EMPIRICAL BAYESIAN APPROACH TO PROCESSING OF MEDICAL DATA

Anton Georgiev, Doctor of Science, Associate Professor at the Technical University of Varna,
georgiev_an@tu-varna.bg

Abstract: In this article are assayed opportunities provided by Bayesian parametric empirical procedure in situations which require reconciliation of identical data, derived from diversified sources of information for medical diagnostic tests, for various diseases, for the effects of new drugs on the health of patients and others. The proposed procedure is based on a discrete approximation of a priori law distribution law distribution.

Keywords: reconciliation of uniform information with medical essence, pooling of uniform information from various sources, synchronizing of data of a medical nature.

EMPIРИЧЕН БЕЙСОВСКИ ПОДХОД ПРИ РАБОТА С МЕДИЦИНСКИ ДАННИ

инж. Антон Славчев Георгиев, д.т.н., доц. в Технически университет - Варна,
georgiev_an@tu-varna.bg

Абстракт: В статията са анализирани възможностите, предоставяни от Бейсовската параметрична емпирична процедура, в ситуации при които е необходимо съвместяване на еднородни данни, постъпващи от диверсифицирани източници на информация за медико-диагностични изследвания, за различни заболявания, за въздействието на нови медикаменти върху здравословното състояние на пациенти. Предложената процедура се основава на дискретна апроксимация на априорното разпределение.

Ключови думи: съвместяване на еднотипна информация с медицински характер, обработка на медицинска информация от еднородни източници, синхронизиране на информация от медицинско естество.

Prelude

What is known [1] „a priori” and „a posteriori” are two of the basic concepts in theory of knowledge. A priori literally means “from before” or “from earlier.” This is because a priori knowledge depends upon what a person can derive from the world without needing to experience it. This is better known as reasoning. Of course, a degree of experience is necessary upon which a priori knowledge can take shape. Let’s look at an example. If you were in a closed room with no windows and someone asked you what the weather was like, you would not be able to answer them with any degree of truth. If you did, then you certainly would not be in possession of a priori knowledge. It would just be impossible to implement the reflections to bring forth intelligent solution. Of course, then, „a posteriori” literally means “from what comes later” or “from what comes after.” This is a reference to experience and using a different kind of reasoning (inductive) to gain knowledge. This kind of knowledge is gained by first having an experience (and the important idea in philosophy is that it is acquired through the five senses) and then using logic and reflection to derive understanding from it.
philosophy, this term is sometimes used interchangeably with empirical knowledge, which is knowledge based on observation. It is believed that "a priori" knowledge base is more reliable than "a posteriori" knowledge base. This may seem counterintuitive, since in the former case someone can just sit inside of a room and base their knowledge on factual evidence while in the latter case someone is having real experiences in the world. But the problem lies in this very fact: everyone’s experiences are subjective and open to interpretation.

The purpose of this article is to seek ways to unite the prior and current experimental data for the purpose of medical research [2]-[6]. In this sense, the ideas in this article are continuation of Ideas in [7], [8], [9].

Introduction

To achieve accurate and precise results from the medical tests is necessary as the high precision [10]-[12] and reliability of equipment [14]-[19], as well appropriate methods for processing of the medical statistical data [22]. Unfortunately, when performing medical research, teams doing survey still rarely applied mathematical methods to integrate information from different sources. Such pooling of information may be useful in tracking various diseases as well as in the study of the impact of new medicinal products on the patient's condition. This fact has its logical explanation. In contrast to technical objects in which the use of a priori information is known practice, in medicine deviation of a posterior result (due to not accurate or incorrectly interpreted of a priori information) could be fatal to the life and health of large groups of people. This circumstance requires seeking and developing sufficiently accurate and reliable mathematical methods. Such methods, which reduce to zero the probability of inappropriate or improper pooling of data coming from different sources. In the present article has been researched and analyzed exactly this kind of methods. Methods are based on the application of conditional probabilities in modelling the process of pooling of a priori data obtained from made until clinical trials and the latest information about the particular clinical trial.

Research Idea and Concept

After having defined function of verisimilitude [8] and is known distribution function $F(t;\theta)$ of the research parameter for aggregate of accumulated same kind data, Bayesian procedure may continue to give the corresponding estimates of the evaluated parameters.

