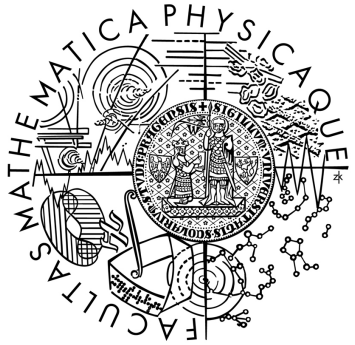
Charles University in Prague
Faculty of Mathematics and Physics

DOCTORAL THESIS



Vladimíra Sečkárová

Cross-entropy based combination of discrete probability distributions for distributed decision making

Department of Probability and Mathematical Statistics

Supervisor of the doctoral thesis:	Ing. Miroslav Kárný, DrSc.
Study programme:	Mathematics
Specialization:	Probability and Mathematical Statistics

Prague 2015

I would like to express my 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 loved ones and my friends for their unlimited support.

This research has been partially supported by GAČR 13-13502S.

I declare that I carried out this doctoral thesis independently, and only with the cited sources, literature and other professional sources.

I understand that my work relates to the rights and obligations under the Act No. 121/2000 Coll., the Copyright Act, as amended, in particular the fact that the Charles University in Prague has the right to conclude a license agreement on the use of this work as a school work pursuant to Section 60 paragraph 1 of the Copyright Act.

In ..... date ..... signature of the author

Název práce: Kombinování diskrétních pravděpodobnostních rozdělení pomocí křížové entropie pro distribuované rozhodování

Autor: Vladimíra Sečkárová

Katedra: Katedra pravděpodobnosti a matematické statistiky

Vedoucí disertační práce: Ing. Miroslav Kárný, DrSc., Ústav teorie informace a automatizace AV ČR, v.v.i.

Abstrakt: Tato práce se zabývá návrhem systematického kombinování diskrétních pravděpodobnostních distribucí založeném na teorii rozhodování a teorii informace, konkrétně na křížové entropii (známé také jako Kullbackova-Leiblerova (KL) divergence). Optimální kombinací je pravděpodobnostní funkce minimimalizující podmíněnou střední hodnotu KL-divergence. Hustota pravděpodobnosti, která se váže k této střední hodnotě, rovněž minimalizuje KL-divergenci za podmínek vztažených k řešenému problému. Ačkoliv je kombinace odvozena pro pravděpodobnostní typ informace na společném nosiči, můžeme ji po transformaci a/anebo rozšíření použít i pro míchání jiných typů informace. Práce také zahrnuje diskuzi o navrhovaném kombinování a sekvenčním zpracování dat, opakujících se datech, ovlivnění výsledků preferencemi mezi zdroji informace a aplikaci na reálná data.

Klíčová slova: teorie distribuovaného rozhodování, Kullbackova-Leiblerova divergence, princip minimální křížové entropie

Title: Cross-entropy based combination of discrete probability distributions for distributed decision making

Author: Vladimíra Sečkárová

Department: Department of Probability and Mathematical Statistics

Supervisor: Ing. Miroslav Kárný, DrSc., The Institute of Information Theory and Automation of the Czech Academy of Sciences

Abstract: In this work we propose a systematic way to combine discrete probability distributions based on decision making theory and theory of information, namely the cross-entropy (also known as the Kullback-Leibler (KL) divergence). The optimal combination is a probability mass function minimizing the conditional expected KL-divergence. The expectation is taken with respect to a probability density function also minimizing the KL divergence under problem-reflecting constraints. Although the combination is derived for the case when sources provided probabilistic type of information on the common support, it can applied to other types of given information by proposed transformation and/or extension. The discussion regarding proposed combining and sequential processing of available data, duplicate data, influence of the results by preferences among sources of information and application on real data are also included.

Keywords: distributed decision making, Kullback-Leibler divergence, minimum cross-entropy principle.

Contents

Introduction

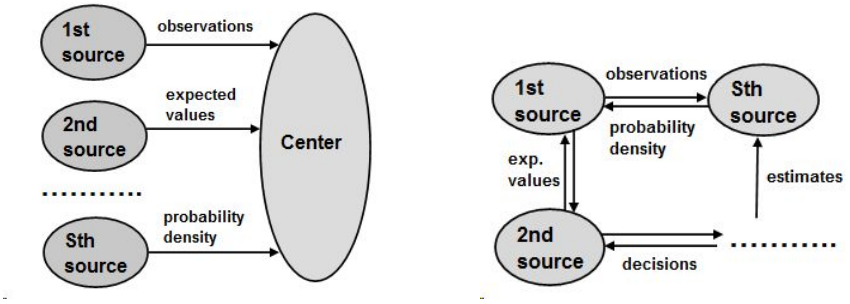
It is in the human's nature to enrich his knowledge by learning from own experience and experience of others. But how important is the information (opinions, ideas, suggestions ...) from others and how should it be incorporated into our knowledge? These questions form the departure point of this thesis, which intends to contribute to their mathematical solutions whenever and wherever appropriate.

Decentralized (distributed) decision making

Decision making (DM) is a complex theory consisting of several parts including information collecting and processing, construction of the set of appropriate decisions, etc., all performed to obtain the final decisions optimal in some a priori determined sense. Handling such complex processes has been approached from two different viewpoints: we can either consider a centralized DM with center as the dominating element, or distributed DM (for examples see Fig. 1). Since "Modern society, with its overwhelming diversity of interests and developments and its ever growing complexity, can no longer be understood and governed by the paradigm of centralized decision making" (Schneeweiss, 2003), we focus on the distributed approach. The recent contributions in distributed DM include, e.g., multi-agent online learning (Xu et al., 2015), estimation over adaptive networks (Tu and Sayed, 2014), estimation in dynamic systems (Carli et al., 2008), task switching (Wagenpfeil et al., 2009).

DM lets us view each *source* of information as a decision maker, an individual with its own past experience and environment, taking actions, focusing on subjective aims, etc. The processes in DM can be done for one decision maker, but the task becomes more interesting when other interacting sources (data sensors, experts) are available and taken into account. Based on the level of cooperation among sources several cooperation scenarios can be exploited (Kárný et al., 2007). In this work we focus on *wise selfish source*, which follows its aims, but knows that without respecting others it can reach less than possible cooperative scenario. For more details on DM with multiple sources see also (Kárný and Guy, 2004).

The issue arising within the considered scenario is that in majority of the developed approaches the final decisions are constructed to serve the whole group of decision makers suppressing the sources' personal aims. While constructing the final decisions, problems might occur due to limited abilities of sources, e.g., in the case when some sources can only provide a partial information about the studied problem. Such sources can then also experience problems with exploiting



(a) Centralized decision making. (b) Distributed decision making.

Fig. 1: Examples of centralized and decentralized (distributed) decision making. On the left: Centralized decision making. On the right: Distributed decision making.

the final decision due their limited abilities (they simply need not understand the combined information). Although the literature in distributed DM is very rich and new approaches are still being developed, this shortcoming has not received much attention. The closest work dealing with this problem can be found in (Kárný et al., 2009). There, the authors suggest that beside the knowledge of the source also its uncertainty and aims should be treated as random variables (see the definition of *behavior* therein). Since these characteristics are likely to change over time, the authors consider a probability distribution over their values. In particular, the probability distribution of each source is expressed in probabilistic form as an 'ideal' probability density function (pdf).

To overcome the addressed issue we assume the probabilistic type of information, too. But instead of determining the 'ideal' pdf we control the acceptance of the final combination of provided information. This is taken as the final decision about compromise among sources cooperating in this way. More details are given in the next paragraphs.

Probabilistic opinion pooling

In this work we assume the existence of a group (set) of sources, providing their opinions about the underlying (studied) problem to each other. The problem relates to a hidden (stochastic) phenomenon, that is not directly observable, but about which an opinion can be formulated. Examples of such phenomena are anticipated elections results, companies contracts and many others. The assumption on obtaining opinions yields a specific decision making process commonly known as opinion pooling. Due to the complexity of the space of possible decisions we consider the probability distributions over this set rather than single values. The final decision (result of pooling) will be then a probability distribution. Since we also assume the sources provided probability distribution, the final decision is a combination of probability distributions.

Combining probabilistic information within a group of sources has been of interest for a long time, see, e.g., (von Neumann and Morgenstern, 2007). Different approaches consider different levels of cooperation among sources and different exploitations of their proposed combination. One can assume the sources are rep-

resented as a group of individuals "who must act together as a team and reach consensus" (DeGroot, 1974). Or, one can consider that sources are "perfectly coherent, rational as decision makers and cooperate in agreeing to adopt a group utility function" (West, 1984).

For combining non-probabilistic type of information see, e.g., method based on coherency (of the combination) defined by the moments (Wisse et al., 2008), or based on ranks and linearly updated scores (Wu et al., 2009). The addressed problem outlined above (cooperating selfish sources) is still insufficiently resolved.

Cross-entropy and combining probability distributions

Many probability combining approaches, including the approach proposed in this thesis, heavily exploit the Kullback-Leibler (KL) divergence (Kullback and Leibler, 1951) (in the literature the term cross-entropy is also used). Based on the order of its arguments we arrive at two basic classes - linear pools and log-linear pools, see, e.g., (Abbas, 2009).

Linear pools consider, e.g., Garcia and Puig (2004), where "expert opinion is represented as a probability and associated with a confidence level that expresses the conviction of the corresponding expert on its own judgement". For linear pool exploiting entropy instead of cross-entropy and related to this thesis see (Sečkárová, 2013).

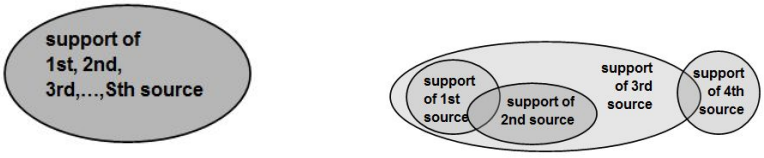
The examples of KL-divergence based log-linear pools include, e.g., reliability estimation (Dedecius and Sečkárová, 2013a), or determination of weights for prior distributions pooling (Rufo et al., 2012).

KL-divergence based combination as compromise capturing individual

The result of any pooling can be viewed as a compromise among considered sources, a combination of sources' opinions, where every included individual has to sacrifice its own aims in order to satisfy group aims. Notice that in the above mentioned works for combining probability distributions (with/without KL-divergence) the source's personal aims has never been treated. The only work known to the author dealing with this issue is (Kárný et al., 2009), where the 'ideal' pdf is included in combining, see comments above.

We deal with this shortcoming by introducing the constraints on acceptance of the desired combination of opinions by individual sources, which still yield a compromise among the sources. This suggestion together with the KL-divergence as dissimilarity measure between individuals and desired combination help us form a (weighted) linear combination of sources probability distributions.

Although the proposed combination is generally built to combine joint probabilities over a common support, it is applicable to combine also partial knowledge and non-probabilistic type of given information. When partial knowledge is available (see Fig. 2), i.e., the sources provide conditional and marginal probabilities, we propose an extension to joint probabilities. If another type of information is provided, such as a subset of possible outcomes or expected values, we propose a transformation into probabilities, which can be then extended, if necessary.



(a) Equal supports. (b) Different supports.

Fig. 2: Distributed decision making: Example of how the supports of sources (decision makers) can vary. On the left: common support. On the right: Different supports.

Throughout the thesis we assume we have a finite number of sources. If we deal with a large group of sources, this can be decomposed into smaller, possibly overlapping, groups and the proposed combination can be applied.

Thus, the proposed approach, representing a unified way for combining probabilistic and non-probabilistic (even partial) type of information, can be exploited in a wide range of problems in distributed decision making.

Brief summary of the proposed combining

In the thesis we consider a group of sources acting within a cooperative scenario. They share their information (opinions) to enhance the description/evidence about the current state of stochastic phenomenon. Our suggestion is to combine provided information so that our general requirement - to serve *each source in the group* while representing a compromise among sources - is satisfied.

In order to obtain the optimal combination of available opinions, their representation in the form of a probability mass function (pmf) is assumed in this work. To define the notion of optimality we exploit the theory of Bayesian decision making (BDM), see (Savage, 1972). BDM searches for a decision rule, a mapping from the set of observations to the set of actions, minimizing the expected loss function. In our case, the available information serves as the observations and the set of possible estimators serves as the set of actions. The searched decision is then a combination of given pmfs and is constructed as an estimator of an unknown pmf minimizing the expected loss function (see step 2 of Alg. 1 below).

Similarly to earlier discussed approaches we adopt the KL-divergence as the loss function (Bernardo, 1979). We show that the optimal estimator is then itself a conditional expectation with respect to an unspecified probability density function (pdf). To obtain the corresponding conditional pdf we again exploit the KL-divergence, namely the minimum cross-entropy principle (Shore and Johnson, 1980), rather than the Bayes rule. Still, both approaches can yield the same result, as outlined by Campenhout and Cover (1981). Finally, we express the optimality conditions imposed on the estimator to be a compromise based on the available pmfs by means of the KL-divergence.

If a source provides conditional pmfs of a subset of considered random variables given the values of the remaining random variables, we construct its joint version by exploiting the minimum cross-entropy principle or the maximum entropy principle. This also applies to the case when some source provides a

marginal pmf of a specific subset of considered random variables (see step 1 of Alg. 1 below). Then, the extended pmfs can be combined by proposed approach and sources obtain the appropriate conditional/marginal version of the optimal combination (see step 3 of Alg. 1).

If a subset of the outcomes of the random vector is available, we perform the transformation into a pmf by using Kronecker delta.

Layout of the work

The construction, properties and examples of use of the optimal estimator form the core of the thesis. It consists of seven chapters briefly summarized below:

- The first chapter contains the definitions of the terms used in the thesis.
- In the second chapter we derive the combination under general setup.
- The third chapter includes the combination when the Dirichlet distribution is assumed. This form of optimal combination of pmfs will be then used in the rest of the thesis. This chapter also includes the properties and illustrative examples of the derived combination.
- The novel contribution of the proposed method lies in handling the sources' diversity, which allows us to treat cases when sources are unable to provide joint pmfs or, e.g., to include variables representing sources' personal aims. The diversity is in particular expressed by marginal and conditional probability distributions, subsets of outcomes and expected values of the underlying random vector. Preparation (extension, transformation) of given probabilistic and non-probabilistic information for combining is of interest in the fourth chapter.
- The proposed combination is then applied to real data examples in the fifth chapter.
- In the sixth chapter a comparison of the proposed approach with a diffusion distributed approach is given. The chapter also includes the detailed description of the second method introduced by Dedecius and Sečkárová (2013b) and brief comparison to the proposed combination.
- The conclusion and future work plans are given in the seventh chapter.

Data: Non-probabilistic and probabilistic information from s sources, $j = 1, \ldots, s < \infty$.
Result: The cross-entropy based combination of pmfs arising from transformation and extension of information given by sources.
begin

1 Prepare the provided information for combining.
for *j^{th} sources (j=1,. . . ,s)* **do**
if *non-probabilistic information is given* **then**
transform it into probabilistic form (see Section 4.2)
end
if *extension of a probabilistic information (a pmf) is needed* **then**
extend given probabilistic information (see Section 4.1)
end
end

2 Combine the (transformed and/or extended) probabilistic information (pmfs) $p_1, \ldots, p_s$ from sources by using (3.3)**, i.e.,**

• Set the prior values of the parameters of the Dirichlet distribution in terms of prior pmfs, e.g., as the arithmetic mean of $p_1, \ldots, p_s$ (see Section 3.1);
• Numerically solve the optimization problem (3.10) to obtain the estimates of the parameters of the Dirichlet distribution yielding the final combination (see Section 3.2.1);
• Alternatively, numerically solve the optimization problem (3.11) to obtain the coefficients in the final combination (see Section 3.2.2).

For general setup see Chapter 2.
For setup with the Dirichlet distribution see Chapter 3.

3 Project the resulting combination on the support of each source.
end

Alg. 1: Proposed cross-entropy based combining.

Chapter 1

Preliminaries

1.1 Kullback-Leibler divergence

Definition 1.1. *Assume two probability measures P and Q on a measurable space X, P absolutely continuous with respect to Q. We define the Kullback-Leibler (KL) divergence as :*

$$\begin{aligned} \mathrm{KLD}(P||Q) = \mathrm{E}_P \ln \frac{P}{Q} = \int_X \ln \frac{P}{Q}\, dP, \quad P \ll Q, \\ = \infty \qquad\qquad\qquad\qquad \text{otherwise,} \end{aligned} \tag{1.1}$$

where E_P denotes the expectation with respect to probability distribution P.

Remark. *If there exists a (Lebesgue, counting...) measure μ defined on X for which the Radon-Nikodym derivatives $p = \frac{dP}{d\mu}$ and $q = \frac{dQ}{d\mu}$ exist (i.e., the pdfs or pmfs exist), the KL divergence simplifies to*

$$\mathrm{KLD}(p||q) = \int_X p(x) \ln \frac{p(x)}{q(x)} \mu(dx). \tag{1.2}$$

KL-divergence is also commonly referred to as cross-entropy. We shall use both terms as synonyms.

Branches of mathematics exploiting KL-divergence include testing statistical hypotheses (Basu et al., 2013), optimization (Fischer, 2010), decision making (Kißlinger and Stummer, 2013), image processing (Villena et al., 2010), statistical inference for methane emissions (Sabolová et al., 2015), etc.

Properties of KL-divergence

The following is trivial to show:

- KL-divergence is a premetric.
- $\mathrm{KLD}(P||Q) \geq 0$ with equality iff $P = Q$ a.e.
- KL-divergence does not satisfy the triangle inequality nor is symmetric.

For more details, see (Kullback, 1997).

Decomposition of KL-divergence

The KL-divergence can be decomposed into:

$$\mathrm{KLD}(P||Q) = -\mathrm{E}_P \ln Q + \mathrm{E}_P \ln P = \mathrm{Kerr}(P||Q) - \mathrm{H}(P), \tag{1.3}$$

where Kerr(.||.) is the Kerridge inaccuracy (Kerridge, 1961) and H(.) is the Shannon entropy (Shannon, 1948).

1.2 Maximum entropy and minimum cross-entropy principles

Assume that we have a prior guess $p_0(x)$ of the pdf $p(x)$ of X. Assume also that we obtained a partial information delimiting a set of all adequate pdfs $p(x)$. Question that arises is how to enrich the prior guess by this piece of information?

Shore and Johnson (1980) axiomatically justified that when certain values or bounds on the expectations of $h(X)$ are given

$$\mathcal{S} = \left\{ p : \int_X p(x)h(x)dx = d \text{ or } \leq d \right\},$$

where $h(.)$ is a given vector function, minimization of $\mathrm{KLD}(p||p_0)$ over $\mathcal{S}$ is the adequate way how to choose a single element p from $\mathcal{S}$. Cover and Thomas (2006) have shown that the outcome minimum cross-entropy often coincides with conditioning by $\mathcal{S}$.

In the special case of uniform p_0, the maximum entropy principle is obtained. The popularity of this principle is based on the fact that entropy is "a unique, unambiguous criterion for 'the amount of uncertainty' (Shannon, 1948), which agrees with our intuitive notions that a flat distribution represents more uncertainty than does a sharply peaked one, and satisfies all other conditions, which make it reasonable", (Jaynes, 1957).

For further details on axiomatic derivation of both principles see (Shore and Johnson, 1980).

1.3 Dirichlet distribution

In this section we focus on the basic properties of this continuous multivariate distribution extensively used in this thesis. For further details see, e.g., (Kotz et al., 2005).

The support of this distribution is an $(n-1)$ dimensional probability simplex $\mathcal{Q}_{sim}$, a set of all n-tuples satisfying

$$Q_{sim} = \left\{ (p_1, \ldots, p_n) : \sum_{i=1}^{n} p_i = 1, 0 \leq p_i \leq 1, i = 1, \ldots, n \right\} \tag{1.4}$$

An n-dimensional random vector q satisfying the simplex property ($q \in Q_{sim}$) is distributed according to the Dirichlet distribution with parameters $(\alpha_1, \ldots, \alpha_n)$ if its pdf has the form:

$$f(q_1, \ldots, q_n, \alpha_1, \ldots, \alpha_n) = \frac{1}{B(\alpha_1, \ldots, \alpha_n)} \prod_{i=1}^{n} q_i^{\alpha_i - 1}, \quad \alpha_i > 0, \quad i = 1, \ldots, n, \tag{1.5}$$

where $B(.)$ is the multivariate beta function

$$B(\alpha_1, \ldots, \alpha_n) = \frac{\prod_{i=1}^{n} \Gamma(\alpha_i)}{\Gamma(\sum_{i=1}^{n} \alpha_i)},$$

$\Gamma(.)$ is the gamma function (Abramowitz and Stegun, 1964).

The corresponding expected value of q_i is

$$\mathrm{E}[q_i] = \frac{\alpha_i}{\sum_{k=1}^{n} \alpha_k}, \quad i = 1, \ldots, n, \tag{1.6}$$

and the expected value of $\ln q_i$ is

$$\mathrm{E}[\ln[q_i]] = \psi(\alpha_i) - \psi\left(\sum_{k=1}^{n} \alpha_k\right), \quad i = 1, \ldots, n, \tag{1.7}$$

where $\psi(.)$ is the digamma (psi) function (Abramowitz and Stegun, 1964)

$$\psi(z) = \frac{d(\ln \Gamma(z))}{dz}.$$

In the thesis we exploit only real positive z, for which the digamma function is well-defined.

Chapter 2

Cross-entropy based optimal combination

In this chapter we derive the optimal combination of sources' pmfs under certain general conditions imposed below. In particular, we express the desired combination as an unknown pmf and construct its estimator meeting specific constraints delimiting what the desired combination is. The key contribution lies in the specification of the constraints so that the optimal combination serves each source as a reasonable evidence about the hidden phenomenon while being a compromise among sources. The definition of constraints based on expected dissimilarity from an unknown combination is given in Section 2.2. Part of this chapter is included in author's accepted contribution, see (Sečkárová, 2015).

2.1 The optimal combination as optimal estimator

Let us have a discrete random vector X with n possible outcomes x_i, $i = 1, \dots, n$, $n < \infty$, where each outcome x_i occurs with an unknown probability q_i. Assume that we obtained s pmfs

$$p_j = (p_{j1}, \dots, p_{jn}), \quad j = 1, \dots, s, \ s < \infty,$$

where p_{ji} is the probability assigned by the j^{th} source to the i^{th} outcome. These p_j are viewed as *observations*. Also assume that no information about the reliability of respective sources is available.

The search for the optimal combination of $p_1, \dots, p_s$ is here interpreted as the search for the optimal estimator $\hat{q}$ of unknown desired combination q based on $p_1, \dots, p_s$. To obtain $\hat{q}$ we exploit the Bayesian decision making (BDM), see (Savage, 1972). Thus, we search for the optimal estimator $\hat{q}$ minimizing the conditional expected value of a loss function given observations $p_1, \dots, p_s$, where the expected value is taken over the set of all possible pmfs q (delimited by appropriate pdf, derived later).

We will proceed as follows:

1. First, we will specify the loss function operating on the set of all possible n-dimensional probability vectors – over the $(n-1)$-dimensional probability simplex Q_{sim}, see (1.4).

2. Then, we determine the optimal estimator with respect to an unspecified conditional pdf $\pi(q|p_1,\ldots,p_s)$. Note that $\pi(q|p_1,\ldots,p_s)$ is a pdf of the n-dimensional random vector taking values in Q_{sim}.

2.1.1 Loss function

To follow the BDM methodology the loss function should express "closeness" of the estimated q and its estimator $\hat{q}$. To follow our setup we particularly need the loss function to measure the dissimilarity between pmfs. According to (Bernardo, 1979), we exploit the KL-divergence (1.1) and search for the optimal estimator $\hat{q}$, which minimizes the conditional expectation of the KL-divergence.

2.1.2 Optimal estimator

Proposition 2.1. *Let $q=(q_1,\ldots,q_n)$ be an n-dimensional random vector taking values in $(n-1)$-dimensional probability simplex Q_{sim}. Let pmfs $p_1,\ldots,p_s \in Q_{sim}$ be the observations about q and let $\tilde{Q} \subseteq Q_{sim}$ be a set of all pmfs $\tilde{q}$ satisfying*

$$\mathrm{E}_{\pi(q|p_1,\ldots,p_s)}[\mathrm{KLD}(q||\tilde{q})|p_1,\ldots,p_s] < \infty, \tag{2.1}$$

where $\mathrm{E}[.|.]$ denotes the conditional expectation with respect to a known conditional pdf $\pi(q|p_1,\ldots,p_s)$.

