Reduced-order Unscented Kalman Filtering with application to parameter identification in large-dimensional systems

Philippe Moireau, Dominique Chapelle

ESAIM: COCV
DOI: 10.1051/cocv/2010006

ESAIM: Control, Optimisation and Calculus of Variations
www.esaim-cocv.org

REDUCED-ORDER UNSCENTED KALMAN FILTERING WITH APPLICATION TO PARAMETER IDENTIFICATION IN LARGE-DIMENSIONAL SYSTEMS

Philippe Moireau[1] and Dominique Chapelle[1]

Abstract. We propose a general reduced-order filtering strategy adapted to Unscented Kalman Filtering for any choice of sampling points distribution. This provides tractable filtering algorithms which can be used with large-dimensional systems when the uncertainty space is of reduced size, and these algorithms only invoke the original dynamical and observation operators, namely, they do not require tangent operator computations, which of course is of considerable benefit when nonlinear operators are considered. The algorithms are derived in discrete time as in the classical UKF formalism – well-adapted to time discretized dynamical equations – and then extended into consistent continuous-time versions. This reduced-order filtering approach can be used in particular for the estimation of parameters in large dynamical systems arising from the discretization of partial differential equations, when state estimation can be handled by an adequate Luenberger observer inspired from feedback control. In this case, we give an analysis of the joint state-parameter estimation procedure based on linearized error, and we illustrate the effectiveness of the approach using a test problem inspired from cardiac biomechanics.

Mathematics Subject Classification. 93E11, 93B30, 35R30, 74H15.

Received February 20, 2009. Revised November 11, 2009.
Published online March 24, 2010.

1. Introduction

Reduced-order Kalman filtering is of major interest for all observation problems in which the uncertainty space is of reduced size compared to the global state space. It allows to factorize the covariance matrix in a form such that the costly computations are performed on a matrix which has the size of the uncertainty space. For models arising from partial differential equations (PDEs), this approach is likely to be the only tractable option to perform Kalman filtering, provided that the uncertainty space can be adequately circumscribed [5,20]. Although it may be difficult to justify for general large-dimensional systems that practical uncertainties directly fall in this category, we have shown in [17] how for joint state-parameter estimation a Luenberger observer approach – see [16] – can be used as first-stage state filter, typically based on robust control strategies, but also potentially taking into account the specific nature of the observer system to use novel feedback laws [18]. In essence, this allows to restrict the uncertainty to the parameter space, hence reduced Kalman filtering – or the H^∞ variant proposed in [7] – can be used in combination with the Luenberger observer to provide a complete joint state-parameter estimation filter.

Keywords and phrases. Filtering, data assimilation, state and parameter estimation, identification in PDEs.

[1] INRIA, B.P. 105, 78153 Le Chesnay Cedex, France. philippe.moireau@inria.fr; dominique.chapelle@inria.fr

When nonlinear operators are present in the dynamics or in the measurement modeling, Unscented Kalman Filtering (UKF) is an attractive alternative to Extended Kalman Filtering (EKF). Whereas EKF uses the differentiation of nonlinear dynamical and observation operators to evaluate the propagation of probability moments, UKF instead employs well-chosen sampling points which are propagated through the nonlinear operators themselves to evaluate the resulting empirical moments. This eliminates the need for tangent operator implementations, and it can be argued that propagated moments are then more accurately approximated [12], see also [23] for a comparison of UKF and EKF in an example of parametric estimation in a mechanical system.

However, UKF seems to have been – so far – restricted to systems of relatively limited sizes due to the lack of reduced-order versions for this approach, as commented on in [22]. The primary purpose of this paper is to present a systematic method for deriving a reduced-order version for any specific UKF procedure, namely for any particular choice of sampling points and associated weights. We also show that the so-called SEIK procedure proposed in [9,19] corresponds to one instance of sampling points in this framework – namely, the simplex distribution – hence, is in fact a reduced-order UKF method. Furthermore, considering the continuous-time version of the UKF as presented in [21] we formulate a corresponding reduced-order version which is the (formal) limit of our time-discrete formulation, which then can be seen as a time-discretization scheme.

We finally specialize this generic reduced-order approach to the case of joint state-parameter estimation in PDE systems, when an adequate Luenberger observer is available for state estimation as proposed in [7,17,18]. This gives a complete state-parameter data assimilation procedure well-suited to handle nonlinearities – with the corresponding continuous-time version. We perform the linearized error analysis for the discrete-time version, while for the continuous-time version we show that the linearized error follows the same dynamics as in [17], hence can be similarly analyzed. We then present numerical examples using a test problem inspired from cardiac biomechanics to illustrate the practical effectiveness of our approach. The estimation accuracy appears to be of similar quality to the assessment results of [17], at a comparable computational cost and without the implementation complications related to operator differentiations.

The outline of the paper is as follows. In Section 2 we give a synthetic presentation of Unscented Kalman Filtering, with a few examples of particularly valuable sampling point distributions. Then, Section 3 presents our main results by demonstrating how reduced filtering can be adequately formulated in the UKF framework, including for a continuous time system. Finally, in Section 4 we discuss an application example of such procedures in the case of joint state-parameter estimation for a biomechanical system, and we use this example to successfully assess our procedures, before providing some concluding remarks in Section 5.

2. Unscented Kalman filtering overview

We recall the classical UKF formulation before introducing the reduced-order version.

2.1. The UKF transform

The Unscented Kalman Filter (UKF) as introduced in [12,13] is based on using well-chosen "interpolation points" in order to propagate the mean and covariance of a random variable with improved accuracy with respect to standard Extended Kalman Filtering (EKF). In order to summarize the rationale for this procedure, let us consider a random variable $X \in \mathbb{R}^p$ of mean $\mathbb{E}(X)$ and covariance matrix $P \in \mathcal{M}_p$. In the sequel, the space of matrices of p rows and q columns will be denoted by $\mathcal{M}_{p,q}$ and we will keep only one index when both are equal. Then, the mean and covariance of the random variable $X_f = f(X)$ for any non-linear function f satisfy:

- **Mean**

$$\begin{aligned}\mathbb{E}(X_f) &= \mathbb{E}(f(\mathbb{E}(X) + \tilde{X})) \\ &= \mathbb{E}\Big(f(\mathbb{E}(X)) + \mathrm{d}f(\mathbb{E}(X)) \cdot \tilde{X} + \tfrac{1}{2}\,\mathrm{d}^2 f(\mathbb{E}(X)) : (\tilde{X} \otimes \tilde{X}) + o(\|\tilde{X}\|^2)\Big) \\ &= f(\mathbb{E}(X)) + \tfrac{1}{2}\,\mathrm{d}^2 f(\mathbb{E}(X)) : P + o\left(\mathbb{E}(\|\tilde{X}\|^2)\right);\end{aligned}$$

- **Covariance**

$$\begin{aligned}\mathrm{Cov}(X_f) &= \mathbb{E}((X_f - \mathbb{E}(X_f))(X_f - \mathbb{E}(X_f))^T) \\ &= \mathbb{E}\Big(\{\ \mathrm{d}f(\mathbb{E}(X)) \cdot \tilde{X} + \tfrac{1}{2}\,\mathrm{d}^2 f(\mathbb{E}(X)) : (\tilde{X} \otimes \tilde{X} - P)\}\{\ldots\}^T\Big) + o\left(\mathbb{E}(\|\tilde{X}\|^2)\right) \\ &= \ \mathrm{d}f(\mathbb{E}(X)) \cdot P \cdot\ \mathrm{d}f(\mathbb{E}(X))^T + o\left(\mathbb{E}(\|\tilde{X}\|^2)\right).\end{aligned}$$

Let us now construct r points

$$X^{(i)} = \mathbb{E}(X) + \tilde{X}^{(i)}, \quad 1 \leq i \leq r, \tag{2.1}$$

associated with r coefficients α_i that together satisfy the following conditions

$$\begin{cases} \displaystyle\sum_{1\leq i\leq r} \alpha_i = 1 \\ E_\alpha(X^*) \stackrel{\text{def}}{=} \displaystyle\sum_{1\leq i\leq r} \alpha_i X^{(i)} = \mathbb{E}(X) \\ \mathrm{Cov}_\alpha(X^*) \stackrel{\text{def}}{=} \displaystyle\sum_{1\leq i\leq r} \alpha_i (X^{(i)} - \mathbb{E}(X)) \cdot (X_i - \mathbb{E}(X))^T = \mathrm{Cov}(X) \end{cases} \tag{2.2}$$

meaning

$$\begin{cases} \displaystyle\sum_{1\leq i\leq r} \alpha_i \tilde{X}^{(i)} = 0, \\ \displaystyle\sum_{1\leq i\leq r} \alpha_i \tilde{X}^{(i)} \cdot \tilde{X}^{(i)\,T} = \mathrm{Cov}(X) \end{cases} \tag{2.3}$$

and propagate them by the non-linear function f

$$X_f^{(i)} = f(X^{(i)}). \tag{2.4}$$

We then verify that the empirical mean

$$\begin{aligned} E_\alpha(X_f^*) &= \sum_{1\leq i\leq r} \alpha_i X_f^{(i)} \\ &= f(\mathbb{E}(X)) + \sum_{1\leq i\leq r} \alpha_i\, \mathrm{d}f(\mathbb{E}(X)) \cdot \tilde{X}^{(i)} + \sum_{1\leq i\leq r} \alpha_i\, \mathrm{d}^2 f(\mathbb{E}(X)) : \tilde{X}^{(i)} \otimes \tilde{X}^{(i)} + o\left(\mathbb{E}(\|\tilde{X}\|^2)\right) \\ &= f(\mathbb{E}(X)) + \ \mathrm{d}^2 f(\mathbb{E}(X)) : P + o\left(\mathbb{E}(\|\tilde{X}\|^2)\right) = \mathbb{E}(X_f) + o\left(\mathbb{E}(\|\tilde{X}\|^2)\right), \end{aligned}$$

and empirical covariance

$$\begin{aligned} \mathrm{Cov}_\alpha(X_f^*) &= \sum_{1\leq i\leq r} \alpha_i (f(X^{(i)}) - E_\alpha(X_f))(f(X^{(i)}) - E_\alpha(X_f))^T \\ &= \sum_{1\leq i\leq r} \alpha_i \left(\ \mathrm{d}f(\mathbb{E}(X)) \cdot \tilde{X}^{(i)} + \ \mathrm{d}^2 f(\mathbb{E}(X)) : (\tilde{X}^{(i)} \otimes \tilde{X}^{(i)} - P) + o\left(\mathbb{E}(\|\tilde{X}\|^2)\right)\right) \left(\ldots\right)^T \\ &= \ \mathrm{d}f(\mathbb{E}(X))\Big(\sum_{1\leq i\leq r} \alpha_i \tilde{X}^{(i)} \tilde{X}^{(i)T}\Big).\, \mathrm{d}f(\mathbb{E}(X)) + o\left(\mathbb{E}(\|\tilde{X}\|^2)\right) \\ &= \ \mathrm{d}f(\mathbb{E}(X)) \cdot P \cdot\ \mathrm{d}f(\mathbb{E}(X)) + o\left(\mathbb{E}\|\tilde{X}\|^2\right) = \mathrm{Cov}(X_f) + o\left(\mathbb{E}(\|\tilde{X}\|^2)\right), \end{aligned}$$

approximate respectively to the second order and first order the mean and covariance of X_f. This explains why UKF is expected to have a better accuracy than EKF, as the order of approximation is similar to that provided by second-order Kalman filters, see [22] and references therein.

We henceforth refer to such points satisfying the constraints (2.2) as *sigma-points* as in [13]. In order to construct such sigma-points, we can start by defining "unitary" sigma-points – denoted by $I^{(i)}$ – meaning sigma-points representing a random vector of zero mean and unit covariance. In the sequel, we use the abusive notation $\sqrt{\mathrm{Cov}(X)}$ in order to express any choice of square matrix C such that

$$\mathrm{Cov}(X) = CC^T.$$

In particular, a Cholesky decomposition or the principal square root of $\mathrm{Cov}(X)$ fulfill this condition. Then, the points defined by

$$X^{(i)} = \mathbb{E}(X) + \tilde{X}^{(i)} = \mathbb{E}(X) + \sqrt{\mathrm{Cov}(X)} I^{(i)} \tag{2.5}$$

satisfy the conditions in (2.2) since

$$\sum_{1\leq i\leq r} \alpha_i \tilde{X}_i \tilde{X}_i^T = \sqrt{\mathrm{Cov}(X)} \cdot \mathbb{1}_p \cdot \sqrt{\mathrm{Cov}(X)}^T = \mathrm{Cov}(X).$$

In the sequel we will denote by $[X^*]$ the matrix obtained by concatenating the column vectors $X^{(i)}$ side by side.

