Pliska Stud. Math. **24** (2015), 181–188

PLISKA
STUDIA MATHEMATICA

# DYNAMIC PARAMETER ESTIMATION BASED ON MINIMUM CROSS-ENTROPY METHOD FOR COMBINING INFORMATION SOURCES*

Vladimíra Sečkárová

When combining information sources, e.g. measuring devices or experts, we deal with two problems: which combining method to choose (linear combination, geometric mean) and how to measure the reliability of the sources, i.e. how to assign the weights to them. Inspired by [5] we introduce a method which overcomes such shortcomings. Proposed method, based on minimization of the Kullback-Leibler divergence with specific constraints, directly combines data, i.e. probability vectors, thus no additional step to obtain the weights is needed. The detailed description of the proposed method and a comparison with recently introduced dynamic diffusion estimation [2], which heavily depends on the determination of the weights, form the core of this contribution.

## 1. Introduction

Statisticians, who would like to use methods for combining information sources providing data about a biological process for example, face several issues: which method to use; how do I know which source is reliable. To solve these issues and to improve the performance of combining methods, especially in the dynamic scenarios, the observations are treated as random variables. Unfortunately, the underlying probability distribution is often fixed and we might get misleading results if the probability distribution does not fit the data well. In this paper, we simply assume each source provides a probability vector, assigning each possible

*This work has been partially supported by the grant GACR 13-13502S, grant svv 2015 No. 260225 and by the special scholarship from ČSOB.

2010 *Mathematics Subject Classification:* 94A17, 62L12

*Key words:* minimum cross-entropy principle, Kullback-Leibler divergence, dynamic diffusion estimation

outcome of random variable a probability (number of outcomes is considered to be finite).

Inspired by the work [5] we will search for the combination of given probability vectors as the minimizer of the expected loss function [1], where the loss function should reflect our demand on working with probability vectors. The Kullback-Leibler (KL) divergence [4], a non-symmetric function measuring the 'distance' of one probability vector from another, is a reasonable choice.

The form of the minimizer, based on minimum cross-entropy principle with constraints, then lets us combine given probability vectors without additional determination of weights. Detailed description of the proposed method is given in Section 2.

The proposed method is then compared to the recently introduced dynamic diffusion estimation (DDE). Although DDE assumes each source computes its estimate based on data from cooperating sources, this approach can be viewed from the centralized point of view – all data are combined by combining element not included in the set of sources. In Section 3. we show that the proposed method and DDE coincide if the random variables in DDE are categorically distributed. Since the weights in DDE are not specified, the proposed method can be exploited as an estimation method for the parameter of the categorical distribution. The differences between results given by considered methods are demonstrated on the example in Section 4.. The basics of DDE and useful formulas for the proposed method can be found in Section 6. – Appendix.

## 2. Minimum cross-entropy based method for combining sources

Let us consider the following scenario (see [5]): let us have $s$ sources, each providing observations as $n$-dimensional probability vectors. Thus from $j^{th}$ source we obtain a probability vector $p_j = (p_{j1}, \dots, p_{jn})$, where $\sum_{i=1}^{n} p_{ji} = 1$ and $p_{ji} \geq 0$ for $i = 1, \dots, n$. The advantage of this approach is that no particular probability distribution is assumed.

We denote the combination of $p_1, \dots, p_s$, an unknown probability vector, by $q$, treat it as a random vector and search for its optimal (explained later) estimate $\hat{q}$. To obtain $\hat{q}$ we minimize the conditional expected value of the KL-divergence (KLD) [4] with respect to the conditional pdf $\pi(q|p_1, \dots, p_s)$ conditioned on $p_1, \dots, p_s$

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

The minimizing element of this expected loss [1] is the conditional expected value

of $q$ with respect to the conditional pdf $\pi(q|p_1,\dots,p_s)$ conditioned on $p_1,\dots,p_s$

$$\hat{q} = \mathrm{E}_{\pi(q|p_1,\dots,p_s)}[q|p_1,\dots,p_s]. \tag{1}$$

Since the estimate (1) heavily depends on the form of the unknown conditional pdf $\pi(q|p_1,\dots,p_s)$ we dedicate the next section to the search of this pdf.

### 2.1. Search for the conditional pdf $\pi(q|p_1,\dots,p_s)$

In this section we determine the conditional pdf $\pi(q|p_1,\dots,p_s)$. We specify what the appropriate conditional pdf has to satisfy and how to choose one pdf among all appropriate pdfs.

