J. R. Statist. Soc. B (1984)
46, *No.* 3, *pp.* 431–439

Outlier Models and Prior Distributions in Bayesian Linear Regression

By MIKE WEST

University of Warwick, UK

[Received January 1983. Revised September 1983]

SUMMARY

Bayesian inference in regression models is considered using heavy-tailed error distributions to accommodate outliers. The particular class of distributions that can be constructed as scale mixtures of normal distributions are examined and use is made of them as both error models and prior distributions in Bayesian linear modelling, including simple regression and more complex hierarchical models with structured priors depending on unknown hyperprior parameters.

Keywords: OUTLIER ACCOMMODATION; BAYESIAN LINEAR MODELS; HIERARCHICAL MODELS; SCALE MIXTURES OF NORMAL DISTRIBUTIONS; SHRINKAGE PRIORS; WEIGHTED LEAST SQUARES

1. INTRODUCTION

The modelling of outliers in nominally normal linear regression models using alternative error distributions which are heavy-tailed relative to the normal provides an automatic means of both detecting and accommodating possibly aberrant observations. Such realistic models do, however, often lead to analytically intractable analyses with complex posterior distributions in several dimensions that are difficult to summarize and understand. In this paper we consider a special yet rather wide class of heavy-tailed, unimodal and symmetric error distributions for which the analyses, though apparently intractable, can be examined in some depth by exploiting certain properties of the assumed error form. The distributions concerned are those that can be constructed as scale mixtures of normal distributions. In his paper concerning location parameters, de Finetti (1961) discusses such distributions and suggests the hypothetical interpretation that

> "each observation is taken using an instrument with normal error, but each time chosen at random from a collection of instruments of different precisions, the distribution of the precisions being that indicated (by the mixing distribution)."

We investigate the basic features of such models in Section 2 along with more general concepts of outlier handling within a Bayesian framework.

In Section 3 we extend the earlier ideas to the specification of simple prior distributions adopting heavy-tailed forms, as in Ramsay and Novick (1980) and, in more complex models, Harrison and Stevens (1976). This approach ensures that any marked inconsistency between the prior and the data is highlighted and that possibly suspect components of prior distributions are discounted when they disagree with likelihoods based on reliable data. Finally in Section 4 the extension is made to more structured priors and hierarchical specifications dependent on unknown hyperprior parameters. In the special case of a simple one-way classification model with a shrinkage prior (Smith, 1973) we discuss the potentially misleading and imprecise inferences that can be arrived at in the standard normal analysis and show how a refinement based on heavy-tailed prior distributions using scale mixtures of normals avoids this.

† *Present address*: Dept of Statistics, University of Warwick, Coventry, CV4 7AL, UK

 0035–9246/84/46431 $2.00

2. ERROR MODELS FOR OUTLIERS IN LINEAR REGRESSION

We consider the linear regression model for scalar observations $y_1, \ldots, y_n$ given by

$$y_r = \mathbf{x}_r^{\mathrm{T}}\beta + \epsilon_r, \quad r = 1, \ldots, n, \tag{2.1}$$

where $\mathbf{x}_1, \ldots, \mathbf{x}_n$ is a set of known p-vectors of regressors, β is the p-vector of regression parameters and $\epsilon_1, \ldots, \epsilon_n$ is a set of zero-mean exchangeable random variables with common distribution continuous on $\mathbb{R}$, unimodal and symmetric. We suppose that this distribution has density $p(\epsilon/\sigma)/\sigma$ where σ is an unknown scale parameter and, in order to accommodate observational outliers, choose $p(\cdot)$ to be heavy-tailed relative to the standard normal density.

2.1. *Influence Functions and Outlier Proneness*

Assume initially that σ is known and equal to unity and that the prior for β is $\pi(\beta)$. The influence of individual observations on the posterior distribution $\pi(\beta \mid D_n)$, where $D_n = \{y_r, \mathbf{x}_r; r = 1, \ldots, n\}$, can be investigated initially by considering the posterior score function, assuming differentiability,

$$\frac{d}{d\beta} \ln \pi(\beta \mid D_n) = \frac{d}{d\beta} \ln \pi(\beta) + \sum_{r=1}^{n} \mathbf{x}_r\, g(y_r - \mathbf{x}_r^{\mathrm{T}}\beta) \tag{2.2}$$

where $g(\epsilon) = -d/d\epsilon \ln p(\epsilon)$ is the influence function of the error density $p(\epsilon)$. From (2.2) the effect that the observation y_r has on the posterior score function is determined by the influence function g and thus, in common with the classical M-estimation approach, we should utilize error densities having bounded influence functions. (Andrews *et al.*, 1972).

