Estimating mixtures of regressions

Merrilee Hurn†

Department of Mathematical Sciences, University of Bath, UK

Ana Justel‡

Department of Statistics, Universidad Autónoma de Madrid, Spain

Christian P. Robert†

Université Paris Dauphine and CREST, Insee, Paris

Summary. In this paper, we show how Bayesian inference for switching regression models and their generalisations can be achieved by the specification of loss functions which overcome the label switching problem common to all mixture models. We also derive an extension to models where the number of components in the mixture is unknown, based on the birth-and-death technique developed in Stephens (2000a). The methods are illustrated on various real datasets.

Keywords: Bayesian inference, birth-and-death process, label switching, logistic regression, loss functions, MCMC algorithms, Poisson regression, switching regression

1. Introduction

The *switching regression* model is well-known in the econometrics literature; it arises when an observed quantity y depends on a vector of covariates x in a linear way

$$y = x'\beta_i + \sigma_i \epsilon \;, \qquad \epsilon \sim g(\epsilon) \tag{1}$$

where the (β_i, σ_i)'s $(i = 1, \cdots, k)$ vary among a set of k possible values with probabilities $p_1, \cdots, p_k$. In other words, assuming a normal distribution on the perturbation ϵ, the conditional distribution of y given x is a mixture of normal distributions

$$y|x \sim p_1 \mathcal{N}(x'\beta_1, \sigma_1^2) + \cdots + p_k \mathcal{N}(x'\beta_k, \sigma_k^2) \;, \tag{2}$$

as illustrated by Figure 1 in a simple regression case.

This model was introduced in Goldfeld and Quandt (1976) and has been mainly studied from a likelihood point of view. Its appeal is clear when considering datasets such as those in Figure 2 *(left)*, which describes the equivalence ratio, that is, the richness of the air-ethanol mix in an engine, against the concentration of nitrous oxide in exhaust, as presented in Hurvich *et al.* (1998), and in Figure 2 *(right)*, which plots the pairs (GNP,CO2) for various countries, where GNP denotes the gross national product per capita in 1996 and CO2 the estimated CO_2 emission per capita the same year. The left graph clearly indicates two

†Supported by EU TMR network ERB–FMRX–CT96–0095 on *"Computational and Statistical Methods for the Analysis of Spatial Data"*.

‡Partially supported by CREST during a visit in May-June 1999 and by DGES Grant PB-970021.

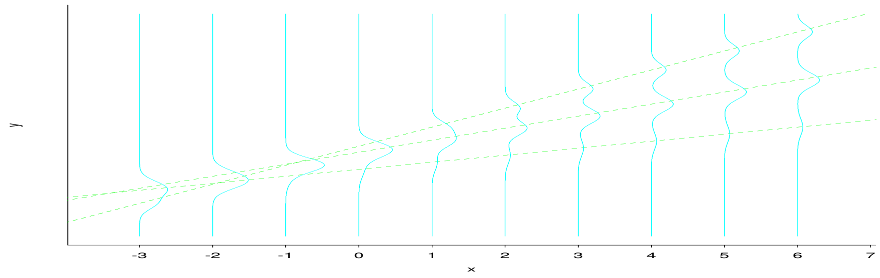

Fig. 1. Conditional distributions of y for several values of x, against the regression lines.

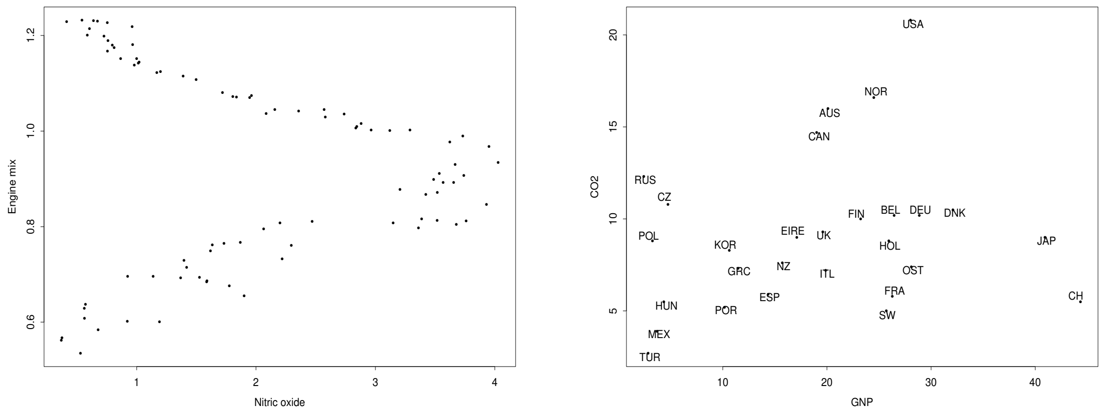

Fig. 2. *(left)* Equivalence ratio against exhaust nitric oxide concentration (*Source:* Hurvich *et al.*, 1998); *(right)* representation of the GNP and CO2 emission levels in 1996 for various countries (*Source:* OECD).

different nitrous oxide concentration dependencies. The data in the right graph is quite spread out, with no clear linear trend; however, there do seem to be several groups for which a linear model would be a reasonable approximation. To identify those groups and the corresponding linear models is of interest for low GNP countries as it may help to clarify on which development path they are embarking.

As well as econometrics and chemometrics, there exist many other areas where heterogeneous covariate dependent populations are found and studied, both at the classification level, namely to identify the homogeneous subpopulations, and at the inference level, namely to estimate the corresponding models. See for instance Bar–Shalom (1978), Quandt and Ramsey (1978), or Kiefer (1980). Extensions of (1) cover time-dependent models such as hidden Markov models and nonlinear switching regressions, for example Hamilton (1989), Shuway and Stoffer (1991), or Billio *et al.* (1999).

The framework is also wider than just simple linear regressions. We can for instance consider mixtures of logistic regressions:

$$P(y = 1|x) = \sum_{i=1}^{k} p_i \, \frac{\exp(x'\beta_i)}{1 + \exp(x'\beta_i)} \;, \tag{3}$$

as seen later in an example describing investment choices based on socio-economic covariates, and mixtures of Poisson regressions,

$$y|x \sim \sum_{i=1}^{k} p_i \, \mathcal{P}(\exp(x'\beta_i)) \;, \tag{4}$$

as illustrated on two datasets relating numbers of accidents and covariates.

The paper is organised as follows: in Section 2 we describe a Bayesian approach to these models and uncover the difficulties which arise in their processing, including the specificities of MCMC algorithms that must be used in such settings and discussing their limitations. In Section 3, we discuss different choices of loss functions depending on different inferential imperatives and designed to overcome these difficulties. In Section 4, we present a natural extension to unknown numbers of components, based on the birth-and-death process technology, for the normal regression as well as the two regression models (3) and (4). We illustrate those methods on several simulated and real datasets.

2. A Bayesian approach

In this section, we briefly review some of the aspects of the Bayesian processing of such models, and describe the inferential difficulties which arise as a result of a lack of identifiability.

2.1. Prior specification

The inferential difficulties we describe below in Section 2.3 are not particular to a specific prior modelling but rather appear as a generic issue. We will thus choose for our simulations and applications standard conjugate priors, that is, independent normal priors on the regression coefficients β_i's, such as $\mathcal{N}(0, 10)$ or $\mathcal{N}(0, 10\sigma_i^2)$, inverse gamma priors on the σ_i^2's for (2), most often an $\mathcal{E}xp(1)$ distribution when dealing with a normalised dataset, and a Dirichlet prior on the weight vector $(p_1, \ldots, p_k)$, $\mathcal{D}_k(1, \ldots, 1)$. This particular choice allows us to use a Gibbs sampler but other choices are equally acceptable since they can be implemented via a Metropolis–Hastings algorithm, as described in the next paragraph.

