Imprecise Hierarchical Dirichlet model with applications

Alessio Benavoli
Dalle Molle Institute for Artificial Intelligence (IDSIA),
Manno, Switzerland, alessio@idsia.ch

Abstract— **Many estimation problems in data fusion involve multiple parameters that can be related in some way by the structure of the problem. This implies that a joint probabilistic model for these parameters should reflect this dependence. In parametric estimation, a Bayesian way to account for this possible dependence is to use hierarchical models, in which data depends on hidden parameters that in turn depend on hyperprior parameters. An issue in this analysis is how to choose the hyperprior in case of lack of prior information. This paper focuses on parametric estimation problems involving multinomial-Dirichlet models and presents a model of prior ignorance for the hyperparameters. This model consists to a set of Dirichlet distributions that expresses a condition of prior ignorance. We analyse the theoretical properties of this model and we apply it to practical fusion problems: (i) the estimate of the packet drop rate in a centralized sensor network; (ii) the estimate of the transition probabilities for a multiple-model algorithm.**

Keywords: imprecise probability, hierarchical models, multinomial-dirichlet distribution, Markov-chain.

I. Introduction

Many estimation problems in data fusion involve multiple parameters that can be related in some way by the structure of the problem. This implies that a joint probabilistic model for these parameters should reflect this dependence. For example, in a centralized sensor networks, it is often important to estimate the reliability of the communication channel, i.e., the probability θ_i that a packed sent from sensor i to the fusion node is received (this information can be used for instance to devise robust smart communication strategies, e.g., [1]). In this case it is reasonable to expect that the estimates of θ_i should be related to each other (if the deploying area of the network is small, we expect that the communication channels between the fusion node and each sensor node have similar properties). We can model this dependence by using a prior distribution in which the θ_i's are viewed as a sample from a common *population distribution*. For the pair sensor i-fusion node the observed data y_i (a binary variable that is one if a packet is received and 0 is it is lost) are used to estimate θ_i. This problem can be modelled hierarchically with a model where the observable outcomes depend on the θ_i's which, in turn, depend probabilistically on further parameters, called hyperparameters. This hierarchy allows to account for the dependence among the θ_i. A hierarchical model has also another advantage. Since the θ_i are estimated globally, the measurements relative to the communication channel from node i to the fusion node also contribute to the estimate of θ_j, i.e., the packet drop rate for the communication channel from node j to the fusion node. Thus, we can produce a reliable estimate of θ_j even in the case we have very few measurements for the communication channel from node j to the fusion node (because indireclty we use the measurements from the other channels).

In this example, data are categorical (more specifically binary) variables and the parameters θ_i for $i = 1, \dots, p$ are probabilities. In this context, it is natural to consider a multinomial likelihood as probabilistic model for the observations y_{ij} given the parameters θ_i and a Dirichlet distribution for the θ_i's given the hyperparameters (i.e., the parameters of the Dirichlet distribution). Because multinomial and Dirichlet are conjugate models, this choice simplifies the computations. To complete the hierarchical model we must select a prior on the hyperparameters.

In case of lack of prior information, a common approach in Bayesian analysis is to select a noninformative distribution as prior for the hyperparameters: for instance a uniform distribution on the hyperparameter space.[1] However, a uniform distribution does not model ignorance but indifference, in the sense that all the elements of the hyperparameter space have the same probability under a uniform distribution. Conversely, ignorance means that we do not know anything about the hyperparameters neither their distribution. The uniform can be one of the possible distributions for the hyperparameters but not the only one.

Since a single probability measure cannot model prior ignorance, the idea proposed in Bayesian robustness [2] is to use a set of probability measures. Examples are: ϵ-contamination models [2], [3]; restricted ϵ-contamination models [4]; intervals of measures [3], [5]; the density ratio class [5], [6], etc. These are *neighbourhood models*, i.e., the set of all distributions that are close (w.r.t. some criterion) to an ideal distribution(the "centre" of the set). Note that this approach is not suitable in case of total lack of prior information, because in this case there is no ideal prior distribution, since no single prior distribution can adequately model the lack of prior information. In case of total lack of prior information, Walley [6] has instead proposed the use of the so-called "near-ignorance" priors. In choosing a set of probability measures to model total prior ignorance, the main aim is to generate lower

[1] If the hyperparameter space is unbounded, then the noninformative distribution is in general improper.

and upper expectations with the property that $\underline{E}(g) = \inf g$ and $\overline{E}(g) = \sup g$ for a specific class of real-valued functions g of interest in the analysis, e.g., mean, variance, credible interval etc. This means that the only information about $E(g)$ is that it belongs to $[\inf g, \sup g]$, which is equivalent to state a condition of complete prior ignorance about the value of g (this is the reason why we said that a single, however noninformative, prior cannot model prior ignorance). However, such condition of prior ignorance can only be imposed on a subset of the possible functions g (for this reason the model is called near-ignorance prior) otherwise it produces vacuous posterior inferences [6, Ch. 5]. This means that a-posteriori we still satisfy $\underline{E}(g|data) = \underline{E}(g) = \inf g$ and $\overline{E}(g|data) = \overline{E}(g) = \sup g$, i.e., we do not learn from data.[2] Based on this idea, Walley [6], [7] has developed near-ignorance prior models for various statistical models. For the Multinomial-Dirichlet conjugate distribution, Walley has derived a model of prior near-ignorance called Imprecise Dirichlet model (IDM). Before stating the goal of this paper, we briefly introduce IDM and explain what is a model of prior near-ignorance

A. Imprecise Dirichlet Model

The *Imprecise Dirichlet Model* (IDM) has been introduced by Walley [8] to draw inferences about the probability distribution of a categorical variable. Consider a variable Y taking values on a finite set $\mathcal{Y}$ of cardinality m and assume that we have a sample of size N of independent and identically distributed outcomes of Y. Our aim is to estimate the probabilities θ_i for $i = 1, \dots, m$, that is the probability that Y takes the i-th value. In other words, we want to estimate a vector on the m-dimensional simplex:

$$\Delta_m(p) = \left\{ (\theta_1, \dots, \theta_m) : \theta_i \geq 0, \quad \sum_{j=1}^{m} \theta_j = 1 \right\}. \quad (1)$$

