PLISKA
STUDIA MATHEMATICA
BULGARICA

ПЛИСКА
БЪЛГАРСКИ
МАТЕМАТИЧЕСКИ
СТУДИИ

Pliska Stud. Math. Bulgar. **22** (2013), 159–168

PLISKA
STUDIA MATHEMATICA
BULGARICA

ON SUPRA-BAYESIAN WEIGHTED COMBINATION OF AVAILABLE DATA DETERMINED BY KERRIDGE INACCURACY AND ENTROPY*

Vladimíra Sečkárová

Abstract. Every process in our environment can be described with a statistical model containing inner properties expressed by parameters. These are usually unknown and the determination of their values is of interest in the statistical branch called parameter estimation. This branch involves many methods solving different estimation cases, e.g. the estimation of location and scale parameters. To obtain the parameter estimate we exploit the data given by data sources. In particular, the estimate is their combination. Improvement of the parameter estimates involve the assignment of the weights to the data sources resulting in a weighted combination of data. Unfortunately this approach brings difficulties regarding the determination of the weights and their subjective affection. In recently introduced Supra-Bayesian approach it is proposed to use the Kerridge inaccuracy and the maximum entropy principle to overcome the problem of subjective influence. In this paper we focus on the derivation of the weights arisen within the Supra-Bayesian approach and on the simulation study of their behaviour and the behaviour of the final estimate.

This work was supported by the grants GACR 13-13502S and SVV 265 315.
2010 *Mathematics Subject Classification:* 94A17.
Key words: Kerridge inaccuracy, maximum entropy principle, parameter estimation.

1. Introduction. Our environment consists of various interesting processes. These can be with a little bit of imagination described with specific statistical models. Related processes can have similar model, but they differ in the inner properties, which are called parameters (e.g. location and scale parameters). These are often unknown. The need to find their values initiated the creation of the statistical branch called parameter estimation, where we try to estimate the true value of the parameters. This branch includes many methods solving different estimation tasks, e.g. to obtain the estimate of the location parameter we can use the least-squares method (see [1]).

The estimation depends heavily on the data available from one or several sources, the resulting estimate is then combination of these data. In order to improve the estimate of considered parameter, or roughly speaking, to obtain the estimate closer to the true, but unknown, value of the parameter, we can assign each of the sources a weight. This weight interprets the credibility of the source and often coheres with a subjective opinion. The estimate is then simply the weighted combination of available data. This type of methods and their generalizations have been developed since the middle of the past century (see [2]), but they often involve limitations regarding the determination of the weights.

If the estimation of a probability density function (pdf) is of interest, several weighting approaches are available (see [3]). In this paper we focus on the case when discrete random variables are considered – on the probability mass function (pmf) estimation. To construct such estimate we use recently introduced Supra-Bayesian approach [4]. We assume, that the data, provided by available data sources, have also the form of pmf. The weights for data sources are then determined without any subjective influence by defining the constraints as the expected values of a particular information divergence. The final estimate, a weighted combination of given pmfs, is obtained by exploiting this information divergence and the constraints.

Since in the Supra-Bayesian approach the constraints are a part of yet unspecified constrained optimization task, values of the weights will heavily depend on arisen Lagrange multipliers. The novelty of this paper consists in the mathematical derivation of the multipliers and the simulation study regarding the behaviour of the multipliers and the final estimate.

The paper is composed as follows: the second section contains the recapitulation of the steps leading to a final estimate, the third section describes the derivation of the Lagrange multipliers. In the fourth part of the paper we provide a simulation study on the behaviour of the Lagrange multipliers and the

final estimate.

2. The estimation based on Supra-Bayesian approach. Recall the problem drafted in the previous section. Let us consider n-dimensional parameter $h = (h(x_1), \ldots, h(x_n))^T$, $n < \infty$, which is $(0,1)^n$-valued and satisfies $\sum_{i=1}^{n} h(x_i) = 1$ (h is a pmf of a discrete random vector X with n possible outcomes). We are interested in estimation of this parameter based on given data $D = (g_1, \ldots, g_s)^T$, $s < \infty$, where g_j denotes a pmf given by j^{th} data source describing the previously mentioned random vector X. To obtain the estimate we will use the Kerridge inaccuracy $\mathrm{K}(.,.)$ defined as $\mathrm{K}(h,g) = -\sum_{i=1}^{n} h(x_i) \log g(x_i)$, where h, g are two pmfs describing a common random vector. In particular, we search for the element minimizing the expected Kerridge inaccuracy in the following sense:

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

where $\pi(h|D)$ stands for the posterior pdf of pmf h based on the data in D and $\mathrm{E}[.|.]$ denotes the conditional expectation with respect to this posterior pdf. Roughly speaking, the expected energy expended in transition from every possible pmf h (from the set H) to a particular pmf $\tilde{h}$ (from the set $\tilde{H}$) is minimal when $\tilde{h}$ has form (1). Sets H and $\tilde{H}$ are not constrained by any additional condition, thus both contain all possible pmfs on $(0,1)^n$ and in fact, they coincide.

In the next part we derive the estimate of the posterior pdf $\pi(h|D)$.

2.1. Determination the posterior probability density function. To compute the estimate (1) we need to determine the posterior probability density function $\pi(h|D)$. In order to do that, we introduce the following task of 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 rationale behind this particular choice of objective function (2) is the following: for determination of the estimate of any pdf under no knowledge available, the maximum entropy principle is satisfactory (see [5]). This principle claims that from the set of all possible pdfs we should choose the one with the highest entropy as the estimate of $\pi(h|D)$. For our purpose this principle is interpreted as an unconstrained optimization task.

In case when the additional information is available, we would like to incorporate it into computation, i.e. we set up the constraints dependent on this information, which leads to a constrained optimization task. Since in our case the pmfs $g_1, \dots, g_s$ are available, we use the information divergence, i.e. the Kerridge inaccuracy, to connect them with h (as depicted in the definition of the set M in (2)). These constraints interpret that for each source the energy needed to change its pmf onto every possible h is bounded. Or in other words, we monitor how far is each of the sources from each h in the sense of expected Kerridge inaccuracy and assume it is bounded.

To determine the estimate of the posterior pdf $\pi(h|D)$ we reformulate the Lagrangian L(.,.) of the optimization task (2) as follows:

$$\mathrm{L}(\widetilde{\pi}(h|D);\boldsymbol{\lambda}(D)) = \int_H \widetilde{\pi}(h|D)\log\left(\frac{\widetilde{\pi}(h|D)}{\frac{\prod_{i=1}^{s} h(x_i)^{(\sum_{j=1}^{s}\lambda_j(D)g_j(x_i)+1)-1}}{Z(\lambda_1(D),\dots,\lambda_s(D))}}\right)\mathrm{d}h$$

$$(3)\qquad -\log Z(\lambda_1(D),\dots,\lambda_s(D))\underbrace{\int_H \widetilde{\pi}(h|D)\mathrm{d}h}_{=1} - \sum_{j=1}^{n}\lambda_j(D)\beta_j(D)$$

$$-\lambda_{s+1}(D)\left(\int_H \widetilde{\pi}(h|D)dh - 1\right),$$

where $Z(\lambda_1(D),\dots,\lambda_s(D))$ is a normalizing constant, $\lambda_j(D) \geq 0$ are Lagrange multipliers, $j = 1,\dots,s+1$ and $\boldsymbol{\lambda}(D) = (\lambda_1(D),\dots,\lambda_s(D))$. We see that the first term is minimal for $\widetilde{\pi}(h|D)$ being the pdf of the Dirichlet distribution with parameters $\sum_{j=1}^{s}\lambda_j(D)g_j(x_i)+1$, $i = 1,\dots,n$. Since $\int_H \widetilde{\pi}(h|D)\mathrm{d}h = 1$, it is obvious that the last term is then equal to zero and is omitted from further computation. The other terms do not depend on $\widetilde{\pi}(h|D)$ and do not influence the minimization. Thus the estimate $\hat{\pi}(h|D)$ in (2) is a pdf of Dirichlet distribution with parameters mentioned above.

2.2. Determination of the parameter estimate. According to the formula (1) and the results of Subsection (2.1) we now construct the estimate $\hat{h}$ of pmf h. In particular, we exploit the formula for the expected value of random vector having Dirichlet distribution and conclude the following:

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

where

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

It is easy to see that the estimate $\hat{h}$ is a weighted combination of the given pmfs $g_1, \ldots, g_s$. To obtain the weights we have to compute the Lagrange multipliers $\lambda_j(D)$, $j = 1 \ldots, s$. As will be shown in the next section, their computation does not require any subjective influence.