Notice first that improper priors, hence standard flat priors, are not acceptable in this setup, because they lead to undefined posterior distributions, whatever the sample size (see, e.g., Robert, 1996), and, secondly, that exchangeable priors on the components, that is, priors that are invariant under permutations of the component labels, produce posterior distributions with identical marginal distributions on all components. This feature is at the root of the inferential difficulties discussed in Section 2.3.

2.2. MCMC algorithms for switching regressions

2.2.1. *Gibbs sampling*

The Gibbs sampler is the most commonly used approach in mixture estimation (Diebolt and Robert, 1994) and consists in augmenting the sample $(x_1, y_1), \cdots, (x_n, y_n)$ from (2) with artificial allocation variables z_i, in order to "de-mix" the observed sample. This allows for simulation of the parameters of each component conditionally on the allocations, that is, taking into account only the observations which have been allocated to this component. For instance, given a mixture of simple regressions,

$$y_i \sim \sum_{j=1}^{k} p_j \mathcal{N}(\beta_{0j} + \beta_{1j} x_i, \tau_j^2)$$

the conditional full distributions involved in this case are simply normal on both β_{0j} and β_{1j} and gamma on τ_j^{-2}.

For logistic regressions (3), a second level of completion is helpful to simulate from the full conditionals since, once the observations have been grouped by allocations, the k densities are of the form

$$\left\{ \prod_{i=1}^{m} \frac{\exp(y_i x_i' \beta)}{1 + \exp(x_i' \beta)} \right\} e^{-\|\beta\|^2 / 2\tau^2} ,$$

which is not directly tractable, but simulation from this distribution can be done via a *slice sampler* (see, e.g., Damien *et al.*, 1999, or Robert and Casella, 1999), leading to truncated normal proposals. In the case of the Poisson regressions (4), the posterior distribution is also untractable directly, being

$$\exp \left\{ \left(\sum_i y_i x_i \right)^t \beta - \sum_i e^{x_i^t \beta} \right\} e^{-\|\beta\|^2 / 2\tau^2} , \tag{5}$$

and a similar slice sampling solution applies.

2.2.2. *Random walk Metropolis-Hastings algorithms*

The intricacy of using the Gibbs sampler plus possibly a slice sampler, while natural, is not necessary. Indeed, a Metropolis–Hastings algorithm based on a properly calibrated Cauchy perturbation may be used for all types of regression models considered here (see Robert and Casella, 1999, for details). For an unconstrained parameterisation θ (that is, after a log-transform on the variances/scale parameters, and a logit-transform on the weights), the proposal is $\tilde{\theta} = \theta^{(t)} + \omega \epsilon_t$, where ϵ_t is multivariate Cauchy and ω calibrated (via a burn-in step) to lead to an acceptance rate of about 0.4. For logistic models, the standard random walk proposal seems to mix faster than the slice sampler (see Altaleb, 1999, for examples) and we observed the same phenomenon for the Poisson regression. In addition, the programming effort associated with the Metropolis–Hastings solution is much lower than with the Gibbs sampler: the same programme applies for all types of regression, once the likelihood and prior are properly modified (Cappé and Robert, 2000).

2.3. Inferential difficulties

Figure 3 *(left)* presents the Metropolis–Hastings output as a projection of the MCMC sample in the (β_0, β_1) plane for the regression coefficients of the simple linear regression $y = \beta_{0i} + x\beta_{1i}$ $(i = 1, 2)$ and the ethanol dataset of Figure 2 *(left)*. The two clusters corresponding to the two components are clearly identified and lead to a good fit for the two regression lines of the model (as they should, given the clear separation between the components in the data).

Unfortunately this ideal situation is rarely encountered, and the general scenario is closer to the picture given in Figure 3 *(right)*, where the regression coefficients for the three components associated with the CO2 dataset take values in the same area of the (β_0, β_1) space. Clearly, the three components are then undistinguishable as a whole (that is, over the sequence of iterations), even though they can be separated at any given iteration.

This phenomenon is often called *label-switching* and has been known for a while, since it is follows from the lack of identifiability of mixture models (see, e.g., Celeux *et al.*, 2000). In fact, as mentioned in Section 2.1, the posterior distribution is invariant under permutation of the labels of a mixture model. This means that a given component of a mixture cannot be individually extracted from the likelihood because the component labels i cannot be distinguished from one another, unless some identifiability constraint is imposed *a priori* (see the discussion of Richardson and Green (1997) for different perspectives on identifying constraints). When the prior implies exchangeability over the parameters (β_i, p_i) or (β_i, σ_i, p_i) $(i = 1, \ldots, k)$, the symmetry of the posterior distribution in i, that is, in terms of the labelling of the parameters for a given number k of regression lines, implies that the corresponding marginal distributions of the (β_i, p_i)'s are all equal. Consequently the posterior marginal means $\mathbb{E}[\beta_i|x, y]$ are all equal and do not provide sensible estimators of the regression lines. Obviously, there are some settings where the different modes in the posterior distribution are so well separated that the MCMC sampler only visits one of the modes and label-switching does not occur, as shown in Figure 3 *(left)*. In such cases, where, *stricto sensus*, the MCMC sampler does not function correctly, the ergodic averages can provide acceptable answers, as shown by Figure 4 *(left)* for the CO2 dataset and a two component mixture modelling.

Celeux *et al.* (2000) considered this problem in the cases of normal and exponential mixture distributions. They showed that the solution based on identifiability constraints, that is, with plug-in estimates for the mixture parameters based on ordered MCMC samples, as in Richardson and Green (1997), does not always provide sensible solutions. This is also true in the case of regression, even when the MCMC samples are well separated against the components. While Figure 4 *(left)* gives the same answers for the unordered and ordered MCMC samplers, Figure 4 *(right)*, which corresponds to a simulated dataset of 185 observations from a 4 component mixture, given on the upper left hand corner of the graph, shows that the two graphs corresponding to the orderings on β_0 and on β_1 differ from the standard averages, which happen to be closer to the true regression lines. (See

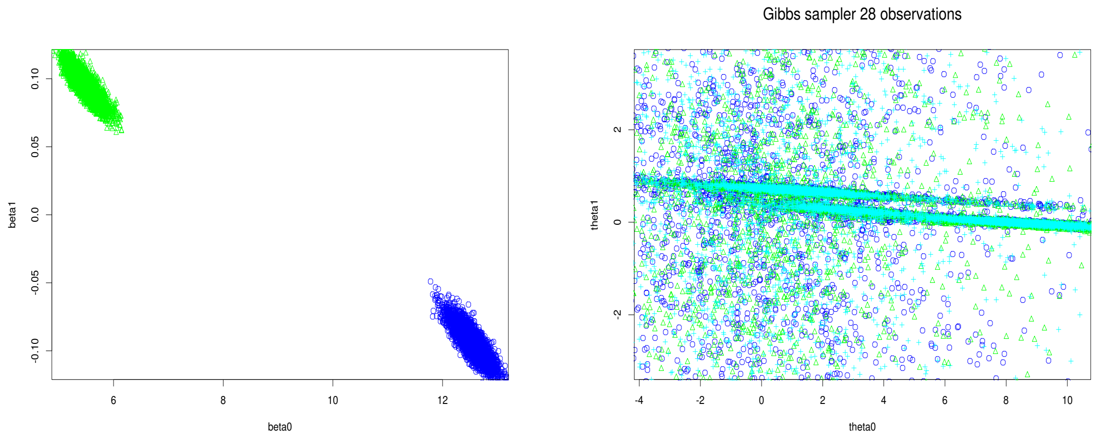

Fig. 3. *(left)* Output of the random walk Metropolis–Hastings sampler projected on the (β_0, β_1) space for the ethanol dataset; *(right)* Output of the Gibbs sampler projected on the (β_0, β_1) space for the CO2 dataset and a 3 component mixture model. *(In both graphs, the MCMC output for each component of the mixture is represented by a different symbol and colour.)*

