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 © 2009 Elsevier B.V. All rights reserved.
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 $0 \leq a_0 \leq 1$ 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 $I(f|g) \geq 0$, 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

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

where r is a given nonnegative constant. If $r \geq r^* = I(g|h)$, 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.

Theorem 1. *$I(r)$ is nonincreasing and convex in r.*

Proof. Let $0 < r_1 \le r_2$. Then $\{f : I(f|h) \le r_1\} \subset \{f : I(f|h) \le r_2\}$. Then $I(r_1) \ge I(r_2)$, and thus, $I(r)$ is nonincreasing.

To show that $I(r)$ is convex in r, let there exist f_1, f_2 such that $I(f_1|g) = I(r_1)$, $I(f_2|g) = I(r_2)$ and $f_\alpha = \alpha f_1 + (1-\alpha) f_2$ for $0 \le \alpha \le 1$. Since $I(\cdot|h)$ is known to be convex in the first argument, $I(f_\alpha|h) \le \alpha I(f_1|h) + (1-\alpha) I(f_2|h) \le \alpha r_1 + (1-\alpha) r_2$. Thus $f_\alpha \in \{f : I(f|h) \le \alpha r_1 + (1-\alpha) r_2\}$. Also,

$$\begin{aligned} I(\alpha r_1 + (1-\alpha) r_2) &\le I(f_\alpha|g) \\ &\le \alpha I(f_1|g) + (1-\alpha) I(f_2|g) \\ &= \alpha I(r_1) + (1-\alpha) I(r_2). \end{aligned}$$

This proves the result. □

The next result shows that the infinite dimensional problem of finding f in (2.1) can be solved by considering a *dual* problem, which is a function of a scalar $\lambda \in \mathscr{R}^+$, and the two solutions are also related.

Theorem 2. *The solution to* (2.1) *is given by*

$$f^*(t) = \frac{(g(t))^{1/(1+\lambda^*)} (h(t))^{\lambda^*/(1+\lambda^*)}}{\int (g(t))^{1/(1+\lambda^*)} (h(t))^{\lambda^*/(1+\lambda^*)}\, dt}, \tag{2.3}$$

where λ^* solves

$$\inf_{\lambda \in \mathscr{R}^+} \left\{ \lambda r + (1+\lambda) \ln \int (g(t))^{1/(1+\lambda)} (h(t))^{\lambda/(1+\lambda)}\, dt \right\}. \tag{2.4}$$

Proof. The Lagrangian for the problem in (2.1) is given by

$$L(f, \lambda) = \int f(t) \ln \frac{f(t)}{g(t)}\, dt + \lambda \left(\int f(t) \ln \frac{f(t)}{h(t)}\, dt - r \right), \tag{2.5}$$

where λ is the Lagrangian multiplier. The associated *dual* problem is given by

$$\sup_{\lambda \in \mathscr{R}^+} \inf_{f \in \mathscr{D}} L(f, \lambda). \tag{2.6}$$

We can write

$$\begin{aligned} L(f, \lambda) &= \int f(t) \ln \frac{f(t)}{g(t)}\, dt + \lambda \left(\int f(t) \ln \frac{f(t)}{h(t)}\, dt - r \right) = -\lambda r + \int f(t) \ln \frac{f(t)}{g(t)}\, dt + \int f(t) \ln \left(\frac{f(t)}{h(t)} \right)^{\lambda} dt \\ &= -\lambda r + \int f(t) \ln \left(\frac{(f(t))^{1+\lambda}}{(g(t) h(t)^{\lambda})} \right) dt = -\lambda r + (1+\lambda) \int f(t) \ln \left(\frac{f(t)}{(g(t))^{1/(1+\lambda)} (h(t))^{\lambda/(1+\lambda)}} \right) dt, \end{aligned}$$

where the second term involves a divergence which is minimum when

$$f(t) = \frac{(g(t))^{1/(1+\lambda)} (h(t))^{\lambda/(1+\lambda)}}{\int (g(t))^{1/(1+\lambda)} (h(t))^{\lambda/(1+\lambda)}\, dt}.$$

Using this f in the infimum in (2.6), the dual problem can be expressed as

$$\sup_{\lambda \in \mathscr{R}^+} \left\{ -\lambda r - (1+\lambda) \ln \int (g(t))^{1/(1+\lambda)} (h(t))^{\lambda/(1+\lambda)}\, dt \right\}. \tag{2.7}$$

The theorem follows (Rockafeller, 1976, 1974; Ben-Tal et al., 1988) by taking the minus sign out of the braces in (2.7). □

The value λ^*, which solves (2.4), can be determined easily by numerical methods. The individual influences of the densities g, h on the final solution f^* are controlled by r through λ^*. The next result shows that the power of h in (2.3), i.e., $\lambda/(1+\lambda) = \eta$, say, is a nonincreasing function of r.

Theorem 3. *Suppose there exists r_0 such that there exists $f_0(t) = (g(t))^{1-\eta} (h(t))^{\eta}/c$ where $c = \int (g(t))^{1-\eta} (h(t))^{\eta}\, dt$, $0 \le \eta \le 1$ and η is chosen as a function of r_0 so that $f_0(t)$ achieves equality in the constraints defined by* (2.2). *In addition, suppose there exists δ_0 such that $|r - r_0| < \delta_0$ for any r. Then η is a nonincreasing function of r.*

Proof. Since $c = \int g(t)^{1-\eta} h(t)^{\eta}\, dt = \int g(t)(h(t)/g(t))^{\eta}\, dt$, we get

$$\begin{aligned} I(f_0|h) &= \int \frac{g(t)^{1-\eta}h(t)^{\eta}}{c} \ln\left(\frac{g(t)^{1-\eta}h(t)^{\eta}/c}{h(t)}\right) dt = \frac{1}{c}\int g(t)^{1-\eta}h(t)^{\eta}\left(\ln\frac{g(t)^{1-\eta}h(t)^{\eta}}{h(t)} - \ln c\right) dt \\ &= \frac{1}{c}\int g(t)^{1-\eta}h(t)^{\eta}\ln\left(\frac{g(t)^{1-\eta}h(t)^{\eta}}{h(t)}\right) dt - \ln c = \frac{1}{c}\int g(t)\left(\frac{h(t)}{g(t)}\right)^{\eta}\ln\left(\frac{g(t)}{h(t)}\right)^{1-\eta} dt - \ln c \\ &= \frac{1-\eta}{c}\int g(t)\left(\frac{h(t)}{g(t)}\right)^{\eta}\ln\left(\frac{g(t)}{h(t)}\right) dt - \ln c. \end{aligned}$$

Hence it follows that

$$\begin{aligned} \frac{\partial r}{\partial \eta} &= \frac{\partial}{\partial \eta} I(f_0|h) \\ &= \frac{\partial}{\partial \eta}\left\{\frac{1-\eta}{c}\int g(t)\left(\frac{h(t)}{g(t)}\right)^{\eta}\ln\left(\frac{g(t)}{h(t)}\right) dt - \ln c\right\} \\ &= \frac{1}{c^2}\left[c\left(-\int g(t)\left(\frac{h(t)}{g(t)}\right)^{\eta}\ln\left(\frac{g(t)}{h(t)}\right) dt - (1-\eta)\int g(t)\left(\frac{h(t)}{g(t)}\right)^{\eta}\left(\ln\frac{h(t)}{g(t)}\right)^2 dt\right)\right. \\ &\quad \left. + (1-\eta)\left(\int g(t)\left(\frac{h(t)}{g(t)}\right)^{\eta}\ln\left(\frac{g(t)}{h(t)}\right) dt\right)^2 - c\int g(t)\left(\frac{h(t)}{g(t)}\right)^{\eta}\ln\left(\frac{h(t)}{g(t)}\right) dt\right] \\ &= \frac{1}{c^2}\left[c\left(-(1-\eta)\int g(t)\left(\frac{h(t)}{g(t)}\right)^{\eta}\left(\ln\frac{g(t)}{h(t)}\right)^2 dt\right) + (1-\eta)\left(\int g(t)\left(\frac{h(t)}{g(t)}\right)^{\eta}\left(\ln\frac{h(t)}{g(t)}\right) dt\right)^2\right] \\ &= -\frac{1-\eta}{c^2}\left[c\left(\int g(t)\left(\frac{h(t)}{g(t)}\right)^{\eta}\left(\ln\frac{h(t)}{g(t)}\right)^2 dt\right) - \left(\int g(t)\left(\frac{h(t)}{g(t)}\right)^{\eta}\left(\ln\frac{h(t)}{g(t)}\right) dt\right)^2\right] \\ &= -(1-\eta)\left[\frac{1}{c}\int g(t)\left(\frac{h(t)}{g(t)}\right)^{\eta}\left(\ln\frac{h(t)}{g(t)}\right)^2 dt - \left(\frac{1}{c}\int g(t)\left(\frac{h(t)}{g(t)}\right)^{\eta}\left(\ln\frac{h(t)}{g(t)}\right) dt\right)^2\right] \\ &= -(1-\eta)\left[\int\left(\ln\frac{h(t)}{g(t)}\right)^2 dP - \left(\int \ln\frac{h(t)}{g(t)}\, dP\right)^2\right] = -(1-\eta)\,\mathrm{var}_P\left(\ln\frac{h(t)}{g(t)}\right) \le 0 \end{aligned}$$

