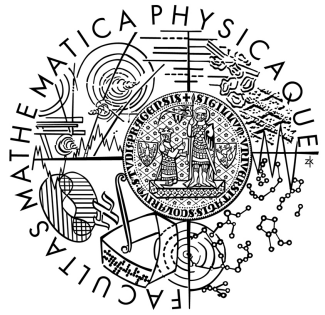
Univerzita Karlova v Praze
Matematicko-fyzikální fakulta

DIPLOMOVÁ PRÁCE



Vladimíra Sečkárová

Supra-bayesovská kombinace pravděpodobnostních distribucí

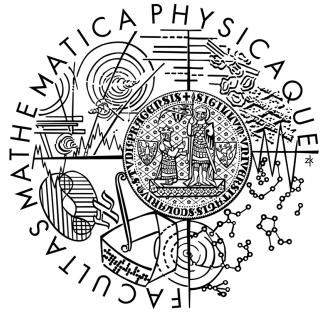
Katedra pravděpodobnosti a matematické statistiky

Vedoucí diplomové práce: Ing. Miroslav Kárný, DrSc

Studijní program: Matematika, Pravděpodobnost, matematická statistika a ekonometrie

Charles University in Prague
Faculty of Mathematics and Physics

DIPLOMA THESIS



Vladimíra Sečkárová

Supra-Bayesian Combination of Probability Distributions

Department of Probability and Mathematical Statistics

Supervisor: Ing. Miroslav Kárný, DrSc

Study Program: Mathematics, Probability, Mathematical Statistics and Econometrics

I would like to express sincere gratitude to my supervisor, Ing. Miroslav Kárný, DrSc, for his immense patience throughout my work on this thesis. I would also like to thank my mum and my friends for their limitless support.

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

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, April 16, 2010 Vladimíra Sečkárová

Contents

Název práce: Supra-bayesovská kombinace pravděpodobnostních distribucí
Autor: Vladimíra Sečkárová
Katedra: Katedra pravděpodobnosti a matematické statistiky
Vedoucí diplomové práce: Ing. Miroslav Kárný, DrSc
e-mail vedoucího: school@utia.cas.cz

Abstrakt: Tato práce sa zaobírá tematikou sdílení pravděpodobnostních informací s využitím Supra-bayesovského přístupu. V 1. kapitole jsou uvedeny všechny potřebné metody a vzorce používány dále v textu. 2. kapitola obsahuje úvod do dané problematiky. Ve 3. kapitole je odvozena hlavní metoda sdílení informace založené na rovnaké doméně. Ve 4. kapitole jsou specifikovány druhy dodané informace, které jsou pak transformovány do pravděpodobnotních termínů a následně rozšířeny na celou doménu. V 5. kapitole jsou zhodnoceny výsledky dosáhnuté v předešlých kapitolách.

Klíčová slova: Bayesovské rozhodování, sdílení pravděpodobnostních informací, Supra-bayesovský přístup

Title: Supra-Bayesian Combination of Probability Distributions
Author: Vladimíra Sečkárová
Department: Department of Probability and Mathematical Statistics
Supervisor: Ing. Miroslav Kárný, DrSc
Supervisor's e-mail address: school@utia.cas.cz

Abstract: In this work we study problems of sharing of probabilistic information by using Supra-Bayesian approach. In 1st Chapter the methods and formulas used in the work are mentioned. 2nd Chapter contains the introduction to discussed topic. In 3rd Chapter the main method of sharing the probabilistic information, which is based on common domain, is derived. In 4th Chapter the types of given knowledge pieces are specified, which are then transformed into probabilistic terms and extended on the whole domain. In 5th Chapter the results from the previous chapters are assessed.

Keywords: Bayesian decision making, sharing of probabilistic information, Supra-Bayesian approach

Chapter 1

Preliminaries

1.1 Random vector

In all following sections we suppose that $m \in \mathbb{N}$, $m < \infty$.

Definition 1.1 (Random variable)**.** *Let* $(\Omega, \mathcal{A}, \mathrm{P})$ *be a probability space,* $(\mathcal{X}, \mathcal{B})$ *be a measurable space. Then the mapping* $X : \Omega \to \mathcal{X}$ *is a* ***random variable*** *taking values in* $(\mathcal{X}, \mathcal{B})$*, if it is measurable, which means:*

$$[X \in B] := \{\omega \in \Omega : X(\omega) \in B\} = X^{-1}(B) \in \mathcal{A} \quad \forall B \in \mathcal{B}.$$

A discrete random variable maps the events to values in a finite or countable set. A continuous random variable maps the events to values in an uncountable set.

Definition 1.2 (Random vector)**.** *For random variables* X_k*,* $k = 1, \ldots, m$*, taking values in* $(\mathcal{X}_k, \mathcal{B}_k)$ *we say that* $\boldsymbol{X} = (X_1, X_2, \ldots, X_m)$ *is a* ***(m-dimensional real-valued) random vector****, if it is a random variable taking values in* $(\times_{k=1}^{m} \mathcal{X}_k, \otimes_{k=1}^{m} \mathcal{B}_k)$.

A discrete random vector is a multidimensional case of mapping events to values in a finite or countable set. In this work we assume the set is always finite. A continuous random vector is a multidimensional case of mapping events to values in an uncountable set.

The following theorem guarantees the measurability of the random vector.

Theorem 1.3 (Collection of random variables)**.** *Let* $(\Omega, \mathcal{A}, \mathrm{P})$ *be a probability space. Let for* $k = 1, 2, \ldots, m$ *the measurable spaces* $(\mathcal{X}_k, \mathcal{B}_k)$ *and random variables* $X_k : (\Omega, \mathcal{A}) \to (\mathcal{X}_k, \mathcal{B}_k)$ *be given. Then* $(X_1, X_2, \ldots, X_m)$ *is a random vector taking values in* $(\times_{k=1}^{m} \mathcal{X}_k, \otimes_{k=1}^{n} \mathcal{B}_k)$.

Proof. See Lachout (2004), Theorem 2.1. □

1.2 Probability mass function, probability density function

Throughout this text we suppose that for each considered space the assumptions of Radon-Nikodým theorem (see Anděl (2007), Theorem 3.1) are satisfied; so for each considered random vector the probability mass function (pmf) or probability density function (pdf), with respect to the corresponding measure, does exist.
Let the probability space $(\Omega, \mathcal{A}, \mathrm{P})$ be given.

Definition 1.4 (Probability mass function for discrete random vector)**.** *Let* $\mathbf{X} = (X_1, \ldots, X_m) : (\Omega, \mathcal{A}) \to (\mathbb{R}^m, \mathcal{B}^m)$ *be a real-valued discrete random vector with possible realizations* $\{\mathbf{x}_i\}_{i=1}^n$, $n < \infty$, $\mathbf{x}_i \in \mathbb{R}^m$. *The distribution of* $\mathbf{X}$ *can be described by the* ***probability mass function (pmf)****, which is defined as:*

$$\begin{aligned} f(\mathbf{x}) &= \mathrm{P}(\mathbf{X} = \mathbf{x}) && \text{if } \mathbf{x} \in \{\mathbf{x}_i\}_{i=1}^n \\ &= 0 && \text{otherwise.} \end{aligned} \tag{1.1}$$

The distribution of discrete random vector can be described also by the probability vector:

$$(f(\mathbf{x}_1), \ldots, f(\mathbf{x}_n)),$$

where

$$\sum_{i=1}^{n} f(\mathbf{x}_i) = 1, \quad f(\mathbf{x}_i) \geq 0, \quad \text{for } i = 1, \ldots, n.$$

The pdf of continuous random vector $\mathbf{X}$ will be, for possible realizations $\mathbf{x} \in \mathbb{R}^m$, denoted by $f(\mathbf{x})$.
In following text it will be clear from context whether $f(\mathbf{x})$ denotes pmf or pdf.

1.3 Expected value of a transformed random vector

In this work we often evaluate the expectation of a transformed random vector. To do this we use the results of Theorem 1.6 based on Theorem 1.5.

Theorem 1.5. *Let* t *be a measurable mapping from* $(\Omega, \mathcal{A}, \mathrm{P})$ *to a measurable space* $(\Lambda, \mathcal{D})$. *Let* g *be a measurable function on* $(\Lambda, \mathcal{D})$ *and* Q *be a measure induced by mapping* t*:* $Q(D) = \mathrm{P}\{t^{-1}(D)\}$ *for* $D \in \mathcal{D}$. *Then, it holds:*

$$\int_{\Omega} g[t(\omega)] \mathrm{dP}(\omega) = \int_{\Lambda} g(t) \mathrm{d}Q(t) \tag{1.2}$$

Proof. Main steps of the proof are given in Anděl (2007). □

Suppose $\mathbf{X}$ is a random vector, the pdf or pmf (in both cases denoted by f) of which exists. Under this assumption, we bring the following theorem:

Theorem 1.6. *If* $t : R^m \to R^1$ *and* $\mathrm{E}|t(\mathbf{X})| < \infty$, *then*
for discrete distribution:

$$\mathrm{E}\ t(\mathbf{X}) = \sum_{i=1}^{n} t(\mathbf{x}_i)\mathrm{P}(\mathbf{X} = \mathbf{x}_i) = \sum_{i=1}^{n} t(\mathbf{x}_i) f(\mathbf{x}_i) \tag{1.3}$$

for continuous distribution:

$$\mathrm{E}\ t(\mathbf{X}) = \int_{R^1} t(\mathbf{x}) f(\mathbf{x}) \mathrm{d}\mathbf{x}. \tag{1.4}$$

Proof. See Dupač and Hušková (2005). □

In this work we also often use the following notation expressing the result of the Theorem 1.6:

$$\mathrm{E}_{f(\mathbf{x})} t(\mathbf{X}), \tag{1.5}$$

which stresses that the expectation of the transformed random vector $t(\mathbf{X})$ is taken with respect to a specific pmf or pdf $f(\mathbf{x})$.

