# A NOTE ON WEIGHTED COMBINATION METHODS FOR PROBABILITY ESTIMATION

Vladimíra Sečkárová*

Institute of Information Theory and Automation of the ASCR,
Pod Vodárenskou věží 4, CZ-182 08 Prague 8, Czech Republic
seckarov@utia.cas.cz

Department of Probability and Mathematical Statistics,
Charles University in Prague, Czech Republic

**Abstract.** To successfully learn from the information provided by available information sources, the choice of automatic method combining them into one aggregate result plays an important role. To respect the reliability in the source's performance each of them is assigned a weight, often subjectively influenced. To overcome this issue, we briefly describe the method based on Bayesian decision theory and elements of information theory. In particular we consider discrete-type information, represented by probability mass functions (pmfs) and obtain an aggregate result, which has also form of pmf. This result of decision making process is found to be a weighted linear combination of available information. Besides the brief description of the novel method, the paper focuses on its comparison with other combination methods. Since we consider the available information and unknown aggregate as pmfs, we mainly focus on the case when the parameter of binomial distribution is of interest and the sources provide appropriate pmfs.

**Keywords:** weighting methods, parameter estimation, Kerridge inaccuracy, maximum entropy principle, binomial distribution

## 1 Introduction

Exploiting available information plays an important role in many parts of mathematics such as parameter estimation, quality control, etc. and their applications. Usually, the processed information originates in many sources. The sources of information can range from experts in particular field to sensors measuring physical variables. To obtain the reliable result of interest based on these data we need to assign each source a weight. This weight should express reliability of a particular source and is usually assigned by an extra expert, thus is subjectively

* This research has been partially supported by GACR 13-13502S and SVV-2013 - 267 315.

influenced. It is worthwhile, especially in complex situations, to prevent the subjectivity. This paper focuses on the objective choice of weights under several commonly acceptable assumptions.

Throughout the paper we assume the sources provide the information about a common random vector having finite amount of realizations. The probability distribution over this random vector depends on an unknown, generally multidimensional, parameter, representing the ideal aggregated information. Our aim, the parameter estimate based on available data, will then be a combination of these data. Useful survey on combination methods from the mathematical and behavioural point of view can be found in [4]. To obtain the aggregate we express the parameter estimation task as a task of decision making and exploit the basic steps of Bayesian decision theory (see e.g. [6], Section VIIID) to compose an optimal decision. The optimality criterion is based on the minimization of an expected loss. The specific loss function we adopt is based on the elements of information theory. The parameter of interest is assumed to be a probability mass function (pmf), i.e. a column vector whose non-negative elements sum to unity, and the data provided by information sources are in the form of pmfs, too. A nice survey on the combination methods using elements of information theory can be found in [1], describing approaches with weights more or less subjectively influenced. To eliminate the subjective influence, we work with the method based on the Kerridge inaccuracy [8] and maximum entropy principle [14], which leads to a final weighted combination with weights determined during the construction of this combination [13]. The weights in the final combination are based on the information included in provided data and thus no subjective influence is added to them.

The main goal of this paper is to compare the proposed method with other methods in the considered field. We focus on the estimation of the parameter in binomial distribution. The methods serving to comparison belong to the group of empirical Bayes methods [11], where the prior distribution is computed from the previous observations.

The paper is organized as follows: the next section provides a brief description of our approach based on Bayesian decision theory and elements of information theory, the third section provides an overview of methods used for comparison, the fourth section gives the resulting estimates obtained by the considered methods based on the same data sample. The fifth section contains conclusion and topics for the future work. The previously published derivations connected with our approach (see [13]) can be found in the Appendix.

## 2 Proposed Method

In this section we briefly introduce a method combining the available information based on elements of information theory. The motivation for this particular choice is based on the aim of elimination of the subjectivity in the weights. For the weights' assignments we focus on the natural part included in provided data, i.e. we exploit the amount of information included in them. This of course requires specific setup described in the following paragraphs.