(where $dP/dt = g(t)(h(t)/g(t))^{\eta}/c$). Since $(\partial\eta/\partial r) = (\partial r/\partial\eta)^{-1}$, it follows that $\partial\eta/\partial r \le 0$. Also, from $I(f_0|h) = r$, if we set $r = 0$ then $f_0 = h$ which means $\eta = 1$; if we set $r = I(g|h)$ (or larger), then $f_0 = g$ which means $\eta = 0$. Thus $\eta \in [0,1]$ is a nonincreasing function of r. □

In the following remarks, we apply the previous results in a Bayesian context. When g, h are both posterior distributions as in Remarks 1–3, we refer to the solution in (2.3) as a 'power posterior'. Recall, the (initial) prior distribution is denoted by $\pi_0(\theta)$.

Remark 1 (*Zellner's quality-adjusted information rule*). If we set the historical posterior as $g(\theta) \propto L(\theta|D_0)\pi_0(\theta)$ and the current posterior as $h(\theta) \propto L(\theta|D)\pi_0(\theta)$, then the power posterior from (2.3) is

$$f^*(\theta) \propto [L(\theta|D_0)]^{1-\eta}[L(\theta|D)]^{\eta}\pi_0(\theta), \tag{2.8}$$

where $0 \le \eta \le 1, \eta = \lambda^*/(1+\lambda^*)$ where λ $(=\lambda^* \in \mathscr{R}^+)$ minimizes

$$\begin{aligned} &\lambda r + (1+\lambda)\ln\int\left[\frac{L(\theta|D_0)\pi_0(\theta)}{\int L(\theta|D_0)\pi_0(\theta)\,d\theta}\right]^{1/(1+\lambda)}\left[\frac{L(\theta|D)\pi_0(\theta)}{\int L(\theta|D)\pi_0(\theta)\,d\theta}\right]^{\lambda/(1+\lambda)} d\theta \\ &= \lambda r + (1+\lambda)\left[\ln\int L(\theta|D_0)^{1/(1+\lambda)}L(\theta|D)^{\lambda/(1+\lambda)}\pi_0(\theta)\,d\theta\right. \\ &\quad \left. - \ln\left(\int L(\theta|D_0)\pi_0(\theta)\,d\theta\right)^{1/(1+\lambda)} - \ln\left(\int L(\theta|D)\pi_0(\theta)\,d\theta\right)^{\lambda/(1+\lambda)}\right] \\ &= \lambda r + (1+\lambda)\ln\int L(\theta|D_0)^{1/(1+\lambda)}L(\theta|D)^{\lambda/(1+\lambda)}\pi_0(\theta)\,d\theta \\ &\quad - \ln\int L(\theta|D_0)\pi_0(\theta)\,d\theta - \lambda\ln\int L(\theta|D)\pi_0(\theta)\,d\theta. \end{aligned} \tag{2.9}$$

Leaving out the third term (which is free of λ), λ $(=\lambda^* \in \mathscr{R}^+)$ minimizes

$$\lambda(r-c)+(1+\lambda)\ln\int L(\theta|D_0)^{1/(1+\lambda)}L(\theta|D)^{\lambda/(1+\lambda)}\pi_0(\theta)\,d\theta, \tag{2.10}$$

where $c=\ln\int L(\theta|D)\pi_0(\theta)\,d\theta$. Setting $\pi_0(\theta)=1,\ g=\pi,\ h=\ell$ in (2.8), yields the quality-adjusted information rule of Zellner when $a+b=1$. Thus Theorem 2 gives us an interpretation of a,b based on 'distance' from the current/historical likelihood, that is, the solution is at a distance of $100(r/r^*)\%$ of $r^*=I(g|h)$ from h. From Theorem 2, the influence of the current and historical data likelihoods on the final solution $f^*(\theta)$ can be controlled by changing r. In calculation, sometimes, it may be easier to use (2.9) directly than (2.10). One would like to choose r a smaller value (than larger) so that more weight is put on current data.

Remark 2 (*Zellner's rule continued*). One can interchange the roles of g and h in Remark 1, which produces from (2.3) as

$$f^*(\theta)\propto[L(\theta|D)]^{1-\eta}[L(\theta|D_0)]^{\eta}\pi_0(\theta), \tag{2.11}$$

where $0\leq\eta\leq 1,\ \eta=\lambda^*/(1+\lambda^*)$, λ^* solves (similar derivation as (2.10))

$$\inf_{\lambda\in\mathscr{R}^+}\left\{\lambda(r-d)+(1+\lambda)\ln\int[L(\theta|D)]^{1/(1+\lambda)}[L(\theta|D_0)]^{\lambda/(1+\lambda)}\pi_0(\theta)\,d\theta\right\}, \tag{2.12}$$

where $d=\ln\int L(\theta|D_0)\pi_0(\theta)\,d\theta$. Here, contrary to Remark 1, one would like to choose r closer to $r^*=I(g|h)$ than to 0 to put more weight on current data. The difference between the posterior distributions in (2.8) and (2.11) is that the former gives more importance on current likelihood whereas the latter puts more importance on the historical likelihood when $\eta\geq 0.5$.

Remark 3 (*Ibrahim and coauthors' power prior*). Another important choice is to set $g(\theta)\propto L(\theta|D)\pi_0(\theta)$, $h(\theta)\propto L(\theta|D)L(\theta|D_0)\pi_0(\theta)$, which gives from (2.3) as

$$f^*(\theta)\propto[L(\theta|D)][L(\theta|D_0)]^{\eta}\pi_0(\theta). \tag{2.13}$$

This power posterior has the same appearance as that of Ibrahim et al. (2003, eq. (2)) with $\eta=a_0$. If $\lambda=\lambda^*$ solves (similar derivation as (2.10))

$$\inf_{\lambda\in\mathscr{R}^+}\left\{\lambda(r-e)+(1+\lambda)\ln\int[L(\theta|D_0)]^{\lambda/(1+\lambda)}[L(\theta|D)]\pi_0(\theta)\,d\theta\right\}, \tag{2.14}$$

where $e=\ln\int L(\theta|D)L(\theta|D_0)\pi_0(\theta)\,d\theta$, then the optimal value of η is $\eta^*=\lambda^*/(1+\lambda^*)$ in (2.13). This η^* is different from the optimal 'guide' value a_g suggested by Ibrahim et al. (2003, eq. 29). The value of λ^* (or η^*) obtained from (2.14) depends on r which lets the user control how close to the historical/current data one would like to be. Thus a_0 in (1.1) can be interpreted as a value that corresponds to the solution that is at a distance of $100(r/r^*)\%$ of $r^*=I(g|h)$ from h.

Remarks 1–3 show that different types of power posterior distributions can be generated by choosing g,h differently in (2.1) and (2.2). Below we consider several statistical models, where we derive the form of the power posterior using (2.8) in the context of Remark 1.

