An Extension of Sigma-Point Kalman Filtering Using Nonlinear Estimator Bases

Timothy J. Wheeler and Andrew K. Packard

Abstract— **This paper investigates the problem of state estimation for nonlinear discrete-time dynamic systems. The estimator is parameterized as a linear combination of chosen basis functions. We seek the parameter that minimizes the mean squared estimation error (MSE); however, computing this objective is intractable. Hence, the MSE is approximated using the Scaled Unscented Transform (SUT), which yields a discrete least-squares optimization problem. Tikhonov regularization is used to avoid overfitting the data supplied by the SUT. A double pendulum example is used to compare this estimation strategy to the Unscented Kalman Filter.**

I. Introduction

In state estimation, the optimal estimate is given by the conditional probability density function of the state given past measurements [3]. Except for a few special cases, this conditional density cannot be computed directly. Hence, we seek a solution that provides a suboptimal estimate at a modest computational cost. To this end, we parameterize the estimator as a linear combination of chosen basis functions and interpret state estimation as weighted statistical linear regression [7], [8]. With mean squared error as the cost, the optimal parameter is a minimizer of a linear least-squares problem. However, computing the problem data involves high-dimensional integrals, so in Sections II-B and II-C, these integrals are approximated by finite sums using the Scaled Unscented Transformation (SUT) [4], [10]. Because the estimator basis is arbitrary, we risk overfitting the data produced by the SUT. Section II-D discusses how to regularize the estimation scheme to avoid overfitting. In the sequel, our estimation scheme is referred to as a Linear Regression Filter (LRF).

In Section III, a double pendulum example is used to compare the performance of the LRF to that of the Unscented Kalman Filter (UKF) [4]. An unknown time-varying parameter is added to the double pendulum model in Section III-C, and the LRF and UKF are compared on this modified system.

In Section IV, we explore the use of kernel ridge regression to minimize the same squared error cost over the regression points provided by the SUT. This allows us to search for estimators in high-dimensional reproducing kernel Hilbert spaces, while maintaining fixed computational complexity. The performance of this kernel-based method is compared to that of the LRF using the example of Section III.

II. Estimation Methodology

A. General Framework (Static Case)

Let W be a random vector taking values in $\mathbb{R}^s$. Given functions $f\colon \mathbb{R}^s \to \mathbb{R}^n$ and $g\colon \mathbb{R}^s \to \mathbb{R}^m$, define the random vectors $X = f(W)$ and $Y = g(W)$. Assume the density of W, denoted p_w, is known and Y is measured. The goal is to estimate X from Y. The estimator is represented in terms of the basis $\mathcal{B} = \{\varphi_1, \dots, \varphi_b\}$, where $\varphi_j\colon \mathbb{R}^m \to \mathbb{R}^n$. Hence, each $\theta \in \mathbb{R}^b$ defines an estimator $\Phi_\theta = \sum_{j=1}^b \theta_j \varphi_j$, whose mean squared error is given by

$$\begin{aligned} J(\theta) &= \mathbb{E}\Big[\|\Phi_\theta(Y) - X\|_2^2\Big] \\ &= \int \|\Phi_\theta(g(w)) - f(w)\|_2^2 \, p_w(w)\,\mathrm{d}w, \end{aligned} \tag{1}$$

and the estimation error variance is given by

$$P_x = \mathbb{E}\bigg[\Big(\Phi_\theta\big(g(W)\big) - f(W)\Big)\Big(\Phi_\theta\big(g(W)\big) - f(W)\Big)^T\bigg].$$

Intuitively, Φ_θ should approximate $f \circ g^{-1}$. However, in a typical estimation problem, $m < n < s$, so g^{-1} does not exist. Therefore, we seek for an estimator that minimizes the mean squared error over the distribution of W.

Note that if f and g are linear, p_w is a Gaussian density, and $\mathcal{B}$ is a basis for all affine functions of the measurement Y, then minimizing (1) yields the Kalman filter.

