Supra-Bayesian Approach to Merging of Incomplete and Incompatible Data

Vladimíra Sečkárová [*]
Department of Adaptive Systems
Institute of Information Theory and Automation
Prague, CZ 182 08
vseckarova@gmail.com

Abstract

In practice we often deal with the problem that the pieces of information given by different sources are often incomplete and even incompatible. In this work we try to solve this problem and present a systematic and unified way how to combine the pieces by using a Supra-Bayesian approach.

1 Introduction

Imagine you want to describe a behavior of a system. The data you want to use to fulfill your task are provided by some sources. Since the sources can describe just a part of the system, the data you got can be incomplete. Or even worse: some or every of the sources gave the data in different form. How to proceed in solving your task then?

1.1 Basic idea

The idea we use is simple: decompose the main task into particular "subtasks" by following this pattern:

- take all possible data pieces from the sources about their domains (a domain: a part of system that the source describes, represented e.g. by random variables),
- focus on one source S and find its neighbors (a neighbor: a source, domain of which has a nonempty intersection with the domain of S).

  For correct comprehension of the reader the Figure 1 describes the considered relations.

To follow the idea of "subtasks" we can now introduce the task of improving the knowledge of source S by using what is given. In fact it means to construct the optimal estimate of pmf (probability mass function - if domains are represented by discrete random vectors) describing S's domain based on a priori given data. The explanation how the system can be described by the subtasks is given below.

In "subtask", the construction of the optimal estimate of pmf describing the unification of all considered domains is included; which also means that we will need to find the optimal merger of given pieces of information. This optimal estimate will be then projected on the domain of source S and the aim of the "subtasks" will be satisfied. That is just a simple layout of the topic discussed in this work.

The problem of combining the heterogeneous sources has been discussed in many works, e.g. in [1], [2] and [3]. The outline of the method proposed here and also the mathematical notation

[*] Department of Probability and Mathematical Statistics, Faculty of Mathematics and Physics, Charles University in Prague

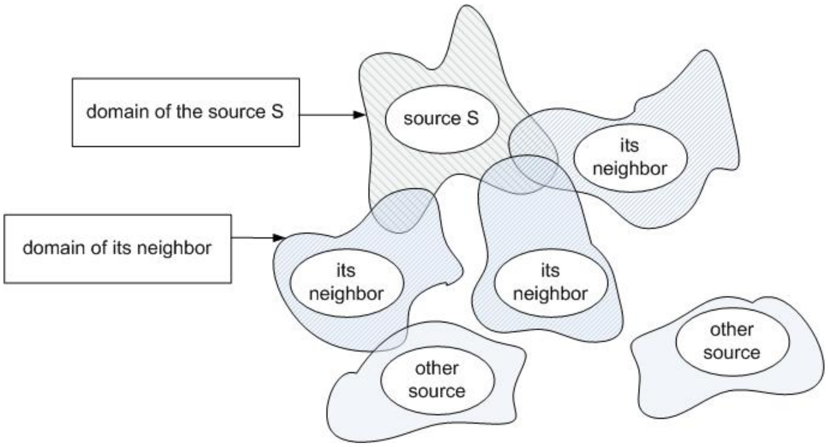


Figure 1: A relation between the source S, its neighbors and other sources

can be found in [4]. The contribution of this work consists of a brief introduction of the method suitable for any form of the initial situation. That is at the end we will come to just one formula handling different types of input data. Also the example and some numerical representation of used constants is given.

2 Proposed method and final estimate

If we introduce following notation:

- a domain: (discrete) random vector $\mathbf{X}$; with finite set of realizations: $\{\mathbf{x}_1, \ldots, \mathbf{x}_n\}$,

- $(g_j(\mathbf{x}_1), \ldots, g_j(\mathbf{x}_n)) = g_j$ - vector of probabilities given by j^{th} source, $j = 1, \ldots, s < \infty$

- $D = (g_1^T, \ldots, g_s^T)^T$ matrix of knowledge pieces given by the source S and its $s-1$ neighbors,

- $h = (h(\mathbf{x}_1), \ldots, h(\mathbf{x}_n))$ the unknown merger (unknown vector of probabilities),

- $H = \{h : \sum_{i=1}^{n} h(\mathbf{x}_i) = 1,\ h(\mathbf{x}_i) \geq 0, i = 1, \ldots, n\}$ - domain of pmfs h, g_j, $j = 1, \ldots, s$,

- L(. , .) – loss function (see [5]),

- K(. , .) – Kerridge inaccuracy (see [6]),

- ${}^O\hat{h}$ the optimal estimate of h (the optimal merger of the given knowledge pieces),

then the considered situation can be described as follows:

- every source:
  - operates on a random vector (with finite number of realizations), which is a part of $\mathbf{X}$,
  - provides a probabilistic or non-probabilistic information (about its domain) denoted by g_j, influenced by skills, decisions and possible actions of particular source,
- and the main task is: find the optimal vector of probabilities ${}^O\hat{h} = ({}^O\hat{h}(\mathbf{x}_1), \ldots, {}^O\hat{h}(\mathbf{x}_n))$ describing the random vector $\mathbf{X}$.