2.2. Normal model

Suppose iid normally distributed observations are available where the current data set is $D=\{(y_1,\ldots,y_n):y_i\sim N(\theta,1)\}$ and the historical data set is $D_0=\{(y_{10},\ldots,y_{n_00}):y_{i0}\sim N(\theta,1)\}$ $(\theta\in\mathscr{R})$. Assuming $\pi_0(\theta)=1,\forall\theta$, we get after simplifications $L(\theta|D_0)\pi_0(\theta)/\int L(\theta|D_0)\pi_0(\theta)\,d\theta=N(\overline{y}_0,1/n_0)$. Similarly, $L(\theta|D)\pi_0(\theta)/\int L(\theta|D)\pi_0(\theta)\,d\theta=N(\overline{y},1/n)$. Using algebra it follows that

$$I\left(N\left(\overline{y}_0,\frac{1}{n_0}\right)\middle|N\left(\overline{y},\frac{1}{n}\right)\right)=\frac{1}{2}\left\{\ln\left(\frac{n_0}{n}\right)-1+\frac{n}{n_0}+n(\overline{y}-\overline{y}_0)^2\right\}$$

$(=r^*$, say) which is $=n(\overline{y}-\overline{y}_0)^2/2$ if $n=n_0$, and $=0$ if $\overline{y}=\overline{y}_0$ as well $(g=h)$. Here the power posterior in (2.8) turns out to be

$$\tilde{c}\left(N\left(\overline{y},\frac{1}{n}\right)\right)^{\lambda/(1+\lambda)}\left(N\left(\overline{y}_0,\frac{1}{n_0}\right)\right)^{1/(1+\lambda)}=N\left(\frac{n\lambda\overline{y}+n_0\overline{y}_0}{n\lambda+n_0},\frac{1+\lambda}{n\lambda+n_0}\right), \tag{2.15}$$

where $\tilde{c}$ is the normalizing constant and the right side is obtained after simplification.

Writing $\eta=\lambda/(1+\lambda)$, we can write the solution in (2.15) as

$$N\left(\frac{\eta n\overline{y}+(1-\eta)n_0\overline{y}_0}{\eta n+(1-\eta)n_0},\frac{1}{\eta n+(1-\eta)n_0}\right). \tag{2.16}$$

Mean and variance of the normal distribution in (2.16) can be written as $w\overline{y}+(1-w)\overline{y}_0$ and $w(1/n)+(1-w)(1/n_0)$ (where $w=\eta n/(\eta n+(1-\eta)n_0)$), respectively. Thus the mean in (2.16) is a weighted average of the means of current ($\eta=1$ or $r=0$) and historical ($\eta=0$ or $r=r^*$) data sets. Similarly for the variance. When $n=n_0$, (2.16) reduces to $N(\eta\overline{y}+(1-\eta)\overline{y}_0,1/n)$.

From (2.9), the optimal $\lambda \geq 0$ is the value λ^* that minimizes

$$\lambda r + \frac{(1+\lambda)}{2}\left[(1-\eta)\ln(n_0) + \eta\ln(n) - \ln(a+b) - \frac{ab(\overline{y}-\overline{y}_0)^2}{(a+b)}\right], \tag{2.17}$$

where $a = n_0(1-\eta),\ b = n\eta$.

When $n = n_0$, we have $a+b = n$. If we also set $r = (n/(n+n_0))r^* = r^*/2$, further simplification from (2.17) shows that the optimum λ must solve the equation $(1+\lambda)^2 = 2$, which gives $\lambda = 0.4138$ (or $\eta = 0.2927$), for any $n = n_0, \overline{y}, \overline{y}_0$. See Section 2.6 for values of η for other possible choices of r and Section 5.1 for simulation to study the effect of n, n_0 on the values of η.

2.3. Exponential model

Suppose the current data are iid exponentially distributed with mean $1/\theta$ and are denoted by $D = \{(y_1,\ldots,y_n) : y_i \sim \text{exponential}\,(1/\theta)\}$. Similarly, the historical data are denoted by $D_0 = \{(y_{10},\ldots,y_{n_0 0}) : y_{i0} \sim \text{exponential}\,(1/\theta)\}$. Assuming $\pi_0(\theta) = 1/\theta, \forall\theta$, we get the optimal power posterior in (2.8) is proportional to

$$(\theta^{n_0} e^{-\theta\sum y_{i0}})^{1/(\lambda+1)}(\theta^{n_1} e^{-\theta\sum y_{i1}})^{\lambda/(\lambda+1)}\frac{1}{\theta} \propto \theta^{\eta n_1 + (1-\eta)n_0 - 1}\exp[-\theta(\eta n_1\overline{y}_1 + (1-\eta)n_0\overline{y}_0)]$$

(where $\eta = \lambda/(1+\lambda)$), which is a gamma distribution with shape parameter $(\lambda n_1 + n_0)/(1+\lambda) = \eta n_1 + (1-\eta)n_0$ and the scale parameter $(\lambda n_1\overline{y}_1 + n_0\overline{y}_0)/(1+\lambda) = \eta n_1\overline{y}_1 + (1-\eta)n_0\overline{y}_0$, and using (2.10) $\lambda = \lambda^*$ solves

$$\inf_{\lambda\geq 0}\left\{\lambda\left(r - \frac{\Gamma(n_1)}{(n_1\overline{y}_1)^{n_1}}\right) + (\lambda+1)\ln\left(\frac{\Gamma((\lambda n_1 + n_0)/(1+\lambda))}{[(\lambda n_1\overline{y}_1 + n_0\overline{y}_0)/(1+\lambda)]^{(\lambda n_1 + n_0)/(1+\lambda)}}\right)\right\}$$

since $c = \int \theta^{n_1} e^{-\theta\sum y_i}(1/\theta)\,d\theta = \Gamma(n_1)/(n_1\overline{y}_1)^{n_1}$.

2.4. Normal linear model

Consider the historical data as $D_0 : \boldsymbol{y}_0 = \boldsymbol{X}_0\boldsymbol{\beta} + \boldsymbol{\epsilon}_0$ where $\boldsymbol{\epsilon}_0 \sim \boldsymbol{N}_{n_0}(\boldsymbol{0}, \sigma_0^2\boldsymbol{I})$, $\boldsymbol{X}_0$ is $n_0 \times p$ of rank p, and $\boldsymbol{\beta}$ is $p \times 1$. We also assume that σ_0^2 is known, and $\pi_0(\boldsymbol{\beta}) = 1$. Similarly, suppose the current data are $D : \boldsymbol{y}_1 = \boldsymbol{X}_1\boldsymbol{\beta} + \boldsymbol{\epsilon}_1$ where $\boldsymbol{\epsilon}_1 \sim \boldsymbol{N}_n(\boldsymbol{0}, \sigma_1^2\boldsymbol{I})$ and σ_1^2 is known. Here

$$L(\boldsymbol{\beta}|D_0) = \frac{1}{(2\pi)^{n_0/2}\sigma_0^{n_0}}\exp\left\{-\frac{1}{2\sigma_0^2}(\boldsymbol{y}_0 - \boldsymbol{X}_0\boldsymbol{\beta})'(\boldsymbol{y}_0 - \boldsymbol{X}_0\boldsymbol{\beta})\right\},$$
$$L(\boldsymbol{\beta}|D_1) = \frac{1}{(2\pi)^{n_1/2}\sigma_1^{n_1}}\exp\left\{-\frac{1}{2\sigma_1^2}(\boldsymbol{y}_1 - \boldsymbol{X}_1\boldsymbol{\beta})'(\boldsymbol{y}_1 - \boldsymbol{X}_1\boldsymbol{\beta})\right\}. \tag{2.18}$$