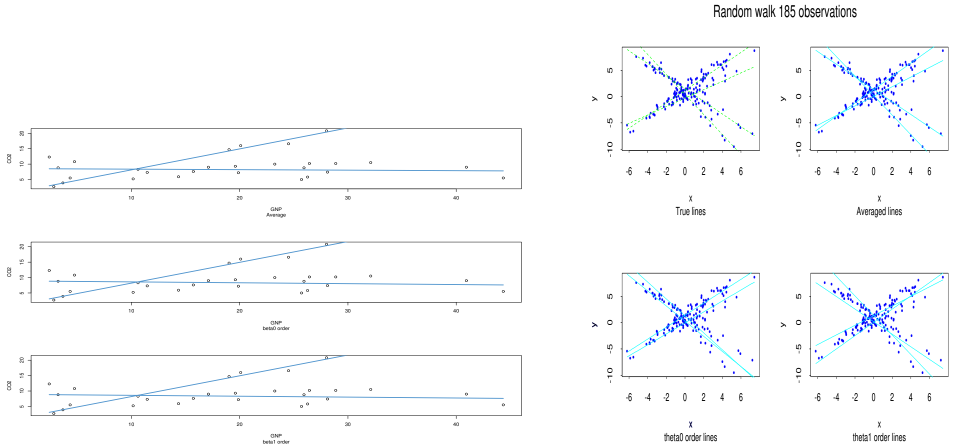

Fig. 4. *(left)* Estimated regressions lines for the CO2 dataset with a 2 component mixture model, based on a random walk Metropolis–Hastings algorithm: the leftmost graph is based on averages component by component, the central graph on averages after reordering in β_0 and the rightmost graph on averages after reordering in β_1; *(right)* same graph for a simulated dataset of 185 observations and a 4 component mixture given on the upper left corner, for a random walk Metropolis–Hastings algorithm.

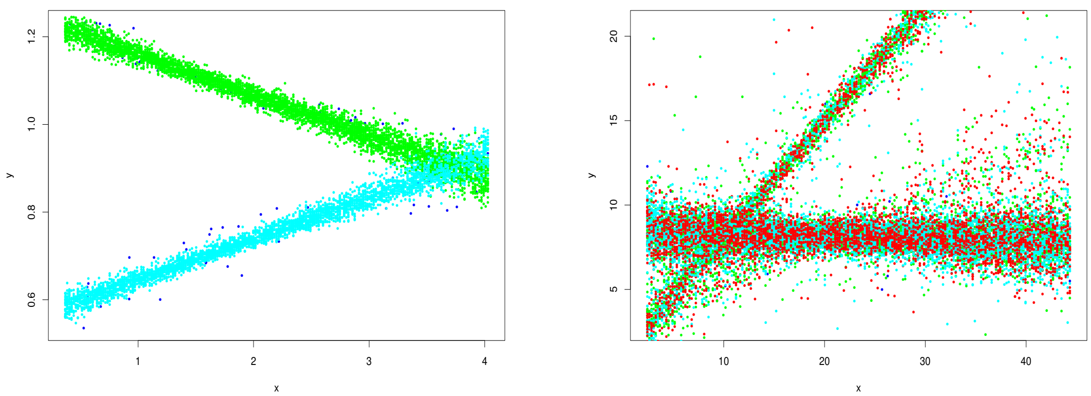

Fig. 5. *(left)* Sampling representation of the posterior distribution on the regression lines $y = \beta_{0i} + x\beta_{1i}$ $(i = 1, 2)$ for the ethanol dataset and a 2 component mixture model, obtained by selecting points at random on these lines along iterations; *(right)* Same graph for the CO2 dataset and a 3 component mixture model.

Figure 6 *(left)* for a plot of the corresponding MCMC samples.)

2.4. *Information content of the MCMC output*

The above statement of the problem does not imply that inference cannot be made in this case, since the MCMC output contains sufficient information on the regression lines and the parameters of the model. This is illustrated by Figure 5, which gives a sampling representation of the posterior distribution of the regression lines $y = \beta_{0i} + x\beta_{1i} \quad (i = 1, \ldots, k)$ for the two datasets of Section 1. This approximation is obtained the following way: at each iteration t of the MCMC sampler, five x's are selected at random in the range $[x_{\min}, x_{\max}]$ and the points

$$(x, \beta_{0i}^{(t)} + x\beta_{1i}^{(t)})$$

are represented on the graph, where the component i is chosen with probability $p_i^{(t)}$. A different colour is associated with each component i and Figure 5 *(right)* shows a strong mixing over colours, which means again that MCMC approximations to the posterior expectations $\mathbb{E}[\beta_i|x, y]$ will perform poorly. However, it also exhibits an equally strong stability in the location of the points, which means that the two main regression lines can be located almost by sight and therefore that inference is still meaningful in this setting. In this regard there is little difference between the left and right graphs in Figure 5, even though the MCMC parameter samples are as dissimilar as possible (one being extremely mixed and the other not mixing at all over the labels).

One can notice that a natural byproduct of this sampling graphical device is to provide an evaluation of the variation of the regression lines and, hence, a Monte Carlo confidence band by selecting the 95% of points closest to the lines.

The inferential difficulty thus lies in formalising the derivation of representative regressions lines from figures like Figure 3 *(right)*. The main goal of this paper is therefore to present solutions to this problem that are based on appropriate loss functions.

3. Loss Based Inference

Given the overall unreliability of the *ad hoc* average and of ordered estimates, we concentrate on the specification of loss functions $L(\xi, \hat{\xi})$ for the various inferential questions which arise in this setting. Following Celeux *et al.* (2000), these loss functions will not rely on the labeling of the components and so will not be affected by the lack of identifiability. Therefore, the corresponding Bayes estimator

$$\hat{\xi}^* = \arg\min_{\hat{\xi}} \mathbb{E}_{\xi|x,y} L(\xi, \hat{\xi}). \tag{6}$$

is equally unaffected by label indexes. (See Stephens (2000b) for a cluster-like approach in the mixture setting.)

For many choices of L, it is not be possible to find $\hat{\xi}^*$ explicitely. However, for a large class of losses, it is computationally feasible to use the two-step approach of Rue (1995): The first step is to use MCMC ergodic averages to approximate $\mathbb{E}_{\xi|x,y} L(\xi, \hat{\xi})$ for a given $\hat{\xi}$. The second step is to perform the optimisation of the estimated expected losses over $\hat{\xi}$. In order for this approach to be computationally viable, a practical restriction on the loss function $L(\xi, \hat{\xi})$ is that it is possible to separate the ξ and $\hat{\xi}$ terms in $L(\xi, \hat{\xi})$ in the sense that all evaluations of the estimated $\mathbb{E}_{\xi|x,y} L(\xi, \hat{\xi})$ with respect to different values of $\hat{\xi}$ can be performed using a single common MCMC sample from the posterior of $\xi|x$. We will restrict ourselves to loss functions for which this requirement holds, and explain how the estimate is obtained in each of the three estimation problems we consider.

3.1. Estimating the mixture of regression lines

A natural choice for $L(\xi, \hat{\xi})$ is to use a loss function which considers the difference between two density functions, one using the parameters ξ and the other the parameters $\hat{\xi}$. The set of k regression lines defines a normal mixture density at any particular x value (as illustrated in Figure 1). Given that we do not make assumptions on the distribution of the regressors x, a natural solution advocated in Dupuis and Robert (1999) is to sum the distances between the conditional distributions, $d(f(\cdot|x, \xi), f(\cdot|x, \hat{\xi}))$, over the observed x values. For the density difference itself, the symmetrised Kullback-Leibler distance can be used

$$L(\xi, \hat{\xi}) = \sum_x \int_{\mathcal{R}} \left[f_\xi(y) \log \frac{f_\xi(y)}{f_{\hat{\xi}}(y)} + f_{\hat{\xi}}(y) \log \frac{f_{\hat{\xi}}(y)}{f_\xi(y)} \right] dy\,, \tag{7}$$

where f_ξ denotes the (conditional) density of the k-component regression mixture (the regressor variables x are omitted for ease of notation). Notice that this loss function does not involve the allocation variables which only appear in the completed density used by the Gibbs sampler. In the case of the qualitative regression models (3) and (4), the only difference is that the Kullback-Leibler distance (7) is modified as

$$L(\xi, \hat{\xi}) = \sum_x \sum_y \left[f_\xi(y|x) \log \frac{f_\xi(y|x)}{f_{\hat{\xi}}(y|x)} + f_{\hat{\xi}}(y|x) \log \frac{f_{\hat{\xi}}(y)}{f_\xi(y|x)} \right]\,, \tag{8}$$

to account for the discrete nature of the observation y.