2.1 Basic case

In this subsection we derive the optimal estimate (optimal merger) of given data pieces. Our assumptions are:

- common domain: all of the picked sources have the same domain,
- data in probabilistic form: every source gave the data in the form of probability vector about realizations of the considered random vector $\mathbf{X}$,

- random inputs: any data piece is modelled as a random variable.

Next we will focus on the Bayesian methodology, on which the proposed construction is based. We are looking for the solution of

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

where $\mathrm{E}_{\pi(h|D)}[.|.]$ is the conditional expectation with respect to the posterior pdf (probability density function) $\pi(h|D)$ and the Kerridge inaccuracy was taken as the loss function (see [6]). By assuming of the applicability of Fubini's theorem we will find that the solution of task (1) has the form:

$$^{O}\hat{h} = \mathrm{E}_{\pi(h|D)}[h|D] \in \operatorname{Arg}\min_{\hat{h}\in\hat{H}} \mathrm{K}[\mathrm{E}_{\pi(h|D)}(h|D), \hat{h}] \tag{2}$$

Estimate $^{O}\hat{h}$ requires the knowledge of the posterior pdf $\pi(h|D)$, which we do not have at this moment. To solve this problem we take all possible pdfs $\pi(h|D)$ satisfying following constraints:

- j^{th} source accepts h as its representative if h is close to the pmf g_j given by the j^{th} source, numerically it means that the conditional expectation of Kerridge inaccuracy of g_j on h is smaller than or equal to a positive finite value $\beta_j(D)$:

$$\mathrm{E}_{\pi(h|D)}[\mathrm{K}(g_j,h)|D] \leq \beta_j(D). \tag{3}$$

We will use the maximum entropy principle (see [7]) to choose among pdfs $\pi(h|D)$ meeting (3):

- the Lagrangian $\mathbf{L}$ of arising optimization task is (after several steps of evaluation):

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

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

  where $Z(\lambda_1(D),\ldots,\lambda_s(D))$ is a normalizing factor and $\lambda_j(D)$ are Kuhn-Tucker multipliers, $j=1,\ldots,s$,

- minimum of the Lagrangian is reached for $\pi(h|D) = {}^{O}\pi(h|D)$ a.e. (explicit form of which is a pdf of Dirichlet distribution), because the first part is Kullback-Leibler divergence of $\pi(h|D)$ on $^{O}\pi(h|D)$ (see [8]), which is minimal for $\pi(h|D) =^{O} \pi(h|D)$ a.e. and the remaining part of Lagrangian does not depend on $\pi(h|D)$ and does not influence the minimization.

From the properties of Dirichlet distribution we get expected value, i.e. the final merger:

$$\mathrm{E}_{^{O}\pi(h|D)}(h(\mathbf{x}_i)|D) = {}^{O}\hat{h}(\mathbf{x}_i) = \lambda_0^*(D) + \sum_{j=1}^{s}\lambda_j^*(D)g_j(\mathbf{x}_i), \quad i=1,\ldots,n, \tag{4}$$

where

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

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

with $\lambda_j(D)$ chosen so that (3) is met.

2.2 General case

Since every of the sources has its own abilities, it is almost impossible to satisfy the assumptions considered in Subsection 2.1:

- the domains of the considered group of sources can be different from picked source S (but have at least one random variable in common with it),

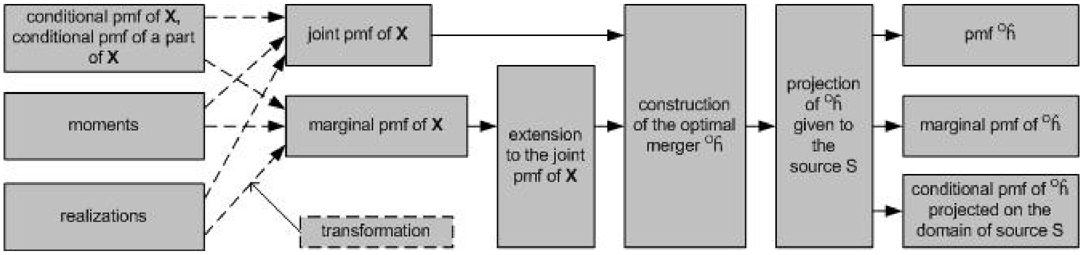

conditional pmf of X, conditional pmf of a part of X
moments
realizations
joint pmf of X
marginal pmf of X
transformation
extension to the joint pmf of X
construction of the optimal merger ${}^O\hat{h}$
projection of ${}^O\hat{h}$ given to the source S
pmf ${}^O\hat{h}$
marginal pmf of ${}^O\hat{h}$
conditional pmf of ${}^O\hat{h}$ projected on the domain of source S