Let us start with a parameter estimation task, where the parameter $h$ has the form of pmf, belonging to the probabilistic simplex. To obtain its estimate $\hat{h}$ we exploit Bayesian decision theory and look for the estimate minimizing the expected loss function. We select the loss function as a function computing the inaccuracy between a pair of pmfs - the Kerridge inaccuracy $\mathrm{K}(.,.)$ (see [3]). The estimate $\hat{h}$ then coincides with the conditional expectation $\mathrm{E}[.|.]$ with respect to the posterior probability density function (pdf) $\pi(h|D)$ of the unknown parameter $h$ (an optimal aggregate) conditioned on available data $D = (g_1, \ldots, g_s)^T$ formed by pmfs $g_j$ given by $s$ sources, $s < \infty$:

$$\begin{aligned} \hat{h} &= \arg\min_{\tilde{h}\in\widetilde{H}} \mathrm{E}_{\pi(h|D)}[\mathrm{K}(h,\tilde{h})|D] \\ &= \arg\min_{\tilde{h}\in\widetilde{H}} \mathrm{K}(\mathrm{E}_{\pi(h|D)}[h|D],\tilde{h}) = \mathrm{E}_{\pi(h|D)}[h|D], \end{aligned} \tag{1}$$

where $\widetilde{H} = \left\{(\tilde{h}(x_1),\ldots,\tilde{h}(x_n)) : \sum_{i=1}^n \tilde{h}(x_i) = 1, \tilde{h}(x_i) > 0,\ i = 1\ldots,n\right\}$ and $g_j \in \widetilde{H}$, $j = 1,\ldots,s$. Sources describe a common random vector $X$ having possible outcomes $\{x_i\}_{i=1}^n$, $n < \infty$, i.e. provide the probabilities $g_j(x_i) = P_j(X = x_i)$, $j = 1,\ldots,s$.

To compute the estimate (1) we need to determine the posterior pdf $\pi(h|D)$, which is yet unknown. To determine its form, we exploit the maximum entropy principle [14]. It leads to the convex optimization:

$$\hat{\pi}(h|D) = \arg\min_{\widetilde{\pi}(h|D)\in\mathrm{M}} \left[\int_H \widetilde{\pi}(h|D)\log\widetilde{\pi}(h|D)\mathrm{d}h\right], \tag{2}$$

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

The constraints in M express the assumption the $j^{th}$ source will accept $h$ as a compromise (optimal aggregate) if it serves as a good approximation of $j^{th}$ pmf. According to [3] we expect that from the Bayesian point of view the Kerridge inaccuracy employed in the set M should reach low values for good approximations. The optional scalar $\beta_j(D)$ reflects "tolerance" of $j^{th}$ source to accept $h$ as an approximation of its opinion $g_j$.

The optimization results in pdf of Dirichlet distribution (see Subsection 6.1) and the final point estimate $\hat{h}$ of $h$ is the expected value of this distribution and has the form of a weighted combination of given pmfs $g_j$ (see Subsection 6.3):

$$\mathrm{E}_{\hat{\pi}(h|D)}(h(x_i)|D) = \hat{h}(x_i) = \lambda_0^*(D) + \sum_{j=1}^s \lambda_j^*(D) g_j(x_i), \quad i = 1,\ldots,n, \tag{3}$$

where

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

The last step of the described method involves the computation of the weights, which heavily depends on evaluation of Kuhn-Tucker multipliers $\lambda_j(D) \geq 0$ arising in the optimization task (2). The straightforward derivation of the multipliers can be found in Subsection 6.2. The problem, which has not been solved yet, is that the multipliers still depend on upper bounds $\beta_j(D)$ for expected Kerridge inaccuracies in (2). Here, we leave each $\beta_j(D)$ free and inspect the behaviour of the estimator (3) as a function of corresponding $\lambda_j(D)$, $j = 1, \ldots, s$. A promising objective solution is being elaborated in Section 4 and suggests the upper bounds (linearly shifted - see (9)) as the mean Kerridge inaccuracy in the following form:

$$\beta_k^*(D) = \frac{\sum_{j=1}^{s} \mathrm{K}(g_k, g_j)}{s} = \mathrm{K}(g_k, h_{\mathrm{data}}), \quad k = 1, \ldots, s, \tag{4}$$

where $h_{\mathrm{data}}(x_i) = \sqrt[s]{\prod_{j=1}^{s} g_j(x_i)}, \quad i = 1, \ldots, n$, i.e. geometric pool of opinions is taken as an aggregate acceptable by all information sources. Thus we can see, that the weights exploit the information captured in the provided data. Since they depend on the choice of the upper bounds $\beta_j(D)$, we study this connection in Section 4.

## 3 Empirical Bayes Methods

In this section we briefly introduce the methods we use for comparison in Section 4. As mentioned earlier, we try to avoid a subjective influence on the weights used in the aggregation process. The first idea, when there is no extra expert assigning the weights to available sources, is to use the equal weights. This approach can be found e.g. in [1].

