# Density-ratio robustness in dynamic state estimation

Alessio Benavoli and Marco Zaffalon

*Istituto Dalle Molle di Studi sull'Intelligenza Artificiale (IDSIA), Galleria 2, CH-6928 Manno (Lugano), Switzerland*

**Abstract**

The filtering problem is addressed by taking into account imprecision in the knowledge about the probabilistic relationships involved. Imprecision is modelled in this paper by a particular closed convex set of probabilities that is known with the name of *density ratio class* or *constant odds-ratio* (COR) model. The contributions of this paper are the following. First, we shall define an optimality criterion based on the squared-loss function for the estimates derived from a general closed convex set of distributions. Second, after revising the properties of the density ratio class in the context of parametric estimation, we shall extend these properties to state estimation accounting for system dynamics. Furthermore, for the case in which the nominal density of the COR model is a multivariate Gaussian, we shall derive closed-form solutions for the set of optimal estimates and for the credible region. Third, we discuss how to perform Monte Carlo integrations to compute lower and upper expectations from a COR set of densities. Then we shall derive a procedure that, employing Monte Carlo sampling techniques, allows us to propagate in time both the lower and upper state expectation functionals and, thus, to derive an efficient solution of the filtering problem. Finally, we empirically compare the proposed estimator with the Kalman filter. This shows that our solution is more robust to the presence of modelling errors in the system and that, hence, appears to be a more realistic approach than the Kalman filter in such a case.

*Keywords:* Coherent lower expectations, density ratio class, maximality, robustness, Kalman filter.

## 1. Introduction

This paper deals with the problem of estimating the state of a discrete-time stochastic dynamical system on the basis of observations. One way of approaching this problem is to assume that the dynamics, the initial condition, and the observations are corrupted by noises with *known distributions* and then to find the conditional distribution of the state given the past observations. This is the so-called Bayesian state estimation approach.

If the dynamics and observations are linear functions of the state and the noise contributors are assumed to be Gaussian, it is well known that the optimal solution of the Bayesian state estimation problem is the Kalman filter (KF), see for instance [1]. In the non-linear/non-Gaussian case, an analytic solution of Bayesian state estimation is in general not available in closed form and a numerical or analytical approximation is required. The extended KF is the most known analytical approximation of the Bayesian state estimation problem for non-linear systems [1]. Conversely, among the numerical techniques, the ones used most frequently are based on Monte Carlo sampling methods (see, for instance, [2, 3, 4]).

A common trait to these techniques is that they assume that the distributions associated with the prior, state transition, and likelihood functions are perfectly known. However, in many practical cases, our information about the system to be modelled may not allow us to characterise these functions with single (precise) distributions. For example, in the Gaussian case, we may only be able to determine an interval that

*Email address:* alessio@idsia.ch, zaffalon@idsia.ch (Alessio Benavoli and Marco Zaffalon)

contains the mean of the Gaussian distribution or, in more general cases, we may only be able to state that the distribution of the noise belongs to some set of distributions.

This leads to alternative models of representation of uncertainty based on a set of probability distributions and, thus, to robust filtering. The most explored techniques for robust filtering are $H_\infty$ [5, 6], $H_2$ [7] and set-membership estimation [8, 9]. These techniques deal mainly with two kinds of uncertainties: norm-bounded parametric uncertainty and/or bounded uncertainty in the noise statistics or in the noise intensity.

Alternative robust filtering methods are based on a $p$-box representation of the set of probability distributions, see for instance [10, 11]. Other approaches to robust filtering are the set-valued Kalman filter [12] or the projection-based approach [13] that model the initial state uncertainty as a convex set of probability distributions. On the other hand, in [14], both system and measurement noise are modelled with convex sets of probability density functions by also assuming that these convex sets are polytopes (here polytope means the convex hull of a finite number of distributions). Another possibility to deal with imprecision is to robustify the KF estimate by computing credible regions for the estimate based on Chebyshev-like inequalities [15, 16].

In a recent paper [17] we have proposed a new more general approach to robust filtering that instead focuses attention on the use of closed convex sets of distributions to model the imprecision in the knowledge about the system parameters and probabilistic relationships involved. The uncertainty models mentioned before can in fact be seen as special cases of closed convex sets of distributions.

This new approach has been derived in the context of Walley's theory of *coherent lower previsions* [18, 19], which is also referred to as *Imprecise Probability*.[1] In [18], it is proved that a convex set of probability distributions can be equivalently characterised by the upper (or lower) expectation functional that it generates as the upper (lower) envelope of the expectations obtained from the distributions in such a set. Hence, the imprecision in the system model can equivalently be expressed in terms of lower/upper expectations.

In [17], by exploiting this equivalence, we have thus derived a solution of the state estimation problem which essentially consists of propagating in time both the lower and upper state expectations over the set of assumed probability distributions. This general solution has a structure that resembles the standard Bayesian solution to state estimation and, in fact, it reduces to it in the case the sets of probability distributions for initial state, measurement equation and state dynamics collapse to single distributions and, thus, the lower and upper expectation functionals coincide (in this case there is no imprecision).

The fact that we work with the lower envelopes of the set of probability distributions is an important difference between our work and the usual approaches in the literature for state estimation with a closed convex set of probability distributions, e.g., [14], which consist of directly processing the distributions in the set. In those approaches, an essential assumption is to require the closed convex set of probability distributions to be a polytope with a finite sets of vertices (in this context vertex means an extreme point of the set of distributions or an extreme point of a membership-set). Then a Bayesian estimator is derived by element-wise processing the vertices of the polytopes associated with the prior (or to the previously computed posterior), likelihood and state transition models. A drawback of this approach is that the number of vertices needed to characterise the convex sets increases exponentially fast over the number of time steps [14]. This problem is overcome in our model by working directly with lower envelopes as we need not explicitly compute the vertices. This nevertheless, our approach guarantees that the conclusions drawn are equivalent [18] to those we should obtain by element-wise processing the distributions in the closed convex sets. This together with the possibility of dealing with more general models of uncertainty are the main contributions of [17].

In [17] we have specialised this general solution of the filtering problem to the so-called linear-vacuous mixture model, which is a family made of convex combinations of a known nominal distribution (e.g., a Gaussian) with arbitrary distributions [20].

The objective of this paper is to specialise the work in [17] to another family of closed convex sets of distributions known with the names of *density ratio class* [21, 22], *interval of measures* [23] or *constant odds-ratio* (COR) model [18, Sec. 2.9.4], which are very useful in robust estimation.

[1] In this context, traditional probability theory, which models uncertainty by using a single probability distribution, is referred to as *precise probability*.

The closed convex set of distributions represented by a COR model has the following form

$$\mathcal{P} = \{p : \ (1-\epsilon)p_0(x) \leq p(x) \leq p_0(x) \ \forall x\},$$

i.e., it is the set of unnormalised probability density functions that are upper bounded by the known nominal density $p_0$ (e.g., Gaussian) and lower bounded by the scaled version $(1-\epsilon)p_0$ of the nominal density.[2] Here, $\epsilon \in (0,1)$ is called imprecision parameter, since it determines the degree of imprecision. Notice in fact that for $\epsilon = 0$ there is no imprecision, since $\mathcal{P}$ includes the single density $p_0$. The COR model has the following characteristics:

1. it is easy to elicit, since only the parameter $\epsilon$ and the density $p_0(x)$ must be specified;
2. it is robust, since it allows for a wide variety of density shapes (unimodal and multimodal), but it is not too imprecise (the tail behaviour is determined by $p_0(x)$);
3. the posterior inferences derived from COR models are computationally tractable.

This paper derives the optimal solution (w.r.t. the squared-loss function) to the state estimation problem in the case the uncertainty on initial state, measurement equation and state dynamics are modelled through COR sets of densities. Similarly to what happens in set-membership estimation, it will result that this optimal solution is a set, and in particular a convex set.

The contributions of this paper are the following. First, after revising the results in [17], we shall define an optimality criterion based on the squared-loss function for the estimates derived from a general closed convex set of distributions. We shall prove that for an estimate to be optimal (undominated under the squared-loss function) it must belong to a closed convex set which is determined by the lower and upper posterior means determined from the closed convex set of distributions. Second, after revising the properties of the COR models [21, 22] in the context of parametric estimation, we shall extend these properties to state estimation accounting for system dynamics. Furthermore, for the case in which the nominal density of the COR model is a multivariate Gaussian, we shall derive closed-form solutions for the set of optimal estimates and for the credible (i.e., Bayesian confidence) region.

Third, we shall discuss how to perform Monte Carlo integrations to compute lower and upper expectations from a COR set of densities. It will be shown that for COR models to compute lower and upper expectations no optimisation (minimisation or maximisation) is necessary. Inferences can in fact be drawn by solving integral equations numerically. By exploiting this property, we shall derive a procedure that employing Monte Carlo sampling techniques allows us to propagate in time both the lower and upper state expectation functionals and, thus, to derive an efficient solution of the filtering problem.

Finally, we empirically compare the proposed COR-based estimator with the KF and show that our solution is more robust to modelling errors and that, hence, it outperforms the KF in such a case.

### Notation

Upper case letters $X, Y$ are used to denote the variables, lower case letters $x, y$ the values of the variables. Calligraphic upper case letters $\mathcal{X}, \mathcal{Y}$ denote subsets of $\mathbb{R}^k$. $E(\cdot)$ denotes the standard expectation operator, while $\underline{E}(\cdot)$ and $\overline{E}(\cdot)$ denote the lower and, respectively, upper expectation operators. A subscript is used to denote the time instant, as in $X_k$. The observation variable at time $k$ is denoted by $Y_k$ and $\tilde{y}_k$ denotes the actual observed value of $Y_k$ at time $k$. $\delta_{\{\tilde{y}_k\}}$ denotes a Dirac's delta on the observation $\tilde{y}_k$.

[2] A COR (density ratio class) model is a special case of interval measures. In the latter in fact the lower density is not necessarily a scaled version of the upper bound.

## 2. Bayesian filtering method

Let us summarise the basic principles of Bayesian filtering (for a wider treatment of filtering theory see [1]). Its goal is the estimation of the state variables of a discrete-time nonlinear system which is "excited" by a sequence of random vectors. It is assumed that nonlinear combinations of the state variables corrupted by noise are observed. We have thus

$$\left\{ \begin{array}{rcl} x_{t+1} & = & f(t, x_t) + w_t \\ y_t & = & h(t, x_t) + v_t, \end{array} \right. \tag{1}$$

where $t$ is the time, $x_t \in \mathbb{R}^n$ is the state vector at time $t$, $w_t \in \mathbb{R}^n$ is the process noise, $y_t \in \mathbb{R}^m$ is the measurement vector, $v_t \in \mathbb{R}^m$ is the measurement noise and $f(\cdot)$ and $h(\cdot)$ are known nonlinear functions. Having observed a finite sequence $\tilde{y}^t = \{\tilde{y}_1, \ldots, \tilde{y}_t\}$ of measurements, we may, in general, seek for an estimate of an entire sequence of states $x^t = \{x_0, \ldots, x_t\}$.

In the Bayesian framework, all relevant information on $x^t = \{x_0, \ldots, x_t\}$ at time $t$ is included in the posterior distribution $p(x^t|\tilde{y}^t)$. In general, a Markov assumption is made to model the system, which implies the following independence conditions:

$$p(x_t|x^{t-1}) \;\; = \;\; p(x_t|x_{t-1}), \quad p(\tilde{y}^t|x^t) \;=\; \prod_{k=1}^{t} p(\tilde{y}_k|x_k). \tag{2}$$

Using these assumptions the probability density function (PDF) over all states can be written simply as:

$$p(x^t|\tilde{y}^t) = \frac{p(x^{t-1}|\tilde{y}^{t-1})p(x_t|x_{t-1})p(\tilde{y}_t|x_t)}{p(y_t|\tilde{y}^{t-1})}. \tag{3}$$

In many applications, we are interested in estimating $p(x_t|y^t)$, one of the marginals of the above PDF. This is the so-called *Bayesian filtering problem*. We have

$$\begin{aligned} p(x_t|\tilde{y}^t) &= \frac{p(x_t|\tilde{y}^{t-1})}{p(\tilde{y}_t|\tilde{y}^{t-1})} p(\tilde{y}_t|x_t) \\ &= \int_{x_{t-1}} dx_{t-1} \; \frac{p(x_t|x_{t-1})p(y_t|x_t)p(x_{t-1}|\tilde{y}^{t-1})}{p(\tilde{y}_t|\tilde{y}^{t-1})}. \end{aligned} \tag{4}$$

From (3) and (4), we see that both $p(x^t|\tilde{y}^t)$ and $p(x_t|\tilde{y}^t)$ can be obtained recursively. Once $p(x_t|\tilde{y}^t)$ has been computed, it is possible to compute the expected value $E_{X_t}[g|\tilde{y}^t]$ w.r.t. $p(x_t|\tilde{y}^t)$ for any function $g(x_t)$ of interest.

In the following, we rewrite the solution of the Bayesian filtering problem in a non-recursive form. This form will be useful to extend the Bayesian filtering approach to the case in which uncertainty is modelled through sets of distributions or, equivalently, lower/upper expectations. Therefore, assume that instead of $p(x_t|\tilde{y}^t)$ we are interested in computing directly $E[g|\tilde{y}^t]$, i.e., the posterior expectation of some function of interest $g$ of $X_t$ given the sequence of observations $\tilde{y}^t$.

*Theorem 1. Assume that*

$$E_{X^t,Y^t}[\delta_{\{\tilde{y}^t\}}] > 0, \tag{5}$$

*and that $E_{Y_k}[\delta_{\{\tilde{y}_k\}}|x_k]$ is well defined*[3] *for all $x_k$ and $k = 1, \ldots, t$, where $\delta_{\{\tilde{y}^t\}} = \prod_{i=1}^{t} \delta_{\{\tilde{y}_i\}}$ and $\delta_{\{\tilde{y}_i\}}$ is a Dirac's delta on the observation $\tilde{y}_i$. Then for any absolutely integrable function $g : \mathcal{X}_t \to \mathbb{R}$, the expected value $E[g|\tilde{y}^t]$ is the unique solution $\mu$ of:*

$$E_{X_0}\left[E_{X_1}\left[E_{Y_1}\left[\ldots E_{X_t}\left[E_{Y_t}[(g-\mu)\delta_{\{\tilde{y}^t\}}|X_t]\right]X_{t-1}\right]\ldots\right]X_1\right]X_0\right] = 0. \tag{6}$$

[3] This implies that $p(y_t|x_t)$ is bounded and continuous in a neighbourhood of $\tilde{y}_t$, see for instance [24, Ch.1].