B. Simplifying Approximations

The cost (1) is simplified by approximating p_w as a weighted sum of Dirac delta functions. Let $S = \{w_j\}_{j=1}^\sigma$ be a subset of $\mathbb{R}^s$ and $\mathcal{A} = \{\alpha_j\}_{j=1}^\sigma$ be a set of nonnegative weights such that p_w is approximated by $\hat{p}_w = \sum_{j=1}^\sigma \alpha_j \delta_{w_j}$. Substituting $\hat{p}_w$ for p_w in the cost (1) yields

$$\begin{aligned} \hat{J}(\theta) &= \sum_{j=1}^{\sigma} \alpha_j \|\Phi_\theta(g(w_j)) - f(w_j)\|_2^2 \\ &= \sum_{j=1}^{\sigma} \alpha_j \left\| \sum_{\ell=1}^{b} \theta_\ell \varphi_\ell(g(w_j)) - f(w_j) \right\|_2^2 \\ &= \|G\theta - F\|_2^2, \end{aligned} \tag{2}$$

where $G \in \mathbb{R}^{\sigma n \times b}$ and $F \in \mathbb{R}^{\sigma n}$. Hence, finding $\hat{\theta}$ that minimizes $\hat{J}$ is a linear least-squares problem. The resulting

We thank NASA for supporting this work (Grant NNX07AC40A) and Eilyan Bitar for many helpful discussions.

T. Wheeler is with the Department of Mechanical Engineering, University of California, Berkeley (email: twheeler@berkeley.edu).

A. Packard is a professor in the Department of Mechanical Engineering, University of California, Berkeley (email: apackard@berkeley.edu).

978-1-4244-4524-0/09/$25.00 ©2009 AACC

estimate of X given the measurement $Y = y$ is $\hat{x} = \Phi_{\hat{\theta}}(y)$, and the approximate estimation error variance is

$$P_{\hat{x}} = \sum_{j=1}^{\sigma} \alpha_j \Big(\Phi_{\hat{\theta}}(g(w_j)) - f(w_j)\Big)\Big(\Phi_{\hat{\theta}}(g(w_j)) - f(w_j)\Big)^T.$$

C. Choosing Regression Points

Of course, the resulting estimator $\Phi_{\hat{\theta}}$ depends on the particular choice of regression points $\mathcal{S}$ and weights $\mathcal{A}$. In this paper, we use the Scaled Unscented Transformation (SUT) [5], [10] to approximate p_w. This method achieves good performance with only $2s+1$ regression points. However, in order for (2) to be be overdetermined (i.e., more regression points than parameters), there must be more than b/m regression points.

If W can be decomposed as $W = [W_1\,;\,W_2]$, where W_1 and W_2 are independent, then the density factors as $p_w = p_{w_1}p_{w_2}$. A reasonable method to generate more regression points is to apply the SUT to p_{w_1} and p_{w_2} separately to get $(\mathcal{S}_1, \mathcal{A}_1)$ and $(\mathcal{S}_2, \mathcal{A}_2)$ and take the set of regression points to be $\mathcal{S} = \mathcal{S}_1 \times \mathcal{S}_2$. For each $(w_i, w_j) \in \mathcal{S}$, define the corresponding weight to be $\alpha_i\alpha_j$, where $\alpha_i \in \mathcal{A}_1$ and $\alpha_j \in \mathcal{A}_2$. If W_1 and W_2 have dimensions s_1 and s_2, this method yields a set of $4s_1s_2 + 2(s_1 + s_2) + 1$ regression points, whose weights are still statistically meaningful.

D. Regularization

The basis $\mathcal{B}$ should be rich enough that $\Phi_\theta \circ g$ approximates f well over the distribution of W for some $\theta \in \mathbb{R}^b$. However, if the basis is too rich, the estimator Φ_θ may be highly nonnlinear and may overfit the relatively small number of regression points provided by the SUT. To allow for a sufficiently complex basis $\mathcal{B}$ while avoiding overfitting, we use Tikhonov regularization [2]. The regularized cost function is

$$\hat{J}_L(\theta) = \|G\theta - F\|_2^2 + \lambda \|L\theta\|_2^2, \tag{3}$$

where $\lambda > 0$ and $L \in \mathbb{R}^{b \times b}$. Since the optimal solution is $\hat{\theta} = (G^TG + \lambda L^TL)^{-1}G^TF$, the matrix L also makes computing $\hat{\theta}$ better conditioned.