#### 2.1.1. Constraints on the conditional pdf $\pi(q|p_1,\dots,p_s)$

We again exploit work [5], where the constraints were represented by the expected KL-divergences from $p_j$ to $q$ with respect to the conditional pdf $\pi(q,p_1,\dots,p_s)$. These expectations were bounded, unfortunately the bounds were not determined exactly. Thus the resulting combination was dependent on their values. To overcome this shortcoming we consider the equalities among the expected values of the KL-divergence which no longer allow any freedom in constraints:

$$\begin{aligned} \mathrm{E}_{\pi(q|p_1,\dots,p_s)}[\mathrm{KLD}(p_s||q)|p_1,\dots,p_s] &= \mathrm{E}_{\pi(q|p_1,\dots,p_s)}[\mathrm{KLD}(p_j||q)|p_1,\dots,p_s] \\ & j = 1,\dots,s-1. \end{aligned} \tag{2}$$

Pdfs satisfying these constraints control the relation between provided probability vectors and the unknown vector $q$ simultaneously among all sources.

#### 2.1.2. Choice of the prior distribution

Choice of the prior pdf $\pi_0(q)$ is based on the fact that we would like to model the probability vector $q$. Thus the prior pdf $\pi_0(q)$ will be the pdf of the Dirichlet distribution.

#### 2.1.3. Form of the conditional pdf $\pi(q|p_1,\dots,p_s)$

To obtain $\pi(q|p_1,\dots,p_s)$ we exploit minimum cross-entropy principle [6] (instead of maximum entropy principle, [5]). We choose conditional pdf $\pi(q|p_1,\dots,p_s)$ that solves the following problem:

$$\begin{gathered} \min_{\pi(q|p_1,\dots,p_s)} \mathrm{KLD}(\pi(q|p_1,\dots,p_s)||\pi_0(q)) \\ \text{with respect to the constraints (2)}, \end{gathered} \tag{3}$$

where $\pi_0(q)$ is the pdf of the Dirichlet distribution with parameters $\nu_{01},\dots,\nu_{0n}$.

Minimizing conditional pdf is the pdf of the Dirichlet distribution for any Dirichlet prior pdf, see (12). Thus it is satisfactory to perform the minimization

over the set of all admissible $\nu_i$ (generally: $\nu_i > 0$, $i = 1, \dots, n$). Values of the parameters $\nu_1, \dots, \nu_n$ for nonlinear optimization task (3) can be determined numerically using e.g. Matlab.

Unlike in [5], where the maximum entropy principle was considered, minimum cross-entropy allows us to change the prior values $\nu_{01}, \dots, \nu_{0n}$ with each time step, which makes the proposed method useful for dynamic scenarios.

The relation to the prior value of the parameters $\nu_{01}, \dots, \nu_{0n}$ is expressed by the following formula (see (12)):

$$\nu_i = \nu_{0i} + \sum_{j=1}^{s} \lambda_j (p_{ji} - p_{si}), \quad i = 1, \dots, n, \tag{4}$$

where the Lagrange multipliers $\lambda_j$, see Section 6.2., can be also obtained numerically.

### 2.2. The combination of $p_1, \dots, p_s$ represented by the estimate $\hat{q}$

The estimate $\hat{q}$ in (1) representing the final weighted combination of $p_1, \dots, p_s$ has, based on the results of Section 2.1. and properties of the Dirichlet distribution, the following form:

$$\hat{q}_i = \frac{\nu_i}{\sum_{i=1}^{n} \nu_i} = \frac{\nu_{0i} + \sum_{j=1}^{s} \lambda_j (p_{ji} - p_{si})}{\sum_{i=1}^{n} \nu_{0i}}, \quad i = 1, \dots, n, \tag{5}$$

because $\sum_{i=1}^{n} \nu_i = \sum_{i=1}^{n} \nu_{0i}$, see (4).

Obtained estimate $\hat{q}$ of $q$ is optimal in the sense that the conditional pdf leading to $\hat{q}$ solves (3).

## 3. Dynamic diffusion estimation for categorical distribution

Setup considered in Section 2. coincides with dynamic diffusion estimation (DDE) (see [2] and Section 6.1.) when the underlying probability distribution of the random variable $Y$ is categorical with $n$ possible categories. In such case the parameter $\theta$ is an $n$-dimensional vector of probabilities $P(X = i) = q_i$, $i = 1, \dots, n$. Conjugate prior distribution is the Dirichlet distribution. Thus the prior and the posterior pdf are both pdfs of the Dirichlet distribution, prior pdf with hyperparameters $\nu_{01}, \dots, \nu_{0n}$ and posterior pdf with hyperparameters $\nu_1, \dots, \nu_n$.

Suppose we obtain from each node the point estimate $(P_j(X = 1), \dots, P_j(X =$

$n)) = (p_{j1}, \ldots, p_{jn})$ of $q$, $j = 1, \ldots, s$. Then the estimate $\hat{q}$ of $q$ has the form (11):

$$\hat{q}^* = \sum_{j=1}^{s} a_j p_j, \tag{6}$$

where the weights $a_j$ are unspecified, often chosen as uniform: $a_j = 1/s$, $j = 1, \ldots, s$.

