Least squares approximation in Bayesian analysis

MICHEL MOUCHART* and LÉOPOLD SIMAR**

*Université Catholique de Louvain

**Facultes Universitaires Saint Louis-Bruxelles

SUMMARY

The paper presents in a simple and unified framework the Least-Squares approximation of posterior expectations. Particular structures of the sampling process and of the prior distribution are used to organize and to generalize previous results. The two basic structures are obtained by considering unbiased estimators and exchangeable processes. These ideas are applied to the estimation of the mean. Sufficient reduction of the data is analysed when only the Least-Squares approximation is involved.

Keywords: LINEAR BAYES, LEAST SQUARES, CREDIBILITY THEORY

1. INTRODUCTION

1.1. *General Formulation*

Consider a random vector (θ',x') with $\theta \in R^q$ and $x \in R^p$. In what follows, x will typically represent (functions of) observations and θ will represent either (functions of) parameters or future observations. In all of this paper, (θ',x') is assumed to be square-integrable.

In Bayesian analysis, attention is often directed toward computing the posterior expectation $E(\theta|x)$. This is, *e.g.*, the Bayesian decision rule under quadratic loss. In this paper we consider simple (*i.e.* linear) approximations of $E(\theta|x)$; they will be denoted $\hat{E}(\theta|x)$. Under the Least-Squares (L.S.) criterion, the best linear approximation of $E(\theta|x)$ is also the best linear approximation of θ. In this second interpretation, $\hat{E}(\theta|x)$ may be viewed as a "best linear estimator of θ". Doob (1953) suggests that $E(\theta|x)$ be called the best L.S. approximation of θ and $\hat{E}(\theta|x)$ the wide-sense version of $E(\theta|x)$.

In order to write explicitly $\hat{E}(\theta|x)$, we partition the vector of the expectations and the variance-covariance matrix in the following way:

$$\mathrm{E}\begin{pmatrix}\theta\\ x\end{pmatrix} = \begin{pmatrix}\mathrm{E}(\theta)\\ \mathrm{E}(x)\end{pmatrix} = \begin{pmatrix}\mathrm{m}\\ \mathrm{E}(x)\end{pmatrix} \tag{1.1}$$

$$\mathrm{V}\begin{pmatrix}\theta\\ x\end{pmatrix} = \begin{pmatrix}\mathrm{V}_{\theta\theta} & \mathrm{V}_{\theta x}\\ \mathrm{V}_{x\theta} & \mathrm{V}_{xx}\end{pmatrix} \tag{1.2}$$

Under the Least-Squares criterion, the best linear approximation of $E(\theta \mid x)$ is known to be:

$$\hat{\mathrm{E}}(\theta \mid x) = m + \mathrm{V}_{\theta x}\, \mathrm{V}_{xx}^{-1}(x\text{-}E(x)). \tag{1.3}$$

It is important to point out that formula (1.3) is valid irrespective of the form of the distribution of (θ,x); the only restriction being the existence of second-order moments. Reading (1.3) component-wise, we conclude that each component of $\hat{\mathrm{E}}(\theta \mid x)$ is also the L.S. approximation of the corresponding component of θ (or of $\mathrm{E}(\theta \mid x)$); more formally:

$$\hat{\mathrm{E}}(\theta \mid x) = \begin{pmatrix}\hat{\mathrm{E}}(\theta_1 \mid x)\\ \hat{\mathrm{E}}(\theta_2 \mid x)\end{pmatrix} \tag{1.4}$$

Formula (1.3) is computationally simple and needs only the specification of the first two moments; this gives it some properties of robustness.

When θ represents parameters of the sampling distribution, this specification will often rely on the following decomposition, based on averaging over sampling moments:

$$E(x) = E_\theta E(x \mid \theta) \tag{1.5}$$

$$V_{xx} = E_\theta V(x \mid \theta) + V_\theta\, E(x \mid \theta) \underset{\text{def}}{=} V_1 + V_0 \tag{1.6}$$

$$V_{\theta x} = E_\theta[\theta E(x' \mid \theta)] - E(\theta)E(x') = \mathrm{cov}\big(\theta, E(x' \mid \theta)\big) \tag{1.7}$$

Apart from the ease of computation and the aspect of robustness, the accuracy of the L.S. approximation is often crucial. Let us introduce

$$\eta = \theta - \hat{E}(\theta \mid x) = (\theta - m) - V_{\theta x} V_{xx}^{-1} \left(x - E(x)\right) \tag{1.8}$$

