Bayesian analysis of Poisson mixtures

Valérie Viallefont*
CREST et INSERM, France.

Sylvia Richardson†
INSERM, France.

Peter J. Green‡
University of Bristol, UK.

Abstract The modelling of rare events via a Poisson distribution sometimes reveals substantial over-dispersion, indicating that some unexplained discontinuity arises in the data. We suggest modelling this over-dispersion by a Poisson mixture. In a hierarchical Bayesian model, the posterior distributions of the unknown quantities in the mixture (number of components, weights, and Poisson parameters) will be estimated by MCMC algorithms, including reversible jump algorithms which permits varying the dimension of the mixture. We will focus on the difficulty of finding a weakly informative prior for the Poisson parameters : different priors will be detailed and compared. Then, the performances of different moves created for changing dimension will be investigated. The model is extended by the introduction of covariates, with homogeneous or heterogeneous effect. Simulated data sets will be designed for the different comparisons, and the model will finally be illustrated on real data.

Key words: Reversible jump, Heterogeneity, Markov Chain Monte Carlo, Poisson mixtures, Poisson distribution, Bayesian inference.

1 Introduction

Over-dispersion of rare events is a common phenomenon which is often tackled using a hierarchical mixed model formulation in which the Poisson parameters are given a distribution at a second level.

This approach has been used, for example, in disease mapping when the number of observed cases per area is small (Clayton and Bernardinelli [1]). Here, we suggest modelling

*Address for correspondence:
INSERM U.170, 16 avenue Paul Vaillant Couturier, 94807 Villejuif Cedex, France.
Email: `vialle@vjf.inserm.fr`
†INSERM U.170, 16 avenue Paul Vaillant Couturier, Villejuif, France.
Email: `richardson@vjf.inserm.fr`
‡Department of Mathematics, University of Bristol, Bristol BS8 1TW, UK.
Email: `P.J.Green@bristol.ac.uk`.

over-dispersion for rare events using Poisson mixtures with a variable "number of components", k, formalised as: $y_i \sim \sum_{j=1}^{k} w_j f(.|\lambda_j)$, where $i = 1, \ldots, n$ index the observations, w denotes the proportions (with $\sum_j w_j = 1$), and λ the vector of parameters (strictly positive) λ_j of the Poisson distribution $f(.|\lambda_j)$.

Inference in mixture models has been approached either in a classical maximum likelihood approach (McLachlan and Basford [2]) or a in Bayesian framework, coupled with MCMC techniques. Our work on Poisson mixtures belongs to this second stream and its recent developments. We are interested here in exploring mixtures with an unknown number of Poisson distributions. The reversible jump algorithm has been introduced by Green ([3]) and permits a fully Bayesian approach when the dimension of the parameter space varies. Its use for modelling mixtures with unknown number of components has been firstly developed by Richardson and Green ([4]) in the context of Gaussian mixtures. It has also been used by Gruet et al. for exponential mixtures([5]), and for the analysis of a Poisson mixture in finance by Dellaportas et al. ([6]).

The model will be made explicit in Section 2, where will also be detailed different prior distributions for the Poisson parameters. Their influence on posterior results will be shown in Section 3. We will describe and compare different ways of implementing reversible jump moves in Section 4. In section 5, the model is extended by the introduction of covariates, before being applied to a real data set in Section 6.

2 Hierarchical Poisson mixture models and algorithm

2.1 Model

A mixture of k Poisson distributions, $y_i \sim \sum_{j=1}^{k} w_j f(.|\lambda_j)$, can be equivalently formulated in a hierarchical manner using a latent variable z_i representing the allocation of each observation y_i to one of the components. The mixture model can thus be written:

$$f(y_i|\lambda, z_i = j) = \text{Poisson}(\lambda_j) \quad \text{with } P(z_i = j) = w_j. \tag{1}$$

This is the lower level of a hierarchical mixture model. Since we place ourselves in a Bayesian framework, the higher levels deal with priors. Prior distributions of the parameters and hyperparameters are chosen with the view of being as uninformative as possible.

- For the numbers of components, k, we choose a Uniform $\{1, \ldots, \eta\}$ as a prior distribution, where η is a fixed constant representing the maximal number of components allowed in the mixture. This is a convenient choice, the value of η being chosen large enough in our experiments so that it does not represent a constraint.

- For the weights w, we follow the classical choice of a Dirichlet prior with a single parameter δ. This gives symmetric priors for each w_j which is a way of being uninformative, the choice $\delta = 1$ corresponding to uniformity over the simplex.

- In order to ensure identifiability, we impose an ordering constraint on the λ's: $\lambda_1 < \lambda_2 < \ldots < \lambda_k$. Subject to this, the Poisson parameters $\{\lambda_j\}_{j=1,\ldots,k}$ will be distributed a priori either as independent variables or jointly modelled. We generically denote this prior by $p(\lambda|\theta, k)$, with hyperparameter θ, the precise choice of which will be detailed in Sections 2.2 and 2.3.

Here and throughout the paper we use $p(.|.)$ to denote a generic conditional distribution.
In summary, the prior structure is specified by:

$$\begin{aligned} k &\sim \text{Uniform}\{1,\ldots,\eta\}, \eta = 30 \\ w &\sim \text{Dirichlet}(\delta,\delta,\ldots,\delta), \delta = 1 \\ \lambda &\sim p(.|\theta,k),\ \lambda_1 < \lambda_2 < \ldots < \lambda_k \end{aligned} \tag{2}$$

The graphical structure of the Poisson mixture is represented in Figure 1 as a Directed Acyclic Graph (DAG).

[Figure 1 about here.]

Using the additional conditional independence assumptions which are represented in the graph, the joint distribution of all the variables is thus given by:

$$p(k|\eta)p(w|\delta,k)p(z|k,w)p(\lambda|\theta,k)p(y|\lambda,z). \tag{3}$$

MCMC algorithms will be used to simulate from the posterior joint distribution of the parameters.

2.2 Prior model with independent λ's

In this case, we assume that the λ_js are independent conditional on θ, each with prior density $p(\lambda_j|\theta)$, and ordered. The prior joint density for λ is then given by :

$$p(\lambda|\theta,k) = k!\ p(\lambda_1|\theta)\ldots p(\lambda_k|\theta)\mathtt{I}_{\lambda_1\le\lambda_2\le\ldots\le\lambda_k}(\lambda),$$

where $\mathtt{I}$ denotes the indicator function, and where the $k!$ term arises from this ordering constraint indicator.

We will consider two types of distributions for the λ_js with respective hyperparameter $\theta = (\alpha,\beta)$ and $\theta = (A,B)$:

$$\lambda_j \sim \text{Gamma}(\alpha,\beta) \text{ or } \lambda_j \sim \text{Uniform}[A,B].$$

Due to its being conjugate to the Poisson distribution, the Gamma distribution is often chosen as a prior for Poisson parameters. In this case, the full conditional for λ_j (i.e. its conditional distribution given all the other parameters) is a truncated Gamma distribution, namely:

$$\lambda_j|\ldots \sim \text{Gamma}(\alpha + \sum_{i|z_i=j} y_i, \beta + n_j)\mathtt{I}_{[\lambda_{j-1},\lambda_{j+1}]}(\lambda_j) \text{ with } n_j = \#\{i|z_i = j\}.$$

This conditional distribution is easily simulated by using a Metropolis algorithm with a $\text{Gamma}(\alpha + \sum_{i|z_i=j} y_i, \beta + n_j)$ proposal, rejecting if the constraint is not satisfied.

One standard way of setting a $\text{Gamma}(\alpha,\beta)$ as a weakly informative prior is to choose small values for its two parameters (such as 0.01 for example, as chosen by Dellaportas et al. [6]). However, such a distribution has a peak in the neighbourhood of zero and this might encourage the inclusion of components with very small Poisson parameters, which would be hard to identify in general. We are thus confronted with the difficulty of finding a weakly informative prior in our case. Indeed, we found in our experiments that changing the values of α and β lead to different posterior distributions, for λ, w and k.

Central to this difficulty is the fact that λ is both the mean and the variance of the Poisson distribution. There is no equivalent to the approach of defining weakly informative priors in normal mixtures, adopted by Richardson and Green ([4]), through separating the prior for the component's location (a Gaussian related to the range) from that for the component's variance (assumed to be an inverse Gamma with fixed α and random β). In fact, we found that using an inverse Gamma with random β leads to overfitting by increasing the number of components, hence, the default option that we adopt here - the sensitivity of which will be explored on a simulated three-components mixture in Section 3.1 - consists of a Gamma(α, β) whose parameters are: α greater than 1 (in order to avoid the exponential shape without reducing too much the coefficient of variation of the Gamma distribution), say 1.1, and β set such that the prior mean, α/β, is equal to the midrange of the observed data.

