Journal of Statistical Planning and Inference 139 (2009) 4051 – 4063

Contents lists available at ScienceDirect

Journal of Statistical Planning and Inference

journal homepage: www.elsevier.com/locate/jspi

Optimal use of historical information

Bhaskar Bhattacharya[1]

Department of Mathematics, Southern Illinois University Carbondale, IL 62901-4408, USA

ARTICLE INFO

Article history:
Received 17 July 2008
Received in revised form
27 April 2009
Accepted 13 May 2009
Available online 22 May 2009

Keywords:
Bayesian
Efficient rules
Historical data
Kullback–Leibler divergence
Optimization
Posterior
Power prior
Quality-adjusted rule

ABSTRACT

When historical data are available, incorporating them in an optimal way into the current data analysis can improve the quality of statistical inference. In Bayesian analysis, one can achieve this by using quality-adjusted priors of Zellner, or using power priors of Ibrahim and coauthors. These rules are constructed by raising the prior and/or the sample likelihood to some exponent values, which act as measures of compatibility of their quality or proximity of historical data to current data. This paper presents a general, optimum procedure that unifies these rules and is derived by minimizing a Kullback–Leibler divergence under a divergence constraint. We show that the exponent values are directly related to the divergence constraint set by the user and investigate the effect of this choice theoretically and also through sensitivity analysis. We show that this approach yields '100% efficient' information processing rules in the sense of Zellner. Monte Carlo experiments are conducted to investigate the effect of historical and current sample sizes on the optimum rule. Finally, we illustrate these methods by applying them on real data sets.

1. Introduction

Incorporating useful information from past similar studies can enhance the quality of current data analysis. In many cancer and AIDS clinical trials, current studies often use treatments that are very similar to or slight modifications of the treatments used in previous studies. The number of students enrolled each year at a university changes for various reasons. When analyzing the current enrollment data it would be useful to incorporate the information from previous years. We refer to data arising from previous similar studies as *historical data*.

In Bayesian analysis one specifies priors to utilize the available information. For a given prior π and a sample likelihood ℓ, the posterior distribution using Bayes' theorem is given by $f \propto \pi\ell$. Zellner (1997a,b, 2002) (see other references therein) recognized that the quality of the inputs, the prior and the sample information, might vary. He suggested updating the information rule (posterior distribution) f by $\tilde{f} \propto q_1(\pi)q_2(\ell)$, where $q_1(\pi)$ is the 'quality-adjusted' prior and $q_2(\ell)$ is the 'quality-adjusted' likelihood function. One choice is to use $q_1(\pi) \propto \pi^a$, $q_2(\ell) \propto \ell^b$, $0 \le a, b \le 1$. Thus, for example, when the prior (sample likelihood) information is of very low quality, one can use $a = 0$ ($b = 0$), and then $\tilde{f} \propto \ell^b$ ($\tilde{f} \propto \pi^a$). Quality-adjusted posteriors are also desirable when it is appropriate to weigh the prior and sample information differently (Zellner, 2002). Often raising the prior and/or the likelihood to a fractional power corresponds to their higher dispersion. Thus quality-adjusted rules are useful when one associates 'low quality' with 'higher dispersion' in contrast to the situation when the prior and/or the likelihood are of very high quality ($a = b = 1$).

E-mail address: bhaskar@math.siu.edu
[1] Supported in part by NSF grant DMS - 0706041.

0378-3758/$ - see front matter
doi:10.1016/j.jspi.2009.05.009

Another way to incorporate available past information in current data analysis is to use an informative prior. One kind of informative priors are *power priors*. Suppose we denote the historical data set by D_0 and the likelihood of D_0 by $L(\theta, D_0)$. Further, let $\pi_0(\theta)$ be the (initial) prior distribution for θ before D_0 is observed. Then the power prior is given by $(L(\theta, D_0))^{a_0}\pi_0(\theta)$ where $0 \le a_0 \le 1$ is a scalar parameter that controls the influence of the historical data on the current data. Under the power prior, the posterior distribution of θ is given by

$$\pi(\theta|D, D_0, a_0) = L(\theta, D)(L(\theta, D_0))^{a_0}\pi_0(\theta), \tag{1.1}$$

where $L(\theta, D)$ is the likelihood of the current data set D. Ibrahim et al. (2003) show that the power priors are optimal in the sense that the distribution in (1.1) minimizes a convex sum of Kullback–Leibler divergence between two posterior densities: one based on 'pooled historical and current data' ($a_0 = 1$) and the other based on 'not using the historical data at all' ($a_0 = 0$). It is also shown that the power priors are 100% efficient according to the optimal information processing rules of Zellner (1988), thus, the ratio of the output to input information is 1.