Clearly, η has zero mean and is uncorrelated with x. Often, one would like to analyse the accuracy of the approximation *for a given* $\boldsymbol{x}$. We have the following posterior moments:

$$E(\eta \mid x) = E(\theta \mid x) - \hat{E}(\theta \mid x) \tag{1.9}$$

$$V(\eta \mid x) = V(\theta \mid x). \tag{1.10}$$

Unfortunately, these quantities are generally at least as difficult to compute as $E(\theta \mid x)$ itself. However, $V(\eta)$ is easily computed from (1.8)

$$V(\eta) = V_{\theta\theta} - V_{\theta x} V_{xx}^{-1} V_{x\theta} \tag{1.11}$$

This formula again depends only on second moments of (θ, x) (directly computable from prior and sampling moments when θ represents a parameter). We now decompose $V(\eta)$ as follows:

$$V(\eta) = E_x V(\eta \mid x) + V_x E(\eta \mid x). \tag{1.12}$$

The dispersion of η has therefore two components: by (1.10) the first one is due to the average posterior variance of θ and the second one comes, by (1.9), from the possible non-linearity of $E(\theta \mid x)$. Therefore $V(\eta)$ gives an upper bound for the average posterior variance of θ:

$$E_x V(\theta \mid x) \leq V(\eta) \tag{1.13}$$

where $\leq$ is written in the sense of positive-definite, symmetric (P.D.S) matrices, and with equality if and only if the true regression is linear $\left(i.e.\ E(\theta \mid x) = \hat{E}(\theta \mid x)\right)$.

In particular,

$$V(\eta) = V(\theta \mid x) \text{ for any } x$$

if and only if the true regression of θ on x is

(i) linear $\left(E(\theta \mid x) = \hat{E}(\theta \mid x)\ \ a.s.\right)$

(ii) homoscedastic ($V(\theta \mid x)$ constant *a.s.*).

As this is the case for the normal distribution, one may interpret the L.S. approximation as adjusting an overall normal distribution on (θ,x) with identical first two moments. In other words, the formula used to compute $\hat{E}(\theta \mid x)$ (*i.e.* (1.3)) and $V(\eta)$ (*i.e.* (1.11)) may be viewed as the conditional mean and variance of that normal approximation. This feature has been demonstrated by Doob (1953 - Chap. 1) and Hartigan (1969); both suggested the terminology of "linear expectation" for $\hat{E}(\theta \mid x)$; Hartigan also suggested "linear variance" for $V(\eta)$, and even used the notation "$V(\theta \mid x)$", but this appears to be ambiguous and will not be used here.

From a decision point of view, let us consider $\hat{E}(\theta \mid x)$ as a decision rule. From (1.8), $V(\eta)$ appears as the Bayesian mean-squared error matrix of $\hat{E}(\theta \mid x)$:

$$M S E(\hat{E}(\theta \mid x)) \equiv E(\theta - \hat{E}(\theta \mid x))\ (\theta - \hat{E}(\theta \mid x))' = V(\eta). \tag{1.14}$$

Under a quadratic loss associated with a decision rule $t = t(x)$:

$$\ell(t,\theta) = (t - \theta)' A(t-\theta) \qquad A : \text{SPDS} \tag{1.15}$$

the Bayesian risk associated with $\hat{E}(\theta \mid x)$ is:

$$R(\hat{E}(\theta \mid x)) \equiv E\, \ell(\hat{E}(\theta \mid x),\theta) = tr\, A\, V(\eta). \tag{1.16}$$

In any case, $V(\eta)$ will determine the decisional accuracy of $\hat{E}(\theta \mid x)$.

1.2. *General Comments and Objectives of the Paper*

We developed an interest in L.S. approximations when supervising a student's thesis on credibility theory (Bouchat (1977)). We then became aware that the idea of L.S. approximation to Bayesian solutions had been widely used in various fields of applications with different terminologies and striking duplication of results. It has been used since 1920 in actuarial sciences under the heading of credibility theory. An overview may be found in Bühlman (1970), de Vijlder (1975) or Kahn (1975). Recent developments are also due to Bülhman (1971) and Jewel (1974 a, b, c). Hartigan (1969) and Goldstein (1975 a, b, 1976), under the heading of linear Bayes methods, analyse the L.S. approximations in various particular statistical problems. Stone (1963) and Dickey (1969) arrive at similar methods when looking for robust Bayesian procedures.

A recurrent theme in the above literature considers whether the L.S. approximations is exact or not, *i.e.* whether or not $\hat{E}(\theta \mid x) = E(\theta \mid x)$. Bailey (1950) and Mayerson (1964) have shown that particular combinations of prior probability and likelihood yield exact credibility for the mean of a process. Jewel (1974 b, c) extended these results for the exponential family under natural-conjugate prior. Kagan, Linnik and Rao (1973, addendum B) give conditions for the linearity of Bayes estimators. Recently, Diaconis and Ylvisaker (1979) have characterized conjugate prior measures through the property of linear posterior expectation of the mean of the process. By so doing they not only extend previous results on exact L.S. approximations, but they also linked this problem to the admissibility of linear estimator under quadratic loss. (See also Kagan *et al* (1973, Chap.7).) A somewhat different approach is to characterize joint distributions (on (θ,x)) having linear expectation $E(\theta \mid x)$. Thus, Lukacs and Laha (1964, Chap. 6) give a necessary and sufficient condition in terms of characteristic functions.

