Univerzita Karlova v Praze
Matematicko-fyzikální fakulta

DIPLOMOVÁ PRÁCE

Evgeny Kalenkovich

Sdílení pravděpodobnostní informace bayesovských agentů

Katedra pravděpodobnosti a matematické statistiky

Vedoucí diplomové práce: Ing. Miroslav Kárný, DrSc
Studijní program: Matematika, Finanční a pojistná matematika

Charles University in Prague
Faculty of Mathematics and Physics

MASTER'S THESIS

Evgeny Kalenkovich

Sharing of probabilistic information of Bayesian agents

Department of Probability and Mathematical Statistics

Supervisor: Ing. Miroslav Kárný, DrSc
Study Program: Mathematics, Insurance and Financial Mathematics

I would like to express sincere gratitude to my supervisor Ing. Miroslav Kárný, DrSc, for his immense patience throughout my work on this thesis.

This research was partially supported by GACR 102/08/0567 and MSMT CR 2C06001.

I hereby certify that I wrote the thesis myself, using only the referenced sources. I grant to Charles University permission to reproduce and to distribute publicly paper or electronic copies of this thesis and to grant others the right to do so.

Prague, August 6, 2010 Evgeny Kalenkovich

Název práce: Sdílení pravděpodobnostní informace bayesovských agentů
Autor: Evgeny Kalenkovich
Katedra: Katedra pravděpodobnosti a matematické statistiky
Vedoucí diplomové práce: Ing. Miroslav Kárný, DrSc
e-mail vedoucího: `school@utia.cas.cz`

Abstrakt: Potřeba kombinovat pravděpodobnostní rozdělení v mnoha problémech teorie rozhodování. V této práci navazujeme na články [14] a [15], ve kterých se protlačuje supra Bayesovský přístup [9]. Je uvedena metoda pro kombinování konečných diskrétních rozdělení stejně. Taktýž způsob, jak zacházet s neúplnou informací a ohraničenými spojitými rozdělenimi.
V diskrétním případě náš přístup je v duchu, ale liší se v několika klíčových bodech od práce [20]. Výsledkem je posunutý aritmetický průměr vektorů pravděpodobností, což je odlišné od obvyklého (viz [9]) aritmetického průměru.

Klíčová slova: kombinace pravděpodobnostních rozdělení, Bayesovské rozhodování, sdílení pravděpodobnostních informací, supra Bayesovské slučování

Title: Sharing of probabilistic information of Bayesian agents
Author: Evgeny Kalenkovich
Department: Department of Probability and Mathematical Statistics
Supervisor: Ing. Miroslav Kárný, DrSc
Supervisor's e-mail address: `school@utia.cas.cz`

Abstract: A need for combining probability distribution arises in many decision-theoretical problems. In this work we follow articles [14] and [15] in pursuing the supra Bayesian approach [9]. A method for combining finite discrete distributions is introduced, as well as a way to deal with incomplete information and bounded continuous distributions.
In the discrete case our approach is along the lines of, but different at a few key points from the thesis [20]. The result is a shifted arithmetic mean of pmfs, which is discrepant from the usual arithmetic pooling (see [9] for details).

Keywords: Bayesian decision-making, sharing of probabilistic information, combining probability distributions, supra Bayesian merging

Contents

Introduction

Have you heard of a psychological term ‘groupthink’? Probably not. According to [4] it was described by Irving Janis after studying several major US foreign policy failures. A very reasonable question that drove his research was, how could a group of well-informed, well-educated people with well-beyond-average intellectual capacities (namely J.F. Kennedy and his advisers) allow for ‘The Bay of Pigs’ to happen. They may not have been able to anticipate that defeat would bring the whole world to the brink of nuclear war, it *would* have sounded as a stretch. But sending a little over 1000 briefly trained Cuban exiles to overpower Fidel Castro, even with all the preceding military actions must have sounded as a sheer idiocy to anyone with at least common sense. So how could all those people who disposed of much more that common sense and had virtually unlimited information sources come to a decision that was so obviously wrong?

Janis defined groupthink as ’a mode of thinking that people engage in when they are deeply involved in a cohesive in-group, when the members’ strivings for unanimity override their motivation to realistically appraise alternative courses of action.’ In plain English: if a group of people caring about each other is discussing an arbitrary matter and attempting to find a consensus, each member will usually subconsciously suppress their creativity, critical thinking, sometimes even common sense, because each member strives for conformity and is averted to antagonizing the group. Everyone assesses disagreement as too risky to act upon (a member is afraid that disagreement as any deviation from the popular opinion may bring embarrassment, that may be seen as ‘stupid’ and suffer consequences) and keeps silent trying to guess what the consensus. It is important that everyone in the group acts in this way, which will most likely lead to a decision that no one likes, no one finds reasonable, but is not so far off as to make someone speak their mind. That is exactly how you end up in a movie theater with a bunch of friends watching the movie not one of you actually wanted to see. This is how groupthink leads to a false consensus that is agreed upon by everybody, but is consciously bad for everyone.

This vicious circle is easily broken when there is at least one member of the group who is brave or careless enough to speak their mind. If one does, then usually everybody or at least many follow and this ignites an actual discussion. We sigh with relief as we just avoided groupthink, but then watch another problem rise. With a variety of opinions that differ marginally substantially it is almost impossible to reach a consensus. Then we have to either set a strict set of rules that define consensus (e.g. voting system), or pick a leader who will not participate actively in the discussion and will declare a consensus afterwards.

The solution we push forward in this thesis is for Kennedy to assign a distance from each one's opinion to the consensus before the session based on his experience. Then he can either exclude himself from the discussion and then find the best consensus from those meeting the distance assignments, or alternatively he can join the discussion (and be the one to break the groupthink), but in this case he has to assign distance from his opinion to the future consensus beforehand as well. Our analogue of breaking groupthink is assuming (demanding) that everyone shares their opinion.

The idea behind the text above is to give our reader an idea of how complicated and how important is it to study group decision-making. Briefly presented above are some of very well-researched and thus 'easy' problems that we have to face every day and that can have enormous consequences. And nobody actually knows how to solve them. And there are many, many more.

Complexity of the group behavior is vast and limitless. Fortunately as mathematicians we don't have to deal with real people with all their complexity and sophistication. And when we do deal with them we reduce them to a set of number and symbols, which is much easier to analyze and work with. We will assume that opinions we have to analyze are already in that form.

The whole part of decision-making theory (see [7] for more on the topic) is dedicated to so-called multi-agent systems, which represent an interacting group. We assume that there is a finite number of so-called agents that express their 'opinions.' which is any information about the problem, or rather the variables we are interested in, in a numerical form. 'The variables we are interested in' can be weather forecast, stock market models or control systems among others. Our goal is to find the best opinion on the basis of opinions we have.

Easy as it may sound (probably not) it is not an easy problem to solve. There has been a lot of methods tried to solve it and a lot of effort put into it. Most of those works concluded that in case opinions are expressed in a form of probability distributions then under some controversial conditions an arithmetic pooling of those distributions is the way to go. A thorough overview of what had been done was written by Genest and Zidek [9]. Competing with them would be pointless so we will mention only a few articles that are in our opinion have some relevance to the present work.

One of the thread in the research is called knowledge elicitation (see for example [19]), which is well-elaborated, but heavily depends on the skills of 'elicitation expert.' Another important thread would be Bayesian decision networks presented for example in [12] and [6]. Though deeply developed Bayesian network provide very ad hoc solution and not provide a systematic approach. Genest and Zidek in [9] describe what they call the supra Bayesian approach which introduces a so-called supra Bayesian, who is a virtual supra agent that has a model of agents opinions and therefore can apply the Bayesian paradigm to finding a pool of opinions. As this is a virtual agent her model should be constructed by somehow relating agents' opinions and this unknown model.

In more recent works [15],[14] the supra Bayesian approach was pushed forward and fairly self-contained and systematic approach was introduced. Unfortunately both of the mentioned articles suffered from a faulty assumption it the core of the approach. This thesis was at first supposed to be based on those two articles and has therefore deviated from the initial guidelines. Co-author of both of the mentioned articles (and the advisor of this thesis) and his colleagues then suggested a different approach that was elaborated for the discrete case in [20]. This work consists mainly of some deeper elaboration, clearing some shortcomings and extension to the continuous case of the thesis [20].

Below we give the layout of the present thesis.

In the section 1 the basics of the method of opinion pooling are introduced. The method is shown for the case where all agents consider the same variables with the same joint support and describe them providing their joint probability mass functions (hereafter *pmf*s for short).

In the section 2 some of the assumption of the previous section are relaxed. Namely we allow supports to be different,then opinions to be expressed in a form of generalized moments of the considered variables and finally we allow agents to provide only conditional pmfs on only part of the variables.

In the section 3 we treat the case where all agents consider the same variables with the same bounded joint support and describe them providing their joint pdfs.

In the section 4 as its name suggests we give a conclusion, an overview of what has been done, what has not been done and what is yet to be done.

In the appendix A we provide some necessary notions and theorems from the information theory and the calculus of variations that are not covered by the standard course at the faculty of mathematics and physics.

1 Basics of method

In this section we describe the basics of our method for pooling probabilistic models. For future reference we will call this method **supra Bayesian merging.** The method is shown for the simplest case, in which we assume that all agents

(a) share one domain, that is they consider the same set of variables,

(b) consider the same set as the support of their common domain,

(c) consider their common domain to be a finite discrete random vector.

(d) describe behavior of their common domain by means of a joint probability mass function (pmf).

Above and hereafter we use *pmf* as an abbreviation for *probability mass function*,

The following is a step-by-step derivation of the method and is therefore somewhat verbose, more so than was initially planned. While wordiness of this section is perfectly justified as a price paid for a detailed explanation, it does hurt transparency of the text. In order to remove this problem there is a subsection 1.6 containing recapitulation of the section. Feel free to leap right to it and turn a few pages back for details whenever needed. Mentioned subsection has for convenience reasons the same structure as the part of this section preceding it.

1.1 Cooperation structure

We begin by defining a 'Bayesian agent' for it is the cornerstone of the whole cooperation structure.

Definition A **Bayesian agent** is an entity characterized by its **probabilistic information,** which is a pmf

$$q(\boldsymbol{x}), \quad \boldsymbol{x} \in supp \subset \mathbb{R}_K, \text{[1][2]}$$

[1] As there is no unified definition of the support of a measure we emphasize that we use the definition that yiels a unique set as the support. For a measure μ we define its support as $supp(\mu) = \overline{\{\zeta : \exists \text{ an open neigborhood } U \text{ of } \zeta \text{ such that } \mu(U) > 0\}}$. Overline denotes closure of a set.

[2] Symbol $\mathbb{R}_K$ denotes the K-dimensional real space.

of a finite discrete real random vector

$$\boldsymbol{X} = (X_1, X_2, \dots, X_K)$$

with the length $K \in \mathbb{N}$ [3] and the support *supp* of a finite cardinality. Random vector $\boldsymbol{X}$ is called agent's **domain.**

We will use notations $q(\boldsymbol{x}_n)$ and $q_n = q(\boldsymbol{x}_n)$ as interchangeable for all pmfs throughout the work.

This definition of a Bayesian agent satisfies assumptions (c) and (d). Let's now say we have $S \in \mathbb{N}$ Bayesian agents indexed by numbers from 1 to S with a shared domain

$$\boldsymbol{X} = (X_1, \dots, X_K),$$

which has a pmf $q^{(s)}$ according to the s^{th} agent and the support

$$\text{s}upp = \{\boldsymbol{x}_1, \boldsymbol{x}_2, \dots, \boldsymbol{x}_N\}$$

according to any agent.

We will use notations $q(\boldsymbol{x}_n)$ and $q_n = q(\boldsymbol{x}_n)$ as interchangeable for all pmfs throughout the work.

We assume that both S and N are greater or equal to 2. Now we have the remaining assumptions (a) and (b) satisfied. Denote p the true pmf (see A.2 for details on what 'true' means in this context) of $\boldsymbol{X}$. Our goal is to find an estimate $\hat{p}$ of p incorporating all the probabilistic information on our hands.

1.2 Supra Bayesian approach

Bayesian inference is an approach to parameter estimation based mainly on treating parameters as random. Assume we know that a vector parameter $\boldsymbol{\theta}$ from a parametric space $\boldsymbol{\Theta} \in \mathcal{B}_N$ [4] is a random vector with the so-called prior distribution, which has a density function $\pi(\boldsymbol{\theta})$. We then observe a random vector $\boldsymbol{X}$ which has a pdf (or pmf) $f(\boldsymbol{x}|\boldsymbol{\theta})$ under the condition that value of the parameter is $\boldsymbol{\theta}$. Applying Bayes' theorem (formal statement

[3] Symbol $\mathbb{N}$ denotes the set of all positive integers.

[4] Symbol $\mathcal{B}_N$ denotes the set of Borel-measurable subsets of $\mathbb{R}_N$

and proof can be found for example in [1] or any other basic textbook on statistics) we get a posterior pdf after the observation $\boldsymbol{X} = \boldsymbol{x}$ is made:

$$\pi(\boldsymbol{\theta}|\boldsymbol{x}) = \frac{f(\boldsymbol{x}|\boldsymbol{\theta})\pi(\boldsymbol{\theta})}{\int_{\boldsymbol{\Theta}} f(\boldsymbol{x}|\boldsymbol{\theta})\pi(\boldsymbol{\theta})\, d\boldsymbol{\theta}}.$$

Calculated in this way posterior distribution can then be used as an updated prior distribution.

Remark One might have noticed that we allow $\boldsymbol{X}$ to be a discrete random vector, but insist on $\boldsymbol{\theta}$ being continuous. It is uncustomary to presume otherwise unless nature of the parameter in question dictates so. It would be strange to assume that variance of a normally distributed random value may take only discrete values. At the same time parameter of a Binomial distribution (the number of Bernoulli trials) is obviously discrete. Usually if the parameter can be continuous (i.e. at least theoretically can take values in a non-null Borel set) then we assume it is. Something similar is used in the described supra Bayesian approach.

With Bayesian approach we allowed certain values (parameters of a distribution) to be random and thus have their own probability distribution. By taking supra Bayesian approach we make one step further and allow distributions to be random. We treat s^{th} agent's probabilistic information $q^{(s)}$ as a realization of a random vector $\boldsymbol{q}^{(s)}$ and the true distribution p as a realization of a random vector $\boldsymbol{p}$.

Remark As with the Bayesian approach there is a question whether allowing randomness is justified 'philosophically.' A recent critique can be found for example in [23]. The decision-making justification of the Bayesian inference was given by Wald (see [3] for details).

The mathematics might work, but is there any relation to reality? Usual way to answer this in case of Bayesian approach is to use subjective probability: if we don't know something then it's as good as random to us. This plea might be used in the case of supra Bayesian approach, especially since two approaches are essentially the same. Each agent's subjective distribution is indeed subjective and can therefore be treated as random.

Or we could realize that any model, probabilistic included, is derived from certain observations in a broader sense, which include all our obtained knowledge, and thus mentioned model is a random statistic, and thus is a random variable.

Or we can see a subjective distribution as a noisy version of the true distribution, random deviation from which is brought about by both the fact that agent's judgment is imperfect and that she cannot possibly include all the relevant variables while constructing her model.

The true distribution is of course random since we do not know it. One could picture God not only throwing a dice for each and every single atomic event, but also picking a dice from a giant bag full of different dice first. This way, not only the event is random, but its distribution is random as well.

These 'philosophical' kinds of interpretations or justifications of any assumption might be interesting to discuss and helpful in understanding, but they are essentially irrelevant. From the theoretical point of view the math works and it is the only thing that matters. From the practical point of view these interpretations are pretty much bogus: the method either works or not. In either case we cannot know whether some particular assumption was wrong or it was something else and we do not care. We do not know whether gravity exists, we cannot know that, but it makes sense mathematically and seems to work in real-life calculations and that is pretty much all that is important. Obviously there are different opinions on the matter.

