Reliable hidden Markov model filtering through coherent lower previsions

Alessio Benavoli
IDSIA
Lugano, Switzerland.
alessio@idsia.ch

Marco Zaffalon
IDSIA
Lugano, Switzerland
zaffalon@idsia.ch

Enrique Miranda
Universidad de Oviedo
Oviedo, Spain
mirandaenrique@uniovi.es

Abstract – *We extend Hidden Markov Models for continuous variables taking into account imprecision in our knowledge about the probabilistic relationships involved. To achieve that, we consider sets of probabilities, also called coherent lower previsions. In addition to the general formulation, we study in detail a particular case of interest: linear-vacuous mixtures. We also show, in a practical case, that our extension outperforms the Kalman filter when modelling errors are present in the system.*

Keywords: continuous Hidden Markov Models, coherent lower previsions, epistemic irrelevance, marginal extension, Kalman filter.

1 Introduction

State estimation problems arise in many fields such as signal processing, communications, tracking and automatic control. In all these problems there is a model for which we observe some function of a parameter set of interest, and where the observations may be corrupted by noise. If the random process to be estimated is modelled by a state variable approach, the estimation problem is referred to as *Bayesian state estimation*. In this case, often a Markov process representation is used to model the random process and, thus, the estimation problem reduces to the estimate of the state of a continuous Hidden Markov Model (HMM).

The major criticism of Bayesian estimation is its "sensitivity" to the choice of the model assumptions, in the sense that the estimate can depend strongly on the choice of the prior, the likelihood and the loss/utility functions. One possible solution for this problem is the so-called *Bayesian sensitivity analysis* or Bayesian robustness approach [1]. Its basic idea is to check the robustness of the estimate by defining a wide class of prior distributions and likelihood functions, to combine each pair by Bayes's rule to form a class of posterior distributions and to check whether these posteriors lead to the same conclusions. Depending on the answer we declare that the model is robust or that the conclusions from any particular Bayesian model are unreliable. Another possibility to deal with the imprecise knowledge of the elements of the model is to consider alternative models of uncertainty, such as Choquet capacities, belief functions and possibility measures. All these models represent uncertainty through lower and upper probabilities, and can be all regarded as special cases of Walley's *coherent lower previsions* [7]. Walley's theory of *coherent lower previsions*, which is usually referred to as *Imprecise Probability* (IP), provides a very general model of uncertain knowledge. In addition, IP models have a clear interpretation in terms of a subject's behaviour. The theory of coherent lower previsions includes as a particular case the approach considered in Bayesian sensitivity analysis: if we specify a number of precise models, they determine coherent lower previsions by taking their lower envelopes. It is also more general [7]. For instance, a difference comes when modeling the notion of independence: with a Bayesian sensitivity analysis we must require that all the admissible models carry the notion of independence; with coherent lower previsions there are a number of less restrictive possibilities. Among these, we shall consider the notion of *epistemic irrelevance*. In this paper, we study the problem of estimating the state of a Hidden Markov Model when we do not have enough information to describe the prior, the state transition and the likelihood models with precise probabilities. Instead, we shall model our uncertainty about the variables of interest by means of coherent lower previsions. To this end, we derive a solution of the state estimation problem for HMM for the general case of coherent lower previsions in infinite spaces (see [2] for inference for HMM in finite spaces). This rule is then specialised for a special class of coherent lower previsions, called *linear-vacuous mixtures*. For this particular class, we empirically compare the proposed estimator with the Kalman filter and show that our solution is more robust to modelling errors and that, hence, it outperforms the Kalman filter in such a case. We also discuss the self-consistency of the proposed solution with the initial assessments.

2 Bayesian estimation for HMM

The aim 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

$$\left\{ \begin{array}{rcl} \mathbf{x}(t+1) & = & \mathbf{f}(t,\mathbf{x}(t)) + \mathbf{w}(t) \\ \mathbf{y}(t) & = & \mathbf{h}(t,\mathbf{x}(t)) + \mathbf{v}(t), \end{array} \right. \tag{1}$$

where t is the time, $\mathbf{x}(t) \in \mathbb{R}^n$ is the state vector, $\mathbf{w}(t) \in \mathbb{R}^n$ is the process-noise, $\mathbf{y}(t) \in \mathbb{R}^m$ is the measurement vector, $\mathbf{v}(t) \in \mathbb{R}^m$ is the measurement noise and $f(\cdot)$ and $h(\cdot)$ are known nonlinear functions. Having observed a finite sequence $\tilde{\mathbf{y}}^t = \{\tilde{\mathbf{y}}(1), \ldots, \tilde{\mathbf{y}}(t)\}$[1] of measurements, one may, in general, seek an estimate of an entire sequence of states $\mathbf{x}^t = \{\mathbf{x}(0), \ldots, \mathbf{x}(t)\}$.

In the Bayesian framework, all relevant information on $\mathbf{x}^t = \{\mathbf{x}(0), \ldots, \mathbf{x}(t)\}$ at time t is included in the posterior distribution $p(\mathbf{x}^t|\tilde{\mathbf{y}}^t)$. A Markov assumption is made to model the system and, thus, the estimation problem reduces to the estimate of the state of a HMM. A consequence of this assumption is that the following independence conditions hold:

$$\begin{array}{rcl} p(\mathbf{x}(t)|\mathbf{x}^{t-1}) & = & p(\mathbf{x}(t)|\mathbf{x}(t-1)), \\ p(\mathbf{y}^t|\mathbf{x}^t) & = & \prod_{k=1}^{t} p(\mathbf{y}(k)|\mathbf{x}(k)). \end{array}$$

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

$$p(\mathbf{x}^t|\tilde{\mathbf{y}}^t) = \frac{p(\mathbf{x}^{t-1}|\tilde{\mathbf{y}}^{t-1})p(\mathbf{x}(t)|\mathbf{x}(t-1))p(\tilde{\mathbf{y}}(t)|\mathbf{x}(t))}{p(\tilde{\mathbf{y}}(t)|\tilde{\mathbf{y}}^{t-1})}. \tag{2}$$

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

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

From (2) and (3), we see that both $p(\mathbf{x}^t|\tilde{\mathbf{y}}^t)$ and $p(\mathbf{x}(t)|\tilde{\mathbf{y}}^t)$ can be obtained recursively. Once $p(\mathbf{x}(t)|\tilde{\mathbf{y}}^t)$ has been computed, it is possible to compute the expected value $E[g(\mathbf{x}(t))|\mathbf{y}^t]$ w.r.t. $p(\mathbf{x}(t)|\mathbf{y}^t)$ for any function of interest $g(\mathbf{x}(t))$. A particular case of (1) is

$$\left\{ \begin{array}{rcl} \mathbf{x}(t+1) & = & \mathbf{A}(t)\mathbf{x}(t) + \mathbf{w}(t) \\ \mathbf{y}(t) & = & \mathbf{C}(t)\mathbf{x}(t) + \mathbf{v}(t), \end{array} \right. \tag{4}$$

with $\mathbf{w}(t) \sim \mathcal{N}(\mathbf{0}, \mathbf{Q}(t))$, $\mathbf{v}(t) \sim \mathcal{N}(\mathbf{0}, \mathbf{R}(t))$, $\mathbf{x}(0) \sim \mathcal{N}(\hat{\mathbf{x}}(0), \mathbf{P}(0))$, and where the matrices $\mathbf{A}(t)$, $\mathbf{C}(t)$, $\mathbf{Q}(t)$, $\mathbf{R}(t)$ are assumed to be known. Then the conditional PDF $p(\mathbf{x}(t)|\tilde{\mathbf{y}}^t)$ is also Gaussian $\mathcal{N}(\hat{\mathbf{x}}(t), \mathbf{P}(t))$ where

$$\left\{ \begin{array}{l} \hat{\mathbf{x}}(t) = \mathbf{A}(t)\hat{\mathbf{x}}(t-1) + \mathbf{L}(t)[\mathbf{y}(t) - \mathbf{C}(t)\mathbf{A}(t)\hat{\mathbf{x}}(t-1)] \\ \mathbf{P}(t) = \mathbf{A}(t)\mathbf{P}(t)\mathbf{A}'(t) + \mathbf{Q}(t) - \mathbf{L}(t)\mathbf{S}(t)\mathbf{L}'(t) \\ \mathbf{S}(t) = \mathbf{C}(t)[\mathbf{A}(t)\mathbf{P}(t)\mathbf{A}'(t) + \mathbf{Q}(t)]\mathbf{C}'(t) + \mathbf{R}(t) \\ \mathbf{L}(t) = [\mathbf{A}(t)\mathbf{P}(t)\mathbf{A}'(t) + \mathbf{Q}(t)]\mathbf{C}'(t)\mathbf{S}^{-1}(t). \end{array} \right.$$

These are the equations of Kalman filter (KF).