We also consider another form of prior for the λ's, that of a Uniform distribution on the data range. The full conditional for each λ_j is in this case:

$$p(\lambda_j|\ldots) \propto e^{-n_j\lambda_j}\lambda_j^{\sum_{i|z_i=j} y_i}\,\mathrm{I}_{[A,B]}(\lambda_j)\mathrm{I}_{[\lambda_{j-1},\lambda_{j+1}]}(\lambda_j).$$

2.3 Prior model with dependent λ's

This form of prior for the mixture parameters has been introduced by Roeder and Wasserman ([7]) for normal mixtures. It gives an explicit way of placing an informative prior on the distance between two Poisson parameters for two components that are consecutive in terms of their values of λ_j. Here the hyperparameter θ is equal to a positive constant M and the prior model for the λs is given by:

$$p(\lambda|M,k) = p(\lambda_1)p(\lambda_2|\lambda_1, M)p(\lambda_3|\lambda_2, M)\ldots p(\lambda_k|\lambda_{k-1}, M)$$

where $p(\lambda_j|\lambda_{j-1}, M)$ is a Normal centered at λ_{j-1} and with variance M^2, truncated to be greater than λ_{j-1} (denoted $\mathrm{N}_+(\lambda_{j-1}, M^2)$), and $p(\lambda_1) \propto 1$, with λ_1 positive.

The full conditional for λ_j is then:

$$p(\lambda_j|\ldots) \propto e^{-n_j\lambda_j}\lambda_j^{\sum_{i|z_i=j} y_i}\,e^{-\frac{(\lambda_j-\lambda_{j-1})^2+(\lambda_j-\lambda_{j+1})^2}{2M^2}}\,\mathrm{I}_{[\lambda_{j-1},\lambda_{j+1}]}(\lambda_j).$$

One important advantage of this type of model is that the hyperparameter M, which controls the distance between two components (*i.e.* the distance between their means), as well as their prior variability, is explicit. This parameter M is of central importance for the posterior distribution of the number of components. For example, if M is very small compared to the anticipated distance between two consecutive λs, there will be a tendency to fit intermediate components between the "true" ones; and hence to find a posterior distribution favouring higher k's. One drawback of this formulation could be the asymmetric role that it gives to λ_1. We have not tested whether this choice is influential, in comparison, say, to initialising the separation by λ_k. Other "separating" joint priors could be defined in the spirit of those used by Green ([3]) for change point analysis.

Roeder and Wasserman's advice concerning the value of M is to choose $M = 5$, which will be one of the choice investigated here. In our sensitivity analysis, we will also choose $M = 10$ and $M = 1$.

2.4 General MCMC algorithm

MCMC implementation consists of making a series of moves which update each set of parameters and preserve detailed balance. Altogether these moves constitute a sweep of the algorithm, *i.e.* one transition of the Markov chain. The main steps are similar to the ones used in the normal mixture case ([4]):

(a) update of k (if accepted, "jump" from one space of parameters to another one, of different dimension) and of the corresponding weights, means and allocations

(b) update of the weights

(c) update of the Poisson parameters

(d) update of the allocations

Moves (b), (c) and (d) are standard, performed by using the Gibbs sampler, or a Metropolis random walk for the case of independent λs with Uniform priors.

Updating k is the most delicate step, and in a separate section, we shall investigate two alternative approaches, respectively denoted "birth and death" or "split and merge", and several variants of the latter approach. These will be detailed in Section 4.

Whatever the proposal, a Metropolis acceptance ratio concerning all simultaneous changes of k, λ, w (their dimension changes) and the allocations z is computed as $\min(1, A)$, where A includes terms coming from joint distribution, proposal ratio and Jacobian of the transformation used in the proposal (see Green [3] for a generic description of reversible jump algorithms, and the Appendix for the formulas in the Poisson mixture case).

3 Statistical performance and comparisons on two simulated data sets

We first examine the performance of the different prior models for the Poisson parameters in terms of influence on the number of components and on the parameters.

3.1 Comparison of the different models

The influence of the different prior models for the λ's will be studied on simulated mixtures of two and three components:
$y_i \sim 0.5$ Poisson(2) + 0.5 Poisson(8) which is a straightforward case of a well separated mixture, with $n = 600$ observations non-randomly allocated.
$y_i \sim 0.25$ Poisson(0.5) + 0.5 Poisson(2) + 0.25 Poisson(5) , which is a more challenging case, with $n = 400$ observations non-randomly allocated.

All results shown correspond to 200000 MCMC sweeps, after 100000 for burn-in.

They correspond to the implementation of a particular move proposal called SM2 and detailed in Section 4.2.

In the case of a two-component mixture, the observed data were in the interval $[0, 17]$, leading us to consider as prior for λ either a Uniform distribution on this interval or a Gamma with parameters α=1.1 and β=0.13.

In terms of the posterior distribution of k, results (Figure 2) are very similar with both the Uniform and the default Gamma options; $k = 2$ corresponds to the mode of $p(k|y)$, but the posterior distribution of k is fairly spread out. Note that the model with dependent λ's shows a clearer peak at $k = 2$, with a posterior distribution of k less spread out. In Table 1 are indicated for $k = 2$ the posterior mean of the mixture parameter, as well as the posterior mean of the deviance which can be informally viewed as a measure of fit (Spiegelhalter [8]). Precisely, at each sweep of the algorithm, values of k, λ's and z are produced, for which a posterior deviance can be calculated. Averaging over the sweeps for each value of k gives a mean posterior deviance which is indicated in Table 1 for $k = 2$.

[Table 1 about here.]

As we can see there is no difference in the mixture parameter estimates arising for the three prior models which all recover satisfactorily the values of the simulated parameters. There is no difference either in the mean posterior deviance.

[Figure 2 about here.]

For the three-component mixture, observed data were in the interval [0,13], thus our standard prior options for the Poisson independent parameters were either a Gamma(1.1, 0.17) or a Uniform on $[0, 13]$. The results given in Table 2 and Figure 3 allow comparisons between different variants for the two independent priors for λ and the dependent one.

[Table 2 about here.]

Identifying the simulated number of components is more difficult in this case. With independent priors for λs, the posterior mode for k is either 6 for the Gamma prior or 5 for the Uniform prior and neither give a satisfactory picture of $p(k|y)$. Using the dependent prior for λ leads a clearer results with a posterior for k less spread out, and a much higher posterior probability that $k = 3$ when $M = 5$.

It is interesting to note that in contrast to $p(k|y)$, the value of the mean posterior deviance gives interpretable indications on the number of components. Indeed, with all prior models, the deviance is substantially reduced between $k = 2$ and $k = 3$ and does not change more than marginally when k is larger than 3.

[Figure 3 about here.]

The last plots of Figures 2 and 3 show the fit of the simulated mixture by its predicted distribution when using a Uniform prior for λ. These were not changed when we used either of the prior options. The good coincidence between solid and dotted lines indicates that the fit observed conditionally on $k = 2$ in Figure 2 and $k = 3$ in Figure 3 (dotted lines in the two Figures) is not changed when increasing the number of components (the overall fit is represented by the solid line).

Regarding the Gamma prior case, we have so far fixed $\alpha = 1.1$, so as to avoid the exponential shape. This might be regarded as arbitrary and, in order to justify this choice, we have also performed the analysis with a Gamma(0.5, β) and a Gamma(0.05, β) priors where the βs are still calculated so that α/β is equal to the midrange of the data, giving respectively, $\beta = 0.077$ and $\beta = 0.0077$. Results are shown in Table 2 and Figure 3: the main changes concern the posterior distribution of k (maximal for $k = 3$ or $k = 4$ when $\alpha = 0.05$), and the values of λ : small values are preferred when α is decreased, leading to a change in the weights. The deviance is very slightly reduced compared to the one found with $\alpha = 1.1$, revealing that an alternative mixture with smaller parameters could also fit the simulated mixture.

When using the dependent prior for λ, the results shown for $M = 1$ and $M = 10$ clearly illustrate how much the posterior distribution of k is depending on M, with an overfitting of many components when M is small.

This section illustrates the difficulty of choosing a prior distribution for the Poisson parameters, and the clear influence of the hyperparameters of this distribution on both the posterior distributions of the mixture parameters and that of k. This difficulty also occurred in the Normal mixture case (see Richardson and Green [4] and the following discussion), in particular concerning the choice for the precision κ of the normal prior distribution for the means. We have found, in our context of Poisson mixtures, that it is more difficult to be weakly informative than for normal mixtures, using either a dependent or independent prior distribution for λ.

In the model with dependent prior for λ, the parameter M has the advantage of an easy interpretation and we would favour this prior model.

4 Different jump proposals

As with any MCMC algorithm, there is a considerable freedom in the design of the moves. Here, we are concerned with investigating the performance of different ways of updating the number of components. Indeed, poorly designed jump moves could lead to a deterioration of the convergence of the algorithm. In the normal mixture case, Richardson and Green [4] defined updating mechanisms based on the intuitive idea of exploiting the current state of the chain, *i.e.* proposing only <u>local</u> changes to few components. Specifically, moves consisting of either splitting of one component or merging of two components, while preserving some characteristic of their overall shape, are designed. Here we want to compare this approach to a more "blind" one which does not use the current state of the chain and simply proposes the birth of a new components from the prior. Furthermore, within the family of split and merge moves, we also want to examine the possible influence on the mixing of different ways of defining the split and merge mechanism in this particular case of parameters that have to satisfy a positivity constraint (this is an obvious difference with the normal mixture case).

