7. Markov chain Monte Carlo methods for sequence segmentation

- So far:
  - time-series data
  - squared error or Bernoulli likelihood
  - dynamic programming
- Now:
  - Sequences of events
  - Poisson likelihood
  - Markov chain Monte Carlo methods

Marko Salmenkivi and Heikki Mannila, Using Markov chain Monte Carlo and dynamic programming for event sequence data. Knowledge and Information Systems 7(3),267-288.

Marko Salmenkivi and Heikki Mannila, Piecewise constant modeling of sequential data using reversible jump Markov chain Monte Carlo. In Wang, J., Zaki, M. J., Toivonen, H., Shasha, D.(Eds.), Data Mining in Bioinformatics, Springer-Verlag London Ltd., 2005.

Marko Salmenkivi, Juha Kere, and Heikki Mannila, Genome segmentation using piecewise constant intensity models and reversible jump MCMC. Bioinformatics (European Conference on Computational Biology) 18 (2), 211-218, 2002.

Heikki Mannila and Marko Salmenkivi, Finding simple intensity descriptions from event sequence data. Proceedings of 7th ACM SIGKDD (KDD'2001) Conference on Knowledge Discovery and Data Mining, 341-346, San Francisco, CA, August 2001.

Sequences of events

- A discrete alphabet Γ: event types
- Events: pairs (A, t), where $A \in \Gamma$ and t is a real number
- Telecommunications alarm management
- Biostatistics
- User interface design

Telecommunications alarm management

- Alarms: indications that something is wrong
- Often not visible to the users
- "Cooling fan temperature on node 123 too high"
- "Bit error rate over τ on link 456"
- 100–1000 different event types
- 1000–10^6 events/day

Biological events on sequences

- DNA sequence of length 100,000,000
- Occurrences of genes, regulatory regions, interesting motifs
- What are the connections between these occurrences?

Words in documents

- A document is a sequence of words
- Certain words reoccur
- Word occurrences have connections to each other

Different types of analysis

- Plots, histograms
- Episodes: local phenomena that occur often
- Here: the intensity of events

Intensity modeling

- A single event type
- Waiting time not longer than t:
  $G(t) = Prob(T \leq t)$
- $g(t)$: density function of G
- $\overline{G}(t) = 1 - G(t)$ expresses the probability that the waiting time is longer than t
- $\overline{G}(t)$ is called a $survival\ function$.

Intensity function

- the $intensity\ function\ \lambda(t)$ is defined as follows

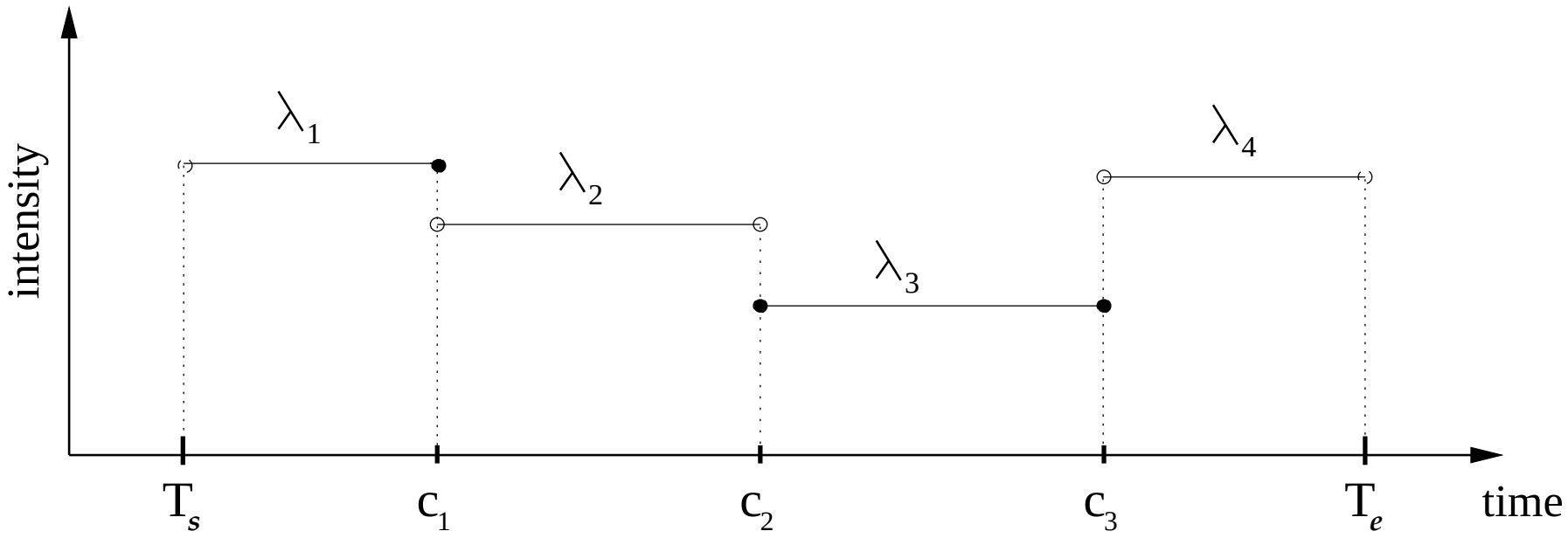
$$\lambda(t) = \frac{g(t)}{\overline{G}(t)}$$