A Bayesian approach consists in assuming a categorical distribution for the data $Ca(y|\theta_1, \dots, \theta_m)$, i.e.:

$$Ca(y|\theta_1, \dots, \theta_m) \propto \theta_i,$$

if the category i is observed, i.e., if $y = y_i$. If there are N i.i.d. observations, then the likelihood is a multinomial distribution (a product of categorical distributions), i.e.:

$$M(data|\theta_1, \dots, \theta_m) \propto \theta_1^{n_1} \cdots \theta_m^{n_m},$$

where n_i is the number of observations for the category i and $\sum_{i=1}^{m} n_i = N$. As prior distribution for the vector of variables $(\theta_1, \dots, \theta_m)$, it is common to assume a Dirichlet distribution

$$\frac{\Gamma(s)}{\prod_{i=1}^{m} \Gamma(st_i)} \prod_{i=1}^{m} \theta_i^{st_i - 1},$$

where $\Gamma(\cdot)$ is the Gamma function, $s > 0$ is the prior strength and $t_i > 0$ (with $\sum_{i=1}^{m} t_i = 1$) is the prior mean of θ_i.[3] The goal is to compute the posterior expectation of θ_i given the measurements. Note that the Dirichlet distribution depends on the parameters s, a positive real value, and $(t_1, \dots, t_m)$, a vector of positive real numbers which satisfy $\sum_{i=1}^{m} t_i = 1$. In case of lack of prior information, an issue in Bayesian analysis is how to choose these parameters to reflect this condition of prior ignorance. To address this issue, Walley has proposed IDM, which considers the set of all possible Dirichlet distributions in the simplex $\Delta_m(p)$:

$$\left\{ \frac{\Gamma(s)}{\prod_{i=1}^{m} \Gamma(st_i)} \prod_{i=1}^{m} \theta_i^{st_i - 1} : \; t_i > 0, \; \sum_{i=1}^{m} t_i = 1 \right\}. \quad (2)$$

For a fixed value s, this is the set of all Dirichlet distributions obtained by letting $(t_1, \dots, t_m)$ to freely vary in $\Delta_m(t)$. Walley has proven that IDM is a model of prior "near-ignorance" in the sense that it provides vacuous prior inferences for the probabilities $P(Y = y_i)$ for $i = 1, \dots, m$. In fact, since $P(Y = y_i) = E[\theta_i] = t_i$, and t_i is free to vary in $\Delta_m(t)$, this means that $P(Y = y_i)$ is vacuous, which implies:

$$\underline{E}[\theta_i] = 0, \quad \overline{E}[\theta_i] = 1. \quad (3)$$

This is a condition of prior ignorance on the mean. Thus, this means that the prior mean of θ_i is unknown, but this does not hold for all functions of $\theta_1, \dots, \theta_m$, for example

$$\underline{E}[\theta_i\theta_j] = 0, \quad \overline{E}[\theta_i\theta_j] = \frac{1}{4}\frac{s}{s+1}, \quad (4)$$

while a prior ignorance model for $\theta_i\theta_j$ would have upper expectation equal to $1/4$. Walley has shown that prior ignorance can only be imposed on a subset of the possible functions of $\theta_1, \dots, \theta_m$ otherwise it produces vacuous posterior inferences [6, Ch. 5], which means that we do not learn from data (for this reason the model is called near-ignorance). However, near-ignorance guarantees prior ignorance for many of the inferences of interest in parametric estimation (e.g., mean, cumulative distribution etc.) and, at the same time, allows to learn from data and converges to the "truth" (be consistent in the terminology of Bayesian asymptotic analysis) at the increase of the number of observations.[4] Walley [9] has also proven that, besides near-ignorance, IDM satisfies several other desiderata for a model of prior ignorance.

Symmetry principle (SP): *if we are ignorant a priori about θ_i, then we have no reason to favour one possible outcome of Y to another, and therefore our probability model on Y should be symmetric.*

Embedding principle (EP): *for each event $A \subset \mathcal{Y}$, the probability assigned to A should not depend on the possibility space $\mathcal{Y}$ in which A is embedded. In particular, the probability assigned a priori to the event A should be invariant w.r.t. refinements and coarsenings of $\mathcal{Y}$.*

[2] Note that the only model that satisfies prior ignorance w.r.t. all the functions g is the set of all probability distributions. Since this set is equivalent to the set of Dirac's deltas (i.e., they produce the same lower and upper expectations), the set of posteriors is again the set of all Dirac's deltas and, thus, we do not learn from data.

[3] Commonly the Dirichlet parameters are denoted as α_i, which are equal to st_i in our notation.

[4] A full model of prior ignorance cannot learn from data [9] as explained in the introduction.

Near-ignorance, SP and EP hold for any model on the simplex which satisfies $E[\theta_i] = t_i$ for $i = 1, \ldots, m$ with $(t_1, \ldots, t_m)$ are free to vary in $\Delta_m(t)$ [9]. In fact, since $P(Y = Y_i) = E[\theta_i] = t_i$, this implies that the probability of the event A is:

$$P(A) = \sum_{y_i \in A} P(Y = y_i) = \sum_{i:\ y_i \in A} t_i,$$

and since t_i are free to vary in $\Delta_m(t)$, it follows that $\underline{P}(A) = 0$ and $\overline{P}(A) = 1$, i.e., the lower and upper probabilities of the event A do not depend on $\mathcal{Y}$. The uniform distribution satisfies SP but not EP since $P(Y = y_i) = 1/|\mathcal{Y}|$, where $|\mathcal{Y}|$ denotes the cardinality of $|\mathcal{Y}|$.[5]

Representation Invariance Principle (RIP): for each event $A \subset \mathcal{Y}$, the posterior inferences of A should be invariant w.r.t. refinements and coarsenings of $\mathcal{Y}$.

The posterior mean of θ_i relative to a Multinomial-Dirichlet conjugate model are:

$$E[\theta_i | n_1, \ldots, n_m] = \frac{n_i + st_i}{N + s}, \tag{5}$$

where n_i is the number of observations for the i-th category and $N = \sum_{i=1}^{m} n_i$. Hence, the lower and upper posterior mean derived from IDM can simply be obtained by

$$\begin{array}{cccc} \frac{n_i+st_i}{N+s} & \overset{t_i \to 0}{=} & \frac{n_i}{N+s} & = & \underline{E}[\theta_i | n_1, \ldots, n_m], \\ \frac{n_i+st_i}{N+s} & \overset{t_i \to 1}{=} & \frac{n_i+s}{N+s} & = & \overline{E}[\theta_i | n_1, \ldots, n_m]. \end{array} \tag{6}$$

Hence, since

$$P(A|data) = \sum_{y_i \in A} P(Y = y_i|data) = \sum_{i:\ y_i \in A} \frac{n_i + st_i}{N + s},$$

the lower $\underline{P}(A|data) = \sum_{i:\ y_i \in A} n_i/(N + s)$ and upper $\overline{P}(A|data) = \sum_{i:\ y_i \in A} (n_i + s)/(N + s)$ posterior probability of A do not depend on refinements and coarsenings of $\mathcal{Y}$. IDM thus satisfies RIP. Observe that the uniform distribution (i.e., $s = |\mathcal{Y}|$ and $t_i = 1/|\mathcal{Y}|$) does not, since $P(A|data) = \sum_{i:\ y_i \in A} (n_i + 1)/(N + |\mathcal{Y}|)$. In general, RIP holds if the lower and upper posterior expectations of the event A do not depend on the number of categories m [9].

