

A Nonparametric Adaptive Nonlinear Statistical Filter

Michael Busch and Jeff Moehlis

Abstract—We use statistical learning methods to construct an adaptive state estimator for nonlinear stochastic systems. Optimal state estimation, in the form of a Kalman filter, requires knowledge of the system’s process and measurement uncertainty. We propose that these uncertainties can be estimated from (conditioned on) past observed data, and without making any assumptions of the system’s prior distribution. The system’s prior distribution at each time step is constructed from an ensemble of least-squares estimates on sub-sampled sets of the data via jackknife sampling. As new data is acquired, the state estimates, process uncertainty, and measurement uncertainty are updated accordingly, as described in this manuscript.

I. INTRODUCTION

Let us consider a continuous nonlinear model that contains model uncertainty in the form of a stochastic forcing term, and is measured at discrete instances of time t_k :

$$dx = f(t, x)dt + \sqrt{Q}dw, \quad (1a)$$

$$y(t_k) = h(x(t_k)) + \sqrt{R}N(0, 1), \quad (1b)$$

where $x \in \mathbb{R}^n$ represents the state of the system, $f(t, x) : \mathbb{R} \times \mathbb{R}^n \rightarrow \mathbb{R}^n$ is the deterministic evolution of the states, $h(x) : \mathbb{R}^n \rightarrow \mathbb{R}^m$ is a function that maps x to the discrete-time measured output $y \in \mathbb{R}^m$, dw describes a vector Wiener process with mean zero and unit variance, and $N(0, 1)$ represents a normally distributed random variable with zero mean and unit variance. It is also noted that the covariance matrices Q and R are positive semi-definite symmetric matrices, and their square roots exist and can be computed using a singular value decomposition [1].

Since its discovery, the Kalman filter, in both its linear and nonlinear forms, has been an effective model-based noise filter that relies on an assumed known deterministic model with additive noise [2]. However, when either parameters of the model or noise variances are unknown, which is common in tasks where model identification and state estimation must occur simultaneously, the Kalman filter is likely to diverge [3], [4].

To prevent divergence, various tuning procedures exist for finding the best estimates of process and measurement noise for given a Kalman filter [5]–[9]. Kalman filter tuning typically involves minimizing the measurement error over iterations of the Kalman filter, with the process and measurement covariances as the free variables. For nonlinear systems, this type of tuning procedure requires that at each optimization iteration, the gradient of a complete time sequence of

Kalman filter iterations is taken with respect to all of the free variables. Hence, these methods are computationally costly, and are susceptible to converging to suboptimal local minima of their objective functions.

Adaptive algorithms have also been developed to allow the Kalman filter to converge on the correct noise values in an online manner [10]–[20]. Much effort has been given to developing adaptive methods for nonlinear systems because online computation is in the spirit of the Kalman filter. Adaptive methods for linear systems have seen much success over the years [18], but their formulation is limited to the linear case and does not extend to nonlinear systems in general. For nonlinear systems, adaptive strategies have been implemented for the main variants of the nonlinear Kalman filter: the extended Kalman filter (EKF) [15], and the unscented Kalman filter (UKF) [10], [13], [14]. However, for the adaptive EKF and UKF methods, convergence performance has yet to be rigorously generalized, and is sensitive to the initial estimates of the unknown parameters.

To overcome these challenges associated with implementing adaptive nonlinear Kalman filters, we propose that the unknown state and parameter distributions of the given model can be estimated by an ensemble of least-squares regression (LSQ) estimates on the known data. Jackknife sampling methods [21]–[23] can be used to generate the ensemble of LSQ estimates [24], and this ensemble generation procedure can then be made adaptive (in a Markov-Chain sense) by taking advantage of how jackknife sampling assimilates newly acquired data into the model. The formula for a statistical Kalman filter can then be used to infer the unknown process uncertainty and measurement noise covariance matrices from ensemble estimates at each step. After the unknown quantities of the stochastic model have converged, the adaptive procedure can be stopped, and a standard nonlinear Kalman filter can be implemented to take over the state estimation process.

Although our approach is supported by the theory behind ensemble Kalman filtering (EnKF) [25], [26], our adaptive method of assimilating the data is original, as well as our application of jackknife sampling to generate ensemble members. Particle filters and the EnKF both make assumptions on the sampling distribution of states, and typically rely on Markov-Chain Monte Carlo (MCMC) simulation to generate ensemble members and deduce ensemble statistics of the states. We show that by using LSQ estimation in conjunction with jackknife sampling of the known data, a sampling distribution and ensemble statistics can be acquired without making any assumptions of the sampling distribution nor having to run a high number of MCMC simulations.

M. Busch is a graduate student of Mechanical Engineering, University of California, Santa Barbara, USA mbsch@enr.ucsb.edu

J. Moehlis is with the Department of Mechanical Engineering, University of California, Santa Barbara, USA moehlis@enr.ucsb.edu

Furthermore, we describe how our adaptive method can be implemented in a parallel setting, and with a fixed number of computations at each update step.

Therefore, the aim of this manuscript is to implement the techniques of jackknife variance estimators as they apply to least squares estimators, to construct an adaptive, non-parametric, and computationally efficient statistical nonlinear filter. To motivate our jackknife sampling LSQ approach for generating ensemble statistics, we shall present an overview of the derivation of a statistical Kalman filter in Section II. In Section III we present a description of jackknife sampling methods, an adaptive jackknife sampling approach to assimilating new data, and how jackknife sampling can be used with LSQ estimation. We combine the results of Sections II and III to construct a procedure in Section IV for estimating the process and measurement noise of the model (1). We present an example application in Section V to demonstrate the efficacy of our adaptive jackknife filter, and we summarize our conclusions and directions for future work in Section VI.

II. ENSEMBLE KALMAN FILTERING

The ultimate objective of this manuscript is to develop a procedure to optimally estimate the states of a noisy nonlinear state-space model, when only a model structure and some observed measurements are provided. Since only a model structure is assumed, we shall attempt to estimate model parameters while simultaneously estimating model states. Optimal state estimation is often performed using a Kalman filter, which has many linear and nonlinear variants. For reasons that will be discussed later, we shall focus our attention on the EnKF and its formulation [2], [25], [26]. Here, we will describe the EnKF in order to motivate our approach for estimating the unknown model parameters in the next section.