3 Coherent lower previsions

Let us briefly introduce the formalism of coherent lower previsions we shall use later in the paper. We refer to [7] for a detailed account of the theory.

Given a possibility space Ω, a *gamble* is a bounded real-valued function on Ω. The set of all gambles on Ω is usually denoted by $\mathcal{L}(\Omega)$. A *Coherent Lower Prevision* (CLP) on a linear subset K of $\mathcal{L}(\Omega)$ is a functional satisfying the following three conditions:

(C1) $\underline{E}(f) \geq \inf_{\omega} f(\omega)\ \forall f \in K$.

(C2) $\underline{E}(\lambda f) = \lambda \underline{E}(f)\ \forall \lambda \geq 0, f \in K$.

(C3) $\underline{E}(f+g) \geq \underline{E}(f) + \underline{E}(g)$ for all $f, g \in K$.

From a lower prevision $\underline{E}$ on K we can always derive a so-called *upper prevision* $\overline{E}$ on $-K$, by means of the equation $\overline{E}(\cdot) = -\underline{E}(-\cdot)$. Because of this relationship, we shall focus on lower previsions only. When the domain K is a linear space of gambles and (C3) holds with equality for every pair $f, g \in K$, the coherent lower prevision $\underline{E}$ is called a *linear prevision*, and is usually denoted by E. A linear prevision is the expectation with respect to its restriction to events, which is a finitely additive probability. Hence, for discrete variables:

$$E(f) = \sum_x f(x)p(x).$$

We can equivalently define CLP in terms of linear previsions. Let $\mathbb{P}(\Omega)$ be the set of linear previsions on $\mathcal{L}(\Omega)$; then $\underline{E}$ is a CLP if and only if

$$\underline{E}(f) = \min\{E(f) : E \in \mathbb{P}(\Omega), E(g) \geq \underline{E}(g)\ \forall g \in K\}$$

for every f in the domain K of $\underline{E}$. Hence, CLPs can be regarded as modelling the imprecise knowledge about a linear prevision: we can simply consider a set $\mathcal{M}$ of possible candidate linear previsions and summarise it with its lower envelope, which is a CLP.

Consider now variables $Z_1, \ldots, Z_m$ taking values in respective spaces $\mathcal{Z}_1, \ldots, \mathcal{Z}_m$. For every $J \subseteq \{1, \ldots, m\}$, let $Z_J = (Z_j)_{j \in J}$ and $\mathcal{Z}_J = \times_{j \in J} \mathcal{Z}_j$. In particular, we denote by $\mathcal{Z}^m := \mathcal{Z}_{\{1,\ldots,m\}}$, our possibility space in the remainder of the paper. Given disjoint subsets U, O, with $O \neq \emptyset$, of $\{1, \ldots, m\}$, we denote $\underline{E}_{Z_O}(\cdot|Z_U)$ the *conditional lower prevision* that, for every gamble f on $\mathcal{Z}_{O \cup U}$ and every $z \in \mathcal{Z}_U$, yields the

[1]The tilde superscript is used in order to distinguish a generic realization $\mathbf{y}$ of the measurement vector and the observed one $\tilde{\mathbf{y}}$.

value $\underline{E}_{Z_O}(f|z)$, which is the lower prevision for the gamble f, if we knew that the variable Z_U took the value z. We can equivalently define $\underline{E}_{Z_O}(\cdot|Z_U)$ on the set of gambles f on $\mathcal{Z}^m$ which only depend on the values in $\mathcal{Z}_{O\cup U}$, in the sense that $f(z_1) = f(z_2)$ when the *projections* $\pi_{O\cup U}(z_1)$, $\pi_{O\cup U}(z_2)$ coincide. These gambles are called $\mathcal{Z}_{O\cup U}$-measurable. A conditional lower prevision $\underline{E}_{Z_O}(\cdot|Z_U)$ defined on the linear space $K_{O\cup U}$ of $\mathcal{Z}_{O\cup U}$-measurable gambles is called *separately coherent* when the following three conditions hold for every $z \in \mathcal{Z}_U$:

(SC1) $\underline{E}_{Z_O}(f|z) \geq \inf_{\omega\in\pi_U^{-1}(z)} f(\omega)$.

(SC2) $\underline{E}_{Z_O}(\lambda f|z) = \lambda \underline{E}_{Z_O}(f|z)\ \forall f \in K_{O\cup U}, \lambda \geq 0$.

(SC3) $\underline{E}_{Z_O}(f+g|z) \geq \underline{E}_{Z_O}(f|z) + \underline{E}_{Z_O}(g|z)$ $\forall f, g \in K_{O\cup U}$.

When (SC3) holds with equality for every $f, g \in K_{O\cup U}$, $\underline{E}_{Z_O}(\cdot|Z_U)$ is a conditional (*linear*) prevision. These are conditional expectations with respect to finitely additive probabilities. Separate coherence is equivalent to being the lower envelope of the dominating conditional previsions.

Given $\underline{E}_{Z_{O_1}}(\cdot|Z_{U_1}), \ldots, \underline{E}_{Z_{O_k}}(\cdot|Z_{U_k})$ with domains $K_{O_1\cup U_1}, \ldots, K_{O_k\cup U_k}$, they are *jointly coherent* when for every $f_j \in K_{O_j\cup U_j}, j = 1, \ldots, k$, $j_0 \in \{1, \ldots, k\}, z_o \in \mathcal{Z}_{U_{j_0}}$ and $f_{j_0} \in K_{O_{j_0}\cup U_{j_0}}$,

$$\sup \sum_{j=1}^{k} (f_j - \underline{E}_{Z_{O_j}}(f_j|Z_{U_j})) - I_{\pi^{-1}_{U_{j_0}}(z_0)}(f_{j_0} - \underline{E}_{Z_{O_{j_0}}}(f_{j_0}|z_0))$$

is non-negative on some B in $\left\{\pi^{-1}_{U_{j_0}}(z_0)\right\} \cup \bigcup S_j(f_j)$, where $S_j(f_j) = \{\pi^{-1}_{U_j}(z) : f_j I_{\pi^{-1}_{U_j}(z)} \neq 0\}$ and $I_{\pi^{-1}_{U_{j_0}}(z_0)}$ is the indicator function[2] of $\pi^{-1}_{U_{j_0}}(z_0)$. Coherence means that the assessments implied by $\underline{E}_{Z_{O_1}}(\cdot|Z_{U_1}), \ldots, \underline{E}_{Z_{O_k}}(\cdot|Z_{U_k})$ are consistent. It implies (but it is also stronger than it) that there is a coherent lower prevision $\underline{E}$ on $\mathcal{L}(\mathcal{Z}^m)$ which is compatible with all the assessments, in the sense that it is coherent with each of them. A sufficient condition for coherence is that $U_1 = \emptyset$ and $U_j = \cup_{\ell=1}^{j-1}(O_\ell \cup U_\ell)$ for $j = 2, \ldots, k$. In that case, the smallest coherent lower prevision $\underline{E}$ which is coherent with $\underline{E}_{Z_{O_1}}(\cdot|Z_{U_1}), \ldots, \underline{E}_{Z_{O_k}}(\cdot|Z_{U_k})$ is called their *marginal extension*, and is given by

$$\underline{E}(f) = \underline{E}_{Z_{O_1}}(\underline{E}_{Z_{O_2}}(\ldots(\underline{E}_{Z_{O_k}}(f|Z_{U_k}))\ldots|Z_{U_2})).$$

