Bayesian Analysis of Network Traffic Data

Olivier Cappé

Ecole Nationale Supérieure des Télécommunications,

Dpt. TSI / CNRS URA 820, Paris, France

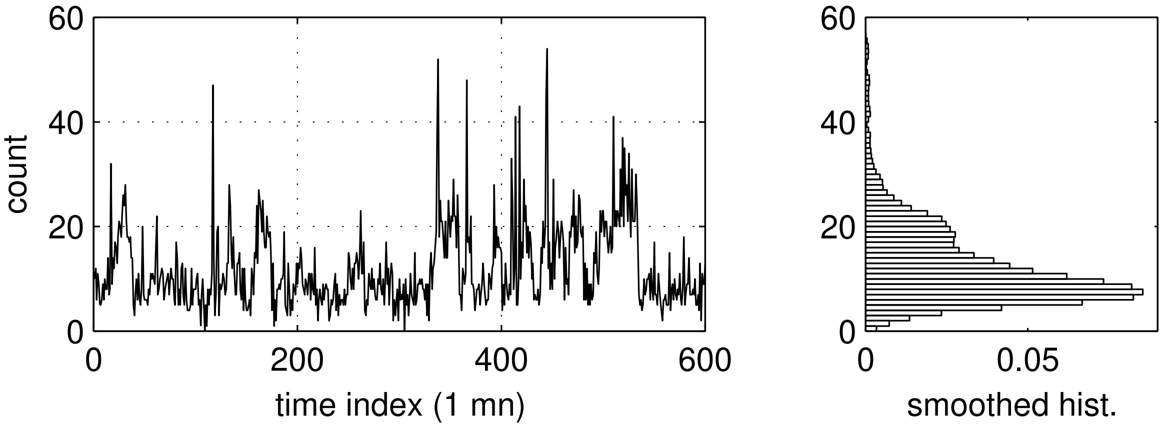
`<cappe@tsi.enst.fr>`

Outline

1. Data and motivations
2. Bayesian model description
3. MCMC sampling technique
4. Analysis of traffic traces

(1) Data under consideration

Overdispersed count data which exhibits both abrupt changes and level clustering



100 ms aggregated block packet counts, with smoothed histogram

Application

→ local analysis of aggregated network traffic data

(1) Network traffic data: some elements

Ethernet / Internet setting

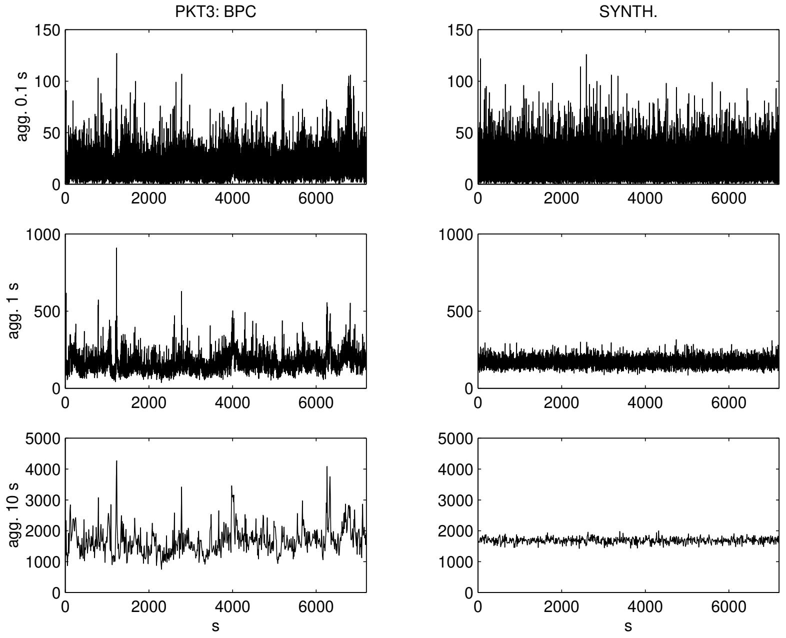
The available information, recorded on a per-packet basis, includes the time-stamp, but also the size of the packet, the protocol, the source and destination addresses, etc.

$\rightarrow$ Many freely available datasets (since the early 90s)

Common features: Long-term effects

(Paxson & Floyd, 1995) (Willinger, Taqqu, Leland, Wilson, 1995)...

1. Persistence of burstiness over different aggregation horizons



Packet counts with 0.1, 1 and 10s aggregation for actual data and *iid* simulation

(**2.** Heavy tailed marginal distribution)

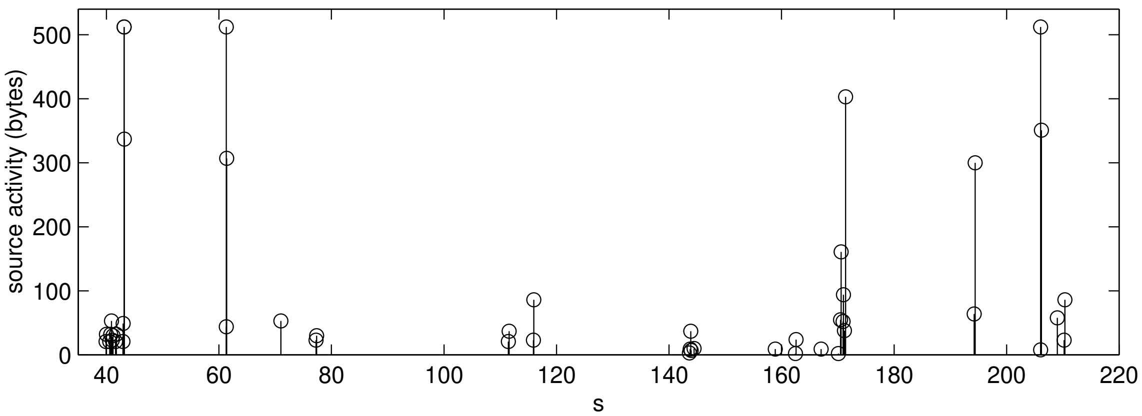
(1) Network traffic data: some elements

Conjectured origin of long-term behavior

(Willinger, Taqqu, Sherman & Wilson, 1995)

- Sources are independent stationary ON/OFF processes with independent heavy-tailed ON/OFF periods
- Aggregation in time and number of independent sources $\uparrow$

$\Rightarrow$ Normalized count process converges to long range dependent process



Example of activity on a single source/destination pair

$\longrightarrow$ Estimation of the "level of activity" from the data is an open question