The idea is to treat all the probabilistic information we obtained from our agents as an input data. Let's denote it as a data matrix

$$Q = \begin{pmatrix} q^{(1)}(\boldsymbol{x}_1) & \dots & q^{(1)}(\boldsymbol{x}_N) \\ \vdots & \ddots & \vdots \\ q^{(S)}(\boldsymbol{x}_1) & \dots & q^{(S)}(\boldsymbol{x}_N) \end{pmatrix}.$$

Its n^{th} column contains different probabilities that $\boldsymbol{X} = \boldsymbol{x}_n$ as viewed by all S agents. And its s^{th} row containg s^{th} agent's pmf. Denote $\boldsymbol{Q}$ the random counterpart of Q. An estimate $\hat{p}$ is a statistic based on the data matrix $\boldsymbol{Q}$, i.e. a measurable mapping from the set of all possible inputs

$$\begin{aligned} \mathcal{L}(S) = \big\{Q : \quad & Q \in \mathbb{R}_{S\times N},{}^{5} \\ & \wedge\ q_{s,n} \geq 0 \quad s = 1, \dots, S, n = 1, \dots, N, \\ & \wedge\ \sum_{n=1}^{N} q_{s,n} = 1 \quad s = 1, \dots, S\big\} \end{aligned}$$

to the set of all possible 'outputs' (pmfs of $\boldsymbol{X}$)

$$\mathcal{L}(1) = \big\{p : p \in \mathbb{R}_N \wedge p_n \geq 0 \quad n = 1, \dots, N \wedge \sum_{n=1}^{N} p_n = 1\big\}.$$

Denote this statistic as h. We must find h such that $\hat{p} = h(Q)$ is as close to the true distribution $\boldsymbol{p}$ as we can make it. To do so we have to decide on what ‘close’ means and then find an optimal h.

1.3 Optimal estimator of the true distribution in case its posterior distribution is known

It was shown in [17] by using intuitive arguments that the Kullback-Leibler divergence is a right way to measure difference between probability distributions. Later, in [22], the principle of minimum Kullback-Leibler divergence was derived axiomatically for certain cases. In this work we will rather use the Kerridge inaccuracy (see A.6 for the definition and basic properties), which leads to the same results, but has a simpler explicit formula. It is defined as the sum of the Kullback-Leibler divergence from some distribution to the true one and the entropy of the latter. In our case the Kerridge inaccuracy $\mathrm{K}\left(p \,\|\, \hat{p}\right)$ from $\hat{p}$ to p is

$$\begin{aligned}\mathrm{K}\left(p \,\|\, \hat{p}\right) = \mathrm{D}\left(p\|\hat{p}\right) + \mathrm{H}(p) &= \sum_{n=1}^{N} p(\boldsymbol{x}_n) \log \frac{p(\boldsymbol{x}_n)}{\hat{p}(\boldsymbol{x}_n)} - \sum_{n=1}^{N} p(\boldsymbol{x}_n) \log p(\boldsymbol{x}_n) \\ &= -\sum_{n=1}^{N} p(\boldsymbol{x}_n) \log \hat{p}(\boldsymbol{x}_n),\end{aligned}$$

where $\mathrm{D}\left(p\|\hat{p}\right)$ and $\mathrm{H}(p)$ stand for the Kullback-Leibler divergence from $\hat{p}$ to p and the entropy of p respectively. Obviously when the first argument in the Kerridge inaccuracy (or, which is the same, in the Kullback-Leibler divergence) is interpreted as the true distribution it is fixed and so it is absolutely the same whether to choose one or the other when looking for an optimal estimate. The Kerridge inaccuracy has, however, a formula that is somewhat more pleasant to work with and so we will stick with this measure.

We cannot observe the true distribution p and we cannot therefore minimize $\mathrm{K}\left(p \,\|\, \hat{p}\right)$ directly. Instead, we will put a weaker optimality condition on the function h and call it optimal if for any given realization Q of the input data matrix $\boldsymbol{Q}$ it minimizes the expected Kerridge inaccuracy from $\hat{p} = h(\boldsymbol{Q})$ to $\boldsymbol{p}$, in other word we have to find h such that

$$\mathrm{E}[\mathrm{K}\left(\boldsymbol{p} \,\|\, h(\boldsymbol{Q})\right) | \boldsymbol{Q} = Q] = \min_{\tilde{p} \in \mathcal{L}(1)} \mathrm{E}[\mathrm{K}\left(\boldsymbol{p} \,\|\, \tilde{p}\right) | \boldsymbol{Q} = Q].^{6}$$

[6] Symbol E stands for expected value.

We assume that $\boldsymbol{p}$ can possess any value from $\mathcal{L}(1)$, which has infinite cardinality and the Borel measure of zero hence we have to conclude that $\boldsymbol{p}$ is neither discrete nor continuous. Consider a random vector

$$\dot{\boldsymbol{p}} = (\boldsymbol{p}(\boldsymbol{x}_1), \ldots, \boldsymbol{p}(\boldsymbol{x}_{N-1})),$$

which is simply the vector $\boldsymbol{p}$ without the last element. Hereafter any variable with a dot above it denotes its 'undotted' counterpart cut at the last element.

The vector $\dot{\boldsymbol{p}}$ takes values in the set

$$\dot{\mathcal{L}}(1) = \{\dot{\boldsymbol{x}} \,:\, \dot{\boldsymbol{x}} \in \mathbb{R}_{N-1} \,\wedge\, \boldsymbol{x} \in \mathcal{L}(1)\},$$

which is a non-null Borel set. Therefore we can assume that $\dot{\boldsymbol{p}}$ is a continuous variable with conditional pdf $\dot{\pi}_{\dot{\boldsymbol{p}}|Q}(\dot{p})$ given $\boldsymbol{Q} = Q$. Let's additionally assume that $\dot{\pi}^2_{\dot{\boldsymbol{p}}|Q}$ is integrable as well. This is merely a technical assumption, which will later allow us to differentiate the entropy $\mathrm{H}(\dot{\pi}_{\dot{\boldsymbol{p}}|Q})$ with respect to $\dot{\pi}_{\dot{\boldsymbol{p}}|Q}$.

Because working with $N-1$-dimensional vectors is counterintuitive and writing dots everywhere all the time is both handful and distracting we will avoid this unnecessary unrest by accepting the following convention: for each $\dot{g} : \mathbb{R}_{N-1} \to \mathbb{R}$

$$\int\limits_{\mathcal{L}(1)} g(p)\pi_{\boldsymbol{p}|Q}(p)\,dp = \int\limits_{\dot{\mathcal{L}}(1)} \dot{g}(\dot{p})\dot{\pi}_{\dot{\boldsymbol{p}}|Q}(\dot{p}),$$

where $g(\boldsymbol{x}) = \dot{g}(\dot{\boldsymbol{x}})$ for each $\boldsymbol{x} \in \mathcal{L}(1)$. Then we can write

$$\begin{aligned}
&\mathrm{E}[\mathrm{K}\left(\boldsymbol{p} \,\|\, \tilde{p}\right) | \boldsymbol{Q} = Q] \\
&\quad= \int\limits_{\mathcal{L}(1)} \mathrm{K}\left(p \,\|\, \tilde{p}\right)\ \pi_{\boldsymbol{p}|Q}(p)\,dp \\
&\quad= \int\limits_{\mathcal{L}(1)} \left(-\sum_{n=1}^{N} p_n \log \tilde{p}_n\right) \pi_{\boldsymbol{p}|Q}(p)\,dp \\
&\quad= -\sum_{n=1}^{N} \log \tilde{p}_n \left(\int\limits_{\mathcal{L}(1)} p_n\, \pi_{\boldsymbol{p}|Q}(p)\,dp\right) \\
&\quad= -\sum_{n=1}^{N} \log \tilde{p}_n\ \mathrm{E}_{\pi_{\boldsymbol{p}|Q}}(\boldsymbol{p}_n | \boldsymbol{Q} = Q) \\
&\quad= \mathrm{K}\left(\mathrm{E}_{\pi_{\boldsymbol{p}|Q}}(\boldsymbol{p} | \boldsymbol{Q} = Q) \,\|\, \tilde{p}\right).
\end{aligned}$$

Gibbs' inequality (see A.3 for details) states that for any two probability distribution the Kullback-Leibler divergence from one to the other is alway greater or equal to zero with equality if and only if the distributions are equal to each other. As a corollary of this inequality the Kerridge inaccuracy $\mathrm{K}\left(\mathrm{E}_{\pi_{\boldsymbol{p}|Q}}(\boldsymbol{p}|\boldsymbol{Q}=Q) \parallel \tilde{p}\right)$ is minimized for $\tilde{p} \in \mathcal{L}(1)$ at $\tilde{p} = \mathrm{E}_{\pi_{\boldsymbol{p}|Q}}(\boldsymbol{p}|\boldsymbol{Q}=Q)$. Thus, for an optimal estimator function h we can write

$$h(Q) = \hat{p} = \mathrm{E}_{\pi_{\boldsymbol{p}|Q}}(\boldsymbol{p}|\boldsymbol{Q}=Q).$$

Note that the result is the same as the least mean square error estimator. Also note that the result is absolutely useless to us unless we construct the pdf $\pi_{\boldsymbol{p}|Q}$.

1.4 Construction of posterior pdf of the true distribution

To construct the pdf $\pi_{\boldsymbol{p}|Q}$ we will apply the principle of maximum entropy, which says that under given constraints we are to choose the distribution with the highest entropy. Recall that for a continuous distribution with the pdf $f(\boldsymbol{x})$ its entropy is defined as

$$\mathrm{H}(f) = -\,\mathrm{E}_f \log f = -\int f(\boldsymbol{x}) \log f(\boldsymbol{x})\, d\boldsymbol{x}. \tag{1}$$

As yet we don't actually have any information that we can use as constraints to put on $\pi_{p|Q}$. One possible solution would be to find the most entropic $\pi_{p|Q}$ without any constraints. It can be shown that the principle of maximum entropy would yield a uniform distribution in this case (to be more precise $\dot{\boldsymbol{p}}$ would be uniform, but let's call $\boldsymbol{p}$ uniform as well). Expected value of a uniform distribution on $\mathcal{L}(1)$ is a uniform distribution on $supp$, which means that we might have as well applied the principle of indifference (see subsection A.2 for details on this principle) right away. And that kind of beats the whole point of using and incorporating all the information available to us.

The solution that we've chosen adds some ad hocery to the otherwise self-contained method. We assume that we are able to evaluate agents' reality-simulating skills in the terms of the expected Kerridge inaccuracy from the true distribution $\boldsymbol{p}$ to their subjective distribution $q^{(s)}$ with respect to the posterior pdf $\pi_{p|Q}$. That means we assume we know for each $1 \leq s \leq S$ a positive constant $\gamma^{(s)}$ such that

$$\mathrm{E}_{\pi_{p|Q}}[\mathrm{K}\left(\boldsymbol{q}^{(s)} \parallel \boldsymbol{p}\right) |\boldsymbol{Q} = Q] = \gamma^{(s)}. \tag{2}$$

As a corollary of the Gibbs' inequality (see A.3) the Kerridge inaccuracy is uniquely minimized for the subjective (second) argument when both arguments are equal. Therefore for the last assumption to be achievable the following inequality should hold for each $s = 1, \ldots, S$:

$$\begin{aligned}\gamma^{(s)} &= \mathrm{E}_{\pi_{p|Q}}[\mathrm{K}\left(\boldsymbol{q}^{(s)} \parallel \boldsymbol{p}\right) |\boldsymbol{Q} = Q] \geq \mathrm{E}_{\pi_{p|Q}}[\mathrm{K}\left(\boldsymbol{q}^{(s)} \parallel \boldsymbol{q}^{(s)}\right) |\boldsymbol{Q} = Q] \\ &= \mathrm{K}\left(q^{(s)} \parallel q^{(s)}\right) = \mathrm{H}(q^{(s)}).\end{aligned}$$

Also, equality in the last inequality would mean demanding that $\boldsymbol{p} = q^{(s)}$, which would make the whole merging pointless. Therefore we assume that for each $s = 1, \ldots, S$:

$$\gamma^{(s)} > \mathrm{H}\left(q^{(s)}\right). \tag{3}$$

Let's denote $L_2(\mathcal{L}(1))$ the set of all real-valued functions f on the set $\mathcal{L}(1)$, for which $\dot{f}^2$ is a Lebesgue-integrable function on $\dot{\mathcal{L}}(1)$ (for obvious reasons we will not differentiate between functions that are equal almost everywhere on this set.) Also, denote for $f \in L_2(\mathcal{L}(1))$

$$\begin{aligned}h_0(f) &= \int\limits_{\mathcal{L}(1)} f(p)\,dp - 1 \\ h_s(f) &= \int\limits_{\mathcal{L}(1)} f(p)\mathrm{K}\left(q^{(s)} \parallel p\right)\,dp - \gamma_s,\ s = 1, \ldots, S.\end{aligned}$$

Then applying the principle of maximum entropy we solve for f the following optimization problem

$$\begin{aligned}\text{maximize } & H(f) \\ \text{subject to } & \quad f \in L_2(\mathcal{L}(1)), \\ & \wedge\ f(p) \geq 0,\ p \in \mathcal{L}(1), \\ & \wedge\ h_s(f) = 0,\ s = 0, \ldots, S..\end{aligned} \tag{*}$$

Denote M the admissible set of the optimization problem above.

We will apply the method of Lagrange multipliers to the optimization problem (*). To do so we will

1. establish that M is a convex set,
2. establish that entropy is a concave function on M,
3. find Frechet derivatives of the entropy $H(f)$ and of the functions $h_s(f)$ with respect to f for each $s = 0, \ldots, S$.

First two facts guarantee that any local extremum of $H(f)$ on M is a global extremum, the derivatives in the third item will allow us to search for stationary points of the Lagrangian functional

$$L(f, \boldsymbol{\lambda}) = L(f, \lambda_0, \lambda_1, \ldots, \lambda_S) = H(f) - \sum_{s=0}^{S} \lambda_s h_s(f). \tag{4}$$

That M is convex is obvious since $L_2(\mathbb{R}_{N-1})$ is a linear vector space, a convex linear combination of nonnegative real numbers is a nonnegative real number, $h_s(f)$ are linear in f.

Proposition 1.1 *Entropy is a concave functional on the set of all pdfs supported on $\mathcal{L}(1)$.*

Proof To prove that entropy is concave we have to show for any two pdfs f, g on $\mathcal{L}(1)$ and for any α in $(0,1), \beta = 1 - \alpha$ that

$$\mathrm{H}(\alpha f + \beta g) \geq \alpha \mathrm{H}(f) + \beta \mathrm{H} g. \tag{5}$$

Let $\boldsymbol{X}, \boldsymbol{Y}$ be two continuous random vectors with pdfs f, g respectively. Let U be uniformly distributed on an open interval $(0,1)$ and let

$$\boldsymbol{Z} = \boldsymbol{X}\, \mathbb{I}_{(0;\alpha)}(U) + \boldsymbol{Y}\, \mathbb{I}_{(\alpha;1)}(U).$$

[7]

Then vectors $(\boldsymbol{Z}, U), (Z)$ have pdfs

$$\begin{aligned} h_{\boldsymbol{Z},U}(p,u) &= \left[f(p)\, \mathbb{I}_{(0;\alpha)}(u) + g(p)\, \mathbb{I}_{(\alpha;1)}(u)\right] \mathbb{I}_{\mathcal{L}(1)}(p), \\ h_{\boldsymbol{Z}}(p) &= [\alpha f(p) + \beta g(p)]\, \mathbb{I}_{\mathcal{L}(1)}(p) \end{aligned}$$

respectively.

[7] Symbol $\mathbb{I}_A$ denotes an indicator function of the set A.