We can present different choices of sigma-points useful in practice:

- **canonical sigma-points** ($r = 2p$): aligned with the canonical base (e_i) of the space with associated coefficients $\alpha_i = \frac{1}{2p}$

$$I^{(i)} = \begin{cases} \sqrt{p}\, e_i, & \text{for } 1 \leq i \leq p \\ -\sqrt{p}\, e_{i-r}, & \text{for } r+1 \leq i \leq 2p; \end{cases} \tag{2.6}$$

- **star sigma-points** ($r = 2p+1$): the origin is added to the previous canonical points

$$I^{(i)} = \begin{cases} \sqrt{p}\, e_i, & \text{for } 1 \leq i \leq p \\ -\sqrt{p}\, e_{i-r}, & \text{for } p+1 \leq i \leq 2p \\ 0 & \text{for } i = 2p+1; \end{cases} \tag{2.7}$$

- **simplex sigma-points** ($r = p+1$): this represents the smallest number of necessary sigma-points, which are located on a regular polyhedron of radius $\sqrt{p}$. These points can be constructed recursively with a procedure similar to that described in [10], namely,

$$I^{(i)} = \sqrt{p}\, \tilde{I}_r^{(i)},$$

where the vectors $\tilde{I}_r^{(i)}$ are the columns of the matrix noted $[\tilde{I}_r^*]$ recursively defined by

$$\begin{cases} [\tilde{I}_1^*] = \left(-\dfrac{1}{\sqrt{2\alpha}} \quad \dfrac{1}{\sqrt{2\alpha}}\right), \quad \alpha = \dfrac{p}{p+1} \\ [\tilde{I}_d^*] = \begin{pmatrix} & & 0 \\ & [\tilde{I}_{d-1}^{(i)}] & \vdots \\ & & 0 \\ \dfrac{1}{\sqrt{\alpha d(d+1)}} & \cdots & \dfrac{1}{\sqrt{\alpha d(d+1)}} \quad \dfrac{-d}{\sqrt{\alpha d(d+1)}} \end{pmatrix}, \quad 2 \leq d \leq p. \end{cases}$$

The weights α_i are chosen all equal ($\alpha_i = \frac{1}{p+1}$) since every point is located on a regular polyhedron around the mean.

Remark 2.1 (sigma-points linear constraints)**.** Note that each family of sigma-points is associated with a set of linear relations linking the column vectors $I^{(i)}$ with each other. More specifically there are $r-p$ such relations. In all cases, we have that the mean of these column vectors is zero, since they are centered. This is the only link for simplex sigma-points, but the other families above have additional straightforward linear relations. These relations can be algebraically summarized in the form

$$[I^*].R = 0, \quad R \in \mathcal{M}_{r,r-p}, \tag{2.8}$$

and they – of course – equivalently correspond to linear constraints that each row of $[I^*]$ should satisfy. Note further that these linear constraints also apply to sigma-points with arbitrary mean and covariance in the form

$$[X^* - E_\alpha(X^*)].R = 0, \tag{2.9}$$

directly deduced from (2.5).

Remark 2.2 (average distance to the mean)**.** Note that in all the above choices the sigma-points are located at a distance $\sqrt{p}$ from the origin (excepting the origin itself for the star sigma-points). This is consistent with the simple identity

$$\begin{aligned}\mathbb{E}(\|X - \mathbb{E}(X)\|^2) &= E((X - \mathbb{E}(X))^T(X - \mathbb{E}(X))) \\ &= \mathbb{E}(\mathrm{tr}((X - \mathbb{E}(X))^T(X - \mathbb{E}(X)))) = \mathbb{E}(\mathrm{tr}((X - \mathbb{E}(X))(X - \mathbb{E}(X))^T)) \\ &= \mathbb{E}(\mathrm{tr}\, P),\end{aligned}$$

applied with the identity covariance matrix, since we are considering unitary points. Some authors have proposed sigma-points which do not satisfy this distance property, in particular for star sigma-points for which increasing the weight of the center-point allows to preserve the unit covariance while moving the other points further away from the mean [11]. The main motivation for this would to better represent some particular probability distributions – namely, up to higher-order moments. Yet, such particular distributions are unlikely to be preserved through the dynamical process – and through the observation operator – hence we take the risk of interpolating non-linear operators with points unduly distant from the expected value as discussed in [15].

2.2. **The UKF filter**

We now consider a discrete-time finite dimensional nonlinear dynamical system

$$X_{n+1} = A_{n+1|n}(X_n),$$

where $A_{n+1|n}(.)$ is the so-called transition operator. For this system we suppose that – even with an adequate model of the system dynamics – some uncertainties remain on the initial condition $X_{n=0}$. Nevertheless, we have some additional information on the system which is provided by measurements at each time step, written in the form

$$Z_n = H_n(X_n) + \chi_n,$$

where $H_n(.)$ is a nonlinear observation operator and χ_n the corresponding noise of covariance W_n associated with the measurement process.

The principle of the UKF filter is to replace the means and covariances of the Kalman Filter by the empirical means and covariances propagated by the dynamical operator A during the prediction, and by the observation operator H during the correction. This leads to the following algorithm:

- **Prediction:**

$$\begin{cases} \hat{X}_n^{(i)+} = \hat{X}_n^+ + \sqrt{P_n^+}\, I^{(i)} \\ \hat{X}_{n+1}^- = E_\alpha(A_{n+1|n}(\hat{X}_n^{*+})) \\ P_{n+1}^- = \mathrm{Cov}_\alpha(A_{n+1|n}(\hat{X}_n^{*+})); \end{cases} \tag{2.10a}$$

- **Correction:**

$$\begin{cases} \hat{X}_{n+1}^{(i)-} = \hat{X}_{n+1}^- + \sqrt{P_{n+1}^-}\, I^{(i)} \\ Z_{n+1}^{(i)} = H_{n+1}(\hat{X}_{n+1}^{(i)-}) \\ P_\alpha^{x\tilde{z}} = \mathrm{Cov}_\alpha(X_{n+1}^{*-}, Z_{n+1}^*) \\ P_\alpha^{\tilde{z}} = W_{n+1} + \mathrm{Cov}_\alpha(Z_{n+1}^*, Z_{n+1}^*) \\ \hat{K}_{n+1} = P_\alpha^{x\tilde{z}}(P_\alpha^{\tilde{z}})^{-1} \\ \hat{X}_{n+1}^+ = \hat{X}_n^- + \hat{K}_{n+1}(Z_{n+1} - E_\alpha(Z_{n+1}^*)) \\ P_{n+1}^+ = P_{n+1}^- - P_\alpha^{x\tilde{z}}(P_\alpha^{\tilde{z}})^{-1}(P_\alpha^{x\tilde{z}})^T. \end{cases} \tag{2.10b}$$

Remark 2.3 (model noise)**.** In the most general case, the model $A_{n+1|n}(.)$ can also be supposed not to be perfect and for example additive Gaussian model noise $\omega_n \in \mathcal{N}(0, Q_n)$ may be considered, *i.e.*

$$X_{n+1} = A_{n+1|n}(X_n) + B_{n+1|n}\omega_n.$$

Then, only the *a priori* covariance is modified during the prediction step in the form

$$P_{n+1}^- = \mathrm{Cov}_\alpha(A_{n+1|n}(\hat{X}_n^{*+})) + B_{n+1|n}Q_nB_{n+1|n}^T.$$

3. Reduced-order UKF

Assuming that P is of reduced rank p – typically much smaller than the dimension of the space d – the basic idea in reduced-order filtering is, in essence, to be able to manipulate covariance matrices in the factorized form

$$P = LU^{-1}L^T, \tag{3.1}$$

where U – in the group of invertible matrices $\mathcal{GL}_p$ – is of much smaller size than $P \in \mathcal{M}_d$ and represents the main uncertainties in the system. What is crucial here is to be able to perform all computations on L and U without needing to compute P as such, see *e.g.* [22] and references therein.

3.1. Matrix-based formulation of the empirical covariances

Consider some sigma-points $(V^{(i)})_{1\leq i\leq r}$ in $\mathbb{R}^p$ – not necessarily unitary but of zero empirical mean (called "centered" sigma-points) – associated with some coefficients $(\alpha) = (\,\alpha_1 \cdots \alpha_r\,)^T$. Then, we define the matrix of these sigma-points denoted by $[V^*] \in \mathcal{M}_{p,r}$. We have

$$\mathrm{Cov}_\alpha(V^*) = [V^*]D_\alpha[V^*]^T,$$

with $D_\alpha = \mathrm{diag}(\alpha_1, \ldots, \alpha_r) \in \mathcal{M}_r$. It is then possible to obtain a factorized form of the empirical covariance of any sigma-points set in $\mathbb{R}^d$ – assumed to represent an arbitrary random variable of covariance of rank p – respecting the same construction rules as the $V^{(i)}$, namely, the linear constraints discussed in Remark 2.1.

Proposition 3.1. *Let $X \in \mathbb{R}^d$ be a random variable with mean and covariance represented by the empirical mean and covariance of the sigma-points $(X^{(i)})_{1\leq i\leq r}$ respecting the construction rules of the $(V^{(i)})_{1\leq i\leq r}$, then we have the identity*

$$\mathrm{Cov}(X) = [X^*]D_\alpha[V^*]^T([V^*]D_\alpha[V^*]^T)^{-1}[V^*]D_\alpha[X^*]^T.$$

Proof. Since the sigma-points $V^{(i)}$ have zero empirical mean, then, if we denote by $[(1)]$ the matrix with all coefficients equal to 1, we have $[(1)]D_\alpha[V^*]^T = 0$, and consequently

$$[X^*]D_\alpha[V^*]^T = [X^* - E_\alpha(X^*)]D_\alpha[V^*]^T.$$

Furthermore, the new sigma-points $X^{(i)}$ verify the same linear constraints as the $V^{(i)}$, namely, $r-p$ linear constraints that the row vectors (in $\mathbb{R}^r$) of $[V^*]$ and $[X^* - E_\alpha(X^*)]$ satisfy. Hence, we can use the fact that the p rows of $[V^*]$ make up a basis of the subspace of all vectors in $\mathbb{R}^r$ which satisfy these constraints, and we infer

$$\exists Q \in \mathcal{M}_{p,d}, \quad [X^* - E_\alpha(X^*)]^T = [V^*]^T \cdot Q.$$

Therefore,

$$\begin{aligned}
\mathrm{Cov}_\alpha(X^*) &= [X^* - E_\alpha(X^*)]D_\alpha[X^* - E_\alpha(X^*)]^T \\
&= [X^* - E_\alpha(X^*)]D_\alpha[V^*]^T Q \\
&= [X^* - E_\alpha(X^*)]D_\alpha[V^*]^T([V^*]D_\alpha[V^*]^T)^{-1}[V^*]D_\alpha[V^*]^T Q \\
&= [X^* - E_\alpha(X^*)]D_\alpha[V^*]^T([V^*]D_\alpha[V^*]^T)^{-1}[V^*]D_\alpha[X^* - E_\alpha(X^*)]^T \\
&= [X^*]D_\alpha[V^*]^T([V^*]D_\alpha[V^*]^T)^{-1}[V^*]D_\alpha[X^*]^T.
\end{aligned}$$

□

Remark 3.1. If the initial sigma-points are some unitary sigma-points $I^{(i)}$, then they verify by definition

$$[I^*]D_\alpha[I^*]^T = \mathrm{Cov}_\alpha(I^*) = \mathbb{1},$$

and the previous proposition reduces to

$$\mathrm{Cov}_\alpha(X^*) = [X^*]D_\alpha[I^*]^T \cdot [I^*]D_\alpha[X^*]^T.$$

Conversely, this allows to define unitary sigma-points from the $V^{(i)}$ by

$$[I^*] = ([V^*]D_\alpha[V^*]^T)^{-\frac{1}{2}}[V^*]. \tag{3.2}$$

Remark 3.2. Consider two sets of centered unitary sigma-points $I_1^{(i)}$ and $I_2^{(i)}$ constructed by the same rule. Then, as above we have

$$[I_1^*]^T = [I_2^*]^T Q,$$

and

$$\mathbb{1} = [I_1^*]D_\alpha[I_1^*]^T = Q^T[I_2^*]D_\alpha[I_2^*]^T Q = Q^T Q,$$

hence $[I_2^*] = Q[I_1^*]$, with Q unitary. This is consistent with the fact that all possible square roots to be used in (2.5) differ by the right-multiplication of a unitary matrix.

3.2. **Simplex case**

In this section, we focus on the simplex distribution. Let us consider now the iteration n, where we have computed the mean $\hat{X}_n^+$ and covariance P_n^+ in the form

$$P_n^+ = L_n U_n^{-1} L_n^T.$$

In the UKF procedure, we start by constructing a sampling of sigma-points

$$\hat{X}_n^{(i)^+} = \hat{X}_n^+ + L_n \sqrt{U_n^{-1}} I^{(i)},$$

for a good choice of unitary sigma-points, computed from a given sampling $V^{(i)}$.