Figure 2: Graphic layout of the methodology

- given information describing the domain of particular source can have other forms (not only the probability vectors):
  1) (conditional) expectations,
  2) "classic" data,
  3) (conditional) marginal probability vectors.

To construct the optimal merger of given information we will use the results from the previous section:

- first we transform the given information into probabilistic terms (probability vectors),
- then we take the unification of the domains of considered sources and extend the probabilistic information from the previous step onto this set.

For graphic layout of the proposed methodology see Figure 2.

Now all the assumptions of Subsection 2.1 are satisfied and we can create the merger (4). If we introduce the decomposition of source's domain considering the unification of all domains as follows:

$$\mathbf{X} = (\mathbf{U}, \mathbf{F}, \mathbf{P}) = \begin{cases} \mathbf{U} & \text{random variables unconsidered by particular source} \\ \mathbf{F} & \text{random variables describing source's ignorance (uncertainty)} \\ \mathbf{P} & \text{random variables describing source's past history (knowledge)} \end{cases}$$

we will come to following results:

- when conditional pmf $g_j(\mathbf{f}_i|\mathbf{p}_i)$ on a subset is given ($\mathbf{U}_i$ non-void):
  ${}^O\hat{h}(\mathbf{x}_i) = \lambda_0^*(D) + \sum_{j=1}^s \lambda_j^*(D)^O\hat{h}(\mathbf{u}_i|\mathbf{f}_i,\mathbf{p}_i)g_j(\mathbf{f}_i|\mathbf{p}_i)^O\hat{h}(\mathbf{p}_i)$, for $i = 1, \ldots, n$,
- when conditional pmf $g_j(\mathbf{f}_i|\mathbf{p}_i)$ on whole set is given ($\mathbf{U}_i$ void):
  ${}^O\hat{h}(\mathbf{x}_i) = \lambda_0^*(D) + \sum_{j=1}^s \lambda_j^*(D)g_j(\mathbf{f}_i|\mathbf{p}_i)^O\hat{h}(\mathbf{p}_i)$, for $i = 1, \ldots, n$,
- when marginal pmf $g_j(\mathbf{p}_i)$ is given ($\mathbf{U}_i$ non-void, $\mathbf{F}_i$ void):
  ${}^O\hat{h}(\mathbf{x}_i) = \lambda_0^*(D) + \sum_{j=1}^s \lambda_j^*(D)^O\hat{h}(\mathbf{u}_i|\mathbf{p}_i)g_j(\mathbf{p}_i)$, for $i = 1, \ldots, n$.

The last step: project the result back on source's domain by computing adequate marginal (conditional) pmfs.

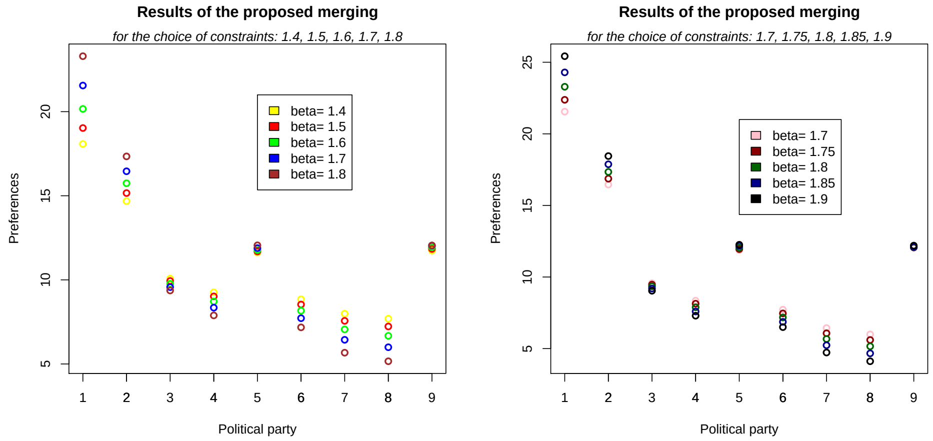
3 Simple example, numerical estimate of bounds $\beta_j(D)$

