The Annals of Statistics
1976, Vol. 4, No. 5, 823-835

CHARACTERIZATION OF PRIOR DISTRIBUTIONS AND SOLUTION TO A COMPOUND DECISION PROBLEM

BY C. RADHAKRISHNA RAO

Indian Statistical Institute

Let $y = \theta + e$ where θ and e are independent random variables so that the regression of y on θ is linear and the conditional distribution of y given θ is homoscedastic. We find prior distributions of θ which induce a linear regression of θ on y. If in addition, the conditional distribution of θ given y is homoscedastic (or weakly so), then θ has a normal distribution. The result is generalized to the Gauss-Markoff model $\mathbf{Y} = \mathbf{X}\boldsymbol{\theta} + \boldsymbol{\epsilon}$ where $\boldsymbol{\theta}$ and $\boldsymbol{\epsilon}$ are independent vector random variables.

Suppose $\bar{y}_i$ is the average of p observations drawn from the ith normal population with mean θ_i and variance σ_0^2 for $i = 1, \cdots, k$, and the problem is the simultaneous estimation of $\theta_1, \cdots, \theta_k$. An estimator alternative to that of James and Stein is obtained and shown to have some advantage.

1. Introduction. The paper is concerned with the following type of problems. There is an unobservable measurement θ on an individual, but observations $y_1, \cdots, y_p$ (the value of p may be unity) may be obtained such that

$$y_i = \alpha\theta + e_i, \qquad i = 1, \cdots, p, \tag{1.1}$$

where e_i are in the nature of errors which do not depend on θ. The measurement θ (or $\alpha\theta$) may be the true value of a characteristic of an individual, and e_i may be measurement errors in repeated trials. How does one estimate (or predict) θ given $y_1, \cdots, y_p$?

There are essentially three approaches. One is the *pivotal statistic* approach of Fisher, where a statistic involving observed and unobserved variables

$$T = f(\theta, y_1, \cdots, y_p) \tag{1.2}$$

is found such that the distribution of T does not involve any unknown element. Inferences on θ are drawn by using the known distribution of T and observed values of $y_1, \cdots, y_p$ without making any further assumptions. The procedure is valid, although the inferential statements on θ in terms of *fiducial probability* advocated by Fisher was subject to some logical criticism.

The second is *straight Bayes*, where a proper or improper prior distribution is imposed on θ and the conditional distribution of θ given $y_1, \cdots, y_p$ is computed (see Lindley, 1971).

The third is *empirical Bayes*, where only a class of prior distributions (sometimes the class of all distributions) is assumed for θ and the conditional distribution (or just expectation) of θ given $y_1, \cdots, y_p$ is estimated (by substituting

Received July 1974.

AMS 1970 *subject classifications.* Primary 62C10; Secondary 62C25.

Key words and phrases. Linear regression, empirical Bayes, prior distribution, characterization problems, simultaneous estimation, compound decision.

estimates for any unknown parameters involved). The success of the method depends on the information available in $y_1, \cdots, y_p$ on the unknown parameters. See Fairfield Smith (1936), Rao (1952, page 329), Rao (1953) and Rao (1973, page 337) for a discussion based on the setup (1.1), and Robbins (1955) for a general theory and other examples.

In the present paper we examine the empirical Bayes approach a little further in the problem mentioned as well as in more general problems. A number of results on characterization of probability distributions which are of wider interest are obtained.

More specifically, the class of prior distributions of θ, which induce a linear regression of θ on $y_1, \cdots, y_p$ under the model (1.1) is obtained. If this happens, under some additional conditions, it is necessary and sufficient that the prior distribution of θ is normal and also that $y_1, \cdots, y_p$ are normally distributed. The results are extended to the Gauss–Markoff model, $\mathbf{Y} = \mathbf{X}\boldsymbol{\theta} + \boldsymbol{\varepsilon}$ involving a vector parameter $\boldsymbol{\theta}$. These results establish the pivotal role of normal prior for $\boldsymbol{\theta}$ if the direct linear regression estimate of $\boldsymbol{\theta}$ on $\mathbf{Y}$ is claimed as optimum.

In the problem of simultaneous estimation of a number of unknowns $\theta_1, \cdots, \theta_k$ under a quadratic loss function, an estimator alternative to that of James and Stein (1961) has been proposed, which is invariant for translations and which seems to have certain advantages.

It has been suggested by a referee that some of the key results which are used in proving the characterization theorems of the paper may be mentioned in the introduction. These are the following:

THEOREM 1.1 (Darmois, 1953, Skitovic, 1954). *Let $X_1, \cdots, X_n$ be independent random variables* (rv's) *taking values in R^1 and suppose that the linear combinations*

$$L_1 = a_1X_1 + \cdots + a_nX_n \qquad \text{and} \qquad L_2 = b_1X_1 + \cdots + b_nX_n$$

are independent. Then for each $i = 1, \cdots, n$ such that $a_i \neq 0$, $b_i \neq 0$, we have that X_i is normally distributed.

THEOREM 1.2 (Ghurye and Olkin, 1962). *Let X_i in Theorem* 1.1 *be* rv's *in R^k and a_i, b_i be constant nonsingular matrices. Then X_i is k-variate normal, $i = 1, \cdots, n$.*

THEOREM 1.3 (Marcinkiewicz, 1938). *If $e^{P(t)}$ is a characteristic function of an* rv, *with $P(t)$ as a polynomial, then the degree of $P(t)$ is utmost two.*

THEOREM 1.4 (Cramér). *If X and Y are independent* rv's *such that $X + Y$ is normally distributed, then X and Y are each normally distributed.*

A theorem of general interest relating to the solution of a functional equation which is repeatedly used in the present paper is as follows.

