A Bayesian Approach for Simultaneous Segmentation and Classification of Count Data

Olivier Cappé

Abstract

A Bayesian approach is proposed that provides a concise description of a series of counts under the form of homogeneous consecutive data segments which are classified based their marginal distribution. Due to the flexibility of the corresponding model, carrying out the actual inference turns out to be a complex task for which an original combination of several Markov Chain Monte Carlo (MCMC) simulation tools is developed. The proposed MCMC sampler makes use of reversible jump moves to achieve communication between models with different numbers of both segments and classes. A large section of the paper is devoted to the discussion of the results obtained on a medium duration section (a few minutes) of a publicly available teletraffic trace taken from the Internet traffic archive.

EDICS: 2-SYSM *System Modeling and Representation* (Special Issue on Monte Carlo Methods for Statistical Signal Processing edited by P. M. Djuric and S J. Godsill)

I. INTRODUCTION

Modeling and analysis of discrete-valued time series plays an important role in various domains which range from social and environmental sciences (monitoring of road traffic for instance) to telecommunication engineering (control and prediction of the amount of data transmitted over a computer network, which we shall refer to as "teletraffic" in the following).

In this contribution, we address the task of simultaneously segmenting and classifying an observed record of discrete-time count data. Segmentation means finding contiguous regions of the data that can be considered as homogeneous, whereas classification aims at identifying various characteristic levels in the data. Typical statistical models for these two purposes are, respectively, changepoint models (segmentation) and mixture models (classification). In achieving this goal of joint segmentation and classification, our may concern will be flexibility.

A first requirement is that the discrete-valued discrete-time nature of the observation be specifically taken into account. In particular, the fact that the observations corresponds to time-aggregated counts obtained from a point process must be accounted for (see section II-A for details).

Flexibility also has some more subtle implications concerning the prior assumptions that can validly be made about the typical durations of segments: In discrete time changepoint models, the most common option consists in assuming that the durations of the stationary segments are a priori independent and geometrically distributed. This has the advantage that, conditionally on the model parameters, posterior inference regarding the positions of the changes can be carried out efficiently using deterministic algorithms [1], [2]. For such models, joint parameter and change locations estimation can be carried out using methods based on Markov Chain Monte Carlo (MCMC) simulations [2], [3], [4]. It is shown in [5] that if the segments are labelled (or belong to a known number of "regimes" following the author's terminology) and with the same prior assumption on the duration of the segments, the model is fully equivalent to a so-called Hidden Markov Model (HMM) [6] for which efficient block simulation strategies are available. In the following however, we will not be using this remark as we consider a prior structure which is more flexible than the HMM in the sense that it allows for more widely dispersed segment durations (see figure 4 of section II-B).

Flexibility finally implies that key quantities including the plausible numbers of classes and segments needed to represent the data should be inferred by the analysis and not predefined by prior expert analysis. To this end, we follow the same approach as that used for the HMM model by [7] which

Address for correspondence: Olivier Cappé <cappe@tsi.enst.fr>, ENST Dpt. TSI / CNRS URA 820, 46 rue Barrault, 75634 Paris cedex 13, France. fax +33 1 45 88 79 35, voice +33 1 45 81 71 11.

is based on the reversible jump methodology introduced by [8]. As far as the MCMC machinery is concerned, the proposed approach differs from both [4], [9] or [10] and [11], [7] by the fact that it implies two different types of dimension changing moves that are nested.

To illustrate this approach, we consider its application to the analysis of a section of aggregated teletraffic data taken from one of the traces available from the Internet Traffic Archive (located at `http://ita.ee.lbl.gov/`). This is a difficult and challenging application because there is currently no available model which can adequately represent such complex data which results from the superposition of a large number of inhomogeneous and non-stationary data flows, especially without knowledge of routing information [13], [12], [14], [15]. The aim here is definitely not to the model the physical process that generates the data but rather to obtain meaningful summaries (or "stylized facts" following the terminology used in econometrics) from traffic traces. Such summaries could be used to monitor the traffic or serve as development tools, as in [16] which describes the application of a graphical stylization tool known as "textured plots" to source level traces in order to validate the so-called "self-similar" traffic model. Note finally that contrary to references such as [13] or [12] which focus on the characterization of the scaling properties and large scale behavior of traffic, the goal of our analysis will be to obtain a detailed description of the statistical behavior of short sections of the data. There is a priori no link between the two approaches since modeling the long term dependence effects does not require to consider some statistical features of the traffic such as marginal distribution or short term correlations, which are rather considered as nuisance parameters. Conversely, we will see in section IV that our approach, which is in some sense a non parametric smoothing method, provides results that cannot be extrapolated to infer the long term behavior of the data.

The rest of the paper is organized as follows: The Bayesian model is presented in section II; The associated MCMC sampler is described in section III, and section IV discusses results obtained on a teletraffic data.

II. Bayesian model

A. Data and parameterization

We consider that the data to be analyzed consists of a section $\{n_t\}_{t=1,\dots,T}$ of length T of a discrete-time count process ($n_t \in \mathbb{N}$). To this observable data is associated a latent (or unobservable) structure which simultaneously defines the segmentation $(K, b_2, \dots b_K)$ and classification $(M, l_1, \dots, l_K)$ configuration according to Fig. 1. The *segments* are hypothetical homogeneous sections of the data which are unambiguously defined by the number of segments K and the segment boundaries $b_1, \dots b_{K+1}$, where the k-th segment extends from indexes b_k to $b_{k+1} - 1$ (inclusive). In accordance with this parameterization, the very first (b_1) and last (b_{K+1}) segment boundaries are set to one and $T + 1$ respectively. The *classes* define different types of statistical characteristics corresponding to the data sections defined by the segments. The classes are parametric and are defined (up to a permutation of the order in which the classes are numbered) by the number of classes M, and the class parameters $(\kappa_m, \pi_m)_{1 \le m \le M}$ (see below for the meaning of these parameters). Each segment is linked to a particular class by means of an attached label l_k which takes values in the range $1, \dots M$.

[Figure 1 about here.]

To implement the idea that the segments should represent homogeneous regions of the data, we will assume that given the parameters and the latent data $\Theta = (M, \kappa_1, \dots, \kappa_M, \pi_1, \dots, \pi_M, K, l_1, \dots, l_K, b_2, \dots b_K)$, the counts n_t are conditionally independent with a marginal distribution which depends first, on the section to which they belong, and, second, on the label attached to this section.

The choice of the marginal distribution is an issue that must be treated with some care. In many works that deal with discrete-valued data such as [17] for a semi-parametric model, [8] for a changepoint model, [18] for a parametric model, the (conditionally) Poisson assumption is used. The problem is that discrete time count data generally arises from the observation of the number of events associated with an underlying continuous time process. In general, the event patterns associated with the underlying continuous time process are less regular than those of a Poisson process and often are non-homogeneous (in time). In such cases, aggregation (or time averaging) has the effect of increasing the variability of the measured counts with respect to the Poisson distribution with identical mean. This observation is well established in actuarial sciences for instance [19] and will be clearly illustrated when analyzing teletraffic data in section IV. To cope with this increase in variability, we will assume that given the latent structure Θ, the observed counts n_t are independent with the *negative Binomial* distribution $\mathrm{Neg-Binomial}(\kappa_\mathrm{m}, \pi_\mathrm{m})$, where m is the label associated with the segment that contains t (*i.e.* l_k where k is such that $b_k \leq t < b_{k+1}$). Thus, each class corresponds to a different negative binomial distribution characterized by the two parameters κ and π:

$$\mathrm{Neg-Binomial}(\mathrm{n}|\kappa, \pi) = \left(\begin{array}{c} n+\kappa-1 \\ \kappa-1 \end{array} \right) \pi^\kappa (1-\pi)^\mathrm{n} \quad \mathrm{for\ n} \in \mathbb{N}$$

where $\pi \in (0, 1)$ and $\kappa > 0$ are both treated as continuously varying parameters. The above parameterization is reminiscent of the interpretation of the negative binomial distribution as a waiting time distribution (number of failures before the κth success for independent Bernoulli trials with probability of success π - assuming that κ is an integer). In the model under consideration, this interpretation is not applicable and it would be more natural to use as parameters the mean $\mu = \kappa(1-\pi)/\pi$ and the variance over-dispersion ratio $\gamma = 1/\pi$ (ratio of the variance to that of a Poison distribution with the same mean). We will however use the parameterization defined by κ and π which is the most attractive from a computational point of view (see section III-C). The transformations

$$\pi(x) = \frac{x}{1+x} \quad \text{and} \quad \beta(x) = \frac{x}{1-x}, \tag{1}$$

will also be needed in the following.

