*Technical Report* **06/93**, (August 30, 1993).
*Presidencia de la Generalidad. Caballeros 9, 46001 - Valencia, Spain.*
*Tel. (34)(6) 386.6138, Fax (34)(6) 386.3626, e-mail: bernardo@mac.uv.es*

# Optimizing Prediction with Hierarchical Models: Bayesian Clustering

JOSÉ M. BERNARDO
*Universidad de Valencia, Spain*
*Presidencia de la Generalidad Valenciana, Spain*

SUMMARY

A frequent statistical problem is that of predicting a set of quantities given the values of some covariates, and the information provided by a training sample. These prediction problems are often structured with hierarchical models that make use of the similarities existing within classes of the population. Hierarchical models are typically based on a 'natural' definition of the clustering which defines the hierarchy, which is context dependent. However, there is no assurance that this 'natural' clustering is optimal in any sense for the stated prediction purposes. In this paper we explore the this issue by treating the choice of the clustering which defines the hierarchy as a formal decision problem. Actually, the methodology described may be seen as describing a large class of new clustering algorithms. The application which motivated this research is briefly described. The argument lies entirely within the Bayesian framework.

*Keywords:* BAYESIAN PREDICTION; HIERARCHICAL MODELLING; ELECTION FORECASTING; LOGARITHMIC DIVERGENCE; PROPER SCORING RULES; SIMULATED ANNEALING.

## 1. INTRODUCTION

Dennis Lindley taught me that interesting problems often come from interesting applications. Furthermore, he has always championed the use of Bayesian analysis in practice, specially when this has social implications. Thus, when I was asked to prepare a paper for a book in his honour, I thought it would be specially appropriate to describe some research which originated on a socially interesting area, –politics–, and may be used to broaden the applications of one of the methodologies he pioneered, –hierarchical linear models–.

## 2. THE PREDICTION PROBLEM

Let $\Omega$ be a set of $N$ elements, let $\boldsymbol{y}$ be a, possibly multivariate, *quantity of interest* which is defined for each of those elements, and suppose that we are interested in some, possibly multivariate, function

$$\boldsymbol{t} = \boldsymbol{t}(\boldsymbol{y}_1, \ldots, \boldsymbol{y}_N)$$

José M. Bernardo is Professor of Statistics at the University of Valencia, and Adviser for Decision Analysis to the President of the State of Valencia. This paper will appear in *Aspects of Uncertainty, a Tribute to D. V. Lindley* (P. R. Freeman and A. F. M. Smith, eds.) New York: Wiley, 1994.

of the values of these vectors over $\Omega$. Suppose, furthermore, that a vector $\boldsymbol{x}$ of covariates is also defined, that its values $\{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N\}$ are known for all the elements is $\Omega$, and that a random *training sample*

$$\boldsymbol{z}_n = \{(\boldsymbol{x}_i, \boldsymbol{y}_i), i = 1, \ldots, n\},$$

which consists of $n$ pairs of vectors $(\boldsymbol{x}, \boldsymbol{y})$, has been obtained. From a Bayesian viewpoint, we are interested in the predictive distribution

$$p(\boldsymbol{t} \,|\, \boldsymbol{z}_n, \boldsymbol{x}_{n+1}, \ldots, \boldsymbol{x}_N).$$

If the set $\Omega$ could be partitioned into a class $\boldsymbol{C} = \{C_i, i \in I\}$ of disjoint sets such that within each $C_i$ the relationship between $\boldsymbol{y}$ and $\boldsymbol{x}$ could easily be modelled, it would be natural to use a hierarchical model of the general form

$$\begin{gathered} p(\boldsymbol{y}_j \,|\, \boldsymbol{x}_j, \boldsymbol{\theta}_{i[j]}), \quad \forall j \in C_i \\ p(\boldsymbol{\theta} \,|\, \boldsymbol{\varphi}) \\ p(\boldsymbol{\varphi}) \end{gathered} \tag{1}$$

where $i[j]$ idenfifies the class $C_i$ to which the $j$-th element belongs, $p(\boldsymbol{y} \,|\, \boldsymbol{x}, \boldsymbol{\theta}_i)$ is a conditional probability density, totally specified by $\boldsymbol{\theta}_i$, which models the stochastic relationship between $\boldsymbol{y}$ and $\boldsymbol{x}$ within $C_i$, $p(\boldsymbol{\theta} \,|\, \boldsymbol{\varphi})$ describes the possible interrelation among the behaviour of the different classes, and $p(\boldsymbol{\varphi})$ specifies the prior information which is available about such interrelation.