THEOREM 1.5 (Khatri and Rao, 1972, also Kagan, Linnik and Rao, 1973, page 471). *Let φ_i be a continuous complex valued function of a real p_i-vector variable and $\mathbf{A}_i$, $\mathbf{B}_i$ be matrices of orders $p \times p_i$ and $m \times p_i$ respectively, $i = 1, \cdots, s$,*

such that

$$\sum_1^s \varphi_i(\mathbf{A}_i'\mathbf{t} + \mathbf{B}_i'\mathbf{v}) = C(\mathbf{t}) + D(\mathbf{u}) .$$

(i) *If* $R(\mathbf{A}_i) = R(\mathbf{B}_i) = p_i$, $i = 1, \cdots, s$, *then* $C(\mathbf{t})$ *and* $D(\mathbf{t})$ *are polynomials.*

(ii) *If* $R(\mathbf{A}_i) = p_i$, $i = 1, \cdots, s$, *then* $C(\mathbf{t})$ *is a polynomial and nothing can be said about* $D(\mathbf{u})$ *without any assumption on* $R(\mathbf{B}_i)$.

[In (i) and (ii), $R(X)$ denotes the rank of X. For a more general version of the theorem, see Khatri and Rao, 1972, and also Kagan, Linnik and Rao, 1973, which will be referred to as KLR in the rest of the paper.]

2. Some theorems on characterizations and applications. Let (X_1, X_2) be a bivariate random variable. The conditional distribution of X_2 given X_1 is said to be *homoscedastic* if it depends on X_1 only through the conditional expectation, more precisely if the variables $X_2 - E(X_2 | X_1)$ and X_1 are independently distributed. In such a situation (X_1, X_2) has the structure

$$\begin{aligned} X_2 &= g(u) + e \\ X_1 &= u \end{aligned} \tag{2.1}$$

where u and e are independent random variables. The conditional distribution of X_2 given X_1 is said to be *weakly homoscedastic* if the conditional variance is independent of X_1. Note that the existence of the second moments of (X_1, X_2) is assumed in the definition of weak homoscedasticity but not for homoscedasticity. We prove the following general theorem:

THEOREM 2.1. *Let* (X_1, X_2) *be a bivariate random variable and let nonzero constants* α *and* β *exist such that*

$$\alpha\beta \neq 1 , \tag{2.2}$$

$$X_2 - \alpha X_1 \quad \textit{and} \quad X_1 \quad \textit{are independent,} \tag{2.3}$$

and

$$X_1 - \beta X_2 \quad \textit{and} \quad X_2 \quad \textit{are independent.} \tag{2.4}$$

Then (X_1, X_2) *has a bivariate normal distribution.*

PROOF. Let $Y_1 = X_1$ and $Y_2 = X_2 - \alpha X_1$. Then from (2.3), Y_1 and Y_2 are independent, and from (2.4)

$$L_1 = \alpha Y_1 + Y_2 \qquad \text{and} \qquad L_2 = (1 - \alpha\beta)Y_1 - \beta Y_2$$

are independent. Then by Theorem 1.1, under the conditions of Theorem 2.1, Y_1 and Y_2 are normal and hence (X_1, X_2) has a bivariate normal distribution.

NOTE 1. If $\alpha = 0$, $\beta \neq 0$, then we can only assert that X_1 and X_2 are independent and the marginal distribution of X_2 is normal.

NOTE 2. If $\alpha\beta = 1$, then $X_2 - \alpha X_1$ is degenerate, but nothing can be said about the distribution of X_1.

COROLLARY. *If* $E(X_2 \mid X_1) = \alpha_0 + \alpha X_1$, $E(X_1 \mid X_2) = \beta_0 + \beta X_2$ *and the conditional distribution of each variable given the other depends on the other only in the expression for the mean, then* (X_1, X_2) *has a bivariate normal distribution.*

THEOREM 2.2. *Let* (X_1, X_2) *be a bivariate* rv *such that* $E(X_1) = E(X_2) = 0$, $E(X_2 \mid X_1) = X_1$ *and the conditional distribution of* X_2 *given* X_1 *is homoscedastic. Further let* $f(t_1, t_2)$ *denote the* ch.f. *of* (X_1, X_2). *Then:*

(i) $E(X_1 \mid X_2) = \beta X_2$, $\beta \neq 0$ *or* 1, iff

$$f(t, 0) = c[f(-t, t)]^a \tag{2.5}$$

for some $a > 0$, *in any interval of* t *where* $f(t, 0)$ *and* $f(-t, t)$ *do not vanish simultaneously and the constant* c *depends on the interval.*

(ii) $E(X_1 \mid X_2) = \beta X_2$, $\beta \neq 0$ *or* 1, *and the conditional distribution of* X_1 *given* X_2 *is homoscedastic or weakly homoscedastic* iff (X_1, X_2) *is bivariate normal.*

PROOF OF (i). Let g and h be ch.f.'s of X_1 and $X_2 - X_1$, respectively. By hypothesis $X_2 - X_1$ and X_1 are independent. Then

$$f(t_1, t_2) = g(t_1 + t_2)h(t_2) .$$

Since $E(X_1 \mid X_2) = \beta X_2$, by applying Lemma 1.1.3 of KLR (page 11),

$$(1 - \beta)g'h = \beta gh'$$

where primes denote derivatives, which has the solution (2.5). This proves the necessity of (2.5). Sufficiency is easily established.

The result (ii) under homoscedasticity is already established in Theorem 2.1. We shall establish the result under weak homoscedasticity. Applying Lemma 1.1.3 of KLR (page 11), we obtain the conditions

$$g'h = \beta(gh)' \quad \text{for linearity}$$
$$g''h = -\sigma^2 gh + \beta^2(gh)'' \quad \text{for weak homoscedasticity}$$

where σ^2 is conditional variance. From these equations it follows that

$$(1 - \beta)(\log g)'' = -\sigma^2 \quad \text{near the origin.}$$

If $\sigma^2 \neq 0$, then $\log g$ is quadratic in t near the origin, and hence g is the ch.f. of a normal distribution. Then, so is h and hence (X_1, X_2) is bivariate normal. If $\sigma^2 = 0$, the distribution of (X_1, X_2) is degenerate. Thus the result (ii) is proved. It may be seen that when $\beta = 0$ or $\beta = 1$, either g or h is degenerate.

