Supra-Bayesian Combination of Probability Distributions

Vladimíra Sečkárová [1,2]

[1] Institute of Information Theory and Automation, Prague, Czech Republic
[2] Charles University in Prague, Faculty of Mathematics and Physics, Department of Probability and Mathematical Statistics, Czech Republic

Abstract

In common life we often need to take every information into account. In this work, handling of different types of given information or knowledge is addressed. The idea of the treatment is to find a suitable model describing the provided knowledge pieces. To successfully solve this task, we use the Supra-Bayesian approach.

I. Introduction - problem formalization

Decision making is an integral, often unrecognized part of our life. If we consider a parametric (population) model, which has one or more unknown parameters, the statistical analysis helps us to gain information from past experience. In order to make the conclusions about them we adopt Bayesian approach where the unknown parameters are treated as random variables.

Bayesian approach used in the parameter estimation is well-elaborated only when the knowledge pieces are given as "ordinary" crisp data ("single values"). No systematic treatment of incompletely compatible knowledge pieces have been given yet. In recently published papers [1] and [2] it is suggested that a Supra-Bayesian approach, see [3], could give a systematic solution. This approach expresses the task of combining the given knowledge pieces as the task of constructing a posterior probability mass function (pmf) or probability density function (pdf) for a fictitious decision maker by using Bayes' theorem. The given knowledge pieces are used as a random data and the ideal merger is estimated as unknown parameter. Both works use this Supra-Bayesian approach, but they differ in relating knowledge pieces to the ideal merger, called "supra-model". Results in these works are promising, but we will not get a Bayesian rule from constructed optimal merger, when "ordinary" data (data values) and parametric model are used.

In this work we try to remove this problem and construct a generally applicable merger for discrete case – the considered sources deal with discrete quantities.

We number the available sources and we focus on the first source. We introduce the task of improving its knowledge. To do this we use the knowledge pieces given by its neighbors. The idea of solving this task consists of two main steps:

1 assume that the first source and all its neighbors provide the knowledge pieces about the same domain and that this knowledge pieces has the form of pmf of the discrete random vector (of the domain).

   The original task of improving the first source's knowledge is now equal to the construction of optimal estimate of pmf based on given pmfs. Which is in fact equal to construction of the optimal merger of given knowledge pieces.

2 solve the situation when other forms of knowledge pieces (also incompletely compatible) are given.

   We focus on how to transform and extend other forms of knowledge pieces into pmfs, so the previously mentioned merger can be used. In the end we project the result back on the domain of each source.

II. The construction of the optimal estimate (of the optimal merger)

Notation used in the text:

- a source: i.e. human being,
- a domain: (discrete) random vector (with finite set of realizations),
- a knowledge piece: information about a part of or whole random vector given by specific source,
- a neighbor: source, domain of which has a nonempty intersection with domain of the first source,
- $\mathbf{X}$ - a (discrete) random vector (domain),
- $\mathbf{x}_1, \ldots, \mathbf{x}_n$ realizations of considered random vector, $n < \infty$,
- $(g_j(\mathbf{x}_1), \ldots, g_j(\mathbf{x}_n)) = g_j$ - vector of probabilities given by j^{th} source, $j = 1, \ldots, s$,
- $D = (g_1^T, \ldots, g_s^T)^T$ matrix of knowledge pieces given by the first source and its $s-1$ neighbors,
- $h = (h(\mathbf{x}_1), \ldots, h(\mathbf{x}_n))$ the unknown merger (unknown vector of probabilities),
- $H = \{h : \sum_{i=1}^{n} h(\mathbf{x}_i) = 1,\ h(\mathbf{x}_i) \geq 0, i = 1, \ldots, n\}$,
- $\widehat{H}$ is a set of all possible decisions (conclusions) about the parameter h, $\widehat{H} \subseteq H$,
- $\mathrm{L}(.,.)$ - loss function (see [4]), $\mathrm{K}(.,.)$ - Kerridge inaccuracy (see [5]),
- ${}^{O}\hat{h}$ the optimal estimate of h (the optimal merger of given knowledge pieces).

The solution of proposed estimation task is found as (see [4]): $\text{Arg}\min_{\hat{h}\in\widehat{H}} \int_H \mathrm{L}(h,\hat{h})\pi(h|D)\mathrm{d}h$.

Furthermore, if we use Kerridge inaccuracy as a loss function (see Subsection 1.), the solution reduces to the following task:

$$\text{Arg}\min_{\hat{h}\in\hat{H}} \mathrm{E}_{\pi(h|D)}[\mathrm{K}(h,\hat{h})|D], \tag{1}$$

where the used notation means conditional expected value with respect to (yet not constructed) pdf $\pi(h|D)$.