B. Priors and hyperparameters

The complete prior structure is plotted in figure 2 using the standard graphical model conventions (circles denote random variables and boxes contain fixed hyperparameters).

[Figure 2 about here.]

The prior on M is geometric with rate $\tau_M \in (0, 1)$ (i.e. $P(M = m) = (1 - \tau_M)/\tau_M \times \tau_M^m$ for $m \leq 1$). This choice does not favor *a priori* any specific value of the number of classes and is simply meant to enforce a complexity penalty (favoring parsimonious models).

[Figure 3 about here.]

Specifying a proper prior for the parameters κ_m and π_m of each class is an absolute requirement in this type of model because of the possible outcome of empty classes for which the posterior would be improper – see [20] which discusses this issue for mixture models. We use independent priors for each of the couple of class-dependent parameters (κ_m, π_m). It is however not sensible to assume that κ_m and π_m themselves are independent even for the purpose of a non informative analysis. In order to illustrate this latter point we randomly drew 5000 segments, with length between 5 and 60 points (representing from 5 seconds to one minute of traffic), in the complete traffic trace from which the data analyzed in section IV is extracted. For each segment, the maximum likelihood estimate of the negative binomial parameters was approached using a few steps of the coordinate ascent method and plotted in log-log coordinates. Figure 3 shows that while it is not unrealistic to assume that the mean $\mu = \kappa(1-\pi)/\pi$ and the dispersion π are independent, the same assumption directly applied to κ and π would totally contradict the nature of the model. Fig. 3-(a) indeed reveals that κ and π are strongly interdependent. The choice of independent priors on κ and π, especially if those priors are vague, would result in most of the a priori mass being put on either very small or very large values of the mean μ. As a consequence, we assume that a priori,

$$p(\mu_m, \pi_m) = \text{Gamma}(\mu_\text{m}|\alpha_\mu, \beta_\mu)\ \text{Beta}(\pi_\text{m}|\alpha_\pi, \beta_\pi) \tag{2}$$

which gives, after application of the transformation $\kappa_m = \mu_m \pi_m/(1 - \pi_m)$,

$$p(\kappa_1, \pi_1, \ldots, \kappa_M, \pi_M|M) = \prod_{m=1}^{M} \text{Gamma}(\kappa_\text{m}|\alpha_\mu, \beta_\mu/\beta(\pi_\text{m}))\text{Beta}(\pi_\text{m}|\alpha_\pi, \beta_\pi) \tag{3}$$

The choice of (2) is motivated by the fact that the beta distribution is the conjugate prior associated to the negative binomial likelihood for the parameter π_m [21]. Although this is no more true of (3), it is nonetheless possible to use efficient simulation procedures based quasi-Gibbs updates as described in section III-C.

Because (3) is entirely symmetric, any permutation of the way the classes are numbered leave the posterior unchanged. As a consequence, the output of the MCMC sampler for a fixed value of M should really be interpreted as samples *from a set* of parameters $(\kappa_k, \pi_k)_{1\leq k\leq M}$ (for a fixed value of M). Following the suggestions of [22] and [23], when the classes need to be unambiguously identified (for instance for inference about the class dependent parameters), the sampler outcomes are ordered by post-processing using a classification rule. In the present work, a simple ordering based on the mean value $\mu_m = \kappa_m(1 - \pi_m)/\pi_m$ of each class was found to be sufficient (performing equivalently to the clustering approaches of [23] applied to μ_m and π_m jointly). The reason for this good behavior of such a simple classifier is that the classes are well separated with respect to their mean values when conditioning on plausible (and in particular not too large) values of the number of classes as shown in figure 9 (section IV). There are of course other options which include imposing identifiability constraints on the class parameters so as to ensure that the ordering of the classes is indeed defined unambiguously – see [24], [11], [18] and [25] for a more detailed account on this point.

We next assume that

$$\begin{aligned} P(K = k|M) &\propto \tau_K^{k-M} \quad \text{for } k \geq M, & (4)\\ P(b_2, \ldots, b_K|K) &= \begin{pmatrix} T-1 \\ K-1 \end{pmatrix}^{-1}, & (5)\\ P(l_1, \ldots, l_K|K, M) &= [M(M-1)^{K-1}]^{-1}, & (6) \end{aligned}$$

where $\propto$ means "proportional to" (up to the normalizing constant which ensures that the distribution sums to 1). The prior on K is geometric as for the number of class, where the constraint $K \geq M$ is imposed because the number of classes would necessarily be ill-defined in cases where there are fewer segments than classes. Note however that this constraint is not sufficient to prevent the appearance of empty classes in the course of the simulations. In practice, empty classes occur only rarely since their appearance is penalized trough the labelling prior given by (6). The geometric prior on M is intended to allow for large number of segments (which typically occurs when analyzing large sections of the data) with a high a priori uncertainty on the number of segments.

(5) corresponds to the assumption that the segments boundaries $b_2, \ldots b_K$ cover the available range of time indexes (2 to T) uniformly. The corresponding prior distribution is that of an ordered draw in $\{2, \ldots T\}$ *without replacement* because of the constraint $b_k > b_{k-1}$ for $k = 1, \ldots, K$ which guarantees that the segments are indeed well defined. Similarly, the label sequence has an uniform a priori distribution over all the $M(M-1)^{(K-1)}$ valid configurations, which are such that no adjacent segments share the same label.

[Figure 4 about here.]

An important point is that (4) and (5) correspond to a prior on the segmentation structure (conditionally on the number of classes) which is less informative than the Bernoulli [10], [4] or Markovian [5] priors commonly used in changepoint analysis. As noted by [5] when the hyperparameters are fixed, using a Markovian prior for the segmentation boundaries is equivalent to Hidden Markov modeling, which is computationally attractive [18]. However, the segment durations are then a priori distributed according to a geometric distribution which implies in particular that long segments are (a priori) unlikely. We thus follow the suggestion of [8] penalizing only the number of segments and not the segment pattern. Accordingly, this prior structure allows for much longer segments than the Markovian prior as illustrated in figure 4. Note that a different and interesting solution would consist of using a parametric assumption for the segment duration as in [9].

In the present study, τ_M, α_μ, β_μ, α_π, β_π and τ_K are treated as hyperparameters and are set to fixed values. Because we would like to be as noninformative as possible particularly for model characteristics that have a possible influence on the segmentation outcome, we choose to set τ_M and τ_K very close to one. It turns out that when τ_M and τ_K are greater than 0.9, their precise values have no significant influence on the results for the data considered in section IV, and that moreover, one can as well use $\tau_M = \tau_K = 1$. This seemingly counterintuitive result is interesting because it reveals the interplay between the prior hypotheses: For the data under consideration, the dimension penalty doesn't come from the priors of M and K but rather from those of the segment and label configurations given by (5) and (6). For the remaining parameters we selected the following values, $\alpha_\mu = 0.1$, $\beta_\mu = 1\text{e}-4$, $\alpha_\pi = 1$, $\beta_\pi = 1$ which correspond to distributions of μ and π that are distinctively more dispersed than the empirical distributions shown in figure 3-(b).

III. MCMC sampler

The MCMC sampling strategy consists of a systematic scan through five types of moves:

1. Updating the segment boundaries
2. Updating the segment labels
3. Creating or removing segments
4. Updating the parameters of each class
5. Modifying the number of classes

Move 2 and 4 use standard Gibbs and/or Metropolis-Hastings updating proposals. Move 1 is also of Gibbs type, following the suggestion of [3] for general changepoint models. The two remaining moves (3 and 5) make use of the reversible jump Metropolis-Hastings scheme introduced by Green [8]. Move 3, which consists in modifying the segmentation by adding or removing one or two segments, is technically comparable to the solutions used in [10] or [9] for analyzing ion channel signals with unknown segmentation (although the latter paper uses continuous change locations). Move 5 which aims at modifying the number of classes (by splitting one class in two or merging together two different classes) is more involved because it necessarily implies a simultaneous modification of the number of classes and of the number of segments.