The role played by $g(\cdot)$ in the Bayesian analysis is investigated further by O'Hagan (1979) in the simple location model where $\beta = \theta$, a scalar, and $\mathbf{x}_r = 1$ for all r. From a Bayesian viewpoint an outlying observation y_n, say, is accommodated if the posterior distribution function $\Pi(\theta \mid D_n)$ converges to $\Pi(\theta \mid D_{n-1})$ for all θ as $\mid y_n \mid$ tends to infinity and this is achieved using outlier-prone error distributions (O'Hagan, 1979), such as the Student t form.

2.2. *Scale Mixtures of Normals*

We can write the influence function of the chosen density in the form

$$g(\epsilon) = h(\epsilon)\,.\,\epsilon \tag{2.3}$$

where $h(\epsilon)$ is non-negative, non-increasing and symmetric about zero. If $p(\epsilon)$ is outlier prone then $h(\epsilon)$ decays faster than $1/\mid \epsilon \mid$ as $\mid \epsilon \mid$ increases. Using this factorization in the score (2.2) we have

$$\frac{d}{d\beta} \ln \pi(\beta \mid D_n) = \frac{d}{d\beta} \ln \pi(\beta) + \sum_{r=1}^{n} \mathbf{x}_r\,.\,h(\epsilon_r)\,.\,\epsilon_r \tag{2.4}$$

where $\epsilon_r = y_r - \mathbf{x}_r^{\mathrm{T}}\beta$; the term involving the data now looks just as if we had taken the ϵ_r to be independently normally distributed with variances λ_r^{-1} where $\lambda_r = h(\epsilon_r)$. This form can be used to provide simple iterative schemes for calculating posterior modes but further useful interpretation is obtained using a result of Chu (1973) that, due to the symmetry of $p(\epsilon)$, we can write

$$p(\epsilon) = \int_0^{\infty} \lambda^{\frac{1}{2}}\, \phi(\lambda^{\frac{1}{2}}\epsilon)\,.\,f(\lambda)\, d\lambda \tag{2.5}$$

where $\phi(\cdot)$ is the standard normal p.d.f. and f is some function on $(0, \infty)$. If f is a density then $p(\epsilon)$ is a scale mixture of normal densities having the conditional specification $(\epsilon \mid \lambda) \sim N(0, \lambda^{-1})$ and λ having prior density $f(\lambda)$. In this case it is easily shown that $E[\lambda \mid \epsilon] = h(\epsilon)$. Andrews and Mallows (1974) discuss conditions under which f is a density and give several examples. One

further wide class, which includes the double exponential, is the exponential power family with $p(\epsilon) \propto \exp(-|\epsilon|^a)$, $0 < a < 2$, for which (2.5) holds with $f(\lambda) \propto \lambda^{-\frac{1}{2}} q(\lambda)$, and $q(\lambda)$ the density of the stable distribution of index $a/2$.

3. OUTLIERS IN SIMPLE LINEAR REGRESSION

3.1. *Known Scale Parameter*

Consider the model (2.1) with $p(\epsilon)$ having the form (2.3) and scale parameter σ known. We can now view the errors as being conditionally independent normal with

$$(\epsilon_r \mid \lambda_r) \sim N(0, \sigma^2 . \lambda_r^{-1}), \quad r = 1, \ldots, n, \tag{3.1}$$

ϵ_r independent of λ_s, $r \neq s$, and $\lambda_1, \ldots, \lambda_n$ independent with common density $f(\cdot)$. Let $\Lambda = \{\lambda_1, \ldots, \lambda_n\}$ and, as before, $D_n = \{y_r, \mathbf{x}_r; r = 1, \ldots, n\}$. The conditional normality suggests a conjugate normal prior for β, $\beta \sim N(\mathbf{b}_0, \sigma^2 B_0)$, say, where $\mathbf{b}_0$ and B_0 are known. Then, directly, we have

$$(\beta \mid D_n, \Lambda) \sim N(\mathbf{b}(\Lambda), \sigma^2 B(\Lambda)), \tag{3.2}$$

where

$$B(\Lambda)^{-1} = B_0^{-1} + \sum_{r=1}^{n} \lambda_r \mathbf{x}_r \mathbf{x}_r^{\mathrm{T}}, \tag{3.3}$$

and

$$\mathbf{b}(\Lambda) = B(\Lambda)(B_0^{-1} b_0 + \sum_{r=1}^{n} \lambda_r \mathbf{x}_r y_r). \tag{3.4}$$