During the revision of this paper we also became aware of recent results by Goel and DeGroot (1979) and by Goel (1979) characterizing linearity of posterior expectations in linear regression and in a scale parameter family.

This problem of exact approximation will not be pursued further in this paper. Instead our main objective is to present in a simple and unified framework previous results otherwise stated in particular contexts. By so doing, we simplify unnecessarily complicated results and remove ambiguities (which possibly induced errors).

The unifying argument is given in the general formulation of the previous section and is essentially summarized in the formulae giving $\hat{E}(\theta \mid x)$ and $V(\eta)$ (*i.e.* (1.3) and (1.11)). In this very simple framework, the presentation is organized according to particular structures of the first two moments (1.1) and (1.2): focusing attention on these particular structures induces natural generalizations of previous results and clarifies the role of the given assumptions. In particular, it may suggest suitable transformations (of the observations or of the parameters) in order to take advantage of specific structures (both in the prior information and in the sampling process).

Finally we systematically analyse the case of several parameters. It appears that treating each parameter individually or treating all parameters together does not affect the computation of $\hat{E}(\theta \mid x)$ or of the diagonal elements of $V(\eta)$. However, the role of the simultaneity in the inference shows up in the off-diagonal elements of $V(\eta)$ and so affects its inverse, associated with the concept of precision.

In Section 2, we consider two particular structures induced by the use of unbiased estimators and by some properties of exchangeability in the sampling process. It is shown that those cases provide peculiar forms of the

L.S. approximation under more general conditions than previously presented. The last section addresses itself to the question of sufficient reduction of the data when *only* the L.S. approximation is involved.

2. PARTICULAR STRUCTURES

Formula (1.3) gives a rather general framework to treat L.S. approximation in Bayesian analysis. For instance, in non-parametric situations θ may be a finite-dimensional characteristic of an infinite dimensional parameter (*viz.* the distribution function of the observation). Similarly, x may be either a full sample result or a statistic defined on a more complete sample result. Note however that the specification of $V_{\theta x}$ and V_{xx} is not always easy. It is then important to choose a suitable statistic x carefully and to take advantage of the particular structure of both the prior information and the sampling process. The object of this section is to analyse two particular structures which prove to be basic for the L.S. approximations.

2.1. *Use of Unbiased Estimator*

Suppose that we first reduce the sample to an unbiased estimator of θ:

$$E(x|\theta) = \theta. \tag{2.1}$$

Clearley, in this particular case, $p = q$. This structure implies that:

$$E(x) = E(\theta) = m \tag{2.2}$$

$$V_{\theta x} = V_{\theta\theta} = V\,E(x|\theta) = V_0. \tag{2.3}$$

Then $\hat{E}(\theta|x)$ may be written as

$$\hat{E}(\theta|x) = V_1(V_1+V_0)^{-1}m + V_0(V_1+V_0)^{-1}x \tag{2.4}$$

If V_0 and V_1 are both regular, this simplifies to

$$\hat{E}(\theta|x) = (V_0^{-1}+V_1^{-1})^{-1}\,[V_0^{-1}m + V_1^{-1}x] \tag{2.5}$$

$\hat{E}(\theta|x)$ appears as a weighted matrix average between $E(\theta)$ and x; *i.e.* $\hat{E}(\theta|x)$ has the form:

$$\hat{E}(\theta|x) = Am + (I - A)x. \tag{2.6}$$

Note that this derivation is very easy and is implied only by the property of unbiasedness. It has appeared frequently in the literature, in particular for the case $p = q = 1$ with θ being the population mean and x the sample mean. This formula is familiar for the Bayesian inference on the mean of a normal process where, in this case, $\hat{E}(\theta|x) = E(\theta|x)$ (see *e.g.* Raiffa and Schlaifer (1961)) or in credibility theory (see *e.g.* Bühlmann (1970)).