If in particular all the conditional lower previsions are linear, this is the only compatible joint. On the other hand, if we have a coherent lower prevision $\underline{E}$ on $\mathcal{L}(\mathcal{X}^m)$ and a separately coherent conditional lower prevision $\underline{E}_{Z_O}(\cdot|Z_U)$ on $K_{O\cup U}$, a necessary condition for their joint coherence is that for every gamble f in $K_{O\cup U}$ and every $z \in \mathcal{Z}_U$:

$$\underline{E}(I_{\pi_U^{-1}(z)}(f - \underline{E}_{Z_O}(f|z))) = 0. \qquad \text{(GBR)}$$

This is called the *Generalised Bayes Rule*, because it becomes Bayes' rule when we have linear previsions. When $\underline{E}(z) > 0$, there is only one value of $\underline{E}_{Z_O}(f|z)$ which satisfies (GBR) with $\underline{E}$, and therefore the conditional lower prevision is uniquely determined by $\underline{E}$ and the notion of coherence.

4 Generalisation of Bayesian state estimation

We can rephrase Bayesian state estimation in the formalism we have just introduced. The aim of Bayesian state estimation is to compute the conditional linear prevision $E_{\mathbf{X}(t)}[g(\mathbf{x}(t))|\tilde{\mathbf{y}}^t]$, where the notation $\mathbf{X}$ has been introduced to distinguish between random vectors and their generic realisations $\mathbf{x}$, while $\tilde{\mathbf{y}}$ denotes the actual observations. It is assumed that the value $\mathbf{x}(t)$ belongs to $\mathbb{R}^n$ and that $\mathbf{y}(t)$ belongs to $\mathbb{R}^m$ for every t. However, in order to have bounded gambles, which are the basis of Walley's theory, hereafter we assume $\mathbf{x}(t) \in \mathcal{X}_t$ and $\mathbf{y}(t) \in \mathcal{Y}_t$ for each t, where $\mathcal{X}_t$ and $\mathcal{Y}_t$ are closed subsets of $\mathbb{R}^n$ and $\mathbb{R}^m$, respectively. The solution of the estimation problem can be obtained by applying GBR

$$E_{\mathbf{X}(t)}[g|\tilde{\mathbf{y}}^t] = \mu \quad s.t. \quad E_{\mathbf{X}^t,\mathbf{Y}^t}[I_{\tilde{\mathbf{y}}^t} \cdot (g - \mu)] = 0, \quad (5)$$

taking into account that there is a unique μ satisfying this equation when $E_{\mathbf{X}^t,\mathbf{Y}^t}[I_{\tilde{\mathbf{y}}^t}] = E_{\mathbf{Y}^t}[I_{\tilde{\mathbf{y}}^t}] > 0$. However, since the probability that a continuous random variable assumes a particular value is zero, in our context $E_{\mathbf{Y}^t}[I_{\tilde{\mathbf{y}}^t}] = Pr(\mathbf{Y}^t = \tilde{\mathbf{y}}^t) = 0$ and therefore GBR does not have a unique solution. A way to overcome this problem in precise probability is to regard the measurements $\tilde{\mathbf{y}}(t)$ for any $t \geq 1$ as idealisations of discrete events $\tilde{\mathbf{y}}_d(t) = \mathcal{B}(\tilde{\mathbf{y}}(t), \delta)$, where $\mathcal{B}(\tilde{\mathbf{y}}(t), \delta)$ are nested neighbourhoods of $\tilde{\mathbf{y}}(t)$ which decrease to the limit $\tilde{\mathbf{y}}(t)$ as their radius $\delta > 0$ decreases to zero. This makes also sense in practice because of the finitary precision of the instruments. Hence, we can now assume that $E_{\mathbf{Y}^t_d}[I_{\tilde{\mathbf{y}}^t_d}] > 0$ and apply (GBR), solving thus Eq. (5). As discussed in Section 2, the gambles we are interested in, $g \in \mathcal{L}(\mathcal{X}^t \times \mathcal{Y}^t)$, are $\mathcal{X}_t$-measurable. Furthermore, to make things compatible with Section 3, we assume that $g(\mathbf{x}(t))$ is a bounded real-valued function. When this is not the case, we solve the problem for each component of $g(\mathbf{x}(t))$. Using the linearity property of linear previsions the second equation in (5) can be rewritten as

$$\begin{aligned} 0 &= E_{\mathbf{X}^t,\mathbf{Y}^t_d}[I_{\tilde{\mathbf{y}}^t} g] - \mu\ E_{\mathbf{X}^t,\mathbf{Y}^t_d}[I_{\tilde{\mathbf{y}}^t}] \\ &= E_{\mathbf{X}^t,\mathbf{Y}^t_d}[I_{\tilde{\mathbf{y}}^t} g] - \mu\ E_{\mathbf{Y}^t_d}[I_{\tilde{\mathbf{y}}^t}], \end{aligned} \quad (6)$$

[2] A real-valued function on a domain is called the *indicator function* of a given subset of this domain if it takes the value one in the elements of the subset and zero outside.

and the right-hand side is equal to

$$E_{\mathbf{X}^{t-1},\mathbf{Y}_d^{t-1}}[E_{\mathbf{X}(t)}[E_{\mathbf{Y}_d(t)}[I_{\tilde{\mathbf{y}}^t} g|\mathbf{X}(t), \mathbf{X}^{t-1}, \mathbf{Y}_d^{t-1}] \\ |\mathbf{X}(t-1), \mathbf{X}^{t-2}, \mathbf{Y}_d^{t-1}]] - \mu\ E_{\mathbf{Y}_d^t}[I_{\tilde{\mathbf{y}}^t}], \quad (7)$$

By exploiting the independence assumptions discussed in Section 2 and the fact that the gamble of interest $g(\mathbf{x}(t))$ is a function of $\mathbf{x}(t)$ only, from Eqs. (6)–(7) it can be derived that

$$0 = E_{\mathbf{X}(t)}\left[E_{\mathbf{Y}_d(t)}\left[I_{\tilde{\mathbf{y}}_d(t)} g \middle| \mathbf{X}(t)\right] \middle| \tilde{\mathbf{y}}_d^{t-1}\right] E_{\mathbf{Y}_d^{t-1}}\left[I_{\tilde{\mathbf{y}}_d^{t-1}}\right] \\ - \mu\ E_{\mathbf{Y}_d^t}[I_{\tilde{\mathbf{y}}^t}]. \quad (8)$$

Since we are assuming $E_{\mathbf{Y}_d^t}[I_{\tilde{\mathbf{y}}_d^t}] > 0$, Eq. (8) can be solved with respect to μ, and we obtain

$$\mu = \frac{E_{\mathbf{X}(t)}\left[g\ E_{\mathbf{Y}_d(t)}\left[I_{\tilde{\mathbf{y}}_d(t)} \middle| \mathbf{X}(t)\right] \middle| \tilde{\mathbf{y}}_d^{t-1}\right]}{E_{\mathbf{Y}_d(t)}\left[I_{\tilde{\mathbf{y}}_d(t)} \middle| \tilde{\mathbf{y}}_d^{t-1}\right]}. \quad (9)$$

Finally, assuming some regularity conditions [7] and that the radius of neighbourhoods does not depend on $\mathbf{x}^t$, for $\delta \to 0$ one gets Bayes' rule for conditional PDF, i.e. that $E_{\mathbf{X}(t)}[g|\tilde{\mathbf{y}}^t]$ is equal to

$$\frac{\int_{\mathbf{x}(t)} g(\mathbf{x}(t))\ p(\mathbf{x}(t)|\tilde{\mathbf{y}}^{t-1}) p(\tilde{\mathbf{y}}(t)|\mathbf{x}(t)) d\mathbf{x}(t)}{p(\tilde{\mathbf{y}}(t)|\tilde{\mathbf{y}}^{t-1})}.$$