■

*Proof: Let us start from the inner part of (6). Since* $E_{Y_t}[h|x_t] = \int h(x_t, y^t)p(y_t|x_t)dy_t$ *for any absolutely integrable function* $h$ *and given values of* $x_t$ *and* $y^{t-1}$*, in case* $h = (g-\mu)\delta_{\{\tilde{y}^t\}}$ *one has*

$$f_t(x_t, y^{t-1}, \mu) = E_{Y_t}[(g-\mu)\delta_{\{\tilde{y}^t\}}|x_t] = (g(x_t)-\mu)\cdot\delta_{\{\tilde{y}^{t-1}\}}(y^{t-1})\cdot p(\tilde{y}_t|x_t),$$

*which follows from the assumption that* $E_{Y_t}[\delta_{\{\tilde{y}_t\}}|x_t]$ *is well defined. Thus, one has:*

$$\begin{array}{rcl} E_{X_t}\left[f_t\middle|x_{t-1}\right] & = & \delta_{\{\tilde{y}^{t-1}\}}(y^{t-1})\cdot\int(g(x_t)-\mu)p(x_t|x_{t-1})p(\tilde{y}_t|x_t)dx_t \\ & = & \delta_{\{\tilde{y}^{t-1}\}}(y^{t-1})\cdot g_{t-1}(x_{t-1},\mu), \end{array}$$

*given* $x_{t-1}$ *and* $y^{t-1}$*. By proceeding in this way from time* $t-1$ *to time* 1*, one gets*

$$E_{X_1}\left[f_1\middle|x_0\right] = \int g_1(x_1,\mu)p(x_1|x_0)p(\tilde{y}_1|x_0)dx_1 = g_0(x_0,\mu),$$

*given* $x_0$ *and at time* 0*,*

$$E_{X_0}\left[g_0\right] = \int g_0(x_0,\mu)p(x_0)dx_0.$$

*Hence, it results that:*

$$E_{X_0}\left[g_0\right] = \int\int\cdots\int(g(x_t)-\mu)p(x_0)\prod_{i=1}^{t}p(x_i|x_{i-1})p(\tilde{y}_i|x_i)dx_t dx_{t-1}\cdots dx_0 = 0.$$

*By exploiting the additivity (linearity) property of the integrals and by solving the above equation w.r.t.* $\mu$*, one gets*

$$\mu = \frac{\int\int\cdots\int g(x_t)p(x_0)\prod_{i=1}^{t}p(x_i|x_{i-1})p(\tilde{y}_i|x_i)dx_t dx_{t-1}\cdots dx_0}{\int\int\cdots\int p(x_0)\prod_{i=1}^{t}p(x_i|x_{i-1})p(\tilde{y}_i|x_i)dx_t dx_{t-1}\cdots dx_0} = E[g|\tilde{y}^t],$$

*where the denominator is positive because of (5). The last equality follows straightforwardly from (3) and change of the integration order.* ■

Observe that the Dirac's deltas in (6) are introduced for conditioning the joint on the observed values of $Y^t$, i.e, $\{\tilde{y}_1,\ldots,\tilde{y}_t\}$, by exploiting the fact that $E_{Y_k}[\delta_{\{Y_k=\tilde{y}_k\}}|x_k] = p(\tilde{y}_k|x_k)$.[4] In fact, since all variables $Y_k$ are observed, the conditional density $p(y_k|x_k)$ must be evaluated at the observed value $\tilde{y}_k$. This follows from the properties of Bayesian conditioning for probability density functions. Notice that the relation $E_{Y_k}[\delta_{\{Y_k=\tilde{y}_k\}}|x_k] = p(\tilde{y}_k|x_k)$ is a consequence of a limiting procedure; for a more rigorous definition of this limiting procedure we point the reader to [18, Sec. 6.10.4] and to Example 1.

## 3. Set of distributions

In many practical problems, the uncertainty cannot be adequately quantified by using a single probability distribution. Consider for instance the model in (1). The nonlinear functions $f$ and $h$ may be not perfectly known and/or the available information on the noises may be not enough to specify a single probability distribution for $w$ and $v$. In this context, we say that the probabilistic knowledge is *imprecise*. A way to probabilistically describe such imprecision is to consider all possible distributions that are compatible with the available information on the system to be modelled. This is the approach followed by Walley in [18]. In [18], it is proved that a convex set of probability distributions can equivalently be characterised by

[4] Thus, here the Dirac's delta has to be interpreted in relation to how it affects $p(y_k|x_k)$ when it is integrated w.r.t. it.

the lower/upper expectation functional that it generates as the lower/upper envelope of the expectations obtained from the distributions in such a set. Hence, the imprecision in the system model can equivalently be expressed in terms of lower/upper expectations.

Given, for instance, the set of distributions $\mathcal{P}$ that describe the imprecision on the probabilistic knowledge of a variable $X \in \mathcal{X}$ and any function $g : \mathcal{X} \to \mathbb{R}$, one can define its lower and upper expectations w.r.t. $\mathcal{P}$ by

$$\begin{aligned} \underline{E}_X(g) &= \inf_{p_x \in \mathcal{P}} \int_X g(x) p_X(x) dx, \\ \overline{E}_X(g) &= \sup_{p_x \in \mathcal{P}} \int_X g(x) p_X(x) dx, \end{aligned} \tag{7}$$

where $p_X(\cdot)$ is the PDF (w.r.t. the Lebesgue measure) of $X$ under one of the possible distributions in $\mathcal{P}$, assuming it exists. In other words, the set $\mathcal{P}$ can then be characterised by the upper and lower expectations, $\overline{E}_X(g)$ and $\underline{E}_X(g)$, generated as the supremum and infimum of

$$E_X(g) = \int_X g(x) p_X(x) dx$$

over the probability measures in $\mathcal{P}$. From (7), it can be verified that $\underline{E}_X(g) = -\overline{E}_X(-g)$ and, thus, $\underline{E}_X$ fully describes $\mathcal{P}$. It can also be verified that $\underline{E}_X$ satisfies the following properties:

**(C1)** $\underline{E}_X(g_1) \geq \inf_x g_1$,

**(C2)** $\underline{E}_X(\lambda g_1) = \lambda \underline{E}_X(g_1)$,

**(C3)** $\underline{E}_X(g_1 + g_2) \geq \underline{E}_X(g_1) + \underline{E}_X(g_2)$,

for any $\lambda > 0$ and bounded scalar functions $g_1(\cdot), g_2(\cdot)$. Observe that (C1)–(C3) are the generalisation of the axioms of probability to lower expectation functionals. They specify which are the properties that the functional $\underline{E}_X$ has to satisfy to be a so-called *coherent lower prevision* (CLP). See [18, Ch. 2] for details about the behavioural implications of (C1)–(C3).

Conversely, given a functional $\underline{E}_X(g)$ that satisfies (C1)–(C3), it is possible to define a closed convex set $\mathcal{P}$ of (finitely additive) probabilities that generates the lower expectation $\underline{E}_X(g)$, for any $g(\cdot)$. This is proved in [18] and establishes a one-to-one correspondence between closed convex sets of probabilities and coherent lower previsions. Observe that the definitions (7) and the properties (C1)–(C3) can straightforwardly be extended to the conditional case, i.e., $\underline{E}_X(g_1|y) \geq \inf_{\mathcal{X}\times\{y\}} g_1$, $\underline{E}_X(\lambda g_1|y) = \lambda \underline{E}_X(g_1|y)$ and $\underline{E}_X(g_1+g_2|y) \geq \underline{E}_X(g_1|y) + \underline{E}_X(g_2|y)$ for any $\lambda > 0$, bounded scalar functions $g_1(\cdot), g_2(\cdot) : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$ and for any value $y$ of the conditional variable $Y$.

CLPs are very expressive models and offer a great flexibility to the modeller. They allow to model a state of *complete ignorance* about the value $x$ of $X$ by using the so-called *vacuous model*, i.e., $\underline{E}_X[g] = \inf g$ and $\overline{E}_X[g] = \sup g$, which corresponds to consider as $\mathcal{P}$ the set of all possible probabilities. On the other side, they reduce to standard probabilistic models in case $\mathcal{P}$ includes only a single probability and, thus, $\overline{E}_X[g] = \underline{E}_X[g] = E_X[g]$. All the intermediate degrees of imprecision between the single probability case and the set of all probabilities (complete ignorance) can be addressed by using suitable CLP-based models. In Section 6 we shall present a CLP based model that is useful in state estimation. Before doing that, we revise results derived in [17] which allow us to extend the general Bayesian solution of the state estimation problem to the case in which the uncertainty is modelled through sets of probabilities or, equivalently, CLPs.

## 4. Robust filtering through coherent lower expectations models

In Section 2, we have revised the Bayesian approach to filtering. The aim of Bayesian state estimation is to compute the conditional expectation of some function of interest $g$ of $X_t$ given the observations

$\tilde{y}^t = \{\tilde{y}_1, \tilde{y}_2, \ldots, \tilde{y}_t\}$, $E_{X_t}[g|\tilde{y}^t]$. Assume that the available information does not allow us to specify a unique probability describing each source of uncertainty in the dynamical system. We can then use sets of probabilities or, equivalently, CLPs to model the available knowledge. In this and the following sections, we shall assume that $X_k \in \mathcal{X}_k$ and $Y_k \in \mathcal{Y}_k$ for each $k$, where $\mathcal{X}_k$ and $\mathcal{Y}_k$ are convex subsets of $\mathbb{R}^n$ and, respectively, $\mathbb{R}^m$.

Consider CLPs for the initial state $\underline{E}_{X_0}$, the system dynamics $\underline{E}_{X_k}[\cdot|X_{k-1}]$ and the observation process $\underline{E}_{Y_k}[\cdot|X_k]$ for $k = 1, \ldots, t$. How can we derive the conditional CLP $\underline{E}_{X_t}[\cdot|\tilde{y}^t]$?

*Theorem 2. Assume that the CLPs $\underline{E}_{X_0}$, $\underline{E}_{X_k}[\cdot|X_{k-1}]$ and $\underline{E}_{Y_k}[\cdot|X_k]$ are known for $k = 1, \ldots, t$. Furthermore, assume that, for each $k = 1, \ldots, t$, $X^{k-2}$ and $Y^{k-1}$ are epistemically irrelevant to $X_k$ given $X_{k-1}$ and that $X^{k-1}$ and $Y^{k-1}$ are irrelevant to $Y_k$ given $X_k$, meaning that*

$$\underline{E}_{X_k}[h_1|x^{k-1}y^{k-1}] = \underline{E}_{X_k}[h_1(x^{k-2}, y^{k-1}, \cdot)|x_{k-1}], \tag{8}$$

$$\underline{E}_{Y_k}[h_2|x^k, y^{k-1}] = \underline{E}_{Y_k}[h_2(x^{k-1}, y^{k-1}, \cdot)|x_k], \tag{9}$$

*for any bounded scalar functions $h_1 : \mathcal{X}^k \times \mathcal{Y}^{k-1} \to \mathbb{R}$, $h_2 : \mathcal{X}^k \times \mathcal{Y}^k \to \mathbb{R}$ and given $x^k, y^{k-1}$. Assume also that*

$$\underline{E}_{X^t,Y^t}[\delta_{\{\tilde{y}^t\}}] > 0, \tag{10}$$

*and that $E^P_{Y_k}[\delta_{\{\tilde{y}_k\}}|x_k]$ is well defined for each $P$ in the closed convex of probabilities associated to $\underline{E}_{Y_k}[\cdot|x_k]$ and for all $x_k$ and $k = 1 \ldots, t$. Then, given the sequence of measurements $\tilde{y}^t = \{\tilde{y}_1, \tilde{y}_2, \ldots, \tilde{y}_t\}$, the posterior CLP $\underline{E}_{X_t}[g|\tilde{y}^t]$ for any bounded scalar function $g : \mathcal{X}_t \to \mathbb{R}$ is equal to the unique value $\mu \in \mathbb{R}$ that satisfies the following equation:*

$$0 = \underline{E}_{X^t,Y^t}\Big[\delta_{\{\tilde{y}^t\}} \cdot (g - \mu)\Big], \tag{11}$$

*where the above joint CLP is given by:*

$$\underline{E}_{X_0}\Big[\underline{E}_{X_1}\Big[\underline{E}_{Y_1}\Big[\ldots \underline{E}_{X_t}\Big[\underline{E}_{Y_t}\Big[\delta_{\{\tilde{y}^t\}} \cdot (g - \mu)\Big|X_t\Big]\Big|X_{t-1}\Big] \ldots \Big|X_1\Big]\Big|X_0\Big]\Big]. \tag{12}$$

■

Equation (11) is called Generalised Bayes Rule [18, Ch. 6]. The proof of Theorem 2 can be found in [17, Th. 2].[5] Intuitively, the result follows straightforwardly from (6) by replacing standard expectations with lower expectations. The conditions in (8)–(9) generalise the Markov conditions (2). The condition (10) ensures that the Generalised Bayes Rule is applicable (in other words that the denominator of Bayes' rule is positive) for any probability in the closed convex set associated to $\underline{E}_{X^t,Y^t}$.

It is worth to point out that to compute $\underline{E}_{X_t}[g|\tilde{y}^t]$ in the imprecise case, we cannot in general derive a recursive solution as in the Bayesian case. This is a consequence of the fact that CLPs are not additive (property (C3)), see also [17, Sec. 4]. In other words, to compute $\underline{E}_{X_t}[g|\tilde{y}^t]$ at any time $t$ it is necessary to go through the joint and to find the value of $\mu$ which solves (12). This means that the computational complexity to compute $\underline{E}_{X_t}[g|\tilde{y}^t]$ increases with time. In [25], it has been shown that, for discrete state variables, the computational complexity for solving (12) increases only linearly with time. The problem is that the constant of proportionality is quadratic in the number of elements of the possibility space of the state (this number is finite in the discrete case). However, for continuous variables, such number is infinite and, thus, this result cannot be applied (this is not surprising since we know that apart from few cases, the exact solution of the filtering problem, even in the standard Bayesian case, is in general infinite dimensional in the continuous case). An approximation is thus necessary. In this paper, we shall show that using a discretisation approach similar to the one used in Monte Carlo (MC) sampling methods, it is possible to derive an approximate solution whose complexity increases linearly with time.

[5] Observe that the proof in [17] has been obtained by assuming that the observation variables are discretised. Intuitively, we can see Theorem 2 as the limit of this result when the size of the discretisation interval goes to zero.

## 5. Optimality criterion and decision making

In the Bayesian setting, it is well known that the posterior mean of $X$, i.e., $\hat{x} = E[X|y]$, is the value that minimises the scalar squared error loss function $(X-x)^T(X-x)$, i.e.,

$$E[X|y] = \arg\min_x E_{X,Y}[(X-x)^T(X-x)],$$

where $E_{X,Y}[\cdot]$ is the joint expectation w.r.t. the variables $X, Y$. In the case of CLPs, one computes lower and upper posterior expectations of $X$, i.e., $\underline{E}[X|y]$ and $\overline{E}[X|y]$. Are these values optimal in some sense?

To answer this question, we must specify an optimality criterion for CLPs. In this paper, we shall use the maximality criterion proposed by Walley [18, Sec. 3.9.2]. Under maximality, we say that an estimator $\hat{x}_2$ dominates (is preferred to) $\hat{x}_1$ under the squared loss if for all densities $p_{X,Y}$ in the convex set $\mathcal{P}$, it holds that $E^p_{X,Y}((X-\hat{x}_1)^T(X-\hat{x}_1)) > E^p_{X,Y}((X-\hat{x}_2)^T(X-\hat{x}_2))$ or, equivalently, if

$$E^p_{X,Y}((X-\hat{x}_1)^T(X-\hat{x}_1) - (X-\hat{x}_2)^T(X-\hat{x}_2)) > 0 \ \ \forall p_{X,Y} \in \mathcal{P}, \tag{13}$$

where $E^p_{X,Y}$ denotes the expectation w.r.t. the density $p_{X,Y}$. A necessary and sufficient condition for (13) to be satisfied is that

$$\underline{E}_{X,Y}[(X-\hat{x}_1)^T(X-\hat{x}_1) - (X-\hat{x}_2)^T(X-\hat{x}_2)] > 0. \tag{14}$$

In the maximality criterion, estimators are compared w.r.t. the same probability, and thus $\hat{x}_2$ is said to dominate $\hat{x}_1$ if (13) is satisfied for each probability in the convex set. This is a straightforward generalisation of the Bayesian decision criterion to set of probabilities.