Then, the pmf $\hat{q}$ minimizing the conditional expectation of the KL-divergence

$$\mathrm{E}_{\pi(q|p_1,\ldots,p_s)}[\mathrm{KLD}(q||\tilde{q})|p_1,\ldots,p_s] \tag{2.2}$$

over the set $\tilde{Q}$ is

$$\hat{q} = \mathrm{E}_{\pi(q|p_1,\ldots,p_s)}[q|p_1,\ldots,p_s]. \tag{2.3}$$

Proof. The expected loss function has the following form:

$$\begin{aligned}
&\mathrm{E}_{\pi(q|p_1,\ldots,p_s)}[\mathrm{KLD}(q||\tilde{q})|p_1,\ldots,p_s] && (2.4)\\
&= \int_{\tilde{Q}} \pi(q|p_1,\ldots,p_s) \sum_{i=1}^{n} q_i \ln \frac{q_i}{\tilde{q}_i} dq \\
&= \int_{\tilde{Q}} \pi(q|p_1,\ldots,p_s) \sum_{i=1}^{n} q_i \ln q_i dq - \int_{\tilde{Q}} \pi(q|p_1,\ldots,p_s) \sum_{i=1}^{n} q_i \ln \tilde{q}_i dq \\
&= -\mathrm{E}_{\pi(q|p_1,\ldots,p_s)}[\mathrm{H}(q)|p_1,\ldots,p_s] - \sum_{i=1}^{n} \int_{\tilde{Q}} \pi(q|p_1,\ldots,p_s) q_i \ln \tilde{q}_i dq && (2.5)\\
&= -\mathrm{E}_{\pi(q|p_1,\ldots,p_s)}[\mathrm{H}(q)|p_1,\ldots,p_s] + \mathrm{Kerr}\left(\mathrm{E}_{\pi(q|p_1,\ldots,p_s)}(q|p_1,\ldots,p_s)||\tilde{q}\right), && (2.6)
\end{aligned}$$

where $\mathrm{Kerr}(.||.)$ is the Kerridge inaccuracy and $\mathrm{H}(.)$ is the entropy, c.f. Section 1.1.

The property (2.1) allows us to use the Fubini theorem and exchange the summation and the integral in (2.5).

The first term in (2.6) does not depend on $\tilde{q}$, thus the minimization of the expected loss (2.2) reduces to the minimization of the second term containing the Kerridge inaccuracy. The Kerridge inaccuracy reaches its minimum when the arguments are equal a.e., thus the minimizing pmf is

$$\mathrm{E}_{\pi(q|p_1,\dots,p_s)}(q|p_1,\dots,p_s) = \hat{q}.$$

□

The estimator $\hat{q}$ represents our optimal combination of $p_1,\dots,p_s$ and the desired decision rule of our BDM task. Commonly, the decision rule assigning $p_1,\dots,p_s$ minimizer of the conditional expected value of a specific loss function is also known as the Bayes decision rule. For the estimation task, it is also called the Bayesian estimator.

The next step in determination of the optimal combination $\hat{q}$ is the choice of the conditional pdf $\pi(q|p_1,\dots,p_s)$.

2.2 Minimum cross-entropy based conditional pdf

In this section we construct the conditional pdf $\pi(q|p_1,\dots,p_s)$ required in the combination $\hat{q}$ defined by (2.3). Since we are working within the BDM framework, let us assume the prior guess $\pi_0(q)$ about $\pi(q|p_1,\dots,p_s)$, acting on $(n-1)$-dimensional probability simplex Q_{sim}, is available. When also the likelihood function is available, we would exploit the Bayes rule to obtain $\pi(q|p_1,\dots,p_s)$. Our considered problem, however, does not provide the likelihood function. Instead of the Bayes rule, we exploit the minimum cross-entropy principle introduced in Section 1.2.

According to the minimum cross-entropy principle, we search for the conditional pdf $\pi(q|p_1,\dots,p_s)$ minimizing the KL-divergence of $\pi(q|p_1,\dots,p_s)$ from $\pi_0(q)$ with respect to constraints given by the expected values of functions of q. We, in particular, introduce the constraints on unknown pmf q by the KL-divergence between observations p_j and q. The definition of these constraints, in which we delimit the set of π-admissible pmfs q attributing to the optimal estimator $\hat{q}$, is the key contribution of this section.

Then, we derive the formula of the conditional pdf $\pi(q|p_1,\dots,p_s)$ satisfying these constraints and minimizing the above KL-divergence.

2.2.1 Kullback-Leibler divergence based constraints

The definition of constraints in the minimum cross-entropy principle is the most important part not only in the search for the conditional pdf $\pi(q|p_1,\dots,p_s)$, but consequently in determination of the optimal combination $\hat{q}$. To obtain conditional pdf $\pi(q|p_1,\dots,p_s)$, we first relate each provided pmf to desired combination q.

Since sources work within cooperative scenario, that is they are willing to share their pmfs and exploit the optimal combination, every desired combination q in Q_{sim} can be also viewed as a compromise among sources. To objectify how well the compromise q performs from j^{th} source point of view, we would like to measure the dissimilarity between p_j and q. Since both are pmfs, we again exploit the KL-divergence, c.f. (1.1), as an appropriate dissimilarity measure. To formalize the supported *selfishness*, we would like to measure how well q performs as an approximation of p_j. Thus, p_j enters the KL-divergence as the first argument and q as the second argument: $KLD(p_j||q)$. Since q is unknown, we focus on the dissimilarity defined by the expected value of the KL-divergence with respect to $\pi(q|p_1,\ldots,p_s)$: $\mathrm{E}_{\pi(q|p_1,\ldots,p_s)}[\mathrm{KLD}(p_j||q)|p_1,\ldots,p_s]$.

To relate the sources and their pmfs $p_1,\ldots,p_s$ we consider a *wise selfish* cooperative scenario among sources. In particular, we assume that the j^{th} source is aware that other sources influence the combination, but selfishly would like its dissimilarity to be less than or at most equal to the dissimilarities of other sources

$$\mathrm{E}_{\pi(q|p_1,\ldots,p_s)}[\mathrm{KLD}(p_j||q)|p_1,\ldots,p_s] \leq \mathrm{E}_{\pi(q|p_1,\ldots,p_s)}[\mathrm{KLD}(p_l||q)|p_1,\ldots,p_s],$$

$l,j=1,\ldots,s$, $l\neq j$. A closely related problem has been studied in (Sečkárová, 2013).

The optimal combination $\hat{q}$, as the estimator of compromise q based on the conditional pdf $\pi(q|p_1,\ldots,p_s)$, should serve all sources in the group. Since we have no prior information about the quality ("reliability") preferences among respective sources, we have to impose the equality constraint. These informal considerations motivate the formal definition of constraints delimiting a subset $Q_\pi \subset Q_{sim}$. Pmfs in Q_π are perceived as appropriate pfms non-contradicting information about desired combination provided by the group of undistinguishable information sources $j=1,\ldots,s$.

Definition 2.2. *Let the conditional pdf $\pi(q|p_1,\ldots,p_s)$ satisfy the following constraints on the expected value of the KL-divergence:*

$$\mathrm{E}_{\pi(q|p_1,\ldots,p_s)}[\mathrm{KLD}(p_j||q)|p_1,\ldots,p_s] = \mathrm{E}_{\pi(q|p_1,\ldots,p_s)}[\mathrm{KLD}(p_s||q)|p_1,\ldots,p_s], \quad j=1,\ldots,s-1, \tag{2.7}$$

where

$$\mathrm{E}_{\pi(q|p_1,\ldots,p_s)}[\mathrm{KLD}(p_j||q)|p_1,\ldots,p_s] < \infty,\ j=1,\ldots,s.$$

Pmfs q, included in the subset Q_π of $(n-1)$-dimensional probability simplex Q_{sim} (1.4) satisfying the above constraints, will be called π-admissible pmfs.

Remark. *The right-hand side of constraints in (2.7) is defined with respect to the s^{th} source. Since we assumed the equality among conditional expectations, any source $j=1,\ldots,s-1$ can be used on the right-hand side of (2.7).*

Let $\mathcal{S}$ denote the set of all conditional pdfs $\pi(q|p_1,\ldots,p_s)$ satisfying (2.7) (each having specific support Q_π). We now inspect the cardinality of the set $\mathcal{S}$ in order to recognize whether $\mathcal{S}$ is not empty.

Let us rewrite the equations (2.7) in the Definition 2.2 as follows:

$$-H(p_j) + \int_{Q_{sim}} \pi(q|p_1,\dots,p_s) \sum_{i=1}^{n} p_{ji} \ln \frac{1}{q_i} dq \\ = -H(p_s) + \int_{Q_{sim}} \pi(q|p_1,\dots,p_s) \sum_{i=1}^{n} p_{si} \ln \frac{1}{q_i} dq,$$

yielding

$$-H(p_j) + H(p_s) = \int_{Q_{sim}} \pi(q|p_1,\dots,p_s) \sum_{i=1}^{n} (p_{si} - p_{ji}) \ln \frac{1}{q_i} dq \\ = \sum_{i=1}^{n} (p_{ji} - p_{si}) \mathrm{E}_{\pi(q|p_1,\dots,p_s)} \left[\ln q_i | p_1,\dots,p_s\right],$$

where $j = 1,\dots,s-1$ and H(.) is the entropy. Thus, the equations in (2.7) coincide with the following system of linear equations:

$$-H(p_j) + H(p_s) = \sum_{i=1}^{n} (p_{ji} - p_{si}) a_i, \quad j = 1,\dots,s-1, \tag{2.8}$$

where

$$a_i = \mathrm{E}_{\pi(q|p_1,\dots,p_s)} \left[\ln q_i | p_1,\dots,p_s\right], \quad i = 1,\dots,n. \tag{2.9}$$

If $s - 1 < n$, there are infinitely many vectors

$$a = (a_1,\dots,a_n),$$

solving the system (2.8), and thus infinitely many conditional pdfs $\pi(q|p_1,\dots,p_s)$ satisfying (2.9).

For $n \leq s-1$ the system (2.8) need not to have a solution. If solution of (2.8) exists, then it can be plugged into (2.9) and yields many pdfs $\pi(q|p_1,\dots,p_s)$. If no solution exists, we require differences of the left- and right-hand side in (2.7) to be close to zero in an assumed sense, e.g., the least-squares one adopted below. Thus, we focus on the approximate solution a of the system (2.8) obtained by method of least squares. Note that we search for approximate solution within the set of negative n-tuples, since each a_i, $i = 1,\dots,n$, denotes the expected value of a logarithm of a random variable taking values in $[0,1]$. We obtain a unique vector a and, as before, many conditional pdfs $\pi(q|p_1,\dots,p_s)$ can satisfy (2.9). Thus, the set $\mathcal{S}^{\bullet}$ of all pdfs $\pi(q|p_1,\dots,p_s)$ satisfying constraints (2.7) in the least-squares sense, is non-empty. In the following section we discuss which pdf from $\mathcal{S}^{\bullet}$ to choose and derive its formula.

2.2.2 Conditional pdf of π-admissible vectors

In the previous section we introduced the notion of π-admissible pmfs q by means of the expected KL-divergences between observations p_j and unknown desired combination q. These expectations are taken with respect to the yet unspecified

conditional pdf $\pi(q|p_1,\ldots,p_s)$ in a non-empty, typically infinite-dimensional, set $\mathcal{S}$.

The choice of a single pdf from $\mathcal{S}$, the last step in the search for the conditional pdf $\pi(q|p_1,\ldots,p_s)$ describing π-admissible pmfs q, is based on the minimum cross-entropy principle introduced in Section 1.2. That is, we choose the conditional pdf $\pi(q|p_1,\ldots,p_s)$ satisfying (2.7) (at least in the least-squares sense) and minimizing the KL-divergence of $\pi(q|p_1,\ldots,p_s)$ from a known prior pdf $\pi_0(q)$.

Since the domain of the prior pdf $\pi_0(q)$ is the $(n-1)$-dimensional probability simplex Q_{sim}, $\pi_0(q)$ can be expressed as a function of $p_1,\ldots,p_s$. To highlight this property the prior pdf will be from now on denoted by $\pi_0(q|p_1,\ldots,p_s)$. This pdf generally does not fulfill the constraints (2.7) with $\pi_0(q|p_1,\ldots,p_s)$ in the role of $\pi(q|p_1,\ldots,p_s)$.

First we study the uniqueness of the posterior pdf $\pi(q|p_1,\ldots,p_s)$ resulting from the minimum cross-entropy principle.

Proposition 2.3. *Let $\mathcal{S}^\bullet$ denote the set of all conditional pdfs $\pi(q|p_1,\ldots,p_s)$ satisfying (2.7) in the least-squares sense and assume*

$$\mathcal{S}^* = \mathcal{S}^\bullet \cap \left\{\pi(q|p_1,\ldots,p_s) : \mathrm{KLD}\left(\pi(q|p_1,\ldots,p_s)\middle|\middle|\pi_0(q|p_1,\ldots,p_s)\right) < \infty\right\}$$

and known prior pdf $\pi_0(q|p_1,\ldots,p_s)$ having its support Q_{sim}. Then, the pdf $\hat{\pi}(q|p_1,\ldots,p_s)$, solving the following minimization problem

$$\min_{\pi(q|p_1,\ldots,p_s)\in\mathcal{S}^*} \mathrm{KLD}\left(\pi(q|p_1,\ldots,p_s)\middle|\middle|\pi_0(q|p_1,\ldots,p_s)\right), \tag{2.10}$$

is unique.

Proof. To show that the minimizer of (2.10) in the set $\mathcal{S}^*$ is unique, we first inspect the convexity of $\mathcal{S}^*$.

Let $(a_1^*,\ldots,a_n^*)$ denote the solution of linear system (2.8) resulting from least-squares. For pdfs $\pi_1,\pi_2\in\mathcal{S}^\bullet$ we have

$$a_i^* = \mathrm{E}_{\pi_1}[\ln q_i|p_1,\ldots,p_s] = \mathrm{E}_{\pi_2}[\ln q_i|p_1,\ldots,p_s], \quad i=1,\ldots,n,$$

where $\pi_l = \pi_l(q|p_1,\ldots,p_s)$, $l=1,2$. For their convex combination $\pi^* = \alpha\pi_1+\beta\pi_2$, where $\alpha,\beta\geq 0$, $\alpha+\beta=1$, the following holds

$$\alpha\pi_1+\beta\pi_2 \geq 0, \qquad \int_{Q_{sim}} (\alpha\pi_1+\beta\pi_2)dq = \alpha+\beta = 1$$

and

$$\mathrm{E}_{\pi^*}[\ln q_i|p_1,\ldots,p_s] = \int_{Q_{sim}} (\alpha\pi_1+\beta\pi_2)\ln q_i dq = \alpha a_i^* + \beta a_i^* = a_i^*,$$

$i=1,\ldots,n$. Thus, $\pi^*\in\mathcal{S}^\bullet$ and $\mathcal{S}^\bullet$ is convex set of pdfs.

Since we have

$$KLD\left(\pi_l\middle|\middle|\pi_0(q|p_1,\ldots,p_s)\right) < \infty, \quad l=1,2,$$

based on strict concavity of the entropy H(.), see (Cover and Thomas, 2006), we obtain that also

$$\begin{aligned} KLD\left(\pi^*\middle|\middle|\pi_0(q|p_1,\dots,p_s)\right) &= \int_{Q_{sim}} (\alpha\pi_1+\beta\pi_2)\ln\frac{\alpha\pi_1+\beta\pi_2}{\pi_0(q|p_1,\dots,p_s)}dq \\ &= -H(\alpha\pi_1+\beta\pi_2) - \int_{Q_{sim}} (\alpha\pi_1+\beta\pi_2)\ln\pi_0(q|p_1,\dots,p_s)dq \\ &< -\alpha H(\pi_1) - \beta H(\pi_2) \\ &\quad - \int_{Q_{sim}} \alpha\pi_1\ln\pi_0(q|p_1,\dots,p_s)dq - \int_{Q_{sim}} \beta\pi_2\ln\pi_0(q|p_1,\dots,p_s)dq \\ &= KLD\left(\pi_1\middle|\middle|\pi_0(q|p_1,\dots,p_s)\right) + KLD\left(\pi_2\middle|\middle|\pi_0(q|p_1,\dots,p_s)\right) < \infty. \end{aligned}$$

The set $\mathcal{S}^*$, as an intersection of two convex sets, is also a convex set.

Since the KL-divergence is strictly convex in its first argument and is bounded below by zero, its unique infimum exists on a non-empty convex set of pdfs.

We first examine whether or not the infimum can be on the boundary of the set $\mathcal{S}^*$. We focus on pdfs with shrinking support, in particular when pdfs approach the Dirac delta function. Since every Dirac delta function can be viewed as a limit of (in this case truncated) normal multivariate distributions $N^T(\mu,\sigma^2)$ with appropriate mean and variance σ^2 converging to zero we obtain that

$$\lim_{\sigma^2\to 0} \mathrm{KLD}\left(N^T(\mu,\sigma^2)\middle|\middle|\pi_0(q|p_1,\dots,p_s)\right) = \infty.$$

Thus, the infimum can not be reached on the boundary of the set $\mathcal{S}^*$ and KL-divergence reaches on the set $\mathcal{S}^*$ its minimum $\hat{\pi}$. This minimum is unique; otherwise there would exist another pdf $\hat{\pi}_2$ minimizing the KLD and satisfying:

$$\omega = \mathrm{KLD}(\hat{\pi}||\pi_0) = \mathrm{KLD}(\hat{\pi}_2||\pi_0).$$

Strict convexity implies for $\alpha+\beta=1$, $\alpha,\beta\geq 0$ that

$$\mathrm{KLD}(\alpha\hat{\pi}+\beta\hat{\pi}_2||\pi_0) < \alpha\mathrm{KLD}(\hat{\pi}||\pi_0) + \beta\mathrm{KLD}(\hat{\pi}_2||\pi_0) = \omega,$$

which is in contradiction with $\hat{\pi}_1$, $\hat{\pi}_2$ being different minimizers. □

In the next proposition we finally derive the conditional pdf $\pi(q|p_1,\dots,p_s)$. To obtain the explicit formula we assume that the constraints (2.7) can be met.

Proposition 2.4. *Let Q_{sim} be the $(n-1)$-dimensional simplex and $\pi(q|p_1,\dots,p_s)$ be a known prior pdf acting on Q_{sim}. Let the set of all pdfs satisfying* (2.7) *be non-empty. The conditional pdf $\hat{\pi}(q|p_1,\dots,p_s)$ minimizing the KL-divergence in* (2.10) *under the constraints* (2.7) *is*

$$\hat{\pi}(q|p_1,\dots,p_s) \propto \pi_0(q|p_1,\dots,p_s)\prod_{i=1}^{n} q_i^{\sum_{j=1}^{s-1}\lambda_j(p_{ji}-p_{si})}. \tag{2.11}$$

where λ_j are the Lagrange multipliers chosen so that (2.7) *is met.*

Proof. Minimization of (2.10) with respect to equality constraints in (2.7) forms the task of constrained non-linear optimization. In order to determine the form of the conditional pdf $\hat{\pi}(q|p_1,\ldots,p_s)$ we can exploit Lagrange multipliers.

The Lagrangian of the considered optimization task looks as follows:

$$
\begin{aligned}
&KLD\left(\pi(q|p_1,\ldots,p_s)\middle|\middle|\pi_0(q|p_1,\ldots,p_s)\right)\\
&\quad+\sum_{j=1}^{s-1}\lambda_j\left(\mathrm{E}_{\pi(q|p_1,\ldots,p_s)}[\mathrm{KLD}(p_j||q)|p_1,\ldots,p_s]-\mathrm{E}_{\pi(q|p_1,\ldots,p_s)}[\mathrm{KLD}(p_s||q)|p_1,\ldots,p_s]\right)\\
&=\int_{Q_{sim}}\pi(q|p_1,\ldots,p_s)\ln\frac{\pi(q|p_1,\ldots,p_s)}{\pi_0(q|p_1,\ldots,p_s)}dq\\
&\quad+\sum_{j=1}^{s-1}\lambda_j\left(\int_{Q_{sim}}\pi(q|p_1,\ldots,p_s)\sum_{i=1}^{n}\left(p_{ji}\ln\frac{p_{ji}}{q_i}-p_{si}\ln\frac{p_{si}}{q_i}\right)dq\right)\\
&=\int_{Q_{sim}}\pi(q|p_1,\ldots,p_s)\ln\frac{\pi(q|p_1,\ldots,p_s)}{\pi_0(q|p_1,\ldots,p_s)}dq\\
&\quad+\sum_{j=1}^{s-1}\lambda_j\left(\int_{Q_{sim}}\pi(q|p_1,\ldots,p_s)\sum_{i=1}^{n}(p_{ji}-p_{si})\ln\frac{1}{q_i}dq\right)\\
&+\sum_{j=1}^{s-1}\Big(-H(p_j)+H(p_s)\Big)\int_{Q_{sim}}\pi(q|p_1,\ldots,p_s)dq\\
&=\int_{Q_{sim}}\pi(q|p_1,\ldots,p_s)\ln\frac{\pi(q|p_1,\ldots,p_s)}{\pi_0(q|p_1,\ldots,p_s)}dq\\
&\quad+\int_{Q_{sim}}\pi(q|p_1,\ldots,p_s)\sum_{i=1}^{n}\ln\frac{1}{q_i^{\sum_{j=1}^{s-1}\lambda_j(p_{ji}-p_{si})}}dq+\sum_{j=1}^{s-1}\lambda_j\Big(-H(p_j)+H(p_s)\Big)\\
&=\int_{Q_{sim}}\pi(q|p_1,\ldots,p_s)\ln\frac{\pi(q|p_1,\ldots,p_s)}{\pi_0(q|p_1,\ldots,p_s)\prod_{i=1}^{n}q_i^{\sum_{j=1}^{s-1}\lambda_j(p_{ji}-p_{si})}}dq \qquad (2.12)\\
&\quad+\sum_{j=1}^{s-1}\lambda_j\Big(-H(p_j)+H(p_s)\Big).
\end{aligned}
$$

Fubini theorem again allows us to exchange the summation and the integral. The last term does not depend on the $\pi(q|p_1,\ldots,p_s)$. The first term is the shifted KL-divergence and its minimization yields the resulting pdf (2.11). □

In the present chapter, we have derived the estimator $\hat{q}$, equation (2.3), depending on conditional pdf $\pi(q|p_1,\ldots,p_s)$ (2.11) (and on prior pdf $\pi_0(q|p_1,\ldots,p_s)$). To show straightforward use of $\hat{q}$ for combining provided pmfs, we exploit pdf of the Dirichlet distribution. This choice is supported by the fact that both pdfs, $\pi_0(q|p_1,\ldots,p_s)$ and $\pi(q|p_1,\ldots,p_s)$, are pdfs over the probability simplex and that (2.11) can be easily applied. The derivation of $\hat{q}$ in the case that $\pi_0(q|p_1,\ldots,p_s)$ is the pdf of the Dirichlet distribution and the determination of the Lagrange multipliers is treated in the next chapter.

Chapter 3

Optimal combination for Dirichlet prior and its properties

The results in Chapter 2 were derived for arbitrary prior pdf $\pi_0(q|p_1,\ldots,p_s)$. In this chapter, we derive the conditional pdf $\pi(q|p_1,\ldots,p_s)$ in (2.11) and the optimal estimator $\hat{q}$ of desired combination q in (2.3) for computationally advantageous prior pdf $\pi_0(q|p_1,\ldots,p_s)$ – a pdf of the Dirichlet distribution.

In order to obtain values of $\hat{q}$ for given pmfs $p_1,\ldots,p_s$ we deal with the constrained non-linear minimization task. To solve it, we exploit penalty optimization and form an unconstrained version. The behavior of the minimized function is studied and the relation of the parameters of the Dirichlet distribution to the proposed estimator $\hat{q}$ is presented.

Then, we discuss how the proposed estimator $\hat{q}$ deals with duplicate observations, with sequentially combined pmfs and with a change of preferences among sources.

Part of this chapter is included in author's accepted contribution (Sečkárová, 2015).

3.1 The optimal combination for Dirichlet prior

Proposition 3.1. *Let q be a random vector taking values in $(n-1)$-dimensional probability simplex Q_{sim}. Let the prior pdf $\pi_0(q|p_1,\ldots,p_s)$ be the pdf of the Dirichlet distribution $Dir(\nu_{01},\ldots,\nu_{0n})$:*

$$\pi_0(q|p_1,\ldots,p_s)=\frac{1}{B(\nu_{01},\ldots,\nu_{0n})}\prod_{i=1}^{n}q_i^{\nu_{0i}-1},\quad \nu_{0i}>0. \tag{3.1}$$

Then, the resulting posterior distribution $\pi(q|p_1,\ldots,p_s)$ minimizing the KL-divergence of $\pi(q|p_1,\ldots,p_s)$ from $\pi_0(q|p_1,\ldots,p_s)$ with respect to the constraints (2.7) *is the pdf of the Dirichlet distribution with parameters*

$$\hat{\nu}_i=\nu_{0i}+\sum_{j=1}^{s-1}\lambda_j\left(p_{ji}-p_{si}\right), \tag{3.2}$$

and the estimator $\hat{q}$ of the unknown q has the form

$$\hat{q}_i = \mathrm{E}_{\pi(q|p_1,\dots,p_s)}(q_i|p_1,\dots,p_s) = \frac{\nu_{0i}}{\sum_{k=1}^{n}\nu_{0k}} + \frac{\sum_{j=1}^{s-1}\lambda_j\,(p_{ji}-p_{si})}{\sum_{k=1}^{n}\nu_{0k}}, \tag{3.3}$$

where $i = 1,\dots,n$ and $\lambda_1,\dots,\lambda_{s-1}$ are the Lagrange multipliers determined so that $(s-1)$ constraints (2.7) are met.

