Working Paper 98-61
Statistics and Econometrics Series 29
Septiembre 1998

Departamento de Estadística y Econometría
Universidad Carlos III de Madrid
Calle Madrid, 126
28903 Getafe (Spain)
Fax (34)91 624-9849

HETEROGENEITY AND MODEL UNCERTAINTY IN BAYESIAN REGRESSION MODELS

Justel, A. and Peña, D.*

Abstract

Data heterogeneity appears when the sample comes from at least two different populations. We analyze three types of situations. The first and simplest case corresponds to the situation in which the majority of the data comes form a central model and a few isolated observations comes from a contaminating distribution. Then the data from the contaminating distribution are called outliers and they have been studied in depth in the statistical literature. The second case corresponds to the situation in which we still have a central model but the heterogeneous data may appears in clusters of outliers which mask each other. This is the multiple outlier problem which is much more difficult to handle and it has understood and analyzed in the last few years. The few Bayesian contributions to this problem are presented. The third case corresponds to the situation in which we do not have a central model but instead different groups of data have been generated by different models. When the data is multivariate normal, this problem has been analyzed by mixture models under the name of cluster analysis but a challenging area of research is to develop a general methodology to apply this multiple model approach to other statistical problems. Heterogeneity implies in general an increase in the uncertainty of predictions, and in this paper a procedure to measure this effect is proposed.

Key Words

cluster analysis; influential data; masking; mixture model; outliers; predictive distributions; robust estimation.

*Justel, Department of Mathematics, Universidad Autónoma de Madrid, e-mail: ana.justel@.uam.es; Peña, Department of Statistics and Econometrics, Universidad Carlos III de Madrid, e-mail: dpena@est-econ.uc3m.es. We acknowledge support for this work from DGES, A. Justel for the Grant PB97-0021 and D. Peña for the Grant PB96-0111.

1. INTRODUCTION

Since the beginning of data analysis it was found that real data is often contaminated by heterogeneous observations or outliers. Outliers have been found even in small set of data coming from controlled experiments (see for instance Stigler, 1973, 1986). It is well known that the presence of a few extreme outliers can distort completely the result of the statistical analysis and make the Bayesian inference very inefficient. In spite of the seminal paper by Box and Tiao (1968), the study of outliers has not attracted many interest in the Bayesian literature. For instance in the 1997 Current index of statistics (CIS) out of 1151 references leading with outliers only 67 (5.8 %) either use Bayesian methods or refer to them.

In the last ten years in which large data sets are becoming more common due to the increasing computer power available, it has been found that outliers appear often in clusters, and then the methods derived to deal with a few isolated outliers are unable to detect them. This problem is called masking and again it has been mainly studied from the frequentist approach. For instance, going again to the 1997 CIS, out of the 22 papers leading with masking only two use the Bayesian approach.

Today many data sets are huge and heterogeneous: the computer has made possible to take measurements of many variables with almost no cost at short intervals in an automatic way. For instance, we find data sets of thousands of variables and millions of observations in astronomy (see Fayyad et al., 1996 for a description of some of these huge data set), quality control (in many chemical processes data is recorded every second or ten seconds of many production variables), finance (the stock transactions are collected at each pulse), business (all the purchases made in some period of time by millions of credit car users), and so on. These huge data sets creates new problems for statistical analysis, because none of the usual textbook hypothesis are expected to be true. We expect clustering of outliers and masking, nonstationarity, dependent observations, selection bias and errors in variables, as well as other measurement problems (see Hand, 1998 for an excellent description of these problems). A consequence of this is that different models are supposed to hold in different regions of the parameter space and also at each point we have several different models which can generate the data. The Bayesian paradigm is a flexible tool in order to model this type of situations although it may require some adjustment in order to represent some of the complicated and messy data set which we will be dealing with in the next future.

An important consequence of heterogeneity is model uncertainty. If the observations in the sample can be generated by different models, this will increase the uncertainty of the forecast of a future observation. To be specific, suppose that we assume that future data can be generated by a set of models $M_1, ..., M_m$ with probabilities $w_1, ..., w_m$. Then the forecast of a new observation will be given by

$$p(y) = \sum w_i p(y/M_i)$$

and the variability in the mixture distribution $p(y)$ will be in general larger than the variability of a central single distribution $p(y/M)$, which is usually considered in standard statistical applications.

In this work we review the Bayesian contributions to deal with heterogeneity in the linear regression model. Bernardo and Smith (1994) and O'Hagan (1994) are general references and good introductions to this problem. The paper is organized as follows. In section 2 we review briefly the methods developed for dealing with isolated outliers in linear models; in section 3 we discuss masking in regression and in section 4 we introduce the general heterogeneity problem and its relationship to Bayesian clustering. Finally, in section 5 we comment on the implication of heterogeneity in increasing model uncertainty for forecasting, and a statistic to measure this effect is suggested.

2. SINGLE OUTLIERS AND INFLUENTIAL DATA

We consider the usual regression model

$$y_i = \boldsymbol{x}_i \boldsymbol{\beta} + u_i \qquad i = 1, \ldots, n, \tag{2.1}$$

where $\boldsymbol{y} = (y_1, \ldots, y_n)'$ is a vector of responses, $\boldsymbol{X} = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n)'$ is a full rank $n \times p$ matrix of independent variables, $\boldsymbol{\beta}$ is a p-vector of unknown parameters and $\boldsymbol{u}$ is a vector of non observable random perturbations.

The Bayesian methods for outlier and influential data identification can be classified into two groups: i) diagnostic methods; and ii) robust methods. These two approaches differ in the way they assume that the data have been generated. The diagnostic methods consider a central model and try to find observations which have a small probability of being generated by it. They do not establish the mechanism which generates outliers. The robust methods incorporates an alternative model which can generate aberrant observations; for instance in regression the usual assumed hypothesis of normality is changed to the assumption of a heavy tail error distribution.

2.1. Diagnostic methods

The diagnostic methods assume a central model for data generation and outliers are considered as observation with small probability of being generated by this central model. Therefore they are identified by looking at the predictive densities $p(y_i \mid \boldsymbol{y}_{(i)})$, where $\boldsymbol{y}_{(i)}$ means that the data point y_i is deleted from the sample and analyzed as a new observation. If y_i is a single outlier, the probability to predict y_i given the rest of the sample is very low. This procedure for outlier detection is known as *the ordinate of the conditional predictive density method*, and was introduced by Geisser (1980).