- the "immediate" probability of occurrence at time t, given again that no event occurred before t.
- Thus $\lambda(t)dt$ expresses the probability of exactly one event during a "short" time interval dt.
- The intensity function uniquely determines the distribution function $G(t)$.

Piecewise constant intensity functions

$$\lambda(t) = \begin{cases} \lambda_1 & \text{if } T_s \leq t < c_1, \\ \lambda_2 & \text{if } c_1 \leq t < c_2, \\ \vdots & \vdots \\ \lambda_i & \text{if } c_{i-1} \leq t \leq T_e, \\ 0 & \text{elsewhere} \end{cases}$$

- $\{T_s, T_e\} \in \mathbf{R}$: the start and end times of the observation period
- $\{\lambda_1, ..., \lambda_i\} \in \mathbf{R}^+$ are the intensity values in i pieces
- $\{c_1, ..., c_{i-1}\} \in [T_s, T_e]$ are the *change points* of the function


intensity
λ1
λ2
λ3
λ4
Ts
c1
c2
c3
Te
time

Poisson likelihood

$S_e = \{(e, t_1), \ldots, (e, t_n)\},$ where $t_i \in [T_s, T_e],$ for all $i = 1, \ldots, n.$

$$L(S_e \mid \lambda) = \prod_{j=1}^{k} \lambda(t_j) \cdot \exp(- \int_{t_{j-1}}^{t_j} \lambda(t)\, dt) \cdot \exp\,(- \int_{t_k}^{T_e} \lambda(t)\, dt),$$

$$L(S_e \mid \lambda) = \exp\,(- \int_{T_s}^{T_e} \lambda(t)\, dt) \cdot \prod_{j=1}^{k} \lambda(t_j).$$

Bayesian approach

$$P(\theta, Y) = P(\theta)P(D|\theta).$$

$P(\theta)$: the $prior\ distribution$ of θ

$P(D|\theta)$ the $likelihood$ of the data D.

$$P(\theta|D) = \frac{P(\theta)P(D|\theta)}{P(D)} = \frac{P(\theta)P(D|\theta)}{\int_\theta P(\theta)P(D|\theta)d\theta}.$$

the $posterior\ distribution$ of the parameter θ

Goal

- Find out the expectation of a function $g(\theta)$ of the parameter θ over the posterior distribution $P(\theta|D)$
- E.g., find the posterior average of the number of pieces
- Analytical integration is not possible due to the complexity of the density functions
- Use Markov chain Monte Carlo (MCMC) methods
- Sample $x_1, \ldots, x_N$ from π, the posterior distribution of the parameters

$$E_\pi g(X) = \int g(\theta)\pi(\theta)\ d\theta \approx \frac{1}{N}\ \sum_{i=1}^{N} g(x_i).$$