The marginal posterior distribution is, of course, directly available as

$$\pi(\beta \mid D_n) \propto \pi(\beta) \prod_{r=1}^{n} p\,[(y_r - \mathbf{x}_r^{\mathrm{T}}\beta)/\sigma]$$

and using (2.2), the posterior mode(s) β^* are easily found to satisfy

$$\beta^* = \mathbf{b}(\Lambda^*) \tag{3.5}$$

where

$$\Lambda^* = \{\lambda_1^*, \ldots, \lambda_n^*\},$$

with

$$\lambda_r^* = h\,[(y_r - \mathbf{x}_r^{\mathrm{T}}\beta^*)/\sigma] = E[\lambda_r \mid D_n, \beta = \beta^*]. \tag{3.6}$$

Notice that an alternative representation of $\pi(\beta \mid D_n)$ is as a mixture of the normal distributions (3.2) with respect to the posterior for $(\Lambda \mid D_n)$, which is generally complex. The asymptotic approximation to $\pi(\beta \mid D_n)$ is a normal form with mode β^* and covariance matrix $G^{-1}(\beta^*)$ where $G(\beta)$ is the information matrix at β given by

$$G(\beta) = -\frac{d^2}{d\beta d\beta^{\mathrm{T}}} \ln \pi(\beta \mid D_n),$$

$$= \sigma^{-2} \left\{ B_0^{-1} + \sum_{r=1}^{h} \mathbf{x}_r \mathbf{x}_r^{\mathrm{T}} g'\,[(y_r - \mathbf{x}_r^{\mathrm{T}}\beta)/\sigma] \right\}.$$

Further, since $g(\epsilon) = h(\epsilon)\epsilon$ we have $g'(\epsilon) = h(\epsilon) + h'(\epsilon)\epsilon$, hence

$$G(\beta^*) = \sigma^{-2} B(\Lambda^*)^{-1} + \sigma^{-2} \sum_{r=1}^{n} \mathbf{x}_r \mathbf{x}_r^{\mathrm{T}} h'(\epsilon_r^*) \epsilon_r^* \tag{3.7}$$

where $\epsilon_r^* = (y_r - \mathbf{x}_r^{\mathrm{T}} \beta^*)/\sigma$. Now using the fact that, for outlier-prone models, $h(\epsilon)$ is a decreasing function of $|\epsilon|$, we have $h'(\epsilon)\epsilon < 0$ and so the second term of (3.7) is negative definite. Therefore

$$G(\beta^*)^{-1} = \sigma^2 B(\Lambda^*) + V \tag{3.8}$$

where V is positive definite. Equation (3.8) reflects the fact that, although the marginal mode β^* is of the form of the conditional mode $\mathbf{b}(\Lambda)$ evaluated at Λ^*, the corresponding covariance matrix $B(\Lambda^*)$ underestimates the uncertainty in $\pi(\beta \mid D_n)$; the addition of the extra term V corrects for this.

Finally notice that, if we adopt an improper uniform reference prior for β by setting $B_0^{-1} = 0$, then $\mathbf{b}(\Lambda)$ and $B(\Lambda)$ are the generalized or weighted least squares vector and covariance matrix respectively and β^* of (3.5) is the usual M-estimate for the model.

