BAYESIAN STATISTICS 4, pp. 61–77
J. M. Bernardo, J. O. Berger, A. P. Dawid and A. F. M. Smith, (Eds.)

Robust Sequential Prediction from Non-random Samples: the Election Night Forecasting Case*

JOSÉ M. BERNARDO and F. JAVIER GIRÓN
Generalitat Valenciana, Spain and *Universidad de Málaga, Spain*

SUMMARY

On Election Night, returns from polling stations occur in a highly non-random manner, thus posing special difficulties in forecasting the final result. Using a data base which contains the results of past elections for all polling stations, a robust hierarchical multivariate regression model is set up which uses the available returns as a training sample and the outcome of the campaign surveys as a prior. This model produces accurate predictions of the final results, even with only a fraction of the returns, and it is extremely robust against data transmission errors.

Keywords: HIERARCHICAL BAYESIAN REGRESSION; PREDICTIVE POSTERIOR DISTRIBUTIONS; ROBUST BAYESIAN METHODS.

1. THE PROBLEM

Consider a situation where, on election night, one is requested to produce a sequence of forecasts of the final result, based on incoming returns. Unfortunately, one cannot treat the available results at a given time as a random sample from all polling stations; indeed, returns from small rural communities typically come in early, with a vote distribution which is far removed from the overall vote distribution.

Naturally, one expects a certain geographical consistency among elections in the sense that areas with, say, a proportionally high socialist vote in the last election will still have a proportionally high socialist vote in the present election. Since the results of the past election are available for each polling station, each incoming result may be compared with the corresponding result in the past election in order to learn about the direction and magnitude of the swing for each party. Combining the results already known with a prediction of those yet to come, based on an estimation of the swings, one may hope to produce accurate forecasts of the final results.

Since the whole process is done in real time, with very limited checking possibilities, it is of paramount importance that the forecast procedure (i) should deal appropriately with missing data, since reports from some polling stations may be very delayed, and (ii) should be fairly robust against the influence of potentially misleading data, such as clerical mistakes in the actual typing of the incoming data, or in the identification of the corresponding polling station.

* This paper has been prepared with partial financial help from project number PB87-0607-C02-01/02 of the *Programa Sectorial de Promoción General del Conocimiento* granted by the *Ministerio de Educación y Ciencia*, Spain. Professor José M. Bernardo is on leave of absence from the *Departamento de Estadística e I.O., Universidad de Valencia*, Spain.

In this paper, we offer a possible answer to the problem described. Section 2 describes a solution in terms of a hierarchical linear model with heavy tailed error distributions. In Section 3, we develop the required theory as an extension of the normal hierarchical model; in Section 4, this theory is applied to the proposed model. Section 5 provides an example of the behaviour of the solution, using data from the last (1989) Spanish general election, where intentional "errors" have been planted in order to test the robustness of the procedure. Finally, Section 6 includes additional discussion and identifies areas for future research.

2. THE MODEL

In the Spanish electoral system, a certain number of parliamentary seats are assigned to each province, roughly proportional to its population, and those seats are allocated to the competing parties using a corrected proportional system known as the Jefferson-d'Hondt algorithm (see e.g., Bernardo, 1984, for details). Moreover, because of important regional differences deeply rooted in history, electoral data in a given region are only mildly relevant to a different region. Thus, a sensible strategy for the analysis of Spanish electoral data is to proceed province by province, leaving for a final step the combination of the different provincial predictions into a final overall forecast.

Let r_{ijkl} be the proportion of the valid vote which was obtained in the last election by party i in polling station j, of electoral district k, in county l of a given province. Here, $i = 1, \ldots, p$, where p is the number of studied parties, $j = 1, \ldots, n_{kl}$, where n_{kl} is the number of polling stations in district k of county l; $k = 1, \ldots, n_l$, where n_l is the number of electoral districts in county l, and $l = 1, \ldots, m$, where m is the number of counties (*municipios*) in the province. Thus, we will be dealing with a total of

$$N = \sum_{l=1}^{m} \sum_{k=1}^{n_l} n_{kl}$$

polling stations in the province, distributed over m counties. For convenience, let $\boldsymbol{r}$ generically denote the p-dimensional vector which contains the past results of a given polling station.

Similarly, let y_{ijkl} be the proportion of the valid vote which party i obtains in the present election in polling station j, of electoral district k, in county l of the province under study. As before, let $\boldsymbol{y}$ generically denote the p-dimensional vector which contains the incoming results of a given polling station.

At any given moment, only some of the $\boldsymbol{y}$'s, say $\boldsymbol{y}_1, \ldots, \boldsymbol{y}_n$, $0 \leq n \leq N$, will be known. An estimate of the final distribution of the vote $\boldsymbol{z} = \{z_1, \ldots, z_p\}$ will be given by

$$\hat{\boldsymbol{z}} = \sum_{i=1}^{n} \omega_i \boldsymbol{y}_i + \sum_{i=n+1}^{N} \omega_i \hat{\boldsymbol{y}}_i, \qquad \sum_{i=1}^{N} \omega_i = 1,$$

where the ω's are the relative weights of the polling stations, in terms of number of voters, and the $\hat{\boldsymbol{y}}_j$'s are estimates of the $N - n$ unobserved $\boldsymbol{y}$'s, to be obtained from the n observed results.

Within each electoral district, one may expect similar political behaviour, so that it seems plausible to assume that the observed swings should be exchangeable, i.e.,

$$\boldsymbol{y}_{jkl} - \boldsymbol{r}_{jkl} = \boldsymbol{\alpha}_{kl} + \boldsymbol{e}_{jkl}, \qquad j = 1, \ldots, n_{kl};$$

where the α's describe the average swings within each electoral district and where, for robustness, the e's should be assumed to be from a heavy tailed error distribution.

Moreover, electoral districts may safely be assumed to be exchangeable within each county, so that

$$\alpha_{kl} = \beta_l + u_{kl}, \qquad k = 1, \ldots, n_l,$$

where the β's describe the average swings within each county and where, again for robustness, the u's should be assumed to be from a heavy tailed error distribution.

Finally, county swings may be assumed to be exchangeable within the province, and thus

$$\beta_l = \gamma + v_l, \qquad l = 1, \ldots, m;$$

where γ describes the average expected swing within the province, which will be assumed to be known from the last campaign survey. Again, for robustness, the distribution of the v's should have heavy tails.

In Section 4, we shall make the specific calculations assuming that e, u and v have p-variate Cauchy distributions, centered at the origin and with known precision matrices P_α, P_β and P_γ which, in practice, are estimated from the swings recorded between the last two elections held. The model may however be easily extended to the far more general class of elliptical symmetric distributions.

