14th International Conference on Information Fusion
Chicago, Illinois, USA, July 5-8, 2011

Marginalized Sigma-Point Filtering

Fredrik Sandblom
Electrical & Electronics Engineering
Volvo 3P - Product Development
Gothenburg, Sweden
Email: fredrik.sandblom.2@volvo.com

Lennart Svensson
Department of Signals and Systems
Chalmers University of Technology
Gothenburg, Sweden
Email: lennart.svensson@chalmers.se

***Abstract*—In this paper we present a method for estimating mean and covariance of a transformed Gaussian random variable. The method is based on evaluations of the transforming function and resembles the unscented transform or Gauss–Hermite integration in that aspect. However, the information provided by the evaluations is used in a Bayesian framework to form a posterior description of the transforming function. Estimates are then derived by marginalizing the function from the analytical expression of the mean and covariance. An estimation algorithm, based on the assumption that the transforming function is constructed by Hermite polynomials, is presented and compared to the cubature rule and the unscented transform. Contrary to the unscented transform, the resulting approximation of the covariance matrix are guaranteed to be positive-semidefinite and the algorithm performs much better than the cubature rule for the evaluated scenario.**

Keywords: Numerical integration, Sigma point filtering, Kalman filtering, Bayesian estimation, Moment matching.

I. Introduction

Calculating the mean and covariance of stochastic variables is central in many estimation tasks dealing with non-deterministic components. One example is the recursive state estimation carried out by nonlinear filters, where the posterior mean and covariance are used to characterize the distribution.

The general Bayesian solution to the state estimation problem involves integration of probability density functions, integrals which are rarely mathematically tractable. The family of *Gaussian filters* solves the recursive estimation problem under the assumption that the posterior distribution is approximately Gaussian. The equations used to compute the posterior mean and covariance under this assumption are those of the linear minimum mean squared error (LMMSE) estimator, which coincides with the well known Kalman filter for linear, Gaussian systems [1].

A variety of Gaussian filters have been proposed to cope with non-linear models [2], and the derivative-free filters [3], [4], [5], [6], [7] are particularly useful; with little or no adjustment, they can be applied to a wide range of problems. These filters use a transformed set of deterministically chosen points, often referred to as *sigma-points*, to calculate the mean and covariance directly from the propagated points. It has been shown in [8] that the unscented transform [3], [4] realizes the fully symmetric integration formula presented in [9], and an extensive analysis of the numerical integration perspective on Gaussian filters is given in [10].

The unscented transform calculates the noncentral second moment using the same integration rule used for calculating the mean, which can lead to covariance matrix estimates which are not positive-semidefinite. This behavior was overcome with the recent introduction of the cubature integration rule [7], which is a special case of the unscented transform that performs better than other methods of comparable complexity [7], [11]. Unfortunately, the robustness comes at the expense of using a less accurate integration rule. Another curiosity is that the mean and covariance are computed based on different assumptions on the underlying mapping.

In this paper we *model the transforming function as a stochastic process* and use the transformed sigma-points to learn the process. To be more explicit, the function is described as a linear combination of Hermite polynomials, for which expressions for the mean and covariance are well known. The coefficients are given a hierarchical prior and the posterior distribution of these coefficients is computed, conditioned on the transformed sigma-points. The desired mean and covariance can then be calculated analytically by marginalizing the influence of the coefficients.

There are several reasons to derive sigma-point algorithms using Bayesian techniques. First, the expression for the covariance matrix estimate is based on the analytical expression rather than a numerical approach, hence the estimate is always positive-semidefinite and the relation to the mean is clear. Second, the model assumptions become clearly visible through the prior distribution. Third, Bayesian methods are generally well performing in the sense that they are admissible under relatively loose assumptions [12] and that they are optimal when the performance is averaged over the prior. Finally, we know that the key to improve performance is the choice of the prior. Although designing a prior can be difficult, we believe such choices are better made explicitly rather than implicitly. To illustrate this, we present a family of priors that results in the cubature and the unscented transform rules. It is shown that the presented algorithm outperforms the cubature rule in the evaluated scenario, using a simple prior. More specifically, we appear to provide more robust covariance estimates, when the underlying polynomials are not completely linear.

II. Problem formulation

Consider a transformation, $g : \mathbb{R}^n \rightarrow \mathbb{R}^m$ and a stochastic variable $\mathbf{x} \in \mathbb{R}^n$ with probability density function

$$\mathbf{x} \sim \mathcal{N}(\boldsymbol{\mu}_\mathbf{x}, \boldsymbol{P}_\mathbf{x}),$$

where g, $\boldsymbol{\mu}_\mathbf{x}$ and $\boldsymbol{P}_\mathbf{x}$ are all known. We wish to calculate the mean and covariance of the transformed variable $\mathbf{y} \in \mathbb{R}^m$:

$$\mathbf{y} = g(\mathbf{x}).$$

These moments are given by the integral expressions

$$\mathbb{E}[\mathbf{y}] = \int_{\mathbb{R}^n} \mathcal{N}(\mathbf{x}\,; \boldsymbol{\mu}_\mathbf{x}, \boldsymbol{P}_\mathbf{x}) g(\mathbf{x}) d\mathbf{x} \tag{1}$$

$$\begin{aligned}\mathrm{Cov}(\mathbf{y}) &= \int_{\mathbb{R}^n} \mathcal{N}(\mathbf{x}; \boldsymbol{\mu}_\mathbf{x}, \boldsymbol{P}_\mathbf{x}) \big[g(\mathbf{x}) - \mathbb{E}[g(\mathbf{x})]\big] \big[g(\mathbf{x}) - \mathbb{E}[g(\mathbf{x})]\big]^T d\mathbf{x} \\ &= \int_{\mathbb{R}^n} \mathcal{N}(\mathbf{x}; \boldsymbol{\mu}_\mathbf{x}, \boldsymbol{P}_\mathbf{x}) g(\mathbf{x}) g(\mathbf{x})^T d\mathbf{x} - \mathbb{E}[\mathbf{y}]\mathbb{E}[\mathbf{y}]^T. \end{aligned} \tag{2}$$

Expressing the solutions to these integrals on a closed form is often impossible for transformations encountered in practice. Sigma-point methods provide approximate solutions to these integrals, and have proven to be useful with respect to both performance and simplicity.