Hence, $E_{\mathbf{X}(t)}[\cdot|\tilde{\mathbf{y}}^t]$ is a linear functional which is completely determined by the PDFs $p(\mathbf{x}(t)|\tilde{\mathbf{y}}^{t-1})$ and $p(\tilde{\mathbf{y}}(t)|\mathbf{x}(t))$. Consider now the case in which the available information does not allow us to specify a unique probability measure describing each source of uncertainty in the dynamical system. We can then use coherent lower previsions to model the available knowledge. Consider a coherent lower prevision $\underline{E}_{\mathbf{X}(0)}$ and separately coherent $\underline{E}_{\mathbf{X}(t)}(\cdot|\mathbf{X}(t-1))$, $\underline{E}_{\mathbf{Y}_d(t)}(\cdot|\mathbf{X}(t))$ for $t \geq 1$, and let us derive from them a separately coherent conditional lower prevision $\underline{E}_{\mathbf{X}(t)}[g|\tilde{\mathbf{y}}_d^t]$. As in (5), this can be done by applying (GBR) on $\underline{E}_{\mathbf{X}^t,\mathbf{Y}_d^t}$:

$$\underline{E}_{\mathbf{X}(t)}[g|\tilde{\mathbf{y}}_d^t] = \mu \ \ s.t. \ \ \underline{E}_{\mathbf{X}^t,\mathbf{Y}_d^t}[I_{\tilde{\mathbf{y}}_d^t} \cdot (g-\mu)] = 0. \quad (10)$$

Since the measurements have been assumed discrete, there is a unique solution of Eq. (10) if $\underline{E}_{\mathbf{X}^t,\mathbf{Y}_d^t}[I_{\tilde{\mathbf{y}}_d^t}] > 0$. In the sequel, it is assumed that $\mathbf{X}^{t-1}$ and $\mathbf{Y}_d^{t-1}$ are epistemically irrelevant to $\mathbf{X}(t)$ given $\mathbf{X}(t-1)$ and that $\mathbf{X}^{t-1}$ and $\mathbf{Y}_d^{t-1}$ are irrelevant to $\mathbf{Y}_d(t)$ given $\mathbf{X}(t)$, i.e., there exist $\underline{E}_{\mathbf{X}(t)}[\cdot|\mathbf{x}(t-1)]$ and $\underline{E}_{\mathbf{Y}_d(t)}[\cdot|\mathbf{x}(t)]$ such that:

$$\begin{aligned} \underline{E}_{\mathbf{X}(t)}\left[h_1 \middle| \mathbf{x}(t-1)\right] &= \underline{E}_{\mathbf{X}(t)}\left[h_1 \middle| \mathbf{x}^{t-1}, \mathbf{y}_d^{t-1}\right] \\ \underline{E}_{\mathbf{Y}_d(t)}\left[h_2 \middle| \mathbf{x}(t)\right] &= \underline{E}_{\mathbf{Y}_d(t)}\left[h_2 \middle| \mathbf{x}^t, \mathbf{y}_d^{t-1}\right], \end{aligned}$$

for each $h_1 \in \mathcal{L}(\mathcal{X}^t \times \mathcal{Y}^{t-1})$, $h_2 \in \mathcal{L}(\mathcal{X}^t \times \mathcal{Y}^t)$, $\mathbf{x}^t$ and $\mathbf{y}_d^{t-1}$. The joint $\underline{E}_{\mathbf{X}^t,\mathbf{Y}_d^t}$ in (10) can be obtained from $\underline{E}_{\mathbf{X}(0)}$, $\underline{E}_{\mathbf{X}(t)}(\cdot|\mathbf{X}(t-1))$, $\underline{E}_{\mathbf{Y}_d(t)}(\cdot|\mathbf{X}(t))$, for each $t \geq 1$, by marginal extension, and can be written as

$$\underline{E}_{\mathbf{X}(0)}[\underline{E}_{\mathbf{X}(1)}[\underline{E}_{\mathbf{Y}_d(1)}[\dots \underline{E}_{\mathbf{X}(t)}[\underline{E}_{\mathbf{Y}_d(t)}[\cdot|\mathbf{X}(t)]|\mathbf{X}(t-1)] \\ \dots|\mathbf{X}(1)]|\mathbf{X}(0)]], \quad (11)$$

taking also into account that (i) the gamble of interest g in $I_{\tilde{\mathbf{y}}_d^t}(g-\mu)$ is a function of $\mathbf{x}(t)$ only ; and (ii) the gamble $I_{\tilde{\mathbf{y}}_d^t} = I_{\tilde{\mathbf{y}}_d(1)} \cdots I_{\tilde{\mathbf{y}}_d(t)}$ depends on the values of the measurements $\tilde{\mathbf{y}}_d^t$ only. If we now apply conditions (C2) and (SC2), the recursivity of the marginal extension and introduce the notation $g_t(\mathbf{x}(t),\mu) = g(\mathbf{x}(t)) - \mu$, Eq. (10) can be conveniently rewritten as

$$0 = \underline{E}_{\mathbf{X}^{t-1},\mathbf{Y}_d^{t-1}}\left[I_{\tilde{\mathbf{y}}_d^{t-1}}\ g_{t-1}\right], \quad (12)$$

where

$$g_{t-1} = \underline{E}_{\mathbf{X}(t)}\Big[g_t\Big(I_{\{g_t \geq 0\}} \underline{E}_{\mathbf{Y}_d(t)}\left[I_{\tilde{\mathbf{y}}_d(t)} \middle| \mathbf{X}(t)\right] \\ + I_{\{g_t < 0\}} \overline{E}_{\mathbf{Y}_d(t)}\left[I_{\tilde{\mathbf{y}}_d(t)} \middle| \mathbf{X}(t)\right]\Big) \Big| \mathbf{X}(t-1)\Big]. \quad (13)$$

If we proceed recursively in this way, we obtain $\underline{E}_{\mathbf{x}(0)}[g_0] = 0$, where

$$g_0 = \underline{E}_{\mathbf{X}(1)}\Big[g_1\Big(I_{\{g_1 \geq 0\}} \underline{E}_{\mathbf{Y}_d(1)}\left[I_{\tilde{\mathbf{y}}_d(1)} \middle| \mathbf{X}(1)\right] \\ + I_{\{g_1 < 0\}} \overline{E}_{\mathbf{Y}_d(1)}\left[I_{\tilde{\mathbf{y}}_d(1)} \middle| \mathbf{X}(1)\right]\Big) \Big| \mathbf{X}(0)\Big]. \quad (14)$$

By comparing (12) with (7)–(9), we see that when we use coherent lower previsions we cannot derive an expression for the conditional similar to Eq. (9). This is due to the fact that coherent lower previsions are super-additive instead of linear.

We can also make the following observations about the model presented in Eqs. (12)–(14):

1. As we show in the Appendix, $\underline{E}_{\mathbf{X}^t,\mathbf{Y}_d^t}$ is coherent with all the initial assessments. Moreover, it is the smallest, or least-committal, coherent lower prevision which is compatible with the local models and the additional hypothesis of epistemic irrelevance.
2. Moreover, $\underline{E}_{\mathbf{X}(t)}[g|\tilde{\mathbf{y}}_d^t]$ is not only coherent with $\underline{E}_{\mathbf{X}^t,\mathbf{Y}_d^t}$, but also with all the local assessments we have used to derive $\underline{E}_{\mathbf{X}^t,\mathbf{Y}_d^t}$. This is also proven in the Appendix.