From these assumptions, one may obtain the joint posterior distribution of the average swings of the electoral districts, i.e.,

$$p(\alpha_1, \ldots, \alpha_{n_m} \mid y_1, \ldots, y_n, r_1, \ldots, r_N)$$

and thus, one may compute the posterior predictive distribution

$$p(z \mid y_1, \ldots, y_n, r_1, \ldots, r_N)$$

of the final distribution of the vote,

$$z = \sum_{i=1}^{n} \omega_i y_i + \sum_{i=n+1}^{N} \omega_i(\alpha_i + r_i), \qquad \sum_{i=1}^{N} \omega_i = 1,$$

where, for each i, α_i is the swing which corresponds to the electoral district to which the polling station i belongs.

A final transformation, using the d'Hondt algorithm, $s = \text{Hondt}[z]$, which associates a partition

$$s = \{s_1, \ldots, s_p\}, \qquad s_1 + \cdots + s_p = S$$

among the p parties of the S seats allocated to the province as a function of the vote distribution z, may then be used to obtain a predictive posterior distribution

$$p(s \mid y_1, \ldots, y_n, r_1, \ldots, r_N) \tag{2.1}$$

over the possible distributions among the p parties of the S disputed seats.

The predictive distributions thus obtained from each province may finally be combined to obtain the desired final result, i.e., a predictive distribution over the possible Parliamentary seat configurations.

3. ROBUST HIERARCHICAL LINEAR MODELS

One of the most useful models in Bayesian practice is the Normal Hierarchical Linear Model (NHLM) developed by Lindley and Smith (1972) and Smith (1973). In their model the assumption of normality was essential for the derivation of the exact posterior distributions of the parameters of every hierarchy and the corresponding predictive likelihoods. Within this setup, all the distributions involved were normal and, accordingly, the computation of all parameters in these distributions was straightforward. However, the usefulness of the model was limited, to a great extent, by the assumption of independent normal errors in every stage of the hierarchy. In this section,

(i) We first generalize the NHLM model to a multivariate setting, to be denoted NMHLM, in a form which may be extended to more general error structures.

(ii) We then generalize that model to a Multivariate Hierarchical Linear Model (MHLM) with rather general error structures, in a form which retains the main features of the NMHLM.

(iii) Next, we show that the MHLM is weakly robust, in a sense to be made precise later, which, loosely speaking, means that the usual NMHLM estimates of the parameters in every stage are distribution independent for a large class of error structures.

(iv) We then develop the theory, and give exact distributional results, for error structures which may be written as scale mixtures of matrix-normal distributions.

(v) Finally, we give more precise results for the subclass of Student's matrix-variate t distributions.

These results generalize the standard multivariate linear model and also extend some previous work by Zellner (1976) for the usual linear regression model.

A k-stage general multivariate normal hierarchical linear model MNHLM, which generalizes the usual univariate model, is given by the following equations, each representing the conditional distribution of one hyperparameter given the next in the hierarchy. It is supposed that the last stage hyperparameter, Θ_k, is known.

$$\begin{aligned} \boldsymbol{Y} \mid \Theta_1 &\sim N(\boldsymbol{A}_1\Theta_1, \boldsymbol{C}_1 \otimes \Sigma) \\ \Theta_i \mid \Theta_{i+1} &\sim N(\boldsymbol{A}_{i+1}\Theta_{i+1}, \boldsymbol{C}_{i+1} \otimes \Sigma); \qquad i = 1, \ldots k-1. \end{aligned} \tag{3.1}$$

In these equations $\boldsymbol{Y}$ is an $n \times p$ matrix which represents the observed data, the Θ_i's are the i-th stage hyperparameter matrices of dimensions $n_i \times p$ and the $\boldsymbol{A}_i$'s are design matrices of dimensions $n_{i-1} \times n_i$ (assuming that $n_0 = n$). The $\boldsymbol{C}_i$'s are positive definite matrices of dimensions $n_{i-1} \times n_{i-1}$ and, finally, Σ is a $p \times p$ positive definite matrix. The matrix of means for the conditional matrix-normal distribution at stage i is $\boldsymbol{A}_i\Theta_i$ and the corresponding covariance matrix is $\boldsymbol{C}_i \otimes \Sigma$, where $\otimes$ denotes the Kronecker product of matrices.

From this model, using standard properties of the matrix-normal distributions, one may derive the marginal distribution of the hyperparameter Θ_i, which is given by

$$\Theta_i \sim N(\boldsymbol{B}_{ik}\Theta_k, \boldsymbol{P}_i \otimes \Sigma), \qquad i = 1, \ldots k-1,$$

where

$$\boldsymbol{B}_{ij} = \boldsymbol{A}_{i+1} \cdots \boldsymbol{A}_j, \qquad i < j;$$

$$\boldsymbol{P}_i = \boldsymbol{C}_{i+1} + \sum_{j=i+1}^{k-1} \boldsymbol{B}_{ij}\boldsymbol{C}_{j+1}\boldsymbol{B}_{ij}'.$$

The predictive distribution of $\boldsymbol{Y}$ given Θ_i is

$$\boldsymbol{Y} \mid \Theta_i \sim N(\boldsymbol{A}_i^*\Theta_i, \boldsymbol{Q}_i \otimes \Sigma),$$

where

$$A_i^* = A_0 A_1 \cdots A_i \quad \text{with} \quad A_0 = I;$$
$$Q_i = \sum_{j=0}^{i-1} A_j^* C_{j+1} A_i^{*\prime}.$$

From this, the posterior distribution of Θ_i given the data Y, $\{A_i\}$ and $\{C_i\}$ is

$$\Theta_i \,|\, Y \sim N(D_i d_i, D_i \otimes \Sigma),$$

with

$$D_i^{-1} = A_i^{*\prime} Q_i^{-1} A_i^* + P_i^{-1};$$
$$d_i = A_i^{*\prime} Q_i^{-1} Y + P_i^{-1} B_{ik} \Theta_k.$$

In order to prove the basic result of this section, the MNHLM (3.1) can be more usefully written in the form

$$\begin{aligned} Y &= A_1 \Theta_1 + U_1 \\ \Theta_i &= A_{i+1} \Theta_{i+1} + U_{i+1}; \qquad i = 1, \ldots k-1, \end{aligned} \tag{3.2}$$

where the matrix of error terms U_i are assumed independent $N(O, C_i \otimes \Sigma)$ or, equivalently, that the matrix $U = (U_1, \ldots, U_k)$ is distributed as

$$\begin{pmatrix} U_1 \\ \vdots \\ U_k \end{pmatrix} \sim N\left[\begin{pmatrix} O \\ \vdots \\ O \end{pmatrix}; \begin{pmatrix} C_1 & \cdots & O \\ \vdots & \ddots & \vdots \\ O & \cdots & C_k \end{pmatrix} \otimes \Sigma \right]. \tag{3.3}$$

Predictive distributions for future data Z following the linear model

$$Z = W_1 \Theta_1 + U_W, \qquad U_W \sim N(O, C_W \otimes \Sigma), \tag{3.4}$$

where Z is a $m \times p$ matrix and U_W is independent of the matrix U, can now be easily derived. Indeed, from properties of the matrix-normal distributions it follows that

$$Z \,|\, Y \sim N(W D_1 d_1, (W D_i W' + C_W) \otimes \Sigma). \tag{3.5}$$

Suppose now that the error vector U is distributed according to the scale mixture

$$U \sim \int N(0, C \otimes \Lambda)\, dF(\Lambda), \tag{3.6}$$

where C represents the matrix whose diagonal elements are the matrices C_i and the remaining elements are zero matrices of the appropriate dimensions, i.e., the diagonal covariance matrix of equation (3.3), and $F(\Lambda)$ is any matrix-distribution with support in the class of positive definite $p \times p$ matrices. Clearly, the usual MNHLM (3.2) can be viewed as choosing a degenerate distribution at $\Lambda = \Sigma$ for F, while, for example, the hypothesis of U being distributed as a matrix-variate Student t distribution is equivalent to F being distributed as an inverted-Wishart distribution with appropriate parameters.