A. Model Uncertainty Propagation

Let us consider a general nonlinear model that contains model uncertainty in the form of a stochastic forcing term

$$dx = f(t, x)dt + g(x)dq, \quad (2)$$

where x represents the state of the system, $f(t, x)$ gives the deterministic evolution of the states, $g(x)$ is a function that may depend on the states, and $dq = \sqrt{Q}dw$ describes a vector Wiener process with mean zero and covariance matrix $Q\delta(t)$. It is noted as a technical detail that since $g(x)$ is not an explicit function of dq , the Ito interpretation is used [27], and $\int_{t_{k-1}}^{t_k} dw = \sqrt{t_k - t_{k-1}}N(0, 1)$. Thus, one can integrate (2) from t_{k-1} to t_k to obtain the distribution of $x(t_k)$ when the distribution $x(t_{k-1})$ is known. The probability distribution of $x(t_k)$ for a given initial point $x(t_{k-1})$ is

$$x(t_k) = F(t_k, x(t_{k-1})) + g(x(t_{k-1}))\sqrt{Q\Delta t_k}N(0, 1), \quad (3)$$

where $\Delta t_k = (t_k - t_{k-1})$, and $F(t_k, x(t_{k-1}))$ is the evolution operator that deterministically maps x from time t_{k-1} to t_k according to the $dx = f(t, x)dt$ part of (2).

When $g(x)dq$ is normally distributed and forms a Markov process, it is shown in [25] that it is possible to derive the Fokker-Planck equation to describe the time evolution of the probability density function $p(x, t)$ of the model state:

$$\frac{\partial p(x, t)}{\partial t} + \sum_i \frac{\partial (f_i(t, x)p(t, x))}{\partial x_i} = \frac{1}{2} \sum_{i,j} \frac{\partial^2 (gQg^T)_{ij}}{\partial x_i \partial x_j}, \quad (4)$$

where $f_i(t, x)$ is the i^{th} component of $f(t, x)$, and gQg^T is the covariance matrix for the model errors at time t .

The EnKF, as discussed in [25] and [26], applies a Markov Chain Monte Carlo Method (MCMC) to solve (4). The probability density $p(x, t)$ is represented by an ensemble of N model states $x^{(i)}$ for $i \in \{1, \dots, N\}$, and the ensemble prediction, by integrating model states forward according to (3), is equivalent to using a MCMC method to solve (4). Hence, there is no need to find an explicit form for the solution $p(x, t)$ of (4) because $p(x, t)$ can be sufficiently described by its ensemble statistics.

Since we assume no prior knowledge of the function $g(x)$, we shall simplify matters and take $g(x) = I_{n \times n}$ so that all of the model uncertainty is spatially invariant and entirely attributed to the process noise. Furthermore, we shall assume discrete measurements y_k at times t_k , which have their own uncertainty that we shall assume to be normally distributed. For the remainder of the manuscript we shall assume the continuous-discrete stochastic model defined by (1).

B. A General Statistical Kalman Filter

To help explain how the Kalman filter is implemented from an ensemble of nonlinear system realizations, we first introduce the Kalman filter. The following description for a general Kalman filter closely follows [2], and is consistent with the EnKF of [26]. However, [26] assumes a linear mapping from the states to the outputs, but here we want to allow for nonlinear mappings, too.

We define the following variables at the discrete time instance t_k of the latest measurement $y(t_k)$:

- $x(t_k)$ = true state value,
- $\hat{x}^-(t_k)$ = state estimate prior to measurement,
- $\hat{x}^+(t_k)$ = posterior state estimate,
- $P^-(t_k) = E[(x(t_k) - \hat{x}^-(t_k))([\dots])^T]$,
- $P^+(t_k) = E[(x(t_k) - \hat{x}^+(t_k))([\dots])^T]$,

where the $[\dots]$ is shorthand notation for the term immediately to the left of it, so that the covariance matrices are written as $E[(z)([\dots])^T] = E[(z)(z)^T]$. For notational convenience, we shall momentarily omit any explicit dependence on t_k because all of the variables are understood to be implicitly evaluated at the same time instance t_k .

As described in [2], the Kalman filter is defined as

$$\hat{x}^+ = \hat{x}^- + K(y - \bar{y}), \quad (5)$$

$$P_x^+ = P_x^- - KP_{xy}^T, \quad (6)$$

$$K = P_{xy}P_y^{-1}. \quad (7)$$

Here, the notation P_{ab} denotes the cross-covariance of random variables a and b . This choice of K in (7) minimizes

the variance of the state estimates in (6), and (5) is an unbiased estimator of the model states (i.e., $\widehat{x}^\pm = \bar{x}$). We remark that (5) has the Markov Property, and is only true when the Markov Property is true for each of its elements. In the next section we will discuss how one can use the ensemble output statistics to appropriately estimate \bar{y} , and prevent measurement bias from affecting the state estimate in (5).

C. Ensemble estimation of P_x and P_y

When integrating an ensemble of points forward in time according to (3), the state covariance matrix P_x^- depends on the distribution of those deterministic points and the stochastic forcing term. For notational convenience, let us denote $\widehat{x}^- = F(t_k, x(t_{k-1}))$. Since the ensemble mean is unbiased so that $\widehat{x}^- = \bar{x}$, then an approximation for the prior ensemble covariance $P_x^-(t_k)$ becomes

$$\begin{aligned} P_x^- &= E \left[(x - \widehat{x}^-)(\dots)^T \right] \\ &= E \left[(\widehat{x}^- - \widehat{x}^- + \sqrt{Q\Delta t_k}N(0,1))(\dots)^T \right] \\ &= \frac{1}{N-1} \sum_{i=1}^N \left[(\widehat{x}_{(i)}^- - \widehat{x}^-)(\dots)^T \right] + Q\Delta t_k \\ &= \widehat{P}_x^- + Q\Delta t_k, \end{aligned} \quad (8)$$

where \widehat{P}_x^- is the sample ensemble covariance of the state prior distribution, and

$$\widehat{x}^- = \frac{1}{N} \sum_{i=1}^N \widehat{x}_{(i)}^-$$

for the collection of N ensemble members $\widehat{x}_{(i)}^-$.

