{x}_1, \ldots, \mathbf{x}_n$ of hidden states. Denote the covariance of the observation noise at time t by $\mathbf{S}_t$. Then,

$$\mathbf{x}_t|\mathbf{x}_{t-1} \sim \mathcal{N}\left(\mathbf{A}^T\mathbf{x}_{t-1} + \mathbf{b}, \mathbf{Q}\right), \quad t > 1, \tag{1}$$

$$\mathbf{y}_t|\mathbf{x}_t, \mathbf{S}_t \sim \mathcal{N}\left(\mathbf{C}^T\mathbf{x}_t + \mathbf{d}, \mathbf{S}_t\right), \tag{2}$$

where $\mathcal{N}(\mu, \mathbf{\Sigma})$ denotes a multivariate Gaussian distribution with mean μ and covariance $\mathbf{\Sigma}$. Both (1) and (2) resemble the traditional probabilistic KF model. The key difference is that the covariance $\mathbf{S}_t$ of the observation noise is no longer fixed a priori.

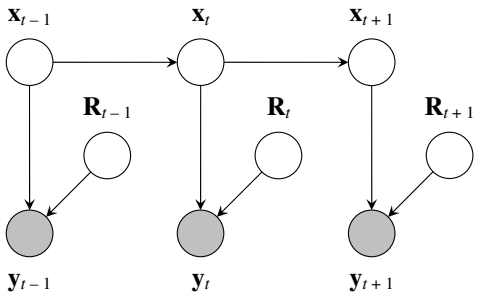

Fig. 1. The ORKF as a 3-slice DBN.

The observation noise is treated as a hidden variable and hence is non-constant. Instead, it is sampled from a probability distribution at each time step. Namely,

$$\mathbf{S}_t^{-1} \sim \mathcal{W}\left(\mathbf{R}^{-1}/s, s\right), \quad t > 1, \tag{3}$$

where $\mathcal{W}(\mathbf{\Lambda}, \nu)$ denotes a Wishart distribution [24] with $d \times d$ precision matrix $\mathbf{\Lambda} \succ \mathbf{0}$ and $\nu > d - 1$ degrees of freedom. The mean of the distribution is $\nu\mathbf{\Lambda}$, and ν quantifies how tightly the distribution is concentrated around its mean.

The reader should take a moment to examine (2) and (3). These equations describe the process via which observations are generated from the underlying state sequence. If we multiply both of them, marginalize over $\mathbf{S}_t$ and apply the matrix determinant lemma, we obtain

$$\begin{aligned} p\left(\mathbf{y}_t|\mathbf{x}_t\right) &= \int_{\mathbf{S}_t \succ \mathbf{0}} p\left(\mathbf{y}_t|\mathbf{x}_t, \mathbf{S}_t\right) p\left(\mathbf{S}_t\right) \mathrm{d}\mathbf{S}_t \\ &\propto \left(1 + \delta_t^T \mathbf{R}^{-1} \delta_t / s\right)^{-\frac{s+1}{2}}, \end{aligned} \tag{4}$$

where we have defined

$$\delta_t = \mathbf{y}_t - \mathbf{C}^T\mathbf{x}_t - \mathbf{d}.$$

An observation $\mathbf{y}_t$ is thus Student t-distributed [25] given its corresponding hidden state $\mathbf{x}_t$.

The Student-t is a sub-exponential distribution which has much heavier tails than the Gaussian. It spreads its probability mass more evenly across observation space and further

away from the mode, assigning outliers a non-negligible probability. By defining the noise model as in (2) and (3), outliers need not be explicitly pre-filtered or treated as a special case. Instead, they will be taken care of naturally within the Bayesian filtering framework.

IV. Inference

A. Approximate filtering

Inference is the process of estimating the posterior distribution over hidden states, given a sequence of observations. Within the context of sequential data models, filtering consists on finding the posterior given all data up to and including the current time step. The filtered posterior distribution is defined as

$$\alpha\left(\mathbf{x}_t\right) \triangleq p\left(\mathbf{x}_t|\mathbf{y}_1,\ldots,\mathbf{y}_t\right). \tag{5}$$

In the standard KF, α is Gaussian and can be evaluated in closed form. This no longer holds for the robust version presented here. The non-Gaussian noise model renders the filtered posterior analytically intractable. Nevertheless, a tractable approximation can still be derived by taking a structured variational [26] approach.

Bayesian filtering is a well-known estimation problem. The exact posterior in (5) can be evaluated, at least in theory, by alternating between the *prediction* step and the *update* step. The familiar recursive filtering equation

$$\alpha\left(\mathbf{x}_t\right) \propto p\left(\mathbf{y}_t|\mathbf{x}_t\right) \underbrace{\int p\left(\mathbf{x}_t|\mathbf{x}_{t-1}\right)\alpha\left(\mathbf{x}_{t-1}\right)\mathrm{d}\mathbf{x}_{t-1}}_{p(\mathbf{x}_t|\mathbf{y}_1,\ldots,\mathbf{y}_{t-1})} \tag{6}$$

gives the exact solution to the filtering problem for the ORKF. The prediction step consists of convoluting $\alpha(\mathbf{x}_{t-1})$ with $p(\mathbf{x}_t|\mathbf{x}_{t-1})$, the transition likelihood function, whereas the update step consists of multiplying the result by $p(\mathbf{y}_t|\mathbf{x}_t)$, the conditional data likelihood function.