1.4 Optimal estimate as minimizer of conditional expectation of loss function

Bayesian methods represent one of the basic approaches to solutions of statistical problems. In Hušková (1985), the basics of their usage in estimation theory are introduced. Suppose that:

- $\mathbf{X}$ is a random vector with possible realizations $\{\mathbf{x}_i\}_{i=1}^n$, $n < \infty$,
- $h = (h(\mathbf{x}_1), \ldots, h(\mathbf{x}_n))$ is an unknown parameter from a nonempty set $H = \{h : \quad \sum_{i=1}^n h(\mathbf{x}_i) = 1, \quad h(\mathbf{x}_i) \geq 0, \quad i = 1, \ldots, n\}$,
- $g_j = (g_j(\mathbf{x}_1), \ldots, g_j(\mathbf{x}_n))$ is a random vector taking values in a non empty set $G_j = \{g_j : \quad \sum_{i=1}^n g_j(\mathbf{x}_i) = 1, \quad g_j(\mathbf{x}_i) \geq 0, \quad i = 1, \ldots, n\}, \quad j = 1, \ldots, s$,

- $(g_1^T, \ldots, g_s^T)^T \in \times_{j=1}^{s} G_j$ is $(s \times n)$ matrix consisting of random vectors g_j, (^{T} denotes transposition),
- D is a possible realization of the above – $(s \times n)$ matrix,
- the conditional pdf $\pi(h|D)$ does exist,
- $\widehat{H}$ is a set of all possible decisions (conclusions) about the parameter h, $\widehat{H} \subseteq H$,
- $\hat{h}$ is an element of $\widehat{H}$,
- $L(h, \hat{h}) : H \times \widehat{H} \to \mathbb{R}^1$ is a loss function; it expresses how much loss we sustain by accepting the decision $\hat{h}$ when the true value of parameter is h,
- there exists $k > -\infty$ that: $L(h, \hat{h}) \geq k \ \forall h \in H$ and $\forall \hat{h} \in \widehat{H}$,
- $\delta : \times_{j=1}^{s} G_j \to \widehat{H}$ is a decision function,
- $\delta(D)$ for $D \in \times_{j=1}^{s} G_j$ is a decision about parameter h if $(g_1^T, \ldots, g_s^T)^T = D$,
- $\mathrm{R}(h, \delta)$ is the risk for pertaining the decision function δ if the true value of the parameter is h risk is defined as:

$$\mathrm{R}(h, \delta) = \mathrm{E}[\mathrm{L}(h, \delta)|h]$$

- Δ is a set of all Bayesian decision functions δ for which $\mathrm{R}(h, \delta) < \infty$.

In Hušková (1985) it is shown that if we find a value ${}^{O}\delta(D)$ (D is fixed) satisfying:

$$\begin{aligned} {}^{O}\delta(D) &= \mathrm{Arg} \min_{\delta(D) \in \widehat{H}} \mathrm{E}[\mathrm{L}(h, \delta\left((g_1^T, \ldots, g_s^T)^T\right))|(g_1^T, \ldots, g_s^T)^T = D] \\ &= \mathrm{Arg} \min_{\delta(D) \in \widehat{H}} \mathrm{E}[\mathrm{L}(h, \delta(D))] = \mathrm{Arg} \min_{\delta(D) \in \widehat{H}} \int_H \mathrm{L}(h, \delta(D)) \pi(h|D) \mathrm{d}h, \end{aligned} \tag{1.6}$$

then corresponding ${}^{O}\delta$ is the optimal Bayesian decision function. Meaning value of ${}^{O}\delta(D)$ can be found as minimizer of conditional expectation of loss function with respect to posterior pdf $\pi(h|D)$ – that means under assumption $(g_1^T, \ldots, g_s^T)^T = D$. This conclusion was made under assumption h is a random vector.

In Hušková (1985) we also find that the task of estimating an unknown parameter can be interpreted as a statistical decision problem (H, Δ, R) (see Hušková (1985)), where the set of possible decisions $\widehat{H}$ about h coincides with H. Decision function δ provides the estimate $\hat{h}$ of the unknown parameter h and Δ is a set of estimates of parameter h. The loss function $\mathrm{L}(h, \delta)$ expresses the inaccuracy of estimate δ on true value of h.

1.5 Maximum entropy principle as proper method for inference

Suppose that the random vector has a set of possible outcomes $\mathbf{x}_i$ with unknown probabilities $q(\mathbf{x}_i)$, $q = \{q(\mathbf{x}_i)\}_i$, and we know the constraints on q: values of certain expectations $\sum_i q(\mathbf{x}_i) f_k(\mathbf{x}_i)$ or bounds on these values. Suppose then you need to choose $q^* = \{q^*(\mathbf{x}_i)\}_i$ that is in some sense the best estimate of q given what we know. Usually there remains an infinite set of distributions that are not ruled out by the constraints. Question that arises which one we will choose.

The principle of maximum entropy states that from the set of all distributions satisfying the constraints we should choose the one with largest entropy; the distribution for which is $-\sum_i q^*(\mathbf{x}_i) \log q^*(\mathbf{x}_i)$ maximal. Maximization of the entropy as a general inference procedure was firstly proposed in Jaynes (1957). Despite the success of the principle, it remained controversial: the controversy appeared in the foundations of the principle, because it was usually justified on the basis of entropy's unique properties. None of the justifications of the maximum entropy principle mentioned in Shore and Johnson (1980) is based on a formal description of what is required of a method for taking information into account. Since the principle is asserted as a general method of inductive inference, it is reasonable to require that different ways of using it while taking some information into account should lead to consistent results. This requirement is formalized in four consistency axioms. They are all based on one fundamental principle: if a problem can be solved in more than one way, the result should be consistent. They can informally be phrased as follows:

i) *Uniqueness*: The result should be unique.

ii) *Invariance*: The choice of coordinate system should not matter.

iii) *System of Independence*: It should not matter whether one accounts for independent information about independent systems separately in terms of different distributions or together in terms of a joint distribution.

iv) *Subset Independence*: It should not matter whether one treats an independent subset of system states in terms of a separate conditional or in terms of the full system distribution.

The axioms are stated in terms of an abstract information operator, they make no reference to information measures.

In Shore and Johnson (1980) is then proved that: given a new constraint information there is only one distribution satisfying the considered constraints that can be chosen

by a procedure satisfying the consistency axioms; this unique distribution maximizes the entropy.

1.6 Basic properties of Kerridge inaccuracy

Suppose $p = \{p(\mathbf{x}_i)\}_i$ are probabilities of outcomes $\mathbf{x}_i$ of an experiment provided by some information source, while $q = \{q(\mathbf{x}_i)\}_i$ are true probabilities of these outcomes, $\sum_i q(\mathbf{x}_i) = \sum_i p(\mathbf{x}_i) = 1$. In Kerridge (1961) it is shown that inaccuracy of the opinion p can be measured by:

$$\mathrm{K}(q, p) = -\sum_i q(\mathbf{x}_i) \log p(\mathbf{x}_i).$$

The key property of Kerridge inaccuracy, which will be useful for us, is:
The value of $\mathrm{K}(q, p)$ is minimal, for fixed q, when $q(\mathbf{x}_i) = p(\mathbf{x}_i)\ \forall i$, that is:

$$q = \mathrm{Arg}\ \min_p \mathrm{K}(q, p). \tag{1.7}$$

For details see Kerridge (1961).

1.7 Basic properties of Kullback-Leibler divergence

Kullback-Leibler divergence, also known as relative entropy, between two pdfs $f(\mathbf{x})$ and $g(\mathbf{x})$ of a continuous random vector $\mathbf{X}$, is defined as:

$$D(p||q) = \int p(\mathbf{x}) \log \frac{p(\mathbf{x})}{q(\mathbf{x})} \mathrm{d}\mathbf{x}. \tag{1.8}$$

It is commonly used in statistics as a measure of similarity between two probability distributions. The divergence satisfies three basic properties:

- *Self-similarity*: $D(f||f) = 0$,
- *Self-identification*: $D(f||g) = 0$ only if $f = g$ *a.e.*,
- *Positivity*: $D(f||g) \geq 0\ \forall\ f,\ g$.

For more details, see Kullback (1997) and Kullback and Leibler (1951).

1.8 Dirichlet distribution

Suppose we have n ($n < \infty$) different realizations $\mathbf{x}_1, \ldots, \mathbf{x}_n$ of discrete random vector $\mathbf{X}$. Let $h(\mathbf{x}_i)$ denote a probability of $\mathbf{x}_i$. Suppose that we observed $\mathbf{X}$ k times. Let Y_i denote the number of realizations of $\mathbf{x}_i$, which appeared in k observations. The joint distribution of random vector $(Y_1, \ldots, Y_n)$ is the multinomial distribution $M(k; h(\mathbf{x}_1), \ldots, h(\mathbf{x}_n))$ with parameters $h(\mathbf{x}_1), \ldots, h(\mathbf{x}_n)$, which are given by $\sum_{i=1}^{n} h(\mathbf{x}_i) = 1$, $h(\mathbf{x}_i) \geq 0$. The conjugate prior distribution of the parameters of the multinomial distribution is the (continuous) Dirichlet distribution, which is a multivariate generalization of the beta distribution.
The pdf of this distribution for variables $h(\mathbf{x}_1), \ldots, h(\mathbf{x}_n)$, specified by parameters $\nu_1, \ldots, \nu_n$, is given as follows:

$$f(h(\mathbf{x}_1), \ldots, h(\mathbf{x}_n)) = \frac{1}{Z(\nu_1, \ldots, \nu_n)} \prod_{i=1}^{n} h(\mathbf{x}_i)^{\nu_i - 1}. \tag{1.9}$$

where $h(\mathbf{x}_1), \ldots, h(\mathbf{x}_n) \geq 0$, $h(\mathbf{x}_1) + \ldots + h(\mathbf{x}_n) = 1$ and $\nu_1, \ldots, \nu_n > 0$.
The parameters ν_i can be interpreted as "prior observation counts" for events $\mathbf{x}_i$ governed by $h(\mathbf{x}_i)$.
The normalizing factor is

$$Z(\nu_1, \ldots, \nu_n) = \frac{\prod_{i=1}^{n} \Gamma(\nu_i)}{\Gamma(\nu_0)},$$

where Γ denotes Euler gamma function and

$$\nu_0 = \sum_{i=1}^{n} \nu_i.$$

The mean and variance of Dirichlet distribution are:

$$\mathrm{E}\left[h(\mathbf{x}_i)\right] = \frac{\nu_i}{\nu_0} \tag{1.10}$$

$$\mathrm{Var}\left[h(\mathbf{x}_i)\right] = \frac{\nu_i(\nu_0 - \nu_i)}{\nu_0^2(\nu_0 + 1)}. \tag{1.11}$$

Chapter 2

Introduction

Consider a group of knowledge sources (e.g. human beings, ...), where each of them provides a knowledge piece over its domain (quantities considered by it) and wants to improve its knowledge using the knowledge coming from others. The question that arises: how to do this?
When solving this task we focus on the first source. It does not restrict the generality of the solution as the proposed method can be then used for any other source.

