Ann. Inst. Statist. Math.
Vol. 51, No. 2, 201–215 (1999)

LINEAR BAYES AND OPTIMAL ESTIMATION

V. P. GODAMBE

Department of Statistics and Actuarial Science, University of Waterloo, Waterloo, Ontario, Canada N2L 3G1

(Received August 26, 1996; revised September 8, 1997)

Abstract. In non-Bayesian statistics, it is often realistic to replace a full distributional assumption by a much weaker assumption about its first few moments; such as for instance, mean and variance. Along the same lines in Bayesian statistics one may wish to replace a completely specified prior distribution by an assumption about just a few moments of the distribution. To deal with such Bayesian semi-parametric models defined only by a few moments, Hartigan (1969, *J. Roy. Statist. Soc. Ser. B*, **31**, 440–454) put forward linear Bayes methodology. By now it has become a standard tool in Bayesian analysis. In this paper we formulate an alternative methodology based on the theory of optimum estimating functions. This alternative methodology is shown to be more readily applicable and efficient in common problems, than the linear Bayes methodology mentioned above.

Key words and phrases: Bayes methodology, conditioning, estimating functions, linearity, optimality.

1. Introduction

To extend the theory of optimum estimating functions to semi-parametric Bayesian models we need a generalized version of a theorem of Godambe and Thompson (1989). We first very briefly state the theorem and then give the needed generalization.

Deemphasizing mathematical details, let $\mathcal{X} = \{x\}$ be an abstract sample space, $\mathcal{P} = \{p\}$ a class of probability distributions on $\mathcal{X}$. Further $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_m)$ is an m dimensional parameter with real components θ_r defined on $\mathcal{P}$; $\Omega = \{\boldsymbol{\theta}(p) : p \in \mathcal{P}\}$. The estimating function theory provides estimation of the unknown parameter $\boldsymbol{\theta}$, on the basis of the data x as follows: We define *elementary estimating functions* h_j as real functions on $\mathcal{X} \times \Omega$ such that, under the distribution $p \in \mathcal{P}$, the expectation of h_j conditional on some partitioning $\mathcal{X}_j$ of the sample space $\mathcal{X}$, $\mathcal{E}(h_j \mid \mathcal{X}_j) = 0$, $j = 1, \ldots, k$. Here and subsequently, for brevity, a σ-field generated by a partition is simply referred to as a 'partition'. Thus h_j is said to be *unbiased* conditionally on $\mathcal{X}_j$. Now let

$$\boldsymbol{g} = (g_1, \ldots, g_m) \tag{1.1}$$

be an estimating function such that

$$g_r = \sum_{j=1}^{k} h_j a_{jr}$$

where a_{jr} are any functions of (x, θ) which are measurable w.r.t. $\mathcal{X}_j$, $j = 1, \ldots, k$, $r = 1, \ldots, m$;

$$\mathcal{G} = \{\boldsymbol{g}\}. \tag{1.2}$$

Further we introduce the estimating function

$$\boldsymbol{g}^* = (g_1^*, \ldots, g_m^*) \tag{1.3}$$

where

$$g_r^* = \sum_{j=1}^{k} h_j a_{jr}^*$$

with $a_{jr}^* = \mathcal{E}\{(\partial h_j/\partial\theta_r) \mid \mathcal{X}_j\}/\mathcal{E}\{(h_j^2) \mid \mathcal{X}_j\}$ assuming the derivative exists, $r = 1, \ldots, m$. Note the estimating function $\boldsymbol{g}^* \in \mathcal{G}$ in (1.2). Denoting $h_j a_{jr}^*$ by h_{jr} the elementary estimating functions h_j, $j = 1, \ldots, k$ are said to be *mutually orthogonal* if $\mathcal{E}(h_{jr} h_{j'r'} \mid \mathcal{X}_j) = 0$, $j \neq j'$, $j, j' = 1, \ldots, k$; $r, r' = 1, \ldots, m$. Now with the estimating function $\boldsymbol{g}$ in (1.1) we define two matrices:

$$J = \|\mathcal{E}(g_r g_{r'})\|, \qquad H = \|\mathcal{E}(\partial g_r/\partial\theta_{r'})\|,$$

$r, r' = 1, \ldots, m$. Similarly with the estimating function $\boldsymbol{g}^*$ in (1.3) we have matrices J^* and H^*. With the usual notation if A^T and A^+ denote the transpose and the generalized inverse respectively of matrix A, Godambe and Thompson (1989) theorem, can be stated as follows.

THEOREM 1.1. *If the elementary estimating functions h_j, $j = 1, \ldots, k$ are 'mutually orthogonal', in the class $\mathcal{G}$ in* (1.2), *the estimating function $\boldsymbol{g}^*$ given by* (1.3) *is optimal in the sense that the matrix*

$$J - H(H^*)^+ J^* (H^{*T})^+ H^T \tag{1.4}$$

is positive semidefinite, for all $\boldsymbol{g} \in \mathcal{G}$.

The estimate of $\boldsymbol{\theta}$ is obtained by solving the equation $\boldsymbol{g}^* = 0$ for the observed value of x.

A generalization of the above theorem is obtained when the parameter $\boldsymbol{\theta}$ has a distribution; now p denotes a joint distribution of $(x, \boldsymbol{\theta})$. Again $\mathcal{P} = \{p\}$. In this *new setup* the partitionings $\mathcal{X}_j$ (of the sample space $\mathcal{X}$) of the original theorem are to be replaced by the partitionings of $\mathcal{X} \times \Omega$; all the 'expectations' being replaced by the 'expectations w.r.t. the joint distribution p of $(x, \boldsymbol{\theta})$'. Of course when a partition of $\mathcal{X} \times \Omega$ is given by 'holding $\boldsymbol{\theta}$ fixed', $\mathcal{E}(\cdot \mid \boldsymbol{\theta})$ reduces to the expectation

in the original theorem. It is now understood that some elementary estimating functions h_j, $j = 1, \ldots, k$ would be exclusively functions of the parameter $\boldsymbol{\theta}$. The 'new setup' implies further extentions (E) below:

(E) The estimating functions g in (1.1), including g^* of (1.3) are now obtained from functions h_j, with coefficients 'a_{jr}' that are measurable functions w.r.t. the partitionings of $\mathcal{X} \times \Omega$ corresponding to h_j, $j = 1, \ldots, k$; $r = 1, \ldots, m$. Accordingly we have the definition of orthogonality of the estimating functions h_j, $j = 1, \ldots, k$. As before in (1.2), the class $\mathcal{G} = \{\boldsymbol{g}\}$. However now, the matrices corresponding to J and H of the preceeding setup are denoted by J_1 and H_1 respectively for the clarity of presentation.