Since we avoided the derivatives in minimization of the Lagrangian (3), the values of its Lagrange multipliers are still unknown. In the next section we derive the multipliers $\lambda_j(D)$ for the final estimate (4).

3. Determination of the Lagrange multipliers. Recall the optimization task (2), in particular the constraints

$$\mathrm{E}_{\pi(h|D)}(\mathrm{K}(g_j, h)|D) \leq \beta_j(D), \ j = 1, \ldots, s.$$

In order to obtain the Lagrange multipliers $\lambda_j(D)$, $j = 1, \ldots, s$, we have to set the upper bounds $\beta_j(D)$ on the expected values for each source. Before we actually do that, we provide a straightforward derivation of the multipliers. We compute the first derivatives of Lagrangian (3) 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 Lagrangian. Here, we omit the first and the last term of considered Lagrangian from differentiation, since they are already minimized. The first derivative with respect to λ_k looks then as follows:

$$\begin{aligned} &\frac{\partial}{\partial \lambda_k} \left(-\log Z(\lambda_1(D), \ldots, \lambda_s(D)) - \sum_j \lambda_j(D) \beta_j(D) \right) \\ &= \frac{\partial}{\partial \lambda_k} \left(-\log \frac{\prod \Gamma(1 + \sum_j \lambda_j(D) g_k(x_i))}{\Gamma(n + \sum_j \lambda_j(D))} \right) - \beta_k(D) \end{aligned}$$

$$= -\sum_i \psi\left(1+\sum_j \lambda_j(D) g_k(x_i)\right) * g_k(x_i) + \psi\left(n+\sum_j \lambda_j(D)\right) * 1 - \beta_k(D)$$

(5)

$$= -\sum_i \psi_i * g_k(x_i) + \psi_0 - \beta_k(D) \quad \forall \lambda_j, j = 1, \ldots, s,$$

where ψ is the digamma function.

We obtain the following system of nonlinear equations (using one-sided inverse – left inverse):

$$\begin{aligned} -\boldsymbol{P}_{(s\times n)}\boldsymbol{\psi}_{(n\times 1)} + \boldsymbol{\psi}_{0,(s\times 1)} &= \boldsymbol{\beta}_{(s\times 1)} \\ -\boldsymbol{P}_{(s\times n)}\boldsymbol{\psi}_{(n\times 1)} &= \boldsymbol{\beta}_{(s\times 1)} - \boldsymbol{\psi}_{0,(s\times 1)} \\ I_n\boldsymbol{\psi}_{(n\times 1)} &= -\boldsymbol{P}^{-1}_{\text{left},(n\times s)}\left(\boldsymbol{\beta}_{(s\times 1)} - \boldsymbol{\psi}_{0,(s\times 1)}\right) \\ \boldsymbol{\psi}_{(n\times 1)} &= -\boldsymbol{P}^{-1}_{\text{left},(n\times s)}\boldsymbol{\beta}^*_{(s\times 1)}. \end{aligned}$$

That is:

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

To obtain the multipliers we use the inverse digamma function:

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

The final matrix interpretation is:

$$(\boldsymbol{P}^T)_{(n\times s)}\boldsymbol{\lambda}_{s\times 1} = (\psi^{-1}(-\boldsymbol{P}^{-1}_{\text{left},(n\times s)}\boldsymbol{\beta}^*_{(s\times 1)}))_{(n\times 1)} - \mathbf{1}_{(n\times 1)}$$

$$\boldsymbol{\lambda}_{(s\times 1)} = (\boldsymbol{P}^T)^{-1}_{\text{left},(s\times n)}\left((\psi^{-1}(-\boldsymbol{P}^{-1}_{\text{left},(n\times s)}\boldsymbol{\beta}^*_{(s\times 1)}))_{(n\times 1)} - \mathbf{1}_{(n\times 1)}\right). \tag{6}$$

Since the derivation is done, we now discuss the choice of the upper bounds $\beta_j(D)$. In particular we set the value of $\beta^*_k(D)$ because ψ_0 in (5) is unknown.