3.2. *Unknown Scale Parameter*

When σ is unknown the conditional normality of the errors again suggests a conjugate analysis as follows. For convenience let $\phi = \sigma^{-2}$. Then the conjugate joint prior for β and ϕ is the normal/gamma form given by

$$(\beta \mid \phi) \sim N(\mathbf{b}_0, \phi^{-1} B_0)$$

and

$$\phi \sim d_0^{-1} \chi^2_{c_0},$$

for some c_0, $d_0 > 0$. Then routine calculation (De Groot, 1970, Section 11.10) leads to the joint posterior of the form

$$(\beta \mid D_n, \phi, \Lambda) \sim N(\mathbf{b}(\Lambda), \phi^{-1} B(\Lambda)) \tag{3.9}$$

and

$$(\phi \mid D_n), \Lambda \sim d(\Lambda)^{-1} \chi^2_{c_1}, \tag{3.10}$$

where $\mathbf{b}(\Lambda)$, $B(\Lambda)$ are as in Section 3.1, $c_1 = c_0 + n$ and, if $\mathbf{Y} = (y_1, \ldots, y_n)$ and $X = (\mathbf{x}_1, \ldots, \mathbf{x}_n)^{\mathrm{T}}$, then

$$d(\Lambda) = d_0 + (\mathbf{Y} - X\mathbf{b}(\Lambda))^{\mathrm{T}} \mathbf{Y} + (\mathbf{b}_0 - \mathbf{b}(\Lambda))^{\mathrm{T}} B_0^{-1} \mathbf{b}_0.$$

The mode of (3.10) is $\phi^*(\Lambda) = (c_1 - 2)/d(\Lambda)$ and the discussion of Section 3.1 suggests the use of the asymptotic normal approximation to $\pi(\beta \mid D_n, \phi)$ with mode $\beta^* = \mathbf{b}(\Lambda^*)$ and covariance matrix $G(\beta^*)^{-1}$ of (3.7) with $\sigma^{-2} = \phi$, where, now, $\Lambda^* = \{\lambda_1^*, \ldots, \lambda_n^*\}$ and

$$\lambda_r^* = h\,[(y_r - \mathbf{x}_r^{\mathrm{T}} \beta^*)/\sigma^*] = E\,[\lambda_r \mid D_n, \beta = \beta^*, \sigma = \sigma^*], \tag{3.11}$$

with $\sigma^{*2} = \phi^*(\Lambda^*)^{-1} = d(\Lambda^*)/(c_1 - 2)$.

Finally consider the case of the improper reference prior for β and ϕ given by $\pi(\beta, \phi) \propto \phi^{-1}$. Then, again using standard results, the posterior is of the form of (3.9) and (3.10) but now with $\mathbf{b}(\Lambda)$ and $B(\Lambda)$ as the generalized least squares vector and covariance matrix respectively,

$$c_1 = n - p \quad \text{and} \quad d(\Lambda) = \sum_{r=1}^{n} \lambda_r (y_r - \mathbf{x}_r^{\mathrm{T}} \mathbf{b}(\Lambda))^2,$$

the usual weighted residual sum of squares.

3.3. *Data Analysis using Student t Models*

Relles and Rogers (1977) report encouraging results for analyses of the location problem using mixtures of Student t distributions. In the following simple examples we use t distributions with $k > 0$ degrees of freedom to illustrate the above analysis. In this case we have $h(\epsilon) = (k+1)/(k+\epsilon^2)$ and so the maximum value of λ_r^* in (3.6) is $1 + k^{-1}$, when the standardized residual is zero. Further the score $g(\epsilon)$ has its turning points at $\pm\sqrt{k}$ and thus $g'(\epsilon)$ is negative when $\epsilon^2 > k$. For such values of the residuals the contribution of the corresponding observation to the information matrix in (3.7) is then negative, and the observation becomes doubtful. The calculations are performed using a standard regression package utilizing an iterative weighted least squares routine to compute β^*, σ^{*2} and Λ^* of (3.5), (3.11) and (3.6) respectively, in each case using an improper reference prior.

Now, due to the redescending form of $g(\epsilon)$ in the Student t model, the likelihood may be multimodal. Experience with the simple location problem indicates that this is, however, a remote possibility with a realistic percentage (< 20 per cent, say) of possible outliers. If there are two (or more) sizeable groups of observations providing conflicting information, bimodality may result and convergence to one mode will lead to many of the weights λ_r^* being small. In this case the form of the model, in particular, linearity on the chosen scale and symmetry, should be questioned. Finally, the analysis should be viewed as explorative with the sensitivity to values of k examined using two or three different values. No single analysis is definitive and the values chosen here are simply for illustration.