The prediction step is straightforward. Assume that, up to time $t-1$, the filtered posterior $\alpha(\mathbf{x}_{t-1})$ is a Gaussian density. That is, let

$$\mathbf{x}_{t-1}|\mathbf{y}_1,\ldots,\mathbf{y}_{t-1} \sim \mathcal{N}\left(\mu_{t-1},\boldsymbol{\Sigma}_{t-1}\right).$$

Then, the integral in the right-hand side of (6) can be readily solved to give

$$\mathbf{x}_t|\mathbf{y}_1,\ldots,\mathbf{y}_{t-1} \sim \mathcal{N}\left(\mathbf{m}_{t-1},\mathbf{P}_{t-1}\right),$$

where the predictive mean and covariance

$$\mathbf{m}_{t-1} = \mathbf{A}^T\mu_{t-1} + \mathbf{b}, \tag{7}$$

$$\mathbf{P}_{t-1} = \mathbf{A}^T\boldsymbol{\Sigma}_{t-1}\mathbf{A} + \mathbf{Q}, \tag{8}$$

are identical to the standard KF case.

The update step is slightly more involved. Ideally, the updated posterior should be a Gaussian density. This way, we could apply (7) and (8), replacing $t-1$ by t, and the prediction-update cycle would repeat itself. Unfortunately, (4) shows that $\mathbf{y}_t|\mathbf{x}_t$ is Student t-distributed, and this translates into $\alpha(\mathbf{x}_t)$ being non-Gaussian. However, we can still approximate it as a Gaussian in some optimal sense. Particularly, the relative entropy or Kullback-Leibler (KL) divergence [27] is a natural distance metric between probability distributions. Hence we select an approximate Gaussian posterior so as to minimize its KL divergence with respect to the true posterior.

B. The update step

Let $\hat{\alpha}$ be the approximate filtered posterior. We set $\hat{\alpha}(\mathbf{x}_t)$ to a Gaussian probability density function with mean μ_t and covariance $\boldsymbol{\Sigma}_t$ such that

$$\mathrm{KL}\left(\hat{\alpha}\|\alpha\right) = \int \hat{\alpha}\left(\mathbf{x}_t\right)\ln\frac{\hat{\alpha}\left(\mathbf{x}_t\right)}{\alpha\left(\mathbf{x}_t\right)}\mathrm{d}\mathbf{x}_t$$

is minimal. Minimizing this divergence equates to maximizing a local variational lower bound on $\ln p(\mathbf{y}_t)$, the marginal data log-likelihood. The bound attains its maximum when

$$\ln\hat{\alpha}\left(\mathbf{x}_t\right) = \left\langle\ln p\left(\mathbf{y}_t|\mathbf{x}_t,\mathbf{S}_t\right)\right\rangle + \ln p\left(\mathbf{x}_t|\mathbf{y}_1,\ldots,\mathbf{y}_{t-1}\right) + \ldots$$

with $\langle\cdot\rangle$ denoting expectation with respect to the posterior distribution over $\mathbf{S}_t$.

Algebraic manipulation of the above expression gives the desired result. Define $\boldsymbol{\Gamma}_t^{-1} = \langle\mathbf{S}_t^{-1}\rangle$ and define the Kalman gain matrix

$$\mathbf{K}_t = \left(\mathbf{C}^T\mathbf{P}_{t-1}\mathbf{C} + \boldsymbol{\Gamma}_t\right)^{-1}\mathbf{C}^T\mathbf{P}_{t-1}.$$

Then, the optimal values for the posterior mean and covariance are given by

$$\mu_t = \mathbf{m}_{t-1} + \mathbf{K}_t^T\left(\mathbf{y}_t - \mathbf{C}^T\mathbf{m}_{t-1}\right), \tag{9}$$

$$\boldsymbol{\Sigma}_t = \mathbf{P}_{t-1}\left(\mathbf{I} - \mathbf{C}\mathbf{K}_t\right). \tag{10}$$

Notice that these equations are almost identical to the standard KF update equations. The only difference is that the observation noise covariance $\mathbf{R}$ in the Kalman gain matrix is replaced by $\boldsymbol{\Gamma}_t$, its expected value.

It remains to derive an expression for the $\mathbf{S}_t$ posterior. Maximization of the variational lower bound reveals that $\mathbf{S}_t^{-1}$ is Wishart-distributed,

$$\mathbf{S}_t^{-1}|\mathbf{y}_1,\ldots,\mathbf{y}_t \sim \mathcal{W}\left(\boldsymbol{\Gamma}_t^{-1}/\nu_t,\nu_t\right),$$

with $\nu_t = s+1$ degrees of freedom and inverse precision

$$\boldsymbol{\Gamma}_t = \frac{s\mathbf{R} + \left\langle\delta_t\delta_t^T\right\rangle}{s+1}, \tag{11}$$

where $\langle\cdot\rangle$ denotes expectation with respect to $\hat{\alpha}$. Taking into account that $\hat{\alpha}$ is Gaussian, it is easy to show that

$$\left\langle\delta_t\delta_t^T\right\rangle = \left(\mathbf{y}_t - \mathbf{C}^T\mu_t - \mathbf{d}\right)\left(\mathbf{y}_t - \mathbf{C}^T\mu_t - \mathbf{d}\right)^T + \mathbf{C}^T\boldsymbol{\Sigma}_t\mathbf{C}.$$

Notice that $\boldsymbol{\Gamma}_t$ is a convex combination of the nominal noise covariance $\mathbf{R}$ and the expected outer product matrix. In addition, $\boldsymbol{\Gamma}_t$ converges to $\mathbf{R}$ as $s\to\infty$ and thus the ORKF reduces to the traditional KF in the limit of an infinitely precise noise distribution.

Equations (9), (10) and (11) are coupled via the Kalman gain and outer product matrices. No closed-form solution exists and therefore they must be solved iteratively. We alternate between estimating $\hat{\alpha}$ and estimating the posterior over

Algorithm 1 Outlier-robust Kalman filtering.