With this notation we can state the following theorem

Theorem 3.1 *If the random matrix $\boldsymbol{U}$ is distributed according to (3.6), then*

i) the marginal distribution of Θ_i is

$$\Theta_i \sim \int N(\boldsymbol{B}_{ik}\Theta_k, \boldsymbol{P}_i \otimes \Lambda)\, dF(\Lambda) \qquad i = 1, \ldots k-1;$$

ii) the predictive distribution of $\boldsymbol{Y}$ given Θ_i is

$$\boldsymbol{Y} \mid \Theta_i \sim \int N(\boldsymbol{A}_i^*\Theta_i, \boldsymbol{Q}_i \otimes \Lambda)\, dF(\Lambda \mid \Theta_i), \qquad i = 1, \ldots k-1;$$

where the posterior distribution of Λ given Θ_i, $F(\Lambda \mid \Theta_i)$, is given by

$$dF(\Lambda \mid \Theta_i) \propto |\Lambda|^{-n_i/2} \exp\Big\{ -\frac{1}{2} tr\Lambda^{-1}(\Theta_i - \boldsymbol{B}_{ik}\Theta_k)'\boldsymbol{P}_i^{-1}(\Theta_i - \boldsymbol{B}_{ik}\Theta_k)\Big\} dF(\Lambda);$$

iii) the posterior distribution of Θ_i given the data $\boldsymbol{Y}$ is

$$\Theta_i \mid \boldsymbol{Y} \sim \int N(\boldsymbol{D}_i\boldsymbol{d}_i, \boldsymbol{D}_i \otimes \Lambda)\, dF(\Lambda \mid \boldsymbol{Y}), \qquad i = 1, \ldots k-1;$$

where the posterior distribution of Λ given $\boldsymbol{Y}$, $F(\Lambda \mid \boldsymbol{Y})$, is given by

$$dF(\Lambda \mid \boldsymbol{Y}) \propto |\Lambda|^{-n/2} \exp\Big\{ -\frac{1}{2} tr\Lambda^{-1}(\boldsymbol{Y} - \boldsymbol{A}_k^*\Theta_k)'\boldsymbol{Q}_k^{-1}(\boldsymbol{Y} - \boldsymbol{A}_k^*\Theta_k)\Big\} dF(\Lambda).$$

Proof. The main idea is, simply, to work conditionally on the scale hyperparameter Λ and, then, apply the results of the MNHLM stated above.

Conditionally on Λ, the error matrices $\boldsymbol{U}_i$ are independent and normally distributed as $\boldsymbol{U}_i \sim N(\boldsymbol{O}, \boldsymbol{C}_i \otimes \Lambda)$; therefore, with the same notation as above, we have

$$\Theta_i \mid \Lambda \sim N(\boldsymbol{B}_{ik}\Theta_k, \boldsymbol{P}_i \otimes \Lambda),$$
$$\boldsymbol{Y} \mid \Theta_i, \Lambda \sim N(\boldsymbol{A}_i^*\Theta_i, \boldsymbol{Q}_i \otimes \Lambda),$$

and

$$\Theta_i \mid \boldsymbol{Y}, \Lambda \sim N(\boldsymbol{D}_i\boldsymbol{d}_i, \boldsymbol{D}_i \otimes \Lambda); \qquad i = 1, \ldots, k.$$

Now, by Bayes theorem,

$$\frac{dF(\Lambda \mid \Theta_i)}{dF(\Lambda)} \propto g(\Theta_i \mid \Lambda), \quad \frac{dF(\Lambda \mid \boldsymbol{Y})}{dF(\Lambda)} \propto h(\boldsymbol{Y} \mid \Lambda),$$

where $g(\Theta_i \mid \Lambda)$ and $h(\boldsymbol{Y} \mid \Lambda)$ represent the conditional densities of Θ_i given Λ and $\boldsymbol{Y}$ given Λ, which are $N(\boldsymbol{B}_{ik}\Theta_k, \boldsymbol{P}_i \otimes \Lambda)$ and $N(\boldsymbol{A}_k^*\Theta_k, \boldsymbol{Q}_k \otimes \Lambda)$, respectively.

From this, by integrating out the scale hyperparameter Λ with respect to the corresponding distribution, we obtain the stated results. ◁

The theorem shows that all distributions involved are also scale mixtures of matrix-normal distributions. In particular, the most interesting distributions are the posteriors of the hyperparameters at every stage given the data, i.e., $\Theta_i \mid \boldsymbol{Y}$. These distributions turn out to be just a scale mixture of matrix-normals. This implies that the usual modal estimator of the Θ_i's, i.e., the mode of the posterior distribution, which is also the matrix of means for those F's with finite first moments, is $\boldsymbol{D}_i\boldsymbol{d}_i$, whatever the prior distribution F of Λ. In this sense,

these estimates are robust, that is, they do not depend on F. However, other parameters and characteristics of these distributions such as the H.P.D. regions for the hyperparameters in the hierarchy depend on the distribution F of Λ.

Note that from this theorem and formula (3.5) we can also compute the predictive distribution of future data $\boldsymbol{Z}$ generated by the model (3.4), which is also a scale mixture.

$$\boldsymbol{Z} \mid \boldsymbol{Y} \sim \int N(\boldsymbol{W}\boldsymbol{D}_1\boldsymbol{d}_1, (\boldsymbol{W}\boldsymbol{D}_1\boldsymbol{W}' + \boldsymbol{C}_W) \otimes \Lambda)\, dF(\Lambda \mid \boldsymbol{Y}). \tag{3.7}$$

More precise results can be derived for the special case in which the $\boldsymbol{U}$ matrix is distributed as a matrix-variate Student t. For the definition of the matrix-variate Student t, we follow the same notation as in Box and Tiao (1973, Chapter 8).