*Theorem 3. A necessary and sufficient condition for $\hat{x}_2$ to be undominated under maximality is:*

$$\hat{x}_2 \in \mathcal{X}^* = \left\{ \int x p(x|y)dx : \ p \in \mathcal{P} \right\}, \tag{15}$$

*where $\mathcal{P}$ is the closed convex set of probabilities associated to $\underline{E}_{X,Y}$.*[6] *Furthermore, $\mathcal{X}^*$ is a convex subset of $\mathcal{X}$.* ■

*Proof: Condition (14) is satisfied if for all $p \in \mathcal{P}$ it holds that:*

$$\begin{array}{rcl} 0 & < & \int_{\mathcal{Y}} \int_{\mathcal{X}} \left[(X-\hat{x}_1)^T(X-\hat{x}_1) - (X-\hat{x}_2)^T(X-\hat{x}_2)\right] p(x,y)dxdy \\ & = & \int_{\mathcal{Y}} \int_{\mathcal{X}} \left[(X-\hat{x}_1)^T(X-\hat{x}_1) - (X-\hat{x}_2)^T(X-\hat{x}_2)\right] p(x|y)p(y)dxdy. \end{array} \tag{16}$$

*Fixed $p$ the above inequality is satisfied for all $\hat{x}_1$ if*

$$\hat{x}_2 = \int_{\mathcal{X}} x p(x|y)dx = E_p[X|y],$$

*where $E_p[X|y]$ is the posterior mean computed w.r.t. $p \in \mathcal{P}$. This follows from the fact that fixed $p$ the estimate which minimises the squared loss is the posterior mean. However, the estimate $\hat{x}_2$ to dominate $\hat{x}_1$ under the criterion (14) must satisfy (16) for each $p \in \mathcal{P}$ (not only for the density $p$ such that $\hat{x}_2 = E_p[X|y]$). It is clear that any $\hat{x}_1$ inside $\mathcal{X}^*$ cannot be dominated. In fact, by considering the $p$ such that $\hat{x}_1 = E_p[X|y]$, the right hand side of (16) is negative for any $\hat{x}_2 \neq \hat{x}_1$. Thus, a sufficient condition for $\hat{x}_2$ to be undominated is that $\hat{x}_2 \in \mathcal{X}^*$.*

*To prove that the condition $\hat{x}_2 \in \mathcal{X}^*$ is also necessary, we must show that given $x_1 \notin \mathcal{X}^*$ there exists $\hat{x}_2 \in \mathcal{X}^*$ such that (16) holds for any $p \in \mathcal{P}$.*

*Consider a given $p \in \mathcal{P}$, then*

$$\begin{array}{rcl} 0 & \leq & \int_{\mathcal{Y}} \int_{\mathcal{X}} \left[(X-\hat{x}_1)^T(X-\hat{x}_1) - (X-\hat{x}_2)^T(X-\hat{x}_2)\right] p(x,y)dxdy \\ & = & \int_{\mathcal{Y}} \left[(\hat{x}^*-\hat{x}_1)^T(\hat{x}^*-\hat{x}_1) - (\hat{x}^*-\hat{x}_2)^T(\hat{x}^*-\hat{x}_2)\right] p(y)dy, \end{array} \tag{17}$$

[6] In general it is not true that the set of maximal undominated actions coincides with the union of the optimal Bayesian actions obtained by minimising the loss function w.r.t. each probability in $\mathcal{P}$, see [18, Sec. 3.9.5] for details.

*where $\hat{x}^* = E_p[X|y] \in \mathcal{X}^*$. This follows from the property of the quadratic form $(\cdot)^T(\cdot)$. By selecting $\hat{x}_2$ to be the euclidean orthogonal projection of $\hat{x}_1$ on $\mathcal{X}^*$ and by noticing that:*

$$\begin{aligned}&(\hat{x}^* - \hat{x}_1)^T(\hat{x}^* - \hat{x}_1) = (\hat{x}^* - \hat{x}_2 + \hat{x}_2 - \hat{x}_1)^T(\hat{x}^* - \hat{x}_2 + \hat{x}_2 - \hat{x}_1)\\ &= ||\hat{x}^* - \hat{x}_2||^2 + ||\hat{x}_2 - \hat{x}_1||^2 + 2||\hat{x}^* - \hat{x}_2|| \cdot ||\hat{x}_2 - \hat{x}_1|| \cos(\theta)\\ &\leq ||\hat{x}^* - \hat{x}_2||^2 + ||\hat{x}_2 - \hat{x}_1||^2,\end{aligned}$$

*where $||x_i - x_j||^2 = (x_i - x_j)^T(x_i - x_j)$, $\theta$ is the angle between the vectors $\hat{x}_2 - \hat{x}_1$ and $\hat{x}^* - \hat{x}_2$, which is greater than or equal to $\pi/2$[7] since $\mathcal{X}^*$ is convex (see the last part of the proof for the convexity of $\mathcal{X}^*$). Figure 1 explains the geometry of the above inequality in case $\mathcal{X}^*$ is a circle.*

*Therefore, it follows that*

$$(\hat{x}^* - \hat{x}_1)^T(\hat{x}^* - \hat{x}_1) - (\hat{x}^* - \hat{x}_2)^T(\hat{x}^* - \hat{x}_2) \geq (\hat{x}_2 - \hat{x}_1)(\hat{x}_2 - \hat{x}_1)^T > 0,$$

*and, thus, (17) holds for any value of $y$. Thus, the projection of $\hat{x}_1$ on $\mathcal{X}^*$ is an estimate that dominates $\hat{x}_1$ for any $p \in \mathcal{P}$.*

*The convexity of $\mathcal{X}^*$ follows from the fact that $\mathcal{P}$ is a closed convex set. In fact consider the convex combination $\alpha\hat{x}_a + (1-\alpha)\hat{x}_b$ with $\hat{x}_a, \hat{x}_b \in \mathcal{X}^*$ and $\alpha \in (0,1)$, then*

$$\alpha\hat{x}_a + (1-\alpha)\hat{x}_b = \int_{\mathcal{X}} x(\alpha p_a(x|y) + (1-\alpha)p_b(x|y))dx,$$

*where $p_a(x|y), p_b(x|y)$ are the densities whose posterior means are $\hat{x}_a, \hat{x}_b$. Thus, being $(\alpha p_a(x|y) + (1-\alpha)p_b(x|y)) \in \mathcal{P}$ for each $\alpha \in (0,1)$, it follows that $\alpha\hat{x}_a + (1-\alpha)\hat{x}_b \in \mathcal{X}^*$ for each $\alpha \in (0,1)$.* ■

It should be noticed that in the scalar case $\mathcal{X}^*$ reduces to the interval $[\underline{E}[X|y], \overline{E}[X|y]]$. In the vectorial case, the set $\mathcal{X}^*$ is included in the box $[\underline{E}[X|y], \overline{E}[X|y]]$.

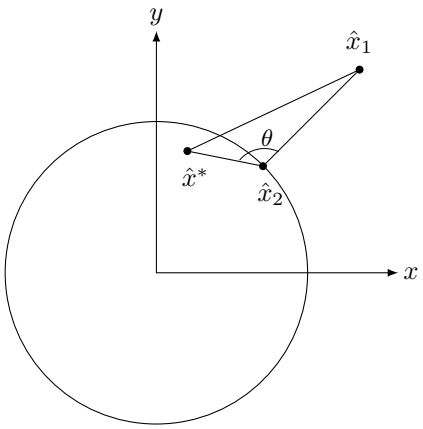



Figure 1: Dominating estimator.

## 6. Constant odds-ratio model

Consider a bounded scalar function $g$ on $\mathcal{X}$ and define the lower expectation of $g$, denoted by $\underline{E}(g)$, as the unique solution $\mu$ of

$$(1-\epsilon)E_0((g-\mu)^+) - E_0((g-\mu)^-) = 0, \tag{18}$$

where $(g-\mu)^+ = \max(g-\mu, 0)$ and $(g-\mu)^- = -\min(g-\mu, 0)$ are the positive and, respectively, negative part of $g-\mu$, $E_0(\cdot)$ is the expectation w.r.t. some nominal probability $P_0$, and the constant $\epsilon$ is a design

[7] See for instance [26, Th. 1.2.4.] for the proof that $\cos(\theta) \leq 0$.

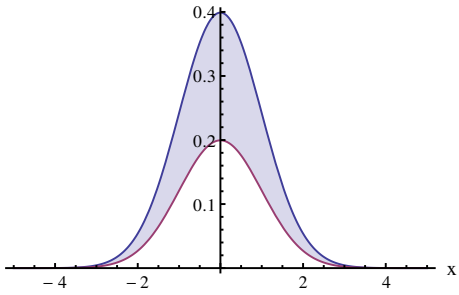



Figure 2: Set of densities (filled area) defined by the COR model.

parameter belonging to $(0, 1)$. The resulting lower expectation is called the *constant odds-ratio* (COR) model [18, Sec. 2.9.4]. To define (18), the modeller is therefore required to choose $E_0(\cdot)$ and to specify the value of $\epsilon \in (0, 1)$.

What is the set of probabilities associated to (18)? Assume that $E_0(g) = \int_{\mathcal{X}} g(x)p_0(x)dx$ where $p_0(x)$ is the PDF (w.r.t. the Lebesgue measure) of $X$ under $E_0$, assuming it exists. By setting $\mu = \underline{E}(g)$, the expression in (18) can then be rewritten as follows:

$$\int_{\mathcal{X}} (g - \underline{E}(g))^{+}(1 - \epsilon)p_0(x)dx - \int_{\mathcal{X}} (g - \underline{E}(g))^{-}p_0(x)dx = 0, \tag{19}$$

or, equivalently,

$$\inf_{(1-\epsilon)p_0(x)<p(x)<p_0(x)} \int_{\mathcal{X}} (g(x) - \underline{E}(g))p(x)dx = 0. \tag{20}$$

In fact, the lower of the above integral is simply obtained by selecting $p(x) = (1 - \epsilon)p_0(x)$ in the region of the space $\{x : g(x) - \underline{E}(g) \geq 0\}$ and $p(x) = p_0(x)$ in the region $\{x : g(x) - \underline{E}(g) < 0\}$. The expression (20) can also be rewritten as:

$$\underline{E}(g) = \inf_{(1-\epsilon)p_0(x)<p(x)<p_0(x)} \frac{\int_{\mathcal{X}} g(x)p(x)dx}{\int_{\mathcal{X}} p(x)dx}, \tag{21}$$

although the expression (19) is operatively more convenient, since it transforms a minimisation (infimum) problem in an equation. From Expressions (20)–(21) it follows that the COR model accounts for imprecision in the knowledge of the nominal PDF $p_0$ and this imprecision can equivalently be represented by the following set of unnormalised densities:

$$\mathcal{P}_0 = \left\{ p : \ (1 - \epsilon) \leq \frac{p(x)}{p_0(x)} \leq 1 \right\}. \tag{22}$$

For this reason the COR model is also known with the name of *density ratio class* [21, 22] and also as *interval of measures* [23]. Figure 2 shows the bound $(1 - \epsilon)p_0(x) \leq p(x) \leq p_0(x)$ in case $p_0(x) = \mathcal{N}(x; 0, 1)$, $\epsilon = 0.5$ and $x \in \mathbb{R}$.

Consider the special case $g = I_{\{B\}}$, where $B \subseteq \mathcal{X}$ and $I_{\{B\}}$ is the indicator function of $B$ (i.e., $I_{\{B\}}(x) = 1$ if $x \in B$ and zero otherwise), then from (19) it follows that:

$$\underline{P}(B) = \underline{E}(I_{\{B\}}) = \frac{(1 - \epsilon)P_0(B)}{1 - \epsilon P_0(B)}, \quad \overline{P}(B) = \overline{E}(I_{\{B\}}) = \frac{P_0(B)}{1 - \epsilon + \epsilon P_0(B)}. \tag{23}$$

The equality $\underline{P}(B) = \underline{E}(I_{\{B\}})$ (respectively $\overline{P}(B) = \overline{E}(I_{\{B\}})$) follows from the fact that the expectation of an indicator over a measurable subset of $\mathcal{X}$ gives the probability of such subset and, thus, $\underline{E}(I_{\{B\}})$ ($\overline{E}(I_{\{B\}})$) gives the lower (upper) probability of $B$. The COR model thus provides a simple formula (23) to compute the lower and upper probability of subsets of $\mathcal{X}$.

Notice also that for $B = \mathcal{X}$ one gets correctly $\underline{E}(I_{\{B\}}) = \overline{E}(I_{\{B\}}) = 1$ and, furthermore, that (23) satisfies all the properties (C1)–(C3) defined in Section 3 when $g_1$ and $g_2$ are indicator functions over subsets of $\mathcal{X}$. Therefore, (23) defines a consistent (or, more precisely, a *coherent*) lower probability.[8]

In case the size of $B$ is small, i.e., $P_0(B) \approx 0$, (23) reduces to

$$\underline{P}(B) \approx (1-\epsilon)P_0(B), \quad \overline{P}(B) \approx (1-\epsilon)^{-1}P_0(B),$$

which allows us to give a more direct interpretation of the COR model under the point of view of a modeller. The model (23) can in fact be used to account for the following kind of uncertainty. Assume that we specify a Gaussian density as nominal model $p_0(x)$ and we use this density to compute a 95% credible (Bayesian confidence) region for the value of the variable $X$. However, we are not very confident that the probability that the true value of the variable belongs to this set with probability 0.95, but we consider the possibility (e.g., $\epsilon \in (0,1)$) that this probability can be between $(1-\epsilon)0.95$ and $(1-\epsilon)^{-1}0.95$. Furthermore, if we impose the constraints (C1)–(C3), which ensure that the set of probabilities bounded by $\underline{P}(B)$ and $\overline{P}(B)$ for each $B \subseteq \mathcal{X}$ is coherent, we obtain the bounds (23) that can then be extend to all the bounded scalar functions $g$ to finally obtain (19). The work in [27] discusses a different elicitation procedure for COR models based on quantiles.

The COR model is of interest in robust filtering, since it allows to model the imprecision by simply specifying bounds on the PDF. Furthermore, these bounds are defined by only two quantities: $1-\epsilon$ which determines the degree of imprecision and the nominal density $p_0$. The set $\mathcal{P}_0$ in (22) can include unimodal (multimodal) densities whose maximum(s) can get farther from zero as the value of $\epsilon$ gets larger. Thus, the COR model allows us to account for a variety of shapes that are in general critical for robustness in state estimation. Despite it allows for this wide variety of shapes, the COR model is not too imprecise. The tail behaviour is in fact determined by the nominal density $p_0(x)$, which essentially allows to mainly restrict the imprecision to the high density region of the support of $p_0(x)$.

An example of bimodal distribution included in the COR model of Figure 2 is shown in Figure 3. Notice that the imprecision is mainly restricted to the $3\sigma$ interval of the Gaussian.

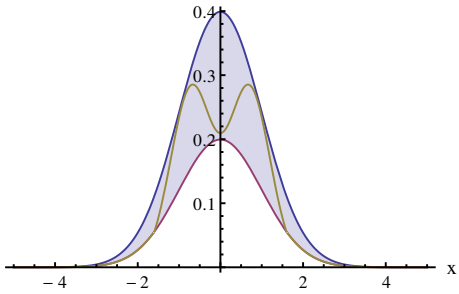



Figure 3: Bimodal density included in the COR model.

### *6.1. Properties of the COR model: updating and prediction*

The COR model has several nice properties [22] that we review in the following two subsections.

Assume that $\mathcal{P}_0$ in (22) expresses our prior information on $X$ and consider as likelihood model another COR model defined by the set of densities:

$$\mathcal{P}_{y|x} = \left\{ p : \ (1-\epsilon_m) \leq \frac{p(y|x)}{p_0(y|x)} \leq 1 \right\}, \tag{24}$$

[8] Observe that the lower expectation (19) satisfies (C1)–(C3) also for all bounded scalar function $g_1$ and $g_2$ and, thus, it is also coherent.

for any value $x$ of $X$, where $p_0(y|x)$ is the nominal density that defines the COR model and the constant $\epsilon_m \in (0,1)$ is the imprecision which, for simplicity, we assume not depending on $x$. How can we compute the posterior CLP of $g : \mathcal{X} \to \mathbb{R}$ given $Y = \tilde{y}$?