Theorem 2.2 provides the answer to the question raised about the prior distribution of θ in the model

$$y = \theta + e$$

where θ and e are independent, which induces a linear regression of θ on y. In particular we have the following results.

(a) If the regression of θ on y is linear and e has a normal distribution, then the distribution of θ is also normal by an application of the result (2.5).

(b) If in addition to linear regression, the conditional distribution of θ given y is homoscedastic or weakly so, then the distributions of both θ and e are normal.

THEOREM 2.3. *Let* (X_1, X_2) *be a bivariate* rv *such that* $E(X_1) = E(X_2) = 0$ *and* $E(X_2 \mid X_1) = X_1$. *Further let* $Y = X_1 + X_3$ *where* X_3 *is independent of* (X_1, X_2). *Then* $E(X_2 \mid Y) = \beta Y$, $\beta \neq 0$ *or* 1, iff *the* ch.f.*'s* φ_1 *and* φ_3 *of* X_1 *and* X_3 *satisfy the relationship*

$$\varphi_1(t) = c[\varphi_3(t)]^a \tag{2.6}$$

for some $a > 0$, *in any interval of* t *where* φ_1 *and* φ_3 *do not vanish simultaneously, where* c *is a constant depending on the interval.*

The proof is omitted as the result can be obtained by a direct application of the condition of linearity of regression given in Lemma 1.1.3 of KLR (page 11).

The result (2.6) is interesting since it shows that the nature of regression is altered if the independent variable is subject to an independent error, unless the ch.f.'s of the independent variable and the error satisfy a certain relationship.

In Theorems 2.2 and 2.3, the condition $E(X_2 \mid X_1) = X_1$ can be replaced by $E(X_2 \mid X_1) = \alpha X_1$, $\alpha \neq 0$. In such a case we can consider the variables X_2 and $Y = \alpha X_1$ and apply the results of the theorems.

3. The Gauss–Markoff model. Let us consider the Gauss–Markoff model

$$E(\mathbf{Y} \mid \boldsymbol{\theta}) = \mathbf{X}\boldsymbol{\theta} \tag{3.1}$$

with the additional condition that the conditional distribution of $\mathbf{Y}$ given $\boldsymbol{\theta}$ is homoscedastic, i.e., $\mathbf{Y}$ has the structure

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\theta} + \boldsymbol{\epsilon} \tag{3.2}$$

where $\boldsymbol{\theta}$ and $\boldsymbol{\epsilon}$ are independent. The problem is to find prior distributions of $\boldsymbol{\theta}$ which enable us to reverse the relationship (3.1) and make the following types of statements:

$$E(\boldsymbol{\theta} \mid \mathbf{Y}) = \mathbf{BY}\,. \tag{3.3}$$

$$E(\boldsymbol{\theta} \mid \mathbf{Y}) = \mathbf{BY} \quad \text{and the conditional distribution of } \boldsymbol{\theta} \text{ given } \mathbf{Y} \text{ is homoscedastic.} \tag{3.4}$$

$$E(\boldsymbol{\theta} \mid \mathbf{Y}) = \mathbf{BY} \quad \text{and the conditional distribution of } \boldsymbol{\theta} \text{ given } \mathbf{Y} \text{ is weakly homoscedastic.} \tag{3.5}$$

In Section 2, we solved the problem when both $\boldsymbol{\theta}$ and $\mathbf{Y}$ are one-dimensional variables. The same kinds of results hold when $\boldsymbol{\theta}$ and $\mathbf{Y}$ are vectors. We show that for (3.4) and (3.5) to hold, $\boldsymbol{\theta}$ (or more specifically $\mathbf{X}\boldsymbol{\theta}$) should have a multivariate normal distribution (m.n.d.) while for (3.3) a simple relationship between the ch.f.'s of $\boldsymbol{\theta}$ and $\boldsymbol{\epsilon}$ is sufficient. The following theorems are proved:

THEOREM 3.1. *Let* $\mathbf{Y} = \mathbf{X}\boldsymbol{\theta} + \boldsymbol{\varepsilon}$, *where* $\boldsymbol{\theta}$ *and* $\boldsymbol{\varepsilon}$ *are independent and* $\boldsymbol{\theta} = \mathbf{BY} + \boldsymbol{\eta}$, *where* $\mathbf{Y}$ *and* $\boldsymbol{\eta}$ *are independent, and* $\mathbf{X}$, $\mathbf{B}$ *are given matrices. Then the following hold*:

(i) $(\mathbf{X}\boldsymbol{\eta}, \mathbf{B}\boldsymbol{\varepsilon})$ (i.e., $\mathbf{X}\boldsymbol{\eta}$ *and* $\mathbf{B}\boldsymbol{\varepsilon}$ *jointly*) *has an* m.n.d.

(ii) *If* $\boldsymbol{\varepsilon}$ *has a nonsingular distribution, then* $(\mathbf{X}\boldsymbol{\theta}, \mathbf{BY})$ *has an* m.n.d. *If rank* X *is equal to the number of components of the vector* $\boldsymbol{\theta}$, *then* $(\boldsymbol{\theta}, \mathbf{BY})$ *has an* m.n.d.

(A problem of the type mentioned in Theorem 3.1 was partially investigated by Fisk (1970). Unfortunately, his results do not seem to be correct.)