Our suggestion is to exploit the setup in Section 2. and the combination (5), since it can be easily transformed into the form (6):

$$\hat{p}_i = \sum_{j=1}^{s} \frac{\frac{\nu_{0j}}{s p_{ji}} + \lambda_j \left(1 - \frac{p_{si}}{p_{ji}}\right)}{\sum_{i=1}^{n} \nu_{0i}} p_{ji}, \tag{7}$$

when $p_{ji} \neq 0$, $i = 1, \ldots, n$, $j = 1, \ldots, s$.

In both Sections 2. and 3. the time index can be easily added to (5) and (6).

## 4. Example – Proposed method vs. DDE in combining probability vectors

This example compares two previously described methods, the proposed method for combining probability vectors (see Section 2.) and the method for combining point estimates (see Sections 6.1. and 3.).

Suppose we have 3 sources/nodes ($s = 3$) providing 3-dimensional probability vectors ($n = 3$). Values of probability vectors/point estimates at time instant $t = 1, \ldots, 50$ were obtained from the 'true' probability vector ($q_1 = 0.56, q_2 = 0.22, q_3 = 0.22$) by adding small noise $\epsilon$ ($|\epsilon| < 0.1$) to the probabilities $q_1$ and $q_2$. Resulting combinations (5) and (6) (with time independent uniform weights: $a_j = 1/s$, $j = 1, \ldots, s$) are shown in the Fig.1 on the left.

The results in case when the third source was corrupted, meaning his 'true' probability vector was ($q_1 = 0.04, q_2 = 0.35, q_3 = 0.61$), are shown in the Fig.1 on the right.

## 5. Conclusion and future work

The presented method based on minimum cross-entropy principle and specific constraints brings a simplification into combining information sources. A lot of combining methods, treating data as random variables, consider particular probability distribution. We work with probability vectors (discrete case, finite number of outcomes) with no assumption on the underlying probability distribution. Another positive contribution of the proposed method is that the final combination is obtained without additional step to compute weights for the sources within dynamic setting. Our suggestion is to use the proposed method in the combine step

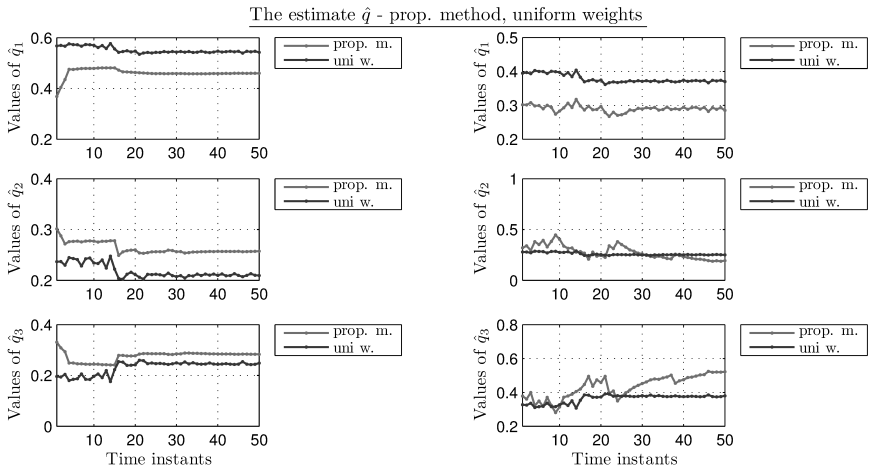


Figure 1: Combinations obtained by the proposed method and by DDE with uniform weights. On the left: no corrupted sources. On the right: the third source is corrupted.

of the recently introduced dynamic diffusion estimation in distributed networks [2].

In future work we would like to inspect the relation between $\hat{p}$ and $\nu$ in (5) for possible exact form of the weights after the values of $\nu$ will be determined numerically.

## 6. Appendix

### 6.1. Dynamic diffusion estimation (DDE)

Here we list basic ideas and formulas for the dynamic diffusion estimation in distributed networks. For more details see [2].

Let $y$ be an observed variable, $\theta$ be an unknown fixed parameter. DDE considers pdfs of probability distributions belonging to the exponential family:

$$f(y|\theta) = h(y)g(\theta)\exp[\eta(\theta)T(y)], \tag{8}$$

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

To easily incorporate data obtained at each time step the sequential Bayes rule is exploited. A prior pdf, conjugate to (8), is

$$\pi(\theta|\xi,\omega) = q(\xi,\omega)g(\theta)^{\omega}\exp[\eta(\theta)\xi], \tag{9}$$

where $\xi$ and $\omega$ are the hyperparameters.