To find the measurement covariance P_y and cross-covariance P_{xy} , the process noise can be made an explicit term by taking a series expansion of $h(x(t_k))$ about \widehat{x}^- at time t_k :

$$h(x(t_k)) = h(\widehat{x}^- + \sqrt{Q\Delta t_k}N(0,1)) \quad (9)$$

$$\begin{aligned} &= h(\widehat{x}^-) + Dh_{\widehat{x}^-} \sqrt{Q\Delta t_k}N(0,1) \\ &+ \sum_{n=2}^{\infty} \frac{1}{n!} D^n h_{\widehat{x}^-} (\sqrt{Q\Delta t_k}N(0,1))^n, \end{aligned} \quad (10)$$

where $D^n h_{x^{(i)}}$ represents the n^{th} vector derivative of h about

the point $x_{(i)}(t_k)$. From this we obtain

$$\begin{aligned} P_y &= E \left[(y - \bar{y})(\dots)^T \right] \\ &= E \left[(h(x(t_k)) - \overline{h(x(t_k))} + \sqrt{R}N(0,1))(\dots)^T \right] \\ &= E \left[(h(\widehat{x}^-) + Dh_{\widehat{x}^-} \sqrt{Q\Delta t_k}N(0,1) - \overline{h(\widehat{x}^-)} \right. \\ &\quad \left. + \sqrt{R}N(0,1))(\dots)^T \right] \\ &= \frac{1}{N-1} \sum_{i=1}^N \left[(h(\widehat{x}_{(i)}^-) - \overline{h(\widehat{x}_{(i)}^-)})(\dots)^T \right] \\ &+ \frac{1}{N} \sum_{i=1}^N Dh_{\widehat{x}_{(i)}^-} Q\Delta t_k Dh_{\widehat{x}_{(i)}^-}^T + R \\ &= \widehat{P}_y + \widehat{Q}_y + R, \end{aligned} \quad (11)$$

where \widehat{P}_y is the sample ensemble covariance of the measurements and \widehat{Q}_y comes from the stochastic forcing term. Similarly, one finds the cross covariance to be

$$\begin{aligned} P_{xy} &= E \left[(x - \bar{x})(y - \bar{y})^T \right] \\ &= E \left[(\widehat{x}^- - \bar{x})(h(\widehat{x}^-) - \overline{h(\widehat{x}^-)})^T \right] \\ &= \frac{1}{N-1} \sum_{i=1}^N \left[(\widehat{x}_{(i)}^- - \overline{\widehat{x}_{(i)}^-})(h(\widehat{x}_{(i)}^-) - \overline{h(\widehat{x}_{(i)}^-)})^T \right] \\ &= \widehat{P}_{xy}, \end{aligned} \quad (12)$$

where \widehat{P}_{xy} is the sample ensemble cross-covariance and all additive noise terms vanish because they are mutually uncorrelated.

By substituting equations (8), (11), and (12) into equations (7) and (6), one obtains

$$K = \widehat{P}_{xy}(\widehat{P}_y + \widehat{Q}_y + R)^{-1} \quad (13)$$

$$\widehat{x}^+ = \widehat{x}^- + K(y - \bar{y}) \quad (14)$$

$$P_x^+ = \widehat{P}_x^- + Q\Delta t_k - K(\widehat{P}_y + \widehat{Q}_y + R)K^T. \quad (15)$$

However, to implement this nonlinear statistical filter, we need to have quantities for Q and R , which we propose can be estimated directly from the data, and without making any assumptions on their sampling distribution. We did make assumptions that the process and measurement noise terms are Gaussian, which we will find in the next section, is actually consistent with a least squares parameter estimation strategy.

III. ENSEMBLE GENERATION AND ADAPTIVE UPDATE

In order to implement a statistical Kalman filter, we need to obtain ensemble estimates of the state and output distributions. To do this, we can take a statistical sample of those distributions via *jackknife sampling* [21]–[23], which has been shown to be a robust and computationally efficient way of estimating the sample distribution of a given population. By mapping the data to the state-space via LSQ estimation, we shall jackknife sample the known data in order to obtain the underlying sample distribution of the states and model parameters. The points that define the sample distribution of

the states and model parameters are then treated as ensemble members for the statistical Kalman filter. We shall first explain jackknife sampling, its consistency properties and an adaptive update rule, and then apply jackknife sampling to LSQ estimation.

A. Jackknife Sampling

Suppose we are given a sequence of n data measurements $D_n = \{Y_1, \dots, Y_n\}$, where $Y_i = (y_i, t_i)$ is defined for an observed response vector y_i from a known input sequence of t_i values. For the moment, let us fix the number of available data points n and choose some fixed positive integer d . We shall describe the *delete- d jackknife* estimator [22], [23], which estimates the sample distribution of parameters by aggregating the least squares estimates on randomly chosen subsets of $r = n - d$ data points. Let S_r be the collection of subsets of $\{1, \dots, n\}$ that have size r . For $s = \{i_1, \dots, i_r\} \in S_r$, let $\hat{\theta}_s = \hat{\theta}(Y_{i_1}, \dots, Y_{i_r})$. The delete- d jackknife estimator of $\text{var}(\theta_n)$ is defined as

$$v_n = \frac{r}{dN} \sum_{s \in S_r} \left(\hat{\theta}_s - \theta_n \right) \left(\hat{\theta}_s - \theta_n \right)^T, \quad (16)$$

where $N = \binom{n}{d}$, and θ_n is the parameter estimate that explains all of the available n data points. For a finite set of measurements, we can approximate θ_n by the arithmetic average of subsample means, which we call the jackknife estimate $\hat{\theta}_n$, and define

$$\tilde{v}_n = \frac{r}{dN} \sum_{s \in S_r} \left(\hat{\theta}_s - \hat{\theta}_n \right) \left(\hat{\theta}_s - \hat{\theta}_n \right)^T \quad (17)$$

with

$$\hat{\theta}_n = \frac{1}{N} \sum_{s \in S_r} \hat{\theta}_s.$$

When N is very large, the number of computations can be reduced by implementing techniques from survey sampling. For instance, take a simple random sample (without replacement) of size m from S_r (i.e., $S_m \subset S_r$). Compute $\hat{\theta}_s$ for $s \in S_m$, and use

$$v_n^s = \frac{r}{dm} \sum_{s \in S_m} \left(\hat{\theta}_s - \theta_n \right) \left(\hat{\theta}_s - \theta_n \right)^T \quad (18)$$

and

$$\tilde{v}_n^s = \frac{r}{dm} \sum_{s \in S_m} \left(\hat{\theta}_s - \hat{\theta}_n \right) \left(\hat{\theta}_s - \hat{\theta}_n \right)^T \quad (19)$$

with

$$\hat{\theta}_n = \frac{1}{m} \sum_{s \in S_m} \hat{\theta}_s.$$

to approximate v_n and \tilde{v}_n , respectively. These approximations are called the *jackknife-sampling variance estimators* (JSVE's) [22], [23], and m is the second-stage sample size. It is also noted that the pre-factor terms $r/(dN)$ and $r/(dm)$ are explained in [22], [23], and mitigate the bias associated with estimating the variance from a finite sample.