Example 3.1. Lindley (1979) discusses the estimation of the location parameter of a Student $t-5$ distribution with $\sigma = 1$ based on the observations $-1, -0.3, -0.1, 0.4, 0.9, 1.6, 3$. The posterior is unimodal with mode $\beta^* = 0.495$ as opposed to the arithmetic mean 0.643, and $B(\Lambda^*) = 0.185$ with $G^{-1}(\beta^*) = 0.203$. Finally the weights λ_r^* are given by

$$0.83, 1.07, 1.12, 1.20, 1.16, 0.97, 0.53$$

indicating, as suggested by Lindley, that the final observation is extreme, providing a negative contribution to the information $G(\beta^*)$ with all others being positive.

Example 3.2. Cook *et al.* (1982) discuss a simple straight line regression with an error model that permits only one outlier, adopting a maximum likelihood analysis. They conclude that, on this basis, the observation numbered 9 is the estimated outlier although that numbered 11 is also suspect. In an analysis based on a Student $t-5$ model our fitted regression line is very similar to that of Cook *et al.* and the weights λ_r^* are all greater than 1.1 but for $\lambda_9 = 0.591$ and $\lambda_{11} = 0.593$, so that observations 9 and 11 both provide a negative contribution to the information $G(\beta^*)$.

4. STRUCTURED PRIOR MODELS

Ramsay and Novick (1980) discuss the use of heavy-tailed non-normal distributions for both the error distribution and the prior is the model of Section 3. An extreme example of a heavy-tailed prior for β is, of course, the improper uniform reference prior which is designed to be consistent with any body of observed data. We shall now investigate the use of proper priors which, unlike the normal, have the effect that marked inconsistencies between the prior and reliable data does not lead to the data being discounted in favour of the prior.

4.1. *Simple Priors for Regression Parameters*

Ramsay and Novick (1980) create heavy-tailed priors by modifying the usual normal form in the same way as they modify likelihoods. Thus if we wish to modify our nominally normal prior $\beta \sim N(\mathbf{b}_0, \sigma^2 B_0)$, Ramsay and Novick suggest that we use instead that (possibly improper) prior having score

$$\frac{d}{d\beta} \ln \pi(\beta) = \sigma^{-2}\, h[d(\beta)]\, .B_0^{-1}(\mathbf{b}_0 - \beta)$$

where $d^2(\beta) = \sigma^{-2}(\beta - \mathbf{b}_0)^T B_0^{-1} (\beta - \mathbf{b}_0)$ and $h(u)$ is a decreasing function on $(0, \infty)$ that decays faster than $1/u$. As an example suppose that $h(u) = (k+p)/(k+u^2)$, $k > 0$, in which case $\pi(\beta)$ is a multivariate Student t density with k degrees of freedom. Clearly, by analogy with Section 2.2, we can obtain such priors by supposing the existence of a scalar random variable $\gamma > 0$, say, with density $f(\gamma)$, such that $(\beta \mid \gamma) \sim N(\mathbf{b}_0, \sigma^2 \gamma^{-1} B_0)$. Thus, if

$$q(d) = \int_0^\infty \gamma^{\frac{1}{2}} \phi(\gamma^{\frac{1}{2}} d) . f(\gamma) d\gamma$$

is the density of an outlier-prone distribution then the marginal prior for β has built-in protection against misspecification of $\mathbf{b}_0$ and B_0 in the sense that reliable data that conflicts with the prior is not discounted in favour of the prior. The variable γ can now be included in the set of precision parameters Λ and the analysis of Section 3 performed using the proper, conditionally normal and conjugate prior.

Of course this approach leads to the discrediting of the entire prior in cases of conflict between the prior and a reliable likelihood. In many problems this is not the behaviour we want since much of our prior specification, in terms of $\mathbf{b}_0$ and $\mathbf{B}_0$, may be quite valid with only certain components being in doubt. Consider as a simple example the case of exchangeable, and uncorrelated β_i, $i = 1, \ldots, p$ with $\mathbf{b}_0 = \mathbf{1}b$ and $B_0 = I$. What should we now conclude if an outlier-prone analysis of a data set with an improper uniform prior provides a posterior for β favouring values near b for β_i, $i = 1, \ldots, p-1$, but far from b for θ_p? The modification mentioned above, giving β an elliptically (in this case spherically) symmetric heavy-tailed prior, is not reasonable since it is only the pth component of the prior that disagrees with the data and should, perhaps, be discredited. The solution is to treat the β_i as independent with individual heavy-tailed priors, in this case with the same prior due to the exchangeability assumption. Using normal scale mixtures we would then have a set of p independent variables $\gamma_1, \ldots, \gamma_p$ with common density $f(\cdot)$ such that $(\beta_i \mid \gamma_i) \sim N(b, \sigma^2 \gamma_i^{-1})$, $i = 1, \ldots, p$, and β_i independent of γ_j for $i \neq j$. The analysis will again follow the lines of Section 3 with the set of γ_i variables included in Λ.