An interesting computational remark here is that the log-likelihood of i.i.d. negative binomial observations can be computed in two different ways:

$$\log p(n_1, \dots, n_T | \kappa, \pi) = T\kappa \log(\pi) + S\log(1-\pi) - T\log\left(\Gamma(\kappa)\right) + \sum_{t=1}^{T} \log\left(\Gamma(n_t + \kappa)\right), \tag{7}$$

where $S = \sum_{t=1}^{T} n_t$, or

$$\log p(n_1, \dots, n_T | \kappa, \pi) = T\kappa \log(\pi) + S\log(1-\pi) + \sum_{r=1}^{R} C_r \log(\kappa + r - 1), \tag{8}$$

where $R = \max\{n_1, \dots, n_T\}$, $C_r = \#\{1 \leq t \leq T : n_t \geq r\}$ and with the convention that the sum is null if $R = 0$ (that is if all counts are zero). Eq. (8) is very efficient when the observed counts are small (i.e. when $R_T \ll T$), especially when it is needed to evaluate the log-likelihood for several configurations of the parameters κ and π because the rank statistics C_r are computed only once. On the other hand, (7) should be systematically preferred when the observed counts are large (a few hundreds or more) which usually makes the computation of the rank statistics C_r (for $r = 1$ to R) very penalizing. In the rest of the paper, we assume that the observed counts exceed only rarely a few hundred and thus use the form of Eq. (8) whenever the associated computation load can be expected to be lower. Note that in practice, as the length and position of the segments are both unknown and variable, it is more efficient to check for each segment if either the length or the maximum observed count is greater so as to choose between (7) or (8).

A. *Updating the segment boundaries*

The segment boundaries $b_2, \dots b_K$ are updated using a systematic scan Gibbs move. The full conditional distribution for b_k is given by

$$\begin{aligned} P(b_k = t_0 | \cdots) \;\propto\; & \left\{ \prod_{t=b_{k-1}}^{t_0-1} (1-\pi_{l_{k-1}})^{n_t} \Gamma(n_t + \kappa_{l_{k-1}}) \right\} \left(\frac{\pi_{l_{k-1}}^{\kappa_{l_{k-1}}}}{\Gamma(\kappa_{l_{k-1}})} \right)^{(t_0-b_{k-1})} \\ & \times \left\{ \prod_{t=t_0}^{b_{k+1}-1} (1-\pi_{l_k})^{n_t} \Gamma(n_t + \kappa_{l_k}) \right\} \left(\frac{\pi_{l_k}^{\kappa_{l_k}}}{\Gamma(\kappa_{l_k})} \right)^{(t_0-b_k)} \end{aligned} \tag{9}$$

for $b_{k-1} < t_0 < b_{k+1}$.

B. *Updating the segment labels*

Here again a systematic Gibbs move is used, where the full conditional distribution for l_k is

$$P(l_k = m | \cdots) \propto \pi_m^{N^{(k)}\kappa_m} (1-\pi_m)^{S^{(k)}} \prod_{r=1}^{R^{(k)}} (\kappa_m + r - 1)^{C_r^{(k)}} \times \mathbb{I}_{m \neq l_{k-1}} \mathbb{I}_{m \neq l_{k+1}} \tag{10}$$

where $N^{(k)}$ is the number of data points in segment k (that is $N^{(k)} = b_{k+1} - b_k$), $S^{(k)}$ the sum of these points, $R^{(k)}$ the maximum value, $C_r^{(m)}$ are the rank statistics (the number of points greater or equal to r) and $\mathbb{I}$ denote the indicator function. Note that for the first ($k = 1$) and last ($k = K$) segments, only one of the two constraints in (10) is active because there are no left (resp. right) adjacent segment. Because of this model constraint, that no neighboring labels should be alike, the above Gibbs scheme is clearly not applicable when there are only two classes ($M = 2$) because all moves would be rejected. For this particular case however, there are only two valid complete label sequences whatever the number of segments. Thus when $M = 2$, the complete sequence $(l_1, \dots, l_K)$ is drawn directly in a block.

C. *Updating the parameters of each class*

The parameters $\pi_1, \dots, \pi_M$ are conditionally independent with full conditional distribution given by

$$p(\pi_m | \cdots) \propto \mathrm{Beta}(\pi_\mathrm{m} | \kappa_\mathrm{m} \bar{\mathrm{N}}^{(\mathrm{m})} + \alpha_\pi, \bar{\mathrm{S}}^{(\mathrm{m})} + \beta_\pi)\, \mathrm{Gamma}(\kappa_\mathrm{m} | \alpha_\mu, \beta_\mu / \beta(\pi_\mathrm{m})) \tag{11}$$

where $\bar{N}^{(m)}$ is the number of data points classified within class m and $\bar{S}^{(m)}$ denotes the sum of these points. Where the phrase "classified within class m" should be interpreted as belonging to a segment whose label is m (contrary to the corresponding quantities in (10) which are computed from a single segment of data).

The first term in (11) corresponds to the product of the likelihood by the marginal prior on π_m while the second term corresponds to the prior on κ_m given π_m. In practical situations (for noninformative analysis at least), the variations of the second term with π_m are rather small. An efficient simulation procedure thus consist of using a Metropolis Hastings step where the proposed update π_* is distributed according to a $\mathrm{Beta}(\pi_*|\kappa_\mathrm{m}\bar{\mathrm{N}}^{(\mathrm{m})}+\alpha_\pi, \bar{\mathrm{S}}^{(\mathrm{m})}+\beta_\pi)$ distribution and accepted with probability $\min(1,A)$ where

$$A = (\beta(\pi_*)/\beta(\pi_m))^{-\alpha_\mu}\, \mathrm{e}^{-\kappa_\mathrm{m}(\beta_\mu/\beta(\pi_*)-\beta_\mu/\beta(\pi_\mathrm{m}))}$$

Note that with the choice of the hyperparameters α_μ and β_μ made in section IV, the accept/reject correction is almost unneeded as the rejection rate is very much less than 1 percent.

The full conditional distribution for κ_m is

$$p(\kappa_m|\cdots) \propto (\kappa_m)^{\alpha_\mu-1}\left\{\prod_{r=1}^{\bar{R}^{(m)}}(\kappa_m+r-1)^{\bar{C}_r^{(m)}}\right\}\mathrm{e}^{-[\beta_\mu/\beta(\pi_\mathrm{m})+\bar{\mathrm{N}}^{(\mathrm{m})}\log(1/\pi_\mathrm{m})]\kappa_\mathrm{m}} \tag{12}$$

where $\bar{R}^{(m)}$ denotes the maximum value of the data points classified within class m, and $\bar{C}_r^{(m)}$ are the corresponding rank statistics. Empirically the full conditional given by (12) appears to be closely fitted by a Gamma distribution. To take profit of this remark we proceed as in [26] by using a single step of the Metropolis-Hastings algorithm with a Gamma proposal tuned to match the mode and the log-curvature of the full conditional. Differentiation of (12) yields

$$\frac{d\log p(\kappa_m|\cdots)}{d\kappa_m} = -(\beta_\mu/\beta(\pi_m)+\bar{N}^{(m)}\log(1/\pi_m)) + \frac{\alpha_\mu-1}{\kappa_m} + \sum_{r=1}^{\bar{R}^{(m)}}\frac{\bar{C}_r^{(m)}}{\kappa_m+r-1} \tag{13}$$

$$\frac{d^2\log p(\kappa_m|\cdots)}{d\kappa_m^2} = -\left(\frac{\alpha_\mu-1}{(\kappa_m)^2} + \sum_{r=1}^{\bar{R}^{(m)}}\frac{\bar{C}_r^{(m)}}{(\kappa_m+r-1)^2}\right) \tag{14}$$

This second expression indicates that the logarithm of the full conditional distribution is a strictly convex function if $\bar{R}^{(m)} \geq 1$. As in [26], we thus use (13)- (14) the following way:

- Starting from the moment estimate $\kappa_m = \bar{S}^{(m)}/\bar{N}^{(m)}\pi_m/(1-\pi_m)$, perform a few Newton steps to find the mode κ (in the following simulations, only one iteration is used).
- Compute the log-curvature at the mode $w = -\frac{d^2\log p(\kappa|\cdots)}{d\kappa^2}$ according to (14).
- Compute the parameters of a Gamma distribution with mode and log-spread matched to κ an w with parameters $\alpha = 1+\kappa^2 w$ and $\beta = \kappa w$
- Use a $\mathrm{Gamma}(\alpha,\beta)$ distributed proposal κ_* which is accepted with probability $\min(1,A)$ where

$$A = \left(\frac{\kappa_*}{\kappa_m}\right)^{\alpha_\mu-\alpha}\prod_{r=1}^{\bar{R}^{(m)}}\left(\frac{\kappa_*+r-1}{\kappa_m+r-1}\right)^{\bar{C}_r^{(m)}}\mathrm{e}^{-[\beta_\mu/\beta(\pi_\mathrm{m})+\bar{\mathrm{N}}^{(\mathrm{m})}\log(1/\pi_\mathrm{m})-\beta](\kappa_*-\kappa_\mathrm{m})} \tag{15}$$

In practice, the probability of rejection for the above proposal scheme is about 1-2% which reflects the fact that (12) is closely matched to the $\mathrm{Gamma}(\alpha,\beta)$ distribution.

