BAYESIAN STATISTICS 7, pp. 000–000
J. M. Bernardo, M. J. Bayarri, J. O. Berger, A. P. Dawid,
D. Heckerman, A. F. M. Smith and M. West (Eds.)
© Oxford University Press, 2003

Identifying Mixtures of Regression Equations by the SAR procedure

DANIEL PEÑA, JULIO RODRÍGUEZ
Universidad Carlos III de Madrid, Spain
dpena@est-econ.uc3m.es puerta@est-econ.uc3m.es

GEORGE C. TIAO
University of Chicago, USA
gct@gsb.uchicago.edu

SUMMARY

A procedure for identifying data heterogeneity when fitting regression models is presented. The method is based on the SAR procedure, developed by Peña and Tiao (2002), and has three steps. First, the sample is cleaned for large outliers by using a discrepancy measure based on predictive ordinates. Second, observations are split into small homogeneous groups by using a link function derived from cross validation predictive distributions. Third, these small groups are then iteratively enlarged by incorporating observations homogeneous to those in the group. In this way the possible piece-wise regression equations are found. Examples are shown to illustrate the performance of the procedure for finding mixtures of regressions due to the presence of outliers or to switching regression models. A Monte Carlo study of the power of the proposed procedure is presented.

Keywords: OUTLIERS; PREDICTIVE DISTRIBUTION; STRUCTURAL CHANGE; SWITCHING REGRESSION.

1. INTRODUCTION

Data heterogeneity when a regression model is fitted implies that some kind of segmentations or clustering exits among the sample units. This problem has been studied under two main approaches: (1) outlier detection and/or robust regression estimation and (2) switching and/or structural change regression. From the Bayesian point of view these two situations can be formulated as mixture estimation problems with different types of mixture distributions. For the outlier problem Box and Tiao (1968) proposed a scale contaminated normal model for the noise distribution and develop a Bayesian estimation procedure for this model. Several other outlier detection and robust Bayesian estimation methods have been proposed based on mixtures and heavy tails distributions. The estimation of these mixture models can be carried out by Markov Chain Monte Carlo (MC^2) methods, but the standard Gibbs sampling algorithm presents serious convergence problems when there are groups of masked outliers (see Justel and Peña, 1996).

A different source of heterogeneity was introduced by Quandt (1958) who proposed a model in which the data is assumed to follow the regression equation $y_i = \boldsymbol{x}_i'\boldsymbol{\beta}_1 + u_i$ for $i = 1, ..., n_1$ and a different regression equation $y_i = \boldsymbol{x}_i'\boldsymbol{\beta}_2 + u_i$ for $i = n_1 + 1, ..., n$ where $\boldsymbol{\beta}_1 \neq \boldsymbol{\beta}_2$. The key problem in this model is to identify the change point between the two regimes and estimate the parameters (see for instance Schweder, 1976) and this problem has been extensively studied under the name of regression under a structural change. A comparison of some of the procedures developed to estimate the change point can be found in Andrews et al (1996). A generalization of this model is to assume that each observation can be generated with some unknown probability by one of the two regressions models $y_i = \boldsymbol{x}_i'\boldsymbol{\beta}_j + \sigma_j u_i$, where $j = 1, 2$ and u_j is $N(0, 1)$. This situation may correspond to the case in which a categorical variable is omitted in the model, like gender, month, or day of the week. If the categorical variable only has effect on the intercept we have the standard missing attribute case, whereas if the regression coefficients depend on the omitted categorical variable we have the switching regression problem. In the general case of G possible regimes, or groups of data, and assuming normality, the model is

$$y_i/\boldsymbol{x}_i \sim \sum_{g=1}^{G} \alpha_g N(\boldsymbol{x}_i'\boldsymbol{\beta}_g, \sigma_g^2), \tag{1}$$

where $\alpha_g \geq 0$ and $\sum_{g=1}^{G} \alpha_g = 1$. This model has been studied extensively both from the Bayesian and the Likelihood point of view. See Aitkin (2001) for some comparisons. When the number of groups is known and we have some reasonable initial estimate for the parameters in the different regimes the model can be estimated by MC^2 methods. However, when this information is not available the estimation of this model is a difficult problem. First, the components of the mixture have identification problems, Celeux et al. (2000). Second, when one of the groups has a much smaller variance than the others, and behaves like a set of high leverage outliers, a false convergence of the Gibbs sampling procedure may occur and some modifications are required to achieve convergence (see Justel and Peña, 2001). Third, more experience is required about the performance of the available procedures proposed to deal with the Bayesian estimation of these models, see for instance Richarson and Green (1997) and Stephens (2000), when the number of parameters and the number of groups are large and we do not have good initial estimates to start the algorithm. A recent analysis with Bayesian MC^2 methods can be found in Hurn, Justel and Roberts (2002).

In this paper we present a different approach to solve the heterogeneity problem. The approach is exploratory and is based on the SAR procedure developed by Peña and Tiao (2002). It is designed to deal with any type of heterogeneity, including outliers and switching regressions, and it can be used either as a starting point in a MC^2 estimation algorithm or as a general model building procedure.

The paper is organized as follows. In section 2 the SAR procedure applied to the regression model is briefly summarized. A key ingredient of the procedure is the link function, and the properties of this function in the regression case are studied in section 3. Section 4 presents simple simulated examples to illustrate the algorithm and discusses its implementation. Section 5 shows that the proposed procedure can provide the same solution as sophisticated MC^2 algorithms; that it is able to succeed in cases in which MC^2 algorithms are found to fail; and that it can be applied in large data mining problems. Section 6 includes a simulation study about the size and power of the procedure. Section 7 contains some brief concluding remarks.

2. THE SAR PROCEDURE IN REGRESSION

Given a sample of n points $(y_i, \boldsymbol{x}_i')$, $i = 1, ..., n$, where $\boldsymbol{x}_i$ is a $p \times 1$ vector, let $\boldsymbol{Y} = (y_1, ..., y_n)'$ be the vector of responses and $\boldsymbol{X} = (\boldsymbol{x}_1, ..., \boldsymbol{x}_n)'$ the $n \times p$ full rank matrix of explanatory variables. The SAR (split and recombine) procedure is an iterative method to identify possible clusters in a sample. It can be applied for all sorts of data heterogeneity and includes three basic steps. First, large isolated outliers are identified and deleted from the sample. Second, the remaining data is split iteratively into homogeneous groups of some minimum fixed size, called basic sets. Third, each basic set is allowed to grow by recombining points one by one into it, as long as they are found homogeneous with the observations already in the group. The complement of the enlarged group will now be treated as a new sample for a repeated application of the three-step split and recombine procedure. This process will lead to a set of possible different grouping of the sample, called possible data configurations (PDCs). These solutions can be explored by a model selection criteria (see Peña, Rodríguez and Tiao, 2002) or they can be used as starting point for a MC^2 algorithm for mixtures.

The SAR procedure is based on three statistics and a link function. The first statistic is based on the predictive ordinate, $p(y_i|\boldsymbol{Y}_{(i)})$, that has been used by several authors for outlier detection, see Box (1980), Geisser (1980, 1987), Pettit and Smith (1985), Pettit (1990), Peña and Tiao (1992), Peña and Guttman (1993), and others. The notation $\boldsymbol{Y}_{(i)}$ indicates that observation y_i has been deleted from the data set $\boldsymbol{Y}$. The justification of this measure is easy to see by introducing a dummy variable δ_i that is equal to 1 when observation y_i has been generated by the same model that has generated the data in $\boldsymbol{Y}_{(i)}$ and 0 otherwise. Then, we have that

$$P(\delta_i = 0|y_i, \boldsymbol{Y}_{(i)}) = 1 - kP(y_i|\delta_i = 1, \boldsymbol{Y}_{(i)})P(\delta_i = 1) \tag{2}$$

where k is a constant. Calling $p(y_i|\boldsymbol{Y}_{(i)})$ the predictive density of y_i when it is generated by the same model that generates $\boldsymbol{Y}_{(i)}$ we see that the smaller the predictive ordinate the larger the probability that the observation has been generated by a model different from the one that generates the rest of the data. The first statistic we define is the standardized predictive ordinate $c_0(i)$, given by

$$c_0(i) = -2\ln\left\{\frac{p(y_i|\boldsymbol{Y}_{(i)})}{p(\widehat{y}_{i(i)}|\boldsymbol{Y}_{(i)})}\right\}, \tag{3}$$

where $p(y_i|\boldsymbol{Y}_{(i)})$ is the predictive distribution of y_i given data $\boldsymbol{Y}_{(i)}$, and $\widehat{y}_{i(i)} = E(y_i|\boldsymbol{Y}_{(i)})$ is the expected value of this distribution. This standardized predictive ordinate is computed first to test for outliers, and second to test if a new point can be incorporated to a group formed by data $\boldsymbol{Y}_{(i)}$. The second statistic we define, $c_1(i)$, is given by

$$c_1(i) = \max_{y_j}(c_1(i|j)) = \max_{y_j}\left[-2\ln\left\{\frac{p(y_i|\boldsymbol{Y}_{(ij)})}{p(\widehat{y}_{i(ij)}|\boldsymbol{Y}_{(ij)})}\right\}\right] \tag{4}$$

where $\boldsymbol{Y}_{(ij)}$ represents a set of data in which observations y_i and y_j are deleted, $p(y_i|\boldsymbol{Y}_{(ij)})$ is the predictive distribution of y_i given data $\boldsymbol{Y}_{(ij)}$, and $\widehat{y}_{i(ij)} = E(y_i|\boldsymbol{Y}_{(ij)})$. The maximum is taken with respect all the observations in $\boldsymbol{Y}_{(i)}$ and indicates the minimum standardized value of the predictive distribution of y_i that can be obtained by deleting a point from $\boldsymbol{Y}_{(i)}$. The third statistic we use is the maximum standardized change in the predictive distribution obtained by deleting an observation from the set $\boldsymbol{Y}_{(i)}$, and is given by

$$d_1(i) = c_1(i) - c_0(i)$$

and this statistic is used jointly with $c_0(i)$ to identify masked outliers from the sample. Finally, the binary relationship implied by $c_1(i)$ provides a link function, $l(i)$, between observation y_i and the observation y_j which produces the maximum in (4): observation y_j is the one that, when deleted, will make y_i the most discrepant with the rest of the data. This link function is defined by

$$l(i) = j \;\text{ if }\; y_j = \arg\max_{y_k}(c_1(i|k))$$

that is, the point $(y_i, \boldsymbol{x}_i')$ is linked to $(y_j, \boldsymbol{x}_j')$ if deleting this last point from the sample produces the maximum discrepancy between the observation y_i and the remaining $n-2$ observations. We will call the point, $(y_j, \boldsymbol{x}_j)$, the discriminator of $(y_i, \boldsymbol{x}_i)$. Note that (1) each sample point must have a discriminator; (2) a discriminator is, in some metric implied by the model, an extreme point in the set $\boldsymbol{Y}_{(i)}$; (3) the relationship defined by the link function is neither symmetric nor transitive.