PROOF OF (i). By hypothesis

$$E[\exp(i\mathbf{t}_1'(\mathbf{X}\boldsymbol{\theta} - \mathbf{CY}) + i\mathbf{t}_2\mathbf{Y})] = E[\exp i\mathbf{t}_1'(\mathbf{X}\boldsymbol{\theta} - \mathbf{CY})] \cdot E[\exp i\mathbf{t}_2'\mathbf{Y}], \tag{3.6}$$

where $\mathbf{C} = \mathbf{XB}$. Writing $\mathbf{Y} = \mathbf{X}\boldsymbol{\theta} + \boldsymbol{\varepsilon}$ in (3.6) and denoting the log ch.f. of $\mathbf{X}\boldsymbol{\theta}$ and $\boldsymbol{\varepsilon}$ by f and g respectively, (3.6) becomes

$$f((\mathbf{I} - \mathbf{C}')\mathbf{t}_1 + \mathbf{t}_2) + g(-\mathbf{C}'\mathbf{t}_1 + \mathbf{t}_2) = A_1(\mathbf{t}_1) + A_2(\mathbf{t}_2), \tag{3.7}$$

where A_1 and A_2 are suitably defined functions. Now applying Theorem 1.5, we find that $A_1(\mathbf{t}_1)$ is a polynomial of degree 2 utmost. (Note that the same cannot be said about $A_2(\mathbf{t}_2)$ since the ranks of $\mathbf{C}$ and $\mathbf{I} - \mathbf{C}$ may not be full.) But $A_1(\mathbf{t}_1)$ is the log ch.f. of $\mathbf{X}\boldsymbol{\eta}$, and hence $\mathbf{X}\boldsymbol{\eta} = (\mathbf{I} - \mathbf{C})\mathbf{X}\boldsymbol{\theta} - \mathbf{C}\boldsymbol{\varepsilon}$ has an m.n.d. Since $\boldsymbol{\theta}$ and $\boldsymbol{\varepsilon}$ are independent, by Theorem 1.4,

$$(\mathbf{I} - \mathbf{C})\mathbf{X}\boldsymbol{\theta} \quad \text{and} \quad \mathbf{C}\boldsymbol{\varepsilon} \tag{3.8}$$

have m.n.d.'s. Similarly $\mathbf{B}\boldsymbol{\varepsilon}$ has an m.n.d. Since $\boldsymbol{\theta}$ and $\boldsymbol{\varepsilon}$ are independent, $(\mathbf{I} - \mathbf{C})\mathbf{X}\boldsymbol{\theta}$ and $\mathbf{B}\boldsymbol{\varepsilon}$ are independent. Then $\mathbf{X}\boldsymbol{\eta} = (\mathbf{I} - \mathbf{C})\mathbf{X}\boldsymbol{\theta} - \mathbf{XB}\boldsymbol{\varepsilon}$ and $\mathbf{B}\boldsymbol{\varepsilon}$ are jointly m.n. which proves (i).

To prove (ii), let there exist a vector $\mathbf{b}$ such that $(\mathbf{I} - \mathbf{C})'\mathbf{b} = \mathbf{O}$. Substituting $\mathbf{t}_1 = \mathbf{bv}$ and $\mathbf{t}_2 = \mathbf{C'bu}$ in (3.7), we have the equation

$$g(\mathbf{C'b}(u - v)) = D_1(u) + D_2(v). \tag{3.9}$$

Then $g(\mathbf{C'b}u)$ is linear in u. Since $g(\mathbf{C'b}u)$ is the log ch.f. of $\mathbf{b'C}\boldsymbol{\varepsilon}$, it follows that $\mathbf{b'C}\boldsymbol{\varepsilon}$ is degenerate contrary to assumption. Hence $\mathbf{b'C} = \mathbf{O} = \mathbf{b}'$ by the choice of $\mathbf{b}$, which implies that $(\mathbf{I} - \mathbf{C})$ has a full rank. Then from (3.8), $\mathbf{X}\boldsymbol{\theta}$ has an m.n.d. Since $\mathbf{X}\boldsymbol{\theta}$ and $\mathbf{B}\boldsymbol{\varepsilon}$ are independent and have m.n.d.'s it follows that $\mathbf{X}\boldsymbol{\theta}$ and $\mathbf{BY}$ $(= \mathbf{BX}\boldsymbol{\theta} + \mathbf{B}\boldsymbol{\varepsilon})$ are jointly m.n. The rest of the results in (ii) of Theorem 3.1 follow easily.

THEOREM 3.2. *Let* $\mathbf{Y} = \mathbf{X}\boldsymbol{\theta} + \boldsymbol{\varepsilon}$ *be the Gauss–Markoff model as in* (3.2). *Further let*

$$E(\boldsymbol{\theta} \mid \mathbf{Y}) = \mathbf{BY} \quad \text{and} \quad D(\boldsymbol{\theta} \mid \mathbf{Y}) = \boldsymbol{\Sigma} \quad (\text{independent of } \mathbf{Y}) \tag{3.10}$$

where D *denotes the variance-covariance matrix. Then* $(\boldsymbol{\theta}, \mathbf{BY})$ *is* m.n. *if no linear combination of* $\boldsymbol{\varepsilon}$ *is degenerate and rank* $\mathbf{X}$ *is equal to the number of components of* $\boldsymbol{\theta}$.

To prove the theorem we need the following lemma which is a generalization of Lemma 1.1.3 of KLR (page 11).

LEMMA 3.1. *Let* $f(\mathbf{t}, \mathbf{u})$ *be the joint* ch.f. *of* $(\mathbf{X}_1, \mathbf{X}_2)$, *where* $\mathbf{X}_1$ *and* $\mathbf{X}_2$ *are vector random variables, which may be of different dimensions.*

(i) *If the first order moments exist (in which case the mean values may be taken as zero), and* $E(\mathbf{X}_2 \mid \mathbf{X}_1) = \mathbf{B}\mathbf{X}_1$ *then*

$$\frac{\partial}{\partial \mathbf{u}} f(\mathbf{t}, \mathbf{u})\Big|_{\mathbf{u}=\mathbf{O}} = \mathbf{B}\frac{\partial}{\partial \mathbf{t}} f(\mathbf{t}, \mathbf{O}) \tag{3.11}$$

where the functions involved are vectors of derivatives (see Rao, 1973, pages 71–72).

(ii) *If the second order moments exist,* $E(\mathbf{X}_2 \mid \mathbf{X}_1) = \mathbf{B}\mathbf{X}_1$ *and* $D(\mathbf{X}_2 \mid \mathbf{X}_1) = \mathbf{\Sigma}$ *independent of* $\mathbf{X}_1$, *then*

$$\frac{\partial^2}{\partial \mathbf{u}^2} f(\mathbf{t}, \mathbf{u})\Big|_{\mathbf{u}=\mathbf{O}} = -f(\mathbf{t}, \mathbf{O})\mathbf{\Sigma} + \mathbf{B}\frac{\partial^2}{\partial \mathbf{t}^2} f(\mathbf{t}, \mathbf{O})\mathbf{B}' . \tag{3.12}$$

The results are established on the same lines as in Lemma 1.1.3 of KLR (page 11).