4.1 Jumps via birth and death process

This "blind" update of the number of components proceeds by simply proposing either to delete a component at random or to create a new component with parameters (λ_* and w_*)

sampled from the priors.

Then all the observations are reallocated among the $k+1$ or $k-1$ components, the weights are normalised and the vector of λs ordered. Finally an acceptance ratio is computed to accept this change or not (see Appendix).

This process will be referred to as "birth and death" move, and labelled "BD".

4.2 Split and merge jumps

In the spirit of Richardson and Green [4], we will use the matching of zero and first order moments when proposing to split one component with parameters (w_j, λ_j) into two new ones with parameters (w_{j_1}, λ_{j_1}) and (w_{j_2}, λ_{j_2}). Hence, the proposed parameters will always satisfy:

$$w_{j_1} + w_{j_2} = w_j,$$

$$w_{j_1}\lambda_{j_1} + w_{j_2}\lambda_{j_2} = w_j\lambda_j,$$

and respect the ordering constraint $\lambda_{j-1} < \lambda_{j_1} < \lambda_{j_2} < \lambda_{j+1}$, where λ_{j-1} and λ_{j+1} are respectively the means of the previous and next components to the j^{th} one.

Reversibly, we will propose to combine two consecutive components.

We will now specify three ways of proposing a split, which are based on different intuitions of how to enforce the positivity constraint on Poisson parameters.

(1) Adding a λ_j dependent quantity is a first way for splitting a component with parameter λ_j into two components with parameters λ_{j_1} and λ_{j_2}. This is labelled "SM1" in the following. We will first generate two variables:
$u_1 \sim Beta(2,2)$ and $u_2 \sim Exp(2)$, then define:

$$\begin{cases} w_{j_1} = w_j u_1 \\ w_{j_2} = w_j(1-u_1) \\ \lambda_{j_1} = \lambda_j - \lambda_j^{1/2} w_{j_1} u_2/u_1 \\ \lambda_{j_2} = \lambda_j + \lambda_j^{1/2} w_{j_1} u_2/(1-u_1) \end{cases}$$

If λ_{j_1} is not positive or if the ordering constraint $(\lambda_{j-1} < \lambda_{j_1} < \lambda_{j_2} < \lambda_{j+1})$ is not respected, we reject this move. Otherwise, we calculate a Metropolis acceptance ratio.

(2) Adding a "neighbours of λ_j" dependent quantity : this second way of splitting the j^{th} component satisfies automatically the positivity and order constraint: $\lambda_{j-1} < \lambda_{j_1} < \lambda_{j_2} < \lambda_{j+1}$. This is labelled "SM2" in the following. Again, we use two auxiliary variables : $u_1 \sim U(0,1), u_2 \sim U(0,1)$, then define:

$$\begin{cases} w_{j_1} = w_j u_1 \\ w_{j_2} = w_j(1-u_1) \\ \lambda_{j_1} = \lambda_j - c u_2(1-u_1) \\ \lambda_{j_2} = \lambda_j + c u_2 u_1 \end{cases}$$

where

$$c = \begin{cases} \min\{(\lambda_j - \lambda_{j-1})/(1-u_1), (\lambda_{j+1} - \lambda_j)/u_1\} & \text{if } 1 < j < k \\ \min\{(\lambda_1/(1-u_1), (\lambda_2 - \lambda_1)/u_1\} & \text{if } 1 = j < k \\ (\lambda_j - \lambda_{j-1})/(1-u_1) & \text{if } 1 < j = k, \text{ and} \\ \lambda_1/(1-u_1) & \text{if } 1 = j = k. \end{cases}$$

Note that c is always positive.

(3) Multiplying λ_j by a variable u_1 with values in $[0, 1]$ is a third way of defining $\lambda_{j_1} < \lambda_j$, labelled "SM3". Here again, we need two auxiliary variables :
$u_1 \sim Beta(2, 2)$ and $u_2 = u_1 * Beta(2, 2/10)$, so that $\lambda_{j_2} > \lambda_{j_1}$. We thus define:

$$\begin{cases} w_{j_1} = w_j u_1 \\ w_{j_2} = w_j(1-u_1) \\ \lambda_{j_1} = \lambda_j u_2/u_1 \\ \lambda_{j_2} = \lambda_j(1-u_2)/(1-u_1) \end{cases}$$

If the ordering constraint $(\lambda_{j-1} < \lambda_{j_1} < \lambda_{j_2} < \lambda_{j+1})$ is not respected, we reject this move.

The Metropolis acceptance ratio for the three split and merge moves defined above is given in the Appendix.

Note that, in the split and merge moves, the ordering constraint is used not only to identify the components, but also to define a move. This was not true in the birth and death process, where the created component has parameters drawn from their prior independently to the order, and is ordered afterwards. The influence of the ordering constraint on the MCMC performance has been discussed in the normal mixture context (see the discussion of the Richardson and Green paper [4]) and alternative approaches which avoid labelling the components have been proposed, the consequent non-identifiability leading to several equivalent posterior modes (Stephens [9], Celeux et al. [10]). A simulated tempering algorithm has been proposed to favour a visit of all the modes by Celeux, Hurn and Robert [10]. In our case, the labelling is natural as it corresponds to the unique parameter; for simplicity we did not carry out the procedure proposed by Stephens ([9]).

4.3 Comparisons of different algorithms for changing the number of components

Here, we are interested in comparing the MCMC performance of our different implementations of the jump moves for changing k. For a overview of different methods to control the convergence of a MCMC algorithm, see Robert et al. ([11]).

We will use the three-component mixture for the comparisons. The different jump moves are compared in two graphical manners: Figure 4 shows the cumulated frequencies for the posterior distribution of k. We see a similar pattern for the four moves implemented, with stability after 50000 sweeps, reflecting satisfactory convergence. A close visual inspection seems to show that the SM2 move has the shortest burn-in.

[Figure 4 about here.]

Figure 5 illustrates the successive allocation of each observations to one or the other of the components : the three subgroups are correctly identified no difference between the four implemented moves.

[Figure 5 about here.]

The acceptance rate for the dimension change was about 25% for both the BD and SM1 moves and around 17% and 40% for SM2 and SM3 respectively, when a Uniform prior was used for λ.

We have also compared the four jump moves by quantifying their efficiency in estimating expectations of certain functions of the realised chain by ergodic averages. It is well-known (see, for example, Sokal [12]) that the asymptotic variance of such an average in a stationary process exceeds that for independent sampling by a factor

$$\tau = \sum_{l=-\infty}^{\infty} \rho_l,$$

known as the integrated autocorrelation time. This quantity may be estimated from the realised chain by various methods – we use Sokal's adaptive truncated periodogram estimator, recommended and adapted by Besag and Green ([13]). This was computed for samplers using each jump move, for two different outputs of the chain, the number of components k and the deviance; Tables 3 and 4 summarise the results.

[Table 3 about here.]

[Table 4 about here.]

What is first noticeable from the graphs and the tables is that the performance of the "blind" birth and death move is not markedly worse than the others. Moreover, there is little difference between all the moves, the SM2 move corresponding in most cases to the lowest integrated autocorrelation time. In terms of CPU time, the BD move is as expected the slowest since it involves updating all the allocations which is a time consuming calculation.

5 Using covariate information

5.1 Model with covariates

Hierarchical Poisson models with covariates are commonly used. In epidemiology, for example, one considers typically the number y_i of observed cases in area i for some health outcome and its relationship to the demographic structure of the population, as well as to potential risk factors.

The simplest model involves only the expected number of cases E_i for each unit (calculated on the basis of the age $\times$ sex distribution of area i) and no additional parameter. In this case the mixture model becomes:

$$f(y_i|\lambda, z_i = j, E_i) = \text{Poisson}(\lambda_j E_i) \quad \text{with } P(z_i = j) = w_j. \tag{4}$$

More generally, we might want to take into account a number of covariates $X_1, X_2, \ldots, X_p$. In such case, we have to estimate regression parameters, denoted generically by γ, that measure the effect of each covariate. The joint distribution of the hierarchical model which generalises (3) is given by:

$$p(k|\eta)p(w|\delta,k)p(z|k,w)p(\lambda|\theta,k)p(\gamma|\sigma)p(y|\lambda,z,E,X,\gamma), \tag{5}$$

where γ is either a vector or a matrix of regression coefficients, and σ denotes the hyperparameters of the prior distribution of γ.

5.1.1 Homogeneous covariate effect

Let us first consider the case where each covariate X_l is supposed to have a similar effect among all the mixture components. The mixture model with p covariates can be written as:

$$f(y_i|z_i = j, \lambda, E, \gamma, X) = \text{Poisson}(\lambda_j E_i e^{\sum_{l=1}^{p} \gamma_l X_{il}}), \text{ with } P(z_i = j) = w_j; \tag{6}$$

where γ is a p-vector whose elements γ_l are supposed to be independent. Precisely, the prior distribution for each element of γ is chosen as a normal $\mathrm{N}(0, \sigma^2)$, independently for each l. Note that λ_j might be expressed as $e^{\log(\lambda_j)}$, where $\log(\lambda_j)$ represents the intercept in the Poisson loglinear regression. The model (6) is then a mixture of Poisson loglinear regressions, where the components have different intercepts. In the context of mixed models, Watier, Richardson and Green ([14]) have explored the case of random intercept modelled by a mixture of Gaussians.