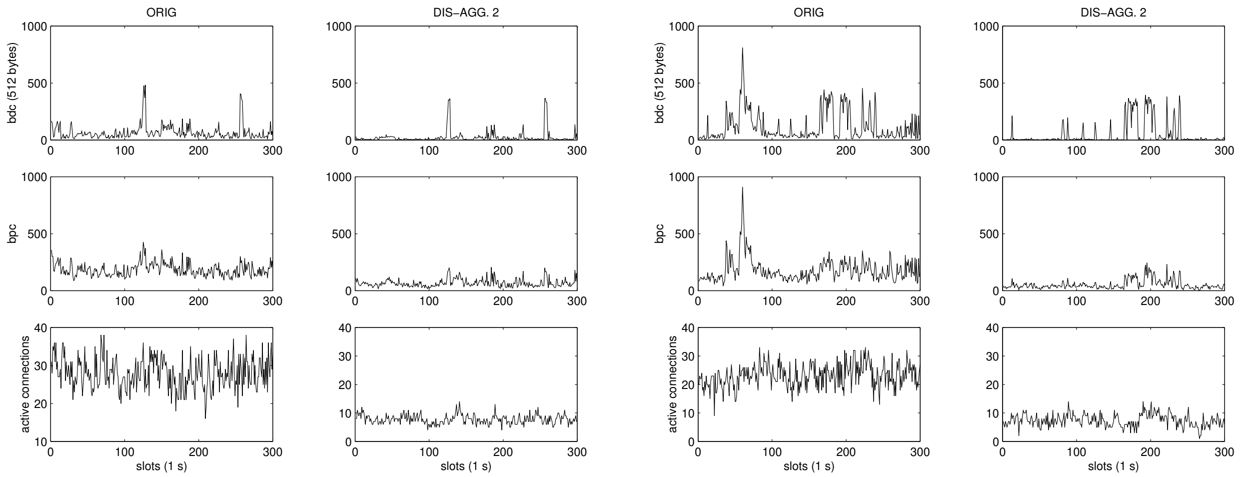
(1) Network traffic data: some elements

Significance of the available data (?)

"What does it mean in this setting to pick the "best-fitting" model for a given traffic trace, when the underlying data set keeps changing over time, across the same network, across different networks, and so on?" - W. Willinger & V. Paxon in discussion of (Resnick, 1997)

What should be analyzed (?)

Dataset: Trace LBL PKT-3 of (Paxson & Floyd, 1995), 2 hours of wide-area TCP traffic, recorded jan. 1994 (moderate activity).



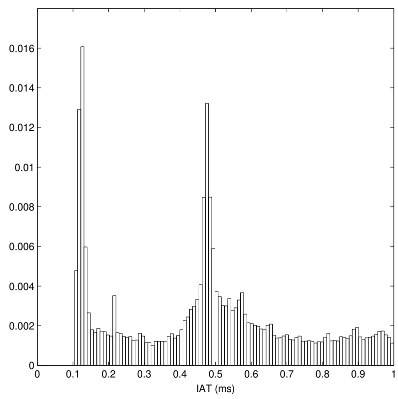
(1) Goals of the present study

$\Rightarrow$ **Local analysis of (time) aggregated data***

Why using aggregated data?

Aggregation: cumulative sum of the number of packets (Block Packet Count) or of the size of the packets (Byte Data) received during a fixed time slot (usually, from 1 ms to 1 s)

- Smoothing of "non-random" effects
- Huge size of the data sets: 150 packets per second for the trace considered here, may go up to several thousands of packets per second for faster and more active links (see Discussion of Resnick, 1997).



Local modeling

Analyze short segments of data (a few minutes?)

* Which does not necessarily have any implication for traffic control and performance analysis!

(2) Choice of the model

Goals

1. Classification: Sort the data points, according to some criterion ($\approx$ mixture model)
$\rightarrow$ Number of classes should be kept as low as possible (ON/OFF approximation?)

2. Segmentation: Find homogeneous time periods ($\approx$ changepoint model)
$\rightarrow$ Assumptions concerning the duration of the segments should not be too restrictive (heavy-tailed?)

$*$**. Robustness to aggregation:**
$\rightarrow$ Overdispersed marginal distribution (when compared to Poisson)
$\rightarrow$ Parametric modeling of the duration of the segment should be avoided (choice of the time-slot is arbitrary)

Open questions

- Is the visual impression of abrupt changes supported by the data?
- How many "classes" are needed? (are ON/OFF models plausible?)
- What is the influence of the aggregation horizon and the number of active sources?

(2) Related works (Bayesian framework)

(Stephens, 1994) Gibbs sampling for changepoint model (fixed number of changes)

(Green, 1995) Application of “reversible jumps” to models with unknown number of changes

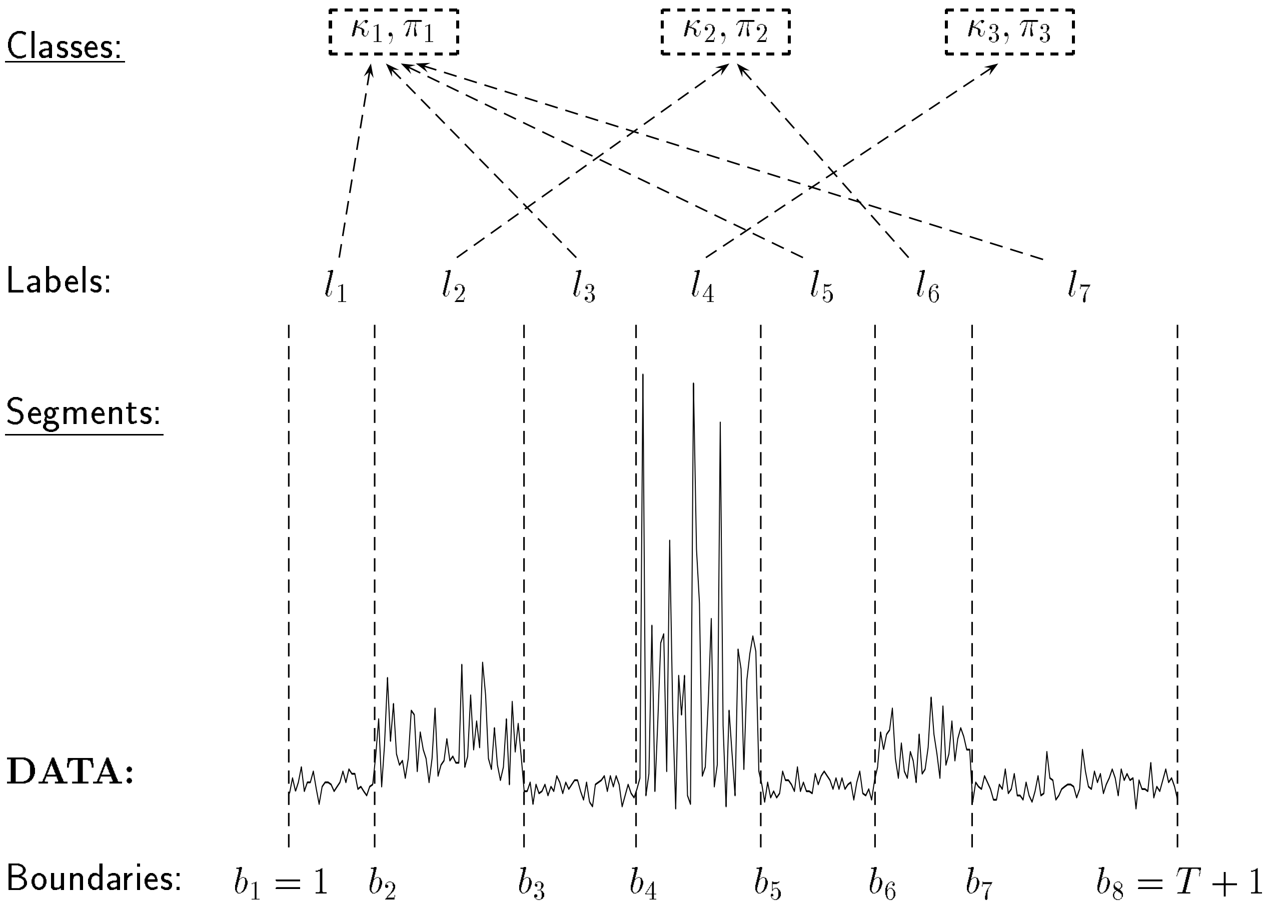
(Arjas & Heikkinen, 1997) Application to non-parametric estimation of the intensity of a doubly stochastic Poisson process