Theorem 3.2. *If $\boldsymbol{U} \sim t(\boldsymbol{O}, \boldsymbol{C}, \boldsymbol{S}; \nu)$ with dispersion matrix $\boldsymbol{C} \otimes \boldsymbol{S}$ and ν degrees of freedom, then*

(i) the posterior distribution of Θ_i given $\boldsymbol{Y}$ is

$$\Theta_i \mid \boldsymbol{Y} \sim t_{n_i p}(\boldsymbol{D}_i\boldsymbol{d}_i, \boldsymbol{D}_i, (\boldsymbol{S} + \boldsymbol{T}); \nu + n),$$

where the matrix $\boldsymbol{T} = (\boldsymbol{Y} - \boldsymbol{A}_k^*\Theta_k)'\boldsymbol{Q}_k^{-1}(\boldsymbol{Y} - \boldsymbol{A}_k^*\Theta_k)$;

(ii) the posterior distribution of Λ is an inverted-Wishart,

$$\Lambda \mid \boldsymbol{Y} \sim InW(\boldsymbol{S} + \boldsymbol{T}, \nu + n).$$

(iii) the predictive distribution of $\boldsymbol{Z} = \boldsymbol{W}_1\Theta_1 + \boldsymbol{U}_W$ is

$$\boldsymbol{Z} \mid \boldsymbol{Y} \sim t_{mp}(\boldsymbol{W}\boldsymbol{D}_1\boldsymbol{d}_1, (\boldsymbol{W}\boldsymbol{D}_1\boldsymbol{W}' + \boldsymbol{C}_W), \boldsymbol{S} + \boldsymbol{T}; \nu + n).$$

Proof. The first result is a simple consequence of the fact that a matrix-variate Student t distribution is a scale mixture of matrix-variate normals. More precisely, if $\boldsymbol{U} \sim t(\boldsymbol{O}, \boldsymbol{C}, \boldsymbol{S}; \nu)$, then $\boldsymbol{U}$ is the mixture given by (3.6), with $F \sim InW(\boldsymbol{S}, \nu)$.

From this representation and Theorem 3.1. iii), we obtain that the inverted-Wishart family for Λ is a conjugate one. In fact,

$$\begin{aligned}\frac{dF(\Lambda \mid \boldsymbol{Y})}{d\Lambda} &\propto |\Lambda|^{-n/2} \exp\left\{-\frac{1}{2}\mathrm{tr}\Lambda^{-1}\boldsymbol{T}\right\} \cdot |\Lambda|^{-(\nu/2+p)} \exp\left\{-\frac{1}{2}\mathrm{tr}\Lambda^{-1}\boldsymbol{S}\right\} \\ &\propto |\Lambda|^{-((\nu+n)/2+p)} \exp\left\{-\frac{1}{2}\mathrm{tr}\Lambda^{-1}(\boldsymbol{T} + \boldsymbol{S})\right\};\end{aligned}$$

and (ii) follows. Finally, substitution of (ii) into (3.7) establishes (iii). ◁

4. PREDICTIVE POSTERIOR DISTRIBUTIONS OF INTEREST

In this section we specialize the results just established to the particular case of the model described in Section 2. In order to derive the predictive distribution of the random quantity $\boldsymbol{z}$ let us introduce some useful notation. Let $\boldsymbol{Y}$ denote the full $N \times p$ matrix whose rows are the vectors $\boldsymbol{y}_i$ of observed and potentially observed results, as defined in Section 2. Partition this matrix into the already observed part $\boldsymbol{y}_1, \ldots, \boldsymbol{y}_n$, i.e., the $n \times p$ matrix $\boldsymbol{Y}_1$ and the unobserved part, the $(N - n) \times p$ matrix $\boldsymbol{Y}_2$ formed with the remaining $N - n$ rows of $\boldsymbol{Y}$. Let $\boldsymbol{R}$ denote the $N \times p$ matrix whose rows are the vectors $\boldsymbol{r}_i$ of past results and $\boldsymbol{R}_1$, $\boldsymbol{R}_2$ the corresponding partitions. By $\boldsymbol{X}$ we denote the matrix of swings, i.e., $\boldsymbol{X} = \boldsymbol{Y} - \boldsymbol{R}$ with $\boldsymbol{X}_1$,

$\boldsymbol{X}_2$ representing the corresponding partitions. Finally, let ω be the row vector of weights $(\omega_1, \ldots, \omega_N)$ and ω_1 and ω_2 the corresponding partition.

With this notation the model presented in Section 2, which in a sense is similar to a random effect model with missing data, can be written as a hierarchical model in three stages as follows

$$\begin{aligned} \boldsymbol{X}_1 &= \boldsymbol{A}_1\Theta_1 + \boldsymbol{U}_1, \\ \Theta_1 &= \boldsymbol{A}_2\Theta_2 + \boldsymbol{U}_2, \\ \Theta_2 &= \boldsymbol{A}_3\Theta_3 + \boldsymbol{U}_3; \end{aligned} \tag{4.1}$$

where $\boldsymbol{X}_1$ is a $n \times p$ matrix of known data, whose rows are of the form $\boldsymbol{y}_{jkl} - \boldsymbol{r}_{jkl}$ for those indexes corresponding to the observed data $\boldsymbol{y}_1, \ldots, \boldsymbol{y}_n$, Θ_1 is an $N \times p$ matrix whose rows are the p-dimensional vectors $\boldsymbol{\alpha}_{kl}$, Θ_2 is an $m \times p$ matrix whose rows are the p-dimensional vectors $\boldsymbol{\beta}_l$ and, finally, Θ_3 is the p-dimensional row vector $\boldsymbol{\gamma}$. The matrices $\boldsymbol{A}_i$ for $i = 1, 2, 3$ have special forms; in fact $\boldsymbol{A}_1$ is an $n \times N$ matrix whose rows are N-dimensional unit vectors, with the one in the place that matches the polling station in district k of county l from which the data arose. $\boldsymbol{A}_2$ is an $N \times m$ matrix whose rows are m-dimensional units vectors, as follows: the first n_1 rows are equal to the unit vector $\boldsymbol{e}_1$, the next n_2 rows are equal to the unit vector $\boldsymbol{e}_2$, and so on, so that the last n_m rows are equal to the unit vector $\boldsymbol{e}_m$. Finally, the $m \times 1$ matrix $\boldsymbol{A}_3$ is the m-dimensional column vector $(1, \ldots, 1)$.

The main objective is to obtain the predictive distribution of $\boldsymbol{z}$ given the observed data $\boldsymbol{y}_1, \ldots, \boldsymbol{y}_n$ and the results from the last election $\boldsymbol{r}_1, \ldots, \boldsymbol{r}_N$. From this, using the d'Hondt algorithm, it is easy to obtain the predictive distribution of the seats among the p parties.

The first step is to derive the posterior of the $\boldsymbol{\alpha}$'s or, equivalently, the posterior of Θ_1 given $\boldsymbol{Y}$ or, equivalently, $\boldsymbol{X}_1$.