The existence of prior correlation causes a problem. Suppose that our assessed covariance matrix B_0 is not diagonal and that we require the same sort of componentwise protection against misspecification. The most obvious method of dealing with this is to "rotate" the prior to orthogonality as follows. The quadratic form in the nominal prior, $(\beta - \mathbf{b}_0)^T B_0^{-1} (\beta - \mathbf{b}_0)$ can be written as $\Sigma (z_r - \mathbf{a}_r^T \beta)^2$, where $\mathbf{a}_r^T$ is the rth row of the $p \times p$ matrix A satisfying $AB_0A^T = I$, and z_r is the rth element of $\mathbf{z} = A\mathbf{b}_0$, $r = 1, \ldots, p$. Thus $A\beta \sim N(\mathbf{z}, I)$ and the corresponding componentwise heavy-tailed form is obtained by replacing the identity covariance matrix with a diagonal form having the variables $\gamma_1, \ldots, \gamma_p$ as diagonal elements. If Γ denotes this matrix then we have $(\beta \mid \Gamma) \sim N(\mathbf{b}_0, C\Gamma C^T)$ where $C = A^{-1}$. Of course it is useful to view the vector $\mathbf{z}$ as a vector of quasi-observations with "prior" regression matrix A in order to use standard regression packages for calculation. Notice that this form allows comparison of A and $X = (\mathbf{x}_1, \ldots, \mathbf{x}_n)^T$; the prior is potentially influential if the rows of A are extreme by comparison with the rows of the original regression matrix X.

Example 4.1. Ramsay and Novick (1980) discuss the analysis of a multiple regression concerning 29 sets of observations on measures of performance in each of three educational tests. We denote the response variable by Y, the two regressors by X_1 and X_2. After lengthy discussion Ramsay and Novick arrive at nominal prior for β, σ^2 as given in Section 3.2 and, following the above discussion, the corresponding quantities $\mathbf{z}$ and A are given by

$$\mathbf{z} = \begin{pmatrix} 43 \\ 0.31 \\ 0.31 \end{pmatrix}, \quad A = \begin{pmatrix} 0.45 & 74.4 & 74.4 \\ 0 & -6.3 & 6.3 \\ -0.12 & 0 & 0 \end{pmatrix}$$

By comparison, the data $\{y_i, (1, x_{1i}, x_{2i})\}$, $i = 1, \ldots, 29$, are such that

$$74 < y_i < 143,\ 65 < x_{1i} < 135,\ 78 < x_{2i} < 129.$$

So at a glance we can see that the first row of A is reasonably consistent with the data although the corresponding element of $\mathbf{z}$ is outside the range of the observations. The other two components come from a completely different part of the design space, however, indicating a qualitative difference between the prior information and that provided by the data.

However, an analysis using Student $t-5$ models for the data and the components of the prior provides no indication of conflict between the data and the prior; the prior specification is consistent with the experimental results and there are no observational outliers. This is confirmed by further analyses with different values of the degrees of freedom parameter for the models; the estimated regression coefficients varied only slightly as k varied indicating that there is really no need to doubt normality.

4.2. *Hierarchical Models*

The ideas underlying the componentwise analysis above are relevant to the application of hierarchical linear models (Lindley and Smith, 1972). Harrison and Stevens (1976) use contaminated normal models to characterize observational outliers and changes in the structure of time-series in a dynamic linear model framework. These changes in structure can be viewed as corresponding to the occurrence of "outliers" in the lower (unobservable) levels of a hierarchical model and hence the need for heavy-tailed distributions in these lower levels. The notion is important in general models but will only be discussed here in connection with the simple yet illuminating example of a one-way classification with a "shrinkage" prior.