(Chib, 1997) Block Gibbs sampling strategy (requires geometric priors on the segment durations)

(Stark, 1997) Application to ion channel data (jump in the mean in Gaussian noise): Discrete changepoints, continuous levels, moves with analytic marginalization (state space of the MCMC sampling chain reduced to the indexes of the changes)

(Hodgson, 1998) Same application: Two levels, continuous changepoints, sojourn times with parametric (gamma) distribution

(2) Parameterization of the problem



Distributional assumption

$$n_t | (b_k \leq t < b_{k+1}, l_k, \kappa_{l_k}, \pi_{l_k}) \sim \mathrm{Neg-Binomial}(n_t | \kappa_{l_k}, \pi_{l_k})$$

where

$$\mathrm{Neg-Binomial}(n|\kappa, \pi) = \begin{pmatrix} n + \kappa - 1 \\ \kappa - 1 \end{pmatrix} \pi^\kappa (1 - \pi)^n$$

(2) Discussion of the parameterization

Constraints

Unambiguous definition of $K \Rightarrow b_k \neq b_{k+1}$ and $l_k \neq l_{k+1}$

Remark: When M is small the label sequence is severely constrained (in particular for $M = 2$)

Labels and boundaries versus affectation variables

The use of $i_t = l_{\{k_0 : b_{k_0} \leq t < b_{k_0+1}\}}$ (for $t = 1, \ldots, T$) avoids variable dimension parameters for the classification part but poses some simulation problems

Conditional distribution

Allows for overdispersed (compared to Poisson) distribution

$$\mathrm{Neg-Binomial}(n|\kappa, \pi) = \int \mathrm{Poisson}(n|\lambda)\mathrm{Gamma}(\lambda|\alpha, \beta) d\lambda$$

with $\alpha = \kappa, \beta = \pi/(1-\pi)$.
$\Rightarrow$ Suitable for aggregated data, but other choices are possible...

(2) Priors and hyperparameters

1. Classification (unordered)

Number of classes $p(M) \propto \tau_M^M$ for $M \geq 1$

Parameters of each class

$$p(\kappa_1, \pi_1, \ldots, \kappa_M, \pi_M | M) = \prod_{m=1}^{M} \mathrm{Gamma}(\kappa_m | \alpha_\kappa, \beta_\kappa) \mathrm{Beta}(\pi_m | \alpha_\pi, \beta_\pi)$$

2. Segmentation

Number of segments $p(K|M) \propto \tau_K^K$ for $K \geq M$

Segment boundaries

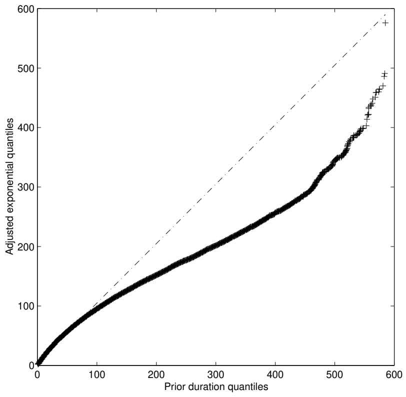
$$p(b_2, \ldots, b_K | K) = 1 / \binom{T-1}{K-1}$$

with the convention $b_1 = 1, b_{K+1} = T + 1$.

Segment labels $p(l_1, \ldots, l_K | K, M) = 1/[M(M-1)^{K-1}]$ with the constraint that $l_k \neq l_{k+1}$ for $k = 2, \ldots, K$.

(2) Discussion of the prior structure

Duration of the segments



Follows (Green, 1995)
$\rightarrow$ Allows for more dispersed segment durations than approaches based on Markovian assumptions (Chib, 1997)

Figure: Q-Q plot (10 000 draws of the segments conditionally to $M = 2$ for $T = 600$, $\tau_K = 0.9$ compared with adjusted exponential)

Number of classes

The constraint $K \geq M$ is not sufficient to prevent the appearance of empty classes
$\rightarrow$ Empty classes are penalized only by the prior on the label sequence

Parameters of each class

"Simple" parameterization with independent priors (semi-conjugate prior for π_k) $+$ unordered simulation of the class parameters
Other options... (Richardson & Green, Robert & Titterington, 1997)

(3) Description of the MCMC sampler

Hybrid Markov Chain Monte Carlo algorithm with systematic cycling through five different types of moves:

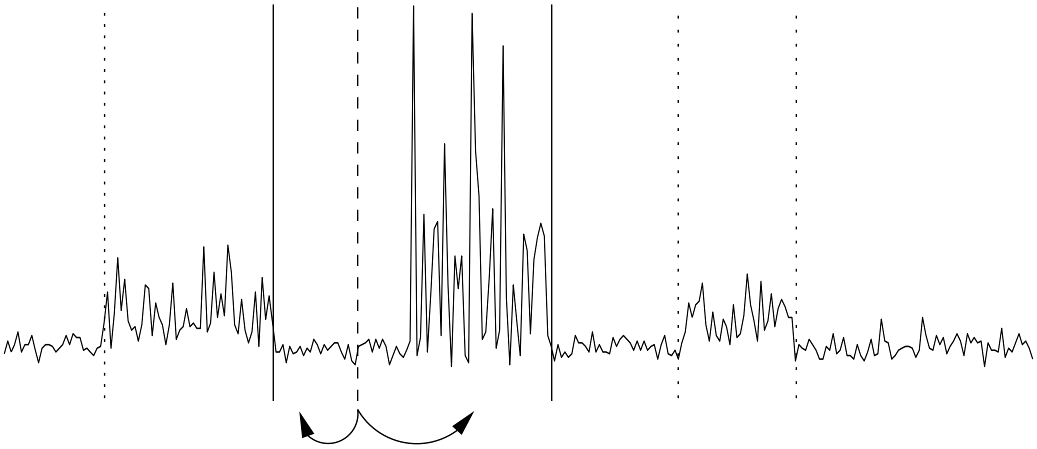
1. Updating the segment boundaries (Gibbs)
2. Updating the segment labels (Gibbs)
3. Creating or removing segments (Reversible jump MH, with specific $K \leftrightarrow K+2$ move when $M = 2$)
4. Updating the parameters of each class (Gibbs and MH)
5. Modifying the number of classes (Reversible jump MH)*

Moves 3 and 5 are based on the reversible jump approach of (Green, 1995).