There exists a vector $(\boldsymbol{Z}', U')$ such that $\boldsymbol{Z}'$ and $\boldsymbol{Z}$ and U' and U have the same distributions respectively, but $\boldsymbol{Z}'$ and U' are independent and thus have pdf

$$h_{\boldsymbol{Z}',U'}(p,u) = [\alpha f(p) + \beta g(p)]\, \mathbb{I}_{\mathcal{L}(1)}(p)\, \mathbb{I}_{(0,1)}(u).$$

Starting with the fact that Kullback-Leibler divergence is nonnegative we acquire the desired inequality (5):

$$\begin{aligned}
0 \leq & \mathrm{D}\left(h_{\boldsymbol{Z},U} \| h_{\boldsymbol{Z}',U'}\right) \\
= & \int\limits_{\mathcal{L}(1)\times(0;1)} h_{\boldsymbol{Z},U}(p,u) \log \frac{h_{\boldsymbol{Z},U}(p,u)}{h_{\boldsymbol{Z}',U'}(p,u)}\, dp\, du \\
= & \int\limits_{\mathcal{L}(1)\times(0;1)} h_{\boldsymbol{Z},U}(p,u) \log h_{\boldsymbol{Z},U}(p,u)\, dp\, du \\
& - \int\limits_{\mathcal{L}(1)\times(0;1)} h_{\boldsymbol{Z},U}(p,u) \log h_{\boldsymbol{Z}',U'}(p,u)\, dp\, du \\
= & \int\limits_{\mathcal{L}(1)\times(0;1)} (f(p)\, \mathbb{I}_{(0;\alpha)}(u) + g(p)\, \mathbb{I}_{(\alpha;1)}(u)) \times \\
& \qquad\qquad \times \log(f(p)\, \mathbb{I}_{(0;\alpha)}(u) + g(p)\, \mathbb{I}_{(\alpha;1)}(u))\, dp\, du \qquad (6) \\
& - \int\limits_{\mathcal{L}(1)\times(0;1)} (f(p)\, \mathbb{I}_{(0;\alpha)}(u) + g(p)\, \mathbb{I}_{(\alpha;1)}(u)) \log(\alpha f(p) + \beta g(p))\, dp\, du \qquad (7)
\end{aligned}$$

Integral (6) can be rewritten as

$$\begin{aligned}
& \int\limits_{\mathcal{L}(1)\times(0;\alpha)} f(p) \log f(p)\, dp\, du + \int\limits_{\mathcal{L}(1)\times(\alpha,1)} g(p) \log g(p)\, dp\, du \\
= & \alpha \int\limits_{\mathcal{L}(1)} f(p) \log f(p)\, dp + \beta \int\limits_{\mathcal{L}(1)} g(p) \log g(p)\, dp \\
= & -\left(\alpha \mathrm{H}(f) + \beta \mathrm{H}(g)\right).
\end{aligned}$$

Integral (7) can be rewritten as

$$
\begin{aligned}
&\int\limits_{\mathcal{L}(1)\times(0;\alpha)} f(p)\log(\alpha f(p)+\beta g(p))\,dp\,du\\
&\quad+\int\limits_{\mathcal{L}(1)\times(\alpha,1)} g(p)\log(\alpha f(p)+\beta g(p))\,dp\,du\\
&=\alpha\int\limits_{\mathcal{L}(1)} f(p)\log(\alpha f(p)+\beta g(p))\,dp+\beta\int\limits_{\mathcal{L}(1)} g(p)\log(\alpha f(p)+\beta g(p))\,dp\\
&=\int\limits_{\mathcal{L}(1)} (\alpha f(p)+\beta g(p))\log(\alpha f(p)+\beta g(p))\,dp\\
&=-\operatorname{H}(\alpha f+\beta g).
\end{aligned}
$$

Consequently

$$0 \le \operatorname{D}\left(h_{\boldsymbol{Z},U}\|h_{\boldsymbol{Z}',U'}\right)=\operatorname{H}(\alpha f+\beta g)-\alpha\operatorname{H}(f)+\beta\operatorname{H}(g),$$

thus inequality (5) holds and entropy is indeed concave. ■

For each $s=0,1,\ldots,S$ the function h_s is linear in f and thus is Frechet differentiable with

$$\frac{\partial h_0(f)}{\partial f}=1,\quad \frac{\partial h_s(f)}{\partial f}=\operatorname{K}\left(q^{(s)}\parallel p\right).$$

To even consider finding the Frechet derivative of entropy we need entropy to be defined on some open neighborhood of each pdf, which is not true for entropy defined as it is. First, any open neighborhood of any pdf obviously contains functions integral of which doesn't equal to 1. Second, any open neighborhood of any pdf contains functions that are not nonnegative, i.e. that are negative on a set of positive measure. These inconveniences are easily removed by defining entropy using the usual formula (1) for any integrable function on $\mathcal{L}(1)$ with the convention that $x\log x=0$ for any $x<0$. Let's call this new functional **generalized entropy.**

If Frechet differential of a functional exists it equals to the Gateaux differential of this functional. So let's first find Gateaux differential of generalized entropy. Let f,h be elements $L_2(\mathcal{L}(1)), f\ge 0$ and let α be a real number then

$$\begin{aligned}\partial \mathrm{H}(f,h) &= \lim_{\alpha\to 0}\frac{1}{\alpha}(\mathrm{H}(f+\alpha h)-\mathrm{H}(f))\\ &= \frac{\partial}{\partial\alpha}\mathrm{H}(f+\alpha h)\bigg|_{\alpha=0}\\ &= \frac{\partial}{\partial\alpha}\left[-\int_{\mathcal{L}(1)}(f+\alpha h)(p)\log(f+\alpha h)(p)\,dp\right]\Bigg|_{\alpha=0}\end{aligned}$$

Assuming that we can change order of integration and differentiation and using the fact that $(x\log x)' = \log x + 1$ we get

$$\partial \mathrm{H}(f,h) = -\int_{\mathcal{L}(1)}(\log f(p)+1)h(p)\,dp.$$

Thus if Frechet derivative of entropy $\frac{\partial}{\partial f}\mathrm{H}(f)$ exists it equals $-\log f - 1$. Instead of proving that it does exist we will take the usual (see for example [5]) approach of assuming it does and then showing that the result obtained with the use of the method of Lagrange multipliers is indeed a global extremum.

Let's assume that there are such f and $\boldsymbol{\lambda}$ that the Lagrangian functional (4) has a stationary point at $(f,\boldsymbol{\lambda})$. Then differentiating $L(f,\boldsymbol{\lambda})$ with respect to f and setting the obtained result equal to zero for $f=\pi_{\boldsymbol{p}|Q}$ we get

$$-\log\pi_{\boldsymbol{p}|Q}(p) - 1 + \lambda_0 + \sum_{s=1}^{S}\lambda_s \mathrm{K}\left(q^{(s)}\parallel p\right) = 0$$

$$\Downarrow$$

$$\pi_{\boldsymbol{p}|Q}(p) = \exp\left[\lambda_0 - 1 + \sum_{s=1}^{S}\lambda_s \mathrm{K}\left(q^{(s)}\parallel p\right)\right]. \tag{8}$$

Proposition 1.2 *Let there exist real numbers $\lambda_0,\lambda_1,\ldots,\lambda_S$ such that*

$$\pi_{\boldsymbol{p}|Q}(p) = \exp\left[\lambda_0 - 1 + \sum_{s=1}^{S}\lambda_s \mathrm{K}\left(q^{(s)}\parallel p\right)\right]$$

is an element of the admissible set M of the problem $(*)$. Then for any admissible function f from this set

$$\mathrm{H}(\pi_{\boldsymbol{p}|Q}) \geq \mathrm{H}(f)$$

with equality if and only if $\pi_{\boldsymbol{p}|Q} = f$.

Proof Let f a function from the set M, then

$$\begin{aligned}
\mathrm{H}(f) &= \mathrm{K}\left(f \parallel \pi_{\boldsymbol{p}|Q}\right) - \mathrm{D}\left(f \| \pi_{\boldsymbol{p}|Q}\right) \\
&\leq \mathrm{K}\left(f \parallel \pi_{\boldsymbol{p}|Q}\right) \\
&= -\int\limits_{\mathcal{L}(1)} f(p) \log \pi_{\boldsymbol{p}|Q}(p)\, dp \\
&= -\int\limits_{\mathcal{L}(1)} f(p) \Big[\lambda_0 - 1 + \sum_{s=1}^{S} \lambda_s \mathrm{K}\left(q^{(s)} \parallel p\right)\Big]\, dp \\
&= \lambda_0 - 1 + \sum_{s=0}^{S} \lambda_s \gamma_s \\
&= -\int\limits_{\mathcal{L}(1)} \pi_{\boldsymbol{p}|Q}(p) \Big[\lambda_0 - 1 + \sum_{s=1}^{S} \lambda_s \mathrm{K}\left(q^{(s)} \parallel p\right)\Big]\, dp \\
&= -\int\limits_{\mathcal{L}(1)} \pi_{\boldsymbol{p}|Q}(p) \log \pi_{\boldsymbol{p}|Q}(p)\, dp \\
&= \mathrm{H}(\pi_{\boldsymbol{p}|Q}).
\end{aligned}$$

Therefore $\mathrm{H}(\pi_{\boldsymbol{p}|Q}) \geq \mathrm{H}(f)$. Equality $\mathrm{D}\left(f \| \pi_{\boldsymbol{p}|Q}\right) = 0$ in Gibbs' inequality holds if and only if $\pi_{\boldsymbol{p}|Q} = f$. Therefore the same is true about equality $\mathrm{H}(\pi_{\boldsymbol{p}|Q}) = \mathrm{H}(f)$ and consequently $\mathrm{H}(\pi_{\boldsymbol{p}|Q})$ uniquely maximizes entropy on the set M. ∎

Optimization problem (*) has thereby transformed into the following set of simultaneous equations:

$$\begin{aligned}
&\text{solve for } \lambda_0, \lambda_1, \ldots, \lambda_S \\
&\qquad h_s(\pi_{\boldsymbol{p}|Q}) = 0 \quad \forall s = 0, 1, \ldots, S, \qquad (9) \\
&\text{where} \quad \pi_{\boldsymbol{p}|Q}(p) = \exp\Big[\lambda_0 - 1 + \sum_{s=1}^{S} \lambda_s \mathrm{K}\left(q^{(s)} \parallel p\right)\Big].
\end{aligned}$$

1.5 Final estimator of the true distribution

Below is the definition of the Dirichlet distribution, which is known to Bayesians as the conjugate prior of the parameters of the multinomial distribution.

Definition Random vector $\boldsymbol{Y} = (Y_1, Y_2, \ldots, Y_N), N \geq 2$, is said to have the **Dirichlet distribution** with positive parameters $\alpha_1, \alpha_2, \ldots, \alpha_N$ (denoted $\boldsymbol{Y} \sim Dir(\boldsymbol{\alpha})$) if the vector $\dot{\boldsymbol{Y}} = (Y_1, Y_2, \ldots, Y_{N-1})$ has pdf

$$f_{\boldsymbol{Y}}(\boldsymbol{y}) = f_{\dot{\boldsymbol{Y}}}(\dot{\boldsymbol{y}}) = C \prod_{n=1}^{N} y_n^{\alpha_n - 1}$$

for $y_n \geq 0, n = 1, 2, \ldots N, \sum_{n=1}^{N} y_n = 1$ and $f_{\dot{\boldsymbol{Y}}}(\dot{\boldsymbol{y}}) = 0$ everywhere else. The constant C is a normalizing multiplier.

Recall also that if $\boldsymbol{Y} \sim Dir(\boldsymbol{\alpha})$ then

$$\mathrm{E}\,\boldsymbol{Y} = \frac{1}{\sum_{n=1}^{N} \alpha_n} (\alpha_1, \alpha_2, \ldots, \alpha_N). \tag{10}$$

Let's show that if (9) has a solution then $\pi_{\boldsymbol{p}|Q}$ is a pdf of a Dirichlet distribution.

Proposition 1.3 *Let $\lambda_0, \lambda_1, \ldots, \lambda_S$ be a solution of* (9) *then*

$$\pi_{\boldsymbol{p}|Q}(p) = \exp\left[\lambda_0 - 1 + \sum_{s=1}^{S} \lambda_s \mathrm{K}\left(q^{(s)} \parallel p\right)\right] = C \prod_{n=1}^{N} p_n^{\alpha_n - 1},$$

where $C = e^{\lambda_0 - 1}$ and

$$\alpha_n = 1 + \sum_{s=1}^{S} \lambda_s q_n^{(s)} > 0, n = 1, \ldots, N. \tag{11}$$

Proof Proof of this proposition is based on a straight-forward evaluation.

$$\begin{aligned}
\pi_{\boldsymbol{p}|Q}(p) &= \exp\left[\lambda_0 - 1 + \sum_{s=1}^{S} \lambda_s \mathrm{K}\left(q^{(s)} \parallel p\right)\right] \\
&= \exp\left[\lambda_0 - 1 - \sum_{s=1}^{S} \sum_{n=1}^{N} \lambda_s q_n^{(s)} \log p_n\right] \\
&= e^{\lambda_0 - 1} \prod_{n=1}^{N} \exp\left[\log p_n \sum_{s=1}^{S} \lambda_s q_n^{(s)}\right] \\
&= e^{\lambda_0 - 1} \prod_{n=1}^{N} p_n^{\left[1 + \sum_{s=1}^{S} \lambda_s q_n^{(s)}\right] - 1} \\
&= C \prod_{n=1}^{N} p_n^{\alpha_n - 1}.
\end{aligned}$$

If there existed n such that $\alpha_n \leq 0$ then $\int_{\mathcal{L}(1)} \pi_{\boldsymbol{p}|Q}(p)$ would equal to infinity (consequence of the fact that x^α is integrable on $(0;\varepsilon)$ for $\varepsilon > 0$ if and only if $\alpha > -1$). Therefore α_n is positive for each $n = 1, 2, \ldots, N$. Positiveness of $C = e^{\lambda_0 - 1}$ is obvious. ∎

In the subsection 1.3 we established that $\mathrm{E}_{\pi_{\boldsymbol{p}|Q}}(\boldsymbol{p}|\boldsymbol{Q} = Q)$ is an optimal estimator of the true distribution p. In the previous subsection we by turn constructed $\pi_{\boldsymbol{p}|Q}$ as the solution of (9). Later in the proposition 1.3 we showed that optimal distribution of $\boldsymbol{p}$ given $\boldsymbol{Q} = Q$ is the Dirichlet distribution. Combining these facts and using (10) we obtain the estimator of the true distribution as

$$\hat{p} = \mathrm{E}_{\pi_{\boldsymbol{p}|Q}}(\boldsymbol{p}|\boldsymbol{Q} = Q) = \frac{1}{\sum_{n=1}^{N} \alpha_n}(\alpha_1, \alpha_2, \ldots, \alpha_N),$$

where αs are defined as in the proposition 1.3.

Denote

$$\begin{aligned} q^{(0)} &= (\frac{1}{N}, \ldots, \frac{1}{N}), \\ \omega_0 &= \frac{N}{N + \sum_{s=1}^{S}\sum_{n=1}^{N} \lambda_s q_n^{(s)}} = \frac{N}{N + \sum_{s=1}^{S} \lambda_s}, && (12) \\ \omega_s &= \frac{\lambda_s}{N + \sum_{s=1}^{S}\sum_{n=1}^{N} \lambda_s q_n^{(s)}} = \frac{\lambda_s}{N + \sum_{s=1}^{S} \lambda_s}, \; s = 1, \ldots, S. && (13) \end{aligned}$$

Then $\sum_{s=0}^{S} \omega_s = 1$ and

$$\hat{p} = \sum_{s=0}^{S} \omega_s q^{(s)}.$$

This way it is apparent that our method yields a weighted arithmetic mean of input pmfs $q^{(s)}$ and a uniform pmf $q^{(0)}$ as the result of pooling.