5.1.2 Heterogeneous covariate effect

We suppose now that a discontinuity can arise in the covariate effect, *i.e.* that the effect of the covariates is different among the components.

The set of Poisson regression coefficients, γ, becomes a $l \times k$-matrix, and the generalised model is now written as:

$$f(y_i|z_i = j, \lambda, E, \gamma, X) = \text{Poisson}(\lambda_j E_i e^{\sum_l \gamma_{jl} X_{il}}), \text{ with } P(z_i = j) = w_j. \tag{7}$$

Here again, the prior for each coefficient γ_{jl} will be a centred normal with variance σ^2, independently for each (j, l).

5.2 Modifications in the algorithm

At each sweep, the coefficients will be updated by using a random walk Metropolis, for each element of γ consecutively.

In the case of heterogeneous coefficients, creating a new component involves the simultaneous increase of the dimension of vectors w, λ and also γ. If the birth and death process was used to change the dimension of the parameter space, then the new component will have parameters $(w_*, \lambda_*, \gamma_*)$ where elements of γ_* could be proposed from their prior distribution, together with a new vector of allocation z'. Here, we choose to use the split and merge moves : when component j is split, the vector γ_j is transformed and for each l in $1, \ldots, p$, we propose to change γ_{jl} into $(\gamma_{j_1 l}, \gamma_{j_2 l})$ so that :

$$\gamma_{j_1 l} = \gamma_{jl} + v_l \gamma_{jl}$$

$$\gamma_{j_2 l} = \gamma_{jl} - v_l \gamma_{jl}$$

where each v_l is generated from a centred normal with variance σ_v^2. Reversibly, for the merge move, we calculate γ_{jl} as the mean of $(\gamma_{j_1 l}, \gamma_{j_2 l})$, for each l in $1, \ldots, p$ (see the Appendix for the extra terms introduced in the Metropolis acceptance ratio).

5.3 Simulated examples

We simulated two types of two-component Poisson mixture with covariates, and analysed those by using each of the two models : heterogeneous (7) or homogeneous (6). This is justified since, in general, we would have no idea whether the covariates have an effect or not, nor whether this potential effect is the same among the subgroups.

Data set I has been simulated as a two-component Poisson mixture, under the heterogeneous model (7). Data set II is a two-component Poisson mixture, with homogeneous covariate effect (model (6)).

The simulation regression parameters are:

Data set I		Data set II	
Component 1	Component 2	Component 1	Component 2
$w_1 = 0.4$	$w_2 = 0.6$	$w_1 = 0.4$	$w_2 = 0.6$
$\lambda_1 = 1.2$	$\lambda_2 = 4.3$	$\lambda_1 = 1.2$	$\lambda_2 = 4.3$
$\gamma_{1,1} = 0.2$	$\gamma_{2,1} = 0.8$	$\gamma_1 = 0.2$	
$\gamma_{1,2} = 0.5$	$\gamma_{2,2} = -0.3$	$\gamma_2 = 0.5$	

We generated 200 Poisson counts by first simulating a 200×2-matrix X with independent standard normal elements, representing the independent X_1 and X_2 covariates, and secondly a 200-vector y where

$$f(y_i | z_i = j, X_{1i}, X_{2i}, \lambda, \gamma) = \text{Poisson}(\lambda_j \ \exp(\gamma_{j,1} X_{1i} + \gamma_{j,2} X_{2i})),$$

with $\gamma_{j,l} = \gamma_l \ \forall j$, for $1 \leq l \leq 2$, in the homogeneous model.

We fixed:

$$z_i = \begin{cases} 1 & \text{if } 1 \leq i \leq 80 \\ 2 & \text{if } 81 \leq i \leq 200. \end{cases}$$

The reversible jump algorithm implemented used the SM2 move presented in Section 4.2.

Posterior means for w, λ, γ, obtained on 200000 MCMC sweeps after 50000 for burn-in, are reported in Table 5 for the two analysis models corresponding to homogeneous or heterogeneous effects of the covariates. Thus, the diagonal elements in Table 5 correspond to the analysis model which is compatible with the true model generating the data, while the off-diagonal elements correspond to different cases of misspecification of the model.

In all the analyses, the prior for γ_l or $\gamma_{j,l}$ is a centred Normal with $\sigma^2 = 25$, and we used as prior structure for the λs the model with independent λs defined in Section 2.2, whose prior is a Uniform on the range of the data. Fits of predicted values are shown in Figure 7.

[Table 5 about here.]

In the case of the latent existence of two well separated groups, like data set I and II are drawn from, the posterior distribution of k exhibits a clear mode at $k = 2$ irrespective of the specification of the regression model (heterogeneous or homogeneous). Thus, Table 5 illustrates the results conditionally on $k = 2$.

Let us first comment on the results of Table 5 : for the two data sets, as expected, when the "true model" was used to analyse the data (diagonal elements in Table 5), the posterior expectations for w, λ and γ are close to their true values.

However, when the regression model used to analyse the data is misspecified (off-diagonal elements in Table 5), an alternative solution is found. When data set I is analysed with the assumption of homogeneous regression effect, we find that the estimated regression coefficients are close to the true ones for the *second* component, which has the higher weight. Consequently, the estimated value of λ_1 ($E(\lambda_1|y) = 0.54$) is shrunk in order to separate out this component. The weights show that some data of the first component must be misclassified and this is confirmed by the allocation map (shown Figure 6, bottom left). On the other hand, all data in the second component are well allocated.

[Figure 6 about here.]

On the contrary, if we analyse data set II with the model of heterogeneous effect (which is unnecessarily over-parameterised), this does not change the weights, nor the allocation map, the λ are still well estimated but the regression coefficients in the two components are not well separated.

Unsurprisingly, the fit seems to be better under the heterogeneous assumption for the data set I (see Figure 7). For the data set II, the fit is only slightly modified by the assumption of heterogeneous coefficients.

[Figure 7 about here.]

The hypothesis of an heterogeneous covariate effect, which introduces more regression parameters in the model (and hence causes slower mixing), can be checked a posteriori: in case where the posterior distribution of k indicates that there is more than one Poisson regression, we can assess how the estimated regression coefficients differ from one component to another one. For example, the coefficients γ_2 estimated for data set II have posterior means that are close to each other and to the one estimated in the homogeneity case, while the γ_1s have non-separated covering intervals.

We believe that the hypothesis of heterogeneous covariate effect, that substantially increases the complexity of the model, the algorithm, and the interpretation, has to be considered with caution and in the light of whether there is a clear improvement of either the fit or the stability of allocation of the observations. A further point to consider is whether the posterior distributions of the estimated regression coefficients are clearly separated.

6 Application

Data on which we apply this model come from INRETS (Institut National de Recherche sur les Transports et leur Securité) and concern the traffic on 118 road junctions in France. Previous analyses have been published by Brenac [15].

The observed numbers of road accidents (Y) in these 118 road junctions have been recorded. Covariates entering the regression are:

E; the period on which accidents have been recorded (in number of years).

Ts; Daily traffic flow on secondary road.

Tm; Daily traffic flow on main road.

The variable E can be seen as the expected number of accidents, and we include it in the Poisson regression as an offset variable. The other variables (Ts and Tm) have been standardised.

The aim of the study was to determine which of these 118 were the most dangerous and hence deserve priority in being equipped: this aim can be formulated as the need to discriminate between two subgroups of road junctions: "less dangerous" / "dangerous". The standard Poisson regression first performed by Brenac showed substantial over-dispersion, which was taken into account by introducing a coefficient of extra-Poisson variability following Breslow ([16]).

We first fit a model with homogeneous covariate effect, which is written as:

$$f(Y_i|z_i, E, Ts, Tm, \lambda, \gamma) = \text{Poisson}(\lambda_j \; E_i \; \exp(\gamma_{Ts} \; Ts_i + \gamma_{Tm} \; Tm_i))$$

and where the priors are chosen as:

$k \sim \text{Uniform}\{1, \ldots, 30\}$, $\lambda_j \sim \text{Uniform}[0, 31]$ and $\gamma_l \sim \text{Normal}(0, 25)$.

We use the SM2 move for changing the dimension and ran 300000 MCMC sweeps after 100000 for burn-in.

The posterior distribution of k, shown in Figure 8, is clearly peaked on $k = 3$. Figure 8 also shows the matching of the histogram by the predicted values, that indicates a good fit, as well as the posterior densities of the Poisson regression coefficients for Ts and Tm.

[Figure 8 about here.]

Table 6 gives the posterior means of the parameters, conditionally on $k = 3$.

[Table 6 about here.]

The identification of a small ($w_3 = 7\%$) subgroup of road junctions where λ is much higher illustrates that there exist some road junctions where the risk of accidents is increased, independently from the traffic parameters, a subgroup that should be re-equipped as a priority.