* Not described in COMPSTAT-98 paper

(3) 1. Updating the segment boundaries

Gibbs cycle, drawing $b_k | b_{k-1}, b_{k+1}, l_{k-1}, l_k$ for $2 \leq k \leq T$ (Stephens, 1994)



$$P(B_k = t_0 | b_{k-1}, b_{k+1}, l_{k-1}, l_k) \propto$$

$$\left(\frac{\pi_{l_{k-1}}^{\kappa_{l_{k-1}}}}{\Gamma(\kappa_{l_{k-1}})} \right)^{(t_0 - b_{k-1})} \prod_{t=b_{k-1}}^{t_0 - 1} (1 - \pi_{l_{k-1}})^{n_t} \Gamma(n_t + \kappa_{l_{k-1}})$$

$$\times \left(\frac{\pi_{l_k}^{\kappa_{l_k}}}{\Gamma(\kappa_{l_k})} \right)^{(b_{k+1} - t_0)} \prod_{t=t_0}^{b_{k+1} - 1} (1 - \pi_{l_k})^{n_t} \Gamma(n_t + \kappa_{l_k})$$

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

(3) 2. Updating the segment labels

Gibbs cycle, computing $p(l_k|l_{k-1}, l_{k+1}, b_k, b_{k+1}, M, \kappa_{1:M}, \pi_{1:M})$ for $k = 1, \dots K$

Usually more efficient to use the alternative form of the likelihood:

$$P(L_k = m|b_k, b_{k+1}, M, \kappa_{1:M}, \pi_{1:M}) \propto$$
$$\pi_m^{N^{(k)}\kappa_m}(1-\pi_m)^{S^{(k)}} \prod_{r=1}^{R^{(k)}} (\kappa_m + r - 1)^{C_r^{(k)}}$$

where $N^{(k)}$ is the number of data points in segment k ($N^{(k)} = b_{k+1} - b_k$), $S^{(k)}$ the sum of these points, $R^{(k)}$ the maximum value, $C_r^{(m)}$ are the rank statistics (number of points greater or equal to r).

Remark: Mixing problems when M is small (locked configurations)
$\rightarrow$ for $M = 2$, block Gibbs sampling is used (two feasible configurations)

l_1	l_2	l_3	l_4	l_5	l_6
1	2	3	1	2	1

Example of locked label configurations for $M = 3$: l_2 and l_3 are “locked” (but are still susceptible of being modified by the Gibbs cycle if the value of l_1 first changes)

(3) 3. Updating the parameters of each class

"Probability of success" π_m

$\pi_1, \ldots, \pi_M$, are updated by a Gibbs cycle drawing from the full conditional

$$\pi_m | \cdots \sim \mathrm{Beta}(\kappa_m \bar{N}^{(m)} + \alpha_\pi, \bar{S}^{(m)} + \beta_\pi)$$

with $\bar{N}^{(m)} = \#\{1 \leq t \leq T : i_t = m\}$, $\bar{S}^{(m)} = \sum_{\{1 \leq t \leq T : i_t = m\}} n_t$

Shape parameter κ_m

The full conditional writes

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

where $\bar{R}^{(m)} = \max_{\{1 \leq t \leq T : i_t = m\}} n_t$,
$\bar{C}_r^{(m)} = \#\{1 \leq t \leq T : i_t = m, n_t \geq r\}$

$\rightarrow$ Independent Metropolis-Hastings kernels with adjusted Gamma proposal distributions*

---

* See COMPSTAT-98 paper for more detailed description of the proposal scheme

(3) 4. Creating or removing segments (split and merge)

Merge: $K+1 \to K$

1. Select the Merge move with probability $P_M(K+1)$
2. Draw a segment k in $\{1,\dots,K\}$ with probability $1/K$
3. Draw the label of the merged segment l'_k in $\{1,\dots,M\}$

The move is rejected if $l'_k = l_{k-1}$ or $l'_k = l_{k+2}$

Split: $K \to K+1$

1. Select the Split-2 move with probability $P_S(K)$
2. Draw a segment k in $\{1,\dots,K\}$ with probability $(b_{k+1} - b_k - 1)/(T-K)$
3. Draw a new sub-segment boundary b'_{k+1} in $\{b_k+1,\dots,b_{k+1}-1\}$
4. Draw independently in $\{1,\dots,M\}$ two labels for the new subsegments

The move is rejected if any two successive labels in the sequence $(l_{k-1}, l'_k, l'_{k+1}, l_{k+1})$ are identical

(3) 5. Modifying the number of classes (merge: $M+1 \to M$)

1. Select class merging move with probability $P_{CM}(M+1)$
2. Draw the index m of the first class to be merged in $\{1, \ldots, M\}$

Let L_m and L_{m+1} denote the number of segments associated with each of the two classes to be merged, the parameters of the merged class are computed as weighted geometric means

$$\begin{aligned}
\kappa'_m &= \left(\kappa_m^{L_m} \kappa_{m+1}^{L_{m+1}}\right)^{\frac{1}{L_m + L_{m+1}}} \\
\pi'_m &= \pi\left(\left[\beta(\pi_m)^{L_m} \beta(\pi_{m+1})^{L_{m+1}}\right]^{\frac{1}{L_m + L_{m+1}}}\right)
\end{aligned}$$

where $\pi(x) = \frac{x}{1+x}$, $\beta(x) = \frac{x}{1-x}$ (and the inverse scale parameter $\beta \in [0, +\infty)$)

Remark: Merging two classes usually induces a reduction of the number of segments

(3) 5. Modifying the number of classes (split: $M \to M+1$)

1. Select class splitting move with probability $P_{CS}(M)$.
2. Select the index m of the first class to be splitted in $\{1, \ldots, M\}$ with probability L_m/K
3. For each segment $k_1, \ldots k_{L_m}$ with label m,
   (a) Draw a number of additional sub-segments H_{k_i} with truncated $\mathrm{Geom}(\tau_B(b_{k_i+1} - b_{k_i})/T)$ distribution
   (b) Draw the new sub-segments frontiers (ordered draw without replacement) if $H_{k_i} \geq 1$
   (c) Draw a valid sequence of the labels $m', (m+1)'$
4. Compute perturbated class parameters

$$\begin{array}{rclrcl} \kappa'_m & = & \kappa_m \rho_\kappa^{L'_{m+1}} & \kappa'_{m+1} & = & \kappa_m \rho_\kappa^{-L'_m} \\ \pi'_m & = & \pi(\beta(\kappa_m)\rho_\pi^{L'_{m+1}}) & \pi'_{m+1} & = & \pi(\beta(\kappa_m)\rho_\pi^{-L'_m}) \end{array}$$

where L'_m and L'_{m+1} are the number of segments associated to each of the two new classes, and the perturbations ρ_κ and ρ_π are drawn under the $\mathrm{Gamma}(g, g)$ distribution (with $g \gg 1$)

Remark: The split above move usually results in an increased number of classes

(3) Probability of acceptance for the class splitting move

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

$$A = \text{likelihood ratio}$$

$$(\mathbf{A1}) \quad \times \tau_M \tau_K^{(K'-K)} \frac{\binom{T-1}{K-1} M(M-1)^{(K-1)}}{\binom{T-1}{K'-1} (M+1) M^{(K'-1)}}$$

$$(\mathbf{A2.1}) \quad \times \frac{\beta_\kappa^{\alpha_\kappa}}{\Gamma(\alpha_\kappa)} \left(\frac{\kappa'_m \kappa'_{m+1}}{\kappa_m} \right)^{\alpha_\kappa - 1} \mathrm{e}^{-\beta_\kappa ((\kappa'_m + \kappa'_{m+1}) - \kappa_m)}$$

$$(\mathbf{A2.2}) \quad \times \frac{\Gamma(\alpha_\pi + \beta_\pi)}{\Gamma(\alpha_\pi)\Gamma(\beta_\pi)} \left(\frac{\pi'_m \pi'_{m+1}}{\pi_m} \right)^{\alpha_\pi - 1} \left(\frac{(1-\pi'_m)(1-\pi'_{m+1})}{1-\pi_m} \right)^{\beta_\pi - 1}$$