2.1 Outline of the method

The treatment of improving of the first source's knowledge discussed in this work is based three important steps:

1. *Use knowledge pieces from specific sources*

   It is reasonable to use the knowledge from sources, which are somehow connected with the first source: concretely, their domains have a non empty intersection with the domain of the first source. Such sources will be called neighbors.

2. *Focus on merging of the given knowledge pieces*

   If the domains of considered sources are the same and knowledge pieces provided by sources have the same (later specified) form, then construct the **optimal merger** of them. Since the domains of considered sources are not necessarily the same, follow these steps:

   – consider the union of domains of the first source and its neighbors,

   – extend the available knowledge pieces on this union,

– construct the optimal merger of extended knowledge pieces.

(The constructed merger is in fact the estimate of pmf or pdf of the common domain.)

3. *Project the constructed estimate on the domain of the first source and give this projection back to it.*

2.2 Construction of the optimal merger: used theory and existing results

For construction of the optimal merger, we interpret it as an estimate of the "objective" pdf or pmf on the union of sources' domains. For estimation, we use decision theory as a general method that suits well to our purpose and it is often used for solving estimation problems. In combination with Bayesian methodology it is an useful tool for making conclusions under uncertainty: for construction of Bayesian decision function the unknown (inaccessible) quantities are treated as random variables and the joint distribution of them based on available data (knowledge pieces) is constructed and analyzed. If we take the knowledge pieces from the first source and its neighbors as given data, then by inserting them into this distribution various characteristics of resulting conditional (posterior) distribution can be evaluated.
Previously mentioned way is well established and elaborated if data are given as concrete (observed) values. It is obvious that data also can be of another form. For instance, marginal or conditional distributions describing "ordinary" data or their generalized moments. The question that arises is: how to exploit knowledge pieces given in another form?

No systematic treatment of incompletely compatible knowledge pieces have been given yet. In recently published papers Kárný, M and Guy, T.V. and Bodini, A. and Ruggeri, F. (2009) and Kárný (2009) it is suggested that a Supra-Bayesian approach, see Genest and Zidek (1986), could give a systematic solution. This approach expresses the task of combining the given knowledge pieces as the task of constructing a posterior pmf or pdf for a fictitious decision maker by using Bayes' theorem (see Hušková (1985)). The given knowledge pieces are used as a random data and the ideal merger to be estimate as unknown parameter. Both works Kárný, M and Guy, T.V. and Bodini, A. and Ruggeri, F. (2009) and Kárný (2009) use this Supra-Bayesian approach, but they differ in relating knowledge pieces to the ideal merger, called "supra-model". Results in these works are promising, but suffering from the following problems:
i) we will not get a Bayesian rule from constructed optimal merger, when "ordinary"

joint pdf conditional pdf moments realizations	of	a set of all r.v. $\mathbf{X}$ a subset of all r.v. $\mathbf{X}$

Table 2.1: Possible forms of knowledge pieces

data (data values) and parametric model are used, and
ii) the resulting merger is given by an implicit formula, solvability of which has not been established.

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

2.3 Towards decision-theoretical problem formalization

Let the domain of each source be represented by a set of discrete random variables. Without loss of generality, we again focus on the first source.

Definition 2.1 (Neighbor of the first source)**.** *A neighbor of the first source is a source, whose set of considered random variables has a non empty intersection with the set considered by the first source. The first source and its neighbors are labeled by j, $j = 1, \ldots, s$.*

Assumption 2.2. *Every source from the group of sources has a finite amount of neighbors.*

In our case it holds: $s < \infty$. The aim of the first source – improvement of its knowledge – is reached by using the knowledge pieces from its neighbors. All available knowledge pieces are somehow processed and the result of this processing (projected on the first source's domain) is given back to the first source (see Section 2.1 and Figure 2.1).

Let the union of sets of random variables considered by s sources be a finite collection of discrete random variables denoted by $\mathbf{X} = (X_1, \ldots, X_m)$, possible realizations of which are denoted by $\mathbf{x}_1, \ldots, \mathbf{x}_n$, $n < \infty$. We consider following forms of knowledge pieces provided by sources:

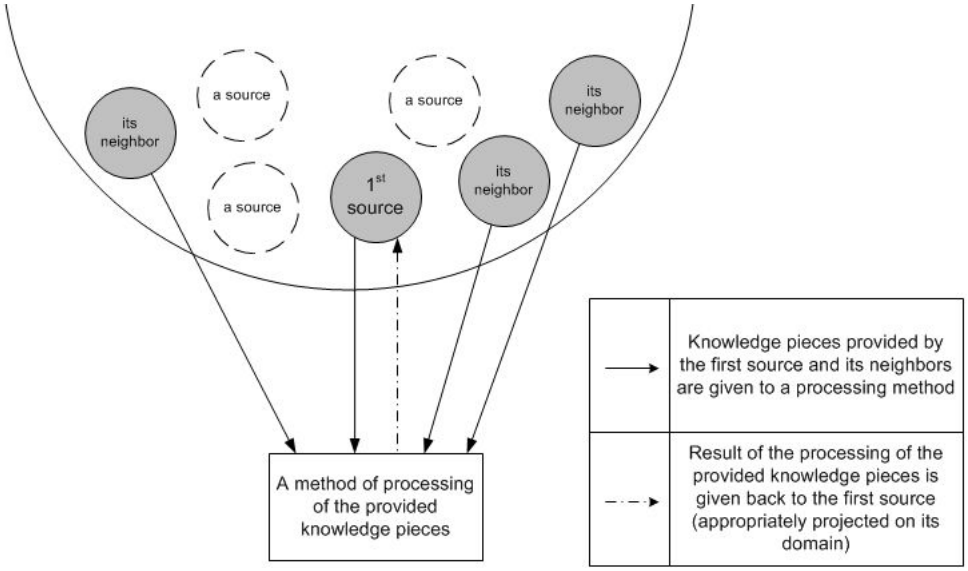


Figure 2.1: Simple graphic layout of proposed method

Now we focus on the part of improving first source's knowledge, where the optimal merger of available knowledge pieces is constructed.
We suppose that a joint pmf describing $\mathbf{X}$ exists, see Section 1.2. It is denoted by h and represented by a probability vector $h = (h(\mathbf{x}_1), \ldots, h(\mathbf{x}_n))$, where:

$$\sum_{i=1}^{n} h(\mathbf{x}_i) = 1, \quad h(\mathbf{x}_i) \geq 0, \quad i = 1, \ldots, n. \tag{2.1}$$

Since we have knowledge pieces describing all or a part of $\mathbf{X}$ and we consider the unknown pmf h as the random vector: we can express our task as a statistical decision problem (see Section 1.4). By using Bayesian decision theory we construct the optimal estimate ${}^{O}\hat{h}$ of h. The resulting pmf ${}^{O}\hat{h}$ is the optimal merger of available knowledge pieces. Projection of this estimate on the variables considered by the first source will be then used for improvement of the knowledge of this source.

The following chapters contains details of the construction of the optimal merger: In Chapter 3, we derive the optimal merger (an optimal estimate ${}^{O}\hat{h}$ of h) under assumption that each of the s considered sources (both the first source and its neighbors) gives the knowledge piece about $\mathbf{X}$ in the form of a joint pmf. For jth source and considered discrete case, it is represented by a probability vector: $g_j = (g_j(\mathbf{x}_1), \ldots, g_j(\mathbf{x}_n))$, $j = 1, \ldots, s$. It sounds reasonably, that the construction of ${}^{O}\hat{h}$ would be easier, when