In [28], it was shown that

- ([28] Theorem 1) $\text{var}(v_n) = o(n^{-2})$,

- ([28] Theorem 2) $0 \leq \text{var}(v_n^s) - \text{var}(v_n) = O(m^{-1}\tau_n)$, for $\tau_n = E[(\theta_n - \theta)^4]$.

We remark that $\text{var}(v_n^s)$, $\text{var}(v_n)$, and $E[(\theta_n - \theta)^4]$ are well defined for jointly distributed random variables [29], and are only needed here to prove asymptotic consistency of jackknife sampled distributions.

The authors of [28] also show that choosing $m = n^\delta$ for some $\delta \geq 1$ is sufficient and has the same number of computations as the delete-1 jackknife estimator. If m is much smaller than N , sampling with replacement for the second-stage sample will produce almost the same estimator as sampling without replacement, which further simplifies the sampling procedure and is nearly identical to bootstrap sampling. It is also important to note that these results do not necessarily rely on $m^{-1} \sum_{s \in S_m} \theta_n \rightarrow \theta$ as $m \rightarrow \infty$, which is a convergence result that we will further discuss next.

B. Adaptive Jackknife Variance Estimator

Although the estimates are conditioned on past data, we see that the ensemble jackknife estimates abide by the Markov property in the sense that they only rely on the previous ensemble measurement and the current ensemble measurement. When tracking only the mean and variance of the distribution, all of the previous ensemble members may be forgotten, as their statistics are sufficiently captured by the mean and variance.

Suppose another measurement is collected so that there are now a total of $n + 1$ data points, and for computational reasons we want the values of r and m to remain the same as before. When constructing the basic form of our adaptive equations, it is important to define the mean and variance of the linear combination of two uncorrelated random variables X_1 and X_2 . For $\mu_1 = E[X_1]$, $\mu_2 = E[X_2]$, $v_1 = \text{var}(X_1)$, $v_2 = \text{var}(X_2)$, and two constants $a_1, a_2 \in \mathbf{R}$ such that

$$X_3 = a_1 X_1 + a_2 X_2,$$

then

$$E[X_3] = a_1 \mu_1 + a_2 \mu_2, \quad (20)$$

$$\text{var}(X_3) = a_1^2 v_1 + a_2^2 v_2. \quad (21)$$

In this context, each jackknife estimate $\hat{\theta}_n$ can be view as a combination of jackknife estimates $\hat{\theta}_{n \in s}$ that include the n^{th} data point, and those that do not $\hat{\theta}_{n \notin s}$:

$$\hat{\theta}_n = a_1 \hat{\theta}_{n \in s} + a_2 \hat{\theta}_{n \notin s}, \quad (22)$$

where $a_1 + a_2 = 1$. It is also assumed that $\hat{\theta}_{n \in s}$ and $\hat{\theta}_{n \notin s}$ are uncorrelated, which is intuitively justified by the fact that the noise contributing to the n^{th} data point is uncorrelated with the noise contributing to any of the previous $n - 1$ data points.

The values a_1 and a_2 in (22) represent the relative likelihoods of occurrence for the two types of jackknife estimates $\hat{\theta}_{n \in s}$ and $\hat{\theta}_{n \notin s}$, respectively. If we temporarily remove the n^{th} data point from the data set, we see that there are $\binom{n-1}{r}$ possible unique jackknife estimates $\hat{\theta}_{n \notin s}$ that can be

obtained from r data points. Moreover, it becomes apparent that $\widehat{\theta}_{n \notin s} = \widehat{\theta}_{n-1}$. Since there are $\binom{n}{r}$ total possible unique jackknife estimates of $\widehat{\theta}_n$, the likelihood of reselecting an estimate $\widehat{\theta}_n$ is $\binom{n-1}{r} \binom{n}{r}^{-1} = 1 - r/n$. Hence, one obtains

$$a_1 = r/n \text{ and } a_2 = 1 - r/n. \quad (23)$$

By substituting equations (22) and (23) into (20), and observing that $\widehat{\theta}_{n \notin s} = \widehat{\theta}_{n-1}$, the adaptive jackknife sample mean estimator is defined to be

$$\widehat{\theta}_n = \frac{r}{n} \widehat{\theta}_{n \in s} + \left(1 - \frac{r}{n}\right) \widehat{\theta}_{n-1}, \quad (24)$$

where

$$\widehat{\theta}_{n \in s} = \frac{1}{m} \sum_{s \in S_m^+} \widehat{\theta}_s,$$

and

$$S_m^+ = \{s \in S_m \mid n+1 \in s = \{i_1, \dots, i_m\}\}.$$

Similarly, the jackknife sample variance update is obtained by substituting equations (22) and (23) into (21). By again observing that $\widehat{\theta}_{n \notin s} = \widehat{\theta}_{n-1}$, one gets

$$\tilde{v}_n^s = \left(\frac{r}{n}\right)^2 \tilde{v}_{n \in s}^s + \left(1 - \frac{r}{n}\right)^2 \tilde{v}_{n-1}^s, \quad (25)$$

where

$$\tilde{v}_{n \in s}^s = \frac{r}{(d+1)m} \sum_{s \in S_m^+} \left(\widehat{\theta}_s - \widehat{\theta}_n\right) \left(\widehat{\theta}_s - \widehat{\theta}_n\right)^T.$$

Equation (24) inherits the convergence properties of its respective constituent terms $\widehat{\theta}_{n \in s}$ and $\widehat{\theta}_{n \notin s}$, because each of those constituent terms have identical convergence properties and (24) is a convex combination of its constituent terms. The same reasoning about convergence applies to (25) and its constituent terms $\tilde{v}_{n \in s}^s$ and $\tilde{v}_{n \notin s}^s$, as well. Furthermore, since we are effectively keeping track of a running average of second-stage m samples, the total number of second-stage samples acquired at measurement number $n = n_0 + k$ is $m_n = m_0 + km$, where m_0 is the number of second-samples used to estimate the first n_0 measurements. By choosing $m_0 = n_0$, then the condition $m_n = n^\delta$ for some $\delta \geq 1$ is satisfied, and the variance estimate of v_n has the same accuracy as the delete-1 jackknife, but for a fixed number of computations at each increment of n .