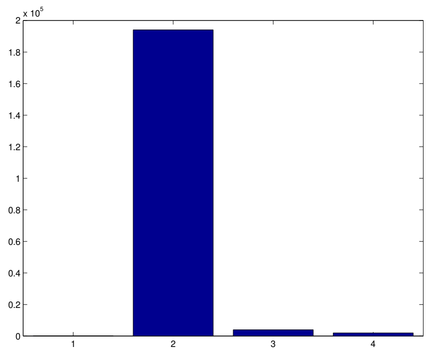
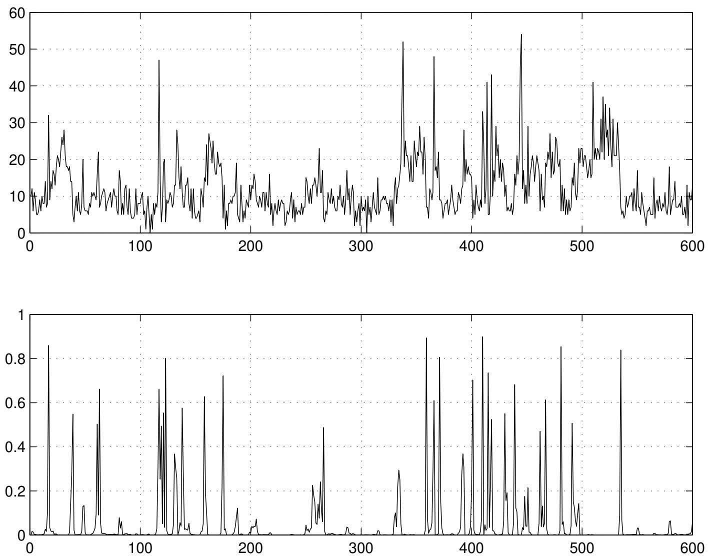
$$(\mathbf{A3}) \quad \times \frac{P_{CM}(M+1)K}{P_{CS}(M) M L_m}$$

$$(\mathbf{A4}) \quad \times \prod_{i=1}^{L_m} \frac{2}{1 - \tau_B (b_{k_i+1} - b_{k_i})/T} \left(\frac{T}{\tau_B (b_{k_i+1} - b_{k_i})} \right)^{H_{k_i}} \binom{b_{k_{i+1}} - b_{k_i} - 1}{H_{k_i}}$$

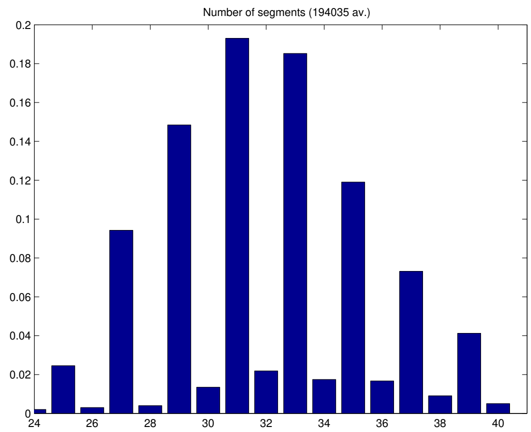
$$(\mathbf{A5}) \quad \times (\kappa'_m \beta(\pi'_m))^{\frac{L'_{m+1} - 1}{L'_m + L'_{m+1}}} (\kappa'_{m+1} \beta(\pi'_{m+1}))^{\frac{L'_m + 1}{L'_m + L'_{m+1}}} (L'_m + L'_{m+1})^2 \frac{(1-\pi'_m)^2 (1-\pi'_{m+1})^2}{(1-\pi_m)^2}$$

$$(\mathbf{A6}) \quad \times \left(\frac{g_\kappa^{g_\kappa}}{\Gamma(g_\kappa)} \rho_\kappa^{(g_\kappa - 1)} \mathrm{e}^{-g_\kappa \rho_\kappa} \frac{g_\pi^{g_\pi}}{\Gamma(g_\pi)} \rho_\pi^{(g_\pi - 1)} \mathrm{e}^{-g_\pi \rho_\pi} \right)^{-1}$$

(4) Analysis of traffic data: 60s of 0.1s aggregated data (packet count)

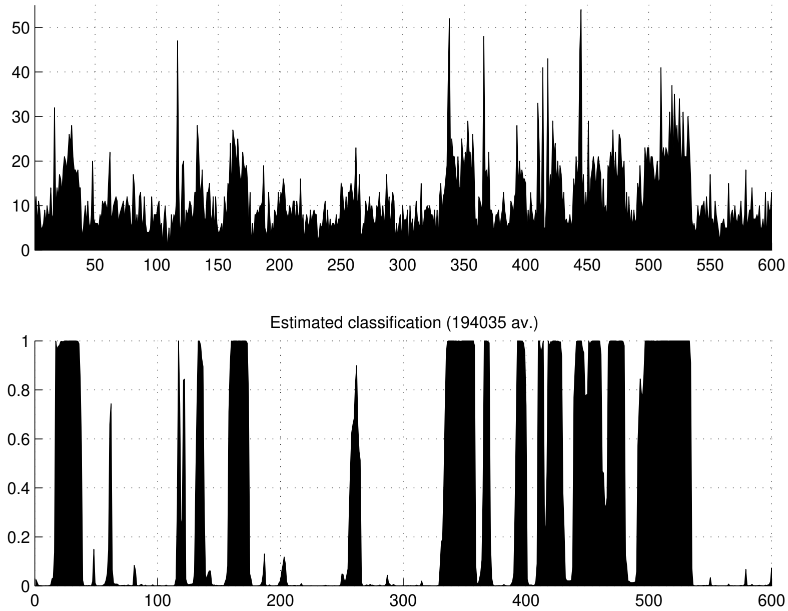



Data and posteriors of the number of classes and the presence of a boundary (200 000 sampler iterations)

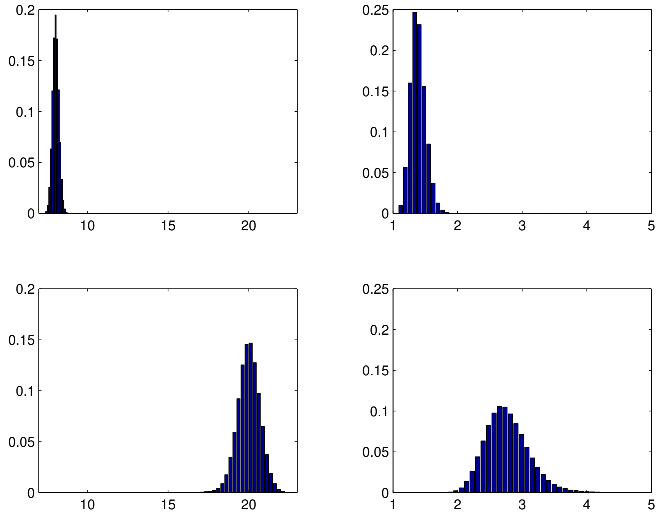


Posterior of the number of segments conditioning on $M = 2$ (190 000 av.)

(4) 60 s of 0.1s aggregated data (cont.)