From Theorem 3.2, for $k = 3$ we have

$$\begin{aligned} \boldsymbol{D}_1^{-1} &= \boldsymbol{A}_1'\boldsymbol{C}_1^{-1}\boldsymbol{A}_1 + (\boldsymbol{C}_2 + \boldsymbol{A}_2\boldsymbol{C}_3\boldsymbol{A}_2')^{-1} \\ \boldsymbol{d}_1 &= \boldsymbol{A}_1'\boldsymbol{C}_1^{-1}\boldsymbol{X}_1 + (\boldsymbol{C}_2 + \boldsymbol{A}_2\boldsymbol{C}_3\boldsymbol{A}_2')^{-1}\boldsymbol{A}_2\boldsymbol{A}_3\boldsymbol{\gamma}. \end{aligned}$$

The computation of $\boldsymbol{D}^{-1}$ involves the inversion of an $N \times N$ matrix. Using standard matrix identities, $\boldsymbol{D}^{-1}$ can also be written in the form

$$\boldsymbol{D}_1^{-1} = \boldsymbol{A}_1'\boldsymbol{C}_1^{-1}\boldsymbol{A}_1 + \boldsymbol{C}_2^{-1} - \boldsymbol{C}_2^{-1}\boldsymbol{A}_2(\boldsymbol{A}_2'\boldsymbol{C}_2^{-1}\boldsymbol{A}_2 + \boldsymbol{C}_3^{-1})^{-1}\boldsymbol{A}_2'\boldsymbol{C}_2^{-1}$$

which may be computationally more efficient when the matrix $\boldsymbol{C}_2$ is diagonal and m, as in our case, is much smaller than N.

Further simplification in the formulae and subsequent computations result from the hypothesis of exchangeability of the swings formulated in Section 2. This implies that the matrices $\boldsymbol{C}_i$ are of the form $k_i\boldsymbol{I}$, where k_i are positive constants and $\boldsymbol{I}$ are identity matrices of the appropiate dimensions.

Now, the predictive model for future observations is

$$\boldsymbol{X}_2 = \boldsymbol{Y}_2 - \boldsymbol{R}_2 = \boldsymbol{W}\Theta_1 + \boldsymbol{U}_W, \qquad \boldsymbol{U}_W \sim N(\boldsymbol{O}, \boldsymbol{C}_W \otimes \boldsymbol{S});$$

where $\boldsymbol{W}$ is the $(N - n) \times N$ matrix whose rows are N-dimensional unit vectors that have exactly the same meaning as those of matrix $\boldsymbol{A}_1$.

Then, using the results of the preceding section, the predictive ditribution of $\boldsymbol{Y}_2$ given the data $\boldsymbol{Y}_1$ and $\boldsymbol{R}$ is

$$\boldsymbol{Y}_2 \sim t_{(N-n)p}(\boldsymbol{R}_2 + \boldsymbol{W}\boldsymbol{D}_1\boldsymbol{d}_1, \boldsymbol{W}\boldsymbol{D}_1\boldsymbol{W}' + \boldsymbol{C}_W, \boldsymbol{S} + (\boldsymbol{Y}_1 - 1\boldsymbol{\gamma})'\boldsymbol{Q}_3^{-1}(\boldsymbol{Y}_1 - 1\boldsymbol{\gamma}); \nu + n)$$

due to the fact that the matrix $A_3^* = 1$, where 1 is an n column vector with all entries equal to 1.

From this distribution, using properties of the matrix-variate Student t, the posterior of z which is a linear combination of Y_2 is

$$z \mid Y_1, R \sim t_{1p}(\omega_1 Y_1 + \omega_2 R_2 + \omega_2 W D_1 d_1, \\ \omega_2(W D_1 W' + C_W)\omega_2', S + (Y_1 - 1\gamma)' Q_3^{-1}(Y_1 - 1\gamma); \nu + n).$$

This matrix-variate t is, in fact, a multivariate Student t distribution, so that, in the notation of Section 2,

$$p(z \mid y_1, \ldots; y_n, r_1, \ldots, r_N) = \mathrm{St}_p(z \mid m_z, S_z, \nu + n) \tag{4.2}$$

i.e., a p-dimensional Student t, with mean

$$m_z = \omega_1 Y_1 + \omega_2 R_2 + \omega_2 W D_1 d_1,$$

dispersion matrix,

$$\frac{\omega_2(W D_1 W' + C_W)\omega_2'}{\nu + n}(S + (Y_1 - 1\gamma)' Q_3^{-1}(Y_1 - 1\gamma);$$

and $\nu + n$ degrees of freedom.

5. A CASE STUDY: THE 1989 SPANISH GENERAL ELECTION

The methodology described in Section 4 has been tested using the results, for the Province of Valencia, of the last two elections which have been held in Spain, namely the European Parliamentary Elections of June 1989, and the Spanish General Elections of October 1989.

The Province of Valencia has N = 1566 polling stations, distributed among m = 264 counties. The number n_l of electoral districts whithin each county varies between 1 and 19, and the number n_{kl} of polling stations within each electoral district varies between 1 and 57.

The outcome of the October General Election for the p = 5 parties with parliamentary representation in Valencia has been predicted, pretending that their returns are partially unknown, and using the June European Elections as the database. The parties considered were PSOE (socialist), PP (conservative), CDS (liberal), UV (conservative regionalist) and IU (communist).

	5%			20%			90%			**Final**
	Mean	Dev.	Error	Mean	Dev.	Error	Mean	Dev.	Error	
PSOE	40.08	0.46	–0.43	40.39	0.40	–0.13	40.50	0.16	–0.02	40.52
PP	23.72	0.49	–0.40	24.19	0.45	0.07	24.19	0.18	0.07	24.12
CDS	6.28	0.36	–0.20	6.33	0.33	–0.15	6.49	0.13	0.01	6.49
UV	11.88	0.50	0.44	11.62	0.46	0.17	11.42	0.17	–0.02	11.45
IU	10.05	0.40	0.03	9.93	0.37	–0.09	10.01	0.14	–0.02	10.02

Table 1. *Evolution of the percentages of valid votes.*

For several proportions of known returns (5%, 20% and 90% of the total number of votes), Table 1 shows the means and standard deviations of the marginal posterior distributions of

the percentages of valid votes obtained by each of the five parties. The absolute error of the means with respect to the final result actually obtained are also quoted.

It is fairly impressive to observe that, with only 5% of the returns, the absolute errors of the posterior modes are all smaller than 0.5%, and that those errors drop to about 0.15% with just 20% of the returns, a proportion of the vote which is usually available about two hours after the polling stations close. With 90% of the returns, we are able to quote a "practically final" result without having to wait for the small proportion of returns which typically get delayed for one reason or another; indeed, the errors all drop below 0.1% and, on election night, vote percentages are never quoted to more than one decimal place.

In Table 2, we show the evolution, as the proportion of the returns grows, of the posterior probability distribution over the possible allocation of the S=16 disputed seats.

PSOE	PP	CDS	UV	IU	5%	20%	90%	**Final**
8	4	1	2	1	0.476	0.665	0.799	1.000
7	4	1	2	2	0.521	0.324	0.201	0.000
7	5	1	2	1	0.003	0.010	0.000	0.000

Table 2. *Evolution of the probability distribution over seat partitions.*

Interestingly, two seat distributions, namely {8, 4, 1, 2, 1} and {7, 4, 1, 2, 2}, have a relatively large probability from the very beginning. This gives advance warning of the fact that, because of the intrinsically discontinuous features of the d'Hondt algorithm, the last seat is going to be allocated by a few number of votes, to either the socialists or the communists. In fact, the socialists won that seat, but, had the communists obtained 1,667 more votes (they obtained 118,567) they would have won that seat.