Now we have the following generalization of the previously stated Theorem 1.1.

THEOREM 1.2. *Assume the interpretation of estimating function g^*, the class $\mathcal{G}$, the orthogonality and the matrices J_1 and H_1 as in* (E) *above. Now if the elementary estimating functions h_j, $j = 1, \ldots, k$ are mutually orthogonal then in $\mathcal{G}$, g^* is optimal in the sense that the matrix*

$$J_1 - H_1(H_1^*)^+ J_1^*(H_1^{*T})^+ H_1^T$$

is positive semidefinite for all $g \in \mathcal{G}$.

The proof of Theorem 1.2 is along the same lines as that of the previous Theorem 1.1; following is an outline. It is easy to check

$$\|\mathcal{E}(\partial g_r/\partial\theta_{r'})\| = \|\mathcal{E}(g_r g_{r'}^*)\|, \qquad r, r' = 1, \ldots, m,$$

because of the orthogonality of the estimating functions h_j, $j = 1, \ldots, k$. The above equality, as shown in Godambe and Thompson (1987) implies positive semidefiniteness of the matrix $J_1 - H_1(H_1^*)^+ J_1^*(H_1^{*T})^+ H_1^T$ hence the proof. As before the estimate of $\boldsymbol{\theta}$ is obtained by solving the equation $g^* = 0$ for the given x.

In case of a scalar parameter $\boldsymbol{\theta} = \theta$, the optimality of the estimating function $\boldsymbol{g}^* = g^*$ in Theorem 1.2 is equivalent to the inequality,

$$\mathcal{E}\left\{g^*/\mathcal{E}\left(\frac{\partial g^*}{\partial\theta}\right)\right\}^2 \leq \mathcal{E}\left\{g/\mathcal{E}\left(\frac{\partial g}{\partial\theta}\right)\right\}^2 \tag{1.5}$$

for all distributions $p \in \mathcal{P}$ and estimating functions $g \in \mathcal{G}$ in (1.2). The inequality (1.5) has interesting implications: Let for the joint distribution p of (x, θ), $p(\theta \mid x)$ denote the posterior density of θ given $x = (x_1, \ldots, x_n)$. With certain restriction on the class $\mathcal{P} = \{p\}$ and the class of estimating functions $\mathcal{G}$ in (1.2), yet covering the illustrations of Section 2, we have the following theorem.

THEOREM 1.3. *If the estimating function g^* is optimal, that is if it satisfies the criterion* (1.5), *then*

(1) $\text{Corr}^2\{g^*, \partial \log p(\theta \mid x)/\partial\theta\} \geq \text{Corr}^2\{g, \partial \log p(\theta \mid x)/\partial\theta\}$,

(2) $\mathcal{E}\{g^* - \partial \log p(\theta \mid x)/\partial\theta\}^2 \leq \mathcal{E}\{g - \partial \log p(\theta \mid x)/\partial\theta\}^2$,
for all distributions $p \in \mathcal{P}$ *and estimating functions* $g \in \mathcal{G}$.

The proof of the above theorem is given in the Appendix. It is along the lines of similar previous results of Godambe and Thompson (1987) and Godambe and Heyde ((1987), p. 232). Further comments on Theorem 1.3 above are given in Section 3.

We have formally introduced optimal estimation via estimating functions, through Theorem 1.2. Before illustrating its applications for semiparametric models, we here very briefly introduce the linear Bayes method in common use.

For a scaler parameter $\boldsymbol{\theta} = \theta$ and the sample space $\mathcal{X} = R^n$, $x = (x_1, \ldots, x_n)$, the linear Bayes method (Hartigan (1969), West and Harrison (1989, p. 136)) can be described as follows. The estimate

$$\hat{\theta} = a_0^* + \sum_{i=1}^{n} a_i^* x_i$$

of θ is said to be linear Bayes if

$$\mathcal{E}\left(\theta - a_0 - \sum_{i=1}^{n} a_i x_i\right)^2, \tag{1.6}$$

is minimized for the variations of a_i at $a_i = a_i^*$, $i = 0, \ldots, n$; the expectation in (1.6) being taken w.r.t. the *joint distribution* of θ and x_i, $i = 1, \ldots, n$.

The two methods of estimation, one the linear Bayes, just described and the other obtained via optimal estimating functions would be compared in Section 3.

The examples in Section 2, illustrating the applications of the optimal estimating functions for parameter estimation also illustrate how they can be used for *forecasting* or predicting future observations.

2. Illustrations

Here the sample space $\mathcal{X} = R^n$, $x = (x_1, \ldots, x_n)$ and $\Omega = R^1$. The class $\mathcal{P}$ of distributions p on $\mathcal{X} \times \Omega$ is given as follows. For every fixed θ, the variates $x_1, \ldots, x_n$ are independently distributed with common mean θ and variance $\sigma^2(\theta)$, *a known function* of θ. This is the case in generalized linear models (McCullagh and Nelder (1989)). Further we assume that θ is so distributed that its mean and variance are fixed (known) namely θ_0 and v_0.

The elementary estimating functions h in the present case are given by $x_i - \theta$, $i = 1, \ldots, n$ and $\theta - \theta_0$. The partition of $\mathcal{X} \times \Omega$ in case of $x_i - \theta$ is generated by 'θ', $i = 1, \ldots, n$ and for $\theta - \theta_0$ it is $\mathcal{X} \times \Omega$ itself, i.e. no partition. Thus

$$\mathcal{E}\{(x_i - \theta) \mid \theta\} = 0, \quad i = 1, \ldots, n; \qquad \mathcal{E}(\theta - \theta_0) = 0.$$

Note that the functions $x_i - \theta$, $i = 1, \ldots, n$ and $\theta - \theta_0$ are 'mutually orthogonal'. Hence the optimal estimating function for θ in the present case is given by

$$g^* = -\left\{\sum_{i=1}^{n} \frac{x_i - \theta}{\sigma^2(\theta)}\right\} + \frac{\theta - \theta_0}{v_0}. \tag{2.1}$$