Given a specific partition $\boldsymbol{C}$, the desired predictive density $p(\boldsymbol{t} \,|\, \boldsymbol{z}_n, \boldsymbol{x}_{n+1}, \ldots, \boldsymbol{x}_N)$ may be computed by:

(i) deriving the posterior distribution of the $\boldsymbol{\theta}_i$'s,

$$p(\boldsymbol{\theta} \,|\, \boldsymbol{z}_n, \boldsymbol{C}) \propto \int \prod_{j=1}^{n} p(\boldsymbol{y}_j \,|\, \boldsymbol{x}_j, \boldsymbol{\theta}_{i[j]}) p(\boldsymbol{\theta} \,|\, \boldsymbol{\varphi})\, p(\boldsymbol{\varphi})\, d\boldsymbol{\varphi}; \tag{2}$$

(ii) using this to obtain the conditional predictive distribution of the unknown $\boldsymbol{y}$'s,

$$p(\boldsymbol{y}_{n+1}, \ldots, \boldsymbol{y}_N \,|\, \boldsymbol{x}_{n+1}, \ldots, \boldsymbol{x}_N, \boldsymbol{z}_n, \boldsymbol{C}) = \int \prod_{j=n+1}^{N} p(\boldsymbol{y}_j \,|\, \boldsymbol{x}_j, \boldsymbol{\theta}_{i[j]}) p(\boldsymbol{\theta} \,|\, \boldsymbol{z}_n, \boldsymbol{C})\, d\boldsymbol{\theta}; \tag{3}$$

(iii) computing the desired predictive density

$$p(\boldsymbol{t} \,|\, \boldsymbol{z}_n, \boldsymbol{x}_{n+1}, \ldots, \boldsymbol{x}_N, \boldsymbol{C}) = f[\boldsymbol{y}_1, \ldots, \boldsymbol{y}_n,\, p(\boldsymbol{y}_{n+1}, \ldots, \boldsymbol{y}_N \,|\, \boldsymbol{x}_{n+1}, \ldots, \boldsymbol{x}_N, \boldsymbol{z}_n)] \tag{4}$$

of the function of interest $\boldsymbol{t}$ as a well-defined probability transformation $f$ of the joint predictive distribution of the unknown $\boldsymbol{y}$'s, given the appropriate covariate values $\{\boldsymbol{x}_{n+1}, \ldots, \boldsymbol{x}_N\}$ and the known $\boldsymbol{y}$ values $\{\boldsymbol{y}_1, \ldots, \boldsymbol{y}_n\}$.

This solution is obviously dependent on the particular choice of the partition $\boldsymbol{C}$. In this paper, we consider the choice of $\boldsymbol{C}$ as a formal decision problem, propose a solution, which actually provides a new class of (Bayesian) clustering algorithms, and succinctly describe the case study, –Mexican State elections–, which actually motivated this research.

## 3. THE DECISION PROBLEM

The choice of the partition $\boldsymbol{C}$ may be seen as a decision problem where the decision space is the class of the $2^N$ parts of $\Omega$ , and the relevant uncertain elements are the unknown value of the quantity of interest $\boldsymbol{t}$, and the actual values of the training sample $\boldsymbol{z}_n$. Hence, to complete the specification of the decision problem, we have to define a utility function $u[\boldsymbol{C}, (\boldsymbol{t}, \boldsymbol{z}_n)]$ which measures, for each pair $(\boldsymbol{t}, \boldsymbol{z}_n)$, the desirability of the particular partition $\boldsymbol{C}$ used to build a hierarchical model designed to provide inferences about the value of $\boldsymbol{t}$, given the information provided by $\boldsymbol{z}_n$.