Since equal weights do not reflect the reliability of the sources at all, we focus on different type of methods, namely, methods exploiting previously obtained data. The next section will bring the comparison of our method with group of empirical Bayes methods. These methods exploit the Bayes' formula in order to get the estimate of an unknown parameter. Their advantage lies in using the prior distribution based on the data available from the previous time instants, rather than choosing a specific prior distribution. When the prior information enters the Bayes' formula the final estimate is a weighted combination of available data.

The empirical Bayes methods are well-applicable in case when the estimation of a multidimensional parameter being a pmf is of interest. We consider four different approaches in this field, i.e. Griffin-Krutchkoff's estimator [7], Copas' second estimator [5], Lemon's estimator [10] and smooth incomplete beta estimator [12]. Formulas belonging to the mentioned estimators, using a common notation, can be found in [12]. Griffin-Krutchkoff's estimator provide a linear optimal estimator, where the optimality origins in minimizing the risk based on squared error loss. Similar situation considered Copas proposing an estimator, which is again assumed to be linear, it minimizes mean squared loss and guarantees a minimax estimate. Lemon's estimator uses a mean value of specifically chosen functions reflecting the current and previous data as posterior pmf. In particular the estimator focuses on a conditional probability of the modelled variable conditioned by the unknown parameter while plugging in some estimate of this parameter. All three methods can be easily applied to estimation of the parameter of binomial distribution. Finally, we inspect a smooth incomplete beta estimator, derived particularly for the case of binomial distribution. Here, it is suggested to use a function based on incomplete beta function. The parameter of interest and the final estimate have both form of pmf. In all cases the resulting pmf is a weighted combination of available data, thus these methods are the perfect choice for comparison with our method in (3).

The difference between the above group of methods and method in (3) is in the approach to the available data. To obtain the final estimate the former use the empirical prior distribution based on the previous observations. The latter does not use any prior information and combines data pieces at once.

## 4 Comparison

In this section a comparison of the proposed method and empirical Bayes methods is given. Assume we are interested in estimation of the probability $p \in (0,1)$ of success and the probability of failure $1-p$ in $N$ independent trials modeled by binomial distribution.

### 4.1 Illustrative Example

Thus in the case of the empirical Bayes methods we are looking for the estimate $\hat{p}$ (at the same time for the estimate $1-\hat{p}$) of an unknown parameter $p$ of random variable $Y$ distributed according to $Bi(N,p)$. The probability of $k$ successes in $N$ independent trials is then $P(Y=k) = \binom{N}{k} p^k (1-p)^{N-k}$. Let us assume we observe $N$ trials at $s$ time instants. Each time we obtain the number of successes $y_j$ and failures $N-y_j$. To obtain the aggregate $\hat{p}$ in empirical Bayes methods we use the binomial fractions $y_j/N$, which can be viewed as empirical estimates of $p$. We can then also get the estimate of probability of failures simply by computing $1-\hat{p}$.

To apply the method proposed in (3) we take a look at the considered situation from a different perspective. The unknown probability of success and failure form a 2-dimensional unknown parameter $h$. Let $X$ denote a random variable having two realizations ($n=2$), namely, $\{x_1, x_2\}$ ={success,failure} and thus $h = (h(x_1), h(x_2))^T$. Also assume we have $s$ sources providing pmfs $g_j$, $j = 1, \ldots, s$. We realize that according to the notation in the previous paragraph we have

$$\begin{aligned} h =&(h(x_1), h(x_2))^T = (p, 1-p)^T \\ \hat{h} =&(\hat{h}(x_1), \hat{h}(x_2))^T = (\hat{p}, 1-\hat{p})^T \\ g_j =&(g_j(x_1), g_j(x_2))^T = (y_j/N, (N-y_j)/N)^T, \quad j = 1, \ldots, s. \end{aligned}$$

Now we can focus on how the previously mentioned methods work. We generate four random values, number of successes, from $Bi(10, 1/3)$. That is, we have $s = 4$ and for each $j = 1, \ldots, 4$ we can compute the binomial fractions $y_j/10$ and their counterparts $(10 - y_j)/10$ to get pmfs $g_j$. The data are the following:

$$D = \begin{pmatrix} \left(\frac{y_1}{10}, \frac{10-y_1}{10}\right) \\ \cdots \\ \cdots \\ \left(\frac{y_4}{10}, \frac{10-y_4}{10}\right) \end{pmatrix} = \begin{pmatrix} (g_1(x_1), g_1(x_2)) \\ \cdots \\ \cdots \\ (g_4(x_1), g_4(x_2)) \end{pmatrix} = \begin{pmatrix} (0.3, 0.7) \\ (0.4, 0.6) \\ (0.2, 0.8) \\ (0.1, 0.9) \end{pmatrix}$$