Assume that we have been given values of ωs. Let's show that there is a one-to-one correspondence between the vectors $\boldsymbol{\omega}$ and $\boldsymbol{\lambda}$ unless $\omega_0 = 0$ and find conditions that $\boldsymbol{\omega}$ has to meet for the corresponding $\boldsymbol{\lambda}$ to solve (9). We omit ω_0 from the consideration as $\omega_0 = 1 - \sum_{s=1}^{S} \omega_s$ and do the same with λ_0 as it can be derived from other λs. We then define $\boldsymbol{\omega}$ as $(\omega_1, \ldots, \omega_S)$ and redefine $\boldsymbol{\lambda}$ as $(\lambda_1, \ldots, \lambda_S)$.

For each $s = 1, \dots, S$

$$\omega_s = \frac{\lambda_s}{N + \sum_{s=1}^{S} \lambda_s} \Leftrightarrow N\omega_s + \omega_s \sum_{s=1}^{S} \lambda_s = \lambda_s$$

or, in the vector notation

$$N\boldsymbol{\omega} + \boldsymbol{\omega}\mathbf{1}^T\boldsymbol{\lambda} = \boldsymbol{\lambda} \Leftrightarrow \left(\boldsymbol{I}_s - \boldsymbol{\omega}\mathbf{1}^T\right)\boldsymbol{\lambda} = N\boldsymbol{\omega}.^8$$

Determinant of the matrix $\boldsymbol{I}_s - \boldsymbol{\omega}\mathbf{1}^T$ equals to the determinant of $\boldsymbol{I}_1 - \mathbf{1}^T\boldsymbol{\omega}$ according to the Sylvester's determinant theorem (see for example [2] for reference), which equals to $1 - \sum_{s=1}^{S} \omega_s = \omega_0$. Therefore (13) defines a one-to-one correspondence between $\{\boldsymbol{\lambda} \in \mathbb{R}_S\}$ and the set $\{\boldsymbol{\omega} \in \mathbb{R}_S : \sum_{s=1}^{S} \omega_s \neq 1\}$. From now on we assume that $\omega_0 = \sum_{s=1}^{S} \omega_s \neq 0$.

Now, with regard to the conditions $h_s(\pi_{\boldsymbol{p}|Q}) = 0 \quad \forall s = 0, 1, \dots, S$. For $s = 0$ it simply means that $\pi_{\boldsymbol{p}|Q}$ is a pdf. In the proposition 1.3 we established a necessary condition for this to be true as $\alpha_n = 1 + \sum_{s=1}^{S} \lambda_s q_n^{(s)} > 0, n = 1, \dots, N$, absolutely the same reasoning can be used to show that the latter inequalities constitute a sufficient condition for $h_0(\pi_{\boldsymbol{p}|Q})$ to equal to 0 as well.

Recall that for $s = 1, \dots, S$ the condition $h_s(\pi_{\boldsymbol{p}|Q})$ means that

$$\mathrm{E}_{\pi_{p|Q}}[\mathrm{K}\left(\boldsymbol{q}^{(s)} \parallel \boldsymbol{p}\right) | \boldsymbol{Q} = Q] = \gamma^{(s)}. \tag{14}$$

In the subsection 1.3 we showed that

$$\mathrm{E}[\mathrm{K}\left(\boldsymbol{p} \parallel \tilde{p}\right) | \boldsymbol{Q} = Q] = \mathrm{K}\left(\mathrm{E}_{\pi_{\boldsymbol{p}|Q}}(\boldsymbol{p}|\boldsymbol{Q} = Q) \parallel \tilde{p}\right).$$

It is tempting to presume that similarly

$$\mathrm{E}_{\pi_{p|Q}}[\mathrm{K}\left(\boldsymbol{q}^{(s)} \parallel \boldsymbol{p}\right) | \boldsymbol{Q} = Q] = \mathrm{K}\left(\boldsymbol{q}^{(s)} \parallel E_{\pi_{p|Q}}\boldsymbol{p}\right) \tag{15}$$

and thus (14) can be rewritten as

$$\mathrm{K}\left(q^{(s)} \parallel \sum_{s=0}^{S} \omega_s q^{(s)}\right) = \gamma^{(s)}, s = 1, \dots, S.$$

Unfortunately (and obviously) (15) does not hold, therefore we have to look for ω such that $\omega_0 \neq 0$ and $\boldsymbol{\lambda} = N\left(\boldsymbol{I}_s - \boldsymbol{\omega}\mathbf{1}^T\right)^{-1}\boldsymbol{\omega}$ solves (9).

[8] For an arbitrary matrix A the symbol A^T denotes the transposition of A.

1.6 Recapitulation

1.6.1 Cooperation structure

Assume we have S Bayesian agents that consider K-dimensional random vector $\boldsymbol{X} = (X_1, \dots, X_K)$ with N possible realizations $\boldsymbol{x}_1, \dots, \boldsymbol{x}_N$. According to s^{th} agent $\boldsymbol{X}$ has a pmf $q^{(s)}$. Denote p the true pmf of $\boldsymbol{X}$.

1.6.2 Supra Bayesian approach

Assume that $q^{(s)}, p$ are realizations of random vectors $\boldsymbol{q}^{(s)}, \boldsymbol{p}$. Denote $\boldsymbol{Q}$ the random matrix $\{\boldsymbol{q}_n^{(s)}\}_{1 \leq s \leq S, 1 \leq n \leq N}$.

1.6.3 Optimal estimator of the true distribution in case its posterior distribution is known

We assume that given $\boldsymbol{Q} = Q$ the vector $\dot{\boldsymbol{p}} = (p_1, \dots, p_n)$ is continuously distributed with the pdf $\dot{\pi}_{\boldsymbol{p}|Q}$ supported on $\mathcal{L}(1) = \{(p_1, \dots, p_{N-1}) \in \mathbb{R}_{N-1} : \sum_{n=1}^{N-1} p_n < 1\}$. For clarity sake we don't write any dots bearing in mind that it is $\dot{\boldsymbol{p}}$ that is continuous, not $\boldsymbol{p}$, and so forth.

Denote $\hat{p} = \mathrm{E}_{\pi_{\boldsymbol{p}|Q}}(\boldsymbol{p}|\boldsymbol{Q} = Q)$. Then $\hat{p}$ is an optimal estimator of $\boldsymbol{p}$ in the sense that it minimizes for $\tilde{p}$ the expected conditional Kerridge inaccuracy form $\tilde{p}$ to $\boldsymbol{p}$ given $\boldsymbol{Q} = Q$, i.e. $\mathrm{E}[\mathrm{K}\left(\boldsymbol{p} \parallel \tilde{p}\right) | \boldsymbol{Q} = Q]$.

1.6.4 Construction of posterior pdf of the true distribution

To utilize the latter result we need to construct $\pi_{\boldsymbol{p}|Q}$. Assume there are known constants $\gamma_s, s = 1, \dots, S$ such that

$$\mathrm{E}_{\pi_{p|Q}}[\mathrm{K}\left(\boldsymbol{q}^{(s)} \parallel \boldsymbol{p}\right) | \boldsymbol{Q} = Q] = \gamma^{(s)}, s = 1, \dots, S.$$

Using this equalities as constraints we search for $\pi_{\boldsymbol{p}|Q}$ as a pdf supported on $\mathcal{L}(1)$ with the highest entropy. We then show that such a pdf exists if and only if there is a solution of the following set of simultaneous equations:

$$\begin{aligned}
&\text{solve for } \lambda_0, \lambda_1, \dots, \lambda_S \\
&\qquad\qquad h_s(\pi_{\boldsymbol{p}|Q}) = 0 \quad \forall s = 0, 1, \dots, S, \qquad\qquad (16) \\
&\text{where} \quad \pi_{\boldsymbol{p}|Q}(p) = \exp\left[\lambda_0 - 1 + \sum_{s=1}^{S} \lambda_s \mathrm{K}\left(q^{(s)} \parallel p\right)\right].
\end{aligned}$$

1.6.5 Final estimator of the true distribution

We show that if (16) has a solution than the corresponding pdf $\pi_{\boldsymbol{p}|Q}$ is a pdf of a Dirichlet distribution with the parameters $\alpha_n = 1 + \sum_{s=1}^{S} \lambda_s q_n^{(s)}, n = 1, \ldots, N$. Bearing that in mind and combing the results from the previous two section we obtain the final estimator

$$\hat{p} = \mathrm{E}_{\pi_{\boldsymbol{p}|Q}}(\boldsymbol{p}|\boldsymbol{Q} = Q) = \frac{1}{\sum_{n=1}^{N} \alpha_n}(\alpha_1, \alpha_2, \ldots, \alpha_N) = \sum_{s=0}^{S} \omega_s q^{(s)},$$

where $q^{(0)} = (\frac{1}{N}, \ldots, \frac{1}{N}), \omega_0 = \frac{N}{N + \sum_{s=1}^{S} \lambda_s}, \omega_s = \frac{\lambda_s}{N + \sum_{s=1}^{S} \lambda_s},\ s = 1, \ldots, S.$

We then show that $\sum_{s=1}^{S} \omega_s \neq 0$ and that for such ωs the corresponding vector λ can be found as $\boldsymbol{\lambda} = N \left(\boldsymbol{I}_s - \boldsymbol{\omega} \mathbf{1}^T\right)^{-1} \boldsymbol{\omega}$. So alternatively we can look for the vector of weights $\boldsymbol{\omega}$ such that the corresponding $\boldsymbol{\lambda}$ solve the set of simultaneous equations from the previous subsection.

2 Several relaxations of assumptions

The supra Bayesian merging is considerably demanding in assumptions made about input models. In this section we make an attempt at relaxing some of those assumptions and making certain guidelines on what to do in case some conditions cannot be met. We will not however remove the assumption that each agent considers her domain to be a finite discrete random vector as it is the topic of the section 3.

The idea behind each relaxation is to extend and/or transform input data to make it compliant with the supra Bayesian merging described in the section 1.

To avoid unnecessary repetition in following subsections we assume that the same cooperation structure as in the section 1 except for explicitly stated differences. The same notation is used as well.

2.1 Different supports

It is somewhat bit unreasonable to assume that all agents assign positive probabilities to the same possible realizations, i.e. that the set they believe to be the support of their common domain $\boldsymbol{X}$ is the same for each agent. Let's relax the assumption of the common support and allow s^{th} agent to have her own support $supp^{(s)}$.

After doing so we need to decide what support the true distribution of $\boldsymbol{X}$ has. For two reasons we define the support of $\boldsymbol{X}$ as

$$supp = \cup_{s=1}^{S} supp^{(s)} .$$

First reason is conceptual: however small is our belief in s^{th} agent's skills it should be still positive and no value that she thinks is possible should be dismissed. If we have absolutely no belief in her skills then we should exclude her from the consideration completely.

Second reason is a more technical one, but has the same roots. During the supra Bayesian merging we assume that there is a know constant γ_s such that

$$\mathrm{E}_{\pi_{p|Q}}[\mathrm{K}\left(\boldsymbol{q}^{(s)} \parallel \boldsymbol{p}\right) |\boldsymbol{Q} = Q] = \gamma^{(s)}.$$

This assumption is never met unless $supp^{(s)} \subset supp$, because otherwise $\mathrm{K}\left(\boldsymbol{q}^{(s)} \parallel \boldsymbol{p}\right) = \infty$.

We then extend the pmf $q^{(s)}$ onto $supp$ in a 'natural' way by defining extended pmf ${}^{ext}q^{(s)}$ as

$$
{}^{ext}q^{(s)}(\boldsymbol{x}) = \begin{cases} g^{(s)}(\boldsymbol{x}), & \text{if } \boldsymbol{x} \in supp^{(s)} \\ 0, & \text{otherwise.} \end{cases}
$$

and then use extended pmfs instead of the original ones

Remaining part of merging takes place without changes.

2.2 Generalized moments of the domain as an input

Another possibility is that an agent instead of specifying the individual probabilities of all possible outcome of $\boldsymbol{X}$ provides only $L \in \mathbb{N}$ generalized moments of $\boldsymbol{X}$. That is she provides L measurable transformations $\phi_l(\boldsymbol{X})$, $l = 1, \ldots, L$, of $\boldsymbol{X}$, L constants a_l, $l = 1, \ldots, L$, and she asserts that

$$
\mathrm{E}\,\phi_l(X) = a_l,\, l = 1, \ldots, L. \tag{17}
$$

Let's assume that it is only the first agent who failed to comply with the assumptions of the supra Bayesian merging and provided the above generalized moments. To construct a pmf ${}^{con}q^{(1)}$ to use as her input pmf we apply the principle of maximum entropy, that is we search for ${}^{con}q^{(1)}$ as the solution of the following optimization problem:

$$
\begin{aligned}
&\text{maximize } H(q^{(1)}) \\
&\text{subject to} \quad \sum_{n=1}^{N} q^{(1)}(\boldsymbol{x}_n) = 1, \\
&\qquad\qquad\quad \wedge \mathrm{E}\,\phi_l(X) = a_l,\, l = 1, \ldots, L, \\
&\qquad\qquad\quad \wedge q^{(1)} > 0.
\end{aligned}
$$

Lagrangian function of this optimization problem

$$
\begin{aligned}
L(q^{(1)}, \boldsymbol{\lambda}') &= L(q^{(1)}, \lambda'_0, \lambda'_1 \ldots, \lambda'_L) \\
&= H(q^{(1)}) - \lambda'_0(\sum_{n=1}^{N} q^{(1)}(\boldsymbol{x}_n) - 1) - \sum_{l=1}^{L} \lambda'_l(\mathrm{E}_{q^{(1)}}\,\phi_l(X) - a_l)
\end{aligned}
$$

has a saturation point at the point $({}^{con}q^{(1)}, \boldsymbol{\lambda}')$ such that

$$
\frac{\partial L}{\partial q^{(1)}}({}^{con}q^{(1)}, \boldsymbol{\lambda}') = -log^{con}q^{(1)} - \mathbf{1} - \lambda'_0\mathbf{1} - \sum_{l=1}^{L} \lambda'_l \phi_l({}^{con}q^{(1)}) = 0
$$

and ${}^{con}q^{(1)}$ satisfies the constraints of the problem.

Concavity of the maximized function (the fact that entropy is a concave function on the set of pmfs on a given support can be proven in a way similar to the way it was proven for pdfs in the section 1) and linearity of the constraints guarantee that if there is such a point $({}^{con}q^{(1)}, \boldsymbol{\lambda}')$ then under the above constraints $H(q^{(1)})$ is globally maximized at the point

$${}^{con}q^{(1)}(\boldsymbol{x}_n) = C' \exp\Big[\sum_{l=1}^{L} \lambda'_l \phi_l(\boldsymbol{x}_n)\Big],\ n = 1, \ldots, N,$$

where the normalizing constant C' is chosen to make sure that ${}^{con}q^{(1)}$ is indeed a pmf (that it is a probability vector). If no such point exists then we dismiss the first agent since her model is inconsistent within itself of is not compliant with the support $supp$. In both cases instead of dismissing her completely we can 'split' her into several 'subagents' assigning different groups of moments to each of them.

2.3 Different domains and conditional pmfs

It would be favorable if we could incorporate models that consider only part of the domain $\boldsymbol{X}$ and model this part by means of a conditional pmf. Again, assume that it is only the first agent who considers only the vector $\boldsymbol{X}^{(1)} = (X_1, X_2, \ldots, X_{K^{(1)}})$ for some number $K^{(1)} \leq K$. Suppose also that she provides a conditional pmf

$$q^{(1)}_{\boldsymbol{X}^2|\boldsymbol{X}^1}(\boldsymbol{x}^1, \boldsymbol{x}^2), \boldsymbol{x}^{(1)} \in supp^{(1)},$$

where

$$\begin{aligned}
\boldsymbol{X}^1 &= (X_1, \ldots, X_{K^1}) \text{ for some } 0 \leq K^1 < K^{(1)},\\
\boldsymbol{X}^2 &= (X_{K^1+1} \ \ldots, X_{K^{(1)}})\\
\boldsymbol{x}^{(1)} &= (x_1, \ldots, x_{K^{(1)}}),\\
\boldsymbol{x}^1 &= (x_1, \ldots, x_{K^1}),\\
\boldsymbol{x}^2 &= (x_{K^1+1} \ \ldots, x_{K^{(1)}}) \text{ and}\\
supp^{(1)} &\subseteq \{\boldsymbol{x}^{(1)} : \exists \boldsymbol{x} \in supp\}. \qquad (18)
\end{aligned}$$

Failure to comply with the last condition would make the first model inconsistent (see the first subsection for details) with the other $S-1$ models or rather. In this case either the first agent should be excused from the merging or $supp$ should be extended so this condition is met. Note that we allowed K^1 to be zero and $K^{(1)}$ to be equal to K in order to include the cases when the first agent provides a joint pmf of $\boldsymbol{X}^{(1)}$ or considers the whole vector $\boldsymbol{X}$ as merely special cases.