E. Application to Dynamic Systems

Consider a discrete-time dynamic system of the form

$$\begin{aligned} X_{k+1} &= f(X_k, V_k), & X_0 &\sim \mathcal{N}(x_0, P_{x_0}), \\ Y_k &= g(X_k, V_k), & V_k &\sim \mathcal{N}(0, P_{v_k}), \end{aligned} \tag{4}$$

where V_k and X_k are independent for all $k \geq 0$. One time step of this system fits the general framework of Section II-A with $W = [X_k\,;\,V_k]$, $X = X_{k+1}$, and $Y = Y_k$. The LRF is iteratively applied to the system (4) as follows:

1) Use the SUT to get regression points that approximate the densities of X_k and V_k.
2) Use the regression points to compute G and F.
3) Find $\hat{\theta}$ that minimizes $\hat{J}_L$.
4) Measure $Y_k = y_k$.
5) Estimate $\hat{x}_{k+1}$ and $P_{\hat{x}_{k+1}}$ using $\Phi_{\hat{\theta}}$ and y_k.
6) Assume $X_{k+1} \sim \mathcal{N}(\hat{x}_{k+1}, P_{\hat{x}_{k+1}})$.
7) Increment k and repeat Step 1.

III. Double Pendulum Example

A. System Dynamics & Discretization

A simple double pendulum system is used to compare the effectiveness of the LRF to that of the UKF. Assuming that each link has unit mass and unit length, the equations of motion are $\dot{\theta}_1 = \omega_1$, $\dot{\theta}_2 = \omega_2$,

$$\dot{\omega}_1 = \frac{\omega_1^2\gamma\beta + g\sin\theta_2\beta + \omega_2^2\gamma - 2g\sin\theta_1}{2 - \beta^2} + d_1$$

$$\dot{\omega}_2 = \frac{\omega_2^2\gamma\beta - 2(g\sin\theta_1\beta + \omega_1^2\gamma - g\sin\theta_2)}{\beta^2 - 2} + d_2,$$

where $g = 9.81$, $\beta = \cos(\theta_2 - \theta_1)$, and $\gamma = \sin(\theta_2 - \theta_1)$. Here, d_1 and d_2 are exogenous disturbances. The measured outputs are $y_1 = \theta_1 + n_1$ and $y_2 = \omega_2 - \omega_1 + n_2$, where n_1 and n_2 are measurement noises. To express this model in the form of (4), define $X = [\theta_1\ \omega_1\ \theta_2\ \omega_2]^T$, $Y = [y_1\ y_2]^T$, and $V = [d_1\ \ d_2\ \ n_1\ \ n_2]^T$. Also, let f and g be such that $\dot{X} = f(X, V)$ and $Y = g(X, V)$. We define f_h to be the discretization of f using the fourth-order Runge-Kutta scheme with stepsize h. Hence, if $X_k = x(kh)$ and $Y_k = y(kh)$ for $k \in \mathbb{N}$, the discretized system has the desired form of equations (4). The unknown initial condition of the system is modeled as a random vector $X_0 \sim \mathcal{N}(x_0, P_{x_0})$ and the unknown disturbance V is modeled as an i.i.d. random sequence $V_k \sim \mathcal{N}(0, P_v)$.