The conditional predictive density can be seen as the ratio of two predictive densities,

$$p(y_i \mid \boldsymbol{y}_{(i)}) = \frac{p(\boldsymbol{y})}{p(\boldsymbol{y}_{(i)})},$$

and involves the predictive distribution $p(\boldsymbol{y})$ that was suggested by Box (1980) as a general diagnostic tool for any statistical model. This idea has been also explored by Pettit and Smith (1985) and Pettit (1990).

The conditional predictive ordinate is connected with the classical studentized residual test for outlier detection. With non informative priors, Pettit (1990) shows that

$$p(y_i \mid \boldsymbol{y}_{(i)}) = c\, s_{(i)}^{-1}(1-h_i)^{1/2}\left(1+\frac{t_i^2}{n-p-1}\right)^{-\frac{n-p}{2}}, \tag{2.2}$$

where t_i is the studentized residual

$$t_i = \frac{y_i - \boldsymbol{x}_i'\hat{\boldsymbol{\beta}}}{s_{(i)}(1-h_i)^{1/2}}, \tag{2.3}$$

$\hat{\boldsymbol{\beta}} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{y}$ is the least square estimate, $s_{(i)}^2 = \sum(y_j - \boldsymbol{x}_j'\hat{\boldsymbol{\beta}}_{(i)})/(n-p-1)$ is the unbiased residual variance estimate when y_i is deleted, and h_i is the leverage of the observation, that is, the $i-th$ diagonal element of the matrix $\boldsymbol{H} = \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'$, given by $h_i = \boldsymbol{x}_i'(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{x}_i$. Then data with large studentized residual have a small conditional predictive ordinate (2.2) and will be detected as outliers. An advantage of the conditional predictive ordinate method is that observations with high leverage (h_i is bounded by 1) will have small conditional predictive ordinate (2.2), independently that they are outliers or not. This is deduced from the studentized residual expression

$$t_i = \frac{(1-h_i)^{1/2} e_{i(i)}}{s_i},$$

where $e_{i(i)} = y_i - \boldsymbol{x}_i'\hat{\boldsymbol{\beta}}_{(i)}$ is the least square residual after deleting y_i in the regression estimation. When h_i goes to 1, the studentized residual goes to zero, independently that the $i-th$ data is an outlier ($e_{i(i)}$ is large) or a good data ($e_{i(i)}$ is small). In this case, the $i-th$ data point is very far from the rest in the independent variables space, and it is call an *influential data*. Note that the Bayesian measure is able to detect both outliers and influential points whereas the studentized residual will be unable to detect high leverage outliers.

The Bayesian approach for the identification of influential points is to measure the change of a relevant distribution when the point under consideration is deleted. Johnson and Geisser (1983, 1985) and Geisser (1985) proposed the Kullback-Leibler divergence (Kullback and Leibler, 1951) to measure the distance between the predictive distribution when deleting one data, $p(\boldsymbol{y}_{(i)})$, and the predictive with all the sample , $p(\boldsymbol{y})$, and proved that their measure is asymptotically equivalent to the sum of the Cook's statistic (Cook, 1977) and a convex function of the studentized residuals. The relationship between Cook's statistics and the studentized residual (2.3) is given by the formula

$$D_i = \frac{(n-p)}{p} \frac{t_i^2}{(n-p-1+t_i^2)} \frac{h_i}{1-h_i}.$$

Then it is easy to see, that the Cook's statistic will be large for influential outliers and small for good data. Another approach is proposed by Pettit and Smith (1985) and by Guttman and Peña (1988, 1993). These late authors proposed to compare by the Kullback-Leibler divergence the posterior parameter distributions, with and without the observations. They proved that changes in the posterior distribution of $\boldsymbol{\beta}$ are also function of the Cook's statistic, as it is derived from the expression of the Kullback-Leibler divergence

$$d(p(\boldsymbol{\beta} \mid \boldsymbol{y}_{(i)}), p(\boldsymbol{\beta} \mid \boldsymbol{y})) = \frac{pD_{(i)}^2}{2} + \frac{s_{(i)}^2}{2s^2}\left(p + \frac{h_{ii}}{1-h_{ii}}\right) + \frac{s^2}{2s_{(i)}^2}(p-h_{ii}) - p,$$

where $pD_{(i)} = (\hat{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}_{(i)})'\boldsymbol{X}_{(i)}'\boldsymbol{X}_{(i)}(\hat{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}_{(i)})/s_{(i)}^2$. These authors also proved that changes in the posterior distribution of σ^2 can be interpreted as an outlier measure depending on the studentized residuals t_i and the standardized residuals r_i. Finally, the changes in the joint posterior distribution of the two parameters are combinations of the influence measures on the posterior distribution of $\boldsymbol{\beta}$ and of the outlier measure. Girón, Martínez and Morcillo (1992) proposed to consider an observation as influential when it does not belong to the highest predictive density region $p(y \mid y_{(i)})$, and estimation influential with respect to a set of parameters when it does not belong to the highest posterior distribution region. They applied these ideas to regression models and showed the relationship of the proposed procedure with the Kalman Filter. Kass, Tierney and Kadane (1989) also suggested some influence measures based on deleting one observation. They use asymptotic methods to study the changes in some functions of interest. Using Decision Theory ideas Kempthorne (1986) and Carlin and Polson (1991) analyzed changes in the Bayes risk to identify influential points.

Note that all the proposals mentioned for the single outlier and influential data identification can be easily extended to the problem of group identification, but they require that the number and the position of the outliers are known.

2.2. Robust methods

The robust methods propose a model for the generation of all the data, including possible outliers. Then the estimation is carried out using all the sample, but the model used reduces the weight of the outliers in the estimation. There are two ways to obtain this effect. The first one is to assume a heavy tail distribution. The second to assume a mixture of distributions: a central one which generates the good points and an alternative one which is responsible for the outliers. In practice both are similar because the justification of using a heavy tail distribution is that the central model is contaminated but an unspecified distribution with heavy tails and this property is transmitted to the final mixture distribution.

Several heavy tail distribution have been suggested for regression problems. Box and Tiao (1973) proposed the power exponential family. West (1984) suggested to use heavy tail distributions that can be decomposed in a mixture of normal with different scales. It includes some well known families like the Student-t, the stables, the logistic and the double exponential. An advantage of this family is that it is possible to study the posterior parameter distributions by exploring some properties of the errors, which is not always the case with general heavy tail distributions. Fernández and Steel (1998) have proposed skewed student distributions which can also used for this purpose.