1: **function** FILTER(μ_{t-1}, $\mathbf{\Sigma}_{t-1}$, $\mathbf{y}_t$)
2: $\quad\mathbf{m} \leftarrow \mathbf{A}^T\mu_{t-1} + \mathbf{b}$ ▷ Predict
3: $\quad\mathbf{P} \leftarrow \mathbf{A}^T\mathbf{\Sigma}_{t-1}\mathbf{A} + \mathbf{Q}$
4: $\quad\mathbf{\Gamma} \leftarrow \mathbf{R}$
5: $\quad$**repeat**
6: $\quad\quad\mathbf{K} \leftarrow \left(\mathbf{C}^T\mathbf{P}\mathbf{C} + \mathbf{\Gamma}\right)^{-1}\mathbf{C}^T\mathbf{P}$ ▷ Update
7: $\quad\quad\mu_t \leftarrow \mathbf{m} + \mathbf{K}^T\left(\mathbf{y}_t - \mathbf{C}^T\mathbf{m} - \mathbf{d}\right)$
8: $\quad\quad\mathbf{\Sigma}_t \leftarrow \mathbf{K}^T\mathbf{\Gamma}\mathbf{K} + \left(\mathbf{I} - \mathbf{K}^T\mathbf{C}^T\right)\mathbf{P}\left(\mathbf{I} - \mathbf{C}\mathbf{K}\right)$
9: $\quad\quad\delta \leftarrow \mathbf{y}_t - \mathbf{C}^T\mu_t - \mathbf{d}$

$$\mathbf{\Gamma} \leftarrow \frac{s\mathbf{R} + \delta\delta^T + \mathbf{C}^T\mathbf{\Sigma}_t\mathbf{C}}{s+1}$$

10: $\quad$**until converged**
11: $\quad$**return** $\mu_t, \mathbf{\Sigma}_t$
12: **end function**

$\mathbf{S}_t$ until reaching a fixed point. Convergence and optimality are guaranteed since the variational lower bound is convex with respect to μ_t, $\mathbf{\Sigma}_t$ and $\mathbf{\Gamma}_t$.

Algorithm 1 shows pseudo-code for approximate recursive filtering by the ORKF model. The algorithm takes the mean μ_{t-1} and covariance $\mathbf{\Sigma}_{t-1}$ of the hidden state at time $t-1$, along with the data point $\mathbf{y}_t$ at time t, and returns the posterior hidden state mean μ_t and covariance $\mathbf{\Sigma}_t$ at time t. The reader should note the resemblance of Algorithm 1 to the standard KF forward recursions. For increased numerical stability, we use Joseph's form [28] in line 8 when updating the posterior hidden state covariance. To assess convergence, we monitor the innovation likelihood.

C. A constrained form of the update step

The ORKF model generalizes at least two previous approaches in the literature. The Kalman filter for robust outlier detection of [20], [21] and the recursive noise adaptive Kalman filter introduced in [22] are both special cases of the model we propose here. We can recover either of the two approaches by imposing specific constraints on the form of the posterior inverse precision matrix in (11).

Suppose we constrain $\mathbf{\Gamma}_t$ in (11) to be proportional to $\mathbf{R}$. Then $\mathbf{\Gamma}_t$ must be of the form $\mathbf{R}/w_t$ with $w_t > 0$. The set S of all such matrices is a linear subspace of the positive definite cone. Hence, the restriction of $\mathbf{\Gamma}_t$ to S can be found via linear projection. The resulting scale factor is

$$w_t = \frac{ds + d}{ds + \delta_t^T\mathbf{R}^{-1}\delta_t + \mathrm{Tr}\left[\mathbf{\Sigma}_t\mathbf{C}\mathbf{R}^{-1}\mathbf{C}^T\right]},$$

which is exactly the same as the weight update formulae in equation (12) of [20] and equation (11) of [21] with $a_{w_k,0} = b_{w_k,0} = ds$.

It is interesting to note that the authors of [20], [21] make the same modeling assumptions that we make. Although it is not mentioned in their work, they also implicitly assume heavy-tailed observation noise. Their noise is Student-t distributed in a similar fashion as in (4) albeit with a different parametrization. Then why is the approach of [20], [21] a special case of the one presented here? The explanation lies in the variational approximation. The authors of [20], [21] assume a Gaussian-Gamma form for the variational posterior distribution, which is more restrictive than the Gaussian-Wishart posterior we adopt here. Intuitively, the Gamma distribution has only two free parameters whereas the Wishart has a total of $d(d+1)/2 + 1$ parameters and hence is much more flexible.

Suppose we now constrain $\mathbf{\Gamma}_t$ to be diagonal. Furthermore, suppose that the parameter $\mathbf{R}$ is now time-varying. At each time step we set $\mathbf{R}_t = \rho\mathbf{\Gamma}_{t-1}$ with ρ a fixed constant such that $0 < \rho \leq 1$. Doing so is perfectly valid since $\mathbf{R}_t$ depends *deterministically* on $\mathbf{S}_{t-1}$ via its expectation[1]. Then the ORKF model becomes equivalent to the approach of [22]. As in the work of [20], [21], the only difference lies in the parametrization of the noise posterior.

D. Fixed-interval smoothing

Algorithm 1 can be extended to perform state smoothing. Fixed-interval smoothing takes into account future observations to improve the estimate of the current observation. Suppose we have access to a sequence $\mathbf{y}_{n+1}, \ldots, \mathbf{y}_{n+k}$ of k measurements after the nth data point. How can we use this information to improve our current estimate of $\mathbf{x}_n$?