We then compute the propagated sigma-points $A(\hat{X}_n^{(i)+})$ and we set

$$\hat{X}_{n+1}^- = E_\alpha(A(\hat{X}_n^{(i)+})),$$

with the empirical covariance satisfying

$$P_{n+1}^- = [A(\hat{X}_n^{*+})] D_\alpha [V^*]^T ([V^*] D_\alpha [V^*]^T)^{-1} [V^*] D_\alpha [A(\hat{X}_n^{*+})]^T, \tag{3.3}$$

due to Proposition 3.1, since there is no linear constraint associated with simplex points (other than the mean identity). This means that we can use as sigma-points

$$\hat{X}_{n+1}^{(i)-} = A(\hat{X}_n^{(i)+}).$$

Defining

$$L_{n+1} = [\hat{X}_{n+1}^{*-}] D_\alpha [V^*]^T, \quad P_\alpha^V = [V^*] D_\alpha [V^*]^T,$$

it implies that

$$P_{n+1}^- = L_{n+1} (P_\alpha^V)^{-1} L_{n+1}^T.$$

In order to take into account the correction step, we then first compute the observation points

$$Z_{n+1}^{(i)} = H(\hat{X}_{n+1}^{(i)-}),$$

and, from the UKF algorithm, we get the mean and empirical covariance in the correction step with

$$\begin{aligned} \mathrm{Cov}_\alpha(\tilde{X}, \tilde{Z}) &= \sum_{1 \le i \le r} \alpha_i (\hat{X}_{n+1}^{(i)-} - \hat{X}_{n+1}^-)(Z_{n+1}^{(i)} - E_\alpha(Z_{n+1}^*))^T \\ &= L_{n+1} ([V^*] D_\alpha [V^*]^T)^{-1} \{HL\}_{n+1}^T, \end{aligned} \tag{3.4}$$

where

$$\{HL\}_{n+1} = [H(\hat{X}_{n+1}^*)] D_\alpha [V^*]^T.$$

In addition

$$\begin{aligned} P_\alpha^{\tilde{z}} &= W_{n+1} + \sum_{1 \le i \le r} \alpha_i (Z_{n+1}^{(i)} - E_\alpha(Z_{n+1}))(Z_{n+1}^{(i)} - E_\alpha(Z_{n+1}^*))^T \\ &= W_{n+1} + \{HL\}_{n+1} ([V^*] D_\alpha [V^*]^T)^{-1} \{HL\}_{n+1}^T. \end{aligned} \tag{3.5}$$

Introducing

$$U_{n+1} = P_\alpha^V + \{HL\}_{n+1}^T W_{n+1}^{-1} \{HL\}_{n+1}, \tag{3.6}$$

it is not difficult using classical Kalman filter algebraic manipulations to obtain

$$\begin{aligned}
\hat{X}^+_{n+1} &= \hat{X}^-_{n+1} + L_{n+1}(P^v_\alpha)^{-1}\{HL\}^T_{n+1}(W_{n+1} + \{HL\}_{n+1}(P^v_\alpha)^{-1}\{HL\}^T_{n+1})^{-1}(Z_{n+1} - E_\alpha(Z^*_{n+1})) \\
&= \hat{X}^-_{n+1} + L_{n+1}U^{-1}_{n+1}U_{n+1}(P^v_\alpha)^{-1}\{HL\}^T_{n+1}(W_{n+1} + \{HL\}_{n+1}(P^v_\alpha)^{-1}\{HL\}^T_{n+1})^{-1}(Z_{n+1} - E_\alpha(Z^*_{n+1})) \\
&= \hat{X}^-_{n+1} + L_{n+1}U^{-1}_{n+1}(\mathbb{1} + \{HL\}^T_{n+1}W^{-1}_{n+1}\{HL\}_{n+1}(P^v_\alpha)^{-1})\{HL\}^T_{n+1} \\
&\quad \times (W_{n+1} + \{HL\}_{n+1}(P^v_\alpha)^{-1}\{HL\}^T_{n+1})^{-1}(Z_{n+1} - E_\alpha(Z^*_{n+1})) \\
&= \hat{X}^-_{n+1} + L_{n+1}U^{-1}_{n+1}\{HL\}^T_{n+1}W^{-1}_{n+1}(W_{n+1} + \{HL\}_{n+1}(P^v_\alpha)^{-1}\{HL\}^T_{n+1}) \\
&\quad \times (W_{n+1} + \{HL\}_{n+1}(P^v_\alpha)^{-1}\{HL\}^T_{n+1})^{-1}(Z_{n+1} - E_\alpha(Z^*_{n+1})) \\
&= \hat{X}^-_{n+1} + L_{n+1}U^{-1}_{n+1}\{HL\}^T_{n+1}W^{-1}_{n+1}(Z_{n+1} - E_\alpha(Z^*_{n+1})).
\end{aligned}$$

We finally obtain the correction covariance

$$P^+_{n+1} = L_{n+1}U^{-1}_{n+1}L^T_{n+1},$$

using the classical matrix inversion lemma which we recall without proof (based on the same kind of manipulation as above) for completeness.

Lemma 3.2 (matrix inversion lemma)**.** *Let M_1, M_{12}, M_{21}, M_2 be matrices with M_1, M_2 and $M_2 - M_{21}M_1^{-1}M_{12}$ invertible, then $M_1 - M_{12}M_2^{-1}M_{21}$ is invertible and verifies*

$$(M_1 - M_{12}M_2^{-1}M_{21})^{-1} = M_1^{-1} + M_1^{-1}M_{12}(M_2 - M_{21}M_1^{-1}M_{12})^{-1}M_{21}M_1^{-1}.$$

Remark 3.3 (optional resampling)**.** As presented in Section 2.2, the UKF procedure classically uses a resampling after the prediction step, unlike what we just discussed in this reduced version. In fact, we can further comment on the fact that no resampling is needed in the simplex case. Indeed, using Proposition 3.1, the resampled points would have the following expression

$$\begin{aligned}
[\hat{X}^{*-}_{n+1}] &= [\hat{X}^-_{n+1}] + [A(\hat{X}^{*+}_n)]D_\alpha[V^*]^T([V^*]D_\alpha[V^*]^T)^{-1/2}[I^*] \\
&= [\hat{X}^-_{n+1}] + L_{n+1}([V^*]D_\alpha[V^*]^T)^{-1}[V^*].
\end{aligned}$$

Let us then construct some examples of $[V^*]$ in the case where $\alpha_i = \frac{1}{p+1}$. Since the $V^{(i)}$ have to be centered, each line of $[V^*]$ is orthogonal to the vector $(1 \dots 1)^T \in \mathbb{R}^{p+1}$ denoted by (1) in the sequel. We can use as the rows of $[V^*]$ the row vectors ℓ_i such that

$$\ell_i^T = e_i - \frac{\langle (1), e_i \rangle}{\langle (1), e_\natural \rangle} e_\natural,$$

with $(e_i)_{1 \le i \le p+1}$ the vectors of the canonical base of $\mathbb{R}^{p+1}$ and $e_\natural$ a vector such that $\langle (1), e_\natural \rangle \neq 0$. Taking $e_\natural = \sum_{1 \le i \le p+1} e_i$ we obtain

$$[V^*] = \begin{pmatrix} 1 & & 0 & 0 \\ & \ddots & & \\ 0 & & 1 & 0 \end{pmatrix} - \frac{1}{p+1}\begin{pmatrix} 1 & \cdots & 1 \\ \vdots & \vdots & \vdots \\ 1 & \cdots & 1 \end{pmatrix}. \tag{3.7}$$

It is also possible to choose $e_\natural = e_{p+1}$. Then

$$[V^*] = \begin{pmatrix} 1 & & 0 & -1 \\ & \ddots & & \\ 0 & & 1 & -1 \end{pmatrix}.$$

In both cases we can show the identity

$$[V^*]^T([V^*]D_\alpha[V^*]^T)^{-1}[V^*] = \begin{pmatrix} 1 & & 0 \\ & \ddots & \\ 0 & & 1 \end{pmatrix} - \frac{1}{p+1}\begin{pmatrix} 1 & \cdots & 1 \\ \vdots & \vdots & \vdots \\ 1 & \cdots & 1 \end{pmatrix},$$

hence

$$L_{n+1}([V^*]D_\alpha[V^*]^T)^{-1}[V^*] = [A(\hat{X}_n^{*+}) - \hat{X}_{n+1}^-],$$

so that

$$\hat{X}_{n+1}^{(i)-} = A(\hat{X}_n^{(i)+}).$$

With other choices of vectors $V^{(i)}$, the only effect of the resampling would be to rotate the points around the mean along the "covariance ellipsoid" as discussed in Remark 3.2. This explains why we do not introduce any resampling at the analysis stage.

It is also worthwhile noting that the so-called SEIK filter formulated in [19] is equivalent to this reduced UKF algorithm. In fact, the SEIK procedure uses in the resampling and covariance factorization a matrix T which is nothing but the transposed of $[V^*]$ with the particular choice of sigma-points corresponding to (3.7). Hence, although at first the introduction of this specific matrix may seem somewhat arbitrary – and raise questions on some other possibly better choices – we can see here through the scaling (3.2) that this choice corresponds to adequate unitary sampling points, and that any other valid choice would correspond to rotated sampling points. Of course, similar observations hold for the choice of $[V^*]$ which do not have the intrinsic character of unitary sigma-points – meaning that the corresponding empirical covariance is arbitrary. Nevertheless, the advantage of the $[V^*]$ points – or the row vectors of T in [19] – compared to the unitary sampling points is in the ease of their constructions from the sampling rules.

Algorithm summary. Given adequate sampling rules, precompute the corresponding $[I^*]$,

- **Sampling**:

$$\begin{cases} C_n = \sqrt{U_n^{-1}} \\ \hat{X}_n^{(i)+} = \hat{X}_n^+ + L_n C_n I^{(i)}, \quad 1 \le i \le p+1; \end{cases} \tag{3.8a}$$

- **Prediction**:

$$\begin{cases} \hat{X}_{n+1}^- = E_\alpha(A(\hat{X}_n^{*+})) \\ \hat{X}_{n+1}^{(i)-} = \left| \begin{array}{ll} \hat{X}_{n+1}^- + [A(\hat{X}_n^{*+}) - \hat{X}_{n+1}^-]D_\alpha^{\frac{1}{2}} I^{(i)} & \text{with resampling} \\ \text{or} & \\ A(\hat{X}_n^{(i)+}) & \text{without resampling} \end{array} \right. \\ L_{n+1} = [X_{n+1}^{*-}]D_\alpha[V^*]^T \in \mathcal{M}_{d,p} \\ P_{n+1}^- = L_{n+1}(P_\alpha^V)^{-1}L_{n+1}^T; \end{cases} \tag{3.8b}$$

- **Correction**:

$$\begin{cases} Z_{n+1}^{(i)} = H(\hat{X}_{n+1}^{(i)-}) \\ \{HL\}_{n+1} = [Z_{n+1}^*]D_\alpha[V^*]^T \\ U_{n+1} = \mathbb{1} + \{HL\}_{n+1}^T W_{n+1}^{-1}\{HL\}_{n+1} \in \mathcal{M}_p \\ \hat{X}_{n+1}^+ = \hat{X}_{n+1}^- + L_{n+1}U_{n+1}^{-1}\{HL\}_{n+1}^T W_{n+1}^{-1}(Z_{n+1} - E_\alpha(Z_{n+1}^*)) \\ P_{n+1}^+ = L_{n+1}U_{n+1}^{-1}L_{n+1}^T. \end{cases} \tag{3.8c}$$

Remark 3.4 (model noise in reduced-order UKF)**.** Referring to Remark 2.3, model noise makes reduced-order strategies difficult to apply since it changes the rank of the *a priori* covariance. Nevertheless, assuming that the number of significant singular values remains limited we can use a Frobenius-norm projection to eliminate the other (small) values, followed by an adequate resampling of the sigma-points according to this projected covariance. This type of projection-resampling strategy will be explained in greater detail in the next section when addressing general choices of sigma-points, since the rank of the *a priori* covariance may also change due to the number of sigma-points considered.

3.3. Generalized case

We now propose an extension of this reduced order UKF strategy to choices of sigma-points other than the simplex sampling. In such cases, the sigma-points distributions must satisfy some constraints in addition to the mean identity, as summarized in (2.8) and (2.9), in adequacy with the fact that r centered points span a space of dimension p. In the above argument, this implies that the identity (3.3) no longer holds, because the propagated points do not satisfy these additional constraints in general. For instance, we can easily imagine that star sigma-points do not preserve their specific distribution – and span a space of dimension larger than p, indeed – after propagation through a non-linear operator.