This is coherent with the results found by Brenac : 6 road junctions were identified as the 5% of the sites where the residuals showed an excess of risk. Five of these 6 appear among the 8 road junctions allocated to the third component when using modal classification, the 6th one being mostly allocated to the second component. Hence, the Poisson mixture gives a natural classification which can be used for identifying outliers instead of choosing an arbitrary cut-off.

A model with heterogeneous covariate effect has also been tried, but in this case, we found a clear lack of stability for the coefficients λs and, symmetrically, of the corresponding γs, with no indication of an improvement of the fit given by predicted values. Figure 9 illustrates this unstability during the course of the algorithm. Moreover, the 95% credibility intervals of the regression parameters are not well separated for any of the two covariates (results not shown).

[Figure 9 about here.]

7 Discussion

The models and algorithms presented successfully permit the assessment of the existence of discontinuities in the underlying structure of counts data, and estimation of the number of latent subgroups.

Concerning the different algorithms implemented for changing k via reversible jump moves, we have shown that a blind birth and death move gives convergence of similar quality as split and merge moves. Nevertheless, the CPU time needed when using the birth and death is higher, and this cannot be ignored, especially in the case of a large number of observations. This is due to the fact that all the observations are reallocated among the component at each proposition of birth or death. Other variants of birth and death move which would preserve part of the allocations could be envisaged. We have not explored this further, our conclusions being that fine tuning in this case would not lead to clear improvement of performances.

Recall that the prior distribution for the parameters of a mixture cannot be completely noninformative. We have shown that, even when using priors that are weakly informative, there is a notable influence on the posterior estimation of the number of subgroups. In a recent paper, Wasserman [17] proposes to use Jeffrey's priors for the parameters of any mixture model with a fixed number of components, conditioning on there being no empty components. This way of being noninformative is not directly applicable to the case of variable k since it requires the non-straightforward evaluation of normalising constants. This could be an interesting direction for future research.

The components have been labelled and ordered following the values of λ. In the case where no covariates are involved, this is coherent with the assumption of a discontinuity that would take the form of unobserved subgroups with different expectations, each Poisson distribution being characterised by a unique parameter λ. We have also seen that the dependent prior permits describing a priori how distant these expectations are.

We are aware that, when covariates are introduced, the question of labelling is more delicate. Let us consider the case where there is one covariate, which has a different effect among two components, but that, once this effect is identified, the two components have very close Poisson rates λ. It is clear that a model with homogeneous effect will not fit these data adequately. But neither will a model with heterogeneous effects which assumes, via the ordering constraint, that the λs are different. Indeed, there is substantial overlap between posterior distributions of the λs, and the difficulty in distinguishing the λs will give bimodal posterior distributions for the regression coefficients.

Other approaches to presentation of posterior inferences exist, where components are not ordered, and a search for the modes of the joint posterior distribution is made. This is the approach developed by Stephens [9] in the context of mixture of distributions, with unordered components. Using Stephen's approach, Hurn, Justel and Robert ([18]) have explored the estimation of mixture of regressions. But we stress that *joint* inference about component and covariate parameters is not affected by such issues, it is only when we seek to make *marginal* inferences that the difficulty of interpretation arises.

Our limited experience with a model including heterogeneous covariate effect leads us to conclude on a cautious note. The increasing flexibility of the model is offset by problems of non-identifiability which should not be discounted. We believe that interpretation has to be paired with prior knowledge.

Acknowledgments

We would like to thank Sylvain Lassarre and Thierry Brenac for providing the INRETS data set. For financial support, we want to acknowledge the TMR network contract C.E.ERB-FMRX-CT960095 (VV) and an EPSRC visiting fellowship (SR).

References

[1] D. Clayton and L. Bernardinelli. Bayesian methods for mapping disease risk. In *Geographical and environmental epidemiology : methods for small-area studies*, pages 205–220. Oxford University Press, London, 1992.

[2] G. J. McLachlan and K. E. Basford. *Mixture models : inference and applications to clustering*. Dekker, New York, 1988.

[3] P.J. Green. Reversible jump Markov Chain Monte Carlo computation and Bayesian model determination. *Biometrika*, 82:711–732, 1995.

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

[5] M.A.Gruet, A.Philippe, and C.P. Robert. MCMC control spreadsheets for exponential mixture estimation. *Journal of Computational and graphical Statistics*, 8:298–317, 1999.

[6] P. Dellaportas, D. Karlis, and E.Xekalaki. Bayesian analysis of finite poisson mixtures. Technical report, Athens University of Economics and Business, 1997.

[7] K. Roeder and L. Wasserman. Practical Bayesian density estimation using mixtures of normals. *Journal of the American Statistical Association*, 92:894–902, 1997.

[8] D. Spiegelhalter, N.G. Best, and B.P. Carlin. Bayesian deviance, the effective number of parameters, and the comparison of arbitrary complex models. Unpublished manuscript, 1998.

[9] M. Stephens. Bayesian analysis of mixtures with an unknown number of components — an alternative to reversible jump methods, 1998.

[10] G. Celeux, M.A. Hurn, and C.P. Robert. Computational and inferential difficulties with mixture posterior distributions. *Journal of the American Statistical Association*, 2000. to appear.

[11] C. P. Robert. *Discretization and MCMC convergence assessment*, volume 135 of *Lecture notes in statistics*. Springer-Verlag, 1998.

[12] A.D. Sokal. Monte carlo methods in statistical mechanics : fondations and new algorithms. Cours de troisième cycle de la physique en Suisse Romande, Lausanne, 1989.

[13] J. Besag and P.J. Green. Spatial statistics and Bayesian computation. *Journal of the Royal Statistical Society (Series B)*, 55:25–37, 1993.

[14] L. Watier, S. Richardson, and P.J. Green. Using gaussian mixtures with unknown number of components for mixed model estimation. In *14th International Workshop on Statistical Modelling*, Graz, Austria, 1999.

[15] T. Brenac. Accidents en carrefour sur routes nationales, modélisation du nombre d'accidents prédictible sur un carrefour et exemples d'applications. Technical Report 185, INRETS, Arcueil, 1994.

[16] N.E. Breslow. Extra-poisson variation in log-linear models. *Applied Statistics*, 33(1):38–44, 1984.

[17] L. Wasserman. Asymptotic inference for mixture models using data dependent priors. Technical Report 677, Carnegie Mellon University, 1998.

[18] M.A. Hurn, A. Justel, and C.P. Robert. Estimating mixtures of regressions. Technical report, CREST, 2000.

Appendix

Metropolis acceptance ratio

Let us consider the case where there is no covariate, and define the states $x = (w, \lambda, k, z)$ and $x' = (w', \lambda', k', z')$, with $k' = k + 1$. The elements of vector u are the auxiliary variables needed in the reversible jump so that dimensions of (x, u) and x' match.

The Metropolis acceptance ratio is
P(accept $x \to x'$) = $\min(1, A)$ or P(accept $x' \to x$) = $\min(1, A^{-1})$, with

$$A = \frac{p(k')p(w'|\delta, k')p(\lambda'|\theta, k')p(z'|k', w')p(y|\lambda', z')}{p(k)p(w|\delta, k)p(\lambda|\theta, k)p(z|k, w)p(y|\lambda, z)}$$

$$\times \frac{pr(x' \to (x, u))}{pr((x, u) \to (x'))p(u)} \times \left| \frac{\partial(w', \lambda', k', z')}{\partial(w, \lambda, k, z, u)} \right|.$$

where the Jacobian in the last term will be denoted by $|J|$ and $pr(a \to b)$ denotes the probability of proposing b while in a.

In the following, we will detail the terms involved in A for the case where the prior f for λ is independent. The value of A depends on the type of move implemented in the reversible jump. Throughout, $P(alloc)$ denotes the probability of a specific allocation among the components, and b_k and d_k refer to the probabilities of choosing to increase and decrease by 1 the number of components k.