Proof. This proposition is a special case of Proposition 2.4. By inserting (3.1) into (2.11) we immediately obtain the pdf of the Dirichlet distribution with parameters equal to (3.2). The explicit formula of $\hat{\pi}(q|p_1,\dots,p_s)$ is

$$\begin{aligned}\hat{\pi}(q|p_1,\dots,p_s) =& \frac{1}{B\left(\nu_{01}+\sum_{j=1}^{s-1}\lambda_j(p_{j1}-p_{s1}),\dots,\nu_{0n}+\sum_{j=1}^{s-1}\lambda_j(p_{jn}-p_{sn})\right)} \\ &\times\prod_{i=1}^{n} q_i^{\nu_{0i}+\sum_{j=1}^{s-1}\lambda_j(p_{ji}-p_{si})-1}.\end{aligned} \tag{3.4}$$

The estimator $\hat{q}$, the conditional expectation (2.3) with respect to the derived conditional pdf, has the form (see Section 1.3)

$$\hat{q}_i = \mathrm{E}_{\hat{\pi}(q|p_1,\dots,p_s)}(q_i|p_1,\dots,p_s) = \frac{\hat{\nu}_i}{\sum_{k=1}^{n}\hat{\nu}_k}, \quad i = 1,\dots,n,$$

where vector $(\hat{\nu}_1,\dots,\hat{\nu}_n)$ denotes the parameters of $\hat{\pi}(q|p_1,\dots,p_s)$. The following property

$$\sum_{i=1}^{n}\hat{\nu}_i = \sum_{i=1}^{n}\left(\nu_{0i}+\sum_{j=1}^{s-1}\lambda_j\,(p_{ji}-p_{si})\right) = \sum_{i=1}^{n}\nu_{0i}, \tag{3.5}$$

then yields the final formula (3.3) for $\hat{q}$.

According to Proposition 2.3 there exists a unique solution to the considered minimization within the set of $\mathcal{S}^*$. The positivity of $\hat{\nu}_i$, $i = 1,\dots,n$, follows from the requirement that $\hat{\pi}(q|p_1,\dots,p_s)$ has to be a pdf. □

Remark. *In this case the pdf $\pi(q|p_1,\dots,p_s)$ is absolutely continuous with respect to $\pi_0(q|p_1,\dots,p_s)$ as $(n-1)$-dimensional simplex Q_{sim} is the support for both pdfs.*

It is somewhat surprising that the equation (3.2) combines simultaneously both, the parameters of the Dirichlet distribution and pmfs $p_1,\dots,p_s$. When sources provide only pmfs, then these can be viewed as individual guess for $\nu_1,\dots,\nu_n$ when $\sum_{i=1}^{n}\nu_i = \sum_{i=1}^{n}\nu_{0i} = 1$. By plugging it into (3.3) we obtain

$$\hat{q}_i = p_{0i} + \sum_{j=1}^{s-1}\lambda_j p_{ji} + \left(-\sum_{j=1}^{s-1}\lambda_j\right)p_{si}, \tag{3.6}$$

where prior pmf $(p_{01},\dots,p_{0n})$, generally $p_{0i} = \frac{\nu_{0i}}{\sum_{i=1}^{n}\nu_{0i}}$, coincides with $(\nu_{01},\dots,\nu_{0n})$, a part of $\hat{q}$ induced by prior pdf $\pi_0(q|p_1,\dots,p_s)$.

Remind, that we focus on combining sources' (experts') opinions, where the prior information about the studied problem may not be available. For the prior guess on $(p_{01}, \ldots, p_{0n})$ one should then exploit provided pmfs $p_1, \ldots, p_s$, see beginning of Section 2.2.2. Based on the additive nature of the derived optimal estimator $\hat{q}$ and the considered relation between $(\nu_{01}, \ldots, \nu_{0n})$ and $(p_{01}, \ldots, p_{0n})$ in (3.6), we focus on the weighted linear combination of $p_1, \ldots, p_s$

$$\nu_{0i} = \sum_{j=1}^{s} w_{0j} p_{ji}, \tag{3.7}$$

with weights $w_{01}, \ldots, w_{0s}$ expressing preferences among considered sources. Preferences can be assigned by delegated person or depend on other available information, e.g., sources' prior information about parameters of the Dirichlet distribution. The constraints (equality of the expected KL-divergences) should then be modified accordingly.

Throughout the thesis we consider no preferences among sources, unless explicitly stated. We thus set w_{0j} equal and, following the discussion above, we obtain the arithmetic mean of $p_1, \ldots, p_s$ as the prior guess on parameters of the Dirichlet distribution

$$\nu_{0i} = \frac{\sum_{j=1}^{s} p_{ji}}{s} \quad i = 1, \ldots, n. \tag{3.8}$$

Based on the choice of prior pdf $\pi_0(q|p_1, \ldots, p_s)$ we were able to determine that the conditional pdf $\hat{\pi}(q|p_1, \ldots, p_s)$ is pdf of the Dirichlet distribution with parameters $\hat{\nu}_1, \ldots, \hat{\nu}_n$. In order to fully determine $\hat{\pi}(q|p_1, \ldots, p_s)$ and the optimal combination (3.3), it is sufficient to search for $\hat{\nu}_1, \ldots, \hat{\nu}_n$ meeting (2.7) and minimizing (2.10) instead of searching through the set of all possible conditional pdfs. This is practically applicable whenever the number n of possible realizations of discrete random vector X (and thus the number of parameters $\nu_{01}, \ldots, \nu_{0n}$ and $\hat{\nu}_1, \ldots, \hat{\nu}_n$) is smaller or comparable with the number s of information sources. The technicalities of this search are given in the next section.

3.2 Search for parameters and coefficients yielding proposed combination

In this section we focus on properties of the derived combination in terms of $\hat{\nu}_1, \ldots, \hat{\nu}_n$ and $\lambda_1, \ldots, \lambda_{s-1}$ when pmfs $p_1, \ldots, p_s$ are given and $\sum_{i=1}^{n} \nu_{0i} = 1$.

3.2.1 Numerical search for parameters of the Dirichlet distribution

Minimization of the KL-divergence in (2.10), when constraints in (2.7) are met, forms a task of non-linear optimization with constraints represented by the system of $(s-1)$ equalities with n unknown parameters $\nu_1, \ldots, \nu_n$. According to

Proposition 3.1, $\pi(q|p_1,\ldots,p_s)$ is now pdf of the Dirichlet distribution. The constraints

$$\mathrm{E}_{\pi(q|\nu_1,\ldots,\nu_n)}[\mathrm{KLD}(p_j||q)|p_1,\ldots,p_s] = \mathrm{E}_{\pi(q|\nu_1,\ldots,\nu_n)}[\mathrm{KLD}(p_s||q)|p_1,\ldots,p_s],$$

$j = 1,\ldots,s-1$, are in the considered case represented by the following system of equations (see (2.8) and (2.9))

$$-H(p_j)+H(p_s) = \sum_{i=1}^{n}(p_{ji}-p_{si})\underbrace{\left(\psi(\nu_i)-\psi\left(\sum_{i=1}^{n}\nu_i\right)\right)}_{\mathrm{E}_{\pi(q|p_1,\ldots,p_s)}(\ln q_i)} \quad j=1,\ldots,s-1, \tag{3.9}$$

where $\nu_1,\ldots,\nu_n$ are unknown parameters of pdf $\pi(q|p_1,\ldots,p_s)$ and $\psi(.)$ is the digamma (psi) function, see Section 1.3.

To solve (2.10) with $Dir(\nu_{01},\ldots,\nu_{0n})$ as prior pdf $\pi_0(q|p_1,\ldots,p_s)$ under (3.9) we exploit the penalty function approximation, thoroughly treated, e.g., by (Boyd and Vandenberghe, 2004). It allows us to reformulate the hardly tractable constrained minimization task as a series of easier unconstrained minimization tasks. Generally, a penalty function, which reaches its minimum when constraints are met, is multiplied by the penalty parameter b, say,

$$b = 100,\ 1000,\ 10000,\ 100000,\ 1000000,$$

and added to the original minimized function.

We exploit the squared penalty function, which is in harmony with least-squares handling of potentially inconsistent constraints (3.9). We then deal with a following unconstrained minimization task for each b:

$$\begin{aligned}\arg\min_{\nu_1,\ldots,\nu_n} &\mathrm{KLD}\left(\pi(q|\nu_1,\ldots,\nu_n)\middle|\middle|\pi_0(q|\nu_{01},\ldots,\nu_{0n})\right)\\ &+ b\times\sum_{j=1}^{s-1}\left[-H(p_j)+H(p_s)-\sum_{i=1}^{n}(p_{ji}-p_{si})\left(\psi(\nu_i)-\psi\left(\sum_{k=1}^{n}\nu_k\right)\right)\right]^2\\ \text{w.r.t. } &\nu_i>0 \quad i=1,\ldots,n, \quad \sum_{i=1}^{n}\nu_i=\sum_{i=1}^{n}\nu_{0i}=1,\end{aligned} \tag{3.10}$$

where $\pi(q|p_1,\ldots,p_s)$ and $\pi_0(q|p_1,\ldots,p_s)$ are pdfs of the Dirichlet distribution with parameters $\nu_i = \nu_i(p_{1i},\ldots,p_{si})$ and $\nu_{0i} = \nu_{0i}(p_{1i},\ldots,p_{si})$ respectively, and with $(\nu_{01},\ldots,\nu_{0n})$ based on the arithmetic mean of $p_1,\ldots,p_s$.

We have shown theoretically in Proposition 2.3 that pdf $\hat{\pi}(q|p_1,\ldots,p_s)$ (in the present case represented by n-tuple $\hat{\nu}_1,\ldots,\hat{\nu}_n$) solving (2.10) and meeting (2.7) exists and is unique. Since we now rely on numerical approach, in the following examples we demonstrate the behavior of new minimized function (3.10) for $n \geq s-1$ and $n < s-1$.

In the following examples Matlab software was used to obtain the figures capturing the behavior of the minimized function. To obtain $\hat{\nu}_1,\ldots,\hat{\nu}_n$ (and thus $\hat{q}$) we exploited the standard Matlab function *fmincon* (for details see `http://www.mathworks.com/help/optim/ug/fmincon.html`).

Remark. *When no data and no prior information is available, the maximum entropy principle (being special case of the minimum cross-entropy principle) states that we should exploit the pmf with the largest entropy - pmf of the uniform distribution. When data is available, the minimizing pmf (optimal combination) will shift from the pmf of the uniform distribution towards the data, still having the largest possible entropy. Since the combination* (3.3) *is generally not a convex combination, the optimal combination* $\hat{q}$ *can lie outside of the range of pmfs* $p_1, \ldots, p_s$ *– with higher entropy than any of them.*

Example 3.2. *This low-dimensional example illustrates how the penalty based optimization behaves in the simplest case. Let sources provide the following pmfs:*

$$p_1 = [0.9, 0.1]$$
$$p_2 = [0.55, 0.45].$$

Assume that the prior pmf p_0 *is* $(0.725, 0.275)$*, the arithmetic mean of* p_1 *and* p_2*. Fig. 3.1 shows the behavior of the minimized function* (3.10) *and logarithm of these values for different values of penalty parameter* b*. The example indicates that the optimized function can be very flat around minimum (Fig. 3.1 on the left) but minimum is clearly visible on logarithmic version of the optimized function (Fig. 3.1 on the right). The optimal combination* $\hat{q}$ *of* p_1 *and* p_2 *lies within the range of provided pmfs and is shifted towards the source with higher entropy (providing* p_2*):*

$$\hat{q} = (0.602, 0.398).$$

Let us now have sources with the exact same entropy but opposite opinions

$$p_1 = [0.6, 0.4]$$
$$p_2 = [0.4, 0.6].$$

Prior pmf is again their arithmetic mean: $p_0 = (0.5, 0.5)$*. As one would expect, their optimal combination lies in the 'middle' of their range:*

$$\hat{q} = (0.5, 0.5).$$

Example 3.3. *This example provides an insight into multivariate cases and shows behavior of the penalty-based optimization for the lowest possible number of sources* s *and higher number of outcomes* n*. Let sources provide the following pmfs:*

$$p_1 = [0.65, 0.2, 0.15]$$
$$p_2 = [0.55, 0.15, 0.3].$$

Consider the prior pmf as the arithmetic mean of p_1 *and* p_2*:* $p_0 = (0.6, 0.175, 0.225)$*. The logarithm of minimized function* (3.10)*, for better visibility of the minimizing element, is given in Fig. 3.2 (top). It is complemented by a zoomed view on a smaller set of* (ν_1, ν_2, ν_3)*, see Fig. 3.2 (bottom). Note that the constraint* $\sum_{i=1}^{3} \nu_i = 1$ *determines* ν_3*. Value of the optimal combination of* p_1*,* p_2*,* p_3 *lies within the range of provided pmfs and is shifted towards the source with higher entropy (providing* p_2*):* $\hat{q} = (0.57, 0.16, 0.27)$*.*

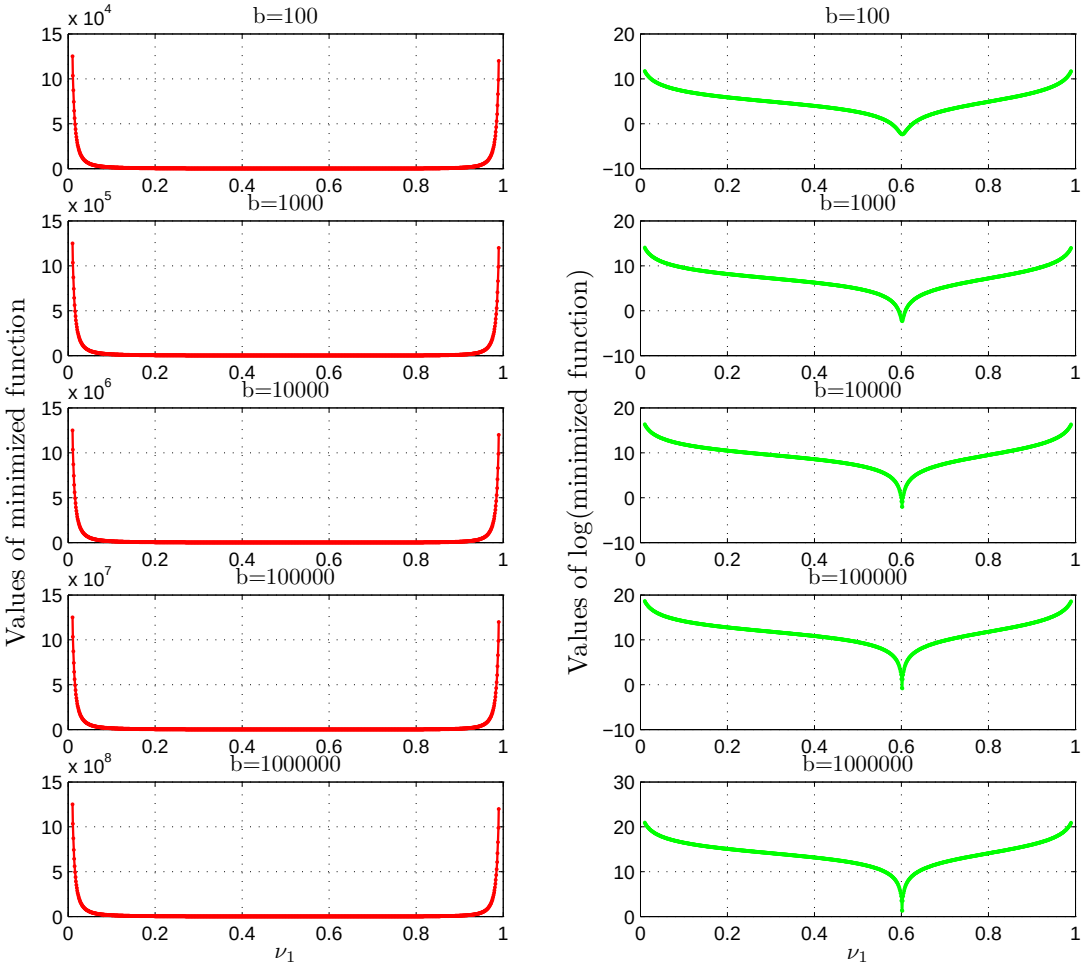


Fig. 3.1: On the left: behavior of the minimized function (3.10). On the right: behavior of the logarithm of the minimized function (3.10). The results correspond with different values of b and $n = 2$, $s = 2$. Only dependence on ν_1 is shown, as $\nu_2 = 1 - \nu_1$.

Example 3.4. *This example provides a non-trivial but still understandable multivariate case. Let sources provide the following pmfs:*

$$p_1 = [0.5, 0.15, 0.35]$$
$$p_2 = [0.55, 0.25, 0.2]$$
$$p_3 = [0.6, 0.1, 0.3]$$
$$p_4 = [0.33, 0.33, 0.33]$$

Consider the prior pmf $p_0 = (0.495, 0.2, 0.295)$ *as the arithmetic mean of* $p_1, \ldots, p_4$. *Fig. 3.3 captures the behavior of* (3.10) *for* $n = 3$, $s = 4$, *where the parameter* $\nu_4 > 0$ *is determined by the requirement* $\sum_{i=1}^{n} \nu_i = 1$. *This example indicates the desirable property: numerical behavior is similar to cases with more freedom. The minimization led to the optimal combination* $\hat{q} = (0.38, 0.29, 0.33)$.

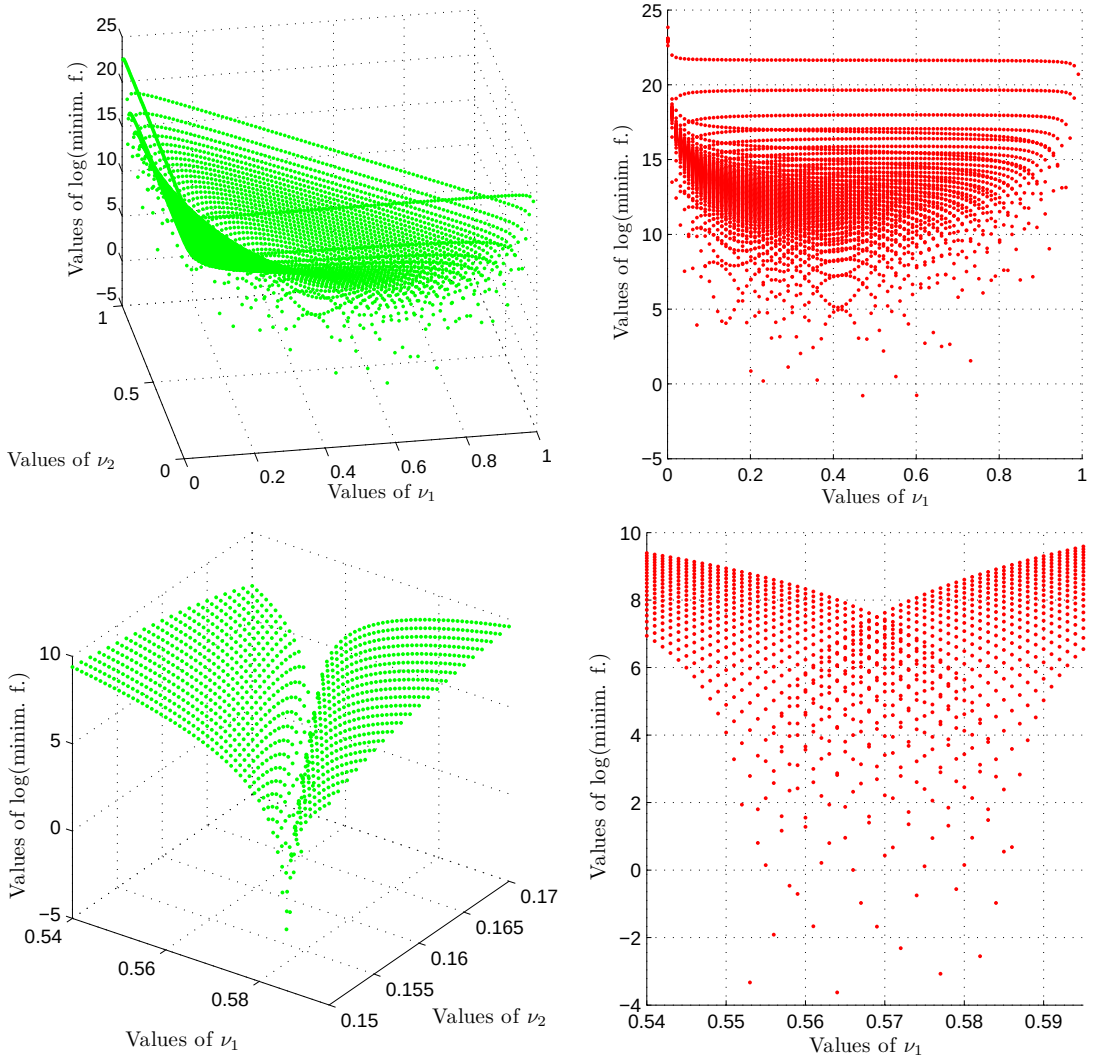


Fig. 3.2: Top: Behavior of the logarithm of the minimized function (3.10), where $b = 1000000$, $n = 3$, $s = 2$ (different angles of view). Bottom: Behavior of the same function on a smaller set of ν_1, ν_2 (different angles of view).

3.2.2 Numerical search for coefficients in proposed combination

Commonly the number of sources s is finite while the number of possible outcomes of X denoted by n can be very large (the case $s \ll n$ is the generic one). Then, the optimization over the Lagrange multipliers is the only viable option.

If $n \gg s$ or if we are interested in coefficients $\lambda_1, \ldots, \lambda_{s-1}$, we can, based on relation (3.2), minimize function (3.10) with respect to $\lambda_1, \ldots, \lambda_{s-1}$ instead of $\nu_1, \ldots, \nu_n$. In this section we focus on several technicalities that occur in such case.

To obtain $\lambda_1, \ldots, \lambda_{s-1}$ we again exploit squared penalty function. In particu-

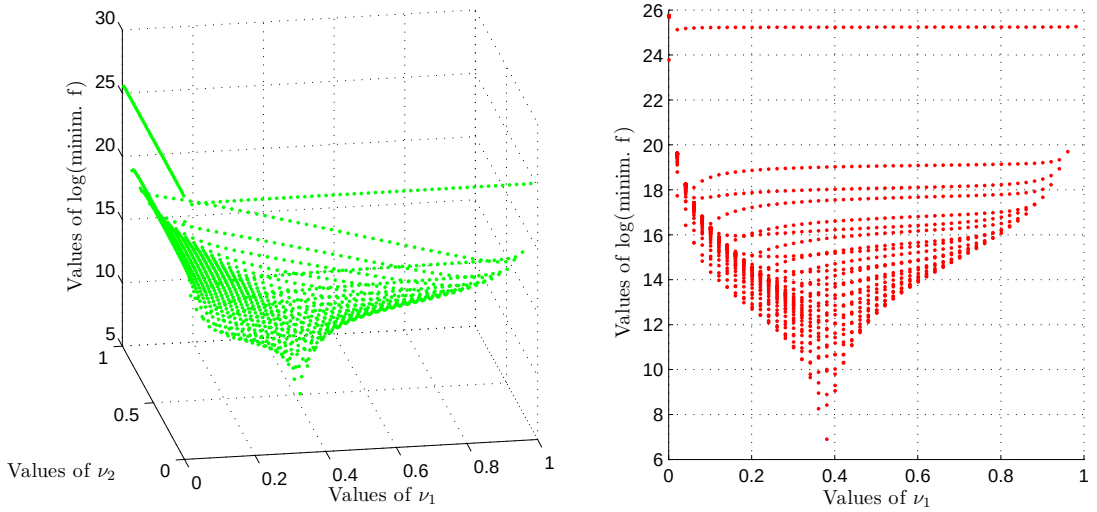


Fig. 3.3: Behavior of the logarithm of the minimized function (3.10), where $b = 1000000$, $n = 3$, $s = 4$ (different angles of view).

lar, we numerically search for $\lambda_1, \ldots, \lambda_{s-1}$ minimizing

$$
\begin{aligned}
& KLD\left(\pi(q|\nu_i(\lambda_1,\ldots,\lambda_{s-1}), i = 1,\ldots,n)\middle|\middle|\pi_0(q|p_1,\ldots,p_s)\right) \\
& + b \times \sum_{j=1}^{s-1} \Bigg[- H(p_j) + H(p_s) \\
& \quad - \sum_{i=1}^{n} (p_{ji} - p_{si}) \left(\psi(\nu_i(\lambda_1,\ldots,\lambda_{s-1})) - \psi\left(\sum_{k=1}^{n} (\nu_k(\lambda_1,\ldots,\lambda_{s-1})) \right) \right) \Bigg]^2 \\
& \text{w.r.t. } \nu_i(\lambda_1,\ldots,\lambda_{s-1}) > 0 \quad i = 1,\ldots,n, \qquad (3.11)
\end{aligned}
$$

where $\pi(q|\nu_i(\lambda_1,\ldots,\lambda_{s-1}), i = 1,\ldots,n)$ and $\pi_0(q|p_1,\ldots,p_s)$ are pdfs of the Dirichlet distribution with parameters

$$\nu_i(\lambda_1,\ldots,\lambda_{s-1}) = p_{0i} + \sum_{j=1}^{s-1} \lambda_j p_{ji} + \left(- \sum_{j=1}^{s-1} \lambda_j \right) p_{si},$$

and $\nu_{0i} = p_{0i}$, respectively. The parameters $\nu_{01}, \ldots, \nu_{0n}$ are based on arithmetic mean of $p_1, \ldots, p_s$.

Consequence of relation between $\hat{\nu}_1, \ldots, \hat{\nu}_n$ and $\lambda_1, \ldots, \lambda_{s-1}$

Based on results from Proposition 3.1 we now study how the relation (3.2), plugged into (3.10) and yielding (3.11), influences the search for $\lambda_1, \ldots, \lambda_{s-1}$. The uniqueness of $\hat{\nu}_1, \ldots, \hat{\nu}_n$, the assumption $\sum_{i=1}^{n} \hat{\nu}_i = 1$ and relation (3.2) yield two cases:

1. First, if $n > s - 1$, the exact solution may not exist. However, we can still obtain the approximation of the solution by using, e.g., least squares. This idea supports our choice of the quadratic penalty function in Section 3.2.2.