In order to compute $\underline{E}_{\mathbf{X}(t)}[g(\mathbf{x}(t))|\tilde{\mathbf{y}}_d^t]$, it is necessary to propagate back to time the functional $g(\mathbf{x}(t)) - \mu$ until the initial state is reached, and then to find the value of μ which satisfies $\underline{E}_{\mathbf{x}(0)}[g_0] = 0$. Each step can be very heavy from a computational point view. Moreover, the computational complexity increases up to time linearly. Possible ways to overcome this computational issue are: (i) to find classes of CLPs for which the computation of (12)–(14) is feasible; (ii) to truncate the recursion after N steps in the past by finding a CLP which approximates $\underline{E}_{\mathbf{X}(t-N)}[g|\tilde{\mathbf{y}}_d^{t-N}]$. Concerning the first point, one idea would be to use a particular case of coherent lower previsions: *2-monotone lower previsions* [3], which have a number of useful properties: they are closed under convex combinations; and their value for a generic gamble can simply be obtained by applying the Choquet integral w.r.t. their restriction to events. An example of 2-monotone lower previsions, for which the solution of (12)–(14) is affordable, is the linear-vacuous mixture model in Section 5.

4.1 Decision making and estimation

Let us briefly discuss the decision making approach to estimation. The Bayesian methodology provides the estimate which minimises the expected posterior risk. If in particular we consider a squared error loss risk, the Bayesian estimate is the mean of the posterior distribution. This estimation is provided in general together with its *credibility region*: a $100(1-\alpha)$ credibility region for a scalar random variable x is a region χ such that $E(I_{\{x\in\chi\}}) = Pr(x \in \chi) = 1-\alpha$, where $Pr(\cdot)$ is the posterior distribution. When we consider sets of probabilities, we deal with lower and upper expectations and, thus, with interval-valued expectations $[\underline{E}(\cdot), \overline{E}(\cdot)]$, leading to the problem of decision making under imprecision [7]. A consequence of imprecision is that, in general, we must abandon the idea of choosing a unique value for the estimate. With this in mind, the path followed in this paper is that of extending the Bayesian decision making approach to the IP framework by calculating the lower $\underline{E}(\mathbf{x}(t))$ and upper $\overline{E}(\mathbf{x}(t))$ means and an IP version of the credibility region. In particular, the IP credibility region is evaluated by seeking for the minimum volume region χ such that $\underline{E}(I_{\{x\in\chi\}}) > 1-\alpha$. It is easy to see that, in the precise case, the IP credibility region coincides with the Bayesian one and that $\overline{E}(\mathbf{x}(t)) = \underline{E}(\mathbf{x}(t)) = \hat{\mathbf{x}}(t)$.

5 Linear-vacuous mixture model

Assume now that the knowledge on the initial state and state evolution process is modeled by *linear-vacuous mixtures*:

$$\underline{E}_{\mathbf{x}(0)}(g) = \epsilon_x \int\limits_{\mathbf{x}(0)} g(\mathbf{x}(0))\ \mathcal{N}(\mathbf{x}(0); \hat{\mathbf{x}}(0), \mathbf{P}(0))\ d\mathbf{x}(0) + (1-\epsilon_x) \inf_{\mathbf{x}(0)} g(\mathbf{x}(0)), \quad (15)$$

$$\begin{aligned}&\underline{E}_{\mathbf{x}(t)}(g|\mathbf{x}(t-1)) \\ &= \epsilon_w \int\limits_{\mathbf{x}(t)} g(\mathbf{x}(t))\ \mathcal{N}(\mathbf{x}(t); \mathbf{A}\mathbf{x}(t-1), \mathbf{Q})\ d\mathbf{x}(t) \\ &+ (1-\epsilon_w) \inf_{\mathbf{x}(t)} g(\mathbf{x}(t)),\end{aligned} \quad (16)$$

where the scalars ϵ_w and ϵ_x belong to $[0,1]$. Furthermore, assume that discrete measurements of the state are available and that the uncertainty on the measurement process can be represented with a *linear prevision*[3] $E_{\mathbf{Y}_d(t)}(h|\mathbf{x}(t))$ given by

$$\sum_{\mathbf{y}_d(t)} h(\mathbf{y}_d(t)) \int\limits_{\mathbf{y}_d(t)} I_{\mathbf{y}_d(t)}(\mathbf{u}(t)) \mathcal{N}(\mathbf{u}(t); \mathbf{C}\mathbf{x}(t), \mathbf{R})\ d\mathbf{u}(t). \quad (17)$$

This generalises the model given in Eq. (4) to the linear-vacuous mixtures and can be used for example to model the imprecision of the linear time-invariant system (4) but where the process noise is $\mathbf{w}_\epsilon(t) = \epsilon_w \mathbf{w}(t) + (1-\epsilon_w)\mathbf{n}(t)$ and $\mathbf{x}(0) \sim \epsilon_x \mathcal{N}(\hat{\mathbf{x}}(0), \mathbf{P}(0)) + (1-\epsilon_x)\mathbf{u}(t)$, and the noises $\mathbf{n}(t)$ and $\mathbf{u}(t)$ are assumed to have unknown distributions (not necessarily constant w.r.t. time). Note that the model which characterises both $\mathbf{w}_\epsilon(t)$ and $\mathbf{x}(0)$ is the so-called *ϵ-contamination* which has been widely used in Bayesian robustness [1]. The correspondence between this system and (15)–(17) follows from the fact that for $\mathbf{w}_\epsilon(t)$ the ϵ-contamination model implies that $\underline{E}_{\mathbf{W}_\epsilon(t)} = \epsilon_w \int_{\mathbf{w}(t)} g(\mathbf{w}(t))\ \mathcal{N}(\mathbf{w}(t); \mathbf{0}, \mathbf{Q})\ d\mathbf{w}(t) + (1-\epsilon_w) \inf_{\mathbf{w}(t)} g(\mathbf{w}(t))$ [7]. Hence, we can exploit a result from [5] to prove that the knowledge of $\underline{E}_{\mathbf{W}_\epsilon(t)}$ and the fact $\mathbf{x}(t+1) = \mathbf{A}\mathbf{x}(t) + \mathbf{w}_\epsilon(t)$ together imply (16). This holds also for $\mathbf{x}_0$ and extends to CLPs a well-known result from the precise case. Assuming that the discretization step δ in $\mathbf{y}_d(t) = \mathcal{B}(\mathbf{y}(t), \delta)$ is small, the last integral can be approximated as $\nu(\delta)\mathcal{N}(\mathbf{y}(t); \mathbf{C}\mathbf{x}(t), \mathbf{R})$ where $\nu(\delta) > 0$ is the Lebesgue measure of $\mathcal{B}(\mathbf{y}(t), \delta)$, which has been assumed independent of $\tilde{\mathbf{y}}(t)$. Hence,

$$E_{\mathbf{Y}_d(t)}(h|\mathbf{x}(t)) \approx \nu(\delta) \sum_{\mathbf{y}_d(t)} h(\mathbf{y}_d(t))\ \mathcal{N}(\mathbf{y}(t); \mathbf{C}\mathbf{x}(t), \mathbf{R}).$$

Recall that our aim is to solve (10). Applying the results in (12)–(14), the target conditional CLP can be calculated by solving w.r.t. μ the following equation:

$$\underline{E}_{\mathbf{x}(0)}(g_0) \approx \epsilon_x \int\limits_{\mathbf{x}(0)} g_0(\mathbf{x}(0), \mu)\ \mathcal{N}(\mathbf{x}(0); \hat{\mathbf{x}}(0), \mathbf{P}(0)) + (1-\epsilon_x) \inf_{\mathbf{x}(0)} g_0(\mathbf{x}(0), \mu) = 0, \quad (18)$$

where $g_0(\mathbf{x}(0), \mu)$ can be obtained recursively by

$$\begin{aligned}&g_{k-1}(\mathbf{x}(k-1), \mu) \\ &= \nu(\delta)\epsilon_w \int_{\mathbf{x}(k)} g_k(\mathbf{x}(k), \mu) \mathcal{N}(\mathbf{x}(k); \mathbf{A}\mathbf{x}(k-1), \mathbf{Q}) \\ &\qquad \cdot \mathcal{N}(\tilde{\mathbf{y}}(k); \mathbf{C}\mathbf{x}(k), \mathbf{R})\ d\mathbf{x}(k) \\ &+ \nu(\delta)(1-\epsilon_w) \inf_{\mathbf{x}(k)} g_k(\mathbf{x}(k), \mu) \mathcal{N}(\tilde{\mathbf{y}}(k); \mathbf{C}\mathbf{x}(k), \mathbf{R}),\end{aligned}$$

[3]The assumption of a probabilistic model for the measurement process instead of an imprecise model, like the ones for the initial state and the state evolution process, is used just to simplify the derivations. However, the derivation presented in Section 4 is general and allows also for imprecise measurements.

for $k = 1, \dots, t$, with the final condition $g_t(\mathbf{x}(t), \mu) = g(\mathbf{x}(t)) - \mu$ and with $\tilde{\mathbf{y}}_d(t) = \mathcal{B}(\tilde{\mathbf{y}}(t), \delta)$.