We take a structured variational approximation similar to that in [29]. The complete data log-likelihood for a window of k data points is given by

$$\mathcal{L} = \sum_{t=n+1}^{n+k} \ln p\left(\mathbf{x}_t|\mathbf{x}_{t-1}\right) + \ln p\left(\mathbf{y}_t|\mathbf{x}_t, \mathbf{S}_t\right) + \ln p\left(\mathbf{S}_t\right). \quad (12)$$

We define a variational joint posterior distribution q that factorizes over both subsets of hidden variables, $\{\mathbf{x}_{n+1}, \ldots, \mathbf{x}_{n+k}\}$ and $\{\mathbf{S}_{n+1} \ldots, \mathbf{S}_{n+k}\}$, so that states and noise estimates decouple. Maximizing the expected value of $\mathcal{L}$ in (12) with respect to q gives the solution to the smoothing problem for the ORKF.

Pseudo-code for approximate fixed-interval smoothing in the ORKF model is shown in Algorithm 2. The algorithm updates the mean μ_n and covariance $\mathbf{\Sigma}_n$ of the hidden state estimate at time n by incorporating information from k future observations. Algorithm 2 bears much likeness to the standard Rauch-Tung-Striebel (RTS) smoother. Forward recursions are followed by backward recursions until the state and noise estimates settle at a fixed point. The sensor noise covariance is replaced by $\mathbf{\Gamma}_t$ at each time step during the forward pass, and these values are updated according to (11) during the backward pass.

V. LEARNING

A. Maximum likelihood parameter estimation

All the parameters of the ORKF can be fitted according to the Maximum Likelihood (ML) criterion via the Expectation Maximization (EM) algorithm [30]. The EM algorithm is

[1]The situation would be completely different if $\mathbf{R}_t$ depended on $\mathbf{S}_{t-1}$ directly. For instance, if $\mathbf{R}_t = \rho\mathbf{S}_{t-1}$ then $\mathbf{R}_t$ would be a random variable and hence Algorithm 1 would no longer be valid.

Algorithm 2 Outlier-robust Kalman smoothing.

function SMOOTH(μ_n, $\mathbf{\Sigma}_n$, $\mathbf{y}_n$, k)
  repeat
    for $t = n+1, \ldots, n+k$ **do**
      $\mu_t \leftarrow \mathbf{A}^T \mu_{t-1} + \mathbf{b}$
      $\mathbf{\Sigma}_t \leftarrow \mathbf{A}^T \mathbf{\Sigma}_{t-1} \mathbf{A} + \mathbf{Q}$
      $\mathbf{K} \leftarrow \left(\mathbf{C}^T \mathbf{\Sigma}_t \mathbf{C} + \mathbf{\Gamma}_t\right)^{-1} \mathbf{C}^T \mathbf{\Sigma}_t$
      $\mu_t \leftarrow \mu_t + \mathbf{K}^T \left(\mathbf{y}_t - \mathbf{C}^T \mu_t - \mathbf{d}\right)$
      $\mathbf{\Sigma}_t \leftarrow \mathbf{K}^T \mathbf{\Gamma}_t \mathbf{K} + \left(\mathbf{I} - \mathbf{K}^T \mathbf{C}^T\right) \mathbf{\Sigma}_t \left(\mathbf{I} - \mathbf{C}\mathbf{K}\right)$
    end for
    for $t = n+k, \ldots, n+1$ **do**
      $\delta \leftarrow \mathbf{y}_t - \mathbf{C}^T \mu_t - \mathbf{d}$

$$\mathbf{\Gamma}_t \leftarrow \frac{s\mathbf{R} + \delta\delta^T + \mathbf{C}^T \mathbf{\Sigma}_t \mathbf{C}}{s+1}$$

      $\mathbf{J} \leftarrow \left(\mathbf{A}^T \mathbf{\Sigma}_{t-1} \mathbf{A} + \mathbf{Q}\right)^{-1} \mathbf{A}^T \mathbf{\Sigma}_{t-1}$
      $\mu_{t-1} \leftarrow \mu_{t-1} + \mathbf{J}^T \left(\mu_t - \mathbf{A}^T \mu_{t-1} - \mathbf{b}\right)$

$$\mathbf{\Sigma}_{t-1} \leftarrow \mathbf{J}^T \left(\mathbf{Q} + \mathbf{\Sigma}_t\right) \mathbf{J} + \left(\mathbf{I} - \mathbf{J}^T \mathbf{A}^T\right) \mathbf{\Sigma}_{t-1} \left(\mathbf{I} - \mathbf{A}\mathbf{J}\right)$$

    end for
  until converged
  return $\mu_n, \mathbf{\Sigma}_n$
end function

a well-established method for parameter estimation in statistical models involving hidden variables. It proceeds by alternating between the expectation step and the maximization step to optimize the log-likelihood of the data in a coordinate-wise fashion. The expectation step consists of inferring the posterior distribution over hidden variables and can be carried out with Algorithm 2. The maximization step optimizes the data log-likelihood with respect to the model parameters. In this section we give the parameter update equations for the maximization step.

The state dynamics parameters $\mathbf{A}$ and $\mathbf{Q}$ in (1) can be estimated as shown in [31]. The state bias $\mathbf{b}$, which models a constant drift, can be estimated by concatenating it with $\mathbf{A}$ and augmenting the state vector with an extra dimension. The sensor bias term $\mathbf{d}$ in (2), which models systematic measurement errors, is given by a weighted average,

$$\mathbf{d} = \left(\sum_{t=1}^{n} \mathbf{\Gamma}_t^{-1}\right)^{-1} \left(\sum_{t=1}^{n} \mathbf{\Gamma}_t^{-1} \left(\mathbf{y}_t - \mathbf{C}^T \mu_t\right)\right),$$

where each term is weighted by the precision $\mathbf{\Gamma}_t$ of its corresponding measurement.