The average measure of accuracy $V(\eta)$, given in (1.11) becomes

$$V(\eta) = (V_0^{-1} + V_1^{-1})^{-1} \tag{2.7}$$

It is illuminating to write down the upper bound for the average posterior variance, in (1.13), in terms of "mean" precisions (where "mean" stands in the sense of harmonic mean, *i.e.* the inverse of the expectation of the inverse)

$$[EV(\theta|x)]^{-1} \geq [V(\theta)]^{-1} + [EV(x|\theta)]^{-1} \tag{2.8}$$

Thus the "mean" posterior precision is at least equal to the prior precision plus the "mean" sampling precision, with equality if and only if $E(\theta|x)$ is linear in x. This addition of precision is familiar (with equality) for the Bayesian inference on the mean of a normal process. In the scalar case, (2.8) has also been derived by Finucan (1971). Note however that this rule of additive precision should not be used componentwise unless $V(\theta)$ and $V(x|\theta)$ are both diagonal, which is fairly unusual.

In the light of formula (2.6) it may be illuminating to rewrite (2.7) as follows:

$$V(\eta) = A\, V_0\, A' + (I - A)\, V_1\, (I-A)'. \tag{2.9}$$

This fact has been noticed by Stone (1963) for the estimation of a mean in the one-dimensional case (with x being the sample mean).

Suppose one is ready to specify the functional form of the sampling distribution but that the computation of $E(\theta|x)$ is difficult or that robustness w.r.t. the prior specification is desired. In such a case, Rao-Blackwellization may be useful. Let s be a sufficient statistic and $x^* = E(x|s,\theta) = E(x|s)$. Then $\hat{E}(\theta|x^*)$ will improve $\hat{E}(\theta|x)$ in the following sense. Let starred symbols be associated with x^* instead of x. Clearly $V_0^* = V_0$; furthermore $V_1^* \leq V_1$ by Rao-Blackwell's theorem. Therefore, from (2.7), $V(\eta^*) \leq V(\eta)$.

2.2. *Exchangeability*

2.2.1. *Introduction*

We now consider exchangeable processes, i.e. processes where the finite dimensional distributions are invariant under permutation of indices (see *e.g.* Hewitt and Savage (1955)). This class of processes generalizes the class of I.I.D. processes and also includes the mixtures of I.I.D. processes. Thus these processes arise naturally when nuisance parameters are integrated out so as to get marginalized likelihood (and prior distribution) on the parameters of interest alone. Integration of part of the parameters may also be motivated by paying attention to robustness: in a two parameter problem, for instance, the prior distribution $D(\theta_2|\theta_1)$ may be rather easily assigned while on θ_1 a more robust procedure may be preferred, *e.g.* by assigning only the first moment of θ_1.

Here we concentrate attention on the first two moments of a finite sequence $x = (x_1,\ldots,x_n)$ generated by such a process. In this case, $p = n$, the sample size, and q is arbitrary. We first analyse the implications of exchangeability only on the first moment, then on the first two moments: we shall call these processes first-order and second-order exchangeable.

These processes will give characterization of L.S. approximations similar to (2.4) and (2.5).

2.2.2. *First-order exchangeability*

For expository purposes, it is convenient, and not restrictive, to specify the first component of θ as the sampling expectation of the process. First-order exchangeability is then characterized by

$$E(x|\theta) = \theta_1 \mathbf{1} \tag{2.10}$$

where $\mathbf{1} = (1\ 1\ldots1)' \in \mathrm{R}^n$.

Let us decompose $E(\theta)$ and $V(\theta)$ as follows:

$$E(\theta) = [m_i] \qquad i = 1,\ldots,q \tag{2.11}$$

$$V_{\theta\theta} = [v_{ij}] = [v_1 \ldots v_q] \qquad i,j = 1,\ldots,q \tag{2.12}$$

where v_i is the *i-th* column of $V_{\theta\theta}$. First order exchangeability implies

$$E(x) = m_1 \mathbf{1} \tag{2.13}$$

$$V_{\theta x} = v_1 \mathbf{1}' \tag{2.14}$$

$$V_{xx} = V_1 + v_{11}\mathbf{1}\mathbf{1}' \tag{2.15}$$

The L.S. approximation now becomes

$$\hat{E}(\theta \mid x) = m + [1 + v_{11}\mathbf{1}' V_1^{-1}\mathbf{1}]^{-1}\ v_1\mathbf{1}' V_1^{-1}(x - m_1\mathbf{1}) \tag{2.16}$$

and the average measure of precision (1.11) becomes

$$\begin{aligned} V(\eta) &= V_{\theta\theta} - [1 + v_{11}\mathbf{1}' V_1^{-1}\ \mathbf{1}]^{-1}\ \mathbf{1}' V_1^{-1}\mathbf{1}\ v_1 v_1' \\ &= V_{\theta\theta} - [v_{11} + (\mathbf{1}' V_1^{-1}\mathbf{1})^{-1}]^{-1} v_1 v_1' \end{aligned} \tag{2.17}$$

This involves a rather peculiar rule of additive precision analogue to (2.8) (for details, see appendix):

$$[EV(\theta \mid x)]^{-1} \geq V^{-1}_{\theta\theta} + \mathbf{1}' V_1^{-1}\mathbf{1} \begin{bmatrix} 1 & 0 & \dots & 0 \\ 0 & 0 & \dots & 0 \\ \vdots & \vdots & & \vdots \\ 0 & 0 & \dots & 0 \end{bmatrix} \tag{2.18}$$

with equality if and only if $E(\theta \mid x)$ is linear in x.