Since, by assumption, we are only interested in predicting $\boldsymbol{t}$ given $\boldsymbol{z}_n$, the utility function should only depend on the *reported* predictive distribution for $\boldsymbol{t}$, say $q_t(.\,|\,\boldsymbol{z}_n, \boldsymbol{C})$, and the actual value of $\boldsymbol{t}$, i.e., should be of the form

$$u[\boldsymbol{C}, (\boldsymbol{t}, \boldsymbol{z}_n)] = s[q_t(.\,|\,\boldsymbol{z}_n, \boldsymbol{C}), \boldsymbol{t}]. \tag{5}$$

The function $s$ is known is the literature as a *score function*, and it is natural to assume that it should be *proper*, i.e., such that its expected value should be maximized if, and only if, the reported prediction *is* the predictive distribution $p_t(.\,|\,\boldsymbol{z}_n, \boldsymbol{x}_{n+1}, \ldots, \boldsymbol{x}_N, \boldsymbol{C})$. Furthermore, in a pure inferential situation, one may want the utility of the prediction to depend only on the probability density it attaches to the true value of $\boldsymbol{t}$. In this case (Bernardo, 1979), the score function must be of the form

$$s[q_t(.\,|\,\boldsymbol{z}_n, \boldsymbol{C}), \boldsymbol{t}] = A \log[p(\boldsymbol{t}\,|\,\boldsymbol{z}_n, \boldsymbol{x}_{n+1}, \ldots, \boldsymbol{x}_N, \boldsymbol{C})] + B, \quad A > 0. \tag{6}$$

Although, in our applications, we have always worked with this particular utility function, the algorithms we are about to describe may naturally be used with *any* utility function $u[\boldsymbol{C}, (\boldsymbol{t}, \boldsymbol{z}_n)]$.

For a given utility function $u$ and sample size $n$ the optimal choice of $\boldsymbol{C}$ is obviously that which maximizes the expected utility

$$u^*[\boldsymbol{C}\,|\,n] = \int\int u[\boldsymbol{C}, (\boldsymbol{t}, \boldsymbol{z}_n)]\, p(\boldsymbol{t}, \boldsymbol{z}_n)\, d\boldsymbol{t}\, d\boldsymbol{z}_n. \tag{7}$$

An analytic expression for $u^*[\boldsymbol{C}\,|\,n]$ is hardly ever attainable. However, it is not difficult to obtain a numerical approximation. Indeed, using Monte Carlo to approximate the outer integral, the value of $u^*[\boldsymbol{C}\,|\,m]$, for $m < n$ may be expressed as

$$u^*[\boldsymbol{C}\,|\,m] \approx \frac{1}{k}\sum_{l=1}^{k} \int u[\boldsymbol{C}, \boldsymbol{z}_{m(l)}, \boldsymbol{t})]\, p(\boldsymbol{t}\,|\,\boldsymbol{z}_{m(l)})\, d\boldsymbol{t}, \tag{8}$$

where $\boldsymbol{z}_{m(l)}$ is one of $k$ random subselections of size $m < n$ from $\boldsymbol{z}_n$. This, in turn, may be approximated by

$$u^*[\boldsymbol{C}\,|\,m] \approx \frac{1}{k}\sum_{l=1}^{k} \frac{1}{n_s}\sum_{j=1}^{n_s} u[\boldsymbol{C}, \boldsymbol{z}_{m(l)}, \boldsymbol{t}_j)], \tag{9}$$

where $t_j$ is one of $n_j$ simulations obtained, possibly by Gibbs sampling, from $p(\boldsymbol{t}\,|\,\boldsymbol{z}_{m(l)})$.