Denote additionally $\boldsymbol{X}^3 = (X_{K^{(1)}+1} \ldots, X_K)$ and $\boldsymbol{x}^3 = (x_{K^{(1)}+1} \ldots, x_K)$. Then we can represent our notation graphically, which will prevent the possible confusion:

$$(\underbrace{\underbrace{X_1, \ldots, X_{K^1}}_{\boldsymbol{X}^1}, \underbrace{X_{K^1+1}, \ldots, X_{K^{(1)}}}_{\boldsymbol{X}^2}}_{\boldsymbol{X}^{(1)}}, \underbrace{X_{K^{(1)}+1}, \ldots, X_K}_{\boldsymbol{X}^3}).$$

The same scheme works if $\boldsymbol{X}$ is replaced with $\boldsymbol{x}$ at every instance and also for any other letters.

We need to extend the pmf $q^{(1)}_{\boldsymbol{X}^2|\boldsymbol{X}^1}$ onto the whole domain $\boldsymbol{X}$ and onto the whole support $supp$ in such a way that the resulting pmf ${}^{ext}q^{(1)}$ is both close to the original pmf $q^{(1)}_{\boldsymbol{X}^2|\boldsymbol{X}^1}$ and does not contain any unnecessary 'junk' information.

Let's denote marginal and conditional marginal pmfs for an arbitrary pmf q of $\boldsymbol{X}$ with the support $supp$ at an arbitrary point $\boldsymbol{x} \in supp$

$$\begin{aligned} q_{\boldsymbol{X}^2|\boldsymbol{X}^1}(\boldsymbol{x}^{(1)}) &= \sum_{\boldsymbol{y} \in supp: \boldsymbol{y}^{(1)} = \boldsymbol{x}^{(1)}} q(\boldsymbol{y}) / \sum_{\boldsymbol{y} \in supp: \boldsymbol{y}^1 = \boldsymbol{x}^1} q(\boldsymbol{y}), \\ q_{\boldsymbol{X}^1}(\boldsymbol{x}^1) &= \sum_{\boldsymbol{y} \in supp: \boldsymbol{y}^1 = \boldsymbol{x}^1} q(\boldsymbol{y}). \end{aligned}$$

Pmf $q_{\boldsymbol{X}^3|\boldsymbol{X}^{(1)}}$ is defined in the same manner as $q_{\boldsymbol{X}^2|\boldsymbol{X}^1}$. Using these definitions we can postulate the condition of the resulting pmf ${}^{ext}q^{(1)}$ being close to the original pmf $q^{(1)}_{\boldsymbol{X}^2|\boldsymbol{X}^1}$ as

$$ {}^{ext}q^{(1)}_{\boldsymbol{X}^2|\boldsymbol{X}^1} = q^{(1)}_{\boldsymbol{X}^2|\boldsymbol{X}^1} \qquad \forall \boldsymbol{x}^{(1)} \in supp^{(1)} . \tag{19}$$

Now for the requirement of inserting as little 'junk' information as possible. If the first agent had to find this desired extension alone then it would be reasonable to apply the principle of maximum entropy and search for ${}^{ext}q^{(1)}$ as for the most entropic pmf of $\boldsymbol{X}$ supported on $supp$ and meeting the requirement (19). However, as we have $S-1$ other models of $\boldsymbol{X}$ at our disposition it would be wasteful not to use them. Therefore we will look for ${}^{ext}q^{(1)}$ closest to the true distribution p. Closest - as already usual for us - in the terms of the expected Kerridge inaccuracy given $\boldsymbol{Q} = Q$ where $q^{(1)} = {}^{ext}q^{(1)}$.

Formally we look for ${}^{ext}q^{(1)}$ as the solution of the following optimization problem:

$$\begin{aligned} \text{minimize} \quad & \mathrm{E}[\mathrm{K}\left(\boldsymbol{p} \,\|\, q\right)|\boldsymbol{Q} = Q] \\ \text{subject to} \quad & \sum_{n=1}^{N} q(\boldsymbol{x}_n) = 1, \\ & \wedge\, q_{\boldsymbol{X}^2|\boldsymbol{X}^1}(\boldsymbol{x}^{(1)}) = q^{(1)}_{\boldsymbol{X}^2|\boldsymbol{X}^1}(\boldsymbol{x}^{(1)}) \quad \forall \boldsymbol{x}^{(1)} \in supp^{(1)}, \\ & \wedge\, q > 0. \end{aligned}$$

Hereafter we write for shortness sake $\{\boldsymbol{x}^{(1)} : \exists \boldsymbol{x} \in supp\}$ meaning the set

$$\{\boldsymbol{x}^{(1)} : \exists \boldsymbol{x}^3 \text{ such that } \boldsymbol{x} \in supp\}.$$

Same idea applies to other similar sets as well.

Proposition 2.1 *The solution of the above problem if the posterior pdf $\pi_{\boldsymbol{p}|Q}$ is given and $supp^{(1)} \subseteq \{\boldsymbol{x}^{(1)} : \exists \boldsymbol{x} \in supp\}$ is the pmf ${}^{ext}q^{(1)}$ defined for $\boldsymbol{x} \in supp$ as follows:*

$${}^{ext}q^{(1)}(\boldsymbol{x}) = \begin{cases} \hat{p}_{\boldsymbol{X}^3|\boldsymbol{X}^{(1)}}(\boldsymbol{x}) q^{(1)}_{\boldsymbol{X}^2|\boldsymbol{X}^1}(\boldsymbol{x}^{(1)}) \hat{p}_{\boldsymbol{X}^1}(\boldsymbol{x}^1), & \textit{if } \boldsymbol{x}^{(1)} \in supp^{(1)}, \\ \hat{p}(\boldsymbol{x}) & \textit{if } \boldsymbol{x}^{(1)} \notin supp^{(1)}, \end{cases}$$

where $\hat{p} = \mathrm{E}_{\pi_{\boldsymbol{p}|Q}}(\boldsymbol{p}|\boldsymbol{Q} = Q)$ is an optimal estimator of the true distribution p from the subsection 1.3.

Proof Let's first show that ${}^{ext}q^{(1)}$ is indeed a pmf with the support $supp$. That ${}^{ext}q^{(1)}$ is positive is obvious from its definition since $\hat{p}$ is supported on $supp$.

$$\begin{aligned} \sum_{\boldsymbol{x} \in supp} {}^{ext}q^{(1)}(\boldsymbol{x}) = & \sum_{\boldsymbol{x} \in supp : \boldsymbol{x}^{(1)} \notin supp^{(1)}} \hat{p}(\boldsymbol{x}) \\ & + \sum_{\boldsymbol{x} \in supp : \boldsymbol{x}^{(1)} \in supp^{(1)}} \hat{p}_{\boldsymbol{X}^3|\boldsymbol{X}^{(1)}}(\boldsymbol{x}) q^{(1)}_{\boldsymbol{X}^2|\boldsymbol{X}^1}(\boldsymbol{x}^{(1)}) \hat{p}_{\boldsymbol{X}^1}(\boldsymbol{x}^1). \end{aligned}$$

Let's transform the second term.

$$\begin{aligned}
&\sum_{\boldsymbol{x}\in supp:\boldsymbol{x}^{(1)}\in supp^{(1)}} \hat{p}_{\boldsymbol{X}^3|\boldsymbol{X}^{(1)}}(\boldsymbol{x}) q^{(1)}_{\boldsymbol{X}^2|\boldsymbol{X}^1}(\boldsymbol{x}^{(1)})\hat{p}_{\boldsymbol{X}^1}(\boldsymbol{x}^1)\\
&\quad = \sum_{\boldsymbol{x}^{(1)}\in supp^{(1)}} \sum_{\boldsymbol{x}^3:(\boldsymbol{x}^{(1)},\boldsymbol{x}^3)\in supp} \hat{p}_{\boldsymbol{X}^3|\boldsymbol{X}^{(1)}}(\boldsymbol{x}) q^{(1)}_{\boldsymbol{X}^2|\boldsymbol{X}^1}(\boldsymbol{x}^{(1)})\hat{p}_{\boldsymbol{X}^1}(\boldsymbol{x}^1)\\
&\quad = \sum_{\boldsymbol{x}^{(1)}\in supp^{(1)}} q^{(1)}_{\boldsymbol{X}^2|\boldsymbol{X}^1}(\boldsymbol{x}^{(1)})\hat{p}_{\boldsymbol{X}^1}(\boldsymbol{x}^1)
\end{aligned}$$

In the last equality we used the fact that a conditional pmf is a probability vector for any given fixed value of the condition.

Denote $supp^1 = \{\boldsymbol{x}^1 : \exists \boldsymbol{x}^{(1)} \in supp^{(1)}\}$ then

$$\begin{aligned}
\sum_{\boldsymbol{x}^{(1)}\in supp^{(1)}} q^{(1)}_{\boldsymbol{X}^2|\boldsymbol{X}^1}(\boldsymbol{x}^{(1)})\hat{p}_{\boldsymbol{X}^1}(\boldsymbol{x}^1) &= \sum_{\boldsymbol{x}^1\in supp^1} \sum_{\boldsymbol{x}^2:(\boldsymbol{x}^1,\boldsymbol{x}^2)\in supp^{(1)}} q^{(1)}_{\boldsymbol{X}^2|\boldsymbol{X}^1}(\boldsymbol{x}^{(1)})\hat{p}_{\boldsymbol{X}^1}(\boldsymbol{x}^1)\\
&= \sum_{\boldsymbol{x}^1\in supp^1} \hat{p}_{\boldsymbol{X}^1}(\boldsymbol{x}^1)\\
&= \sum_{\boldsymbol{x}^1\in supp^1} \sum_{\boldsymbol{y}\in supp:\boldsymbol{y}^1\in supp_1} \hat{p}(\boldsymbol{y})\\
&= \sum_{\boldsymbol{x}\in supp:\boldsymbol{x}^{(1)}\in supp^{(1)}} \hat{p}(\boldsymbol{x}).
\end{aligned}$$

Therefore

$$\begin{aligned}
\sum_{\boldsymbol{x}\in supp} {}^{ext}q^{(1)}(\boldsymbol{x}) &= \sum_{\boldsymbol{x}\in supp:\boldsymbol{x}^{(1)}\notin supp^{(1)}} \hat{p}(\boldsymbol{x}) + \sum_{\boldsymbol{x}\in supp:\boldsymbol{x}^{(1)}\in supp^{(1)}} \hat{p}(\boldsymbol{x})\\
&= \sum_{\boldsymbol{x}\in supp} \hat{p}(\boldsymbol{x}) = 1
\end{aligned}$$

and ${}^{ext}q^{(1)}$ is indeed a pmf supported on $supp$.

Similar transformations can be used to show that for any $\boldsymbol{x}^{(1)} \in supp^{(1)}$

$$ {}^{ext}q^{(1)}_{\boldsymbol{X}^{(1)}}(\boldsymbol{x}^{(1)}) = q^{(1)}_{\boldsymbol{X}^2|\boldsymbol{X}^1}(\boldsymbol{x}^{(1)})\hat{p}(\boldsymbol{x}^1), \qquad {}^{ext}q^{(1)}_{\boldsymbol{X}^1}(\boldsymbol{x}_1) = \hat{p}(\boldsymbol{x}_1).$$

Consequently

$$ {}^{ext}q^{(1)}_{\boldsymbol{X}^2|\boldsymbol{X}^1} = q^{(1)}_{\boldsymbol{X}^2|\boldsymbol{X}^1}(\boldsymbol{x}^{(1)})$$

ant thus ${}^{ext}q^{(1)}$ satisfies the condition (19).

Recall that in the section 1 we showed that

$$\mathrm{E}[\mathrm{K}\left(\boldsymbol{p} \,\|\, q\right) | \boldsymbol{Q} = Q] = \mathrm{K}\left(\mathrm{E}_{\pi_{\boldsymbol{p}|Q}}(\boldsymbol{p}|\boldsymbol{Q} = Q) \,\|\, q\right) = \mathrm{K}\left(\hat{p} \,\|\, q\right).$$

Let q be an arbitrary pmf of $\boldsymbol{X}$ supported on $supp$ and satisfying the condition (19). Then

$$\begin{aligned}
\mathrm{K}\left(\hat{p} \,\|\, q\right) &= \sum_{\boldsymbol{x} \in supp} \hat{p}(\boldsymbol{x}) \log q(\boldsymbol{x}) \\
&= \sum_{\boldsymbol{x} \in supp} \hat{p}_{\boldsymbol{X}^3|\boldsymbol{X}^{(1)}}(\boldsymbol{x}) \hat{p}_{\boldsymbol{X}^2|\boldsymbol{X}^1}(\boldsymbol{x}^{(1)}) \hat{p}_{\boldsymbol{X}^1}(\boldsymbol{x}^1) \times \\
&\qquad\qquad\qquad \times \log\left(q_{\boldsymbol{X}^3|\boldsymbol{X}^{(1)}}(\boldsymbol{x}) q_{\boldsymbol{X}^2|\boldsymbol{X}^1}(\boldsymbol{x}^{(1)}) q_{\boldsymbol{X}^1}(\boldsymbol{x}^1)\right) \\
&= \sum_{\boldsymbol{x} \in supp} \hat{p}_{\boldsymbol{X}^3|\boldsymbol{X}^{(1)}}(\boldsymbol{x}) \hat{p}_{\boldsymbol{X}^2|\boldsymbol{X}^1}(\boldsymbol{x}^{(1)}) \hat{p}_{\boldsymbol{X}^1}(\boldsymbol{x}^1) \log q_{\boldsymbol{X}^3|\boldsymbol{X}^{(1)}}(\boldsymbol{x}) \\
&\quad + \sum_{\boldsymbol{x} \in supp} \hat{p}_{\boldsymbol{X}^3|\boldsymbol{X}^{(1)}}(\boldsymbol{x}) \hat{p}_{\boldsymbol{X}^2|\boldsymbol{X}^1}(\boldsymbol{x}^{(1)}) \hat{p}_{\boldsymbol{X}^1}(\boldsymbol{x}^1) \log q_{\boldsymbol{X}^2|\boldsymbol{X}^1}(\boldsymbol{x}^{(1)}) \\
&\quad + \sum_{\boldsymbol{x} \in supp} \hat{p}_{\boldsymbol{X}^3|\boldsymbol{X}^{(1)}}(\boldsymbol{x}) \hat{p}_{\boldsymbol{X}^2|\boldsymbol{X}^1}(\boldsymbol{x}^{(1)}) \hat{p}_{\boldsymbol{X}^1}(\boldsymbol{x}^1) \log q_{\boldsymbol{X}^1}(\boldsymbol{x}^1) \\
&= \sum_{\boldsymbol{x} \in supp} \hat{p}_{\boldsymbol{X}^3|\boldsymbol{X}^{(1)}}(\boldsymbol{x}) \hat{p}_{\boldsymbol{X}^2|\boldsymbol{X}^1}(\boldsymbol{x}^{(1)}) \hat{p}_{\boldsymbol{X}^1}(\boldsymbol{x}^1) \log q_{\boldsymbol{X}^3|\boldsymbol{X}^{(1)}}(\boldsymbol{x}) \\
&\quad + \sum_{\boldsymbol{x}^{(1)} : \exists \boldsymbol{x} \in supp} \hat{p}_{\boldsymbol{X}^2|\boldsymbol{X}^1}(\boldsymbol{x}^{(1)}) \hat{p}_{\boldsymbol{X}^1}(\boldsymbol{x}^1) \log q_{\boldsymbol{X}^2|\boldsymbol{X}^1}(\boldsymbol{x}^{(1)}) \\
&\quad + \sum_{\boldsymbol{x}_1 : \exists \boldsymbol{x} \in supp} \hat{p}_{\boldsymbol{X}^1}(\boldsymbol{x}^1) \log q_{\boldsymbol{X}^1}(\boldsymbol{x}^1) \\
&= \sum_{\boldsymbol{x}^{(1)} : \exists \boldsymbol{x} \in supp} \hat{p}_{\boldsymbol{X}^2|\boldsymbol{X}^1}(\boldsymbol{x}^{(1)}) \hat{p}_{\boldsymbol{X}^1}(\boldsymbol{x}^1) \\
&\qquad\qquad\qquad \sum_{\boldsymbol{x}^3 : (\boldsymbol{x}^{(1)}, \boldsymbol{x}^3) \in supp} \hat{p}_{\boldsymbol{X}^3|\boldsymbol{X}^{(1)}}(\boldsymbol{x}) \log q_{\boldsymbol{X}^3|\boldsymbol{X}^{(1)}}(\boldsymbol{x}) \\
&\quad + \sum_{\boldsymbol{x}^1 : \exists \boldsymbol{x} \in supp} \hat{p}_{\boldsymbol{X}^1}(\boldsymbol{x}^1) \sum_{\boldsymbol{x}^2 : \exists (\boldsymbol{x}^1, \boldsymbol{x}^2, \boldsymbol{x}^3) \in supp} \hat{p}_{\boldsymbol{X}^2|\boldsymbol{X}^1}(\boldsymbol{x}^{(1)}) \log q_{\boldsymbol{X}^2|\boldsymbol{X}^1}(\boldsymbol{x}^{(1)}) \\
&\quad + \sum_{\boldsymbol{x}_1 : \exists \boldsymbol{x} \in supp} \hat{p}_{\boldsymbol{X}^1}(\boldsymbol{x}^1) \log q_{\boldsymbol{X}^1}(\boldsymbol{x}^1)
\end{aligned} \tag{20}$$