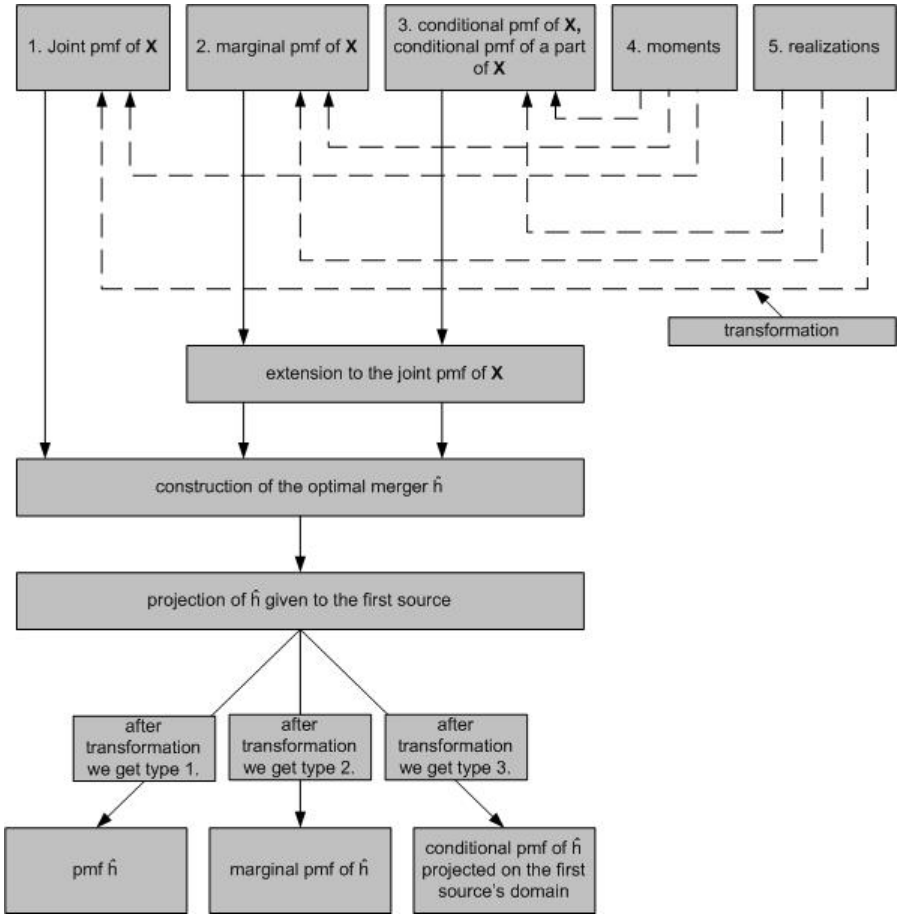


Figure 2.2: Graphic layout of the methodology discussed in Chapters 3 and 4

knowledge pieces are of the same form as h: the given knowledge pieces are joint pmfs of $\mathbf{X}$.
In Chapter 4, we focus on transformation and extension of knowledge pieces provided in other forms then the joint pmf of $\mathbf{X}$, see Table 2.1. If the moments or realizations are given, they are expressed in probabilistic terms: concretely as marginal or conditional pmfs. Once all given knowledge pieces are expressed as pmfs, we extend them to the joint pmf of $\mathbf{X}$ – so the construction of ${}^{O}\hat{h}$ discussed in Chapter 3 is possible.

Graphic layout of the methodology discussed in Chapters 3 and 4 is in Figure 2.2.

Chapter 3

Construction of the optimal merger

In this chapter we assume, that each of the s considered sources (the first source and its neighbors) provides its own description of distribution of discrete random vector $\mathbf{X}$ in the form of joint pmf, concretely as the probability vector

$$(g_j(\mathbf{x}_1), \ldots, g_j(\mathbf{x}_n)) = g_j, \quad j = 1, \ldots, s.$$

Under this assumption on provided knowledge pieces we construct their optimal merger.

As briefly mentioned in the Section 2.2, if we consider the optimal merger as an optimal estimate ${}^{O}\hat{h}$ of the "objective" pmf h of $\mathbf{X}$ (existence of such pmf is assumed, see Section 1.2), then the task of constructing the optimal merger becomes the task of constructing the optimal estimate ${}^{O}\hat{h}$ of h, based on provided knowledge pieces. By adopting the assumptions given in Section 1.4 (the notation used in Section 1.4 coincides with the notation introduced so far), the task of estimation of h can be interpreted as a statistical decision problem, for which the solution is introduced in (1.6). We also realize that the set $\widehat{H} = \{\hat{h} : \quad \sum_{i=1}^{n} \hat{h}(\mathbf{x}_i) = 1, \quad , \hat{h}(\mathbf{x}_i) \geq 0, \quad i = 1, \ldots, n\}$ coincides with the range of possible decision functions Δ introduced in Section 1.4. The solution of such estimation task is found as:

$$\mathrm{Arg}\min_{\hat{h}\in\widehat{H}} \int_H \mathrm{L}(h, \hat{h})\pi(h|D)\mathrm{d}h. \tag{3.1}$$

Furthermore, if we use Kerridge inaccuracy as a loss function (see Section 3.1), the solution reduces to the following task:

$$\mathrm{Arg}\min_{\hat{h}\in\widehat{H}} \mathrm{E}_{\pi(h|D)}[\mathrm{K}(h, \hat{h})|D]. \tag{3.2}$$

The used notation for expectation is explained in Section 1.3.

Proposition 3.1. *Let μ, η be the measures defined on measure spaces $(\times_{k=1}^{m}\mathcal{X}_k, \otimes_{k=1}^{m}\mathcal{B}_k)$, $(H, \mathcal{H})$, so that $(\times_{k=1}^{m}\mathcal{X}_k, \otimes_{k=1}^{m}\mathcal{B}_k, \mu)$, $(H, \mathcal{H}, \eta)$ are σ-finite measure spaces. Under assumption that:*

$$\int_{H\times(\times_{k=1}^{m}\mathcal{X}_k)} \pi(h(\mathbf{x})|D)h(\mathbf{x})\log\hat{h}(\mathbf{x})\mathrm{d}(\mu\times\eta) < \infty, \tag{3.3}$$

the solution of task (3.2) for $\hat{H} = H$ (see Section 1.4) has the form:

$$^{O}\hat{h} = E_{\pi(h|D)}(h|D). \tag{3.4}$$

Proof. The assumption of σ-finiteness of considered measure spaces is satisfied, for example, for the following choice of measures: μ – counting measure, η – Lebesgue measure. The proof is made for this case of our direct interest.
Since the assumptions of Fubini's theorem (see Folland (1999)) – σ-finiteness and (3.3) – we can rewrite the task (3.2) as follows:

$$\begin{aligned}
\operatorname{Arg}\min_{\hat{h}\in\hat{H}} \mathrm{E}_{\pi(h|D)}[\mathrm{K}(h,\hat{h})|D] &= \operatorname{Arg}\min_{\hat{h}\in\hat{H}} \int_H \pi(h|D)\mathrm{K}(h,\hat{h})\mathrm{d}h \\
&= \operatorname{Arg}\min_{\hat{h}\in\hat{H}} \int_H \pi(h|D)\sum_{i=1}^{n} h(\mathbf{x}_i)\log\hat{h}(\mathbf{x}_i)\mathrm{d}h \\
&= \operatorname{Arg}\min_{\hat{h}\in\hat{H}} \sum_{i=1}^{n}\left(\int_H \pi(h(\mathbf{x}_i)|D)h(\mathbf{x}_i)\mathrm{d}h\right)\log\hat{h}(\mathbf{x}_i) \\
&= \operatorname{Arg}\min_{\hat{h}\in\hat{H}} \sum_{i=1}^{n}\left(\mathrm{E}_{\pi(h(\mathbf{x}_i)|D)}(h(\mathbf{x}_i)|D)\right)\log\hat{h}(\mathbf{x}_i) \\
&= \operatorname{Arg}\min_{\hat{h}\in\hat{H}} \mathrm{K}(\mathrm{E}_{\pi(h|D)}(h|D),\hat{h}).
\end{aligned} \tag{3.5}$$

If we denote the optimal solution of the task (3.2) by $^{O}\hat{h}$ and since we know from the Section 1.6, that:

$$\operatorname{Arg}\min_{\hat{h}\in\hat{H}} \mathrm{K}(\mathrm{E}_{\pi(h|D)}(h|D),\hat{h}) = \mathrm{E}_{\pi(h|D)}(h|D),$$

we get:

$$^{O}\hat{h} = \operatorname{Arg}\min_{\hat{h}\in\hat{H}} \mathrm{E}_{\pi(h|D)}[\mathrm{K}(h,\hat{h})|D] = \operatorname{Arg}\min_{\hat{h}\in\hat{H}} \mathrm{K}(\mathrm{E}_{\pi(h|D)}(h|D),\hat{h}) = \mathrm{E}_{\pi(h|D)}(h|D).$$

□

If no other constraints are given, $\mathrm{E}_{\pi(h|D)}(h|D)$ is the merger searched for. If some additional constraints are given, we are looking for the solution of the optimization task:

$$\min_{\hat{h}\in\hat{H}} \mathrm{K}(\mathrm{E}_{\pi(h|D)}(h|D), \hat{h})$$
$$\text{s.t. given constraints.}$$

In both cases we need to evaluate $\mathrm{E}_{\pi(h|D)}(h|D)$, which is a conditional expectation with respect to the yet unspecified posterior pdf $\pi(h|D)$ of the unknown random vector h. In the following section the considered choice of the loss function is discussed. After a preparatory Section 3.2 the posterior pdf $\pi(h|D)$ is constructed (see Section 3.3). In Section 3.4, the optimal estimate of h based on the constructed posterior pdf $\pi(h|D)$ is derived.

3.1 Kerridge inaccuracy as loss function

Here we outline the reason why we have chosen Kerridge inaccuracy (see Section 1.6) as the loss function when estimating the multivariate parameter h, which is now a probability vector, by using Bayesian decision theory (see Section 1.4).

In the discussion, we assume that $\widehat{H} = H$.

Firstly assume that probability vector $h = (h(\mathbf{x}_1), \ldots, h(\mathbf{x}_n))$ is given. Then the optimal estimate has to satisfy (according to Section 1.4):

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

and since h is given and $H = \widehat{H}$, we also know that the minimum is reached for ${}^{O}\hat{h} = h$. For the set of all loss functions reaching the finite minimum for ${}^{O}\hat{h}$ it is shown in Bernardo (1979), that the Kerridge inaccuracy $\mathrm{K}(h,{}^{O}\hat{h}) = -\sum_{i=1}^{n} h(\mathbf{x}_i) \log {}^{O}\hat{h}(\mathbf{x}_i)$ is a representative of this set of loss functions.