In the linear regression case calling $\widehat{\boldsymbol{\beta}} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{Y}$ and $s^2 = \boldsymbol{e}'\boldsymbol{e}/(n-p)$, where $\boldsymbol{e} = (\boldsymbol{I} - \boldsymbol{H})\boldsymbol{Y}$ and $\boldsymbol{H} = \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'$ is the idempotent projection matrix, and assuming the standard non informative prior for the vector of parameters $\boldsymbol{\theta} = (\boldsymbol{\beta}, \sigma)$, $p(\boldsymbol{\theta}) \propto \sigma^{-1}$, then, given $\boldsymbol{Y}$, the predictive distribution for a future observation y_f from the linear model (see e. g. Box and Tiao, 1973), is the standard univariate t distribution with ν degrees of freedom

$$p(y_f|\boldsymbol{Y}) = \frac{1}{B\left(\frac{1}{2}, \frac{\nu}{2}\right)\sqrt{\nu}} s^{-1}(1+h_f)^{-1/2}\left(1 + \frac{t_f^2}{\nu}\right)^{-\frac{(\nu+1)}{2}}$$

where $t_f^2 = (y_f - \widehat{y}_f)^2/s^2(1+h_f)$, $\widehat{y}_f = \boldsymbol{x}_f'\widehat{\boldsymbol{\beta}}$, $h_f = \boldsymbol{x}_f'(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{x}_f$, and $\boldsymbol{x}_f$ is the future explanatory vector. For $c_0(i)$, we let $y_f = y_i$, $\boldsymbol{x}_f = \boldsymbol{x}_i$, and we have that

$$\ell n[p(y_i|\boldsymbol{Y}_{(i)})] = k - \frac{1}{2}\ln[s_{(i)}^2(1 + h_{i(i)})] - \left(\frac{\nu_0+1}{2}\right)\ln\left(1 + \frac{t_i^2}{\nu_0}\right)$$

where $\nu_0 = (n-p-1), h_{i(i)} = h_f, t_i = t_f$ in which $\widehat{y}_{i(i)} = \boldsymbol{x}_i'\widehat{\boldsymbol{\beta}}_{(i)}$ is the predictive mean of y_i, and $\widehat{\boldsymbol{\beta}}_{(i)}$ and $s_{(i)}^2$ are the estimates based on $\boldsymbol{Y}_{(i)}$ and $\boldsymbol{X}_{(i)}$ where $\boldsymbol{X}_{(i)}$ is obtained from $\boldsymbol{X}$ by deleting $\boldsymbol{x}_i$. Since $\ln[p(\hat{y}_{i(i)}|\boldsymbol{Y}_{(i)})] = k - (1/2)\ln\left[s_{(i)}^2(1 + h_{i(i)})\right]$, we can write

$$c_0(i) = (\nu_0 + 1)\ln\left(1 + \frac{t_i^2}{\nu_0}\right). \tag{5}$$

The statistic t_i is the studentized residual, or standardized predictive residual, that is normally used for testing an individual outlier in regression. Expression (5) shows that $c_0(i)$ is a monotonic transformation of t_i^2 and, for large n, these two measures will be equivalent.

In the same way, letting $y_f = y_i$, $\boldsymbol{x}_f = \boldsymbol{x}_i$, and deleting y_i and y_j from $\boldsymbol{Y}$ to become $\boldsymbol{Y}_{(ij)}$, we obtain $c_1(i)$ as

$$c_1(i) = (\nu_1 + 1)\max_{y_j}\left[\ln\left\{1 + \frac{t_{i(ij)}^2}{\nu_1}\right\}\right] \tag{6}$$

Here $\nu_1 = (n-p-2)$ and $t_{i(ij)}^2$ is given by $t_{i(ij)}^2 = (y_i - \widehat{y}_{i(ij)})^2/s_{(ij)}^2(1 + h_{i(ij)})$, where $h_{i(ij)} = h_f$, $\widehat{y}_{i(ij)} = \boldsymbol{x}_i'\widehat{\boldsymbol{\beta}}_{(ij)}$ is the predictive mean of y_i, and $\widehat{\boldsymbol{\beta}}_{(ij)}$ and $s_{(ij)}^2$ are the estimates based on $\boldsymbol{Y}_{(ij)}$ and $\boldsymbol{X}_{(ij)}$. For large n, $c_1(i)$ will be equivalent to $\max_{y_j} t_{i(ij)}^2$.

The SAR procedure consists of the following three basic steps (see Peña and Tiao, 2002, for further justifications, tables of the statistics and further details):

1. Check the sample for outliers by computing $c_0 = \max c_0(i)$ and $d_1 = \max d_1(i)$ and test for outliers according to the null distribution of first c_0 and then d_1. The critical values of these statistics have been obtained by Monte Carlo. Delete the outliers and repeat the checking until either no more outliers are found or a set of size$= p + h$, is obtained, where h is a parameter to control the degrees of freedom of the fitted model that is usually taken, for n small or moderate, equal to $ln(n-p)$.

2. Split the remaining sample by putting in the same group all the observations which share the same discriminator. Points that are either outliers or discriminators are considered as isolated observations and they are not incorporated into the groups. Thus: (1) If $l(i) = j$ and $l(k) = j$ then (i, k) are put in the same group. (2) discriminators and treated as isolated points. (3) observations in subgroups of sizes smaller than $p + h$ are treated as isolated points. Once the new groups are found, we go back to 1 and through repeated application of steps 1 and 2, we split the original sample S_0 into m basic sets $(B_1, ..., B_m)$ and isolated points consisting of outliers, discriminators and observations from undersized groups in the splitting process.

3. The recombining process starts by selecting a basic set, B_i, fitting the model to the data in this basic set, and checking the observations in the complementary set $\overline{B}_i$ for homogeneity with respect to the basic set one at a time. The checking is done with the statistic $c_0(i)$ conditional on data in B_i. If the minimum of $c_0(i)$ for all the points in $\overline{B}_i$ is below the 99th percentile of the distribution of this statistic, then the corresponding observation is incorporated into the basic set. Then the model is refitted to the newly enlarged basic set including a new observation and the process of checking the remaining observations is repeated. When no more observations can be incorporated into the initial set, the enlarged set is considered as a first level final homogeneous group F_i. This enlarging process is repeated for each of the m basic sets. The result will be a set of first level final groups $(F_1, ..., F_m)$. Now, select a first level final group F_i and conditional on this group the complementary set $\overline{F}_i$ is analyzed as a new sample and steps 1-3 are applied. Repeating this process, the algorithm is continued until we obtain all the non-redundant PDCs originated from the set of basic sets. Note that each PDC is a partition of the sample into disjoint subsets.

The key tool of the procedure is the link function, which defines the discriminator for each sample point. In the next section we analyze its properties in the regression case.

3. THE PROPERTIES OF THE LINK FUNCTION

Let $\boldsymbol{e} = \boldsymbol{Y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}} = (e_1, ..., e_n)'$ be the residuals of the regression fitted to the sample $(\boldsymbol{Y}, \boldsymbol{X})$ and h_{ij} the $ij - th$ element of the projection matrix $\boldsymbol{H} = \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'$. It is shown in the appendix that we can write

$$t^2_{i(ij)} = \frac{e_i^2 + (1 - h_{jj})^{-1}[h_{ij}^2 e_j^2 (1 - h_{jj})^{-1} + 2e_i h_{ij} e_j]}{a_{ii} - c(1 - h_{jj})^{-1}[dh_{ij}^2 + b_{ii} e_j^2 + 2e_i h_{ij} e_j]}$$

where $c = (n-p-2)^{-1}$ and $d = (n-p)s^2$ are constants and $b_{ii} = (1-h_{ii})$, $a_{ii} = c[db_{ii} - e_i^2]$. This expression shows that for a given observation, $(y_i, \boldsymbol{x}_i')$, the statistic $t^2_{i(ij)}$ is maximized if the discriminator point $(y_j, \boldsymbol{x}_j')$ is chosen so that: (1) the product $e_i h_{ij} e_j$ is positive, (2) the measure of leverage of the discriminator $(1 - h_{jj})^{-1}$ is as large as possible, (3) the predictive residual $e_j(1-h_{jj})^{-1}$ is, in absolute value, as large as possible. Dividing both terms by $s^2(1-h_{ii})$ and calling $z_{ij} = h_{ij}(1 - h_{jj})^{-1/2}(1 - h_{ii})^{-1/2}$ and $r_j = e_j[s^2(1 - h_{jj})]^{-1/2}$ to the standardized residuals, we have that

$$t^2_{i(ij)} = \frac{r_i^2 + [z_{ij}^2 r_j^2 + 2r_i r_j z_{ij}]}{c[(n-p) - r_i^2 - (n-p)z_{ij}^2 - r_j^2 - 2r_i r_j z_{ij}]}$$

and it is straightforward to check that, assuming $r_i r_j z_{ij}$ is positive, the partial derivatives of this function with respect to the new variables r_j and z_{ij}, that measure outlyingness and leverage, are positive. Thus, the discriminator should be a point with large standardized residual, r_j, and large leverage, as measured by z_{ij}. It is interesting to note that these measures appear in a natural way in the analysis of outlying observations in linear models using standard Bayesian procedures, see Peña and Guttman, 1993. Note that we can write, in the space of the explanatory variables, $h_{ij} = h_{ii}^{1/2} h_{jj}^{1/2} \cos\theta$, and, therefore, the cross leverage h_{ij} will be large for a fixed value of h_{ii} if the leverage of the discriminator, h_{jj}, is large and $\cos\theta$ close to one.