*Theorem 4. Consider the COR model in (18), its associated set of prior densities $\mathcal{P}_0$ in (22) and the COR likelihood model defined by (24). Assume that:*

$$\int_{\mathcal{Y}} (\delta_{\{\tilde{y}\}}(y) - \alpha)^+ h(x,y)dy = h(x,\tilde{y}), \tag{25}$$

*for each $x \in \mathcal{X}$, for any positive scalar $\alpha$ and continuous and bounded scalar function $h$ on $\mathcal{X} \times \mathcal{Y}$. Assume also that $\underline{E}_X(p_0(\tilde{y}|x)) > 0$ and that $p_0(\tilde{y}|x)$ is continuous and bounded for any $x \in \mathcal{X}$ and given $Y = \tilde{y}$. The posterior lower expectation of $g : \mathcal{X} \to \mathbb{R}$ given $Y = \tilde{y}$, denoted by $\underline{E}(g|\tilde{y})$, is the unique solution $\mu$ of*

$$\int (g-\mu)^+(1-\epsilon)(1-\epsilon_m)p_1(x|\tilde{y})dx + \int (g-\mu)^- \frac{1}{1-\epsilon_m} p_1(x|\tilde{y})dx = 0, \tag{26}$$

*where $p_1(x|\tilde{y}) = p_0(\tilde{y}|x)p_0(x)/\int p_0(\tilde{y}|x)p_0(x)dx$.* ■

*Proof: By Theorem 2, it follows that $\underline{E}(g|\tilde{y})$ is the unique solution $\mu$ of*

$$\underline{E}_X(\underline{E}_Y(\delta_{\{\tilde{y}\}} \cdot (g-\mu)|X)) = 0. \tag{27}$$

*Since $(g-\mu)$ is a function of $X$ only, it follows that:*

$$\underline{E}_Y(\delta_{\{\tilde{y}\}} \cdot (g-\mu)|X) = (g-\mu)^+ \underline{E}_Y(\delta_{\{\tilde{y}\}}|X) + (g-\mu)^- \overline{E}_Y(\delta_{\{\tilde{y}\}}|X).$$

*Since $\underline{E}_Y(\cdot|X)$ is a COR model, $\underline{E}_Y(\delta_{\{\tilde{y}\}}|X)$ can be computed as follows:*

$$\int_{\mathcal{Y}} (\delta_{\{\tilde{y}\}}(y) - \underline{E}_Y(\delta_{\{\tilde{y}\}}|x))^+ (1-\epsilon_m)p_0(y|x)dy - \int_{\mathcal{Y}} (\delta_{\{\tilde{y}\}}(y) - \underline{E}_Y(\delta_{\{\tilde{y}\}}|x))^- p_0(y|x)dy = 0, \tag{28}$$

*for each value $x$ of $X$. Observe that*

$$\begin{array}{rcl} -\int_{\mathcal{Y}} (\delta_{\{\tilde{y}\}}(y) - \underline{E}_Y(\delta_{\{\tilde{y}\}}|x))^- p_0(y|x)dy & = & \int_{\mathcal{Y}} (\delta_{\{\tilde{y}\}}(y) - \underline{E}_Y(\delta_{\{\tilde{y}\}}|x)) p_0(y|x)dy \\ & - & \int_{\mathcal{Y}} (\delta_{\{\tilde{y}\}}(y) - \underline{E}_Y(\delta_{\{\tilde{y}\}}|x))^+ p_0(y|x)dy. \end{array}$$

*From (25) with $h(x,y) = p(y|x)$, it then follows that (28) is equivalent to*

$$(1-\epsilon_m)p_0(\tilde{y}|x) - \underline{E}_Y(\delta_{\{\tilde{y}\}}|x) \int_{\mathcal{Y}} p_0(y|x)dy = 0,$$

*and, thus,*

$$\underline{E}_Y(\delta_{\{\tilde{y}\}}|x) = (1-\epsilon_m)p_0(\tilde{y}|x).$$

*The upper expectation $\overline{E}_Y(\delta_{\{\tilde{y}\}}|x)$ can be obtained by considering a piecewise density that is equal to $p_0(y|x)$ in the region where $\delta_{\{\tilde{y}\}}(y) - \overline{E}_Y(\delta_{\{\tilde{y}\}}|x)$ is positive and to $(1-\epsilon_m)p_0(y|x)$ in the region where $\delta_{\{\tilde{y}\}}(y) - \overline{E}_Y(\delta_{\{\tilde{y}\}}|x)$ is negative, i.e.,*

$$\int_{\mathcal{Y}} (\delta_{\{\tilde{y}\}}(y) - \overline{E}_Y(\delta_{\{\tilde{y}\}}|x))^+ p_0(y|x)dy - (1-\epsilon_m) \int_{\mathcal{Y}} (\delta_{\{\tilde{y}\}}(y) - \overline{E}_Y(\delta_{\{\tilde{y}\}}|x))^- p_0(y|x)dy = 0,$$

*and, thus,*

$$\overline{E}_Y(\delta_{\{\tilde{y}\}}|x) = (1-\epsilon_m)^{-1} p_0(\tilde{y}|x).$$

*By replacing the lower and upper in (27), one gets*

$$\underline{E}_X\left[(g-\mu)^+(1-\epsilon_m)p_0(\tilde{y}|x) + (g-\mu)^-(1-\epsilon_m)^{-1}p_0(\tilde{y}|x)\right] = 0. \tag{29}$$

*Since $\underline{E}_X$ is COR model, (26) follows straightforwardly from the definition (19) dividing by $\int_{\mathcal{X}} p_0(x)p_0(\tilde{y}|x)dx$, which is positive because of the assumption $\underline{E}(p_0(\tilde{y}|x)) > 0$.* ■

Theorem 4 shows that, after observing $Y = \tilde{y}$ and in the case $\epsilon_m = 0$, the set of priors $\mathcal{P}_0$ is updated to a model of the same form:

$$\mathcal{P}_1 = \{p : \ (1-\epsilon)p_1(x) \leq p(x) \leq p_1(x)\}, \tag{30}$$

but with $p_1(x|\tilde{y}) = p_0(\tilde{y}|x)p_0(x)/\int p_0(\tilde{y}|x)p_0(x)dx$ (the case $\epsilon_m = 0$ was proved in [22]). Therefore, we must only update the nominal density from $p_0(x)$ to $p_1(x|\tilde{y})$ in order to obtain the updated COR model after observing $Y = \tilde{y}$. In case $\epsilon_m > 0$, the posterior model is again a COR model but in which the lower bound for the densities now is scaled by $(1-\epsilon)(1-\epsilon_m)^2$.

One could wonder if the assumption (25) is too restrictive. The following example shows that for instance in case the "nascent" delta function (i.e., the limiting sequence of densities that generates the Dirac's delta) is a Gaussian pulse, then (25) holds.

*Example 1. Consider the following integral*

$$\int \left(\delta_{\{0\}} - \alpha\right)^+ h(y)dy = \lim_{\sigma\to 0} \int \left(\frac{1}{\sqrt{2\pi}\sigma}\exp\left(-\frac{y^2}{2\sigma^2}\right) - \alpha\right)^+ h(y)dy.$$

*For $\sigma < \alpha\sqrt{2\pi}$, the inequality*

$$\frac{1}{\sqrt{2\pi}\sigma}\exp\left(-\frac{y^2}{2\sigma^2}\right) \geq \alpha,$$

*is satisfied for any*

$$-\sigma\sqrt{2\left(-\ln\left(\alpha\sqrt{2\pi}\sigma\right)\right)} \leq y \leq \sigma\sqrt{2\left(-\ln\left(\alpha\sqrt{2\pi}\sigma\right)\right)}.$$

*By using the above bounds as integration limits, one gets*

$$\lim_{\sigma\to 0} \int_{-\sigma\sqrt{2\left(-\ln\left(\alpha\sqrt{2\pi}\sigma\right)\right)}}^{\sigma\sqrt{2\left(-\ln\left(\alpha\sqrt{2\pi}\sigma\right)\right)}} \left(\frac{1}{\sqrt{2\pi}\sigma}\exp\left(-\frac{y^2}{2\sigma^2}\right) - \alpha\right) h(y)dy.$$

*For $\sigma$ suitably small and assuming that $h(y)$ is continuous and bounded around zero [24, Ch. 1], one has:*

$$\lim_{\sigma\to 0} h(0) \int_{-\sigma\sqrt{2\left(-\ln\left(\alpha\sqrt{2\pi}\sigma\right)\right)}}^{\sigma\sqrt{2\left(-\ln\left(\alpha\sqrt{2\pi}\sigma\right)\right)}} \left(\frac{1}{\sqrt{2\pi}\sigma}\exp\left(-\frac{y^2}{2\sigma^2}\right) - \alpha\right) dy.$$

*By a change of variables, $z = y/\sigma$, one gets*

$$\lim_{\sigma\to 0} h(0) \left[\int_{-\sqrt{2\left(-\ln\left(\alpha\sqrt{2\pi}\sigma\right)\right)}}^{\sqrt{2\left(-\ln\left(\alpha\sqrt{2\pi}\sigma\right)\right)}} \frac{1}{\sqrt{2\pi}}\exp\left(-\frac{z^2}{2}\right) dz - \int_{-\sqrt{2\left(-\ln\left(\alpha\sqrt{2\pi}\sigma\right)\right)}}^{\sqrt{2\left(-\ln\left(\alpha\sqrt{2\pi}\sigma\right)\right)}} \alpha\sigma dz\right].$$

*By noticing that $-\ln(\sigma) \to \infty$ for $\sigma \to 0$ and that $-\sigma\ln(\sigma) \to 0$ for $\sigma \to 0$, the first integral tends to 1 while the second integral to zero. Thus the above limit is equal to $h(0)$.* ■

The following theorem shows that, conversely, the prediction step does not preserve the structure of the COR model.

*Theorem 5. Consider the following unconditional and conditional COR models*

$$\inf_{(1-\epsilon_0)p_0(x_0)<p(x_0)<p_0(x_0)} \int_{\mathcal{X}_0} (g(x_0) - \underline{E}_{X_0}(g))p(x_0)dx_0 = 0, \tag{31}$$

*and*

$$\inf_{(1-\epsilon_1)p_1(x_1|X_0)<p(x_1|X_0)<p_1(x_1|X_0)} \int_{\mathcal{X}_1} (g'(x_1) - \underline{E}_{X_1}(g'|X_0))p(x_1|x_0)dx_1 = 0, \tag{32}$$

*where* $g : \mathcal{X}_0 \to \mathbb{R}$, $g' : \mathcal{X}_1 \to \mathbb{R}$ *and* $\epsilon_0, \epsilon_1 \in (0,1)$. *The lower expectation* $\underline{E}_{X_1}(g') := \underline{E}_{X_0}(\underline{E}_{X_1}(g'|X_0))$ *is the unique solution of*

$$\inf_{(1-\epsilon_0)(1-\epsilon_1)p_0(x_0)p_1(x_1|x_0)<p(x_0)p(x_1|x_0)<p_0(x_0)p_1(x_1|x_0)} \frac{\int_{\mathcal{X}_0} dx_0}{\frac{1}{\int_{\mathcal{X}_1} p(x_1|x_0)dx_1} \int_{\mathcal{X}_1} (g'(x_1) - \underline{E}_{X_0}(\underline{E}_{X_1}(g'|X_0)))p(x_1|x_0)p(x_0)dx_1 = 0.} \tag{33}$$

■

*Proof: Rewrite (32) as follows*

$$\underline{E}_{X_1}(g'|x_0) = \inf_{(1-\epsilon_1)p_1(x_1|x_0)<p(x_1|x_0)<p_1(x_1|x_0)} \frac{\int_{\mathcal{X}_1} g'(x_1)p(x_1|x_0)dx_1}{\int_{\mathcal{X}_1} p(x_1|x_0)dx_1}; \tag{34}$$

*replace g in (31) with* $\underline{E}_{X_1}(g'|X_0)$ *to obtain*

$$\inf_{(1-\epsilon_0)p_0(x_0)<p(x_0)<p_0(x_0)} \frac{\int_{\mathcal{X}_0} dx_0}{\frac{1}{\int_{\mathcal{X}_1} p(x_1|x_0)dx_1} \int_{\mathcal{X}_1} (g'(x_1) - \underline{E}_{X_0}(\underline{E}_{X_1}(g'|X_0)))p(x_1|x_0)p(x_0)dx_1 = 0,} \inf_{(1-\epsilon_1)p_1(x_1|x_0)<p(x_1|x_0)<p_1(x_1|x_0)} \tag{35}$$

*which is equivalent to (33).* ■

The case in which $\epsilon_1 = 0$ has been proved in [28, Ch. 4]. It can be observed that, because of the term $1/(\int_{\mathcal{X}_1} p(x_1|x_0)dx_1)$, (33) is not a COR model. However, this term is essential to prevent that the imprecision grows in time.

On the other hand, since (33) has not the structure of a COR model, it implies that no recursive solution is available for the COR model.

## 7. Properties of the COR model in the Gaussian case

In this section, we discuss some properties of the COR model in the case the nominal model is a Gaussian density.

*Theorem 6. Given the COR model defined by*

$$\mathcal{P}_0 = \{p : \ (1-\epsilon)\mathcal{N}(x; \hat{x}_0, P_0) \leq p(x) \leq \mathcal{N}(x; \hat{x}_0, P_0)\}, \tag{36}$$

*the set of posterior means* $\mathcal{X}^*$ *defined in (15) is the following ellipsoid:*

$$\mathcal{X}^* = \left\{(x - x_0)^T P_0^{-1}(x - x_0) \leq \hat{\gamma}^2\right\}. \tag{37}$$

$\hat{\gamma}$ *denotes the solution of*

$$\gamma = \epsilon[\phi(\gamma) + \gamma\Phi(\gamma)], \tag{38}$$

*where* $\phi(\cdot)$ *and* $\Phi(\cdot)$ *are respectively the standard Gaussian density and the standard cumulate distribution function.*

*The lower and upper probability of measurable subsets of* $B \subseteq \mathcal{X}$ *are given by:*

$$\begin{aligned} \underline{P}(B) &= \frac{(1-\epsilon)P_0(B)}{1-\epsilon P_0(B)}, \\ \overline{P}(B) &= \frac{P_0(B)}{(1-\epsilon)+\epsilon P_0(B)}. \end{aligned} \tag{39}$$

*The minimum volume ellipsoid that has lower probability* $1-\alpha$ *of including the true value of* $X$ *has the following shape:*

$$\xi = \left\{x : \ (x - \hat{x}_0)^T(\rho(\alpha)P_0)^{-1}(x - \hat{x}_0) \leq 1\right\}, \tag{40}$$

*where the scaling factor $\rho(\alpha) > 0$ ensures that the probability of $x$ to be in $\xi$ is at least $1-\alpha$.* ■

*Proof: Consider the eigenvalue-eigenvector decomposition $P_0 = V\Lambda^{-1}V^T$ with $|V| = 1$ and $VV^T = \mathbb{I}$, where $\mathbb{I}$ is the identity matrix, and define $z = V^T\Lambda^{-\frac{1}{2}}x$, $z_0 = V^T\Lambda^{-\frac{1}{2}}\hat{x}_0$, $dz \propto dx$, then*

$$\begin{aligned}\mathcal{X}^* &= \left\{\mu : \int(x-\mu)p(x)dx = 0,\ \forall\ (1-\epsilon)\mathcal{N}(x;\hat{x}_0,P_0) \le p(x) \le \mathcal{N}(x;\hat{x}_0,P_0)\right\}\\ &= \left\{\mu : \int(x-\mu)q(x)\mathcal{N}(x;\hat{x}_0,P_0)dx = 0,\ \forall\ (1-\epsilon) \le q(x) \le 1\right\}\\ &= \left\{\mu_0 : \int(z-\mu_o)q(\Lambda^{\frac{1}{2}}Vz)\mathcal{N}(z;z_0,\mathbb{I})dz = 0\ ,\ \forall\ (1-\epsilon) \le q(\Lambda^{\frac{1}{2}}Vz) \le 1\right\},\end{aligned} \tag{41}$$