The second way is to accept the normality assumption for most of the data and assume an alternative distribution for the outliers. Then the lack of homogeneity in the sample is modeled with a mixture of distributions. In this model, it is assumed that the data may come from a central distribution with high probability, $(1-\alpha)$, and from a contaminated distribution with low probability, α. Two main outlier identification tools are used : 1) the posterior distribution for each point coming from the alternative distribution, given a particular generation mechanism for the rest of the sample; and 2) the Bayes factor to compare predictive distributions with different models. The cases more studied are those introduced by Tukey (1960) of mixtures of normals for the error distribution. The first one is the normal scale contamination model, Box and Tiao (1968) (SC model), where the data follow a model with error distributions

$$u_i \sim (1-\alpha)\ N(0,\sigma^2) + \alpha\ N(0,k^2\sigma^2) \qquad i=1,\ldots,n.$$

The second one is the normal level-shift model, by Guttman (1973) and Abraham and Box (1978) (LS model), where the error distributions are

$$u_i \sim (1-\alpha)\ N(0,\sigma^2) + \alpha\ N(\lambda,\sigma^2) \qquad i=1,\ldots,n.$$

The third is the additive model with m outliers, by Guttman, Dutter and Freeman (1978) (AD model). It supposes that there are m outliers in the sample (m is fixed by analyzing the model for

$m = 0, 1, \ldots)$ and the error distributions are $u_{i_j} \sim N(\lambda_j, \sigma^2)$, for $j = 1, \ldots, m$, and $u_{i_j} \sim N(0, \sigma^2)$, otherwise. These models can be combined and, for instance, Eddy (1980) has proposed a combination of the Box and Tiao (1968) and Abraham and Box (1978) models.

In general, these proposals assume that the regression model is written as

$$y_i \mid \boldsymbol{x}_i \sim (1-\alpha)\ f_1(y_i) + \alpha f_2(y_i),$$

where f_1 is the central model and f_2 is a contaminating one. If we assume that α is known, the ML estimation of this model can be carried out by the EM algorithm, as shown by Aitkin and Tunnicliffe-Wilson (1980). The EM algorithm can be seen as introducing a set of unobserved classification variables $\boldsymbol{\delta} = (\delta_1, \ldots, \delta_n)'$, defined as $\delta_i = 1$ when y_i is generated by the alternative distribution, and $\delta_i = 0$ otherwise. Then we substitute these variables for their expectations and estimate the parameters given the values of these variables.

In the Bayesian approach we want to compute the posterior distribution of the parameters given the data. This is also simplified if we introduce the classification variables and compute the posterior distribution $p(\boldsymbol{\beta}, \sigma^2, \boldsymbol{\delta} \mid \boldsymbol{y})$. The data y_i will be called an outlier when the marginal probability $p_i = p(\delta_i = 1 \mid \boldsymbol{y})$ is greater than 0.5. Thus, α is the prior probability that any observation is an outlier. Calling $A(r)$ to the event "r particular δ_i variables are equal to one and the remaining $n-r$ are zero", the posterior distribution of $\boldsymbol{\beta}$ is

$$p(\boldsymbol{\beta} \mid \boldsymbol{y}) = \sum_r p(A(r) \mid \boldsymbol{y})\, p(\boldsymbol{\beta} \mid A(r), \boldsymbol{y}),$$

where the weights $p(A(r) \mid \boldsymbol{y})$ are the posterior probabilities of all the possible configurations $A(r)$. With the usual reference priors for $\boldsymbol{\beta}$ and σ^2, $p(\boldsymbol{\beta}, \sigma^2) \propto \sigma^{-2}$, and assuming that k and α in the SC model or α and λ in the LS model, are known, these probabilities can be found in Freeman (1980). Eddy (1980) indicated that the mean of the distribution of $p(\boldsymbol{\beta} \mid A(r), \boldsymbol{y})$ in the three models can be seen as weighted least square estimates.

To identify the outliers we can use the weights $p(A(r) \mid \boldsymbol{y})$. In the particular case of a single outlier in the sample, the probabilities are

$$p(A_i(1) \mid \boldsymbol{y}) \propto \omega |\mathbf{X}'\mathbf{V}_{(1)}\mathbf{X}|^{-1/2} s_{(i)}^{-v}, \tag{2.4}$$

where $A_i(r)$ means that δ_i is equal to one, that is, y_i is one of the r contaminated data. The values of v and ω depend on the model: $v = n-p$ and $\omega = \alpha/k(1-\alpha)$ for the SC model, $v = n-p-1$ and $\omega = \alpha/(1-\alpha)$ for the LS model, and $v = n-p-1$ and $\omega = 1$ for the AD model. In the general case the probability of an observation to be an outliers is given by $p_i = \sum_r p(A_i(r) \mid \boldsymbol{y})$. This probability requires to compute the probabilities for all the 2^n possible combinations. We will see in the next subsection an alternative and feasible way to compute these probabilities by using MCMC methods.

The second method to identify outliers is to use Bayes factors. With the Bayes factor, and applying the Jeffreys rule (Jeffreys, 1961), it is possible to compare the predictive distribution for a model with only one outlier with the predictive distribution for an outlier free model. In this case the Bayes factor can be expressed as

$$F_{10}(i) = \frac{p(y \mid A_i(1))}{p(y \mid A(0))}.$$

Pettit (1992) extended the use of the Bayes factor to improper prior distributions by using the Spiegelhalter and Smith (1982) method of finding imaginary observations subsets of minimum size.

Peña and Guttman (1993) compared these approaches and showed that the posterior probability of a particular set of data to be outlier with the LS or AD model is inversely proportional to the ordinate of the predictive density, so that both approaches can be considered as equivalent.

2.3. Outlier detection with Gibbs Sampling

Bayesian analysis of outlier problems using the Gibbs sampler was initialized by Verdinelli and Wasserman (1991) for i.i.d. data. Their procedure was generalized by Justel and Peña (1996a) to the case of outliers in regression models. They considered the Box and Tiao (1968) model with the reference priors mentioned before, but assume that the contamination parameter α is unknown and use a $Beta(\gamma_1, \gamma_2)$ as prior distribution for this parameter. Gibbs sampling avoids the 2^n necessary computations to obtain the marginal posterior probabilities p_i.