If the dimension is changed by a birth and death process, then $u = (w_*, \lambda_*)$ where $w_* \sim \text{Beta}(1, k)$ with density denoted by g and λ_* simulated from its prior : $p(\lambda_*) = f(\lambda_*|\theta)$. We calculate $|J| = (1 - w_*)^{k-1}$. Thus :

$$A = \frac{p(k+1)}{p(k)} \times \frac{w_*^{\delta-1}(1-w_*)^{k(\delta-1)}}{B(\delta,k\delta)} \times (k+1)f(\lambda_*|\theta) \times \frac{p(z'|k',w')p(y|\lambda',z')}{p(z|k,w)p(y|\lambda,z)}$$
$$\times \frac{d_{k+1}P(alloc\ \ in\ \ k)}{(k+1)b_k P(alloc\ \ in\ \ k+1)} \times g(w*)^{-1} f(\lambda_*|\theta)^{-1} \times |J|$$
$$= \frac{p(k+1)}{p(k)} \times \frac{w_*^{\delta-1}(1-w_*)^{k(\delta-1)}}{B(\delta,k\delta)} \times \prod_{i=1}^{n} \frac{\sum_{j=1}^{k'} w'_j \exp(-\lambda'_j)\lambda_j'^{y_i}}{\sum_{j=1}^{k} w_j \exp(-\lambda_j)\lambda_j^{y_i}} \times \frac{d_{k+1}}{b_k} \times B(1, k)$$

If the dimension is changed by a split and merge move, then $u = (u_1, u_2)$ with u_1 and u_2 being simulated from distributions with densities g_1 and g_2 respectively. If component j is split into components (j_1, j_2) :

$$A = \frac{p(k+1)}{p(k)} \times \frac{w_{j_1}^{\delta-1+l_1} w_{j_2}^{\delta-1+l_2}}{w_j^{\delta-1+l_1+l_2} B(\delta, k\delta)} \times (k+1) \frac{f(\lambda_{j_1}|\theta) f(\lambda_{j_2}|\theta)}{f(\lambda_j|\theta)} \times \frac{p(y|\lambda', z')}{p(y|\lambda, z)}$$

$$\times \frac{d_{k+1}}{b_k P_{alloc}} \times (g_1(u_1) \times g_2(u_2))^{-1} \times |J|$$

where

split/merge 1	$\lvert J \rvert = \frac{\sqrt{\lambda_j} w_j^2}{1-u_1}$,	$u_1 \sim \text{Beta}(2,2)$,	$u_2 \sim \text{Exp}(2)$
split/merge 2	$\lvert J \rvert = c w_j$,	$u_1 \sim \text{Uniform}(0,1)$,	$u_2 \sim \text{Uniform}(0,1)$
split/merge 3	$\lvert J \rvert = \frac{\lambda_j w_j}{1-u_1}$,	$u_1 \sim \text{Beta}(2,2)$,	$u_2 = u_1 \text{Beta}(2,2)$.

If covariates are included in the model with heterogeneous effect, we simulate $u = (u_1, u_2, v_1, ..., v_p)$ where each $v_l \sim \mathrm{N}(0, \sigma_v^2)$, then propose, within a split move, to change γ_{jl} into $(\gamma_{j_1 l}, \gamma_{j_2 l})$ for $1 \leq l \leq p$. The extra terms that have to be introduced in A are :
$\times \frac{\prod_{l=1}^p p(\gamma_{j_1 l}|\sigma^2)) p(\gamma_{j_2 l}|\sigma^2)}{p(\gamma_{jl}|\sigma^2)}$ entering in the joint density
$\times \prod_{l=1}^p p(u_l|\sigma_u^2))$ entering in the proposal terms
and $\times \prod_{l=1}^p 2\gamma_{jl}$ in the Jacobian.

Figures

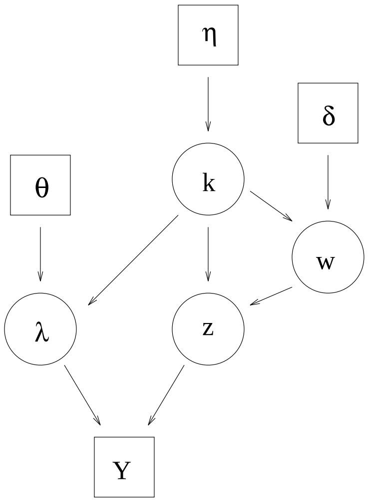

Figure 1: Directed Acyclic Graph for the Poisson mixture model

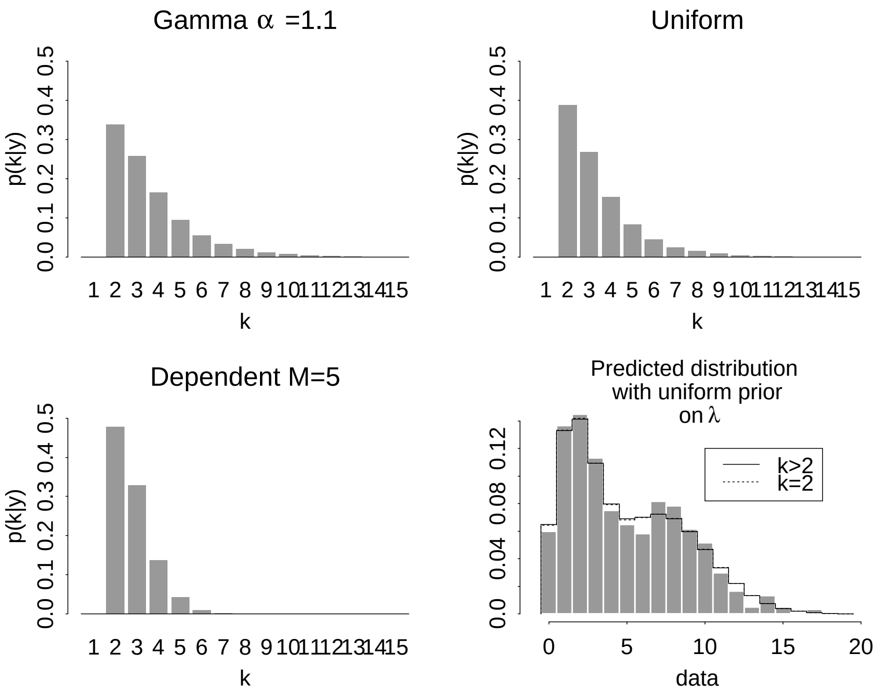

Figure 2: Posterior distributions of k under different prior distributions for λ (left and top right), and fit (bottom right) of the simulated two-component Poisson mixture by its predicted distribution when using a Uniform prior for λ.

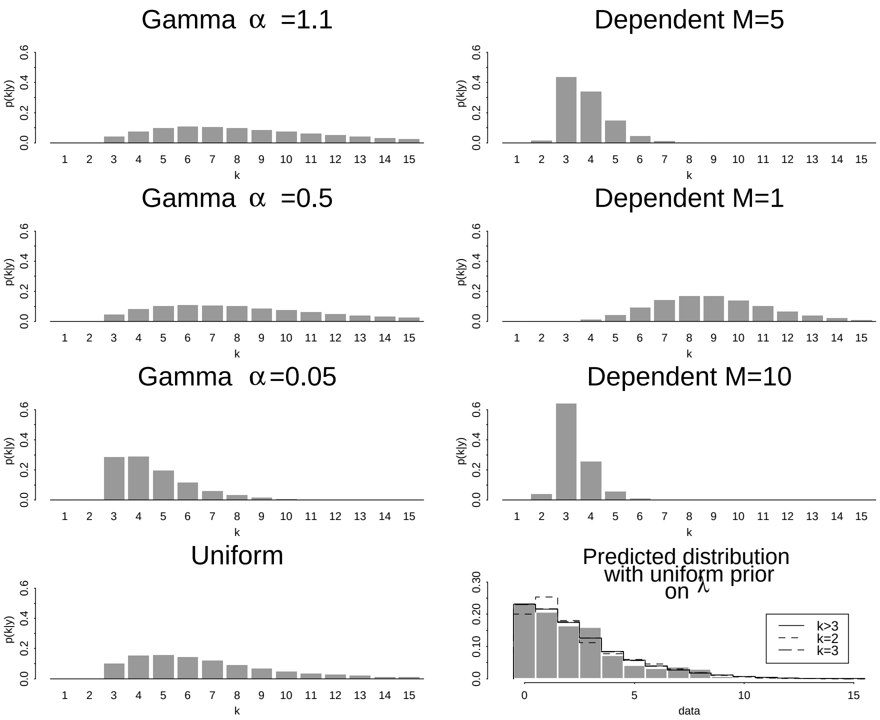

Figure 3: Posterior distribution of k for different prior distributions of λ, and fit (bottom right) of the simulated three-component Poisson mixture by its predicted distribution when using a Uniform prior for λ.