On the other hand if $\sigma^2(\theta)$ is 'not known' but $\mathcal{E}\{\sigma^2(\theta)\} = \sigma_0^2$ is 'known' we could use for each of the functions $x_i - \theta$, $i = 1, \ldots, n$ and $\theta - \theta_0$ the partition $\mathcal{X} \times \Omega$, i.e. no partition. Note again the functions $x_i - \theta$, $i = 1, \ldots, n$ and $\theta - \theta_0$ are 'mutually orthogonal' and the optimum estimating function for θ is given by

$$g_0^* = -\left\{\sum_{i=1}^{n} \frac{x_i - \theta}{\sigma_0^2}\right\} + \frac{(\theta - \theta_0)}{v_0}. \tag{2.2}$$

In the above illustration if x_i is a binomial variate, equations (2.1) and (2.2) will be equivalent to

$$g^* = -\frac{\sum_{i=1}^{n}(x_i - \theta)}{\theta(1-\theta)} + \frac{\theta - \theta_0}{v_0}; \tag{2.3}$$

$$g_0^* = -\frac{\sum_{i=1}^{n}(x_i - \theta)}{\theta_0(1-\theta_0) - v_0} + \frac{\theta - \theta_0}{v_0}. \tag{2.4}$$

It is easy to see that the solution of the estimating equation $g_0^* = 0$ obtained from (2.4) coincides with the *posterior expectation* θ conditional on $x_1, \ldots, x_n$, $\mathcal{E}(\theta \mid x_1, \ldots, x_n)$, in case θ comes from a conjugate family of distributions. More generally the solution of the estimating equation $g_0^* = 0$ obtained from (2.2) would coincide with the 'posterior expectation' $\mathcal{E}(\theta \mid x_1, \ldots, x_n)$ provided we restrict to a class of distributions $\mathcal{P} = \{p\}$ for which the following conditions hold. (i) Conditional on θ, $x_1, \ldots, x_n$ are independent with common mean θ and variance $\sigma^2(\theta)$. (ii) The distribution of θ is such that $\mathcal{E}(\theta) = \theta_0$, $\mathcal{E}\{\sigma^2(\theta)\} = \sigma_0^2$ and $\mathcal{E}(\theta - \theta_0)^2 = v_0$, θ_0, σ_0, v_0 being all known. (iii) The posterior expectation of θ conditional on $x_1, \ldots, x_n$ that is $\mathcal{E}(\theta \mid x_1, \ldots, x_n) = \alpha\bar{x} + \beta$ where $\bar{x}$ is the mean of $x_1, \ldots, x_n$, α, β being independent of x's. This general result follows from Ericson (1969). The conditions (i), (ii), (iii) would be satisfied for instance, if $x_1, \ldots, x_n$ came from a one parameter (θ) exponential family with mean θ and θ comes from a corresponding conjugate family of distributions (Diaconis and Ylvisaker (1979)). Of course there could be many other distributions than the ones just mentioned satisfying conditions (i), (ii), (iii) above.

Finally we consider an illustration from stochastic processes. Again as before the sample space $\mathcal{X} = R^n$ and the parameter space $\Omega = R^1$. Let $x_1, x_2, \ldots, x_n$ be a branching process with $x_0 = 1$. Here x_i can be written as a sum of x_{i-1} variates which conditionally on x_{i-1} are independently and identically distributed as x_1. Assuming $\mathcal{E}(x_1 \mid \theta) = \theta$, we have $\mathcal{E}(x_i \mid x_1, \ldots, x_{i-1}, \theta) = \theta x_{i-1}$, $i = 1, \ldots, n$. Now our elementary estimating functions h are given by $(x_i - \theta x_{i-1})$, $i = 1, \ldots, n$ and $\theta - \theta_0$, where as before θ_0 is the known mean of the prior distribution of θ. The partition of $\mathcal{X} \times \Omega$, for the function $(x_i - \theta x_{i-1})$ is generated by $(x_1, \ldots, x_{i-1}; \theta)$, $i = 1, \ldots, n$ and for $(\theta - \theta_0)$ it is $\mathcal{X} \times \Omega$ itself. As said before by 'partition' we mean the σ-field generated by the partition. Note the elementary estimating functions $(x_i - \theta x_{i-1})$; $i = 1, \ldots, n$, $(\theta - \theta_0)$ are mutually orthogonal. Now let $\mathcal{E}\{(x_1 - \theta)^2 \mid \theta\} = \sigma^2(\theta)$, $\mathcal{E}(\theta - \theta_0)^2 = v_0$ where as before σ^2 is a known function of θ and v_0 is known. In the present case $\mathcal{E}\{(x_i - \theta x_{i-1})^2 \mid x_1, \ldots, x_{i-1}, \theta\} = \sigma^2(\theta) x_{i-1}$. Hence the optimum estimating function for θ is given by

$$-\sum_{i=1}^{n} \frac{(x_i - \theta x_{i-1})}{\sigma^2(\theta)} + \frac{(\theta - \theta_0)}{v_0}.$$

The above estimating function excluding the Bayesian factor $\{(\theta - \theta_0)/v_0\}$ was obtained previously by Godambe (1985).

The Theorem 1.2 given in Section 1 can also be used for forecasting a future value of a random variate. This is briefly illustrated with the example of branching process discussed in the preceeding paragraph. Suppose one is interested in forecasting a future value say x_{n+1}, on the basis of the observed previous values (sample) $x_1, \ldots, x_n$. To do this, one just has to write elementary estimating function $h = x_{n+1} - \theta x_n$, *in addition* to those previously mentioned namely, $x_i - \theta x_{i-1}$, $i = 1, \ldots, n$ and $\theta - \theta_0$. These provide two jointly optimum estimating functions for estimating θ and x_{n+1}:

$$g_1^* = -\sum_{i=1}^{n} \frac{x_i - \theta x_{i-1}}{\sigma^2(\theta)} + \frac{\theta - \theta_0}{v_0}; \qquad g_2^* = \frac{x_{n+1} - \theta x_n}{\sigma^2(\theta)}.$$

Let $g_1^* = g^* + g_2^*$. Then the estimate or forecast for x_{n+1} is given by $\hat{x}_{n+1} = \hat{\theta} x_n$ where $\hat{\theta}$ is given by the solution of the equation $g^*(\hat{\theta}) = 0$, for the observed values of $x_1, \ldots, x_n$.