In order to understand better the behavior of the link function suppose that we want to find the discriminator of a point y_i with non negative residual, $e_i \geq 0$. As the $t^2_{i(ij)}$ statistic is invariant to translations, suppose for simplification that all the variables are measured on deviations to the mean so that the regression equation has no constant term. We consider the following three possible cases:

(1) Suppose $x_i = 0$ so that $h_{ii} = 0 = z_{ij}$. Then $t^2_{i(ij)} = r_i^2/c[(n-p) - r_i^2 - r_j^2]$, and the discriminator will be the point with largest value of the standardized residual r_j^2.

(2) Suppose $x_i > 0$ but $e_i = 0 = r_i$ and the point analyzed lies on the regression line. Then $t^2_{i(ij)} = [(n-p) - (n-p)z_{ij}^2 - r_j^2]^{-1} z_{ij}^2 r_j^2/c$ and the point will be linked to the value than maximizes this expression. Note again that the discriminator will be a high leverage point with large standardized residual.

(3) Suppose $x_i \gg 0$ and $e_i \gg 0$. Then the discriminator must: (a) have $r_i r_j z_{ij} > 0$ which implies that either the residual is also positive and $\theta < \pi/2$ or the residual is negative and $\pi/2 < \theta < \pi$, and (b) have large values of the leverage and the standardized residual.

For instance, consider the simple regression through the origin, with $\sum x_i = \sum y_i = 0$ and $\sum x_i^2 = 1$ so that $h_{ij} = x_i x_j$, if an observation has an x value larger than the mean and the residual is positive the discriminator will be either: (1) a point with positive residual and also $x_j > 0$, or (2) a point with negative residual and $x_j < 0$.

Figure 1 shows a sample of simulated data from an homogeneous regression and the links obtained in this sample. The discriminators are indicated by: (1) an arrow and (2) the vertex of the cone of lines that go to all the points that are linked to the discriminator. Only eight points are discriminators. If we now put together points with the same discriminator and consider as a group a set of at least 3 points we will obtain four groups. Note that, as indicated in the previous rules, the discriminator is a point with either a large residual or a large leverage or a combination of both effects. For instance, points 8 and 24 have high leverage, whereas points 7 and 20 have a relatively large residual. Note that points are linked to discriminators located according to the previous rules.

We now illustrate the behavior of the link function in the four heterogeneous cases of simple regression considered in Figure 2: (a) concentrated contamination; (b) switching regression; (c) mixture of two regressions with omitted categorical variable and (d) structural change. Figure 2 shows the data indicated by numbers and the results of applying the link function to the four data sets. Only a few points in each case are discriminators and they are indicated as before by: (1) an arrow and (2) the vertex of the cone of lines that go to all the points that are linked to the discriminator. In case (a) only five points are discriminators and putting together those observations with the same discriminator we obtain five basic sets. The link function puts all the outliers in the same basic set. Note that points are linked to discriminators that are extreme according to the previous rules. In case (b) the sample is split into five groups induced by the five discriminators. A discriminator is always an extreme point with large leverage and large residual, and points are linked to the discriminator with the same residual sign and at the same side of the mean of the explanatory variable. Again, the sample is split into homogeneous groups. A similar situation occurs in case (c) in which seven discriminators exists. Finally, in case (d) four discriminators are found and, as before, points are linked to extreme observations (high leverage and or high residual) located in the same region of the space as determined by the fitted regression line. In the four examples the sample is split into homogeneous groups and this

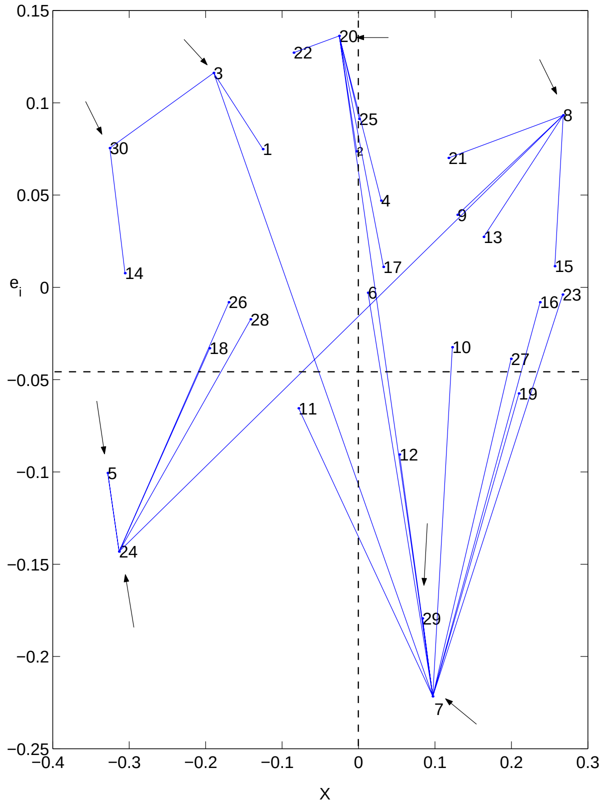


Figure 1. *Results of applying the link function in the homogeneous regression case*

is the pattern found in other SAR applications (see Peña and Tiao, 2002, and Peña, Rodríguez and Tiao, 2002).

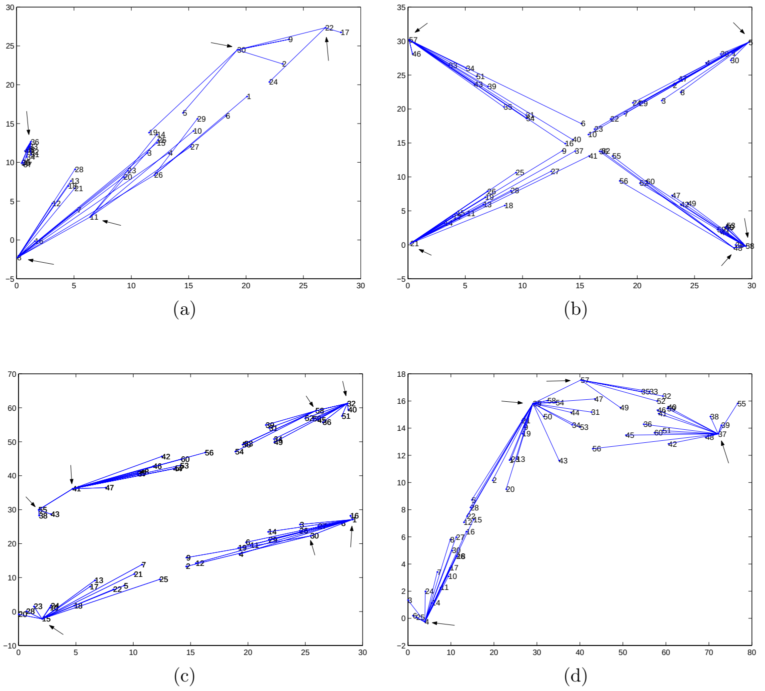
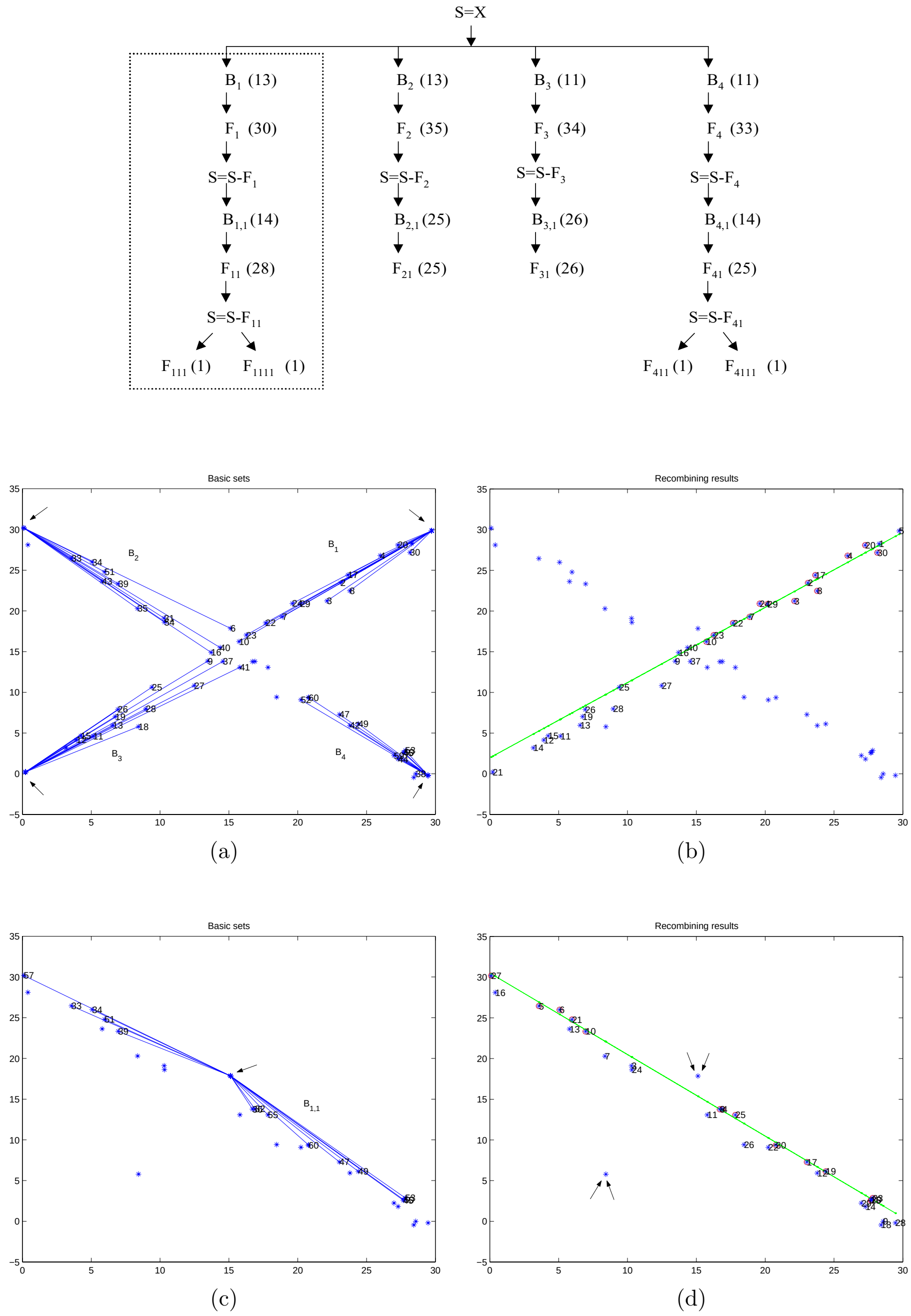