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

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

where $\pi(h|D)$ is the posterior pdf of the possible values of $h \in H$ (see Section 1.4). Putting these statements together, we get:

$${}^{O}\hat{h} \in \mathrm{Arg}\min_{\hat{h}\in\hat{H}} \mathrm{E}_{\pi(h|D)}\mathrm{K}(h, \hat{h}). \tag{3.6}$$

3.2 Task of non-linear programming with constraints

For constructing the posterior pdf $\pi(h|D)$, we need to solve the following task of non-linear programming:

$$\min_{\pi(h|D)\in \mathrm{M}} \int_H \pi(h|D) \log \pi(h|D) \mathrm{d}h. \tag{3.7}$$

The set M of all admissible solutions is:

$$\mathrm{M} = \begin{array}{l} \left\{\pi(h|D): \quad \int_H \pi(h|D) \sum_{i=1}^n g_j(\mathbf{x}_i) \log h(\mathbf{x}_i) \mathrm{d}h - \beta_j(D) \leq 0, \quad j = 1, \ldots, s, \right. \\ \left. \int_H \pi(h|D) \mathrm{d}h - 1 = 0, \quad \pi(h|D) \geq 0 \quad \forall h \in H \right\}, \end{array}$$

where $\beta_j(D)$ are given constants.
The idea of solving this task of non-linear programming is to add some constraints to the utility function and then minimize the new utility function subject to remaining constraints.
In our case, we add the first s constraints to the utility function $\int_H \pi(h|D) \log \pi(h|D) \mathrm{d}h$.
For the original task (3.7), we choose appropriate set $\widetilde{\mathrm{M}} \supset \mathrm{M}$ and we define:

- $\widehat{\mathrm{M}} = \left\{\pi(h|D): \quad \int_H \pi(h|D) \mathrm{d}h - 1 = 0, \quad \pi(h|D) \geq 0 \quad \forall h \in H\right\}$
- $\boldsymbol{\lambda}(D) = (\lambda_1(D), \ldots, \lambda_s(D)) \in \mathbb{R}_+^s$
- the Lagrangian:

$$\begin{aligned} \mathrm{L}(\pi(h|D); \boldsymbol{\lambda}(D)) &= \int_H \pi(h|D) \log \pi(h|D) \mathrm{d}h + \sum_{j=1}^s \lambda_j(D) \left[\mathrm{E}_{\pi(h|D)} \mathrm{K}(g_j, h) - \beta_j(D)\right] \\ &= \int_H \pi(h|D) \log \pi(h|D) \mathrm{d}h \\ &+ \sum_{j=1}^s \lambda_j(D) \left[\int_H \pi(h|D) \sum_{i=1}^n g_j(\mathbf{x}_i) \log h(\mathbf{x}_i) \mathrm{d}h - \beta_j(D)\right] \end{aligned} \tag{3.8}$$

  on the set $(\widetilde{\mathrm{M}} \cap \widehat{\mathrm{M}}) \times \mathbb{R}_+^s$.

If we want to find the minimizer of the original task (3.7) as the minimizer of the task:

$$\mathrm{Arg} \min_{\pi(h|D)} \mathrm{L}(\pi(h|D); \boldsymbol{\lambda}(D)), \text{ where } \pi(h|D) \in (\widetilde{\mathrm{M}} \cap \widehat{\mathrm{M}}), \tag{3.9}$$

we need to set the values of $\boldsymbol{\lambda}(D)$. To do this we use global optimality conditions (GOC), see Lachout (2007).

Definition 3.2 (Global optimality conditions)**.** *Let* ${}^{O}\pi(h|D)$ *take real values for each* $h \in H$, ${}^{O}\boldsymbol{\lambda}(D) \in \mathbb{R}^s_+$ *and* $\widetilde{\mathrm{M}} \supset \mathrm{M}$. *Then,* $({}^{O}\pi(h|D), {}^{O}\boldsymbol{\lambda}(D))$ *satisfies the global conditions of optimality (GOC) for the task* (3.7) *on the set* $(\widetilde{\mathrm{M}} \cap \widehat{\mathrm{M}}) \times \mathbb{R}^s_+$ *if* $({}^{O}\pi(h|D); {}^{O}\boldsymbol{\lambda}(D))$ *is a saddlepoint of Lagrangian* (3.8) *on the set* $(\widetilde{\mathrm{M}} \cap \widehat{\mathrm{M}}) \times \mathbb{R}^s_+$. *It means that the following holds:*

- ${}^{O}\pi(h|D) \in (\widetilde{\mathrm{M}} \cap \widehat{\mathrm{M}})$, ${}^{O}\boldsymbol{\lambda}(D) \in \mathcal{R}^s_+$

- $\forall \quad \pi(h|D) \in (\widetilde{\mathrm{M}} \cap \widehat{\mathrm{M}}), \quad \boldsymbol{\lambda}(D) \in \mathcal{R}^s_+$

$$\mathrm{L}(\pi(h|D); {}^{O}\boldsymbol{\lambda}(D)) \geq \mathrm{L}({}^{O}\pi(h|D); {}^{O}\boldsymbol{\lambda}(D)) \geq \mathrm{L}({}^{O}\pi(h|D); \boldsymbol{\lambda}(D)).$$

The following theorem brings the relation between the solution satisfying (GOC) and the solution of the original task (3.7).

Theorem 3.3. *Let* ${}^{O}\pi(h|D)$ *take real values for each* $h \in H$ *and, for* $(\widetilde{\mathrm{M}} \cap \widehat{\mathrm{M}})$, *let it hold that all functions in task* (3.7) *are real-valued functions on* $\widetilde{\mathrm{M}}$. *If there exists* ${}^{O}\boldsymbol{\lambda}(D) \in \mathbb{R}^s_+$ *that* $({}^{O}\pi(h|D), {}^{O}\boldsymbol{\lambda}(D))$ *satisfies (GOC) on the set* $(\widetilde{\mathrm{M}} \cap \widehat{\mathrm{M}}) \times \mathbb{R}^s_+$, *then the global minimum of the original task* (3.7) *is reached in* ${}^{O}\pi(h|D)$.

Proof. See Lachout (2007). □

3.3 Construction of posterior pdf $\pi(h|D)$

Since the set of all possible posterior pdfs $\pi(h|D)$ is large, to choose the optimal one we put some additional conditions on the form of $\pi(h|D)$. The considered set will diminish and from the remaining possible posterior pdfs we choose the one with the highest entropy (see Section 1.5). Following the assumptions mentioned in Section 1.5, we define the constraints on the posterior pdf: j^{th} source takes h as its representative if h is close to the pmf g_j (vector of probabilities) given by j^{th} source, meaning the conditional expectation of Kerridge inaccuracy of g_j on h is smaller than or equal to some positive finite value $\beta_j(D)$:

$$\mathrm{E}_{\pi(h|D)}[\mathrm{K}(g_j, h)|D] \leq \beta_j(D). \tag{3.10}$$

From the set of possible posterior pdfs of h satisfying constraints (3.10) we choose the one with maximum entropy. Which means we are looking for solution of the following optimization task:

$$\mathrm{Arg} \max_{\pi(h|D) \in \mathrm{M}} - \int_H \pi(h|D) \log \pi(h|D) \mathrm{d}h, \tag{3.11}$$

where

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

Proposition 3.4 (Optimal posterior pdf)**.** *Let all constraints in* (3.11) *be active. Then, the optimal solution of the optimization task* (3.11) *is:*

$$^{O}\pi(h|D) = \frac{1}{Z(\lambda_1(D), ..., \lambda_s(D))} \prod_{i=1}^{n} h(\mathbf{x}_i)^{\sum_{j=1}^{s} \lambda_j(D) g_j(\mathbf{x}_i)}, \tag{3.12}$$

where

$$Z(\lambda_1(D), ..., \lambda_s(D)) > 0$$

and

$$\lambda_j(D) > 0 \quad j = 1, \ldots, s.$$

Proof. The solution of the task (3.11) can be found also as the solution of:

$$\mathrm{Arg} \min_{\pi(h|D) \in \mathrm{M}} \int_H \pi(h|D) \log \pi(h|D)\mathrm{d}h \tag{3.13}$$

where

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

Since the considered optimization task (3.13) is a task of nonlinear programming, to find its minimizer $^{O}\pi(h|D)$ we will use the results given in Lachout (see Section 3.2).

If we assume, that the conditions of the applicability of Fubini's theorem (see Folland (1999)) are satisfied, we can rewrite the Lagrangian $\mathrm{L}(\pi(h|D); \boldsymbol{\lambda}(D))$ (see formula (3.8)) as follows:

$$\int_H \pi(h|D) \log \pi(h|D)\mathrm{d}h + \lambda_1(D) \left(\mathrm{E}_{\pi(h|D)}(\mathrm{K}(g_1, h)|D) - \beta_1(D) \right) + \ldots$$
$$+ \lambda_s(D) \left(\mathrm{E}_{\pi(h|D)}(\mathrm{K}(g_s, h)|D) - \beta_s(D) \right)$$

$$= \int_H \pi(h|D) \log \pi(h|D)\mathrm{d}h + \sum_{j=1}^{s} \lambda_j(D) [\mathrm{E}_{\pi(h|D)}(\mathrm{K}(g_j, h)|D) - \beta_j(D)]$$

$$= \int_H \pi(h|D) \log \pi(h|D) \mathrm{d}h - \overbrace{\sum_{j=1}^{s} \int_H}^{\text{Fubini}} \lambda_j(D) \pi(h|D) \sum_{i=1}^{n} g_j(\mathbf{x}_i) \log(h(\mathbf{x}_i)) \mathrm{d}h$$
$$- \sum_{j=1}^{n} \lambda_j(D) \beta_j(D)$$

$$= \int_H \left(\pi(h|D) \log \pi(h|D) - \pi(h|D) \sum_{i=1}^{n} \sum_{j=1}^{s} \lambda_j(D) g_j(\mathbf{x}_i) \log(h(\mathbf{x}_i)) \right) \mathrm{d}h$$
$$- \sum_{j=1}^{n} \lambda_j(D) \beta_j(D)$$

$$= \int_H \pi(h|D) \left(\log \pi(h|D) - \sum_{i=1}^{n} \log(h(\mathbf{x}_i))^{\sum_{j=1}^{s} \lambda_j(D) g_j(\mathbf{x}_i)} \pm \log Z(\lambda_1(D), \ldots, \lambda_s(D)) \right) \mathrm{d}h$$
$$- \sum_{j=1}^{n} \lambda_j(D) \beta_j(D)$$

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

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

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

We see, its minimum is reached for $\pi(h|D) =^O \pi(h|D)$ a.e., because:

- the first part $(\mathrm{D}(\pi(h)|D)||^O\pi((h)|D)$, which is Kullback-Leibler divergence of $\pi(h|D)$ on ${}^O\pi(h|D))$, is minimal for $\pi(h|D) =^O \pi(h|D)$ a.e. (see Section 1.7).

- the remaining part of Lagrangian does not depend on $\pi(h|D)$ and does not influence the minimization.