The elementary estimating functions h considered above are linear in x's and θ. For nonlinear functions h, let in the above illustration of the branching process, the mean of the distribution of x_1 be known, say μ, and the unknown variance (for the consistency of notation) be θ. Now to estimate θ, $h_i = (x_i - \mu x_{i-1})^2 - \theta x_{i-1}$, $i = 1, \ldots, n$. Here, as in Godambe (1985), the optimum estimating function g^* depends on the skewness of the distribution of x_1.

3. A comparison with linear Bayes estimation

As stated at the end of Section 1, the linear Bayes estimate is obtained by minimizing the expectation in (1.6)

$$\mathcal{E}\left(\theta - a_0 - \sum_{i=1}^{n} a_i x_i\right)^2$$

for variations of a_i, $i = 0, \ldots, n$. The above expectation is w.r.t. the joint distribution of $x_1, \ldots, x_n$ and θ. Let this joint distribution be such that conditional on θ, $x_i = 1, \ldots, n$ are independent with common mean θ and variance $\sigma^2(\theta)$ and marginally for θ, $\mathcal{E}(\theta) = \theta_0$, $\mathcal{E}(\theta - \theta_0)^2 = v_0$ and $\mathcal{E}\{\sigma^2(\theta)\} = \sigma_0^2$, θ_0, v_0, σ_0 being all known. For this distribution the linear Bayes estimation (Hartigan (1969, p. 448)) coincides with the solution of the estimating equation $g_0^* = 0$ obtained from (2.2). This estimate, as noted in the preceding section, actually equals the posterior expectation $\mathcal{E}(\theta \mid x_1, \ldots, x_n)$ for a class of joint distributions of θ and x's.

The distinction between the two methods of estimation, one given by the linear Bayes and the other by the optimal estimating functions is best brought about by the estimating function g^* in (2.1). *When $\sigma^2(\theta)$ is a known function of θ and when $\mathcal{E}\{\sigma^2(\theta)\} = \sigma_0^2$ is also known*, according to the optimal estimating

function theory, estimation given by $g^* = 0$ is to be preferred to estimation given by $g_0^* = 0$ where g^* and g_0^* are given by (2.1) and (2.2) respectively. For, when $\sigma^2(\theta)$ is a known function, as stated in Section 2, the estimating function g^* is 'optimal' in the class of estimating functions $\mathcal{G}$ in (1.2). Particularly g^* is better than g_0^* for $g_0^* \in \mathcal{G}$. Below we elaborate on this optimality criterion.

The properties (1) and (2) of Theorem 1.3 of Section 1 clearly suggest that the optimum estimating function g^* in "some sense" is *closer* than any other estimating function g in the class $\mathcal{G}$, to $\partial \log p(\theta \mid x_1, \ldots, x_n)/\partial\theta$. For an intuitive illustration let p be a distribution such that conditional on θ, the distribution of $x_1, \ldots, x_n$ belongs to an exponential family with mean θ, $\pi(\theta)$ being the marginal distribution of θ. Now the term $\{\cdot\}$ of g^* in (2.1) is the 'score function'. This however is not true about the term $\{\cdot\}$ of g_0^* in (2.2). Further

$$\partial \log p(\theta \mid x_1, \ldots, x_n)/\partial\theta = \{\text{score function}\} + \partial \log \pi(\theta)/\partial\theta.$$

In each of the estimating functions g^*, g_0^* and $\partial \log p(\theta \mid x_1, \ldots, x_n)/\partial\theta$, for large sample size n, the term $\{\cdot\}$ would generally dominate the remaining term. Hence g^* is 'closer' than g_0^* to $\partial \log p(\theta \mid x_1, \ldots, x_n)/\partial\theta$. That is generally the solution of the equation $g^* = 0$ would tend to approximate the *mode* of the posterior distribution $p(\theta \mid x_1, \ldots, x_n)$. On the other hand, as seen in the preceeding section, for a restricted class of distributions $\mathcal{P} = \{p\}$, the solution of the equation $g_0^* = 0$ would provide the *mean* of the posterior distribution.

The score function played a central role in the theory of estimating functions, right from its inception. In recent years the theory was directed to find appropriate substitutes for the score function, in case of nuisance parameters and semiparametric models through conditional and quasi-score functions (Godambe (1976, 1985); Lindsay (1982)). Along the line, it is natural that the theory be directed to find appropriate substitute for the posterior score, for semiparametric Bayesian models.

Now for an estimation problem of primarily decision theoretic nature, the 'mean' rather than the 'mode' of the posterior distribution of the parameter θ could be more relevent. As seen before the former is approximated by the linear Bayes methodology while the latter is approximated by the estimating function methodology. Apart from this distinction we also note that as an immediate upshot of the Bayesian methodology follows the principle of 'conditioning on the entire data x' and using *exclusively* the posterior distribution $p(\theta \mid x)$ for estimation. In respect of conditioning, the estimating function theory is in a sense more flexible. The different elementary estimating functions are conditined on different partitions of $\mathcal{X} \times \Omega$. As such, conditioning on the entire data x, is not of much relevance here. Yet under appropriate conditions, the optimal estimation is based 'exclusively' on the posterior distribution $p(\theta \mid x)$, as in case of the Bayesian methodology. This is illustrated by the following very simple, though rather extreme example.

Let the joint distribution p of (x, θ) be such that conditionally on θ, the density of x is $f(x \mid \theta)$ and the prior density of θ is $\pi(\theta)$. Suppose further that we start with elementary estimating functions

$$h_1 = \alpha(\theta)\frac{\partial \log f(x \mid \theta)}{\partial\theta}, \qquad h_2 = c\frac{\partial \log \pi(\theta)}{\partial\theta} \tag{3.1}$$

where α is any function of θ and c is any constant. Granting some conditions (Ghosh (1993)) on the prior density $\pi(\theta)$ and the conditional density $f(x \mid \theta)$, we have

$$\mathcal{E}(h_2) = 0; \qquad \mathcal{E}(h_1 \mid \theta) = 0$$

and

$$\mathcal{E}\left\{\frac{\partial \log \pi(\theta)}{\partial \theta}\right\}^2 = -\mathcal{E}\left\{\frac{\partial^2 \log \pi(\theta)}{\partial \theta^2}\right\};$$