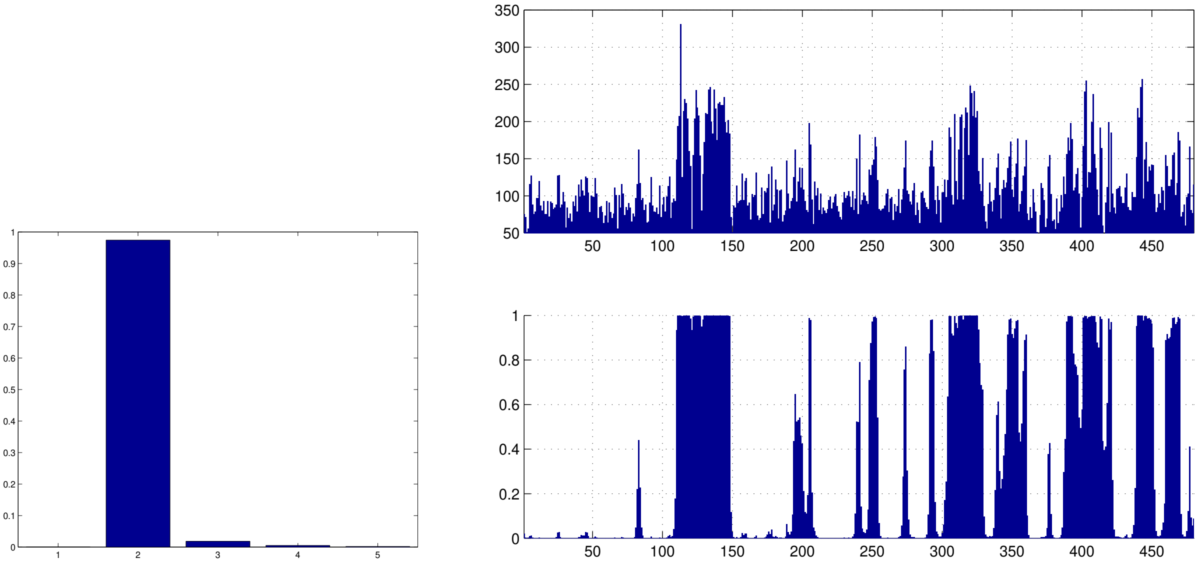


Data and posterior classification (for $M = 2$)

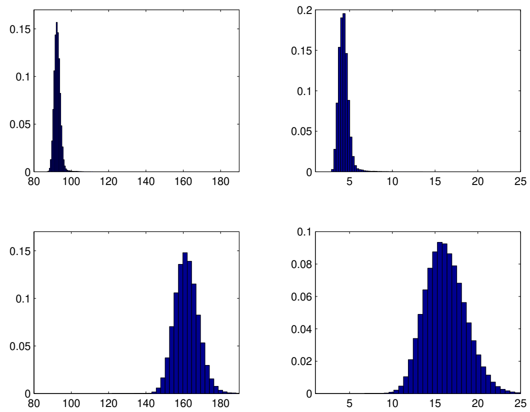


Posteriors of the mean ($\kappa(1-\pi)/\pi$) and the dispersion ($1/\pi$) for the two classes, conditioning on $M = 2$

(4) 8mn of 1s aggregated packet data

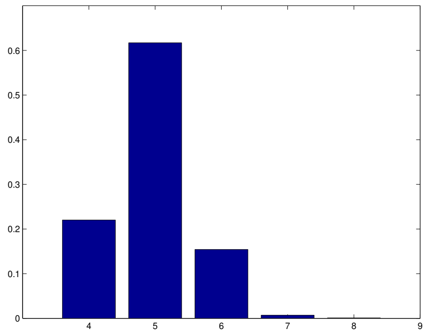
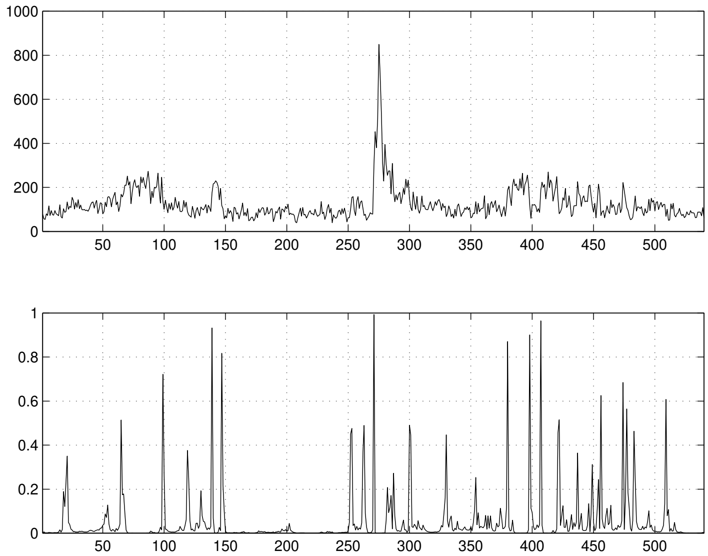


Data with posteriors of the number of classes and the classification (200 000 sampler iterations)



Posteriors of the mean ($\kappa(1-\pi)/\pi$) and the dispersion ($1/\pi$) for the two classes, conditioning on $M = 2$ (195 000 av.)

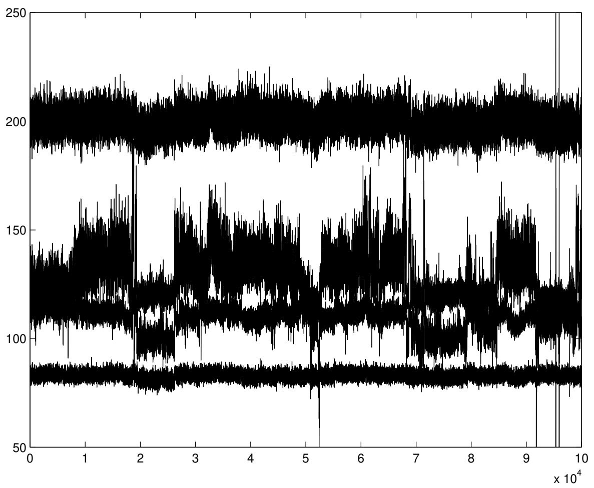
(4) 9mn of 1s aggregated packet data

Data and posteriors for num. class. and bnd. (200 000 it.)

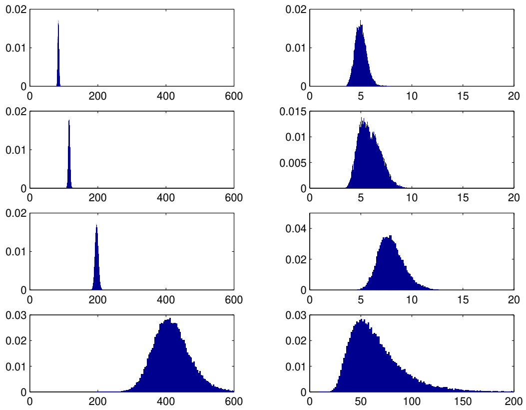
→ When considering longer intervals, more than 2 classes may be needed (but segments still are mostly attributed to a single class)

Rmk: Interesting sampler artifact (?)