Since ${}^O\pi(h|D)$ is taking real values and if we adopt the assumption of existence of ${}^O\boldsymbol{\lambda}_j(D) = (\lambda_1(D), \ldots, \lambda_s(D))$, so that $({}^O\pi(h|D), {}^O\boldsymbol{\lambda}_j(D))$ satisfies the global optimality conditions (see Definition 3.2) for the task (3.7), then the assumptions of Theorem 3.3 are satisfied and ${}^O\pi(h|D)$ is the minimizer of task (3.13) and ${}^O\pi(h|D)$ solves the task (3.11). □

3.4 Merging, construction of the estimate ${}^O\hat{h}$ of h

Proposition 3.5 (The optimal estimate ${}^O\hat{h}$)**.** *Let us define $\nu_0, \nu_1, \ldots, \nu_n$ as:*

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

$$\nu_0 = \sum_{i=1}^{n} \nu_i$$

and the normalizing constant $Z(\lambda_1(D), \ldots, \lambda_s(D))$ from the formula (3.12) as:

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

Here, $\lambda_j(D) =^O \lambda_j(D) > 0$, $j = 1, \ldots, s$, from Proposition 3.4.

Then, the optimal estimate ${}^O\hat{h}$ of h has the form:

$$\mathrm{E}_{{}^O\pi(h|D)}(h(\mathbf{x}_i)|D) =^O \hat{h}(\mathbf{x}_i) = \lambda_0^*(D) + \sum_{j=1}^{s} \lambda_j^*(D) g_j(\mathbf{x}_i), \tag{3.14}$$

where

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

$$\lambda_j^*(D) = \frac{\lambda_j(D)}{n + \sum_{j=1}^{s} \lambda_j(D)},$$

$$n\lambda_0^*(D) + \sum_{j=1}^{s} \lambda_j^*(D) = 1 \tag{3.15}$$

$$\lambda_j^*(D) > 0, \quad j = 0, \ldots, n. \tag{3.16}$$

Proof. Since we derived the optimal posterior pdf (see Proposition 3.4 in Section 3.3) we can now evaluate the conditional expectations $\mathrm{E}_{\pi(h|D)}(h(\mathbf{x}_i)|D)$ of

$$\mathrm{E}_{\pi(h|D)}(h|D) = (\mathrm{E}_{\pi(h|D)}(h(\mathbf{x}_1)|D), \ldots, \mathrm{E}_{\pi(h|D)}(h(\mathbf{x}_n)|D)).$$

The optimal posterior pdf has the form:

$$\begin{aligned} {}^{O}\pi(h|D) &= \frac{1}{Z(\lambda_1(D), ..., \lambda_s(D))} \prod_{i=1}^{n} h(\mathbf{x}_i)^{\sum_{j=1}^{s} \lambda_j(D) g_j(\mathbf{x}_i)} \\ &= \frac{1}{\frac{\prod_{i=1}^{k} \Gamma(\nu_i)}{\Gamma(\nu_0)}} \prod_{i=1}^{n} h(\mathbf{x}_i)^{\nu_i - 1}. \end{aligned} \tag{3.17}$$

Since we know, that

$$h(\mathbf{x}_i) \geq 0 \text{ for } i = 1 \ldots, n,$$

$$\sum_{i=1}^{n} h(\mathbf{x}_i) = 1,$$

and

$$\nu_i = 1 + \sum_{j=1}^{s} \lambda_j(D) g_j(\mathbf{x}_i) > 0 \text{ for } i = 1, \ldots, n,$$

because $\lambda_j(D) > 0$ for $j = 1, \ldots, s$ and $g_j(\mathbf{x}_i) \geq 0$, then (3.17) is a pdf of Dirichlet distribution $Dir(h(\mathbf{x}_1), \ldots, h(\mathbf{x}_n); \nu_1, \ldots, \nu_n)$ (see Section 1.8). By using the properties

of Dirichlet distribution (see Section 1.8 formula (1.10)) we get:

$$
\begin{aligned}
{}^{O}\hat{h}(\mathbf{x}_i) = \mathrm{E}_{O_{\pi(h|D)}}(h(\mathbf{x}_i)|D) = \frac{\nu_i}{\nu_0} &= \frac{1+\sum_{j=1}^{s}\lambda_j(D)g_j(\mathbf{x}_i)}{\sum_{i=1}^{n}[1+\sum_{j=1}^{s}\lambda_j(D)g_j(\mathbf{x}_i)]} \\
&= \frac{1+\sum_{j=1}^{s}\lambda_j(D)g_j(\mathbf{x}_i)}{n+\sum_{i=1}^{n}\sum_{j=1}^{s}\lambda_j(D)g_j(\mathbf{x}_i)} = \frac{1+\sum_{j=1}^{s}\lambda_j(D)g_j(\mathbf{x}_i)}{n+\sum_{j=1}^{s}[\lambda_j(D)\underbrace{\sum_{i=1}^{n}g_j(\mathbf{x}_i)}_{=1}]} \\
&= \frac{1}{n+\sum_{j=1}^{s}\lambda_j(D)} + \sum_{j=1}^{s}\frac{\lambda_j(D)}{n+\sum_{j=1}^{s}\lambda_j(D)}g_j(\mathbf{x}_i) \\
&= \lambda_0^*(D) + \sum_{j=1}^{s}\lambda_j(D)\lambda_0^*(D)g_j(\mathbf{x}_i) = \lambda_0^*(D) + \sum_{j=1}^{s}\lambda_j^*(D)g_j(\mathbf{x}_i).
\end{aligned}
\tag{3.18}
$$

Then, ${}^{O}\hat{h} = ({}^{O}\hat{h}(\mathbf{x}_1), \ldots, {}^{O}\hat{h}(\mathbf{x}_n))$ is the optimal estimate of pdf h of $\mathbf{X}$.
Since ${}^{O}\hat{h}$ is a pmf, we see that:

$$
\begin{aligned}
\sum_{i=1}^{n}{}^{O}\hat{h}(\mathbf{x}_i) &= \sum_{i=1}^{n}\lambda_0^*(D) + \sum_{i=1}^{n}\sum_{j=1}^{s}\lambda_j^*(D)g_j(\mathbf{x}_i) \\
&= n\lambda_0^*(D) + \sum_{j=1}^{s}\lambda_j^*(D)\underbrace{\sum_{i=1}^{n}g_j(\mathbf{x}_i)}_{=1} = n\lambda_0^*(D) + \sum_{j=1}^{s}\lambda_j^*(D) = 1,
\end{aligned}
\tag{3.19}
$$

and for $j = 1, \ldots, s$

$$
\lambda_j(D) > 0 \text{ from Proposition (3.4)} \quad \begin{cases} \lambda_j^*(D) = \lambda_j(D)/(n+\sum_{j=1}^{s}\lambda_j(D)) > 0 \\ \\ \lambda_0^*(D) = 1/(n+\sum_{j=1}^{s}\lambda_j(D)) > 0. \end{cases} .
\tag{3.20}
$$

□

Chapter 4

Extension of the other forms of given information

In previous chapter we assumed that every source gives the piece of information in the form of the joint pmf of a collection of discrete random variables $\mathbf{X}$ (concretely as a probability vector of possible realizations $\mathbf{x}_1, \ldots, \mathbf{x}_n$). But this is very restrictive condition on the sources. Even if each of s sources uses all its abilities and past experience to provide some knowledge, it is highly probable the source will provide the knowledge about a subset of $\mathbf{X}$ or about a conditional relation of parts of $\mathbf{X}$. In this chapter possible forms of the given knowledge pieces are presented and their transformation into joint pmf of $\mathbf{X}$ (into a probability vector of possible realizations $\mathbf{x}_1, \ldots, \mathbf{x}_n$), useful for merging (see the previous chapter), is discussed.
Let:

- $\mathbf{P}_j$ denote part of $\mathbf{X}$, which describes the j^{th} source's past experience; $\mathbf{p}$ belongs to the set of possible realizations of $\mathbf{P}_j$,
- $\mathbf{F}_j$ denote a part of $\mathbf{X}$, which describes the j^{th} source's ignorance; $\mathbf{f}$ belongs to the set of possible realizations of $\mathbf{F}_j$,
- $\mathbf{U}_j$ denote a part of $\mathbf{X}$, that is unconsidered by the j^{th} source; $\mathbf{u}$ belongs to the set of possible realizations of $\mathbf{U}_j$.

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

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

2) a concrete realization (value) of $\mathbf{F}_j \subset \mathbf{X}$ on a part $\mathbf{P}_j \subset \mathbf{X}$, $(\mathbf{F}_j \cup \mathbf{P}_j) \subseteq \mathbf{X}$,

   or

   a concrete realization of $\mathbf{P}_j \subseteq \mathbf{X}$,

3) conditional pmf (in the form of probability vector) of $\mathbf{F}_j$ on $\mathbf{P}_j$, where $(\mathbf{F}_j \cup \mathbf{P}_j) \subseteq \mathbf{X}$, denoted by $g_j(\mathbf{f}|\mathbf{p})$

4) joint pmf (in the form of probability vector) of $\mathbf{P}_j \subset \mathbf{X}$ (marginal pmf of $\mathbf{X}$), denoted by $g_j(\mathbf{p})$

4.1 Unification of data, mapping moments and values on probabilities

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

4.1.1 Moments given

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

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

  denoted by:

  $$\mathrm{E}_{g_j(\mathbf{f}|\mathbf{p})}(\phi(\mathbf{F}_j, \mathbf{P}_j)|\mathbf{P}_j) = \psi(\mathbf{P}_j), \tag{4.1}$$

  where ϕ, ψ are functions specified by the source and the expectation is taken with respect to a, yet unspecified, pmf $g_j(\mathbf{f}|\mathbf{p})$, existence of which is assumed (see Section 1.2).

- moments of $\mathbf{P}_j \subseteq \mathbf{X}$,

  denoted by :

  $$\mathrm{E}_{g_j(\mathbf{p})}(\phi(\mathbf{P}_j)) = \psi, \tag{4.2}$$

  where ϕ and ψ are a function and a value specified by the source and the expectation is taken with respect to a, yet unspecified, pmf $g_j(\mathbf{p})$, existence of which is assumed (see Section 1.2).