The upper picture in Fig. 1 shows the behaviour of the Kuhn-Tucker multipliers $\lambda_j(D)$ depending on the values of $\beta^*_j(D)$, $j = 1, \ldots, 4$, see (4). We decrease linearly shifted bounds $\beta^*_j(D)$ (see (9)) as follows:

$$\beta^*_{j,l}(D) = \beta^*_j(D) \times (0.85 - l \times 0.0084) \text{ for instants } l = 1, \ldots, 100, \qquad (5)$$

$$\beta^*_{j,1}(D) = \mathrm{K}(g_j, h_{\mathrm{data}}). \qquad (6)$$

In case of dynamic setup, where with each time point we obtain a new data, the empirical Bayes methods update the estimate by new data. In case of our method, the data $g_{j,t}$ in (3) can be viewed as the estimate given by $j^{th}$ source based on its data up to time point $t$. In the next time step, a new estimate $g_{j,t+1}$ is obtained and again, formula (3) is used to combine all available $g_{j,t+1}$, $j = 1, \ldots, s$.

The bottom picture in Fig. 1 brings the final estimate (final aggregate) $\hat{h}$ computed by method in (3) with changing value of $\lambda_j(D)$. Also the source with the highest and the lowest entropy are drawn.

The resulting estimates $\hat{h}$ are given in the Fig. 2. They were obtained using equal weights (EW) and the following methods: Griffin-Krutchkoff's estimator (GK), Copas' second estimator (Co), Lemon's estimator (Le) and a smooth incomplete beta estimator (BE), all briefly introduced in the previous section. We can see that even under a small number of available data our method and Copas' second estimator performed quite well regarding the information, that the data were drawn from binomial distribution with probability of success equal to 1/3 (probability of failure is thus 2/3). In particular, Copas' estimator coincides with the estimator based on equal weights, the mean value of drawn data. The results obtained from Lemon's, Griffin-Krutchkoff's estimators and smooth incomplete beta estimator differ from the true value of $h = (h(x_1), h(x_2))^T$, but are closer to what we would naturally expect from obtained data, which can be misleading for small sample cases. A case with larger sample is studied in Subsection 4.2.

At the end we note that while the empirical Bayes methods exploited the fact that we focus on binomial distribution, our method (3) do not need the information about the original distribution to obtain $\hat{h}$. This predestinates our method to be a normative method for estimation of pmfs.

### 4.2 Monte Carlo Simulations

In this subsection we study the behaviour of considered methods in Monte Carlo study. We assume the same setup as introduced in Subsection 4.1, thus we gen-