Estimating the sensor gain matrix $\mathbf{C}$ is fairly involved. The reason is that the negative log-likelihood,

$$-\ell = \frac{1}{2} \sum_{t=1}^{n} \delta_t^T \mathbf{\Gamma}_t^{-1} \delta_t + \mathrm{Tr}\left[\mathbf{\Sigma}_t \mathbf{C} \mathbf{\Gamma}_t^{-1} \mathbf{C}^T\right],$$

contains terms quadratic in $\mathbf{C}$ with both left- and right-hand side matrix products and hence the sufficient statistics cannot be factored out. An approach using Schur complements, similar to the one for solving the Lyapunov equation, could potentially yield a closed-form solution. However, the size of the problem still grows quadratically with the size of the data set. Therefore, as a constant-time alternative, we propose applying the Newton-Raphson method. The first and second derivatives of ℓ are readily evaluated,

$$-\nabla \ell = \sum_{t=1}^{n} -\mu_t \left(\mathbf{y}_t - \mathbf{C}^T \mu_t - \mathbf{d}\right)^T \mathbf{\Gamma}_t^{-1} + \Sigma_t \mathbf{C} \mathbf{\Gamma}_t^{-1},$$

$$-\nabla^2 \ell = \sum_{t=1}^{n} \left(\mu_t \mu_t^T + \mathbf{\Sigma}_t\right) \otimes \mathbf{\Gamma}_t^{-1},$$

where $\otimes$ denotes the Kronecker product. Then, $\mathbf{C}$ is updated as $\mathrm{vec}(\mathbf{C}) - \nabla^2 \ell^{-1} \mathrm{vec}(\nabla \ell)$, where $\mathrm{vec}(\cdot)$ denotes vertical concatenation of the columns of a matrix.

Updating the parameters of the noise model in (3) is straightforward. The optimal noise covariance $\mathbf{R}$ equals the parallel combination of the expected noise covariances,

$$\mathbf{R} = \left(\frac{1}{n} \sum_{t=1}^{n} \mathbf{\Gamma}_t^{-1}\right)^{-1}.$$

The scalar number of degrees of freedom s is the (unique) solution to the following equation,

$$\frac{1}{n} \sum_{t-1}^{n} \ln |\mathbf{\Gamma}_t|^{-1} - \ln \left|\frac{1}{n} \sum_{t=1}^{n} \mathbf{\Gamma}_t^{-1}\right| + \sum_{i=1}^{d} \ln \frac{s}{2} - \psi\left(\frac{s+1-i}{2}\right) - \sum_{i=1}^{d} \frac{1}{n} \sum_{t=1}^{n} \ln \frac{\nu_t}{2} - \psi\left(\frac{\nu_t + 1 - i}{2}\right) = 0,$$

which can be found by line search. The digamma function ψ can be evaluated via partial fraction series [32].

VI. Results

A. Synthetic data

We tested the performance of the ORKF with synthetic data. The data was generated from a noisy oscillator model with two hidden states and $d = 3$ observable outputs. The state dynamics are given by

$$\mathbf{A} = \begin{bmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{bmatrix} \quad \mathbf{b} = \begin{bmatrix} 0 \\ 0 \end{bmatrix} \quad \mathbf{Q} = \begin{bmatrix} \sigma^2 & 0 \\ 0 & \sigma^2 \end{bmatrix},$$

where $\theta = 2\pi/100$ and $\sigma = 1/2$, and

$$\mathbf{C} = \begin{bmatrix} 0 & 1 & 1 \\ 0 & 0 & 1 \end{bmatrix} \quad \mathbf{d}^T = \begin{bmatrix} 0 & 0 & 0 \end{bmatrix}.$$

defines the mapping from states to outputs.

Noise was added according to two different criteria. The first criterion is based on the assumption that outliers are statistically independent from one another. The data point at time t is sampled as follows

$$\mathbf{y}_t \sim \mathcal{N}\left(\mathbf{C}^T \mathbf{x}_t + \mathbf{d}, \mathbf{R}_t\right),$$

where $\mathbf{R}_t$ is a diagonal matrix. The ith diagonal element $r_{i,t}$ of $\mathbf{R}_t$ is set to 1 with probability ϵ and to $\lambda^2 \gg 1$ with probability $1 - \epsilon$, that is

$$p\left(r_{i,t} = 1\right) = \epsilon, \quad p\left(r_{i,t} = \lambda^2\right) = 1 - \epsilon.$$

The second criterion assumes that outliers are correlated through time. In this case, the diagonal elements of $\mathbf{R}_t$ are sampled from a hidden Markov model with transition probability ϵ and emission states $\{1, \lambda^2\}$, that is

$$p\left(r_{j,t} = \lambda^2 | r_{i,t-1} = 1\right) = \epsilon/d,$$

and similarly for the remaining probabilities.

Two data sets of size 10^3 were generated. The first data set was sampled under the assumption of statistical independence amongst outliers and was labeled as the INDEP data set. The second was sampled assuming correlated outliers and was labeled XCORR. Both were drawn with $\epsilon = 1/5$ and $\lambda = 20$. The Manhattan norm[2] and the squared Euclidean norm of the estimation error $\mathbf{x}_t - \mu_t$ was stored for each data point and the process was repeated 10 times. Table I shows the results of running the fixed-interval smoothing routine in Algorithm 2 with order $k = 10$. The entries in the table show the average (and maximum) error norm. Statistics were taken over all 10 trials.

TABLE I
ERROR STATISTICS FOR SYNTHETIC DATA

	Norm	Ting et al.	ORKF
INDEP	Manhattan	1.0053 (5.5217)	0.8953 (4.0769)
	Euclidean	0.9166 (4.1601)	0.8077 (3.2524)
XCORR	Manhattan	3.4048 (22.1300)	2.5683 (12.9278)
	Euclidean	3.5850 (15.7714)	2.5863 (10.9912)

Table I clearly shows that the ORKF model we propose here consistently outperforms the approach of Ting et al. presented in [20], [21]. The difference is not as noticeable for the INDEP data set since outliers tend to occur at random and there are often many gaps in between. Still, the average error norm for the ORKF is always 9% lower than for the algorithm of [20], [21]. The difference in performance becomes apparent for the XCORR data set due to the temporal dependency amongst outliers. In this case the ORKF performs at least 24% better, on average, with respect to both error metrics. These differences are statistically significant for p-values of 1% or greater.