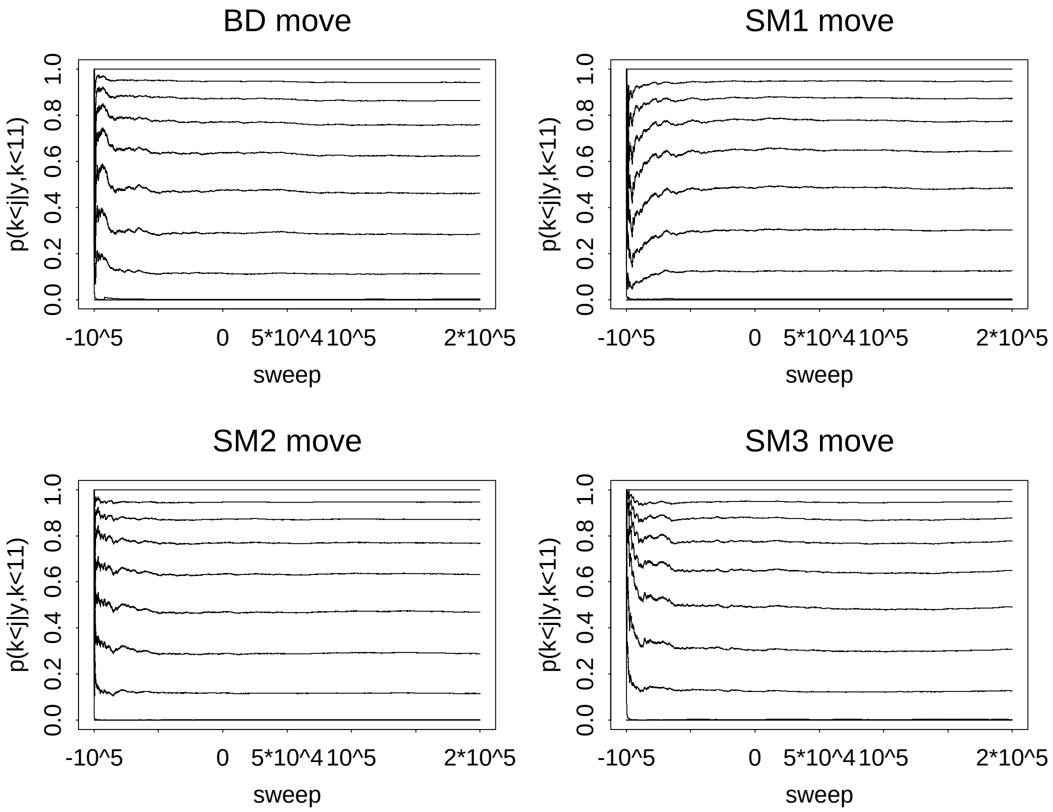

Figure 4: Graphical comparison of the MCMC performance of the different jump moves on the simulated data for the three-components Poisson mixture: plots of cumulative occupancy fractions for the posterior distribution of k, with a Uniform prior for λs.

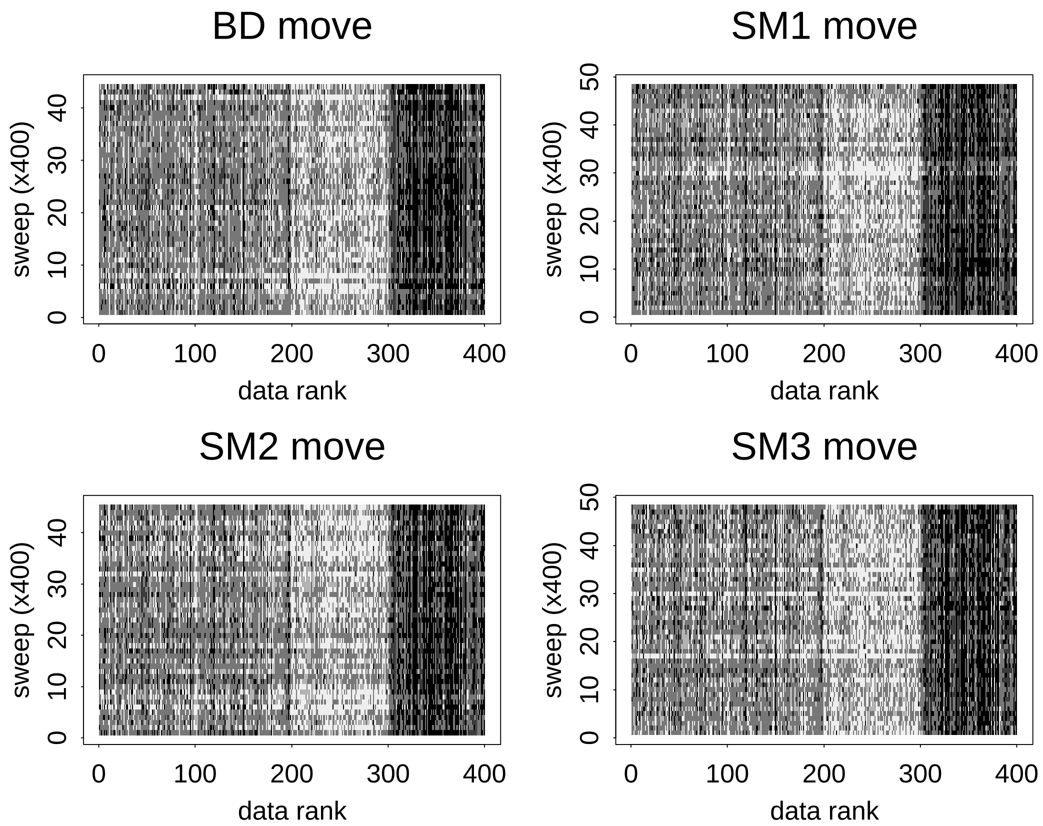

Figure 5: Allocation maps of the observations simulated from the three-component Poisson mixture, conditioning on $k = 3$, with a Uniform prior for λs

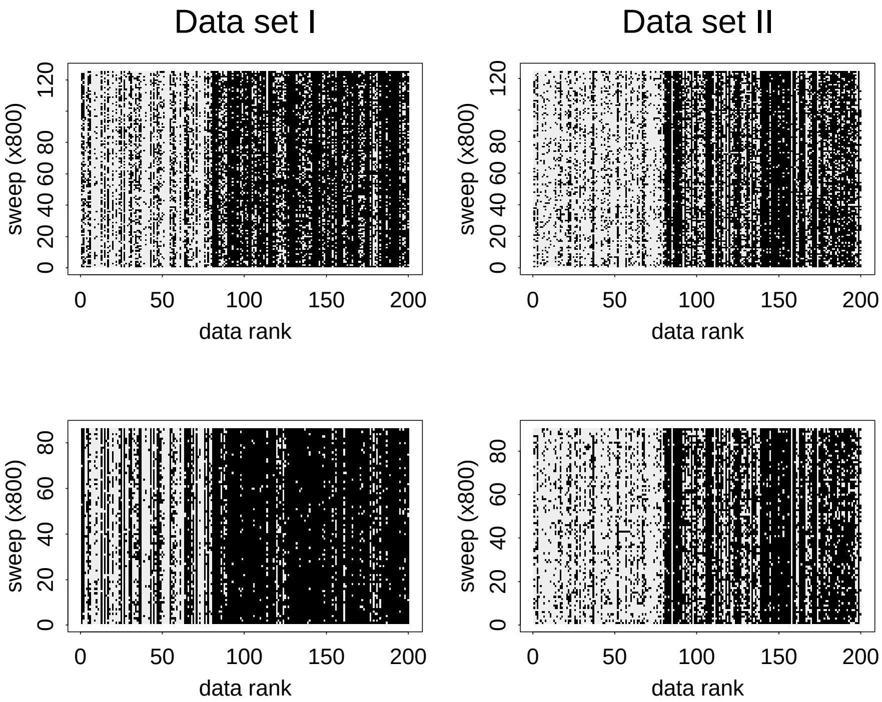

Figure 6: Allocation map for data sets I (left) and II (right) given under the two different assumptions regarding the covariates effect: heterogeneous (top row) or homogeneous (bottom row).

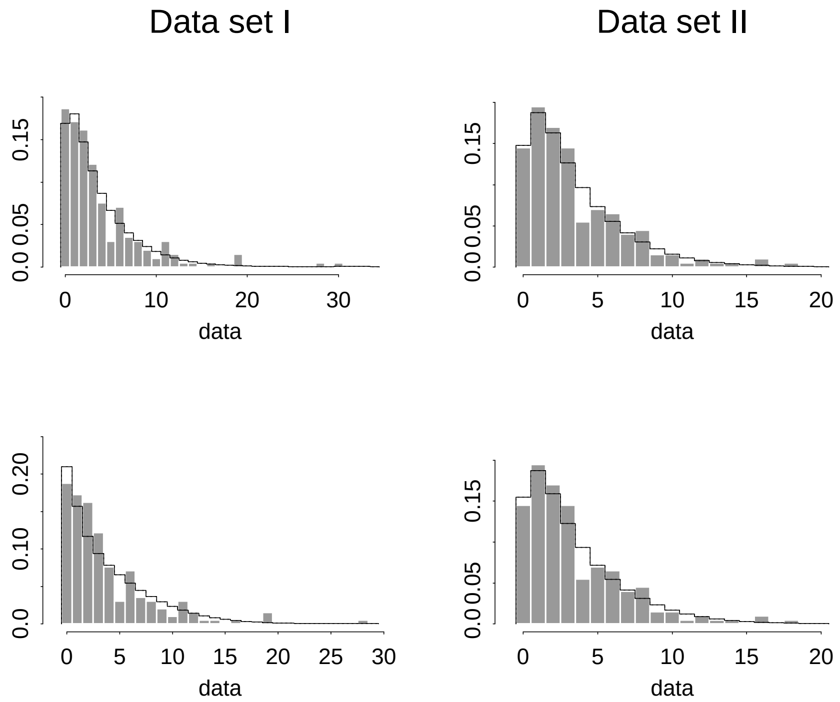

Figure 7: Fit of the histogram of observed data by the predicted values, for the two simulated data sets and the two different assumptions regarding the covariates effect: heterogeneous (top row) or homogeneous (bottom row).