Learning/Convergence Principle (LCP): for each event $A \subset \mathcal{Y}$, there exists $\overline{N}$ such that for $N \geq \overline{N}$ the posterior inferences about A should not be vacuous. Moreover, for $N \to \infty$, the posterior inferences should converge to $\lim_{N \to \infty} n_A/N$, where n_A is the number of occurrences of the event A in the N observations [10].

IDM satisfies learning and convergence because

$$P(A|data) = \sum_{y_i \in A} P(Y = y_i|data) = \sum_{i:\ y_i \in A} \frac{n_i + st_i}{N + s} \to \frac{n_i}{N}$$

for $N \to \infty$ and, thus, the effect of the prior vanishes at the increase of the number of observations.

Observe that IDM satisfies all the above principles and also the coherence (CP) and likelihood (LP) principles [8], [11]. Another important characteristic of the IDM is its computational tractability, which follows by the conjugacy between the categorical (multinomial) and Dirichlet distributions for i.i.d. observations.

B. Contribution of the paper

Consider now the case in which we have categorical observations $\mathbf{y}_i$ for each group (population) $i = 1, \ldots, p$, i.e., each $\mathbf{y}_i$ is a vector, i.e., $\mathbf{y}_i = (y_{i1}, \ldots, y_{im})$ being m the number of categories and where $y_{ij} = 1$ if the j-th category is observed in the group i and 0 otherwise. We assume that we have a sample of size n_i of independent and identically distributed outcomes of $\mathbf{y}_i$ for each group (population) $i = 1, \ldots, p$ and a total of $N = \sum_{i=1}^{p} n_i$ observations. In this case the population parameter $\boldsymbol{\theta}_i$ is also a vector $\boldsymbol{\theta}_i = (\theta_{i1}, \ldots, \theta_{im})$ of category probabilities which satisfies $\sum_j \theta_{ij} = 1$. We assume that $\mathbf{y}_i$ is distributed according to a categorical (multinomial) distribution with parameter $\boldsymbol{\theta}_i$, i.e., the likelihood is $M(\mathbf{y}_i|\boldsymbol{\theta}_i)$ a multinomial distribution. By conjugacy we further assume that $\boldsymbol{\theta}_i$ is generated from a Dirichlet distribution $Dir(\boldsymbol{\theta}_i, \alpha_1, \ldots, \alpha_m)$ with parameters $\alpha_j > 0$. The quantity $n_0 = \sum_k \alpha_k$ represents the number of pseudo-observations (prior sample size or prior strength) and $y_{0j} = \alpha_j / \sum_k \alpha_k$ represents the j-th prior pseudo-observation. To complete the hierarchical model we need to specify a prior distribution over the prior parameters $n_0, \mathbf{y}_0 > 0$ with $\mathbf{y}_0 = (y_{01}, \ldots, y_{mj})$. In this paper, we assume that n_0 is fixed and, thus, we place a prior only on $\mathbf{y}_0$. The aim of this paper is to extend IDM to hierarchical models to obtain an Imprecise Hierarchical Dirichlet model (IHDM). In the next sections, we will derive IHDM and we investigate its properties.

II. Imprecise Hierarchical Dirichlet model

The quantity $n_0 = \sum_k \alpha_k$ is assumed to be known thus, to complete the hierarchical model, only a prior model on $\mathbf{y}_0$ must be chosen. By the properties of the Multinomial-Dirichlet model, it follows that [12]

$$\begin{aligned} &\prod_{i=1}^{p} M(\mathbf{y}_i|\boldsymbol{\theta}_i) Dir(\boldsymbol{\theta}_i, n_0 y_{01}, \ldots, n_0 y_{0m}) \\ &\propto \prod_{i=1}^{p} \Gamma(n_0) \prod_{j=1}^{m} \frac{\theta_{ij}^{n_{ij} + n_0 y_{0j} - 1}}{\Gamma(n_0 y_{0j})}, \end{aligned} \tag{7}$$

where n_{ij} is the number of time the category j in the group i has been observed. We can now marginalize $\boldsymbol{\theta}_i$ in (7). From the property of the Multinomial-Dirichlet model, it follows that the resulting PDF is [12]:

$$\begin{aligned} &p(\mathbf{y}_1, \ldots, \mathbf{y}_p|\mathbf{y}_0) \\ &= \int \prod_{i=1}^{p} M(\mathbf{y}_i|\boldsymbol{\theta}_i) Dir(\boldsymbol{\theta}_i, n_0 y_{01}, \ldots, n_0 y_{0m}) d\boldsymbol{\theta}_1 \ldots d\boldsymbol{\theta}_p \\ &\propto \prod_{i=1}^{p} \frac{\prod_{j=1}^{m} \Gamma(n_0 y_{0j} + n_{ij})}{\Gamma(n_0 + n_i)} \frac{\Gamma(n_0)}{\prod_{j=1}^{m} \Gamma(n_0 y_{0j})}. \end{aligned} \tag{8}$$

[5] Consider a bag which includes red and blue balls. In a condition of prior ignorance our probability that the next drawn is red should not depend on the way we organize the possibility space, for instance alternatively as red, light blue and dark blue. If we use the uniform distribution this probability is $1/2$ for the first case and $1/3$ for the second. EP is not satisfied.

By exploiting the following properties of the Gamma function,

$$\begin{aligned}\Gamma(n_0 y_{0j}+n_{ij}) &= \prod_{l=1}^{n_{ij}}(n_0 y_{0j}+l-1)\Gamma(n_0 y_{0j}),\\ \Gamma(n_0+n_i) &= \prod_{l=1}^{n_i}(n_0+l-1)\Gamma(n_0),\end{aligned}$$

then (8) can be rewritten as:

$$p(\mathbf{y}_1,\ldots,\mathbf{y}_p|\mathbf{y}_0) \propto \prod_{i=1}^{p} \frac{\prod_{j=1}^{m}\prod_{l=1}^{n_{ij}}(n_0 y_{0j}+l-1)}{\prod_{l=1}^{n_i}(n_0+l-1)}. \qquad (9)$$