Proposition II..1. *Let μ, η be the measures defined on measure spaces $(\times_{k=1}^m \mathcal{X}_k, \otimes_{k=1}^m \mathcal{B}_k)$, $(H,\mathcal{H})$, so that $(\times_{k=1}^m \mathcal{X}_k, \otimes_{k=1}^m \mathcal{B}_k, \mu)$, $(H,\mathcal{H},\eta)$ are σ-finite measure spaces. Under assumption that:*

$$\int_{H\times(\times_{k=1}^m \mathcal{X}_k)} \pi(h(\mathbf{x})|D)h(\mathbf{x})\log\hat{h}(\mathbf{x})\mathrm{d}(\mu\times\eta) < \infty, \tag{2}$$

the solution of task (1) *for $\widehat{H} = H$ has the form:*

$${}^{O}\hat{h} = E_{\pi(h|D)}(h|D). \tag{3}$$

This can be seen by straightforward evaluation, by using Fubini's theorem (see [6]) and properties of Kerridge inaccuracy.

1. Kerridge inaccuracy as a loss function

Assume that probability vector h is given. Then the optimal estimate has to satisfy (according to Bayesian approach):

$${}^{O}\hat{h} \in \text{Arg}\min_{\hat{h}\in\hat{H}} \mathrm{L}(h,\hat{h})$$

and since h is given and $H = \widehat{H}$, we also know that the minimum is reached for ${}^{O}\hat{h} = h$.
For the set of all loss functions reaching the finite minimum for ${}^{O}\hat{h}$ it is shown in [7], that the Kerridge inaccuracy $\mathrm{K}(h,{}^{O}\hat{h}) = -\sum_{i=1}^n h(\mathbf{x}_i)\log {}^{O}\hat{h}(\mathbf{x}_i)$ is a representative of this set of loss functions.

When h is unknown then according to the Bayesian set-up the optimal estimate is found as:

$${}^{O}\hat{h} \in \text{Arg}\min_{\hat{h}\in\hat{H}} \mathrm{E}_{\pi(h|D)}\mathrm{L}(h,\hat{h}),$$

where $\pi(h|D)$ is the posterior pdf of the possible values of $h \in H$. Putting these statements together, we get:

$${}^{O}\hat{h} \in \text{Arg}\min_{\hat{h}\in\hat{H}} \mathrm{E}_{\pi(h|D)}\mathrm{K}(h,\hat{h}). \tag{4}$$

2. Construction of posterior pdf

Since the set of all possible posterior pdfs $\pi(h|D)$ is large, to choose the optimal one we put some additional conditions on the form of $\pi(h|D)$. The considered set will diminish and from the remaining possible posterior pdfs we choose the one with the highest entropy (see [8]). We define the constraints on the posterior pdf:

- j^{th} source takes h as its representative if h is close to the pmf g_j (vector of probabilities) given by j^{th} source,

  meaning the conditional expectation of Kerridge inaccuracy of g_j on h is smaller than or equal to some positive finite value $\beta_j(D)$:

$$\mathrm{E}_{\pi(h|D)}[\mathrm{K}(g_j,h)|D] \le \beta_j(D). \tag{5}$$

From the set of possible posterior pdfs of h satisfying constraints (5) we choose the one with maximum entropy. Which means we are looking for solution of the following optimization task:

$$\text{Arg}\max_{\pi(h|D)\in\mathrm{M}} -\int_H \pi(h|D)\log\pi(h|D)\mathrm{d}h, \tag{6}$$

where $\mathrm{M} = \{\pi(h|D) : \mathrm{E}_{\pi(h|D)}(\mathrm{K}(g_j,h)|D) - \beta_j(D) \le 0,\ j=1,\dots,s,\ \int_H \pi(h|D)\mathrm{d}h - 1 = 0\}$.

Proposition II..2 (Optimal posterior pdf)**.** *Let all constraints in* (6) *be active. Then, the optimal solution of the optimization task* (6) *is:*

$${}^{O}\pi(h|D) = \frac{1}{Z(\lambda_1(D),...,\lambda_s(D))}\prod_{i=1}^n h(\mathbf{x}_i)^{\sum_{j=1}^s \lambda_j(D)g_j(\mathbf{x}_i)}, \tag{7}$$

where $Z(\lambda_1(D),...,\lambda_s(D)) > 0$ *and* $\lambda_j(D) > 0 \quad j=1,\dots,s.$

Main steps of the proof are:

1. equivalently rewrite the task (6) as:

$$\text{Arg}\min_{\pi(h|D)\in \text{M}}\int_H \pi(h|D)\log\pi(h|D)\text{d}h, \tag{8}$$

$\text{M}=\{\pi(h|D):\ \text{E}_{\pi(h|D)}(\text{K}(g_j,h)|D)-\beta_j(D)\le 0,\ j=1,\dots,s,\ \int_H \pi(h|D)\text{d}h-1=0\}$.