Figure 2. *Behavior of the link function in simple linear regression in the four cases: (a) group of masked outliers; (b) switching regression; (c) mixture of two regressions with omitted categorical variable; (d) regression under an structural change.*

4. THE PROPOSED ALGORITHM

The algorithm proposed for finding heterogeneity in regression models is the SAR algorithm briefly presented in Section 2. Here we will just illustrate it by using the data from Figure 2(b). Some of the intermediate results when applying the algorithm are indicated in Figure 3. First, four basic stets are found (see Figure 3(a)); second, the enlarging of the first basic set incorporates all the observations from the first regime (see Figure 3(b)); third, conditioning on this result the complementary part is split and a basic set is found (see Figure 3(c)); fourth the set is enlarged and the rest of the available observations are incorporated except for two points indicated by arrows in Figure 3(d) that are identified as outliers.



(a)

(b)

(c)

(d)

Figure 3. *Flow chart of the SAR procedure in the switching regression. (a) Basic sets; (b) enlarging of the first basic set; (c) Splitting of the complementary part, one basic set is found; (d) enlarging of the basic set and the two outliers detected, that are indicated by arrows.*

The two regression lines and the two outliers found in Figures 3(a)-(d) is one of the possible data configurations, (PDCs) provided by the algorithm, because when starting from a different basic set some of the points may be allocated in a slight different way. Figure 3, top, shows a flow chart of the four paths generated by each of the four basic sets. The path illustrated in Figures 3(a)-(d) is the first one, and it is presented in a dotted rectangle. Starting from the first basic set of 13 observations (B_1) the set is enlarged to include 30 observations (F_1). Then the complementary part of F_1 is analyzed and a basic set $B_{1,1}$ of 14 observations is found. This basic set grows to include 28 observations (F_{11}). The complementary part of this groups contains two observations that are considered as outliers (groups F_{111} and F_{1111}, each of one observation). The four PDCs finally found are shown in Figure 4. These four PDCs only differ in the way in which the doubtful observations in the intersection of the two lines are allocated. In the first PDC these observations are all included in the regression with positive slope and no outliers are found. In the second PDC, they are included in the regression with negative slope. In the third and fourth PDCs two of the points are considered as outliers and the remaining doubtful points are allocated: (1) in the third PDC to the regression with positive slope and (2) in the fourth to the one with negative slope. The selection of the best PDC can be made by : (1) Using the BIC criterion; (2) Fitting two regression lines to each of the two clear groups and then computing the posterior probabilities of each observation in the intersection belonging to each of the two regression lines.

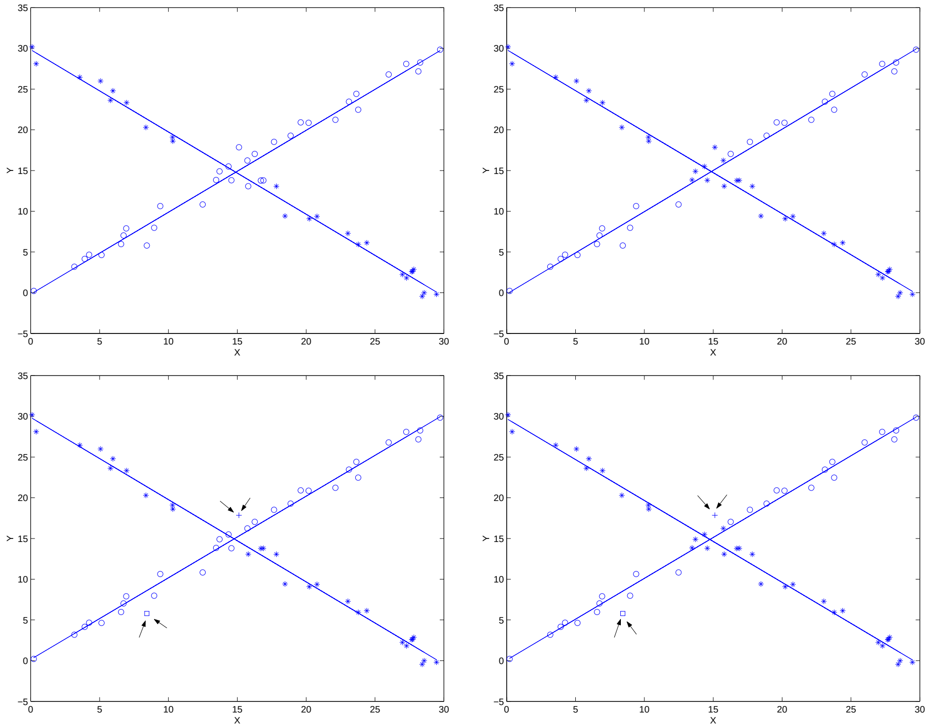
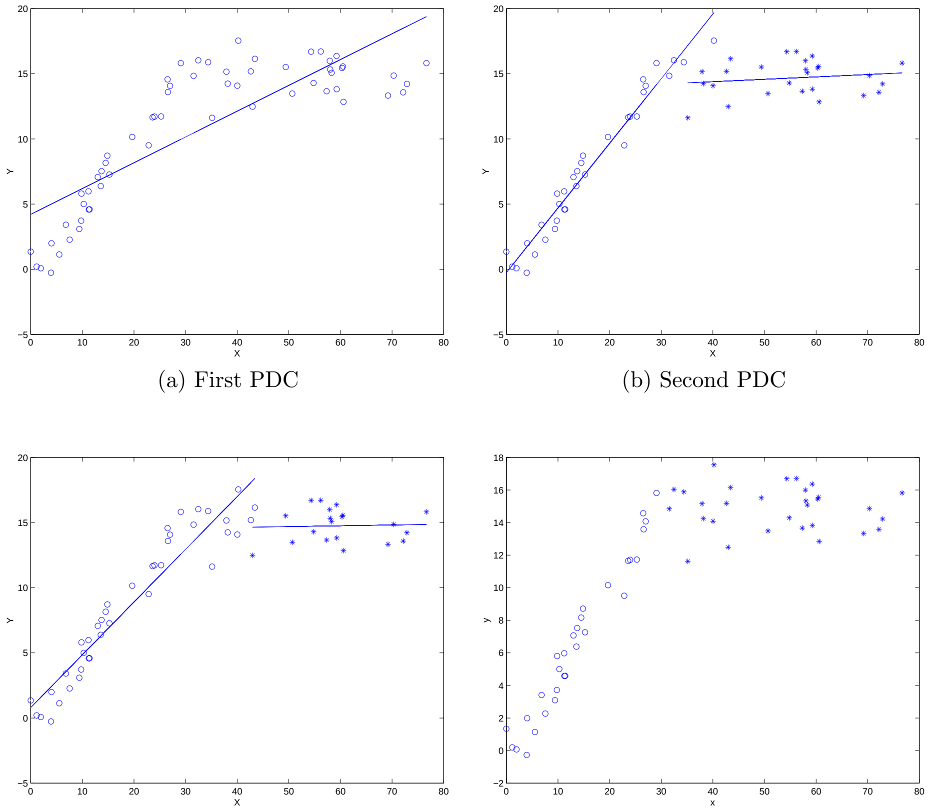


Figure 4. *The four PDCs found for data from Figure 2 (b) in the switching regression problem.*

In the three other cases shown in Figure 1 also more than one PDC may be obtained. In case (a) two PDCs are found: the first considers the sample as homogeneous and the second splits the data into the homogeneous group and the outliers. In case (c) only one PDC is obtained and this is the correct solution. In case (d) three possible PDCs, shown in Figure 5, are found. The interesting ones are the second and the third that differ in the way in which observations in the intersection of the two regimes are allocated. The method makes clear that some of these observations can be generated by both models. The true data generation pattern for this example is shown in Figure 5(d).



(a) First PDC

(b) Second PDC

(c) Third PDC

(d) True generated mixture

Figure 5. *PDCs for the structural chage data From 1(d). The two interesting solutions (2nd and 3rd) differ in the way the data around the change point are allocated between the two groups.*

Some remarks about the procedure are in order:

Remark 1. If p is large and n/p is small, let us say smaller than 20, many points can be extreme and many discriminators may be found, leading to many small initial groups. This will make the procedure slower and less powerful for the identification of small heterogeneous groups, as too many points will be deleted in the first split. In this case it is faster and safer to delete only discriminators of order m, that are defined as data points that are discriminators for at least m points in the sample. In Peña, Rodríguez and Tiao (2002) these discriminators are defined and analyzed for cluster problems in high dimensions.