*where $\mu = \Lambda^{-\frac{1}{2}}V^T\mu_0$. Since the bounds in $(1-\epsilon) \le q(\Lambda^{\frac{1}{2}}Vz) \le 1$ do not depend on $z$ , the value at which $q(\cdot)$ is computed does not matter: we can thus replace $q(\Lambda^{\frac{1}{2}}Vz)$ with $q(z)$. Assume for the moment that $\mathcal{X} = \mathbb{R}^2$. Consider then the first component of $z-\mu_o$, i.e., the scalar $z_1-\mu_{o1}$, and assume we want to maximise it. Select then*

$$q(z_1,z_2) = q(z_2|z_1)q(z_1) = q(z_2|z_1)\left(I_{\{z_1-\mu_{o1}\ge 0\}} + I_{\{z_1-\mu_{o1}<0\}}(1-\epsilon)\right),$$

*where $q(z_1,z_2)$ has to satisfy $1-\epsilon \le q(z_1,z_2) \le 1$. Thus, consider*

$$\begin{aligned}0 &= \int(z_1-\mu_{o1})\left(I_{\{z_1-\mu_{o1}\ge 0\}}(z_1) + I_{\{z_1-\mu_{o1}<0\}}(z_1)\cdot(1-\epsilon)\right)\mathcal{N}(z_1;z_{01},1)\\ &\cdot \int q(z_2|z_1)\mathcal{N}(z_2;z_{02},1)dz_2dz_1\\ &= \int(z_1-\mu_{o1})I_{\{z_1-\mu_{o1}\ge 0\}}(z_1)\cdot\mathcal{N}(z_1;z_{01},1)\int q(z_2|z_1)\mathcal{N}(z_2;z_{02},1)dz_2dz_1\\ &+ \int(z_1-\mu_{o1})I_{\{z_1-\mu_{o1}<0\}}(z_1)\cdot(1-\epsilon)\mathcal{N}(z_1;z_{01},1)\int q(z_2|z_1)\mathcal{N}(z_2;z_{02},1)dz_2dz_1.\end{aligned} \tag{42}$$

*Observe that to maximise the above integral, one should maximise the integral $\int q(z_2|z_1)\mathcal{N}(z_2;z_{02},1)$ for the values $z_1$ such that $z_1-\mu_{o1} \ge 0$ and to minimise it for the values $z_1$ such that $z_1-\mu_{o1} < 0$. This means to select $q(z_2|z_1) = 1$ in the first case, and $q(z_2|z_1) = 1-\epsilon$ in the second case. However, in the second case $q(z_2|z_1)$ cannot be equal to $1-\epsilon$ otherwise $1-\epsilon \le q(z_1,z_2) \le 1$ is not satisfied. In other words, $q(z_2|z_1) = q(z_2) = 1$ in both cases. This gives the maximum of $E[Z_1]$, since in this case $\int q(z_2|z_1)\mathcal{N}(z_2;z_{02},1)dz_2 = 1$ and, thus, the joint COR model reduces to the single variable COR model of the variable $Z_1$. Thus, for any other value of $q(z_2|z_1)$ the upper expectation $\overline{E}[Z_1]$ cannot be greater than the upper expectation computed for the case $q(z_2|z_1) = 1$, because $\overline{E}[Z_1]$ is the upper expectation corresponding to the univariate case (only $Z_1$ is considered). The value $\mu_{o1}$ which solves the above equation in the case $q(z_2|z_1) = 1$ can be computed by:*

$$0 = \int(z_1-\mu_{o1})\left(I_{\{z_1-\mu_{o1}\ge 0\}} + I_{\{z_1-\mu_{o1}<0\}}(1-\epsilon)\right)\mathcal{N}(z_1;z_{01},1)dz_1,$$

*which gives the maximum of $\mu_{o1}$. The above equation can be rewritten as:*

$$\begin{aligned}0 &= \int_{-\infty}^{\mu_{o1}}(z_1-\mu_{o1})\mathcal{N}(z_1;z_{01},1)(1-\epsilon)dz_1\\ &+ \int_{\mu_{o1}}^{\infty}(z_1-\mu_{o1})\mathcal{N}(z_1;z_{01},1)dz_1\\ &= \int_{-\infty}^{(\mu_{o1}-z_{01})}(u+z_{01}-\mu_{o1})\mathcal{N}(u;0,1)(1-\epsilon)du\\ &+ \int_{-(\mu_{o1}-z_{01})}^{\infty}(u+z_{01}-\mu_{o1})\mathcal{N}(u;0,1)du\\ &= \int_{-\infty}^{\gamma}(u-\gamma)\mathcal{N}(u;0,1)(1-\epsilon)du\\ &+ \int_{\gamma}^{\infty}(u-\gamma)\mathcal{N}(u;0,1)du,\end{aligned}$$

*where $u = z_1 - z_{01}$ (change of variable) and $\gamma = \mu_{o1} - z_{01}$. Hence, it follows that*

$$0 = -(1-\epsilon)\phi(\gamma) - (1-\epsilon)\gamma\Phi(\gamma) + \phi(\gamma) - \gamma(1-\Phi(\gamma))$$

*or, equivalently $\gamma = \epsilon[\phi(\gamma) + \gamma\Phi(\gamma)]$. Let $\hat{\gamma}$ be the value that solves the above expression, then from $\gamma = \mu_{o1} - z_{01}$ it follows that*

$$\overline{E}(Z_1) = \mu_{o1} = z_{01} + \hat{\gamma}. \tag{43}$$

*For $q(z_2|z_1) = q(z_2) = 1$ the last equality in (41) for $z_2$ becomes*

$$\int (z_2 - \mu_{o2}) \mathcal{N}(z_2; z_{02}, 1) dz_2 = 0,$$

*and it is satisfied if $\mu_{o2} = z_{02}$.*

*Therefore $\overline{E}(Z_1) = z_{01} + \hat{\gamma}$ and $E(Z_2) = z_{02}$. The lower for $E(Z_1)$ can be determined in a similar way $\underline{E}(Z_1) = z_{01} - \hat{\gamma}$ and $E(Z_2) = z_{02}$. By changing the roles of $Z_1$ and $Z_2$ one gets $E(Z_1) = z_{01}$ and $\overline{E}(Z_2) = z_{02} + \hat{\gamma}$ and, respectively, $E(Z_1) = z_{01}$ and $\underline{E}(Z_2) = z_{02} - \hat{\gamma}$. Considering the transformation $\mu = \Lambda^{\frac{1}{2}} V \mu_o$, one gets four points belonging to the border of $\mathcal{X}^*$. Notice that $\mathbb{I} = WW^T$ for any matrix $W$ such that $WW^T = \mathbb{I}$ and $|det(W)| = 1$. Then $\mathcal{N}(z; z_0, \mathbb{I})$ in (41) can equivalently be rewritten as $\mathcal{N}(z; z_0, WW^T)$. Repeating the derivations after (41) for the vector $z' = Wz$, one can find other two orthogonal directions (determined by the rows of $W$) w.r.t. which the previous four transformed points $(Wz_{01} \pm W\hat{\gamma}, Wz_{02})$ and $(Wz_{01}, Wz_{02} \pm W\hat{\gamma})$ are still extremes for the transformed domain $z'$. This can be repeated for any transformation $W$. Hence, it follows that in the $Z$ domain the set $\mathcal{Z}^*$ defined by the last equation in (41) is a circle centred at $z_0$, i.e., $(z - z_0)^T W^T W (z - z_0) \leq \hat{\gamma}^2$, and, thus, it becomes the ellipsoid (37) in the original domain $x$ after the transformation $x = \Lambda^{\frac{1}{2}} V z$.*

*The case with more than two dimensions can be treated in a similar way by rewriting $q(z_1, z_2, \ldots, z_n) = q(z_2, \ldots, z_n|z_1) q(z_1)$ and proceeding as before.*

*Equations (39) hold for any COR model; they follow from (23). The last part of the theorem, i.e., (40), follows from the fact that for $\epsilon = 0$, given $\alpha$ there exists $\rho(\alpha)$, such that the ellipsoid (40) is the minimum volume region that has probability $\alpha$ of including the true value of $X$ (this holds since the nominal density is Gaussian). Let us call this ellipsoid $\xi_1$. For $\epsilon > 0$, the minimum volume ellipsoid in (40) has to include $\xi_1$. The fact that it has the same eigenvector-eigenvalue decomposition of $\xi_1$ follows by the fact that the level curves of the lower and upper density in $\mathcal{P}_0$ are given by (40) as in the part of the proof that has proved (37).*

■

Figure 4 shows the value of $\hat{\gamma}$ as a function of $1 - \epsilon$. It should be pointed out that $\hat{\gamma}$ goes to infinity for $(1 - \epsilon) \to 0$. Figure 5 shows the shape of $\mathcal{X}^*$ in a two-dimensional case with $\epsilon = 0.5$, $\hat{\gamma} \approx 0.276$. $\hat{x}_0 = 0$ and

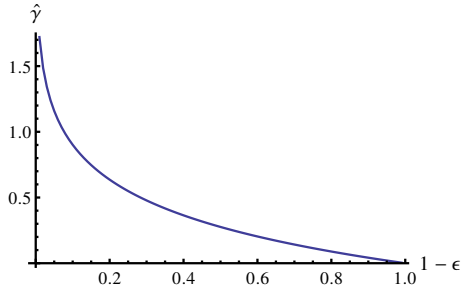



Figure 4: Values of $\hat{\gamma}$ as a function of $1 - \epsilon$.

$$P_0 = V \Lambda V' = \begin{bmatrix} 0.5 & -0.87 \\ -0.87 & -0.5 \end{bmatrix} \begin{bmatrix} 1 & 0 \\ 0 & 4 \end{bmatrix} \begin{bmatrix} 0.5 & -0.87 \\ -0.87 & -0.5 \end{bmatrix}.$$

Notice that, since $V$ is an orthonormal matrix ($det(V) = -1$), it corresponds to a rotation (60 degrees in the figure) in $\mathbb{R}^2$. Table 1 reports the values of $\rho(\alpha)$ that ensures the ellipsoid (40), in the standard bivariate case, i.e., $x^T x \leq \rho(\alpha)$, to include the true value of $X$ with lower probability equal to $1 - \alpha = 0.95$ for different values of $\epsilon$. For $\epsilon = 0$ (no imprecision), $\rho(\alpha)$ is the $(100\alpha)$th percentile of the chi-square distribution with $n = 2$ degrees of freedom. Some comments to Theorem 6:

- Although the ellipsoid $\mathcal{X}^*$ has the same shape of a credible set, it has a different meaning. It represents the uncertainty on the value of the mean $E(X)$ due to the fact that our knowledge on $X$ is imprecise,

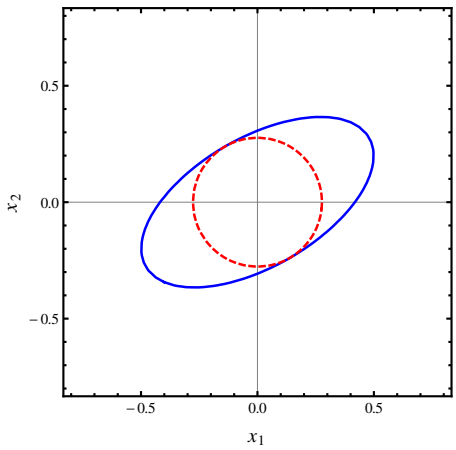



Figure 5: The set in the $Z$-plane is in red-dashed, while the set in the $X$-plane is in blue. The latter corresponds to $\mathcal{X}^*$.

| $\epsilon$ | $\rho(\alpha)$ | $\epsilon$ | $\rho(\alpha)$ |
|---|---|---|---|
| 0 | 5.99 | 0.6 | 7.68 |
| 0.1 | 6.15 | 0.7 | 8.39 |
| 0.2 | 6.47 | 0.8 | 9.22 |
| 0.3 | 6.64 | 0.9 | 10.60 |
| 0.4 | 6.89 | 0.99 | 15.08 |
| 0.5 | 7.31 | 0.999 | 19.46 |

Table 1: Scaling factor vs. imprecision for the standard bivariate ellipsoid $x^T x \leq \rho(\alpha)$ with $1-\alpha = 0.95$.

and, thus, represented through a set of probabilities. In the case $\epsilon = 0$ (no imprecision), this set reduce to a single point, that is the mean of the Gaussian nominal density.

- Conversely, the ellipsoid (40) is a credible ellipsoid. It represents the region of the space that has at least probability $1-\alpha$ of including the true value of $X$. In absence of imprecision $\epsilon = 0$ it reduces to the credible ellipsoid of the nominal Gaussian density.

The next corollary specialises Theorem 4 to the case in which the nominal density in the COR model is a Gaussian.

*Corollary 1. Assume that*

$$\mathcal{P}_0 = \{p : \ (1-\epsilon)\mathcal{N}(x; \hat{x}_0, P_0) \leq p(x) \leq \mathcal{N}(x; \hat{x}_0, P_0)\}\,, \tag{44}$$

*and that the likelihood model is also COR:*

$$\mathcal{P}_{y|x} = \{p : \ (1-\epsilon_m)\mathcal{N}(y; Cx, R) \leq p(y|x) \leq \mathcal{N}(y; Cx, R)\}\,. \tag{45}$$

*Having observed $Y = \tilde{y}$, the set of posteriors that determines the COR model is*

$$\mathcal{P}_{x|\tilde{y}} = \left\{p : \ (1-\epsilon)(1-\epsilon_m)^2\mathcal{N}(x; \hat{x}_1, P_1) \leq p(x|\tilde{y}) \leq \mathcal{N}(x; \hat{x}_1, P_1)\right\}, \tag{46}$$

*where $\hat{x}_1 = P_1\left(P_0^{-1}\hat{x}_0 + C^T R^{-1}\tilde{y}\right)$ and $P_1^{-1} = P_0^{-1} + C^T R^{-1} C$.* ■

*Proof: This follows from Theorem 4 and properties of the Gaussian density.* ■

From the expression (46), one could wonder what happens increasing the number of observations. If $n$ further observations of $x$ are available, the posterior COR model becomes:[9]

$$\mathcal{P}_{x|\tilde{y}^n} = \left\{p : \ (1-\epsilon)(1-\epsilon_m)^{2n}\mathcal{N}(x; \hat{x}_n, P_n) \leq p(x|\tilde{y}) \leq \mathcal{N}(x; \hat{x}_n, P_n)\right\}, \tag{47}$$

[9]The expression is valid by assuming the observations are epistemically independent given $X_0$. The irrelevance conditions in Theorem 2 must hold in both directions, see for instance [25] for more details.

where $\hat{x}_n = P_n\left(P_{n-1}^{-1}\hat{x}_{n-1} + C^T R^{-1}\tilde{y}_n\right)$ and $P_n^{-1} = P_{n-1}^{-1} + C^T R^{-1} C$. Notice that the imprecision grows as $(1-\epsilon_m)^{2n}$, while the variance of the Gaussian decreases as $1/\sqrt{n}$. Figure 6 plots the value of $\hat{\gamma}/\sqrt{n}$ for $\epsilon_m \in \{0.9, 0.8, 0.7, 0.6\}$ as a function of $n$. It can be noticed that the volume of $\mathcal{X}^*$ decreases with $n$. This is important because it means that the decrease of the variance is stronger than the increase of the imprecision. In other words, the uncertainty on the value of $X$, i.e., the volume of $\mathcal{X}^*$, is going to decrease at the accumulation of the evidence.

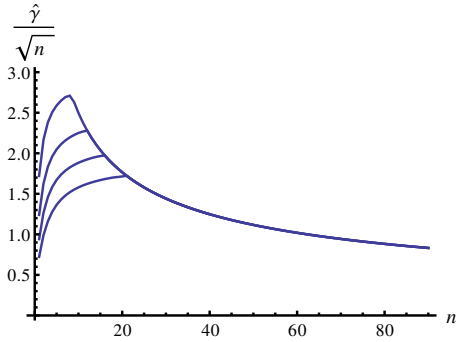



Figure 6: Values of $\hat{\gamma}/\sqrt{n}$ as a function of $n$ for $\epsilon = 1$ and $\epsilon_m \in \{0.9, 0.8, 0.7, 0.6\}$ (from the top to the bottom curve).