Then, the following steps, forming the base of DDE, are performed: adapt step and/or combine step. In adapt step the hyperparameters of $j^{th}$ source are updated by new set of data $\{y_k\}$, where $k \in \mathcal{N}_j$ and $\mathcal{N}_j$ is the set of indices of other nodes cooperating with $j$ including $j$. In combine step the nodes' updated hyperparameters and/or point estimates $\hat{\theta}_k$ are combined.

### 6.1.1. Combine step when point estimates of $\hat{\theta}$ are given

DDE assumes decentralized scenario, where $j^{th}$ node updates its point estimate using point estimates of its neighbours (see [3]):

$$\hat{\theta}_j^* = \sum_{k \in \mathcal{N}_j} a_{jk} \hat{\theta}_k, \tag{10}$$

where $a_{jk}$ are weights assigned by $j^{th}$ source to its neighbours.

Since the method proposed in Section 2. assumes centralized scenario, we assume a collecting element (a device or another expert - not included in the original set of nodes) collects all point estimates and combines them. In case we have $s$ nodes the formula (10) looks as follows

$$\hat{\theta}_j^* = \sum_{k=1}^{s} a_k \hat{\theta}_k, \tag{11}$$

where $a_{j,t}$ are weights assigned by the collecting element to the nodes $j = 1, \ldots, s$.

## 6.2. Minimizer of the constrained minimum cross-entropy

The Lagrangian of the optimization task (3) is

$$\begin{aligned}
&\int_Q \pi(q|p_1,\ldots,p_s) \ln \frac{\pi(q|p_1,\ldots,p_s)}{\frac{1}{B(\nu_{01},\ldots,\nu_{0n})} \prod_{i=1}^n q_i^{\nu_{0i}-1} \prod_{i=1}^n q_i^{\sum_{j=1}^s \lambda_j (p_{ji}-p_{si})}} dq \\
&+ \sum_{j=1}^{s} \lambda_j \left(H(p_j) - H(p_s)\right) \\
&\pm \ln \frac{1}{B(\nu_{01} + \sum_j \lambda_j (p_{ji}-p_{si}), \ldots, \nu_{0n} + \sum_j \lambda_j (p_{ji}-p_{si}))} \\
= &\int_Q \pi(q|p_1,\ldots,p_s) \\
&\times \ln \frac{\pi(q|p_1,\ldots,p_s)}{\frac{1}{B(\nu_{01}+\sum_j \lambda_j (p_{ji}-p_{si}),\ldots,\nu_{0n}+\sum_j \lambda_j (p_{ji}-p_{si}))} \prod_{i=1}^n q_i^{\nu_{0i}+\sum_{j=1}^s \lambda_j (p_{ji}-p_{si})-1}} dq
\end{aligned}$$

$$+\ln\frac{\frac{1}{B(\nu_{01}+\sum_j\lambda_j(p_{ji}-p_{si}),\dots,\nu_{0n}+\sum_j\lambda_j(p_{ji}-p_{si}))}}{\frac{1}{B(\nu_{01},\dots,\nu_{0n})}}+\sum_{j=1}^{s}\lambda_j\left(H(p_j)-H(p_s)\right)$$

where $\lambda_j$, $j=1\dots,s$, are the Lagrange multipliers and $H(.)$ is the entropy.

Its minimizer

$$\hat{\pi}(q|p_1,\dots,p_s)=\frac{1}{B(\nu_{01}+\sum_j\lambda_j(p_{ji}-p_{si}),\dots,\nu_{0n}+\sum_j\lambda_j(p_{ji}-p_{si}))}\times\prod_{i=1}^{n}q_i^{\nu_{0i}+\sum_{j=1}^{s}\lambda_j(p_{ji}-p_{si})-1} \tag{12}$$

is the pdf of the Dirichlet distribution with parameters (4).

## REFERENCES


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

[2] K. Dedecius, V. Sečkárová. Dynamic Diffusion Estimation in Exponential Family Models. *IEEE Signal Processing Letters*, **20**, *11* (2013), 1114–1117.

[3] S. Frühwirth-Schnatter. Finite Mixture and Markov Switching Models (Springer Series in Statistics). 1st ed. Springer, Aug. 2006.

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

[5] V. Sečkárová On Supra-Bayesian Weighted Combination Of Available Data Determined By Kerridge Inaccuracy And Entropy. *Pliska Stud. Math. Bulgar.*, **22** (2013), 159–168.

[6] 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 of the CAS*
*Pod Vodárenskou věží 4*
*CZ-182 08 Prague 8*
*e-mail:*seckarov@utia.cas.cz

*Department of Probability*
*and Mathematical Statistics*
*Charles University in Prague*
*Sokolovká 83*
*Prague*