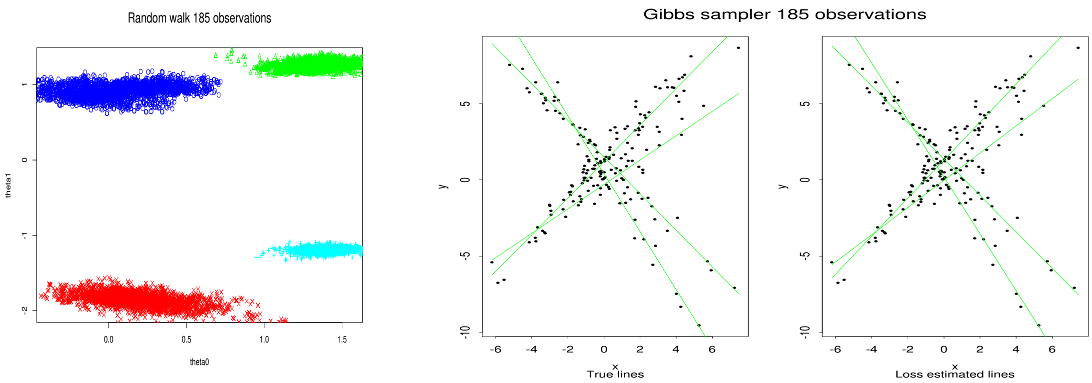

Fig. 6. *(left)* Random walk MCMC sample associated with a simulated dataset of 185 observations in the (β_0, β_1) space for a 4 component mixture model; *(right)* Corresponding estimates of the regression lines obtained by the loss function *(right)* along with the true lines *(left)* (the dataset is represented on each plot).

Notice that the choice of the Kullback-Leibler distance is defended for several reasons in Bernardo and Smith (1994), reasons ranging from information theory to asymptotics, including the fact that this loss function is invariant under reparameterisation, unlike the L_2 distance for instance. An additional incentive for chosing this distance is that it is invariant under relabelling. Moreover, it lends itself to the two-step "estimation-then-maximisation" approach in that we can decompose the posterior expected loss as follows:

$$\begin{aligned} \mathbb{E}_{\xi|x,y}\left[\sum_x \int_{\mathcal{R}} \left(f_\xi(y) \log \frac{f_\xi(y)}{f_{\hat\xi}(y)} + f_{\hat\xi}(y) \log \frac{f_{\hat\xi}(y)}{f_\xi(y)} \right) dy \right] = \\ \sum_x \int_{\mathcal{R}} \Big(\mathbb{E}_{\xi|x,y}\left[f_\xi(y) \log f_\xi(y)\right] - \log f_{\hat\xi}(y) \mathbb{E}_{\xi|x,y}\left[f_\xi(y)\right] \\ + f_{\hat\xi}(y) \log f_{\hat\xi}(y) - f_{\hat\xi}(y) \mathbb{E}_{\xi|x,y}\left[\log f_\xi(y)\right] \Big)\, dy, \end{aligned} \tag{9}$$

assuming that the order of integration may be interchanged. Considering the four terms on the right hand side of this equation individually, the first can be seen to be irrelevant for our purposes since it does not depend on $\hat\xi$. The second and fourth terms both involve posterior expectation, and we propose using the Markov chain Monte Carlo step to estimate these values on a series of y-grids (one for each of the x_i's in the summation). We can then use numerical integration for the second, third and fourth terms on these same grids of y's to evaluate the estimated expected posterior loss (less a constant independent of $\hat\xi$) for any value of $\hat\xi$, that is, each evaluation requires n numerical integrations. The optimisation step cannot be performed explicitely, and we have to resort to simulated annealing as in Celeux *et al.* (2000). The computational burden grows with the number of distinct observed x values but is feasible for all of our examples.

As a first illustration, Figure 6 depicts the MCMC sample corresponding to the simulated dataset of Figure 4 *(right)*, plus the comparison between the estimated and the true regression lines. As can be seen on the right, the fit is particularly good, but Figure 6 *(left)* shows that the MCMC clusters hardly overlap.

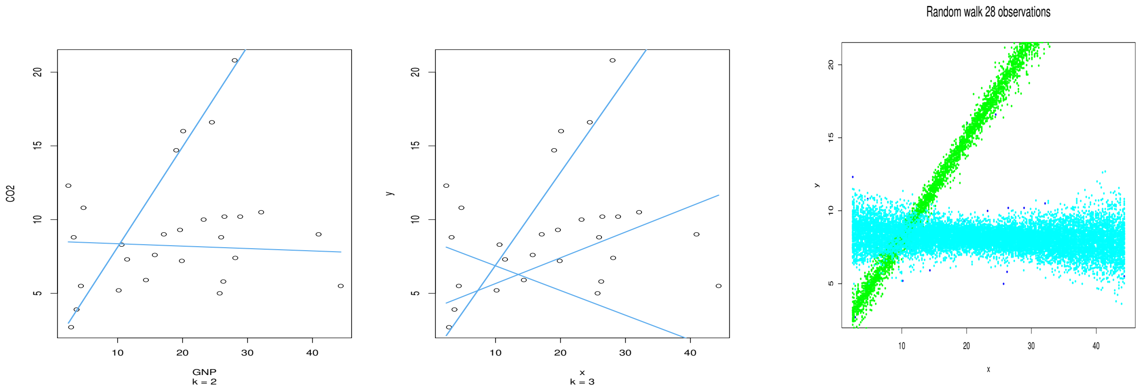

Fig. 7. *(left)* Comparisons of the loss estimates for 2 and 3 component mixtures for the CO2 dataset and a random walk MCMC sample; *(right)* Repartition of the derived sample of the regression lines for the 2 component mixture model.

Moving on to the CO2 dataset introduced in the Introduction, Figure 7 *(left)* compares the estimators obtained when fitting a 2 component and a 3 component model to the data. The 2 component model is more satisfactory, in the sense that it corresponds to the posterior spread of the regression lines in Figure 7 *(right)*, obtained by the sampling device explained in the Introduction, while two of the regression lines proposed in the 3 component case are almost identical. (Besides, the estimated weight of the third component is equal to 0.007.)

3.2. Estimating clusters

There are applications where the primary information of interest is not the regression lines, but *clusters*, that is, the groupings of allocations of observations. Here again we need a loss function for which the arbitrariness of the component labels is immaterial. Recalling that the notation $z_i = l$ denotes that observation i is allocated to line l, we suggest using a function which compares configurations z and $\hat{z}$ by noting whether pairs of observations i and j are allocated to the same component under one set of allocations but to different components under the other set:

$$L(z, \hat{z}) = \sum_{i<j} \left\{ \mathbb{I}_{[z_i=z_j]}(1 - \mathbb{I}_{[\hat{z}_i=\hat{z}_j]}) + \mathbb{I}_{[\hat{z}_i=\hat{z}_j]}(1 - \mathbb{I}_{[z_i=z_j]}) \right\} . \tag{10}$$

Since the posterior expectation of $\mathbb{I}_{[z_i=z_j]}$ is the posterior probability of this event,

$$\mathbb{E}_{z|x,y} L(z, \hat{z}) = \sum_{i<j} \left\{ P_{z|x,y}(z_i = z_j)(1 - \mathbb{I}_{[\hat{z}_i=\hat{z}_j]}) + \mathbb{I}_{[\hat{z}_i=\hat{z}_j]} P_{z|x,y}(z_i \neq z_j) \right\} . \tag{11}$$

The z_i's are automatically generated as part of the Gibbs sampler steps, but it is straightforward to add an extra Data Augmentation step to the other samplers, without jeopardizing the stability properties of both the MCMC and the z_i chains. Notice moreover that this Data Augmentation step is fairly generic and hence costless, as opposed to the Gibbs parameter step, which always involves specific programming. §

§All the `C` programs used in this paper are available upon request from the authors.