Finally, consider Theorem 5 in case $p_0(x_0) = \mathcal{N}(x_0; \hat{x}_0, P_0)$, $p_1(x_1|x_0) = \mathcal{N}(x_1; Ax_0, Q)$, $\epsilon_0 = \epsilon_1 = \epsilon$ and $g' : \mathcal{X} \to \mathbb{R}$, i.e.,

$$\begin{array}{l} \inf\limits_{(1-\epsilon)<q(x_0)<1} \int_{\mathcal{X}_0} q(x_0)dx_0 \inf\limits_{(1-\epsilon)<q(x_1|x_0)<1} \frac{1}{\int_{\mathcal{X}_1} q(x_1|x_0)\mathcal{N}(x_1; Ax_0, Q)dx_1} \\ \int_{\mathcal{X}_1}(g' - \underline{E}_{X_0}(\underline{E}_{X_1}(g'|x_0)))q(x_1|x_0)\mathcal{N}(x_1; Ax_0, Q)\mathcal{N}(x_0; \hat{x}_0, P_0)dx_1 = 0. \end{array} \tag{48}$$

From Theorem 5, it follows that prediction does not preserve the structure of the COR model either in the Gaussian case. This means that the results derived in Theorem 6 cannot be extended to the predictive model (48) and, thus, applied to the filtering problem. However, since the only informative part in the COR model (48) is represented by the joint density $\mathcal{N}(x_1; Ax_0, Q)\mathcal{N}(x_0; \hat{x}_0, P_0)$, our conjecture is that the set of posterior means $\mathcal{X}^*$ and the credible ellipsoid has the same shape as in Theorem 6, i.e., their directions are determined by the covariance matrix. This conjecture, that we intend to prove in future work, is confirmed by numerical simulations as it will be shown in Section 10.1.

The fact that the prediction does not preserve the structure of the COR model means also that a recursive solution for the filtering problem does not exist in the case COR sets are employed to model the state dynamics and the measurement equation.

There are thus two avenues that we can follow. The first is to outer-approximate the predictions drawn with (48) with the predictions obtained from an approximating COR model in order to keep a recursive structure of the filter. For this purpose, we can consider the COR model whose set of means includes that of (48) or, whose $(1-\alpha)\%$ ellipsoid includes that of (48) (the previous conjecture can be very useful for this purpose). By imposing these inclusion constraints one can derive the quantities that define a COR model, the mean and the variance of the Gaussian nominal density and the imprecision factor $\epsilon$.

The second possibility that we shall follow in this paper is to solve the filtering problem by using the general approach presented in Theorem 2. In other words, we abandon the idea of solving the filtering problem recursively and, at each time step, we compute the lower posterior expectation directly from the joint model (12). However, in order to compute (12), we need to address two issues. The first is to evaluate integrals numerically. An algorithm to perform such computation, which exploits Monte Carlo integration methods, is discussed in the next section (a general overview about Monte Carlo methods can be found in [29]). The second is to compute the set $\mathcal{X}^*$ of optimal Bayesian estimates and the relative credible region. Since the prediction step does not preserve the structure of the COR model, we cannot exploit Theorem 6. Therefore, we must compute these sets numerically as described in Section 9.

## 8. Monte Carlo integration

In robust filtering with sets of probabilities, the goal is to compute lower and upper expectations of a scalar function $g$ (we shall discuss the vectorial case in the next section). Consider for instance (48), how can one find $\underline{E}_{X_0}(\underline{E}_{X_1}(g'|x_0))$? The following algorithm describes the steps necessary to compute numerically $\underline{E}_{X_0}(\underline{E}_{X_1}(g'|x_0))$.

1. Set the sample sizes $n_0$, $n_1$ and the scaling factors $\beta_1, \beta_2 > 1$.
2. Sample $x_0^{(i)} \sim \mathcal{N}(x_0; \hat{x}_0, \beta_1 P_0)$ for $i = 1, \ldots, n_0$.
3. Sample $x_1^{(j_i)} \sim \mathcal{N}(x_1; Ax_0^{(i)}, \beta_2 Q)$ for $j_i = 1, \ldots, n_1$ and $i = 1, \ldots, n_0$.
4. Fix $i$ and set a numerical value for $\mu_1^{(i)} = \underline{E}_{X_1}(g'|x_0^{(i)})$ and for each $j_i = 1, \ldots, n_1$ evaluate the sign of $g'(x_1^{(j_i)}) - \mu_1^{(i)}$:
   (a) if $g'(x_1^{(j_i)}) - \mu_1^{(i)} \geq 0$ set $q(x_1^{(j_i)}|x_0^{(i)}) = 1 - \epsilon$;
   (b) else set $q(x_1^{(j_i)}|x_0^{(i)}) = 1$.
5. By applying a bisection method repeat steps 4(a)–4(b) to solve w.r.t. $\mu_1^{(i)}$ the following equation:
$$0 = \sum_{j=1}^{n_1} (g'(x_1^{(j_i)}) - \mu_1^{(i)}) q(x_1^{(j_i)}|x_0^{(i)}) \frac{\mathcal{N}(x_1^{(j_i)}; x_0^{(i)}, Q)}{\mathcal{N}(x_1^{(j_i)}; x_0^{(i)}, \beta_2 Q)}.$$
6. Repeat steps 4–5 to find $\hat{\mu}^{(i)}$, the root of the above equation, for each $i = 1, \ldots, n_0$.
7. For each $i = 1, \ldots, n_0$, define $g(x_0^{(i)}) = \mu_1^{(i)}$ and evaluate the sign of $g(x_0^{(i)}) - \mu$:
   (a) if $g(x_0^{(i)}) - \mu \geq 0$ set $q(x_0^{(i)}) = 1 - \epsilon$;
   (b) else set $q(x_0^{(i)}) = 1$.
8. By applying a bisection solve w.r.t. $\mu$:
$$0 = \sum_{i=1}^{n_0} (g(x_0^{(i)}) - \mu) q(x_0^{(i)}) \frac{\mathcal{N}(x_0^{(i)}; \hat{x}_0, P_0)}{\mathcal{N}(x_0^{(i)}; \hat{x}_0, \beta_1 P_0)}.$$
9. Increase $n_0$, $n_1$ and/or the scaling factors $\beta_1, \beta_2 > 1$ and repeat the previous steps up to the moment the value of $\hat{\mu}$ converges (its variations are below the prescribed level of accuracy).

The final solution $\hat{\mu}$ gives $\underline{E}_{X_0}(\underline{E}_{X_1}(g'|x_0))$. To compute the upper expectation one can exploit the fact that $\overline{E}_{X_0}(\overline{E}_{X_1}(g'|x_0)) = -\underline{E}_{X_0}(\underline{E}_{X_1}(-g'|x_0))$. Notice that the factors $\beta_1, \beta_2 > 1$ are used to increase the variance of the sampling distributions in order to speed up the convergence rate for fixed values of $n_0$, $n_1$ (in fact as it has been shown in Table 1, at the increasing of $\epsilon$ the mass of the COR models spreads in space). The procedure described in the previous algorithm can be generalised to any other case we can meet working with COR models.

## 9. Vector-valued and unbounded functions

In the above section we have defined properties of the COR model by considering the expectations of scalar bounded functions $g$. In the filtering problem, we are even interested to compute the lower (upper) expectation of $g = X$, which is unbounded and a vector in the multivariate case ($n > 1$). This means to solve a minimisation (maximisation) with a multi-objective unbounded cost function. In order to extend the previous results to vector-valued unbounded functions $g$, the idea is first to transform $g$ to an unbounded scalar function by multiplying its components for a weighting vector, e.g., $v^T x$ where $v \in \mathbb{R}^n$ and then to truncate it in a bounded region of $B \subset \mathbb{R}^n$, i.e., $v^T x I_{\{B\}}(x)$. In the bi-dimensional case, one can for instance consider $v = [\cos(\theta),\ \sin(\theta)]$ for $\theta \in [0, 2\pi)$. Fixed a direction $\theta$, we can then determine the minimum $r_m(\theta)$ and maximum $r_M(\theta)$ (depending on $\theta$) of $r = v^T x I_{\{B\}}(x)$:

$$\begin{cases} \overline{E}[(\cos(\theta)X_1 + \sin(\theta)X_2)I_{\{B\}}] &= r_M(\theta), \\ \underline{E}[(\cos(\theta)X_1 + \sin(\theta)X_2)I_{\{B\}}] &= r_m(\theta). \end{cases} \tag{49}$$

Let $E_p$ be the expectation w.r.t. a generic density $p$ in the COR set, from the above constraints, since $E_p[(\cos(\theta)X_1 + \sin(\theta)X_2)I_{\{B\}}] = \cos(\theta)E_p(X_1 I_{\{B\}}) + \sin(\theta)E_p(X_2 I_{\{B\}})$, it follows that:

$$\begin{cases} \cos(\theta)E_p(X_1 I_{\{B\}}) + \sin(\theta)E_p(X_2 I_{\{B\}}) &\leq r_M(\theta), \\ \cos(\theta)E_p(X_1 I_{\{B\}}) + \sin(\theta)E_p(X_2 I_{\{B\}}) &\geq r_m(\theta). \end{cases} \tag{50}$$

By taking the limit of the truncation so that $B = \mathbb{R}^n$ and, thus, $I_{\{B\}} = 1$,[10] the above equalities (the case in which $\leq$ and $\geq$ are strict) define the two tangent planes to $\mathcal{X}^*$ (defined in (15)) orthogonal to direction determined by $v$ and, thus, the inequalities determine a bounded region of $\mathbb{R}^n$ that include all the points of $\mathcal{X}^*$. By varying $\theta \in [0, 2\pi)$ and, thus, changing the direction, we can obtain an approximation of $\mathcal{X}^*$.

## 10. Practical implementation of the COR filter in the linear Gaussian case

The problem in the COR models based filtering is that no recursive solution exists. Hence, to compute the lower and upper posterior expectations, one has to go through the joint model at each time steps (as described in Theorem 2). Unfortunately, for the latter approach, one should notice that Theorem 2 in case of continuous state variables gives only a theoretical solution of the filtering problem (an infinite dimensional solution). A discretisation (approximation) of the state is thus necessary for practical implementations. As in the Bayesian case, one could discretise the state just in the regions of the space which have higher probability of including the true state (this is the approach followed in Monte Carlo methods). However, since no recursive solution is available for COR models, we cannot determine such regions recursively (for instance by using the sequential importance resampling algorithm, see for instance [30]). In other words, the solution of the filtering problem requires to sample a joint model, whose number of states increases in time. In the following, we describe a method to perform such sampling locally and efficiently by exploiting the observability of the dynamical system.

Consider the following three COR models for initial state, state dynamics and measurement equations:

$$\begin{aligned} \mathcal{P}_{X_0} &= \{(1-\epsilon_0)\mathcal{N}(x_0; \hat{x}_0, P_0) \leq p(x_0) \leq \mathcal{N}(x_0; \hat{x}_0, P_0)\}, \\ \mathcal{P}_{X_{k+1}|X_k} &= \{(1-\epsilon_s)\mathcal{N}(x_{k+1}; Ax_k, Q) \leq p(x_{k+1}|x_k) \leq \mathcal{N}(x_{k+1}; Ax_k, Q)\}, \\ \mathcal{P}_{Y_k|X_k} &= \{(1-\epsilon_m)\mathcal{N}(y_k; Cx_k, R) \leq p(y_k|x_k) \leq \mathcal{N}(y_k; Cx_k, R)\}. \end{aligned} \tag{51}$$

Observe that for $\epsilon_0 = \epsilon_m = \epsilon_s = 0$ (no imprecision) we shall be back to the KF case. We assume that

- the imprecision parameters $\epsilon$ and the matrices $A, C, P_0, Q, R$ are time-invariant;[11]

[10] We assume that this limit exists, it is finite and well defined. We intend to prove in future work that this holds provided that the unbounded function $g$ is absolutely integrable w.r.t. the densities in the COR set. This condition is verified for the Gaussian COR model considered later.

[11] The extension to the time-variant case is straightforward.

- the pair of matrices $(A, C)$ is observable being $\nu$ the observability index.[12]

The observability of the pair $(A, C)$ implies that the value of any state $x_k$ can be estimated from the system outputs $y_k$ that have been observed through the time interval $(k, k+\nu]$. In other words, $\nu$ observations are necessary and sufficient to determine an estimate of the components of the state $x_k$.

Assume for the moment that there is not imprecision $\epsilon_0 = \epsilon_m = \epsilon_s = 0$ and that the goal is to compute $E_{X_t}[g|\tilde{y}^t]$ directly (non recursively) from the joint $E_{X^t,Y^t}$. To achieve this goal we exploit the observability of the pair $(A, C)$ to sample locally from the joint $E_{X^t,Y^t}$. In particular, we split the time interval $[0, t]$ in $\lfloor t/\nu \rfloor$ ($\lfloor \cdot \rfloor$ denotes the floor function) observable parts, so that the states $x_k$ can be estimated from the observations in the interval $(k, k+\nu]$ for any $k = 0, \ldots, t$. In this way, each $\nu$ time instants we can sample the state $x_k$ and then propagate these samples up to the next observable state at time $k+\nu$ and so on. This means that we decompose the joint $E_{X^t,Y^t}[(g-\mu)\delta_{\{\tilde{y}^t\}}]$ in (3) as follows:

$$E_{X_0}\Bigg[E_{X_1}\Big[\cdots E_{X^{t-2\nu+3:t-\nu+1},Y^{t-2\nu+1:t-\nu}}\big[E_{X^{t-\nu+2:t},Y^{t-\nu+1:t}}[(g-\mu)\delta_{\{\tilde{y}^t\}}|X_{t-\nu+1}] \Big| X^{t-2\nu+2}\big]\cdots\Big|X_0\Big]\Bigg], \tag{52}$$

where $X^{t-\nu+2:t}$ denotes the sequence of states from time $t-\nu+2$ to time $t$ (similar for $Y$) and where $\nu$ is the observability index. In other words, we have decomposed the joint in $\lfloor t/\nu \rfloor$ conditional and observable parts, i.e., each part includes the minimum number of observations to estimate all the components of the state. Now consider the inner conditional joint $E_{X^{t-\nu+2:t},Y^{t-\nu+1:t}}[(g-\mu)\delta_{\{\tilde{y}^t\}}|x_{t-\nu+1}]$ for $X_{t-\nu+1} = x_{t-\nu+1}$ and $g : \mathcal{X}_t \to \mathbb{R}$, which is equal to

$$\begin{array}{l}\mathcal{N}(\tilde{y}_{t-\mu+1}; Cx_{t-\nu+1}, R)\int \mathcal{N}(x_{t-\nu+2}; Ax_{t-\mu+1}, Q)\mathcal{N}(\tilde{y}_{t-\mu+2}; Cx_{t-\nu+2}, R)dx_{t-\nu+2} \\ \cdots \int \mathcal{N}(x_{t-1}; Ax_{t-2}, Q)\mathcal{N}(\tilde{y}_{t-1}; Cx_{t-1}, R)dx_{t-1} \\ \cdot \int (g(x_t)-\mu)\mathcal{N}(x_t; Ax_t, Q)\mathcal{N}(\tilde{y}_t; Cx_t, R)dx_t.\end{array} \tag{53}$$