Markov chains

- A set of states $x, y, \ldots$
- Moves between states
- Transition function $T(x, y)$: the probability of being in state y at time $i + 1$, if one is in state x at time i
- $\sum_y T(x, y) = 1$ (one moves to some state)
- Stationary distribution:
  - a probability distribution $f(x)$ over the states
  - such that if f is the distribution now, then f is the distribution after one more step
- Stationary distribution exists and is unique under fairly mild conditions

Metropolis-Hastings algorithm

- How to form a Markov chain that has a given function f as the equilibrium distribution?
- The set of possible states is given (the set of all possible parameter values)
- For states x and y let $T(x, y) =$ probability of moving from state x to state y
- The reversibility condition

$$f(x)T(x, y) = f(y)T(y, x)$$

- Equal amounts of mass flow from x to y and from y to x

Reversibility

- Assume $f(x)T(x,y) = f(y)T(y,x)$. Then f is the stationary distribution of T.
- Why?
- $$Pr(\text{ at state } x \text{ after one more step }) = \sum_y f(y)T(y,x)$$
- Reversibility implies

$$\sum_y f(y)T(y,x) = \sum_y f(x)T(x,y) = f(x)\sum_y T(x,y) = f(x)$$

How to find the function $T(x, y)$?

- Can be difficult
- Let $S(x, y)$ be a candidate function for T
- assume $f(x)S(x, y) > f(y)S(y, x)$, i.e., S is not OK
- let us accept a move from x to y with probability $\alpha(x, y)$ that corrects the imbalance:

$$\alpha(x, y) = \min(1, \frac{f(y)S(y, x)}{f(x)S(x, y)})$$

- Then

$$f(x)S(x, y)\alpha(x, y) = f(y)S(y, x) = f(y)S(y, x)\alpha(y, x)$$

- I.e., $T(x, y) = S(x, y)\alpha(x, y)$ is a good function!

Metropolis-Hastings algorithm

- start from a random point x
- generate a new point y with probability $S(x, y)$
- accept y with probability $\alpha(x, y)$
- otherwise, x is the next point, also

Requirements

- Must be able to compute $f(x)/f(y)$ for all pairs of states x, y
- In the Bayesian approach, this can be done
- $$P(\theta_1|D) = \frac{P(\theta_1)P(D|\theta_1)}{\int_\phi P(\phi)P(D|\phi)d\phi}.$$
- Thus
$$\frac{P(\theta_1|D)}{P(\theta_2|D)} = \frac{P(\theta_1)P(D|\theta_1)}{P(\theta_2)P(D|\theta_2)}$$
i.e., the ratio of the quantities (prior*likelihood)

Markov chain Monte Carlo methods

- Metropolis-Hastings algorithm is only one of the possible methods
- Very powerful technique in general
- Arbitrary likelihood functions and priors can (at least in principle) be used
- Convergence is sometimes an issue

Variable dimension

- What if the parameter space has components of different dimensions?
- Different numbers of constant parts in the piecewise constant function
- Green's generalization

Reversible jump MCMC (Green)

- current state x, dimension m
- proposed state y, dimension $n = m + u$
- $y = g(x, rand(u))$
- $$\alpha = \min(1, \frac{f(\theta')\ q(\theta', \theta)}{f(\theta)\ q(\theta', \theta)} \left| \frac{\partial g(\theta, u)}{\partial \theta\ \partial u)} \right|)$$
- Jacobian of the transformation of the random variables θ and u.
- the jeversible jump Metropolis Hastings algorithm, or reversible jumb Markov chain Monte Carlo (RJMCMC).

MCMC for piecewise constant intensity models

- Lots of details in getting the implementation to run
- Priors:
  - number of pieces $k \sim$ Geom(γ);
  - levels of functions $\lambda_j \sim$ Uniform(α, β);
  - change times $c_j \sim$ Uniform(T_s, T_e).