Fig. 2 shows the first half of one of the sequences from the XCORR data set. The upper plot shows the raw data points and ground truth together with the second hidden state as estimated by both models. The bottom plot highlights all the locations in which outliers occur and shows the level of noise (i.e. the second diagonal element of $\mathbf{\Gamma}_t$) inferred by each model. Both plots include 95% confidence bounds.

The figure shows how the algorithm of Ting et al. finds it difficult to track the state of the system. The estimated state (green line in the top plot) is heavily influenced by the large number of outliers and often deviates strongly from the true state (blue line). In contrast, the ORKF (in red) successfully follows the hidden state despite the high level of noise. During the long burst of outliers around $t = 20$, it draws information from the third sensor to aid inference and maintain the track. In addition, the level of noise detected by the ORKF (the red line in the bottom plot) closely matches the actual locations of the outliers (yellow dots), whereas the noise inferred by the model of [20], [21] (green line) bears no resemblance to their true positions.

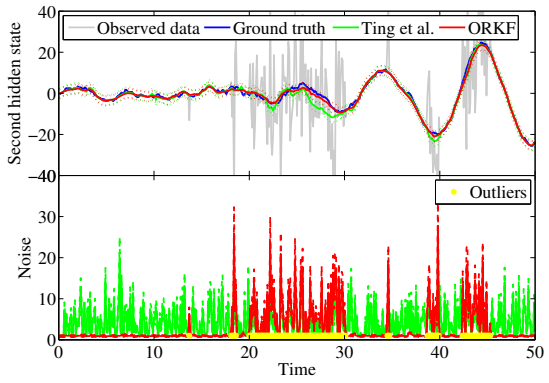

Fig. 2. Second hidden state and noise. See text for details.

We made no comparisons against the model of Särkkä and Nummenmaa since no smoothing algorithm is given in [22]. Although an algorithm for *filtering* is given, we found that filtering alone is not enough to cope with the high volume of outliers. Furthermore, we found that the algorithm of [22] may sometimes diverge under certain conditions, especially if the noise characteristics change abruptly and burst noise persists for an extended period of time.

B. GPS data set

The ORKF was also tested on real data. The Segway sensor platform shown in Fig. 3 was mounted on the back of a utility car and driven around the city of Sydney. The platform is equipped with a NovAtel Global Positioning System (GPS) receiver and a Honeywell Inertial Measurement Unit (IMU) as well as lasers and cameras. For our experiments, we recorded the position estimate from a Synchronized Position Attitude Navigation (SPAN) system. The system returns the raw non-differential GPS readings at an average rate of 1 Hz and the IMU acceleration and gyro measurements at an average rate of 50 Hz. It also fuses both modalities to produce an estimate of the position of the platform.

Fig. 3. The Segway mobile sensor platform.

[2]The Manhattan, or taxicab norm of $\mathbf{x} \in \mathbb{R}^d$ is defined as $\sum_{i=1}^{d} |x_i|$.

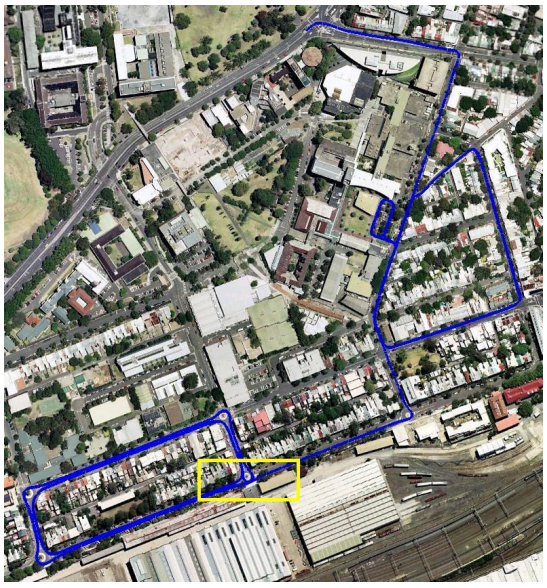

Fig. 4. Vehicle trajectory superimposed on an aerial photograph.

The platform was driven for 20 minutes around Darlington, a residential suburb in the inner-west part of Sydney. Fig. 4 shows the trajectory superimposed on an aerial photograph of the area. A total of 1301 GPS data points were stored. Each data point consists of latitude, longitude and altitude values and their corresponding uncertainty estimates, which are plotted in Fig. 5. The average latitude and longitude uncertainty is slightly over 5 meters.

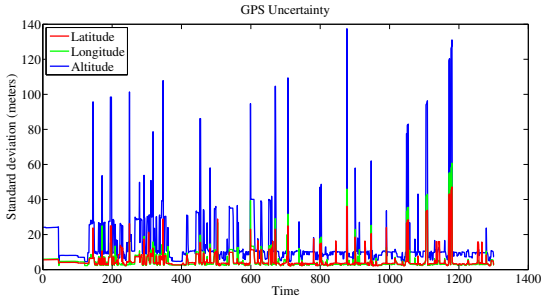

Fig. 5. Measurement uncertainty as provided by the GPS receiver.