A. The sigma-point approach

The family of sigma-point filters use integral approximations on the form of a weighted sum:

$$\int_{\mathbb{R}^n} \mathcal{N}(\mathbf{x}\,; \boldsymbol{\mu}_\mathbf{x}, \boldsymbol{P}_\mathbf{x}) g(\mathbf{x}) d\mathbf{x} \approx \sum_{i=0}^{2n} w_i g(\mathbf{x}^i). \tag{3}$$

The so-called sigma-points, $\{\mathbf{x}^0, \ldots, \mathbf{x}^{2n}\}$, and the associated weights, w_i, are chosen according to a deterministic scheme. For the unscented transform and the cubature rule, they are:

$$\mathbf{x}^0 = \quad \mathbb{E}[\mathbf{x}] \tag{4}$$

$$\mathbf{x}^i = \begin{cases} \mathbb{E}[\mathbf{x}] + \left(\sqrt{\frac{n}{(1-w_0)}\boldsymbol{P}_\mathbf{x}}\right)_i, & 1 \leq i \leq n \\ \mathbb{E}[\mathbf{x}] - \left(\sqrt{\frac{n}{(1-w_0)}\boldsymbol{P}_\mathbf{x}}\right)_{i-2n/2}, & n < i \leq 2n \end{cases} \tag{5}$$

$$w_i = \frac{1-w_0}{2n}, \tag{6}$$

where $i = 1, \ldots, 2n$ and $\left(\sqrt{\boldsymbol{P}_\mathbf{x}}\right)_i$ is the i^{th} column of the matrix square root such that $\sqrt{\boldsymbol{P}_\mathbf{x}}\sqrt{\boldsymbol{P}_\mathbf{x}}^T = \boldsymbol{P}_\mathbf{x}$. When $\mathbf{x}$ is Gaussian, the suggested setting for the unscented transform [4] is to use $w_0 = 1 - n/3$, whereas the cubature rule is obtained by setting $w_0 = 0$, effectively removing $\mathbf{x}^0$ from the set. This integral approximation strategy, applied to equation (1), yields the estimator

$$\mathbb{E}[g(\mathbf{x})] \approx \sum_{i=0}^{2n} w_i g(\mathbf{x}^i) \triangleq \bar{\mathbf{y}}. \tag{7}$$

The covariance matrix estimate, $\hat{\boldsymbol{P}}_\mathbf{y}$, is usually expressed in terms of the weighted sum of squares, but we prefer to view it on the form (2) to make the dual use of the integral approximation clear:

$$\begin{aligned}\mathrm{Cov}(\mathbf{y}) &\approx \sum_{i=0}^{2n} w_i [g(\mathbf{x}^i) - \bar{\mathbf{y}}][g(\mathbf{x}^i) - \bar{\mathbf{y}}]^T \\ &= \sum_{i=0}^{2n} w_i g(\mathbf{x}^i) g(\mathbf{x}^i)^T - \sum_{i=0}^{2n} w_i g(\mathbf{x}^i) \bar{\mathbf{y}}^T \\ &\quad - \bar{\mathbf{y}} \sum_{i=0}^{2n} w_i g(\mathbf{x}^i)^T + \sum_{i=0}^{2n} w_i \bar{\mathbf{y}} \bar{\mathbf{y}}^T \\ &= \sum_{i=0}^{2n} w_i g(\mathbf{x}^i) g(\mathbf{x}^i)^T - \bar{\mathbf{y}} \bar{\mathbf{y}}^T \triangleq \hat{\boldsymbol{P}}_\mathbf{y}. \end{aligned} \tag{8}$$

If the mean (7) is correctly calculated using a *minimum* number of points, the covariance matrix estimate (8) will in general not be exact. In fact, with negative weights it may not even be positive-semidefinite.

III. Proposed idea

Even though the transforming function g is known, we model it as a stochastic process with a prior distribution $\pi(g)$. Apart from the prior, the only available information is the evaluated points, $\chi = \left[\mathbf{x}^0, \ldots, \mathbf{x}^{2n}\right]$, and the function values at these points, $\mathbf{z} = \left[g(\mathbf{x}^0), \ldots, g(\mathbf{x}^{2n})\right]$. The knowledge about g is summarized in the posterior distribution $p(g|\mathbf{z}, \chi)$, from which we intend to compute the moments of interest.

The mean, expressed as a function of the transformation, g, is denoted by

$$\bar{\mathbf{y}}(g) = \int \mathcal{N}(\mathbf{x}\,; \boldsymbol{\mu}_\mathbf{x}, \boldsymbol{P}_\mathbf{x}) g(\mathbf{x}) d\mathbf{x}, \tag{9}$$

and the corresponding covariance matrix by

$$P_\mathbf{y}(g) = \int \mathcal{N}(\mathbf{x}\,; \boldsymbol{\mu}_\mathbf{x}, \boldsymbol{P}_\mathbf{x}) [g(\mathbf{x}) - \bar{\mathbf{y}}(g)][g(\mathbf{x}) - \bar{\mathbf{y}}(g)]^T d\mathbf{x}. \tag{10}$$

The expressions for the desired mean and covariance of $\mathbf{y}$, given $\mathbf{z}$ and χ, are given by marginalization:

$$\begin{aligned}\bar{\mathbf{y}}_\pi &= \mathbb{E}[\mathbf{y}|\mathbf{z}, \chi] \\ &= \int \bar{\mathbf{y}}(g) p(g|\mathbf{z}, \chi) dg \end{aligned} \tag{11}$$

$$\begin{aligned}\boldsymbol{P}_{\mathbf{y},\pi} &= \mathbb{E}\left[\boldsymbol{P}_\mathbf{y}(g) \,\middle|\mathbf{z}, \chi\right] \\ &= \int \boldsymbol{P}_\mathbf{y}(g) p(g|\mathbf{z}, \chi) dg. \end{aligned} \tag{12}$$

The idea is to use a prior for which the integrals in (11) and (12) have closed form solutions. In this paper we focus on one such prior, presented in Section IV, where g is assumed to belong to the family of Hermite polynomials.

IV. Hermite polynomials

We assume that the function g can be constructed from Hermite polynomials since it leads to very simple expressions for the mean, $\bar{\mathbf{y}}(g)$, and covariance, $P_{\mathbf{y}}(g)$. Additionally, it facilitates a comparison with other sigma-point methods, which typically calculate the integral (7) exactly for certain polynomials.

A. Scalar transformations

The transformation $g : \mathbb{R}^1 \to \mathbb{R}^1$ determined by a linear combination of Hermite polynomials up to order p can be written

$$y = \theta_0 H_0 + \sum_{k=1}^{p} \theta_k H_k(x),$$

where H_k is given by (52). Because of the properties of scaled Hermite polynomials [see Appendix A] it is trivial to calculate the expected value and the variance:

$$\begin{aligned} \mathbb{E}[y] &= \theta_0 \\ \text{Var}(y) &= \sum_{k=1}^{p} \theta_k^2 k! \,. \end{aligned}$$

For example, if $x \sim \mathcal{N}(0,1)$ and $y = x + x^2$, the expected value is 1 and the variance is $1+2 = 3$, as $y = H_0 + H_1 + H_2$.

B. Stochastic decoupling

The useful properties of the Hermite polynomials, described in Appendix A, only hold for multivariate variables if the elements are uncorrelated. Therefore, a stochastic decoupling procedure similar to the approach in [5] is proposed. Instead of studying

$$\mathbf{y} = g(\mathbf{x}), \quad \mathbf{x} \sim \mathcal{N}(\boldsymbol{\mu}_{\mathbf{x}}, \boldsymbol{P}_{\mathbf{x}}), \tag{13}$$

we introduce $\tilde{\mathbf{x}} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_n)$ and set

$$\mathbf{y} = \tilde{g}(\tilde{\mathbf{x}}) \triangleq g(\boldsymbol{\mu}_{\mathbf{x}} + \sqrt{\boldsymbol{P}_{\mathbf{x}}}\,\tilde{\mathbf{x}}), \tag{14}$$

which has the same distribution as the original $\mathbf{y}$ in (13). The algorithm described in Section V-D comprises this adaptation. In the following sections we assume $\mathbf{x} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ for simplicity.