C. Least Squares Parameter Estimator

The previous sections established general results for the convergence in sample-variance for a parameter estimate without any mention of the parameter estimator. Since we want to make no assumptions about the parameter's prior distribution, we shall choose the well known LSQ estimator. Fortunately, the LSQ estimator naturally produces a normally distributed parameter estimate [24], which is consistent with the assumed uncertainty terms in the stochastic model (1a) and (1b).

Suppose we have, again, a sequence of n data measurements $D_n = \{Y_1, \dots, Y_n\}$, where $Y_i = (y_i, t_i)$, as defined earlier. Adopting much of the notation from [24], we consider

a general nonlinear model to describe an observed sequence of data

$$y_i = H(t_i, \theta) + \sigma e_i, \quad i = 1, \dots, n, \quad (26)$$

where θ is a vector of unknown constant parameters, $H(t, \theta)$ is a nonlinear function in θ , the e_i 's are independent and identically distributed (i.i.d.) unobservable random variables with mean zero and variance one, and σ is the unknown error standard deviation. It is also noted that the error terms define the measurement residuals $r_i = (y_i - H(t_i, \theta)) = \sigma e_i$.

A LSQ parameter estimator finds an estimate $\widehat{\theta}_n$ of the parameters that minimizes the mean squared error (MSE) for a model over all available data points

$$\widehat{\theta}_n = \underset{\theta}{\operatorname{argmin}} \frac{1}{n} \sum_{i=1}^n (y_i - H(t_i, \theta))^2, \quad (27)$$

which effectively minimizes σ in the model (26). In relation to the SDE model (2), one finds that $H(t, \theta) = h(F(t, x(T)))$ when $\theta = x(T)$ for some fixed point in time T .

We remark that the solution to (27) also minimizes the sample variance of the $\widehat{\theta}_n$ estimate's residuals $\operatorname{var}(\widehat{r}_n)$. When using all of the data points, the solution to (27) is only one point estimate of the parameters. With only one point estimate of the parameters $\widehat{\theta}_n$, there is no knowledge about how sensitive the parameters are to the data, or equivalently, what the variance estimate is of the parameters (i.e. $\operatorname{var}(\widehat{\theta}_n)$) that produced the given realization of the data. Jackknife variance estimation, such as the JSVE, provides a way of aggregating parameter estimates without making any prior assumptions about the distribution of $\widehat{\theta}$ (i.e., JSVEs are *nonparametric* estimators).

From the given data realization D_n , we can implement a delete- d jackknife sampling of D_n to generate a sample distribution of D , which directly gives us a sample distribution of θ by running the LSQ estimator on each jackknife sample of D_n . This approach is rigorously studied in [24] (and references therein), which specifically describes the asymptotic consistency properties of the LSQ estimator and its jackknife variance estimator in nonlinear models. For the jackknife estimate $\widehat{\theta}_n$ of θ_n , it was found in [24] that consistency and asymptotic normality of $\widehat{\theta}_n$ can be established, as well as the consistency of the jackknife variance estimator of the asymptotic covariance matrix of $\widehat{\theta}_n$. The results are summarized here for the delete-1 jackknife, as originally presented in [24], and can easily be extended to the delete- d case using the results of the previous sections.

- ([24] Theorems 1 and 2) For a LSQ estimator θ_n conditioned on n data points, then $\theta_n \rightarrow \theta$ almost surely (a.s.), and the distribution of a sequence of consistent LSQ estimators θ_n is asymptotically normally distributed.
- ([24] Lemma 3) Let $\widehat{\theta}_s$, for $i = \{1, \dots, n\}$, be the collection of delete-1 jackknife samples of the LSQ

estimates of θ_n . Then

$$\max_{i \leq n} \left\| \hat{\theta}_{ni} - \theta \right\| \rightarrow 0 \quad \text{a.s.} \quad (28)$$

- ([24] Theorem 4) The jackknife variance estimator is consistent, by proving that $n(\hat{v}_n - v_n) \rightarrow 0$ a.s.

Therefore, a jackknifed sampling of least squares estimates allows us to estimate a prior distribution of parameters for a nonlinear model without having to implement MCMC methods. An added benefit of the LSQ jackknife sampling procedure is that the estimated parameter distribution will asymptotically be normally distributed. Ensuring that the distributions are normal is essential to the performance of the EnKF, since the EnKF only uses the first two moments of the ensemble distribution. Furthermore, the adaptive scheme in the previous section provides a computationally efficient way of assimilating new data into the statistical model.

IV. POSTERIOR ESTIMATION VIA ENSEMBLE FILTERING

In previous sections, we saw how to use ensemble filtering to construct posterior estimates of a distribution's mean and covariance without having to implement MCMC methods. However, the ensemble filtering requires knowledge of the process noise, measurement noise, and the mean and covariance of the prior distribution. When these prior quantities are known, ensemble filtering can be implemented to further reduce the computational cost of assimilating new data. Without prior knowledge of model parameters or model noise distributions, we propose that one can implement jackknife estimation methods to initialize the stochastic model such that ensemble filtering can take over the posterior parameter and state estimation process once it produces posterior estimates that agree with that of the adaptive jackknife method.