Nevertheless, if we wish to keep the rank (or rank approximation) fixed, we can of course project the propagated empirical covariance onto a matrix of rank p. If this projection is performed in the Frobenius norm, this amounts to computing the pth-order singular value decomposition (SVD) of the matrix. For the *a priori* covariance matrix, we thus seek the SVD of

$$P_\alpha = [A(\hat{X}_n^{*+}) - E_\alpha(A(\hat{X}_n^{*+}))]D_\alpha[A(\hat{X}_n^{*+}) - E_\alpha(A(\hat{X}_n^{*+}))]^T \in \mathcal{M}_d,$$

which is of rank at most $r-1$. Setting $M = [A(\hat{X}_n^{*+}) - E_\alpha(A(\hat{X}_n^{*+}))]D_\alpha^{\frac{1}{2}}$, we can show that the SVD of $P_\alpha = MM^T$ can be performed by diagonalizing the associated Grammian matrix

$$G_\alpha = M^T M \in \mathcal{M}_r,$$

in the form

$$G_\alpha = \Upsilon_r \Sigma_r \Upsilon_r^T, \text{ with } \Upsilon_r^T \Upsilon_r = \mathbb{1},$$

with the (positive) eigenvalues stored in decreasing order in the diagonal matrix Σ_r. Then the normalized eigenvectors of P_α are the column vectors of $M\Upsilon_r\Sigma_r^{-\frac{1}{2}}$ since

$$\begin{cases} MM^T(M\Upsilon_r\Sigma_r^{-\frac{1}{2}}) = M\Upsilon_r\Sigma_r^{\frac{1}{2}} = (M\Upsilon_r\Sigma_r^{-\frac{1}{2}})\Sigma_r \\ \Sigma_r^{-\frac{1}{2}}\Upsilon_r^T M^T M\Upsilon_r\Sigma_r^{-\frac{1}{2}} = \mathbb{1}. \end{cases}$$

Hence,

$$P_\alpha = (M\Upsilon_r\Sigma_r^{-\frac{1}{2}})\Sigma_r(\Sigma_r^{-\frac{1}{2}}\Upsilon_r^T M^T) = M\Upsilon_r\Upsilon_r^T M^T,$$

and we obtain the p-th-order SVD by

$$P_{\text{SVD}} = M\Upsilon_p\Upsilon_p^T M^T$$

where Υ_p contains only the first p columns of Υ_r. We can now easily regenerate r sigma-points satisfying the required construction rules by using some unitary sigma-points in the sampling equation

$$\begin{aligned}[\hat{X}^{*-}_{n+1}] &= [E_\alpha(A(\hat{X}^{*+}_n))] + M\Upsilon_p[I^*] \\ &= [E_\alpha(A(\hat{X}^{*+}_n))] + [A(\hat{X}^{*+}_n) - E_\alpha(A(\hat{X}^{*+}_n))]D_\alpha^{1/2}\Upsilon_p[I^*],\end{aligned}$$

which correspond by definition to the p-th-order *a priori* covariance matrix

$$P^-_{n+1} \overset{def}{=} [\hat{X}^{*-}_{n+1}]D_\alpha[V^*]^T([V^*]D_\alpha[V^*]^T)^{-1}[V^*]D_\alpha[\hat{X}^{*-}_{n+1}]^T,$$

and to the empirical mean

$$\hat{X}^-_{n+1} = E_\alpha(A(\hat{X}^{*+}_n)).$$

We now need to propagate the new sigma-points $\hat{X}^{(i)-}_{n+1}$ through the observation operator. Using the same notation as in Section 2 we can write the filter in the form

$$\hat{K}_{n+1} = P_\alpha^{\check{x}\check{z}}(P_\alpha^{\check{z}})^{-1},$$

and we will compute this operator using the matrix inversion lemma to obtain a tractable algorithm. To this end we introduce the following compact notation

$$[\tilde{X}] = [\hat{X}^*_{n+1} - \hat{X}^-_{n+1}], \quad [\tilde{Z}] = [Z^*_{n+1} - E_\alpha(Z^*_{n+1})],$$

and we then have

$$\begin{aligned}\hat{K}_{n+1} &= [\tilde{X}]D_\alpha[\tilde{Z}]^T(W_{n+1} + [\tilde{Z}]D_\alpha[\tilde{Z}]^T)^{-1} \\ &= [\tilde{X}]D_\alpha[\tilde{Z}]^T\Big(W^{-1}_{n+1} - W^{-1}_{n+1}[\tilde{Z}](D_\alpha^{-1} + [\tilde{Z}]^TW^{-1}_{n+1}[\tilde{Z}])^{-1}[\tilde{Z}]^TW^{-1}_{n+1}\Big) \\ &= [\tilde{X}]D_\alpha\Big(\mathbb{1}_r - [\tilde{Z}]^TW^{-1}_{n+1}[\tilde{Z}](D_\alpha^{-1} + [\tilde{Z}]^TW^{-1}_{n+1}[\tilde{Z}])^{-1}\Big)[\tilde{Z}]^TW^{-1}_{n+1}.\end{aligned}$$

Let us now set

$$D_{\mathfrak{m}} = [\tilde{Z}]^TW^{-1}_{n+1}[\tilde{Z}] \in \mathcal{M}_r,$$

which – unlike for $P_\alpha^{\check{z}}$ – can be computed in practice, since its dimension is equal to the number of sigma-points. We thus have

$$\begin{aligned}\hat{K}_{n+1} &= [\tilde{X}]D_\alpha(\mathbb{1} - D_{\mathfrak{m}}(D_\alpha^{-1} + D_{\mathfrak{m}})^{-1})[\tilde{Z}]^TW^{-1}_{n+1}, \\ &= [\tilde{X}](D_\alpha - D_\alpha(D_{\mathfrak{m}}^{-1} + D_\alpha)^{-1}D_\alpha)[\tilde{Z}]^TW^{-1}_{n+1},\end{aligned} \tag{3.9}$$

and, by the same argument as in Proposition 3.1, this gain can also be written in the form

$$\hat{K}_{n+1} = L_{n+1}(P_\alpha^V)^{-1}[V^*](D_\alpha - D_\alpha(D_{\mathfrak{m}}^{-1} + D_\alpha)^{-1}D_\alpha)[\tilde{Z}]^TW^{-1}_{n+1}, \tag{3.10}$$

with

$$L_{n+1} = [\hat{X}^{*-}_{n+1}]D_\alpha[V^*]^T.$$

Note that the term $[\tilde{Z}]^T$ in (3.9) cannot be treated in the same manner since the sigma-points propagated by the observation operator do not satisfy the original constraints. In addition to the gain, we also need to compute the *a posteriori* covariance matrix in order to resample at the next step. We have

$$\begin{aligned}P^+_{n+1} &= P^-_{n+1} - P_\alpha^{\check{x}\check{z}}(P_\alpha^{\check{z}})^{-1}(P_\alpha^{\check{x}\check{z}})^T \\ &= P^-_{n+1} - [\tilde{X}]D_\alpha(D_{\mathfrak{m}} - D_{\mathfrak{m}}(D_\alpha^{-1} + D_{\mathfrak{m}})^{-1}D_{\mathfrak{m}})D_\alpha[\tilde{X}]^T.\end{aligned}$$

We now use the matrix inversion lemma to simplify

$$\Big(D_{\mathfrak{m}} - D_{\mathfrak{m}}(D_\alpha^{-1} + D_{\mathfrak{m}})^{-1} D_{\mathfrak{m}}\Big)^{-1} = D_{\mathfrak{m}}^{-1} + D_\alpha,$$

and obtain as for the filter

$$\begin{aligned} P_{n+1}^+ &= [\tilde{X}](D_\alpha - D_\alpha(D_{\mathfrak{m}}^{-1} + D_\alpha)^{-1} D_\alpha)[\tilde{X}]^T \\ &= L_{n+1}(P_\alpha^V)^{-1}[V^*](D_\alpha - D_\alpha(D_{\mathfrak{m}}^{-1} + D_\alpha)^{-1} D_\alpha)[V^*]^T(P_\alpha^V)^{-1} L_{n+1}^T. \end{aligned} \tag{3.11}$$

The advantage of this last form is that we can again write

$$P_{n+1}^+ = L_{n+1} U_{n+1}^{-1} L_{n+1}^T,$$

with

$$\begin{aligned} U_{n+1}^{-1} &= (P_\alpha^V)^{-1}[V^*](D_\alpha - D_\alpha(D_{\mathfrak{m}}^{-1} + D_\alpha)^{-1} D_\alpha)[V^*]^T(P_\alpha^V)^{-1} \\ &= (P_\alpha^V)^{-1} - (P_\alpha^V)^{-1}[V^*] D_\alpha (D_{\mathfrak{m}}^{-1} + D_\alpha)^{-1} D_\alpha [V^*]^T (P_\alpha^V)^{-1}. \end{aligned}$$

Hence, defining $D_V \in \mathcal{M}_r$ as

$$D_V = D_\alpha [V^*]^T (P_\alpha^V)^{-1} [V^*] D_\alpha, \tag{3.12}$$

we can simplify – with another application of the matrix inversion lemma

$$U_{n+1} = P_\alpha^V + [V^*] D_\alpha (D_{\mathfrak{m}}^{-1} + D_\alpha - D_V)^{-1} D_\alpha [V^*]^T. \tag{3.13}$$

The factorized form $P_{n+1}^+ = L_{n+1} U_{n+1}^{-1} L_{n+1}^T$ means that the *a posteriori* covariance matrix is already of rank p. This is because the sigma-points $\hat{X}_{n+1}^{*-}$ directly appear in the covariance, without any nonlinear operator applied on them. Hence, this ensures that we can iterate at the next step.

Algorithm summary. Given adequate sampling rules, precompute the corresponding $[V^*]$, $P_\alpha^V = [V^*] D_\alpha [V^*]^T$, $[I^*] = ([V^*] D_\alpha [V^*]^T)^{-\frac{1}{2}} [V^*]$, and $D_V = D_\alpha [V^*]^T (P_\alpha^V)^{-1} [V^*] D_\alpha$.

- **Sampling**:

$$\begin{cases} C_n = \sqrt{U_n^{-1}} \\ \hat{X}_n^{(i)+} = \hat{X}_n^+ + L_n C_n I^{(i)}, \quad 1 \le i \le r; \end{cases} \tag{3.14a}$$

- **Prediction**:

$$\begin{cases} \hat{X}_{n+1}^- = E_\alpha(A(\hat{X}_n^{*+})) \\ \hat{X}_{n+1}^{(i)-} = \hat{X}_{n+1}^- + [A(\hat{X}_n^{*+}) - \hat{X}_{n+1}^-] D_\alpha^{\frac{1}{2}} \Upsilon_p I^{(i)}, \quad \text{resampling with SVD} \\ L_{n+1} = [X_{n+1}^{*-}] D_\alpha [V^*]^T \in \mathcal{M}_{d,p} \\ P_{n+1}^- = L_{n+1}(P_\alpha^V)^{-1} L_{n+1}^T; \end{cases} \tag{3.14b}$$

- **Correction**:

$$\begin{cases} [\tilde{Z}] = [H(\hat{X}_{n+1}^*) - E_\alpha(H(\hat{X}_{n+1}^*))] \\ D_{\mathfrak{m}} = [\tilde{Z}]^T W_{n+1}^{-1} [\tilde{Z}] \in \mathcal{M}_r \\ U_{n+1} = P_\alpha^V + [V^*] D_\alpha \big(D_{\mathfrak{m}}^{-1} + D_\alpha - D_V\big)^{-1} D_\alpha [V^*]^T \in \mathcal{M}_p \\ \hat{K}_{n+1} = L_{n+1}(P_\alpha^V)^{-1}[V^*](D_\alpha - D_\alpha(D_{\mathfrak{m}}^{-1} + D_\alpha)^{-1} D_\alpha)[\tilde{Z}]^T W_{n+1}^{-1} \\ \hat{X}_{n+1}^+ = \hat{X}_{n+1}^- + \hat{K}_{n+1}(Z_{n+1} - E_\alpha(H(\hat{X}_{n+1}^*))) \\ P_{n+1}^+ = L_{n+1} U_{n+1}^{-1} L_{n+1}^T. \end{cases} \tag{3.14c}$$

Note that in practice $\hat{K}_{n+1}$ does not need to be assembled, as directly computing the product $[\tilde{Z}]^T W_{n+1}^{-1}(Z_{n+1} - E_\alpha(H(\hat{X}^*_{n+1})))$ in the correction is more effective.