2. we assume that there exists

$${}^O\boldsymbol{\lambda}(D)=({}^O\lambda_1(D),\dots,{}^O\lambda_s(D))\in\mathbb{R}^s_+$$

that $({}^O\pi(h|D),{}^O\boldsymbol{\lambda}(D))$ satisfies Global Optimality Conditions (see [9]); then the global minimum of the original task (6) is reached in ${}^O\pi(h|D)$.

3. by using the theory of nonlinear programming (see [9]) and by assuming of the applicability of Fubini's theorem we rewrite the Lagrangian of the task (8) as follows:

$\text{L}(\pi(h|D);\boldsymbol{\lambda}(D))=$

$=\int_H \pi(h|D)\log\pi(h|D)\text{d}h+\lambda_1(D)\left(\text{E}_{\pi(h|D)}(\text{K}(g_1,h)|D)-\beta_1(D)\right)+\dots$

$+\lambda_s(D)\left(\text{E}_{\pi(h|D)}(\text{K}(g_s,h)|D)-\beta_s(D)\right)$

$$=\int_H \pi(h|D)\log\left(\frac{\frac{\pi(h|D)}{\prod_{i=1}^s h(\mathbf{x}_i)^{\sum_{j=1}^s\lambda_j(D)g_j(\mathbf{x}_i)}}}{\frac{1}{Z(\lambda_1(D),\dots,\lambda_s(D))}}\right)\text{d}h-\log Z(\lambda_1(D),\dots,\lambda_s(D))\underbrace{\int_H\pi(h|D)\text{d}h}_{=1}-\sum_{j=1}^n\lambda_j(D)\beta_j(D)$$

$$=\text{D}(\pi(h|D)||{}^O\pi(h|D))-\log Z(\lambda_1(D),\dots,\lambda_s(D))-\sum_{j=1}^n\lambda_j(D)\beta_j(D)$$

4. minimum of the Lagrangian is reached for $\pi(h|D)={}^O\pi(h|D)$ a.e., because the first part $\text{D}(\pi(h|D)||{}^O\pi(h|D))$, which is Kullback-Leibler divergence of $\pi(h|D)$ on ${}^O\pi(h|D)$, is minimal for $\pi(h|D)=^O\pi(h|D)$ a.e. and the remaining part of Lagrangian does not depend on $\pi(h|D)$ and does not influence the minimization.

3. Merging, construction of the optimal estimate

Proposition II..3 (The optimal estimate ${}^O\hat{h}$)**.** *Let us define $\nu_0,\nu_1,\dots,\nu_n$ as:*

$$\nu_i=1+\sum_{j=1}^s\lambda_j(D)g_j(\mathbf{x}_i),\quad i=1,\dots,n,\quad \nu_0=\sum_{i=1}^n\nu_i$$

and the normalizing constant $Z(\lambda_1(D),\dots,\lambda_s(D))$ from the formula (7) *as:*

$$Z(\lambda_1(D),\dots,\lambda_s(D))=\frac{\prod_{i=1}^k\Gamma(\nu_i)}{\Gamma(\nu_0)}.$$

Here, $\lambda_j(D)={}^O\lambda_j(D)>0$, $j=1,\dots,s$, from Proposition II..2. Then, the optimal estimate ${}^O\hat{h}$ of h has the form:

$$\text{E}_{{}^O\pi(h|D)}(h(\mathbf{x}_i)|D)={}^O\hat{h}(\mathbf{x}_i)=\lambda_0^*(D)+\sum_{j=1}^s\lambda_j^*(D)g_j(\mathbf{x}_i), \tag{9}$$

where

$\lambda_0^*(D)=\frac{1}{n+\sum_{j=1}^s\lambda_j(D)}$, $\lambda_j^*(D)=\frac{\lambda_j(D)}{n+\sum_{j=1}^s\lambda_j(D)}$, $n\lambda_0^*(D)+\sum_{j=1}^s\lambda_j^*(D)=1$, $\lambda_j^*(D)>0$, $j=0,\dots,n$.

Main steps of the proof:

1. the optimal posterior pdf has the form:

$${}^O\pi(h|D)=\frac{1}{Z(\lambda_1(D),...,\lambda_s(D))}\prod_{i=1}^n h(\mathbf{x}_i)^{\sum_{j=1}^s\lambda_j(D)g_j(\mathbf{x}_i)}=\frac{1}{\frac{\prod_{i=1}^k\Gamma(\nu_i)}{\Gamma(\nu_0)}}\prod_{i=1}^n h(\mathbf{x}_i)^{\nu_i-1}. \tag{10}$$

Since we know, that: $h(\mathbf{x}_i)\ge 0$ for $i=1\dots,n$, $\sum_{i=1}^n h(\mathbf{x}_i)=1$

and: $\nu_i=1+\sum_{j=1}^s\lambda_j(D)g_j(\mathbf{x}_i)>0$ for $i=1,\dots,n$, because: $\lambda_j(D)>0$ for $j=1,\dots,s$ and $g_j(\mathbf{x}_i)\ge 0$,

then (10) is a pdf of Dirichlet distribution $Dir(h(\mathbf{x}_1),\dots,h(\mathbf{x}_n);\nu_1,\dots,\nu_n)$.