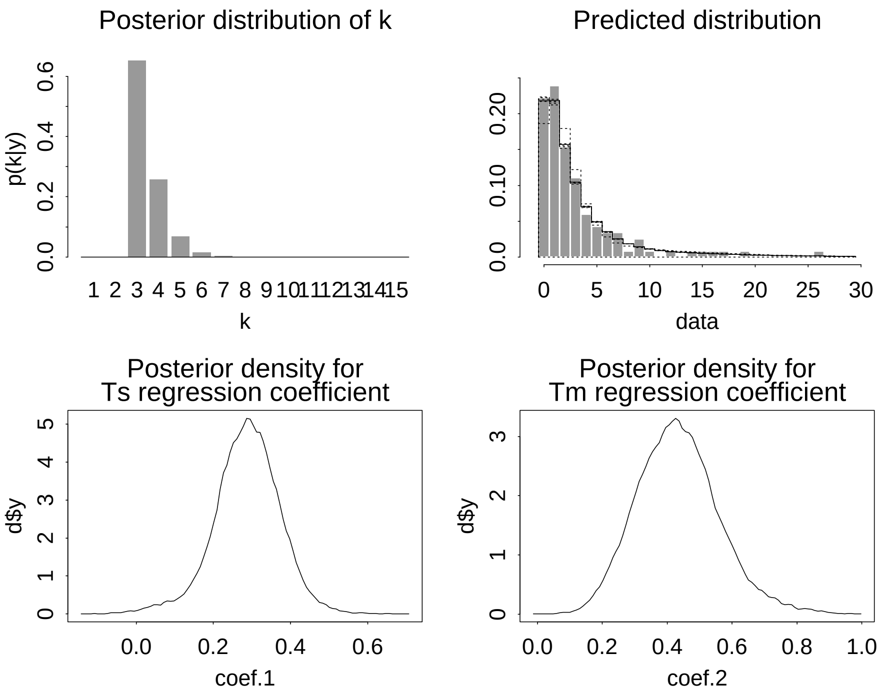

Figure 8: Posterior distribution of k, posterior fit of the predicted values, and posterior densities of the regression parameters.

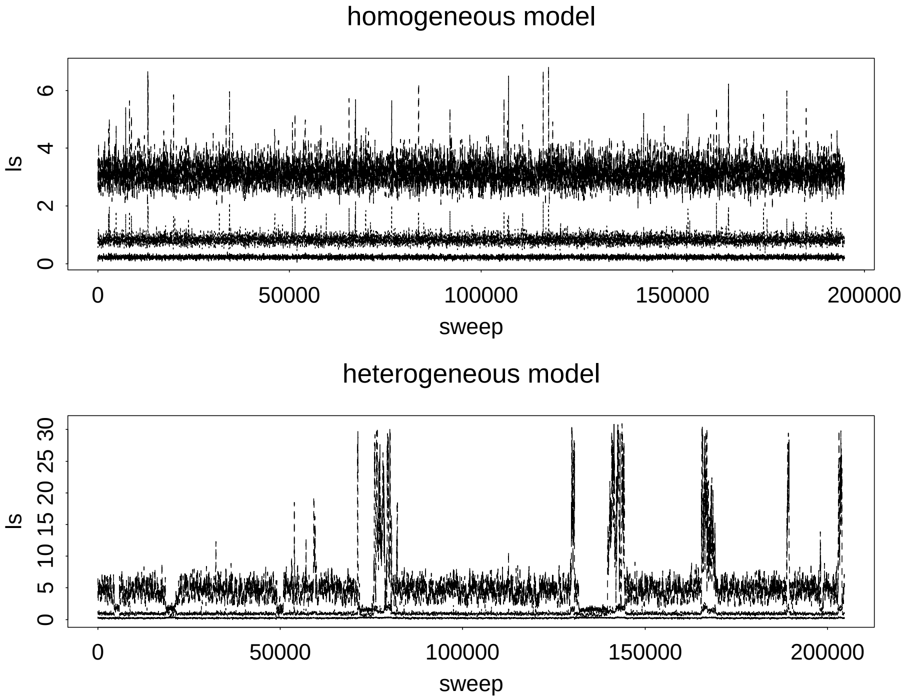

Figure 9: Values of λ_1, λ_2 and λ_3 when $k = 3$ with homogeneous (top) and heterogeneous (bottom) covariate effect model.

Tables

Table 1: Posterior estimates for the two-component Poisson mixture

λ prior form	Mixture parameters$\vert k=2$				Deviance
	$\widehat{w_1}$	$\widehat{w_2}$	$\widehat{\lambda_1}$	$\widehat{\lambda_2}$	when $k=2$
Gamma	0.505	0.495	2.06	7.84	2478.6
Uniform	0.505	0.495	2.07	7.85	2478.5
Dependent $M=5$	0.505	0.495	2.07	7.84	2478.8

Table 2: Posterior estimates for the three-component Poisson mixture

λ prior form	Mixture parameters$\|k=3$						Deviance		
	$\widehat{w_1}$	$\widehat{w_2}$	$\widehat{w_3}$	$\widehat{\lambda_1}$	$\widehat{\lambda_2}$	$\widehat{\lambda_3}$	$k=2$	$k=3$	$k=4$
Gamma									
$\alpha=1.1$	0.275	0.531	0.194	0.47	2.19	5.79	1376.3	1296.5	1295.2
$\alpha=0.5$	0.218	0.582	0.200	0.30	2.03	5.75	1373.7	1284.2	1285.0
$\alpha=0.05$	0.138	0.646	0.216	0.05	1.82	5.62	1370.6	1264.1	1260.7
Uniform	0.279	0.533	0.188	0.47	2.23	5.86	1374.3	1295.5	1293.5
Dependent									
$M=5$	0.277	0.531	0.192	0.46	2.20	5.80	1373.4	1296.2	1294.7
$M=1$	0.357	0.415	0.228	0.64	2.44	5.08	1369.8	1323.7	1314.7
$M=10$	0.275	0.537	0.188	0.45	2.21	5.87	1373.7	1294.5	1293.5

Table 3: Sokal integrated autocorrelation time calculated on the last 20000 iterations of a 300000 run (including 100000 for burn-in), with Uniform priors for λs

	BD	SM1	SM2	SM3
Sokal				
for k	5.66	8.74	5.87	10.32
for deviance	2.10	1.98	1.61	1.62
CPU time (seconds)	576	439	442	442

Table 4: Sokal integrated autocorrelation time calculated on the last 20000 iterations of a 300000 run (including 100000 for burn-in), with dependent λs prior

	BD	SM1	SM2	SM3
Sokal				
for k	2.95	1.95	1.69	2.26
for deviance	5.24	3.17	2.54	3.59
CPU time (seconds)	357	310	309	311

Table 5: Posterior summaries (posterior means for the mixture parameter, posterior means and 95% credibility intervals for the regression coefficients) under the different analysis models, conditionally on $k = 2$.

	Data set I (heterogeneous)		Data set II (homogeneous)	
	Component 1	Component 2	Component 1	Component 2
heterogeneous	$p(k=2\|y) = 0.997$		$p(k=2\|y) = 0.87$	
	$w_1 = 0.40$	$w_2 = 0.60$	$w_1 = 0.48$	$w_2 = 0.52$
	$\lambda_1 = 1.18$	$\lambda_2 = 4.32$	$\lambda_1 = 1.41$	$\lambda_2 = 4.66$
	$\gamma_{1,1} = 0.32$	$\gamma_{2,1} = 0.73$	$\gamma_{1,1} = 0.07$	$\gamma_{2,1} = 0.26$
	$[0.03, 0.64]$	$[0.65, 0.81]$	$[-0.15, 0.29]$	$[0.15, 0.37]$
	$\gamma_{1,2} = 0.43$	$\gamma_{2,2} = -0.32$	$\gamma_{1,2} = 0.44$	$\gamma_{2,2} = 0.46$
	$[0.22, 0.64]$	$[-0.40, -0.25]$	$[0.19, 0.71]$	$[0.35, 0.57]$
homogeneous	$p(k=2\|y) = 0.68$		$p(k=2\|y) = 0.34$	
	$w_1 = 0.29$	$w_2 = 0.71$	$w_1 = 0.48$	$w_2 = 0.52$
	$\lambda_1 = 0.54$	$\lambda_2 = 4.29$	$\lambda_1 = 1.35$	$\lambda_2 = 4.76$
	$\gamma_1 = 0.72$		$\gamma_1 = 0.22$	
	$[0.65, 0.80]$		$[0.12, 0.32]$	
	$\gamma_2 = -0.30$		$\gamma_2 = 0.46$	
	$[-0.37, -0.23]$		$[0.35, 0.57]$	

Table 6: Road junctions data set : results under hypothesis of homogeneous effect conditional on $k = 3$. The first column gives the results of a simple Poisson regression (conditional on $k = 1$).

	Component 1	Component 2	Component 3
	$w_1 = 0.59$	$w_2 = 0.34$	$w_3 = 0.07$
$\lambda = 0.60$	$\lambda_1 = 0.23$	$\lambda_2 = 0.84$	$\lambda_3 = 3.14$
$\gamma_{Ts} = 0.26 \quad [0.16, 0.36]$		$\gamma_{Ts} = 0.28 \quad [0.09, 0.45]$	
$\gamma_{Tm} = 0.50 \quad [0.39, 0.62]$		$\gamma_{Tm} = 0.43 \quad [0.20, 0.71]$	