2. Second, when $n \leq s-1$, we deal with the underdetermined system yielding infinitely many $(s-1)$-tuples $\lambda_1, \ldots, \lambda_{s-1}$ minimizing (3.11).

Thus, if we perform the optimization directly with respect to $\lambda_1, \ldots, \lambda_{s-1}$, their values might be ambiguous while the value of the optimal combination will stay unchanged.

The following examples show the behavior of the function (3.11) with respect to $\lambda_1, \ldots, \lambda_{s-1}$.

Example 3.5. *Let sources provide the pmfs given in Example 3.2:*

$$p_1 = [0.9, 0.1]$$
$$p_2 = [0.55, 0.45].$$

As the examples in Section 3.2.1 suggested, the minimized function (3.11) *is flat on the set of* $\nu_1, \ldots, \nu_n$ *satisfying* $\sum_{i=1}^{n} \nu_i = 1$ *and* $\nu_i > 0$, $i = 1, \ldots, n$, *see Fig. 3.1 on the left. We obtain the same results when studying the dependency of this function on coefficient* λ_1, *see Fig. 3.4 on the left. Using logarithm of the minimized function, the minimum is clearly visible, see Fig. 3.4 on the right. The obtained optimal combination* $\hat{q} = (0.602, 0.398)$ *coincides with result in Example 3.2.*

Example 3.6. *In this example we demonstrate the above discussed consequence of relation* (3.2) *between* $\lambda_1, \ldots, \lambda_{s-1}$ *and* $\nu_1, \ldots, \nu_n$ *in multivariate case.*

First, let three sources $(s = 3)$ *provide the following pmfs* $(n = 3)$*:*

$$p_1 = [0.75, 0.15, 0.1]$$
$$p_2 = [0.6, 0.25, 0.15]$$
$$p_3 = [0.1, 0.2, 0.7].$$

Fig. 3.5 (top) captures the logarithm of the minimized function (3.11) *for prior guess* p_0 *chosen as arithmetic mean of* p_1, p_2 *and* p_3*:* $(0.48, 0.2, 0.32)$. *The unique minimizer* $(\lambda_1, \lambda_2) = (-0.29, 0.18)$, *yielding optimal combination of* p_1, p_2, p_3*:* $\hat{q} = (0.38, 0.22, 0.39)$, *is visible (see discussion for case* $n > s-1$*).*

Second, let three sources $(s = 3)$ *provide the following pmfs* $(n = 2)$*:*

$$p_1 = [0.85, 0.15]$$
$$p_2 = [0.7, 0.3]$$
$$p_3 = [0.4, 0.6]$$

Fig. 3.5 (bottom) captures the logarithm of the minimized function (3.11) *for prior guess* p_0 *again chosen as the arithmetic mean of* p_1, p_2 *and* p_3*:* $(0.65, 0.35)$. *We clearly see the ambiguity in pairs* (λ_1, λ_2) *(see discussion for case* $n \leq s-1$*). However, the optimal combination of* p_1, p_2, p_3 *is unique:* $\hat{q} = (0.55, 0.45)$.

Remark. *The optimal estimator* $\hat{q}$ *of the desired combination* q *proposed in* (3.3) *was derived under the assumption that all elements of the prior guess* $\nu_{01}, \ldots, \nu_{0n}$ *are positive. We also assumed that all considered sources have the same support*

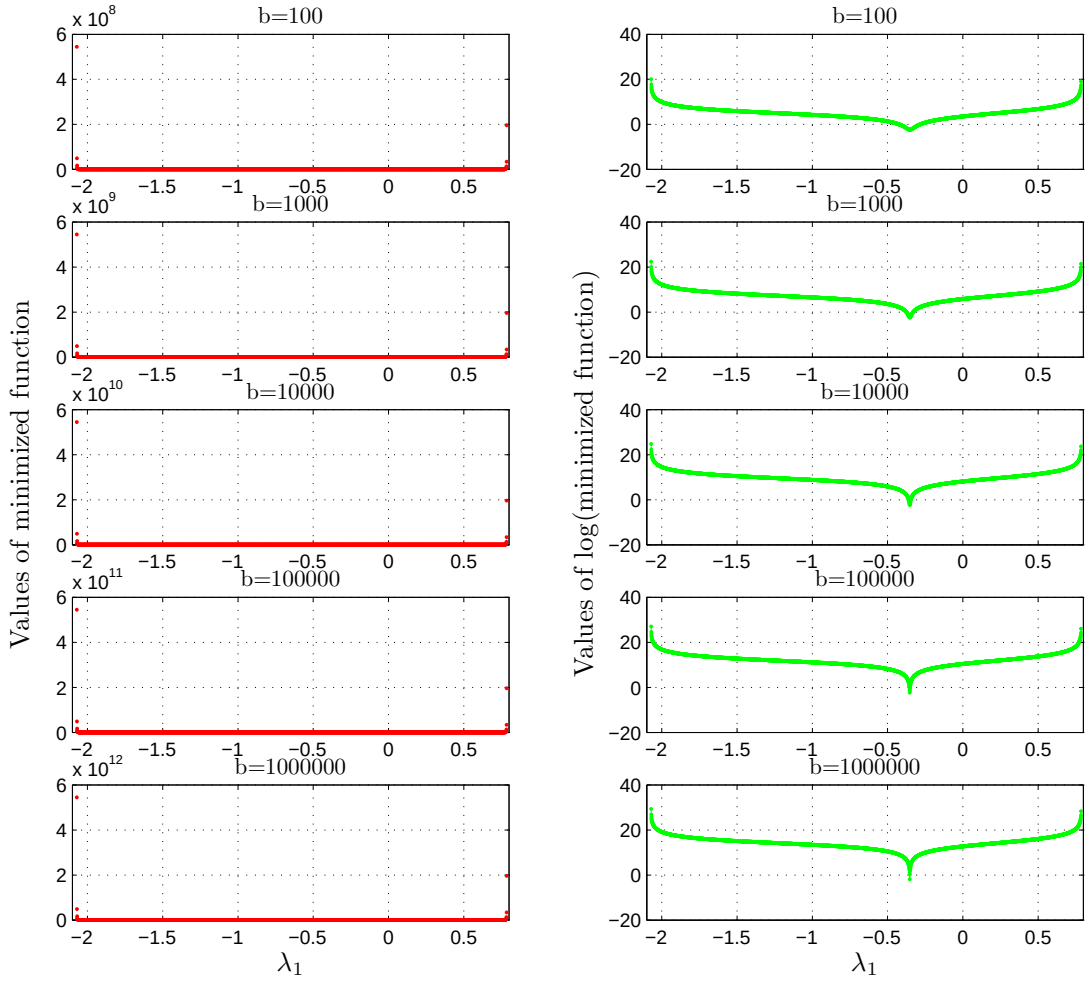


Fig. 3.4: On the left: Behavior of the minimized function (3.11). On the right: Behavior of the logarithm of the minimized function (3.11). Different values of b, $n = 2$, $s = 2$.

and thus provided p_{ji} are positive too, $i = 1, \ldots, n$, $j = 1, \ldots, s$. Then, the arithmetic mean of provided pmfs $p_1, \ldots, p_s$ serves as a correct prior guess of $\nu_{01}, \ldots, \nu_{0n}$. In the case there is index i such that $p_{ji} = 0$ for all j, we suggest to set ν_{0i} as a small positive value (e.g., $\leq 10^{-5}$).

Remark. *Throughout this thesis we assume the dimension n of $q = (q_1, \ldots, q_n)$ is finite. The extension to countably many elements of q involves the changes in exploited Dirichlet distribution $Dir(\nu_{01}, \ldots, \nu_{0n})$. Let us consider the following reparametrization: $dp_{0i} = \nu_{0i}$, where p_{0i} can be viewed as a prior guess on q_i, $i = 1, \ldots, n$, and $d < \infty$ as a concentration parameter around p_{0i}: $\sum_{i=1}^{n} \nu_{0i} = d < \infty$. For all choices of this type, limit of the combination (3.3) for $n \to \infty$ exists and provides a continuous extension of the finite case. It corresponds to the extension when we consider a Dirichlet process $(P(A_1), \ldots, P(A_l)) \sim Dir(dP_0(A_1), \ldots, dP_0(A_l))$, see (Ferguson, 1973). Here, P_0 is a probability measure on the measurable space $(\Lambda, \sigma(\Lambda))$ and $\{A_1, \ldots, A_l\}$ are pairwise disjoint partitions, $\cup_{k=1}^{l} A_k = \Lambda$. This will allow to relate the proposed combination (3.3)*

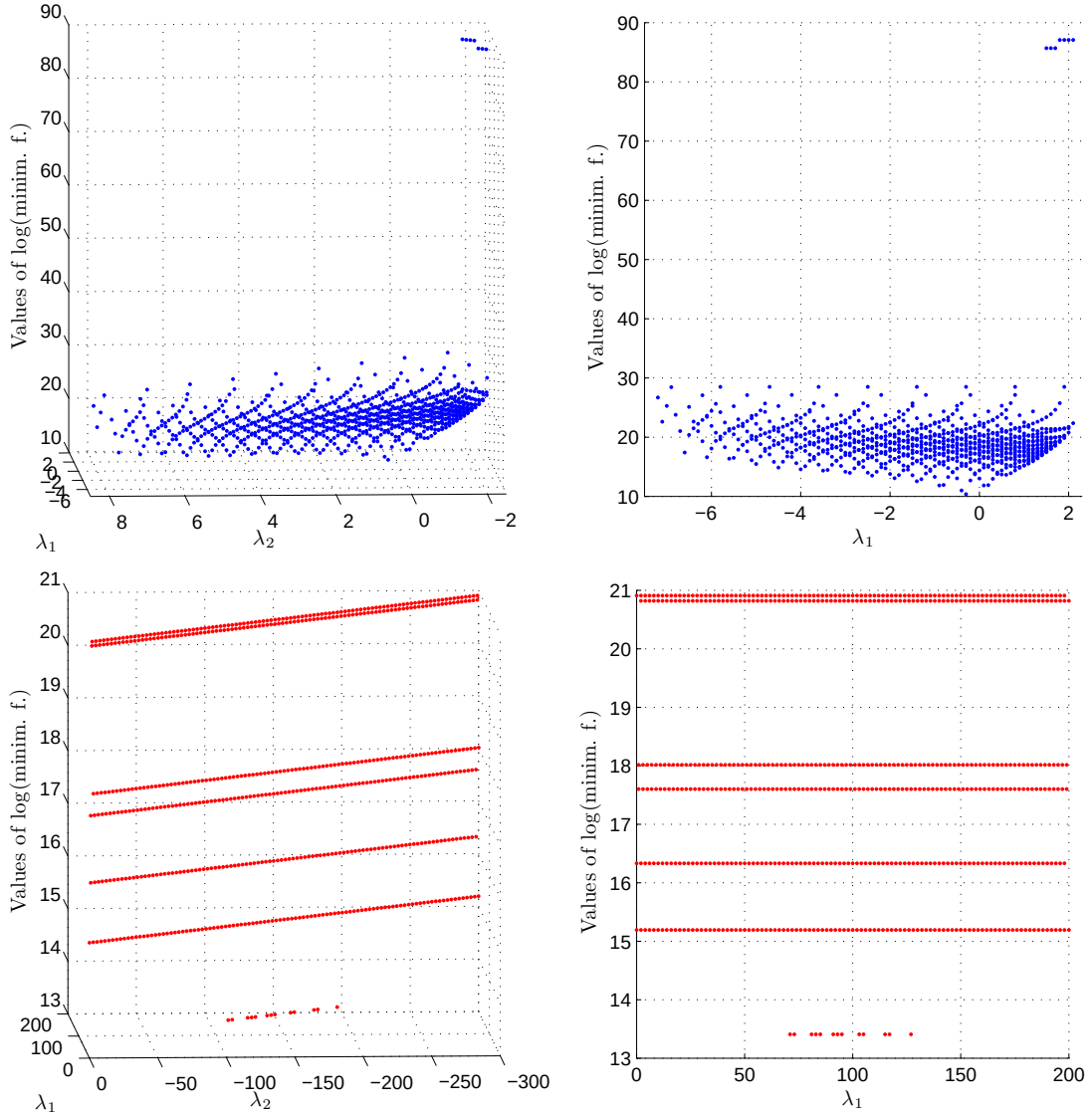


Fig. 3.5: Top: Behavior of the logarithm of the minimized function (3.11) with respect to λ_1 and λ_2, $b = 10000000$, $n = 3$ and $s = 3$ (different angles of view). The minimum is clearly visible. Bottom: Behavior of the logarithm of the minimized function (3.11) with respect to λ_1 and λ_2, $b = 10000000$, $n = 2$ and $s = 3$ (different angles of view). Many minimizing pairs of λ_1 and λ_2, yielding a unique value of the optimal combination of provided pmfs, exist.

and pdfs describing finite-dimensional real space. Such approach makes our methodology widely applicable and is of further interest in future work.

3.3 Additional properties of proposed combination

In this section we discuss three interesting features of the derived optimal estimator $\hat{q}$:

1. Since the formula (3.3) includes differences between $p_1, \ldots, p_{s-1}$ and p_s we first inspect how the value of the final combination changes with duplicate observations (Section 3.7).

2. Second, we focus on the difference between optimal combination obtained by data processed at once (static case) and processed sequentially (referred to as the dynamic case).

3. Recall the discussion after Proposition 3.1 regarding the occurrence of parameters $\nu_{01}, \ldots, \nu_{0n}$ and pmfs $p_1, \ldots, p_s$ in formula (3.3) of the estimator $\hat{q}$. When no other information about sources is available, we assumed that $\sum_{i=1}^{n} \nu_{0i} = 1$ and the arithmetic mean was chosen as prior guess on $(\nu_{01}, \ldots, \nu_{0n})$. We study how the optimal combination $\hat{q}$ changes when pmfs $p_1, \ldots, p_s$ also the preferences about sources (allowing $\sum_{i=1}^{n} \nu_{0i} \neq 1$) are given.

3.3.1 Duplicate observations

In this section we study the change in optimal combination (3.3) when duplicate pmfs occur.

Proposition 3.7. *Let $\lambda_1, \ldots, \lambda_{s-1}, \lambda_s$ be the coefficients in the combination $\hat{q}$ of $p_1, \ldots, p_s, p_{s+1}$. Then, for a fixed prior pmf p_0, the combination of $p_1, \ldots, p_s, p_{s+1}$, where $p_{s+1,i} = p_{k,i}$ for some $k \in \{1, \ldots, s\}$, $i = 1, \ldots, n$, coincides with $\hat{q}$ evaluated with omission of p_{s+1} and unchanged p_0.*

Proof. Without loss of generality assume that pmf of $(s+1)^{st}$ source coincides with pmf od s^{th} source. The dissimilarity constraints look as follows:

$$\mathrm{E}_{\pi(q|p_1,\ldots,p_s,p_{s+1})}\left[\mathrm{KLD}(p_j|q)|p_1, \ldots, p_s\right] = \mathrm{E}_{\pi(q|p_1,\ldots,p_s,p_{s+1})}\left[\mathrm{KLD}(p_{s+1}|q)|p_1, \ldots, p_s\right]$$

$j = 1, \ldots, s$. The optimal combination $\hat{q}$ based on $p_1, \ldots, p_s, p_{s+1}$ for $\sum_{i=1}^{n} \nu_{0i} = 1$ has the following form:

$$\begin{aligned}\hat{q}_i &= p_{0i} + \sum_{j=1}^{s} \lambda_j (p_{ji} - p_{(s+1)i}) \\ &= p_{0i} + \sum_{j=1}^{s-1} \lambda_j (p_{ji} - p_{si}) + \lambda_s (p_{si} - p_{si}).\end{aligned}$$

The right-hand side of the last equation includes only $(p_{ji} - p_{si})$, $j = 1, \ldots, s-1$, with coefficients $\lambda_1, \ldots, \lambda_{s-1}$. Thus, for fixed prior pmf p_0, the optimization (3.11) based on $s+1$ pmfs where $p_{si} = p_{s+1,i}$, $i = 1, \ldots, n$, coincides with optimization based on s pmfs, i.e., without the repeated pmf.

□

Remark. *Let us focus on k^{th} source, $k \neq s$. The additivity property of combination* (3.3) *implies that if other s_1 sources gave the same pmf p_k, then the coefficient of each source equals $\frac{\lambda_k}{s_1}$.*

Remark. *It may seem strange that repeated sources' opinion are not taken more "seriously", with a higher weight. This is consequence of the fact that individual sources are not qualified by a weight reflecting their reliability. When such a weighting will be introduced, the coincidence of opinions can be taken into account and distinguished from cheating by repetitions of the same opinion.*

The current solution without weights is of a conservative type. It qualifies all repetitions as "cheating" and prevents overweighing of such source.

The choice of p_0 as arithmetic mean allows plausible deviation from this conservative treatment as it takes into account all, even repetitive, sources.

Example 3.8. *To illustrate the effect of duplication let us have* 2 *sources giving the following pmfs:*

$$p_1 = [0.75, 0.15, 0.1]$$
$$p_2 = [0.7, 0.1, 0.2].$$

Prior guess is p_0 is the arithmetic mean of p_1 and p_2, $\sum_{i=1}^{n} \nu_{0i} = 1$. The optimal combination lies within the range of given pmfs - $\hat{q} = (0.71, 0.11, 0.18)$. Consider now 3 *sources as follows:*

$$p_1 = [0.75, 0.15, 0.1]$$
$$p_2 = [0.75, 0.15, 0.1]$$
$$p_3 = [0.7, 0.1, 0.2]$$

Despite the first and the second source provided the same pmf the resulting pmf is again $(0.71, 0.11, 0.18)$.

3.3.2 Dynamic case

In the previous sections we assumed that the given data were combined in a batch. In this section we would like to extend the use of the proposed combining to the case when pmfs are processed sequentially, i.e., at each time instant $t = 1, \ldots, T$, $T < \infty$ a set of s_t new pmfs is available.

According to the combining proposed in Proposition 3.1, we should with every new set of data join the old data and new data together, which might be computationally inefficient. Thus, we suggest to connect it to the value of the optimal estimator $\hat{q}$ from the previous time step in the following way:

a) combine pmfs $p_{t,1}, \ldots, p_{t,s_t}$ available at time step t using (3.3) to obtain $\hat{q}_t^*$,

b) then combine $\hat{q}_t^*$ and $\hat{q}_{t-1}$ from the previous step $t-1$ using again (3.3) to obtain final $\hat{q}_t$.

The value of the optimal combination when data are processed according to the proposed idea is (consider $T = 2$):

At time step t = 1:

$$\hat{q}_{1,i} = \underbrace{\frac{1}{s_1}\sum_{j=1}^{s_1} p_{1,ji}}_{p_{1,0i}} + \sum_{j=1}^{s_1-1} \lambda_{1,j}(p_{1,ji} - p_{1,si}). \tag{3.12}$$

At time step t = 2:

$$\begin{aligned}\hat{q}^*_{2,i} =& \frac{1}{s_2}\sum_{j=1}^{s_2} p_{2,ji} + \sum_{j=1}^{s_2-1} \lambda^*_{2,j}(p_{2,ji} - p_{2,si}),\\ \hat{q}_{2,i} =& \underbrace{\frac{1}{2}(\hat{q}^*_{2,i} + \hat{q}_{t-1,i})}_{p_{t,0i}} + \lambda_{2,1}(\hat{q}^*_{2,i} - \hat{q}_{t-1,i}),\end{aligned} \tag{3.13}$$

where $\hat{q}_1$ is the combination of $p_1, \ldots, p_{s_1}$ based on (3.3) with coefficients $\lambda_{1,j}$, $j = 1, \ldots, s_1$ (we assume $\sum_{i=1}^{n} \nu_{t,0i} = 1$ for all t).

The combination $\hat{q}^*_2$ is based on $p_1, \ldots, p_{s_2}$ with coefficients $\lambda^*_{2,j}$, $j = 1, \ldots, s_2$, $s_2 = 2$. Finally, $\hat{q}_{2,i}$ is the combination of $\hat{q}^*_2$ and $\hat{q}_1$ again based on (3.3) with coefficient $\lambda_{2,1}$ (we combine only two pmfs).

If we process all given data at once, the resulting value of the combination is:

$$\begin{aligned}\hat{q}_{all,i} =& \underbrace{\frac{1}{s}\sum_{j=1}^{s} p_{ji}}_{p_{0i}} + \sum_{j=1}^{s-1} \lambda_j(p_{ji} - p_{si})\\ =& \sum_{j=1}^{s-1}\left(\frac{1}{s} + \lambda_j\right)p_{ji} + \left(\frac{1}{s} - \sum_{j=1}^{s-1}\lambda_j\right)p_{si}\end{aligned} \tag{3.14}$$

where $s = s_1 + \ldots + s_T$.

By inserting the combination $\hat{q}_{t-1}$ into $\hat{q}_t$ (3.13) we generally obtain different value of the combination to the value obtained by combining s sources directly (3.14). The comparison of both ways is given in the following example.

Example 3.9. *Let the sources provide the following pmfs:*

time step $t = 1$

$$p_1 = [0.7, 0.3, 0]$$
$$p_2 = [0.3, 0.1, 0.6]$$

time step $t = 2$

$$p_3 = [0.35, 0.55, 0.1]$$
$$p_4 = [0.4, 0.15, 0.45].$$

By exploiting the combination (3.3) *and ideas in* (3.13) *and* (3.14) *we obtained*

the following results at time step $t = 2$:

$$\begin{aligned} \text{arithmetic mean of all pmfs} &: (0.4375, 0.2750, 0.2875) \\ \text{combining by proposed idea (3.13)} &: (0.3796, 0.3231, 0.2974) \\ \text{combining all pmfs at once (3.14)} &: (0.3862, 0.3207, 0.2931) \end{aligned}$$

In this simple case, we obtained similar results for both approaches, which is encouraging for further investigation of proposed dynamic combining (3.13).

The question how well the sequential incorporation of data influences the final value of the optimal combination $\hat{q}$ is of interest for further research. Similar problem regarding the approximations in the field of recursive Bayesian estimation has been discussed recently, see (Kárný, 2014).

3.3.3 Prior probabilities influenced by preferences

The relation between the parameters $\hat{\nu}_1, \ldots, \hat{\nu}_n$ and provided pmfs $p_1, \ldots, p_s$ given in (3.2) motivated the choice $\sum_{i=1}^{n} \nu_{0i} = 1$. With no preferences among sources the arithmetic mean of $p_1, \ldots, p_s$ was chosen to serve as the prior guess $(\nu_{01}, \ldots, \nu_{0n})$.

In this section, we assume that sources can be assigned prior coefficients $w_{0j} > 0$ reflecting their importance as suggested in (3.7),

$$\nu_{0i} = \sum_{j=1}^{s} w_{0j} p_{ji},$$

and yielding

$$\sum_{i=1}^{n} \nu_{0i} = \sum_{i=1}^{n} \sum_{j=1}^{s} w_{0j} p_{ji} = \sum_{j=1}^{s} w_{0j}.$$

Such coefficients can reflect, e.g., the number of observations on which p_j is based.

Besides the prior values $\nu_{01}, \ldots, \nu_{0n}$, it is also desirable that the constraints (2.7) will be affected by these weights. In particular, we approach the combination to more important sources by requiring

$$w_j \mathrm{E}\left[\mathrm{KLD}(p_j||q)|p_1, \ldots, p_s\right] = w_s \mathrm{E}\left[\mathrm{KLD}(p_s||q)|p_1, \ldots, p_s\right],$$

$j = 1, \ldots, s-1$, yielding the following weighted counterpart of constraints (3.9):

$$-w_j H(p_j) + w_s H(p_s) = \sum_{i=1}^{n} (p_{ji} w_j - p_{si} w_s) \left[\psi(\nu_i) - \psi\left(\sum_{k=1}^{n} \nu_k\right)\right].$$

Then, the optimal estimator $\hat{q}$ of desired combination q is

$$\hat{q}_i = p_{0i} + \sum_{j=1}^{s-1} \frac{\lambda_j}{\sum_{l=1}^{s} w_l} (w_j p_{ji} - w_s p_{si}), \quad i = 1, \ldots, n,$$

where

$$p_{0i} = \frac{\nu_{0i}}{\sum_{k=1}^{n} \nu_{0k}} = \sum_{j=1}^{s} \frac{w_j p_{ji}}{\sum_{l=1}^{s} w_l}.$$

In the presented discussion, we have assumed that the elements of pmf provided by source j, $j = 1, \ldots, s$, have the same weights $w_{ji} = const$, $i = 1, \ldots, n$. If needed, also an element-dependent version w_{ji}, $i = 1, \ldots, n$, can be simply exploited similarly to the method developed in (Guiaşu, 1971).

Remark. *In this remark we focus on case when we change the sum of weights while keeping the proportion among weights the same. Consider the following two sets of weights*

$$\sum_{j=1}^{s} w_j = d \quad \text{and} \quad \sum_{j=1}^{s} w_j^* = \sum_{j=1}^{s} k w_j = d^*.$$

The prior guesses on the parameters of the Dirichlet distribution then look as follows:

$$\text{for } w_1, \ldots, w_s: \quad \nu_{0i} = \sum_{j=1}^{s} w_j p_{ji}$$

$$\text{for } w_1^*, \ldots, w_s^*: \quad \nu_{0i}^* = \sum_{j=1}^{s} w_j^* p_{ji}$$

From the formula (3.3) *for $\hat{q}$ derived for weights w^**

$$\hat{q}_i = \sum_{j=1}^{s} \frac{k w_j p_{ji}}{\sum_{l=1}^{s} k w_l} + \sum_{j=1}^{s-1} \frac{\lambda_j}{\sum_{l=1}^{s} k w_l} (k w_j p_{ji} - k w_s p_{si}),$$

where $k = \frac{d^*}{d}$, it might seem that both optimal estimators, based on w and w^*, coincide.