Tables 1 and 2 are the product of a very realistic simulation. The numbers appear to be very stable even if the sampling mechanism in the simulation is heavily biased, as when the returns are introduced by city size. The next Valencia State Elections will be held on May 26th, 1991; that night, will be the *première* of this model in *real* time.

6. DISCUSSION

The multivariate normal model NMHLM developed in Section 3 is a natural extension of the usual NHLM; indeed, this is just the particular case which obtains when $p = 1$ and the matrix S is an scalar equal to 1. As defined in (3.1), our multivariate model imposes some restrictions on the structure of the global covariance matrix but, this is what makes possible the derivation of simple formulae for the posterior distributions of the parameters and for the predictive distributions of future observations, all of which are matrix-variate-normal. Moreover, within this setting it is also possible, as we have demonstrated, to extend the model to error structures generated by scale mixtures of matrix-variate-normals. Actually, this may be futher extended to the class of elliptically symmetric distributions, which contains the class of scale mixtures of matrix-variate-normals as a particular case; this will be reported elsewhere. Without the restrictions we have imposed on the covariance structure, further progress on the general model seems difficult.

One additional characteristic of this hierarchical model, that we have not developed in this paper but merits careful attention, is the possibility of sequential updating of the hyperparameters, in a Kalman-like fashion, when the observational errors are assumed to be conditionally independent given the scale matrix hyperparameter. The possibility of combining

the flexibility of modelling the data according to a hierarchical model, with the computational advantages of the sequential characteristics of the Kalman filter deserves, we believe, some attention and further research.

As shown in our motivating example, the use of sophisticated Bayesian modelling in forecasting may provide qualitatively different answers, to the point of modifying the possible uses of the forecast.

REFERENCES

Bernardo, J. M. (1984) Monitoring the 1982 Spanish Socialist victory: a Bayesian analysis. *J. Amer. Statist. Assoc.* **79**, 510–515.

Box, G. E. P. and Tiao, G. C. (1973) *Bayesian Inference in Statistical Analysis*. Reading, MA: Addison-Wesley.

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

Smith, A. F. M. (1973). A general Bayesian linear model. *J. Roy. Statist. Soc. B* **35**, 67–75.

Zellner, A. (1976) Bayesian and non-Bayesian analysis of the regression model with multivariate Student-t error terms. *J. Amer. Statist. Assoc.* **71**, 400–405.

APPENDIX

Tables 3 and 4 below describe, with the notation used in Tables 1 and 2, what actually happened in the Province of Valencia on election night, May 26th, 1991, when $S = 37$ State Parliament seats were being contested.

	5%			20%			90%			**Final**
	Mean	Dev.	Error	Mean	Dev.	Error	Mean	Dev.	Error	
PSOE	41.5	3.6	–1.0	41.6	2.6	–0.9	42.4	2.2	–0.1	42.5
PP	23.5	3.1	0.0	23.4	2.8	–0.1	23.5	1.9	0.0	23.5
CDS	4.4	1.4	1.9	4.8	0.5	2.3	2.9	0.5	0.4	2.5
UV	14.4	2.3	–2.0	13.6	1.3	–2.8	16.0	2.0	–0.4	16.4
IU	9.2	2.0	0.9	9.4	2.2	1.1	8.6	1.9	0.3	8.3

Table 3. *Evolution of the percentages of valid votes.*

PSOE	PP	CDS	UV	IU	5%	20%	90%	**Final**
18	10	0	6	3	0.06	0.02	0.82	1.00
18	9	0	7	3	0.03	0.02	0.04	0.00
17	10	2	5	3	0.03	0.47	0.01	0.00
17	9	2	5	4	0.03	0.17	0.01	0.00
17	10	1	6	3	0.36	0.02	0.01	0.00
18	9	1	6	3	0.11	0.02	0.01	0.00

Table 4. *Evolution of the probability distribution over seat partitions.*

It is easily appreciated by comparison that both the standard deviations of the marginal posteriors, and the actual estimation errors, were far larger in real life than in the example. A general explanation lies in the fact that state elections have a far larger local component

than national elections, so that variances within strata were far larger, specially with the regionalists (UV). Moreover, the liberals (CDS) performed very badly in this election (motivating the resignation from their leadership of former prime minister Adolfo Suarez); this poor performance was very inhomogeneous, however, thus adding to the inflated variances. Nevertheless, essentially accurate final predictions were made with 60% of the returns, and this was done over two hours before any other forecaster was able to produce a decent approximation to the final results.

DISCUSSION

L. R. PERICCHI (*Universidad Simón Bolívar, Venezuela*)

This paper addresses a problem that has captured statisticians' attention in the past. It is one of these public problems where the case for sophisticated statistical techniques, and moreover the case for the Bayesian approach, is put to the test: quick and accurate forecasts are demanded.

The proposal described here has some characteristics in common with previous approaches and some novel improvements. In general this article raises issues of modelling and robustness.

The problem is one on which there is substantial prior information from different sources, like past elections, surveys, etc. Also, exchangeability relationships in a hierarchy are natural. Furthermore, the objective is one of prediction in the form of a probability distribution of the possible configurations of the parliament. Thus, not surprisingly, this paper, as previous articles on the same subject, Brown and Payne (1975, 1984) and Bernardo (1984), have obtained shrinkage estimators, "borrowing strength", setting the problem as a Bayesian Hierarchical Linear model. Bernardo and Girón in the present article get closer to the Brown and Payne modelling than that of Bernardo (1984), since they resort to modelling directly the "swings" rather than modelling the log-odds of the multinomial probabilities. All this, coupled with the great amount of prior information, offers the possibility of very accurate predictions from the very begining of the exercise.

A limitation of the model, as has been pointed out by the authors, is the lack of sequential updating. The incoming data is highly structured —there is certainly a bias of order of declaration— producing a trend rather than a random ordering. This prompts the need for sequential updating in a dynamic model that may be in place just before the election, as the authors confirmed in their verbal reply to the discussion.

The second limitation is in our opinion of even greater importance and that is the lack of "strong" robustness (see below), protecting against unbounded influence of wrong information of counts and/or wrong classification of polling stations; i.e. gross errors or atypical data should not influence unduly the general prediction of the swings. The usual hierarchical normal model has been found extremely sensitive to gross errors, possibly producing large shrinkages in the wrong direction.

At this point a short general discussion is in order. The term 'Bayesian Robustness' covers a wide field within which it can have quite different meanings. The first meaning begins with the recognition of the inevitability of imprecision of probability specifications. Even this first approach admits two different interpretations (that have similarities but also important differences). One is the "*sensitivy analysis*" interpretation (Berger, 1990), which is widely known. The second is the *upper and lower probability* interpretation. The latter is a more radical departure from precise analysis, which rejects the usual axiomatic foundations and derives directly the lower probability from its own axioms for rational behaviour, (Walley, 1990). The second meaning of robustness is closer to the Huber-Hampel notion of

assuming models (likelihoods and/or priors) that avoid unbounded influence of assumptions, but still work with a single probability model. The present paper uses this second meaning of robustness.