2. by using the properties of Dirichlet distribution we get:

${}^O\hat{h}(\mathbf{x}_i)=\text{E}_{{}^O\pi(h|D)}(h(\mathbf{x}_i)|D)=\frac{\nu_i}{\nu_0}=\frac{1+\sum_{j=1}^s\lambda_j(D)g_j(\mathbf{x}_i)}{\sum_{i=1}^n[1+\sum_{j=1}^s\lambda_j(D)g_j(\mathbf{x}_i)]}=\frac{1}{n+\sum_{j=1}^s\lambda_j(D)}+\sum_{j=1}^s\frac{\lambda_j(D)}{n+\sum_{j=1}^s\lambda_j(D)}g_j(\mathbf{x}_i)$, which is the optimal estimate we are looking for.

III. Extension of the other forms of given information

In previous section we assumed that every source gives the piece of information in the form of the joint pmf of a collection of discrete random variables $\mathbf{X}$ (concretely as a probability vector of possible realizations $\mathbf{x}_1, \ldots, \mathbf{x}_n$). In this section other possible forms of the given knowledge pieces are presented and their transformation into joint pmf of $\mathbf{X}$ (into a probability vector of possible realizations $\mathbf{x}_1, \ldots, \mathbf{x}_n$), useful for merging (see the previous section), is discussed.
Let:

- $\mathbf{P}_j$ denote part of $\mathbf{X}$, which describes the j^{th} source's past experience (random variables, realizations of which describe the random past history for a particular source); $\mathbf{p}$ belongs to the set of possible realizations of $\mathbf{P}_j$,
- $\mathbf{F}_j$ denote a part of $\mathbf{X}$, which describes the j^{th} source's ignorance (random variables with unknown realizations of the future for a particular source); $\mathbf{f}$ belongs to the set of possible realizations of $\mathbf{F}_j$,
- $\mathbf{U}_j$ denote a part of $\mathbf{X}$, that is unconsidered by the j^{th} source; $\mathbf{u}$ belongs to the set of possible realizations of $\mathbf{U}_j$.

Considered forms of knowledge pieces given by the j^{th} source are:

1) moments:
   – conditional moments of $\mathbf{F}_j \subset \mathbf{X}$ on a part $\mathbf{P}_j \subset \mathbf{X}$, $(\mathbf{F}_j \cup \mathbf{P}_j) \subseteq \mathbf{X}$,
   – moments of $\mathbf{P}_j \subseteq \mathbf{X}$
2) a concrete realization (value) of $\mathbf{F}_j \subset \mathbf{X}$ on a part $\mathbf{P}_j \subset \mathbf{X}$, $(\mathbf{F}_j \cup \mathbf{P}_j) \subseteq \mathbf{X}$, or
   a concrete realization of $\mathbf{P}_j \subseteq \mathbf{X}$,
3) conditional pmf (in the form of probability vector) of $\mathbf{F}_j$ on $\mathbf{P}_j$, where $(\mathbf{F}_j \cup \mathbf{P}_j) \subseteq \mathbf{X}$, denoted by $g_j(\mathbf{f}|\mathbf{p})$
4) joint pmf (in the form of probability vector) of $\mathbf{P}_j \subset \mathbf{X}$ (marginal pmf of $\mathbf{X}$), denoted by $g_j(\mathbf{p})$

Since the aim of this section is to construct the joint pmf of $\mathbf{X}$, we need to transform type 1) and 2) of given knowledge pieces into probabilistic terms.

1. *Moments given*

Possible types of moments, the j^{th} source can provide, are:

- conditional moments of $\mathbf{F}_j \subset \mathbf{X}$ on a part $\mathbf{P}_j \subset \mathbf{X}$, $(\mathbf{F}_j \cup \mathbf{P}_j) \subseteq \mathbf{X}$, denoted by:

$$\mathrm{E}_{g_j(\mathbf{f}|\mathbf{p})}(\phi(\mathbf{F}_j, \mathbf{P}_j)|\mathbf{P}_j) = \psi(\mathbf{P}_j), \tag{11}$$

  where ϕ, ψ are functions specified by the source and the expectation is taken with respect to a, yet unspecified, pmf $g_j(\mathbf{f}|\mathbf{p})$, existence of which is assumed.
- moments of $\mathbf{P}_j \subseteq \mathbf{X}$, denoted by :

$$\mathrm{E}_{g_j(\mathbf{p})}(\phi(\mathbf{P}_j)) = \psi, \tag{12}$$

  where ϕ and ψ are a function and a value specified by the source and the expectation is taken with respect to a, yet unspecified, pmf $g_j(\mathbf{p})$, existence of which is assumed.