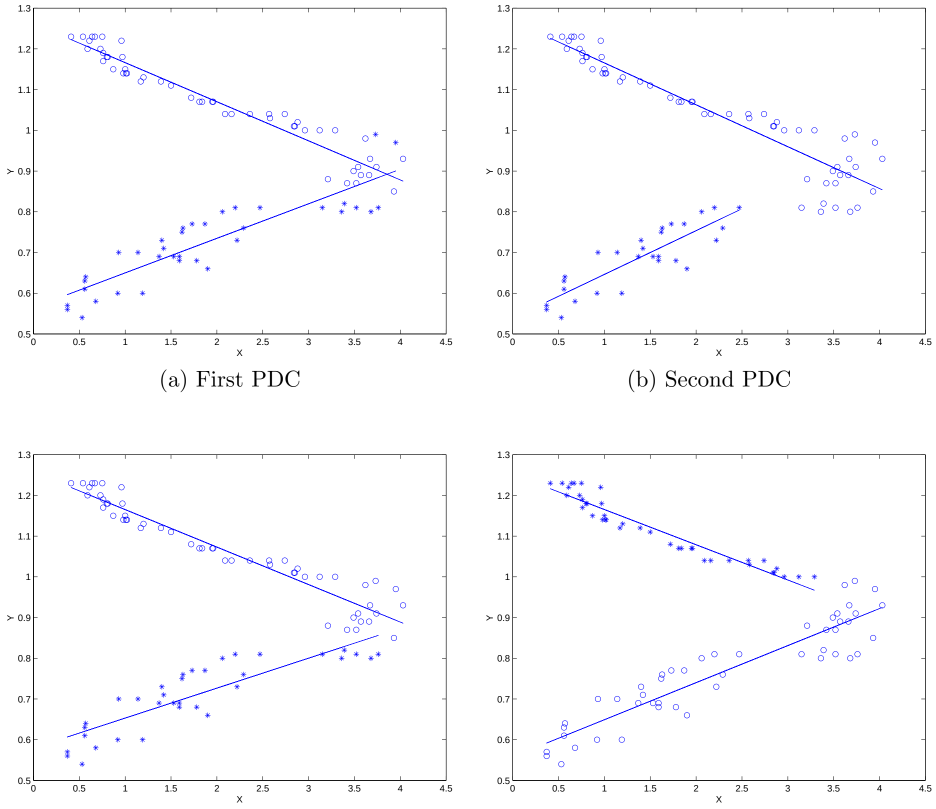
Remark 2. The procedure evaluates the relative size of statistics c_0 and d_1 by their sampling distribution which has been obtained by simulation. Although we believe that this type of crossbreeding between Bayesian and frequentist ideas enriches statistics, the procedure could be made easily completely Bayesian by working either with the posterior probabilities or the Bayes factors. In order to do so we have to introduce an alternative model to explain how the point y_i under consideration could be heterogeneous with the group. A simple solution is to assume as alternative distribution the scale normal contaminated model and use (2) to compute the posterior probabilities or the Bayes factors.

Remark 3. The solutions obtained by the SAR procedure are not very sensitive to the choice of the parameter h and therefore to the minimal size, $p + h$, of the basic set. When h is small the procedure obtains a large number of basic sets and this increases the power to find small heterogeneous groups of masking outliers, but the number of PDC increases and makes the procedure slower for large data sets. If we expect that heterogeneity will be due to the possible existence of two or more regressions, we can choose h moderately large.

5. EXAMPLES

In this section three examples are shown to illustrate the performance of the SAR algorithm and to compare it with other approaches. The first one is a real data set that has already be analyzed in the Bayesian literature. We show that the SAR procedure leads to the same solution obtained by MC^2 methods. The second example is a masking outlier example, and there the SAR procedure succeeds where standard MC^2 methods may fail completely. The third example is presented to illustrate the usefulness of the SAR procedure as a data mining tool in large high dimensional regression data sets.

Example 1. Figure 6 shows the ethanol data set, which relates the equivalence ratio, a measure of the richness of the air-ethanol mix for burning ethanol in a single-cylinder automobile test to the engine concentration of nitric oxide in engine exhaust (normalized by engine work) (Brinkman, 1981). Hurvich, Simonoff and Tsai (1998) analyze this data set with nonparametric regression and Hurn, Justel and Robert (2001) use Stephens (2000) birth-and-death proposal for carrying out MC^2 estimation of mixture switching regression models. Figure 6 shows the four PDCs found by the SAR procedure. The four PDCs differ in the way in which doubtful points are allocated: the two regimes are clearly identified in the four cases, but there are a large uncertainty on the allocation of the observations in the intersections of the two lines.



(a) First PDC

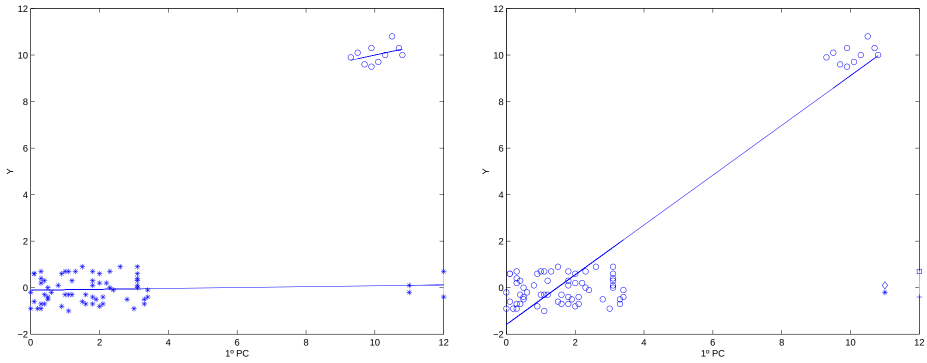
(b) Second PDC

(c) Third PDC

(d) Fourth PDC

Figure 6. *The four PDCs found for the Ethanol data when y= equivalence ratio and x= nitric oxide concentration.*

Example 2. An interesting classical example of masking is the artificial data generated by Hawkins, Bradu and Kass (1984). The model includes 75 data points of one response and 3 explanatory variables. The data is generated in such a way that the first 10 data points are high leverage outliers, whereas the next four are good observations with high leverage. Traditional methods of outlier detection fail in this case due to the high leverage problem and Justel and Peña (1996) showed that Gibbs Sampling also fails and it is unable to identify the outliers, even after 30,000 iterations. These authors showed that the lack of convergence in the algorithm is not a problem of the outlier model considered, as the same lack of convergence was found in all the outliers models included in the study, including a nonparametric hierarchical model based on Direchlet processes and later Justel and Peña (2001) introduced modifications in the MC chain to solve this problem. Figure 7 shows the two PDCs found with the SAR algorithm in this problem. The first one is the right one, and is the one chosen by the BIC criterion. The second one is the wrong one obtained by the standard application of the Gibbs Sampling algorithm.



(a) First PDC

(b) Second PDC

Figure 7. *The two PDCs for the HBK data.*

Example 3. In this example we consider a mixture of two regressions with omitted categorical variable in relatively large dimension data. The sample has been generated by the model

$$y = \beta_0 + \boldsymbol{\beta}_1' \boldsymbol{x} + \beta_2 z + u.$$

Here $\boldsymbol{x}$ has dimension 20, $u \sim N(0,1)$, and 400 values are generated for the first regression with $z = 0$, and 100 values for the second regression with $z = 1$. The parameter values have been chosen as $\beta_0 = 0$, $\boldsymbol{\beta}_1' = -1_{20}' = (-1, ..., -1)$ and $\beta_2 = 90$, and the values of the explanatory variables are independent random drawings from a uniform distribution. For the first regression the range of the explanatory variables is $(0, 10)$ so that $\boldsymbol{x}|(z = 0) \sim [U(0, 10)]^{20}$ whereas for the second the range is $(9, 10)$ so that $\boldsymbol{x}|(z = 1) \sim [U(9, 10)]^{20}$. These values have been chosen so that the standard residual plots from the fitted regression do not provide any evidence of heterogeneity. The results of the application of the SAR procedure to this data set with different values of h are indicated in Table 1.

If a small value of h is chosen, see the case $h = 10$, three PDCs are obtained that only differ in the allocation of two points as outliers. The first one assigns correctly 392 out of 400 to the first regression and 99 observations out of 100 to the second regression, and the remaining 9 points are isolated outliers. The second PDC assigns correctly 392 out of the 400 to the first regression and the whole 100 observations to the second and, again, the remaining 8 points are considered as outliers. Finally, the third PDC assigns correctly 391 out of 400 to the first regression and the 100 observations to the second and includes 9 isolated outliers. For moderate

Table 1. *PDCs for data in example 3 for several h values.*

h	PDC1			PDC2			PDC3			Time
	$z=0$	$z=1$	IA	$z=0$	$z=1$	IA	$z=0$	$z=1$	IA	
10	392	99	9	392	100	8	391	100	9	161s
25	400	100	0	–	–	–	–	–	–	118s
50	400	100	0	–	–	–	–	–	–	118s
75	500	0	0	–	–	–	–	–	–	93s

values of h, see the rows for $h = 25$ and or $50,$ the SAR procedure obtains only one PDC with all observations correctly classified and without outliers. For large values of $h,$ see the row of $h = 75,$ the minimum size of the basic group is 75+21=96, and the procedure obtain only one PDC with all observations in one group and the data set is found homogeneous. Note that this is to be expected when the size of the basic group is similar or larger than the group of heterogeneous observations. The table includes the running time in seconds of the program (written in Matlab) for several values of h. We have also tried $h = 6,$ that corresponds to the rule $ln(n-p)$. Nine PDCs configuration are found. Eight of them identify clearly the two groups and include between 391 and 394 observation of the first group in the first regression and between 99 and 100 of the second in the second regression. These 8 PDCs only differ in the number of outliers that goes from 6 to 10. The 9th PDC also finds two groups of 393 and 105 observations plus 2 outliers, but the second group includes 5 points from the first that produce strong biases in the estimation of the regression coefficients. Thus for large n we recomend that h should be chosen so that the size of the basic set is similar to the expected size of the smallest heterogeneous groups we want to detect.