After simplifications one can find $g = L(\boldsymbol{X}\boldsymbol{\beta}|D_0)\pi_0(\boldsymbol{\beta})/\int L(\boldsymbol{\beta}|D_0)\pi_0(\boldsymbol{\beta})\,d\boldsymbol{\beta} = \boldsymbol{N}_p(\hat{\boldsymbol{\beta}}_0, \sigma_0^2(\boldsymbol{X}_0'\boldsymbol{X}_0)^{-1})$, and, $h = L(\boldsymbol{\beta}|D)\pi_0(\boldsymbol{\beta})/\int L(\boldsymbol{\beta}|D)\pi_0(\boldsymbol{\beta})\,d\boldsymbol{\beta} = \boldsymbol{N}_p(\hat{\boldsymbol{\beta}}_1, \sigma_1^2(\boldsymbol{X}_1'\boldsymbol{X}_1)^{-1})$ where $\hat{\boldsymbol{\beta}}_0 = (\boldsymbol{X}_0'\boldsymbol{X}_0)^{-1}\boldsymbol{X}_0'\boldsymbol{y}_0$ and $\hat{\boldsymbol{\beta}}_1 = (\boldsymbol{X}_1'\boldsymbol{X}_1)^{-1}\boldsymbol{X}_1'\boldsymbol{y}_1$. Then it follows by algebra that $r^* = I(g|h)$, where ($|\boldsymbol{A}|$= determinant of matrix$\boldsymbol{A}$, $\text{tr}(\boldsymbol{A})$ = trace of matrix $\boldsymbol{A}$)

$$r^* = \frac{1}{2}\left(\text{tr}\left(\frac{\sigma_0^2}{\sigma_1^2}(\boldsymbol{X}_0'\boldsymbol{X}_0)^{-1}(\boldsymbol{X}_1'\boldsymbol{X}_1)\right) - \log\left|\frac{\sigma_0^2}{\sigma_1^2}(\boldsymbol{X}_0'\boldsymbol{X}_0)^{-1}(\boldsymbol{X}_1'\boldsymbol{X}_1)\right| - p\right) + \frac{1}{2\sigma_1^2}(\hat{\boldsymbol{\beta}}_0 - \hat{\boldsymbol{\beta}}_1)'(\boldsymbol{X}_1'\boldsymbol{X}_1)(\hat{\boldsymbol{\beta}}_0 - \hat{\boldsymbol{\beta}}_1).$$

Using (2.18) in (2.8) and simplifying, the power posterior is given by a multivariate normal distribution $\boldsymbol{N}_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ where

$$\boldsymbol{\mu} = \boldsymbol{\Gamma}\hat{\boldsymbol{\beta}}_0 + (\boldsymbol{I} - \boldsymbol{\Gamma})\hat{\boldsymbol{\beta}}_1, \quad \boldsymbol{\Gamma} = (a_0\boldsymbol{X}_0'\boldsymbol{X}_0 + a_1\boldsymbol{X}_1'\boldsymbol{X}_1)^{-1}a_0(\boldsymbol{X}_0'\boldsymbol{X}_0), \quad \boldsymbol{\Sigma} = (a_0\boldsymbol{X}_0'\boldsymbol{X}_0 + a_1\boldsymbol{X}_1'\boldsymbol{X}_1)^{-1},$$

where $a_0 = (1-\eta)/\sigma_0^2$, $a_1 = \eta/\sigma_1^2$, $\eta^* = \lambda^*/(\lambda^*+1)$. The mean $\boldsymbol{\mu}$ (covariance $\boldsymbol{\Sigma}$) is a linear combination of the means (covariances) of g and h. Using (2.18) and the form of $\boldsymbol{N}_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ above, it follows from (2.10) that λ^* solves the dual problem

$$\inf_{\lambda\geq 0}\left\{\lambda(r + C) - \frac{\lambda+1}{2}(D + E + F)\right\}, \tag{2.19}$$

where

$$C = \frac{1}{2\sigma_1^2}(\boldsymbol{y}_1'\boldsymbol{y}_1 - \boldsymbol{y}_1'\boldsymbol{X}_1(\boldsymbol{X}_1'\boldsymbol{X}_1)^{-1}\boldsymbol{X}_1'\boldsymbol{y}_1) + \frac{1}{2}\ln|\boldsymbol{X}_1'\boldsymbol{X}_1| + \frac{n_1 - p}{2}\ln 2\pi\sigma_1^2,$$
$$D = n_0(1-\eta)\ln\sigma_0^2 + n_1\eta\ln\sigma_1^2 + (n_1\eta + n_0(1-\eta) - p)\ln(2\pi),$$
$$E = (a_0\boldsymbol{y}_0'\boldsymbol{y}_0 + a_1\boldsymbol{y}_1'\boldsymbol{y}_1) + \log|(a_0\boldsymbol{X}_0'\boldsymbol{X}_0 + a_1\boldsymbol{X}_1'\boldsymbol{X}_1)|,$$
$$F = (a_0\boldsymbol{X}_0'\boldsymbol{y}_0 + a_1\boldsymbol{X}_1'\boldsymbol{y}_1)'(a_0\boldsymbol{X}_0'\boldsymbol{X}_0 + a_1\boldsymbol{X}_1'\boldsymbol{X}_1)^{-1}(a_0\boldsymbol{X}_0'\boldsymbol{y}_0 + a_1\boldsymbol{X}_1'\boldsymbol{y}_1),$$

where $\eta = \lambda/(\lambda+1)$. Note D, E, F depend on η.

2.5. Logistic regression model

Consider a binary response variable Y and an explanatory variable $\boldsymbol{X} = (X_1 = 1, X_2, \ldots, X_p)$ which are related by the relation $P(Y=1)/P(Y=0) = e^{\sum_{i=1}^{p} \beta_i x_i}$, where $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)$ are parameters. Assume $\pi_0(\boldsymbol{\beta}) = 1$. In this example we use superscripts 0:historical, and 1:current with y, X variables. The historical data are $D_0 = \{\boldsymbol{y}^0, \boldsymbol{X}^0\}$ where $\boldsymbol{X}^0$ is $n_0 \times p$, and $\boldsymbol{y}^0$ is $n_0 \times 1$. The current data are $D_1 = \{\boldsymbol{y}^1, \boldsymbol{X}^1\}$ where $\boldsymbol{X}^1$ is $n_1 \times p$, and $\boldsymbol{y}^1$ is $n_1 \times 1$. The likelihoods are

$$L(\boldsymbol{\beta}|D_0) = \frac{e^{\sum_{i=1}^{n_0} \left(\sum_{j=1}^{p} \beta_j x_{ij}^0\right) y_i^0}}{\prod_{i=1}^{n_0} (1 + e^{\sum_{j=1}^{p} \beta_j x_{ij}^0})}, \quad L(\boldsymbol{\beta}|D_1) = \frac{e^{\sum_{i=1}^{n_1} (\sum_{j=1}^{p} \beta_j x_{ij}^1) y_i^1}}{\prod_{i=1}^{n_1} \left(1 + e^{\sum_{j=1}^{p} \beta_j x_{ij}^1}\right)}. \tag{2.20}$$

Setting $g = L(\boldsymbol{\beta}|D_0)/\int L(\boldsymbol{\beta}|D_0)\, d\boldsymbol{\beta}$, and, $h = L(\boldsymbol{\beta}|D)/\int L(\boldsymbol{\beta}|D)\, d\boldsymbol{\beta}$ it follows by algebra that $r^* = I(g|h)$, where using (2.20),

$$r^* = \frac{c_1}{c_0} + \frac{1}{c_0} \int \frac{e^{\sum_{i=1}^{n_0} (\sum_{j=1}^{p} \beta_j x_{ij}^0) y_i^0}}{\prod_{i=1}^{n_0} (1 + e^{\sum_{j=1}^{p} \beta_j x_{ij}^0})} \left[\sum_{j=1}^{p} \left(\sum_{i=1}^{n_0} x_{ij}^0 y_i^0 - \sum_{i=1}^{n_1} x_{ij}^1 y_i^1 \right) \beta_j + \sum_{i=1}^{n_1} \ln(1 + e^{\sum_{j=1}^{p} \beta_j x_{ij}^1}) - \sum_{i=1}^{n_0} \ln(1 + e^{\sum_{j=1}^{p} \beta_j x_{ij}^0}) \right] d\boldsymbol{\beta},$$

where $c_0 = \int L(\boldsymbol{\beta}|D_0)\, d\boldsymbol{\beta}$, $c_1 = \int L(\boldsymbol{\beta}|D_1)\, d\boldsymbol{\beta}$.