The usual, normal model is given in Smith (1973) as

$$(y_{ij} \mid \beta_i) \sim N(\beta_i, \sigma^2), \quad j = 1, \ldots, n_i, \tag{4.1}$$

with

$$\beta_i \sim N(\mu, \sigma^2\tau^2), \quad i = 1, \ldots, p. \tag{4.2}$$

where τ^2 is known. We shall, for convenience, adopt a uniform reference prior for μ, and will consider three different analyses of the model. In each analysis the posterior modes of the normal posterior distributions for the β_i are shrinkage estimators of the form

$$\beta_i^* = \alpha_i \bar{\beta}_i + (1 - \alpha_i)\bar{\mu}, \tag{4.3}$$

where $0 < \alpha_i < 1$,

$$\bar{\beta}_i = \left[\sum_{j=1}^{n_i} a_{ij}\right]^{-1} \sum_{j=1}^{n_i} a_{ij} y_{ij}, \tag{4.4}$$

and

$$\bar{\mu} = \left[\sum_{i=1}^{p} b_i\right]^{-1} \sum_{i=1}^{p} b_i \bar{\beta}_i, \tag{4.5}$$

with the a_{ij} and b_i to be specified. The analyses are as follows.

(i) The standard normal model above has $a_{ij} = 1$ and $b_i = n_i(n_i\tau^2 + 1)^{-1}$ for all i, j. So $\bar{\beta}_i$ is the arithmetic mean of the observations in the ith group, $\bar{\mu}$ is a weighted average of these with the weights simply accounting for the different sample sizes n_i, and $\alpha_i = \tau^2 b_i$.

(ii) To accommodate observational outliers we use the analysis of Section 3, modifying the likelihood to read

$$(y_{ij} \mid \beta_i, \lambda_{ij}) \sim N(\beta_i, \sigma^2\lambda_{ij}^{-1}), \quad j = 1, \ldots, n_i, \quad i = 1, \ldots, p,$$

where the λ_{ij} are independent with common density $f(\cdot)$. Outliers will then be protected against as in Section 3 and, again, iterative calculations will provide β_i^* and the corresponding values of

$$\lambda_{ij}^* = E[\lambda_{ij} \mid \mathbf{Y}, \beta = \beta^*].$$

To examine the form of β_i^* assume that the λ_{ij} are known then we have

$$a_{ij} = \lambda_{ij} \text{ and, if } q_i = \sum_{j=1}^{n_i} \lambda_{ij}, (j = 1, \ldots, n_i), b_i = q_i(q_i\tau^2 + 1)^{-1} \text{ with } \alpha_i = \tau^2 b_i.$$

So not only do the λ_{ij} discount outliers in the calculation of the $\bar{\beta}_i$ in (4.4), but, from (4.5), the influence of $\bar{\beta}_i$ on the overall mean estimate $\bar{\mu}$ is proportional to the total precision in the ith group, q_i.

(iii) Suppose in the analysis of (ii) the values of $\bar{\beta}_i$, $i = 1, \ldots, p-1$, are closely grouped but β_p is quite separate from this group indicating that the pth group mean is extreme by comparison with the other, similar, values. In this case we would like to note that β_p is atypical but, as a result of the model, the overall mean estimate $\bar{\mu}$ is unduly shifted towards $\bar{\mu}_p$ and so, in the shrinkage in (4.3), the very features that we are interested in detecting—the differences between the β_i—tend to be obscured. This "overshrinkage" as a result of the prior (4.2) can be avoided by using a heavy-tailed prior specified by

$$(\beta_i \mid \gamma_i) \sim N(0, \sigma^2\tau^2\gamma_i^{-1}) \quad i = 1, \ldots, p,$$

with the γ_i being independent with a common prior $f(\cdot)$. Again the analysis of Section 3 can be used to calculate the β_i^* and corresponding γ_i^* but for now assume the γ_i to be known. Then, returning to (4.3)-(4.5), we have a_{ij} and q_i as in (ii), $b_i = \gamma_i q_i(q_i\tau^2 + \gamma_i)^{-1}$ and $\alpha_i = \tau^2 b_i/\gamma_i$. As a result the atypical $\bar{\beta}_p$, having a small γ_p, will be discounted in the calculation of $\bar{\mu}$ and, since α_i increases with decreasing γ_i, β_p^* will be shrunk much less towards $\bar{\mu}$ than the other, homogeneous, β_i^*.

Example 4.2. The data of Table 1 is a subset of observations measuring the effect of the application of a sulphur treatment in reducing scab disease of potatoes taken from Cochran and Cox (1957, p. 97). The original experiment actually concerned two factors, the time and level of the application, but we consider the data here as a simple comparison of six treatments to illustrate the above analysis with $p = 6$ and $n_i = 4$ for all i.