(3.2)

$$\mathcal{E}\left\{\left(\frac{\partial \log f(x \mid \theta)}{\partial \theta}\right)^2 \,\middle|\, \theta\right\} = -\mathcal{E}\left\{\left(\frac{\partial^2 \log f(x \mid \theta)}{\partial \theta^2}\right) \,\middle|\, \theta\right\}.$$

Thus the partition of $\mathcal{X} \times \Omega$ obtained by holding 'θ fixed' in case of h_1 and $\mathcal{X} \times \Omega$ itself, in case of h_2 yields the optimum estimating function

$$g^* = h_1 \frac{\mathcal{E}(\partial h_1/\partial \theta \mid \theta)}{\mathcal{E}(h_1^2 \mid \theta)} + h_2 \frac{\mathcal{E}(\partial h_2/\partial \theta)}{\mathcal{E}(h_2^2)} \tag{3.3}$$

$$= -\frac{\partial \log f(x \mid \theta)}{\partial \theta} - \frac{\partial \log \pi(\theta)}{\partial \theta}$$

because of (3.2). If as before $p(\theta \mid x)$ denotes the posterior density of θ given x, then from (3.3) we have

$$-g^* = \frac{\partial \log p(\theta \mid x)}{\partial \theta}. \tag{3.4}$$

Note that the equation (3.4) will not follow if in case of the elementary estimating function h_1, the partitioning given by 'θ fixed' is not used and θ is allowed to vary. For instance let $f(x \mid \theta) = \exp\{\theta x - \psi(\theta)\}$ and $\pi(\theta) = \exp\{\mu\theta - \nu\psi(\theta)\}$ where ψ is a known function of θ and μ, ν are known constants. In this case we may start with the elementary estimating functions

$$h_1 = \alpha(\theta)\{x - \psi'(\theta)\}, \qquad h_2 = c\{\mu - \nu\psi'(\theta)\}$$

as in (3.1). Note $\mathcal{E}(h_1 \mid \theta) = 0$ and $\mathcal{E}(h_2) = 0$. The optimum estimating function in the present case is given by

$$-g^* = (x + \mu) - (\nu + 1)\psi'(\theta)$$

$$= \frac{\partial \log p(\theta \mid x)}{\partial \theta}$$

as in (3.4).

4. Efficiencies and confidence/Bayes interval estimation

In this section initially the efficiencies of the estimating functions g^* and g_0^* in (2.1) and (2.2) are compared. Subsequently, *confidence/Bayes* intervals based on the estimating functions g^* and g_0^* are defined. The 'lengths' of these intervals are related to the efficiencies of g^* and g_0^*.

As usual the efficiency of an estimating function g is given by the inverse of the r.h.s. of (1.5):

$$\mathcal{E}ff.(g) = \frac{\{\mathcal{E}(\partial g/\partial\theta)\}^2}{\mathcal{E}(g^2)}.$$

Thus

$$\text{(4.1)} \qquad \mathcal{E}ff.(g^*) = \left[n\mathcal{E}\left\{\frac{1}{\sigma^2(\theta)}\right\} + \frac{1}{v_0}\right]; \qquad \mathcal{E}ff.(g_0^*) = \left(\frac{n}{\sigma_0^2} + \frac{1}{v_0}\right).$$

Now since $\sigma_0^2 = \mathcal{E}\{\sigma^2(\theta)\}$, and since $\mathcal{E}\{\sigma^2(\theta)\}\mathcal{E}\{\sigma^2(\theta)\}^{-1} > 1$, we have in (4.1) $\mathcal{E}ff.(g^*) \geq \mathcal{E}ff.(g_0^*)$, a conclusion also implied by the optimality of the estimating function g^*. For large sample sizes n, from (4.1), the ratio

$$\text{(4.2)} \qquad \frac{\mathcal{E}ff.(g^*)}{\mathcal{E}ff.(g_0^*)} = \sigma_0^2\mathcal{E}\left\{\frac{1}{\sigma^2(\theta)}\right\}.$$

If the parameter θ has a completely specified prior distribution and the random variate x conditional on θ has a specified parametric distribution, Bayes shortest intervals for θ could be obtained, *conditional on the observed value of* x (Godambe (1961)). These intervals are of course, based on the posterior distribution of the parameter θ, conditional on x. In the present case, since no completely specified prior distribution for θ, nor a parametric distribution for x given θ, is assumed, the posterior distribution for θ is undefined. Hence we define confidence/Bayes intervals, based on Chebyshev's inequality as follows.

The variances of the estimating function g^* and g_0^* in (2.1) and (2.2) with respect to the joint distribution p of (x, θ) are given by

$$\text{(4.3)} \qquad v(g^*) = n\mathcal{E}\left\{\frac{1}{\sigma^2(\theta)}\right\} + \frac{1}{v_0}, \qquad v(g_0^*) = \frac{n}{\sigma_0^2} + \frac{1}{v_0}.$$

Further, by Chebyshev's inequality, with sufficiently large (x, θ)-joint probability, hold the two inequalities,

$$\text{(4.4)} \qquad \left|\frac{g^*}{\sqrt{v(g^*)}}\right| \leq k; \qquad \left|\frac{g_0^*}{\sqrt{v(g_0^*)}}\right| \leq k$$

for an appropriately large k. If the inversion of the two inequalities in (4.4) provides intervals for θ, they are called *confidence/Bayes* intervals based on the estimating functions g^* and g_0^* respectively.

Now the length ℓ_0^* of the confidence/Bayes interval based on the estimating function g_0^* is directly obtained from (4.3) and (4.4), as $\ell_0^* =$

$2k(n)^{-1/2}[\mathcal{E}\{\sigma^2(\theta)\}]^{1/2}$. Further, applying one step Taylor expansion to two functions $g^* \pm k\{v(g^*)\}^{1/2}$, and using (4.3), provide for large sample sizes n, the length of the confidence/Bayes interval based on the estimating function g^* as, $\ell^* \simeq 2k(n)^{-1/2}[\mathcal{E}\{1/\sigma^2(\theta)\}]^{1/2}\sigma^2(\theta)$; this depends on θ. Here replacing $\{1/\sigma^2(\theta)\}$ by its expectation, $\mathcal{E}\{1/\sigma^2(\theta)\}$ as an *approximation* leads to the ratio

$$(\ell_0^*/\ell^*)^2 \simeq [\mathcal{E}\{\sigma^2(\theta)\}][\mathcal{E}\{1/\sigma^2(\theta)\}]. \tag{4.5}$$

The approximation just mentioned would appear to be adequate for θ's with high prior probability, under a sharply defined prior distribution. This is supported by the illustrations in the next paragraph. Now it follows from (4.2) and (4.5) that

$$(\ell_0^*/\ell^*)^2 \simeq \frac{\mathcal{E}ff.(g^*)}{\mathcal{E}ff.(g_0^*)}. \tag{4.6}$$

As noted before, since g^* is the optimum estimating function, in (4.6), r.h.s. ≥ 1. It then follows that the length of the interval ℓ^* based on g^* is often shorter than ℓ_0^*, the length of the interval based on g_0^*. However, to compute the confidence/Bayes interval based the estimating function g^* one has to know $\mathcal{E}\{1/\sigma^2(\theta)\}$.