Using (2.10), the power posterior is obtained by selecting $\lambda = \lambda^*$ which minimizes

$$\lambda(r - c_1) + (1+\lambda) \ln \int \frac{e^{\sum_{j=1}^{p} ((1/(1+\lambda)) \sum_{i=1}^{n_0} x_{ij}^0 y_i^0 - (\lambda/(1+\lambda)) \sum_{i=1}^{n_1} x_{ij}^1 y_i^1) \beta_j}}{\left(\prod_{i=1}^{n_0} (1 + e^{\sum_{j=1}^{p} \beta_j x_{ij}^0})\right)^{1/(1+\lambda)} \left(\prod_{i=1}^{n_1} (1 + e^{\sum_{j=1}^{p} \beta_j x_{ij}^1})\right)^{\lambda/(1+\lambda)}} d\boldsymbol{\beta}. \tag{2.21}$$

Using $\eta = \lambda/(1+\lambda)$, the power posterior is given by

$$\boldsymbol{W} = \tilde{c} \left(\frac{e^{\sum_{i=1}^{n_0} (\sum_{j=1}^{p} \beta_j x_{ij}^0) y_i^0}}{\prod_{i=1}^{n_0} (1 + e^{\sum_{j=1}^{p} \beta_j x_{ij}^0})} \right)^{1-\eta} \left(\frac{e^{\sum_{i=1}^{n_1} (\sum_{j=1}^{p} \beta_j x_{ij}^1) y_i^1}}{\prod_{i=1}^{n_1} (1 + e^{\sum_{j=1}^{p} \beta_j x_{ij}^1})} \right)^{\eta}, \tag{2.22}$$

where $\tilde{c}$ is a normalizing constant.

We will choose estimates of β_j's ($\hat{\boldsymbol{\beta}} = (\hat{\beta}_1, \ldots, \hat{\beta}_k)$) by maximizing $\boldsymbol{W}$ in (2.22). Considering $\ln \boldsymbol{W}$, and setting $\partial \ln \boldsymbol{W}/\partial \boldsymbol{\beta} = 0$, we get

$$(1-\eta) \sum_{i=1}^{n_0} x_{ij}^0 y_i^0 + \eta \sum_{i=1}^{n_1} x_{ij}^1 y_i^1 - (1-\eta) \sum_{i=1}^{n_0} x_{ij}^0 \hat{\pi}_i^0 - \eta \sum_{i=1}^{n_1} x_{ij}^1 \hat{\pi}_i^1 = 0, \quad j = 1, \ldots, p, \tag{2.23}$$

where $\hat{\pi}_i^j = \exp(\sum_{k=1}^{p} \hat{\beta}_k x_{ik}^j)/[1 + \exp(\sum_{k=1}^{p} \hat{\beta}_k x_{ik}^j)]$, $j = 0, 1$.

The estimators $\hat{\boldsymbol{\beta}}$ have large sample normal distribution with estimated covariance matrix given by the negative of the inverse of the second derivative matrix of $\boldsymbol{W}$. This is given by

$$\boldsymbol{H} = \{(1-\eta) \boldsymbol{X}_0' (\mathrm{Diag}[\pi_{i0}(1-\pi_{i0})]) \boldsymbol{X}_0 + \eta \boldsymbol{X}_1' (\mathrm{Diag}[\pi_{i1}(1-\pi_{i1})]) \boldsymbol{X}_1\}^{-1},$$

where $\mathrm{Diag}[\pi_{i0}(1-\pi_{i0})]$ denotes the $n_0 \times n_0$ matrix with diagonal entries as $\pi_{i0}(1-\pi_{i0})$; similarly, for $\mathrm{Diag}[\pi_{i1}(1-\pi_{i1})]$; $\pi_{i0} = \exp(\sum_{j=1}^{p} \beta_j x_{ij}^0)/(1 + \exp(\sum_{j=1}^{p} \beta_j x_{ij}^0))$, $\pi_{i1} = \exp(\sum_{j=1}^{p} \beta_j x_{ij}^1)/(1 + \exp(\sum_{j=1}^{p} \beta_j x_{ij}^1))$.

The nonlinear equations in (2.23) can be solved by Newton–Raphson method. Starting with an initial value $\boldsymbol{\beta}^{(0)}$, the iterates are given by

$$\boldsymbol{\beta}^{(t+1)} = \boldsymbol{\beta}^{(t)} + \{\boldsymbol{H}^{(t)}\}^{-1} \boldsymbol{q}^{(t)}, \quad t \geq 0,$$

where

$$\boldsymbol{H}^{(t)} = (1-\eta) \boldsymbol{X}_0' (\mathrm{Diag}[\pi_{i0}^t (1-\pi_{i0}^t)]) \boldsymbol{X}_0 + \eta \boldsymbol{X}_1' (\mathrm{Diag}[\pi_{i1}^t (1-\pi_{i1}^t)]) \boldsymbol{X}_1,$$
$$\boldsymbol{q}^{(t)} = (1-\eta) \boldsymbol{X}_0' (\boldsymbol{y}^0 - \boldsymbol{\pi}_0^t) + \eta \boldsymbol{X}_1' (\boldsymbol{y}^1 - \boldsymbol{\pi}_1^t),$$

where $\boldsymbol{\pi}_0^t = (\pi_{i0}^t)$, $\pi_{i0}^t = \exp(\sum_{j=1}^{p} \beta_j^t x_{ij}^0)/(1 + \exp(\sum_{j=1}^{p} \beta_j^t x_{ij}^0))$ and $\boldsymbol{\pi}_1^t = (\pi_{i1}^t)$, $\pi_{i1}^t = \exp(\sum_{j=1}^{p} \beta_j^t x_{ij}^1)/(1 + \exp(\sum_{j=1}^{p} \beta_j^t x_{ij}^1))$. Using the estimates of $\boldsymbol{\beta}$ and $\boldsymbol{H}$ from the iteration, one can construct confidence intervals for parameters β_i.

2.6. The choice of r

In the context of Remark 1, the choice of r, as in the assessment of quality-of-life in clinical trials (see, e.g., Cox et al., 1992; Zhao and Tsiatis, 1999a,b; Korn, 1993; Chen and Sen, 2001), is often subjective and is based on the perception of the experimenter. In

Table 1
Exponents (η) of current data likelihoods for different r (= distance between historical and current) in normal model with mean (μ) and variance (σ^2) of the solution.

r	λ	η	μ	σ^2
0	∞	1.0	10.00	0.100
100	2.47	0.71	14.47	0.078
200	1.16	0.54	16.32	0.068
300	0.58	0.37	17.75	0.061
400	0.24	0.19	18.94	0.055
500	0.00	0.00	20.00	0.050

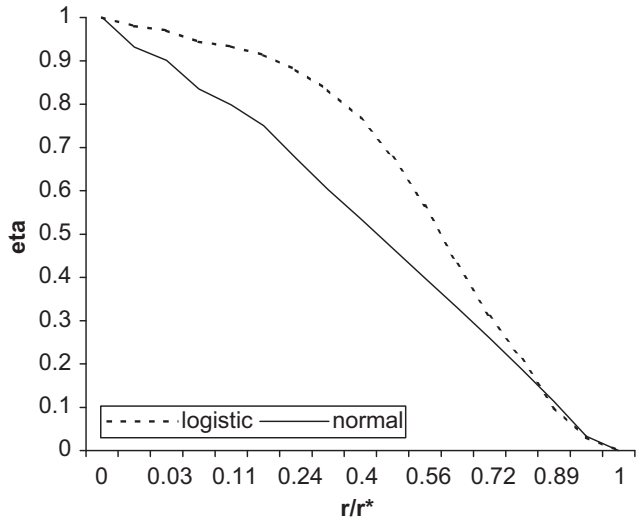

Fig. 1. Graph of eta (exponent of current) versus r/r^* (proportional distance between current and historical).

extreme cases when the historical data are not of very good quality, or could be inappropriate for some reason, one might put more emphasis on current data, and use $r = 0.01r^*$ e.g., where $r^* = I(g|h)$. If the current data are not reliable for some reason, then one might use $r = 0.99r^*$ e.g. (see also Zellner, 2002). Except these cases, the value of r would be set somewhere in between $[0.01r^*, 0.99r^*]$ based on the quality (or appropriateness) of the historical data.

Typically, one would like to be much closer to the current data than to the historical. Some suggestions for choosing r are: (1) $r=[pn/(n+n_0)]r^*, p \in [0.1, 1.0]$, where n_0, n are the sample sizes of the historical (g) and current (h) data; (2) $r=[ps_0^2/(s_0^2+s^2)]r^*, p \in [0.1, 1.0]$, where s_0^2, s^2 are the sample variances of the historical and current data, respectively. If we follow (1) with $n = n_0$ for normal model of Section 2.2, we get $r = pr^*/2$, and simplification from (2.17) shows that the optimum λ must solve the equation $p\lambda^2 + 2p\lambda + p - 2 = 0$, which gives $\eta = \lambda/(1+\lambda) = 1 - \sqrt{p/2}$ after discarding the negative solution for λ. Thus, for example, when $p = 0.5, 0.125$, then $\eta = 0.5, 0.75$, respectively, for any $n = n_0, \overline{y}, \overline{y}_0$. A solution $f^*(t) \propto [g(t)]^{\sqrt{p/2}}[h(t)]^{1-\sqrt{p/2}}$ (see (2.3)) is within $100(p/2)\%$ of r^* from h.