B. Numerical Results

Define $x_0 = [3.1,\ 3.1,\ 0.8,\ -5.5]^T$, $P_{x_0} = 0.7I$, and $P_v = \mathrm{diag}([1,\ 1,\ 0.5,\ 0.5])$. Spanning the set of all third-order polynomials, the basis is given by

$$\mathcal{B} = \{y_1^{\alpha_1}y_2^{\alpha_2}e_j \mid \alpha_1, \alpha_2 \geq 0, \alpha_1 + \alpha_2 \leq 3, j = 1, \ldots, 4\},$$

where e_i is the ith column of the 4-by-4 identity matrix. The regularization parameters λ and L are chosen to minimize the contribution of high-degree polynomials in the estimator. In particular, we take $\lambda = 1$ and $L_{jj} = (\alpha_1 + \alpha_2)^{0.5}$, where $\varphi_j(y) = y_1^{\alpha_1}y_2^{\alpha_2}e_i$. The off-diagonal entries of L are zero.

We integrate the system for $N = 80$ time steps with a stepsize of $h = 0.025$. Figure 1 shows that LRF and UKF produce similar results for a single trial. However, Figure 2 shows the norm of the estimation error (i.e., $\|x_k - \hat{x}_k\|_2$) averaged over 50 independent trials. From this figure, we see that the LRF usually outperforms the UKF. Because the LRF with no regularization consistently exhibits poor performance, only the regularized case is considered here.

C. Estimating an Unknown Parameter

To further demonstrate the performance of the LRF, we consider the double pendulum example with an unknown time-varying friction parameter c_k. The modified system dynamics are

$$X_{k+1} = f_h(X_k, V_k) + [0,\ c_k\Delta_{\omega_k},\ 0,\ -c_k\Delta_{\omega_k}]^T,$$

where $\Delta_{\omega_k} := X_k^{(2)} - X_k^{(1)}$. To apply the LRF and UKF, we extend the state vector to $X_k^{\mathrm{ext}} = [X_k; c_k]$ and model the dynamics of c_k as $c_{k+1} = c_k + \mu_k^2$, where $\mu_k \sim \mathcal{N}(0, 0.05)$.

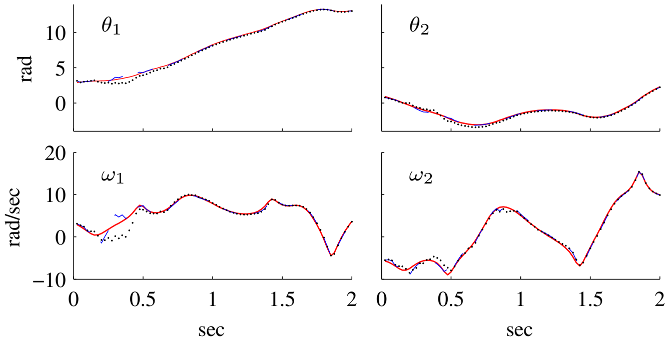

Fig. 1. States of the double pendulum and their estimated values. Legend: True (red, —), LRF (blue, – –), UKF (black, $\cdots$).

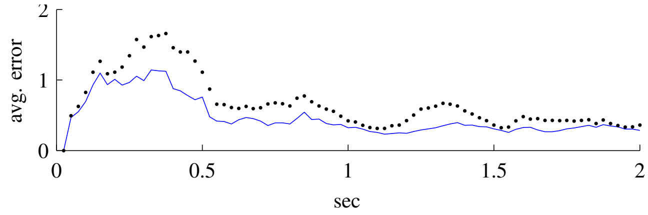

Fig. 2. Estimation error $\|x_k - \hat{x}_k\|_2$ averaged over 50 independent simulations. Legend: LRF (blue, —), UKF (black, $\cdots$).

Again, the system is simulated for $N = 80$ time steps with a stepsize of $h = 0.025$. Figures 3 and 4 show the results when the true value of c_k varies from 0 to 1 continuously, while Figures 5 and 6 show the results when the true value of c_k varies from 0 to 1 in three discrete steps. The data shown in Figures 4 and 6 are the values of $\left\|[x_k, c_k]^T - \hat{x}_k^{\text{ext}}\right\|_2$ averaged over 50 independent simulations. In both cases, the LRF performs better than the UKF.