Since the expected Kerridge inaccuracy does not have the least upper bound property (its upper bound can rise to infinity), we will focus on a fixed bound based on given pmfs $g_1, \ldots, g_s$. For k^{th} data source we will use the mean value of the Kerridge inaccuracy as follows:

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

where

$$h_{\mathrm{data}}(x_i) = \sqrt[s]{\prod_{j=1}^{s} g_j(x_i)} \quad i = 1, \ldots, n. \tag{7}$$

The behaviour of the multipliers $\lambda_j(D)$, $j = 1, \ldots, s$ and the final combination $\hat{h}$ (estimate of the unknown pmf h) is shown in the next section.

4. Simulation results. In this section we study the behaviour of the Lagrange multipliers $\lambda_j(D)$, $j = 1, \ldots, s$ and the weighted combination (4) based on them. Regarding the information given in the previous section, we set the upper bound for j^{th} source as follows: $\beta_j^*(D) = \mathrm{K}(g_j, h_{\mathrm{data}})$, where h_{data} is defined in (7). We explore the changes in the value of $\lambda_j(D)$ when decreasing these $\beta_j^*(D)$ but keeping the ratio between β_j^* and β_k^*, $j \neq k$, $j, k = 1, \ldots, s$, the same.

4.1. Illustrative example 1. Consider the case of four sources providing the pmfs $g_1, \ldots, g_4$:

$$D = \begin{pmatrix} (g_1(x_1), g_1(x_2)) \\ \ldots \\ \ldots \\ (g_4(x_1), g_4(x_2)) \end{pmatrix} = \begin{pmatrix} (0.45, 0.55) \\ (0.4, 0.6) \\ (0.15, 0.85) \\ (0.1, 0.9) \end{pmatrix}$$

We decrease the bounds in the following way: $\beta_{j,l}^* = \beta_j^* * (1 - l * 0.0099)$ for instants $l = 1, \ldots, 100$, $\beta_{j,1}^* = \mathrm{K}(g_j, h_{\mathrm{data}})$. The results using Matlab built-in functions are shown in Figure 1.

In the upper part of Figure 1. we see that for 100 instants first few λ_j are negative, which is at odds with the property of Lagrange multipliers ($\lambda_j \geq 0$, $j = 1, \ldots, 4$). This is probably caused by the properties of the inverse digamma function in (6).

In the bottom part we notice that in the majority of cases the final weighted combination is closer to the pmfs with higher entropy (g_1 – the highest entropy, g_4

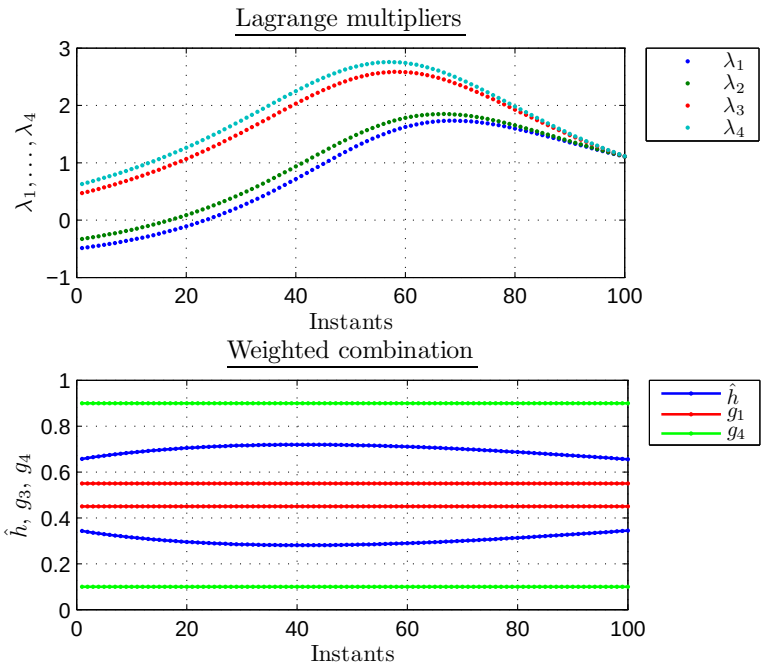

Fig. 1. The behaviour of the $\lambda_j(D)$, $j = 1, \ldots, 4$ and the final weighted combination $\hat{h}(x_i)$, $i = 1, 2$ for 100 instants

– the lowest entropy). There is also a significant difference among these weighted combinations, that can be fixed by choosing different starting point for $\beta_j^*(D)$, $j = 1, \ldots, 4$.