Remark 3.5. It is still possible to introduce a matrix $\{HL\}_n$ in this algorithm to reproduce the structure of the algorithm (3.8). Nevertheless, the expression is not as simple due to the fact that the $Z^{(i)}$ particles do not satisfy the construction rules. In fact, we obtain

$$\{HL\}_{n+1} = [\tilde{Z}](\mathbb{1} - D_\alpha(D_{\mathfrak{m}}^{-1} + D_\alpha)^{-1})(\mathbb{1} + D_V(D_{\mathfrak{m}}^{-1} + D_\alpha - D_V)^{-1})D_\alpha[V^*]^T,$$

giving as in (3.8c)

$$\hat{X}^+_{n+1} = \hat{X}^-_{n+1} + L_{n+1}U_{n+1}^{-1}\{HL\}_{n+1}^T W_{n+1}^{-1}(Z_{n+1} - E_\alpha(Z^*_{n+1})).$$

As a conclusion, this generalizes the reduced order UKF formulation proposed in [19] for any choice of sigma-points. Nevertheless, this general form requires an SVD, an additional sampling and a more complex computation of the *a posteriori* covariance. This increased complexity can be justified in practice if

- the interpolation accuracy through some specific nonlinear operators is improved by some well-suited choices of sigma-points;
- some probability density functions are better described by a richer sampling procedure, see in particular [13].

3.4. **Continuous-time algorithm**

One formal drawback of the UKF formulation is that it is time-discrete in essence due to the resampling procedures – namely, it is difficult to define a natural continuous-time version. Nevertheless, a continuous-time extension was proposed in [21] based on a formal limiting procedure with respect to the time-step as is classical in Kalman theory. Let us define the targeted continuous-time dynamical system and observation process in the very general form

$$\begin{cases} \dot{X} = A(X,t) \\ Z = H(X,t) + \chi \end{cases} \tag{3.15}$$

where χ denotes a white noise of covariance W asymptotically related to the discrete-time Gaussian error according to the classical rule

$$W = \Delta t\, W_n.$$

Introducing the sigma-points defined for all time by

$$\hat{X}^{(i)}(t) = \hat{X}(t) + \sqrt{P(t)}I^{(i)}, \tag{3.16}$$

the continuous algorithm is given by [21]

$$\begin{cases} \dot{\hat{X}} = [A(\hat{X}^*,t)]D_\alpha(1) + \hat{K}(Z - [H(\hat{X}^*,t)]D_\alpha(1)) \\ \dot{P} = [\hat{X}^*]D_V[A(\hat{X}^*,t)]^T + [A(\hat{X}^*,t)]D_V[\hat{X}^*]^T - \hat{K}W\hat{K}^T \end{cases} \tag{3.17}$$

where D_V is defined by (3.12) and

$$\hat{K}(t) = [\hat{X}^*]D_V[H(\hat{X}^*,t)]^T W^{-1}. \tag{3.18}$$

We note that (3.17) does not characterize a classical linear filter form

$$\dot{\hat{X}} = A(\hat{X},t) + \hat{K}(Z - H(\hat{X},t)).$$

In fact, the UKF is a non-linear filter of the most general form $\dot{\hat{X}} = \phi(\hat{X}, Z)$.

We can now formulate the reduced-order version of the continuous-time UKF as we did in discrete-time. Still focusing on the simplex case, we obtain the reduced-order dynamics

$$\begin{cases} \dot{\hat{X}} = [A(\hat{X}^*, t)]D_\alpha(1) + \hat{K}(Z - [H(\hat{X}^*, t)]D_\alpha(1)) \\ \dot{L} = [A(\hat{X}^*, t)]D_\alpha[V^*]^T - \frac{1}{2}L(P_\alpha^V)^{-1}\{HL\}^T W^{-1}\{HL\} \end{cases} \tag{3.19}$$

with

$$\begin{cases} \hat{X}^{(i)} = \hat{X} + L(P_\alpha^V)^{-\frac{1}{2}} I^{(i)} \\ \{HL\} = [H(\hat{X}^*, t)]D_\alpha[V^*]^T \\ \hat{K} = L(P_\alpha^V)^{-1}\{HL\}^T W^{-1}. \end{cases} \tag{3.20}$$

We note, indeed, that the sigma-points are constructed so that the identity $L = [\hat{X}^*]D_\alpha[V^*]^T$ holds. Then, with the covariance decomposition $P(t) = L(P_\alpha^V)^{-1}L^T$ we can check that (3.16)–(3.18) directly follow. We point out that this factorization of the covariance is different from the square root factorization of [21] which is not adapted to reduced-order filtering.

4. Application to parameter identification in large dimensional systems

We now assume that the system we are considering is the result of the discretization of a (non-linear) infinite-dimensional system formally written in a dynamical system form $\dot{x} = \mathcal{A}(x, t)$. The aim of numerical simulations is clearly to approximate this system by an "*in-silico*" version using an adequate space discretization

$$\dot{X} = A(X, t),$$

and time discretization

$$X_{n+1} = A_{n+1|n}(X_n, t_n),$$

meaning that X_n is expected to converge when the space and time discretization steps, Δh and Δt respectively, tend to 0. This classical approach may encounter some difficulties due to a lack of knowledge on the system concerning the initial conditions $x(0)$ and the dynamical operator $\mathcal{A}$. In this paper we restrict the model uncertainties considered to some parameter uncertainties globally collected in a variable denoted by θ. We point out that this modeling of the uncertainties can take into account boundary conditions uncertainties – which are common – using *e.g.* well-suited parameterized Robin boundary conditions [6]. Some of these parameters are spatially distributed, but their spatial discretization is limited since we cannot expect to obtain the same level of spatial localization in the identification as for the state variable x. In fact, the difference between classical numerical analysis and data assimilation is that, here, we expect to use some available measurements $Z = \mathcal{H}x + \chi$ to recover the convergence of the X_n to x despite the uncertainties, and yet, identifiability limits the space discretization of the unknown parameters with respect to the measurements' completeness. Therefore, for given observations $\theta \in \mathbb{R}^p$ is supposed to be limited in size – typically less than 100 – while the dimension N of X is in the range 10^4–10^6. Then, we consider a discretized model in the form

$$\begin{cases} X_{n+1} = A_{n+1|n}(X_n, \theta, t_n) \\ X_{n=0} = X_0 + \zeta^x \\ \theta = \theta_0 + \zeta^\theta \end{cases}$$

and we want to estimate the unknown quantities ζ^x and ζ^θ with the measurements

$$Z_n = H_n X_n + \chi_n,$$

where χ_n contains measurement errors but also space and time discretization errors associated with the conversion of the real measurements in the approximate system variables. Once the problem is thus posed, it is

classical in filtering-based estimation to consider this joint state-parameter system as a state estimation problem for the following augmented system

$$\begin{cases} X^e_{n+1} = \begin{pmatrix} X_{n+1} \\ \theta_{n+1} \end{pmatrix} = \begin{pmatrix} A_{n+1|n}(X_n, \theta_n, t_n) \\ \theta_n \end{pmatrix} = A^e_{n+1|n}(X^e_n, t_n) \\ X^e_{n=0} = X^e_0 + \zeta^e = \begin{pmatrix} X_0 + \zeta^X \\ \theta_0 + \zeta^\theta \end{pmatrix}. \end{cases}$$

In theory, a Kalman filter can be directly derived for the above formulation but the dimension of the initial state X makes this filter intractable in practice. In [17] we developed joint state-parameter filters using a reduced-order filter restricted to the parameter space. In order to do so, we took advantage of the fact that the physical system underlying the state variable X allows to use effective Luenberger filters – also called forward nudging in [4] – to estimate the state part and circumscribe the estimation error on the parameter space where we apply reduced-order Kalman filters. This method was fully described and analyzed in the case of linear systems and extended to non-linear systems using a reduced-order Extended Kalman Filter (EKF) version. Here, our objective is now to use our reduced-order UKF to handle more complex non-linearities.

Let us suppose that we have formulated an asymptotically stable state estimation, and chosen an adequate time-discretization scheme respecting this stability. Here, to simplify the presentation, we consider an explicit form, but everything would also work with implicit schemes. Then, our stable discretization of the state estimator with perfectly known parameters is

$$\bar{X}_{n+1} = A_{n+1|n}(\bar{X}_n, \theta) + \bar{K}^X_{n+1|n}(Z_n - H(\bar{X}_n)),$$

with initial condition $\bar{X}_{n=0} = X_0$, namely, the *a priori* initial state. This – together with zero dynamics for the parameter vector – gives the augmented system $\bar{X}^e$ that we will track in our observer approach, and the only uncertainty in this system corresponds to the parameter initial condition. Therefore, in the reduced rank strategy, it is reasonable to consider an SVD approximation of the covariance matrix of rank equal to the dimension of the parameter space, and this will be substantiated by a detailed mathematical analysis below.

4.1. **Formulation**

When applying the above reduced rank algorithm to the augmented system, the equations can be decomposed on the state and parameter components. In the sequel we focus on simplex-based reduced-order UKF to simplify the equations, but similar derivations could be followed in the general case. Recalling that the observation operator only applies on the state part, this gives the following algorithm:

- **Sampling:**

$$\begin{cases} C_n = \sqrt{U_n^{-1}} \\ \hat{X}^{(i)+}_n = \hat{X}^+_n + L^X_n C^T_n I^{(i)}, \quad 1 \le i \le p+1 \\ \hat{\theta}^{(i)+}_n = \hat{\theta}^+_n + L^\theta_n C^T_n I^{(i)}, \quad 1 \le i \le p+1; \end{cases} \tag{4.1a}$$

- **Prediction:**

$$\begin{cases} \hat{X}^{(i)-}_{n+1} = A_{n+1|n}(\hat{X}^{(i)+}_n, \hat{\theta}^{(i)+}_n) + \bar{K}^X_{n+1|n}(Z_n - H(\hat{X}^{(i)+}_n)) \\ \hat{X}^-_{n+1} = E_\alpha(\hat{X}^{*-}_{n+1}) \\ \hat{\theta}^-_{n+1} = \hat{\theta}^+_n; \end{cases} \tag{4.1b}$$

- **Correction:**

$$\begin{cases} L^X_{n+1} = [\hat{X}^{*-}_{n+1}] D_\alpha [V^*]^T \in \mathcal{M}_{N,p} \\ L^\theta_{n+1} = [\theta^{*-}_{n+1}] D_\alpha [V^*]^T \in \mathcal{M}_p \\ Z^{(i)}_{n+1} = H(X^{(i)-}_{n+1}) \\ \{HL^X\}_{n+1} = [Z^{(i)}_{n+1}] D_\alpha [V^*]^T \\ U_{n+1} = P^V_\alpha + \{HL^X\}^T_{n+1} W^{-1}_{n+1} \{HL^X\}_{n+1} \in \mathcal{M}_p \\ \hat{X}^+_{n+1} = \hat{X}^-_{n+1} + L^X_{n+1} U^{-1}_{n+1} \{HL^X\}^T_{n+1} W^{-1}_{n+1} (Z_{n+1} - E_\alpha(Z^*_{n+1})) \\ \hat{\theta}^+_{n+1} = \hat{\theta}^-_{n+1} + L^\theta_{n+1} U^{-1}_{n+1} \{HL^X\}^T_{n+1} W^{-1}_{n+1} (Z_{n+1} - E_\alpha(Z^*_{n+1})). \end{cases} \tag{4.1c}$$

For the initialization we take

$$L^\theta_0 = \mathbb{1}, \quad L^X_0 = 0, \quad U_0 = \mathrm{Cov}(\zeta_\theta)^{-1},$$

so that the initial covariance is concentrated on the parameters, as intended, with

$$L^\theta_0 U_0^{-1} (L^\theta_0)^T = \mathrm{Cov}(\zeta_\theta).$$

4.2. **Analysis**

4.2.1. *Formulation with linear operators*

We analyze the performance of the algorithm in terms of linearized error. To find the equations driving the linearized error, it is equivalent here to derive the error system in the case of linearized dynamics and observation operator:

$$X_{n+1} = A_{n+1|n} X_n + B_{n+1|n} \theta_n + R_n,$$

and

$$Z_n = H_n X_n + \chi_n.$$

Then the state estimator is

$$\begin{aligned} \bar{X}_{n+1} &= A_{n+1|n} \bar{X}_n + B_{n+1|n} \theta + R_n + \bar{K}^X_{n+1|n} (Z_n - H_n \bar{X}_n) \\ &= A^K_{n+1|n} \bar{X}_n + B_{n+1|n} \theta + R^Z_n, \end{aligned} \tag{4.2}$$

with $A^K_{n+1|n} = A_{n+1|n} - \bar{K}^X_{n+1|n} H_n$ and $R^Z_n = R_n + \bar{K}^X_{n+1|n} Z_n$.