Recall the function (3.10)*:*

$$KLD(\pi(q|p_1, \ldots, p_s)||\pi_0(q|p_1, \ldots, p_s)) = \ln \frac{\Gamma(\sum_{i=1}^{n} \nu_i)}{\Gamma(\sum_{i=1}^{n} \nu_{0i})} + \sum_{i=1}^{n} \ln \frac{\Gamma(\nu_{0i})}{\Gamma(\nu_i)}$$
$$+ \sum_{i=1}^{n} (\nu_i - \nu_{0i}) \left(\psi(\nu_i) - \psi \left(\sum_{k=1}^{n} \nu_k \right) \right)$$

leading to $\hat{q}$ by minimization with respect to $\nu_1, \ldots, \nu_n$ or by minimization of (3.11) *with respect to $\lambda_1, \ldots, \lambda_{s-1}$. Due to non-linearity of this function in prior guess $\nu_{01}, \ldots, \nu_{0n}$ we can not expect the combinations based on $w_1, \ldots, w_s$ and $w_1^* = k w_1, \ldots, w_s^* = k w_s$ to be equal.*

The following example demonstrates the change in $\hat{q}$ with change of d, while keeping the preference ratio among sources the same.

Example 3.10. *Let us have* 2 *sources providing pmfs*

$$p_1 = [0.7, 0.2, 0.1]$$
$$p_2 = [0.6, 0.1, 0.3]$$

having equal weights w_1 *and* w_2*, respectively. Let* $d \in \{0.1, 1, 10\}$*. Then, for the prior guess*

$$p_0 = \frac{w_1 p_1 + w_2 p_2}{\sum_{j=1}^{2} w_j}$$

we obtain the following optimal estimates $\hat{q}$*:*

$$w_1 = w_2 = 0.05, \quad \sum_{j=1}^{2} w_j = 0.1 : \quad \hat{q} = (0.53, 0.19, 0.28)$$

$$w_1 = w_2 = 0.5, \quad \sum_{j=1}^{2} w_j = 1 : \quad \hat{q} = (0.64, 0.14, 0.22)$$

$$w_1 = w_2 = 5, \quad \sum_{j=1}^{2} w_j = 10 : \quad \hat{q} = (0.65, 0.15, 0.20).$$

With higher $\sum_{i=1}^{n} w_i$ *the optimal combination tends towards the arithmetic mean of provided pmfs.*

When the first source has a higher preference (four times higher than the second source: $w_1 = 4w_2$*), the resulting combination, for different sums of weights, becomes:*

$$w_1 = 0.08, w_2 = 0.02, \quad \sum_{j=1}^{2} w_j = 0.1 : \quad \hat{q} = (0.57, 0.30, 0.13)$$

$$w_1 = 0.8, w_2 = 0.2, \quad \sum_{j=1}^{2} w_j = 1 : \quad \hat{q} = (0.59, 0.30, 0.11)$$

$$w_1 = 8, w_2 = 2, \quad \sum_{j=1}^{2} w_j = 10 : \quad \hat{q} = (0.69, 0.25, 0.06).$$

The resulting combination converges towards the first source as $\sum_{i=1}^{n} w_i = \sum_{i=1}^{n} \nu_{0i}$ *grows.*

Chapter 4

Extension and transformation to joint pmf

In the previous chapters we assumed that the opinions about the outcomes of a common discrete random vector X given by sources $1, \ldots, s$ were in the probabilistic form. In particular, we assumed that all sources had the same support and provided joint pmfs $p_{ji} = P_j(X = x_i)$ over outcomes $\{x_i\}_{i=1}^n$ of X,

$$p_{ji} \geq 0, \quad \sum_{i=1}^{n} p_{ji} = 1, \quad i = 1, \ldots, n, \quad j = 1, \ldots, s, \quad n, s < \infty.$$

The optimal combination of $p_1, \ldots, p_s$, derived as optimal estimator $\hat{q}$ of unknown desired combination q, was given in Proposition 3.1.

In this chapter we extend the versatility of combining proposed in Chapter 3 by considering new types of provided information: a probabilistic (but not joint) type and a non-probabilistic type. To show that proposed optimal combination (3.3) is applicable also to other types than joint pmfs, we define extension and transformation of provided information. Their definitions form the contribution of this chapter.

To introduce the notion of probabilistic and non-probabilistic type of information, let the modeled random vector X consist of m random variables

$$X = (X_1, \ldots, X_k, X_{k+1}, \ldots, X_m), \quad m < \infty,$$

again with a finite number of outcomes x_i, $i = 1, \ldots, n < \infty$. Examples of the non-joint probabilistic data are then

- a conditional probability vector of $X_1, \ldots, X_k$ given values of $X_{k+1}, \ldots, X_m$,
- a marginal probability vector of $X_1, \ldots, X_k$.

To obtain the optimal combination $\hat{q}$ we appropriately extend these provided conditional or marginal pmfs to the joint probability vector of X so that formula (3.3) can be applied. If the information is in a non-probabilistic form such as

- a subset of the set of possible outcomes of X,

– an expected values of X is available,

we propose a transformation into probabilistic form. We can then extend the transformed information into the joint probability vector, if needed, and apply (3.3).

4.1 Probabilistic data – extension

The following section includes the suggestion how to extend given conditional pmfs to the joint pmfs and an example demonstrating the behavior of the resulting optimal estimator $\hat{q}$.

4.1.1 Conditional probability

Assume each source provided the conditional pmf

$$\begin{aligned} &p_j(x_1,\ldots,x_{k_j}|x_{k_j+1},\ldots,x_m) \\ &\quad = P_j(X_1=x_1,\ldots,X_{k_j}=x_{k_j}|X_{k_j+1}=x_{k_j+1},\ldots,X_m=x_m) \end{aligned}$$

where $j=1,\ldots,s$. The set of indices $1,\ldots,m$ splits for the j^{th} source on indices pointing to probabilistically described values u_j and values in condition v_j, which allows us to simplify the notation

$$(x_1,\ldots,x_{k_j})=u_j \quad \text{and} \quad (x_{k_j+1},\ldots,x_m)=v_j$$

and to symbolically describe the set of outcomes $\{(x_1,\ldots,x_{k_j+1})\}$ by $\{u_j\}$ and the set of outcomes $\{(x_{k_j+1},\ldots,x_m)\}$ by $\{v_j\}$.

The combination $\hat{q}$ requires joint probabilities and thus we want to construct the extension p_j^* of p_j.

Definition 4.1. *Joint pmf $p_j^*=p_j^*(u_j,v_j)$ is called a reasonable extension of conditional pmf $p_j(u_j|v_j)$ provided by j^{th} source if it satisfies two following properties:*

a) its conditional version coincides with given pmf:

$$p_j(u_j|v_j)=p_j^*(u_j|v_j),$$

b) it is the best approximation of $\hat{q}$ in terms of KL-divergence

$$p_j^*=\arg\min KLD(\hat{q}||p_j^\bullet),$$

with minimization performed over the set of all joint pmfs $p_j^\bullet$ satisfying a),

where $\hat{q}$ is the optimal combination of the processed information sources and u_j, v_j denote the splitting of $(x_1,\ldots,x_m)$ for the j^{th} source.

This is in harmony with Bernardo (1979) and similar idea was used by Kárný et al. (2009).

The exact form of the extension p_j^* is given in the next proposition.

Proposition 4.2. *The reasonable extension $p_j^*(u_j, v_j)$ of a given conditional pmf $p_j(u_j|v_j)$ introduced in Definition 4.1, has the form of chain rule with the unavailable marginal pmf $p_j(v_j)$ replaced by the corresponding marginal pmf $\hat{q}(v_j)$ of $\hat{q}(u_j, v_j)$*

$$p_j^*(u_j, v_j) = p_j(u_j|v_j)\hat{q}(v_j), \tag{4.1}$$

where u_j, v_j denote the splitting of $(x_1, \ldots, x_m)$ for the j^{th} source.

Proof. The KL-divergence (1.1)

$$\begin{aligned} KLD(\hat{q}||p_j^\bullet) &= \sum_{\{u_j\},\{v_j\}} \hat{q}(u_j, v_j) \ln \frac{\hat{q}(u_j, v_j)}{p_j^\bullet(u_j, v_j)} \\ &= -H(\hat{q}) - \sum_{\{u_j\},\{v_j\}} \hat{q}(u_j, v_j) \ln p_j^\bullet(u_j, v_j) \\ &= -H(\hat{q}) - \sum_{\{u_j\},\{v_j\}} \hat{q}(u_j, v_j) \ln p_j(u_j|v_j) - \sum_{\{u_j\},\{v_j\}} \hat{q}(u_j|v_j)\hat{q}(v_j) \ln p_j^\bullet(v_j) \end{aligned}$$

is minimized by $p_j^*(v_j) = \hat{q}(v_j)$, which yields the extension (4.1). □

The reasonable extensions p_j^*, $j = 1, \ldots, s$, defining the joint pmfs of X can be now combined as presented before by (3.3). The only change is that it will lead to implicit relations for $\hat{q}$, whose marginal pfms are used in extensions:

$$p_j^*(u_j, v_j) = p_j(u_j|v_j) \times \hat{q}(v_j).$$

Thus, if the j^{th} source provides a conditional probability with respect to the splitting $x = (u_j, v_j)$ and the s^{th} source gives conditional pmf with respect to the splitting $x = (y_s, z_s)$, the dissimilarity constraint is:

$$\begin{aligned} &H(p_j^*) + \sum_{\{u_j\},\{v_j\}} p_j(u_j|v_j)\hat{q}(v_j) \times \mathrm{E}_{\pi(q|p_1,\ldots,p_s)} \ln q(u_j, v_j) \\ &= H(p_s^*) + \sum_{\{u_s\},\{v_s\}} p_s(u_s|v_s)\hat{q}(v_s) \times \mathrm{E}_{\pi(q|p_1,\ldots,p_s)} \ln q(u_s, v_s). \end{aligned}$$

The solution of the resulting implicit relation (recall (3.7))

$$\hat{q}(u_j, v_j) = \underbrace{\sum_{j=1}^{s} w_{0j} p_j^*(u_j, v_j)}_{p_0(u_j, v_j)} + \sum_{j=1}^{s-1} \lambda_j \Big(p_j(u_j|v_j)\hat{q}(v_j) - p_s(u_s|v_s)\hat{q}(v_s) \Big) \tag{4.2}$$

is non-trivial due to the dependence on marginal pmfs $\hat{q}(v_j)$ and $\hat{q}(v_s)$ of $\hat{q}(x)$. The computation can be approached by successive approximations: by inserting a guess of $\hat{q}$ to the right-hand side of equation and obtaining a new guess on left-hand-side, etc. until reaching a stationary solution. This possibility is not elaborated here in detail. However, the following example indicates how the vital initial guess can be obtained in dynamic scenario.

In the dynamic scenario the solution of the implicit equation $\hat{q}_t(u_j, v_j)$ enters into the constraints via the marginals $\hat{q}_t(v_j)$, which are available only at the end of – inevitably iterative – computation at time instant t. Now, the idea is to use $\hat{q}_{t-1}(v_j)$, obtained from $\hat{q}_{t-1}(u_j, v_j)$, instead of $\hat{q}_t(v_j)$. The evaluation starts at $t = 1$ with pmf of the uniform distribution as $\hat{q}_{t-1}(v_j)$. Then, we apply the dynamic scenario described in Section 3.3.2.

Note that at each time instant we can compute a conditional version of the combination $\hat{q}_t$ respecting the variables considered by particular source. This conditional version can be then used by the particular source in the case it is unable to exploit the joint version. This feature extends the applicability of combining proposed in the previous chapter.

The following example shows the evolution of the optimal combination $\hat{q}_t$ of the desired combination q in time course.

Example 4.3. *Let $X = (X_1, X_2)$ where X_1 and X_2 have possible outcomes $\{1, 2, 3\}$, $(m = 2)$. Assume we have $s = 3$ sources providing conditional pmfs $p_j(u|v)$, for ease of presentation with a common splitting $u = x_1$, $v = x_2$. At time $t = 1$, the first source provides the following:*

$p_1(u\|v)$	$V = 1$	$V = 2$	$V = 3$
$U = 1$	*1/4*	*1/2*	*1/3*
$U = 2$	*3/4*	*1/2*	*1/3*
$U = 1$	*0*	*0*	*1/3*

the second source provides:

$p_2(u\|v)$	$V = 1$	$V = 2$	$V = 3$
$U = 1$	*1/5*	*1/3*	*1/4*
$U = 2$	*2/5*	*1/6*	*1/4*
$U = 3$	*2/5*	*1/2*	*1/2*

and the third source provides:

$p_3(u\|v)$	$V = 1$	$V = 2$	$V = 3$
$U = 1$	*0*	*1/8*	*1/2*
$U = 2$	*1/2*	*5/8*	*1/4*
$U = 3$	*1/2*	*1/4*	*1/4*

The outcomes $x = (u, v)$ are ordered pairs with the following structure:

- $x_1 = (1, 1)$, $x_2 = (1, 2)$, $x_3 = (1, 3)$
- $x_4 = (2, 1)$, $x_5 = (2, 2)$, $x_6 = (2, 3)$
- $x_7 = (3, 1)$, $x_8 = (3, 2)$, $x_9 = (3, 3)$.

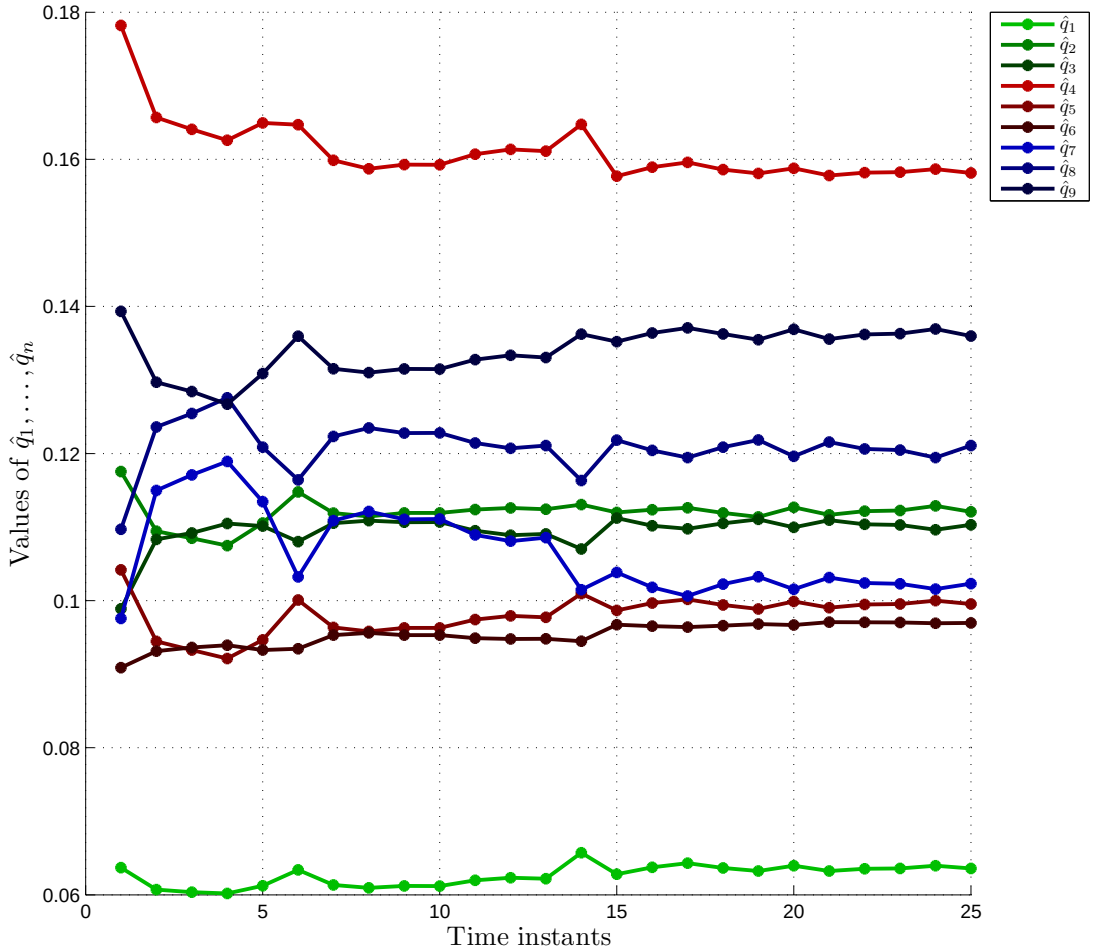


Fig. 4.1: The optimal combination $\hat{q}$ of conditional pmfs after proposed extension, $b = 10000$, $s = 3$, $n = 9$, $t = 1, \ldots, 25$.

We consider a dynamic scenario and at each time instant t, $t = 1, \ldots, 25$, *we generate new vector of conditional probabilities for each source by adding and subtracting a small noise of order less than* 10^{-3} *to provided pmfs, while keeping the properties of conditional pmf satisfied. In order to obtain the joint pmfs we exploit the extension* (4.1) *and the idea proposed for the dynamic case by* (3.13). *The behavior of the optimal combination* $\hat{q}$ *is shown in Fig. 4.1.*

After the last step of computation the resulting conditional pmf (projected back to sources) is:

$\hat{q}_T(u\|v)$	$V = 1$	$V = 2$	$V = 3$
$U = 1$	*0.2*	*0.34*	*0.32*
$U = 2$	*0.49*	*0.3*	*0.28*
$U = 3$	*0.31*	*0.36*	*0.4*

4.1.2 Marginal probability

Assume the j^{th} source provides the marginal pmf

$$p_j(x_1, \ldots, x_{k_j}) = P_j(X_1 = x_1, \ldots, X_{k_j} = x_{k_j}),$$

where $k_j < m$. Then, a similar need for extension as in Section 4.1.1 is faced. Again, denote $u_j = (x_1, \ldots, x_{k_j})$ and $v_j = (x_{k_j+1}, \ldots, x_m)$ splitting of $x = (u_j, v_j)$ for the j^{th} source. To construct the extension to the joint pmf we exploit the same idea as in the above section.

Definition 4.4. *Joint pmf $p_j^* = p_j^*(u_j, v_j)$ is called a reasonable extension of marginal pmf $p_j(v_j)$ provided by the j^{th} source if*

- *a) its marginal version coincides with given marginal pmf: $p_j^*(v_j) = p_j(v_j)$,*
- *b) it minimizes the Kullback-Leibler divergence $KLD(\hat{q}|p_j^\bullet)$ among all pmfs $p_j^\bullet$ satisfying a),*

where $\hat{q}$ is the optimal combination of the processed information sources and u_j, v_j stand for the splitting of $(x_1, \ldots, x_m)$ for the j^{th} source.

Similarly to Proposition 4.2 we can obtain the exact form of the extension

$$p_j^*(u_j, v_j) = \hat{q}(u_j|v_j) \times p_j(v_j),$$

where $\hat{q}(u_j|v_j)$ is derived from $\hat{q}(u_j, v_j)$ and implicit equation similar to (4.2) is to be solved.

For approximations and dynamic scenarios, we can exploit a similar idea as in Section 4.1.1. Thus, at time instant t we use $\hat{q}_{t-1}(u_j|v_j)$ instead of $\hat{q}_t(u_j|v_j)$.

4.2 Non-probabilistic form – transformation

In this section we discuss the non-probabilistic form of provided information - when subsets of the set of possible outcomes or requirements on expected value of X are given.

4.2.1 Subsets of the set of outcomes

Subset of outcomes with specified probabilities

Assume that source provides a subset of possible outcomes together with the probability of each outcome such that the sum of these probabilities is one. In this case, we can directly proceed to the combining after extension of given pmf, if needed.

Subset of outcomes with unspecified probabilities

Let us assume j^{th} source provides the following information: 'The possible outcomes of X are x_k and x_l' ($x_k \neq x_l$, $k \neq l$). Since the source does not specify

which outcome is more probable, we assume both can appear with the same probability. Regarding other outcomes, we assume they are unlikely to appear and thus the probability of any element in the set $\{x_1, \ldots, x_n\} \setminus \{x_k, x_l\}$ is zero. The information given by j^{th} source can be then represented by the following pmf:

$$p_j = (0, \ldots, \underbrace{1/2}_{k}, \ldots, 0, \underbrace{1/2}_{l}, \ldots, 0).$$

This holds for any subset of cardinality $n_1 \in \{1, \ldots, n\}$, where n denotes the number of all possible (different) outcomes of X. Each of the considered outcomes is then assigned probability $1/n_1$. Now, we can apply the above introduced extension or combine pmfs directly.

A single numerical value

Let us assume that we have n possible outcomes of random vector X and n sources, $(s = n)$. Assume also each source suggests one outcome (a crisp) and, for simplicity assume that j^{th} source provides x_j, $j = 1, \ldots, s$. We exploit the Kronecker delta,

$$\begin{aligned} p_j(x) = \delta_{x,x_{j,given}} &= 1 \text{ if } x = x_{j,given} \\ &= 0 \text{ otherwise,} \end{aligned} \tag{4.3}$$

to transform the numerical value given by j^{th} source denoted by $x_{j,given}$ into pmf. That is

$$p_j = (0, \ldots, 0, \underbrace{1}_{x_j,given}, 0, \ldots, 0),$$

with probability one assigned to outcome $x_{j,given}$, $j = 1, \ldots, s$.

A single numerical value: properties of the optimal combination

In this case, the corresponding entropies $H(p_1), \ldots, H(p_n)$ are equal (zero). For the dissimilarity constraints in the static case the following holds:

$$H(p_j) + \sum_{i=1}^{n} p_{ji} \left(\psi(\nu_i) - \psi\left(\sum_{k=1}^{n} \nu_k \right) \right) = H(p_s) + \sum_{i=1}^{n} p_{si} \left(\psi(\nu_i) - \psi\left(\sum_{k=1}^{n} \nu_k \right) \right)$$

$$\underbrace{p_j(x_{j,given})}_{1} \psi\left(\nu_{j,given}\right) - n\psi\left(\sum_{k=1}^{n} \nu_k \right) = \underbrace{p_s(x_{s,given})}_{1} \psi(\nu_{s,given}) - n\psi\left(\sum_{k=1}^{n} \nu_k \right).$$

Thus, we have

$$\psi(\nu_{j,given}) = \psi(\nu_{s,given}), \quad j = 1, \ldots, s-1,$$

yielding

$$\nu_k = \nu_l, \quad k, l = 1 \ldots, n, \tag{4.4}$$

because the digamma function $\psi(.)$ is a continuous increasing function on $\mathcal{R}_+$. The following then holds for parameters $(\hat{\nu}_1, \ldots, \hat{\nu}_n)$ yielding $\hat{q}$:

$$\begin{aligned}\hat{\nu}_i &= \nu_{0i} + \sum_{j=1}^{s-1} \lambda_j (p_{ji} - p_{si}) = \nu_{0i} + \sum_{j=1}^{s-1} \lambda_j \delta_{x_i, x_{j,given}}, \\ \hat{\nu}_n &= \nu_{0n} + \sum_{j=1}^{s-1} \lambda_j (p_{jn} - p_{sn}) = \nu_{0n} - \sum_{i=1}^{n-1} \sum_{j=1}^{s-1} \lambda_j \delta_{x_i, x_{j,given}}. \end{aligned} \tag{4.5}$$

According to the relation (4.4) we obtain that

$$\sum_{j=1}^{s-1} \lambda_j \delta_{p_{ji}, p_{si}} = \nu_{01} + \sum_{j=1}^{s-1} \lambda_j \delta_{p_{j1}, p_{s1}} - \nu_{0i}, \quad i = 1, \ldots, n-1. \tag{4.6}$$

By plugging (4.6) into the equation for ν_n in (4.5) we obtain

$$\nu_{01} + \sum_{j=1}^{s-1} \lambda_j \delta_{p_{j1}, p_{s1}} = \nu_{0n} - \sum_{i=1}^{n-1} \left(\nu_{01} + \sum_{j=1}^{s-1} \lambda_j \delta_{p_{j1}, p_{s1}} - \nu_{0i} \right),$$

and

$$\sum_{j=1}^{s-1} \lambda_j \delta_{p_{j1}, p_{s1}} = \frac{\sum_{i=1}^{n} \nu_{0i}}{n} - n \frac{\nu_{01}}{n} = \frac{1}{n} - \nu_{01}. \tag{4.7}$$

Thus, in any case of prior guesses $\nu_{01}, \ldots, \nu_{0n}$ we are able to compute the exact values of $\lambda_1, \ldots, \lambda_{s-1}$, while satisfying the equality relation (4.4). We demonstrate the derived relation together with the case of duplicate pmfs in the following example.

Example 4.5. *Let us have random variable X with three possible outcomes $(n = 3)$ and sources $(s = 3)$ providing following pmfs:*

$$\begin{aligned} p_1 &= [1, 0, 0], \\ p_2 &= [0, 1, 0], \\ p_3 &= [0, 0, 1]. \end{aligned}$$

The prior guess $\nu_{01}, \ldots, \nu_{0n}$ is still considered as the arithmetic mean of available pmfs

$$\nu_{0i} = \frac{1}{3}, \quad i = 1, \ldots, n.$$

According to (4.4) *we obtain that*

$$\lambda_1 = \lambda_2 = 0.$$

The optimal combination of p_1, p_2 and p_3 is a pmf of the uniform distribution $\hat{q} = (0.33, 0.33, 0.33)$.