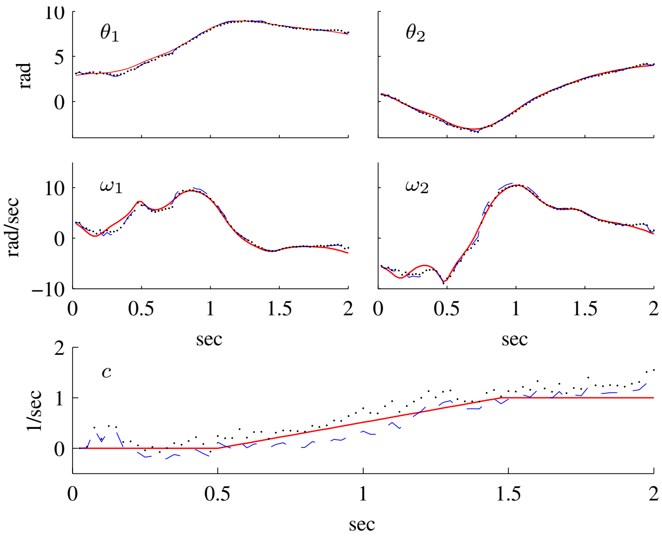

Fig. 3. States of the double pendulum with friction and their estimated values. Legend: True (red, —), LRF (blue, – –), UKF (black, $\cdots$).

IV. Kernel Ridge Regression

Computing $\hat{\theta}$ that minimizes $\hat{J}_L$ involves solving a σn-by-b least-squares problem. Hence, the computational complexity of the LRF increases with the size of the basis $\mathcal{B}$.

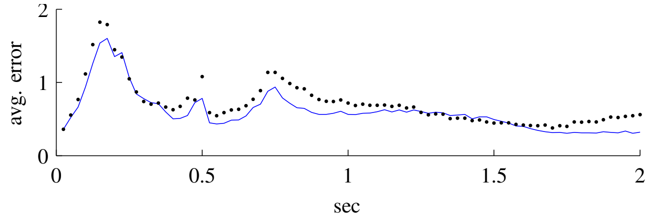

Fig. 4. Estimation error $\left\|[x_k, c_k]^T - \hat{x}_k^{\text{ext}}\right\|_2$ averaged over 50 independent simulations. (continuously varying friction parameter) Legend: LRF (blue, —), UKF (black, $\cdots$).

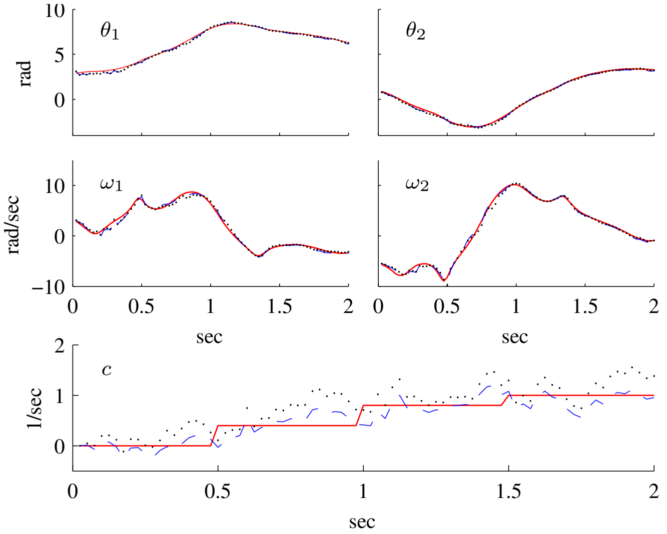

Fig. 5. States of the double pendulum with friction and their estimated values. Legend: True (red, —), LRF (blue, – –), UKF (black, $\cdots$).

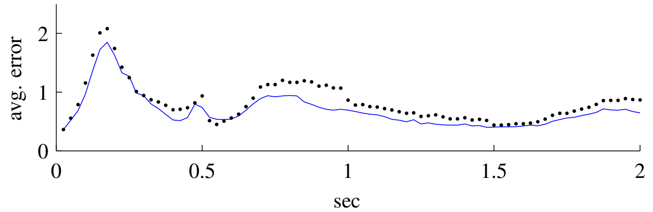

Fig. 6. Estimation error $\left\|[x_k, c_k]^T - \hat{x}_k^{\text{ext}}\right\|_2$ averaged over 50 independent simulations. (discretely varying friction parameter) Legend: LRF (blue, —), UKF (black, $\cdots$).