Based on the specified results in [7], [8], from the N series of tests $T^{(1)}\ldots T^{(N)}$, by volume, respectively $n_1\ldots n_N$, the information is represented by a vector $T^{(i)}$.

Each vector $T^{(i)}$ combines in itself two vectors: vector of the empirical information of the respective packet data $T^{(i)}$ and the vector of censoring $T^{C(i)}$. The last two vectors have a volume, respectively $d_i$ and $k_i$, and the sum of their volume determines the volume of the vector $T^{(i)}$, i.e. $n_i=d_i+k_i$ [8], [9]. Empirical Bayesian assessment of the evaluated parameter $R(t;\theta)=1-F(t;\theta)$ of available medical data should be established for the sample $T^{(N)}$, taking into account the results of tests made for the previous $N-1$ series. As initial information we assume that are available data contained empirical data to impact treatment of $n$ patient. In the testing of each $j$-th patient can be accumulated single-objective information (for example, the impact of a medicament or for the effect of the application of a therapy, etc.) or has been collected two-parametric information (e.g., a second time parameter). It may be accumulated and three-parametric information: the impact of treatment on the time of reporting the outcome of treatment and parameters of censoring results. For illustration, assume that is available two-parametric information like: moment in which impact is recorded (or is not accounted for desired
Impact of a drug (therapy) $T'_j$, and data for censoring $T_{ij}$, $(j=1, ..., n)$. In the general case, in the absence of censorship, the moments in which is accounted the impact of the drug (treatment) $T'_j$, are mutually-independent, evenly distributed random variables, with distribution function $F(t; \theta)$ [23]. The sample $T=(T_1, T_2, ..., T_n)$, corresponding to the general plan of the test, has the look:

$$T=(T'_{j1}, T'_{j2}, ..., T'_{jd}, T_{111}, T_{112}, T_{113}, ..., T_{11k}), \quad d+k=n$$

where $J=(j_1, j_2, ..., j_d)$ is a plurality of containing the numbers of patients in whom it is established the expected effect of treatment, and plurality of $I=(i_1, i_2, ..., i_k)$ contains the number of patients whose treatment data were censored. For the purposes of parametric Bayesian evaluation, whole multitude of the values $T=(T'_1, T'_2, ..., T'_n, T_{111}, T_{112}, T_{113}, ..., T_{11n})$ is modelled using a random vector. The function of the distribution of the multiplicity of random values is conditionally distribution $T_{111}, T_{112}, T_{113}, ..., T_{11n}$. The condition is that the values of the moments in which has been established (not identified) the desired effect of the drug (treatment) satisfy the condition $T'_{j1}>T_{111}, T'_{j2}>T_{112}, ..., T'_{jd}>T_{11k}$, and the data gathered on the impact satisfy the condition $T'_{j1}\leq T_{111}, T'_{j2}\leq T_{112}, ..., T'_{jd}\leq T_{11d}$. The conditional distribution $T_{111}, T_{112}, T_{113}, ..., T_{11n}$, does not depend on $\theta$ and can depend solely on the values $T'_{j1}, T'_{j2}, ..., T'_{jd}$. In this case, if there is a joint density distribution of values $T_{11j}$, then the conditional density distribution in force ratio type:

$$P(T_{111}, T_{112}, T_{113}, ..., T_{11n} \mid T'_{1}, T'_{2}, ..., T'_{n})=P(T_{111}, T_{112}, T_{113}, ..., T_{11n} \mid T'_{j1}, T'_{j2}, ..., T'_{jd}),$$

where:

$$T^*_{j1} \leq T_{11j}, \quad \forall j \in J,$$

$$T^*_{i} > T_{11i}, \quad \forall i \in I.$$  

Plan for the collection, accumulation and processing of empirical information with censorship, for which is satisfied the formula (5), in specialized literature is called the plan with uninformative censoring. It should be noted also that the ratio (5) is satisfied if the dataset $T'_1, T'_2, ..., T'_n$ and the set of data for censoring $T_{111}, T_{112}, T_{113}, ..., T_{11n}$ are mutually independent data packets, i.e. when the value of $T^*_j$ is an independent variable and in terms of $j$ and in terms of all values $T_{11i}$ ($i, j=1, ..., n$). This assumption is absolutely correct in studying the impact of treatment, because the data obtained can be represented by Boolean model - the treatment is influenced positively and it is found that the patient's condition has improved or treatment is not impacted and does not establish a change in the patient's condition. It is only necessary to note which of the two events was occurred. The thus obtained sample will be in the type $T=(T_1, T_2, ..., T_n)$, where $T_j = \min \{T^*_j, T_{11j}\}$. Event $T_j = T^*_j$ ($T_j = T_{11j}$) reflects the occurrence of the event "a positive response to treatment."

The existence of parametric family allows through the aggregated expressions in the Bayesian theory (1) and (2), to be written verisimilitude function $l(T \mid \theta)$ for a random sample $T$. Whereupon, based on statistics $\Phi$, this function can be written as (4)

$$l(\theta \mid T) = K(T)l_0(\theta; \Phi).$$
Evaluation, which is interesting for us, can be obtained, if known posterior density parameter $\theta$ of the test results obtained from the $N$-th series. Then, in the basic Bayesian theorems, is obtained:

$$h(\theta|T^{(N)}) = h(\theta|\Phi^{(N)}) \sim h(\theta)l_0(\theta; \Phi^{(N)}). \quad (5)$$

In the common case, if we assume that a priori density $h(\theta)$ is not known, it is necessary to carry out approximation. For this purpose we assume that there exist assessments $\hat{\theta}^{(1)}, \hat{\theta}^{(2)}, \ldots, \hat{\theta}^{(i)}$ of each $\hat{\theta}^{(i)}$ of which was obtained for relevant sample $T^{(i)}$ (such as $i = 1+N$). To be obtained the assessments can be used all available for this purpose range of statistical methods. Thus, for example using the method of maximum verisimilitude, the components of the assessment $\hat{\theta}^{(i)}$ are determined by means the decisions of the system of equations:

$$\frac{\partial l(\theta; \Phi^{(i)})}{\partial \theta_j} = 0 \Rightarrow \hat{\theta}_j = \theta_j, \; j = 1 \div m \quad , \quad (6)$$

where $m$ is the dimensionality of the vector $\theta$. In this case suitable can be a standard Bayesian method, according to which, the assessment $\hat{\theta}^{(i)}_j$ is determined as the mean posterior for the sample $T^{(i)}$:

$$\hat{\theta}^{(i)}_j = \frac{1}{\beta_j} \int \phi_j l_0(\theta; \Phi^{(i)}) h_1(\theta) d\theta, \; j = 1 \div m, \quad (7)$$

where: $\beta_i$ is a normalization factor;

$h_i(\theta)$ is priori density of the parameter $\theta$, which may exist if it is produced in the testing of the $i$-th series medical tests. If a priori information about the $i$-th series is missing, as a priori density $h_i(\theta)$ of the parameter $\theta$ may be used as the "informative" density of Harold Jeffreys [24], [25].

To approximate the $h(\theta)$ in the formulas (5) can be used non-parametric discrete evaluation of the density of which:

$$d\hat{H}(\theta) = \frac{1}{n} \sum_{i=1}^{N} n_i \alpha(\hat{\theta}^{(i)}, \theta), \quad (8)$$

where:

$$\alpha(x, y) = \begin{cases} 1, & \text{ako } x = y \\ 0, & \text{ako } x \neq y \end{cases} \quad , \quad (9)$$

$$n = \sum_{j=1}^{N} n_j. \quad (10)$$

Assessment carried out according (8) summarizes used by Bennett [26] evaluation for the case of different sizes of samples $T^{(i)}$. The corresponding evaluation of the empirical function of a priori distribution $\hat{H}(\theta)$ is a step function with increasing $n_i/n$ for each point $\hat{\theta}^{(i)}$ (for each of the assessments $\hat{\theta}_1 \div \hat{\theta}_{N-1}$). It can find the average value of the posterior point evaluation by approximation of a priori distribution:

---

3 The method, which Bennett offers, is connected to an approximation of a priori function of distribution. For this purpose, are used stepped features that increase by $1/N$ for each of the evaluations.
where \( \Phi^{(N)} \) are required statistical data corresponding to \( T^{(N)} \). By using the formula (8) to be simplifying the formula (11) and be introducing some transformations [23], the formula of the average value of the posterior point estimate (11) gets type:

\[
\hat{R}_c^*(t_0) = \frac{\sum_{i=1}^{N} n_i R(t_0; \hat{\theta}^{(i)}) l_0(\hat{\theta}^{(i)}; \Phi^{(N)})}{\sum_{i=1}^{N} n_i l_0(\hat{\theta}^{(i)}; \Phi^{(N)})}.
\] (12)

In a similar manner can be determined empirical evaluation of a priori dispersion:

\[
\sigma^2_{R_c^*(t_0)} = \frac{1}{\sum_{i=1}^{N} n_i l_0(\hat{\theta}^{(i)}; \Phi^{(N)})} - \hat{R}_c^*(t_0).
\] (13)

The developed procedure allows deriving empirical Bayesian estimates with second, third, fourth, etc. approximations. For this purpose it is necessary to determine the empirical Bayesian evaluation of the parameter \( \hat{\theta}^{(i)} \) for a first approximation:

\[
\hat{\theta}^{(i)}_j = \frac{\sum_{k=1}^{i} n_k \hat{\theta}^{(k)}_j l_0(\hat{\theta}^{(k)}; \Phi^{(i)})}{\sum_{k=1}^{i} n_k l_0(\hat{\theta}^{(k)}; \Phi^{(i)})}.
\] (14)

After which in formulas (12) and (13) instead of \( \hat{\theta}^{(i)} \) is placed the assessment of the first approximation \( \hat{\theta}^{(i)}_j \).

For the parametric family of exponential distributions \( F(t; \lambda) = 1 - \exp(-\lambda t) \), the function of the verisimilitude for statistical sampling \( T^{(i)} \) has the types

\[
l(\lambda|T^{(i)}) \sim l_0(\lambda; d_i; k_i) = \lambda^{d_i} \exp(-\lambda k_i),
\] (15)

where: \( k_i = T^{(i)}_1 + T^{(i)}_2 + \ldots + T^{(i)}_m \), \( d_i \) - number of patients who has not been ascertained improvement status. Considering the formula (15) and be administered formulas (12) and (13) for Bayesian empirical point estimation of the intensity of in the cases in which there has not been established a positive effect of treatment \( \lambda \), and for empirical evaluation of posterior dispersion (if we have an exponential distribution) can be written:

\[
\hat{\lambda}_c^* = \frac{\sum_{i=1}^{N} n_i \hat{\lambda}_i^{d_i} \exp(-\hat{\lambda}_i k_N)}{\sum_{i=1}^{N} n_i \hat{\lambda}_i^{d_i} \exp(-\hat{\lambda}_i k_N)},
\] (16)
\[ \hat{\lambda}_c = \frac{\sum_{i=1}^{N} n_i \hat{\lambda}_i \hat{\lambda}_i^{d \lambda N+1} \exp(-\hat{\lambda}_i k_N)}{\sum_{i=1}^{N} n_i \hat{\lambda}_i^{d \lambda N} \exp(-\hat{\lambda}_i k_N)}, \]

\[ \sigma^2_{\hat{\lambda}_c} = \frac{\sum_{i=1}^{N} n_i \hat{\lambda}_i^{d \lambda N} \exp(-\hat{\lambda}_i k_N)}{\sum_{i=1}^{N} n_i \hat{\lambda}_i^{d \lambda N} \exp(-\hat{\lambda}_i k_N)} - \hat{\lambda}_c^2, \]

\[ \sigma^2_{\hat{\lambda}_c} = \frac{\sum_{i=1}^{N} n_i \hat{\lambda}_i^{d \lambda N+2} \exp(-\hat{\lambda}_i k_N)}{\sum_{i=1}^{N} n_i \hat{\lambda}_i^{d \lambda N} \exp(-\hat{\lambda}_i k_N)} - \hat{\lambda}_c^2. \]