A. Estimating R from Cross-Validation

When implementing the jackknife LSQ estimator, the sampling distribution for θ produces an output distribution for y . However, the measurements are subject to uncertainty, as accounted for in (2), and this uncertainty can be measured as being attributed to the additional *out-of-sample* error. Cross-validation (CV) is a statistical learning technique typically used to evaluate a model by describing its out-of-sample statistics. Typical CV methods involve training a model on a subset S_m of the available data, and then validating (testing) the model on the complement of S_m , which we denote as S_m^c .

The delete- d jackknife variance estimator already removes d data points from the available data before each step of the parameter estimation, which naturally allows us to use those d data points to acquire out-of-sample residual statistics that are indicative of the errors we would see for a future measurement. Furthermore, we can use the delete- d jackknife methodology to obtain jackknife estimates of the residual statistics, except the validation set uses a delete- r jackknife estimate.

For a given jackknife parameter estimate $\hat{\theta}_s$ such that $s \in S_m$, a residual \hat{r}_j is defined for some $j \in S_m^c$ as

$$\hat{r}_j = y_j - H(t_j, \hat{\theta}_s), \quad (29)$$

and the residual statistics defined for a set of μ indices $\{j_1, \dots, j_\mu\} \in S_m^c$, for which $\mu \leq d$, are

$$\begin{aligned} \overline{\hat{r}}_s &= \frac{1}{\mu} \sum_{j \in S_m^c} \hat{r}_j, \\ \hat{\sigma}_s^2 &= MSE(\hat{\theta}_s) = \frac{d}{r\mu} \sum_{j \in S_m^c} \hat{r}_j \hat{r}_j^T, \end{aligned}$$

where $\hat{\sigma}_s^2$ estimates the out-of-sample variance of $\hat{\theta}_s$.

For each $\hat{\theta}_s$ estimate, there exists a corresponding jackknife sample distribution of out-of-sample residual values \hat{r}_j . The jackknife mean of \hat{r} is the measurement bias, and the jackknife variance estimate $\hat{\sigma}_s^2$ captures the uncertainty attributed to both $\hat{\theta}_n$ and the measurement noise $\sqrt{R}N(0, 1)$. Since we obtain m estimates of $\hat{\theta}_s$, we also obtain m sample distributions of the out-of-sample residuals, and the expected residual distribution is described by the arithmetic mean of the m residual distributions (i.e., each residual distribution has equal probability of being the correct residual distribution). The expected jackknife residual statistics are

$$\overline{\hat{r}}_n = \frac{1}{m} \sum_{s \in S_m^c} \hat{r}_s, \quad (30)$$

$$\hat{\sigma}_n^2 = \frac{1}{m^2} \sum_{s \in S_m^c} \hat{\sigma}_s^2. \quad (31)$$

The adaptive rule outlined in Section III can also be applied to obtain

$$\overline{\hat{r}}_n = \frac{r}{n} \hat{r}_{n \in S_m^c} + \left(1 - \frac{r}{n}\right) \overline{\hat{r}}_{n-1} \quad (32)$$

$$\hat{\sigma}_n^2 = \left(\frac{r}{n}\right)^2 \hat{\sigma}_{n \in S_m^c}^2 + \left(1 - \frac{r}{n}\right)^2 \hat{\sigma}_{n-1}^2, \quad (33)$$

where $\hat{r}_{n \in S_m^c}$ and $\hat{\sigma}_{n \in S_m^c}^2$ are defined by (30) and (31) with $n \in S_m^c$.

By taking $P_y = \sigma_n^2$, and

$$\hat{P}_y = \frac{r}{dm} \sum_{s \in S_m} \left(H(t_s, \hat{\theta}_s) - \frac{1}{m} \sum_{s \in S_m} H(t_s, \hat{\theta}_s) \right) ([\dots])^T, \quad (34)$$

then one can solve for R from (11) to get

$$R = \hat{\sigma}_n^2 - \hat{P}_y. \quad (35)$$

Because the LSQ estimator finds a deterministic realization of each $\hat{\theta}_s$ assuming no stochastic forcing, it is noted that when using the definitions (31) and (34), the \hat{Q}_y term of (35) is identically equal to the zero matrix. It is also noted that by combining the results of [28] and [24], both $\hat{\sigma}_n^2$ and \hat{P}_y are each asymptotically consistent, and thus R is asymptotically consistent as well.

One can also account for the measurement bias in (5) to correct the expected output signal

$$\overline{\hat{x}}^+ = \overline{\hat{x}}^- + K \left(y - \bar{y} - \overline{\hat{r}}_n \right). \quad (36)$$

B. Estimating Q from the ensemble filter

When comparing the jackknife LSQ estimator model (16) to the SDE model (2), the parameter vector θ of the LSQ estimator is usually comprised of the SDE state values $x(t)$ at some time t_k :

$$\theta_k = \begin{pmatrix} x(t_k) \\ x_p \end{pmatrix}, \quad (37)$$

where the SDE state values $x(t)$ are augmented by the SDE model parameters x_p having zero deterministic dynamics (i.e., $dx_p = Q_p dw$). It follows from (3) that

$$\begin{bmatrix} x(t_{k+1}) \\ x_p \end{bmatrix} = \begin{bmatrix} F(t_{k+1}, \theta_k) \\ 0 \end{bmatrix} + \sqrt{Q_k \Delta t_k} N(0, 1). \quad (38)$$

Here,

$$Q_k = \begin{bmatrix} Q_t & Q_{tp} \\ Q_{tp} & Q_p \end{bmatrix},$$

where Q_t is the Q from (3), Q_p is the auto-covariance of uncertainty in the parameters, and Q_{tp} represents the cross-covariance between uncertainty in the states and parameters. Together, Q_k defines the process uncertainty of the augmented stochastic model (38).