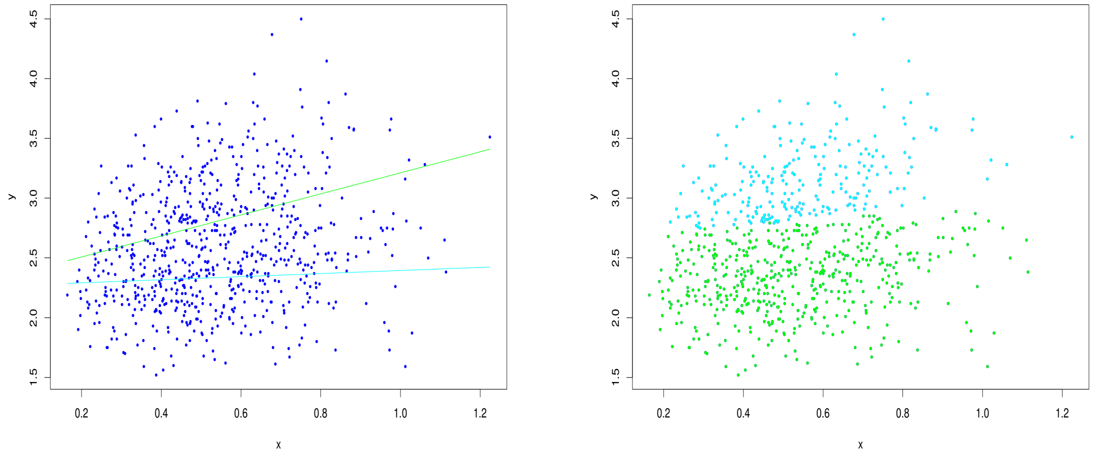

Fig. 8. *(left)* Estimated regression lines for the dataset of Fan and Chen (1999) for a 2 component model via the loss function of §3.1; *(right)* colour representation of the allocations of the observations to components 1 (blue) and 2 (green) via the loss function of §3.2.

Consider, as an illustration of both loss functions presented so far, a dataset from Fan and Chen (1999), which relates daily measurements of the nitrogen pollutant NO_2 in Hong Kong in 1994-1995 and the number of daily hospital admissions for circulation and respiration problems. For two components, the MCMC clusters are clearly separated and the posterior distribution of the regression lines gives two distinct bands. The loss analysis of the MCMC sample is in agreement with this empirical analysis: the estimated lines in Figure 8*(right)* correspond to these two bands, while the dataset is neatly separated into two groups, as shown in Figure 8*(left)*, with a division which is close to the average of the two regression lines. Notice that the estimated weights of the two components are 0.553 and 0.447, respectively. (The fact that one regression line is almost flat is quite interesting in that it suggests there are days where the NO_2 level does not directly affect hospital admissions.)

3.3. Estimating the lines plus allocated observations

A more likely scenario is that the question of interest is actually to identify the lines to which groups of observations belong. By this we mean that we wish to state which subsets of the observations are associated with a regression line with parameters β, σ^2. Notice that in this scenario, there is a redundancy in estimating both the allocations for each observation and the weights associated with each line. We therefore consider only the allocations and the parameters of the different regressions. The parameter of interest is thus η defined to be the set of k $\{\beta, \sigma^2\}$'s, together with $(z_1, \ldots z_n)$.

We again propose a form of Kullback-Leibler distance as a loss function, but now considering the allocation of observation (x_i, y_i), say $z_i = l$, the density in question at x_i is the regression with parameters $\{\beta_l, \sigma_l^2\}_l$ rather than the mixture regression. The difference between these densities is illustrated by comparing Figures 1 and 9, the former showing the

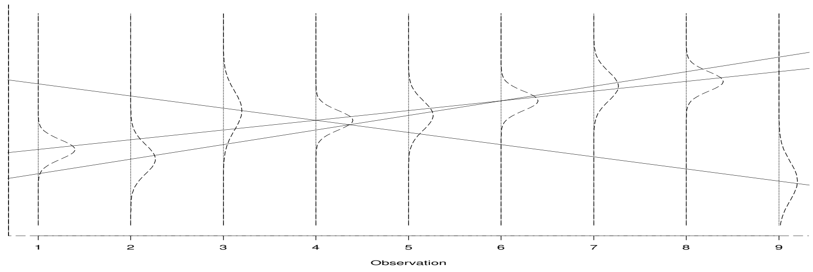

Fig. 9. The 3-component labelled regression mixture conditional distributions of $y|x, z$.

full regression mixture at each x_i, the latter the normal densities corresponding to the z_i-th regression at each labelled x_i. The loss function is therefore

$$L(\eta, \hat{\eta}) = \sum_x \int_{\mathcal{R}} \left[f_\eta(y) \log \frac{f_\eta(y)}{f_{\hat{\eta}}(y)} + f_{\hat{\eta}}(y) \log \frac{f_{\hat{\eta}}(y)}{f_\eta(y)} \right] dy \,, \tag{12}$$

where f_η now denotes the density of the regression line corresponding to the allocation of the corresponding x. In the vocabulary of the EM algorithm, we thus work with the completed likelihood rather than with the observed likelihood. We employ the same Markov chain Monte Carlo ideas as in Section 3.1 to estimate expected posterior values on a grid, and numerical integration and simulation for the optimisation step.

4. Unknown numbers of components

When the degree of heterogeneity in the data is unknown, as is the case for the CO_2 dataset of Section 1, it is unreasonable to postulate a fixed number k of components in the mixtures (2), (3) and (4), and k must be estimated as well. From a Bayesian point of view, this implies using a prior distribution on k: for illustration purposes as well as simplicity's sake, we choose a Poisson $\mathcal{P}(\lambda_1)$ distribution for this prior with $\lambda_1 = 2$ in most examples below. While this complex estimation problem has been addressed for standard mixtures through reversible jump techniques (Richardson and Green, 1997), we opt for the recent alternative proposed in Stephens (2000a) and based on birth-and-death processes. Both approaches are equally valid on theoretical grounds (since they both satisfy the required detailed balance conditions) and our choice is based solely on practical considerations; the birth-and-death process avoids the specification of well-calibrated moves, such as the splits-and-merges of Richardson and Green (1997) and the corresponding computation of the Jacobians of the one-to-one transforms, plus it allows for easier changes of both prior distributions and parameterisation. Another relevant point is that the implementation of the birth-and-death procedure is completely free from the MCMC algorithm chosen for the fixed k steps. In fact, it works even without a supplementary MCMC algorithm. This property allows very modular programming when implementing the method and thus improves portability. (Notice however that the reversible jump algorithm, when restricted to birth and death moves with birth proposals based on the prior distribution, enjoy similar properties. See Cappé *et al.*, 2001, for further connections between both approaches.)