C. The multidimensional transformation

A transformation $g : \mathbb{R}^n \to \mathbb{R}^m$ performed by a linear combination of base functions can be written

$$g(\mathbf{x}; \boldsymbol{\theta}) = \boldsymbol{\theta}^T \mathbf{h}(\mathbf{x}), \tag{15}$$

where the base functions enter the equation through

$$\mathbf{h}(\mathbf{x}) = \left[H_0,\ H_1(x_1), \ldots, H_p(x_1),\ H_1(x_2), \ldots, H_p(x_n)\right]^T. \tag{16}$$

We construct the weight matrix from the vectors $\boldsymbol{\theta}^{i,j}$, describing the transformation from x_i to y_j, and scalars θ_0^j:

$$\boldsymbol{\theta} = \begin{bmatrix} \theta_0^1 & \ldots & \theta_0^j & \ldots & \theta_0^m \\ \boldsymbol{\theta}^{1,1} & \ldots & \boldsymbol{\theta}^{1,j} & \ldots & \boldsymbol{\theta}^{1,m} \\ \vdots & & \vdots & & \vdots \\ \boldsymbol{\theta}^{n,1} & \ldots & \boldsymbol{\theta}^{n,j} & \ldots & \boldsymbol{\theta}^{n,m} \end{bmatrix}. \tag{17}$$

Consequently, $\boldsymbol{\theta}^j$, the j^{th} column of $\boldsymbol{\theta}$, defines the mapping from $\mathbf{x} \in \mathbb{R}^n$ to y_j over the base functions in $\mathbf{h}(\mathbf{x})$:

$$y_j = \theta_0^j H_0 + \sum_{i=1}^{n} \sum_{k=1}^{p} \boldsymbol{\theta}^{i,j}(k) H_k(x_i). \tag{18}$$

The vectors $\boldsymbol{\theta}^{i,j}$ are assumed to be independently generated from a hierarchic model:

$$\boldsymbol{\theta}^{i,j} \sim \mathcal{N}(0, \alpha_j \boldsymbol{P}_\theta^{i,j}). \tag{19}$$

It will be shown in Section V-C that the estimation can be designed such that the prior on θ_0 does not affect the posterior distribution, but for completeness let it be assumed that all scalars θ_0^j are independently drawn from $\mathcal{N}(0, \sigma_{\theta_0}^2)$. The covariance matrix $\text{Cov}(\boldsymbol{\theta}^j) = \alpha_j \boldsymbol{P}_\theta^j$ is therefore block-diagonal:

$$\boldsymbol{P}_\theta^j = \begin{bmatrix} \sigma_{\theta_0}^2 / \alpha_j & \mathbf{0} & \ldots & \mathbf{0} \\ \mathbf{0} & \boldsymbol{P}_\theta^{1,j} & \ddots & \vdots \\ \vdots & \ddots & \ddots & \mathbf{0} \\ \mathbf{0} & \ldots & \mathbf{0} & \boldsymbol{P}_\theta^{n,j} \end{bmatrix}. \tag{20}$$

The hyperparameter, α_j, will be further discussed in Section V-A.

D. Expressions for mean and covariance

The function g is completely described by $\boldsymbol{\theta}$ through equation (15), and we turn our attention to the expressions for $\bar{\mathbf{y}}(\boldsymbol{\theta})$ and $\boldsymbol{P}_{\mathbf{y}}(\boldsymbol{\theta})$. To simplify notation, we introduce the vector

$$\mathbf{w} \triangleq \mathbb{E}[\mathbf{h}(\mathbf{x})] = [1,\ 0, \ldots, 0]^T. \tag{21}$$

For a given polynomial, i.e., one realization of $\boldsymbol{\theta}$, $\mathbf{y}$ has the mean

$$\begin{aligned} \bar{\mathbf{y}}(\boldsymbol{\theta}) &= \boldsymbol{\theta}^T \mathbf{w} \\ &= \left[\theta_0^1, \ldots, \theta_0^m\right]^T, \end{aligned} \tag{22}$$

and the covariance matrix

$$\begin{aligned} &\boldsymbol{P}_{\mathbf{y}}(\boldsymbol{\theta}) \\ &= \int_{\mathbb{R}^n} \mathcal{N}(\mathbf{x}\,; \boldsymbol{\mu}_{\mathbf{x}}, \boldsymbol{P}_{\mathbf{x}}) \left[g(\mathbf{x}; \boldsymbol{\theta}) - \bar{\mathbf{y}}(\boldsymbol{\theta})\right] \left[g(\mathbf{x}; \boldsymbol{\theta}) - \bar{\mathbf{y}}(\boldsymbol{\theta})\right]^T d\mathbf{x} \\ &= \mathbb{E}[[\boldsymbol{\theta}^T \mathbf{h}(\mathbf{x}) - \boldsymbol{\theta}^T \mathbf{w}][\boldsymbol{\theta}^T \mathbf{h}(\mathbf{x}) - \boldsymbol{\theta}^T \mathbf{w}]^T] \\ &= \boldsymbol{\theta}^T \mathbb{E}[[\mathbf{h}(\mathbf{x}) - \mathbf{w}][\mathbf{h}(\mathbf{x}) - \mathbf{w}]^T] \boldsymbol{\theta} \\ &= \boldsymbol{\theta}^T \mathbf{C} \boldsymbol{\theta}. \end{aligned} \tag{23}$$

All off-diagonal elements of $\mathbf{C} \triangleq \mathbb{E}[[\mathbf{h}(\mathbf{x}) - \mathbf{w}][\mathbf{h}(\mathbf{x}) - \mathbf{w}]^T]$ are zero, and the $pn + 1$ elements of the diagonal are given by equations (50) and (51) in Appendix A:

$$\text{diag}(\mathbf{C}) = [0,\ 1!,\ 2!, \ldots,\ p!, \ldots,\ 1!,\ 2!, \ldots,\ p!]^T. \quad (24)$$

Expressions (22) and (23) have been derived for a given parameter vector $\boldsymbol{\theta}$. However, since $\boldsymbol{\theta}$ is modeled as a stochastic variable, we carry out the marginalization in (11) and (12) to form the final estimators:

$$\begin{aligned} \bar{\mathbf{y}}_\pi &= \mathbb{E}\left[\boldsymbol{\theta}^T \middle| \mathbf{z}, \chi\right] \mathbf{w} \quad (25)\\ &= \left\{ \boldsymbol{\mu}_{\theta|\mathbf{z}} \triangleq \mathbb{E}\left[\boldsymbol{\theta} \middle| \mathbf{z}, \chi\right] \right\} = \boldsymbol{\mu}_{\theta|\mathbf{z}}^T \mathbf{w} \\ &= \mathbb{E}\left[[\theta_0^1, \ldots, \theta_0^m]^T \middle| \mathbf{z}, \chi\right] \end{aligned}$$

$$\begin{aligned} \boldsymbol{P}_{\mathbf{y},\pi} &= \mathbb{E}\left[\boldsymbol{\theta}^T \mathbf{C}\, \boldsymbol{\theta} \middle| \mathbf{z}, \chi\right] \quad (26)\\ &= \boldsymbol{\mu}_{\theta|\mathbf{z}}^T \mathbf{C}\, \boldsymbol{\mu}_{\theta|\mathbf{z}} + \mathbb{E}\left[[\boldsymbol{\theta} - \boldsymbol{\mu}_{\theta|\mathbf{z}}]^T \mathbf{C} [\boldsymbol{\theta} - \boldsymbol{\mu}_{\theta|\mathbf{z}}] | \mathbf{z}, \chi\right] \\ &= \boldsymbol{\mu}_{\theta|\mathbf{z}}^T \mathbf{C}\, \boldsymbol{\mu}_{\theta|\mathbf{z}} + \begin{bmatrix} \alpha_1 \text{Tr}\left\{ \boldsymbol{P}_{\theta|\mathbf{z}}^1 \mathbf{C} \right\} & & 0 \\ & \ddots & \\ 0 & & \alpha_m \text{Tr}\left\{ \boldsymbol{P}_{\theta|\mathbf{z}}^m \mathbf{C} \right\} \end{bmatrix}. \end{aligned}$$

Expressions for the conditional mean, $\boldsymbol{\mu}_{\theta|\mathbf{z}}$, and posterior covariance matrices, $\boldsymbol{P}_{\theta|\mathbf{z}}^j$ ($j = 1 \ldots m$), given observations $\mathbf{z}, \chi$, are derived in the following section.