TABLE 1

1	2	3	4	5	6
9	16	10	30	18	17
9	10	4	7	24	7
16	18	4	21	12	16
4	18	5	9	19	17

For illustration, the models used are based on Student $t-2$ distributions with $\tau^2 = 0.25$ so that the within group standard deviation is twice that between groups. Table 2 displays the values of the modes β_i^*, $i = 1, \ldots, 6$, in essentially increasing order, and the estimate σ^{*2} for the analyses (i), (ii) and (iii) and the ordinary least square (*OLS*) analysis.

Note the considerable shrinkage in (i) as compared with the raw *OLS* estimates; in (i) groups 2, 4, 5 and 6 are clearly similar with 1 and 3 somewhat lower. The estimates in (ii) are similar to those in (i) with the exception of group 4; β_4^* is shifted downwards essentially due to the accommodation of the outlier $Y_{41} = 30$ which receives a weight $\lambda_{41}^* = 0.18$. Group 5 is similarly, though less markedly, affected due to the discounting of $Y_{52} = 24$ with $\lambda_{52}^* = 0.51$. In (iii) β_3^* is shifted

TABLE 2

Group	3	1	6	2	4	5	σ^{*2}
OLS	5.8	9.5	14.3	15.5	16.8	18.3	35.2
(i)	9.5	11.4	13.8	14.4	15.0	15.8	30.2
(ii)	9.4	11.2	14.0	14.3	12.5	15.0	20.6
(iii)	8.0	11.9	14.2	14.5	13.1	15.4	19.2

downwards and the other group estimates are shrunk more closely together. In (i) and (ii) then, overshrinkage occurred and is remedied in (iii) in which $\gamma_3^* = 0.41$ with the other group weights being larger than 1. Group 3 is thus atypical with the other five essentially homogeneous.

ACKNOWLEDGEMENT

I am grateful to the referees for their comments on an earlier version of this paper.

REFERENCES

Andrews, D. F., Bickel, P. J., Hampel, F. R., Huber, P. J., Rogers, W. M. and Tukey, J. W. (1972) *Robust Estimates of Location: Survey and Advances*. Princeton: Princeton University Press.

Andrews, D. F. and Mallows, C. L. (1974) Scale mixtures of normal distributions. *J. R. Statist. Soc.* B, **36**, 99–102.

Chu, K. C. (1973) Estimation and detection for linear systems with elliptical random variables. *I.E.E.E. Trans. Aut. Con.*, **18**, 499–505.

Cochran, W. G. and Cox, G. M. (1957) *Experimental Designs*, 2nd Edition. London: Wiley.

Cook, R. D., Holschuh, N. and Weisberg, S. (1982) A note on an alternative outlier model. *J. R. Statist. Soc.* B, **44**, 370–376.

De Finetti, B. (1961) The Bayesian approach to the rejection of outliers. *Proceedings 4th Berkeley Symp. Math. Prob. Statist.*, Volume 1, pp. 199–210. Berkeley, Calif.: University Press.

De Groot, M. H. (1970) *Optimal Statistical Decisions*. McGraw-Hill.

Harrison, P. J. and Stevens, C. F. (1976) *Bayesian Forecasting* (with Discussion). *J. R. Statist. Soc.* B, **38**, 205–247.

Lindley, D. V. (1979) Approximate Bayesian methods. In *Bayesian Statistics*, (J. M. Bernardo *et al.*, eds.) Valencia: University Press.

Lindley, D. V. and Smith, A. F. M. (1972) Bayes estimates for the linear model (with Discussion). *J. R. Statist. Soc.* B, **34**, 1–41.

O'Hagan, A. (1979) On outlier rejection phenomena in Bayes inference. *J. R. Statist. Soc.* B, **41**, 358–367.

Ramsay, J. O. and Novick, M. R. (1980) *PLU* robust Bayesian decision theory: point estimation. *J. Amer. Statist. Ass.*, **75**, 901–907.

Relles, D. A. and Rogers, W. H. (1977) Statisticians are fairly robust estimators of location. *J. Amer. Statist. Ass.*, **72**, 107–111.

Smith, A. F. M. (1973) Bayes estimates in one-way and two-way models. *Biometrika*, **60**, 319–330.