In this case, the observer equations simplify due to the linearity of the operators. During the prediction step, since the empirical mean commutes with the dynamics operator we thus obtain

$$\begin{cases} \hat{X}^-_{n+1} = A^K_{n+1|n} \hat{X}^+_n + B_{n+1|n} \hat{\theta}^+_n + R^Z_n \\ \hat{\theta}^-_{n+1} = \hat{\theta}^+_{n+1}. \end{cases}$$

Then, at the correction stage, since the observation operator is also linear we have $\{HL^X\}_{n+1} = H_{n+1} L^X_{n+1}$, hence

$$\begin{cases} \hat{X}^+_{n+1} = \hat{X}^-_{n+1} + L^X_{n+1} U^{-1}_{n+1} (L^X_{n+1})^T H^T_{n+1} W^{-1}_{n+1} (Z_{n+1} - H_{n+1} \hat{X}^-_{n+1}) \\ \hat{\theta}^+_{n+1} = \hat{\theta}^-_{n+1} + L^\theta_{n+1} U^{-1}_{n+1} (L^X_{n+1})^T H^T_{n+1} W^{-1}_{n+1} (Z_{n+1} - H_{n+1} \hat{X}^-_{n+1}) \end{cases}$$

where L_n^X has the following dynamics

$$\begin{aligned}
L_{n+1}^X &= [\hat{X}_{n+1}^{*-}]D_\alpha[V^*]^T \\
&= [A_{n+1|n}^K \hat{X}_n^{*+} + B\hat{\theta}_n^{*+}]D_\alpha[V^*]^T + R_n^Z(1)^T D_\alpha[V^*]^T \\
&= A_{n+1|n}^K \hat{X}_n^+(1)^T D_\alpha[V^*]^T + A_{n+1|n}^K L_n^X C_n^T[I^*]D_\alpha[V^*]^T \\
&\quad + B_{n+1|n}\hat{\theta}_n^+(1)^T D_\alpha[V^*]^T + B_{n+1|n}L_n^\theta C_n^T[I^*]D_\alpha[V^*]^T \\
&= (A_{n+1|n}^K L_n^X + B_{n+1|n}L_n^\theta)C_n^T[I^*]D_\alpha[V^*]^T,
\end{aligned}$$

meaning

$$L_{n+1}^X = (A_{n+1|n}^K L_n^X + B_{n+1|n}L_n^\theta)C_n^T P_\alpha^{IV}, \tag{4.3}$$

with

$$P_\alpha^{IV} = [I^*]D_\alpha[V^*]^T,$$

and L_n^θ verifies

$$L_{n+1}^\theta = L_n^\theta C_n^T[I^*]D_\alpha[V^*]^T = L_n^\theta C_n^T P_\alpha^{IV}. \tag{4.4}$$

Finally, we also have

$$U_{n+1} = P_\alpha^V + (L_{n+1}^X)^T H_{n+1}^T W_{n+1}^{-1} H_{n+1} L_{n+1}^X.$$

4.2.2. *Error analysis*

Let us consider the errors defined by

$$\begin{cases} \tilde{X}_n^- = \bar{X}_n^- - \hat{X}_n^- \\ \tilde{\theta}_n^- = \theta - \hat{\theta}_n^-. \end{cases}$$

We have,

$$\begin{aligned}
\tilde{X}_{n+1}^- &= A_{n+1|n}^K \bar{X}_n + B_{n+1|n}\theta + R_n^Z - A_{n+1|n}^K \hat{X}_n^+ - B_{n+1|n}\hat{\theta}_n^+ - R_n^Z \\
&= A_{n+1|n}^K \bar{X}_n + B_{n+1|n}\theta - A_{n+1|n}^K \hat{X}_n^- - A_{n+1|n}^K L_n^X U_n^{-1} L_n^{X\,T} H_n^T W_n^{-1}(Z_n - H_n\hat{X}_n^-) \\
&\quad - B_{n+1|n}\hat{\theta}_n^- - B_{n+1|n}L_n^\theta U_n^{-1} L_n^{X\,T} H_n^T W_n^{-1}(Z_n - H_n\hat{X}_n^-) \\
&= A_{n+1|n}^K \tilde{X}_n^- + B_{n+1|n}\tilde{\theta}_n^- - L_{n+1}^X (P_\alpha^{IV})^{-1} C_n L_n^{X\,T} H_n^T W_n^{-1}(Z_n - H_n\hat{X}_n^-),
\end{aligned}$$

and

$$\begin{aligned}
\tilde{\theta}_{n+1}^- &= \theta - \hat{\theta}_n^- - L_n^\theta U_n^{-1} L_n^{X\,T} H_n^T W_n^{-1}(Z_n - H_n\hat{X}_n^-) \\
&= \tilde{\theta}_n^- - L_{n+1}^\theta (P_\alpha^{IV})^{-1} C_n L_n^{X\,T} H_n^T W_n^{-1}(Z_n - H_n\hat{X}_n^-).
\end{aligned}$$

Noting that $L_n^X (L_n^\theta)^{-1}$ follows the dynamics

$$L_{n+1}^X (L_{n+1}^\theta)^{-1} = A_{n+1|n}^K L_n^X (L_n^\theta)^{-1} + B_{n+1|n}, \tag{4.5}$$

which is the dynamics of the sensitivity $\frac{\partial \bar{X}_n}{\partial \theta}$ of the state estimate with respect to the parameter vector (4.2), we introduce the change of variables inspired from [17,24]

$$\eta_n = \tilde{X}_n^- - L_n^X (L^\theta{}_n)^{-1}\tilde{\theta}_n, \tag{4.6}$$

and we have

$$\begin{aligned}\eta_{n+1} &= A^K_{n+1|n}\tilde{X}^-_n + B_{n+1|n}\tilde{\theta}^-_n - L^X_{n+1}(L^\theta_{n+1})^{-1}\tilde{\theta}^-_n \\ &= A^K_{n+1|n}\tilde{X}^-_n + B_{n+1|n}\tilde{\theta}^-_n - \big(A^K_{n+1|n}L^X_n(L^\theta_n)^{-1} + B_{n+1|n}\big)\tilde{\theta}^-_n \\ &= A^K_{n+1|n}\,\eta_n. \end{aligned} \tag{4.7}$$

Therefore, η_n follows exactly the homogeneous dynamics which we assumed to be asymptotically stable.

Defining $P^{\theta-}_n = L^\theta_n(P^V_\alpha)^{-1}{L^\theta_n}^T$, the dynamics of $\tilde{\theta}^-_n$ gives then

$$\begin{aligned}(P^{\theta-}_{n+1})^{-1}\tilde{\theta}^-_{n+1} &= (L^\theta_{n+1})^{-T}P^V_\alpha(L^\theta_{n+1})^{-1}\tilde{\theta}^-_n \\ &\quad - (L^\theta_{n+1})^{-T}P^V_\alpha(P^{IV}_\alpha)^{-1}C_n{L^X_n}^T H^T_n W^{-1}_n\big(H_n(\eta_n + L^X_n(L^\theta{}_n)^{-1}\tilde{\theta}_n) + \epsilon_n + \chi_n\big),\end{aligned}$$

with $\epsilon_n = H_n(X_n - \bar{X}_n)$. In order to focus on the homogeneous part of the dynamics, we denote by $\ell(\eta_n, \chi_n, \epsilon_n)$ the linear terms in these variables appearing in the equations. We can then infer that

$$\begin{aligned}(P^{\theta-}_{n+1})^{-1}\tilde{\theta}^-_{n+1} &= (L^\theta_n)^{-T}C_n^{-1}(P^{IV}_\alpha)^{-T}P^V_\alpha(P^{IV}_\alpha)^{-1}C_n^{-T}(L^\theta_n)^{-1}\tilde{\theta}^-_n + \ell(\eta_n,\chi_n,\epsilon_n) \\ &\quad - (L^\theta_n)^{-T}C_n^{-1}(P^{IV}_\alpha)^{-T}P^V_\alpha(P^{IV}_\alpha)^{-1}C_n{L^X_n}^T H^T_n W^{-1}_n H_n L^X_n(L^\theta{}_n)^{-1}\tilde{\theta}_n \\ &= \Big((L^\theta_n)^{-T}U_n(L^\theta_n)^{-1} - (L^\theta_n)^{-T}{L^X_n}^T H^T_n W^{-1}_n H_n L^X_n(L^\theta{}_n)^{-1}\Big)\tilde{\theta}_n + \ell(\eta_n,\chi_n,\epsilon_n),\end{aligned}$$

recognizing that Proposition 3.1 applied to the $(I^{(i)})$ entails $(P^{IV}_\alpha)^{-T}P^V_\alpha(P^{IV}_\alpha)^{-1} = \mathbb{1}$. Finally, using the definition of U_n we get

$$(P^{\theta-}_{n+1})^{-1}\tilde{\theta}^-_{n+1} = (P^{\theta-}_n)^{-1}\tilde{\theta}^-_n + \ell(\eta_n,\chi_n,\epsilon_n), \tag{4.8}$$

with $\ell(\eta_n,\chi_n,\epsilon_n) = -\big(L^X_n(L^\theta_n)^{-1}\big)^T H^T_n W^{-1}_n(H_n\eta_n + \epsilon_n + \chi_n)$. Therefore, the homogeneous part of (4.8) shows that $(P^{\theta-}_n)^{-1}\tilde{\theta}^-_n$ is conserved over time up to the perturbation terms created by ℓ that we know are small or decreasing to 0.

We remark that the equations describing the errors (4.7) and (4.8) are the discrete-time versions of those obtained in continuous time in [17] for a reduced-order joint parameter-state estimation procedure based on Extended Kalman Filtering (EKF) instead of UKF. This formal similarity is substantiated by the fact that EKF and UKF are equivalent for linear systems, but we emphasize that the result obtained here is interesting in that we have obtained *exact* discrete error equations. This equivalence also shows that in the case of the reduced-order EKF of [17] we would obtain the same exact discrete error equations, namely, (4.7) and (4.8) with $(P^{\theta-}_{n+1})^{-1}$ interpreted as the *a priori* covariance of the parameter vectors, and the same role in the expressions of η and ℓ for the corresponding sensitivity of the state estimator with respect to the parameter.

Therefore, in this convergence analysis, the vanishing of the auxiliary variable η_n is based on the stability of the state estimation operator $A^K_{n+1|n}$, and then the parameter error $\tilde{\theta}_n$ is controlled by the evolution of $P^{\theta-}_n$, which satisfies

$$\begin{aligned}(P^{\theta-}_{n+1})^{-1} &= \big(L^\theta_{n+1}(P^V_\alpha)^{-1}{L^\theta_{n+1}}^T\big)^{-1} \\ &= (L^\theta_n)^{-T}C_n^{-1}(P^{IV}_\alpha)^{-T}P^V_\alpha(P^{IV}_\alpha)^{-1}C_n^{-T}(L^\theta_n)^{-1} \\ &= (L^\theta_n)^{-T}U_n(L^\theta_n)^{-1} \\ &= (P^{\theta-}_n)^{-1} + \big(L^X_n(L^\theta_n)^{-1}\big)^T H^T_n W^{-1}_n H_n\big(L^X_n(L^\theta_n)^{-1}\big).\end{aligned} \tag{4.9}$$

We recognize in this equation the discrete-time evolution of the classical sensitivity grammian

$$P^{\theta-}_{n+1} = U_0 + \sum_{k=1}^{n}\big(L^X_k(L^\theta_k)^{-1}\big)^T H^T_k W^{-1}_k H_k\big(L^X_k(L^\theta_k)^{-1}\big),$$

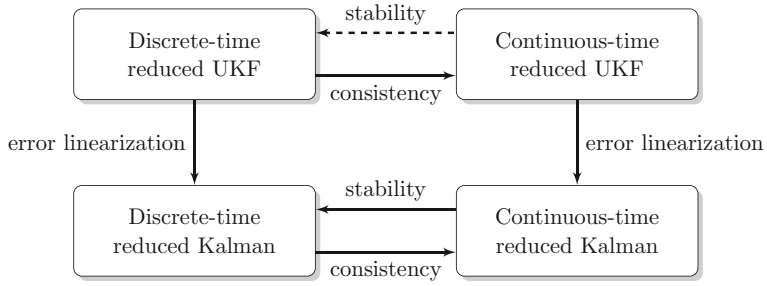

FIGURE 1. Relations between reduced filtering procedures for parameter estimation.

as it arose in continuous time in [17]. Note that this grammian can be numerically computed in the estimation process to assess the identifiability, hence the convergence.