We define a constant acceleration model for the vehicle dynamics. The only output that is directly observable is the GPS position. Hence the model is defined as follows,

$$\mathbf{A} = \begin{bmatrix} \mathbf{I} & \mathbf{0} & \mathbf{0} \\ \Delta t\mathbf{I} & \mathbf{I} & \mathbf{0} \\ \Delta t^2\mathbf{I}/2 & \Delta t\mathbf{I} & \mathbf{I} \end{bmatrix} \quad \mathbf{b} = \begin{bmatrix} \mathbf{0} \\ \mathbf{0} \\ \mathbf{0} \end{bmatrix} \quad \mathbf{d} = \mathbf{0}$$

$$\mathbf{Q} = \begin{bmatrix} \alpha^2\mathbf{I} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \Delta t^2\beta^2\mathbf{I} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \Delta t^4\gamma^2\mathbf{I} \end{bmatrix} \quad \mathbf{C} = \begin{bmatrix} \mathbf{I} & \mathbf{0} & \mathbf{0} \end{bmatrix},$$

where each block is either a 3-element vector or a 3×3 matrix. The GPS sampling period Δt is 1 second and the parameters α, β and γ have units of meters. We fitted these parameters by the ML criterion using all 1301 data points and they converged to $\alpha = 1.0931$, $\beta = 0.5334$ and $\gamma = 0.2789$ meters in 17 iterations of EM. We used a time-varying sensor noise covariance matrix $\mathbf{R}_t$ and set its diagonal elements according to the uncertainty estimates provided by the GPS receiver (Fig. 5).

We ran both the filtering and the smoothing routines in Algorithms 1 and 2. For comparison, we also ran the standard Kalman filter forward and backward recursions. To reduce the impact of outliers on the standard KF, we set a χ^2 gating threshold of 7.82 (a 95% confidence level for dimension 3). The trajectory returned by SPAN, estimated by fusing IMU with GPS, is regarded as the ground truth.

Fig. 6 shows a zoom-in on the data set and the trajectories estimated by both the standard KF (green line) and the ORKF introduced here (red). This particular sector, outlined as a yellow box in Fig. 4, is a roundabout on the corner of Wilson and Codrington St. and is especially noisy due to frequent multipath reflections. Fig. 7 shows the same roundabout with results corresponding to the smoothing case. Fig. 8 shows how the Root Mean Squared (RMS) estimation error evolves through time and includes 5% and 95% percentiles (dotted lines). Table II summarizes these results. All units in the figures and tables are either meters or seconds.

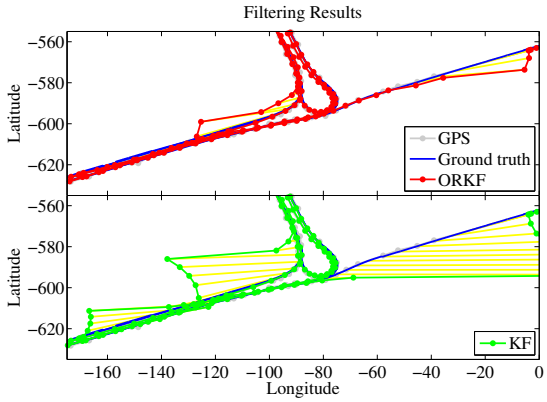

Fig. 6. Zoom-in of the GPS data set and filtering results.

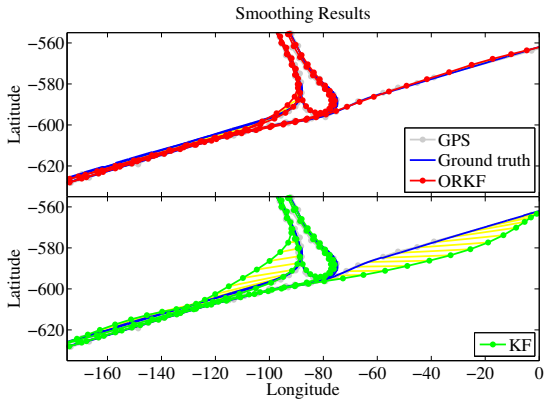

Fig. 7. Zoom-in of the GPS data set and smoothing results.

Fig. 6 shows two occasions in which both the KF and the ORKF momentarily produce fairly large estimation errors. The first occasion corresponds to the tall spike in the upper

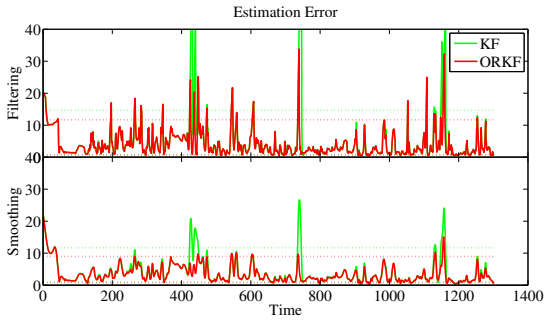

Fig. 8. Filtered and smoothed estimation error (in meters).

TABLE II
RMS ERROR FOR THE GPS DATA SET

	KF	ORKF
Filtering	10.071 (123.08)	5.8898 (33.777)
Smoothing	5.6085 (26.686)	4.4209 (21.499)

plot of Fig. 8 around $t = 750$, when the vehicle approaches the roundabout from the northeast. The estimation error from the standard KF attains its maximum value of 123 meters (Table II) for $t = 746$. In comparison, the maximum error for the ORKF occurs for $t = 739$ and is 33 meters, almost four times less. Additionally, the ORKF performs remarkably well in the smoothing case. Its maximum estimation error of 21 meters occurs at the start of the sequence, for $t = 1$. In contrast, the smoothed KF (bottom plot in Fig. 7) is still off by as much as 26 meters (Fig. 8, Table II) for $t = 740$.

VII. Conclusion

In this paper we have introduced a novel approach for processing sequential data corrupted with outliers. The ORKF model we have proposed is a robust version of the well-known Kalman filter model. Efficient filtering and smoothing algorithms were given which are straightforward modifications of standard KF forward and backward recursions. Experiments with both synthetic and real data have shown the potential of our approach.