To illustrate the above confidence/Bayes interval estimation, let the joint distribution $p(x,\theta) = f(x \mid \theta)\pi(\theta)$ where as before f is the conditional distribution of x given θ and π is the marginal (prior) distribution of θ. Further let $x = (x_1, \ldots, x_n)$ be n variates which conditionally on θ are iid as Poisson with mean θ, $\pi(\theta)$ being a gamma distribution with shape parameter α and scale parameter β. For this distribution p, $\sigma^2(\theta) = \theta$, $\mathcal{E}\{\sigma^2(\theta)\} = (\alpha/\beta) = \sigma_0^2$, $\mathcal{E}\{1/\sigma^2(\theta)\} = \{\beta/(\alpha-1)\}$. Now the approximation $\{1/\sigma^2(\theta)\} \simeq \mathcal{E}\{1/\sigma^2(\theta)\}$ is required to derive inequalities (4.5) from (4.4). However for the present distribution p, the use of the approximation just mentioned is unnecessary. For now the two inequalities in (4.4) themselves, for large samples, imply

$$\left|\frac{\bar{x}-\theta}{\theta}\right| \leq \frac{k}{\sqrt{n}}\left(\frac{\beta}{\alpha-1}\right)^{1/2}; \quad |\bar{x}-\theta| \leq \frac{k}{\sqrt{n}}\left(\frac{\alpha}{\beta}\right)^{1/2} \tag{4.7}$$

respectively. If as before ℓ^* and ℓ_0^* are lengths of the confidence/Bayes intervals based on the estimating functions g^* and g_0^* respectively, from (4.7)

$$\left(\frac{\ell^*}{\ell_0^*}\right)^2 = (\bar{x})^2\left\{\frac{\beta^2}{\alpha(\alpha-1)}\right\}. \tag{4.8}$$

Now the *modal value* of the predictive distribution of $\bar{x}$ is $(\alpha-1)/\beta$. Further for this predictive distribution the standard deviation around the 'mode' is approximately $v_0 + (1/\beta^2)$, v_0 as before being the variance of θ; $v_0 = (\alpha/\beta^2)$. Thus when v_0 is small and β is large, $\bar{x}$ would be within a short interval around the 'mode' with a *large predictive frequency*. (Actually for $\alpha = 1.01$, $\beta = 1$, the predictive frequency with which $\bar{x}$ takes values in a small neighbourhood of the mode, namely $\{(\alpha-1)/\beta\} = .01$, is ≥ 0.95.) For such values of $\bar{x}$, the ratio of the lengths of the confidence/Bayes intervals in (4.8),

$$\left(\frac{\ell^*}{\ell_0^*}\right)^2 \simeq \left(\frac{\alpha-1}{\alpha}\right) \leq 1 \tag{4.9}$$

implying $|\ell^*| \leq |\ell_0^*|$. Further from (4.2), in (4.9), $\{(\alpha - 1)/\alpha\} = \{\mathcal{E}ff.(g_0^*)/\mathcal{E}ff.(g^*)\}$, hence

$$\left(\frac{\ell_0^*}{\ell^*}\right)^2 = \frac{\mathcal{E}ff.(g^*)}{\mathcal{E}ff.(g_0^*)},$$

as in (4.6). Similar results are obtained, if 'x' instead of having a Poisson distribution as in the above example, has a binomial distribution; now the prior distribution gamma is replaced by a beta distribution.

For simplicity of presentation the above illustrations are restricted to single parameter cases. But general arguments can be easily extended to multiparametric situations. For an interesting multiparametric application of the 'generalized version' of Godambe and Thompson (1989) theorem given in Section 1, we refer to Naik-Nimbalkar and Rajarshi (1995). These authors also provide a generalized version of the theorem within the context of state-space models. Within the fully parametric Bayesian model, the optimality of the estimating function given by $\{\partial \log p(\theta \mid x_1, \ldots, x_n)/\partial\theta\}$ that is the derivative of the logarithm of the posterior density was previously established by Ferreira (1982) and Ghosh (1993).

5. Empirical Bayes setup

Again, as in Section 2, let the random variates x_i, $i = 1, \ldots, n$ be distributed independently, this time not with a constant mean θ, but with means θ_i and variances $\sigma^2(\theta_i)$; with $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_n)$,

$$\mathcal{E}(x_i - \theta_i \mid \boldsymbol{\theta}) = 0, \quad \mathcal{E}\{(x_i - \theta_i)^2 \mid \boldsymbol{\theta}\} = \sigma^2(\theta_i), \qquad i = 1, \ldots, n. \tag{5.1}$$

As before $\sigma^2(\theta)$ is a known function of θ. Further let $\theta_1, \ldots, \theta_n$ themselves be distributed independently with mean θ_0 and variance v_0;

$$\mathcal{E}(\theta_i - \theta_0) = 0, \quad \mathcal{E}(\theta_i - \theta_0)^2 = v_0, \qquad i = 1, \ldots, n.$$

For the moment we assume θ_0 and v_0 to be *known*. As in Section 2 the elementary estimating functions h are given by $x_i - \theta_i$ and $\theta_i - \theta_0$, $i = 1, \ldots, n$. The partition of $\mathcal{X} \times \Omega$ in case of $x_i - \theta_i$ is generated by $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_n)$, $i = 1, \ldots, n$ and for $\theta_i - \theta_0$ by $\mathcal{X} \times \Omega$ itself, $i = 1, \ldots, n$. These functions h are mutually *orthogonal*. Hence the jointly optimal estimating equations for estimating θ_i, $i = 1, \ldots, n$ are given by

$$\frac{(x_i - \theta_i)}{\sigma^2(\theta_i)} - \frac{(\theta_i - \theta_0)}{v_0} = 0. \tag{5.2}$$