We next show that by adding duplicate pmf to the given set of pmfs the value of the resulting optimal combination $\hat{q}$ will not change. Note the difference between

Proposition 3.7 and the current situation - now the prior guess $\nu_{01}, \ldots, \nu_{0n}$ (the arithmetic mean of given pmfs) has changed.

If we add $p_4 = (1, 0, 0)$, the resulting combination (where the prior pmf p_0 is the arithmetic mean of $p_1, \ldots, p_4$) is again $\hat{q} = (0.33, 0.33, 0.33)$. This is the consequence of general relations (4.7) *and* (4.6) *which allow us to compute exact values of $\lambda_1, \ldots, \lambda_4$ while keeping* (4.4) *satisfied.*

In case of four sources with the above provided pmfs we obtain:

$$\lambda_1 = \frac{1}{3} - \frac{2}{4},$$
$$\lambda_2 = \lambda_3 = \frac{2}{4} + \left(\frac{1}{3} - \frac{2}{4}\right) - \frac{1}{4} = \frac{1}{3} - \frac{1}{4}.$$

For values $\hat{\nu}_i$, related to outcomes x_i which were not observed, the following holds

$$\hat{\nu}_i = \nu_{0i}. \tag{4.8}$$

Since we $\nu_{01}, \ldots, \nu_{0n}$ are based on the arithmetic mean of given pmfs $p_1, \ldots, p_s$, for computational reasons ($\nu_{0i} > 0$, $i = 1, \ldots, n$) we perturb such ν_{0i} by a small positive value ($\leq 10^{-5}$) as noted in Remark 3.2.2. This suggestion is demonstrated in the following example.

Example 4.6. *Assume that after transformation we obtained the following pmfs:*

$$p_1 = [0, 0, 1]$$
$$p_2 = [0, 1, 0]$$
$$p_3 = [0, 1, 0]$$

As formula (4.8) *suggests, in this case we have $\hat{\nu}_1 = \nu_{01}$. Since throughout the thesis we exploit prior pmf p_0 as the arithmetic mean of given pmfs, this implies $\nu_{01} = 0$ which is in contradiction with constraints in* (3.10). *To solve this we suggest to assign p_{01} a small positive value say $< 10^{-5}$, which ensures the constraints $\nu_{0i} > 0$, $i = 1, \ldots, n$, are satisfied and minimizes the influence of ν_{01}. The optimal combination of p_1, p_2, p_3, computed numerically, is $\hat{q} = (0.00, 0.50, 0.50)$.*

In the case that the range of X is not specified, we can focus on the union of outcomes given by sources. The resulting pmf in the considered illustrative example would be then $\hat{q} = (0.5, 0.5)$.

A single numerical value: connection to Bayesian solution

In the next remark we study the connection of current derivations to the Bayes rule in the situations in which it is applicable.

Remark. *Let us consider a random event, which takes values in the set of n possible values $\{x_i\}_{i=1}^n$. The most general probability distribution covering this case is the categorical distribution $Cat(q_1, \ldots, q_n)$, where $q_i = q(x_i) = P(X = x_i)$. Vector $q = (q_1, \ldots, q_n)$ is an unknown parameter satisfying the properties of pmf: $0 \leq q_i \leq 1$, $\sum_{i=1}^n q_i = 1$. Let us assume we have s independent observations $y_1, \ldots, y_s$. In order to obtain the estimator $\hat{q}$ of q based on $y_1, \ldots, y_s$ we consider*

q as a random vector and exploit the theory of Bayesian estimation, i.e., the Bayes rule

$$\pi(q|y_1,\ldots,y_s) \propto f(y_1,\ldots,y_s|q) \times \pi_0(q)$$

where $\pi_0(q)$ *is the prior pdf,* $f(y_1,\ldots,y_s|q)$ *is the likelihood and* $\pi(q|y_1,\ldots,y_s)$ *is the posterior pdf. In this case the likelihood is*

$$f(y_1,\ldots,y_s|q) = \prod_{j=1}^{s} f(y_j|q) = \prod_{i=1}^{n} q_i^{\sum_{j=1}^{s}\delta_{x_i,y_j}},$$

where $\delta_{x,y}$ *denotes the Kronecker delta.*

The estimator $\hat{q}$ *is then based on the resulting posterior pdf (is a function of posterior pdf).*

Here, we exploit the Dirichlet distribution, a conjugate prior for the categorical distribution (Raïffa and Schlaifer, 1961). Thus, for $Dir(\alpha_{01},\ldots,\alpha_{0n})$ *we obtain:*

$$\pi(q|y_1,\ldots,y_s) \propto \prod_{i=1}^{n} q_i^{\sum_{j=1}^{s}\delta_{x_i,y_j}} \prod_{i=1}^{n} q_i^{\alpha_{0i}-1}.$$

The Bayesian update of parameters of the Dirichlet distribution based on $y_1,\ldots,y_s$ *is thus:*

$$\alpha_i = \alpha_{0i} + \sum_{j=1}^{s} \delta_{x_i,y_j}.$$

In the case that we know that the estimator $\hat{q}$ *is the conditional expectation with respect to* $\pi(q|y_1,\ldots,y_s)$ *the following holds:*

$$\hat{q}_i = E_{\pi(q|y_1,\ldots,y_s)}(q_i|y_1,\ldots,y_s) = \frac{\alpha_i}{\sum_{k=1}^{n}\alpha_k} = \frac{\alpha_{0i} + \sum_{j=1}^{s}\delta_{x_i,y_j}}{\sum_{k=1}^{n}\alpha_{0k} + s},$$

where

$$\sum_{k=1}^{n}\alpha_k = \sum_{k=1}^{n}\alpha_{0k} + \sum_{j=1}^{s}\sum_{k=1}^{n}\delta_{x_k,y_j}.$$

The formula above reminds of the optimal combination $\hat{q}$*. In the case that each source provided a single value from the set of possible values, we can rewrite the formula* (3.3) *for prior pdf* $Dir(\nu_{01},\ldots,\nu_{0n})$ *as follows:*

$$\hat{q}_i = \frac{\nu_{0i} + \sum_{j=1}^{s-1}\lambda_j\delta_{x_i,y_j} - \sum_{j=1}^{s-1}\lambda_j\delta_{x_s,y_j}}{\sum_{k=1}^{n}\nu_{0k}}. \tag{4.9}$$

The main difference is in the sum of parameters: for the posterior pdf obtained by the Bayes rule it is increased by number of sources s (the number of new observations). In (4.9)*, the sum does not change, which coincides with the former explanation that the prior can already be based on obtained data (can be expressed in terms of new arrived data).*

4.2.2 Specified expected value

In this section we assume that sources' opinions are represented by moments of the discrete random vector X. For j^{th} source the notation will be

$$e_j = E_{p_j}\phi(X), \tag{4.10}$$

where p_j is incompletely specified pmf with elements $p_{ji} = P_j(X = x_i)$ and $\phi(.)$ is a given function of X (typically polynomial). In order to transform expectation e_j to pmf p_j we exploit the maximum entropy principle or, if a guess p_{0j} of p_j is available, the minimum cross-entropy principle (see Section 1.2).

Definition 4.7. *Joint pmf p_j is a reasonable representation of expected value e_j given by j^{th} source if*

- *it satisfies relation* (4.10)*;*
- *it maximizes entropy:*

  $$p_j = \arg\max H(p_j^\bullet),$$

  where the minimization is performed over the set of pmfs $p_j^\bullet$ satisfying a).

We choose p_j with the largest entropy among all probability vectors satisfying the relation (4.10) and use it in the optimal combination $\hat{q}$ defined in (3.3).

Chapter 5

Real data application

In this chapter we apply the proposed combination to three different sets of real data: data for decision making in a company, galaxy zoo data and data from European social survey. Each section brings an overview of used data and obtained results.

5.1 Decision making in contract evaluation

In this example we focus on decision making in a company satisfying the definition of a small and medium enterprize. We would like to thank Dr. Ing. Pavel Ettler (COMPUREG Plzeň, s.r.o.) for the data exploited in this section.

The company, owned by 5 partners, focuses on production of industrial devices and information systems. Each owner possesses the same ownership stake and thus, each vote is equivalently included in strategic decision making.

Company deals with orders or potential actions of following type:

A Order in the main domain of the company: built know-how can be used, the assertion of the modification of existing solution is expected, assumed to be the most profitable.

B Order in a new domain: a significant effort in search for a new solution is expected, which makes this order less profitable.

C Long-term care about customer's devices: functioning devices generally influence profitability of contracts A and B; broken devices lower their profitability.

In this case the basic flat-rates, maximum length of the time period between reported fault and beginning of the repair have to be defined.

D One-off repair of the device: broken device before repair - non-profitable, repaired device influences profits of contracts A and B.

The decision, whether to accept or decline the order, is based on several criteria: suggested price, expenses, employee capacity, required deadline. The data obtained from owners is shown in Table 5.1. To help owners to decide and then to

assign the priority to each contract and order we apply the proposed combination (3.3).

5.1.1 Ranking without assigned probabilities

Here, we focus on Table 1 in Table 5.1 containing rankings $P1, \ldots, P5$ of individual partners for following aspects characterizing the above variants A-D:

- REI - relative expected income (% - mutually relative with the variant A taken as the base),
- RD - relative demandingness (% - mutually relative with the variant A taken as the base),
- DP - delivery period (months) of the order,
- Dist - distance (kilometers) to realization place influencing additional (non-monetary) costs related to the given order,
- P1,. . ., P5 - ranking from five partners for each aspect (values on discrete scale {1,. . .,10}, 1-least favorable, 10-most favorable),
- OIR1,. . ., OIR5 - overall individual ranking from five partners (a real number from $[1, 10]$).

Since the overall individual rankings (OIRs) from partners are available, the company would like to know the resulting ranking for each contract and order. Data are given in columns OR1,. . ., OR5 in Table 1 of Table 5.1. Let us consider random variable X as ranking of a particular contract with possible outcomes x equal to individual rankings and unknown pmf $q(x)$. In search for the combination of OIRs we exploit Section 4.2.1. In particular, we interpret the outcomes in the probabilistic form (pmfs $p_1, \ldots, p_5$) and combine them using (3.3). Resulting optimal combination $\hat{q}$ of $p_1, \ldots, p_s$ is the pmf of the uniform distribution, as implied by (4.4). To give back a single ranking for each contract/order we compute the expected value of OIRs with respect to the obtained pmf $\hat{q}$. The result is simply the weighted linear combination of individual rankings for equal weights (partners votes are equivalent) summing to one, i.e., the arithmetic mean of rankings. Resulting expected values for each contract are given in the following table with contract numbers corresponding to numbering in Table 1 of Table 5.1:

No. 1	9.38	No. 2	9.38	No. 3	7.71	No. 4	6.56	No. 5	6.36

We see that contract No. 1 and No. 2 obtained the highest and the same value of expected OIR, thus they are undistinguishable for us at the moment. In the following section we study whether this remains true when we focus on rankings of a particular aspect with assigned probabilities.

Table 1

No.	Type	REI	P1	P2	P3	P4	P5	RD	P1	P2	P3	P4	P5	DP	P1	P2	P3	P4	P5
1	A	1,00	10	10	9	10	10	1,00	10	10	8	10	10	7	10	10	10	10	10
2	A	0,85	10	9	9	7	10	0,70	9	9	7	10	10	5	10	10	9	10	10
3	B	0,15	3	7	8	5	8	0,60	5	10	6	9	9	3	7	8	8	5	10
4	C	0,05/year	2	6	7	7	8	0,10	3	7	10	10	9	0	3	9	5	4	8
5	D	0,20	5	1	7	5	8	0,25	2	8	10	9	9	1	4	5	7	6	9

No.	Type	Dist	P1	P2	P3	P4	P5	**OIR1**	**OIR2**	**OIR3**	**OIR4**	**OIR5**
1	A	600	6	8	9	9	5	9.50	9.75	9.00	8.75	9.88
2	A	120	8	10	10	10	9	9.75	9.50	8.75	9.00	9.94
3	B	10	10	10	10	10	10	6.25	8.75	8.00	7.25	8.31
4	C	80	10	9	6	7	7	4.25	7.00	7.25	7.00	7.31
5	D	250	4	6	9	9	4	4.25	5.75	8.25	6.00	7.56

Table 2

No.	Type	REI	P1	P2	P3	P4	P5	RD	P1	P2	P3	P4	P5	DP	P1	P2	P3	P4	P5
1	A	1,00	10:1.0	10:1.0	9:0.6 10:0.4	10:1.0	10:1.0	1.00	10:1.0	10:1.0	8:1.0	10:1.0	10:1.0	7	10:1.0	10:1.0	10:1.0	10:1.0	10:1.0
2	A	0,85	10:1.0	9:1.0	9:1.0	7:1.0	10:1.0	0.70	8:0.3 9:0.6 10:0.1	9:1.0	7:1.0	10:1.0	10:1.0	5	10:1.0	10:1.0	9:1.0	10:1.0	10:1.0
3	B	0,15	2:0.3 3:0.4 4:0.3	7:1.0	8:1.0	5:1.0	8:1.0	0.60	5:1.0	10:1.0	4:0.1 5:0.1 6:0.6 7:0.1 8:0.1	9:1.0	9:1.0	3	7:1.0	8:1.0	8:1.0	5:1.0	10:1.0
4	C	0,05/year	1:0.2 2:0.5 3:0.3	6:1.0	6:0.2 7:0.6 8:0.2	7:1.0	8:1.0	0.10	3:1.0	7:1.0	10:1.0	10:1.0	9:1.0	-	1:0.2 2:0.2 3:0.2 4:0.2 5:0.2	9:1.0	5:1.0	4:1.0	8:1.0
5	D	0,20	4:0.2 5:0.4 6:0.2	1:1.0	7:1.0	5:1.0	8:0.4 9:0.5 10:0.1	0.25	1:0. 2:0.333 3:0.333	8:1.0	10:1.0	9:1.0	8:0.3 9:0.4 10:0.3	1	4:1.0	5:1.0	7:1.0	6:1.0	9:1.0

No.	Type	**Dist**	P1	P2	P3	P4	P5
1	A	600	6:0.9 7:0.1	8:1.0	8:0.2 9:0.7 10:0.1	9:1.0	2:0.1 3:0.1 4:0.2 5:0.2 6:0.2 7:0.1 8:0.1
2	A	120	8:1.0	10:1.0	10:1.0	10:1.0	9:1.0
3	B	10	10:1.0	10:1.0	10:1.0	10:1.0	10:1.0
4	C	80	10:1.0	9:1.0	6:1.0	7:1.0	7:1.0
5	D	250	3:0.3 4:0.4 5:0.3	6:1.0	9:1.0	7:0.2 8:0.3 9:0.3 10:0.2	4:1.0

Table 5.1: Data for decision making. Above: Table 1 includes aspect-dependent rankings and overall rankings from partners P1,...,P5 and OIR1,...,OIR5 respectively. Below: Table 2 includes aspect-dependent rankings from partners $P1, \ldots, P5$ with assigned probabilities displayed as *ranking:assigned probability*.

5.1.2 Ranking with assigned probabilities

The combining of deterministic opinion brought no added value to standard averaging and did not help in distinguishing contracts No. 1 and No. 2. In this section, partners were more expressive in specifying their rankings of the contract aspects and provided several rankings with probabilities for particular aspect of a particular contract, see Table 2 in Table 5.1. To demonstrate the contributions of the proposed combination, we focus on aspect 'Dist'. Here, the random variable X is again the ranking with possible outcomes $\{1, \ldots, 10\}$. Since some outcomes are assigned zero probability from all partners, we exploit Remark 3.2.2 to combine partners' opinions. Resulting combined pmfs and corresponding expected rankings for contracts with respect to the aspect 'Distance' are given in Table 5.2.

Aspect: Distance											
Ranking	1	2	3	4	5	6	7	8	9	10	Exp. rank.
No. 1	0	0.02	0.07	0.19	0.15	0.23	0.07	0.05	0.21	0.02	6.1
No. 2	0	0	0	0	0	0	0	0.33	0.33	0.33	9
No. 3	0	0	0	0	0	0	0	0	0	1	10
No. 4	0	0	0	0	0	0.25	0.25	0	0.25	0.25	8
No. 5	0	0	0.14	0.10	0.06	0.10	0.14	0.32	0.10	0.03	6.53

Table 5.2: Results after combining data for aspect 'Dist' given in Table 2 of Table 5.1 by proposed combination. The expected values for rankings are also given.

Unlike in previous case, we can now easily distinguish the contracts. With respect to aspect 'Distance' the highest expected ranking obtained contract No. 3 followed by contract No. 2, which in the previous section obtained the highest expected OIR.

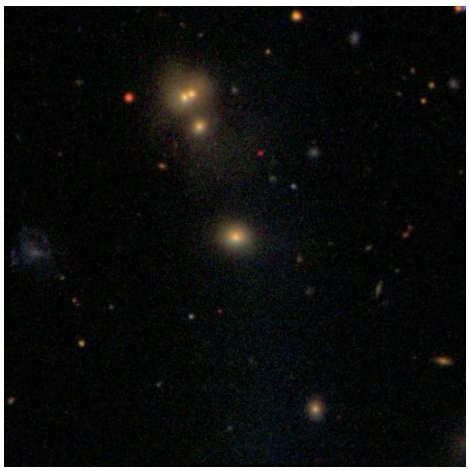
5.2 Galaxy Zoo data

In this section we focus on classification of galaxies in astrophysics. For a long time the catalogues including the classifications for available (documented) galaxies were produced by individuals or small groups of astronomers. The rise of modern technologies provides a larger recording of galaxies (e.g., Sloan Digital Sky Survey - SDSS) and makes the classifying of all elements very time consuming.

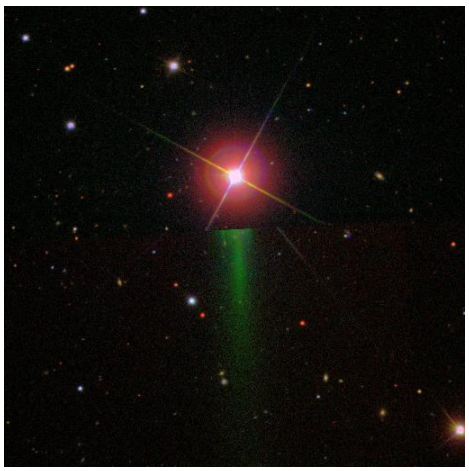
To solve this problem public was allowed to participate in classification. In 2007 a web page including pictures of galaxies from SDSS was launched to let the users from all over the world classify the galaxies (`http://zoo1.galaxyzoo.org/`).

In order to be included in online classification it is optional to register, read a tutorial and pass first (simpler) classification (11 out of 15 galaxies have to be identified correctly). Classifications of unregistered users are denied. Details can be found in Lintott et al. (2008). The current version of the project can be found on `http://zoo2.galaxyzoo.org/`.

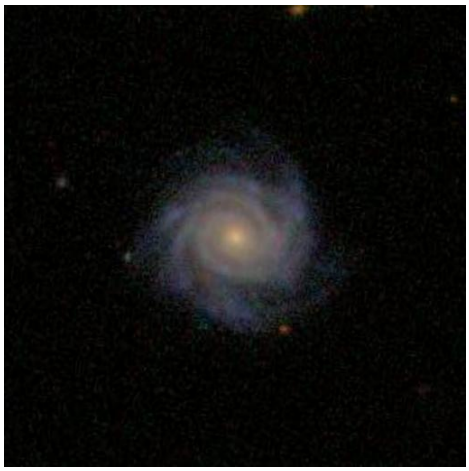
The data release includes several datasets (`http://data.galaxyzoo.org/`). We focus on Table 7 which for each galaxy includes the probabilities of belonging



(a) Elliptical galaxy

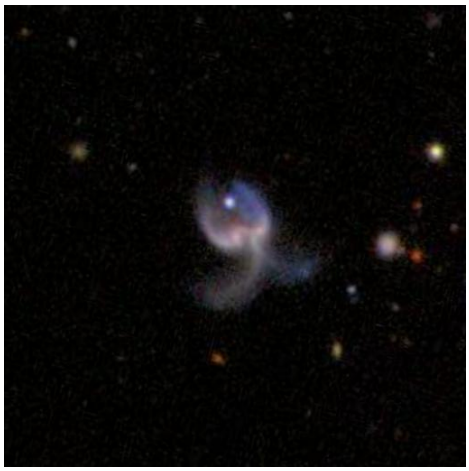


(b) Star/Don't know

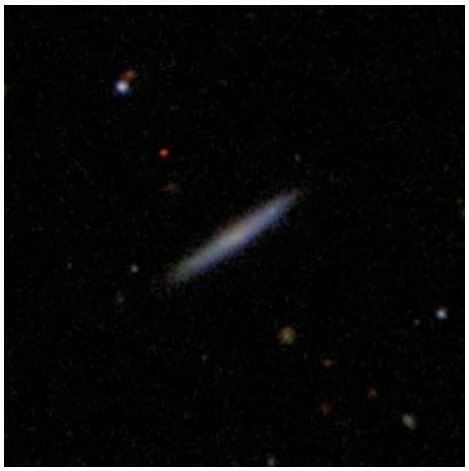


(c) Spiral (clockwise) galaxy

(d) Spiral (anticlockwise) galaxy

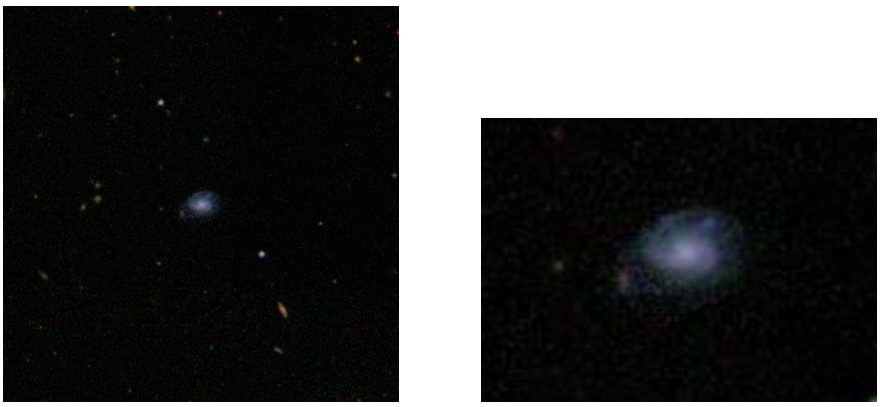


(e) Merger



(f) Edge on

Fig. 5.1: Six different categories for objects (galaxies) in the pictures available on Galaxy Zoo.



(a) Original SDSS image. (b) Image close up.

Fig. 5.2: 'OBJID': 587727180064817249

"OBJID"	"P_EL"	"P_CW"	"P_ACW"	"P_EDGE"	"P_DK"	"P_MG"
587727227300741210	0.4790	0.0170	0.1210	0.3380	0.0450	0
587727225153257596	0.7350	0.0290	0	0.1470	0.0740	0.0150
587730774425665700	0.4790	0	0	0.0140	0.4790	0.0270
587730774962536596	0.8850	0.0190	0	0.0580	0.0190	0.0190
587731186203885750	0.7120	0	0	0.2200	0.0680	0

Table 5.3: Sample of online available galaxy zoo data (Table 7 available online).

to a particular category out of 6 considered categories.

The sample from the downloaded data is shown in Table 5.3 with following abbreviations:

- ObjID: identification number of the galaxy according to SDSS
- P_EL - elliptical galaxy
- P_CW - spiral (clockwise) galaxy
- P_ACW - spiral (anticlockwise) galaxy
- P_EDGE - spiral (edge-on) galaxy
- P_DK - don't know
- P_MG - merger.

Examples of galaxies are given in Fig. 5.1 (images obtained from SDSS).

We focus on the object with 'OBJID' 587727180064817249, see Fig. 5.2. Pmf available for this object in Table 7 of Galaxy Zoo datasets is:

"OBJID"	"P_EL"	"P_CW"	"P_ACW"	"P_EDGE"	"P_DK"	"P_MG"
587727180064817249	0.045	0.023	0.881	0.04	0.01	0

To connect current setup with terms used in Chapter 3 we consider 'galaxy category' as the random variable X having six possible outcomes: "P_EL", "P_CW", "P_ACW", "P_EDGE", "P_DK", "P_MG", $n = 6$.