The application of the Gibbs sampling (see Gelfand and Smith, 1990) is carried out by augmenting the parameter vector with a set of latent (unobserved) classification variables $(\delta_1, \ldots, \delta_n)$. Then the objective of the procedure is to obtain samples from the joint posterior $p(\boldsymbol{\beta}, \boldsymbol{\delta}, \sigma^2, \alpha \mid \boldsymbol{y})$. Starting from an arbitrary vector of initial values, the Gibbs sampler provides a sample of the posterior distribution for all the parameters in the model. It means that when the algorithm converges we will have a sample to be used for the computation of an estimate of $p(\delta_i = 1 \mid \boldsymbol{y})$, for $i = 1, \ldots, n$. The basic requirement for the Gibbs sampler is to be able to draw samples from all the conditional parameter distributions, conditional to the sample and to the other parameters. Justel and Peña (1996a) computed all the necessary conditionals and showed that generation from these distributions is very easy by using random number generators, as the ones described in Devroye (1986) or Ripley (1987).

The full conditional distributions are:

i) The conditional distribution of the vector $\boldsymbol{\beta}$ is $N_p\left(\tilde{\boldsymbol{\beta}}, \sigma^2(\mathbf{X}'\boldsymbol{V}^{-1}\mathbf{X})^{-1}\right)$, where $\tilde{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}'\mathbf{V}^{-1}\boldsymbol{y}$ and $\boldsymbol{V}$ is a diagonal matrix with elements $v_{ii} = 1 + \delta_i(k^2 - 1)$.

ii) The conditional distribution of σ^2 is $Inverted - Gamma$ $\left(n/2, \sum u_i^{*2}/2\right)$, where $u_i^* = (y_i - x_i'\boldsymbol{\beta})/(1 + \delta_i(k - 1))$.

iii) The conditional distribution of α is $Beta$ $(\gamma_1 + \sum \delta_i, \gamma_2 + n - \sum \delta_i)$.

iv) The conditional probability of $\delta_i = 1$ is

$$P(\delta_i = 1 \mid y, \beta, \sigma^2, \alpha) = \left(1 + \left(\frac{1-\alpha}{\alpha}\right) F_{10}(i)\right)^{-1}, \qquad (2.5)$$

where $F_{10} = k \cdot \exp\left(-u_i^2/2\phi^{-1}\sigma^2\right)$ is the Bayes factor and $\phi = 1 - k^{-2}$.

Note that the conditional probability that observation ith is an outlier depends only on the standardized residual u_i^2/σ^2. If the residual is small, $F_{10}(i)$ will be large and the probability (2.5) will be small. The opposite occurs when u_i^2/σ^2 is large.

Although the Gibbs sampler allows for easy computations of the marginal probabilities for each data to be an outlier, Justel and Peña (1996a) showed in several examples that Gibbs sampling fails for outlier detection in some data sets with multiple outliers. This case will be discussed in the next section.

3. MULTIPLE OUTLIERS

The formulas (2.2) and (2.4) can be easily used for single outlier detection, as well as generalized for checking the presence of a particular group of outliers (i.e., see Peña and Guttman, 1993). However, the most relevant problem is when the number and the position of the outliers are unknown, as it is the usual case with real data. In this case, two ideas may be considered: (1) to detect multiple outliers one by one, using single outlier detection procedures; and (2) to identify multiple outliers by computing all the probabilities for the possible outlier groups.

These two possibilities present serious problems in some particular, but not unusual, situations. In one hand, the deleting one by one observation procedures with multiple outliers can be subject to masking. Masking occurs when one outlier observation is not detected because of the presence of other outliers. Also, one good point can be wrongly identified as outlier due to the effect of the outliers, and this is called the swamping problem. The masking appears when there are several very similar outliers, which are also high leverage data. In this case, the studentized residuals tend to be small when they are not all deleted at the same time. Moreover, when the size of the outlier group is large the leverages of these data tend to be small, although they are very far away. Then the conditional predictive ordinate is large and outliers are not identified. Peña and Yohai (1995) proved this fact in the limit case of a group I of n_{I} outliers, $(y_a, \boldsymbol{x}_a')$, where $h_a = \boldsymbol{x}_a'(\boldsymbol{X}_{(\mathrm{I})}'\boldsymbol{X}_{(\mathrm{I})})^1\boldsymbol{x}_a'$. Then the residuals are

$$e_a = \frac{y_a - \boldsymbol{x}_a'\hat{\boldsymbol{\beta}}_{(\mathrm{I})}}{1 + n_{\mathrm{I}}\, h_a}.$$

If h_a is large, the residuals are small and they do not change if only one data is deleted (n_{I} is substituted by $n_{\mathrm{I}} - 1$). The leverages for all the data in the group I are $h_a/(1 + n_{\mathrm{I}}h_a)$, that tend to be small when n_{I} increases.

On the other hand, the generalization of (2.4) for a particular group of outliers may avoid the masking, but they involve the extensive computations of the 2^n posterior probabilities which correspond to all the possible configurations for the generation of the data.

Some proposals have been suggested to solve the masking problem from a Classical point of view, as the LMS of Rousseeuw (1984), or the methods of Rousseeuw and Zomeren (1990), Hadi and Simonoff (1993), Atkinson (1994), and Peña and Yohai (1995, 1998). However, the masking problem has received very few attention in the Bayesian literature. We only have found the works of Peña and Tiao (1992) and Justel and Peña (1996b).

3.1. Bayesian robustness curves

Peña and Tiao (1992) proposed a method based on stratified sampling to reduce the heavy computations on the multiple outlier detection problem. They suggested two new diagnostic tools: the Bayesian robustness curves BROC and SEBROC. Using the SC model, these curves compare the model with h outliers (M_h) with the outlier free model (M_0). The BROC is defined as the ratio of the posterior probabilities of model M_h and M_0, for different values of the number of outliers h, that is

$$P_{h0} = \frac{P(M_h \mid \boldsymbol{y})}{P(M_0 \mid \boldsymbol{y})} = \binom{n}{h}\left(\frac{\alpha}{1-\alpha}\right)^h F_{h,0},$$

where the Bayes factor is

$$F_{h,0} = \binom{n}{h}^{-1} k^{-h} \sum_r \frac{|\boldsymbol{X}'\boldsymbol{X}|^{1/2}}{|\boldsymbol{X}'\boldsymbol{X} - \phi\boldsymbol{X}_{(r)}'\boldsymbol{X}_{(r)}|^{1/2}} \left(\frac{s^2}{s_{(r)}^2}\right)^{(n-p)/2}. \tag{3.1}$$