4.3. **Continuous-time limit**

In this case, the reduced-order continuous-time equations (3.19) take the specialized form

$$\begin{cases} \dot{\hat{X}} = [A(\hat{X}^*, \theta^*, t)]D_\alpha(1) + L^X(P_\alpha^V)^{-1}\{HL\}^T W^{-1}(Z - [H(\hat{X}^*, t)]D_\alpha(1)) \\ \dot{\hat{\theta}} = L^\theta(P_\alpha^V)^{-1}\{HL\}^T W^{-1}(Z - [H(\hat{X}^*, t)]D_\alpha(1)) \\ \dot{L^X} = [A(\hat{X}^*, \theta^*, t)]D_\alpha[V^*]^T - \frac{1}{2}L^X(P_\alpha^V)^{-1}\{HL^X\}^T W^{-1}\{HL^X\} \\ \dot{L^\theta} = -\frac{1}{2}L^\theta(P_\alpha^V)^{-1}\{HL^X\}^T W^{-1}\{HL^X\}. \end{cases} \tag{4.10}$$

This gives in the linearization

$$\begin{cases} \dot{\hat{X}} = A_K\hat{X} + B\theta + R_Z + L^X(P_\alpha^V)^{-1}{L^X}^T H^T W^{-1}(Z - H\hat{X}) \\ \dot{\hat{\theta}} = L^\theta(P_\alpha^V)^{-1}{L^X}^T H^T W^{-1}(Z - H\hat{X}) \\ \dot{L^X} = AL^X + BL^\theta - \frac{1}{2}L^X(P_\alpha^V)^{-1}{L^X}^T H^T W^{-1} HL^X \\ \dot{L^\theta} = -\frac{1}{2}L^\theta(P_\alpha^V)^{-1}L^X H^T W^{-1} L^X \end{cases} \tag{4.11}$$

hence, we can easily verify that

$$\begin{cases} \dot{\hat{X}} = A_K\hat{X} + B\hat{\theta} + R_Z + \big(L^X(L^\theta)^{-1}\big)\big(L^\theta(P_\alpha^V)^{-1}{L^\theta}^T\big)\big(L^X(L^\theta)^{-1}\big)^T H^T W^{-1}(Z - H\hat{X}) \\ \dot{\hat{\theta}} = \big(L^\theta(P_\alpha^V)^{-1}{L^\theta}^T\big)\big(L^X(L^\theta)^{-1}\big)^T H^T W^{-1}(Z - H\hat{X}) \\ \dfrac{\mathrm{d}}{\mathrm{d}t}\big(L^X(L^\theta)^{-1}\big) = AL^X(L^\theta)^{-1} + B \\ \dfrac{\mathrm{d}}{\mathrm{d}t}\big((L^\theta)^{-T}P_\alpha^V(L^\theta)^{-1}\big) = \big(L^X(L^\theta)^{-1}\big)^T H^T W^{-1} H\big(L^X(L^\theta)^{-1}\big)^T \end{cases} \tag{4.12}$$

which are exactly the equations of the observer taking into account the parametric sensitivity and covariance as seen in [17].

We emphasize that we have now completed the substantiation of the relations summarized in Figure 1 for reduced filtering procedures applied to parameter estimation, in combination with previously designed effective state estimators. Namely, we have shown that:

- The discrete-time reduced-order UKF filter applied to parameter estimation corresponds to a time discretization scheme for a non-trivial continuous version described in (4.10). This discretization is consistent by construction – by the arguments of [21].

- On the bottom edge of the diagram we can easily establish a parallel consistency relation between discrete-time reduced-order Kalman filter applied to parameter estimation in linear systems and the continuous-time version analyzed in [17]. Furthermore, the stability of the linear estimator error (4.12) ensures the stability of the time discretization scheme provided by the discrete-time version.
- This continuous-time reduced-order UKF filter gives an asymptotically stable linearized error system (under the observability conditions stated in [17]), hence it is effective for perturbations of sufficiently small amplitudes. Therefore, the discrete-time reduced-order UKF filter applied to parameter estimation for the (continuous-time) reference system is justified by combining a standard stability-consistency argument with a linearized error analysis. This also justifies the dashed arrow in the diagram.

4.4. **Example**

In order to illustrate the efficiency of our strategy we will assess it with a test problem inspired from cardiac biomechanics. More specifically, our objective will be to estimate contractility parameters in a model of cardiac mechanical contraction using velocity measurements. The motivation of such a test lies in an increasing demand from clinicians to quantify the degree of damage in cardiac tissues caused by an infarct, using available measurements such as tagged MRI [2]. This test problem was already considered in [17,18] and relies on the classical dynamical principle of equilibrium written in a weak form

$$\int_\Omega \rho \frac{\mathrm{d}\underline{\dot{y}}}{\mathrm{d}t} \cdot \underline{v}\,\mathrm{d}\Omega = \int_\Omega \rho \underline{\dot{y}} \cdot \underline{v}\,\mathrm{d}\Omega, \quad \forall \underline{v} \tag{4.13}$$

$$\int_\Omega \rho \frac{\mathrm{d}\underline{\dot{y}}}{\mathrm{d}t} \cdot \underline{v}\,\mathrm{d}\Omega = -\int_\Omega \underline{\underline{\Sigma}}(\underline{y}, \underline{\dot{y}}) : \mathrm{d}_{\underline{y}}\underline{\underline{e}} \cdot \underline{v}\,\mathrm{d}\Omega + \int_\Omega \underline{f} \cdot \underline{v}\,\mathrm{d}\Omega, \quad \forall \underline{v}. \tag{4.14}$$

As a constitutive law defining the second Piola-Kirchhoff stress tensor $\underline{\underline{\Sigma}}$ in this formulation we will use:

- a hyperelastic material given by the Ciarlet-Geymonat potential

$$W^e(\underline{\underline{e}}) = \kappa_1(J_1 - 3) + \kappa_2(J_2 - 3) + \kappa_3(J - 1) - \kappa_3 \ln J,$$

  where $\underline{\underline{e}}$ denotes the nonlinear Green-Lagrange strain tensor and J_1, J_2, J the classical reduced invariants, see *e.g.* [14];
- a viscous term corresponding to the pseudo-potential

$$W^v(\underline{\underline{e}}) = \beta\ \mathrm{tr}(\underline{\underline{\dot{e}}}^2);$$

- a prestress term associated with the virtual work expression

$$F^{PS}(\underline{v}) = \sum_{1 \le i \le 17} \int_{\Omega_i^{AHA}} \theta_i \sigma_0 w(\|\underline{CM}\| - ct)\ \mathrm{tr}(\,\mathrm{d}_{\underline{y}}\underline{\underline{e}} \cdot \underline{v})\,\mathrm{d}\Omega,$$

  which carries the effect of an isotropic stress field created by the electrical activation, itself represented by the given profile w shown in Figure 3 resulting in a spherical wave of prestress of initiation center C and wave speed c. The quantity σ_0 denotes a constant scaling coefficient, and the parameters θ_i some dimensionless contractility values to be estimated in some predefined regions Ω_i^{AHA}. Typically, a normal value for a contractility coefficient would be 1, but the actual value may be reduced in some regions as the result of some pathological conditions such as ischemia and infarcts.

  The domain considered is shown in Figure 2, and its characteristic dimensions are comparable to those of a human left ventricle, which justifies the cardiac terminology "apex and base" to refer to the two extremities of the domain. The system is clamped over the planar surface at the base.

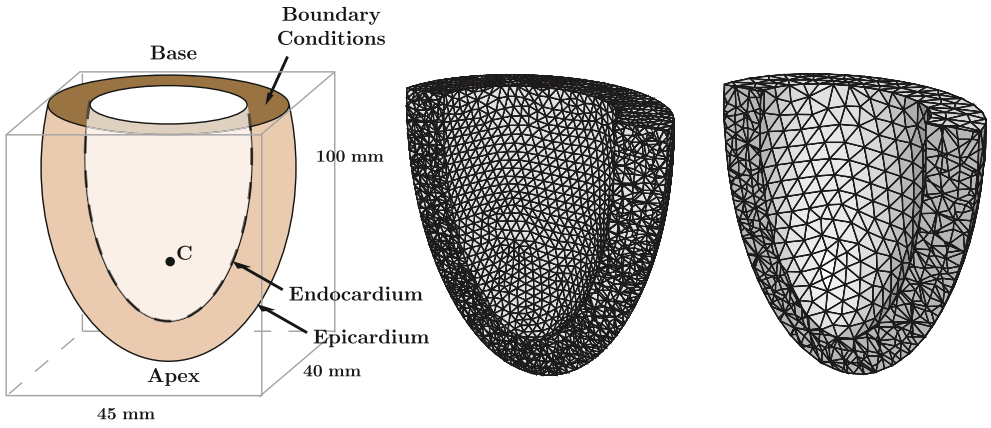

FIGURE 2. (a) Model geometry – (b) reference mesh – (c) observer mesh.

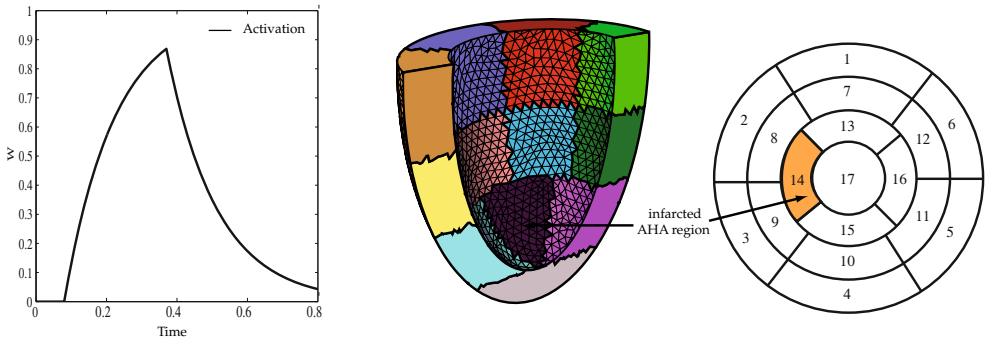

FIGURE 3. Electrical activation profile (left) – AHA subdivision (right).

Figures 2 and 3 also show:

- the parametric regions corresponding to the AHA-advocated subdivision [1];
- the meshes used in the numerical simulations, corresponding to about 30 000 degrees of freedom for the reference simulation – used to generate the synthetic measurements – and about 6000 degrees of freedom for the observer system;
- the initiation center C of the spherical activation wave (the wave speed is taken as $c = 0.5$ m·s^{-1}, which means that it takes 0.2 s for the wave to reach the base).

In our numerical simulations we have used the values

$$\rho = 10^3 \text{ kg·m}^{-3}, \quad \kappa_1 = 2.2 \times 10^3 \text{ Pa}, \quad \kappa_2 = 33 \text{ Pa}, \quad \kappa_3 = 2 \times 10^4 \text{ Pa}, \quad \beta = 0.68 \text{ Pa·s}, \tag{4.15}$$

which give some physiologically-relevant values for the resulting volume variations in the cavity. Standard $P1$-Lagrange finite elements were used for spatial discretization [3], and the energy-preserving Newmark scheme was employed for time discretization [14], with a time step of 1 ms.

Remark 4.1 (time discretization of the Luenberger observer)**.** As emphasized in the introduction of Section 4, we need the time discretization of the state observer to preserve the asymptotic stability intrinsically provided by the Luenberger filter in continuous time. In our numerical experiments we noted that a simple trapezoidal rule consistent with the Newmark scheme is adequate when the system contains sufficient internal damping –

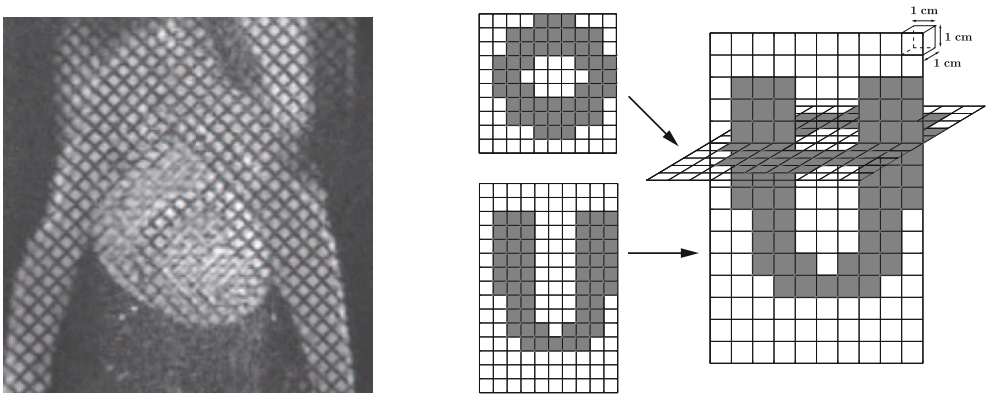

FIGURE 4. Tagged MRI (left) – Measurement cells in synthetic data (right).

as is the case for biological tissues – and provided the time step is reasonably small. In other cases some carefully devised numerical viscosity may be used to recover the asymptotic stability as in [8].