Note that for the (harmonic) mean of the posterior precisions the sampling improves the lower bound of the element corresponding to θ_1, the mean of the process, only and for θ_1 this improvement is given by the element (1,1) of (2.18):

$$\left[E\ V(\theta \mid x)\right]^{-1}_{11} \geq (v_{11} - v_{12}V_{22}^{-1}v_{21})^{-1} + \mathbf{1}' V_1^{-1}\mathbf{1} \tag{2.19}$$

where $V_{\theta\theta}$ has been partitioned as follows:

$$V_{\theta\theta} = \begin{bmatrix} v_{11} & v_{12} \\ v_{21} & V_{22} \end{bmatrix} \tag{2.20}$$

If, from the start, the model had been marginalized on θ_1, formulae (2.16) and (2.17) would have given:

$$\hat{E}(\theta_1 \mid x) = [v_{11}^{-1} + \mathbf{1}' V_1^{-1}]^{-1}[(m/v_{11}) + \mathbf{1}' V_1^{-1}x] \tag{2.21}$$

$$V(\eta_1) = [v_{11}^{-1} + \mathbf{1}' V_1^{-1}\mathbf{1}]^{-1} \tag{2.22}$$

The role of the simultaneity of the θ_i's may be appreciated by comparing the inverse of (2.22) and the expression (2.19): they are equivalent if, a priori, θ_1 is uncorrelated with the other θ_i's (*i.e.* $v_{12} = 0$).

2.2.3. *Second-order exchangeability*

As for first-order exchangeability, we specify the first three components of θ as follows:

$$\theta_1 = E(x_i|\theta) \qquad i=1,\ldots,n \tag{2.23}$$

$$\theta_2 = V(x_i|\theta) \qquad i=1,\ldots,n \tag{2.24}$$

$$\theta_3 = \operatorname{cov}(x_i,x_j|\theta) \qquad i,j = 1,\ldots,n \quad i\neq j. \tag{2.25}$$

Second-order exchangeability is characterized by the following two conditions:

$$\mathrm{E}(x|\theta) = \theta_1\mathbf{1} \tag{2.26}$$

$$V(x|\theta) = (\theta_2-\theta_3)I_{(n)}+\theta_3\mathbf{1}\mathbf{1}' \tag{2.27}$$

where, again, $\mathbf{1} = (1,1,\ldots 1)' \in \mathbf{R}^n$ and (θ_2,θ_3) are restricted by:

$$(-\theta_2/n\text{-}1)<\theta_3<\theta_2. \tag{2.28}$$

Like $V(x|\theta)$, $V_1 = EV(x|\theta)$ and V_{xx} have the same structure as an intraclass correlation matrix. In particular:

$$V_{xx} = (m_2 - m_3)\, I_{(n)} + (m_3+v_{11})\, \mathbf{1}\mathbf{1}' \tag{2.29}$$

Formula (2.16) specializes then as follows:

$$\hat{\mathrm{E}}(\theta|x) = m+[m_2+(n\text{-}1)m_3+nv_{11}]^{-1}v_1\mathbf{1}'[x-m_1\mathbf{1}] \tag{2.30}$$

We note that (2.30) is a linear function of x, the sample mean, ($x = n^{-1}\mathbf{1}'x$); thus the L.S. *approximation of* θ (*or of* $E(\theta|x)$) *by* x *depends on* x *only*. This will be further analysed in Section 3. As this dependence is linear, we conclude:

$$\hat{E}(\theta|x) = \hat{\mathrm{E}}(\theta|x). \tag{2.31}$$

An alternative proof of (2.31) would run as follows. Since:

$$V(\bar{x}) = n^{-2}\mathbf{1}'V_{xx}\mathbf{1} = n^{-1}[m_2 - m_3 + n(m_3 + \nu_{11})], \tag{2.32}$$

formula (2.30) may be rewritten as follows:

$$\hat{E}(\theta \mid x) = \hat{E}(\theta \mid \bar{x}) = m + \frac{\bar{x} - m_1\nu_1}{V(\bar{x})} \tag{2.33}$$

where, evidently, $\nu_1 = V_{\theta\bar{x}}$.