The sum in (3.1) is over the $\binom{n}{h}$ possible configurations of h outliers and $n - h$ good data, and $s_{(r)}^2$ is a residual sum of squares given in Box and Tiao (1968). The BROC curve provides information about the number of outliers, however it is not able to identify masked outliers. The alternative in these cases is to use the Sequential Bayesian Robustness Curve (SEBROC) that is, for each h, the ratio

$$S_{h,h-1} = \frac{P_{h,0}}{P_{h-1,0}}.$$

The key point of the proposal in Peña and Tiao (1992) is to use stratified sampling ideas to reduce the $\binom{n}{h}$ computations for $P_{h,0}$, or the $\binom{n}{h-1}+\binom{n}{h}$ for $S_{h,h-1}$, without loss of efficiency. The method consists on:

1. If the $i-th$ observation is an outlier, all the elements
$$d(i,j)=\frac{p(A_{i,j}(2)\mid y)}{p(A(0)\mid y)}-\frac{p(A_i(1)\mid y)}{p(A(0)\mid y)}\frac{p(A_j(1)\mid y)}{p(A(0)\mid y)}$$
will take high values. Divide the sample in two parts, one of size n_1 that holds the potential outliers and another of size $n-n_1$ that holds the possible good data.
2. Using that
$$\binom{n}{h}=\sum_{r=0}^{h}\binom{n_1}{r}\binom{n-n_1}{h-r},$$
compute the $\binom{n}{h}$ combinations in which h of the n data are deleted by computing all the combinations on the group of size n_1, but only a small sample on the group of size $n-n_1$. For instance, if $n_1=10$, $n_2=20$ and $h=3$, compute the $\binom{10}{3}$ combinations in which three data are deleted from the n_1, the $\binom{10}{2}$ combinations in which two data are deleted from the n_1 and one randomly selected from the n_2 , the $\binom{10}{1}$ combinations in which one data is deleted from the n_1 combined with a random sample of the $\binom{20}{2}$ combinations of good data, and one small random sample of the $\binom{20}{3}$ possibilities of deleting good data.

3.2. Adaptive Gibbs Sampling

The proposal of Justel and Peña (1996b) is based on an adaptive Gibbs sampling algorithm (AGSA). When the outliers are isolated the Gibbs sampler works very well, however in strong masking cases the algorithm fails and outliers may not be detected when the convergence seems to be reached. A key factor to explain the lack of convergence in these cases seems to be the effect of the leverage in the estimation of linear regression models. When high leverage outliers which cause masking are classified as good data in the initial vector $\boldsymbol{\delta}^{(0)}$, the probabilities that these points are identified as outliers depend on the initial residuals $u_i^{(0)}=y_i-\boldsymbol{x}_i'\boldsymbol{\beta}^{(0)}$, where $\boldsymbol{\beta}^{(0)}$ is the mean of the conditional distribution given $\boldsymbol{\delta}^{(0)}$. For large k, the residuals $u_i^{(0)}$ will be small if the leverages are high, and these decrease with the number of outliers. Therefore, for high leverage outliers the residuals $u_i^{(0)}$ will be close to zero and the probabilities (2.5) will also be close to zero. On the other hand, when the masked outliers are not classified as good data in the initial vector $\boldsymbol{\delta}^{(0)}$, the out-of-sample residuals $u_i^{(0)}$ will be large and the probability (2.5) will be close to one. Therefore, the set of outliers will be detected in the next iteration only when all of them are classified as such in the drawing from the conditional distribution (2.5).

The solution to this problem begins with the correct initial classification of the group of masked outliers. Justel and Peña (1996b) proposed to compute the posterior probabilities of each observation being an outlier with the AGSA. The idea is to use the Gibbs sampler to find an outlier free subset. Then to split the sample and adapt the initial conditions to incorporate this information about possible outliers. When running the Gibbs sampling with these initial conditions it converges very quick to the posterior distributions. The splitting mechanism is based on the eigenstructure of the δ's covariance matrix estimated with the Gibbs sampler output. This matrix exploits the dependency structure among the observations generated by masking. The eigenvectors associated to the non zero eigenvalues provide information about which data are outlier candidates. The result is an adaptive method divided in three stages:

i) Standard Gibbs sampler: The Gibbs sampling is initialized by classifying a few data as good observations. Then the algorithm is run until the outlier probability series are stable.

ii) Outlier free subset identification: The covariance matrix of the classification variables is estimated with the Gibbs output from the first stage. The outlier free subset contains the observations with non null coefficients on the eigenvectors associated to the non zero eigenvalues and the observations with high marginal probability.

iii) Estimation: The Gibbs sampling is initialized by classifying as good the data in the outlier free subset. Then the algorithm is run until the outlier probability series are stable and all the posterior distributions are estimated with the Gibbs sampling output.

The procedure can be used automatically and includes: (1) a criterion for initial conditions selection without any prior information; and (2) a method to be used for grouping data based on the covariance matrix. Its application to some of the most frequently used examples in multiple outlier detection shows that it is able to unmask outliers in samples where other methods fail.

4. THE GENERAL CASE

The general heterogeneity case corresponds to a situation in which each point can be generated by a different model. To be specific, suppose that we have a set of models $M_1, \ldots, M_m$ such that M_j implies that $F(y \mid x)$ is $N(\beta_j' x, \sigma_j^2)$, that is, the data come from different regression models with different regression parameters and error variances. Associated with each of these models are prior probabilities ω_j, where $\sum \omega_j = 1$. When we know which model generates each observation, and we assume the prior covariances between coefficients of different equations are zero, we have the seemingly unrelated regression of Zellner (1971). When the prior covariance matrix is not block diagonal, then we have the shrinkage estimates by Lindley and Smith (1972).

Model heterogeneity may seem to be related to the problem of model selection, where we have a set $(M_1, M_2, ..., M_m)$ of possible models and we want to select the one which is more compatible with the data. The problem has a straightforward solution achieved by computing the posterior probabilities

$$p(M_i \mid \boldsymbol{D}) = \frac{p(\boldsymbol{D} \mid M_i)p(M_i)}{\sum p(\boldsymbol{D} \mid M_i)p(M_i)}$$

where $\boldsymbol{D}$ is the sample data. The specification of $p(M_i)$ requires that we have a partition of the model space, that is the models must be incompatible, and in model selection this is not the case in general. This is obvious when some models are nested, as when selecting between a linear or a quadratic regression. In general, the alternative non nested models that we are considering have some degree of overlapping, because they have been chosen to explain the same data set. However, the problem of overlapping models does not appear in the heterogeneity case in which we do not intend to select a model, rather we assume that we have several models and the problem is to identify the observations generated from each model and to use this information for estimation and forecasting.