Equation (9) may be used to obtain an approximation to the expected utility of any given partition $C$. By construction, the optimal partition will agglomerate the elements of $\Omega$ in a form which is most efficient if one is to predict $\boldsymbol{t}$ given $\boldsymbol{z}_n$. However, the practical determination of the optimal $\boldsymbol{C}$ is far from trivial.

## 4. THE CLUSTERING ALGORITHM

In practical situations, where $N$ may be several thousands, an exhaustive search among all partitions $\boldsymbol{C}$ is obviously not feasible. However, the use of an agglomerative procedure to obtain a sensible initial solution, followed by an application of a simulated annealing search procedure, leads to practical solutions in a reasonable computing time.

In the aglomerative initial step, we start from the partition which consists of all the $N$ elements as classes with a single element, and proceed to a systematic agglomeration until the expected utility is not increased by the process. The following, is a pseudocode for this procedure.

```
$\boldsymbol{C} := \{\text{all elements in}\Omega\}$
**repeat**
    **for** $i$:=1 **to** $N$
        **for** $j$:=$i+1$ **to** $N$
            **begin**
                $C^* := C \ominus (i,j), \quad \{C_i \to C_i \cup C_j)\}$
                **if** $u^*[\boldsymbol{C}^*] > u^*[\boldsymbol{C}]$ **then** $\boldsymbol{C} := \boldsymbol{C}^*$
            **end**
**until** No_Change
```

The result of this algorithm may then be used as an initial solution for a simulated annealing procedure. Simulated annealing is an algorithm of random optimization which uses as a heuristic base the process of obtaining pure crystals (annealing), where the material is slowly cooled, giving time at each step for the atomic structure of the crystal to reach its lowest energy level at the current temperature. The method was described by Kirkpatrick, Gelatt and Vechhi (1983) and has seen some statistical applications, such as Lundy (1985) and Haines (1987). The algorithm is special in that, at each iteration, one may move with positive probability to solutions with lower values of the function to maximize, rather than directly jumping to the point with the highest value within the neighborhood, thus drastically reducing the chances of getting trapped in local maxima. The following, is a pseudocode for this procedure.

```
**get** Initial_Solution $\boldsymbol{C}_0$, Initial_Temperature $t_0$, Initial_Distance $d_0$;
$\boldsymbol{C} := \boldsymbol{C}_0; t := t_0; d := d_0;$
**repeat**
**while** (**not** $d$-Finished) **do**
    **begin**
        **while** (**not** $t$-Optimized) **do**
            **begin**
                Choose_Random($\boldsymbol{C}_i \mid d$)
                $\delta := u^*[\boldsymbol{C}_i] - u^*[\boldsymbol{C}_0]$
                **if** $(\delta \geq 0)$ **then** $\boldsymbol{C} := \boldsymbol{C}_i$
                **else if** (exp$\{-\delta/t\} \leq$ Random) **then** $\boldsymbol{C} := \boldsymbol{C}_i$
            **end**;
        $t := t/2$
    **end**;
Reduce_Distance($d$)
**until** $d < \varepsilon$
```

In the annealing procedure, the distance among two partitions is defined as the number of different classes it contains.

## 5. AN APPLICATION TO ELECTION FORECASTING

Consider a situation where, on election night, one is requested to produce a sequence of forecasts of the final result, based on incoming returns. 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, in each of a set of appropriately chosen strata, one may hope to produce accurate forecasts of the final results.

In Bernardo and Girón (1992), a hierarchical prediction model for this problem was developed, using electoral districts within counties as a 'natural' partition for a three stage hierarchy, and the results were successfully applied some weeks later to the Valencia State Elections. One may wonder, however, whether the geographical clustering used in the definition of the hierarchical model is optimal for the stated prediction purposes.