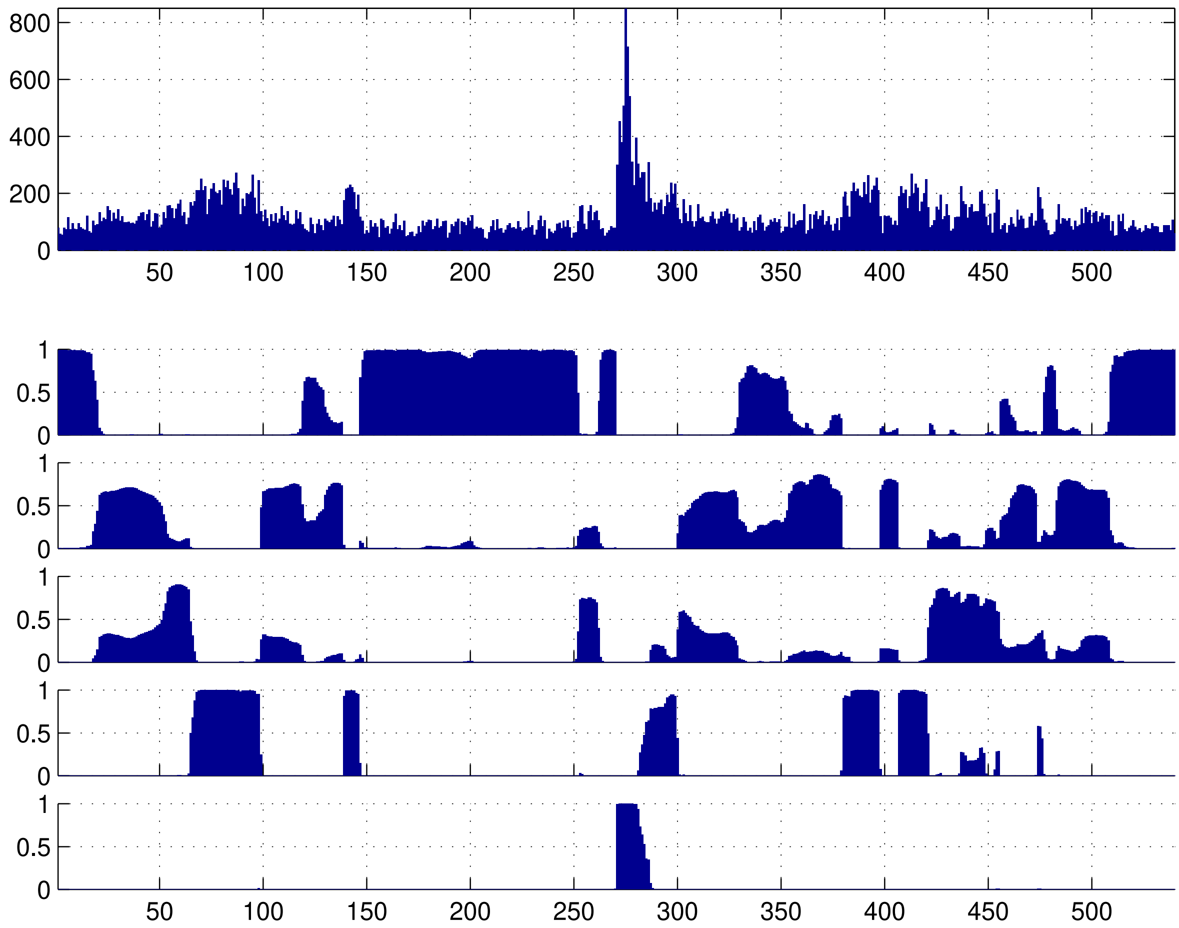


"Sample paths" for classes 1 to 4 when $M = 5$

(4) 9mn of 1s aggregated data (cont.)

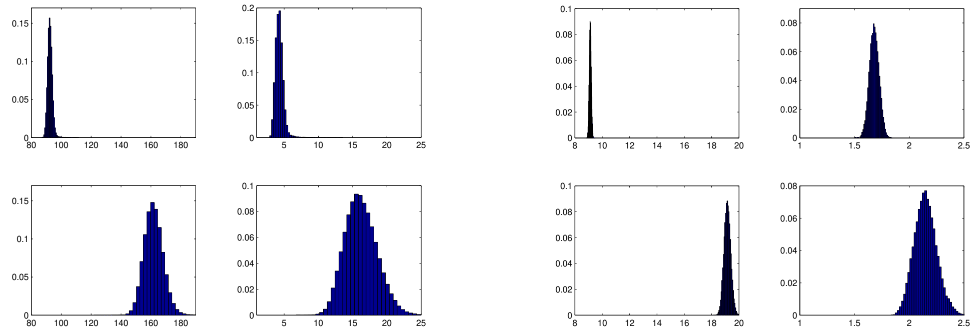


Posteriors of the mean $(\kappa(1-\pi)/\pi)$ and the dispersion $(1/\pi)$ for the two classes, conditioning on $M = 4$ (145 000 av.)

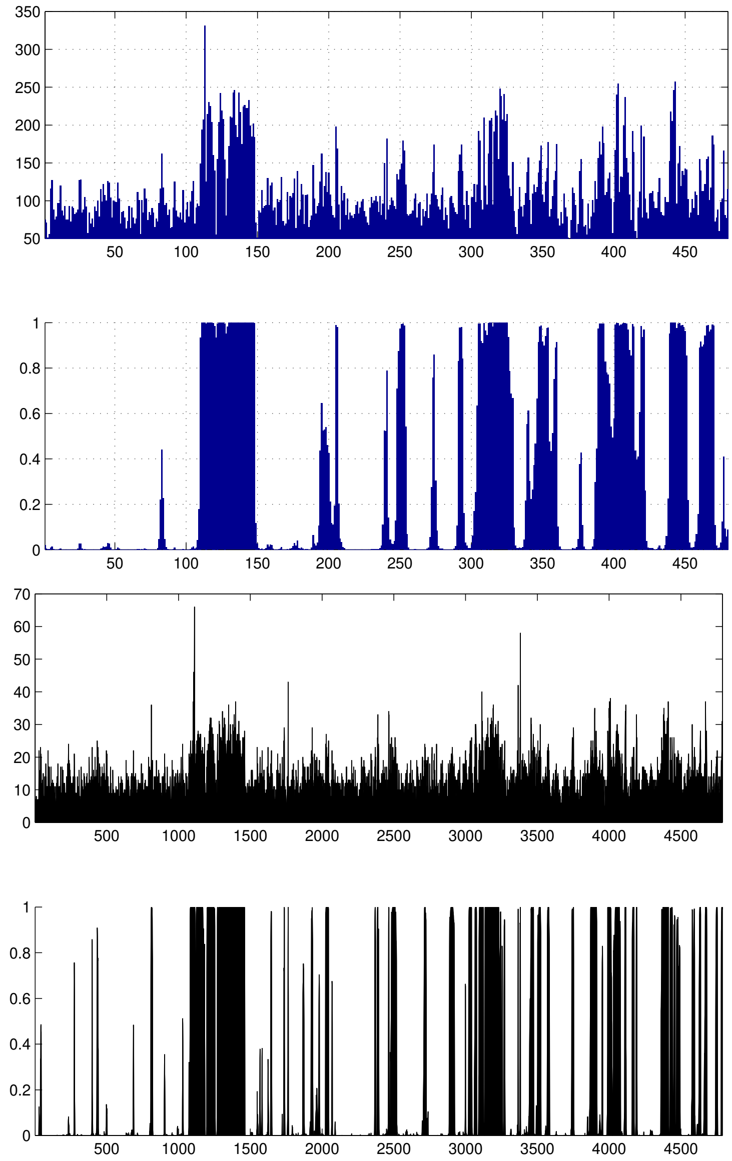


Data and posterior classification, conditioning on $M = 4$ (145 000 iterations)

(4) Influence of aggregation (data as in xxiii)