For $v = 0, 1, \cdots, V$

$t \leftarrow v$

Run till $t > v + 1$

1. Compute $\delta_j(\beta) = \dfrac{L(y|\beta_j)}{L(y)} \dfrac{\lambda_0}{\lambda_1}$
2. $\delta(\beta) \leftarrow \sum_{j=1}^{k} \delta_j(\beta_j)$, $\mu \leftarrow \lambda_0 + \delta(\beta)$, $u \sim \mathcal{U}([0,1])$
3. $t \leftarrow t - u \log(u)$
4. With probability $\delta(\beta)/\mu$

   Remove component j with probability $\delta_j(\beta)/\delta(\beta)$

   $k \leftarrow k - 1$

   $p_\ell \leftarrow p_\ell/(1 - p_j)$ $(\ell \neq j)$

   Otherwise,

   Add component j from the prior $\pi(\beta)$

   $p_j \sim \mathcal{B}e(\gamma, k_\gamma)$

   $p_\ell \leftarrow p_\ell(1 - p_j)$ $(\ell \neq j)$

   $k \leftarrow k + 1$
5. Run I MCMC(k, β, p)

Fig. 10. Algorithmic representation of the birth-and-death process. The parameter λ_1 is the mean of the Poisson prior on k and γ is the parameter of the Dirichlet prior on p. The MCMC(k, β, p) function is a generic MCMC algorithm for a fixed k.

4.1. Birth-and-death processes

We will not go into details about the birth-and-death procedure since this is a straightforward adaptation of Stephens' (2000a) algorithm. Figure 10 gives a description of the method in an program-like manner. The method depends on a single parameter λ_0, which is the birth rate in the point process. We found in practice that better mixing was associated with higher values of λ_0, as should be expected, but this is at the expense of running more simulations between two monitoring (integer) instants. The influence of the number of iterations I where an MCMC sampler is run for the current value of k (Step 5.) is small and we chose $I = 10$ in our experiments.

The algorithms perform quite satisfactorily in our case and produces output where label switching is naturally very high (since the meaning of component 1, say, is vacuous, given that the number of components varies from one iteration to one another). The inferential issues addressed in Section 3 are thus reproduced in the current setup and we study the implementation of the above loss functions in the case of varying k.

Notice first that, from an inferential point of view, there is little difference from the fixed k setting. Indeed, the inference must be conditional on k, as argued in Robert (1997), because a general principle in Bayesian model choice is that parameters appearing in different

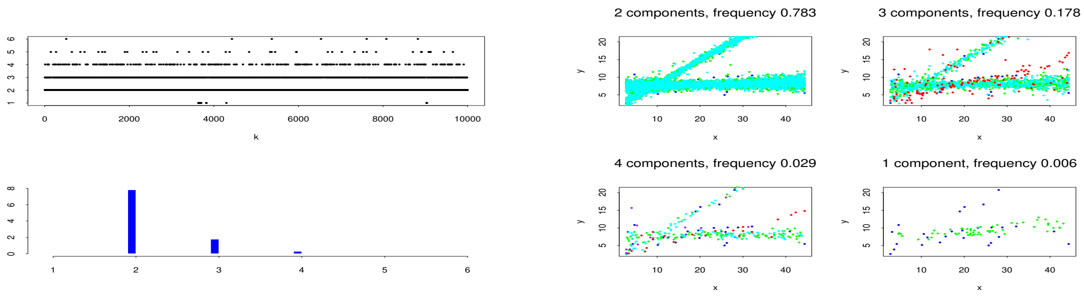

Fig. 11. *(left)* Rawplot of the successive values of k along $10,000$ consecutive iterations of the birth-and-death algorithm and corresponding likelihoods for the CO2 dataset; *(right)* representation of the posterior distributions of the regression lines for the four most likely values of k.

models must be considered as separate entities. Therefore, the analysis for each value of k is as in the fixed k case. Obviously, running the analysis for varying k also leads to inference about the value of k supported by the data, both through the posterior probabilities and through the predictive plots of the regression lines, approximated as in the fixed k case. For instance, the most probable value of k for the simulated dataset of 185 observations is $k = 4$.

4.2. Normal regression

We reanalysed the CO2 dataset, with the conclusion shown in Figure 11 that $k = 2$ is the solution most favoured by the data (using a Poisson prior on k with mean 2). It is interesting to see that the birth-and-death process creates enough mixing to ensure that k moves at a good rate. (See also how the likelihood [in *green*] closely follows the graph of the k's.) Notice that, in the case $k = 3$ there is a stable third line appearing, which is coherent with the findings of Figure 7 when k is fixed. However, the weight associated with the third regression line is estimated by 0.0069 using the loss of Section 3 and is thus negligible.

4.3. Logistic regression

A mixture of logistic regression models (3) is appropriate for the analysis of a dataset from the INSEE *(Institut National de la Statistique et des Etudes Economiques)* 1992 survey on financial assets. This dataset of 1278 individuals relates the possession of certain financial assets (SICAV), which are mixed portfolios of actions and obligations (??) proposed by banks, to some socioeconomic covariates, from which we selected the inactivity dummy variable (`1` if inactive, `0` if active), the age group (from `1` to `6`), the yearly income, and the overall fortune (both variables have been standardized). The prior on the β_{ij}'s is a normal $\mathcal{N}(0, 100)$ distribution.

Figure 12 *(right)* shows the histograms and the trace of the output of the Gibbs sampler for a fixed number of components equal to 4. It demonstrates the need for a loss function processing since the label-switching phenomenon occurs quite intensely, as demonstrated in particular by the almost identical histograms.

When k varies, the mixing over k induced by the birth-and-death algorithm is quite high, as shown by Figure 12 *(left)*, which also contains the sequence of the loglikelihoods

Table 1. Loss estimates of the logistic coefficients for the INSEE dataset produced by the procedure of §3.1 for the most probable values of k (the numbers of occurences of the corresponding k's are indicated in parentheses, out of a total number of 5000 simulations).

k	p_i	β_{0i}	β_{1i}	β_{2i}	β_{3i}
2	0.750707	4.947922	12.030713	−7.071619	4.142905
(1204)	0.249293	15.330934	−17.102980	4.523014	−1.444494
3	0.713159	6.714042	13.041261	−5.963821	3.590483
(1294)	0.214304	−5.907764	−15.258550	−1.879069	2.081183
	0.072536	1.926905	0.032500	1.811353	0.305916
4	0.510735	5.131563	13.833825	−2.484054	8.002322
(1064)	0.256141	10.304051	3.565791	2.601400	7.830500
	0.195479	−22.490153	−7.773485	−4.394919	−16.251573
	0.037574	0.768561	−24.482456	−8.485442	−1.261904
5	0.686954	5.191444	11.760222	−1.490999	4.755761
(753)	0.193014	−10.976373	−19.461148	2.193058	−6.434572
	0.113680	0.899894	0.034735	0.849712	0.256568
	0.004023	−0.843996	−1.608183	15.548180	−3.690492
	0.002288	6.275197	4.580817	−20.383082	20.850748
6	0.509334	−1.304325	13.492725	2.856363	6.403604
(423)	0.213097	−9.474540	−14.109398	−5.309507	17.080243
	0.182308	5.228880	8.426478	3.101820	−10.239990
	0.052475	5.899200	0.420056	5.975035	−1.794606
	0.024636	8.441083	−0.577892	15.301988	6.705896
	0.018165	24.436231	4.678067	−3.750072	27.574212

at the successive values of the parameters. The shapes of the posterior distributions of the β_{ij}'s, when averaging over k, are very similar to those obtained in Figure 12 *(right)* for $k = 4$. Notice also how the loglikelihood follows more accurately the shape of the k curve in the last part of the iterations, which may indicate a better fit for all values of k due to a stationarisation of the process. Table 1 provides the estimates of the logistic coefficient obtained via the loss function of §3.1. It shows a persistence of the first component for the three values of k, with a negative effect of the yearly income which can be explained as a tendency for people with higher incomes to choose more risky but more profitable financial assets. A similar analysis can be made for the second component when $k = 3$ and the third group when $k = 4$: younger active people with higher incomes are less likely to own such assets. Notice also that the last component for $k = 4$ has a fairly low probability.