A particular case of model heterogeneity is the one in which the response $\boldsymbol{y}$ is a r-dimensional vector, we do not have explanatory variables in the model, and the distributions F are known or are known up to a parameter vector $\boldsymbol{\theta}$. This is the standard clustering problem. The application of mixture models to clustering has a long tradition. See Binder (1978), Titterington et al. (1985), McLachlan and Basford (1998), Bernardo and Girón (1988, 1989), Lavine and West (1992) and Bernardo (1994). In the standard application of cluster analysis the number of components in the mixture, m, is assumed known. Then the model can be estimated by MCMC by introducing latent (unobserved) variables δ_j $(1 \leq j \leq n)$ which indicate the label of the group from which observation j is drawn. Of course, a priori

$$p(\delta_j = i) = \omega_i, \qquad \text{for } i = 1, \ldots, m.$$

This model has been studied by Diebolt and Robert (1994) who proposed a data augmentation algorithm to carry out the estimation and proved that it converges geometrically. They also study the convergence of Gibbs sampling.

In practice the number of components in the mixture is unknown. Then we have four possible approaches. The first one estimates m by a Schwarz criterion (Raftery, 1996). The second one use a Kullblack-Leibler estimate (Mengersen and Robert, 1996). The third one (Nobile, 1994) assumes a prior distribution for m , evaluates the likelihood of the data under each mixture model $p(\boldsymbol{y} \mid m)$ and then uses Bayes Theorem to compute the posterior $p(m \mid \boldsymbol{y})$. Finally, the fourth and more direct approach is to assume that the value of m is unknown and so it is included as an additional parameter to be estimated: we have a problem of Bayesian analysis of mixtures with an unknown number of components. A problem recently analyzed by Richardson and Green (1997). These authors proposed a model in which the joint distribution of all the variables of interest is given by

$$p(m, \boldsymbol{\delta}, \boldsymbol{\omega}, \boldsymbol{\theta}, \boldsymbol{y}) = p(m)p(\boldsymbol{\omega} \mid m)p(\boldsymbol{\delta} \mid \boldsymbol{\omega}, m)p(\boldsymbol{\theta} \mid m)p(\boldsymbol{y} \mid \boldsymbol{\delta}, \boldsymbol{\theta}),$$

which is similar to the model considered by Binder (1978). Briefly, we have a hierarchical model in which first we specify the number of components, m, then the probability of each component, $\boldsymbol{\omega}$, then we decide how many observations we take from each component by specifying the values of the latent variables, $\boldsymbol{\delta}$, then we fix the values of the parameters, $\boldsymbol{\theta}$, given the model m and finally we set the values of the sample given the model from which they are generated.

The authors apply this model to univariate normal mixtures. The prior distribution for the number of components is assumed to be uniform between 1 and a given value $m_{\max}$. The prior probabilities for $\boldsymbol{\omega}$ and the parameters $\boldsymbol{\theta} = (\boldsymbol{\mu}, \boldsymbol{\sigma})$ are the usual ones: for $\boldsymbol{\omega}$ a Dirichlet distribution, for the mean a normal prior and an inverted gamma for the variance.

The estimation of this model is carried out by a reversible jump MCMC, (Green, 1995) which is a Metropolis-Hasting algorithm in which, in addition to the usual Gibbs sampling updating of the parameters $\boldsymbol{\delta}, \boldsymbol{\omega}, \boldsymbol{\theta}$, two further moves are introduced.

(1) Splitting one mixture component into two, or merging two mixture components into one;

(2) The birth or death of an empty component.

At each step a random choice is made between attempting to split or combine. This is done with equal probabilities unless we have just one group (then we always split) or we have reached the maximum number of groups (then we always combine). The combining is carried out by choosing at random two adjacent groups in terms of the current value of their means and merging the observations of both groups into a new group. Splitting is made by random selection of a group and splitting it into two, also at random. The decision between birth and death is also taken randomly with equal probabilities and either a new group is created by sampling the parameters from the prior distribution or it is deleted.

The application of this scheme to the regression case present several problems. First, as in the normal mean case there is an identification problem because the whole model is invariant to permutation of the level of the groups. In the univariate case this is solved by using an increasing order for the means, but in the regression case it is not obvious how to define a clear ordering for the vectors of regression parameters $\boldsymbol{\beta}_j$, $j = 1, \ldots, m$. Second, the splitting and merging of the groups use the natural adjacent idea in the univariate case but there is not a clear way to extend this approach to the vector case in regression. Third we wonder if the same type of problems of convergence for the Gibbs Sampling that we have found in the multiple outlier case can again appears here. In the univariate case the possibility of strong making for the leverage effect of some observations does not appear, but in the regression set up the algorithm may fail for the same reasons shown in Justel and Peña (1996a). Further research is needed to discover if the reversible jump MCMC algorithm can be used with success in regression problems.

5. HETEROGENEITY AND MODEL UNCERTAINTY

Model heterogeneity implies that if we want to forecast the value of a future observation y and: (1) we know that it can be generated by a set of models $M_1, ..., M_m$ with probabilities $w_1, ..., w_m$, (2) we do not know which one will be the correct model, then we have to use the marginal predictive density given by

$$p(y \mid \boldsymbol{D}) = \sum w_i p(y \mid M_i, \boldsymbol{D}).$$

where $\boldsymbol{D}$ stands for data. In standard statistical applications either we have a central model or a model is selected from the sample. Let us call M_0 to this central model and $p(y \mid M_0, \boldsymbol{D})$ to the predictive distribution derived from it. Using $p(y \mid M_0, \boldsymbol{D})$ instead of $p(y \mid \boldsymbol{D})$ will in general underestimate the uncertainty in the forecast. We define the increase in uncertainty due to model heterogeneity by the Kullblack-Leibler distance between the distributions $p(y \mid \boldsymbol{D})$ and $p(y \mid M_0, \boldsymbol{D})$

$$U = \int \ln \frac{p(y \mid \boldsymbol{D}, M_0)}{p(y \mid \boldsymbol{D})} p(y \mid \boldsymbol{D}, M_0) dy$$