With the notation of this paper, a two-stage hierarchical model for this problem is defined by the set of equations

$$\begin{gathered}\boldsymbol{y}_{j[i]} = \boldsymbol{x}_{j[i]} + \boldsymbol{\theta}_i + \varepsilon_{0j[i]}, \quad j \in C_i, \qquad p(\varepsilon_{0j[i]} \,|\, \boldsymbol{\alpha}_0), \quad E[\varepsilon_{0j[i]}] = 0 \\ \boldsymbol{\theta}_i = \boldsymbol{\varphi} + \varepsilon_{1i}, \quad i \in I, \qquad p(\varepsilon_{1i} \,|\, \boldsymbol{\alpha}_1), \quad E[\varepsilon_{1i}] = 0 \\ \pi(\boldsymbol{\varphi}, \boldsymbol{\alpha}_0, \boldsymbol{\alpha}_1)\end{gathered} \tag{10}$$

where $\boldsymbol{y}_{j[i]}$ is the vector which describes the results on the new election in polling station $j$ which belongs to class $C_i$, $\boldsymbol{x}_{j[i]}$ contains the corresponding results in the past election, the error distributions of $\varepsilon_0 = (\varepsilon_{01[1]}, \ldots, )$ and $\varepsilon_1 = (\varepsilon_{11}, \ldots, )$, $p(\varepsilon_0 \,|\, \boldsymbol{\alpha}_0)$ and $p(\varepsilon_1 \,|\, \boldsymbol{\alpha}_1)$, have zero mean and are fully specified by the hiperparameters $\boldsymbol{\alpha}_0$ and $\boldsymbol{\alpha}_1$, and $\pi(\boldsymbol{\varphi}, \boldsymbol{\alpha}_0, \boldsymbol{\alpha}_1)$ is the reference distribution (Berger and Bernardo, 1992) which corresponds to this model.

The function of interest is the probability vector which describes the final results of the new election, i.e.,

$$\boldsymbol{t} = \sum_{i \in I} \sum_{j \in C_i} \beta_{j[i]} \boldsymbol{y}_{j[i]} \tag{11}$$

where $\beta_{j[i]}$ is the (known) proportion of the population which lives in the poling station $j$ of class $C_i$. The posterior distribution of $\boldsymbol{t}$ may be derived using the methods described above.

In this particular application, however, interest is essentially centered on a good estimate of $\boldsymbol{t}$. Given some results from the new election, i.e., the training sample $\boldsymbol{z}_n$, the value of $\boldsymbol{t}$ may be decomposed into its known and unknown parts, so that the expected value of the posterior distribution of $\boldsymbol{t}$ may be written as

$$E[\boldsymbol{t} \,|\, \boldsymbol{z}_n] = \sum_{i \in I} \sum_{j \in \text{ Obs}} \beta_{j[i]} \boldsymbol{y}_{j[i]} + \sum_{i \in I} \sum_{j \in \text{ NoObs}} \beta_{j[i]} E[\boldsymbol{y}_{j[i]} \,|\, \boldsymbol{z}_n], \tag{12}$$

where

$$E[\boldsymbol{y}_{j[i]} \,|\, \boldsymbol{z}_n] = \boldsymbol{x}_{j[i]} + \int \int E[\boldsymbol{\theta}_i \,|\, \boldsymbol{z}_n, \boldsymbol{\alpha}_0, \boldsymbol{\alpha}_1] \, p(\boldsymbol{\alpha}_0, \boldsymbol{\alpha}_1 \,|\, \boldsymbol{z}_n) \, d\boldsymbol{\alpha}_0 \, d\boldsymbol{\alpha}_1. \tag{13}$$

The conditional expectation within the double integral may be analytically found under different sets of conditions. In their seminal paper on hierarchical models, Lindley and Smith (1972) already provided the relevant expressions under normality, when $\boldsymbol{y}$ is univariate. Bernardo and Girón (1992) generalize this to multivariate models with error distributions which may be expressed as scales mixtures of normals; this includes heavy tailed error distributions such as the matrix-variate Student $t$'s. If an analytical expression for the conditional expectation $E[\boldsymbol{\theta}_i \,|\, \boldsymbol{z}_n, \boldsymbol{\alpha}_0, \boldsymbol{\alpha}_1]$ may be found, then an approximation to $E[\boldsymbol{y}_{j[i]} \,|\, \boldsymbol{z}_n]$ may be obtained by using Gibbs sampling to approximate the expectation integral.