By treating each jackknife estimate $\hat{\theta}_s$ as an ensemble estimate $\hat{\theta}_i$, we have N ensemble estimates at time t_k :

$$\hat{x}^{(i)} = F(t_k, \hat{\theta}_{ki}), \quad (39)$$

$$\hat{P}_x = \frac{1}{N} \sum (\hat{x}^{(i)} - \bar{\hat{x}}^{(i)}) ([\dots])^T. \quad (40)$$

Essentially, the jackknife samples represent an ensemble of state estimates via the transformation of (38). In terms of the ensemble filtering framework, the posterior state covariance matrix P_x^+ for state values $x(t_k) = \theta_k$ can be estimated from the jackknife variance estimate v_k , and the prior state covariance matrix P_x^- for the state values $x(t_k) = \theta_{k-1}$ can be estimated by evolving ensemble members backward in time to t_k (similar to the prediction step in UKF) and calculating the ensemble variance at that time step, say \hat{P}_x^- . The justification here is that both P_x^+ and v_n are representations of the state covariance matrix after assimilating new data. By substituting $v_n = \hat{P}_x^+$ into (15) and taking $\hat{\sigma}_n^2 = \hat{P}_y + R$, one can explicitly solve for Q :

$$Q = \frac{1}{\Delta t_k} \left(v_n - \hat{P}_x^- + \hat{P}_{xy} (\hat{\sigma}_n^2 - \hat{Q}_y)^{-1} \hat{P}_{xy}^T \right). \quad (41)$$

C. Discussion

To simplify the computation of (41), the \hat{Q}_y can be omitted from (41), which will yield a pessimistic (i.e., greater in norm) solution for Q since \hat{Q}_y is positive semi-definite and contributes positively to an inverted term. For many applications, including robust control, this is an acceptable approximation.

For highly nonlinear systems, the LSQ procedure may possibly find a region of minima that are located significantly further away from the dominant mode. These types of secondary modes can quickly emerge and cause the \hat{P}_x^- to be large enough to make (41) negative semi-definite. One solution to this problem would be to implement a Gaussian

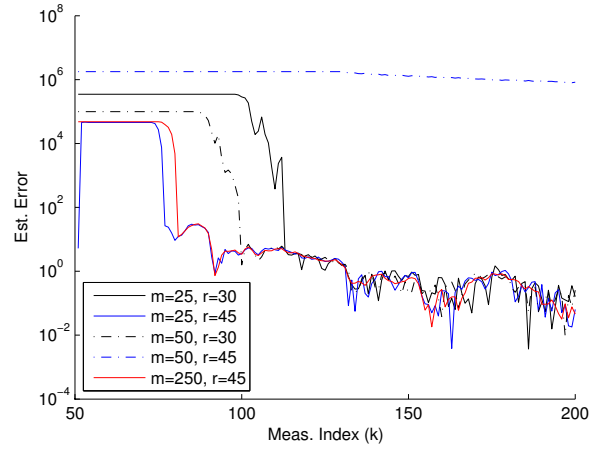


Fig. 1. Adaptive jackknife estimation performance evaluation for a logistic model, with different jackknife parameter values. In all test cases, $n = 50$ and $\mu = n - r$.

mixture model (GMM) on the ensemble of realizations and run the adaptive Kalman filter on the constituent normal distributions of ensemble members. In cases where this approach is too computationally costly, the \hat{P}_x^- term can be omitted from (41) to, again, yield an even more pessimistic solution for Q .

V. EXAMPLE APPLICATION

To demonstrate the performance of the adaptive jackknife estimator, we shall consider a simple logistic model with discrete measurements and additive noise:

$$\begin{pmatrix} dx \\ d\beta \\ dN \end{pmatrix} = \begin{pmatrix} \beta x \left(1 - \frac{x}{N}\right) \\ 0 \\ 0 \end{pmatrix} dt + \sqrt{Q} dw \quad (42)$$

$$y_k = x(t_k) + \sqrt{R} N(0, 1), \quad (43)$$

where $\beta \in \mathbf{R}^+$ is the growth parameter, $N \in \mathbf{R}^+$ is the upper bound of x , and dq and $\sqrt{R}N(0, 1)$ are the noise processes described in Section II. The logistic model defined by (42) and (43) is a common model used to describe the adoption of a behavior or new technology [30], and is known to have well known convergence properties when using a jackknife sampling LSQ variance estimator [24]. It is also noted that the integral of the deterministic part of (42) (i.e., $dx = \beta x \left(1 - \frac{x}{N}\right) dt$) has the solution:

$$x(t) = \frac{Nx(0) \exp \beta t}{N + x(0) (\exp \beta t - 1)}. \quad (44)$$

We simulated a sequence of 200 measurements y_k , at times uniformly distributed on the interval $t = [0, 80]$, with initial values $(x(0), \beta, N) = (1, 0.225, 500)$, and noise covariance matrices $Q = \text{diag}(15, 0.001, 10)$ and $R = 1$. Figure 1 shows the error, in Euclidean norm, between the state estimate of the adaptive jackknife filter and the value of (44) at time t_k . For a fixed *burn-in* period of 50 measurements, we find that the estimate of the augmented state vector converges with a greater number of included measurements r , and

fewer jackknife samples m . With a greater value of r , the adaptive jackknife filter is able to use as many measurements as possible during the burn-in initialization phase, which (i) causes a reduction in the jackknife variance estimate according to (25), and (ii) results in an initial estimate closer to the true value by causing the value r/n to be large. Using fewer jackknife samples (i.e., $m = 25$ vs $m = 50$) seems to also counter-intuitively produce a fast convergence result in this example, because few jackknife samples are needed to accurately represent the uncertainty distributions in the model. Choosing $m = 50$ causes an over-sampling of outliers, and it is not until we have $m = 250$ that the true distribution emerges.

VI. CONCLUSION AND FUTURE WORK

We have shown how one can implement the techniques of jackknife variance estimators as they apply to least squares estimators to construct an adaptive, nonparametric, and computationally efficient statistical nonlinear filter.

One issue that we left as an assumption is that for each jackknife estimate, there exists a solution to the LSQ problem. In fact, this is not a far-fetched assumption to make because bootstrap methods (similar to jackknife sampling) have been shown to efficiently search for the solution to the general LSQ problem [31]. Lastly, we also remark that jackknife sampling LSQ problem is easily broken down to a parallel computation problem, since the LSQ solution for each jackknife sample of the data can be solved independently of each other jackknife sample. Therefore, there is room for future work on this subject to increase computational efficiency, both with respect to improving LSQ estimation and parallelizing each step of the adaptive algorithm.

VII. ACKNOWLEDGMENTS

This work was supported by the Institute for Collaborative Biotechnologies through grant W911NF-09-0001 from the U.S. Army Research Office and by the Army Research Laboratory under cooperative agreement W911NF-09-2-0053 (NS-CTA). The content of the information does not necessarily reflect the position or the policy of the Government, and no official endorsement should be inferred. The U.S. Government is authorized to reproduce and distribute reprints for Government purposes notwithstanding any copyright notice herein.