6. MONTE CARLO ANALYSIS FOR THE CHANGING REGRESSION PROBLEM

We have carried out a Monte Carlo experiment for different structures of mixture regressions. 1000 samples for each combination of parameter values have been generated as follows. First 50 uniform $U(0,10)$ observations have been obtained and these values are used as the explanatory variable in the regression $y = 1.5x + 1.5\epsilon$, where ϵ is $N(0,1)$. Then, 30 observations are generated by $x = d+U(0,15)$ and used as explanatory variables in the regression $y = a+bx+\sigma\epsilon$ where, as before, ϵ is $N(0,1)$. Note that the important parameter d controls whether or not there is a horizontal gap $(H_g = d-10)$ and vertical gap $(V_g = a-15+db)$ between the end point of the first and the beginning point of the two regression lines.

Table 2 shows the frequency of obtaining an approximate correct solution for 13 different parameter values. The first case corresponds to the size of the test, as all the data is generated for the same model. Cases 1-8 correspond to structural change, cases 9-10 to switching regression and cases 11-12 to mixture of two regressions with an omitted dummy variable. The size of the test has been computed with $h = 3,$ similar values have been found for other values of $h \leq 10.$ In the cases of two regression lines without a horizontal or vertical gap usually two solutions are found, but the points in the intersection are attributed either to the first or the second regression. There is no information to identify these points into one of the two groups unless we introduce some separation among them. Thus, in this case the proportion of points correctly identified is smaller than when a gap is introduced..

Table 2 shows a second Monte Carlo experiment to analyze the power of the procedure for finding a concentrated contamination. 1000 samples for each combination of parameter values

Table 2. *Power study for the two regimes and contaminated regression.*

Two regimes regression

case	b	V_g	H_g	σ	95true	90true
0	1.5	0	0	1.5	0.979	1.00
1	-.5	0	0	1	0.217	0.666
2	-.5	0	0	0.5	0.512	0.929
3	-.5	-2.5	5	1	0.836	0.894
4	-.5	-2.5	5	0.5	0.906	0.971
5	-.5	+2.5	5	1	0.904	0.958
6	-.5	+2.5	5	0.5	0.980	1.000
7	.5	-2.5	5	1	0.732	0.767
8	.5	-2.5	5	0.5	0.827	0.869
9	-1.5	0	-10	1	0.220	0.808
10	-1.5	0	-10	0.5	0.435	0.921
11	1	0	-10	1	0.968	0.994
12	1	0	-10	0.5	0.975	0.996

Contaminated regression

case	n_1	n_2	t_0	s_0	95true	90true
1	50	10	1	0.1	0.950	0.980
2	50	10	1	0.5	0.836	0.885
3	50	10	1	1	0.629	0.817
4	50	10	2	0.1	0.982	1.000
5	50	10	2	0.5	0.969	0.994
6	50	10	2	1	0.965	0.992
7	50	10	3	0.1	0.988	0.999
8	50	10	3	0.5	0.973	0.995
9	50	10	3	1	0.976	0.998
10	50	10	4	0.1	0.985	1.000
11	50	10	4	0.5	0.986	0.998
12	50	10	4	1	0.983	0.997

have been generated as follows. First n_1 uniform $U(0, 10)$ observations have been obtained and these values are used as the explanatory variable in the regression $y = 1.5x + \epsilon$, where ϵ is $N(0, 1)$. Then, n_2 observations are obtained from $N((x_0, y_0), s_0\boldsymbol{I})$, where x_0 is a random value from a uniform $U(-2.5, 12.5)$, and

$$y_0 = 1.5x + t_0 s_R \sqrt{(1 - x_0^2)},$$

where s_R is the residual standard deviation in the regression using the first n_1 observations. Note that t_0 represents the standardized size of the outlier and its location is made in agreement with this parameter. Finally the parameter s_0 defines the dispersion of the outliers with respect to its center. The table shows that the power of the SAR procedure in all cases is large, and the procedure seems to overcome the masking problem.

7. CONCLUDING REMARKS

The SAR procedure seems to offer a powerful method for the Bayesian analysis of regression mixture models. The method can be helpful in identifying patterns in heterogeneous regression data, including masked outliers, switching regression, change point problems and other multiple regime situations. Although the main contribution of the procedure is to find structure in the data as an exploratory tool, the selection between the possible data configurations (PDCs) can be done by applying a MC^2 algorithm for model estimation. An important difference of the SAR procedure with respect to alternative methods for finding heterogeneity in regression mixtures is that the mixture components do not have to compete to classify each observation. Thus, more than one possible solution may exist and the procedure will find all solutions coherent with the model structure, under different restrictions implied by the conditioning of the homogeneous enlarged basic sets. This property gives a high robustness to the SAR procedure because when there exist observations that could be assigned to various components of the mixture they are usually not split up between the different mixture components but are allocated in groups to the PDCs, avoiding the well known masking problem in outlier detection. Finally, the ideas presented here can be easily extended to other regression situations as non linear regression problems, heteroskedastic regression models, generalized linear models and multivariate regression models, including simultaneous equation econometric models.

ACKNOWLEDGEMENTS

Daniel Peña and Julio Rodríguez acknowledge support from GRANT BEC 2000-0167, MCYT, Spain, and by the Cátedra BBVA de Calidad.

REFERENCES

Aitkin, M. (2001). Likelihood and Bayesian Analysis of Mixtures. *Statistical Modelling* **1**, 287–304.

Andrews, D. W. K., Lee, I. and Ploberger, W. (1996). Optimal changepoint tests for normal linear regression. *Journal of Econometrics* **70**, 9–38.

Brinkman, N. D. (1981). Ethanol fuel–a single cylinder engine study of efficiency and exhaust emissions. *SAE Trans* **90**, 1410–1427.

Box, G. E. P. (1980). Sampling and Bayesian Inference in Scientific Modelling and Robustness, (with discussion). *J. Roy. Statist. Soc. A* **143**, 383–430.

Box, G. E. P. and Tiao, G. C. (1968). A Bayesian Approach to Some Outliers Problems. *Biometrika* **55**, 119–129.

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

Celeux, G., Hurn, M. and Robert, C. (2000). Computational and inferential difficulties with mixture posterior distributions. *J. Amer. Statist. Assoc.* **95**, 957–970.

Cook, R. D. and Weisberg, S. (1982). *Residuals and influence in regression.* Chapman and Hall.

Geisser, S. (1980). Discussion of a paper by G.E.P. Box. *J. Roy. Statist. Soc. A* **143**, 416-417.

Geisser, S. (1987). Influential observations, diagnosis and discordancy test. *J. Appl. Statist.* **14**, 133-142.

Hawkins, D. M, Bradu, D. and Kass, G. V. (1984). Location of several outliers in multiple regression data using elemental sets. *Technometrics* **26**, 197–208.

Hurn, M., Justel, A. and Robert, C. P. (2002). Estimating mixtures of regressions. *Journal of Computational and Graphical Statistics* (to appear).

Hurvich, C. M., Simonoff, J. S. and Tsai, C. (1998). Smoothing parameter selection in nonparametric regression using an improved Akaike information criterion. *J. Roy. Statist. Soc. B* **60**, 2, 271–293.

Justel, A. and Peña, D. (1996). Gibbs Sampling Will Fail in Outlier Problems with Strong Masking. *Journal of Computational and Graphical Statistics* **5**, 2, 176–189.

Justel, A. and Peña, D. (2001). Bayesian Unmasking in Linear Models. *Computational Statistics and Data Analysis* **36**, 69-84.

Quandt, R. E. (1958). The Estimation of the Parameters of a Linear Regression System Obeying Two Separate Regimes. *J. Amer. Statist. Assoc.* **53**, 873–880.

Peña, D. and Guttman, I. (1993). Comparing Probabilistic Methods for Outlier Detection. *Biometrika* **80**, 603-610.

Peña, D. and Tiao, G. C. (1992). Bayesian Robustness Functions for Linear Models. *Bayesian Statistics 4* J.M. Bernardo et al (eds). Oxford University Press, 365–388.

Peña, D. and Tiao, G. C. (2002). The SAR Procedure: A Diagnostic Analysis of Heterogeneous Data. (submitted).

Peña, D., Rodríguez, J. and Tiao, G. C. (2002). Cluster Analysis by the SAR Procedure. (manuscript).

Pettit, L. I. (1990). The Conditional Predictive Ordinate for the Normal Distribution. *J. Roy. Statist. Soc. B* **52**, 1, 175–184.

Pettit, L. I. and Smith, A. F. M. (1985). Outliers and influential observations in linear models. (with discussion), *Bayesian Statistics 2* (J. M. Bernardo, M. H. DeGroot, D. V. Lindley and A. F. M. Smith, eds.), Amsterdam: North-Holland, 473–494.

Richarson, S. and Green, P. J. (1997). On Bayesian analysis of mixtures with an unknown number of components. *J. Roy. Statist. Soc. B* **59**, 731–758.

Stephens, M. (2000). Bayesian Analysis of Mixture Models with an Unknown Number of Components–An Alternative to Reversible Jump Methods. *The Annals of Statistics* **28**, 40–74.

Schweder, T. (1976). Some 'optimal' methods to detect structural shift or outliers in regression. *J. Amer. Statist. Assoc.* **71**, 491–450.