D. Creating or removing segments

We now come to more elaborate moves which modify the number of model parameters, beginning with K, the number of segments. The move used to update this parameter is a straightforward instance of the general reversible jump approach [8], notably because the dimension varying parameters (number of segments) are discrete, which makes the evaluation of the proposal ratio straightforward. This type of move is thus briefly described in its simplest version.

The "split or merge" mechanism for drawing K starts by selecting at random either the split ($K \to K+1$) or the merge ($K \leftarrow K+1$) alternative. We first consider merging two consecutive segments assuming that the current number of segments is $K+1$:

1. Draw a segment k in $\{1, \ldots, K\}$ with probability $1/K$ (merging will be performed on the segments numbered k and $k+1$),
2. Draw the label l'_k of the merged segment uniformly in $\{1, \ldots, M\}$,
where the quantities denoted with a prime pertain to the lower dimension K. The merge proposal is systematically rejected at this point if either $l'_k = l_{k-1}$ or $l'_k = l_{k+2}$.

Assuming that the current number of segments now is K, the reverse (split) proposal consists of:
1. Draw a segment k in $\{1, \ldots, K\}$ with probability

$$p(k) = (b_{k+1} - b_k - 1)/(T - K),$$

2. Draw a new sub-segment boundary b'_{k+1} in $\{b_k + 1, \ldots, b_{k+1} - 1\}$ with probability $1/(b_{k+1} - b_k - 1)$,
3. Draw independently in $\{1, \ldots, M\}$ two labels for the new sub-segments,
where the quantities denoted with a prime sign now refer to the highest dimension $(K+1)$. The split proposal is systematically rejected if any two successive labels in the sequence $(l_{k-1}, l'_k, l'_{k+1}, l_{k+1})$ are identical. In step 1 above, the use of uneven probabilities was found to be much valuable in making the selection of longer segments more likely as well as preventing the selection of eventual segments of length 1 which cannot be split any further.

The Metropolis-Hastings acceptance probability for the split move is then given by $\min(1, A)$ where

$$A = \text{likelihood ratio} \times \underbrace{\tau_K \frac{K}{(T-K)(M-1)}}_{\textbf{(A1)}} \times \underbrace{M \frac{(T-K)}{K}}_{\textbf{(A2)}} \tag{16}$$

where **(A1)** corresponds to the prior ratio (when going from K to $K+1$ segments), and **(A2)** to the proposal ratio. Note that in computing the likelihood ratio according to (10) it is only necessary to take into account the part of the data whose label is changing as a consequence of the split move. As usual, the reverse move (merge) is accepted with probability $\min(1, A^{-1})$ where A is defined as in (16) (which follows from the remark that after a split move, the quantities b_k, b'_{k+1}, b_{k+1} will be reindexed as b_k, b_{k+1}, b_{k+2}).

As noted by [9] and , the above move becomes inefficient when the number M of classes is small. For the model under consideration, it is easily verified that the previously described split move will be systematically rejected when $M = 2$ except if the splitting occurs for the very first $(k = 1)$ or last $(k = K)$ segment. The solution proposed by [9] and [10] to overcome this limitation consists of devising an "insert or delete" proposal scheme which increases by two the number of segment boundaries by insertion of a new segment in the middle of an existing segment, or conversely reducing by two the number of segment boundaries by merging three consecutive segments together. We adopt a similar solution which is proposed randomly in place of the "split or merge" proposal. The details of the corresponding proposal are omitted since they are similar to the simpler "split or merge" mechanism discussed above. This modification is however only required when M is small (two or three) and could thus be omitted without noticeably reducing the sampler's mixing for the data considered in section IV.

E. Modifying the number of classes

Simulation of the number of classes M is by far the most complex task because a modification of M may imply a complete redefinition of all the latent structure, and in particular of the segmentation.

[Figure 5 about here.]

Indeed, consider the case of figure 1 and assume that a move from dimension $M = 3$ to dimension $M = 2$ is to be proposed following the merge strategy previously adopted. The latent structure obtained after appropriate relabelling and removal of the obsolete segments (those which separate regions of the data corresponding to the same new label) is shown in figure 5. Note that the number of segments has been reduced from seven to three, segments 1 to 3 (resp. 6 to 7) now being merged together in segment 1' (resp. 3'). This process is clearly dependent on the choice of the classes to be merged as grouping classes 2 and 3 would not have required any modification of the segmentation. The most challenging task is of course not going from figure 1 to figure 5, but rather consists of ensuring that the converse move has a non zero probability of being proposed. The following remark proves to be valuable for this purpose: When moving from the configuration of figure 1 to that of 5, it is possible to consider that the only dimension changing parameters are the class parameters $(\kappa_m, \pi_m)_{m=1,\ldots,M}$. This is simply a consequence of the fact that whatever the number K of segments, the boundaries $\{b_k\}_{1\leq k\leq K+1}$ and the labels $\{l_k\}_{1\leq k\leq K}$ can be equivalently reparameterized by an equivalent set of fixed dimension parameters, namely the class indicators $\{i_t\}_{1\leq t\leq T}$ (such that $i_t = l_k$ if $b_k \leq t < b_{k+1}$). This latter parameterization is not convenient for actually simulating the latent structure and is thus not used in the present contribution, but its mere existence shows that the part of the latent structure which pertains to the data segmentation may or may not, depending on what's most convenient, be considered as dimension varying data.

The proposal mechanism once again randomly selects between two alternatives which correspond respectively to merging two classes together and to splitting a single class apart.

We first consider the merge move and denote by $(M+1)$ the current number of classes (as previously, quantities indicated by a prime pertain to the lower dimension which prevails once the merge move has been completed):

1. Draw the index m of the first class to be merged in $\{1,\ldots,M\}$ with probability $1/M$ (the two successive classes indexed by m and $m+1$ will be merged). Let L_m and L_{m+1} denote the number of segments associated with each class. The proposal is rejected at this point if both classes are empty (*i.e.* $L_m = L_{m+1} = 0$).

2. Compute the parameters of the merged class according to

$$\kappa'_m = \sqrt{\kappa_m \kappa_{m+1}} \tag{17}$$

$$\pi'_m = \pi\left(\sqrt{\beta(\pi_m)\beta(\pi_{m+1})}\right) \tag{18}$$

where $\beta(x)$ and $\pi(x)$ denote the reparameterization transformations defined by (1). The intermediate reparameterization shown in (18) simply guarantees that the transformed parameters π'_m lies in the valid range $(0,1)$.

Note that in accordance with figure 5, merging two classes induces, in most cases, a reduction of the number of segments to K', and thus necessitates a renumbering of the labels. Finally, the proposal is systematically rejected if $K' < M$.

Now, assuming that the current number of classes is M, the reverse move consists of:

1. Draw the index m of the class to be split in $\{1,\ldots,M\}$ with probability L_m/K, where L_m is the number of segments with label m.

2. For each segment $k_1,\ldots k_{L_m}$ with label m,

(a) Draw the number of additional sub-segments H_{k_i} with a $\mathrm{Geom}(\tau_{\mathrm{B}}(\mathrm{b}_{\mathrm{k_i}+1} - \mathrm{b}_{\mathrm{k_i}})/\mathrm{T})$ distribution (truncated to $b_{k_i+1} - b_{k_i} - 1$).