V. Calculating the posterior distribution

Closed-form expressions for the mean and covariance of the transformed variable have been derived in the previous section. However, the posterior mean and variance of the elements in $\boldsymbol{\theta}$ are needed in order to evaluate these expressions. Our objective is therefore now to calculate the posterior distribution $p(\boldsymbol{\theta}|\mathbf{z}, \chi)$ and its first two moments.

An exact expression for the posterior distribution is obtained by marginalizing the hierarchical model,

$$p(\boldsymbol{\theta}^j|\mathbf{z}, \chi) = \int p(\boldsymbol{\theta}^j|\alpha_j, \mathbf{z}, \chi) p(\alpha_j|\mathbf{z}, \chi) d\alpha_j, \quad (27)$$

which is usually difficult. A simple yet useful approach is to use a point estimate of α_j. In other words, we set

$$p(\boldsymbol{\theta}^j|\mathbf{z}, \chi) \approx p(\boldsymbol{\theta}^j|\hat{\alpha}_j, \mathbf{z}, \chi), \quad (28)$$

for some estimate $\hat{\alpha}_j$ which is assumed known in this section. The relation between observations $\mathbf{z}$ and parameter vector $\boldsymbol{\theta}$ was established in equation (15) and is linear:

$$\mathbf{z} = \boldsymbol{\theta}^T \mathbf{H}^T(\chi), \quad (29)$$

where the observation matrix is given by:

$$\mathbf{H}(\chi) = \begin{bmatrix} \mathbf{h}^T(\mathbf{x}^0) \\ \vdots \\ \mathbf{h}^T(\mathbf{x}^{2n}) \end{bmatrix}. \quad (30)$$

For notational convenience, we will omit the reference to χ from now on.

Given a Gaussian prior distribution, $\mathcal{N}(\boldsymbol{\theta}^j; \mathbf{0}, \alpha_j \boldsymbol{P}_\theta^j)$, the posterior distribution is also Gaussian with mean and covariance [13]

$$\boldsymbol{\mu}_{\theta|\mathbf{z}}^j = \boldsymbol{P}_\theta^j \mathbf{H}^T \left[\mathbf{H} \boldsymbol{P}_\theta^j \mathbf{H}^T \right]^{-1} \mathbf{z}^j \quad (31)$$

$$\alpha_j \boldsymbol{P}_{\theta|\mathbf{z}}^j = \left(\mathbf{I} - \boldsymbol{P}_\theta^j \mathbf{H}^T \left[\mathbf{H} \boldsymbol{P}_\theta^j \mathbf{H}^T \right]^{-1} \mathbf{H} \right) \alpha_j \boldsymbol{P}_\theta^j, \quad (32)$$

where $\mathbf{z}^j$ is the j^{th} column in $\mathbf{z}^T$. The conditional mean matrix is $\boldsymbol{\mu}_{\theta|\mathbf{z}} = [\boldsymbol{\mu}_{\theta|\mathbf{z}}^1, \boldsymbol{\mu}_{\theta|\mathbf{z}}^2, \ldots, \boldsymbol{\mu}_{\theta|\mathbf{z}}^m]$ [see equation (17)] and estimates (25) and (26) can thus be readily calculated.

If all transformations are treated the same way a priori, i.e., if the covariance matrices $\boldsymbol{P}_\theta^{i,j}$ in (19) do not depend on j, the elements $\text{Tr}\{\boldsymbol{P}_\theta^j \mathbf{C}\}$ are also independent of j. Hence, the superscript j can be dropped and the expression for $\boldsymbol{P}_{\mathbf{y},\pi}$ can be simplified to

$$\boldsymbol{P}_{\mathbf{y},\pi} = \boldsymbol{\mu}_{\theta|\mathbf{z}}^T \mathbf{C} \boldsymbol{\mu}_{\theta|\mathbf{z}} + \begin{bmatrix} \alpha_1 & & 0 \\ & \ddots & \\ 0 & & \alpha_m \end{bmatrix} \text{Tr}\{\boldsymbol{P}_{\theta|\mathbf{z}} \mathbf{C}\}. \quad (33)$$

To simplify notation in the remaining part of the paper, it is assumed that $\boldsymbol{P}_\theta$ and $\boldsymbol{P}_\theta^j$ can be used interchangeably. Furthermore, according to equation (32), $\text{Tr}\{\boldsymbol{P}_{\theta|\mathbf{z}} \mathbf{C}\}$ does not depend on $\mathbf{z}$ and can therefore be calculated in advance.

A. Empirical hierarchical model

The estimates of α_j presented below are derived from the posterior distribution

$$p(\alpha_j|\mathbf{z}) \propto p(\mathbf{z}|\alpha_j) p(\alpha_j). \quad (34)$$

The posterior, on the other hand, relies on expressions for the likelihood $p(\mathbf{z}|\alpha_j)$ and the prior $p(\alpha_j)$ presented next.

1) The likelihood function: In our setting, $\boldsymbol{\theta}^j$ is a Gaussian random variable conditioned on α_j, and the linearly dependent variable $\mathbf{z}^j$ is therefore also Gaussian. However, the mean of $\mathbf{z}$ is unaffected by the hyperparameters as $\mathbb{E}[\mathbf{z}] = \mathbf{0}$. Since we are interested only of the dependence on α_j, we introduce $\tilde{\mathbf{z}}^j = [z_1^j - \theta_0^j, \ldots, z_\rho^j - \theta_0^j]^T$ and, because $\boldsymbol{\theta}$ is zero mean, the likelihood function takes the following simple form:

$$p(\tilde{\mathbf{z}}^j|\alpha_j) = \frac{1}{(2\pi)^{\frac{\rho}{2}} (\alpha_j)^{\frac{\rho}{2}} \sqrt{|\tilde{\mathbf{H}} \tilde{\boldsymbol{P}}_\theta^j \tilde{\mathbf{H}}^T|}}\ e^{-\frac{1}{2\alpha_j} \tilde{\mathbf{z}}^{jT} (\tilde{\mathbf{H}} \tilde{\boldsymbol{P}}_\theta^j \tilde{\mathbf{H}}^T)^{-1} \tilde{\mathbf{z}}^j},$$

where ρ is the number of elements in $\mathbf{z}^j$. The row in $\mathbf{H}$ containing H_0, and the row and column in $\boldsymbol{P}_\theta^j$ concerning θ_0^j, are removed and the adjusted matrices are denoted $\tilde{\mathbf{H}}$ and $\tilde{\boldsymbol{P}}_\theta^j$.