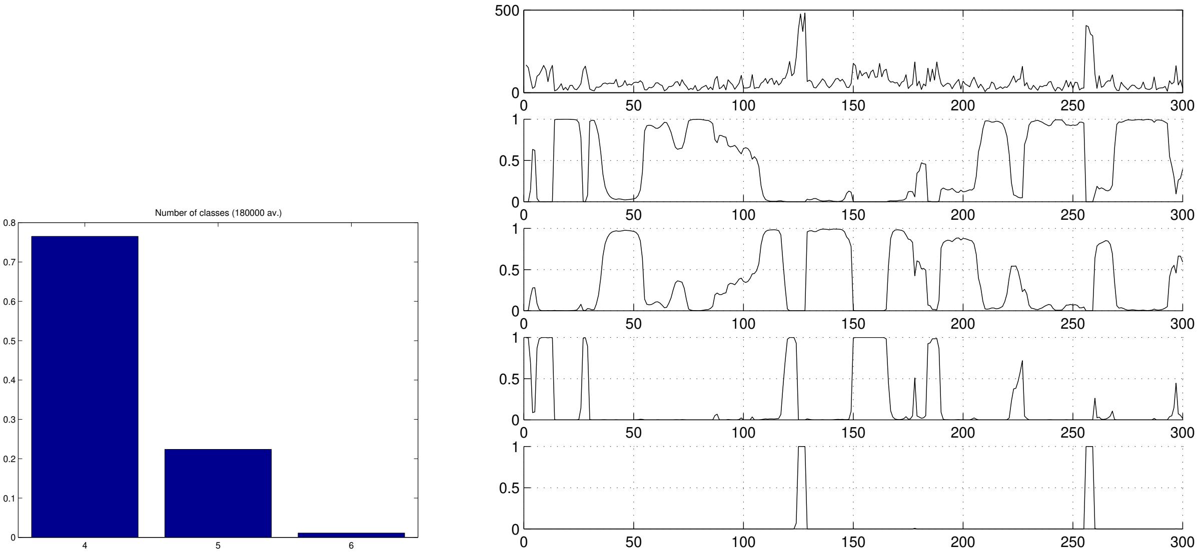


Posteriors of the mean and dispersion; **Left** 1s aggregation, **Right** 0.1s aggregation

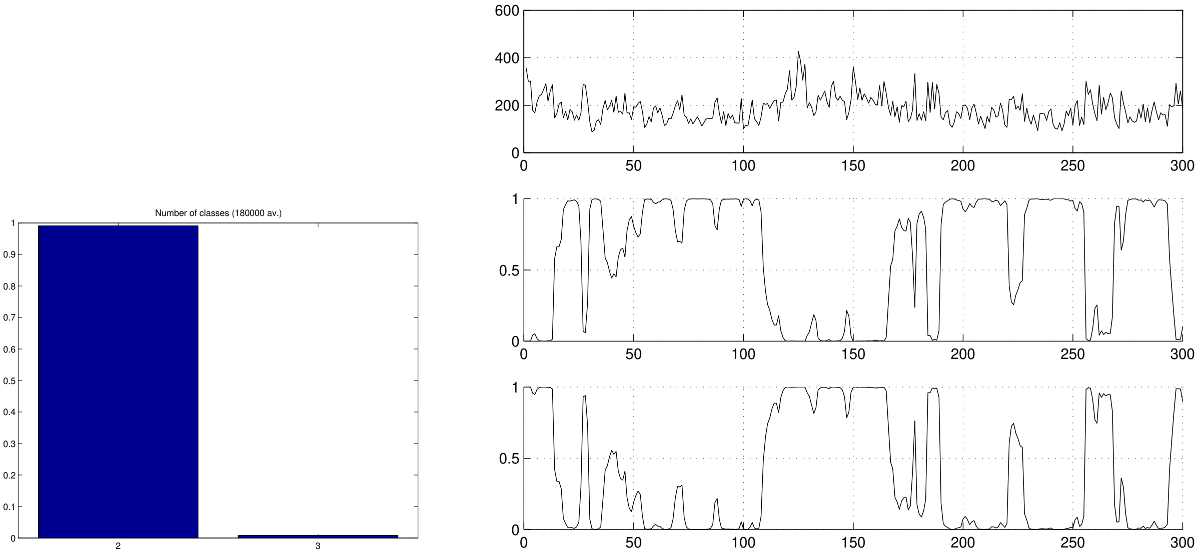


Data and posterior classification; **Top** 1s aggregation, **Bottom** 0.1s aggregation

(4) Analysis of size and packet data (3mn of 1s aggregated data)



Byte data (512 bytes unit)



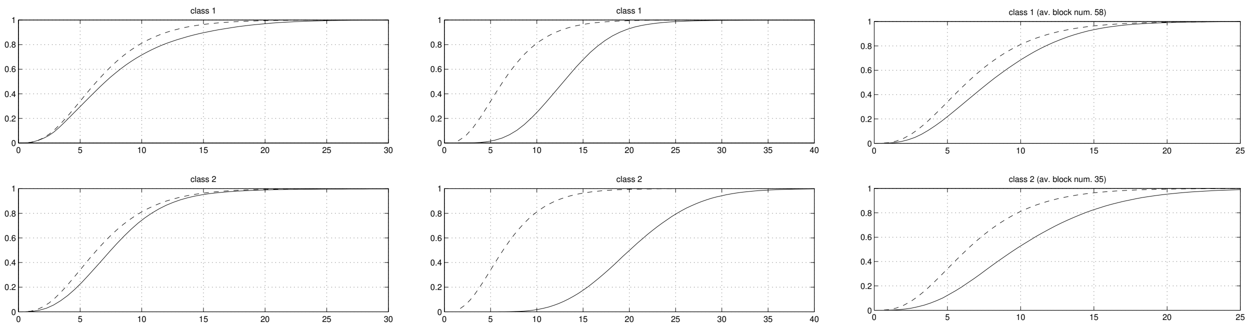
Packet data

(4) Are iid segments plausible?

Test statistic: Conditional (to the segmentation and the class parameters) Kullback divergence estimated on grouped and binned data ($\mathcal{L} = 2$, $\mathcal{B} = 3$)

$$T \sum_{k=1}^{\mathcal{L}^{\mathcal{B}}} \hat{p}_k^{(\mathcal{L},\mathcal{B})} (\log \hat{p}_k^{(\mathcal{L},\mathcal{B})} - \log p_k^{(\mathcal{L},\mathcal{B})})$$

summed over all segments with identical label at each sampler iteration
$\rightarrow$ Conditional distribution (large T) $\chi^2_{\mathcal{L}^{\mathcal{B}}-1}$



Packet data (20-35 segments): xxi (0.1s agg.), xxiii (1s agg., low source aggregation), xxvii (1s agg.)

$\Rightarrow$ iidness seems generally acceptable (except when aggregating few sources over long time intervals)

Conclusions and clues

Model and MCMC sampler

Sampler Specific parameterization for the two classes case, better processing of model/sampler artifacts (not limited to label switching), convergence assesment

Parameterization Allow for informative analysis (mean $\kappa(1-\pi)/\pi$) and overdispersion factor ($1/\pi$) would be a more natural choice, at least for specifying the priors)

Traffic analysis

- Abrupt changes are present in the analyzed data
- “Homogeneous” segments are more dispersed than Poisson
- Modeling the “homogeneous” segments as *iid* requires some sort of aggregation
- Active/Inactive (rather than ON/OFF) models are plausible on horizons of a few minutes (but activity levels vary with time)
- Size and packet count data have different characteristics
- Further investigation is needed concerning links with more active sources...