(b) Draw the new sub-segment's frontiers uniformly if needed (if $H_{k_i} \geq 1$) with probability

$$1/\binom{b_{k_i+1} - b_{k_i} - 1}{H_{k_i}}$$

(c) Draw one of the two valid sequences of the labels m' and $(m+1)'$ with probability 1/2.

Let L'_m and L'_{m+1} denote the number of segments associated to each of the two new classes as a result of this random sub-segmentation.

3. Draw two positive perturbations ρ_κ and ρ_π according to a $\mathrm{Gamma}(\mathrm{g}_\kappa, \mathrm{g}_\kappa)$ distribution, and compute

the new class parameters according to

$$\kappa'_m = \kappa_m \rho_\kappa \tag{19}$$
$$\kappa'_{m+1} = \kappa_m / \rho_\kappa \tag{20}$$
$$\pi'_m = \pi(\beta(\kappa_m)\rho_\pi) \tag{21}$$
$$\pi'_{m+1} = \pi(\beta(\kappa_m)/\rho_\pi) \tag{22}$$

As in the case of the segment splitting move, the inclusion of L_m (the number of segments with label m) in step 1 of the class splitting move avoids splitting empty classes and favors (to some extent) split moves that concerns classes which are representative of a large number of segments.

The class splitting move is accepted with probability $\min(1, A)$ where

$$\begin{array}{rcl}
A & = & \text{likelihood ratio} \\
\textbf{(A1)} & \times & \tau_M \tau_K^{(K'-K)} \dfrac{\begin{pmatrix} T-1 \\ K-1 \end{pmatrix} M(M-1)^{(K-1)}}{\begin{pmatrix} T-1 \\ K'-1 \end{pmatrix} (M+1)M^{(K'-1)}} \\
\textbf{(A2.1)} & \times & \dfrac{\beta_\mu^{\alpha_\mu}}{\Gamma(\alpha_\mu)} \left(\dfrac{\beta(\pi'_m)\beta(\pi'_{m+1})}{\beta(\pi_m)} \right)^{-\alpha_\mu} \left(\dfrac{\kappa'_m \kappa'_{m+1}}{\kappa_m} \right)^{\alpha_\mu - 1} \mathrm{e}^{-\beta_\mu \left(\frac{\kappa'_\mathrm{m}}{\beta(\pi'_\mathrm{m})} + \frac{\kappa'_\mathrm{m+1}}{\beta(\pi'_\mathrm{m+1})} - \frac{\kappa_\mathrm{m}}{\beta(\pi_\mathrm{m})} \right)} \\
\textbf{(A2.2)} & \times & \dfrac{\Gamma(\alpha_\pi + \beta_\pi)}{\Gamma(\alpha_\pi)\Gamma(\beta_\pi)} \left(\dfrac{\pi'_m \pi'_{m+1}}{\pi_m} \right)^{\alpha_\pi - 1} \left(\dfrac{(1-\pi'_m)(1-\pi'_{m+1})}{1-\pi_m} \right)^{\beta_\pi - 1} \\
\textbf{(A3)} & \times & \dfrac{K}{M L_m} \displaystyle\prod_{i=1}^{L_m} \dfrac{2}{1 - \tau_B (b_{k_i+1} - b_{k_i})/T} \left(\dfrac{T}{\tau_B (b_{k_{i+1}} - b_{k_i})} \right)^{H_{k_i}} \begin{pmatrix} b_{k_{i+1}} - b_{k_i} - 1 \\ H_{k_i} \end{pmatrix} \\
\textbf{(A4)} & \times & 4\, \kappa'_{m+1} \beta(\pi'_{m+1}) \dfrac{(1-\pi'_m)^2 (1-\pi'_{m+1})^2}{(1-\pi_m)^2} \\
\textbf{(A5)} & \times & \left(\dfrac{g_\kappa^{g_\kappa}}{\Gamma(g_\kappa)} \rho_\kappa^{(g_\kappa - 1)} \mathrm{e}^{-\mathrm{g}_\kappa \rho_\kappa} \dfrac{\mathrm{g}_\pi^{\mathrm{g}_\pi}}{\Gamma(\mathrm{g}_\pi)} \rho_\pi^{(\mathrm{g}_\pi - 1)} \mathrm{e}^{-\mathrm{g}_\pi \rho_\pi} \right)^{-1}
\end{array}$$

where ρ_κ and ρ_π may be equivalently computed from the class parameters according to

$$\rho_\kappa = \sqrt{\kappa'_m / \kappa'_{m+1}},$$
$$\rho_\pi = \sqrt{\beta(\pi'_m)/\beta(\pi'_{m+1})}.$$

The above acceptance ratio decomposes into the following terms: **(A1)** includes the priors on M, K, $\{b_k\}_{2 \le k \le K}$ and $\{l_k\}_{1 \le k \le K}$; **(A2.1)** and **(A2.2)** correspond to the prior ratio for the parameters of the split class; **(A3)** defines the proposal ratio, with the product term corresponding more specifically to the probability of proposing a particular relabeling when performing the split move; **(A4)** is the Jacobian of the transformation from $(\kappa_m, \pi_m, \rho_\kappa, \rho_\pi)$ to $(\kappa'_m, \pi'_m, \kappa'_{m+1}, \pi'_m)$; Finally, **(A5)** is the inverse of the joint probability density of the random perturbations ρ_κ and ρ_π.

The parameters of the sampler are g_κ, g_π and τ_B. The larger the values of g_κ and g_π, the smaller the perturbations brought to the class parameters during a split move. On the other hand if g_κ and g_π are too large, the presence of the factor **(A5)** in the acceptance ratio will practically prevent acceptance of the merge moves. A value of 100 is used both for g_κ and g_π for all the simulations. We use $\tau_B = 0.9$ that is a value of the order of that of τ_K since τ_B defines the law of the number of additional sub-segments when splitting a hypothetical segment of length T (that is, covering all the data). For shorter segments however, the probability of inserting new sub-segments during a split move steadily decreases as a consequence of the geometric proposal rate for H_k being set to $\tau_B(b_{k+1} - b_k)/T$ which renders the expected number of additional sub-segments proportional to the relative length of the split segment. The reversible jump moves of sections III-D and III-E are arguably less efficient than those used to update the model conditional parameters described in sections III-A- III-C. This is a consequence both of the fact that the moves of sections III-A- III-C indeed correspond to full cycles through all the segments or all the classes, and also that the reversible jump proposals of sections III-D-III-E do not enforce the constraints on the parameters so that a fraction of them are rejected irrespectively of the change in likelihood. Hence, at each iteration, 5 reversible jump moves are attempted both for the number of segments (section III-D) and for the number of classes (section III-E) whereas only one instance of the moves described in sections III-A- III-C is performed. With these sampler settings, the reversible jump acceptance rate on the data considered in the next section varies from 10 to 20% for the moves affecting the number of segments K and are a few percent for the moves which modifies the number of classes M (see section IV-B).

IV. Analysis of traffic data

In this section, we analyze a short section of a traffic trace available from the Internet Traffic Archive which was first described by Paxson and Floyd (1995) (trace labelled "LBL-TCP-3"). This trace captures two hours of TCP (Transmission Control Protocol) activity measured over a wide area Internet gateway.

The raw data consists of a collection of information of a different nature which includes the size of the transferred data packets (because TCP is a variable size protocol), the source and destination addresses, the type of the packet and finally a time stamp for each packet. We refer the reader to [13] for known observations considering this traffic trace and to [14], [15] for discussion of the different ways in which such a traffic trace may be analyzed. In the following, we simply consider the one second aggregated TCP packet counts (number of packets transmitted during one second) measure over a six minutes (360 data points) period.

A. Results

For this data record displayed in figure 8, we consider estimation results obtained by Monte Carlo averages computed from 800 000 iterations of a single instance of the Markov chain sampler described in section III.

[Figure 6 about here.]

[Figure 7 about here.]

[Figure 8 about here.]