In practice, several sensitivity analyses should be conducted using different values of r, some giving small and some giving large weights to the historical data. The results should be interpreted based on all these analyses. The choice of r has to be modified for Remark 2 or 3 where g, h are different from above. To illustrate the role of the choice of r in the form of the power posterior, we consider sensitivity analyses in two models: (1) *Normal model.* In Section 2.1, for the normal model, we consider two samples of sizes $n_0 = 20$ and $n = 10$, each taken from normal distribution with mean θ and variance 1. The sample means are taken to be $\overline{y}_0 = 20$ and $\overline{y} = 10$. Here $r^* = 500$. For different choices of r in $[0, 500]$, values of corresponding optimum λ (and η) are found.

Table 1 lists these values along with the mean and variance of the resulting power posterior. We see that as r increases η decreases as expected from Theorem 3. (2) *Logistic regression model.* In Section 2.4, for the logistic regression model, consider $P(Y=1)/P(Y=0) = e^{\sum_{i=0}^{1}\beta_i x_i}$, where $\beta_0 = 1, \beta_1 = -1$, with $n_0 = 20$ and $n = 10$. For the historical data, let $x_i \sim N(0, 0.5^2)$; for the current data, let $x_i \sim N(0, 1)$. Here $r^* = 6.21$. For different choices of r in $[0, 6.21]$, values of corresponding optimum λ (and η) are found.

Fig. 1 presents graphs of η versus r/r^* for these two models. Both have an S-shaped pattern. For any r/r^*, logistic has higher η values than normal throughout except when $r \approx r^*$. For the normal model, η decreases more sharply than the logistic when $r/r^* \leq 0.5$ (approximately), opposite happens when $r/r^* \geq 0.5$. From Fig. 1, a smaller value of r is desirable than larger ones to keep more weight (larger η) on the current data.

3. Efficient information processing

This section shows that the information rules derived in last two sections are 100% *efficient*, a term coined by Zellner (1988). The optimal information processing rules, as discussed by Zellner (1988, 2002), Ibrahim et al. (2003), are special cases of these rules. In addition, our approach allows inclusion of constraints, which was not considered earlier. Zellner defined the quantity $\Delta[g(\theta)]$ for measuring basic information processing as follows in a Bayesian context

$$\begin{aligned}\Delta[g(\theta)] &= \text{output information} - \text{input information}\\ &= \left\{\int g(\theta)\ln(g(\theta))\,d\theta + \int g(\theta)\ln(m)\,d\theta\right\} - \left\{\int g(\theta)\ln(\pi(\theta))\,d\theta - \int g(\theta)\ln(L(\theta))\,d\theta\right\}\\ &= \int g(\theta)\ln\left(\frac{g(\theta)}{L(\theta)\pi(\theta)/m}\right)d\theta, \end{aligned} \tag{3.1}$$

where $g(\theta)$ is the proper pdf, $m = \int L(\theta)\pi(\theta)\,d\theta$ is the marginal density of the data, and $\pi(\theta)$ is the prior distribution for θ. Zellner defined a rule g to be 100% *efficient* if g minimizes (3.1) and achieves $\Delta[g(\theta)] = 0$, which yields 'output information = input information'. It is clear from (3.1), that the function that minimizes (3.1) is $g^*(\theta) = L(\theta)\pi(\theta)/m$. Thus g^* is a 100% efficient information processing rule.

To show that the solution given in Theorem 2 results in a 100% efficient information processing rule, we consider a weighted version of (3.1). The criterion $\Delta(f)$ we use below is motivated by Zellner (2002, eq. (5)), but is applicable in both Bayesian and frequentist setup. The output information is $E_f(\ln f) = \int f(t)\ln(f(t))\,dt$ and $\int f(t)[\ln(m(g,h,w))]\,dt$, where $m(g,h,w) = \int [g(t)]^w[h(t)]^{1-w}\,dt$. In our situation, the input information is given through the pdfs g and h as $E_f(\ln g)$ and $E_f(\ln h)$, respectively, and we define the total input information as $wE_f(\ln g) + (1-w)E_f(\ln h)$, where $0 \le w \le 1$, w is a weight. Thus, we define the criterion as

$$\Delta(f) = \left\{\int f(t)\ln(f(t))\,dt + \int f(t)[\ln(m(g,h,w))]\,dt\right\} - \left\{w\int f(t)\ln(g(t))\,dt + (1-w)\int f(t)\ln(h(t))\,dt\right\}. \tag{3.2}$$

A pdf f which satisfies $\Delta(f) = 0$ is defined to be '100% *efficient*'. By changing w, one can control the effect of the input information.

Since the solution in (2.3) is obtained by minimizing the corresponding Lagrangian in (2.5), the next Theorem shows that the Lagrangian $(L(f,\lambda))$ and $\Delta(f)$ are related.

Theorem 4. *If we choose* $w = 1/(1+\lambda)$ *in* (3.2), *then*

$$L(f,\lambda) = C_1\Delta(f) + C_2,$$

where C_1, C_2 *are constants free of* f, $C_1 = 1+\lambda > 0$, $C_2 = -\lambda r - (1+\lambda)\ln m$.

Proof. Considering the Lagrangian in (2.5), we minimize

$$\begin{aligned} L(f,\lambda) &= \int f(t)\ln\frac{f(t)}{g(t)}\,dt + \lambda\left(\int f(t)\ln\frac{f(t)}{h(t)}\,dt - r\right)\\ &= (1+\lambda)\int f(t)\ln(f(t))\,dt - \int f(t)\ln(g(t))\,dt - \lambda\int f(t)\ln(h(t))\,dt - \lambda r\\ &= (1+\lambda)\Delta(f) - \lambda r - (1+\lambda)\ln m, \end{aligned} \tag{3.3}$$

where $m = m(g,h,w)$ defined in Section 3. This proves the result. □

The next remark applies the previous result in a Bayesian context.

Remark 4. To generalize Zellner's (2002, eq. (5)) criterion to our setting, we set $g(\theta) = L(\theta|D_0)\pi_0(\theta)/\int L(\theta|D_0)\pi_0(\theta)\,d\theta$, $h(\theta) = L(\theta|D)\pi_0(\theta)/\int L(\theta|D)\pi_0(\theta)\,d\theta$, then using algebra the criterion $\Delta(f)$ in (3.2) reduces to

$$\begin{aligned}\Delta(f) = &\left\{\int f(\theta)\ln(f(\theta))\,d\theta + \ln\left[\int (L(\theta|D_0))^w(L(\theta|D))^{1-w}\pi_0(\theta)\,d\theta\right]\right\}\\ &- \left\{w\int f(\theta)\ln(L(\theta|D_0))\,d\theta + (1-w)\int f(\theta)\ln(L(\theta|D))\,d\theta + \int f(\theta)\ln(\pi_0(\theta))\,d\theta\right\}.\end{aligned}$$

To generalize Ibrahim et al.'s (2003, eq. (23)) criterion to our setting, we set $g(\theta) = L(\theta|D)\pi_0(\theta)/\int L(\theta|D_0)\pi_0(\theta)\,d\theta$, $h(\theta) = L(\theta|D)L(\theta|D_0)\pi_0(\theta)/\int L(\theta|D)L(\theta|D_0)\pi_0(\theta)\,d\theta$, then using algebra the criterion $\Delta(f)$ in (3.2) reduces to

$$\begin{aligned}\Delta(f) = &\left\{\int f(\theta)\ln(f(\theta))\,d\theta + \ln\left[\int (L(\theta|D_0))^{1-w}(L(\theta|D))\pi_0(\theta)\,d\theta\right]\right\}\\ &- \left\{(1-w)\int f(\theta)\ln(L(\theta|D_0))\,d\theta + \int f(\theta)\ln(L(\theta|D))\,d\theta + \int f(\theta)\ln(\pi_0(\theta))\,d\theta\right\}.\end{aligned}$$

Other choices of g and h would yield different criteria.