To prove the main theorem, we observe that

$$f(\mathbf{t}, \mathbf{u}) = E[e^{i\mathbf{t}'\mathbf{Y} + i\mathbf{u}'\mathbf{X}\boldsymbol{\theta}}] = h(\mathbf{t} + \mathbf{u})g(\mathbf{t}) \tag{3.13}$$

where h and g are the ch.f.'s of $\mathbf{X}\boldsymbol{\theta}$ and $\boldsymbol{\varepsilon}$ respectively. Observe that

$$E(\boldsymbol{\theta} \mid \mathbf{Y}) = \mathbf{B}\mathbf{Y} \Rightarrow E(\mathbf{X}\boldsymbol{\theta} \mid \mathbf{Y}) = \mathbf{C}\mathbf{Y} \tag{3.14}$$

$$D(\boldsymbol{\theta} \mid \mathbf{Y}) = \mathbf{\Sigma} \Rightarrow D(\mathbf{X}\boldsymbol{\theta} \mid \mathbf{Y}) = \mathbf{\Lambda} \quad \text{(say)}. \tag{3.15}$$

Then an application of (3.11) and (3.12) gives the two equations

$$g(\mathbf{t})(\mathbf{I} - \mathbf{C})\mathbf{H}_1(\mathbf{t}) = h(\mathbf{t})\mathbf{C}\mathbf{G}_1(\mathbf{t}) \tag{3.16}$$

$$g(\mathbf{t})\mathbf{H}_2(\mathbf{t}) = -g(\mathbf{t})h(\mathbf{t})\mathbf{\Lambda} + \mathbf{C}\mathbf{J}(\mathbf{t})\mathbf{C}' \tag{3.17}$$

where $\mathbf{H}_1$, $\mathbf{G}_1$ are vectors of first derivatives of h, g, and $\mathbf{H}_2$, $\mathbf{J}$ are the matrices of second derivatives of h, hg. Differentiating (3.16) and eliminating $\mathbf{J}$ from (3.17), we obtain the equation

$$(\mathbf{I} - \mathbf{C})\left[\frac{\partial}{\partial \mathbf{t}^2} \log h(\mathbf{t})\right] = -\mathbf{\Lambda} . \tag{3.18}$$

If $\mathbf{b}$ is a vector such that $\mathbf{b}'(\mathbf{I} - \mathbf{C}) = \mathbf{O}$, then from (3.16)

$$\mathbf{b}'\mathbf{C}\mathbf{G}_1(\mathbf{t}) = \mathbf{O} \tag{3.19}$$

which shows that a linear combination of $\boldsymbol{\varepsilon}$ is degenerate contrary to hypothesis. Then $\mathbf{b} = \mathbf{O}$ and $\mathbf{I} - \mathbf{C}$ has full rank, and (3.18) gives

$$\frac{\partial^2}{\partial \mathbf{t}^2}[\log h(\mathbf{t})] = -(\mathbf{I} - \mathbf{C})^{-1}\mathbf{\Lambda} . \tag{3.20}$$

Solving, $\log h(\mathbf{t}) = -\mathbf{t}'(\mathbf{I} - \mathbf{C})^{-1}\mathbf{\Lambda}\mathbf{t}$ which is quadratic in $\mathbf{t}$. Since $h(\mathbf{t})$ is the ch.f. of $\mathbf{X}\boldsymbol{\theta}$, the required result is proved. Using the rank condition on $\mathbf{X}$, we find that $\boldsymbol{\theta}$ itself is m.n. Substituting $h(\mathbf{t}) = \exp(-\mathbf{t}'\mathbf{F}\mathbf{t}/2)$ in (3.16), we have on writing $\mathbf{t} = \mathbf{C}'\mathbf{u}$

$$(\mathbf{I} - \mathbf{C})\mathbf{F}\mathbf{C}'\mathbf{u} = \frac{\partial}{\partial \mathbf{u}}[\log g(\mathbf{C}'\mathbf{u})], \tag{3.21}$$

which shows that $g(\mathbf{C}'\mathbf{u})$ is quadratic in $\mathbf{u}$ or $\mathbf{C}\boldsymbol{\varepsilon}$ has an m.n.d. Then using the rank condition on $\mathbf{X}$, $\mathbf{B}\boldsymbol{\varepsilon}$ has an m.n.d. Since $\boldsymbol{\theta}$ and $\boldsymbol{\varepsilon}$ are independent, $(\boldsymbol{\theta}, \mathbf{B}\boldsymbol{\varepsilon})$ is m.n., which establishes Theorem 3.2.

COROLLARY. *Let* $\mathbf{Y} = \mathbf{X}\boldsymbol{\theta} + \boldsymbol{\varepsilon}$ *be the Gauss–Markoff model as in* (3.2) *and further let* $\boldsymbol{\varepsilon}$ *have a nonsingular* m.n. *distribution. If the regression of* $\boldsymbol{\theta}$ *on* $\mathbf{Y}$ *is linear and rank* $\mathbf{X}$ *is equal to the number of components of* $\boldsymbol{\theta}$, *then it is necessary and sufficient that the prior distribution of* $\boldsymbol{\theta}$ *is* m.n.