The value $g_{k-1}(\mathbf{x}(k-1), \mu)$ is the sum of two terms. The first is the expected value of $g_k(\mathbf{x}(k), \mu)$ w.r.t. a Gaussian and the second is the infimum of $g_k(\mathbf{x}(k), \mu)$, also weighted by a Gaussian. The first term can also be regarded as a linear operator $\mathbb{I}_k[\cdot]$ which operates on the function $g_k(\mathbf{x}(k), \mu)$ and produces a function of $\mathbf{x}(k-1)$ and μ. The second term can be seen as an operator $\mathbb{M}_k[\cdot]$ on the function $g_k(\mathbf{x}(k), \mu)$, but it produces a function of μ only. Hence, at time t, the previous equation can be rewritten as follows:

$$g_{t-1}(\mathbf{x}(t-1), \mu) = \mathbb{I}_t[g_t(\mathbf{x}(t), \mu)] + \mathbb{M}_t[g_t(\mathbf{x}(t), \mu)], \quad (19)$$

and, thus, at the time $t-1$:

$$\begin{aligned} & g_{t-2}(\mathbf{x}(t-2), \mu) \\ & = \mathbb{I}_{t-1}[\mathbb{I}_t[g_t(\mathbf{x}(t), \mu)]] + \mathbb{I}_{t-1}[1]\mathbb{M}_t[g_t(\mathbf{x}(t), \mu)] \\ & + \mathbb{M}_{t-1}\Big[\mathbb{I}_t[g_t(\mathbf{x}(t), \mu)] + \mathbb{M}_t[g_t(\mathbf{x}(t), \mu)]\Big], \end{aligned} \quad (20)$$

using the linearity of $\mathbb{I}$ and the fact that $\mathbb{M}_t[g_t(\mathbf{x}(t), \mu)]$ is a function of μ only. Hence, (18) can be decomposed as

$$\begin{aligned} & \mathbb{I}_0[\mathbb{I}_1[\dots\mathbb{I}_t[\cdot]]] + \mathbb{I}_0[1]\mathbb{M}_1[\cdot] + \mathbb{I}_0[\mathbb{I}_1[1]]\mathbb{M}_2[\cdot] + \\ & + \dots + \mathbb{I}_0[\mathbb{I}_1[\dots\mathbb{I}_{t-1}[1]]]\mathbb{M}_t[\cdot] + \mathbb{M}_0\Big[\mathbb{I}_1[\dots\mathbb{I}_t[\cdot]] \\ & + \mathbb{M}_1[\cdot] + \mathbb{I}_1[1]\mathbb{M}_2[\cdot] + \dots + \mathbb{I}_1[\dots\mathbb{I}_{t-1}[1]]\mathbb{M}_t[\cdot]\Big], \end{aligned} \quad (21)$$

where, for the sake of notation, the arguments of the operators have not been made explicit, but can be recovered from (19)–(20). The operators $\mathbb{I}_0$ and $\mathbb{M}_0$ are slightly different from $\mathbb{I}_k$ and $\mathbb{M}_k$, for $k > 0$, as it can be seen from (18). Let us give some comments on (21). The term $\mathbb{I}_0[\mathbb{I}_1[\dots\mathbb{I}_t[\cdot]]]$ is equal to

$$\begin{aligned} & \nu(\delta)^t \epsilon_x \epsilon_w^t \prod_{k=1}^{t} \mathcal{N}(\tilde{\mathbf{y}}(k); \mathbf{C}\mathbf{A}\hat{\mathbf{x}}(k-1), \mathbf{S}(k)) \\ & \int_{\mathbf{x}(t)} g_t(\mathbf{x}(t), \mu)\, \mathcal{N}(\mathbf{x}(t); \hat{\mathbf{x}}(t), \mathbf{P}(t)), \end{aligned} \quad (22)$$

where $\mathbf{S}(k) = \mathbf{R} + \mathbf{C}\mathbf{P}(k|k-1)\mathbf{C}'$, $\hat{\mathbf{x}}(k)$, $\mathbf{P}(k)$, $\mathbf{P}(k|k-1) = \mathbf{A}\mathbf{P}(k)\mathbf{A}' + \mathbf{Q}$, and can be calculated by using the KF from the prior $\mathcal{N}(\mathbf{x}(0); \hat{\mathbf{x}}(0), \mathbf{P}(0))$. This gives the solution of the estimation problem in the precise case, i.e. $\epsilon_x = \epsilon_w = 1$. The generic term $\mathbb{I}_i[\mathbb{I}_{i+1}[\dots\mathbb{I}_{j-1}[\text{arg}]]]$ with argument [arg] equal to $g_t(\mathbf{x}(t), \mu)$, for $1 \le i < j = t$, or to 1, for $1 \le i < j < t$, is

$$\begin{aligned} & \nu(\delta)^{j-i} \epsilon_w^{j-i} \int_{\mathbf{x}(i)} \dots \int_{\mathbf{x}(j-1)} [\text{arg}] \\ & \mathcal{N}(\mathbf{x}(i); \mathbf{A}\mathbf{x}(i-1), \mathbf{Q})\mathcal{N}(\tilde{\mathbf{y}}(i); \mathbf{C}\mathbf{x}(i), \mathbf{R}) \\ & \cdots \mathcal{N}(\mathbf{x}(j-1); \mathbf{A}\mathbf{x}(j-2), \mathbf{Q})\mathcal{N}(\tilde{\mathbf{y}}(j-1); \mathbf{C}\mathbf{x}(i), \mathbf{R}). \end{aligned} \quad (23)$$

By applying the matrix inversion lemma, Eq. (23) can be simplified as follows:

$$\begin{aligned} & \nu(\delta)^{j-i} \epsilon_w^{j-i} \mathcal{N}(\tilde{\mathbf{y}}(i); \mathbf{C}\mathbf{A}\mathbf{x}(i-1), \mathbf{W}_1) \\ & \cdot \prod_{k=i+1}^{j-1} \mathcal{N}(\tilde{\mathbf{y}}(k); \mathbf{C}\mathbf{A}\hat{\mathbf{x}}^*(k-1), \mathbf{S}^*(k)) \\ & \int_{\mathbf{x}(j-1)} [\text{arg}]\, \mathcal{N}(\mathbf{x}(j-1); \hat{\mathbf{x}}^*(j-1), \mathbf{P}^*(j-1))\, d\mathbf{x}(j-1), \end{aligned} \quad (24)$$

where $\mathbf{W}_1^{-1} = \mathbf{R}^{-1} + \mathbf{C}\mathbf{Q}^{-1}\mathbf{C}'$ and $\hat{\mathbf{x}}^*(k)$, $\mathbf{P}^*(k)$, $\mathbf{S}^*(k)$ can be calculated by using the KF starting from the prior $\mathcal{N}(\mathbf{x}(i); \mathbf{W}_2\mathbf{Q}^{-1}\mathbf{A}\mathbf{x}(i-1) + \mathbf{W}_2\mathbf{C}'\mathbf{R}^{-1}\tilde{\mathbf{y}}(i), \mathbf{W}_2)$ with $\mathbf{W}_2^{-1} = \mathbf{Q}^{-1} + \mathbf{C}'\mathbf{R}^{-1}\mathbf{C}$. For [arg] = 1 the integral in (24) marginalises and is equal to 1. The terms $\mathbb{I}_0[\mathbb{I}_1[\dots\mathbb{I}_j[1]]]$, with $0 < j < t$, are equal to

$$\nu(\delta)^j \epsilon_x \epsilon_w^j \prod_{k=1}^{j} \mathcal{N}(\tilde{\mathbf{y}}(k); \mathbf{C}\mathbf{A}\hat{\mathbf{x}}^*(k-1), \mathbf{S}^*(k)). \quad (25)$$

Note that the constants $\nu(\delta)$ can be dropped out to solve (21). In the sequel, we refer to the algorithm presented in this section as the Linear Gaussian-Vacuous Mixture filter (LGVM).