Ibrahim et al. (2003, eq. (29)) suggest an expression for a_0 to be used as optimal when a single historical data set is available and suggest more research is needed to find the suitable value of a_0. In this paper, we show that a well-justified value of a_0 can be obtained by considering constraints relating divergence from a specified distribution when a single historical data set is available. This gives a clear interpretation of the exponent a_0 of the power prior of Ibrahim et al. (2003) in terms of distances from the current and historical data, see Remark 3. In particular, our approach is to minimize the Kullback–Leibler divergence from the historical posterior to a class of distributions which are a given divergence away from the current posterior. It is up to the user to decide how far from the current posterior he/she would like to be. If the value of this divergence is set at r, then the value a_0 depends on r. In Theorem 2, we demonstrate the effect of the choice r on the final solution. In Section 2.6, we discuss the choice of r and conduct related sensitivity analyses. The methods developed in this article are applicable in Bayesian and frequentist situations. To calibrate the divergence so that one can set r appropriately, one may consider the interesting scheme from McCulloch (1989).

Power priors have been studied by several authors other than those mentioned earlier. Ibrahim and Chen (2000) studied power priors in regression situation. Chen et al. (2000) showed power priors are proper for a wide range of models under some general conditions. Ibrahim et al. (1998) showed how to use historical data for trend test in presence of covariates. Walker et al. (2004) considered Bayesian models where the prior puts positive mass on all Kullback–Leibler neighborhoods of all densities. These neighborhoods are similar to the constraint regions in Section 2; however, these authors consider the second argument of the divergence as unknown whereas in this paper the first argument is considered unknown. Similar neighborhoods appear in other Bayesian nonparametric literature (Wasserman, 1998; Ghosal et al., 1999; Barron et al., 1999).

In Section 2, we show in Remarks 1–3 that both quality-adjusted and power prior rules can be derived as special cases of our procedure. In Section 3, we show that the optimal solutions derived in this paper are 100% efficient by modifying the definitions of Zellner (1997a,b, 2002). In Section 4, we consider simulation studies, which investigate the effect of sample sizes on the exponents of historical and current likelihoods. In Section 5, we apply the methods developed on two real data sets. We end with final comments in Section 6.

2. Optimum solution

Let $\mathscr{D}$ be the set of all probability density functions (pdfs) on $\mathscr{R}$. For pdfs $f, g \in \mathscr{D}$, the Kullback–Leibler divergence, or simply, *divergence* between f and g is defined as

$$I(f|g) = \int f(t)\ln\frac{f(t)}{g(t)}\,dt.$$

It is well known that $I(f|g) \ge 0$, and $=0$ if and only if $f = g$. Here and in the sequel we observe the conventions that $\ln 0 = -\infty$, $\ln(a/0) = +\infty$, $0 \cdot (\pm\infty) = 0$.

2.1. Main results

For given pdfs g, h, we consider the infinite dimensional optimization problem of finding the pdf f which solves

$$\inf I(f|g) \tag{2.1}$$

subject to

$$\{f : I(f|h) \le r\}, \tag{2.2}$$

where r is a given nonnegative constant. If $r \ge r^* = I(g|h)$, then the solution to (2.1) is g, and, if $r = 0$, then the solution to (2.1) is h. Thus, in some sense, the solution, say f^*, is 'between' g and h, and at a divergence r (or less) from h.

If the value of (2.1) is denoted by $I(r)$, the following Theorem states two important properties of $I(r)$. The result and proof of Theorem 1 are similar to Blahut (1974, Theorems 2, 3), who considered the discrete situation; however, our interest is in continuous probability densities.