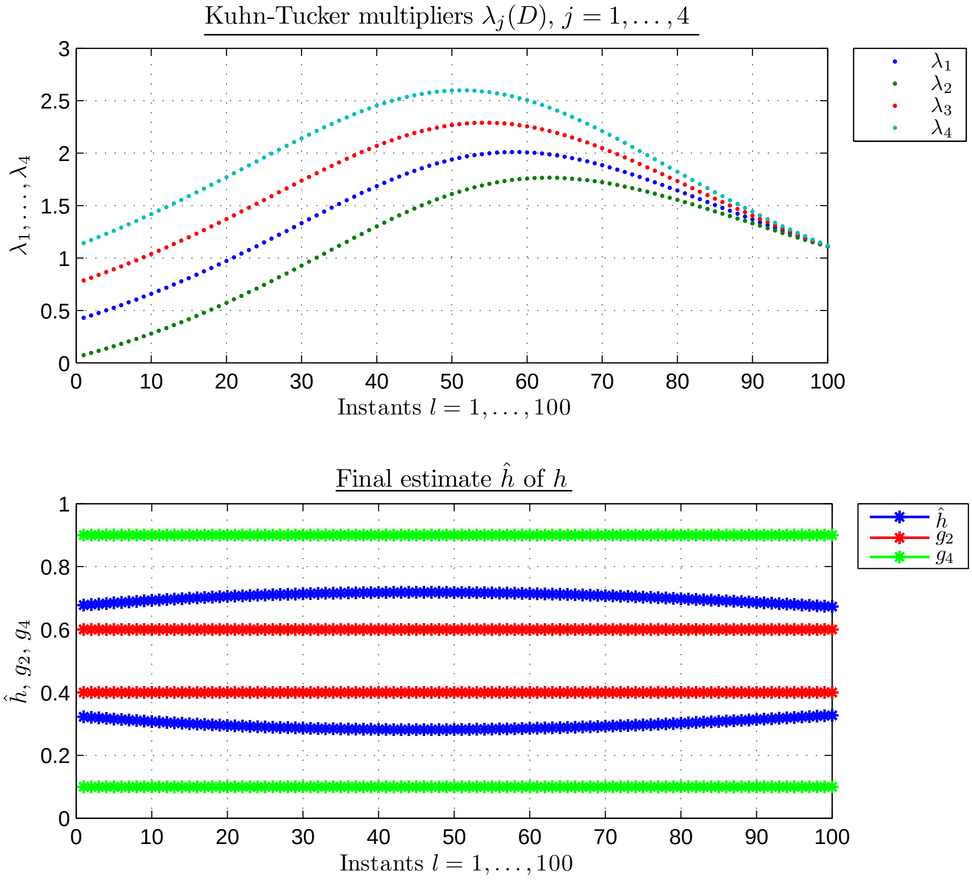

**Fig. 1.** The behaviour of the $\lambda_j(D)$, $j = 1, \dots, 4$ based on 100 different decreasing values of $\beta_j^*(D)$ using (5) and the final weighted combination $\hat{h}(x_i)$, $i = 1, 2$ based on computed $\lambda_j(D)$, $j = 1, \dots, 4$.

erate 10 and 1000 4-tuples from binomial distribution $Bi(10, 1/3)$ and with each new set of data we compute the estimates as in Subsection 4.1. In both cases the upper bounds $\beta_j(D)$ used in our method were with each new set of random values set to $\beta_j^*(D) = \beta_{j,1}^*(D) \times 0.40$, where $\beta_{j,1}^*(D)$ is defined in (6) (see also (4)).

To compare the estimators we are interested in common values as sample mean, computed from all 40 (or 4000) values, mean value and variance of $\hat{h}(x_1)$ obtained from 10 (or 1000) estimates given by considered estimators. We also compute the square error, exploiting the squared distance of the values of par-

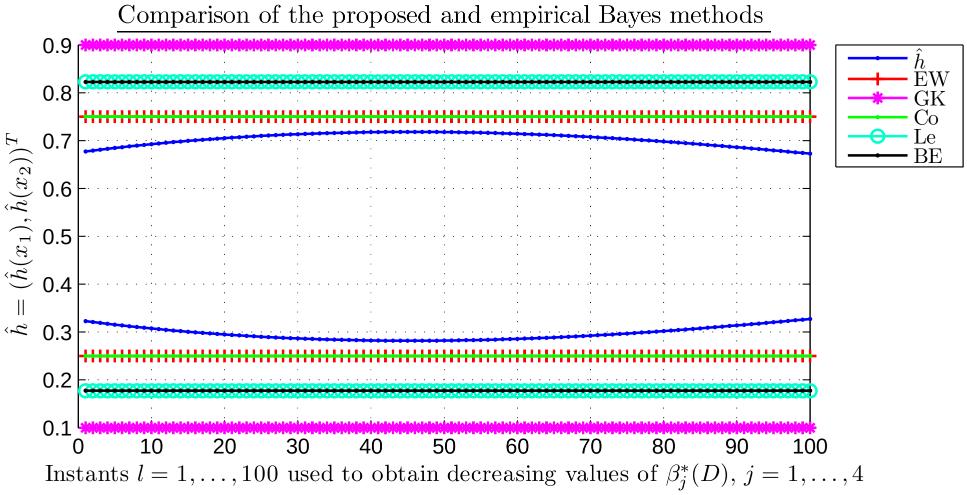

**Fig. 2.** Empirical Bayes methods in comparison with method based on the Kerridge inaccuracy and maximum entropy principle (3) in case when the estimate $\hat{h}(x_1)$ of the parameter $p = h(x_1)$ of binomial distribution is of interest (also its counterpart $\hat{h}(x_2) = 1 - \hat{h}(x_1)$ is drawn). For the proposed method, the estimate was computed for decreasing values of $\beta_j^*(D)$, $j = 1, \ldots, 4$, see (5).