By using again the fact that a conditional pmf is a probability vector for any given fixed value of the condition and by noticing Kerridge inaccuracies in equations above we can rewrite (20) as follows:

$$\begin{aligned}
&\sum_{\boldsymbol{x}^{(1)}:\exists \boldsymbol{x}\in supp} \hat{p}_{\boldsymbol{X}^2|\boldsymbol{X}^1}(\boldsymbol{x}^{(1)})\hat{p}_{\boldsymbol{X}^1}(\boldsymbol{x}^1)\mathrm{K}\left(\hat{p}_{\boldsymbol{X}^3|\boldsymbol{X}^{(1)}}(\boldsymbol{x}^{(1)},*) \parallel q_{\boldsymbol{X}^3|\boldsymbol{X}^{(1)}}(\boldsymbol{x}^{(1)},*)\right)\\
&+ \sum_{\boldsymbol{x}^1\notin supp^1:\exists \boldsymbol{x}\in supp} \hat{p}_{\boldsymbol{X}^1}(\boldsymbol{x}^1)\mathrm{K}\left(\hat{p}_{\boldsymbol{X}^2|\boldsymbol{X}^1}(\boldsymbol{x}^1,*) \parallel q_{\boldsymbol{X}^2|\boldsymbol{X}^1}(\boldsymbol{x}^1,*)\right)\\
&+ \sum_{\boldsymbol{x}^1\in supp^1} \hat{p}_{\boldsymbol{X}^1}(\boldsymbol{x}^1) \sum_{\boldsymbol{x}^2:\exists(\boldsymbol{x}^1,\boldsymbol{x}^2,\boldsymbol{x}^3)\in supp} \hat{p}_{\boldsymbol{X}^2|\boldsymbol{X}^1}(\boldsymbol{x}^{(1)}) \log q^{(1)}_{\boldsymbol{X}^2|\boldsymbol{X}^1}(\boldsymbol{x}^{(1)})\\
&+ \mathrm{K}\left(\hat{p}_{\boldsymbol{X}^1} \parallel q_{\boldsymbol{X}^1}\right) \qquad (21)
\end{aligned}$$

The only parts of the expression (21) that we can influence by the choice of q are the Kerridge inaccuracies. Also, if we change any of the subjective pmfs in those inaccuracies for any other pmf with the same support then the altered q will still be a pmf supported on $supp$ and satisfying the condition (19). Therefore each such pmf can be chosen independently and thus each of the Kerridge inaccuracies in (21) can be minimized independently. Those are minimized by choosing

$$\begin{aligned}
q_{\boldsymbol{X}^3|\boldsymbol{X}^{(1)}}(\boldsymbol{x}^{(1)},*) &= \hat{p}_{\boldsymbol{X}^3|\boldsymbol{X}^{(1)}}(\boldsymbol{x}^{(1)},*),\\
q_{\boldsymbol{X}^2|\boldsymbol{X}^1}(\boldsymbol{x}^1,*) &= \hat{p}_{\boldsymbol{X}^2|\boldsymbol{X}^1}(\boldsymbol{x}^1,*), \boldsymbol{x}^1 \notin supp^1 : \exists \boldsymbol{x} \in supp,\\
q_{\boldsymbol{X}^1} &= \hat{p}_{\boldsymbol{X}^1},
\end{aligned}$$

which leads to minimizing $\mathrm{K}\left(\hat{p} \parallel q\right)$ by choosing

$$q = {}^{ext}q^{(1)}.$$

∎

Extending partial pmf $q^{(1)}_{\boldsymbol{X}^2|\boldsymbol{X}^1}$ as described above make equations in the set (9) implicit. Existence and uniqueness of a solution in this case requires further research.

In the preceding part we assumed that only one agent provided a partial pmf of $\boldsymbol{X}$, which doesn't have to be true, for all we know all agent could do so. In this case we need to decide on what $\boldsymbol{X}$ and $supp$ will be.

A reasonable way to choose $\boldsymbol{X}$ is to pool all the variables considered by all agents into one vector $\boldsymbol{X}$. For supra Bayesian merging to make any sense in this case it would require that agents couldn't be divided into two groups, for which pooling of all considered variables would yield two vectors with no elements in common. If agents can be divided this way then they should be.

For each $s = 1, \dots, S$ subjective partial supports $\mathrm{s}upp^{(s)}$ should be compatible with $\mathrm{s}upp$ in the sense of the condition (18). The support of $\boldsymbol{X}$ should be chosen accordingly. Cartesian product of unions of marginal supports, or minimization in the sense of subsets come into mind.

3 Continuous case

In this section we propose a way to apply supra Bayesian merging to the continuous case, i.e. to the case when agents provide pdfs instead of pmfs. We will still assume that agents share a common domain on a common support. Finite variables in the discrete case are replaced by bounded continuous variables, joint pmfs - with joint pdfs.

We will follow the direction of the section 1 in the beginning. However at the point where in the mentioned chapter we arrived at searching for the most entropic posterior pdf of the true distribution the problem of finding a distribution of a continuous-time random process with the highest entropy will arise. Instead of struggling with the solution of the latter problem we will propose a work-around. It is based on assuming that the true pdf is in fact a simple function on a predefined partition of the support such that the true pdf is constant on each part. We will show that this assumption will reduce the continuous case to a discrete one.

3.1 Cooperation structure

First, we need to redefine a 'Bayesian agent.'

Definition A **Bayesian agent** is an entity characterized by its **probabilistic information,** which is a pdf

$$g(\boldsymbol{x}), \quad \boldsymbol{x} \in supp \subset \mathbb{R}_K,$$

of a bounded continuous real random vector

$$\boldsymbol{X} = (X_1, X_2, \ldots, X_K)$$

with the length $K \in \mathbb{N}$ and the bounded connected open support *supp.* Random vector $\boldsymbol{X}$ is called agent's **domain.**

Assume there are $S \in \mathbb{N}$ (as in the discrete case we assume that $S \geq 2$) Bayesian agents indexed by numbers from 1 to S with a shared domain

$$\boldsymbol{X} = (X_1, \ldots, X_K).$$

Assume that the s^{th} agent provides a pdf $q^{(s)}$ of $\boldsymbol{X}$ support on a bounded support

$$supp \subset \mathbb{R}_K,$$

which is the same for all agents.

Denote ${}^{true}g$ the true pdf of $\boldsymbol{X}$. Our goal is to find an estimate ${}^{true}\hat{g}$ of ${}^{true}g$.

3.2 Supra Bayesian approach

Suppose s^{th} agent's probabilistic information $g^{(s)}$ is a realization of a random process $\boldsymbol{g}^{(s)}$ and the true distribution ${}^{true}g$ is a realization of a random process ${}^{true}\boldsymbol{g}$. All the above processes are continuous-'time' processes indexed by the K-dimensional elements of $supp$.

3.3 Optimal estimator of the true distribution in case its posterior distribution is known

There is only so much precision then can be used and thus required. Our approached is based on knowing how much precision one needs beforehand. We will assume that there is a given partition $(A_1, A_2, \ldots, A_N)$ of $supp$ such that $\cup_{n=1}^{N} A_n = supp$ and the sets $A_1, A_2, \ldots, A_N$ are mutually exclusive. Assume additionally that for each $n = 1, \ldots, N$ the Lebesgue measure $\lambda(A_n)$ is positive. It is not unreasonable to propose that all samples of ${}^{true}\boldsymbol{g}$ are pdf that are simple functions of a form

$$ {}^{true}g(\boldsymbol{x}) = \sum_{n=1}^{N} \mathbb{I}_{A_n}(\boldsymbol{x}) \frac{p_n}{\lambda(A_n)}, \quad \boldsymbol{x} \in supp \tag{22} $$

where $p_n \geq 0$ for each $n = 1, \ldots, N, \sum_{n=1}^{N} = 1$. Denote $\boldsymbol{p} = (\boldsymbol{p}_1, \ldots, \boldsymbol{p}_N)$ the corresponding random vector.

Remark Utilization of the same letters (N and $\boldsymbol{p}$) as in the section 1 is not a mistake, and not a coincidence either, it is a deliberate choice that should emphasize the parallel between two sections. More borrowing like this is to take place below.

Denote $\boldsymbol{G} = (\boldsymbol{g}^{(1)}, \ldots, \boldsymbol{g}^{(N)})$ a random vector (it can also be seen as a matrix with N rows and columns index by the elements of $supp$, the counterpart of the matrix $\boldsymbol{Q}$ from the first section) of input pdfs of which we observed the realization $G = (g^{(1)}, \ldots, g^{(N)})$. As in the section 1 we further assume that $\dot{\boldsymbol{p}} = (p_1, \ldots, p_{N-1})$ has a conditional pdf $\dot{\pi}_{\dot{\boldsymbol{p}}|G}$ given

$\boldsymbol{G} = G$ supported on the set $\dot{\boldsymbol{\mathcal{L}}}(1)$ and then drop dots keeping in mind that the actual continuous vector is $\dot{\boldsymbol{p}}$, not $\boldsymbol{p}$.

We already showed in the subsection 1.3 that in this case the optimal estimator of p is

$$\hat{p} = \mathrm{E}_{\pi_{\boldsymbol{p}|G}}(\boldsymbol{p}|\boldsymbol{G} = G).$$

3.4 Construction of posterior pdf of the true distribution

To construct $\pi_{\boldsymbol{p}|G}$ we assume that we know positive constants $\kappa^{(s)}, s = 1, \ldots, S$ such that

$$\mathrm{E}_{\pi_{p|G}}[\mathrm{K}\left(\boldsymbol{g}^{(s)} \parallel {}^{true}g\right) | \boldsymbol{G} = G] = \kappa^{(s)}. \tag{23}$$

and then apply the principle of maximum entropy.

Similarly to (3) we assume that for each $s = 1, \ldots, S$

$$\kappa^{(s)} > \mathrm{H}\left(g^{(s)}\right).$$

Using (22) we can write

$$\begin{aligned}
\mathrm{K}\left(g^{(s)} \parallel {}^{true}g\right) &= \int_{\boldsymbol{x} \in supp} g^{(s)}(\boldsymbol{x}) \log {}^{true}g(\boldsymbol{x})\, d\boldsymbol{x} \\
&= \int_{\boldsymbol{x} \in supp} g^{(s)}(\boldsymbol{x}) \left(\log \sum_{n=1}^{N} \mathbb{I}_{A_n}(\boldsymbol{x}) \frac{p_n}{\lambda(A_n)} \right) d\boldsymbol{x} \\
&= \sum_{n=1}^{N} \log \frac{p_n}{\lambda(A_n)} \int_{\boldsymbol{x} \in supp} \mathbb{I}_{A_n}(\boldsymbol{x}) g^{(s)}(\boldsymbol{x})\, d\boldsymbol{x}
\end{aligned}$$

Denote

$$q_n^{(s)} = \int_{\boldsymbol{x} \in supp} \mathbb{I}_{A_n}(\boldsymbol{x}) g^{(s)}(\boldsymbol{x})\, d\boldsymbol{x} = \int_{\boldsymbol{x} \in A_n} g^{(s)}(\boldsymbol{x})\, d\boldsymbol{x}.$$

Then for each $s = 1, \ldots, S$ the vector $\boldsymbol{x}^{(s)}$ is a probability vector and

$$\mathrm{K}\left(g^{(s)} \parallel {}^{true}g\right) = \sum_{n=1}^{N} \log \frac{p_n}{\lambda(A_n)} q_n^{(s)} = \sum_{n=1}^{N} q_n^{(s)} \log p_n - \sum_{n=1}^{N} q_n^{(s)} \log \lambda(A_n)$$

So what we look for is the most entropic pdf $\pi_{\boldsymbol{p}|G}$ such that for each $s = 1, \dots, S$

$$\mathrm{E}_{\pi_{p|G}}[\mathrm{K}\left(\boldsymbol{g}^{(s)} \parallel {}^{true}g\right) |\boldsymbol{G} = G] = \gamma^{(s)},$$

where

$$\gamma^{(s)} = \kappa^{(s)} + \sum_{n=1}^{N} q_n^{(s)} \log \lambda(A_n).$$

Let's define for each $s = 1, \dots, S$ a pdf $\tilde{g}^{(s)}$ based on $q^{(s)}$ in the same fashion as p defines ${}^{true}g$ in (22):

$$\tilde{g}^{(s)}(\boldsymbol{x}) = \sum_{n=1}^{N} \mathbb{I}_{A_n}(\boldsymbol{x}) \frac{q_n^{(s)}}{\lambda(A_n)}, \quad \boldsymbol{x} \in supp\,.$$

Then

$$\begin{aligned}
\gamma^{(s)} &= \kappa^{(s)} + \sum_{n=1}^{N} q_n^{(s)} \log \lambda(A_n) \\
&> \mathrm{H}\left(g^{(s)}\right) + \sum_{n=1}^{N} q_n^{(s)} \log \lambda(A_n) \\
&= \int\limits_{\boldsymbol{x} \in supp} g^{(s)}(\boldsymbol{x}) \log g^{(s)}(\boldsymbol{x})\, d\boldsymbol{x} + \sum_{n=1}^{N} q_n^{(s)} \log \lambda(A_n) \\
&= \int\limits_{\boldsymbol{x} \in supp} g^{(s)}(\boldsymbol{x}) \log g^{(s)}(\boldsymbol{x})\, d\boldsymbol{x} - \sum_{n=1}^{N} q_n^{(s)} \log \frac{q_n^{(s)}}{\lambda(A_n)} \\
&\quad + \sum_{n=1}^{N} q_n^{(s)} \log \frac{q_n^{(s)}}{\lambda(A_n)} + \sum_{n=1}^{N} q_n^{(s)} \log \lambda(A_n) \\
&= \int\limits_{\boldsymbol{x} \in supp} g^{(s)}(\boldsymbol{x}) \log g^{(s)}(\boldsymbol{x})\, d\boldsymbol{x} - \int\limits_{\boldsymbol{x} \in supp} g^{(s)}(\boldsymbol{x}) \log \tilde{g}^{(s)}(\boldsymbol{x})\, d\boldsymbol{x} \\
&\quad + \sum_{n=1}^{N} q_n^{(s)} \log q_n^{(s)} \\
&= \int\limits_{\boldsymbol{x} \in supp} g^{(s)}(\boldsymbol{x}) \log \frac{g^{(s)}(\boldsymbol{x})}{\tilde{g}^{(s)}(\boldsymbol{x})}\, d\boldsymbol{x} + \sum_{n=1}^{N} q_n^{(s)} \log q_n^{(s)} \\
&= \mathrm{D}\left(g^{(s)} \| \tilde{g}^{(s)}\right) + \mathrm{H}\left(q^{(s)}\right)
\end{aligned}$$

Applying Gibbs' inequality (see subsection A.2) we get the following inequality

$$\gamma^{(s)} = \mathrm{D}\left(g^{(s)} \| \tilde{g}^{(s)}\right) + \mathrm{H}\left(q^{(s)}\right) \geq \mathrm{H}\left(q^{(s)}\right)$$

and thus acquire the condition (3).