2) A noninformative prior and the posterior distribution: The prior should ensure that the product with the likelihood is a proper probability density. Additionally, to ensure a weak influence of the prior on the posterior distribution, it should be noninformative. One can argue that $p(\alpha_j) \propto 1/\alpha_j$ is a sensibly vague prior [14]. The expression for the posterior distribution is then:

$$p(\tilde{\mathbf{z}}^j|\alpha_j)\, p(\alpha_j) \propto \frac{1}{{\alpha_j}^{\frac{\rho}{2}+1}}\ e^{-\frac{1}{2\alpha_j} d^2}, \quad (35)$$

where $d^2 = \tilde{\mathbf{z}}^{j\,T}(\tilde{\mathbf{H}}\tilde{\boldsymbol{P}}_\theta^j\tilde{\mathbf{H}}^T)^{-1}\tilde{\mathbf{z}}^j$. The above expression is proportionate to the scaled inverse chi-square distribution, so

$$\alpha_{j|\mathbf{z}} \sim \text{inv-}\chi^2(\nu, s^2), \tag{36}$$

with parameters $\nu = \rho$ and $s^2 = d^2/\rho$.

3) Estimates of the hyperparameter: The mean and mode of the scaled inverse chi-square distribution are:

$$\mathbb{E}(\alpha_j) = \frac{\nu}{\nu - 2}s^2 \tag{37}$$

$$\text{mode}(\alpha_j) = \frac{\nu}{\nu + 2}s^2, \tag{38}$$

and can be used as point estimates of α_j in the posterior covariance matrix expression (33). Note that the conditional mean (31) is unaffected by the hyperparameter. The algorithm presented in Section V-D employs the mode, $\hat{\alpha}_j = \text{mode}(\alpha_j)$, as a point estimate of α_j.

B. Estimator performance under the modeling assumption

The posterior mean, $\bar{\mathbf{y}}_\pi$, is by definition an unbiased estimator of $\bar{\mathbf{y}}(\boldsymbol{\theta})$, when averaged over the prior. Further, conditioned on α, the error in the estimate of the mean, $\bar{\mathbf{y}}_\pi - \bar{\mathbf{y}}(\boldsymbol{\theta})$, is a Gaussian random variable with covariance

$$\mathbb{E}\left[[\bar{\mathbf{y}}_\pi - \bar{\mathbf{y}}(\boldsymbol{\theta})][\bar{\mathbf{y}}_\pi - \bar{\mathbf{y}}(\boldsymbol{\theta})]^T\right] = \mathbf{I}_{m\times m}\mathbf{w}\boldsymbol{P}_{\theta|\mathbf{z}}\mathbf{w}^T. \tag{39}$$

The distribution of the elements in $\boldsymbol{P}_{\mathbf{y}}(\boldsymbol{\theta})$ is less trivial; a diagonal element is a weighted sum of chi-square distributed variables, whereas an off-diagonal element is created from products between independent Gaussian random variables. This could be looked upon as a weighted sum of Wishart distributed matrices, created from, $\boldsymbol{\theta}_i$, the *rows* of $\boldsymbol{\theta}$: $\sum_{k=0}^{pn}\mathbf{C}(k+1, k+1)\boldsymbol{\theta}_k^T\boldsymbol{\theta}_k$.

Equation (39) clearly illustrates how uncertainties in $\boldsymbol{\theta}$ affects the estimate. It is desirable to design an estimator such that the above variance equals zero, meaning that for the family of functions described by $p(\boldsymbol{\theta}|\mathbf{z})$, the estimator is correct for every realization. Inserting the right-hand side of equation (32) into (39), we see that

$$\mathbf{w}\left(\mathbf{I} - \boldsymbol{P}_\theta\mathbf{H}^T\left[\mathbf{H}\boldsymbol{P}_\theta\mathbf{H}^T\right]^{-1}\mathbf{H}\right)\boldsymbol{P}_\theta\mathbf{w}^T = 0 \tag{40}$$

if there exists a $\boldsymbol{\lambda}^{\text{opt}}$ such that $\mathbf{w}^T = \mathbf{H}^T\boldsymbol{\lambda}^{\text{opt}}$, and $\mathbf{H}\boldsymbol{P}_\theta\mathbf{H}^T$ is invertible. One could guess that $\boldsymbol{\theta}$ must be known in order for the estimate to be exact, but it is enough to project the uncertainties in $\boldsymbol{\theta}$ onto the plane orthogonal to the vector $\mathbf{w}$. In appendix B it is shown that the selection scheme (4) – (5) attains this projection, which means that $\bar{\mathbf{y}}_\pi = \bar{\mathbf{y}}(\boldsymbol{\theta})$ with probability one.

C. Comparison with other sigma-point methods

The presented framework can be compared to other sigma-point based methods by identifying the weighted sum form (7) of the estimator. The estimate for the mean (25) is re-written:

$$\bar{\mathbf{y}}_\pi = \mathbf{z}\left[\boldsymbol{P}_\theta\mathbf{H}^T\left[\mathbf{H}\boldsymbol{P}_\theta\mathbf{H}^T\right]^{-1}\right]^T\mathbf{w}. \tag{41}$$

This is clearly a weighted sum of the evaluated sigma-points $\mathbf{z}$ with weights

$$\boldsymbol{\lambda} = \left[\mathbf{H}\boldsymbol{P}_\theta\mathbf{H}^{\mathbf{T}}\right]^{-1}\mathbf{H}\boldsymbol{P}_\theta\mathbf{w}. \tag{42}$$

It was stated in Section V-B that there are no posterior uncertainties in $\bar{\mathbf{y}}(\boldsymbol{\theta})$ if $\mathbf{w} = \mathbf{H}^T(\chi)\boldsymbol{\lambda}^{\text{opt}}$, and in this case $\boldsymbol{\lambda} = \boldsymbol{\lambda}^{\text{opt}}$. Whether or not this can be achieved depends on the evaluated points defining the observation matrix (30). Attempting to integrate over a high order polynomial increases the number of elements in the observation matrix, making it harder to find a weight vector.

The definition of the precision of an integration rule is [10]: *A rule is said to have precision p if it integrates monomials up to degree p exactly, but not exactly for some monomials of degree $p+1$.*

The integration rule used by the unscented transform has precision 5, and it is shown in appendix B that selecting evaluation points according to the sigma-point selection scheme (4) – (5) satisfies the projection (40) up to H_5 (H_3, if the cubature points are used). In these cases, as long as $\boldsymbol{P}_\theta^j$ is a full-rank matrix, the proposed estimator for the mean is identical to the sigma-point estimator (7). The explicit model assumptions in the proposed method coincides with the implicit assumptions in the sigma-point filter, and the actual values in the prior no longer affect the result.

The methods still differ in how they calculate the covariance matrix estimate; the unscented transform and the cubature rule both estimate also the covariance integral on the weighted sum form (8). However, the covariance integral of a polynomial of order p needs to be calculated using a rule with precision $2p$ in order to be exact. The proposed method instead relates the mean and the covariance using the transforming function and employs analytical expressions and marginalization to calculate the covariance matrix.