ticular estimator from the sample mean:

$$\text{sq.error} = \sum_{m=1}^{M} (\hat{h}_m(x_1) - \text{ sample mean})^2, \quad M = 10, 1000$$

In case when only 10 4-tuples were generated, the sample mean equals 0.3850 and the results are the following:

| | Proposed method | EW | GK | Co | Le | BE |
|---|---|---|---|---|---|---|
| mean($\hat{h}(x_1)$) | 0.3415 | 0.3850 | 0.4100 | 0.3850 | 0.3939 | 0.3939 |
| var($\hat{h}(x_1)$) | 0.0037 | 0.0068 | 0.0121 | 0.0068 | 0.0061 | 0.0061 |
| sq.error | 0.0526 | 0.0615 | 0.1153 | 0.0615 | 0.0557 | 0.0557 |

We can see that according to the data in the sample, the estimator based on equal weights and Copas' second estimator work well, Lemon's and smooth incomplete Beta estimator are slightly different from the sample mean. If we take the information about the true distribution of generated data, $Bi(10, 1/3)$, our method gives really good estimate of the unknown parameter $h(x_1)$ even for such small sample.

In case of generating 1000 4-tuples (first 10 4-tuples coincide with those used previously) the sample mean is 0.3350. The results for considered methods are the following:

| | Proposed method | EW | GK | Co | Le | BE |
|---|---|---|---|---|---|---|
| mean($\hat{h}(x_1)$) | 0.2994 | 0.3350 | 0.3307 | 0.3344 | 0.3335 | 0.3335 |
| var($\hat{h}(x_1)$) | 0.0028 | 0.0051 | 0.0142 | 0.0062 | 0.0117 | 0.0117 |
| sq.error | 4.0902 | 5.0514 | 14.2110 | 6.1772 | 11.7356 | 11.7222 |

Here we see that all of the considered Bayes estimators perform really well, the estimate $\hat{h}(x_1)$ based on our method is slightly different from the sample mean and the true value $h(x_i)$. On the other hand the variance of $\hat{h}(x_1)$ and the squared error is the lowest among all considered estimators, which after fixing the values of the upper bounds $\beta_j(D)$ can lead estimates based on our estimator being closer to the sample mean and true value of the unknown parameter $h(x_1)$.

## 5 Conclusion and Future Work

In this paper we briefly described the method for combining data based on estimation of an unknown parameter. Both, data and parameter, are being pmfs. This method is based on the Kerridge inaccuracy and maximum entropy principle. The final estimate is a weighted combination of data, where the weights are obtained without any subjective influence, yet are non-trivial. They heavily depend on the Kuhn-Tucker multipliers arising during the computation. The aim of this paper consists in comparison with empirical Bayes methods while considering the binomial distribution and estimation of its parameter – probability of success. The results are satisfactory, even on a very small sample, the proposed method worked really well compared to the empirical Bayes methods. Thus after fixing the value of Kuhn-Tucker multipliers, which is the aim of our future work, the method has a great potential in small sample theory and many other fields of statistics.

## 6 Appendix

### 6.1 Determination of the estimate $\widetilde{\pi}(h|D)$ of the posterior pdf $\pi(h|D)$

To determine the estimate of the posterior pdf $\pi(h|D)$ we focus on the Kuhn-Tucker function of the optimization task (2) and arrange it as follows:

$$\begin{aligned}\mathrm{L}(\widetilde{\pi}(h|D);\boldsymbol{\lambda}(D)) &= \int_H \widetilde{\pi}(h|D)\log\left(\frac{\widetilde{\pi}(h|D)}{\frac{\prod_{i=1}^s h(x_i)^{(\sum_{j=1}^s \lambda_j(D)g_j(x_i)+1)-1}}{Z(\lambda_1(D),\ldots,\lambda_s(D))}}\right)\mathrm{d}h \\ &\quad - \log Z(\lambda_1(D),\ldots,\lambda_s(D))\underbrace{\int_H \widetilde{\pi}(h|D)\mathrm{d}h}_{=1} - \sum_{j=1}^n \lambda_j(D)\beta_j(D) \\ &\quad - \lambda_{s+1}(D)\left(\int_H \widetilde{\pi}(h|D)dh - 1\right),\end{aligned} \tag{7}$$

where $Z(\lambda_1(D), \ldots, \lambda_s(D))$ is a normalizing constant, $\lambda_j(D) \geq 0$ are Kuhn-Tucker multipliers, $j = 1, \ldots, s+1$ and $\boldsymbol{\lambda}(D) = (\lambda_1(D), \ldots, \lambda_s(D))$. According to the properties of the Kullback-Leibler divergence KLD(.,.) [9], the first term is minimal for $\widetilde{\pi}(h|D)$ being the pdf of the Dirichlet distribution with parameters $\sum_{j=1}^{s} \lambda_j(D) g_j(x_i) + 1$, $i = 1, \ldots, n$. The last term of (7) is equal to zero, the rest does not depend on $\widetilde{\pi}(h|D)$ and does not influence the minimization. Thus the estimate $\hat{\pi}(h|D)$ of the posterior pdf $\pi(h|D)$ in (2) is a pdf of Dirichlet distribution with parameters mentioned above.