Imagine there are upcoming elections and we have 9 political parties. Agencies try to ask as much as possible people to discuss the preferences of particular political parties. We now have data from 4 agencies, in the form of particular probabilities of considered parties. According to the notation used in text we have:

- 4 sources: $j = 1, \ldots, 4$,
- 1 random variable X: "preference of a political party",

Voting preferences				
Agency	I	II	III	IV
No. of respondents	1085	1196	11364	713
Polit. party	Voting preferences in %			
1	27	27.8	30.8	30
2	21.2	18.6	18.9	22.5
3	7.5	9.3	15.1	11.5
4	4.3	8.1	8	9
5	16.8	9.9	12.9	13
6	7.4	4.9	5.9	4
7	4.8	3.2	3.2	4
8	3	3.1	3.9	3
9	8	15.1	1.3	3

Table 1: Given data



(a) Concrete results I

(b) Concrete results II

Figure 3: Graphic outputs for different choices of $\beta_j(D) = \beta(D)$, see (3)

- 9 realizations of this random variable: x_i – i^{th} political party will win, $i = 1, \ldots, 9$,
- pieces of information given by the sources: appearance of each realization expressed by percents - see Table 1 (after division by 100 we can consider the given information as probability vectors of realizations).

Now we try to apply the proposed method; the results in Figure 3a and Figure 3b were obtained by using Matlab for different choices of bounds $\beta_j(D) = \beta(D)$ - which means we tried to find one common bound for the constraints in (3).

The Table 2 contains the results printed in Figure 3a, 3b; the penultimate column is containing the original results (o.r.) and in the last column are the results obtained by using relative frequencies (r.f.), i.e., taking all data pieces as one consistent source.

4 Advantages of this method and open problems

Supra-Bayesian approach to combining different types of given information (often, incompletely compatible) introduced in this paper brings following improvement against available techniques:

- non-probabilistic data pieces are treated,
- incompletely compatible data pieces are treated,

	Choices of $\beta(D)$									
polit. party	1.4	1.5	1.6	1.7	1.75	1.8	1.85	1.9	o.r.	r.f.
1	18.1	19.0	20.2	21.6	22.4	23.3	24.3	25.4	22.09	30.22
2	14.7	15.2	15.7	16.5	16.9	17.3	17.9	18.4	20.22	19.22
3	10.1	9.9	9.8	9.6	9.5	9.4	9.2	9.04	16.71	13.86
4	9.3	9.0	8.7	8.3	8.1	7.9	7.6	7.3	10.88	7.78
5	11.6	11.7	11.8	11.9	11.9	12.0	12.2	12.3	11.27	12.95
6	8.8	8.5	8.2	7.7	7.5	7.2	6.9	6.5	4.39	5.84
7	8.0	7.6	7.1	6.4	6.1	5.7	5.2	4.7	2.44	3.36
8	7.7	7.2	6.7	6.0	5.6	5.2	4.7	4.1	4.33	3.72
9	11.7	11.8	11.9	12.0	12.1	12.1	12.1	12.2	7.67	3.04

Table 2: Matlab results

- we get unified Bayesian solution,
- this approach can be applied on every source from the group of sources; the group can be extremely large and distributed.

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

- the unambiguity of the projection of ${}^{O}\hat{h}$ back on source's domain,
- a little search for the choice of the bounds $\beta_j(D)$ in (3), $j = 1, \dots, s$ was done, but still needs improvements,
- and also the proof of existence and unambiguity of Kuhn-Tucker multipliers $\lambda_j(D)$, $j = 1, \dots, s$ is needed.

These problems are definitely topics of a future work.

5 Acknowledgement

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

References

[1] Franz Dietrich. Bayesian group belief. *Soc. Choice Welfare*, 35(4):595–626, 2010.

[2] Olga G. Troyanskaya et al. A Bayesian framework for combining heterogeneous data sources for gene function prediction (in Saccharomyces cerevisiae) . *PNAS*, 100(14):8348–8353, 2003.

[3] M. Kárný. Knowledge elicitation via extension of fragmental knowledge pieces. *Proceedings of the European Control Conference 2009*, pages 1571–1575.

[4] Vladimíra Sečkárová. Supra-Bayesian Combination of Probability Distributions. Master's thesis, Charles University in Prague, Faculty of Mathematics and Physics, Department of Probability and Mathematical Statistics, Sokolovská 83, Prague, 2010.

[5] Morris H. DeGroot. *Optimal statistical decisions.* McGraw-Hill Series in Probability and Statistics. New York etc.: McGraw- Hill Book Company. XVI, 489 p. 139 s. , 1970.

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

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

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