Fig. 6 shows that for the data section under consideration, there is significant evidence in favor of the model corresponding to $M = 4$, with a posterior probability of $M = 0.71$ for the four classes model. Interestingly, the hypothesis of a simple ON/OFF model (with only two classes) is not supported at all by the data. In the following, we thus only consider results conditional upon the value $M = 4$. Figure 7 shows that for the number of segments (conditional upon $M = 4$), the picture is less clear cut and that a point-wise estimate of the number of segment (say $K = 35$ which corresponds to the mode) should be considered as poorly reliable. On the other hand, figure 8 which is obtained by averaging the number of times each label is associated to a particular data point shows that the segmentation model is well supported by the data with the different data segments clearly separated. As the classes are ordered by their mean value μ_m, the high activity regions appear at the bottom of the plot (the corresponding distribution estimates are plotted in figure 9). The burstiness of the data which explains the sparseness of the plot corresponding to high activity class (bottom plot in figure 8) is also a feature revealed by the analysis.

[Figure 9 about here.]

[Figure 10 about here.]

Figures 9 and 10 display the class related summaries with both the posterior distribution of the class-dependent parameters (figure 9) and the class conditional density estimates (figure 10). Figure 9 is an histogram of the simulated class parameters while figure 10 is obtained by averaging the probability density functions corresponding to the four classes for all values of these parameters. Figure 9 shows that there is indeed a good separation of the class parameters based on the means μ_m (left column plots). Another interesting feature of figure 9 is the fact that the classes are more dispersed (i.e. with high values of γ_m) when considering the higher activity levels (classes 3 and 4). Even if this finding is clearly due to a lack of evidence in the case of the fourth class, this trend seems significant when comparing, for example, classes 2 and 3. With values of γ_m of the order of five or more, the obtained class conditional distributions are distinctively more dispersed than one would expect from a Poisson assumption.

B. Mixing issues

[Figure 11 about here.]

[Figure 12 about here.]

An important issue associated with the use of MCMC methods is convergence of the Monte Carlo estimations. Even if we lack space to cover this question in much details (see [27] for a discussion of objective methods for assessing convergence), it is interesting to comment figure 11 which shows the convergence of the posterior probability estimates for the number of classes M as a function of the number of sampler iterations (figure 12 displays a similar picture concerning the number of segments K for 8 quantiles of the empirical posterior of K). Both plots suggest that the MCMC sampler has reached stationarity after 500 000 iterations or so, although the posterior distribution of M still undergoes slight modifications (figure 11) beyond that point. This order of magnitude, which may seem considerable, is indeed fairly common in the MCMC literature [27]. Granted that the model under consideration is quite complex, figures 11 and 12 shows that the proposed sampler is quite efficient for problems of this scale.

Comparing figures 11 and 12 nevertheless indicates that the number of segments (figure 12) tends to stabilize much faster than the number of classes (figure 11). Experiments carried out for longer sections of the data revealed that this problem becomes more salient as the number of segments increases. When the number of segments is larger than one hundred, the number of classes does not change anymore in the course of the iterations. This *mixing problem* (very slow convergence of one of the component of the chain) is certainly a limitation of the method which will be hard to raise (see [9] for a similar finding in a related application). This limitation is not just due to a failure of the sampling strategy but rather reveals the complexity of estimating the number of classes when they are many segments: For $K = 100$ segments for instance, they are about 3.10^{12} more configurations with $M = 5$ classes than with $M = 4$ classes (which also means that when attempting to split from $M = 4$ to $M = 5$ classes, the rightmost part of ratio **(A1)** will be the inverse of that figure – that is extremely small).

C. Analysis of a longer section of data

Analysis of slightly longer sections of data is nonetheless possible and provides interesting insights about the tradeoff between complexity and accurate representation of the data which is achieved by the method. Figures 13 and 14 are the analogous of figures 6 and 8 for a section of data which is twice longer (12 mn, 720 data points). The segment of data considered in section IV-A and plotted in figure 8 corresponds to the right half of figure 14 (right of the vertical dashed line).

[Figure 13 about here.]

[Figure 14 about here.]

The most striking difference when comparing with figure 6 is that the posterior for the number of classes (figure 13) is now more spread out and suggests five or six as the most likely number of classes. The posterior segmentation, conditional on the number of class being $M = 5$, shown in figure 14 indicates that some features are remarkably stable: In particular, focusing on the right part of figure 14, one clearly sees that the fourth and fifth classes in figure 14 corresponds to, respectively, the third and fourth ones in figure 6. Class 3 in figure 14 is also comparable to class 2 in figure 6 (which is confirmed by looking more closely at the estimated values of the class conditional parameters). Finally, class 1 in figure 6 has been splitted up into classes 1 and 2 in figure 14, with the first one (corresponding to the lower mean level) which is mostly selected in the first half of figure 14, that is for the data which was not included in the analysis carried out in section IV-A. The main message here is thus that modeling larger sections of data requires more degrees of freedom, in particular in terms of the possible marginal distributions. This is coherent with the fact that both halves of the data in figure 13 look qualitatively very different.

[Figure 15 about here.]

An interesting point is that the locations of the segment boundaries appear to be very stable. This is also confirmed by figure 15 which shows the posterior for the presence of a segment boundary averaged over all model configurations (including the number of classes and segments as well as the class conditional parameters) for the data shown in figure 14. The fact that figure 15 still shows very well located change points despite the fact that we marginalize over very different models indicate that the presence of abrupt changes is well supported by the data. The most likely number of segments is now $K = 74$, that is slightly less than the twice the mode of figure 7, which is coherent with the fact that there is less activity, and correlatively less changes, in the first half of the section of data shown in figure 14 than in the second one.

Figure 14 also shows that some components like the fifth one have very sparse and irregular time patterns which would be hard to model using simple assumptions like Markovian dependence. Finally, figure 14 clearly indicates that the complexity required to model the behavior of the data depends on the time horizon considered and that the results obtained on limited sections (number of classes in particular) cannot be extrapolated.

V. Conclusions

We have presented a novel approach for the analysis of discrete-time count data which is based on Bayesian modeling and Markov Chain Monte Carlo (MCMC) simulation. We hope to have convinced the reader with the teletraffic example covered in section IV that this approach provides non trivial insightful results when applied to real data. One advantage of the Bayesian approach in this setting are its visual and easily interpretable results such as figures 8 and 10 which may be used to asses the goodness of fit of the model and also suggest possible improvements and/or simplifications. Speed of convergence is certainly a concern for MCMC methods in general, and in the case under study, there was shown to be a practical limit to the complexity of the models that can be handled with the proposed sampling strategy.

Adaptation of this approach to very large scale problems (hundreds of segments and more) is thus an interesting and open question for future research. Solutions that could be considered include tempering schemes [9] and parallel simulations (for models with different number of classes) as advocated by [28]. Further comparison with methods based on variable order hidden Markov modeling, and in particular with techniques which do not rely on data augmentation (ie. the boundaries b_k and labels l_k need not be simulated when proposing dimension changing moves) as in [29] is also of great interest.

References

[1] J. Bai and P. Perron, "Estimating and testing linear models with multiple structural changes," *Econometrica*, vol. 66, pp. 47–78, 1998.