Therefore we're solving the problem that we've already solved in the first section by applying Lagrange multipliers.

3.5 The final estimator of the true distribution

The constructed $\pi_{\boldsymbol{p}|G}$ is a pdf of a Dirichlet distribution and $\hat{p}$ is its expected value. The estimated $\hat{p}$ defines an optimal estimator of ${}^{true}g$ as

$${}^{true}\hat{g} = \sum_{n=1}^{N} \mathbb{I}_{A_n}(\boldsymbol{x}) \frac{\hat{p}_n}{\lambda(A_n)}, \quad \boldsymbol{x} \in supp\,.$$

3.6 Recapitulation

Assume that S agents provided S pdfs $g^{(1)}, \ldots, g^{(S)}$ of the same vector $\boldsymbol{X}$ and supported on the same set $supp$. Denote ${}^{true}g$ the true pdf of $\boldsymbol{X}$. Assume that the required level of precision was set by proposing that

$${}^{true}g(\boldsymbol{x}) = \sum_{n=1}^{N} \mathbb{I}_{A_n}(\boldsymbol{x}) \frac{p_n}{\lambda(A_n)}, \quad \boldsymbol{x} \in supp \tag{24}$$

for some predefined partition $(A_1, \ldots, A_N)$ of $supp$. Denote $\boldsymbol{p} = (\boldsymbol{p}_1, \ldots, \boldsymbol{p}_N)$ the corresponding random vector and $\dot{\pi}_{\dot{\boldsymbol{p}}|G}$ its conditional pdf given $\boldsymbol{G} = G$ where $G = (g^{(1)}, \ldots, g^{(S)})$.

Assume also that we were able to assess agents' skills by finding positive constants $\kappa^{(s)}, s = 1, \ldots, S$ such that

$$\mathrm{E}_{\pi_{p|G}}[\mathrm{K}\left(\boldsymbol{g}^{(s)} \parallel {}^{true}g\right) |\boldsymbol{G} = G] = \kappa^{(s)}. \tag{25}$$

Then an optimal estimator of ${}^{true}g$ can be found in two steps:

1. Apply the supra Bayesian merging on

$$q_n^{(s)} = \int_{\boldsymbol{x} \in A_n} g^{(s)}(\boldsymbol{x})\, d\boldsymbol{x},$$

$$\gamma^{(s)} = \kappa^{(s)} + \sum_{n=1}^{N} \lambda(A_n) q^{(s)}.$$

and calculate $\hat{p}$.

2. Calculate a pdf ${}^{true}\hat{g}$ corresponding to the $\hat{p}$ and satisfying (24) as

$$^{true}\hat{g} = \sum_{n=1}^{N} \mathbb{I}_{A_n}(\boldsymbol{x}) \frac{\hat{p}_n}{\lambda(A_n)}, \quad \boldsymbol{x} \in supp\,.$$

4 Conclusion

In this section we would like to discuss gains and shortcomings of the proposed supra Bayesian merging. Also we propose a way that will possibly lead to elimination of one of the latter and review open problems.

4.1 Gains

- Proposed method of supra Bayesian merging gives a systematic way to approach the problem of statistical model pooling.
- Intuitive approach of arithmetic pooling used widely as ad-hoc method (see [9] for an overview of those methods or [8] for a specific method) was both refined by introducing a necessary shift and given a proper justification.
- In the section 2 we established a way to merge models even in case agents provided us only with partial information: generalized moments or conditional pmf of only some of the considered variables.
- Usually silently ignored question of the support on which pooling is conducted was given proper attention. It may seem to be a small irrelevant thing, but it may lead to mistakes both in theoretical part and, even if mistakes are avoided there, it can still lead to mistakes in implementation. This is due to the fact that outcomes of both the principle of maximum entropy and the principle of minimum cross-entropy (Kullback-Leibler divergence) heavily depend on the support of the optimized distribution.
- One of the mistakes connected with not giving enough attention to support occurred in the propositions 4.1 through 4.3 in [20]. That inaccuracy was fixed and those propositions were unified in the proposition 2.1.
- A way of dealing with the continuous case was suggested.
- Concavity of entropy was proven in this work. And the derivative of entropy was calculated under the assumption that it exists. Proving the latter showed to be unnecessary. Both are quite simple, yet unnecessary for rigorous proves in the first case and for understanding in the case.

While applying the principle of maximum entropy it is usually (see for example [5] or [13]) assumed that differential (continuous-case) entropy is a concave functional and that differentiating the entropy $\mathrm{H}(f)$ with respect to f we get $-\log f$. The first fact is assumed to be true by automatically extending the property of the discrete-case entropy to the continuous-case entropy. That the second fact holds is either assumed using the same logic or it is assumed that the reader is fluent in the calculus of variations. We can only guess as no comments are given. As differential entropy is not a limit of its discrete version simply extending properties to the continuous case is not a sound way to go. Assuming that the reader is familiar with the calculus of variation is at the very least strange.

4.2 Shortcomings

- No obvious extension onto the infinite discrete and unbound continuous case was found. Considering the corresponding limits was conducted, but unsuccessfully.

- In the initial guidelines of this work it was supposed that the method could be applied to the so-called LQ (Linear-Quadratic) scheme. Unfortunately this was unsuccessful.

  Suppose unknown process is controlled by a linear regression with multiple parameters and at each point of time has Normal distribution with known variance and mean that is given by linear combination of regressors and parameters. Each agent is supposed to provide a prior distribution of the parameters which also has Normal distribution. It was conjectured that by pooling of the prior distribution we would get a mixture of Normal distributions. Unfortunately the method to deal with continuous distributions described in this work cannot be applied to this case, therefore the mentioned conjecture is still a conjecture.

4.3 Possible adjustment

A way to get round the second shortcoming is to switch arguments in Kullback-Leibler inaccuracies representing agents' skills. That would mean assuming that we were given positive constants $\gamma'_1, \ldots, \gamma'_S$ such that

$$\mathrm{E}_{\pi_{p|Q}}[\mathrm{K}\left(\boldsymbol{p} \,\|\, \boldsymbol{q}^{(s)}\right) |\boldsymbol{Q} = Q] = \gamma^{(s)}$$

(compare with (2).)

First calculations suggest that in this way the posterior pdf $\hat{\pi}_{p|Q}$ would be of a form

$$\hat{\pi}_{p|Q}(\boldsymbol{p}) = C \prod_{n=1}^{N} e^{-\beta_n p_n}$$

for some positive constants $C, \beta_1, \dots, \beta_N$. Assuming that this is true it would in turn lead to a shifted geometric mean as the final estimator of the true pmf p, which is less intuitive but nevertheless reasonable.

Also

$$\mathrm{E}_{\pi_{p|Q}}[\mathrm{K}\left(\boldsymbol{p} \,\|\, \boldsymbol{q}^{(s)}\right) |\boldsymbol{Q} = Q] = \mathrm{K}\left(\mathrm{E}_{\hat{\pi}_{p|Q}}[\boldsymbol{p}|\boldsymbol{Q} = Q] \,\|\, \boldsymbol{q}^{(s)}\right),$$

which would allow us to look for a $\hat{p}^{final}$ as a shifted geometric mean of $q^{(s)}, s = 1, \dots, S$, such that $\mathrm{K}\left(\hat{p}^{final} \,\|\, q^{(s)}\right) = \gamma_s, s = 1, \dots, S$, and forget about the posterior pdf $\hat{\pi}_{p|Q}$ completely. Apart from obvious numerical advantages, this fact would probably allow for a reasonable way to consider N converging ∞ and thus extend the supra Bayesian merging to infinite discrete random variables.

At the present stage all of the above conjectures are merely what they are - conjectures derived from the author's hunch. There is a need for more calculations and elaboration that constitute a possible direction for the future work.

4.4 Open problems

- Shortcoming described above should be further studied and hopefully removed.

- Constants γ_s in the section 1 were supposed to be provided externally. A systematic way to assess agents' skill is still lacking.

- Solvability and uniqueness of both the set of simultaneous equations and its implicit version as described in 2.3 are yet to be looked into.

A Appendix

This section used to be a part of the section 1, which grew too large and so part of it had to be moved. The purpose of this appendix is to give a short overview of the subjective probability, measures connected with the concept and infinite-dimensional optimization. If 'expected value with respect to a given distribution,' 'Kullback-Leibler divergence,' 'entropy of a probability distribution,' 'Kerridge inaccuracy,' 'Gateaux differential' or 'Frechet differential' sounds unfamiliar then you might consider reading an appropriate part of this section.

A.1 Preliminaries

Though not stated explicitly all random variables in this appendix are assumed to be real-valued unless it makes no difference as in the definition of the Kullbak-Leibler divergence.

All formulas in this section can be expanded onto random vectors by simply writing a $\boldsymbol{x}$ for a multidimensional real variable instead of a one-dimensional x.

A.2 Several distributions of one random variable

Recall that in probability theory **random variable** X is defined as a measurable mapping

$$X : (\Omega, \mathcal{A}, P) \to (S, \mathcal{S}), \tag{26}$$

where $(\Omega, \mathcal{A}, P)$ is a probability space and $(S, \mathcal{S})$ is a measurable space called **observation space.**

The mapping X induces a pushforward measure

$$P_X(B) = P(X^{-1}(B)) \text{ for } B \in \mathcal{S} \tag{27}$$

on the space $(S, \mathcal{S})$, which is called **probability distribution** of X. This definition makes no room for possibility of one random variable to have two different distributions.

However, in information theory and decision-making theory this possibility is allowed. Justification for this is that when we try to apply the notion of a random variable to real-world we might have different opinions about how it is distributed. While we can consider an indicator of tomorrow mean temperature being above zero to be a random variable X defined as follows:

$$X = \begin{cases} 1, & \text{if that temperature is above zero,} \\ 0, & \text{otherwise,} \end{cases}$$

we couldn't possibly *know* on what probability space this random variable is, and, therefore, we couldn't *know* what distribution it has.

Hence, when someone says that X has Bernoulli distribution with parameter, say, 0.5 they merely express their opinion of how likely it is for temperature to be above zero tomorrow. There is no objectivity in that, and there couldn't be any since we cannot observe random variables, but only their realizations. Consequently, someone else might say that X has Bernoulli distribution with parameter 0.1 and be in their right to do so. While it contradicts the definitions of a random variable and a probability distribution as stated above, it still makes sense to refer to X as one random variable with multiple subjective probability distributions. In this case we use the following definitions (sometimes called subjective) of a random variable.

Definition A **random variable** is an unknown outcome of a particular event.

To apply the probability theory we treat such unknown outcomes as random variables by the definition of the probability theory. We attribute properties of these probability-theory variables back to the said outcomes. This approach yields definitions like the one below.

Definition Assume that a random variable (in the sense of the subjective definition) X is treated by someone as a random variable (in the sense of the probability theory) with the distribution P. Then we call P a **probability distribution** in the sense of the subjective definition.

These definitions allow us to work with several opinions about how some random variable is distributed and use the whole probability theory apparatus at the same time. Also, these definitions may be a little confusing. Fortunately (and hopefully) it brings no substantial complications to their application.

Since we allowed random variables to have multiple distributions we cannot simply say 'expected value of X' anymore, but have to say 'expected value of X with respect to distribution P' or 'expected value of X with respect to pdf f_P' instead. To respect this the following notation is used:

$$\mathrm{E}_P\, X \text{ or } \mathrm{E}_{f_P}\, X \text{ respectively.}$$

At the same time as we allow subjective probability distributions we also assume that each random variable does have the true objective distribution that cannot be observed. Let's imagine a lottery that issues one thousand tickets numbered from one to one thousand, only one of which will win tomorrow when the winning number X is announced. Say, we buy one ticket and try to weigh our chances of winning. It is natural to apply what is called the principle of indifference and assign the same probability 1/1000 to each number. If we do so then our subjective probability distribution P_{ours} is

$$P_{subj}(X = k) = 0.001 \text{ for } k = 1, \ldots, 1000.$$

It is not that much improbable that organizers are lazy enough to pick the winning number by simply flipping the coin: if it comes up heads then the winning number is 257, otherwise it is 312. Presumably said organizers assume that their coin is perfect and, therefore their (still subjective) distribution P_{org} is

$$P_{org}(X = k) = \begin{cases} 1/2, & \text{if } k = 257 \text{ or } k = 312, \\ 0, & \text{otherwise.} \end{cases}$$

Now imagine that there is some entity – the nature, God or the giant random number generator – that governs every single event in our world by assigning a probability distribution to its outcome, which is the true distribution in our terms. Assume that this entity 'knows' that the probability of the organizers' coin coming coming up heads is exactly $\sqrt{2}/2$ then the true distribution P^{true} of X is

$$P_{org}(X = k) = \begin{cases} \sqrt{2}/2, & \text{if } k = 257, \\ 1 - \sqrt{2}/2, & \text{if } k = 312, \\ 0, & \text{otherwise.} \end{cases}$$

It might be helpful for understanding the notion of the true distribution to picture God playing dice behind any random outcome. The fact that we do not know the law controlling the random variable in question does not mean there is not one. The true distribution is the one that God uses to determine the yet unknown, neither to us nor to Him, outcome of some event.

Though “it seems hard to sneak a look at God’s cards” we can still observe, infer, and make conjectures about probability distributions of real-life events. That is exactly how different subjective distributions of a common random variable are created.

A.3 Kullback-Leibler divergence

Kullback-Leibler divergence[9] is an asymmetric measure of difference between two probability distributions. It was introduced by Kullback and Leibler in [17] as the mean information contained in a random variable X to discriminate from false hypothesis H_1 that $\mathcal{L}(X) = Q$ to the true hypothesis H_0 that $\mathcal{L}(X) = P$.

Definition Suppose P and Q are mutually absolutely continuous (meaning both $P \ll Q$ and $Q \ll P$ are true) distributions of a random variable X and both measures are absolutely continuous with respect to a common probability measure μ on the same space on which P and Q are defined (the observation space of the variable X). Denote

$$f_P = \frac{dP}{d\mu} \text{ and } f_Q = \frac{dQ}{d\mu}$$

the respective Radon-Nikodym derivatives of P and Q with respect to μ. Call then the **Kullback-Leibler divergence** from Q to P the expected value

$$\mathrm{D}\left(P \| Q\right) = \mathrm{E}_P \log \frac{P}{Q} = \int_{supp} f_P(x) \log \frac{f_P(x)}{f_Q(x)}\, d\mu(x),$$ [10]

where s*upp* is the common support of P and Q. Notation $\mathrm{D}\left(f_P \| f_Q\right)$ is used as an equivalent to $\mathrm{D}\left(P \| Q\right)$.

Note that P and Q having a common support is not an additional assumption but an implication of P and Q being mutually absolutely continuous. The latter assumption is easily alleviated by accepting the following convention based on continuity arguments:

[9] This is the term that articles [15] and [14], which this work roots from, and Wikipedia prefer. If you, however, want to find out more about the subject on the Internet you might wanna consider looking for the *relative entropy* instead.