6 Numerical example

We have performed Monte Carlo simulations in order to show the basic features of the LGVM filter presented in the previous section. These simulations compare the performance of the LGVM with the KF, considering non-Gaussian situations. The one-dimensional model

$$\begin{cases} x(t+1) & = & x(t) + w_\epsilon(t) \\ y(t) & = & x(t) + v(t) \end{cases}$$

has been considered, where $w_\epsilon(t) = \epsilon_w w(t) + (1 - \epsilon_w)n(t)$, $w(t) \sim \mathcal{N}(0, Q)$, $x(0) = \epsilon_x \hat{x}(0) + (1 - \epsilon_x)u(t)$, $\hat{x}(0) \sim N(0, P(0))$ and $v(t) \sim N(0, R)$. Note that, in all simulations, both the LGVM and the KF were designed without assuming the knowledge of the contaminating terms $n(t)$ and $u(t)$. The aim is to investigate the relative sensitivity of the KF and the LGVM to (heavy tailed) disturbances of the nominal Gaussian density. The performance of the filters has been investigated considering different values of the epsilons and different distributions for the contaminating terms. A trajectory of 15 timesteps and a Monte Carlo size of 100 runs are considered. The following cases have been simulated

Case	ϵ_w	PDF for n	Q	R/Q
1	0.95	$5\delta(7)/(1-\epsilon_w)$	0.1	1
2	0.95	$5\delta(7)/(1-\epsilon_w)$	0.1	0.1
3	0.9999	$5\delta(7)/(1-\epsilon_w)$	0.1	1
4	0.95	$\mathcal{N}(0, 125)$	0.1	1

where $\delta(k)$ is 1 when the time t is equal to k and 0 otherwise and ϵ_x was fixed to 1. In the cases 1–3, the trajectory undergoes a jump of 5 units at the time instant $t = 7$. This can be interpreted as an unmodelled manoeuvre. For these cases, the simulation results are shown in Figures 1–3 for a fixed trajectory, i.e., Monte Carlo runs have been performed only w.r.t the measurement's noise realisations. The figures report the true trajectory (TS), the averaged KF's estimate (KF) and the relative 99% credibility interval (Cred KF), the averaged lower (LP) and upper (UP) means and the IP version of credibility interval (Cred IP) as defined in Section 4.1. From Figure 1, it can be noticed that from time 1 to time 6 the KF and the LGVM provide more or less the same credibility interval and the upper and lower means are almost equal and coincide with the KF estimate. At the jump's instant, $t = 7$, the KF estimate is wrong, since the 99% credibility interval does not include the true state. This shows that the KF is not robust to large model errors. On the other hand, the LGVM correctly detects the jump and it is able to enlarge the credibility interval in order to include the true state. Notice also that the difference $\overline{E}(\mathbf{x}(t)) - \underline{E}(\mathbf{x}(t))$ is related to the imprecision present in the system. From the instant $t = 8$ to the end, the true trajectory enters again inside the KF credibility region, since no more jumps occur. Also the LGVM converges towards the true state as it can be seen from the reduction of the size of the credibility interval. However, the convergence rate is slower than that of the KF and depends on the variance-ratio R/Q, as it can be seen comparing Figures 1–2, and on the value of ϵ_w as it can be seen comparing 1 and 3. Since the LGVM is able to deal with all possible contaminating distributions, its slow convergence is expected also for $\epsilon_w = 0.95$. In fact, the set of possible contaminating includes also densities close to Dirac functions, i.e., distributions with zero variance. These are the most difficult contaminations to be filtered out, since they encode a very strong information which needs "several" measurements to be falsified. Conversely, contaminating distributions with large variance are easily filtered out, since the filters prefer the measurements to the information encoded by these distributions. About Figure 3, the fact that LVGM is still robust when $\epsilon_w = 0.9999$ is due to the fact that, at time $t = 7$, the prediction of the KF is so far from the measurement that the value of the precise term (22) is much smaller than the values of the other terms in (25). These results thus show that the LGVM filter outperforms the KF performance when a small value of ϵ_w is selected. In fact, in these cases, LGVM is still robust to unmodelled errors and preserves the fast convergence rate of the KF. Obviously, at the increasing of $\epsilon_w \to 1$ there is a value of ϵ_w for which LGVM and KF almost coincide. In case 4, the contaminating term is a Gaussian with zero mean and variance 125. Thus, the noise w_ϵ is normal-distributed with zero mean and variance $Q_w = \epsilon_w^2 Q + (1 - \epsilon_w)^2 125 \approx 0.4$. The width of the IP version of the 99% credibility interval has been compared with the true 99% credibility interval based on $Q_w = 0.4$. The average ratio between the size of the two intervals are listed hereafter $0.8723(t = 1)$, $1.008(t = 2)$, $1.0027(t = 3)$ and it converges to 1 after $t > 3$. Thus, although the LGVM does not know the contaminating term is able to correctly determine the width of the credibility interval, while the KF can only underestimate its size. Finally, notice that the example discussed here could be generalised to the n-dimensional case. From the practical point of view, the most difficult aspect would be the computation of the operators $\mathbb{M}_k$ whose calculation requires the solution of a minimisation over a vector instead of a scalar.

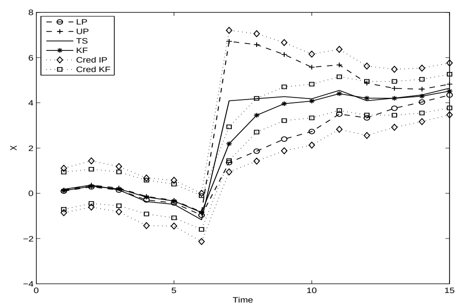

Figure 1: Case 1: $\epsilon_w = 0.95$, $R/Q = 1$.

7 Conclusions

In this paper, we have proposed an extension of Hidden Markov Models that allows for imprecision in our knowledge about the elements of the model, and is arguably more realistic in practical situations. We have also shown, in a practical case, that our extension outperforms the Kalman filter when modelling errors are present in the system. As future prospects, we intend to deepen the comparison with the classical results and to investigate in detail the modelling by means of 2-monotone lower previsions and the truncated recursions discussed in Section 4. Furthermore, we intend to investigate the extension of LGVM to the case where the contaminating distributions are unimodal and/or symmetric. Finally, we plan to generalise the example discussed in Section 6 to the n-dimensional case.

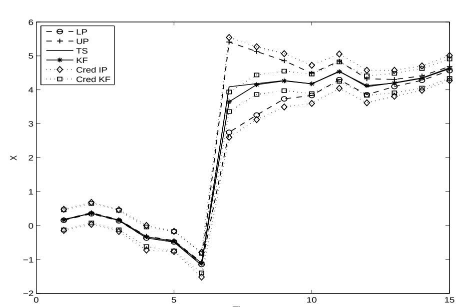

Figure 2: Case 2: $\epsilon_w = 0.95$, $R/Q = 0.1$.