However, if we use kernel ridge regression (KRR) to find an estimator that minimizes the squared error over the regression points provided by the SUT, then the computational complexity of the estimator is fixed. Although the estimator takes values in $\mathbb{R}^n$, we can estimate each of the n components separately. Thus, we present KRR for the scalar-valued case with the understanding that the procedure is repeated for each component.

Let $\mathcal{H}$ be a reproducing kernel Hilbert space (RKHS) with positive definite kernel $\mathbb{K} \colon \mathbb{R}^m \times \mathbb{R}^m \to \mathbb{R}$. Given regression points $\mathcal{S}$ and weights $\mathcal{A}$ from the SUT, our goal is to find $\varphi \in \mathcal{H}$ that minimizes

$$\min_{\varphi \in \mathcal{H}} \frac{1}{2} \sum_{j=1}^{\sigma} \alpha_j \left\|\varphi(g(w_j)) - f(w_j)\right\|_2^2 + \frac{\lambda}{2} \|\varphi\|_{\mathcal{H}}^2 . \quad (5)$$

By the Representer Theorem [6], the optimal $\varphi \in \mathcal{H}$ is of the form $\hat{\varphi}(\,\cdot\,) = \sum_{j=1}^{\sigma} \theta_j \mathbb{K}(\,\cdot\,, g(w_j))$. Note that this expression does not depend on the dimension $\mathcal{H}$. Let $z = [f(w_1), \ldots, f(w_\sigma)]^T$ and $A = \mathrm{diag}(\{\alpha_1, \ldots, \alpha_\sigma\})$. Also, let $K \in \mathbb{R}^{\sigma\times\sigma}$ be given by $K_{ij} = \mathbb{K}(g(w_i), g(w_j))$. Substituting $\hat{\varphi}$, z, A, and K into (5) yields

$$\min_{\theta\in\mathbb{R}^{\sigma}} \frac{1}{2}\,\|A(z-K\theta)\|_2^2 + \frac{\lambda}{2}\theta^T K\theta. \qquad (6)$$

Because $\mathbb{K}$ is a positive definite kernel, the matrix K is positive definite and (6) is a convex quadratic program. It is easily shown that an optimal solution to (6) is

$$\hat{\theta} = (KA^2K + \lambda K)^{-1} KA^2 z \qquad (7)$$

Therefore, applying KRR requires that we solve a σ-by-σ linear system of equations for each of the n components of our estimator, regardless of the dimension of $\mathcal{H}$.

To compare this kernel-based approach with the LRF, we use the inhomogeneous polynomial kernel $\mathbb{K}(x, y) = (x^T y + 1)^d$, where d is the degree [9]. The KRR-based filter is applied to the system of Section III-B with $d = 3$ and $\lambda = 0.1$. Figure 7 shows the state estimates for a single trial and Figure 8 shows the value of $\|x_k - \hat{x}_k\|_2$ averaged over 50 trials. From these figures, it is clear that the LRF performs better than the KRR-based method. The estimates produced by the KRR-based filter are very sensitive to the value of the regularization parameter λ. It is likely that a time-dependent parameter λ_k that depends on the variance of the noises and the estimation error will produce better performance. However, further research is needed to develop such a regularization scheme.

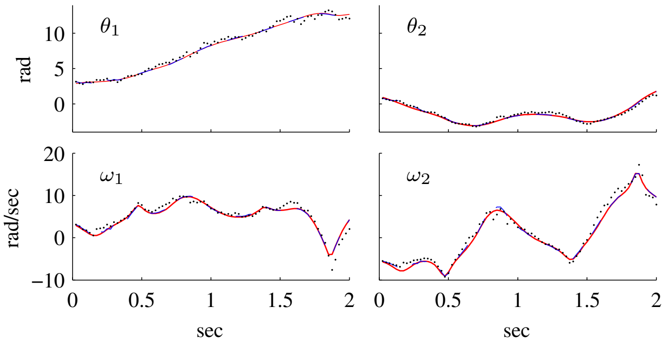

Fig. 7. States of the double pendulum and their estimated values. Legend: True (red, —), LRF (blue, – –), Kernel Ridge Regression (black, · · ·).