D. The marginalized sigma-point estimator

For $\mathbf{x} \in \mathbb{R}^n$, $\mathbf{y} = g(\mathbf{x}) \in \mathbb{R}^m$, $\mathbf{x} \sim \mathcal{N}(\boldsymbol{\mu}_{\mathbf{x}}, \boldsymbol{P}_{\mathbf{x}})$

1) Select a prior covariance matrix $\boldsymbol{\Sigma}$, a diagonal $p \times p$ matrix with at least two nonzero elements ($p = 3$ or 5). Use $\boldsymbol{\Sigma}$ for every instance of $\boldsymbol{P}_\theta^{i,j}$ in equation (20) to create $\boldsymbol{P}_\theta$.

2) Generate sigma-points and calculate all constants:

$$\mathbf{x}^0 = \mathbf{0}_{n\times 1}$$

$$\mathbf{x}^k = \begin{cases} +\left(\sqrt{\frac{n}{(1-w_0)}\mathbf{I}_{n\times n}}\right)_k & ,\ 1 < k \le n \\ -\left(\sqrt{\frac{n}{(1-w_0)}\mathbf{I}_{n\times n}}\right)_{k-n} & ,\ n < k \le 2n \end{cases}$$

$$\chi = \left[\mathbf{x}^0, \mathbf{x}^1, \ldots, \mathbf{x}^{2n}\right].$$

Use $w_0 = 0$ if $p = 3$, or $w_0 = 1 - n/3$ if $p = 5$ (see Section II-A). Matrices $\boldsymbol{P}_{\theta|\mathbf{z}}$, $\mathbf{C}$ and $\mathbf{H}(\chi)$ are constant and can be calculated in advance using equations (20),

(24) and (30) respectively.

3) Propagate the sigma-points:
$$\mathbf{z} = \left[g(\boldsymbol{\mu}_{\mathbf{x}} + \sqrt{\boldsymbol{P}_{\mathbf{x}}}\mathbf{x}^0), \ldots, g(\boldsymbol{\mu}_{\mathbf{x}} + \sqrt{\boldsymbol{P}_{\mathbf{x}}}\mathbf{x}^{2n}) \right].$$

4) Compute the mean, $\bar{\mathbf{y}}_\pi = [\bar{y}_{\pi,1}, \ldots, \bar{y}_{\pi,m}]^T$, using equation (25) and (31):
$$\boldsymbol{\mu}_{\theta|\mathbf{z}} = \boldsymbol{P}_\theta \mathbf{H}^T \left[\mathbf{H} \boldsymbol{P}_\theta \mathbf{H}^T \right]^{-1} \mathbf{z}^T$$
$$\bar{\mathbf{y}}_\pi = \boldsymbol{\mu}_{\theta|\mathbf{z}} \mathbf{w}$$

5) Create an observation matrix $\mathbf{H}(\sqrt{\frac{n}{(1-w_0)}}[1, -1])$ and remove the influence from H_0 (the first column) to obtain $\tilde{\mathbf{H}}$. Estimate the hyperparameters using the known mean:
$$d_{i,j}^2 = \frac{1}{4} \mathbf{z}_j^i \left[\tilde{\mathbf{H}} \boldsymbol{\Sigma} \tilde{\mathbf{H}}^T \right]^{-1} \mathbf{z}_j^{i\,T}$$
$$\hat{\alpha}_j = \frac{1}{2(n+1)} \sum_{i=1}^{n} d_{i,j}^2,$$
where $\mathbf{z}_j^i$ is the j^{th} row in $[g(\mathbf{x}^{2i-1}) - \bar{y}_{\pi,j}, g(\mathbf{x}^{2i}) - \bar{y}_{\pi,j}]$.

6) Calculate the covariance matrix, $\boldsymbol{P}_{\mathbf{y},\pi}$, using equation (33) and the estimates $\hat{\alpha}_j$ from the previous step.

Steps $1-2$ can be prepared, as well as inversion of the matrix product in step 5, whereas steps $3-6$ are executed whenever an estimate is needed.

VI. Examples

The cubature rule is a preferred special case of the unscented transform as the estimated covariance matrix is always positive-definite — a property shared also by the presented method. Further, the results in [7] indicate that the cubature rule performs better than the divided difference filter [5]. Therefore, our main goal is to show how the presented method performs compared to the cubature transform. A transformation relevant to the task of tracking targets using radar is the transformation from polar to Cartesian coordinates, which is also commonly used to illustrate the performance of the unscented transform.

A measure on how much a distribution $q(\mathbf{y})$ differs from a reference distribution $p(\mathbf{y})$, is the Kullback-Leibler (KL) discrimination[1] of q from p [16]:
$$d_{\text{KL}}(p, q) = \int p(\mathbf{y}) \log \frac{p(\mathbf{y})}{q(\mathbf{y})} d\mathbf{y}. \tag{43}$$
This measure was also used in [7] to evaluate the cubature rule, which further motivates using the same approach here. The reference distribution is calculated using Monte Carlo integration, $\int p(x) g(x) dx \approx \sum_{n=1}^{N} g(x_n)$, and $d_{KL}(p, q)$ is evaluated analytically under the assumption that the distributions p and q are Gaussian.

The presented method is implemented using the algorithm in Section V-D, using the same evaluation points as the cubature rule ($\chi = \left[\mathbf{x}^1, \ldots, \mathbf{x}^{2n}\right]$).

[1]Usually referred to as the Kullback-Leibler divergence, although when introduced in [15], the authors used the term "divergence" for the symmetric measure $d_{\text{KL}}(p, q) + d_{\text{KL}}(q, p)$.

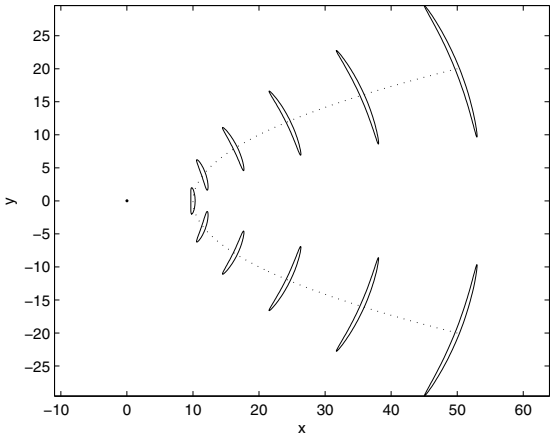

Figure 1. A sensor, situated in the origin, with uncertainties in range and angle measurements observe a target at eleven positions. The "banana-shaped" contours are measurement space covariance contours, transformed to the Cartesian coordinate system.

A. Polar to Cartesian transformation

Let $\mathbf{y} = g(\mathbf{x})$ be the transformation from a polar coordinate system defined in terms of range, r, and azimuth, ψ, to a Cartesian coordinate system:
$$\mathbf{x} = [r, \psi]^T, \quad \mathbf{y} = \begin{bmatrix} x_1 \cos x_2 \\ x_1 \sin x_2 \end{bmatrix}. \tag{44}$$
By modifying the prior, the presented method can be optimized to yield excellent results for a narrow family of transformations. However, this is not a fair comparison and typically not a realistic approach. Instead we use the same prior for the 11 positions in Fig. 1, and for each position we use 8 different azimuth measurement noise variances, σ_ψ^2:
$$\sigma_\psi^2 = [5^2, 10^2, 15^2, 20^2, 25^2, 30^2, 35^2, 40^2](\frac{\pi}{180})^2 \quad [\text{rad}^2]. \tag{45}$$
The range measurement noise variance is constant throughout all evaluations, $\sigma_r^2 = 0.5\ [\text{m}^2]$.