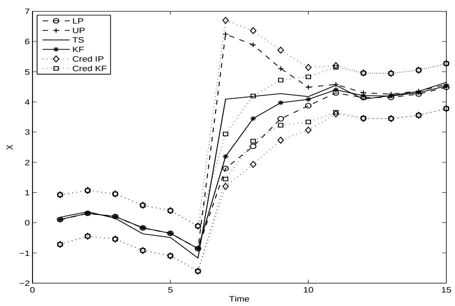

Figure 3: Case 3: $\epsilon_w = 0.9999$, $R/Q = 1$.

Appendix: Coherence results

In this appendix, we prove that the joint and the conditional lower previsions we consider in this paper satisfy the property of coherence introduced in Section 3. To simplify a bit the notation, let $Z_1 = X_0$ and let us define, for every $j \in \{1,\dots,n\}$, $Z_{2j+1} = Y_j, Z_{2j} = X_j$. Let $\mathcal{Z}_k$ be the corresponding possibility space for Z_k. Then the previous assessments can be expressed as

$$\underline{E}_{Z_1}(\cdot), \underline{E}_{Z_2}(\cdot|Z_1), \underline{E}_{Z_{2j+1}}(\cdot|Z_{2j}), j = 1,\dots,n, \\ \underline{E}_{Z_{2k}}(\cdot|Z_{2k-2}), k = 2,\dots,n. \quad (26)$$

Here the domain of $\underline{E}_{Z_{2j+1}}(\cdot|Z_{2j})$ is the set of $\mathcal{Z}_{2j,2j+1}$-measurable gambles, and the domain of $\underline{E}_{Z_{2j}}(\cdot|Z_{2j-2})$ is the set of $\mathcal{Z}_{2j-2,2j}$-measurable gambles. We shall also use the notation $Z^k = Z_1 \times \cdots \times Z_k$ for the product random variable that takes values in the product space $\mathcal{Z}^k = \mathcal{Z}_1 \times \cdots \times \mathcal{Z}_k$. In the language of coherence graphs [6], the collection of conditional lower previsions in Eq. (26) is $A1^+$-representable: each variable $Z_j, j = 1,\dots,2n+1$, appears exactly one time on the left hand side of the conditioning bar, and moreover Z_j is a predecessor of Z_k if and only if $k > j$. As a consequence, we can apply [8, Theorem 2] and deduce that they are coherent with their strong product $\underline{E}$ on $\mathcal{L}(\mathcal{Z}^{2n+1})$. What we set out to prove is that this strong product is precisely the joint lower prevision $\underline{S}$ we have constructed in Eq. (11). For every $j = 3,\dots,2n+1$, let us define the conditional lower prevision $\underline{Q}_{Z_j}(\cdot|Z^{j-1})$ on the set of $\mathcal{Z}^j$-measurable gambles by

$$\underline{Q}_{Z_j}(f|z) = \begin{cases} \underline{E}_{Z_j}(f(z,\cdot)|\pi_{j-1}(z)) & \text{if } j \text{ odd} \\ \underline{E}_{Z_j}(f(z,\cdot)|\pi_{j-2}(z)) & \text{if } j \text{ even,} \end{cases}$$

for every $\mathcal{Z}^j$-measurable gamble f and every $z \in \mathcal{Z}^{j-1}$. Let $\pi^j = \pi_{1,\dots,j}$ to simplify the notation. If we also let $h(j)$ be equal to $j-1$ if j is odd, and equal to $j-2$ if j is even, we would have $\underline{Q}_{Z_j}(f|z) = \underline{E}_{Z_j}(f(z,\cdot)|\pi_{h(j)}(z))$ for every $\mathcal{Z}^j$-measurable gamble f and every $z \in \mathcal{Z}^{j-1}$. Let us show that $\underline{Q}_{Z_j}(\cdot|Z^{j-1})$ is separately coherent for every $j = 1,\dots,2n+1$. This is trivial for $j = 1, 2$, so let us consider $j \in \{3,\dots,2n+1\}$. Since the domain of $\underline{Q}_{Z_j}(\cdot|Z^{j-1})$ is a linear space of gambles for all $j \in \{1,\dots,2n+1\}$, separate coherence is equivalent to conditions (SC1)–(SC3) in Section 3. It is not difficult to show that these conditions follow as a consequence of the separate coherence of $\underline{E}_{Z_j}(\cdot|Z_{\pi_{h(j)}})$. The lower previsions $\underline{Q}_{Z_j}(\cdot|Z^{j-1}), j = 1,\dots,2n+1$, satisfy the hypotheses of the Generalised Marginal Extension Theorem [4, Th. 4]. Hence, their marginal extension is given, for any gamble f on $\mathcal{Z}^{2n+1}$, by

$$\underline{E}(f) = \underline{Q}_{Z_1}(\underline{Q}_{Z_2}(\dots(\underline{Q}_{Z_{2n+1}}(f|Z^{2n}))\dots|Z_1))$$

and this coincides with the joint constructed in Eq. (11). The same theorem also implies that $\underline{E}$ is the smallest coherent lower prevision which is coherent with $\underline{Q}_{Z_j}(\cdot|Z^{j-1}), j = 1,\dots,2n+1$. Let us now assume that $\underline{E}(z) > 0$ for every $z \in \mathcal{Z}_3 \times \cdots \times \mathcal{Z}_{2n+1}$, and define $\underline{E}_{Z_{2n}}(\cdot|Z_3, Z_5,\dots,Z_{2n+1})$ from $\underline{E}$ using regular extension. In the notation used in Section 4, this corresponds to defining $\underline{E}_{X_n}(\cdot|Y_1,\dots,Y_n)$. From the comments after [8, Def. 2], $\underline{E}$ is not only the marginal extension of $\underline{Q}_{Z_j}(\cdot|Z^{j-1}), j = 1,\dots,2n+1$, but also their strong product: this follows from the fact that, in the notation of [8, Def. 2], $A_k \cup I_k = I_k$ for all $k = 1,\dots,2n+1$. Applying [8, Th. 3], we deduce that $\underline{E}, \underline{Q}_{Z_j}(\cdot|Z^{j-1}), j = 1,\dots,2n+1, \underline{E}_{Z_{2n}}(\cdot|Z_3, Z_5,\dots,Z_{2n+1})$ are coherent.

Acknowledgements

This work has been partially supported by the Swiss NSF grants n. 200020-121785/1 and 200020-116674/1 (first and second author), and by the projects TIN2008-06796-C04-01, MTM2007-61193 (third author).

References

[1] J. O. Berger. *Statistical Decision Theory and Bayesian Analysis.* Springer Series in Statistics, New York, 1985.

[2] G. de Cooman, F. Hermans, A. Antonucci, and M. Zaffalon. Epistemic irrelevance in credal networks: the case of imprecise Markov trees. In *ISIPTA '09 – Proceedings of the Sixth International Symposium on Imprecise Probability*, Durham (UK), 2009. Accepted for publication.

[3] G. de Cooman, M. Troffaes, and E Miranda. n-monotone exact functionals. *Journal of Mathematical Analysis and Applications*, 347(1):133–146, 2008.

[4] E. Miranda and G. de Cooman. Marginal extension in the theory of coherent lower previsions. *International Journal of Approximate Reasoning*, 47(1):188–225, 2007.

[5] E. Miranda, G. de Cooman, and I. Couso. Lower previsions induced by multi-valued mappings. *Journal of Statistical Planning and Inference*, 133(1):173–197, 2005.

[6] E. Miranda and M. Zaffalon. Coherence graphs. *Artificial Intelligence*, 173(1):104–144, 2009.

[7] P. Walley. *Statistical Reasoning with Imprecise Probabilities.* Chapman and Hall, New York, 1991.

[8] M. Zaffalon and E. Miranda. Conservative inference rule for uncertain reasoning under incompleteness. *Journal of Artificial Intelligence Research*, 34:757–821, 2009.