Now consider the case when θ_0 and v_0 in (5.2) are *unknown*. To estimate θ_0 and v_0 we have from (5.1), the estimating functions $x_i - \theta_0$, $(x_i - \theta_0)^2 - v_0 - \mathcal{E}\{\sigma^2(\theta_i)\}$, $i = 1, \ldots, n$. Note $\mathcal{E}(x_i - \theta_0) = 0$ and $\mathcal{E}[(x_i - \theta_0)^2 - v_0 - \mathcal{E}\{\sigma^2(\theta_i)\}] = 0$. Further the two sets of estimating functions are mutually orthogonal if for $i = 1, \ldots, n$, (i) $\mathcal{E}\{(x_i - \theta_i)^2 \mid \theta_i\} = \sigma^2$, independent of θ_i, (ii) $\mathcal{E}\{(x_i - \theta_i)^3 \mid \theta_i\} = 0$ and (iii) $\mathcal{E}(\theta_i - \theta_0)^3 = 0$. Of these three assumptions (i) and (ii) are satisfied if given θ_i,

x_i are normal; an assumption often made in the investigation of Empirical Bayes or James-Stein type estimation. The assumption (iii), together with a further assumption (iv) $\mathcal{E}\{(x_i - \theta_i)^4 \mid \theta_i\}$ = constant independent of i provides jointly optimal estimating equations for θ_0 and v_0 in a particularly simple form:

$$\sum_{i=1}^{n}(x_i - \theta_0) = 0; \qquad \sum_{i=1}^{n}\{(x_i - \theta_0)^2 - v_0 - \sigma^2\} = 0. \tag{5.3}$$

From (5.3) we obtain the estimates of θ_0 and v_0 as,

$$\hat{\theta}_0 = \sum_{i=1}^{n} x_i/n; \qquad \hat{v}_0 = \left\{\sum_{i=1}^{n}(x_i - \hat{\theta}_0)^2/n\right\} - \sigma^2 \tag{5.4}$$

where σ^2 is assumed to be known. Further, substitution of the estimates $\hat{\theta}_0$ and $\hat{v}_0$ given by (5.4) in (5.2) provides for $i = 1, \ldots, n$ the estimates

$$\hat{\theta}_i = \left\{\frac{x_i}{\sigma^2} + \frac{\hat{\theta}_0}{\hat{v}_0}\right\} \Big/ \left\{\frac{1}{\sigma^2} + \frac{1}{\hat{v}_0}\right\}. \tag{5.5}$$

For large values of $\hat{v}_0, \hat{\theta}_i \simeq x_i$. The estimators $\hat{\theta}_i$ are clearly Empirical Bayes or James-Stein type estimates.

It is important to note the conditions underlying the estimates $\hat{\theta}_i$ in (5.5). The assumptions (i), (ii), (iv) above are satisfied by any distributions of x_i such that under θ_i, $(x_i - \theta_i)$ are iid and symmetric around '0' for $i = 1, \ldots, n$. The assumption (iii) is generally satisfied by a distribution of θ symmetric around its mean. Now for a given θ_0 and v_0 equations (5.2) are jointly optimal for θ_i, $i = 1, \ldots, n$, so also equations (5.3) are jointly optimal for θ_0 and v_0. Yet the equations (5.2), (5.3) together may not necessarily be jointly optimal for the parameters θ_i $(i = 1, \ldots, n)$, θ_0 and v_0. For, the elementary estimating functions $(x_i - \theta_i)$, $(\theta_i - \theta_0)$ are not orthogonal to estimating functions $(x_i - \theta_0)$, $\{(x_i - \theta_0)^2 - v_0 - \sigma^2\}$. This however need not be a serious concern, for the optimal estimating functions for θ_0 and v_0 namely the left hand sides of the equations (5.3), are approximately 'orthogonal' to the left hand sides of the equations (5.2), for large sample size n. Hence substitutions of the estimates $\hat{\theta}_0$ and $\hat{v}_0$ in the equation (5.2) affect the latters' optimality only a little (Godambe (1991)). This provides justification for the estimates $\hat{\theta}_i$ in (5.5), for large samples n.

An obvious generalization of the estimates $\hat{\theta}_i$ in (5.5) to a situation when we have for $i = 1, \ldots, k$, observations x_{ij}, $j = 1, \ldots, n_i$ is given below. Let under θ_i, x_{ij}, $j = 1, \ldots, n_i$ be distributed independently with the common mean θ_i and variance σ^2. Here the estimates in (5.5) are replaced by

$$\hat{\theta}_i = \left\{\frac{n_i \bar{x}_i}{\sigma^2} + \frac{\hat{\theta}_0}{\hat{v}_0}\right\} \Big/ \left\{\frac{n_i}{\sigma^2} + \frac{1}{\hat{v}_0}\right\}$$

where

$$\hat{\theta}_0 = \sum_{i=1}^{k}\sum_{j=1}^{n_i} x_{ij} \Big/ \sum_{1}^{k} n_i; \qquad \hat{v}_0 = \left\{ \sum_{i=1}^{k}\sum_{j=1}^{n_i} (x_{ij} - \hat{\theta}_0)^2 \Big/ \sum_{1}^{k} n_i \right\} - \sigma^2.$$

Note for large n_i or $\hat{v}_0, \hat{\theta}_i \simeq \bar{x}_i$.

For a previous discussion of mutual relationships between James-Stein estimation and estimating function theory, an interesting reference is Liang and Waclawiw (1990).

Acknowledgements

During my recent visit to the Department of Statistics, University of Poona, I had many valuable comments from U. Naik-Nimbalkar and M. D. Rajarshi on an earlier version of this paper. Helpful remarks from Kunte are acknowledged. I also am indebted to D. Li of our department for important discussions and also for computational help. Comments by M. Thompson led to important clarifications.

Appendix

To prove Theorem 1.3 of Section 1, in addition to the usual regularity assumptions such as existence of the required derivatives we impose following restrictions on the class of distributions $\mathcal{P} = \{p\}$ and the class of estimating functions $\mathcal{G}$ in (1.2):

(a) $p(x, \theta) = f(x \mid \theta)\pi(\theta)$, f being the conditional density of x given θ and π the marginal density of θ. For some numbers a and b, $b > a$, $\pi(\theta) = 0$ for $\theta \geq b$, $\theta \leq a$ and $\pi(\theta) \to 0$ as $\theta \to a$ or b.