In the present case, the modeled measurements are assumed to be given as in [17] by velocities measured in a subpart Ω_m of the domain Ω and sampled by using weight functions $(s_i)_{i=1}^q$ defined on q non-overlapping "measurement cells" within Ω_m. Namely, $\mathcal{H}x = (0\ \mathcal{H}^v)(\underline{y}\ \underline{\dot{y}})^T$ consists of the q three-dimensional vectors given by

$$\underline{z}_i = \int_{\Omega_m} s_i\, \underline{\dot{y}}\, \mathrm{d}\Omega,$$

with normalized sampling functions. The measurement cells are visualized Figure 4 and their sizes typically correspond to tagged-MRI sampling. We point out that displacement measurements could also be considered using a more sophisticated state estimator as described in [18].

As a Luenberger state observer we used the filtering procedure inspired from the DVF (Direct Velocity Feedback) control strategy, as proposed and analyzed in [17]. For the above measurements, this gives the following variational form for the filtering term

$$\gamma \sum_{i=1}^{q} \left(\underline{z}_i - \int_{\Omega_m} s_i\, \underline{\dot{\hat{y}}}\, d\Omega \right) \cdot \int_{\Omega_m} s_i\, \underline{v}\, \mathrm{d}\Omega, \quad \forall \underline{v},$$

with straightforward finite element discretization counterparts in the form $\gamma H^T(Z - H\hat{X})$.

4.4.1. *Test with perfect data*

We start by assessing the procedure using perfect synthetic data produced with the model in which the contractility is reduced to 0.5 in AHA region 14. Then we reinitiate the estimation procedure with all contractility values set to 1, and the only uncertainty in the estimation problem corresponds to parameter uncertainty. The results are displayed in Figure 5, and show the excellent performance of our method.

4.4.2. *Test with multiple uncertainty sources*

We now proceed to test the method with more realistic uncertainties. The measurement noise is assumed to be given by a white noise with standard deviation corresponding to 10% of maximum velocity for a reference sampling rate of 50 ms. When rescaled according to the actual computational time step this corresponds to a standard deviation of about 70% of the maximum velocity. The state initial condition error corresponds to a static displacement obtained by imposing a pressure of 10^3 Pa inside the cavity, which represents a typical

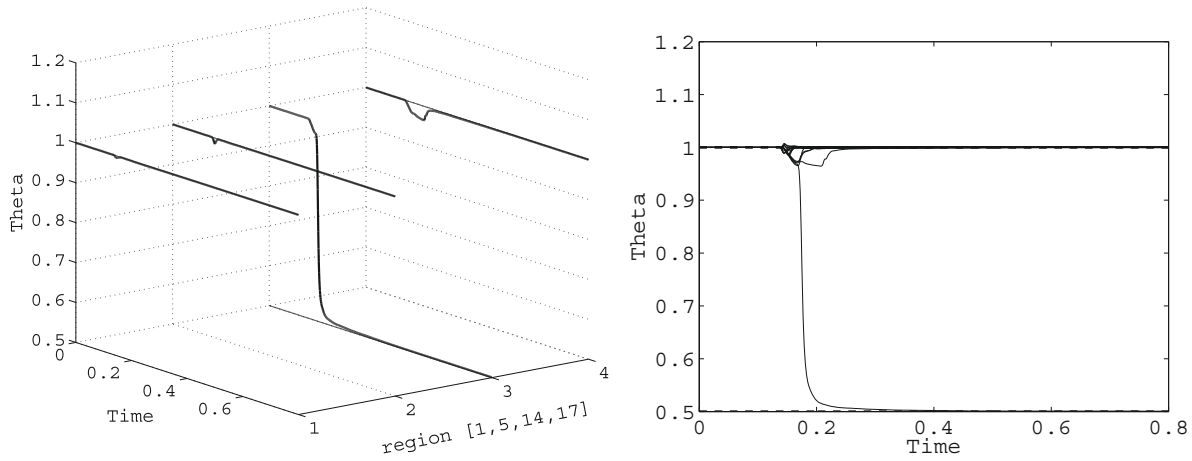

FIGURE 5. Estimation with perfect data: estimated values for four individual regions (left) and for all parameters (right).

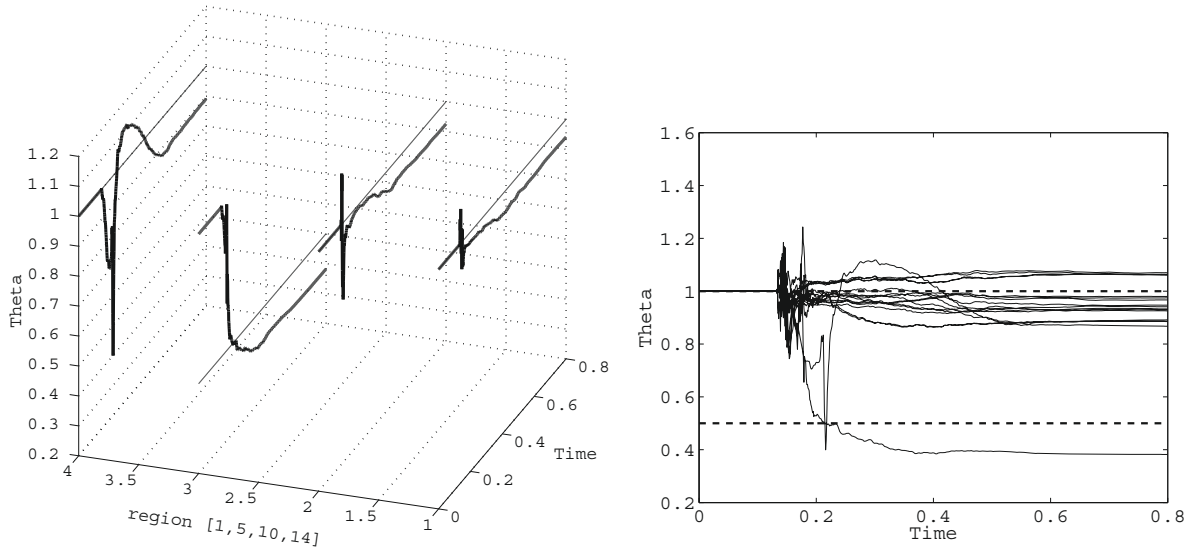

FIGURE 6. Estimation with multiple uncertainties: estimated values for four individual regions (left) and for all parameters (right).

value during the filling stage. The parametric estimation is started with a slight delay of $t_s = 0.1$ to let the state estimator perform a preliminary adjustment.

The results are displayed in Figure 6. The final estimation error is around 10%, which is clearly sufficient to allow the detection of the infarcted region. Nevertheless, the estimation performance is lower than with a linear system – see [17] – which can be explained by the strong coupling between state and parametric estimation in the nonlinear framework, namely, parametric sensitivity depends on the state, unlike in the linear case. Note also that the parameters and state are tightly combined in the prestress term above.

5. Concluding remarks

We have proposed a general reduced-order version of the UKF filtering procedure, allowing to perform tractable computations on a decomposition of the covariance matrices when the uncertainty space is of reduced size compared to a large-dimensional state space. Our reduced-order strategy is valid for any choice of sampling points, hence it encompasses and generalizes the SEIK approach associated with specific simplex distributions. In addition, we have formulated the corresponding continuous-time version of the procedure.

We emphasize that – beyond the expected improved accuracy compared to EKF – a major advantage of UKF lies in that no tangent operators are needed, hence the direct simulation codes can be called in a straightforward manner. Furthermore, as this results in the application of the transition and observation operators on the individual sigma-points, this approach is naturally adapted to parallel computations, with high scalability. We expect this advantage to be of particular value for large-dimensional systems in which the prediction phase is the most computationally intensive step in the filtering procedure.

We have also shown how this reduced-order UKF procedure can be applied to perform parameter estimation when an effective state observer is already available for the system, namely, to construct an adaptive observer in the spirit of [17]. In this context, we also performed a linearized error analysis of the complete state-parameter estimation. Finally, we demonstrated the applicability and performance of the whole methodology using a complex test problem inspired from cardiac biomechanics, as diagnosis assistance in medicine is one example of particularly interesting perspective for this approach.

Acknowledgements. The authors would like to thank Professor Patrick Le Tallec (École Polytechnique) for some interesting discussions regarding the relevance of this approach for mechanical systems.

References

[1] AHA/ACC/SNM, Standardization of cardiac tomographic imaging. *Circulation* **86** (1992) 338–339.
[2] L. Axel, A. Montillo and D. Kim, Tagged magnetic resonance imaging of the heart: a survey. *Med. Image Anal.* **9** (2005) 376.
[3] K.J. Bathe, *Finite Element Procedures*. Prentice-Hall, USA (1996).
[4] J. Blum, F.X. Le Dimet and I.M. Navon, Data assimilation for geophysical fluids, in *Handbook of Numerical Analysis: Computational Methods for the Atmosphere and the Oceans*, R. Temam and J. Tribbia Eds., Elsevier (2008).
[5] M.A. Cane, A. Kaplan, R.N. Miller, B. Tang, E.C. Hackert and A.J. Busalacchi, Mapping tropical Pacific sea level: Data assimilation via a reduced state space Kalman filter. *J. Geophys. Res.* **101** (1996) 22599–22618.
[6] S. Chaabane and M. Jaoua, Identification of Robin coefficients by the means of boundary measurements. *Inv. Prob.* **15** (1999) 1425–1438.
[7] D. Chapelle, P. Moireau and P. Le Tallec, Robust filtering for joint state-parameter estimation in distributed mechanical systems. *DCDS–A* **23** (2009) 65–84.
[8] S. Ervedoza and E. Zuazua, Uniformly exponentially stable approximations for a class of damped systems. *J. Math. Pures Appl.* **91** (2009) 20–48.
[9] I. Hoteit, D.-T. Pham and J. Blum, A simplified reduced order Kalman filtering and application to altimetric data assimilation in Tropical Pacific. *J. Mar. Syst.* **36** (2002) 101–127.
[10] S.J. Julier and J.K. Uhlmann, Reduced Sigma Point Filters for the Propagation of Means and Covariances through Nonlinear Transformations, in *Proc. of IEEE Am. Contr. Conf.*, Anchorage AK, USA, 8–10 May (2002) 887–892.
[11] S.J. Julier and J.K. Uhlmann, The Scaled Unscented Transformation, in *Proc. of IEEE Am. Contr. Conf.*, Anchorage AK, USA, 8–10 May (2002) 4555–4559.
[12] S. Julier, J. Uhlmann and H. Durrant-Whyte, A new approach for filtering nonlinear systems, in *American Control Conference* (1995) 1628–1632.
[13] S. Julier, J. Uhlmann and H. Durrant-Whyte, A new method for the nonlinear transformation of means and covariances in filter and estimators. *IEEE Trans. Automat. Contr.* **45** (2000) 447–482.
[14] P. Le Tallec, Numerical methods for nonlinear three-dimensional elasticity, in *Handbook of Numerical Analysis* **3**, P.G. Ciarlet and J.-L. Lions Eds., Elsevier (1994).
[15] T. Lefebvre, H. Bruyninckx and J. De Schuller, Comments on "A new method for the nonlinear transformation of means and covariances in filters and estimators" [and authors' reply]. *IEEE Trans. Automat. Contr.* **47** (2002) 1406– 1409.
[16] D.G. Luenberger, An introduction to observers. *IEEE Trans. Automat. Contr.* **16** (1971) 596–602.
[17] P. Moireau, D. Chapelle and P. Le Tallec, Joint state and parameter estimation for distributed mechanical systems. *Comput. Meth. Appl. Mech. Eng.* **197** (2008) 659–677.

[18] P. Moireau, D. Chapelle and P. Le Tallec, Filtering for distributed mechanical systems using position measurements: Perspectives in medical imaging. *Inv. Prob.* **25** (2009) 035010.
[19] D.-T. Pham, J. Verron and L. Gourdeau, Filtres de Kalman singuliers évolutifs pour l'assimilation de données en océanographie. *C. R. Acad. Sci. – Ser. IIA* **326** (1998) 255–260.
[20] D.T. Pham, J. Verron and M.C. Roubeaud, A singular evolutive extended Kalman filter for data assimilation in oceanography. *J. Marine Systems* **16** (1998) 323–341.
[21] S. Sarkka, On unscented Kalman filtering for state estimation of continuous-time nonlinear systems. *IEEE Trans. Automat. Contr.* **52** (2007) 1631–1641.
[22] D. Simon, *Optimal State Estimation: Kalman, H^∞, and Nonlinear Approaches*. Wiley-Interscience (2006).
[23] M. Wu and A.W. Smyth, Application of the unscented Kalman filter for real-time nonlinear structural system identification. *Struct. Contr. Health. Monit.* **14** (2006) 971–990.
[24] Q. Zhang and A. Clavel, Adaptive observer with exponential forgetting factor for linear time varying systems, in *Proceedings of the 40th IEEE Conference on Decision and Control* **4** (2001) 3886–3891.