The authors address the need for robustness by replacing the normal errors throughout, by scale mixtures of normal errors. Scale mixtures of normal errors as outlier prone distributions have a long history in Bayesian analyses. They were, perhaps, first proposed as a Bayesian way of dealing with outliers by de Finetti (1961) and have been sucessfully used in static and dynamic linear regression, West (1981, 1984).

Let us note in passing that the class of scale mixture of normals has been considered as a class (in the first meaning of robustness mentioned above) by Moreno and Pericchi (1990). They consider an ε-contaminated model but the base prior π_0 is a scale mixture and the mixing distribution is only assumed to belong to a class H, i.e.

$$\Gamma_{\varepsilon,\pi_0}(H,Q) = \left\{\pi(\theta) = (1-\varepsilon)\int \pi_0(\theta|r)h(dr) + \varepsilon q(\theta), q \in Q, h \in H\right\}$$

Examples of different classes of mixing distributions considered are

$$H_1 = \left\{h(d_r) : \int_0^{r_i} h(d_r) = h_i, i = 1 \ldots n\right\}$$

$$H_2 = \left\{h(d_r) : h(r) \text{ unimodal at } r_0 \text{ and } \int_0^{r_0} h(d_r) = h_0\right\}$$

When π_0 is normal and $\varepsilon = 0$ then $\Gamma(H)$ is the class of scale mixtures of normal distributions with mixing distributions in H. The authors report sensible posterior ranges for probabilities of sets using H_1 and H_2.

Going back to the particular scale mixture of normals considered by Bernardo and Girón, they first conveniently write the usual Multivariate Normal Hierarchical model and by restricting to a common scale matrix (Σ in (3.3) or Λ in (3.6)), they are able to obtain an elegant expression of the posterior distributions (Theorem 3.1.). Furthermore in Theorem 3.2, by specializing to a particular combination of Student-t distributions, they are able to get closed form results. This would be surprising, were it not for Zellner's (1976) conjecture: "similar results (as those for regression) will be found with errors following a matrix Student-t". However, as with Zellner's results the authors get "weak" rather than "strong" robustness, in the sense that the posterior mean turns out to be linear in the observations (and therefore non-robust), although other characteristics of the distributions will be robust. However, "*strong*" robustness is what is required, and some *ad hoc* ways to protect against outlying data (like screening) may be required. Also, approximations on combination of models that yield "strong" robustness may be more useful than exact results. Having said that, we should bear in mind that compromises due to time pressure on election night, may have to be made given the insufficient development of the theory of scale mixtures of normals.

Finally, we remark that the elegant (even if too restricted) development of this paper opens wide possibilities for modelling. We should strive for more theoretical insight in the scale mixture of normals, to guide the assessment. For example O'Hagan's "Credence" theory is still quite incomplete. Moreover, scale mixture of normals offers a much wider choice than just the Student-t, that should be explored. So far Bernardo and Girón have shown us encouraging simulations. Let us wish them well on the actual election night.

A. P. DAWID (*University College London, UK*)

It seems worth emphasising that the "robustness" considered in this paper refers to the invariance of the results (formulae for means) in the face of varying Σ in (3.3) or (what is equivalent) the distribution F of (3.6). This distribution can be thought of either as part of the prior (Σ being a parameter) or, on using (3.6) in (3.2), as part of the model — although note that, in this latter case, the important independence (Markov) properties of the system (3.2) are lost. Relevant theory and formulae for both the general "left-spherical" case and the particular Student-t case may be found in Dawid (1977) — see also Dawid (1981, 1988).

At the presentation of this paper at the meeting, I understood the authors to suggest that the methods also exhibit robustness in the more common sense of insensitivity to extreme data values. One Bayesian approach to this involves modelling with heavy tailed prior and error distributions, as in Dawid (1973), O'Hagan (1979, 1988) —in particular, Student-t forms are often suitable. And indeed, as pointed out at the meeting, the model does allow the possibility of obtaining such distributions for all relevant quantitities. In order to avoid any ambiguity, therefore, it must be clearly realized that, even with this choice, this model does *not* possess robustness against outliers. The Bayesian outlier-robustness theory does not apply because, as mentioned above, after using (3.6) with $F \sim InW(S, \nu)$ the (U_i) are no longer independent. Independence is vital for the heavy-tails theory to work — zero correlation is simply not an acceptable alternative. In fact, since the predictive means under the model turn out to be linear in the data, it is obvious that the methods developed in this paper can *not* be outlier-robust.

S. E. FIENBERG (*York University, Canada*)

As Bernardo and Girón are aware, others have used hierarchical Bayesian models for election night predictions. As far as I am aware the earliest such prediction system was set up in the United States.

In the 1960s a group of statisticians working for the NBC televion network developed a computer-based statistical model for predicting the winner in the U.S. national elections for President (by state) and for individual state elections for Senator and Governor. In a presidential-election year, close to 100 predictions are made, otherwise only half that number are required. The statistical model used can be viewed as a primitive version of a Bayesian hierarchical linear model (with a fair bit of what I. J. Good would call ad hockery) and it predates the work of Lindley and Smith by several years. Primary contributors to the election prediction model development included D. Brillinger, J. Tukey, and D. Wallace. Since the actual model is still proprietary, the following description is somewhat general, and is based on my memory of the system as it operated in the 1970s.

In the 1960s an organization called the News Election Service (NES) was formed through a cooperative effort of the three national television networks and two wire services. NES collects data by precinct, from individual precincts and the 3000 county reporting centers and forwards them to the networks and wire services by county (for more details, see Link, 1989). All networks get the same data at the same time from NES.

For each state, at any point in time, there are data from four sources: (i) a prior estimate of the outcome, (ii) key precincts (chosen by their previous correlation with the actual outcome), (iii) county data, (iv) whole-state data (which are the numbers the networks "oficially" report). The NBC model works with estimates of the swings of the differences between % Republican vote and % Democratic vote (a more elaborate version is used for multiple candidates) *relative* to the difference from some previous election. In addition there is a related model for turnout ratios.

The four sources of data are combined to produce an estimate of $[\%R - \%D]/2$ with

an estimated mean square error based on the sampling variance, historical information, and various bias parameters which can be varied depending on circumstances. A somewhat more elaborate structure is used to accomodate elections involving three or more major candidates. For each race the NBC model requires special settings for 78 different sets of parameters, for biases and variances, turnout adjustment factors, stratification of the state, etc. The model usually involves a geographic stratification of the state into four "substates" based on urban/suburban/rural structure and produces estimates by strata, which are then weighted by turnout to produce statewide estimates.