Since $\mathbf{y}_0$ can vary in the simplex $\mathcal{Y}_0$, we adopt a Dirichlet distribution as prior on $\mathbf{y}_0$:

$$Dir(\mathbf{y}_0; st_1,\ldots,st_m) \propto \prod_{j=1}^{m} y_{0j}^{st_j-1}, \qquad (10)$$

where $s>0$ is the hyper-prior number of pseudo-counts and $\mathbf{t}=(t_1,\ldots,t_m)$ the hyper-prior vector of pseudo-observations. When we are in a condition of prior ignorance about the values of $\mathbf{y}_0$, we can choose the parameters s and $\mathbf{t}$ to reflect this state of prior ignorance by letting $\mathbf{t}$ to vary in the simplex

$$\mathcal{T}=\left\{\begin{array}{l} t_j>0, \quad j=1,\ldots,m,\\ \sum_{j=1}^{m} t_j=1,\end{array}\right. \qquad (11)$$

and s to vary in the interval $[\underline{s},\overline{s}]$ with $0<\underline{s}\leq\overline{s}<\infty$. This means that we adopt an Imprecise Dirichlet Model over the vector of parameters $\mathbf{y}_0$, i.e.,

$$\{Dir(\mathbf{y}_0, st_1,\ldots,st_m): \ s\in[\underline{s},\overline{s}], \quad t\in\mathcal{T}\}. \qquad (12)$$

This is a model of ignorance about $\mathbf{y}_0$ since $E[y_{0j}]=t_j$ and thus:

$$\underline{E}[y_{0j}]=0, \quad \overline{E}[y_{0j}]=1, \quad \forall j=1,\ldots,m,$$

where $\underline{E},\overline{E}$ denote the lower and, respectively upper expectation computed w.r.t. the set of priors in (12). Our information about the mean of $\mathbf{y}_0$ is vacuous, in fact we only known that it belongs to $\mathcal{Y}_0$.

Our aim is to derive meaningfull posterior inferences about the unknown $\mathbf{y}_0$ by exploiting the information in the likelihood (9). Let g denote a map from $\mathcal{Y}_0$ to $\mathbb{R}$, we are thus interested to compute

$$\begin{aligned}\underline{E}[g|\mathbf{y}_1,\ldots,\mathbf{y}_p] &= \inf_{\mathbf{t}\in\mathcal{T},\ s\in[\underline{s},\overline{s}]} E[g|\mathbf{y}_1,\ldots,\mathbf{y}_p],\\ \overline{E}[g|\mathbf{y}_1,\ldots,\mathbf{y}_p] &= \sup_{\mathbf{t}\in\mathcal{T},\ s\in[\underline{s},\overline{s}]} E[g|\mathbf{y}_1,\ldots,\mathbf{y}_p]\end{aligned}$$

where $E[g|\mathbf{y}_1,\ldots,\mathbf{y}_p]$ denotes the posterior expectation of g w.r.t. the posterior $p(\mathbf{y}_0|\mathbf{y}_1,\ldots,\mathbf{y}_p)$ obtained by combining (9) and (10) by Bayes' rule. Typical functions of interest are $g(y_0)=y_{0j}$ and $g(y_0)=I_A(y_0)$ where A is some measurable subset of $\mathcal{Y}_0$ and I_A is the indicator function on A. The following lemma provides conditions which ensure that the lower and upper posterior expectation of real-valued functions g of $\mathbf{y}_0$ are not vacuous a-posteriori.

Lemma 1. *The set of posterior PDFs $p(\mathbf{y}_0|\mathbf{y}_1,\ldots,\mathbf{y}_p)$ obtained by combining (9) and (10) by Bayes' rule are all proper provided that*

$$\forall j=1,\ldots,m, \quad \max_{i=1,\ldots,p} n_{ij}\geq 1, \qquad (13)$$

which means that it is necessary that there is at least one observation for each category j. □

Proof: Consider the likelihood (9) and assume without loss of generality that $n_{1j}\geq 1$ for $j=1,\ldots,m$ and $n_{ij}=0$ for $i>1$ and $j=1,\ldots,m$, then (9) becomes:

$$\begin{aligned}p(\mathbf{y}_1,\ldots,\mathbf{y}_p|\mathbf{y}_0) &\propto \frac{\prod_{j=1}^{m}\prod_{l=1}^{n_{1j}}(n_0 y_{0j}+l-1)}{\prod_{l=1}^{n_1}(n_0+l-1)}\\ &= \frac{\prod_{j=1}^{m} n_0 y_{0j}\prod_{l=2}^{n_{1j}}(n_0 y_{0j}+l-1)}{\prod_{l=1}^{n_1}(n_0+l-1)},\end{aligned} \qquad (14)$$

where in the last equality we have exploited the assumption $n_{1j}\geq 1$. By multiplying the likelihood (14) for the prior Dirichlet prior (10), we obtain the joint

$$\begin{aligned}&\frac{\prod_{j=1}^{m} n_0 y_{0j}\prod_{l=2}^{n_{1j}}(n_0 y_{0j}+l-1)}{\prod_{l=1}^{n_1}(n_0+l-1)}\prod_{j=1}^{m} y_{0j}^{st_j-1}\\ &\propto \frac{\prod_{l=2}^{n_{1j}}(n_0 y_{0j}+l-1)}{\prod_{l=1}^{n_1}(n_0+l-1)}\prod_{j=1}^{m} y_{0j}^{st_j+1-1}\\ &= \frac{\prod_{l=2}^{n_{1j}}(n_0 y_{0j}+l-1)}{\prod_{l=1}^{n_1}(n_0+l-1)} Dir(\mathbf{y}_0, st_1+1,\ldots,st_m+1),\end{aligned} \qquad (15)$$

except for a normalization constant. Now $Dir(\mathbf{y}_0, st_1+1,\ldots,st_m+1)$ is always proper for all values of $\mathbf{t}$ in the simplex $\mathcal{T}$. Thus, the set of posteriors PDF of $\mathbf{y}_0$ given the observations includes only proper PDFs under the assumptions of the Theorem. □

This lemma ensures that the learning/convergence principle holds for the IHDM. In fact, assume that more observations are available, then we can use the set of posteriors in Lemma 1 as new set of priors to compute the set of posteriors that accounts for both the new observations and the ones in (13). Then, because of the Bernstein-von Mises Theorem of convergence [13, Sec. 20.1], if we start with a prior that is proper and positive in the parameter space then we converge to the true value of the parameter for $N\rightarrow\infty$.[6]

Lemma 1 gives conditions to guarantee that the lower and upper posterior inferences of real-valued functions g of $\mathbf{y}_0$ are not vacuous. We have derived these conditions analytically. However, we cannot derive a closed expression for these lower and upper posterior inferences, because in

$$p(\mathbf{y}_0|\mathbf{y}_1,\ldots,\mathbf{y}_p) = \frac{\prod_{i=1}^{p}\frac{\prod_{j=1}^{m}\prod_{l=1}^{n_{ij}}(n_0 y_{0j}+l-1)}{\prod_{l=1}^{n_i}(n_0+l-1)} Dir(\mathbf{y}_0, st_1,\ldots,st_m)}{\int \frac{\prod_{i=1}^{p}\prod_{j=1}^{m}\prod_{l=1}^{n_{ij}}(n_0 y_{0j}+l-1)}{\prod_{l=1}^{n_i}(n_0+l-1)} Dir(\mathbf{y}_0, st_1,\ldots,st_m)d\mathbf{y}_0}$$

[6] The likelihood must be satisfy some basic property for the Theorem to hold, i.e., twice differentiable, continuous etc., see [13, Sec. 20.1].

we cannot find an analytical expression for the denominator. However, the following theorem states which are the values of s and $\mathbf{t}$ for which we obtain the lower and upper posterior mean of $\mathbf{y}_0$. We can thus select these values and then numerically solve

$$\int \mathbf{y}_0 p(\mathbf{y}_0|\mathbf{y}_1,\dots,\mathbf{y}_p)d\mathbf{y}_0,$$

to obtain the lower and upper posterior mean.