4.4. *Poisson regression*

When considering the Poisson regression mixture (4), the setting is very similar to the previous paragraph. We first check the performances of the birth-and-death algorithm on a dataset analysed in Lenk (1999), which relates the monthly unemployment rate with the monthly number of accidents (in thousands) in Michigan, from 1978 to 1987. The model is thus

$$N|\varrho \sim \sum_i p_i \mathcal{P}(\exp\{\beta_{0i} + \beta_{1i}\,\log(\varrho)\})$$

where N denotes the number of accidents and ϱ the corresponding unemployment rate. Since Lenk (1999) noted a possible seasonal effect on the dependence, we postulated a $\mathcal{P}(4)$

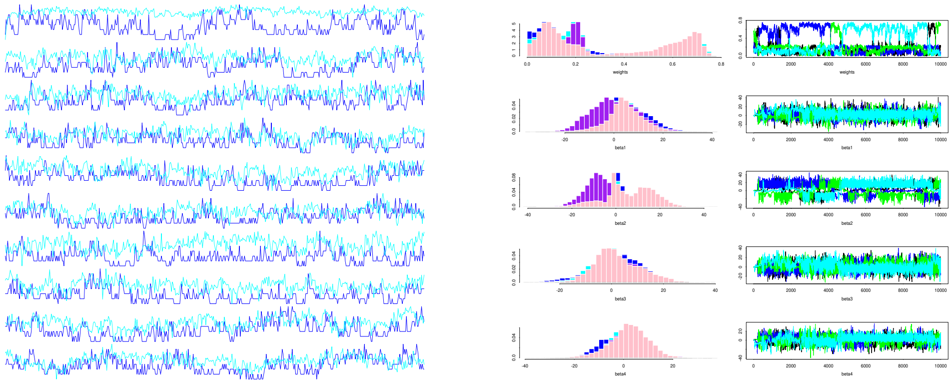

Fig. 12. *(left)* Plot of the successive values of k along $5,000$ consecutive iterations of the birth-and-death algorithm for the INSEE dataset, with in superposition the corresponding value of the log-likelihood; *(right)* histograms and series for the same sample.

prior on the number of components.

The result of this study is that the series of the $k^{(t)}$'s produced by the birth-and-death algorithm is mostly composed of 1's, with a frequency approximately of .91, and of 2's, with a frequency approximately of .09. Examination of the log-likelihood also shows that the Markov chain does not remain in the neighbourhood of a local maximum but explores (to a certain extent) the posterior surface. When $k = 1$, the standard posterior mean is acceptable as estimate and the "plug-in" estimate of the regression function provides a reasonable fit of the dataset which cannot be distinguished from the loss estimate derived from §3.1 in Figure 13. For $k = 2$, the loss estimate produces a second component with a very small weight, 0.001399, which can be ignored.

In this example, we also compared the performances of two versions of the birth-and-death algorithm, that is, when using 50 steps of the Gibbs sampler and 50 steps of the Cauchy random walk sampler at integer times. We observed that the later sampler has a higher mixing rate in terms of exploring the parameter space. This is also true for the next dataset.

We now turn to a second accident dataset, provided by INRETS *(Institut National de Recherche sur les Transports et leur Sécurité)*, which relates the number of accidents at 107 selected crossroads with the average traffic on both the primary (t_p) and the secondary (t_s) roads connecting through these crossroads (Brenac, 1994). The expectation of the Poisson distribution then writes as

$$\exp\{\beta_0 + \beta_1 \log(t_s) + \beta_2 \log(t_p)\}\,. \tag{13}$$

As in Viallefont *et al.* (2001), whose analysis uses reversible jump techniques, the most probable value of k is 3, although $k = 4$ comes as a close second. However, the MCMC sample conditional on $k = 4$ accumulates values near 0 and this case does not lead to higher values of the likelihood. Moreover, for the 3 component case, the third component corresponds to the high risk crossroads, with an increasing risk factor (or *elasticity*) β_1 due to the secondary road and a decreasing elasticity for the primary road, while, for the 4 component case, the elasticities are too high to be realistic from a traffic safety point of view.

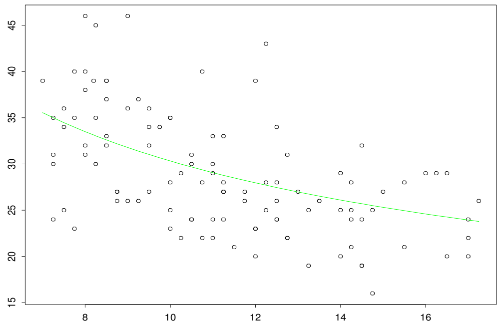

Fig. 13. Regression curve with plug-in loss estimate for the Poisson modelling of the Michigan accident dataset.

Table 2. Estimated parameters of model (13) under the loss function (12), for the two most probable values of k and the INRETS accident dataset (the numbers of occurences of the corresponding k's are indicated in parentheses, out of a total number of 5000 simulations).

Number of components	β_0	β_1	β_2
3	0.271051	0.248952	1.178467
(2530)	1.619090	0.223184	0.405672
	3.025534	0.430442	-0.154674
4	-0.283464	0.452226	1.809185
(1856)	1.019317	0.405510	0.617267
	1.425445	0.269155	2.783191
	1.550998	2.567553	-2.578583

Figure 14 *(left)* gives a contour plot of the posterior expectations of the mixtures of the averages (13), which do not depend on the labelling and is obtained by averaging over the MCMC iterations, producing the expected increase of the number of accidents in both t_p and t_s, while Figure 14 *(right)* inserts a barplot average predictive distribution on the number of accidents for each observed couple (t_s, t_p) with, logicaly, a wider spread for larger values of t_s and t_p. This figure convincingly illustrates the potential of the MCMC sample in providing various interpretations of the estimated model. Figure 15 represents the loss estimates given in Table 2 through contour lines derived from the plug-in expectations (13) and colour allocations of the 107 observations, for both $k = 3$ and $k = 4$. Notice how the different contour systems relate to the overall average of Figure 15 and also the sharp decrease at $t_s = 0$ for most components.

5. Conclusion

While Decision Theory is often criticized on a practical basis, on the ground that actual decision makers cannot build realistic loss functions, we have demonstrated in this paper

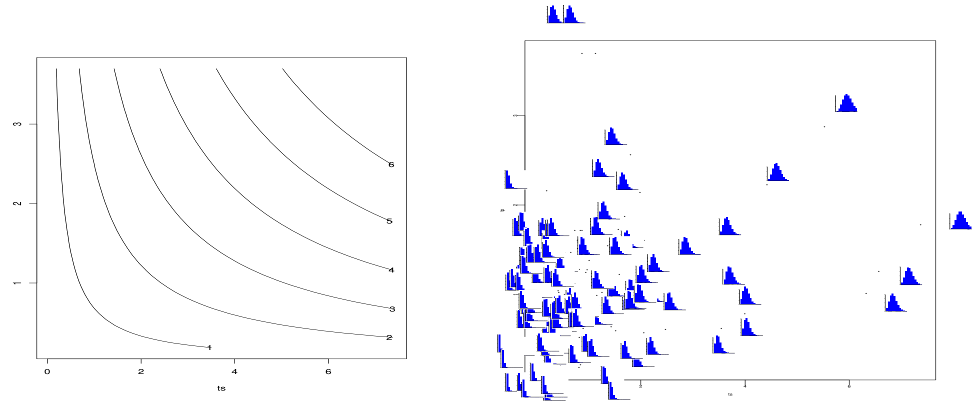

Fig. 14. *(left)* Contour plot of the posterior expectations $\exp(\beta_0 + \beta_1 \log(t_s) + \beta_2 \log(t_p))$ for the INRETS accident dataset; *(right)* corresponding barplot of the average predictive distribution for each observed crossroad (t_s, t_p).