Even with such a computer-based model about a dozen statisticians are required to monitor the flow of data and the model performance. Special attention to the robustness of predictions relative to different historical bases for swings is an important factor, as is collateral information about where the early data are from (e.g., the city of Chicago vs. the Chicago suburbs vs. downstate Illinois).

Getting accurate early predictions is the name of the game in election night forecasting because NBC competes with the other networks on making forecasts. Borrowing strength in the Bayesian-model sense originally gave NBC an advantage over the raw data-based models employed by the other networks. For example, in 1976, NBC called 94 out of 95 races correctly (only the Presidential race in Oregon remained too close to determine) and made several calls of outcomes when the overall percentages favored the eventual loser. In the Texas Presidential race, another network called the Republican candidate as the winner early in the evening at a time when the NBC model was showing the Democratic candidate ahead (but with a large mean square error). Later this call was retracted and NBC was the first to call the Democrat the winner.

The 1980s brought a new phenomenon to U.S. election night predictions: the exit survey of voters (see Link, 1989). As a consequence, the television networks have been able to call most races long before the election polls have closed and before the precinct totals are available. All of the fancy bells and whistles of the kind of Bayesian prediction system designed by Bernardo and Girón or the earlier system designed by NBC have little use in such circumstances, unless the election race is extremely close.

REPLY TO THE DISCUSSION

We are grateful to Professor Pericchi for his valuable comments and for his wish that all worked well on election night. As described in the Appendix above, his wish was reasonably well achieved.

He also refers to the possibility of sequential updating, also mentioned in our final discussion. Assuming, as we do in sections 2 and 4, the hypothesis of exchangeability in the swings —which implies that the C_i matrices in the model are of the form $k_i I$— the derivation of recursive updating equations for the parameters of the posterior of Θ_1 given the data $y_1, \ldots, y_t$, for $t = 1, \ldots, n$, is straightforward. However, no *simple* recursive updating formulae seem to exist for the parameters of the predictive distribution (4.2), due to the complexity of the model (4.1) and to the fact that the order in which data from the polling stations arrive is unknown a priori and, hence, the matrix W used for prediction varies with n in a form which depends on the identity of the new data.

We agree with Pericchi that weak robustness, while being an interesting theoretical extension to the usual hierarchical normal model, may not be enough for detecting gross errors. As we prove in the paper, weak robustness of the posterior mean —which is linear in the observations— is obtained under the error specification given by (3.6), independently of $F(\Lambda)$.

To obtain strong robustness of the estimators, exchangeabilty should be abandoned in favour of independence. Thus, the first equation in model (4.1), should be replaced by

$$\boldsymbol{x}_i = \boldsymbol{a}_i' \Theta_1 + \boldsymbol{u}_i, \qquad i = 1, \ldots, n,$$

where the $\boldsymbol{a}_i'$'s are the rows of matrix $\boldsymbol{A}_1$, and the error matrix $\boldsymbol{U}_i' = (\boldsymbol{u}_1', \ldots, \boldsymbol{u}_n')$ is such that the error vectors $\boldsymbol{u}_i$ are independent and identically distributed as scale mixtures of multivariate normals, i.e., $\boldsymbol{u}_i \sim \int N(0, k_1 \Lambda)\, dF(\Lambda)$.

Unfortunately, under these conditions, no closed form for the posterior is possible, except for the trivial case where $F(\cdot)$ is degenerate at some matrix, say, Σ. In fact, the posterior distribution of Θ_1 given the data is a very complex infinite mixture of matrix-normal distributions. Thus, in order to derive useful robust estimators, we have to resort to approximate methods. One possibility, which has been explored by Rojano (1991) in the context of dynamic linear models, is to update the parameters of the MHLM sequentially, considering one observation at a time, as pointed out above, thus obtaining a simple infinite mixture of matrix-normals, and then to approximate this mixture by a matrix-normal distribution, and proceed sequentially.

Professor Dawid refers again to the fact that the method described is not outlier-robust. Pragmatically, we protected ourselves from extreme outliers by screening out from the forecasting mechanism any values which were more than three standard deviations off under the appropriate predictive distribution, conditional on the information currently variable. Actually, we are developing a sequential robust updating procedure based on an approximate Kalman filter scheme adapted to the hierarchical model, that both detects and accomodates outliers on line.

We are grateful to Professor Fienberg for his detailed description of previous work on election forecasting. We should like however to make a couple of points on his final remarks.

(i) Predicting the winner in a two party race is *far* easier that predicting a parliamentary *seat distribution* among *several* parties.

(ii) In our experience, exit surveys show too much uncontrolled bias to be useful, at least if you have to forecast a seat distribution.

ADDITIONAL REFERENCES IN THE DISCUSSION

Berger, J. O. (1990). Robust Bayesian analysis: sensitivity to the prior. *J. Statist. Planning and Inference* **25**, 303–328.

Brown, P. J. and Payne, C. (1975). Election night forecasting. *J. Roy. Statist. Soc. A* **138**, 463–498.

Brown, P. J. and Payne, C. (1984). Forecasting the 1983 British General Election. *The Statistician* **33**, 217–228.

Dawid, A. P. (1973). Posterior expectations for large observations. *Biometrika* **60**, 664–667.

Dawid, A. P. (1977). Spherical matrix distributions and a multivariate model. *J. Roy. Statist. Soc. B* **39**, 254–261.

Dawid, A. P. (1981). Some matrix-variate distribution theory: notational considerations and a Bayesian application. *Biometrika* **68**, 265–274.

Dawid, A. P. (1988). The infinite regress and its conjugate analysis. *Bayesian Statistics 3* (J. M. Bernardo, M. H. DeGroot, D. V. Lindley and A. F. M. Smith, eds.), Oxford: University Press, 95–110, (with discussion).

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

Link, R. F. (1989). Election night on television. *Statistics: A Guide to the Unknown* (J. M. Tanur *et al.* eds.), Pacific Grove, CA: Wadsworth & Brooks, 104–112.

Moreno, E. and Pericchi, L. R. (1990). An ϵ-contaminated hierarchical model. *Tech. Rep.* Universidad de Granada, Spain.

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

O'Hagan, A. (1988) Modelling with heavy tails. *Bayesian Statistics 3* (J. M. Bernardo, M. H. DeGroot, D. V. Lindley and A. F. M. Smith, eds.), Oxford: University Press, 345–359, (with discussion).

O'Hagan, A. (1990). Outliers and credence for location parameter inference. *J. Amer. Statist. Assoc.* **85**, 172–176.

Rojano, J. C. (1991). *Métodos Bayesianos Aproximados para Mixturas de Distribuciones*. Ph.D. Thesis, University of Málaga, Spain.

Walley, P. (1990). *Statistical Reasoning with Imprecise Probabilities*. London: Chapman and Hall.

West, M. (1981). Robust sequential approximate Bayesian estimation. *J. Roy. Statist. Soc. B* **43**, 157–166.

West, M. (1984). Outlier models and prior distributions in Bayesian linear regressions. *J. Roy. Statist. Soc. B* **46**, 431–439.