[2] D. Barry and J. A. Hartigan, "A Bayesian analysis for change point problems," *J. Amer. Statist. Assoc.*, vol. 88, no. 421, pp. 309–319, 1993.
[3] D. A. Stephens, "Bayesian retrospective multiple changepoint identification," *Appl. Statist.*, vol. 43, pp. 159–178, 1994.
[4] M. Lavielle and E. Lebarbier, "An application of MCMC methods for the multiple change-points problem," *Signal Processing*, vol. 81, no. 1, pp. 39–53, january 2001.
[5] S. Chib, "Estimation and comparison of multiple change point models," *Journal of Econometrics*, vol. 86, pp. 221–241, 1998.
[6] I. L. MacDonald and W. Zucchini, *Hidden Markov models and other models for discrete-valued time series*, Chapman & Hall, 1997.
[7] C. P. Robert, T. Rydén, and D. M. Titterington, "Bayesian inference in hidden Markov models through jump Markov chain Monte Carlo," *J. Royal Statist. Soc. Ser. B*, vol. 62, no. 1, pp. 57–75, 2000.
[8] P. J. Green, "Reversible jump Markov chain Monte Carlo computation and bayesian model determination," *Biometrika*, vol. 82, no. 4, pp. 711–732, 1995.
[9] M. E. A. Hodgson, "A Bayesian restoration of an ion channel signal," *J. Royal Statist. Soc. Ser. B*, vol. 61, pp. 95–114, 1999.
[10] J. A. Stark, W. J. Fitzgerald, and S. B. Hladky, "Multiple-order Markov chain Monte Carlo sampling methods with application to a changepoint model," Tech. Rep. CUED/F-INFENG/TR 302, Departments of Engineering and Pharmacology, University of Cambridge, 1997.
[11] S. Richardson and P. J. Green, "On the Bayesian analysis of mixture with an unknown number of components," *J. Royal Statist. Soc. Ser. B*, vol. 59, no. 4, pp. 731–792, 1997.
[12] W. E. Leland, M. S. Taqqu, W. Willinger, and D. V. Wilson, "On the self-similar nature of ethernet traffic," *Computer Communication Review*, vol. 23, pp. 183–193, 1993.
[13] V. Paxson and S. Floyd, "Wide-area traffic: The failure of Poisson modeling," *IEEE/ACM Transactions on Networking*, vol. 3, pp. 226–244, 1995.
[14] W. Willinger and V. Paxson, "Discussion of the paper by S. I. Resnick," *Annals of Statistics*, vol. 25, no. 5, pp. 1856–1866, 1997.
[15] M. E. Crovella and A. Bestravros, "Self-similarity in world wide web traffic: Evidence and possible causes," *IEEE/ACM Transactions on Networking*, vol. 5, no. 6, pp. 835–846, 1997.
[16] W. Willinger, M. S. Taqqu, R. Sherman, and D. V. Wilson, "Self-similarity through high variability: statistical analysis of ethernet LAN traffic at the source level," *Computer Communication Review*, vol. 25, pp. 100–113, 1995.
[17] E. Arjas and J. Heikkinen, "An algorithm for nonparametric Bayesian estimation of a Poisson intensity," *Computational statistics*, vol. 12, no. 3, pp. 385–402, 1997.
[18] C. P. Robert and M. Titterington, "Reparameterisation strategies for hidden Markov models and Bayesian approaches to maximum likelihood estimation," *Statistics and Computing*, vol. 8, no. 2, pp. 145–158, 1998.
[19] J. Grandell, *Mixed Poisson Processes*, Chapman & Hall, 1997.
[20] J. Diebolt and C. P. Robert, "Estimation of finite mixture by Bayesian sampling," *J. Royal Statist. Soc. Ser. B*, vol. 56, pp. 363–375, 1994.
[21] A. Gelman, J. B. Carlin, H. S. Stern, and D. B. Rubin, *Bayesian data analysis*, Chapman & Hall, 1995.
[22] M. Stephens, "Discussion of the paper by Richardson & Green," *J. Royal Statist. Soc. Ser. B*, vol. 59, no. 4, pp. 731–792, 1997.
[23] G. Celeux, "Bayesian inference for mixture: The label switching problem," in *COMPSTAT 1998 Proceedings in Computational Statistics*, R. Payne and P. Green, Eds. 1998, pp. 227–232, Physica-Verlag.
[24] G. Celeux, M. Hurn, and C. P. Robert, "Computational and inferential difficulties with mixture posterior distributions," *J. Amer. Statist. Assoc.*, vol. 95, pp. 957–970, 2000.
[25] M. Stephens, "Dealing with label-switching in mixture models," *J. Royal Statist. Soc. Ser. B*, vol. 62, pp. 795–809, 2000.
[26] S. Chib, E. Greenberg, and R. Winkelmann, "Posterior simulation and Bayes factors in panel count data models," *Journal of Econometrics*, vol. 86, pp. 33–54, 1998.
[27] C. P. Robert and G. Casella, *Monte Carlo statistical methods*, Springer, 1999.
[28] S. Chib and I. Jeliazkov, "Marginal likelihood from the Metropolis-Hastings output," *J. Amer. Statist. Assoc.*, vol. 96, pp. 270–281, 2001.
[29] O. Cappé, C. Robert, and T. Rydén, "Reversible jump MCMC converging to birth-and-death MCMC and more general continuous time samplers," Submitted, 2001.

List of Figures

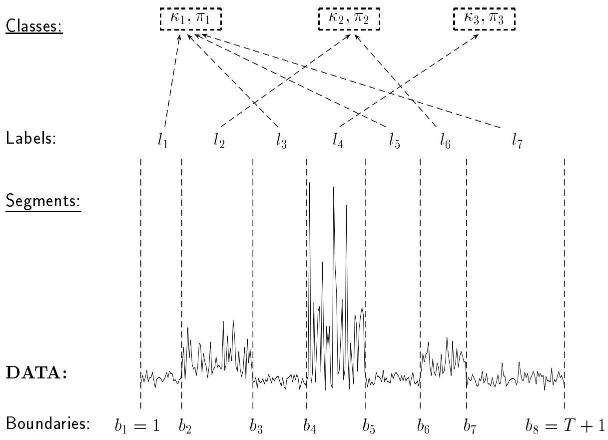

Fig. 1. Example of latent model structure with three classes ($M = 3$) and seven segments ($K = 7$).

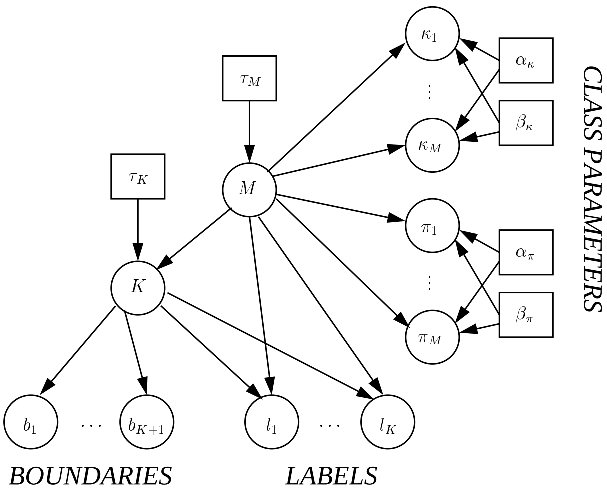

Fig. 2. Graphical representation of the prior structure.

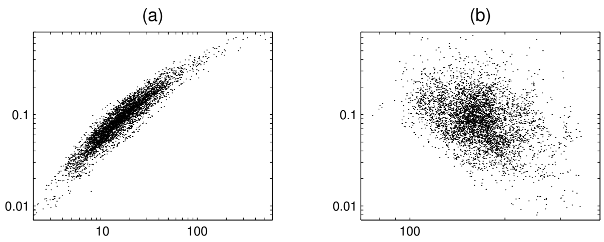

Fig. 3. Negative binomial parameters fitted on 5000 random sections of the data of section IV: **(a)** π as a function of κ; **(a)** π as a function of μ.

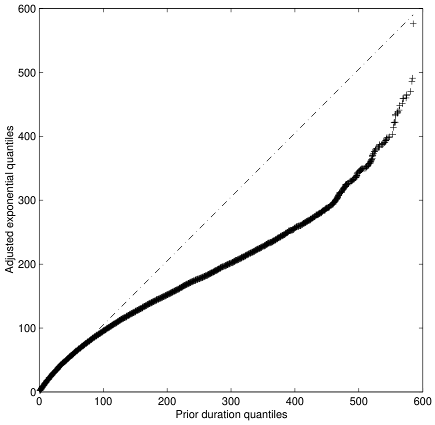

Fig. 4. Q-Q plot of the empirical a priori distribution of the segment duration versus the adjusted exponential distribution (10 000 draws of the segments conditionally on $M = 2$ and $T = 600$ for $\tau_K = 0.9$).

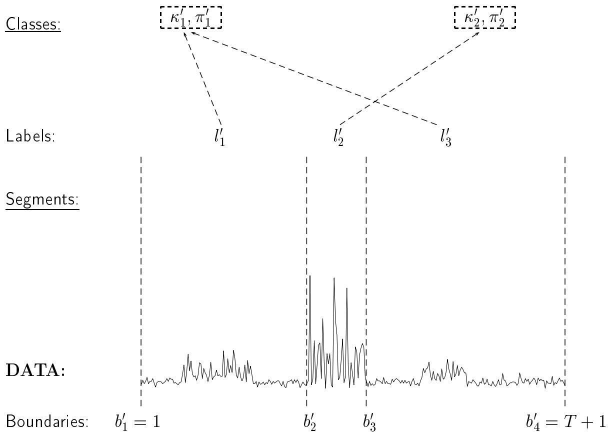

Fig. 5. Latent structure of figure 1 after merging the classes 1 and 2 in class $1'$ and relabelling the class 3 as $2'$.