This measure is positive if both distribution are different and will be equal to zero if they are equal. For instance, let us consider the simplest case of isolated outliers in the Box and Tiao (1968) regression model. Then if we want to forecast the value of a new response variable, y, given the values of the explanatory variables $\boldsymbol{x}$, the predictive distribution $p(y \mid \boldsymbol{D}, M_0)$ is a Student t distribution with mean m_0 and variance v_0. The distribution $p(y \mid \boldsymbol{D})$ will be a mixture of two Student t distributions with the same mean, m_0, variances v_0 and $v_1 = v_0 k^2$ and mixing proportions $(1 - \alpha)$, and α. In order to compute the KL distance we can approximate these distributions by normals with the same mean and variance to obtain

$$U = \frac{1}{2}(\log(v_2/v_0) + v_0/v_2 - 1),$$

where v_2 is the variance of the mixture distribution $p(y \mid \boldsymbol{D})$. Note that as both distributions have the same mean, the KL distance is just the average of the two measures of the relative change in the variances, $\log(v_2/v_0)$ and $(v_0 - v_2)/v_2$. Using that in this model $v_2 = v_0(1 + \alpha(k^2 - 1))$ we obtain that

$$U = \frac{1}{2} \log(1 + \alpha(k^2 - 1)) - \alpha(k^2 - 1)/2(1 + \alpha(k^2 - 1)),$$

and so the increase in uncertainty in the forecast is a monotonic increasing function in α and k^2. We see that the increase in uncertainty depends on the parameter $\lambda = \alpha(k^2 - 1)$. The first derivative of U with respect to λ is

$$\frac{dU}{d\lambda} = \frac{\lambda}{2(1+\lambda)^2}$$

which it is always positive and it is zero at $\lambda = 0$, indicating that a small model heterogeneity has no effect on the uncertainty of the prediction. The inflexion point of the $U(\lambda)$ function can be obtained from

$$\frac{d^2U}{d\lambda^2} = \frac{1-\lambda}{2(1+\lambda)^3}$$

and it is reached for $\lambda = 1$ which corresponds, for instance, to the case $\alpha = .05$ and $k = 6$. From this point, increasing k and/or α by a fixed amount will produce smaller increases in the uncertainty of the prediction.

In the general heterogeneity case, the mean of the distributions $p(y \mid \boldsymbol{D})$ and $p(y \mid M_0, \boldsymbol{D})$ will also be different and the KL distance will depends on the standardized mean difference as well as on the variance changes. An approximation to the KL measure can be computed in closed form by approximating the Student t distributions by normal distributions.

REFERENCES

Abraham, B. and Box, G.E.P. (1978). "Linear models and spurious observations". *Applied Statistics*, 27, 131–138.

Aitkin, M. and Tunnicliffe-Wilson, G. (1980). "Mixture models, outliers, and the EM algorithm". *Technometrics*, 22, 325–31.

Atkinson, A.C. (1994). "Fast very robust methods for the detection of multiple outliers". *Journal of the American Statistical Association*, 89, 1329–1339.

Bernardo, J.M. (1994). "Optimizing prediction with hierarchical models: Bayesian clustering". In *Aspects of uncertainty: A tribute to D.V. Lindley.* Ed. P.R. Freeman and A.F.M. Smith. John Wiley.

Bernardo, J.M. and Girón, F.J. (1988). "A Bayesian analysis of simple mixture problems". *Bayesian Statistics 3*, 67–88. Ed. J.M. Bernardo et al. Oxford University Press.

Bernardo, J.M. and Girón, F.J. (1989). "A Bayesian approach to cluster analysis". *Questiio*, 12, 97–112.

Bernardo, J.M. and Smith, A.F.M. (1994). *Bayesian Theory.* John Wiley.

Binder D. (1978). "Bayesian Clustering". *Biometrika*, 65, 31–38.

Box, G.E.P. (1980). "Sampling and Bayesian inference in scientific modelling and robustness" (with discussion). *Journal of the Royal Statistical Society, A*, 143, 383–430.

Box, G.E.P. and Tiao, C.G. (1968). "A Bayesian approach to some outlier problems". *Biometrika*, 55, 119–129.

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

Carlin, B.P. and Polson, N.G. (1991). "An expected utility approach to influence diagnostics". *Journal of the American Statistical Association,* 86, 1013–1021.

Cook, R.D. (1977). "Detection of influential observations in linear regression". *Technometrics*, 19, 15–18.

Current Index of Statistics. (1997). American Statistical Association.

Devroye, L. (1986). *Non-uniform Random Variate Generation.* Springer-Verlag.

Diebolt, J. and Robert, C. (1994). "Estimation of finite mixture distributions through Bayesian Sampling". *Journal of the Royal Statistical Society, B*, 56, 363–375.

Eddy, W.F. (1980). Discussion to the paper of P.R. Freeman. *Bayesian Statistics 1*, 370–373. Ed. J.M. Bernardo et al. Oxford University Press.

Fayyad, U., Piatestsky-Shapiro, G. and Smyth, P. (1996). "From data mining to knowledge discovery: an overview". In *Advances in Knowledge Discovery and Data Mining.* Ed. U. Fayyad et al. MIT Press.

Fernández, C. and Steel, M. (1998). "On Bayesian modeling of fat tails and skewness". *Journal of the American Statistical Association*, 93, 359–372.

Freeman, P.R. (1980). "On the number of outliers in data from a linear model". *Bayesian Statistics 1*, 349–365. Ed. J.M. Bernardo et al. Oxford University Press.

Geisser, S. (1980). Discussion to the paper of G.E.P. Box. *Journal of the Royal Statistical Society, A*, 143, 416–417.

Geisser, S. (1985). "On the predicting of observables: a selective update". *Bayesian Statistics 2*, 203–230. Ed. J.M. Bernardo et al. North Holland.

Gelfand, A.E. and Smith, A.F.M. (1990). "Sampling-based approaches to calculating marginal densities". *Journal of the American Statistical Association*, 85, 398–409.

Girón, F.J., Martínez, L. and Morcillo, C. (1992). "A Bayesian justification for the analysis of residuals and influence measures". *Bayesian Statistics 4*, 651–660. Ed. J.M. Bernardo et al. Oxford University Press.

Green, P. (1995). "Reversible jump Markov chain Monte Carlo computation and Bayesian model determination". *Biometrika*, 82, 711–732.