### 6.2 Determination of the Kuhn-Tucker multipliers

In this subsection we derive the formula for Kuhn-Tucker multipliers $\lambda_j(D)$ arising in the optimization task (2) and playing the key role in the combination (3). Thus we compute the first derivatives of the Kuhn-Tucker function (7) with respect to $\lambda_j(D)$, $j = 1, \ldots, s$ and set each derivative equal to zero in order to find a minimum of this Kuhn-Tucker function. We omit the first and the last term of considered Kuhn-Tucker function from differentiation. The first term is already minimized - $\hat{\pi}(h|D)$ is a pdf of Dirichlet distribution $Dir(1 + \sum_{j=1}^{s} \lambda_j(D) g_j(x_i), i = 1, \ldots, n)$ and according to the properties of the Kullback-Leibler divergence we have $\mathrm{KLD}(\hat{\pi}(h|D)||\hat{\pi}(h|D)) = 0$. Since $\hat{\pi}(h|D)$ is a pdf, the last term is equal to zero.

The first derivative of (7) with respect to $\lambda_k(D)$ looks then as follows:

$$\begin{aligned}
&\frac{\partial}{\partial \lambda_k}\left(-\log Z(\lambda_1(D), \ldots, \lambda_s(D)) - \sum_j \lambda_j(D)\beta_j(D)\right) \\
&= \frac{\partial}{\partial \lambda_k}\left(-\log \frac{\prod \mathit{\Gamma}(1 + \sum_j \lambda_j(D) g_k(x_i))}{\mathit{\Gamma}(n + \sum_j \lambda_j(D))}\right) - \beta_k(D) \\
&= -\sum_i \psi\left(1 + \sum_j \lambda_j(D) g_k(x_i)\right) g_k(x_i) + \psi\left(n + \sum_j \lambda_j(D)\right) 1 - \beta_k(D) \\
&= -\sum_i \psi_i g_k(x_i) + \psi_0 - \beta_k(D) \quad \forall \lambda_j, j = 1, \ldots, s, \qquad (8)
\end{aligned}$$

where $\psi$ is the digamma function, see [2].

By using one-sided inverse - left inverse - we obtain the following system of nonlinear equations:

$$\begin{aligned}
-D_{(s\times n)}\boldsymbol{\psi}_{(n\times 1)} + \boldsymbol{\psi}_{0,(s\times 1)} &= \boldsymbol{\beta}_{(s\times 1)} \\
-D_{(s\times n)}\boldsymbol{\psi}_{(n\times 1)} &= \boldsymbol{\beta}_{(s\times 1)} - \boldsymbol{\psi}_{0,(s\times 1)} \\
I_n \boldsymbol{\psi}_{(n\times 1)} &= -D^{-1}_{\mathrm{left},(n\times s)}\left(\boldsymbol{\beta}_{(s\times 1)} - \boldsymbol{\psi}_{0,(s\times 1)}\right) \\
\boldsymbol{\psi}_{(n\times 1)} &= -D^{-1}_{\mathrm{left},(n\times s)}\boldsymbol{\beta}^{*}_{(s\times 1)}, \qquad (9)
\end{aligned}$$

where $D = (g_1, \ldots, g_s)^T$ (see Section 2). Thus:

$$\begin{array}{ccc} \psi(1 + \sum_j \lambda_j(D) g_j(x_1)) & = & \sum_j -D^{-1}_{\text{left},1j} \beta^*_j(D) \\ \vdots & \vdots & \vdots \\ \psi(1 + \sum_j \lambda_j(D) g_j(x_n)) & = & \sum_j -D^{-1}_{\text{left},nj} \beta^*_j(D) \end{array}$$

and to obtain the multipliers we use the inverse digamma function:

$$\begin{array}{ccc} \sum_j \lambda_j(D) g_j(x_1) & = & \psi^{-1}(\sum_j -D^{-1}_{\text{left},1j} \beta^*_j(D)) - 1 \\ \vdots & \vdots & \vdots \\ \sum_j \lambda_j(D) g_j(x_n) & = & \psi^{-1}(\sum_j -D^{-1}_{\text{left},nj} \beta^*_j(D)) - 1. \end{array}$$