Theorem 1. *Assume that the condition (13) holds. The lower and upper posterior expectation of y_{0j} are obtained for $s=\overline{s}$, $t_j=0$ and, respectively, $t_j=1$.* □

Proof: This follows by noticing that $g=y_{0j}$ is monotone increasing in $(0,1)$ and the hyperprior $\prod_{j=1}^{m} y_{0j}^{st_j-1}$ puts all the mass in $y_{0j}=0$ for $t_j=0$ and, respectively, $y_{0j}=1$ for $t_j=1$. The mass has its highest value for $s=\overline{s}$. □

A. Estimate of $\boldsymbol{\theta}$

In the previous section, we have derived the lower and upper prior and posterior expectations of $\mathbf{y}_0$. However, we are also interested in computing the lower and upper prior and posterior expectation of the components of $\boldsymbol{\theta}$. For the prior, notice that

$$E[\boldsymbol{\theta}_i|\mathbf{y}_0]=\int \boldsymbol{\theta}_i Dir(\boldsymbol{\theta}_i,n_0y_{01},\dots,n_0y_{0m})d\boldsymbol{\theta}_i=\mathbf{y}_0,$$

this follows from the properties of the Dirichlet distribution and, thus.

$$E[\boldsymbol{\theta}_i]=E[E[\boldsymbol{\theta}_i|\mathbf{y}_0]]=\int \mathbf{y}_0 Dir(\mathbf{y}_0,st_1,\dots,st_m)d\mathbf{y}_0=\mathbf{t}.$$

Then, IHDM is also a model of prior near-ignorance for $\boldsymbol{\theta}_i$, since

$$\underline{E}[\theta_{ij}]=\min t_j=0,\quad \overline{E}[\theta_{ij}]=\max t_j=1.$$

Furthermore, since

$$P[y_{ij}|\boldsymbol{\theta}_i]=\theta_{ij},$$

this follows from the property of the categorical distribution, and $E[\theta_{ij}|\mathbf{y}_0]=y_{0j}$ as observed above, it follows that IHDM also satisfies the Embedding and Symmetry principle defined in Section I-A. In fact, a-priori

$$P[y_{ij}]=E[E[P[y_{ij}|\boldsymbol{\theta}_i]|\mathbf{y}_0]]=t_j,$$

which is free to vary in the simplex.

A-posterior by Bayes' rule, one has that

$$\begin{aligned}&p(\boldsymbol{\theta}|\mathbf{y}_0,\mathbf{y}_1,\dots,\mathbf{y}_p)\\&=\prod_{i=1}^{p} Dir(\boldsymbol{\theta}_i,n_0y_{01}+y_{i1},\dots,n_0y_{0m}+y_{im}),\end{aligned}$$

and, thus,

$$p(\boldsymbol{\theta}|\mathbf{y}_1,\dots,\mathbf{y}_p)=\int p(\boldsymbol{\theta}|\mathbf{y}_0,\mathbf{y}_1,\dots,\mathbf{y}_p)p(\mathbf{y}_0|\mathbf{y}_1,\dots,\mathbf{y}_p)d\mathbf{y}_0,$$

which is the posterior of $\boldsymbol{\theta}$. Since

$$\int \theta_{ij} Dir(\boldsymbol{\theta}_i,n_0y_{01}+y_{i1},\dots,n_0y_{0m}+y_{im})=\frac{n_0y_{0j}+n_{ij}}{n_0+n_i}, \tag{16}$$

we can derive the following result.

Corollary 1. *It holds that*

$$\begin{aligned}&\underline{E}[\theta_{ij}|\mathbf{y}_1,\dots,\mathbf{y}_p]\\&=\underline{E}\left[\frac{n_0y_{0j}+n_{ij}}{n_0+n_i}\middle|\mathbf{y}_1,\dots,\mathbf{y}_p\right]\\&\frac{n_0\underline{E}_{y_0}[y_{0j}|\mathbf{y}_1,\dots,\mathbf{y}_p]+n_{ij}}{n_0+n_i}.\end{aligned} \tag{17}$$

where $\underline{E}[y_{0j}|\mathbf{y}_1,\dots,\mathbf{y}_p]$ is computed in Theorem 1. A similar expression holds for the upper. □

Proof: The result follows straightforwardly from (16). □

From (17) it can be noticed that for $n_i\rightarrow\infty$, $\underline{E}[\theta_{ij}|\mathbf{y}_1,\dots,\mathbf{y}_p]\rightarrow\frac{n_{ij}}{n_i}$ and thus it satisfies the convergence principle defined in Section I-A. It can be shown by a counterexample that RIP does not hold. Consider for instance the case $p=2$, $n_{ij}=1$ for $i=1,2$ and $j=1,2,3$, i.e., $m=3$, in this case $\underline{E}[y_{01}|\mathbf{y}_1,\mathbf{y}_2]\approx 0.295$. If we reduce the number of categories to $m=2$, by putting together the last two categories, i.e., $n_{ij}=1$ for $i=1,2$, $j=1$ and $n_{ij}=2$ for $i=1,2$, $j=2$, we instead obtain $\underline{E}[y_{01}|\mathbf{y}_1,\mathbf{y}_2]\approx 0.354$. The lower means are different. This means that the lower expectation in (17) depends on the number of categories m. This can be also understood by looking at the expression of $p(\mathbf{y}_0|\mathbf{y}_1,\dots,\mathbf{y}_p)$.

III. Differences with the hierarchical Bayesian approach

In the previous section we have seen that, conversely to the Bayesian hierarchical model (based on the uniform prior or any other single prior), the IHDM is a model of prior ignorance. This means that it allows us to start from a very weak information on the parameter of interest. Furthermore, since it satisfies the convergence principle, it allows to converge to the truth at the increase of the evidence. It also satisfies the SP and EP, which cannot both be satisfied by a prior model based on a single density function. Any single prior density contains substantial information about the parameter of interest, because it assigns precise probabilities to hypotheses about this parameter and these have strong implications in the decisions. Conversely, a prior ignorance model starts from a vacuous model, which means that we are not assessing any precise probability to hypotheses. If we are prior ignorant, we should thus use IHDM. In the next section, we will show that the posterior inferences in hierarchical model are very sensitive to the choice of the prior hyperparameter t. This shows, also from a sensitivity analysis point of view, that in case of prior ignorance we should use IHDM instead of a Bayesian hierarchical model based on a single prior density (a precise choice of t).