Guttman, I. (1973). "Care and handling of univariate or multivariate outliers in detecting spuriousity — A Bayesian approach". *Technometrics*, 15, 723–738.

Guttman, I., Dutter, R. and Freeman, P.R. (1978). "Care and handing of univariate outliers in the general linear model to detect spurosity — A Bayesian approach". *Technometrics*, 20, 187–193.

Guttman, I. and Peña, D. (1988). "Outliers and influence: evaluation by posteriors of parameters in the linear model". *Bayesian Statistics 3*, 631–640. Ed. J.M. Bernardo et al. Oxford University Press.

Guttman, I. and Peña, D. (1993). "A Bayesian look at diagnostic in the univariate linear model". *Statistica Sinica*, 3, 367–390.

Hadi, A.S. and Simonoff, J.S. (1993). "Procedures for the identification of multiple outliers in linear models". *Journal of the American Statistical Association*, 88, 1264–1272.

Hand, D.J. (1998). "Data mining: statistics and more?". *The American Statistician*, 52, 112–119.

Jeffreys, H. (1961). *Theory of Probability* (third edition). Oxford University Press.

Johnson, W. and Geisser, S. (1983). "A predictive view of the detection and characterization of influential observations in regression analysis". *Journal of the American Statistical Association*, 78, 137–144.

Johnson, W. and Geisser, S. (1985). "Estimative influence measures for the multivariate general model". *Journal of Statistical Planning and Inference*, 11, 33–56.

Justel, A. and Peña, D. (1996a). "Gibbs sampling will fail in outlier problems with strong masking". *Journal of Computational and Graphical Statistics*, 5, 176–189.

Justel, A. and Peña, D. (1996b). "Bayesian unmasking in linear models". CORE Discussion Paper 9619. Universitè Catholique de Louvain.

Kass, R.F., Tierney, L. and Kadane, J.B. (1989). "Approximate methods for assessing influence and sensitivity in Bayesian analysis". *Biometrika*, 76, 663–674.

Kempthorne, P.J. (1986). "Decision-theoretic measures of influence in regression". *Journal of the Royal Statistical Society, B*, 48, 370–378.

Kullback, S. and Leibler, R.A. (1951). "On information and sufficiency". *Annals of Mathematical Statistics*, 22, 79–86.

Lavine, M. and West, M. (1992). "A Bayesian method for classification and discrimination". *The Canadian Journal of Statistics*, 20, 451–461.

Lindley D.V. and Smith, A.F.M. (1972). "Bayes estimates for the linear model". *The Journal of Royal Statistical Society, B*, 34, 1–18.

McLachlan, G.J. and Basford, K.E. (1998). *Mixture Models: Inference and Applications to Clustering.* Marcel Dekker.

Mengersen, K. and Robert, C. (1996). "Testing for mixtures: a Bayesian entropy approach". In *Bayesian Statistics 5,* 255–276. Ed. J.O. Berger, J.M. Bernardo, A.P. Dawid, D.V. Lindley and A.F.M. Smith. Oxford University Press.

Nobile, A. (1994). Bayesian analysis of finite mixture distributions. *Ph.D. Thesis.* Carnegie Mellon University, Pittsburgh.

O'Hagan, A. (1994). *Kendall's Advanced Theory of Statistics.* Vol. 2B. Bayesian Inference. Edward Arnold.

Peña, D. and Guttman, I. (1993). "Comparing probabilistic methods for outlier detection in linear models". *Biometrika*, 80, 603–610.

Peña, D. and Tiao, G.C. (1992). "Bayesian robustness functions for linear models". *Bayesian Statistics 4*, 365–388. Ed. J.M. Bernardo et al. Oxford University Press.

Peña, D. and Yohai, V.J. (1995). The detection of influential subsets in linear regression by using an influence matrix. *Journal of the Royal Statistical Society, B*, 57, 145–156.

Peña, D. and Yohai, V.J. (1998). "A fast procedure for diagnostics and robust estimation in large data sets" . *Journal of the American Statistical Association*, forthcoming.

Pettit, L.I. (1990). "The conditional predictive ordinate for the normal distribution". *Journal of the Royal Statistical Society, B*, 52, 175–184.

Pettit, L.I. (1992). "Bayes factor for outliers models using the device of imaginary observations". *Journal of the American Statistical Association*, 87, 541–545.

Pettit, L.I. and Smith, A.F.M. (1985). "Outliers and influential observations in linear models". *Bayesian Statistics 2*, 473–494. Ed. J.M. Bernardo et al. Oxford University Press.

Raftery, A.E. (1996). "Hypothesis testing and model selection". In *Markov Chain Monte Carlo in Practice*, 163–188. Ed. W.R. Gilks, S. Richardson and D.J. Spiegelhalter. Chapman and Hall.

Richardson, S. and Green P.J. (1997). "On Bayesian analysis of mixtures with an unknown number of components". *Journal of the Royal Statistical Society, B*, 59, 731–792.

Ripley, B.D. (1987). *Stochastic Simulation.* John Wiley.

Rousseeuw, P.J. (1984). "Least median of squares regression". *Journal of the American Statistical Association*, 79, 871–880.

Rousseeuw, P.J. and van Zomeren, B.C. (1990). "Unmasking multivariate outliers and leverage points". *Journal of the American Statistical Association*, 85, 633–639.

Spiegelhalter, D.J. and Smith, A.F.M. (1982). "Bayes factors for linear and log-linear models with vague prior information". *Journal of the Royal Statistical Society, B*, 44, 377–387.

Stigler, S.M. (1973). "Simon Newcomb, Percy Daniel and the history of robust estimation 1885–1920". *Journal of the American Statistical Association*, 68, 872–879.

Stigler, S.M. (1986). *The History of Statistics.* Harvard University Press.

Titterington, D.M., Smith, A.F.M. and Makov, U.E. (1995). *Statistical Analysis of Finite Mixture Distributions.* John Wiley.

Tukey, J.W. (1960). "A survey of sampling from contaminated distributions". *Contributions to Probability and Statistics: Volume Dedicated to Harold Hotelling.* Oxford University Press.

Verdinelli, I. and Wasserman, L. (1991). "Bayesian analysis of outlier problems using the Gibbs sampler". *Statistics and Computing,* 1, 105–117.

West, M. (1984). "Outlier models and prior distribution in Bayesian linear models". *Journal of the Royal Statistical Society, B*, 46, 431–439.

Zellner, A. (1971). *An introduction to Bayesian Inference in Econometrics.* John Wiley.