VIII. Acknowledgements

The authors would like to thank James Underwood for his assistance during the data collection and Laura Merry for her help with the navigation system.

References

[1] S. Roweis and Z. Ghahramani, "A unifying review of linear-Gaussian models," *Neural Computation*, vol. 11, no. 2, pp. 305–345, 1999.
[2] R. Shumway and D. Stoffer, *Time Series Analysis and its Applications*. Springer-Verlag New York, 2005.
[3] J. Morris, "The Kalman filter: A robust estimator for some classes of linear-quadratic problems," *IEEE Transactions on Information Theory*, vol. 22, pp. 526–534, 1976.
[4] P. Huber, "Robust estimation of a location parameter," *Annals of Mathematical Statistics*, vol. 35, no. 1, pp. 73–101, 1964.
[5] V. Barnett and T. Lewis, *Outliers in Statistical Data*. John Wiley & Sons, 1994.
[6] H. Sorenson and D. Alspach, "Recursive Bayesian estimation using Gaussian sums," *Automatica*, vol. 7, no. 4, pp. 465–479, 1971.
[7] M. West, "Robust sequential approximate Bayesian estimation," *Journal of the Royal Statistical Society, Series B (Methodological)*, vol. 43, no. 2, pp. 157–166, 1981.
[8] R. Meinhold and N. Singpurwalla, "Robustification of Kalman filter models," *Journal of the American Statistical Association*, pp. 479–486, 1989.
[9] W. Wu and A. Kundu, "Recursive filtering with non-Gaussian noises," *IEEE Transactions on Signal Processing*, vol. 44, no. 6, pp. 1454–1468, June 1996.
[10] C. Masreliez and R. Martin, "Robust Bayesian estimation for the linear model and robustifying the Kalman filter," *IEEE Transactions on Automatic Control*, vol. 22, no. 3, pp. 361–371, June 1977.
[11] I. Schick and S. Mitter, "Robust recursive estimation in the presence of heavy-tailed observation noise," *The Annals of Statistics*, vol. 22, no. 2, pp. 1045–1080, June 1994.
[12] Z. Durovic and B. Kovacevic, "Robust estimation with unknown noise statistics," *IEEE Transactions on Automatic Control*, vol. 44, no. 6, pp. 1292–1296, 1999.
[13] G. Kitagawa, "Non-Gaussian state-space modeling of nonstationary time series," *Journal of the American Statistical Association*, vol. 82, no. 400, pp. 1032–1041, 1987.
[14] S. Kramer and H. Sorenson, "Recursive Bayesian estimation using piece-wise constant approximations," *Automatica*, vol. 24, no. 6, pp. 789–801, 1988.
[15] G. Kitagawa, "Monte Carlo filter and smoother for non-Gaussian nonlinear state-space models," *Journal of the American Statistical Association*, vol. 93, pp. 1203–1215, 1996.
[16] N. Gordon, "Approximate non-Gaussian Bayesian estimation and modal consistency," *Journal of the Royal Statistical Society: Series B (Methodological)*, vol. 55, no. 4, pp. 913–918, 1993.
[17] ——, "A hybrid bootstrap filter for target tracking in clutter," *IEEE Transactions on Aerospace and Electronic Systems*, vol. 33, no. 1, pp. 353–358, January 1997.
[18] M. Isard and A. Blake, "Condensation - conditional density propagation for visual tracking," *International Journal of Computer Vision*, vol. 29, no. 1, pp. 5–28, August 1998.
[19] R. van der Merwe, N. de Freitas, A. Doucet, and E. Wan, "The unscented particle filter," in *Advances in Neural Information Processing Systems*, November 2001.
[20] J. Ting, E. Theodorou, and S. Schaal, "A Kalman filter for robust outlier detection," in *Proceedings of the IEEE International Conference on Intelligent Robots and Systems*, 2007.
[21] ——, "Learning an outlier-robust Kalman filter," in *Proceedings of the 2007 European Conference on Machine Learning*, 2007.
[22] S. Sarkka 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.
[23] K. Murphy, "Dynamic Bayesian networks: Representation, inference and learning," Ph.D. dissertation, University of California, Berkeley, 2002.
[24] A. Dawid, "Some matrix-variate distribution theory: Notational considerations and a Bayesian application," *Biometrika*, vol. 68, pp. 265–274, 1981.
[25] W. Gosset, "The probable error of a mean," *Biometrika*, vol. 6, no. 1, pp. 1–25, 1908.
[26] M. Beal, "Variational algorithms for approximate Bayesian inference," Ph.D. dissertation, Gatsby Computational Neuroscience Unit, University College London, 2003.
[27] S. Kullback and R. Leibler, "On information and sufficiency," *Annals of Mathematical Statistics*, vol. 22, no. 1, pp. 79–86, 1951.
[28] R. Bucy and P. Joseph, *Filtering for Stochastic Processes with Applications to Guidance*. American Mathematical Society Chelsea, 2005.
[29] Z. Ghahramani and G. Hinton, "Variational learning for switching state-space models," *Neural Computation*, vol. 12, no. 4, pp. 831–864, 2000.
[30] A. Dempster, N. Laird, and D. Rubin, "Maximum likelihood from incomplete data via the EM algorithm," *Journal of the Royal Statistical Society, Series B*, vol. 39, no. 1, pp. 1–38, 1977.
[31] Z. Ghahramani and G. Hinton, "Parameter estimation for linear dynamical systems," University of Toronto, Tech. Rep. CRG-TR-96-2, February 1996.
[32] M. Abramowitz and I. Stegun, *Handbook of Mathematical Functions, with Formulas, Graphs and Mathematical Tables*. Dover Publications, Incorporated, 1974.