To illustrate the influence of the prior, we present results for two different priors, both assuming a zero-mean Gaussian distribution of $\boldsymbol{\theta}$. The first one is created using a simple assumption; the function can be described as a 2^{nd} order polynomial where the higher order term is relatively small, whereas the second one has been numerically derived to perform well in this scenario. The cubature evaluation points, χ, are used by all three methods and, following the discussion in Section V-C, the prior variance for the mean, θ_0, does not influence the estimate
$$\boldsymbol{\Sigma}_1 = \begin{bmatrix} 1 & 0 & 0 \\ 0 & \frac{1}{100} & 0 \\ 0 & 0 & 0 \end{bmatrix}, \boldsymbol{\Sigma}_2 = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 0.036 & 0 \\ 0 & 0 & 0.0007 \end{bmatrix}. \tag{46}$$
These covariance matrices are used for every instance of $\boldsymbol{P}_\theta^{i,j}$, $i, j \in \{1, 2\}$ in $\boldsymbol{P}_\theta$ [see equation (20)].

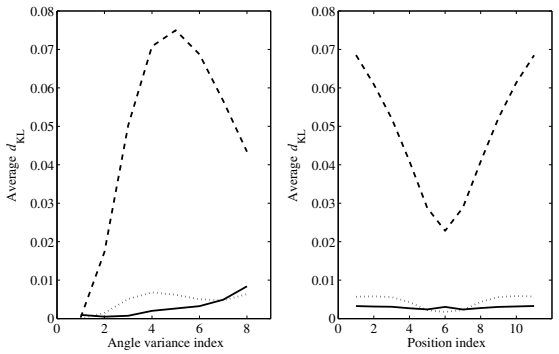

Figure 2. The left figure shows the average Kullback-Leibler discrimination for the different azimuth noise variances, whereas the right figure shows the average Kullback-Leibler discrimination for the positions. The dashed line illustrates the Cubature rule, the dotted line represents the use of $\mathbf{\Sigma}_1$, and the solid line the use of $\mathbf{\Sigma}_2$.

Table I
AVERAGE KULLBACK-LEIBLER DISCRIMINATION

	Average KL-discrimination $[\times 10^{-4}]$
Cubature rule	478
Marginalized, $\mathbf{\Sigma}_1$	45
Marginalized, $\mathbf{\Sigma}_2$	29

The average Kullback-Leibler discrimination is presented in Table I and the mean for each position and noise variance is displayed in Fig. 2. The reference density was calculated using $N = 10^5$ number of samples. The results show that, although all methods perform very good in absolute numbers, the marginalized sigma-point estimator outperforms the Cubature rule using the same points χ.

B. A note on symmetric functions

The cubature rule does not use the sigma-point $\mathbf{x}^0$ in the estimates, thus removing the tendency of the unscented transform to produce non-positive semidefinite covariance matrices when the dimensionality of $\mathbf{x}$ grows. As a consequence, the cubature integration rule with precision 3 is used instead of the rule with precision 5. Another side effect is that functions that are fully symmetric over the covariance contour have no observability regarding covariance, for example:

$$y = x^2, \quad x \sim \mathcal{N}(0, 1). \tag{47}$$

If all propagated points have the same value this will also be the estimate of the mean, i.e., $g(\mathbf{x}^i) = \bar{y}$ for all sigma-points. The variance estimate is then zero:

$$\sum_{i=0}^{2n} w_i [g(\mathbf{x}^i) - \bar{y}][g(\mathbf{x}^i) - \bar{y}]^T = 0.$$

This is rarely the case in real situations, but nevertheless illustrates an undesired behavior. The presented method can safely use $\mathbf{x}^0$, making use of the precision 5 integration rule and reducing the number of functions for which all propagated points take on the same value.

C. A note on model assumptions

The explicit model assumptions in the presented method can explain unexpected estimator behavior. A function similar to the previous example (47) is:

$$y = x_1 x_2, \quad x \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_{2\times 2}). \tag{48}$$

The variance of y is $\mathbb{E}[x_1^2 x_2^2] = 1$, and it is expected of an integration rule of precision $p \geq 4$ to produce an exact estimate. Nevertheless, the sigma-point methods fail to do this due to the unlucky choice[2] of sigma-points; they all evaluate to zero. The prior used in the presented method, however, explicitly excludes cross-terms in the model, so the result should come as no surprise.

VII. CONCLUSIONS

We have presented a derivative-free method for estimating the mean and covariance of a transformed Gaussian-distributed random variable, which has several beneficial properties. In summary, the method:

- is easy to use. The algorithm presented in Section V-D maintains the simplicity of sigma-point filters while ensuring a positive-semidefinite covariance matrix estimate.
- performs well. The results indicate that the method can perform better than the cubature rule for the same observations.
- clarifies assumptions. By assigning a prior and modeling the relation between the mean and covariance, the underlying assumptions are made clear.

The transformation is modeled as a stochastic process and, by evaluating the function, Bayesian inference is used to describe the process. Analytic expressions for the mean and covariance are given for the family of Hermite polynomials, and estimates are calculated using marginalization. Hence, the simplicity of sigma-point filters is maintained while making the underlying assumptions regarding the transforming function clear.

The method is applied to the transformation from polar to Cartesian coordinates and evaluated with respect to the Kullback-Leibler discrimination. The results show that the method is better suited for this transformation than the cubature rule, and that it is possible to find a fixed prior that performs well for the whole scenario.

ACKNOWLEDGEMENT

The authors would like to thank the Swedish Strategic Vehicle Research and Innovation programme and the Intelligent Vehicle Safety Systems programme for sponsoring this work.

[2]It is commonly suggested to use the Cholesky factorization to calculate $\sqrt{P_{\mathbf{x}}}$, in this case yielding the identity matrix.

APPENDIX A
PROPERTIES OF HERMITE POLYNOMIALS

The univariate Hermite polynomials are orthogonal under integration under the Gaussian pdf, i.e., for $x \sim \mathcal{N}(0,1)$,

$$\mathbb{E}[H_i(x)H_j(x)] = \int p(x)H_i(x)H_j(x)dx = \begin{cases} 0 & ,i \neq j \\ i! & ,i = j \end{cases}. \quad (49)$$

It follows that the expected value is zero for all but the 0^{th} polynomial:

$$\mathbb{E}[H_i(x)] = \int p(x)H_i(x)H_0(x)dx = \begin{cases} 0 & ,i \neq j \\ 1 & ,i = 0 \end{cases}. \quad (50)$$

Further, we conclude that, for $[x_1, \dots, x_n]^T \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_{n\times n})$,

$$\begin{aligned} \mathbb{E}[H_i(x_k)H_j(x_l)] &= \int p(\mathbf{x})H_i(x_k)H_j(x_l)d\mathbf{x} \\ &= \begin{cases} 0 & ,i \neq j \cup k \neq l \\ 1 & ,i = j = 0 \; \forall k,l \\ i! & ,i = j \cap k = l \end{cases}. \end{aligned} \quad (51)$$

A simple formula expressing the Hermite polynomials in terms of a random variable $\nu \sim \mathcal{N}(0,1)$ was given in [17]:

$$H_n(x) = \mathbb{E}\left[(x + \nu\sqrt{-1})^n |x\right]. \quad (52)$$

The first six Hermite polynomials are

$$\begin{aligned} &H_0(x) = 1, \quad H_2(x) = x^2 - 1, \quad H_4(x) = x^4 - 6x^2 + 3 \\ &H_1(x) = x, \quad H_3(x) = x^3 - 3x, \quad H_5(x) = x^5 - 10x^3 + 15x. \end{aligned}$$

Scaling the Hermite polynomials to achieve orthogonality when $\sigma_x \neq 1$ is achieved by dividing the argument with the standard deviation: $H_i(x/\sigma_x)$. Expressions for multivariate Hermitian polynomials are described in [17], offering the possibility to extend the framework to model also terms not represented by the univariate Hermite polynomials, i.e., products on the form $y = \prod_{i=1}^{n} x_i^{\kappa_i}$, for $\kappa_i \in \{0, 1, 2, \dots\}$.