APPENDIX

Let $I = (i_1, ..., i_k)$ represent the indices for k observations with vector of responses $\boldsymbol{Y}_I$ and regressors $\boldsymbol{X}_I$. Calling $\widehat{\boldsymbol{\beta}}_{(I)} = (\boldsymbol{X}'_{(I)}\boldsymbol{X}_{(I)})^{-1}\boldsymbol{X}'_{(I)}\boldsymbol{Y}_{(I)}$ to the vector when these observations are deleted from the sample $\boldsymbol{X}, \boldsymbol{Y}$, and using the well known expression (see Cook and Weisberg, 1982) we have that,

$$\widehat{\boldsymbol{\beta}}_{(I)} = \widehat{\boldsymbol{\beta}} - (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}_I(\boldsymbol{I} - \boldsymbol{H}_I)^{-1}\boldsymbol{e}_I,$$

where $\boldsymbol{H}_I = \boldsymbol{X}_I(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'_I$ and $\boldsymbol{e}_I = \boldsymbol{Y}_I - \boldsymbol{X}_I\widehat{\boldsymbol{\beta}}$. Now calling

$$\boldsymbol{e}_{I(I)} = \boldsymbol{Y}_I - \boldsymbol{X}_I\widehat{\boldsymbol{\beta}}_{(I)} = (\boldsymbol{I} - \boldsymbol{H}_I)^{-1}\boldsymbol{e}_I,$$

we have that

$$(n - p - I)s^2_{(I)} = (n-p)s^2 - \boldsymbol{e}'_I(\boldsymbol{I} - \boldsymbol{H}_I)^{-1}\boldsymbol{e}_I$$

and it is easy to show that $\boldsymbol{H}_{I(I)} = \boldsymbol{X}_I(\boldsymbol{X}'_{(I)}\boldsymbol{X}_{(I)})^{-1}\boldsymbol{X}'_{(I)}$ verifies $\boldsymbol{H}_{I(I)} = \boldsymbol{H}_I(\boldsymbol{I} - \boldsymbol{H}_I)^{-1}$. Now, making $\boldsymbol{Y}_I = (y_i, y_j)'$, $I = (i, j)$, using the previous equations and after some straightforward algebra, we found

$$t^2_{i(ij)} = \frac{(e_i + h_{ij}e_j/(1-h_{jj}))^2}{a_{ii} - c/(1-h_{jj})[dh^2_{ij} + b_{ii}e^2_j + 2e_ih_{ij}e_j]}$$

where $c = (n-p-2)^{-1}$ and $d = (n-p)s^2$ are constant and $b_{ii} = (1-h_{ii})$, $a_{ii} = c[db_{ii} - e^2_i]$ are only function of point ith.

To understand better this function, we consider the simple regression through the origin case with $\sum x_i = \sum y_i = 0$ and $\sum x^2_i = 1$ so that $h_{ij} = x_ix_j$. Then, we can write, for each (e_i, x_i), this function as

$$t^2_{i(ij)} = f(e_j, x_j) = \frac{(n-p-2)(e_i + e_jx_ix_j/(1-x^2_j))^2}{(d(1-x^2_i) - e^2_i) - \left(1/(1-x^2_j)\right)(dx^2_ix^2_j + (1-x^2_i)e^2_j + 2e_ie_jx_ix_j)}.$$

It can be proved that if $e_i \neq 0$ and $x_i \neq 0$, then, $f(e_j, x_j) = f(-e_j, -x_j)$ and $f(e_j, -x_j) = f(-e_j, x_j)$ and $(e_j = 0, x_j = 0)$ is a saddle point. If $e_i \neq 0$ or $x_i \neq 0$, then, $f(e_j, x_j) = f(-e_j, -x_j) = f(e_j, -x_j) = f(-e_j, x_j)$ and $(e_j = 0, x_j = 0)$ is a minimum.

DISCUSSION

HAL S. STERN *(Iowa State University, USA)*

1. *Introduction.* Peña, Rodríguez, and Tiao (PRT) present a novel exploratory approach for investigating data heterogeneity when fitting regression models. It is my pleasure as discussant to thank them on behalf of those attending the conference. This introduction provides a brief review of the PRT approach and some preliminary comments. The remaining sections provide more detailed discussion of some important issues.

The authors use the term data heterogeneity to refer to the existence of clusters or subgroups among the sample units. In the regression context the subgroups may be clusters of outlying observations in an otherwise homogeneous population or subpopulations best described by different linear models. The SAR (split and recombine) procedure, described more fully in Peña and Tiao (2002), is used to discover such heterogeneities. The key element in the approach is a 3-step process that identifies a subgroup for which a single regression model seems appropriate. This process is repeated until all such subgroups are identified. The three steps are: (1) outliers are identified using the predictive ordinate $p(y_i|Y_{(i)})$ where $Y_{(i)}$ refers to the vector of responses without observation i; (2) the remaining data is split into small homogeneous groups known as basic sets (each containing at least a specified minimum number of observations) using the concept of "link"ing or discriminator points; (3) these basic sets are enlarged by incorporating all of the observations in the data set that are consistent with the linear model that describes the homogeneous group. Once a basic set is selected and enlarged, the SAR procedure is restarted on the set of all observations that are excluded from the enlarged group. The outcome of the procedure is called a *possible data configuration* (PDC). The outcome depends on the group which is enlarged first thus there may be several different PDCs for a given data set.

The authors cite as their goal to "solve" (my quotes) the heterogeneity problem. Though I quibble below with the idea of solving the heterogeneity problem in this way, it is clear that SAR is a powerful exploratory technique that addresses a number of important heterogeneous regression models. It is interesting to note that the SAR procedure, though proposed by Bayesian researchers and motivated by Bayesian principles, will likely be welcomed by non-Bayesians. The cross-validatory ideas that motivate SAR are familiar in regression diagnostics and there is little for anyone to object to in terms of probability modeling. The remainder of this discussion considers four issues: the types of heterogeneous mixture models considered, appropriate alternatives for comparison with SAR, whether one can solve the heterogeneity problem in this way, and the nature of exploratory analysis.

2. *Mixture of Regressions.* Four heterogeneous regression models are used by the authors to demonstrate the utility of the SAR procedure. These are an outlier contamination model, a structural change model, a switching regression model (subgroups may differ in all coefficients), and a missing categorical variable model (subgroups differ in intercept only). Each of the four can be expressed as a probability model for scalar y_i conditional on covariate vector x_i of the form $y_i \,|\, x_i \sim \sum_j \alpha_j \mathrm{N}(x_i'\beta_j, \sigma_j^2)$. This common form of mixture model misses an important form of heterogeneity. The next paragraph describes a more general regression mixture model. It turns out that the SAR procedure is effective in identifying this form of heterogeneity as well.

One natural way to motivate mixture models is to introduce an indicator z_i for each case, with $z_i = j$ if the ith observation comes from subgroup j (or has level j for an unobserved categorical variable). Then the mixture model for y_i conditional on x_i is obtained by considering the joint model $y_i, z_i \,|\, x_i \sim p(z_i \,|\, x_i)p(y_i \,|\, z_i, x_i)$ and marginalizing over z_i to obtain $y_i \,|\, x_i \sim \sum_j \Pr(z_i = j \,|\, x_i)p(y_i \,|\, z_i, x_i)$. The model considered by the authors is obtained if $\Pr(z_i =$

$j \mid x_i) = \alpha_j$ independent of x_i and a normal linear regression model is used for y_i given z_i and x_i. When viewed in this way however it seems much more natural to allow for the possibility that the latent variable z_i depends on x_i, perhaps with a multinomial model $\Pr(z_i = j \mid x_i) = \exp(x_i'\delta_j)/\sum_k \exp(x_i'\delta_k)$. The multinomial-regression mixture model is used by Peng *et al.* (1996) as a "mixture of experts" and by Morduch and Stern (1997) as an approach to heterogeneity in data regarding family spending in Bangladesh.

The more complicated mixture poses some problems for inference in that the likelihood function may have multiple modes (including some degenerate cases with only enough observations in some subgroups to identify the regression parameters) and the multinomial parameters describing the latent indicators are not well identified. Example 3 in the paper is a mixture of this more complex type because the covariate distribution differs across the subgroups. That example proves that the SAR procedure can handle data heterogeneity of this type.

3. *Alternatives to SAR.*. Several parts of the paper refer to Gibbs sampling or Markov chain Monte Carlo (MCMC) as an alternative to SAR. This is used as a shorthand way to reference a formal analysis of an apparently relevant model, say the outlier contamination model, using MCMC. Of course this is not really a fair comparison at all in the sense that the formal prior-to-posterior analysis provides a very different type of information than is provided by the SAR analysis. It may be difficult to get MCMC algorithms to converge for mixture models but once this is achieved the investigator is rewarded with the usual posterior inferences that Bayesians find so attractive.

As SAR does not offer such inferences it might be better compared to traditional and new clustering techniques. Especially noteworthy are the model-based clustering of Banfield and Raftery (1993) and the work by Liu in this volume. Model-based clustering allows for multivariate Gaussian clusters with varying shape and orientation and thus might pick up the mixtures of linear models contemplated here. Recent work on data visualization is also relevant (Sutherland *et al.* , 2000).

It might also be interesting to explore how well SAR does relative to the traditional approach of statisticians building up from simpler models with the aid of model diagnostics. For example, a traditional linear regression analysis of the authors' example 1 along with residual plots might suggest a quadratic model. Another model-based approach would be to search for modes of the likelihood under a hypothesized heterogeneity or mixture model. One wonders if the different PDCs might show up as different modes in the likelihood.

4. *Solving the Heterogeneity Problem.* The authors clearly identify their procedure as exploratory, suggesting that formal inference might be done after SAR helps to identify a suitable model. It is difficult to argue with such an approach. At times though there are claims about the SAR procedure's ability to find the "correct" solution (at one point noting that using BIC to choose among PDCs would have selected the right answer). There are at least three reasons it would be better to avoid such claims. First, the primary goal of such exploration should be to generate interesting views of the data not to identify the "true" data generating mechanism which is destined to remain unknown in most applications. Second, all of the simulated data sets are from linear regression models or mixtures of linear regressions. These simulations show that SAR can identify heterogeneity and that SAR won't always identify heterogeneity, but the simulations don't address how SAR might perform on data best described with a non-linear relationship. The ethanol data of example 1 looks a bit like data from a quadratic model. Though the piecewise linear model may fit a bit better in that particular data set, this example raises the question as to whether all quadratics should be viewed as mixtures of regressions. A third argument for caution regarding SAR's performance concerns the effect of the minimum

group size parameter. This is a key parameter and the advice provided here is that investigators should choose this parameter to match “the expected size of smallest heterogeneous groups we want to detect”. Thus the quality of SAR’s results will depend mightily on the investigator’s knowledge regarding the expected heterogeneity.