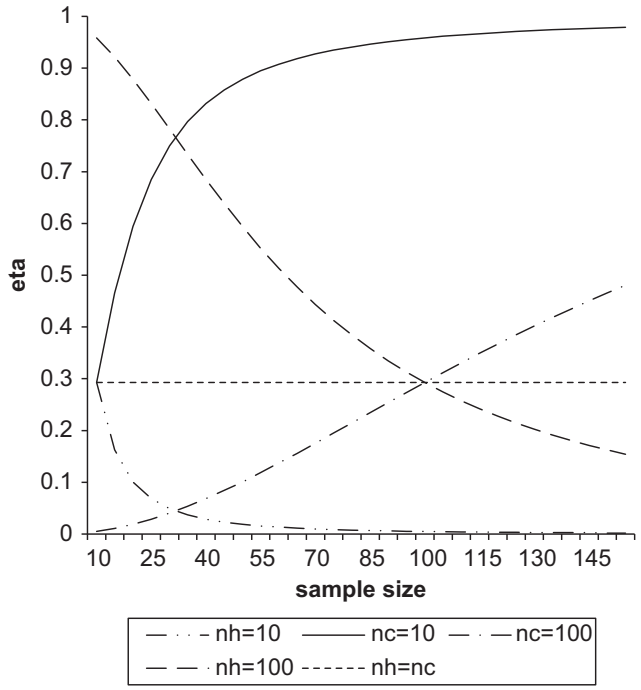

Fig. 2. Effect of sample sizes on eta.

4. Effect of sample sizes on exponent

We conduct Monte Carlo experiments to investigate the effect of sample sizes on exponent values. In the context of Section 2.1, we investigate how the optimal value of η from (2.17) change for different choices of sample sizes of current and historical data when r is kept at a fixed (proportional) distance between g (historical) and h (current). We set $g = N(\overline{y}_0, 1/n_0), h = N(\overline{y}, 1/n)$ and $r = (n/(n+n_0))r^*$, where $r^* = I(g|h)$. We have used the values $\overline{y} = 10, \overline{y}_0 = 20$.

The graphs in Fig. 2 indicate the behavior of η with change in one (or both) of the sample sizes (n: current, n_0: historical). When the sample sizes are equal, we get the curve marked '$nh = nc$', which is a horizontal line at $\eta = 0.2927$, as discussed in Section 2.1. When current data have only $n = 10$ observations ('$nc = 10$') and the historical data size increases (up to 155), η starts at 0.2927 and gets closer to 1 very fast, thus the power posterior shifts importance to the current data from the historical data as $n_0 \to \infty$. However, when $n = 100$ ('$nc = 100$'), then η starts at 0.0052 (for $n_0 = 10$) and increases up to 0.48 (for $n_0 = 155$).

When the historical data size n_0 is kept at a lower level ($nh = 10$), and the current data size n increases (up to 155), then η decreases from 0.2927 to 0.0018. When the historical data size n_0 is kept at $n_0 = 100$ ('$nh = 100$'), and the current data size n increases (up to 155), then η decreases from 0.9586 to 0.1542.

Thus the power posterior has the property of weighing the historical data higher when n_0 is unchanged and n increases. Historical data are heavily discounted when n_0 is very large compared to n. This is pronounced more if we set $r = p(n/(n+n_0))r^*$, $p < 1$. For the value of r chosen in this simulation, it turns out from further simulation (not reported here) that η does not depend on $\overline{y}$, $\overline{y}_0$ values when $\overline{y} \neq \overline{y}_0$. However, for other values of r it does.

5. Examples

Example 1 (*Information analysis of marketshare data*). Company executives from a large packaged foods manufacturer wished to determine which factors influence the market share of one of its products. Data (Kutner et al., 2005, p. 1350) were collected from a national database (Nielson) for 36 consecutive months, September, 1999 to August, 2002. The variables are y = average monthly market share for product (percent), X_1 = average monthly price of product ($), X_2 = presence or absence of discount price during period: 1 if discount, 0 otherwise, and X_3 = presence or absence of package promotion during period: 1 if promotion present, 0 otherwise. Businesses have increasingly used newer technology (e.g. more internet) to advertise their products. Thus we like to analyze this data by considering those after year 2000 as 'current' and those prior to 2000 as 'historical' (see Table 2).

Using the historical data ($n_h = 16$), we get $\hat{y} = 3.53457 - 0.48847X_1 + 0.34291X_2 + 0.15583X_3$ and from the current data ($n_c = 20$) we get $\hat{y} = 2.76845 - 0.20030X_1 + 0.44951X_2 + 0.10013X_3$. Assuming $\sigma_h^2 = s_h^2 = 0.03155$, $\sigma_c^2 = s_c^2 = 0.01645$ as known, from Section 2.4, $r^* = I(g|h) = 24.7592$, where g, h are the historical and the current likelihoods, respectively. From Section 2.6, with $p = 0.15$ we have $0.15(20/36)r^* = 2.06$, and with $p = 0.3$ we have $0.3(20/36)r^* = 4.12$. Considering $r = 2.0, 4.0$, (2.19) is solved by $\lambda = 0.477$ (or, $\eta = 0.323$)and $\lambda = 0.202$ (or, $\eta = 0.168$), respectively. The power posterior estimates, standard errors and 95%

Table 2
Analysis of the market share data.

r	Parameter	Estimates	s.e.	95% confidence interval
Historical	β_0	3.53457	2.9345	(−2.8591, 9.9283)
($r \geq r^*$)	β_1	−0.48847	0.5730	(−1.7369, 0.7600)
	β_2	0.34291	0.0123	(0.3162, 0.3696)
	β_3	0.15583	0.0088	(0.1366, 0.1750)
Power posterior	β_0	3.1981	0.2965	(2.5618, 3.8345)
($r = 2.0$)	β_1	−0.3573	0.0552	(−0.4759, −0.2388)
	β_2	0.3970	0.0057	(0.3847, 0.4093)
	β_3	0.1192	0.0058	(0.1068, 0.1316)
Power posterior	β_0	3.3119	0.4519	(2.3365, 4.2872)
($r = 4.0$)	β_1	−0.3992	0.0858	(−0.5845, −0.2140)
	β_2	0.3771	0.0068	(0.3624, 0.3918)
	β_3	0.1316	0.0068	(0.1168, 0.1463)
Current	β_0	2.76845	0.1583	(2.4329, 3.1040)
($r = 0$)	β_1	−0.20030	0.0279	(−0.2595, −0.1411)
	β_2	0.44951	0.0036	(0.4418, 0.4572)
	β_3	0.10013	0.0035	(0.0927, 0.1076)

Table 3
Analysis of British train accidents data.

r	Parameter	Estimates	s.e.	95% confidence interval
Historical	β_0	4.009	18.231	(−26.772, 54.155)
($r \geq r^*$)	β_1	−0.444	4.323	(−12.127, 7.024)
Power posterior	β_0	13.145	8.768	(−4.041, 30.332)
($r = 0.3$)	β_1	−2.557	1.864	(−6.211, 1.097)
Power posterior	β_0	12.380	8.523	(−4.324, 29.085)
($r = 0.5$)	β_1	−2.399	1.854	(−6.032, 1.234)
Current	β_0	21.756	18.026	(−2.225, 82.142)
($r = 0$)	β_1	−4.269	3.602	(−16.1956, 0.635)

confidence intervals for the parameters are given in Table 3. Between these values $p = 0.15$ or $r = 2$ seems preferable, which produces estimates closer to current data.

Using $r/r^* = 0.08$, the derived value $\eta = 0.323$ produces power posterior estimates by appropriately downweighting the historical data relative to the current data. The derived estimates are within 8% of the 'distance' between the current and the historical posterior estimates from the current posterior estimates.