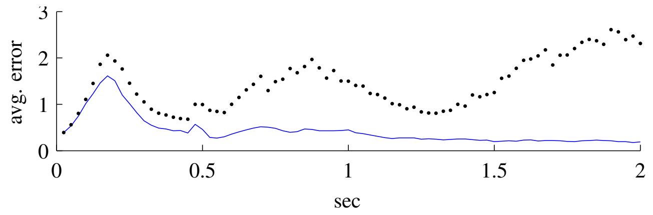

Fig. 8. Estimation error $\|x_k, -\hat{x}_k\|_2$ averaged over 50 independent simulations. Legend: LRF (blue, —), Kernel Ridge Regression (black, · · ·).

V. Conclusions & Future Work

Within the framework of Section II-A, one step of the UKF can be interpreted as weighted statistical regression over a particular set of regression points. Using the same regression points but allowing nonlinear basis functions, our Linear Regression Filter is a natural extension of the UKF. If the nonlinear estimator basis functions are chosen carefully, the LRF achieves better performance than the UKF with a slight increase in cost.

Kernel ridge regression allows us to consider an even larger class of estimators that lie in reproducing kernel Hilbert spaces. However, the performance of this kernel-based approach is very sensitive to the particular value of the regularization parameter used. Further research is required to develop a regularization scheme that takes into account the statistics of the estimation error and noises at each time step.

Because the LRF assumes that $\hat{x}_k$ is a Gaussian random vector at each time step, any algorithm that uses the Kalman filter can be extended to nonlinear systems by simply replacing the Kalman filter with the LRF. For example, the LRF can be applied to nonlinear systems with Markovian switching parameters by replacing the Kalman filter in the Interacting Multiple Model algorithm [1] with the LRF.

References

[1] Y. Bar-Shalom and X.-R. Li, *Estimation and Tracking: Principles, Techniques, and Software.* Boston: Artech House, 1993.

[2] P. C. Hansen, *Rank-Deficient and Discrete Ill-Posed Problems: Numerical Aspects of Linear Inversion.* Philadelphia: SIAM, 1997.

[3] A. H. Jazwinski, *Stochastic Processes and Filtering Theory.* New York: Academic Press, 1970.

[4] S. Julier, J. Uhlmann, and H. Durrant-Whyte, “A new method for the transformation of means and covariances in filters and estimators,” *IEEE Trans. on Automatic Control*, vol. 45, no. 3, 477–482, 2000.

[5] S. J. Julier, “The scaled unscented transformation,” in *Proceedings of the ACC*, Anchorage, AK, 2002, pp. 4555–4559.

[6] G. S. Kimeldorf and G. Wahba, “Some results on Tchebycheffian spline functions,” *Journal of Mathematical Analysis and Applications*, vol. 33, pp. 82–95, 1971.

[7] T. Lefebvre, H. Bruyninckx, and J. De Schuller, “Comment on ‘A new method for the nonlinear transformation of means and covariances in filters and estimators,”’ *IEEE Trans. on Automatic Control,* vol. 47, no. 8, pp. 1406–1409, 2002.

[8] T. Lefebvre, H. Bruyninckx, and J. De Schuller, “Kalman filters for non-linear systems: a comparison of performance,” *International Journal of Control*, vol. 77, no. 7, pp. 639–653, 2004.

[9] B. Schölkopf and A. J. Smola, *Learning with Kernels. Support Vector Machines, Regularization, Optimization, and Beyond.* Cambridge, Mass.: MIT Press, 2002.

[10] R. van der Merwe, “Sigma-point Kalman filters for probabilistic inference in dynamic state space models,” Ph.D. dissertation, OGI School of Science & Engineering at Oregon Health & Science University, 2004.