[10] The logarithm in this definition can be taken to any base as long as the same one is used consistently. We assume that the logarithm is taken to base e.

$$\begin{aligned}
0\log\frac{0}{x} &= 0, && \text{if } x \neq 0;\\
x\log\frac{x}{0} &= \infty, && \text{if } x \neq 0;\\
0\log\frac{0}{0} &= 0.
\end{aligned}$$

The last equality actually has nothing to do with the continuity, its purpose is to allow us to, at least formally, take the integral in the definition A.3 over the whole space S and not bother with writing the region of integration all the time.

A corollary of the second equality is that if support of the assumed distribution is a proper subset of the support the true distribution then the Kullback-Leibler divergence is infinite:

$$supp_Q \subsetneq supp_P \Rightarrow \mathrm{D}\left(P \| Q\right) = \infty.$$

It makes sense in Kullback-Leibler terms because if $supp_Q \subsetneq supp_P$ then we will observe with positive probability values of X that are in $supp_P \setminus supp_Q$ and are only compatible with the true distribution P and, thus, have infinite amount of information to discriminate between the two in favor of P.

If both P and Q are discrete probability measures then

$$\mathrm{D}\left(P \| Q\right) = \sum_{i \in supp} P(x) \log \frac{P(x)}{Q(x)},$$

where $P(x) = P(\{x\}), Q(x) = Q(\{x\})$ and $supp$ is the union of supports of P and Q.

In case P and Q are continuous

$$\mathrm{D}\left(P \| Q\right) = \int\limits_{supp} f_P(x) \log \frac{f_P(x)}{f_Q(x)}\, dx,$$

where f_P and f_Q are probability density functions of P and Q respectively and $supp$ is again the union of supports of P and Q.

It can be shown (proof can be found for example in [5]) with the application of Jensen's inequality that for any two probability measures P and Q

$$\mathrm{D}\left(P \| Q\right) \geq 0 \text{ and } \mathrm{D}\left(Q \| P\right) = 0 \Leftrightarrow P = Q.$$

This theorem is usually referred to as **Gibbs' inequality.**

Furthemore, the Kullback-Leibler divergence is well-defined for infinite discrete and continuous random variables as well as for finite discrete variables (as we will see in the next subsection, this is not true for the entropy, which is not always defined for continuous variables). Other property of the Kullback-Leibler divergence is the one that would be expected from a measure of difference between probability distributions - invariance under one-to-one transformations (also not true for the entropy of continuous variables).

The Kullback-Leibler divergence is not a true distance measure since it is not symmetric and does not satisfy the triangle inequality; it is useful, however, to think of it as a some sort of distance between probability distributions. For example, in case we have to choose a distribution Q out of some set of distributions $\boldsymbol{Q}$ it stands to reason to choose the one that is the 'closest' to the true distribution P. That means that we solve for $\hat{Q}$ the following optimization problem:

$$\hat{Q} = \underset{Q \in \boldsymbol{Q}}{argmin}\, \mathrm{D}\left(P \| Q\right).$$

A.4 Entropy

Entropy of a random variable is a measure of the uncertainty contained in it. If we think that a random variable has some distribution then entropy represents how vague is our knowledge about this random variable. The term was introduced by Shannon in [21]and, thus, sometimes referred to as *Shannon's entropy.*

Definition The **entropy** of a discrete random variable X with support $supp_X$ and probability distribution P is

$$\mathrm{H}(X) = \mathrm{H}(P) = -\sum_{x \in supp_X} P(x) \log P(x).$$

If the support of X has n elements then

$$\mathrm{H}(X) \leq \log n,$$

latter being the entropy of a uniformly distributed variable on the same support.

Let X now be an infinite discrete random variable. It can be shown that if a sequence of discrete finite random variables X_n with supports supp_{X_n} converges to X in distribution at the same time as $supp_{X_n} \nearrow supp_X$ then

$$\lim_{n\to\infty} \mathrm{H}(X_n) = -\sum_{x\in supp} P(x) \log P(x).$$

Therefore, if X is an infinite discrete random variable then the same equation can be used to define its entropy as the one used in the definition above.

For each $x \in$ supp_X probability $P(x)$ lies in the interval $(0;1]$ and therefore $-P(x)\log P(x)$ is a positive finite number. Thus, entropy of a discrete random variable is always defined and

$$\mathrm{H}(X) \geq 0.$$

For a continuous random variable X with a pdf f_X Shannon proposed to define entropy as

$$\mathrm{H}(X) = -\int_{\mathbb{R}} f_X(x) \log f_X(x)\, dx$$

with the convention $0 \log 0 = 0$ based on continuity arguments.

This definition, being derived by simply substituting $\sum$ in the definition at the beginning of this subsection with $\int$, lacks some properties of the discrete-case entropy. It is not always defined, not always non-negative and is not invariant under one-to-one transformations. Consequently, this continuous entropy is often referred to as *differential entropy* to emphasize the difference in properties. One of the solutions to this problem would be to use Kullback-Leibler divergence between the distribution in question and some referential distribution instead of entropy. If the support of the variable is bounded then uniform distribution over the same support is a reasonable choice of referential distribution because it contains the greatest amount of uncertainty. No such universal choice exists in case the support is unbounded.

A.5 Maximum entropy principle

The maximum entropy principle states that under given constraints that represent all our knowledge about a random variable X we should choose the distribution with the greatest entropy (see [10] and [11] - the articles by one of the most prominent apologists of the said principle). Denote $\boldsymbol{P}$ the set of all probability distributions that meet mentioned constraints. Then applying the maximum entropy principle means solving for $\hat{P}$ the following optimization problem:

$$\hat{P} = \underset{P \in \boldsymbol{P}}{argmax}\, H(P).$$

The justification for this principle is that apart from the information we have about X there is no other reason to have any certainty about X. Any distribution that has less entropy contains more unsubstantiated certainty and therefore we should not choose this distribution. The maximum entropy distribution contains the least amount of information other than given by our knowledge of X and therefore is the preferred choice. An axiomatic justification for the principle was given in [22].

'Constraints that represent our knowledge' sounds a little vague so it is worth some elaborating. Usually those constrains have a form of either generalized expected values or bounds on them, e.g.

$$\begin{aligned} \mathrm{E}_P\, X &= \beta_1, \\ \mathrm{E}_P(X - \mathrm{E}\, X)^2 &\leq \beta_2, \\ P(X < 0) = \mathrm{E}_P\, \mathbb{I}_{(-\infty;0)} &= \beta_3, \end{aligned}$$

or something else along those lines.

It is often stated that if we have no information about the variable X then the uniform distribution is the one with the most entropy, which can be seen as a justification for the principle of indifference. This statement, while true, is a little imprecise: it is implied that 1. X has a finite (discrete case) or at least bounded (continuous case) support and 2. we know what this support is or at least can be (we might not know that $supp_P = supp$, but only that $supp_P \subset supp$).This might not help understanding the principle, but is important in applying it since outcome may differ considerably depending on the support constraint. Note that statement $supp_P \subset supp$ can be written in terms of expected values as $\mathrm{E}_P\, \mathbb{I}_{supp} = 1$.

Here is the correct statement about uniform distribution being the most entropic one and two other examples of the maximum entropy principle application.

1. Out of all distributions P with the support $supp_P \subset supp$, where $supp$ is either finite or bounded set, the uniform distribution $U(supp)$ is the one with the maximum entropy.

2. Out of all non-negative distributions with the expected value $1/\lambda$ the one with the most entropy is the exponential distribution $Exp(\lambda)$.
3. Out of all real-valued distributions with the expected value μ and the variance σ^2 the one with the highest entropy is the normal distribution $N(\mu, \sigma^2)$.

A.6 Kerridge inaccuracy

Kerridge inaccuracy is, like Kullback-Leibler divergence, is an asymmetric measure of 'difference' between two probability distributions. It was introduced by Kerridge in the article [16] as a measure combining the amount of mistake made by assuming that the random variable in question has distribution Q when its true distribution is P and uncertainty contained in the true distribution P.

Definition The **Kerridge inaccuracy** $\mathrm{K}\left(P \,\|\, Q\right)$ from a distribution Q to a distribution P is defined as

$$\mathrm{K}\left(P \,\|\, Q\right) = \mathrm{D}\left(P\|Q\right) + \mathrm{H}(P).$$

The Kerridge inaccuracy $\mathrm{K}\left(P \,\|\, Q\right)$ is defined if and only if the right side of the equation above is defined, which means not only that the entropy $\mathrm{H}(P)$ has to be defined, but also that $\mathrm{H}(P) = -\infty$ and $\mathrm{D}\left(P\|Q\right) = \infty$ cannot be true at the same time.

Provided that $\mathrm{K}\left(P \,\|\, Q\right)$ exists and that we accept the following convention:

$$\begin{aligned} &x \log 0 = -\infty, \text{ if } x \neq 0, \\ &0 \log 0 = 0, \end{aligned}$$

we can write in case both Q and P are discrete distributions (denote s*upp* the union of supports of Q and P)

$$\mathrm{K}\left(P \,\|\, Q\right) = -\sum_{x \in supp} P(x) \log Q(x),$$

and if Q and P are continuous distributions with pdfs f_Q and f_P respectively then

$$\mathrm{K}\left(P \,\|\, Q\right) = -\int\limits_{x \in supp} f_P(x) \log f_Q(x)\, dx.$$

Properties of the Kerridge inaccuracy are a corollary of the properties of the Kullback-Leibler divergence and entropy. One that might be worth repetition though is that if the support of an assumed distribution is smaller than the support of the true distribution then the Kerridge inaccuracy (given it is well-defined) from the first one to the second is infinite:

$$supp_Q \subsetneq supp_P \Rightarrow \mathrm{K}\left(P \,\|\, Q\right) = \infty.$$

In some case searching for an optimum distribution may be reformulated as an optimization problem involving minimization of the Kullback-Leibler divergence from Q to P given a certain distribution P and some constraints. Assuming that P has a finite entropy mentioned optimization problem is equivalent to the minimization of the Kerridge inaccuracy from Q to P under the same constraints. The latter statement is an immediate corollary of the definition of the Kerridge measure.

A.7 Infinite-dimensional differentiation and optimization

Applying the principle of maximum entropy in continuous case leads to a bit unusual optimization problems, in which optimization is conducted for a variable which is infinite-dimensional, namely a probability density function. Fortunately the following two definitions and two propositions allow generalization of usual finite-dimensional optimization methods to infinite-dimensional spaces.

Definition Let X be a vector space, Y a normed space, T a transformation defined on a domain $D \subset X$ having range $R \subset Y$. Let $x \in D$ and $h \in X$ be arbitrary elements and let $x + \alpha h \in D$ for all α sufficiently small. If the limit

$$\partial T(x; h) = \lim_{\alpha \to 0} \frac{1}{\alpha}[T(x + \alpha h) - T(x)]$$

exists, it is called the **Gateaux differential** of T at x with increment h. If this limit exists for each $h \in X$, the transformation T is said to be **Gateaux differentiable** at x.

Definition Let X be a vector space, Y a normed space, T a transformation defined on an open domain $D \subset X$ having range $R \subset Y$. If for each $x \in D$

and each $h \in X$ there exists $\partial T(x;h) \in Y$, which is linear and continuous with respect to h such that

$$\lim_{||h|| \to 0} \frac{||T(x+h) - T(x) - \partial T(x;h)||}{||h||} = 0,$$

then T is said to be **Frechet differentiable** at x and $\partial T(x;h)$ is said to be the **Frechet differential** of T at x with increment h. The Frechet differential for a fixed $x \in D$ by the definition can be written as $\partial T(x;h) = \langle A_x, h \rangle$,[11] where A_x is the corresponding unique linear operator. The correspondence $x \to A_x$ defines a transformation, which is called the **Frechet derivative** T' of T.

Proposition A.1 *If the Frechet differential exists then the Gateaux differential exists as well and the two are equal.*

Proposition A.2 *Let the real-valued functional f have a Gateaux differential on a vector space X. A necessary condition for f to have an extremum at $x_0 \in X$ is that $\partial f(x_0;h) = 0$ for all $h \in X$.*

For more on the topic see [18], from which the definitions and the propositions above are taken.

[11] We use notation $\langle *, * \rangle$ for scalar product.

References

[1] Anděl J.: *Základy matematické statistiky*, Matfyzpress, 2009;

[2] Bican, L.: *Lineární algebra a geometrie,* Nakladatelství Academia, 2009;

[3] Bickel, P.J., Doksum, K.A.: *Mathematical Statistics: Basic and Selected Topics, volume I, second edition*, Pearson Prentice-Hall, p. 32, 2009;

[4] Buchanan, D., Huczynski, A.: *Organisational behaviour, introductory text,* Prentice Hall Third Edition, 1997;

[5] Cover, T.M., Thomas, J.A.: *Elements of Information Theory*, John Wiley and Sons, 1991;

[6] Cowell, R.G., Dawid, A.P., Lauritzen, S.L., Spiegelhalter, D.J.: *Probabilistic networks and expert systems*, Springer, 2003, 2nd edition.

[7] DeGroot, M.H.: *Optimal Statistical Decisions,* McGraw-Hill, New York, 1970;

[8] Genest, C., McConway, K.J.: *Allocating the weights in the linear opinion pool*, Journal of Forecasting, vol. 9, no. 1, pp. 53-73, 1990;

[9] Genest, C., Zidek, J.V.: *Combining Probability Distributions: A Critique and an Annotated Bibliography*, Statistical Science, Vol. 1, No. 1., pp. 114-135, 1986;

[10] Jaynes, E. T.: *Information Theory and Statistical Mechanics*, Physical Review Series II, 106 (4), pp. 620-630, 1957;

[11] Jaynes, E.T.: *Where Do We Stand on Maximum Entropy?*, in The Maximum Entropy Formalism, The MIT Press, Cambridge, MA, pp. 15-118, 1979;

[12] Jensen, F.V.: *Bayesian Networks and Decision Graphs*, Springer-Verlag, New York, 2001.

[13] Kapur, J. N., and Kesevan, H. K.: *Entropy optimization principles with applications*, Boston: Academic Press, 1992;

[14] Kárný, M.: *Knowledge elicitation via extension of fragmental knowledge pieces* in ECC'09, Budapest, 2009, accepted;

[15] Kárný, M.,Guy, T.V., Bodini, A., Ruggeri, F.: *Cooperation via sharing of probabilistic elements*, IJCIStudies, vol. II, no. 2, 2009, accepted;

[16] Kerridge, D.F.: *Inaccuracy and Inference*, Journal of the Royal Statistical Society. Series B (Methodological), Vol. 23, No. 1, pp. 184-194, 1961;

[17] Kullback S., Leibler R. A.: *On Information and Sufficiency*, The Annals of Mathematical Statistics, Vol. 22, No. 1, pp. 79-86, March 1951;

[18] Luenberger, D.G.: *Optimization by Vector Space Methods*, John Wiley and Sons, 2009;

[19] O'Hagan, A, Buck, C.E., Daneshkhah, A, Eiser, J.R., Garthwaite, P.H., Jenkinson, D.J., Oakley, J. and Rakow, T.: *Uncertain judgment: eliciting expers' probabilities*, John Wiley & Sons, 2006.

[20] Sečkárová, V.: Supra-Bayesian Combination of Probability Distributions, master's thesis, Charles University in Prague, 2010;

[21] Shannon, C.E.: A Mathematical Theory of Communication, Bell System Technical Journal, vol. 27, pp. 379-423, 623-656, July, October, 1948;

[22] Shore, J. E.; Johnson, R. W.: *Axiomatic Derivation of the Principle of Maximum Entropy and the Principle of Minimum Cross-Entropy*, IEEE Transactions on Information Theory, Volume IT-26, p. 26-37, 1980;

[23] Thompson B.: *A Critique of Bayesian Inference* in Lecture Notes in Statistics, vol. 189, pp. 84-96, 2007.