Define $R_t = R$, $C_t = C$, $z_t = \tilde{y}_t$ and apply the matrix inversion lemma to obtain:

$$\begin{array}{l}\mathcal{N}(x_t; Ax_{t-1}, Q)\ \mathcal{N}(z_t; C_t x_t, R_t) = \mathcal{N}(z_t; C_t Ax_{t-1}, V_t) \\ \cdot\mathcal{N}(x_t; W_t Q^{-1}Ax_{t-1} + W_t C_t^T R_t^{-1} z_t, W_t),\end{array} \tag{54}$$

where $W_t = R_t + C_t Q C_t^T$ and $V_t^{-1} = Q^{-1} + C_t^T R_t^{-1} C_t$. By rewriting

$$\mathcal{N}(\tilde{y}_{t-1}; Cx_{t-1}, R)\mathcal{N}(z_t; C_t Ax_{t-1}, V_t) = \mathcal{N}(z_{t-1}; C_{t-1}x_{t-1}, R_{t-1}),$$

where

$$z_{t-1} = \begin{bmatrix} \tilde{y}_{t-1} \\ \tilde{y}_t \end{bmatrix},\ \ C_{t-1} = \begin{bmatrix} C \\ CA \end{bmatrix}, R_{t-1} = \begin{bmatrix} R & 0 \\ 0 & V_t \end{bmatrix};$$

by applying again (54) to $\mathcal{N}(x_{t-1}; Ax_{t-2}, Q)\mathcal{N}(z_{t-1}; C_{t-1}x_{t-1}, R_{t-1})$ one gets:

$$\mathcal{N}(z_{t-1}; C_{t-1}Ax_{t-2}, V_{t-1}) \cdot \mathcal{N}(x_{t-1}; W_{t-1}Q^{-1}Ax_{t-2} + W_{t-1}C'_{t-1}R_{t-1}^{-1}z_{t-1}, W_{t-1}),$$

with $W_{t-1} = R_{t-1} + C_{t-1}QC_{t-1}^T$ and $V_{t-1}^{-1} = Q^{-1} + C_{t-1}^T R_{t-1}^{-1} C_{t-1}$. By proceeding recursively up to time $t-\nu+1$, one finally gets that (53) is equivalent to:

$$\mathcal{N}(z_{t-\mu+1}; C_{t-\mu+1}x_{t-\nu+1}, R_{t-\mu+1})f(x_{t-\nu+1}, \mu),$$

[12]The observability index is the smallest integer $\nu$ such that the matrix $[C, CA, \ldots, CA^{\nu-1}]^T$ has rank $n$.

where

$$z_{t-\mu+1} = \begin{bmatrix} \tilde{y}_{t-\mu+1} \\ \tilde{y}_{t-\mu+2} \\ \vdots \\ \tilde{y}_t \end{bmatrix}, \quad C_{t-\mu+1} = \begin{bmatrix} C \\ CA \\ \vdots \\ CA^{\mu-1} \end{bmatrix}, R_{t-\mu+1} = \begin{bmatrix} R & 0 \\ 0 & V_{t-\mu+2} \end{bmatrix},$$

$V_{t-\mu+2} = Q^{-1} + C_{t-\mu+2}^T R_{t-\mu+2}^{-1} C_{t-\mu+2}$ and

$$f(x_{t-\nu+1}, \mu) = \int \cdots \int (g(x_t) - \mu) \prod_{i=t-\mu+2}^{t} \mathcal{N}(x_i; W_i Q^{-1} A x_{i-1} + W_i C_i^T R_i^{-1} z_i, W_i) dx_{t-\mu+2} \cdots dx_t. \tag{55}$$

Observe that the matrix $C_{t-\mu+1}$ in $\mathcal{N}(z_{t-\mu+1}; C_{t-\mu+1} x_{t-\nu+1}, R_{t-\mu+1})$ has rank equal to $n$ (it is the observability matrix) and thus $C_{t-\mu+1}^T C_{t-\mu+1}$ is invertible. Hence, from the relationship $z_{t-\mu+1} = C_{t-\mu+1} x_{t-\mu+1} + v_{t-\mu+1}$ with $v_{t-\mu+1} \sim \mathcal{N}(0, \mathcal{R}_{t-\mu+1})$ we can derive that

$$x_{t-\mu+1} = (C_{t-\mu+1}^T C_{t-\mu+1})^{-1} C_{t-\mu+1}^T (z_{t-\mu+1} - v_{t-\mu+1}). \tag{56}$$

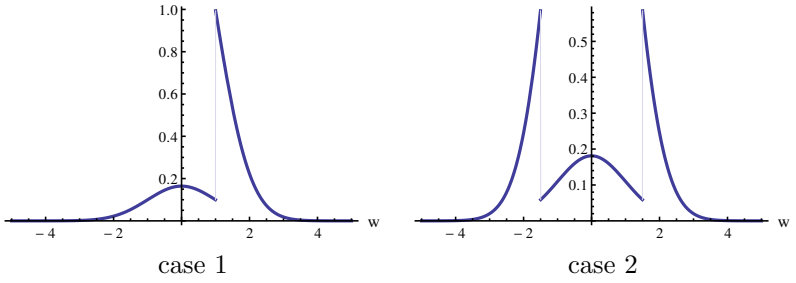
Since $z_{t-\mu+1}$ is known (vector of observations), we can use this relationship to sample values of $x_{t-\mu+1}$ by generating samples of the Gaussian noise $v_{t-\mu+1}$. Then we discretise $f(x_{t-\nu+1}, \mu)$ starting from the sample values obtained from (56). Thus, we repeat the same procedure for the remaining observable parts of the joint from time 0 to time $t-\nu+1$ (the last sampling step will be based on the prior $\mathcal{N}(x_0; \hat{x}_0, P_0)$). Once we have discretised the integrals using the generated samples, the last step consists only to find the unique value of $\mu$ which solves $E_{X^t, Y^t}[(g-\mu)\delta_{\{\tilde{y}^t\}}] = 0$. This value can be easily found by using the bisection method.

In case $\epsilon_0, \epsilon_m, \epsilon_s > 0$ (imprecision), the same sampling strategy described above can be applied to the nominal joint density (that is still Gaussian). Consider for instance the case $t = 3$ and $\nu = 2$, then from the results in Section 7 one has that (12) is equal to:

$$\begin{aligned} 0 = & \inf_{1-\epsilon_0 \le q(x_0) \le 1} \int dx_0 q(x_0) \mathcal{N}(x_0; \hat{x}_0, P_0) \inf_{1-\epsilon_s \le q(x_1|x_0) \le 1} \frac{1}{\int dx_1 q(x_1|x_0) \mathcal{N}(x_1; Ax_0, Q)} \\ & \cdot \int dx_1 q(x_1|x_0) \mathcal{N}(x_1; Ax_0, Q) \inf_{1-\epsilon_m \le q(\tilde{y}_1|x_1) \le (1-\epsilon_m)^{-1}} q(\tilde{y}_1|x_1) \mathcal{N}(\tilde{y}_1; Cx_1, R) \\ & \inf_{1-\epsilon_s \le q(x_2|x_1) \le 1} \frac{1}{\int dx_2 q(x_2|x_1) \mathcal{N}(x_2; Ax_1, Q)} \cdot \int dx_2 q(x_2|x_1) \mathcal{N}(x_2; Ax_1, Q) \\ & \inf_{1-\epsilon_m \le q(\tilde{y}_2|z_2) \le (1-\epsilon_m)^{-1}} q(\tilde{y}_2|x_2) \mathcal{N}(\tilde{y}_2; Cx_2, R) \inf_{1-\epsilon_s \le q(x_3|x_2) \le 1} \frac{1}{\int dx_3 q(x_3|x_2) \mathcal{N}(x_3; Ax_2, Q)} \\ & \cdot \int dx_3 q(x_3|x_2) \mathcal{N}(x_3; Ax_2, Q) \inf_{1-\epsilon_m \le q(\tilde{y}_3|x_3) \le (1-\epsilon_m)^{-1}} q(\tilde{y}_3|x_3)(g(x_3) - \mu) \mathcal{N}(\tilde{y}_3; Cx_3, R). \end{aligned} \tag{57}$$

Since $\mathcal{N}(\tilde{y}_i; Cx_i, R) > 0$ and it does not depend on the $q$, we can take it out from the inf and, thus, apply the transformation described previously to obtain:

$$\begin{aligned} & \inf_{1-\epsilon_0 \le q(x_0) \le 1} \int dx_0 q(x_0) \mathcal{N}(x_0; \hat{x}_0, P_0) \inf_{1-\epsilon_s \le q(x_1|x_0) \le 1} \frac{1}{\int dx_1 q(x_1|x_0) \mathcal{N}(x_1; Ax_0, Q)} \\ & \cdot \int dx_1 q(x_1|x_0) \mathcal{N}(x_1; Ax_0, Q) \inf_{1-\epsilon_m \le q(\tilde{y}_1|x_1) \le (1-\epsilon_m)^{-1}} q(\tilde{y}_1|x_1) \mathcal{N}(\tilde{y}_1; Cx_1, R) \\ & \inf_{1-\epsilon_s \le q(x_2|x_1) \le 1} \frac{1}{\int dx_2 q(x_2|x_1) \mathcal{N}(x_2; Ax_1, Q)} \cdot \int dx_2 q(x_2|x_1) \mathcal{N}(x_2; Ax_1, Q) \\ & \mathcal{N}(\tilde{y}_2; Cx_2, R) \mathcal{N}(\tilde{y}_3; CAx_2, V_3) \inf_{1-\epsilon_m \le q(\tilde{y}_2|z_2) \le (1-\epsilon_m)^{-1}} q(\tilde{y}_2|x_2) \\ & \inf_{1-\epsilon_s \le q(x_3|x_2) \le 1} \frac{1}{\int dx_3 q(x_3|x_2) \mathcal{N}(x_3; Ax_2, Q)} \cdot \int dx_3 q(x_3|x_2) \\ & \mathcal{N}(x_3; W_3 Q^{-1} Ax_2 + W_3 C' R^{-1} \tilde{y}_3, W_3) \inf_{1-\epsilon_m \le q(\tilde{y}_3|x_3) \le (1-\epsilon_m)^{-1}} q(\tilde{y}_3|x_3)(g(x_3) - \mu), \end{aligned} \tag{58}$$




case 1 case 2

Figure 7: Probability density functions for the first component of $w_k$. The first has a discontinuity in 1, the second in $\pm 1.5$. The densities for the second component of $w_k$ are the same.

where $W_3 = R + CQC^T$ and $V_3^{-1} = Q^{-1} + C^T R^{-1} C$. Hence, to discretise (58) we sample $x_0^{(i)} \sim \mathcal{N}(x_0; \hat{x}_0, P_0)$ and we use these samples to generate samples of $x_1$, i.e., $x_1^{(j_i)} \sim \mathcal{N}(x_1; Ax_0^i, Q)$. At time $t = 2$, we stop this nested sampling procedure and we generate samples of $x_2$ directly from $x_2^{(\kappa)} \sim \mathcal{N}(\tilde{y}_2; Cx_2, R)\mathcal{N}(\tilde{y}_3; CAx_2, V_3)$ exploiting the relationship (56). Then we generate samples $x^{(j_\kappa)} \sim \mathcal{N}(x_3; W_3 Q^{-1} A x_2^{(\kappa)} + W_3 C' R^{-1} \tilde{y}_3, W_3)$. Finally, we can apply a procedure similar to the one described in the steps 3–7 in Section 8 to compute the quantity of interest $\hat{\mu} = \underline{E}(g|\hat{y}^n)$. As described in Section 8, the imprecision can be taken into account by increasing the covariances $\beta_1 Q$, $\beta_2 R$, $\beta_3 P_0$ with $\beta_i > 1$ and, thus, by spreading the samples. We can use the knowledge of the credible ellipsoid in (40) to determine the scaling factor for the covariances of each COR model for initial state, dynamics and measurement equation.

Therefore, we exploit the observability index to do a sort of resampling after each $\nu$ time steps and to break down the increasing of the numbers of samples that we should have by applying a nested MC sampling from time 0 to time $t$. In this way, the computational complexity increases only linearly in time.

*10.1. Numerical example*

For the linear Gaussian case case discussed in the previous section, we have performed numerical (Monte Carlo) simulations in order to show the performance of the COR filter and to compare this performance with other known approaches to state estimation.

The true trajectory of the state and measurements are generated by the following dynamical system:

$$\begin{cases} x_{k+1} &= Ax_k + w_k \\ y_k &= Cx_k + v_k \end{cases} \qquad A = \begin{bmatrix} 1 & 1 \\ 0 & 1 \end{bmatrix}, \quad C = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}, \tag{59}$$

where $w_k \sim p(w)$, $x_0 = \hat{x}_0$ with $\hat{x}_0 \sim N(0, P_0)$, $v_k \sim N(0, R)$,

$$P_0 = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}, \quad Q = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}, \quad R = Q,$$

where two kinds of densities $p_T(w_k)$ will be considered for $p(w_k)$ as shown in Figure 7. The first density (case 1) is asymmetric w.r.t. the origin with positive mean 0.9 and variance equal to 1.1. This means that there is a nondeterministic bias in the relationship $x_{k+1} = Ax_k + w_k$. The second density (case 2) has zero mean and variance 2.6. Thus, the variance is greater than the one of the nominal Gaussian density (i.e., $Q$). We call (59) in the case 1 or 2 the true system.

We assume that the modeller does not know the true system. In particular, we consider the case in which the modeller does not know $p_T(w_k)$ but he can specify a bound for it in the form of a COR set of densities:[13]

$$\mathcal{P}_W = \{(1 - \epsilon_s)\mathcal{N}(w_k; 0, Q) \leq p(w_k) \leq \mathcal{N}(w_k; 0, Q)\},$$

[13]Notice that the COR model does not require $p_T(w_k)$ to be stationary (time invariant).

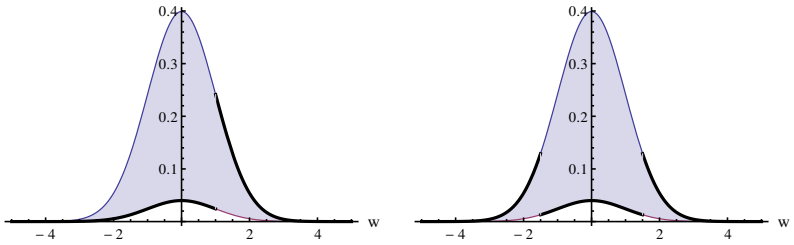



Figure 8: True unnormalised densities for the first component of $w_k$ (in bold) belonging to the COR set of density $\mathcal{P}_W$ with $\epsilon_s = 0.9$ for the two cases in Figure 7.

with $\epsilon_s = 0.9$, see Figure 8. From a modelling point of view, we can see $\mathcal{P}_W$ as a robust model for $w_k$ which is based on the following considerations. First, by specifying $\mathcal{P}_W$, the modeller is stating that he knows that the high density region for $p(w_k)$ is the ellipsoid $w^T Q^{-1} w \leq 11.83$,[14] and that $p(w_k)$ is strictly positive in this region (the unnormalised density is lower bounded by $(1-\epsilon_s)\mathcal{N}(w_k; 0, Q)$). Second, the modeller does not exclude the possibility that $p(w_k)$ is Gaussian, but also allows for a wider variety of density shapes: unimodal and multimodal not necessarily centred at zero. By allowing nonzero mean densities, the modeller is thus considering also cases in which the noise $w_k$ is biased with an unknown bias.

Since $x_{k+1} = Ax_k + w_k$ by a change of variables, it follows that

$$\mathcal{P}_{X_{k+1}|X_k} = \{(1-\epsilon_s)\mathcal{N}(x_{k+1}; Ax_k, Q) \leq p(x_{k+1}|x_k) \leq \mathcal{N}(x_{k+1}; Ax_k, Q)\},$$

which gives the COR set bounds for $p(x_{k+1}|x_k)$ for any $k = 1, \ldots, t$.

The modeller does not know the true density $p_T(w_k)$ but he knows that it belongs to the set $\mathcal{P}_W$. He can thus use the procedure described in Section 10 to compute lower $\underline{E}(g|\tilde{y}^t)$ and upper $\overline{E}(g|\tilde{y}^t)$ bounds for the posterior expectation $E(g|\tilde{y}^t)$ of any function of interest $g$ of $X_k$ at each instant $k = 1, \ldots, t$. In particular, he can compute the posterior set $\mathcal{X}^*$ (the set of estimates that are not dominated under the squared loss) using the procedure described in Section 9 and the minimum volume ellipsoid that has lower probability 0.95 of including the true value of $X_k$ (robust credible ellipsoid).