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

If j^{th} source gives the conditional moments (4.1), the idea for construction of $g_j(\mathbf{f}|\mathbf{p})$ is:

- from the set of all possible conditional pmfs of $\mathbf{F}_j$ conditioned on $\mathbf{P}_j$ (existence of them is assumed, see Section 1.2) construct a set of conditional pmfs satisfying (4.1): $\{g_j^*(\mathbf{f}|\mathbf{p})\}$
- and from $\{g_j^*(\mathbf{f}|\mathbf{p})\}$ choose the conditional pmf with the maximum entropy, it means choose the pmf for which holds:

$$g_j(\mathbf{f}|\mathbf{p}) = \text{ Arg } \max_{\{g_j^*(\mathbf{f}|\mathbf{p})\}} -\sum_{(\mathbf{f},\mathbf{p})} g_j^*(\mathbf{f}|\mathbf{p}) \log g_j^*(\mathbf{f}|\mathbf{p})$$

By applying the same idea on the case, when the j^{th} source gives the moments (4.2), we get $g_j(\mathbf{p})$.

4.1.2 Ordinary data given

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

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

Again we try to express this type of given knowledge pieces in probabilistic terms – the pmf of random variables considered by the j^{th} source.
To do this we use the measure concentrated on one point. It is Kronecker delta:

$$\begin{aligned} \delta_{i,j}^K &= 1 \quad \text{if } i = j \\ &= 0 \quad \text{otherwise.} \end{aligned}$$

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

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

The $g_j(\mathbf{f}|\mathbf{p})$ is a pmf since it satisfies:

$$\sum_{\mathbf{p}} g_j(\mathbf{f}|\mathbf{p}) = \sum_{(\mathbf{f},\mathbf{p})} \delta^K_{(\mathbf{f},\mathbf{p}),\underline{(\mathbf{f},\mathbf{p})}} = 1$$

and

$$g_j(\mathbf{f}|\mathbf{p}) \geq 0$$

for all possible realizations $(\mathbf{f},\mathbf{p})$.
For case, where $\underline{\mathbf{p}}$ is given, we define $g_j(\mathbf{p})$ as $\delta^K_{\mathbf{p},\underline{\mathbf{p}}}$:

$$\begin{aligned} g_j(\mathbf{p}) &= 1 \quad \text{if } \mathbf{p} = \underline{\mathbf{p}} \\ &= 0 \quad \text{otherwise.} \end{aligned}$$

The $g_j(\mathbf{p})$ is a pmf since it satisfies:

$$\sum_{\mathbf{p}} g_j(\mathbf{p}) = \sum_{\mathbf{p}} \delta^K_{\mathbf{p},\underline{\mathbf{p}}} = 1$$

and

$$g_j(\mathbf{p}) \geq 0$$

for all possible realizations $\mathbf{p}$.

4.2 Extension

Since all given knowledge pieces have now the form of pmfs of random variables considered by a particular source: $g_j(\mathbf{f}|\mathbf{p})$ or $g_j(\mathbf{p})$, we can focus on their extension into a joint pmf of $\mathbf{X}$ denoted by ${}^e g_j$ and further in text called extension ${}^e g_j$.
Under the following assumptions (see Section 1.4):

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

  $$g_j(\mathbf{f}_i|\mathbf{p}_i) \geq 0, \quad g_j(\mathbf{p}_i) \geq 0$$

  for $j = 1, \ldots, s$, $i = 1, \ldots, n$,

  $$\sum_{i=1}^{n} g_j(\mathbf{f}_i|\mathbf{p}_i) = 1, \quad \sum_{i=1}^{n} g_j(\mathbf{p}_i) = 1$$

  for $j = 1, \ldots, s$,
- $(s \times n)$ matrix D is a realization of the above matrix,

we introduce the following constraints:

1. The first and intuitively clear assumption on the extension ${}^e g_j$ is: the projection of ${}^e g_j$ on the j^{th} source's domain – ${}^e g_j(\mathbf{f}|\mathbf{p})$ – coincides with $g_j(\mathbf{f}|\mathbf{p})$.

2. The extension ${}^e g_j$ is to be as close as possible to the unknown pmf h (see the beginning of the Chapter 3 - sources provide knowledge pieces about $\mathbf{X}$ in the form of joint pmf, where $\mathbf{X}$ is described by the unknown pmf h). In terms of Bayesian decision theory h is the unknown multivariate random parameter taking values in H, see Section 1.4. We want ${}^e g_j$ to be the minimizer of $\mathrm{E}_{\pi(h|D)}[\mathrm{K}(h, {}^e g_j^*)|D]$, where ${}^e g_j^*$ belongs to a set of all possible pmfs satisfying the constraint 1. denoted by $\{{}^e g_j^*\}$ (see Section 1.4 and the beginning of Chapter 3).

   This requirement means, under assumption of applicability of Fubini's theorem (see the proof of Proposition 3.1), that:

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

   where the global minimum is reached for ${}^e g_j = \mathrm{E}_{\pi(h|D)}(h|D)$, see Section 1.6. In the previous chapter, it is denoted by ${}^O\hat{h}$ (i.e. see Proposition 3.1).

3. The last natural assumption, we already used in previous step, is that ${}^e g_j$ uses all elements of D.

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

4.2.1 Conditional probabilities on a part of random vector

The extension of the knowledge pieces given by the j^{th} source in the form of conditional pmf $g_j(\mathbf{f}|\mathbf{p})$ of $\mathbf{F}_j$ on $\mathbf{P}_j$, $(\mathbf{F}_j \cup \mathbf{P}_j) \subset \mathbf{X}$ is following:

Proposition 4.1. *Let the conditional pmf $g_j(\mathbf{f}|\mathbf{p})$ of $\mathbf{F}_j$ on $\mathbf{P}_j$, $(\mathbf{F}_j \cup \mathbf{P}_j) \subset \mathbf{X}$, be given. Then under the assumption that*

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

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

$$ {}^e g_j(\mathbf{x}_i) =^O \hat{h}(\mathbf{u}_i|\mathbf{f}_i, \mathbf{p}_i) g_j(\mathbf{f}_i|\mathbf{p}_i) {}^O\hat{h}(\mathbf{p}_i), \quad i = 1, \ldots, n, \tag{4.3} $$

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

Proof. In the proof, the following definition of conditional probability is used. The conditional probability of events A_1, A_2, A_3 (under the assumption that $\mathrm{P}(A_2, A_3) > 0$ and $\mathrm{P}(A_3) > 0$) is:

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

The probability of (A_1, A_2, A_3) is then:

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

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

If we realize that:

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

then by assuming of applicability of Fubini's theorem (see Folland (1999)), we can rewrite the task stated in the constraint 2. as follows. By inserting proposed ${}^{e}g_j$ into the minimized expected Kerridge inaccuracy, we get:

$$\mathrm{EK}(h,{}^{e}g_j) = -\int_H \pi(h|D)\sum_{i=1}^{n} h(\mathbf{x}_i)\log^{e} g_j(\mathbf{x}_i)\mathrm{d}h$$

$$= -\int_H \pi(h|D)\sum_{\mathbf{u}}\sum_{\mathbf{f}}\sum_{\mathbf{p}} h(\mathbf{u},\mathbf{f},\mathbf{p})\log\left({}^{O}\hat{h}(\mathbf{u}|\mathbf{f},\mathbf{p})g_j(\mathbf{f}|\mathbf{p}){}^{O}\hat{h}(\mathbf{p})\right)\mathrm{d}h$$

$$\begin{aligned}= &-\int_H \pi(h|D)\sum_{\mathbf{u}}\sum_{\mathbf{f}}\sum_{\mathbf{p}} h(\mathbf{u}|\mathbf{f},\mathbf{p})h(\mathbf{f},\mathbf{p})\log {}^{O}\hat{h}(\mathbf{u}|\mathbf{f},\mathbf{p})\mathrm{d}h\\ &-\int_H \pi(h|D)\sum_{\mathbf{u}}\sum_{\mathbf{f}}\sum_{\mathbf{p}} h(\mathbf{u}|\mathbf{f},\mathbf{p})h(\mathbf{f}|\mathbf{p})h(\mathbf{p})\log g_j(\mathbf{f}|\mathbf{p})\mathrm{d}h\\ &-\int_H \pi(h|D)\sum_{\mathbf{u}}\sum_{\mathbf{f}}\sum_{\mathbf{p}} h(\mathbf{u}|\mathbf{f},\mathbf{p})h(\mathbf{f}|\mathbf{p})h(\mathbf{p})\log {}^{O}\hat{h}(\mathbf{p})\mathrm{d}h\end{aligned}$$

$$= -\sum_{\mathbf{f}}\sum_{\mathbf{p}}\int_H \pi(h(\mathbf{u},\mathbf{f},\mathbf{p})|D)h(\mathbf{f}|\mathbf{p})\mathrm{d}h$$
$$\times\left(\sum_{\mathbf{u}}\int_H \pi(h(\mathbf{u},\mathbf{f},\mathbf{p})|D)h(\mathbf{u}|\mathbf{f},\mathbf{p})\log {}^{O}\hat{h}(\mathbf{u}|\mathbf{f},\mathbf{p})\mathrm{d}h\right)$$

$$-\sum_{\mathbf{u}}\sum_{\mathbf{p}}\int_H \pi(h(\mathbf{u},\mathbf{f},\mathbf{p})|D)h(\mathbf{u}|\mathbf{f},\mathbf{p})h(\mathbf{p})\mathrm{d}h$$
$$\times\left(\sum_{\mathbf{f}}\int_H \pi(h(\mathbf{u},\mathbf{f},\mathbf{p})|D)h(\mathbf{f}|\mathbf{p})\log g_j(\mathbf{f}|\mathbf{p})\mathrm{d}h\right)$$

$$-\sum_{\mathbf{u}}\sum_{\mathbf{f}}\int_H \pi(h(\mathbf{u},\mathbf{f},\mathbf{p})|D)h(\mathbf{u}|\mathbf{f},\mathbf{p})h(\mathbf{f}|\mathbf{p})\mathrm{d}h$$
$$\times\left(\sum_{\mathbf{p}}\int_H \pi(h(\mathbf{u},\mathbf{f},\mathbf{p})|D)h(\mathbf{p})\log^{O}\hat{h}(\mathbf{p})\mathrm{d}h\right).$$