4. Solution to a compound decision problem.

4.1. *Prediction of the unknown true value.* Consider a vector random variable

$$(\theta, y_1, \cdots, y_p) \tag{4.1}$$

where the first component θ is unobservable. However, it is known that $y_i = \theta + e_i$ where $e_1, \cdots, e_p$ are i.i.d. normal variables with mean zero and variance σ_0^2. If the marginal distribution of θ is normal with mean μ and variance σ_b^2, then it is easily seen that

$$E(\theta \mid y_1, \cdots, y_k) = (\sigma_e^2\mu + \sigma_b^2\bar{y})/(\sigma_e^2 + \sigma_b^2), \tag{4.2}$$

$$V(\theta \mid y_1, \cdots, y_k) = \sigma_e^2\sigma_b^2/(\sigma_e^2 + \sigma_b^2), \tag{4.3}$$

where $\sigma_e^2 = \sigma_0^2/p$. If μ, σ_0^2 and σ_b^2 are known, θ may be estimated or predicted by the formula (4.2) which is denoted by $\hat{\theta}$, with the mean square error of prediction (4.3). It is seen that

$$\begin{aligned} \frac{\sigma_e^2\sigma_b^2}{\sigma_e^2 + \sigma_b^2} &\leqq \sigma_e^2, \\ &\to \sigma_e^2 \quad \text{as} \quad \sigma_b^2 \to \infty, \\ &\to 0 \quad \text{as} \quad \sigma_b^2 \to 0. \end{aligned} \tag{4.4}$$

Thus on the criterion of mean square error, $\hat{\theta}$, the direct regression estimator of θ on $\bar{y}$, is better than the inverse regression estimator $\bar{y}$. Now

$$E(\hat{\theta} - \theta)^2 = \sigma_e^2\sigma_b^2/(\sigma_e^2 + \sigma_b^2), \tag{4.5}$$

$$E[(\bar{y} - \theta)^2 \mid \theta] = \sigma_e^2, \tag{4.6}$$

and

$$E[(\hat{\theta} - \theta)^2 \mid \theta] = [E(\hat{\theta} - \theta)^2](\sigma_b^2 + \lambda^2\sigma_e^2)/(\sigma_b^2 + \sigma_e^2), \tag{4.7}$$

where $\lambda = (\theta - \mu)/\sigma_b$. We have the following inequalities:

$$(4.7) < (4.5) < (4.6) \quad \text{if} \quad \lambda < 1 ,$$
$$(4.5) < (4.7) < (4.6) \quad \text{if} \quad 1 < \lambda < [(2\sigma_b^2 + \sigma_e^2)/\sigma_b^2]^{\frac{1}{2}} ,$$

and

$$(4.5) < (4.6) < (4.7) \quad \text{if} \quad \lambda > [(2\sigma_b^2 + \sigma_e^2)/\sigma_b^2]^{\frac{1}{2}} .$$

The expressions (4.6) and (4.7) represent an individual's loss (with a particular value of θ) in estimating the true value of θ by $\bar{y}$ and $\hat{\theta}$ respectively. It appears that for large values of θ, the individual's loss in using $\hat{\theta}$ is larger than that for $\bar{y}$, which calls for some caution in estimating θ by $\hat{\theta}$ in a routine way on a population of individuals although the overall loss (4.5) which is statistician's loss, is minimized. For further comments on the distinction to be made between statistician's loss and individual's loss and a possible way of obtaining a balance between the two, the reader is referred to Rao (1975a, 1975b).

4.2. *Simultaneous estimation when parameters are unknown.* Suppose we have observations on a sample of k individuals from a population described by the model (4.1), with only the y_{ij} observed:

$$\begin{array}{l} (\theta_1), y_{11}, \cdots, y_{p_1 1} \\ \cdot \quad \cdot \quad \cdot \quad \cdot \quad \cdot \\ (\theta_k), y_{1k}, \cdots, y_{p_k k} , \end{array} \tag{4.8}$$

where the sample size may be different for each individual. The problem is to estimate the unobserved values $\theta_1, \cdots, \theta_k$ of these individuals when θ_i are considered as a random sample from a population with mean μ and variance σ_b^2, but the parameters μ, σ_b and σ_0 are unknown. Such a problem was considered in the more general context of selection procedures in genetics by Fairfield Smith (1936), Panse (1946) and the author (Rao, 1953).[1] These provide early examples of compound decision problems. Interest in this area of research is revived after James and Stein (1961) provided a solution under a model involving repeated observations on a fixed set of individuals. We consider both the models in our discussion.

4.2.1. *Super population model.* First, we consider the model (4.1) and assume that the individuals come from a population where θ has a distribution with mean μ and variance σ_b^2. If the parameters are known, the prediction for the ith individual is, using (4.2),

$$\hat{\theta}_i = \bar{y}_{\cdot i} - \frac{\sigma_0^2}{\sigma_0^2 + p_i \sigma_b^2} (\bar{y}_{\cdot i} - \mu) , \qquad i = 1, \cdots, k , \tag{4.9}$$

where $\bar{y}_{\cdot i} = (y_{1i} + \cdots + y_{p_i i})/p_i$. It is easily shown that these estimators

[1] These investigations were based on a suggestion made by R. A. Fisher. The author's paper (Rao, 1953) develops the theory and explains the computational aspects.

provide the minimum expected loss with respect to the loss function

$$E[(t_1 - \theta_1)^2 + \cdots + (t_k - \theta_k)^2] \tag{4.10}$$

for given estimators $t_1, \cdots, t_k$.

When the parameters are unknown we may estimate them from the observed data (4.8) by analysis of variance. The following computations are well known.

Analysis of Variance

Degrees of freedom	Sums of squares	Expectation
Between $(k-1)$	$B = \sum p_i(\bar{y}_{\cdot i} - \bar{y}_{\cdot\cdot})^2$	$\left(p_\cdot - \frac{\sum p_i^2}{p_\cdot}\right)\sigma_b^2 + (k-1)\sigma_0^2$
Within $(p_\cdot - k)$	W (by subtraction)	$(p_\cdot - k)\sigma_0^2$
Total $p_\cdot - 1$	$\sum\sum (y_{ij} - \bar{y}_{\cdot\cdot})^2$	

$$p_\cdot = \sum p_i\,, \qquad \bar{y}_{\cdot i} = (y_{1i} + \cdots + y_{p_i i})/p_i\,, \qquad \bar{y}_{\cdot\cdot} = \sum\sum y_{ij}/p_\cdot\,.$$