Observe that the results of Theorem 6, which provides the analytical expression for $\mathcal{X}^*$ and the credible ellipsoid, hold only for a COR model. Since prediction does not preserve the structure of the COR model, we cannot use this result for the posterior $\underline{E}(g|\tilde{y}^t)$. For this reason, we compute the posterior set $\mathcal{X}^*$ and the credible ellipsoid numerically. For the former, we employ the procedure described in Section 9 with $\theta = \{0, \pi/4, \pi/2, \ldots, 7/4\pi\}$. For the latter, we fix the centre and the directions of the ellipsoid as in the Kalman filter based on the nominal Gaussian system and then we determine numerically the minimum value of the scaling factor $\rho(\alpha)$, as in (40), which ensures that the ellipsoid includes a probability of at least 0.95.

The performance of the COR based filter are compared with respect to the following approaches:

1. the optimal posterior mean and 95% credible ellipsoid that can be obtained by applying particle filter estimation (800 particles) to the true unknown system;

2. the posterior mean and 95% credible ellipsoid obtained with a KF based on the nominal density $\mathcal{N}(w_k; 0, Q)$.

The first approach, that gives the optimal MMSE (minimum mean squared error) estimate $E^*(X_t|\tilde{y}^t)$, is reported as term of comparison but it is not attainable in practice since the modeller does not know $p_T(w_k)$. The second approach is in general suboptimal. KF gives the optimal MMSE estimate only in the case $\epsilon_s = 0$ while, for $\epsilon_s > 0$, it provides the best linear MMSE estimator in the case $E[W_k] = 0$ and $E[W_k W_k^T] = Q$ and a wrong (biased or not calibrated) estimate in all remaining cases, i.e., $E[W_k] \neq 0$ and/or $E[W_k W_k^T] \neq Q$.

[14] This is the 0.9973 probability region for the Gaussian $\mathcal{N}(w_k; 0, Q)$. The lower probability of this region based on $\mathcal{P}_W$ with $\epsilon_s = 0.9$ is 0.973.

| | MSE |
|---|---|
| Opt. | 1.0488 |
| KF | 1.1233 |
| | % |
| COR(Opt.) | 0.98 |
| COR(KF) | 0.99 |

Table 2: MSE and variance of the MSE

For both cases 1 and 2, we have evaluated the performance of the COR filter by Monte Carlo simulations (a trajectory of 12 time steps and a Monte Carlo size of 100 runs). The COR model has been implemented using 50 particles at each time instant.

*10.1.1. Case 1*

Figure 9 shows the outer-approximation of the set $\mathcal{X}^*$ computed for all 12 instants of the trajectory in a single Monte Carlo run. It can be noticed that in all the time steps $\mathcal{X}^*$ includes always the KF estimate and the optimal Bayesian estimate. Since $p_T(w_k)$ is included in the COR set assumed by the modeller, from the theoretical derivations of the previous sections it follows that the optimal MMSE estimate $E^*(X_t|\tilde{y}^t)$ should be always belong to $\mathcal{X}^*$ and, furthermore, $\mathcal{X}^*$ is the minimum volume region that always includes $E^*(X_t|\tilde{y}^t)$. The fact that also the KF is contained in $\mathcal{X}^*$ follows from the consideration that also the nominal Gaussian density is included in the COR set.

It should also be noticed that the true value of the state is not always in $\mathcal{X}^*$. This is correct, since $\mathcal{X}^*$ is not a credible region. It represents the set that include all undominated estimators under the squared loss. In the case $\epsilon_s = 0$ (no imprecision), this set reduce to a single point, the KF posterior estimate. Thus, by providing $\mathcal{X}^*$, the modeller reports the set of the optimal estimators under the squared loss.

In Table 2 we show the MSE for the KF and the optimal Bayesian estimator averaged over the whole trajectory and all 100 Monte Carlo runs. Then we have reported the average number of times the KF and the optimal Bayesian estimates are included in $\mathcal{X}^*$. It can be noticed the inclusion percentage almost coincides with the theoretical value 100%; the difference is due to numerical problems. In fact, by definition, $\mathcal{X}^*$, we compute with our algorithm, includes all the Bayesian estimates that can be obtained by applying Bayesian filtering to one of the densities in the the COR set for initial state, state transition and measurement equation. Observe that, both the Gaussian density and the true densities of the noises in Figure 7 belong to this set and, thus, the KF and optimal Bayesian estimator must be contained in $\mathcal{X}^*$. Conversely, the true trajectory does not have to be contained in this interval. $\mathcal{X}^*$ is not a credible region but it is the interval that includes all the optimal (in the squared error sense) Bayesian estimates. Notice also that, when we only know that the distributions of the noises belong to COR sets we cannot compute the optimal Bayesian estimate. However, we can use our algorithm to compute the region that includes the optimal Bayes estimate. This region is exactly $\mathcal{X}^*$.

*10.2. Case 2*

In this case we have computed the coverage probability of the 95% credible ellipsoid. Table 3 shows this result. It can be noticed that the KF ellipsoid is not calibrated. It includes the true state with a probability of only 0.9, which is less than the expected 0.95. This means that the credible ellipsoid of KF is too small. The KF is in fact using a covariance matrix $Q$ that underestimates the variance of the noise. Conversely, the COR ellipsoid includes the true value of the state with probability of 0.965. This value is (slightly) more than 0.95, which means that the credible ellipsoid of the COR filter is larger than the optimal one. This can be due to the fact that the density $p_T(w_k)$, for case 2, is probably not the most critical one in the COR set and also to the approximation used to compute the ellipsoid for the COR model. In any case, this shows that the inferences based on the COR model are very robust and that it outperforms KF in non-Gaussian and unknown distributions settings.

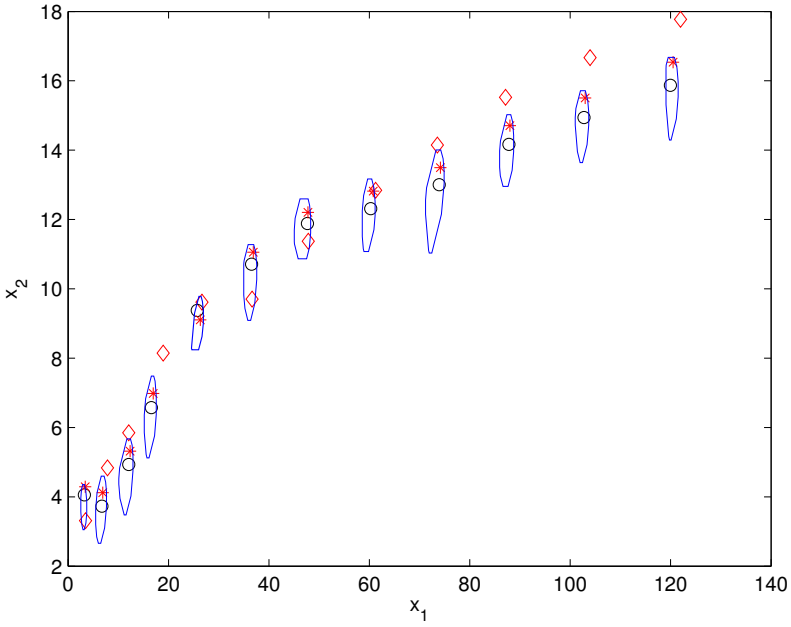



Figure 9: The figure reports the true value of the state (red diamond), the optimal Bayesian estimate obtained with a particle filter that knows the true density of the noise $w_k$ (red star), the KF estimate based on the nominal density (black circle) and the set of the posterior means $\mathcal{X}^*$ (blue).

| | Probability |
|---|---|
| Opt. | 0.95 |
| KF | 0.902 |
| COR | 0.965 |

Table 3: True coverage probability of the theoretical 95% credible ellipsoid.

## 11. Conclusions

In this paper, we have proposed an extension of the classical filtering problem that allows to model imprecision in our knowledge about initial state, system dynamics and measurement equation modelled by means of a closed convex set of probabilities known with the name of density ratio class. The density ratio class model has three main characteristics that make it suitable for robust filtering. First, it is easy to elicit, since only a scalar parameter and a nominal density function must be specified. Second, it is robust, since it allows for a wide variety of density shapes (unimodal and multimodal), but it is not too imprecise (the tail behaviour is fully determined by the nominal density function). Third, the posterior inferences derived by the density ratio class model are computationally tractable.

By exploiting these characteristics, we have derived the solution to the state estimation problem in the case the uncertainty on initial state, measurement equation and state dynamics are modelled through a density ratio class set of densities. We have further shown that the obtained solution is optimal (w.r.t. the squared-loss function) and, thus, that the closed convex set of posterior estimates, that we compute with our algorithm, includes all the Bayesian optimal estimates that we should obtain by first selecting any density in the density ratio class set and then applying Bayesian filtering to compute the posterior estimate.

We have also specialised the density ratio class to the case in which the nominal density is a multivariate Gaussian. For this case, we have derived an efficient algorithm to solve the filtering problem when initial state, system dynamics and measurement equation are modelled by means of density ratio class models. This efficiency is due, in part, to the fact that our algorithm does not need to compute optimisations: the solution method relies on Monte Carlo sampling alone, and hence its complexity is comparable to that of precise-probability approaches.

Finally, we have also shown, in a practical case, that our extension outperforms the Kalman filter when modelling errors are present in the system.

With respect to future prospects, we can devise several lines of investigation. The first might be concerned with deepening the comparison with the classical results. The second might focus on the extension to nonlinear systems. The third might be to include additional information (when available) on the distributions of the noises to reduce the imprecision of the inferences derived by the density ratio class model. For instance, together with bounds for the densities, we might know (i) the moments (e.g., mean and variance) of the noises; (ii) that the densities are unimodal and/or symmetric. The problem is how to efficiently include this information in the density ratio class models.


## Acknowledgements

This work has been partially supported by the Swiss NSF grantsn. 200020-137680/1, 200020-134759/1 and Hasler Foundation grant n. 10030.


## References


[1] A. Jazwinski, *Stochastic processes and filtering theory.* Academic Press; 1st Ed., 1970.

[2] G. Kitagawa, "Monte Carlo filter and smoother for non-Gaussian nonlinear state space models," *Journal of computational and graphical statistics*, vol. 5, no. 1, pp. 1–25, 1996.

[3] M. Arulampalam, S. Maskell, N. Gordon, and T. Clapp, "A tutorial on particle filters for online nonlinear/non-gaussian bayesian tracking," *IEEE Transactions on Signal Processing*, vol. 50, no. 2, pp. 174 –188, 2002.

[4] J. Spall, "Estimation via Markov chain Monte Carlo," *IEEE Control Systems Magazine*, vol. 23, no. 2, pp. 34 – 45, 2003.

[5] M. Fu, C. de Souza, and L. Xie, "$H_\infty$ estimation for uncertain systems," *International Journal of Robust and Nonlinear Control*, vol. 2, no. 2, pp. 87–105, 1992.

[6] P. Bolzern, P. Colaneri, and G. De Nicolao, "Optimal robust filtering for linear systems subject to time-varying parameter perturbations," in *Decision and Control, 1993., Proceedings of the 32nd IEEE Conference on*, pp. 1018 –1023 vol.2, dec 1993.

[7] I. Petersen and D. McFarlane, "Optimal guaranteed cost control and filtering for uncertain linear systems," *Automatic Control, IEEE Transactions on*, vol. 39, pp. 1971 –1977, sep 1994.

[8] F. C. Schweppe, "Recursive state estimation: Unknown but bounded errors and system inputs," in *Adaptive Processes, Sixth Symposium on*, vol. 6, pp. 102 –107, oct. 1967.

[9] D. Bertsekas and I. Rhodes, "Recursive state estimation for a set-membership description of uncertainty," *Automatic Control, IEEE Transactions on*, vol. 16, pp. 117 – 128, apr 1971.
[10] W. Wang and M. Orshansky, "Robust estimation of parametric yield under limited descriptions of uncertainty," in *Proceedings of the 2006 IEEE/ACM international conference on Computer-aided design*, pp. 884–890, ACM New York, NY, USA, 2006.
[11] O. Strauss and S. Destercke, "F-boxes for filtering," in *Proc. of the 7th conference of the European Society for Fuzzy Logic and Technology (EUSFLAT-2011) and LFA-2011.*, (Aix-Les-Bains, France), pp. 935–942, 2011.
[12] D. Morrell and W. Stirling, "Set-valued filtering and smoothing," in *Twenty-Second Asilomar Conference on Signals, Systems and Computers*, vol. 1, 1988.
[13] J. Kenney and W. Stirling, "Nonlinear filtering of convex sets of probability distributions," *Journal of Statistical Planning and Inference*, vol. 105, no. 1, pp. 123–147, 2002.
[14] B. Noack, V. Klumpp, D. Brunn, and U. Hanebeck, "Nonlinear Bayesian estimation with convex sets of probability densities," in *11th International Conference on Information Fusion*, pp. 1–8, 2008.
[15] J. Spall, "The Kantorovich inequality for error analysis of the Kalman filter with unknown noise distributions," *Automatica J. IFAC*, vol. 10, pp. 1513–1517, 1995.
[16] J. Maryak, J. Spall, and B. Heydon, "Use of the Kalman filter for inference in state-space models with unknown noise distributions," *IEEE Transactions on Automatic Control*, vol. 49, no. 1, pp. 87 – 90, 2004.
[17] A. Benavoli, M. Zaffalon, and E. Miranda, "Robust filtering through coherent lower previsions," *Automatic Control, IEEE Transactions on*, vol. 56, pp. 1567 –1581, July 2011.
[18] P. Walley, *Statistical Reasoning with Imprecise Probabilities*. New York: Chapman and Hall, 1991.
[19] E. Miranda, "A survey of the theory of coherent lower previsions," *International Journal of Approximate Reasoning*, vol. 48, no. 2, pp. 628–658, 2008.
[20] J. O. Berger, *Statistical Decision Theory and Bayesian Analysis*. New York: Springer Series in Statistics, 1985.
[21] J. Berger, "Robust bayesian analysis: sensitivity to the prior," *Journal of Statistical Planning and Inference*, vol. 25, no. 3, pp. 303–328, 1990.
[22] L. Wasserman, "Invariance properties of density ratio priors," *The Annals of Statistics*, vol. 20, no. 4, pp. 2177–2182, 1992.
[23] L. DeRoberts and J. Hartigan, "Bayesian inference using intervals of measures," *The Annals of Statistics*, pp. 235–244, 1981.
[24] A. Demidov, *Generalized functions in mathematical physics: main ideas and concepts*, vol. 237. Nova Science Publishers, 2001.
[25] G. De Cooman, F. Hermans, A. Antonucci, and M. Zaffalon, "Epistemic irrelevance in credal nets: the case of imprecise markov trees," *International Journal of Approximate Reasoning*, vol. 51, no. 9, pp. 1029–1052, 2010.
[26] R. Schneider, *Convex bodies: the Brunn-Minkowski theory*, vol. 44. Cambridge Univ Pr, 1993.
[27] S. L. Rinderknecht, M. E. Borsuk, and P. Reichert, "Eliciting density ratio classes," *International Journal of Approximate Reasoning*, vol. 52, no. 6, pp. 792 – 804, 2011.
[28] S. L. Rinderknecht, *Contributions to the use of Imprecise Scientific Knowledge in Decision Support*. ETH Zurich: Ph.D. thesis, 2011.
[29] C. Robert and G. Casella, *Monte Carlo statistical methods*. Springer Verlag, 2004.
[30] A. Doucet, N. De Freitas, and N. Gordon, *Sequential Monte Carlo methods in practice*. Springer Verlag, 2001.