From (2.17) and (2.32), the average measure of accuracy, $V(\eta)$, takes the form:

$$V(\eta) = V_{\theta\theta} - [V(\bar{x})]^{-1}\nu_1\nu_1' \tag{2.34}$$

The rule of additive precision in (2.18) now becomes

$$[EV(\theta \mid x)]^{-1} \geq V_{\theta\theta}^{-1} + \frac{1}{EV(\bar{x} \mid \theta)} \begin{bmatrix} 1 & 0 & \cdots & 0 & 0 \\ 0 & 0 & & & \vdots \\ \vdots & & \ddots & & \vdots \\ \vdots & & & \ddots & \vdots \\ 0 & \cdots & \cdots & \cdots & 0 \end{bmatrix} \tag{2.35}$$

where:

$$EV(\bar{x} \mid \theta) = n^{-1}[m_2 + (n-1)m_3] \tag{2.36}$$

and with equality in (2.35) if and only if $E(\theta \mid x)$ is linear in x. Note that the uncorrelated case (i.e., $\theta_3 = 0$ a.s.) does not provide substantial simplifications.

3. APPLICATION TO THE ESTIMATION OF A POPULATION MEAN

We now consider a sample $x = (x_1, \ldots, x_n)'$ with sample mean $\bar{x} = n^{-1}\mathbf{1}'x$. Let θ be the only parameter of interest. If $E(\theta \mid x)$ is to be approximated by means of $\bar{x}$ alone, use would be made of:

$$\hat{E}(\theta \mid x) = \alpha + \beta x \tag{2.37}$$

where

$$\alpha = E(\theta) - \beta E(x) \tag{2.38}$$

$$\beta = \frac{\text{cov}(\theta,x)}{V(x)} \tag{2.39}$$

and

$$E\ V(\theta \mid x) \leq V(\theta) - \frac{[\text{cov}(\theta,\bar{x}]^2}{V(x)} \tag{2.40}$$

Clearly this approximation is of interest when θ is the population mean in a first-order exchangeable process *i.e.* $E(x_i \mid \theta) = \theta \ \ i = 1,\ldots,n$. In such a case, x is an unbiased estimator of θ: we may therefore pool the results of Sections 2.1 and 2.2.1., namely:

$$E(x) = E(\theta) = m \tag{2.41}$$

$$\text{cov}(\theta,x) = V(\theta) \tag{2.42}$$

$$V(x) = E\ V(x \mid \theta) + V(\theta). \tag{2.43}$$

Therefore:

$$\hat{E}(\theta \mid x) = a\ m + (1\text{-}a)x \tag{2.44}$$

where

$$a = \frac{V(\theta)}{V(\bar{x})} \qquad 1\text{-}a = \frac{E\ V(x \copyright \theta)}{V(x)} \tag{2.45}$$

and

$$E\ V(\theta \mid x) \leq \{V(\theta)^{-1} + [E\ V(x \mid \theta)]^{-1}\}^{-1} \tag{2.46}$$

with equality if and only if $E(\theta \mid x) = \hat{E}(\theta \mid x)$, (a.s.).

In general we also have:

$$E\ V(\theta \mid x) \leq E\ V(\theta \mid x) \tag{2.47}$$

with equality if and only if $E(\theta|x) = E(\theta|\bar{x})$. (a.s.).

Therefore:

$$E\ V(\theta|x) \leq \{[V(\theta)]^{-1} + [E\ V(\bar{x}|\theta)]^{-1}\}^{-1} \tag{2.48}$$

with equality if and only if $E(\theta|x) = \hat{E}(\theta|\bar{x})$. (a.s.). This allows us to state Ericson's (1969) result in the following way: If $E(\theta|x) = \hat{E}(\theta|x)$ (*i.e.* $E(\theta|x)$ is a linear function of x) then $E(\theta|x)$ has the form (2.44) - (2.45). We may also add that $E\ V(\theta|x)$ is *equal* to the r.h.s. of (2.48).

If the process is second-order exchangeable we get an explicit form for $E\ V(\bar{x}|\theta)$ given in (2.36). With this expression, formula (2.44) - (2.45) reproduce Goldstein's (1975, b) Theorem 1 and formula (2.48) corrects his Corollary 1 (ii) (indeed, the l.h.s. of the inequality is actually the (predictive) expectation of the posterior variance and not the posterior variance itself). Note also that in these relationships, the second-order exchangeability adds only an explicit form for $E\ V(x|\theta)$ and insures that $\hat{E}(\theta|x) = \hat{E}(\theta|\bar{x})$.

If we only know that $E(\theta|x) = \hat{E}(\theta|x)$ (*i.e.* $E(\theta|x)$ is a linear function of x) then $E\ V(\theta|x)$ is equal to the r.h.s. of (2.22). In this case, second-order exchangeability guarantees that $E(\theta|x) = \hat{E}(\theta|\bar{x})$ and, therefore, that $E\ V(\theta|x)$ is equal to the r.h.s. of (2.48). This appears in Ericson (1970) where exchangeability is obtained in the context of finite population.