In particular, when the error structure may be assumed to have the simple form

$$D^2[\boldsymbol{\varepsilon}_0 \,|\, h_0, \Sigma] = \frac{1}{h_0}(\boldsymbol{I} \otimes \Sigma), \qquad D^2[\boldsymbol{\varepsilon}_1 \,|\, h_1, \Sigma]] = \frac{1}{h_1}(\boldsymbol{I} \otimes \Sigma), \tag{14}$$

where the $\boldsymbol{I}$'s are identity matrices of appropriate dimensions and $\otimes$ denotes the Kronecker product of matrices, and when the error distribution is expressable as a scale mixture of normals, then the conditional reference reference distribution $\pi(\boldsymbol{\varphi}, \,|\, h_0, h_1, \Sigma)$ is uniform and the first moments of the conditional posterior distribution of the $\boldsymbol{\theta}_i$'s are given by

$$E[\boldsymbol{\theta}_i \,|\, \boldsymbol{z}_n, h_0, h_1, \Sigma] = \frac{n_i h_0 \boldsymbol{r}_{.i} + h_1 \boldsymbol{r}_{..}}{n_i h_0 + h_1} \tag{15}$$

$$D^2[\boldsymbol{\theta}_i \,|\, \boldsymbol{z}_n, h_0, h_1, \Sigma] = \frac{1}{n_i h_0 + h_1}\Sigma, \tag{16}$$

where $n_i$ is the number of polling stations the sample which belong to class $\boldsymbol{C}_i$,

$$\boldsymbol{r}_{.i} = \frac{1}{n_i} \sum_{j \in \boldsymbol{C}_i} \left( \boldsymbol{y}_{j[i]} - \boldsymbol{x}_{j[i]} \right), \quad i \in I \tag{17}$$

are the average sample swings within class $\boldsymbol{C}_i$, and

$$\boldsymbol{r}_{..} = \frac{1}{n} \sum_{j=1}^{n} \boldsymbol{y}_j - \boldsymbol{x}_j = \overline{\boldsymbol{r}}_{.i} \tag{18}$$

is the overall average swing.

Since (14) are the rather natural assumptions of exchangeability within classes, and exchangeability among classes, and (15) remains valid for rather general error distributions, (12), (13), and Gibbs integration over (15) provide together a practical mechanism to implement the model described.

## 6. A CASE STUDY: STATE ELECTIONS IN MEXICO

On February 1993, I was invited by the Mexican authorities to observe their Hidalgo State elections, in order to report on the feasibility of implementing in Mexico the methods developed in Valencia. Although I was not supposed to do any specific analysis of this election, I could not resist the temptation of trying out some methods.

I had taken with me the code of the algorithm I use to select a set of constituencies which, when viewed as a whole, have historically produced, for each election, a result close to the global result. The procedure, which is another application of simulated annealing, is described in Bernardo (1992).

Using the results of the 1989 election in Hidalgo (which were the only available ones), I used that algorithm to select a set of 70 polling stations whose joint behaviour had been similar to that of the State as a whole, and suggested that the local authorities should send agents to those polling stations to report on the phone the corresponding returns as soon as they were counted. A number of practical problems reduced to 58 the total number of results which were available about two hours after the polling stations closed.

In the mean time, I was busy setting up a very simple forecasting model –with no hierarchies included–, programmed in Pascal in a hurry on a resident Macintosh, to forecast the final results based on those early returns. This was in fact the particular case which corresponds to the model described in Section 4, if the partition $C$ is taken to have a single class, namely the whole $\Omega$.