The second term can not be influenced by the choice of ${}^{O}\hat{h}$, since it does not involve marginal or conditional version of ${}^{O}\hat{h}$. The expressions in the brackets () in the first and third term are conditional versions of the Kerridge inaccuracy (see Section 1.6), which are, for an arbitrary condition, uniquely minimized for:

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

and

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

Since the estimate ${}^{O}\hat{h}$ is using all knowledge pieces in D (see Section 3.4), the constraint 3. is satisfied. □

4.2.2 Conditional probabilities on the whole set of random variables

The extension of the knowledge pieces given by the j^{th} source in the form of $g_j(\mathbf{f}|\mathbf{p})$ of $\mathbf{F}_j$ on $\mathbf{P}_j$, $(\mathbf{F}_j \cup \mathbf{P}_j) = \mathbf{X}$, is following:

Proposition 4.2. *Let probability vector $g_j(\mathbf{f}|\mathbf{p})$ of $\mathbf{F}_j$ on $\mathbf{P}_j$, $(\mathbf{F}_j \cup \mathbf{P}_j) = \mathbf{X}$, be given. Then under assumption that*

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

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

$$^{e}g_j(\mathbf{x}_i) = g_j(\mathbf{f}_i|\mathbf{p}_i)^{O}\hat{h}(\mathbf{p}_i), \qquad i = 1, \ldots, n, \tag{4.4}$$

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

Proof. In the proof, the following definition of conditional probability is used.
The conditional elementary probability of A_1, A_2 (under assumption that $P(A_2) > 0$) is:

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

and the joint probability of events (A_1, A_2) is then:

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

By using the same ideas, as were used in the proof of Proposition 4.1, it can be shown, that (4.4) is the unique extension of provided knowledge pieces satisfying the above stated constraints 1., 2., 3.. □

4.2.3 Marginal pmf of random vector

If the knowledge piece is now in the form of a joint pmf $g_j(\mathbf{p})$ of $\mathbf{P}_j \subset \mathbf{X}$, then the extension of it is following:

Proposition 4.3. *Let the probability vector $g_j(\mathbf{p})$ of $\mathbf{P}_j \subset \mathbf{X}$ be given. Then, under the assumption that*

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

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

$$^{e}g_j(\mathbf{x}_i) = {}^{O}\hat{h}(\mathbf{u}_i|\mathbf{p}_i)g_j(\mathbf{p}_i), \qquad i = 1, \ldots, n, \tag{4.5}$$

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

Proof. To prove that (4.5) is the unique extension of $g_j(\mathbf{p})$ to a joint pmf of $\mathbf{X}$ follow the steps of proof of the Proposition 4.2. □

4.3 The optimal merger based on extended knowledge pieces

The optimal merger ${}^{O}\hat{h} = ({}^{O}\hat{h}(\mathbf{x}_1), \ldots, {}^{O}\hat{h}(\mathbf{x}_n))$ derived in Section 3.4 has, according to the extensions derived in Section 4.2, the following forms:

- for the extension constructed in Subsection 4.2.1:

$$ {}^{O}\hat{h}(\mathbf{x}_i) = \lambda_0^*(D) + \sum_{j=1}^{s} \lambda_j^*(D) {}^{O}\hat{h}(\mathbf{u}_i|\mathbf{f}_i, \mathbf{p}_i) g_j(\mathbf{f}_i|\mathbf{p}_i) {}^{O}\hat{h}(\mathbf{p}_i), \tag{4.6} $$

  for $i = 1, \ldots, n$,

- for the extension constructed in Subsection 4.2.2:

$$ {}^{O}\hat{h}(\mathbf{x}_i) = \lambda_0^*(D) + \sum_{j=1}^{s} \lambda_j^*(D)^e g_j(\mathbf{x}_i) = \lambda_0^*(D) + \sum_{j=1}^{s} \lambda_j^*(D) g_j(\mathbf{f}_i|\mathbf{p}_i) {}^{O}\hat{h}(\mathbf{p}_i), \tag{4.7} $$

  for $i = 1, \ldots, n$,

- for the extension constructed in Subsection 4.2.3:

$$ {}^{O}\hat{h}(\mathbf{x}_i) = \lambda_0^*(D) + \sum_{j=1}^{s} \lambda_j^*(D)^e g_j(\mathbf{x}_i) = \lambda_0^*(D) + \sum_{j=1}^{s} \lambda_j^*(D) {}^{O}\hat{h}(\mathbf{u}_i|\mathbf{p}_i) g_j(\mathbf{p}_i), \tag{4.8} $$

  for $i = 1, \ldots, n$.

4.4 Properties of the proposed optimal merger

The optimal merger given in Section 3.4 is not guaranteed to be unique. Moreover, a closed form solution for determining $\lambda_1(D), \ldots, \lambda_s(D)$ does not generally exist.

The optimal merger described in Section 4.3 is not generally unique. It can be seen from following example. If we assume there is only one source providing the knowledge pieces, the choice of "extending factors" on the right hand sides of formulas (4.6), (4.7), (4.8) is ambiguous. On the other hand, even ${}^{O}\hat{h}$ is ambiguous, the projection on source's domain, which is given back to it after merging, is unique.

We conjecture that these properties are valid generally but no proof exists.

Chapter 5

Conclusion

Decision making is an integral, often unrecognized part of our life. However, when, for instance, we are active in financial markets, it is reasonable to take every single piece of available knowledge into account, since the consequences of the decisions made can have a really big financial impact on us.

If we consider a parametric (population) model, which has one or more unknown parameters, the statistical analysis helps us to gain information from past experience. The adopted Bayesian treats the unknown parameters as random variables in order to make the conclusions about them. After using the Bayesian theorem, it means after construction of posterior pdf or pmf of considered random variables, we can report the different characteristics of obtained distribution: the mean, variance, mode ...

Significant restriction that arises while using Bayesian approach to the parameter estimation is that it is well-elaborated only when the knowledge pieces are given as "ordinary" crisp data. There are examples going beyond this state. For instance, in Savchuk and Martz (1994) the Bayes estimators for the true binomial survival probability p, when the prior knowledge (provided by multiple sources) is stated as credibility interval on p. But the general treatment of estimating the parameters of the parametric model, when different forms of knowledge are given, was missing.

In this work, handling of different types of knowledge is addressed. The idea of the treatment is to find a suitable model describing the provided knowledge pieces. To successfully solve this task, we use the Supra-Bayesian approach. We assume that there exists a fictitious decision maker, which considers the given knowledge as random quantities and uses Bayesian methodology to construct the parameters of the above mentioned model: it deals with a "supra" model. This converts the problem to a construction of an appropriate pdf or pmf of considered random quantities and to the construction of the optimal merger of provided knowledge pieces.

The merging of different types of provided knowledge pieces is difficult in the case when the very different properties are described by them. We assume rather general categories of knowledge pieces, for which the optimal merger can be constructed. We assume a set of sources providing knowledge pieces about a discrete random vector, fix one of the sources and introduce the task of improvement of this source's knowledge. The task of improvement of source's knowledge coincides with task of constructing the optimal merger of knowledge pieces provided by the considered source and its neighbors, i.e., sources, for which the intersection of their domains with the domain of considered source is non empty. This approach can be then applied on every source from the group of sources, which can be extremely large and distributed.

Then, we interpret the task of construction of the optimal merger for considered source and its neighbors as a decision task and with the use of Bayesian methods we successfully construct the optimal merger. It is identified with the optimal estimate of the pmf describing knowledge pieces of considered group of neighbors. Firstly we derive the optimal merger, when the provided knowledge pieces are in a "good" form – the form of a joint pmf of the mentioned discrete random vector. Then, we focus on how to transform and then extend the "bad" forms of knowledge pieces into joint pmf of discrete random vector, so the constructed optimal merger can be applied on them. Since the original task was to improve the considered source's knowledge, we project the optimal merger on its domain.

Naturally, we did not discuss many additional questions arising with the derivation of the final formula, i.e., how to get the Lagrangian multipliers or the value of bounds used in (3.10) or the problems with ambiguity of constructed optimal merger (see Section 4.4). They are definitely topics of a future work.

Index

Bibliography

Anděl J, 2007. *Základy matematické statistiky*. **matfyz**press.

Bernardo JM, 1979. Expected information as ecpected utility. *Ann. Stat.* **7**: 686–690.

Dupač V, Hušková M, 2005. *Pravděpodobnost a matematická statistika*. Karolinum.

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

Genest C, Zidek JV, 1986. Combining probability distributions: a critique and an annotated bibliography. With comments, and a rejoinder by the authors. *Stat. Sci.* **1**: 114–148.

Hušková M, 1985. *Bayesovské metódy*. Státní pedagogické nakladatelství.

Jaynes E, 1957. Information theory and statistical mechanics. I, II. .

Kárný M, 2009. Knowledge elicitation via extension of fragmental knowledgepieces .

Kárný, M and Guy, TV and Bodini, A and Ruggeri, F, 2009. Cooperation via sharing of probabilistic elements. *IJCIStudies* **1**: 139–162.

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

Kullback S, 1997. *Information theory and statistics. Reprint of the 2nd ed. '68.* Mineola, NY: Dover Publications, Inc. xvi, 399 p. $ 12.95 .

Kullback S, Leibler R, 1951. On information and sufficiency. *Ann. Math. Stat.* **22**: 79–86.

Lachout P, 2004. *Teorie pravděpodobnosti*. Karolinum.

Lachout P, 2007. Matematické programování. *pracovní text k přednášce ,,EKN011 Optimalizace I"* .

Savchuk V, Martz H, 1994. Bayes reliability estimation using multiple sources of prior knowledge: binomial sampling. *IEEE Transactions on Reliability* **43**: 138–144.

Shore JE, Johnson RW, 1980. Axiomatic derivation of the principle of maximum entropy and the principle of minimum cross-entropy. *IEEE Trans. Inf. Theory* **26**: 26–37.