The probability vectors given in Table 7 available online (a sample from this table is given in Table 5.3) reflect the average of the original votes obtained by

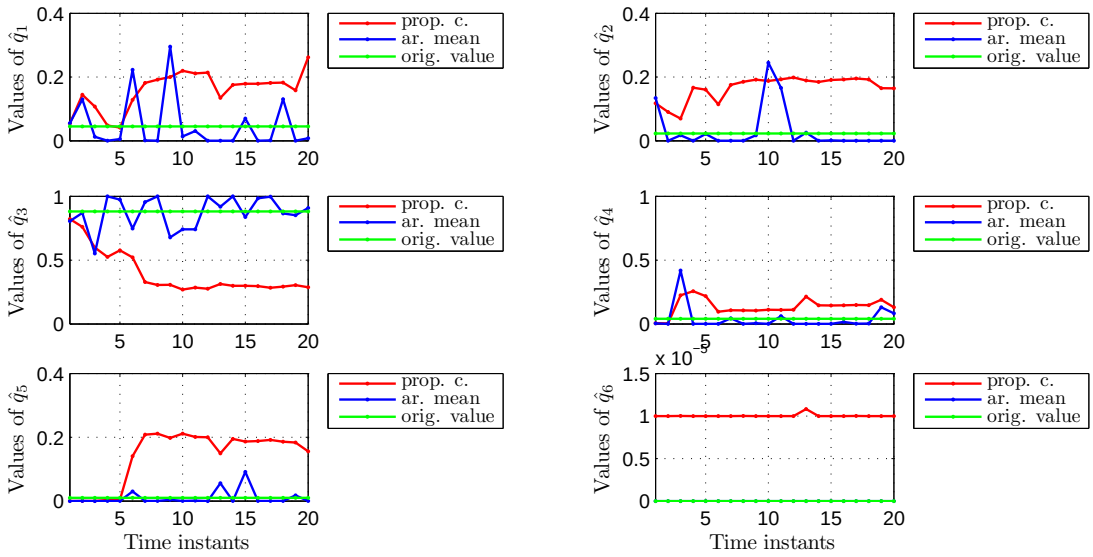


Fig. 5.3: GALAXY data - Resulting combination of simulated pmfs based on proposed combination 'prop. c.' ($b = 100000$) and arithmetic mean of processed data 'ar. mean' for galaxy with 'OBJID' 587727180064817249.

galaxy zoo. However, to study the performance of proposed combination we assume the above given pmf represents the unknown desired combination q

$$q = (0.045, 0.023, 0.881, 0.04, 0.01, 0) \tag{5.1}$$

and simulate a set of pmfs representing votes. Simulated votes will be then combined by using (3.3) to obtain their combination $\hat{q}$.

In the thesis we assumed $\hat{q}$ coincides with $\hat{\nu}$, the vector of parameters of the Dirichlet distribution (see (3.6) and discussion there). To ensure the resulting $\hat{q}$ is comparable with q we simulate pmfs from the Dirichlet distribution with parameters $[\nu_1, \ldots, \nu_n] = [0.045, 0.023, 0.881, 0.04, 0.01, 0]$.

If the original set of votes were available, such large dataset would be processed sequentially. We thus considered several time instants with simulating and processing 5 new pmfs ($s = 5$) at each instant. Then, we applied the theory proposed for dynamic scenarios, see Section 3.3.2.

Note that if the original votes from classifiers were available online, we would exploit Kronecker delta (4.3) to obtain their combination, see Section 4.2.1, while for the elements of simulated pmfs the following holds: $0 \leq p_{ji} \leq 1$, $\sum_{i=1}^{n} p_{ji} = 1$, $j = 1, \ldots, 5$, $i = 1, \ldots, 6$.

Values of the proposed combination based on simulated data are shown in Fig. 5.3 together with arithmetic mean of simulated data for each time step and the original value of pmf (5.1). We see that the combination assigns higher probability to categories with low original probability (categories "P_EL", "P_CW", "P_EDGE", "P_DK"). Consequently, category "P_ACW" was assigned lower probability than its original probability. Since the prior guess $\nu_{01}, \ldots, \nu_{0n}$ is based on arithmetic mean and all elements have to be positive, we exploit Remark 3.2.2 and assign $\nu_{06} = 10^{-5}$, see values of $\hat{q}_6$ in Fig. 5.3. Let us stress that the ob-

tained results are illustrative and surely influenced by errors inherent to sequential treatment and small-sample properties.

5.3 European social survey data

In this section we focus on data available on European social survey (ESS). The basic information, available on their web page `www.europeansocialsurvey.org` and the available documentation, state:

> "ESS is an academically-driven multi-country survey, which has been administered in over 30 countries. Its three aims are, firstly - to monitor and interpret changing public attitudes and values within Europe and to investigate how they interact with Europe's changing institutions, secondly - to advance and consolidate improved methods of cross-national survey measurement in Europe and beyond, and thirdly - to develop a series of European social indicators, including attitudinal indicators."
>
> "36 participating countries altogether have participated in the first six rounds of the ESS: Albania, Austria, Belgium, Bulgaria, Croatia, Cyprus, the Czech Republic, Denmark, Estonia, Finland, France, Germany, Greece, Hungary, Iceland, Ireland, Israel, Italy, Kosovo, Latvia, Lithuania, Luxembourg, Netherlands, Norway, Poland, Portugal, Romania, the Russian Federation, Slovakia, Slovenia, Spain, Sweden, Switzerland, Turkey, Ukraine and the United Kingdom."

This project offers data and documentation by year, country and theme with a wide range of variables: media use and trust, politics, subjective well being, etc. If only particular variables in particular countries and rounds (years) are of interest, the cumulative data wizard can be exploited - "the cumulative file contains data from countries that have been included in the integrated ESS files in two or more rounds." Each set of data also includes a documentation, where the basic information about ESS (introduction, overview, scope and coverage, sampling, etc.) and chosen variables (values, labels) is given.

Happy vs. social meet

We focus on the data from the last round (number 6 - year 2012) in the Czech Republic, see ESS (b). In particular we focus on the following variables:

- happy: 'Taking all things together, how happy would you say you are?' with values and labels:
  - 0 - extremely unhappy,
  - ...,
  - 10 - extremely happy,
  - 77 - refusal to answer,
  - 88 - do not know.

cntry	cname	cseqno	idno	dweight	pspwght	pweight	happy	sclmeet
CZ	ESS1-6e01	50053	1001	1.146	1.063	0.446	9	6
CZ	ESS1-6e01	50054	1002	1.146	1.289	0.446	8	6
CZ	ESS1-6e01	50055	1003	1.146	1.289	0.446	9	5
CZ	ESS1-6e01	50056	1004	1.778	3.041	0.446	7	4

Table 5.4: Sample of SSE data.

- sclmeet: 'How often do you meet socially with friends, relatives or work colleagues?' with values and labels:
  - 1 - never,
  - 2 - less than once a month,
  - 3 - once a month,
  - 4 - several times a month,
  - 5 - once a week,
  - 6 - several times a week,
  - 7 - everyday,
  - 77 - refusal to answer,
  - 88 - do not know,
  - 99 - no answer.

A part of downloaded data is shown in Table 5.4, where

- cseqno: is respondent's sequence number in cumulative dataset
- idno: respondent's identification number
- dweight: design weight
- pspwght: post-stratification weight including design weight
- pweight: population size weight.

The details for weights can be found in the documentation provided together with the downloaded dataset, see ESS (a).

In this case the original data are available (unlike in Section 5.2), thus we can directly apply the optimal combination proposed in Section 3.1. Here, we focus on the relation between the rankings in the original data and consider a random variable X with 4 possible values ($n = 4$):

- x_1 = happy ≥ 5 and sclmeet ≥ 5
- x_2 = happy ≥ 5 and sclmeet < 5
- x_3 = happy < 5 and sclmeet ≥ 5
- x_4 = happy < 5 and sclmeet < 5.

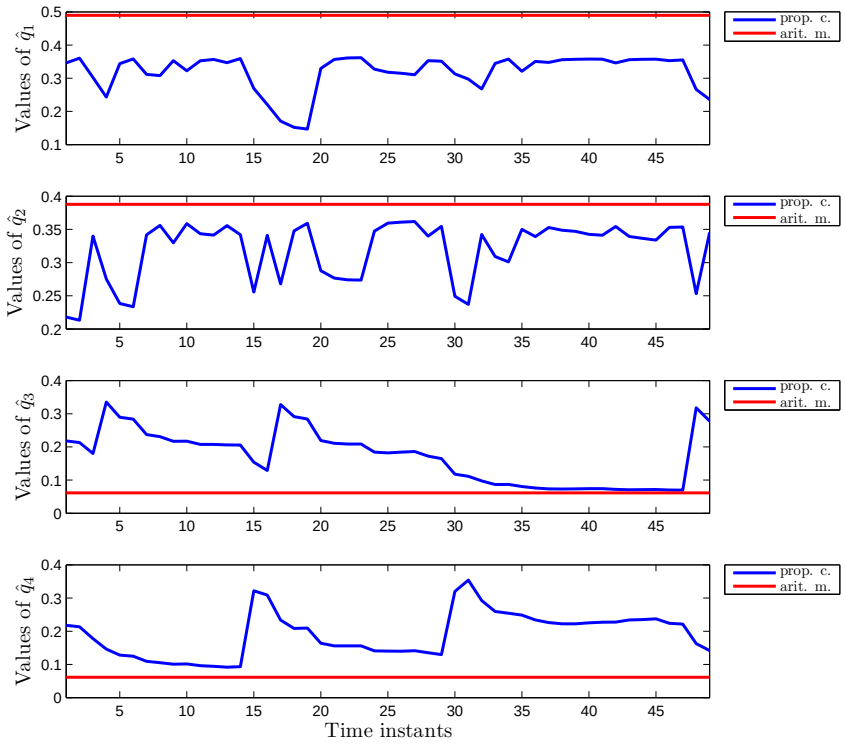


Fig. 5.4: SSE data - the proposed combination (prop. c.) and the arithmetic mean (arit. m.), both based on data from cumulative dataset for the Czech Republic (happy vs. sclmeet).

To apply the proposed combination (3.3) we have to transform data into probabilistic form first, see Section 4.2. Combining all pmfs, expressed as Kronecker delta (4.3), at once would yield a uniform pmf over observed values as the optimal combination (see discussion for single numerical value in Section 4.2.1). Thus, we rather exploit the dynamic scenario and combine each new observation (line) in the dataset with value of the optimal estimator from the previous time step (see Section 3.3.2).

We read first 50 relevant lines from the downloaded cumulative dataset and express them as observations about outcomes $x_1, \ldots, x_4$ (the observations with values 77, 88, 99 were excluded). The values of optimal combination $\hat{q}$, where at each time step t a new pmf and the optimal combination from the previous step $(t-1)$ were combined according to (3.3), are shown in Fig. 5.4 ($\hat{q}_0$ was set as pmf of the uniform distribution). Value of optimal combination at $T = 50$ is:

$$\hat{q}_T = (0.24, 0.34, 0.28, 0.14).$$

We might conclude, that based on answers regarding social meetings and happiness only, people in the Czech Republic in year 2012 were with higher probability happier when having less social interactions.

Chapter 6

Cross-entropy based combination in estimation

In this chapter we compare the combining proposed in (3.3) with the approach introduced in (Dedecius and Sečkárová, 2013b), primarily developed for the dynamic distributed estimation (DDE) in exponential family.

Like cross-entropy based combining, this approach also heavily exploits the KL-divergence. This yields parameter estimator as linear weighted combination of sources' observations and, moreover, combination of sources' estimates. Since each source can have different beliefs in reliability of remaining sources, the weights are assumed to be source-dependent (assigned with respect to the particular source). The following section covers the joint paper (Dedecius and Sečkárová, 2013b).

Then, we relax the assumption on source-dependent weights and exploit the proposed combination for combining estimates provided by sources. We discuss the conditions for both approaches to be comparable.

Part of this discussion is included in author's accepted contribution, see (Sečkárová, 2015).

6.1 Dynamic distributed estimation in exponential family

We address the dynamic distributed estimation of an unknown parameter of interest from noisy measurements by a diffusion network. Each node exchanges information on observations and estimates with its adjacent neighbors and incorporates it locally into its own statistical knowledge. This significantly improves the statistical properties and robustness of the estimation process under regular conditions (Cattivelli et al., 2008). Unlike the consensus algorithms and their variations, e.g., (Olfati-Saber et al., 2007), (Schizas et al., 2008), (Mosquera et al., 2010) and (Jadbabaie et al., 2012), the diffusion algorithms do not require multiple intermediate iterations between two subsequent measurements, see, e.g., (Cattivelli et al., 2008). Furthermore, (Tu and Sayed, 2012) show that the diffusion strategies can outperform the consensus strategies in dynamic environments.

The diffusion solutions are mostly least-squares (LS) oriented, for instance the diffusion least mean squares (LMS) (Lopes and Sayed, 2008),(Liu et al., 2012), recursive least squares (RLS) (Cattivelli et al., 2008) or the Kalman filter (Cattivelli and Sayed, 2010b). Although otherwise sound, they are strongly single-problem oriented and their reformulation for other tasks, e.g., non-LS oriented, is limited or even impossible by nature. The goal of this paper is to overcome this shortcoming. By exploiting the consistent theory of the Bayesian inference, we formulate a new dynamic diffusion estimation method in an abstract way, theoretically independent of a particular model type. The only assumption is its membership in the exponential family. Examples are the normal regression models, Poisson (shot noise) model, Bernoulli, Weibull, Pareto and many other models. We note that the dynamic estimation of a varying parameter coincides with the (Bayesian) parameter tracking. Since the proposed distributed estimation method is rooted in this realm, it is directly possible to use most of the elaborated Bayesian tracking methods, for instance forgetting, e.g., (Peterka, 1981), (Dedecius et al., 2012) and the references therein.

6.1.1 Bayesian estimation in exponential family

Consider discrete-time dynamic modelling of an observed variable y_t determined by an unknown fixed parameter θ and, if exists, a known explanatory variable (e.g., regressor) x_t, where $t = 1, 2, \dots$ are time indices. For the sake of generality, the variables are considered real, possibly multivariate and with compatible dimensions. From the probabilistic viewpoint, the model can be represented by a conditional probability density function (pdf) $f(y_t|x_t, \theta)$. Estimation of θ is based on the knowledge of past data $D_{t-1} = \{y_\tau, x_\tau\}_{\tau=1,\dots,t-1}$ and the prior pdf $\pi(\theta|D_0)$, obtained, e.g., from an expert, based on historical data, or alternatively being a flat noninformative pdf. The Bayesian approach to estimation recursively updates the prior pdf by new data $\{y_t, x_t\}$ via the Bayes' rule (Bernardo and Smith, 2009),

$$\pi(\theta|D_t) \propto f(y_t|x_t, \theta)\pi(\theta|D_{t-1}). \tag{6.1}$$

Here $\propto$ stands for proportionality, i.e., equality up to a normalizing factor. We call (6.1) the *sequential* variant. Equivalently, for time horizon t, the *batch* estimation reads

$$\pi(\theta|D_t) \propto \pi(\theta|D_0) \prod_{\tau=1}^{t} f(y_\tau|x_\tau, \theta). \tag{6.2}$$

Analytical tractability of recursions (6.1)–(6.2) is guaranteed if the model $f(y_t|x_t, \theta)$ is an exponential family distribution and the prior pdf is conjugate to it, as defined below (Bernardo and Smith, 2009) (with time indices dropped):

Definition 6.1 (Exponential family of distributions)**.** *An exponential family of distributions of a variable y with a parameter θ and an explanatory variable x is a family of distributions with pdf of the form*

$$f(y|x, \theta) = h(y, x)g(\theta) \exp\left[\eta(\theta)T(y, x)\right], \tag{6.3}$$

where $h(y,x)$ is a known function, $g(\theta)$ is a known normalization function, $\eta(\theta)$ is a natural parameter and $T(y,x)$ is a sufficient statistic.

Definition 6.2 (Conjugate prior pdf)**.** *A conjugate prior pdf for a parameter θ with the hyperparameters ξ of the same dimension as $T(y,x)$ and $\nu \in \mathbb{R}^+$ has the form*

$$\pi(\theta|\xi,\nu) = q(\xi,\nu)g(\theta)^\nu \exp\left[\eta(\theta)\xi\right], \tag{6.4}$$

where $q(\xi,\nu)$ is a normalization function and $g(\theta)$ has the same form as in the exponential family.

The dimension-preserving sufficient statistic accumulates all statistical knowledge necessary to compute an estimate of θ, regardless of the data sample size. It has the form $T(y_t,x_t)$ for the sequential variant (6.1), and $T(D_t) = \sum_{\tau=1}^{t} T(y_\tau,x_\tau)$ for the product in the batch variant (6.2). The sequential update modifies the prior hyperparameters as follows:

$$\begin{aligned} \xi_t &= \xi_{t-1} + T(y_t,x_t) \\ \nu_t &= \nu_{t-1} + 1, \end{aligned} \tag{6.5}$$

similar rules hold for (6.2). For simplicity, we stick with the sequential variant in the sequel. The modifications for the batch variant are straightforward.

The point estimate of θ can be obtained from the posterior pdf using standard formulas. Usually it is the mean value; sometimes the median or the mode are preferred. The estimation uncertainty is often expressed by the estimator variance.

6.1.2 Diffusion estimation

The diffusion network is an undirected connected graph of N spatially distributed nodes (e.g., sensors). Each node $i = 1,\dots,N$ can directly exchange information with adjacent nodes forming its closed neighborhood $\mathcal{N}_i$ of a cardinality n_i; also $i \in \mathcal{N}_i$. The exchanged information relates to (i) observations (adapt step) and (ii) estimates (combine step). Since the information from the nodes $j \in \mathcal{N}_i$ may have different credibility from the ith node's viewpoint, nonnegative relative weights summing to unity are used to reflect this.

Adapt step

Each network node $i = 1,\dots,N$ employs the same form of an exponential family model $f_i(y_t|x_{i,t},\theta)$ as above. Fixing i and t, we may regard $f_j(y_t|x_{j,t},\theta)$ for $j \in \mathcal{N}_i$ as a complete system of hypotheses about a true model at i. From the ith node's viewpoint, these are valid with probabilities c_{ij}, called weights, summing to unity due to the completeness. The Kullback-Leibler (KL) divergence (Bernardo and Smith, 2009) in the role of the loss function then provides the way to approach the true model by pdf f_i^* as follows:

Proposition 6.3. *Given pdfs f_j with weights $c_{ij}, j \in \mathcal{N}_i$, the best approximating pdf f_i^* optimal in the KL sense, minimizing the cumulative loss*

$$\sum_{j \in \mathcal{N}_i} c_{ij} \, \mathrm{D}\left(f_i^* \middle|\middle| f_j\right)$$

has the form

$$f_i^* \propto \prod_{j \in \mathcal{N}_i} f_j^{c_{ij}}.$$

Proof. By definition of the KL divergence

$$\begin{aligned}
&\sum_{j \in \mathcal{N}_i} c_{ij} \int_{y_t} f_i^*(y_t|\cdot) \log \frac{f_i^*(y_t|\cdot)}{f_j(y_t|\cdot)} \, \mathrm{d}y_t \\
&\quad = \int_{y_t} f_i^*(y_t|\cdot) \log \frac{f_i^*(y_t|\cdot)}{\prod_{j \in \mathcal{N}_i} f_j(y_t|\cdot)^{c_{ij}}} \, \mathrm{d}y_t \\
&\quad = \mathrm{D}\left(f_i^* \middle|\middle| \prod_{j \in \mathcal{N}_i} f_j^{c_{ij}}\right).
\end{aligned}$$

The minimum of the KL divergence is attained when its arguments agree. □

The KL-optimal model f_i^* is given by a geometric mean of available hypothetical models. The initial choice of exponential family models yields the appealing consequence of analytically tractable recursive diffusion update rules similar to (6.5). The Bayes' theorem (6.1) with $f = f_i^*$ updates the hyperparameters according to the following proposition.

Proposition 6.4 (Adapt-posterior pdf)**.** *Given sufficient statistics $T(y_{j,t}, x_{j,t}), j \in \mathcal{N}_i$, the adapt step updates the ith node's hyperparameters $\xi_{i,t-1}$ and $\nu_{i,t-1}$ as follows*

$$\begin{aligned}
\xi_{i,t} &= \xi_{i,t-1} + \sum_{j \in \mathcal{N}_i} c_{ij} T(y_{j,t}, x_{j,t}) \\
\nu_{i,t} &= \nu_{i,t-1} + 1.
\end{aligned} \tag{6.6}$$

The proof is trivial.

Remark. *The KL divergence is a well founded measure of pdfs' dissimilarity (Bernardo and Smith, 2009). The chosen zero-forcing order of its arguments brings the salient feature of analytically tractable computations in the exponential family due to the geometric mean, at the potential cost of variance underestimation. The alternative order (zero-avoiding divergence) would yield the arithmetic average of pdfs, raising computational issues and potential variance overestimation (Bishop, 2006).*

Combine step

The combine step follows the adapt step in order to further improve the statistical properties of individual estimators. We propose two principally different methods, one combining whole adapt-posterior pdfs, the other combining only the point estimates.

Whole adapt-posterior pdfs The ith node combines the adapt-posterior pdfs with the hyperparameters (6.6) of nodes $j \in \mathcal{N}_i$ in the KL-optimal sense prescribed by Proposition 6.3 (with the posterior pdfs π_j and weights a_{ij} in the roles of f_j and c_{ij}). The resulting combine-posterior pdf π_i^* has the following hyperparameters,

$$\begin{aligned} \xi_{i,t}^* &= \sum_{j \in \mathcal{N}_i} a_{ij} \xi_{j,t} \\ \nu_{i,t}^* &= \sum_{j \in \mathcal{N}_i} a_{ij} \nu_{j,t}. \end{aligned} \tag{6.7}$$

These hyperparameters completely characterize the distribution of θ. It is usually very easy to evaluate its moments, quantiles etc. using standard formulas. Furthermore, the combine-posterior pdf can serve as the prior for the next adapt step at $t+1$.

Point estimates If the ith node has access only to the point estimates provided by nodes $j \in \mathcal{N}_i$, for instance the means $\hat{\theta}_{j,t}$ and optionally the related variances $\gamma_{j,t}$, it is possible to directly combine them as follows:

$$\begin{aligned} \hat{\theta}_{i,t}^* &= \sum_{j \in \mathcal{N}_i} a_{ij} \hat{\theta}_{j,t} \\ \gamma_{i,t}^* &= \sum_{j \in \mathcal{N}_i} a_{ij} \left[\gamma_{j,t} + (\hat{\theta}_{j,t} - \hat{\theta}_{i,t}^*)^2 \right]. \end{aligned} \tag{6.8}$$

This approach, motivated by the mixture-based estimation (Frühwirth-Schnatter, 2006), is slightly computationally cheaper, because it avoids intermediate combination of pdfs. The adapt-posterior pdfs remain unmodified and enter $t+1$ as the prior.

Choice of weights

The purpose of weights a_{ij} and c_{ij} is to express the ith node's degree of belief in information from the nodes $j \in \mathcal{N}_i$. For fixed i, both a_{ij} and c_{ij} sum to unity. There exist several (mostly static) methods for their determination, some of them are given in Table 6.1, see, e.g., (Cattivelli and Sayed, 2010a) and references therein. An additional feature of the chosen probabilistic framework is the prospect of theoretically justified information-based methods for dynamic weights. For instance, it is possible to exploit local modelling and sharing of the observations/estimators variances at each node or to measure the fit of the data/estimates using the likelihoods. However, this issue is beyond the main message of the paper and will be addressed in the future.

Method	**Rule**
Uniform	$a_{ij} = 1/n_i$
Laplacian	$a_{ij} = 1/n_{\max}$
Maximum degree	$a_{ij} = 1/N$
Metropolis	$a_{ij} = 1/\max(n_i, n_j)$
Relative degree	$a_{ij} = n_j/\left(\sum_{k\in\mathcal{N}_i} n_k\right)$
Rel. degree-noise variance	$a_{ij} = n_j\sigma_j^2/\left(\sum_{k\in\mathcal{N}_i} n_k\sigma_k^2\right)$

Fig. 6.1: Weights before normalization. $n_i = \text{cardinality}(\mathcal{N}_i)$. The same weights can be used for c_{ij}.

6.1.3 Examples

Diffusion autoregression

Consider a Kth order autoregressive model

$$y_t = x_t'\beta = \sum_{k=1}^{K} y_{t-k}\beta_k + \varepsilon_t, \qquad (y_t \in \mathbb{R}^1),$$

where the explanatory variable $x_t = [y_{t-1}, \ldots, y_{t-K}]' \in \mathbb{R}^K$ is a known column regression vector, $\varepsilon_t \sim \mathcal{N}(0, \sigma^2)$ is additive white noise and $\beta = [\beta_1, \ldots, \beta_K]' \in \mathbb{R}^K$ is a column vector of unknown regression coefficients. Its estimation provides, among others, the least squares (LS) method via the normal equations; the recursive variant is RLS. The same point estimator follows from the Bayesian modelling with $y_t \sim \mathcal{N}(x_t'\beta, \sigma^2)$ and a normal prior distribution for the parameter $\theta \equiv \beta$; the associated uncertainty is an inherent part of the solution. We focus on a bit more complex normal inverse-gamma prior distribution $\mathcal{N}i\mathcal{G}(V, \nu)$, providing an additional advantage of variance estimation with $\theta \equiv \{\beta, \sigma^2\}$. Its hyperparameters standing in the roles of ξ and ν are the extended (symmetric) information matrix $V \in \mathbb{R}^{(K+1)\times(K+1)}$ and the degrees of freedom $\nu \in \mathbb{R}^+$ (Bernardo and Smith, 2009).