IV. Application: packet drop rate estimation

Here we consider a simple example of hierarchical model. Consider a collection of sensors observing a single phenomena through noisy measurements. The sensors collaborate by sharing information with a central node (fusion node) that performs further computations about the phenomena of interest. The messages are exchanged through a wireless channel which is subject to random packet losses. Let y_i be a binary variable which is equal to 1 if the fusion node has received any packet from sensor i and zero otherwise.

A probabilistic way to account for packet drops is by modelling the transmission channel as a discrete-time Markov chain with two states, loss (L) and no-loss (N), and transition probabilities $p(L|L) = \theta^a_{i1}$, $p(N|L) = \theta^a_{i2} = 1 - \theta^a_{i1}$ and $p(N|N) = \theta^b_{i1}$, $p(L|N) = 1 - \theta^b_{i1}$. We assume that the transition probabilities are different for each pair fusion node-sensor. However, since the sensors are deployed in a small area, the probabilities θ^a_{i1} for each pair sensor i-fusion node must be in somehow related (same for θ^b_{i1}). We use a hierarchical model to account for this possible dependence.

Assume for simplicity that there are only two sensors and that each sensor-fusion node performs a transmission test to estimate the parameters $\theta^a_{i1}, \theta^b_{i1}$. For the pair sensor 1 - fusion node, the following list reports the result of each transmission test (1 means that the packet was received, 0 that was lost).

$$\begin{aligned} Data_1 = \quad & \{1,1,1,1,1,1,1,1,1,1,\\ & 0,0,1,1,1,1,1,0,0,0,\\ & 1,1,0,0,0,0,0,1,1,1\}, \end{aligned}$$

the data are to be read consecutively. This is the result for sensor 2:

$$\begin{aligned} Data_2 = \quad & \{1,1,1,0,0,1,1,1,1,1,\\ & 0,0,1,1,0,1,0,1,0,0,\\ & 1,1,0,1,1,1,1,1,0,0\}. \end{aligned}$$

Our goal is to estimate $\theta^a_{i1}, \theta^b_{i1}$ for $i = 1, 2$. We assume θ^a_{i1} and θ^b_{i1} are independent for $i = 1, 2$.

We start to estimate θ^a_{i1} for $i = 1, 2$. From the two datasets we derive that

$$n^a_{11} = 7, n^a_{12} = 3, \qquad n^a_{21} = 4, n^a_{22} = 6,$$

where n^a_{11} (n^a_{21}) is the number of transitions from 0 to 0 for sensor 1 (2) and n^a_{12} (n^a_{22}) is the number of transitions from 0 to 1 for sensor 1 (2).

We assume that $n_0 = 1$, $\underline{s} = \overline{s} = 4$ and, thus, our aim is to estimate y_0.[7] Figure 1 shows the two posterior densities of y_0 for the two extreme cases $t = 0$ (blue) and $t = 1$ (red). Observe that the choice of $t \in (0, 1)$ has not a negligible effect on the posterior of y_0. The lower and upper densities of y_0 obtained for $t = 0$ and $t = 1$ are relatively far apart. This can also be noticed from the lower and upper posterior mean of y_0 which are

$$\underline{E}[y^a_{01}|Data_1, Data_2] = 0.28, \quad \overline{E}[y^a_{01}|Data_1, Data_2] = 0.75,$$

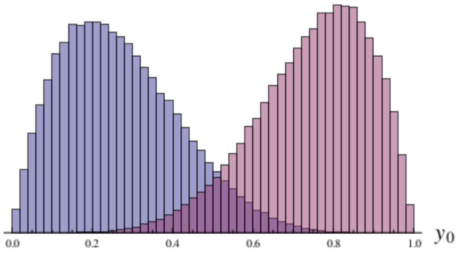

Fig. 1. Histogram of the posterior distribution of y_0 for the two extreme cases $t = 0$ (blue) and $t = 1$ (red). 100000 samples have been generated.

meaning that the posterior mean of y_0 varies in $[0.27, 0.75]$ at the varying of t in $(0, 1)$. We can even compute the robust highest posterior density (HPD) credible interval for y_0, which is defined as the minimum length interval that has lower probability $100(1-\alpha)\%$ of including y_0, i.e.,

$$\min_{b-a:\ [a,b]\subseteq(0,1)} \underline{E}[I_{[a,b]}|Data_1, Data_2] - (1-\alpha) = 0 \quad (18)$$

The 95% robust HPD for y_0 is equal to $[0.04, 0.96]$. Observe that the 95% HPD w.r.t. the prior obtained for $t = 0.5$ and $s = 2$ (a uniform distribution on y_0) is $[0.19, 0.81]$. Thus, in this case, the sensitivity of the posterior inferences to the choice of t is not negligible.

We can also compute the posterior distributions of $\theta^a_{11}, \theta^a_{21}$ for the extreme values $t = 0$ and $t = 1$. By applying Corollary 1, it can be derived that

$$\underline{E}[\theta^a_{11}|Data_1, Data_2] = 0.58, \quad \overline{E}[\theta^a_{11}|Data_1, Data_2] = 0.71,$$

and

$$\underline{E}[\theta^a_{21}|Data_1, Data_2] = 0.36, \quad \overline{E}[\theta^a_{21}|Data_1, Data_2] = 0.50.$$

Note that $E[\theta^a_{i1}|Data_1, Data_2]$ is equal to the posterior probability $P(L|L)$ for sensor i. Thus, for sensor 1, one has that $\underline{P}(L|L) = 0.58$, $\overline{P}(L|L) = 0.71$, which are the lower and upper probabilities of $P(L|L)$.

Similar computations can be performed for $\theta^b_{11}, \theta^b_{21}$ with

$$n^b_{11} = 16, n^b_{12} = 3, \qquad n^b_{21} = 12, n^b_{22} = 14.$$

This time we get

$$\underline{E}[y^a_{01}|Data_1, Data_2] = 0.34, \quad \overline{E}[y^a_{01}|Data_1, Data_2] = 0.8,$$

the two posterior densities of y_0 for the two extreme cases $t = 0$ (blue) and $t = 1$ (red) is shown in Figure 2. We obtain

$$\underline{E}[\theta^b_{11}|Data_1, Data_2] = 0.75, \quad \overline{E}[\theta^b_{11}|Data_1, Data_2] = 0.83,$$

and

$$\underline{E}[\theta^b_{21}|Data_1, Data_2] = 0.74, \quad \overline{E}[\theta^b_{21}|Data_1, Data_2] = 0.84.$$

Hence, we obtain the following matrix of imprecise probabilities for the Markov chain of the channel between sensor 1 and the fusion node:

$$\begin{bmatrix} p(L|L) & p(N|L) \\ p(L|N) & p(N|N) \end{bmatrix} \in \begin{bmatrix} [0.58, 0.71] & [0.39, 0.42] \\ [0.17, 0.25] & [0.75, 0.83] \end{bmatrix}$$

[7] In the binary case, $\mathbf{y}_0$ has only one component.