3. LEAST-SQUARES SUFFICIENCY

In formula (2.31) we have seen a situation where the L.S. approximation depends on x only. One may try to characterize (i.e. to find necessary and sufficient conditions for) situations where L.S. approximation depends on x only. More generally, we may analyze under which conditions the L.S. approximation depends on a transformation of x only. This leads to the concept of "least squares sufficiency". Since $\hat{E}(\theta|x)$ is a linear function of x, one should take care of linear transformation of x only. Hence, the following definition.

Definition Let t: $\mathbf{R}^r \to \mathbf{R}^s$ be a linear transformation of x i.e. $t = Ax$ $(A:sxp)$. Then t is *least-squares sufficient* if and only if $\hat{E}(\theta|x) = \hat{E}(\theta|t(x))$ for any x *(a.s.)*.

Theorem (characterization of L.S. sufficiency)[1]

[1] Comments by A.P. Dawid are gratefully acknowledged as they pointed out an error in a previous version.

Let $t = Ax \qquad (A : s \times p,\ r(A) = s)$
then the following conditions are equivalent:

(i) $\hat{E}(\theta \mid x) = \hat{\mathrm{E}}(\theta \mid t(x))$ almost surely in x,
(ii) $C(A') \supseteq C(V_{xx}^{-1} V_{x\theta})$;
(iii) $\exists\ B\ (q \times s)$ such that $V_{\theta x} V_{xx}^{-1} = BA$,

where $C(\cdot)$ indicates the linear space generated by the columns of a matrix.

Proof
Condition (ii) is clearly equivalent to condition (iii) and condition (i) is equivalent to:

(iv) $$V_{\theta x} V_{xx}^{-1} = V_{\theta x} A'(A V_{xx} A')^{-1} A.$$

Indeed, using a notation similar to that of Section 1 we have:

$$E(t) = A\ E(x) \qquad V_{\theta t} = V_{\theta x} A' \qquad V_{tt} = A V_{xx} A'.$$

As $C(V_{x\theta}) \subseteq C(V_{xx})$, the equivalence between (ii) and (iv) appears clearly once it has been noticed that $A'(A V_{xx} A')^{-1} A V_{xx}$ is a diagonal projection on $C(A')$. Condition (ii) of the theorem gives the geometric motivation of condition (iii) and is indeed equivalent to $t(x_1) = t(x_2) \Rightarrow \hat{\mathrm{E}}(\theta \mid x_1) = \hat{\mathrm{E}}(\theta \mid x_2)$.

Definition The statistic $t = Ax$ is *minimal* L.S. sufficient if and only if $C(A') = C(V_{xx}^{-1} V_{x\theta})$.

In other words, a minimal L.S. sufficient statistic may be constructed from any basis of $C(V_{xx}^{-1} V_{x\theta})$.

As an application we now answer the question considered at the begining of this section: under what condition is $\bar{\mathrm{x}}$ L.S. sufficient? Direct application of the theorem leads to: $\hat{\mathrm{E}}(\theta \mid x) = \hat{\mathrm{E}}(\theta \mid \bar{\mathrm{x}}) \Leftrightarrow \mathbf{1}$ generates the columns of $V_{xx}^{-1} V_{x\theta}$ i.e. $\exists \mathrm{b} \in \mathbf{R}^q$ such that $V_{\theta x} V_{x}^{-1}{}_{x} = b\mathbf{1}'$. Section 2.2 has shown up one such case (with $b = [n V(\bar{\mathrm{x}})]^{-1} v_1$ - see formulae (2.30) and (2.32)).

Appendix: Derivation of (2.18).

Given (2.17), the inverse of $V(\eta)$ may be written as:

$$[V(\eta)]^{-1} = V_{\theta\theta}^{-1}\{I + [1 - [v_{11} + (\mathbf{1}' V_1^{-1} \mathbf{1})^{-1}]^{-1} v_1' V_{\theta\theta}^{-1} v_1]^{-1}$$

$$\cdot [v_{11} + (\mathbf{1}' V_1^{-1} \mathbf{1})^{-1}]^{-1} v_1 v_1' V_{\theta\theta}^{-1}\}.$$

Remember that v_1 is the first column of $V_{\theta\theta}$; this implies:

$$V_{\theta\theta}^{-1} v_1 = \begin{bmatrix} 1 \\ 0 \\ \cdot \\ \cdot \\ \cdot \\ 0 \end{bmatrix}$$

Therefore:

$$v_1' V_{\theta\theta}^{-1} v_1 = v_{11}$$

$$V_{\theta\theta}^{-1} v_1 v_1' V_{\theta\theta}^{-1} = \begin{bmatrix} 1 & 0 & \cdot & \cdot & \cdot & 0 \\ 0 & & & & & \cdot \\ \cdot & & & & & \cdot \\ \cdot & & & & & \cdot \\ 0 & \cdot & \cdot & \cdot & \cdot & 0 \end{bmatrix}$$

from which (2.18) is easily obtained.