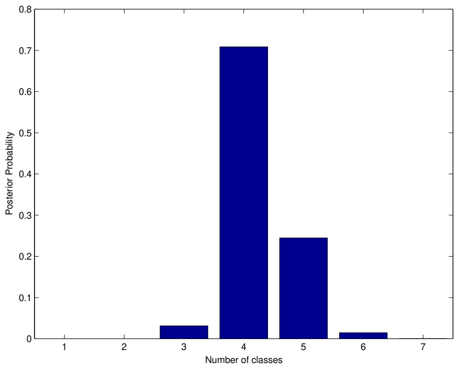

Fig. 6. Estimated posterior for the number of classes.

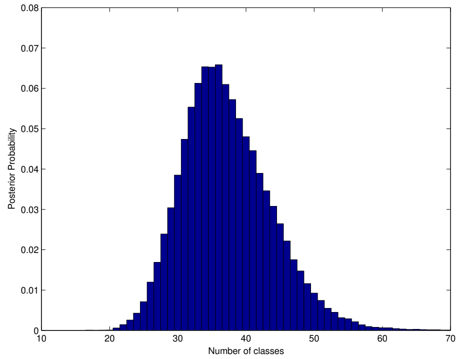

Fig. 7. Estimated posterior for the number of segments conditional upon the number of classes being equal to four.

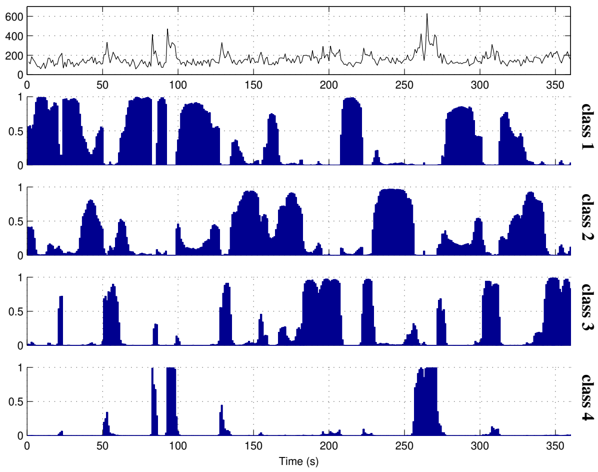

Fig. 8. Data (top plot) with estimated posterior classification conditional upon the number M of classes being equal to four.

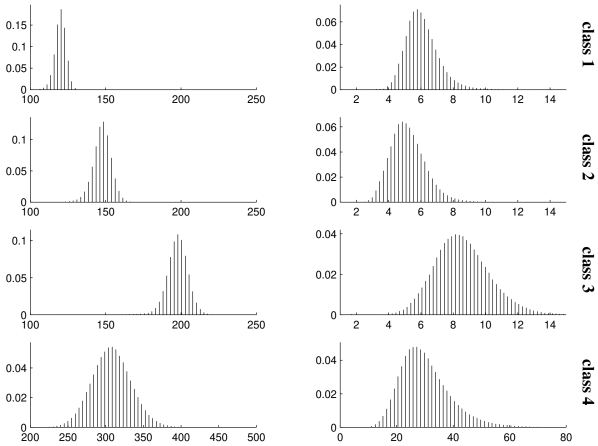

Fig. 9. Estimated posterior distribution of the class characteristics: mean μ_m (left) and over-dispersion ratio γ_m (left) (conditional on $M = 4$). Note the scale changes for the plots pertaining to the fourth class.

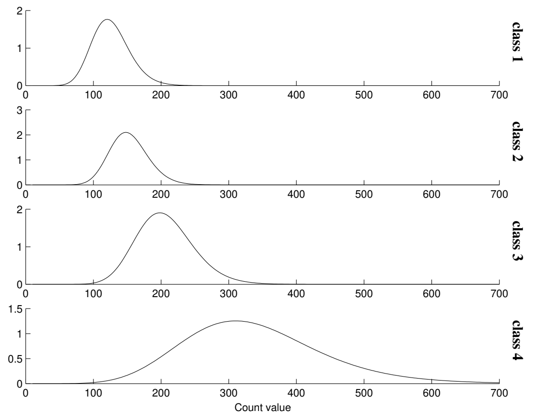

Fig. 10. Density estimates for the four classes conditional upon $M = 4$.

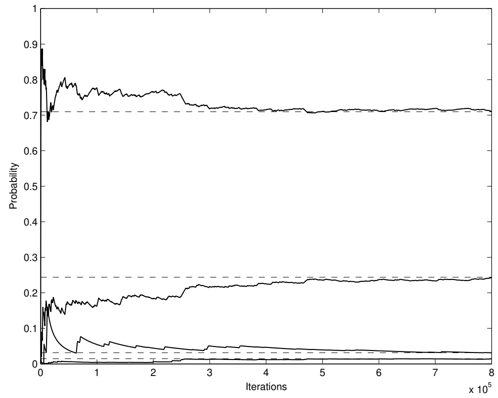

Fig. 11. Convergence of the class probability estimates. Only the estimates corresponding to $M = 3, 4, 5$ and 6 are visible on the plot.

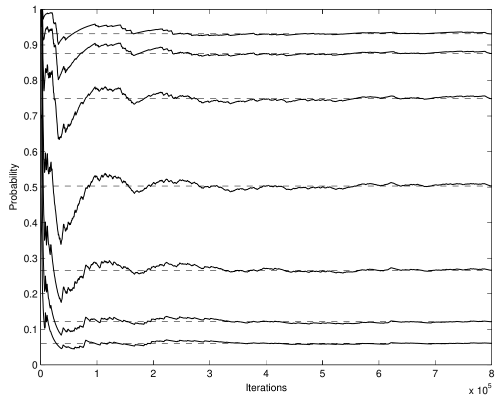

Fig. 12. Convergence of the segment probability estimates corresponding to 8 quantiles.

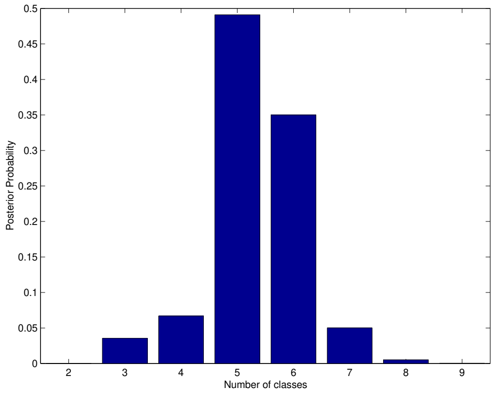

Fig. 13. Estimated posterior for the number of classes (12 mn of data).

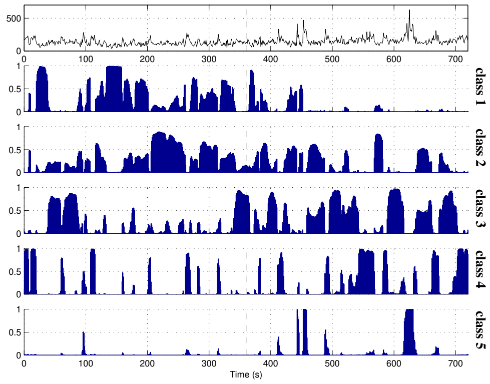

Fig. 14. Data (top plot) with estimated posterior classification conditional upon the number M of classes being equal to five (12 mn of data).

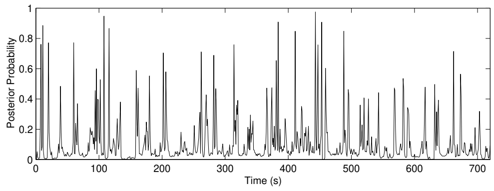

Fig. 15. Estimated posterior for the presence of a segment boundary, marginalizing with respect to all the parameters of the model (12 mn of data).