4.2. Illustrative example 2 In another situation, where

$$D = \begin{pmatrix} (g_1(x_1), g_1(x_2)) \\ \ldots \\ \ldots \\ (g_4(x_1), g_4(x_2)) \end{pmatrix} = \begin{pmatrix} (0.8, 0.2) \\ (0.75, 0.25) \\ (0.55, 0.45) \\ (0.85, 0.15) \end{pmatrix}$$

the results are similar to the results in Subsection (4.1) (see Figure 2). Although there are three sources with obviously lower entropy (g_1, g_2, g_4), the final estimate is still closer to the one with higher entropy (g_3) in the majority of cases.

5. Future work. Since the determination of the Lagrange multipliers in the final weighted combination (4) is still not fully satisfactory, the future work

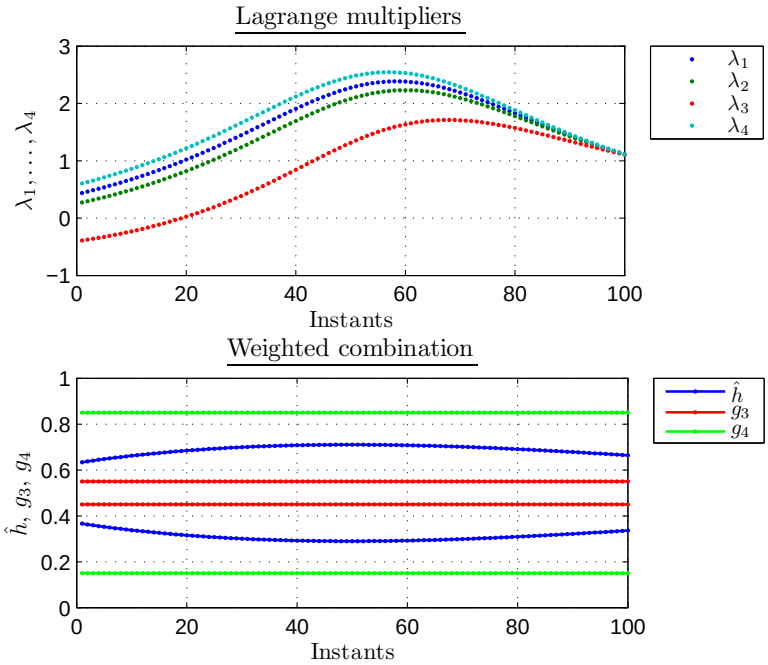

Fig. 2. The behaviour of the λ_j, $j = 1, \ldots, 4$ and the final weighted combination $\hat{h}(x_i)$, $i = 1, 2$ for 100 instants.

involves the reformulation of the optimization task (2) with use of different information divergence, i.e. the Kullback-Leibler divergence, and the minimum cross entropy principle. Also the use of numerical methods is of interest.

REFERENCES

[1] E. H. Lloyd. Least-Squares Estimation of Location and Scale Parameters Using Order Statistics. *Biometrika* **39** (1952), Nos 1/2, 88–95.

[2] M. M. Lentner. Generalized Least-Squares Estimation of a Subvector of Parameters in Randomized Fractional Factorial Experiments. *The Annals of Mathematical Statistics* **40** (1969), No 4, 1344–1352.

[3] Ming-Hui Chen. Importance-Weighted Marginal Bayesian Posterior Density Estimation. *Journal of the American Statistical Association* **89** (1994), No 427, 818–824.

[4] V. Sečkárová Supra-Bayesian Approach to Merging of Incomplete and Incompatible Data. Proceedings of the Decision Making with Multiple Imperfect Decision Makers (Eds Guy Tatiana V., Karny Miroslav, Wolpert David) 24th Annual Conference on Neural Information Processing Systems, 2010.

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

Vladimíra Sečkárová
Department of Adaptive Systems
Institute of Information Theory and Automation
Pod Vodárenskou věží 4
182 08 Prague 8, Czech Republic
`seckarov@utia.cas.cz`

Department of Probability and Mathematical Statistics
Faculty of Mathematics and Physics
Charles University in Prague
Sokolovká 83
186 75 Prague 8, Czech Republic