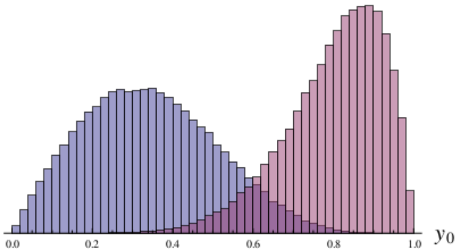

Fig. 2. Histogram of the posterior distribution of y_0 for the two extreme cases $t = 0$ (blue) and $t = 1$ (red). 100000 samples have been generated.

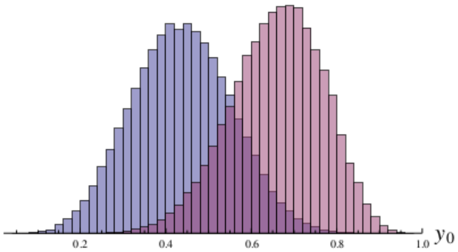

Fig. 3. Histogram of the posterior distribution of y_0 for the two extreme cases $t = 0$ (blue) and $t = 1$ (red). 100000 samples have been generated.

and for the Markov chain of the channel between sensor 2 and the fusion node:

$$\begin{bmatrix} p(L|L) & p(N|L) \\ p(L|N) & p(N|N) \end{bmatrix} \in \begin{bmatrix} [0.36, 0.50] & [0.50, 0.64] \\ [0.16, 0.24] & [0.74, 0.84] \end{bmatrix}$$

Note that $p(N|L) = 1 - p(L|L)$ and $p(L|N) = 1 - p(N|N)$. These matrices lead to an Imprecise Markov Chain, i.e., a Markov chain in which the transition probabilities can vary inside a set of probability measures.

It is interesting to evaluate what happens at the increasing of the number of observations and on the number of sensors. Consider again the estimate of θ^a_{i1} for $i = 1, 2$, but this time assume that

$$n^a_{11} = 70, n^a_{12} = 30, \quad n^a_{21} = 40, n^a_{22} = 60,$$

i.e., we have increased the number of observation but kept the ratio constant, i.e., $n^a_{11}/n^a_{12} = 7/3$ and $n^a_{21}/n^a_{22} = 4/6$. The lower and upper posterior mean of y^a_0 are

$$\underline{E}[y^a_{01}|Data_1, Data_2] = 0.28, \quad \overline{E}[y^a_{01}|Data_1, Data_2] = 0.74.$$

It can be noticed that the increase of the number of observations has almost no effect on the estimate of y^a_0, while it shrinks the imprecision (i.e., the difference between the upper and the lower expectation) of θ^a, i.e.,

$$\underline{E}[\theta^a_{11}|Data_1, Data_2] = 0.68, \quad \overline{E}[\theta^a_{11}|Data_1, Data_2] = 0.70,$$

and

$$\underline{E}[\theta^a_{21}|Data_1, Data_2] = 0.58, \quad \overline{E}[\theta^a_{21}|Data_1, Data_2] = 0.60.$$

Note that the lower and upper expectations almost coincide. To have a similar effect on the estimate of y^a_0 we must increase the number of sensors, for instance

$$\begin{array}{ll} n^a_{11} = 7, n^a_{12} = 3, & n^a_{21} = 4, n^a_{22} = 6, \\ n^a_{31} = 7, n^a_{32} = 3, & n^a_{41} = 7, n^a_{42} = 3, \\ n^a_{51} = 7, n^a_{52} = 3, & n^a_{61} = 7, n^a_{62} = 3. \end{array}$$

Note that the observations of packet drops for the third-sixth channels agree with the ones for the first channel. The lower and upper posterior mean of y^a_0 in this case are

$$\underline{E}[y^a_{01}|Data_1, Data_2] = 0.44, \quad \overline{E}[y^a_{01}|Data_1, Data_2] = 0.66.$$

Thus, the information from the additional four sensors has decreased the imprecision in the estimate of y^a_0. Figure 3 shows the two posterior densities of y_0 for the two extreme cases $t = 0$ (blue) and $t = 1$ (red). Compare it with Figure 1. Summing up, the increase of the number of sensors increases the precision of the estimate of y_0. Instead, the increase of the number of observations for a single pair sensor-fusion node increases the precision of the estimate of θ^a_{i1}.

V. Applications: transition probability estimation in multiple-models

Consider a set of targets that is manoeuvring based on models $m^{(j)} \in \mathcal{M}$ corresponding to the following kinematics

$$\mathbf{x}(t+1) = \mathbf{f}_j(\mathbf{x}(t)) + \mathbf{w}_j(t), \tag{19}$$

where $\mathbf{x} := [x,\ y,\ v,\ h]'$, with x, y Cartesian coordinates of the position, v speed modulus, h heading angle, $\mathbf{w}_j(t)$ zero-mean noise with covariance $\mathbf{Q} = \text{diag}\{0, 0, \sigma^2_v \Delta t, \sigma^2_h \Delta t\}$, with Δt sampling period, and the components of the nonlinear function $\mathbf{f}_j(\mathbf{x}(t))$ are

$$\begin{bmatrix} x(t) + \frac{2v(t)}{\omega_t} \sin\left(\frac{\omega_t \Delta t}{2}\right) \cos\left(h(t) + \frac{\omega_t \Delta t}{2}\right) \\ y(t) + \frac{2v(t)}{\omega_t} \sin\left(\frac{\omega_t \Delta t}{2}\right) \sin\left(h(t) + \frac{\omega_t \Delta t}{2}\right) \\ v(t) \\ h(t) + \omega_t \Delta t \end{bmatrix}, \tag{20}$$

where $\omega_t := \dot{h}(t)$ is the angular speed. This is the *coordinated-turn model* [14]. Accordingly, the model $m^{(j)}$ is completely specified by the value assigned to ω_t. In fact, the inclusion of the angular speed ω_t in the state vector and its estimation are not convenient for short-duration manoeuvres since there is little time for a reliable estimation of ω_t. For $\omega_t = 0$, (19) describes a motion with constant velocity and constant heading (straight motion). Conversely for $\omega_t \neq 0$ it describes a manoeuvre (*turn*) with constant angular speed ω_t, a *left turn* ($\omega_t > 0$) or a *right turn* ($\omega_t < 0$) depending on the sign of ω_t.