In an analogous way, if it is known that \( R(t_0) = \exp(-\hat{\lambda} t_0) \) and taking into account formulas (15), (12) and (13), we can write Bayesian empirical point estimate probability positive effect of the treatment (or the application of the new medicine) and evaluation of posterior dispersion:

\[ \hat{R}_c^*(t_0) = \frac{\sum_{i=1}^{N} n_i \hat{\lambda}_i^{d \lambda N} \exp(-\hat{\lambda}_i k_N)}{\sum_{i=1}^{N} n_i \hat{\lambda}_i^{d \lambda N} \exp(-\hat{\lambda}_i k_N)} \]

\[ \hat{R}_c^*(t_0) = \frac{\sum_{i=1}^{N} n_i \hat{\lambda}_i^{d \lambda N} \exp[-\hat{\lambda}_i (t_0 + k_N)]}{\sum_{i=1}^{N} n_i \hat{\lambda}_i^{d \lambda N} \exp(-\hat{\lambda}_i k_N)} \]

\[ \sigma^2_{\hat{R}_c^*(t_0)} = \frac{\sum_{i=1}^{N} n_i \exp(-2\hat{\lambda}_i t_0) \hat{\lambda}_i^{d \lambda N} \exp(-\hat{\lambda}_i k_N)}{\sum_{i=1}^{N} n_i \hat{\lambda}_i^{d \lambda N} \exp(-\hat{\lambda}_i k_N)} - \hat{R}_c^2(t_0). \]

Final remarks

In most of Bayesian procedures are available for determining the evaluation \( \hat{\lambda}_i (i=1+\ldots+N) \) to be used in the method of maximum verisimilitude. If it is found the solution of the equation of the verisimilitude for this case will be an estimate of type \( \hat{\lambda}_i = d_i/k_i \). It is entirely possible situation where during testing, treatment has impacted positively on all patients. Then that
evaluation \( \hat{\lambda}_i \), obtained by the method of maximum verisimilitude, have a zero value. In cases where a large number of tests have evaluations \( \hat{\lambda}_i = 0 \) strongly reduces the effectiveness of the displayed formulas (17), (19), (21) and (23). I.e. border case, when none of the tests have not been established lack of positive effect (or effect of treatment is positive in all patients, included in all tests) above formulas become inapplicable. Then it is more appropriate if the evaluation \( \hat{\lambda}_i \) is obtained by Bayesian assessment of the sample \( T^{(i)} \), corresponding to the case for "trivial prior information". This assessment can be determined by adopting the even priori distribution for \( \lambda \) in \([0; \infty]\). In this case evaluation is:

\[
\hat{\lambda}_i^* = \frac{d_i + 1}{k_i},
\]

whereby the formulas (17), (19), (21) and (23) are applicable in all data about every possible \( N \) series tests.

The presented procedure is summarized and is based on a discrete approximation of a priori distribution. As a disadvantage of the proposed procedure (the deficiency is common to all procedures using for approximation priori distribution) can be mentioned that it is not possible to determine the interval estimates. Theoretically, to find of \( R_\delta^* \) could use some of the known methods of approximate calculations (using the resulting point estimate \( \hat{R}_{\delta}^*(t_0) \) and mean square deviation \( \sigma_{\delta}^2(\hat{R}_{\delta}^*(t_0)) \), but given the mandatory in medical research accuracy, correctness and precision of the results, methods of approximate calculations must be avoided.

**Conclusion**

The impossibility to be obtained correct estimation of the accuracy of the results obtained sometimes done unsuitable the empirical Bayesian nonparametric methods for use in medical research. This is the reason why in this article emphasis was placed on the ability to use the parametric empirical Bayesian methods. Presented in the article Bayesian approach is an attempt of combining the homogeneous information coming from diversified medical sources of reliable empirical data. For the purpose of research are written formulas for assessing the accuracy and correctness of the posterior medical information. They are based on empirical Bayesian point estimates. The reasons to prefer empirical Bayesian point estimates instead of empirical Bayesian interval estimates are three:

- obtaining of empirical Bayesian point estimates are based on more accessible mathematical tools;
- point estimates are more understandable for professionals who do not deal directly with statistics;
- in the practice empirical Bayesian point estimates were found significantly greater use; this allows easier comparison of data obtained through this approach results with the results of similar studies in the world (which almost always are given by point estimates).

**References**