About 24 hours later, just before the farewell dinner, the provisional official results came in. Table 1, Line 1, contains the official results, in percentage of valid votes of PAN (right wing), PRI (government party), PRD (left wing) and other parties. As it is apparent from Table 1, Line 2, my forecasts were not very good; the mean absolute error (displayed as the loss column in the table, was 3.28. Naturally, as soon as I was back in Valencia, I adapted the hierarchical software which I have been using here. The results (Table 1, Line 3) were certainly far better, but did not quite met the standards I was used to in Spain.

**State of Hidalgo, February 21st, 1993**

| | PAN | PRI | PRD | Others | Loss |
|---|---|---|---|---|---|
| Oficial Results | 8.30 | 80.56 | 5.56 | 5.56 | |
| No hierarchies | 5.5 | 76.8 | 9.3 | 8.4 | 3.28 |
| Districts as clusters | 6.4 | 80.6 | 7.7 | 5.3 | 1.09 |
| Optimal clustering | 8.23 | 80.32 | 6.18 | 5.27 | 0.31 |

**Table 1**. *Comparative methodological analysis.*

On closer inspection, I discovered that the variances within the districts used as clusters in the hierarchical model were far higher than the corresponding variances in Spain. This prompted an investigation on the possible ways to reduce such variances and, naturally, this lead to the general procedures described in this paper.

We used repeated random subselection of size 58 from the last election results in Hidalgo in order to obtain, –using the algorithms described in Section 3–, the 1989 optimal partition of the polling stations. In practice, we made the exangeability assumptions described by (14), assumed Cauchy error distributions, and chose a logarithmic scoring rule. We then used this partition to predict the 1993 election, using the two-stage hirearchical model described in Section 4 and the 58 available polling station results. The results are shown in Table 1, Line 4; it is obvious from them that the research effort did indeed have a practical effect in the Hidalgo data set.

## 7. DISCUSSION

Prediction with hierarchical models is a very wide field. Although very often, the clustering which defines the hierarchy has a natural definition, this is not necessarily optimal from a prediction point of view. If the main object of the model is prediction, it may be worth to explore alternative hierarchies, and the preceding methods provide a promising way to do this.

Moreover, there are other situations where the appropriate clustering is less than obvious. For instance, a model similar to that described here may be used to estimate the total personal income of a country, based on the covariates provided by the census and a training sample which consists of the personal incomes of a random sample of the population and their associated census covariates. The clustering which would be provided by the methods described here may have indeed an intrinsic sociological interest, which goes beyond the stated prediction problem.

Finally, the whole system may be seen as a proposal of a large class of well-defined clustering algorithms, where –as one would expect in any Bayesian solution–, the objectives of the problem are precisely defined. These could be compared with the rather *ad hoc* standard clustering algorithms as explorative data analysis methods used to improve our understanding of complex multivariate data sets.

## REFERENCES

Berger, J. O. and Bernardo, J. M. (1992). On the development of the reference prior method. *Bayesian Statistics 4* (J. M. Bernardo, J. O. Berger, A. P. Dawid and A. F. M. Smith, eds.), Oxford: University Press, 35–60 (with discussion).

Bernardo, J. M. (1979). Expected information as expected utility. *Ann. Statist.* **7**, 686–690.

Bernardo, J. M. (1992). Simulated annealing in Bayesian decision theory. *Computational Statistics* **1** (Y. Dodge and J. Whittaker, eds.) Heidelberg: Physica-Verlag, pp.547–552.

Bernardo, J. M. and Girón, F. J. (1992). Robust sequential prediction form non-random samples: the election night forecasting case. *Bayesian Statistics 4* (J. M. Bernardo, J. O. Berger, A. P. Dawid and A. F. M. Smith, eds.), Oxford: University Press, 61–77, (with discussion).

Kirkpatrick, S., Gelatt, C. D. and Vecchi, M. P. (1983). Optimization by simulated annealing. *Science* **220**, 671–680.

Haines, L. M. (1987). The application of the annealing algorithm to the construction of exact optimal designs for linear regression models. *Technometrics* **29**, 439–447.

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).

Lundy, M. (1985). Applications of the annealing algorithm to combinatorial problems in statistics. *Biometrika* **72**, 191–198.