One method of estimating the parameters μ, σ_b^2, σ_e^2 is to equate $\bar{y}_{\cdot\cdot}$, B and W to their expected values. However, there seems to be some advantage in obtaining biased estimates of σ_b^2 and σ_0^2 through the following equations:

$$B = (k-3)\left[\left(p_\cdot - \frac{\sum p_i^2}{p_\cdot}\right)\frac{\sigma_b^2}{k-1} + \sigma_0^2\right], \tag{4.11}$$
$$W = (p_\cdot - k + 2)\sigma_0^2\,.$$

Denoting the estimates so obtained by $\hat{\sigma}_b^2$ and $\hat{\sigma}_0^2$, and substituting in (4.9), we obtain the empirical Bayes estimators

$$\hat{\theta}_i = \bar{y}_{\cdot i} - \frac{\hat{\sigma}_0^2}{p_i\hat{\sigma}_b^2 + \hat{\sigma}_0^2}(\bar{y}_{\cdot i} - \bar{y}_{\cdot\cdot}) \qquad i = 1, \cdots, k\,. \tag{4.12}$$

The motivation for choosing estimators as in (4.11) comes from the investigation in the case when p_i are all equal to p. In such a case we may write the estimators in the form

$$\hat{\theta}_i = \bar{y}_{\cdot i} - \frac{cW}{B}(\bar{y}_{\cdot i} - \bar{y}_{\cdot\cdot})\,, \qquad i = 1, \cdots, k\,, \tag{4.13}$$

where c is suitably determined. The expected loss associated with the estimators in (4.13) is

$$E\sum\left[(\bar{y}_{\cdot i} - \theta_i) - \frac{cW}{B}(\bar{y}_{\cdot i} - \bar{y}_{\cdot\cdot})\right]^2 \tag{4.14}$$
$$= E[\sum(\bar{y}_{\cdot i} - \theta_i)^2] + \frac{c^2W^2}{pB} - \frac{2cW}{p} + \frac{2cW}{B}\sum\theta_i(\bar{y}_{\cdot i} - \bar{y}_{\cdot\cdot})$$

since $B = p\sum(\bar{y}_{\cdot i} - \bar{y}_{\cdot\cdot})^2$. Observing that W and B are independently distributed as chi-squares with $s = k(p-1)$ and $(k-1)$ d.f., with scale factors

σ_0^2 and $p\sigma_b^2 + \sigma_0^2$, respectively, and that

$$E\frac{\sum \theta_i(\bar{y}_{\cdot i} - \bar{y}_{\cdot\cdot})}{\sum (\bar{y}_{\cdot i} - \bar{y}_{\cdot\cdot})^2} = \frac{p\sigma_b^2}{p\sigma_b^2 + \sigma_0^2} \tag{4.15}$$

the expression (4.14) is found to be

$$\frac{\sigma_0^2}{p}\left[k + \frac{s(s+2)c^2\delta}{(k-3)} - 2cs + 2cs(1-\delta)\right], \tag{4.16}$$

where $\delta = \sigma_0^2/(p\sigma_b^2 + \sigma_0^2)$. The expression (4.16) attains the minimum value

$$\frac{\sigma_0^2}{p}\left[k - \frac{(k-3)s\delta}{s+2}\right] \tag{4.17}$$

by choosing $c = (k-3)/(s+2)$. If we use unbiased estimators of σ_b^2 and σ_0^2 from B and W, the value of c in (4.13) will be $(k-1)/s$, and the loss will be slightly more than (4.17).

It is of interest to compare the losses incurred by using the estimators $\bar{y}_{\cdot i}$ (the traditional averages), $\hat{\theta}_i$ (Bayes when population parameters are known) and $\hat{\hat{\theta}}_i$ (empirical Bayes), in the case where p_i are all equal:

$$E[\textstyle\sum_1^k (\bar{y}_{\cdot i} - \theta_i)^2] = \dfrac{k\sigma_0^2}{p} \tag{4.18}$$

$$E[\textstyle\sum_1^k (\hat{\theta}_i - \theta_i)^2] = \dfrac{k\sigma_0^2\sigma_b^2}{p\sigma_b^2 + \sigma_0^2} \tag{4.19}$$

$$E[\textstyle\sum_1^k (\hat{\hat{\theta}}_i - \theta_i)^2] = \dfrac{\sigma_0^2}{p}\left[k - \dfrac{(k-3)s\delta}{s+2}\right] \tag{4.20}$$

$$= \frac{k\sigma_0^2\sigma_b^2}{p\sigma_b^2 + \sigma_0^2} + \frac{\sigma_0^4}{p\sigma_b^2 + \sigma_0^2}\left[k - \frac{(k-3)s}{s+2}\right]. \tag{4.21}$$

It is seen that (4.19) $\leqq$ (4.21) $\leqq$ (4.18), so that the empirical Bayes is better than the simple averages. The additional loss in using estimates instead of parameters in the Bayes solution is the second expression in (4.21),

$$\frac{\sigma_0^4}{p(\sigma_b^2 + \sigma_0^2)}\left[k - \frac{(k-3)s}{s+2}\right]. \tag{4.22}$$

4.2.2. *Conditional loss* (James–Stein problem). In Section 4.2.1, we computed the overall loss for empirical Bayes estimators under a super population model. We shall now compute the conditional loss, i.e., given the true values $\theta_1, \cdots, \theta_k$, and compare with that of the James–Stein (1961) estimator

$$\tilde{\theta}_i = \left(1 - \frac{aW}{\sum \bar{y}_{\cdot i}^2}\right)\bar{y}_{\cdot i}, \qquad i = 1, \cdots, k, \tag{4.23}$$

where a is a suitably chosen constant. In the James–Stein procedure the estimators are scaled down towards the origin, whereas in the empirical Bayes, the estimators are scaled towards the general observed average. James and Stein

(1961) have shown

$$(4.24)\qquad E[\sum(\tilde{\theta}_i-\theta_i)^2\,|\,\theta_1,\cdots,\theta_k]=\frac{\sigma_0^2}{p}\left[k-\frac{s}{s+2}(k-2)^2\cdot E\frac{1}{k-2+2K_1}\right]$$

where the variable K_1 has a Poisson distribution with mean equal to $\sum p\theta_i^2/2\sigma_0^2$.