Example 2 (*Information analysis of British train accidents*). Agresti (2007, p. 83) lists the number of accidents involving trains alone (collisions, derailments and overruns) and the annual distance traveled in million kilometers in Great Britain between 1975 and 2003. We consider a binary response variable which represents none or at least one accident, and use a predictor variable called 'train-distances', which is distance traveled measured in 100 million kilometers. Train-distance is a measure of railway activity. As stated by Agresti, during the past decade, rail travel has become increasingly privatized in the UK, and some people have expressed fears that accidents have become more likely. Hence, we find it reasonable to analyze this data by considering 1994–2003 as the current data, and 1975–1993 as the historical data. When we consider the historical (0) or the current (1) data separately, using logistic regression we get the estimated probability of an accident

$$\hat{\pi}_0(x) = \frac{\exp(4.0087 - 0.444x)}{1 + \exp(4.0087 - 0.444x)}, \quad \hat{\pi}_1(x) = \frac{\exp(21.756 - 4.269x)}{1 + \exp(21.756 - 4.269x)},$$

where x denotes the train-distances. Here $r^* = I(g|h) = 2.2055$, where g, h are the historical and the current likelihoods, respectively. Using $n_h = 19$, $n_c = 10$ and $p = 1/3$, we have $(1/3)(10/29)r^* = 0.25$, and with $p = 2/3$, we have $(2/3)(10/29)r^* = 0.5$. In this example we considered $r = 0.3, 0.5$. Considering $\pi_0(\theta) = 1$, it turns out that $\lambda = 1.5$ ($\eta = 0.6$) and $\lambda = 0.85$ ($\eta = 0.46$), respectively, minimizes (2.21). Here $p = 1/3$ or $r = 0.3$ seems preferable because of its proximity to the current data. The Newton–Raphson scheme produces

$$\hat{\pi}(x) = \frac{\exp(13.145 - 2.557x)}{1 + \exp(13.145 - 2.557x)}.$$

Calculating $e^{\hat{\beta}_1}$ in each case, we find, when the distance traveled is increased by another 100 million kilometers, the odds of at least another accident is decreased by 64% for the historical data, 1% for the current data, and 8% for the power posterior. These

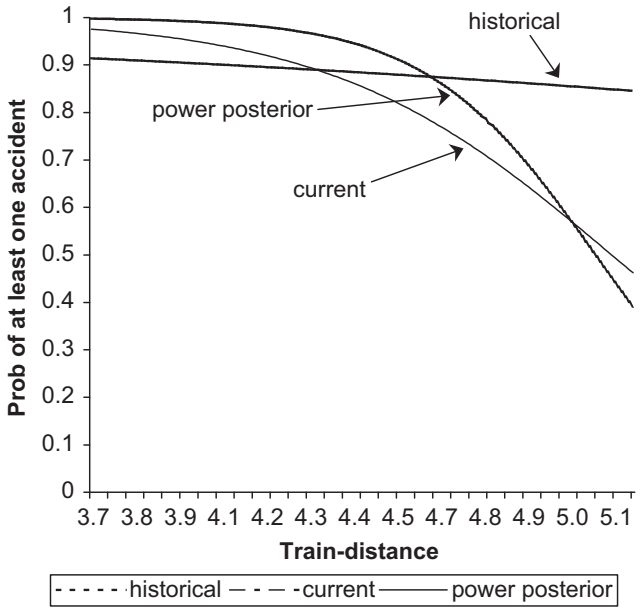

Fig. 3. Logistic regression of British train data.

findings match with those of Agresti, and in addition, they bring out the difference between the current and historical data. The standard errors and the confidence intervals of the estimates are listed in Table 3.

The graphs of estimated probability of accidents are presented in Fig. 3. They confirm that the accident rates have been lower in recent years contrary to the belief. Using $r/r^* = 0.14$, the derived value $\eta = 0.6$ produces power posterior estimates which downweights the historical data substantially, and are within 14% of the 'distance' between the current and the historical posterior estimates from the current posterior estimates.

6. Discussion

The main difficulty in the power prior rule or the quality-adjusted rule is the choice of the exponent values to be used. The notion of an explicit 'distance' of the rule from the historical or current data sets is not directly available. In this paper, we show that these exponent values are functions of r, which give a clear distance (divergence) between the rule and the current/historical posterior. The optimality results of this paper presents a formal justification of these exponent values, and our general approach shows that many other rules can be derived in this way.

To decide which r is the right choice, we performed sensitivity analysis which shows that there is no one 'correct' choice that fits all cases, and it depends on how the related distributions are set. Our procedure produces optimum rules which discounts historical data when its size is much larger than that of the current data. This pattern of behavior is similar to that of the rules generated by the procedure of Ibrahim et al. (2003), although their rules are based on a different optimizing criterion. The choice of a more suitable r is a topic of further investigation of the author.

Acknowledgment

The author thanks the referee for helpful comments.

References

Agresti, A., 2007. An Introduction to Categorical Data Analysis. Wiley, New York.
Barron, A., Schervish, M.J., Wasserman, L., 1999. The consistency of distributions in nonparametric problems. Annals of Statistics 25, 536–561.
Ben-Tal, A., Teboulle, M., Charnes, A., 1988. The role of duality in optimization problems involving entropy functionals with applications to information theory. Journal of Optimization Theory and Applications 58, 209–223.
Blahut, R.E., 1974. Hypothesis testing and information theory. IEEE Transactions on Information Theory 20, 405–417.
Chen, M.-H., Ibrahim, J.G., Shao, Q.-M., 2000. Power prior distributions for generalized linear models. Journal of Statistical Planning and Inference 84, 121–137.
Chen, P.-L., Sen, P.K., 2001. Quality-adjusted survival estimation with periodic observation. Biometrics 57, 868–874.
Cox, D.R., Fitzpatrick, R., Fletcher, A.E., Gore, E.M., Spiegethalter, D.J., Jones, D.R., 1992. Quality of life assessment: Can we keep it simple? Journal of the Royal Statistical Society Series A 155, 353–393.
Ghosal, S., Ghosh, J.K., Ramamoorthi, R.V., 1999. Posterior consistency of Dirichlet mixtures in density estimation. Annals of Statistics 27, 143–158.
Ibrahim, J.G., Chen, M.-H., 2000. Power prior distributions for regression models. Statistical Science 15, 46–60.
Ibrahim, J.G., Chen, M.-H., Sinha, D., 2003. On optimality properties of the power prior. Journal of the American Statistical Association 98, 204–213.
Ibrahim, J.G., Ryan, L.-M., Chen, M.-H., 1998. Use of historical controls to adjust for covariates in trend tests for binary data. Journal of the American Statistical Association 93, 1282–1293.

Korn, E.L., 1993. On estimating the distribution function for quality of life in cancer clinical trials. Biometrika 80, 535–542.
Kutner, M.H., Nachtsheim, C.J., Neter, J., Li, W., 2005. Applied Linear Statistical Models, McGraw-Hill.
McCulloch, R.E., 1989. Local model influence. Journal of the American Statistical Association 84, 473–478.
Rockafeller, R.T., 1976. Duality and stability in extremum problems involving convex functions. Pacific Journal of Mathematics 21, 167–186.
Rockafeller, R.T., 1974. Conjugate duality and optimization. SIAM Regional Conference Series in Applied Mathematics 16, vi+74.
Walker, S., Damien, P., Lenk, P., 2004. On priors with a Kullback–Leibler property. Journal of the American Statistical Association 99, 404–408.
Wasserman, L., 1998. Asymptotic properties of nonparametric Bayesian procedures. In: Dey, D., Muller, P., Sinha, D. (Eds.), Practical Nonparametric and Semiparametric Bayesian Statistics. Springer, New York, pp. 293–304.
Zellner, A., 1988. Optimal Information processing and Bayes' theorem (with discussion). The American Statistician 42, 278–282.
Zellner, A., 1997a. Bayesian Analysis in Econometrics and Statistics. Edward Elgar, Cheltenham, UK.
Zellner, A., 1997b. The Bayesian method of moments (BMOM): theory and applications, advances in econometrics. In: Fomby, T., Hill, R. (Eds.), Applying Maximum Entropy to Econometric Problems, vol. 12. Jai Press, Greenwich, CT, pp. 85–103.
Zellner, A., 2002. Information processing and Bayesian analysis. Journal of Econometrics 107, 41–50.
Zhao, H., Tsiatis, A., 1999a. Testing equality of survival functions of quality-adjusted lifetime. Biometrics 57, 861–867.
Zhao, H., Tsiatis, A., 1999b. Efficient estimation of the distribution of quality-adjusted survival time. Biometrics 55, 1101–1107.