For a further treatment, we transform this type of knowledge pieces into probabilistic terms – probabilities of outcomes of random variables considered by j^{th} source: we focus on construction of $g_j(\mathbf{f}|\mathbf{p})$ or $g_j(\mathbf{p})$.
If j^{th} source gives the conditional moments (11), the idea for construction of $g_j(\mathbf{f}|\mathbf{p})$ is:

- from the set of all possible conditional pmfs of $\mathbf{F}_j$ conditioned on $\mathbf{P}_j$ (existence of them is assumed) construct a set of conditional pmfs satisfying (11): $\{g_j^*(\mathbf{f}|\mathbf{p})\}$
- and from $\{g_j^*(\mathbf{f}|\mathbf{p})\}$ choose the conditional pmf with the maximum entropy, it means choose the pmf for which holds: $g_j(\mathbf{f}|\mathbf{p}) = \text{ Arg } \max_{\{g_j^*(\mathbf{f}|\mathbf{p})\}} -\sum_{(\mathbf{f},\mathbf{p})} g_j^*(\mathbf{f}|\mathbf{p}) \log g_j^*(\mathbf{f}|\mathbf{p})$

By applying the same idea on the case, when the j^{th} source gives the moments (12), we get $g_j(\mathbf{p})$.

2. *Ordinary data given*

In this section, the knowledge pieces, the j^{th} source can provide, are:

- a realization of $\mathbf{F}_j$ conditioned on $\mathbf{P}_j$, where $(\mathbf{F}_j \cup \mathbf{P}_j) \subseteq \mathbf{X}$ is denoted by $\underline{(\mathbf{f}, \mathbf{p})}$
- realization of $\mathbf{P}_j \subseteq \mathbf{X}$ is denoted by $\underline{\mathbf{p}}$

Again we try to express this type of given knowledge pieces in probabilistic terms – the pmf of random variables considered by the j^{th} source.

To do this we use the measure concentrated on one point – Kronecker delta: $\delta_{i,j}^K \begin{array}{ll} = 1 & \text{if } i = j \\ = 0 & \text{otherwise.} \end{array}$

In case, where $\underline{(\mathbf{f}, \mathbf{p})}$ is given, we define $g_j(\mathbf{f}|\mathbf{p})$ as $\delta^K_{(\mathbf{f},\mathbf{p}),\underline{(\mathbf{f},\mathbf{p})}}$:

$$g_j(\mathbf{f}|\mathbf{p}) \begin{array}{ll} = 1 & \text{if } (\mathbf{f}, \mathbf{p}) = \underline{(\mathbf{f}, \mathbf{p})} \\ = 0 & \text{otherwise.} \end{array}$$

The $g_j(\mathbf{f}|\mathbf{p})$ is a pmf since it satisfies:
$\sum_{\mathbf{p}} g_j(\mathbf{f}|\mathbf{p}) = \sum_{(\mathbf{f},\mathbf{p})} \delta^K_{(\mathbf{f},\mathbf{p}),\underline{(\mathbf{f},\mathbf{p})}} = 1$ and $g_j(\mathbf{f}|\mathbf{p}) \geq 0$ for all possible realizations $(\mathbf{f}, \mathbf{p})$.
In case, when $\underline{\mathbf{p}}$ is given, we use the same idea and we define $g_j(\mathbf{p})$ as $\delta^K_{\mathbf{p},\underline{\mathbf{p}}}$.

3. Extension of unified data

Since all given knowledge pieces have now the form of pmfs of random variables considered by a particular source: $g_j(\mathbf{f}|\mathbf{p})$ or $g_j(\mathbf{p})$, we can focus on their extension into a joint pmf of $\mathbf{X}$ denoted by ${}^e g_j$, further in text called extension ${}^e g_j$. Under the following assumptions:

- we consider the unknown pmf h of $\mathbf{X}$ as a random probability vector,
- $\mathbf{p}_i$/$\mathbf{f}_i$/$\mathbf{u}_i$ denote the possible realizations of $\mathbf{p}$/$\mathbf{f}$/$\mathbf{u}$, which are parts of $\mathbf{x}_i$: $\mathbf{x}_i = (\mathbf{u}_i, \mathbf{f}_i, \mathbf{p}_i)$, $i = 1, \ldots, n$,
- $\{\{g_j(\mathbf{f}_i|\mathbf{p}_i)$ or $g_j(\mathbf{p}_i)\}_{j=1,\ldots,s}\}_{i=1,\ldots,n}$ is $(s \times n)$ matrix, where $g_j(\mathbf{f}_i|\mathbf{p}_i)$, $g_j(\mathbf{p}_i)$ are random variables, for which:
  $g_j(\mathbf{f}_i|\mathbf{p}_i) \geq 0, \quad g_j(\mathbf{p}_i) \geq 0 \quad$ for $j = 1, \ldots, s$, $i = 1, \ldots, n$,
  $\sum_{i=1}^n g_j(\mathbf{f}_i|\mathbf{p}_i) = 1, \quad \sum_{i=1}^n g_j(\mathbf{p}_i) = 1 \quad$ for $j = 1, \ldots, s$,
- $(s \times n)$ matrix D is a realization of the above matrix,

we introduce the following constraints:

1. the first and intuitively clear assumption on the extension ${}^e g_j$ is: the projection of ${}^e g_j$ on the j^{th} source's domain – ${}^e g_j(\mathbf{f}|\mathbf{p})$ – coincides with $g_j(\mathbf{f}|\mathbf{p})$.
2. the extension ${}^e g_j$ is to be as close as possible to the unknown pmf h (see the beginning of the Section II. - sources provide knowledge pieces about $\mathbf{X}$ in the form of joint pmf, where $\mathbf{X}$ is described by the unknown pmf h). In terms of Bayesian decision theory h is the unknown multivariate random parameter taking values in H. We want ${}^e g_j$ to be the minimizer of $\mathrm{E}_{\pi(h|D)}[\mathrm{K}(h, {}^e g_j^*)|D]$, where ${}^e g_j^*$ belongs to a set of all possible pmfs satisfying the constraint 1. denoted by $\{{}^e g_j^*\}$.

   This requirement means, under assumption of applicability of Fubini's theorem, that:

   ${}^e g_j = \mathrm{Arg\,min}_{\{{}^e g_j^*\}} \mathrm{E}_{\pi(h|D)} \left(\mathrm{K}(h, {}^e g_j^*)\right) = \mathrm{Arg\,min}_{\{{}^e g_j^*\}} \mathrm{K}(\mathrm{E}_{\pi(h|D)}(h|D), {}^e g_j^*)$,

   where the global minimum is reached for ${}^e g_j = \mathrm{E}_{\pi(h|D)}(h|D)$, see [5]. In the previous section, it is denoted by ${}^O\hat{h}$ (i.e. see Proposition II..1).
3. the last natural assumption, we already used in previous step, is that ${}^e g_j$ uses all elements of D.

The extensions of unified knowledge pieces are discussed in following sections.

4. Conditional probabilities on a part of random vector

Proposition III..1. *Let the conditional pmf $g_j(\mathbf{f}|\mathbf{p})$ of $\mathbf{F}_j$ on $\mathbf{P}_j$, $(\mathbf{F}_j \cup \mathbf{P}_j) \subset \mathbf{X}$, be given. Then under the assumption that*

$$ {}^O\hat{h} = E_{\pi(h|D)}(h|D) $$

the pmf ${}^e g_j$, represented by a probability vector $({}^e g_j(\mathbf{x}_1), \ldots, {}^e g_j(\mathbf{x}_n))$ with:

$$ {}^e g_j(\mathbf{x}_i) = {}^O\, \hat{h}(\mathbf{u}_i|\mathbf{f}_i, \mathbf{p}_i) g_j(\mathbf{f}_i|\mathbf{p}_i) {}^O\hat{h}(\mathbf{p}_i), \quad i = 1, \ldots, n, \tag{13} $$

is the unique extension of $g_j(\mathbf{f}|\mathbf{p})$ meeting the previously mentioned constraints 1., 2., 3..

Proof. In the proof, the following definition of conditional probability is used.

The conditional probability of events A_1, A_2, A_3 (under the assumption that $\mathrm{P}(A_2, A_3) > 0$ and $\mathrm{P}(A_3) > 0$) is:

$\mathrm{P}(A_1|A_2, A_3) = \frac{\mathrm{P}(A_1, A_2, A_3)}{\mathrm{P}(A_2, A_3)}, \quad \mathrm{P}(A_2|A_3) = \frac{\mathrm{P}(A_2, A_3)}{\mathrm{P}(A_3)}$.

The probability of (A_1, A_2, A_3) is then: $\mathrm{P}(A_1, A_2, A_3) = \mathrm{P}(A_1|A_2, A_3)\mathrm{P}(A_2|A_3)\mathrm{P}(A_3)$.

Since the projection of ${}^e g_j$ on the j^{th} source's domain is ${}^e g_j(\mathbf{f}|\mathbf{p}) = g_j(\mathbf{f}|\mathbf{p})$, the constraint 1. is satisfied.

If we realize that:

$\sum_{i=1}^n h(\mathbf{x}_i) \log {}^e g_j(\mathbf{x}_i) = \sum_{i=1}^n h(\mathbf{u}_i, \mathbf{f}_i, \mathbf{p}_i) \log {}^e g_j(\mathbf{u}_i, \mathbf{f}_i, \mathbf{p}_i) = \sum_{\mathbf{u}} \sum_{\mathbf{f}} \sum_{\mathbf{p}} h(\mathbf{u}, \mathbf{f}, \mathbf{p}) \log {}^e g_j(\mathbf{u}, \mathbf{f}, \mathbf{p})$

then by assuming of applicability of Fubini's theorem, we can rewrite the task stated in the constraint 2. as follows. By inserting proposed ${}^e g_j$ into the minimized expected Kerridge inaccuracy, we get:

$$
\begin{aligned}
\mathrm{EK}(h, {}^e g_j) &= -\int_H \pi(h|D) \sum_{i=1}^n h(\mathbf{x}_i) \log {}^e g_j(\mathbf{x}_i) \mathrm{d}h \\
&= -\int_H \pi(h|D) \sum_{\mathbf{u}} \sum_{\mathbf{f}} \sum_{\mathbf{p}} h(\mathbf{u}, \mathbf{f}, \mathbf{p}) \times \log \left({}^O\hat{h}(\mathbf{u}|\mathbf{f}, \mathbf{p}) g_j(\mathbf{f}|\mathbf{p}) {}^O\hat{h}(\mathbf{p}) \right) \mathrm{d}h \\
&= -\sum_{\mathbf{f}} \sum_{\mathbf{p}} \int_H \pi(h(\mathbf{u}, \mathbf{f}, \mathbf{p})|D) h(\mathbf{f}|\mathbf{p}) \mathrm{d}h \times \left(\sum_{\mathbf{u}} \int_H \pi(h(\mathbf{u}, \mathbf{f}, \mathbf{p})|D) h(\mathbf{u}|\mathbf{f}, \mathbf{p}) \log {}^O\hat{h}(\mathbf{u}|\mathbf{f}, \mathbf{p}) \mathrm{d}h \right) \\
&- \sum_{\mathbf{u}} \sum_{\mathbf{p}} \int_H \pi(h(\mathbf{u}, \mathbf{f}, \mathbf{p})|D) h(\mathbf{u}|\mathbf{f}, \mathbf{p}) h(\mathbf{p}) \mathrm{d}h \times \left(\sum_{\mathbf{f}} \int_H \pi(h(\mathbf{u}, \mathbf{f}, \mathbf{p})|D) h(\mathbf{f}|\mathbf{p}) \log g_j(\mathbf{f}|\mathbf{p}) \mathrm{d}h \right) \\
&- \sum_{\mathbf{u}} \sum_{\mathbf{f}} \int_H \pi(h(\mathbf{u}, \mathbf{f}, \mathbf{p})|D) h(\mathbf{u}|\mathbf{f}, \mathbf{p}) h(\mathbf{f}|\mathbf{p}) \mathrm{d}h \times \left(\sum_{\mathbf{p}} \int_H \pi(h(\mathbf{u}, \mathbf{f}, \mathbf{p})|D) h(\mathbf{p}) \log {}^O\hat{h}(\mathbf{p}) \mathrm{d}h \right).
\end{aligned}
$$

The second term can not be influenced by the choice of ${}^{O}\hat{h}$, since it does not involve marginal or conditional version of ${}^{O}\hat{h}$. The expressions in the brackets () in the first and third term are conditional versions of the Kerridge inaccuracy, which are, for an arbitrary condition, uniquely minimized for:

$(h(\mathbf{u}_1|\mathbf{f}_1,\mathbf{p}_1),\ldots,h(\mathbf{u}_n|\mathbf{f}_n,\mathbf{p}_n)) = ({}^{O}\hat{h}(\mathbf{u}_1|\mathbf{f}_1,\mathbf{p}_1),\ldots,{}^{O}\hat{h}(\mathbf{u}_n|\mathbf{f}_n,\mathbf{p}_n)) = ({}^{e}g_j(\mathbf{u}_1|\mathbf{f}_1,\mathbf{p}_1),\ldots,{}^{e}g_j(\mathbf{u}_n|\mathbf{f}_n,\mathbf{p}_n))$

and

$(h(\mathbf{p}_1),\ldots,h(\mathbf{p}_n)) = ({}^{O}\hat{h}(\mathbf{p}_1),\ldots,{}^{O}\hat{h}(\mathbf{p}_n)) = ({}^{e}g_j(\mathbf{p}_1),\ldots,{}^{e}g_j(\mathbf{p}_n))$.

Since the estimate ${}^{O}\hat{h}$ is using all knowledge pieces in D (see Section 3.), the constraint 3. is satisfied. □

5. Conditional probabilities on the whole set of random variables

When $g_j(\mathbf{f}|\mathbf{p})$ of $\mathbf{F}_j$ on $\mathbf{P}_j$, $(\mathbf{F}_j \cup \mathbf{P}_j) = \mathbf{X}$, is given, then under assumption ${}^{O}\hat{h} = \mathrm{E}_{\pi(h|D)}(h|D)$, the unique extension ${}^{e}g_j$, meeting previously mentioned constraints 1., 2., 3., has the form:

$$ {}^{e}g_j(\mathbf{x}_i) = g_j(\mathbf{f}_i|\mathbf{p}_i)\,{}^{O}\hat{h}(\mathbf{p}_i), \qquad i = 1,\ldots,n. \tag{14} $$