ACKNOWLEDGEMENTS

Due to technical problems, the written version circulated at the International meeting on Bayesian statistics was a first draft (of August 1978). In the meantime, useful comments from D.R. Cox, J.B. Kadane and W.J. Rey were received and further work was going on. Consequently, the presentation of the first section has been completely rewritten thanks to the previous comments and those of the official discussants, M. Goldstein and P. Brown.

REFERENCES

BAILEY, A.L. (1950), Credibility Procedures, Laplace's Generalization of Bayes Rule, and the Combination of Collateral Knowledge with Observed Data. *Proceedings of the Casualty Actuarial Society,* **37**, 7-23.

BOUCHAT, A., (1977), *Théorie de la crédibilité: un point de vue non actuariel.* Mémoire du "Diplôme Spécial en Statistique". Université Catholique de Louvain.

BÜHLMANN, H., (1970), *Mathematical Methods in Risk Theory.* Berlin: Springer-Verlag.

— (1971), Credibility Procedures. *Proceedings of the Sixth Berkeley Symposium on Mathematical Statistics and Probability* **1**, 515-525.

DE VIJLDER, FL., (1975), *Introduction aux théories actuarielles de crédibilité.* Office des assureurs de Belgique.

DIACONIS, P. and YLVISAKER, D., (1979), Conjugate Priors for Exponential Families. *Ann. Statist.* **7**, 269-281.

DICKEY, J.M., (1969), Smoothing by Cheating. *Ann. Math. Stat.* **40**, 1477-1482.

DOOB, J.L., (1953), *Stochastic Processes*. New York: Wiley.

ERICSON, W.A., (1969), A note on the Posterior Mean a Population Mean. *J. Roy. Statist. Soc. B,* **31**, 332-334.

— (1970), On the Posterior Mean and Variance of a Population Mean. *J. Amer. Statist. Assoc.* **65**, 649-652.

FINUCAN, H.M., (1971), Posterior precision for Non-Normal Distribution. *J. Roy. Statist. Soc. B,* **33**, 95-97.

GOEL, P.K. (1979), Linear Posterior Expectation in a Scale Parameter Family and the Gamma Distribution. *Technical Report* **163**. Department of Statistics, Carnegie-Mellon University.

GOEL, P., and DeGROOT, H.M., (1979), Only Normal Distribution have Linear Posterior Expectations in Linear Regression. *Technical Report* **157**, Department of Statistics, Carnegie-Mellon University.

GOLDSTEIN, M., (1975a), Approximate Bayes Solutions to Some Non-Parametric Problems. *Ann. Statist.* **3**, 512-517.

— (1975b), A Note on Some Bayesian Non-Parametric Estimates. *Ann. Statist.* **3**, 736-740.

— (1976), Bayesian Analysis of Regression Problems. *Biometrika,* **63**, 51-58.

HARTIGAN, J.A., (1969), Linear Bayesian Methods. *J. Roy. Statist. Soc. B,* **31**, 446-454.

HEWITT, E., and SAVAGE, L., (1955), Symmetric Measures on Cartesian Products. *Trans. Amer. Math. Soc.* **80**, 470-501.

JEWELL, S.W., (1974a), The Credible Distribution. *ASTIN Bulletin* **7**, 237-269.

— (1974b), Credible Means are Exact Bayesian for Exponential Families. *ASTIN Bulletin* **8**, 77-90.

— (1974c), Exact Multidimensional Credibility. *Mitt. der Verein. Schweiz. Versich-Math.* **74**, 193-314.

KAGAN, A., LINNIK, Y.V. and RAO, C.R., (1973), *Characterizations Problems is Mathematical Statistics*. New York: Wiley.

KAHN, P.M., (1975), *Credibility Theory and Application*. New York: Academic Press.

LUKACS, E. and LAHA, R.G. (1964), *Applications of Characteristic Functions*. London: Griffin.

MAYERSON, A.L., (1964), A Bayesian View of Credibility. *Proceedings of the Casualty Actuarial Society,* **51**, 85-104.

RAIFFA, H. and SCHLAIFER, R., (1961), *Applied Statistical Decision Theory*. Harvard: University Press.

STONE, M., (1963), Robustness of Non-Ideal Decision Procedures. *J. Amer. Statist. Assoc.* **58**, 480-486.