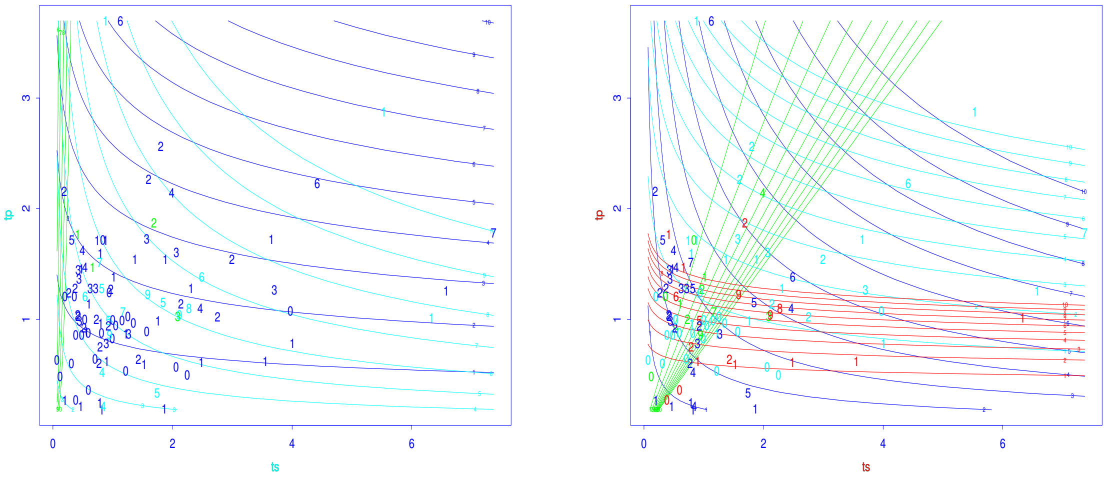

Fig. 15. *(left)* Contour representation of the estimated model conditional on $k = 3$ for the loss function of §3.3. The numbers of accidents are represented in the colour of the contours of the component they are allocated to; *(right)* same representation for $k = 4$.

that formal loss functions based on the Kullback-Leibler functional distance can produce valuable answers in the non-identifiable setting of mixtures of regressions, where standard empirical measures such as the posterior mean do not work. While the simulation steps utilised here to produce a sample from the posterior distribution are standard (even though we are not aware of previous uses of Metropolis-Hastings procedures in such settings), the computational part invoked by the minimisation of the posterior loss is more intense, but feasible.

We have also shown that inference can be drawn about the number of components in the mixture, coupling the simple procedure of Stephens (2000a) with the MCMC algorithms developped here, and that, besides, the loss derivation of estimates also applies in this variable dimension setting by conditioning on k. Examples demonstrated that the mixing behaviour of the samplers was satisfactory, even for several regressors. These various tools, as well as the graphical representations introduced in this paper, can be used for a wide variety of regression problems and of prior distributions, and this at a minimal programming cost. This study indicated in addition that the prior distribution on k is not without effect on the resulting inference: a Poisson $\mathcal{P}(1)$ prior does provide a posterior distribution on k that differs widely from the posterior associated with a Poisson $\mathcal{P}(10)$ prior. This feature is to be expected, given the ill-posed nature of the mixture models. It has, however, a limited influence on the predictive distribution, in the sense that curves such as those in Figure 15 do not vary for both priors.

Acknowledgements

The authors are thankful to the participants to the second TMR workshop in Crete, May 1999, for their comments and in particular to Arnoldo Frigessi (Oslo) for his support. The INSEE dataset was kindly provided by Denis Fougères (CREST) and the INRETS dataset by Valérie Viallefont (CREST), to whom we are extremely grateful, as well as to T. Brenac and S. Lassare (INRETS), who gave us permission to use this second dataset and commented on a previous version. Discussions with M.T. Wells (Cornell), and suggestions from the associate editor and two referees, were equally helpful.

References

Altaleb, A. (1999) Contributions à la théorie des algorithmes MCMC. Ph.D. thesis, Université de Rouen.

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

Billio, M., Monfort, A. and Robert, C.P. (1999) Bayesian estimation of switching ARMA models. *J. Econometrics* **93**, 229–255.

Bar–Shalom, Y. (1978) Tracking methods in a multi-target environment. *IEEE Trans. Automatic Control* **23**, 618–626.

Brenac, T. (1994) Accidents en carrefour sur routes nationales, modélisation du nombre d'accidents prédictible sur un carrefour et exemples d'applications. Rapport INRETS n 185, INRETS, Arcueil.

Cappé, O., and Robert, C.P. (2000) Ten years and still running! *J. Amer. Statist. Assoc.* **95**(4), 1282–1286.

Cappé, O., Robert, C.P., and Rydén, T. (2001) Discrete time and continuous time jump processes with applications to hidden Markov models. Tech. report, CREST, INSEE, Paris.

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

Damien, P., Wakefield, J. and Walker, S. (1999) Gibbs sampling for Bayesian non-conjugate and hierarchical models by using auxiliary variables. *J. Royal Statist. Soc.* (Ser. B) **61**, 331–344.

Diebolt, J., and Robert, C.P. (1994) Estimation of finite mixture distributions by Bayesian sampling. *J. Royal Statist. Soc.* (Ser. B) **56**, 363–375.

Dupuis, J., and Robert, C.P. (1999) Bayesian variable selection in qualitative models by Kullback-Leibler projections. In *Proc. Cagliari Workshop on Model Selection* (W. Racugno, Ed.), 275–305 (with discussion). CNR, Collana Atti di Congressi, Pitagora Editrice, Bologna.

Fan, J. and Chen, J. (1999) One-step local quasi-likelihood estimation. *J. Royal Statist. Soc.* (Ser. B) **61**, 927–943.

Goldfeld, S.M. and Quandt, R.E. (1976) A Markov model for switching regression. *J. Econometrics* **1**, 3–16.

Hamilton, J.D. (1989) A new approach to the economic analysis of nonstationary time series and the business cycle. *Econometrica* **57**, 357–384.

Hurvich, C.M., Simonoff, J.S., and Tsai, C.-L. (1998) Smoothing parameter selection in nonparametric regression using an improved Akaike information criterion. *J. Royal Statist. Soc.* (Ser. B) **60**, 271–294.

Lenk, P. (1999) Bayesian inference for semiparametric regression using a Fourier representation. *J. Royal Statist. Soc.* (Ser. B) **61**, 863–879.

Kiefer, N. (1980) A note on switching regression and logistic discrimination. *Econometrica* **48**. 1065–1069.

Quandt, R.E., and Ramsey, J.B. (1978) Estimating mixtures of Normal distributions and switching regressions (with discussion). *J. Amer. Statist. Assoc.* **73**, 730–752.

Richardson, S., and Green, P.J. (1997) On Bayesian analysis of mixtures with an unknown number of components (with discussion). *J. Royal Statist. Soc.* (Ser. B) **59**, 731–792.

Robert, C.P. (1996) Inference in mixture models. In *Markov Chain Monte Carlo in Practice*, W.R. Gilks, S. Richardson and D.J. Spiegelhalter (Eds.), 441–464. Chapman and Hall, London.

Robert, C.P. (1997) Discussion of "On Bayesian Analysis of Mixtures with an unknown Number of Components" by S. Richardson and P. Green. *J. Royal Statist. Soc.* (Ser. B) **59**, 758–764.

Robert, C.P. and Casella, G. (1999) *Monte Carlo Statistical Methods.* Springer-Verlag, New York.

Rue, H. (1995) New loss functions in Bayesian imaging. *J. Amer. Statist. Assoc.* **90**, 900–908.

Shuway, R.H., and Stoffer, D.S. (1991) Dynamic linear models with switching. *J. Amer. Statist. Assoc.* **81**, 763–769.

Stephens, M. (2000a) Bayesian methods for mixtures of normal distributions - An alternative to reversible jump methods. *Ann. Statist.* **28**, 40–74.

Stephens, M. (2000b) Dealing with label switching in mixture models. *J. Royal Statist. Soc.* (Ser. B) **62**, 795–810.

Viallefont, V., Richardson, S. and Green, P.J. (2001) Bayesian estimation of Poisson mixtures. *J. Nonparametric Statist.* (to appear).