By using the arguments employed by James and Stein it is easily shown that

$$(4.25)\qquad E[\sum(\hat{\theta}_i-\theta_i)^2\,|\,\theta_1,\cdots,\theta_k]=\frac{\sigma_0^2}{p}\left[k-\frac{s}{s+2}(k-3)^2\cdot E\frac{1}{k-3+2K_2}\right],$$

where K_2 is a Poisson variable with mean equal to $Ep(\theta_i-\bar{\theta})^2/2\sigma_0^2$, $\bar{\theta}=(\sum\theta_i)/k$.

The estimators $\hat{\theta}_i$ are translation invariant unlike those of James and Stein, and the overall loss depends on the *variance* of the true values rather than on the *raw sum of squares*, which gives some advantage to the former estimators over the latter. There is, however, some difference in the minimum loss attained in each case, being approximately 3 for empirical Bayes ($\hat{\theta}_i$) and 2 for James–Stein estimators ($\tilde{\theta}_i$). But the loss stays close to 3 in the case of $\hat{\theta}_i$ so long as the variance of the true values (θ_i) is small however large their average may be. But for $\tilde{\theta}_i$, the loss increases with increase in the average value of θ_i.

NOTE 1. In a recent paper, Efron and Morris (1973) considered a number of alternatives to the James–Stein estimator and compared their relative efficiencies. The main inspiration is through empirical Bayes approach as in the earlier work of Fairfield Smith (1936) and Rao (1952, 1953, 1973). Efron and Morris consider a modification of the James–Stein estimator (see equation (7.1) of their paper) which brings the individual estimators closer to the overall average of individual averages rather than to the origin, which comes out naturally when the expectation of θ defined in (4.1) is different from zero. Section 4 of the present paper provides the framework for deriving empirical Bayes estimators where the parameters are unknown and examining their properties.

NOTE 2. In a recent presidential address to the Royal Statistical Society, Finney (1974) obtains the maximum likelihood estimators of $\theta_1,\cdots,\theta_k$ under the super population model of Section 4.2.1, assuming normality for the observations y_{ij} as well as for parameters θ_i. The estimators have the form (4.11) but not the same expression as in (4.11).

NOTE 3. For further discussion of the compound estimation problem including the computation of bias in $\hat{\theta}_i$, a reference may be made to a recent paper by the author (Rao, 1976).

The results of Section 4 of this paper concerned with the computations (4.20), (4.21) and (4.25) supplement the valuable investigations of Efron and Morris, and the approach suggested by Finney to justify James–Stein type estimators without consideration of a loss function.

REFERENCES

[1] DARMOIS, G. (1953). Analyse générale des laisons stochastiques. *Rev. Inst. Internat. Statist.* **21** 2-8.

[2] EFRON, BRADLEY and MORRIS, C. (1973). Stein's estimation rule and its competitors—an empirical Bayes approach. *J. Amer. Statist. Assoc.* **68** 117-130.

[3] FAIRFIELD SMITH, H. (1936). A discriminant function for plant selection. *Ann. Eugenics* (London) **7** 240.

[4] FINNEY, D. J. (1974). Problems, data and inference. *J. Roy. Statist. Soc. Ser. A* **137** 1-23.

[5] FISK, P. R. (1970). A note on the characterization of the multivariate normal distribution. *Ann. Math. Statist.* **41** 486-496.

[6] GHURYE, S. G. and OLKIN, I. (1962). A characterization of the multivariate normal distribution. *Ann. Math. Statist.* **33** 533-541.

[7] JAMES, W. and STEIN, CHARLES (1961). Estimation with quadratic loss. *Fourth Berkeley Symp. Math. Statist. Prob.* **1** 361-379, Univ. of California Press.

[8] KAGAN, A. M., LINNIK, YU. V. and RAO, C. R. (1973). *Characterization Problems in Mathematical Statistics.* Wiley, New York.

[9] KHATRI, C. G. and RAO, C. R. (1972). Functional equations and characterization of probability laws through linear function of random variables. *J. Multivariate Analysis* **2** 162-173.

[10] LINDLEY, D. V. (1971). The estimation of many parameters. *In Foundations of Statistical Inference.* 435-455. Holt, Rinehart and Winston, New York.

[11] MARCINKIEWICZ, J. (1938). Sur les fonctions indépendantes. *Fund. Math.* **31** 86-102.

[12] PANSE, V. G. (1946). An application of the discriminant function for selection in poultry. *J. Genetics* (London) **47** 242.

[13] RAO, C. RADHAKRISHNA (1952). *Advanced Statistical Methods in Biometric Research.* Wiley, New York. Reprinted in 1970, Hafner, New York.

[14] RAO, C. RADHAKRISHNA (1953). Discriminant function for genetic differentiation and selection. *Sankhyā* **12** 229-246.

[15] RAO, C. RADHAKRISHNA (1973). *Linear Statistical Inference and its Applications* (2nd ed.). Wiley, New York.

[16] RAO, C. RADHAKRISHNA (1975a). Simultaneous estimation of parameters in different linear models and applications to biometric problems. *Biometrics* **31** 545-553.

[17] RAO, C. RADHAKRISHNA (1975b). Some thoughts on regression and prediction. *Sankhyā C* **37** 102-120.

[18] RAO, C. RADHAKRISHNA (1976). Simultaneous estimation of parameters—a compound decision problem. *Symposium on Statist. Theory and Related Topics.* Purdue Univ. 1976.

[19] ROBBINS, H. (1955). An empirical Bayes approach to statistics. *Third Berkeley Symp. Math. Statist. Prob.* **1** 157-163.

[20] SKITOVIC, V. P. (1954). Linear forms of independent random variables and the normal distribution law. *Izv. Akad. Nauk. SSSR* **18** 185-200.

INDIAN STATISTICAL INSTITUTE
7, SJS SANSANWAL MARG
NEW DELHI, 110029, INDIA