6. Marginal pmf of random vector

When $g_j(\mathbf{p})$ of $\mathbf{P}_j \subset \mathbf{X}$ is given, then under assumption that ${}^{O}\hat{h} = \mathrm{E}_{\pi(h|D)}(h|D)$ the unique extension ${}^{e}g_j$, meeting the previously mentioned constraints 1., 2., 3.., has the form:

$$ {}^{e}g_j(\mathbf{x}_i) = {}^{O}\hat{h}(\mathbf{u}_i|\mathbf{p}_i)g_j(\mathbf{p}_i), \qquad i = 1,\ldots,n. \tag{15} $$

7. Optimal merger for each considered case

The optimal merger ${}^{O}\hat{h} = ({}^{O}\hat{h}(\mathbf{x}_1),\ldots,{}^{O}\hat{h}(\mathbf{x}_n))$ derived in Subsection 3. has the following forms:

- for the extension constructed in Subsection 4.:
  ${}^{O}\hat{h}(\mathbf{x}_i) = \lambda_0^*(D) + \sum_{j=1}^{s}\lambda_j^*(D)\,{}^{O}\hat{h}(\mathbf{u}_i|\mathbf{f}_i,\mathbf{p}_i)g_j(\mathbf{f}_i|\mathbf{p}_i)\,{}^{O}\hat{h}(\mathbf{p}_i)$, for $i = 1,\ldots,n$,
- for the extension constructed in Subsection 5.:
  ${}^{O}\hat{h}(\mathbf{x}_i) = \lambda_0^*(D) + \sum_{j=1}^{s}\lambda_j^*(D)\,{}^{e}g_j(\mathbf{x}_i) = \lambda_0^*(D) + \sum_{j=1}^{s}\lambda_j^*(D)g_j(\mathbf{f}_i|\mathbf{p}_i)\,{}^{O}\hat{h}(\mathbf{p}_i)$, for $i = 1,\ldots,n$,
- for the extension constructed in Subsection 6.:
  ${}^{O}\hat{h}(\mathbf{x}_i) = \lambda_0^*(D) + \sum_{j=1}^{s}\lambda_j^*(D)\,{}^{e}g_j(\mathbf{x}_i) = \lambda_0^*(D) + \sum_{j=1}^{s}\lambda_j^*(D)\,{}^{O}\hat{h}(\mathbf{u}_i|\mathbf{p}_i)g_j(\mathbf{p}_i)$, for $i = 1,\ldots,n$.

IV. Conclusion - advantages of this method

Method of combining different types of given information (often incompletely compatible) proposed in this paper brings following improvement:

- incompletely compatible probabilistic and also non-probabilistic information is treated,
- unified Bayesian solution,
- scalability: this approach can be applied on every source from the group of sources, which can be extremely large and distributed.

Naturally, we did not discuss many additional questions arising with derivation of the final formula, i.e.:

- unambiguity of the projection of ${}^{O}\hat{h}$ back on source's domain,
- choice of the values $\beta_j(D)$ in (5), $j = 1,\ldots,s$,
- existence and unambiguity of Lagrange multipliers $\lambda_j(D)$, $j = 1,\ldots,n$.

They are definitely topics of a future work.

V. Acknowledgement

This research has been partially supported by GAČR 102/08/0567.

References

[1] Kárný, M and Guy, T.V. and Bodini, A. and Ruggeri, F., "Cooperation via sharing of probabilistic elements," *IJCIStudies*, 1(2):139–162, 2009.

[2] M. Kárný, "Knowledge elicitation via extension of fragmental knowledgepieces," 2009.

[3] C. Genest and J. V. Zidek, "Combining probability distributions: a critique and an annotated bibliography. With comments, and a rejoinder by the authors.," *Stat. Sci.*, 1(1):114–148, 1986.

[4] M. H. DeGroot, *Optimal statistical decisions.*, Wiley-Interscience; Wiley Classics Library. Hoboken, NJ: John Wiley & Sons. xx, 489 p., 1970.

[5] D. Kerridge, "Inaccuracy and inference.," *J. R. Stat. Soc., Ser. B*, 23:184–194, 1961.

[6] G. B. Folland, *Real analysis. Modern techniques and their applications. 2nd ed.*, Pure and Applied Mathematics. A Wiley-Interscience Series of Texts, Monographs, and Tracts. New York, NY: Wiley. xiv, 386 p., 1999.

[7] J. M. Bernardo, "Expected information as ecpected utility.," *Ann. Stat.*, 7:686–690, 1979.

[8] J. E. Shore and R. W. Johnson, "Axiomatic derivation of the principle of maximum entropy and the principle of minimum cross-entropy.," *IEEE Trans. Inf. Theory*, 26:26–37, 1980.

[9] M. S. Bazaraa, H. D. Sherali, and C. M. Shetty, *Nonlinear programming. Theory and algorithms. 2nd ed.*, Hoboken, NJ: John Wiley & Sons. xv, 1993.