(b) Any estimating function $g \in \mathcal{G}$ can be written as $g = g_1 + g_2$ where g_2 is exclusively a function of θ; $g_2 = g_2(\theta)$.

(c) For all distributions $p \in \mathcal{P}$, and all estimating functions $g \in \mathcal{G}$, $\mathcal{E}(g_1 \mid \theta) = 0$.

The conditions (a), (b), (c) above are satisfied in most applications including the illustrations in Section 2.

To prove part (1) of the theorem, we note that because of the conditions (a)–(c),

$$\text{(A.1)} \qquad \mathcal{E}\left(g \frac{\partial \log p(\theta \mid x)}{\partial \theta}\right) = \mathcal{E}\left(g_1 \frac{\partial \log f}{\partial \theta}\right) + \mathcal{E}\left(g_2 \frac{\partial \log \pi}{\partial \theta}\right)$$
$$= -\mathcal{E}\left(\frac{\partial g_1}{\partial \theta}\right) + \int g_2 \frac{\partial \pi}{\partial \theta} d\theta.$$

Now using the condition (a), intergration by parts, gives

$$\text{(A.2)} \qquad \int g_2 \frac{\partial \pi}{\partial \theta} d\theta = -\int \frac{\partial g_2}{\partial \theta} \pi d\theta = -\mathcal{E}\left(\frac{\partial g_2}{\partial \theta}\right).$$

Thus from (A.1) and (A.2) above we have

$$\mathcal{E}\left(g\frac{\partial \log p(\theta \mid x)}{\partial \theta}\right) = -\mathcal{E}\left(\frac{\partial g}{\partial \theta}\right). \tag{A.3}$$

Equation (A.3) above and (1.5) imply part (1) of the theorem.

To prove part (2) of the theorem we note that because of conditions (a), (b), (c),

$$\mathcal{E}\left(g - \frac{\partial \log p(\theta \mid x)}{\partial \theta}\right)^2 = \mathcal{E}\left(g_1 - \frac{\partial \log f(x \mid \theta)}{\partial \theta}\right)^2 + \mathcal{E}\left(g_2 - \frac{\partial \log \pi(\theta)}{\partial \theta}\right)^2. \tag{A.4}$$

Now let g_1^* be the optimum combination of the elementary estimating functions h, which are essentially functions of x and possibly also of θ. Similarly let g_2^* be the optimum combination of the elementary functions h, which are exclusively functions of θ. Further because of the mutual orthogonality of the elementary estimating functions h we have, the optimum estimating function $g^* = g_1^* + g_2^*$. Now in the r.h.s of (A.4) the first term is minimized for $g_1 = g_1^*$ (Godambe and Thompson (1987)). To minimize the second term in the r.h.s of (A.4) we note that

$$\mathcal{E}\left(g_2 - \frac{\partial \log \pi}{\partial \theta}\right)^2 = \mathcal{E}(g_2^2) - 2\mathcal{E}\left(g_2\frac{\partial \log \pi}{\partial \theta}\right) + \mathcal{E}\left(\frac{\partial \log \pi}{\partial \theta}\right)^2. \tag{A.5}$$

As before using condition (a), integration by parts, gives

$$\mathcal{E}\left(g_2\frac{\partial \log \pi}{\partial \theta}\right) = -\mathcal{E}\left(\frac{\partial g_2}{\partial \theta}\right). \tag{A.6}$$

Now let $g_2 = \sum ah$ where a's are some *constants*. From (A.6) it is easy to see that for variations of a's, (A.5) is minimized at $a = a^* = \mathcal{E}(\frac{\partial h}{\partial \theta})/\mathcal{E}(h^2)$; that is (A.5) is minimized for $g_2 = g_2^*$. This proves part (2) of the theorem.

REFERENCES

Diaconis, P. and Ylvisaker, D. (1979). Conjugate priors for exponential families, *Ann. Statist.*, **7**, 269–281.

Ericson, W. A. (1969). A note on posterior mean of population mean, *J. Roy. Statist. Soc. Ser. B*, **31**, 332–334.

Ferreira, P. E. (1982). Estimating equations in presence of prior knowledge, *Biometrika*, **69**, 667–669.

Ghosh, M. (1993). On a Bayesian analog of the theory of estimating function, *C. G. Khatri Memorial Volume Gujarat Statistical Review*, **17A**, 47–52.

Godambe, V. P. (1961). Bayes shortest confidence intervals and admissibility, Circulated at the 33 Session of I.S.I. held at Paris.

Godambe, V. P. (1976). Conditional likelihood and unconditional optimum estimating equations, *Biometrika*, **63**, 277–284.

Godambe, V. P. (1985). The foundations of finite sample estimation in stochastic processes, *Biometrika*, **72**, 419–428.

Godambe, V. P. (1991). Orthogonality of estimating functions and nuisance parameters, *Biometrika*, **78**, 143-151.

Godambe, V. P. and Heyde, C. C. (1987). Quasi-likelihood and optimal estimation, *International Statistical Review*, **55**, 231–244.

Godambe, V. P. and Thompson, M. E. (1987). Logic of least squares—revisited, Tech. Report, Ser. STAT-87-06, University of Waterloo.

Godambe, V. P. and Thompson, M. E. (1989). An extension of quasi-likelihood estimation (with discussion), *J. Statist. Plann. Inference*, **22**, 137–172.

Hartigan, J. A. (1969). Linear Bayesian methods. *J. Roy. Statist. Soc. Ser. B*, **31**, 440–454.

Liang, K.-Y. and Waclawiw, M. A. (1990). Extension of Stein estimating procedure through the use of estimating functions, *J. Amer. Statist. Assoc.*, **85**, 435–440.

Lindsay, B. (1982). Conditional score functions: some optimality results, *Biometrika*, **69**, 503–512.

McCullagh, P. and Nelder, J. (1989). *Generalized Linear Models*, Chapman and Hall, London.

Naik-Nimbalkar, U. V. and Rajarshi, M. D. (1995). Filtering and smoothing via estimating functions, *J. Amer. Statist. Assoc.*, **90**, 301–306.

West, M. and Harrison, J. (1989). *Bayesian Forecasting and Dynamic Models*, Springer, London.