Let us demonstrate the ease of derivation of the diffusion estimator. The model pdf in the vector form reads

$$f(y_t|x_t, \theta) = \frac{\sigma^{-1}}{\sqrt{2\pi}} \exp\left\{-\frac{1}{2\sigma^2}\begin{bmatrix}-1\\ \beta\end{bmatrix}'\begin{bmatrix}y_t\\ x_t\end{bmatrix}\begin{bmatrix}y_t\\ x_t\end{bmatrix}'\begin{bmatrix}-1\\ \beta\end{bmatrix}\right\}.$$

A rearrangement of the terms according to (6.3) reveals the sufficient statistic connected with time t,

$$T(y_t, x_t) = \begin{bmatrix}y_t\\ x_t\end{bmatrix}\begin{bmatrix}y_t\\ x_t\end{bmatrix}'. \qquad (6.9)$$

Hence the update (6.5) of the $\mathcal{N}i\mathcal{G}$ hyperparameters takes the form

$$V_t = V_{t-1} + \begin{bmatrix}y_t\\ x_t\end{bmatrix}\begin{bmatrix}y_t\\ x_t\end{bmatrix}' \qquad \text{and} \qquad \nu_t = \nu_{t-1} + 1.$$

Recall that the autoregressive recursion begins with $t = K + 1$, imposing the initialization with $\nu_K = \nu_0$ and $V_K = V_0$. For $t \geq K+1$, the point estimators of β

and σ^2 are easily reachable after partitioning the matrix V into blocks (Peterka, 1981)

$$V_t \equiv \begin{bmatrix} V_{yy,t} & V'_{yx,t} \\ V_{yx,t} & V_{xx,t} \end{bmatrix}, \qquad V_{yy,t} \in \mathbb{R}^1.$$

Then

$$\hat{\beta}_t = V_{xx,t}^{-1} V_{yx,t} \quad \text{and} \quad \hat{\sigma}_t^2 = \frac{V_{yy,t} - V'_{yx,t} V_{xx,t}^{-1} V_{yx,t}}{\nu_t - K + 2}. \tag{6.10}$$

The diffusion estimator is as follows: The *adapt step* prescribed by Proposition 6.4 has the form

$$\begin{aligned} V_{i,t} &= V_{i,t-1} + \sum_{j \in \mathcal{N}_i} c_{ij} \begin{bmatrix} y_{j,t} \\ x_{j,t} \end{bmatrix} \begin{bmatrix} y_{j,t} \\ x_{j,t} \end{bmatrix}' \\ \nu_{i,t} &= \nu_{i,t-1} + 1. \end{aligned} \tag{6.11}$$

The *combine step* is a direct application of the prescribed rules, too. The first case, the *whole adapt-posterior pdfs* combination using (6.7) and (6.11) reads

$$\begin{aligned} V_{i,t}^* &= \sum_{j \in \mathcal{N}_i} a_{ij} V_{j,t} \\ \nu_{i,t}^* &= \sum_{j \in \mathcal{N}_i} a_{ij} \nu_{j,t}. \end{aligned}$$

The *point estimates* combination puts (6.10) into (6.8).

This diffusion autoregression (with the *point estimates* combine method) coincides with the diffusion RLS proposed by Cattivelli et al. (Cattivelli et al., 2008). Additionally, it provides the noise variance estimator, which can be potentially useful for dynamic determination of the relative degree-noise variance weights (Table 6.1). The notable benefit of the proposed Bayesian approach over the non-Bayesian one lies in the ease and straightforwardness of its application to a chosen problem while still completely retaining all theoretical consistency.

Homogeneous Poisson process

The homogeneous Poisson process (alias homogeneous shot noise) is a random process $\{M_t\}_{t \geq 0}$ of the counts $M_t \in \mathbb{N}_0$, starting with $M_0 = 0$ and with independent stationary Poisson distributed increments satisfying

$$\mathbb{P}\left[M_{t+\tau} - M_t = y_t \middle| \, \lambda\right] = \frac{(\lambda\tau)^{y_t} e^{-\lambda\tau}}{y_t!}, \quad \tau \in \mathbb{N}. \tag{6.12}$$

The *rate* parameter $\lambda \in \mathbb{R}^+$ coincides with the mean and variance of y_t. The process characterizes, e.g., the number of photons or other particles incident on a detector.

Considering the sequential variant with $\tau = 1$ and rewriting (6.12) to the form (6.3) reveals the sufficient statistic

$$T(y_t) = y_t.$$

The conjugate prior for $\theta \equiv \lambda$ is the gamma distribution $\mathcal{G}(\alpha, \beta)$ with shaping hyperparameters $\alpha, \beta > 0$ in the roles of ξ and ν, respectively. Their update (6.5) has the form (Bernardo and Smith, 2009)

$$\begin{aligned} \alpha_t &= \alpha_{t-1} + y_t \\ \beta_t &= \beta_{t-1} + 1. \end{aligned} \tag{6.13}$$

The point estimator of λ is well known to be $\hat{\lambda}_t = \alpha_t/\beta_t$ with the variance $\gamma_t = \alpha_t/\beta_t^2$.

Now we easily derive the diffusion estimator. The *adapt step* according to Proposition 6.4 reads

$$\begin{aligned} \alpha_{i,t} &= \alpha_{i,t-1} + \sum_{j \in \mathcal{N}_i} c_{ij} y_{j,t} \\ \beta_{i,t} &= \beta_{i,t-1} + 1. \end{aligned} \tag{6.14}$$

The *combine step* for *whole adapt-posterior pdfs* (6.7) reads

$$\begin{aligned} \alpha^*_{i,t} &= \sum_{j \in \mathcal{N}_i} a_{ij} \alpha_{j,t} \\ \beta^*_{i,t} &= \sum_{j \in \mathcal{N}_i} a_{ij} \beta_{j,t}. \end{aligned} \tag{6.15}$$

If only the combination of *point estimates* is required, then (6.8) with the above given point estimators is used.

Estimation of Bernoulli process proportions

This example studies the Bernoulli process exploited, e.g., in the queuing theory, reliability analysis and finance. It is a discrete-time stochastic process yielding a sequence of independent identically distributed binary random variables Y_t taking values 0 or 1 (failure or success). It follows the Bernoulli distribution,

$$\mathbb{P}(Y_t = y_t | p) = p^{y_t}(1-p)^{1-y_t}, \qquad y_t \in \{0, 1\},$$

where $p \in [0, 1]$ is the probability of success ($y_t = 1$). Clearly $T(y_t) = y_t$. The conjugate prior for unknown parameter $\theta = p$ is the beta distribution $\mathcal{B}(\alpha, \beta - \alpha)$ with the hyperparameters $\alpha, \beta > 0$ in the roles of ξ and ν, respectively. Their update (6.5) is

$$\begin{aligned} \alpha_t &= \alpha_{t-1} + y_t \\ \beta_t &= \beta_{t-1} + 1. \end{aligned} \tag{6.16}$$

The point estimator is known to be $\hat{p}_t = \alpha_t/\beta_t$ with the variance $\gamma_t = \alpha_t(\beta_t - \alpha_t)/[\beta_t^2(\beta_t + 1)]$.

Note the appealing fact arising from the Bayesian estimation of exponential family models with conjugate priors: the recursions (6.13) and (6.16) are identical although the underlying distributions are not. The diffusion estimation *adapt* and *combine* steps would accordingly agree with (6.14) and (6.15) (or the combination of *point estimates* (6.8)).

6.2 Cross-entropy based combination in diffusion estimation

Let us consider the dynamic diffusion estimation (Section 6.1.2) when the underlying probability distribution of y_t is categorical with n possible categories. In such case the parameter θ and provided estimates $\hat{\theta}_j$, $j = 1, \ldots, s$, coincide with an n-dimensional pmf. For their combination we can exploit the results from Chapter 3. There, the unknown probability vector was denoted by $q = (q_1, \ldots, q_n)$ and the sources' opinions (estimates) were denoted by $p_j = (p_{j1}, \ldots, p_{jn})$, $j = 1, \ldots, s$.

Using this notation the estimate of $\hat{q}$ of q based on $p_1, \ldots, p_s$ has according to (6.8) the following form:

$$\hat{q}_i^* = \sum_{j=1}^{s} a_j p_{ji}, \quad i = 1, \ldots, n, \tag{6.17}$$

where the original weights a_{ij} were displaced by their source-independent versions. Examples of weights a_{ij} (and also a_j) are given in Table 6.1.

We now suggest to exploit the combination (3.3). Since we assumed that prior pmf p_0 is a function of current observations, i.e., their weighted arithmetic mean, we can rewrite the optimal combination $\hat{q}$ similarly to (6.17)

$$\hat{q}_i = \sum_{j=1}^{s-1} (w_{0j} + \lambda_j) p_{ji} + \left(w_{0s} - \sum_{j=1}^{s-1} \lambda_j \right) p_{si}, i = 1, \ldots, n. \tag{6.18}$$

We now compare the optimal combination of pmfs proposed in the thesis and the combination of point estimates based on (6.7) with uniform weights.

Suppose we have 5 sources/nodes ($s = 5$) providing 3-dimensional probability vectors ($n = 3$). Following the idea in Section 5.2 at each time instant we generate data from Dirichlet distribution with parameters

$$(\nu_1, \nu_2, \nu_3) = (0.2, 0.1, 0.7).$$

To obtain the optimal combination we again exploit the dynamic approach described in Section 3.3.2. Resulting combinations of simulated pmfs based on proposed approach (6.18) and DDE (6.17) (with time independent uniform weights: $a_j = 1/s$, $j = 1, \ldots, s$) are shown in the Fig. 6.2 on the left.

The results in the case when the third source is corrupted and its pmfs are simulated from Dirichlet distribution with parameters

$$(\nu_1, \nu_2, \nu_3) = (0.6, 0.2, 0.2),$$

are shown in the Fig. 6.2 on the right.

We see that in the first case (no corrupted sources), the optimal combination is from the beginning stabilized and tends to pmf as flat as possible. In case the third source was corrupted, the resulting combination $\hat{q}$ in (6.18) and combination exploiting combine step of DDE given in (6.17) behave similarly.

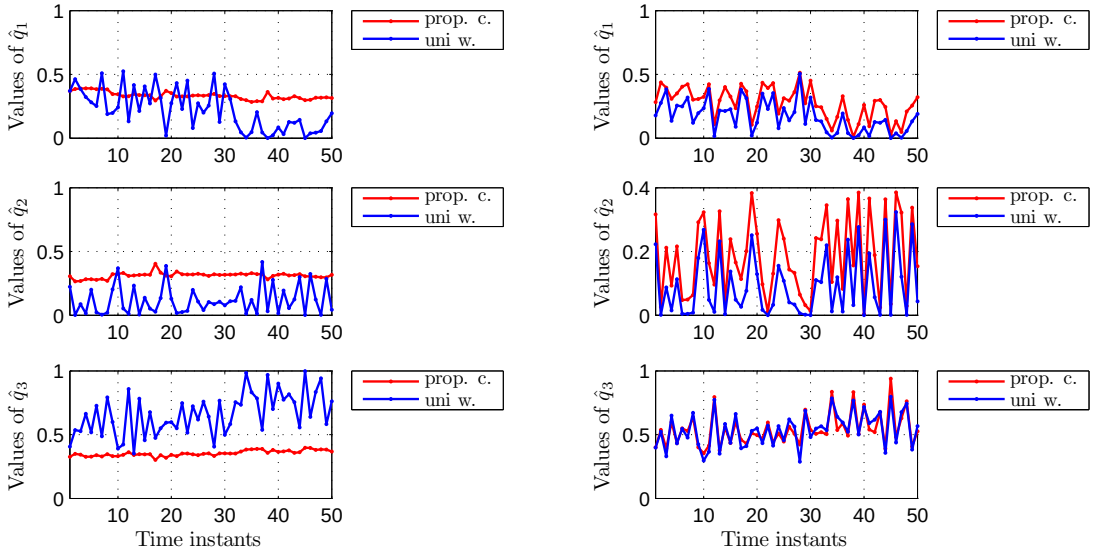


Fig. 6.2: Combinations obtained by the proposed combining and by DDE with uniform weights. On the left: no corrupted sources. On the right: the third source is corrupted.

Chapter 7

Conclusions and future work

The goal of this work was to derive an approach to combine information in distributed decision making (DM) within *wise selfish* cooperative scenario in the group of sources. We focused on the case when sources provided their *opinions* about a stochastic phenomenon. Such situation commonly occurs when sources can not directly observe the process of interest, such as evaluation of the contracts for decision making in companies, elections, etc.

Novelty of proposed combining

In majority of developed approaches within the considered area the final decision, the *compromise*, is constructed in order to serve the whole group, not reflecting sources' personal aims and limited abilities. This motivated us to develop a new approach for combining sources' opinions, which can serve involved individuals.

To the author of this thesis the only known treatment of this issue is a probabilistic one introduced by Kárný et al. (2009). We also consider probabilistic approach; we assume sources provide their opinions about the studied problem as pmfs. Unlike the modeling adopted in Kárný et al. (2009), leading to the combination, which does not reduce to the Bayes rule when possible, we suggest constraints on the acceptance of the desired combination to delimit the set of possible combinations.

The advantage of working with pmfs lies in the treatment of probabilistic and non-probabilistic types of information. The proposed method then represents a unified way how to cover spectrum of sources' characteristics for easy and time effective DM. The construction of the compromise (optimal combination) via the decision making theory with axiomatically justified tools (selected loss, way of extending information) is far from being standard and represents methodological contribution of the thesis.

Brief summary of the thesis

In order to obtain the optimal combination of given pmfs, we expressed the combination as an unknown pmf of a discrete random vector with finite number of outcomes. We then searched for its optimal estimator by exploiting the Bayesian decision making, the minimum cross-entropy principle and the Kullback-Leibler

(KL) divergence (cross-entropy). Under general setup, we obtained the estimator as the conditional expectation with respect to unspecified conditional pdf. Next, we exploited computationally advantageous case when the conditional pdf was the pdf of the Dirichlet distribution. This assumption yielded a weighted linear combination of sources' pmfs and we studied the properties of this combination such as influence of the duplicate data, extension to the dynamic case and influence of sources' preferences. After the derivation we focused on handling

- partial probabilistic information with non-common supports: conditional and marginal pmfs describing the discrete random vector,
- non-probabilistic information: subsets of outcomes and/or expected values of the discrete random vector,

and treated these types of information by the proposed extension and transformation. This treatment is still non-standard and its elaboration to the level allowing to solve variety of applied problems is one of the contributions of this thesis. This was followed by the application to real data for decision making in the company, galaxy identification and social survey. The last part of this thesis was dedicated to another combining method in the area of parameter estimation, also based on the KL-divergence, see Dedecius and Sečkárová (2013b), and its comparison with the proposed combination.

Future work

A diffusion alternative of proposed combining

Although throughout the thesis we assumed the number of sources is finite, the proposed combination can be also applied to large sets of sources by decomposing them into finite (possibly overlapping) groups.

Treatment of implicit equations

In Section 4.1 we suggested extension for obtained conditional/marginal pmfs based on appropriate versions of the optimal combination $\hat{q}$ yielding implicit equations. In the thesis we solved this issue for dynamic scenario by using appropriate versions of $\hat{q}_{t-1}$ instead of $\hat{q}_t$ at time instant t (Section 3.3.2). A solution closely related to the original implicit relation with low computational difficulty is of interest.

Alternative of proposed combining for continuous case

Since we developed a technique to combine pmfs, we are naturally interested in its extension to the continuous case, where sources provide pdfs. A wider use is foreseen, e.g., in mixtures of Kalman filters.

Combination of preferences

The proposed approach in combination with fully probabilistic design (FPD), see (Kárný et al., 2007) and (Kárný and Kroupa, 2012), allows combination of decision objectives within a group. A systematic elaboration requires further research.

Bibliography

European Social Survey (2014). ESS 1-6, European Social Survey Cumulative File, Study Description. Bergen: Norwegian Social Science Data Services.

European Social Survey Cumulative File, ESS 1-6 (2014). Data file edition 1.0. Norwegian Social Science Data Services, Norway – Data Archive and distributor of ESS data. Production date of cumulative dataset: 26.11.2014.

Abbas, A. E. (2009). A Kullback-Leibler View of Linear and Log-Linear Pools. *Decision Analysis 6*(1), 25–37.

Abramowitz, M. and I. Stegun (1964). *Handbook of Mathematical Functions: With Formulas, Graphs, and Mathematical Tables*. Applied Mathematics Series. Dover Publications.

Basu, A., A. Mandal, N. Martin, and L. Pardo (2013). Testing statistical hypotheses based on the density power divergence. *Annals of the Institute of Statistical Mathematics 65*(2), 319–348.

Bernardo, J. and A. Smith (2009). *Bayesian Theory*. Wiley Series in Probability and Statistics. Wiley.

Bernardo, J. M. (1979). Expected information as expected utility. *Ann. Stat. 7*, 686–690.

Bishop, C. (2006). *Pattern Recognition and Machine Learning*. Information Science and Statistics. Springer.

Boyd, S. and L. Vandenberghe (2004). *Convex Optimization*. Cambridge University Press.

Campenhout, J. V. and T. Cover (1981). Maximum Entropy and Conditional Probability. *IEEE Tran. on Inf. Theory 27*(4), 483–489.

Carli, R., A. Chiuso, L. Schenato, and S. Zampieri (2008, May). Distributed kalman filtering based on consensus strategies. *Selected Areas in Communications, IEEE Journal on 26*(4), 622–633.

Cattivelli, F., C. Lopes, and A. Sayed (2008, May). Diffusion Recursive Least-Squares for distributed estimation over Adaptive Networks. *Signal Processing, IEEE Transactions on 56*(5), 1865–1877.

Cattivelli, F. and A. Sayed (2010a, March). Diffusion LMS Strategies for Distributed Estimation. *Signal Processing, IEEE Transactions on 58*(3), 1035–1048.

Cattivelli, F. and A. Sayed (2010b, Sept). Diffusion Strategies for Distributed Kalman Filtering and Smoothing. *Automatic Control, IEEE Transactions on 55*(9), 2069–2084.

Cover, T. M. and J. A. Thomas (2006). *Elements of Information Theory (Wiley Series in Telecommunications and Signal Processing)*. Wiley-Interscience.

Dedecius, K., I. Nagy, and M. Kárný (2012). Parameter tracking with partial forgetting method. *International Journal of Adaptive Control and Signal Processing 26*(1), 1–12.

Dedecius, K. and V. Sečkárová (2013a). Centralized Bayesian reliability modelling with sensor networks. *Mathematical and Computer Modelling of Dynamical Systems 19*(5), 471–482.

Dedecius, K. and V. Sečkárová (2013b). Dynamic Diffusion Estimation in Exponential Family Models. *IEEE Signal Process. Lett. 20*(11), 1114–1117. Accepted also for presentation on IEEE ICASSP 2014.

DeGroot, M. H. (1974). Reaching a consensus. *Journal of the American Statistical Association 69*(345), pp. 118–121.

Ferguson, T. S. (1973, 03). A Bayesian Analysis of Some Nonparametric Problems. *Ann. Statist. 1*(2), 209–230.

Fischer, A. (2010). Quantization and clustering with bregman divergences. *Journal of Multivariate Analysis 101*(9), 2207 – 2221.

Frühwirth-Schnatter, S. (2006). *Finite Mixture and Markov Switching Models: Modeling and Applications to Random Processes*. Springer Series in Statistics. Springer.

Garcia, M. and D. Puig (2004). Robust aggregation of expert opinions based on conflict analysis and resolution. In R. Conejo, M. Urretavizcaya, and J.-L. Pérez de la Cruz (Eds.), *Current Topics in Artificial Intelligence*, Volume 3040 of *Lecture Notes in Computer Science*, pp. 488–497. Springer Berlin Heidelberg.

Guiaşu, S. (1971). Weighted entropy. *Reports on Mathematical Physics 2*(3), 165 – 179.

Jadbabaie, A., P. Molavi, A. Sandroni, and A. Tahbaz-Salehi (2012). Non-Bayesian social learning. *Games and Economic Behavior 76*(1), 210 – 225.

Jaynes, E. T. (1957, May). Information Theory and Statistical Mechanics. *Phys. Rev. 106*, 620–630.

Kárný, M. (2014). Approximate Bayesian recursive estimation. *Information Sciences 289*, 100–111. DOI 10.1016/j.ins.2014.01.048.

Kárný, M. and T. V. Guy (2004). On dynamic decision-making scenarios with multiple participants. In J. Andrýsek, M. Kárný, and J. Kracík (Eds.), *Multiple Participant Decision Making*, Adelaide, pp. 17–28. Advanced Knowledge International.

Kárný, M., T. V. Guy, A. Bodini, and F. Ruggeri (2009). Cooperation via sharing of probabilistic information. *Int. J. of Computational Intelligence Studies 1*(5), 139–162.

Kárný, M., J. Kracík, and T. Guy (2007). Cooperative decision making without facilitator. In B. R. Andrievsky and A. L. Fradkov (Eds.), *IFAC Workshop "Adaptation and Learning in Control and Signal Processing" /9./*. IFAC.

Kárný, M. and T. Kroupa (2012). Axiomatisation of fully probabilistic design. *Information Sciences 186*(1), 105–113.

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

Kißlinger, A.-L. and W. Stummer (2013). Some Decision Procedures Based on Scaled Bregman Distance Surfaces. In F. Nielsen and F. Barbaresco (Eds.), *Geometric Science of Information*, Volume 8085 of *Lecture Notes in Computer Science*, pp. 479–486. Springer Berlin Heidelberg.

Kotz, S., N. Balakrishnan, and N. L. Johnson (2005). *Dirichlet and Inverted Dirichlet Distributions*, pp. 485–527. John Wiley & Sons, Inc.

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

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

Lintott, C. J., K. Schawinski, A. Slosar, K. Land, S. Bamford, D. Thomas, M. J. Raddick, R. C. Nichol, A. Szalay, D. Andreescu, P. Murray, and J. Vandenberg (2008). Galaxy Zoo: morphologies derived from visual inspection of galaxies from the Sloan Digital Sky Survey. *Monthly Notices of the Royal Astronomical Society 389*(3), 1179–1189.

Liu, Y., C. Li, W. K. Tang, and Z. Zhang (2012). Distributed estimation over complex networks. *Information Sciences 197*(0), 91 – 104.

Lopes, C. and A. Sayed (2008, July). Diffusion least-mean squares over adaptive networks: Formulation and performance analysis. *Signal Processing, IEEE Transactions on 56*(7), 3122–3136.

Mosquera, C., R. Lopez-Valcarce, and S. Jayaweera (2010, Feb). Stepsize sequence design for distributed average consensus. *Signal Processing Letters, IEEE 17*(2), 169–172.

Olfati-Saber, R., J. Fax, and R. Murray (2007, Jan). Consensus and cooperation in networked multi-agent systems. *Proceedings of the IEEE 95*(1), 215–233.

Peterka, V. (1981). Bayesian approach to system identification. In P. Eykhoff (Ed.), *Trends and Progress in System Identification*, pp. 239–304. Oxford: Pergamon Press.

Raïffa, H. and R. Schlaifer (1961). *Applied statistical decision theory*. Studies in managerial economics. Division of Research, Graduate School of Business Adminitration, Harvard University.

Rufo, M. J., J. Martín, and C. J. Pérez (2012, 06). Log-linear pool to combine prior distributions: A suggestion for a calibration-based approach. *Bayesian Anal. 7*(2), 411–438.

Sabolová, R., V. Sečkárová, J. Dušek, and M. Stehlík (2015). Entropy based statistical inference for methane emissions released from wetland. *Chemometrics and Intelligent Laboratory Systems 141*(1), 125 – 133.

Savage, L. (1972). *The Foundations of Statistics*. Dover Books on Mathematics Series. Dover Publications.

Schizas, I., A. Ribeiro, and G. Giannakis (2008, Jan). Consensus in Ad Hoc WSNs With Noisy Links - Part I: Distributed Estimation of Deterministic Signals. *Signal Processing, IEEE Transactions on 56*(1), 350–364.

Schneeweiss, C. (2003). *Distributed Decision Making*. Springer.

Sečkárová, V. (2013). On Supra-Bayesian Weighted Combination of Available Data Determined by Kerridge Inaccuracy and Entropy. *Pliska Stud. Math. Bulgar. 22*, 159–168.

Sečkárová, V. (2015). Minimum cross-entropy based weights in dynamic diffusion estimation in exponential family. *Accepted to Pliska Studia Mathematica Bulgarica*.

Shannon, C. (1948, July). A mathematical theory of communication. *Bell System Technical Journal, The 27*(3), 379–423.

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

Tu, S.-Y. and A. Sayed (2012, Dec). Diffusion strategies outperform consensus strategies for distributed estimation over adaptive networks. *Signal Processing, IEEE Transactions on 60*(12), 6217–6234.

Tu, S.-Y. and A. Sayed (2014, March). Distributed decision-making over adaptive networks. *Signal Processing, IEEE Transactions on 62*(5), 1054–1069.

Villena, S., M. Vega, S. Babacan, R. Molina, and A. Katsaggelos (2010, Sept). Using the Kullback-Leibler divergence to combine image priors in super-resolution image reconstruction. In *Image Processing (ICIP), 2010 17th IEEE International Conference on*, pp. 893–896.

von Neumann, J. and O. Morgenstern (2007). *Theory of Games and Economic Behavior (60th Anniversary Commemorative Edition)*. Princeton Classic Editions. Princeton University Press.

Wagenpfeil, J., A. Trachte, T. Hatanaka, M. Fujita, and O. Sawodny (2009, June). Distributed decision making for task switching via a consensus-like algorithm. In *American Control Conference, 2009. ACC '09.*, pp. 5761–5766.

West, M. (1984). Bayesian aggregation. *Journal of the Royal Statistical Society (Ser. A) 147*, 600–607.

Wisse, B., T. Bedford, and J. Quigley (2008). Expert judgement combination using moment methods. *Rel. Eng. & Sys. Safety 93*(5), 675–686.

Wu, S., Y. Bi, X. Zeng, and L. Han (2009, July). Assigning appropriate weights for the linear combination data fusion method in information retrieval. *Inf. Process. Manage. 45*(4), 413–426.

Xu, J., C. Tekin, S. Zhang, and M. van der Schaar (2015, May). Distributed multi-agent online learning based on global feedback. *Signal Processing, IEEE Transactions on 63*(9), 2225–2238.

List of Figures