REFERENCES

- [1] G. Strang, *Introduction to Linear Algebra*. Wellesley, MA: Wellesley-Cambridge Press, 2011.
- [2] D. Simon, *Optimal State Estimation: Kalman, H-infinity, and Nonlinear Approaches*. John Wiley & Sons, 2006.
- [3] F. Schlee, C. Standish, and N. Toda, "Divergence in the Kalman filter.," *AIAA journal*, vol. 5, no. 6, pp. 1114–1120, 1967.
- [4] R. Fitzgerald, "Divergence of the Kalman filter," *Automatic Control, IEEE Transactions on*, vol. 16, no. 6, pp. 736–747, 1971.
- [5] B. M. Åkesson, J. B. Jørgensen, N. K. Poulsen, and S. B. Jørgensen, "A Kalman filter tuning tool for use with model-based process control," 2007.
- [6] Y. Oshman and I. Shaviv, "Optimal tuning of a Kalman filter using genetic algorithms," *Sort*, vol. 6, p. 3, 2000.
- [7] T. D. Powell, "Automated tuning of an extended Kalman filter using the downhill simplex algorithm," *Journal of Guidance, Control, and Dynamics*, vol. 25, no. 5, pp. 901–908, 2002.
- [8] F. Ding, Y. Liu, and B. Bao, "Gradient-based and least-squares-based iterative estimation algorithms for multi-input multi-output systems," *Proceedings of the Institution of Mechanical Engineers, Part I: Journal of Systems and Control Engineering*, vol. 226, no. 1, pp. 43–55, 2012.
- [9] T. K. Lau and K.-w. Lin, "Evolutionary tuning of sigma-point Kalman filters," in *Robotics and Automation (ICRA), 2011 IEEE International Conference on*, pp. 771–776, IEEE, 2011.
- [10] Q. Song and J.-D. Han, "An adaptive UKF algorithm for the state and parameter estimations of a mobile robot," *Acta Automatica Sinica*, vol. 34, no. 1, pp. 72 – 79, 2008.
- [11] V. Fathabadi, M. Shahbazian, K. Salahshour, and L. Jargani, "Comparison of adaptive Kalman filter methods in state estimation of a nonlinear system using asynchronous measurements," in *Proceedings of the World Congress on Engineering and Computer Science*, vol. 2, 2009.
- [12] J. R. Forbes, "Adaptive approaches to nonlinear state estimation for mobile robot localization: an experimental comparison," *Transactions of the Institute of Measurement and Control*, 2012.
- [13] C. Hajiyeve and H. E. Soken, "Robust adaptive Kalman filter for estimation of UAV dynamics in the presence of sensor/actuator faults," *Aerospace Science and Technology*, vol. 28, no. 1, pp. 376 – 383, 2013.
- [14] Z. Jiang, Q. Song, Y. He, and J. Han, "A novel adaptive unscented Kalman filter for nonlinear estimation," in *Decision and Control, 2007 46th IEEE Conference on*, pp. 4293–4298, IEEE, 2007.
- [15] M. Karasalo and X. Hu, "An optimization approach to adaptive Kalman filtering," *Automatica*, vol. 47, no. 8, pp. 1785–1793, 2011.
- [16] S. Kosanam and D. Simon, *Kalman filtering for uncertain noise covariances*. PhD thesis, Cleveland State University, 2004.
- [17] Y. Li and J. Li, "Robust adaptive Kalman filtering for target tracking with unknown observation noise," in *Control and Decision Conference (CCDC), 2012 24th Chinese*, pp. 2075–2080, 2012.
- [18] R. Mehra, "Approaches to adaptive filtering," *Automatic Control, IEEE Transactions on*, vol. 17, no. 5, pp. 693–698, 1972.
- [19] S. Sarkka and J. Hartikainen, "Non-linear noise adaptive Kalman filtering via variational Bayes," in *Machine Learning for Signal Processing (MLSP), 2013 IEEE International Workshop on*, pp. 1–6, IEEE, 2013.
- [20] S. Sarkka and A. Nummenmaa, "Recursive noise adaptive Kalman filtering by variational Bayesian approximations," *Automatic Control, IEEE Transactions on*, vol. 54, no. 3, pp. 596–600, 2009.
- [21] R. G. Miller, "The jackknife-a review," *Biometrika*, vol. 61, no. 1, pp. 1–15, 1974.
- [22] B. Efron and C. Stein, "The jackknife estimate of variance," *The Annals of Statistics*, pp. 586–596, 1981.
- [23] J. Shao, C. J. Wu, et al., "A general theory for jackknife variance estimation," *The Annals of Statistics*, vol. 17, no. 3, pp. 1176–1197, 1989.
- [24] J. Shao, "Consistency of least-squares estimator and its jackknife variance estimator in nonlinear models," *Canadian Journal of Statistics*, vol. 20, no. 4, pp. 415–428, 1992.
- [25] G. Evensen, "Sequential data assimilation with a nonlinear quasi-geostrophic model using Monte Carlo methods to forecast error statistics," *Journal of Geophysical Research: Oceans (1978–2012)*, vol. 99, no. C5, pp. 10143–10162, 1994.
- [26] G. Evensen, "The ensemble Kalman filter: Theoretical formulation and practical implementation," *Ocean Dynamics*, vol. 53, no. 4, pp. 343–367, 2003.
- [27] A. H. Jazwinski, *Stochastic Processes and Filtering Theory*. Courier Dover Publications, 2007.
- [28] J. Shao, "The efficiency and consistency of approximations to the jackknife variance estimators," *Journal of the American Statistical Association*, vol. 84, no. 405, pp. 114–119, 1989.
- [29] G. A. Ghazal and H. Neudecker, "On second-order and fourth-order moments of jointly distributed random matrices: a survey," *Linear Algebra and its Applications*, vol. 321, no. 1, pp. 61–93, 2000.
- [30] H. W. Hethcote, "The mathematics of infectious diseases," *SIAM Review*, vol. 42, pp. 599–653, Dec 2000.
- [31] G. James, D. Witten, T. Hastie, and R. Tibshirani, *An Introduction to Statistical Learning*. New York, NY: Springer, 2013.