The results using matrix notation are the following:

$$\begin{aligned} (D^T)_{(n\times s)} \boldsymbol{\lambda}_{s\times 1} &= (\psi^{-1}(-D^{-1}_{\text{left},(n\times s)} \boldsymbol{\beta}^*_{(s\times 1)}))_{(n\times 1)} - \mathbf{1}_{(n\times 1)} \\ \boldsymbol{\lambda}_{(s\times 1)} &= (D^T)^{-1}_{\text{left},(s\times n)} \left( (\psi^{-1}(-D^{-1}_{\text{left},(n\times s)} \boldsymbol{\beta}^*_{(s\times 1)}))_{(n\times 1)} - \mathbf{1}_{(n\times 1)} \right). \end{aligned} \tag{10}$$

### 6.3 Determination of the final combination

We exploit the formula for the expected value of random vector having Dirichlet distribution $Dir(1 + \sum_{j=1}^{s} \lambda_j(D) g_j(x_i), i = 1, \ldots, n)$ (see Subsection 6.1) and conclude the following:

$$\begin{aligned} \mathrm{E}_{\hat{\pi}(h|D)}(h(x_i)|D) = \hat{h}(x_i) &= \frac{1 + \sum_{j=1}^{n} \lambda_j(D) g_j(x_i)}{\sum_{i=1}^{n} \left(1 + \sum_{j=1}^{s} \lambda_j(D) g_j(x_i)\right)} \\ &= \frac{1 + \sum_{j=1}^{n} \lambda_j(D) g_j(x_i)}{n + \sum_{j=1}^{s} \lambda_j(D)} \end{aligned}$$

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

## References

1. Abbas, A.E. : A Kullback-Leibler View of Linear and Log-Linear Pools. Decision Analysis, vol. 6/1, pp. 25–37 (2009)
2. Abramowitz, M. and Stegun, I. A. (Eds.): Handbook of Mathematical Functions with Formulas, Graphs, and Mathematical Tables. Dover Publications, New York (1972)
3. Bernardo, J.M.: Expected Information as Expected Utility. Ann. Stat., vol. 7, pp. 686–690 (1979)
4. Clemen, R.T. and Winkler, R.L.: Combining Probability Distributions From Experts in Risk Analysis. Risk Analysis, vol. 19/2, pp. 187–203 (1999)
5. Copas, J. B.: Empirical Bayes Methods and the Repeated Use of a Standard. Biometrika, vol. 59, pp. 349–360 (1972)

6. Fine, T.L.: Theories of Probability: An Examination of Foundations. Academic Press, London (1973)
7. Griffin, B.S. and Krutchkoff, R.G.: Optimal Linear Estimators: An Empirical Bayes Version with Application to the Binomial Distribution. Biometrika, vol. 58, pp. 195–201 (1971)
8. Kerridge, D.F.: Inaccuracy and Inference. J. R. Stat. Soc., Ser. B, vol. 23, pp. 184–194 (1961)
9. Kullback, S. and Leibler, R.A.: On information and sufficiency. Ann. Math. Stat., vol. 22, pp. 79-86 (1951)
10. Lemon, G.H. and Krutchkoff, R.G.: An Empirical Bayes Smoothing Technique. Biometrika, vol. 56, pp. 361–365 (1969)
11. Maritz, J.S. : Empirical Bayes Methods. Methuen's Monographs on Applied Probability and Statistics. Methuen and Co Ltd., London (1970)
12. Martz, H.F. and Lian, M.G. : Empirical Bayes Estimation of the Binomial Parameter. Biometrika, vol. 61/3, pp. 517–523 (1974)
13. Sečkárová, V. : On Supra-Bayesian Weighted Combination of Available Data Determined by Kerridge Inaccuracy and Entropy. Pliska Stud. Math. Bulgar., vol. 22, pp. 159-168 (2013)
14. Shore, J.E. and Johnson, R.W.: Axiomatic Derivation of the Principle of Maximum Entropy and the Principle of Minimum Cross-entropy. IEEE Trans. Inf. Theory, vol. 26, pp. 26–37 (1980)