Three candidate models are considered, i.e., $\mathcal{M} = \{m^{(-1)}, m^{(0)}, m^{(+1)}\}$ and the angular speed corresponding to $m^{(\pm j)}$ is assumed to be $\omega^{(\pm j)} = \pm j \cdot 0.15\,\text{rad/s}$, for each $j = 0, 1$.

In a Bayesian multiple-model approach [14]–[16], e.g., the Interacting Multiple Model algorithm, to the filtering problem it is necessary to specify the transition probabilities between the various models. We assume that these transition probabilities are unknown and we use past data to estimate them. The following datasets reports for 5 targets the number of times the target were in the model 0 and the number of transition from model 0 to the other 2 models.

	$m^{(-1)}$	$m^{(0)}$	$m^{(1)}$
target 1	3	4	3
target 2	4	7	1
target 3	4	8	4
target 4	3	3	3
target 5	3	6	3

Our goal is to estimate the transition probabilities

$$P_i(-1|0)=\theta_{i1}, P_i(0|0)=\theta_{i2}, P_i(1|0)=\theta_{i3},$$

for all the targets $i=1,\dots,5$.[8] We assume that there is some relationship between the 5 targets, e.g., some aircraft type, some mission task etc., and, thus, we use a hierarchical approach to estimate the above probabilities. We set $n_0=1$, $\underline{s}=\overline{s}=4$ and, thus, our aim is to estimate $\mathbf{y}_0$. For numerical comparison, first we compute the estimate of $\mathbf{y}_0$ using only data of the first three targets. It results that

$$E[y_{01}]\in[.23,.65], E[y_{02}]\in[.35,.73], E[y_{03}]\in[.26,.62],$$

where the extremes of the intervals denote the lower and, respectively, upper expectation. If we use data of all 5 targets, we instead obtain:

$$E[y_{01}]\in[.34,.59], E[y_{02}]\in[.39,.97], E[y_{03}]\in[.32,.58].$$

It can be observed that the intervals are smaller in the second case because of the additional information. By using the above calculations and Corollary 1, we can derive

$$\begin{array}{l}
P_1(-1|0)\in[.31,.38], P_1(0|0)\in[.39,.56], P_1(1|0)\in[.30,.38],\\
P_2(-1|0)\in[.33,.39], P_2(0|0)\in[.53,.68], P_2(1|0)\in[.14,.20],\\
P_3(-1|0)\in[.27,.32], P_3(0|0)\in[.48,.59], P_3(1|0)\in[.26,.32],\\
P_4(-1|0)\in[.33,.41], P_4(0|0)\in[.35,.53], P_4(1|0)\in[.33,.41],\\
P_5(-1|0)\in[.27,.33], P_5(0|0)\in[.47,.62], P_5(1|0)\in[.27,.33],
\end{array}$$

which are the transition probabilities for each target $i=1,\dots,5$. These probabilities can be employed in the algorithm presented [17] for multiple model estimation with imprecise Markov Chains.

VI. Conclusions

This paper has presented a new prior ignorance model for hierarchical estimation with multinomial-Dirichlet distributions. It has been shown how to use the model to derive posterior inferences on the parameters of interest. Furthermore, we have listed the main properties of the model such symmetry, embedding, learning and convergence principles. As future work we plan to extend this work to the case in which also the prior strength n_0 is unknown by using a set of priors which models prior ignorance on n_0. From a practical point of view, we plan to develop an adaptive multiple-model algorithm that uses the proposed approach to estimate in real-time the transition probabilities which are then employed to compute the state of the targets using for instance the Interacting Multiple Model algorithm. In particular, it is interesting to focus on the stability and convergence properties of this adaptive algorithm.

Acknowledgements

This work has been partially supported by the Swiss NSF grant n. 200020-137680/1

References

[1] G. Battistelli, A. Benavoli, and L. Chisci, "State estimation with remote sensors and intermittent transmissions," *Systems and Control Letters*, vol. 61, no. 1, pp. 155 – 164, 2012.

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

[3] P. Huber, "Robust estimation of a location parameter," *The Annals of Mathematical Statistics*, pp. 73–101, 1964.

[4] S. Sivaganesan and J. Berger, "Ranges of posterior measures for priors with unimodal contaminations," *The Annals of Statistics*, pp. 868–889, 1989.

[5] L. De Robertis and J. Hartigan, "Bayesian inference using intervals of measures," *The Annals of Statistics*, vol. 9, no. 2, pp. 235–244, 1981.

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

[7] L. Pericchi and P. Walley, "Robust Bayesian credible intervals and prior ignorance," *International Statistical Review*, pp. 1–23, 1991.

[8] P. Walley, "Inferences from multinomial data: learning about a bag of marbles," *Journal of the Royal Statistical Society. Series B (Methodological)*, vol. 58, no. 1, pp. 3–57, 1996.

[9] P. Walley, "Measures of uncertainty in expert systems," *Artificial Intelligence*, vol. 83, no. 1, pp. 1–58, 1996.

[10] A. Benavoli and M. Zaffalon, "A model of prior ignorance for inferences in the one-parameter exponential family," *Journal of Statistical Planning and Inference*, vol. 142, no. 7, pp. 1960 – 1979, 2012.

[11] J. Bernard, "An introduction to the imprecise Dirichlet model for multinomial data," *Int. Journal of Approximate Reasoning*, pp. 123–150, 2005.

[12] J. M. Bernardo and A. F. M. Smith, *Bayesian Theory*. Chichester: John Wiley & Sons, 1994.

[13] A. DasGupta, *Asymptotic theory of statistics and probability*. Springer, 2008.

[14] Y. Bar-Shalom, X. Li, and T. Kirubarajan, *Estimation with Applications to Tracking and Navigation*. New York: John Wiley & Sons, 2001.

[15] G. A. Ackerson and K. S. Fu, "On state estimation in switching environments," *IEEE Transactions on Automatic Control*, vol. 15, pp. 10–17, 1970.

[16] C. B. Chang and M. Athans, "State estimation for discrete systems with switching parameters," *IEEE Transactions on Aerospace and Electronic Systems*, vol. 14, pp. 418–425, 1978.

[17] A. Antonucci, A. Benavoli, M. Zaffalon, G. de Cooman, and F. Hermans, "Multiple Model Tracking by Imprecise Markov Trees," in *Proc. 12th Int. Conf. Information Fusion*, (Seattle (USA)), pp. 1767–1774, 2009.

[8] The transition probabilities $P_i(-1|-1), P_i(0|-1), P_i(1|-1)$ and $P_i(-1|1), P_i(0|1), P_i(1|1)$ can be estimated in a similar way.