5. *Concluding Remarks on SAR and Exploratory Analysis..* There are clear roles for exploratory tools in data analysis; such methods help find structure, suggest models, and critique models. The SAR approach seems to offer much in the first two of these roles. Exploratory techniques like SAR can help find structure in new and difficult settings. The mining of large databases and the analysis of large sets of biological data each demonstrate that there is a clear need for exploratory tools in the current scientific environment. The authors mention data mining as a possible application of SAR. The relationships sought in data mining applications are of all types, not just linear, so it would be interesting to see how SAR performs in the data mining context.

In some applications the exploratory analysis may reveal all that we need to know. It is much more common that the outcome of an exploratory analysis is the suggestion of a probabilistic model allowing formal inference regarding the data-generating process. In this regard SAR performs well. The end result of the SAR analysis will often be the suggestion of a suitable mixture model; one must be careful to avoid overstating the significance of such “post hoc” explanations of heterogeneity.

The authors’ conclusions describe plans to extend the application of SAR to finding data heterogeneity in the fitting of generalized linear models, non-linear models, simultaneous equation models, etc. This is clearly an intriguing idea given the success they have had with linear models but there is some risk. One major appeal of the SAR approach to heterogeneity in linear models is that there is a simplicity and intuition which users are likely to find appealing. This may be more difficult to provide if the SAR building blocks are more complex models.

SAR is an exploratory tool worth having in the proverbial “statistician’s toolbox”. Though it doesn’t “solve” the heterogeneity problem it seems capable of providing real insight into the structure of data for a wide range of applications.

P. IGLESIAS and R. B. ARELLANO-VALLE *(Universidad Católia, Chile)*

In this paper an interesting procedure based on SAR is introduced for modelling heteroscedasticity in the context of regression analysis. Such procedure is obtained from the predictive distribution of the normal regression model and the usual non-informative prior distribution. An interesting property is that if the normal regression model is replaced by an elliptical regression model, but the same non-informative prior distribution is considered, then such procedure will remain invariant (Osielwaski and Steel, 1993). On the other hand, if another prior distribution is adopted in order to obtain marginal equivalency with the normal model (Arellano-Valle, Iglesias an Vidal, 2002), then the procedure obtained in this case is also invariant under elliptical models, but will yield different results than those obtained by considering the usual non-informative prior distribution. Although invariance (with respect to the likelihood in this case) can be a desirable property, we would expect that models with heavier or lighter tails than the normal model yield different clusters. We think that this ”lack of robustness” is due to the fact that the procedure is based on the predictive distribution only, that is, the other components of the Bayesian model are not considering. It would be interesting to introduce in the procedure those parameters for which the posterior distributions is not invariant to departures from normality (the precision parameter in this case). Furthermore, if we consider that the usual non-informative prior distribution is not the only reference prior distribution, the following question arise naturally: What is the performance of the proposed procedure when using

Jeffrey's prior or some other reference prior distribution that depends on the order of the model parameters (Bernardo y Smith, 1994)? What is the sensitivity of the procedure to changes on the prior distribution? Finally, there are alternative methods for modelling heteroscedasticity, like the product partition model introduced by Hartigan (1990) which considers all the components of the Bayesian model when obtaining the clustering. Quintana and Iglesias (2002) propose to adopt this approach within the context of decision theory. The idea is to clearly define the purpose of the study (such as estimation, hypothesis testing, outlier detection, etc) and from this to develop a clustering algorithm that depends also on the loss function associated to the decision problem. An interesting problem is to consider justifying the proposed procedure from a decision theoretic viewpoint, doing at the same time a comparative study with other procedures considered in the literature.

CHRISTIAN P. ROBERT *(CEREMADE, Université Paris Dauphine, France)*

While I find the diagnosis tools developped by the authors quite clever and apparently very efficient, I cannot but question the relevance of *exploratory* devices when "exact" procedures are available. Indeed, we studied in Hurn et al. (2002) the performances of a (fully) Bayesian approach to the estimation of the number of components in a mixture of standard [like (1)] and generalized linear models. Using Stephens's (2000) continuous time MCMC algorithm with a simple birth-and-death proposal, we found very satisfactory performances of the algorithm, with no obvious dependance on the starting values (as should be).

Given that such a (fully) Bayesian modelling is possible [and can be implemented in a fairly straightforward manner], it necessarily brings more information that the exploratory SAR procedure, since the later cannot, for instance, classify competing regression lines in terms of their posterior probability or eliminate dubious solutions. While Gibbs sampling indeed has difficulties to escape the "fatal attraction" of leverage points in outlier problems, more hybrid solutions using subsampling *[in a spirit similar to the one developped in this paper]* or tempering (Celeux et al., 2000) should work better.

In addition, the SAR procedure does not strike me as being fundamentally Bayesian, since the predictive $p(y|Y)$ is built on a single normal regression model with conjugate priors, while the actual model is a mixture of normal regression models with or without conjugate priors. So the discriminating factors $c_0(i)$ are only formally related to the Bayesian approach. Also, for other generalized linear models, the contruction of the predictive distribution $p(y|Y)$ is not possible in closed form. It thus seems to me that, unless the authors propose an alternative criterion, the necessary call to an approximative device of a numerical or simulational nature is not fundamentally different (in difficulty level) from the construction of the full MCMC apparatus. I nonetheless find the construction of the discriminators quite interesting in that they may serve as *anchors*, to borrow from Liu et al. (2002) terminology, for constructing more adaptive or more efficient MCMC samplers.

REPLY TO THE DISCUSSION

We first want to thank our official discussant, Hal Stern, for his comments and to Christian Robert, Pilar Iglesias and Reinaldo Arellano-Valle for their contributions to the discussion of our paper.

Our reply to Hal Stern must be brief because we are in agreement with all the points he raised, and we fully share his point of view of the usefulness of the SAR procedure for exploratory data analysis. We appreciate his thoughtful and wise comments on our procedure which will stimulate us to extend it for further applications. We agree that our method should be compared to clustering methods, and in fact we have carried out a comparison of the SAR procedure with and some of traditional and new clustering methods, including k-Means, Mclust, the Projection Pursuit method by Peña and Prieto (2000) and others (see Peña, Rodríguez and Tiao, 2002). The result we have found by an extensive Monte Carlo study is that the SAR procedure seems to have the best performance according to standard criteria to compare cluster methods. In the present paper we have emphasized the exploratory role of the SAR procedure in regression and thus the comments of Harl Stern on our claims on solving the heterogeneity problem are right. However, we also believe that the SAR procedure can be extended to provide a formal structure for inference but this will be the subject of further research.

We agree with Christian Robert that if we had exact procedures we better use them. However the problem is that with complex data set usually there is no exact procedure available. We may assume a model, run MCMC and get an answer but this type of ”exact” procedure when applied with the wrong model can be misleading, as illustrated in the outlier problem we refer to in our paper. We believe that the SAR procedure can help in formulating a reasonable model for the data in hand. Regarding the second point, we assume that a regression model has been fitted to the data and we want to check for homogeneity. If the model was a mixture of regression we could in principle apply the SAR procedure to the mixture to check if it is homogeneous. We agree that a key advantage of the SAR procedure is its simplicity and small computational burden and we will try to keep this feature in its extension to generalized linear models.

The comments and suggestions by P. Iglesias and R.B. Arellano-Valle are very appropriate. We do not expect difference in behavior in the SAR procedure by moderate changes of the prior distribution, but this point deserves a carefully investigation. Also the suggestion to develop cluster algorithms as decision problems is attractive and we will be interested in developments in this area in the future.

ADDITIONAL REFERENCES IN THE DISCUSSION

Arellano-Valle, R. B., Iglesias, P. and Vidal, I. (2002). Bayesian inference for elliptical linear models: Conjugate analysis and model comparison. *In this volume*.

Banfield, J. D. and Raftery, A. E. (1993). Model-based Gaussian and non-Gaussian clustering. *Biometrics* **49**, 803–821.

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

Celeux, G., Hurn, M. and Robert, C.P. (2000) Computational and inferential difficulties with mixtures posterior distributions. *J. Amer. Statist. Assoc.* **95**, 957–979.

Hartigan, J. A. (1990). Partition models. *Comunication in Statistics–Theory and Methods* **19**, 2745–2756.

Liu, J. S., Zhang, J. L., Palumbo, M. L. and Lawrence, C. E. (2002) Bayesian clustering with variable and transformation selections. *In this volume*.

Morduch, J. J. and Stern, H. S. (1997). Using mixture models to detect sex bias in health outcomes in Bangladesh. *J. Econometrics* **77**, 259–276.

Osiewalski, J. and Steel, M. F. J. (1993). Robust Bayesian inference in elliptical regression models. *J. Econometrics* **57**, 345–363.

Peng, F., Jacobs, R. A. and Tanner, M. A. (1996). Bayesian inference in mixture-of-experts and hierarchical mixtures-of-experts models with an application to speech recognition. *J. Amer. Statist. Assoc.* **91**, 953–960.

Peña, D. and Prieto, F.J. (2001). Clustering by Projections. *J. Amer. Statist. Assoc.* **96**, 1433–1445.

Peña, D., Rodríguez, J. and Tiao, G. C. (2002). The SAR Procedure for Cluster Analysis. *Tech. Rep.*, Universidad Carlos III de Madrid, Spain.

Quintana, F. and Iglesias, P. (2002). Nonparametric Bayesian clustering and product partition models. *Tech. Rep.*, Universidad Pontificia, Chile.

Sutherland, P., Rossini, A., Lumley, T., Lewin-Koh, N., Dickerson, J., Cox, Z. and Cook, D.(2000). Orca: A visualization toolkit for high-dimensional data. *J. Comp. Graph. Statist.* **9**, 509–529.