APPENDIX B
OPTIMAL SIGMA-POINTS

In Section V-B it was shown that there are no posterior uncertainties in the estimate of the mean if there exists a $\boldsymbol{\lambda}$ such that

$$\mathbf{H}^T(\chi)\boldsymbol{\lambda} = \mathbf{w}, \quad (53)$$

with $\mathbf{w} = [1,\ 0, \dots,\ 0]^T$. As we shall see, the sigma-point selection scheme (4) - (5) always attains this.

For $x \sim \mathcal{N}(0,1)$ the sigma-points are $\chi = [0, \sqrt{3}, -\sqrt{3}]$ and the observation matrix for Hermite polynomials up to order 5 is:

$$\begin{aligned} \mathbf{H}^T(\chi) &= [\mathbf{h}(0),\ \mathbf{h}(\sqrt{3}),\ \mathbf{h}(-\sqrt{3})] \\ &= \begin{bmatrix} 1 & 1 & 1 \\ 0 & -\sqrt{3} & \sqrt{3} \\ -1 & 2 & 2 \\ 0 & 0 & 0 \\ 3 & -6 & -6 \\ 0 & 6\sqrt{3} & -6\sqrt{3} \end{bmatrix}. \end{aligned} \quad (54)$$

For $\boldsymbol{\lambda} = [\lambda_0, \lambda_1, \dots]^T$ to solve equation (53) we see that:

$$\begin{aligned} 1: &\quad \textstyle\sum_{i=0}^{2n} \lambda_i = 1 &&\text{(from row one)} \\ 2: &\quad \lambda_i = \lambda_j,\ \forall i,j \neq 0 &&\text{(from row two and six)} \\ 3: &\quad \lambda_0 = 4\lambda_i,\ i > 0 &&\text{(from row three and five)} \end{aligned} \quad (55)$$

When the dimensionality of $\mathbf{x}$ increases, no unique elements are added to $\mathbf{H}^T$. When $\mathbf{x} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_{n\times n})$:

$$\mathbf{H}^T(\chi) = \begin{bmatrix} \mathbf{h}(0) & \mathbf{h}(\sqrt{3}) & \mathbf{h}(-\sqrt{3}) & \mathbf{h}(0) & \mathbf{h}(0) & \dots \\ \mathbf{h}(0) & \mathbf{h}(0) & \mathbf{h}(0) & \mathbf{h}(\sqrt{3}) & \mathbf{h}(-\sqrt{3}) & \ddots \\ \vdots & \vdots & \vdots & \ddots & \ddots & \ddots \end{bmatrix}.$$

The third requirement is therefore adjusted to suit the multi-dimensional case: $\lambda_0 = (6 - 2n)\lambda_i$. Substituting λ_i with w_i, these are exactly the criterions (4) - (5), with $w_0 = 1 - n/3$.

The observation matrix associated with the cubature sigma-point selection scheme enjoy the same properties (for $p \leq 3$).

REFERENCES

[1] R. E. Kalman, "A new approach to linear filtering and prediction problems," pp. 24–45, 1960.

[2] K. Ito and K. Xiong, "Gaussian filters for non linear filtering problems," *IEEE Trans. Automatic control*, vol. 45, pp. 910–927, 2000.

[3] S. J. Julier, J. K. Uhlmann, and H. F. Durrant-Whyte, "A new approach for filtering nonlinear systems," in *American Control Conference, 1995. Proceedings of the*, J. K. Uhlmann, Ed., vol. 3, 1995, pp. 1628–1632 vol.3.

[4] S. J. Julier and J. K. Uhlmann, "Unscented filtering and nonlinear estimation," *Proceedings of the IEEE*, vol. 92, no. 3, pp. 401 – 22, 2004.

[5] M. Norgaard, N. K. Poulsen, and O. Ravn, "New developments in state estimation for nonlinear systems," *Automatica*, vol. 36, no. 11, pp. 1627–38, 2000.

[6] I. Arasaratnam, S. Haykin, and R. J. Elliott, "Discrete-time nonlinear filtering algorithms using gauss-hermite quadrature," *Proceedings of the IEEE*, vol. 95, no. 5, pp. 953–977, 2007.

[7] I. Arasaratnam and S. Haykin, "Cubature Kalman Filters," *Automatic Control, IEEE Transactions on*, vol. 54, no. 6, pp. 1254–1269, 2009.

[8] U. Lerner, "Hybrid Bayesian Networks for Reasoning About Complex Systems," Ph.D. dissertation, Stanford University, 2002.

[9] J. McNamee and F. Stenger, "Construction of fully symmetric numerical integration formulas," *Numerische Mathematik*, vol. 10, no. 4, pp. 327–344, 1967, 10.1007/BF02162032.

[10] W. Yuanxin, D. Hu, W. Meiping, and H. Xiaoping, "A numerical-integration perspective on gaussian filters," *IEEE Transactions on Signal Processing*, vol. 54, no. 8, pp. 2910–21, 2006.

[11] C. Fernández-Prades and J. Vilà-Valls, "Bayesian Nonlinear Filtering Using Quadrature and Cubature Rules Applied to Sensor Data Fusion for Positioning," in *IEEE International Conference on Communications*, 2010, pp. 2–6.

[12] C. P. Robert, *The Bayesian Choice: From Decision-Theoretic Foundations to Computational Implementation*, Editors, Ed. Springer, 2007.

[13] S. M. Kay, *Fundamentals of statistical signal processing: estimation theory*. Upper Saddle River, NJ, USA: Prentice-Hall, Inc., 1993.

[14] A. Gelman, J. B. Carlin, H. S. Stern, and D. B. Rubin, *Bayesian Data Analysis, Second Edition (Chapman & Hall/CRC Texts in Statistical Science)*, 2nd ed. Chapman and Hall/CRC, Jul. 2003.

[15] S. Kullback and R. Leibler, "On information and sufficiency," *The Annals of Mathematical Statistics*, pp. 79–86, 1951.

[16] A. R. Runnalls, "Kullback-Leibler Approach to Gaussian Mixture Reduction," *IEEE Transactions on Aerospace and Electronic Systems*, vol. 43, no. 3, pp. 989–999, Jul. 2007.

[17] C. Withers, "A simple expression for the multivariate Hermite polynomials," *Statistics & Probability Letters*, vol. 47, no. 2, pp. 165–169, Apr. 2000.