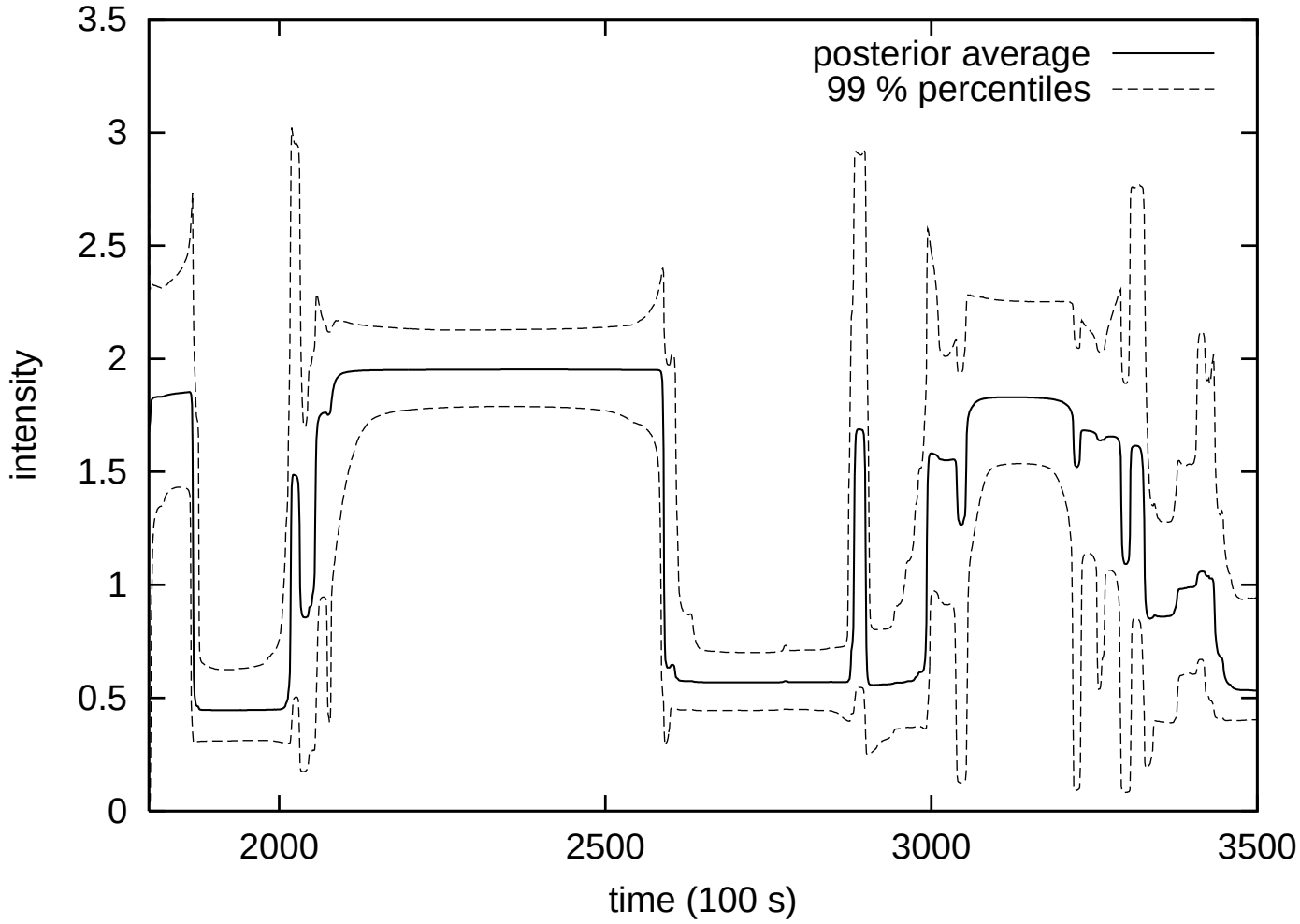
New points

- Modify a level
- Move a starting point of an interval
- Remove one piece
- Add one piece

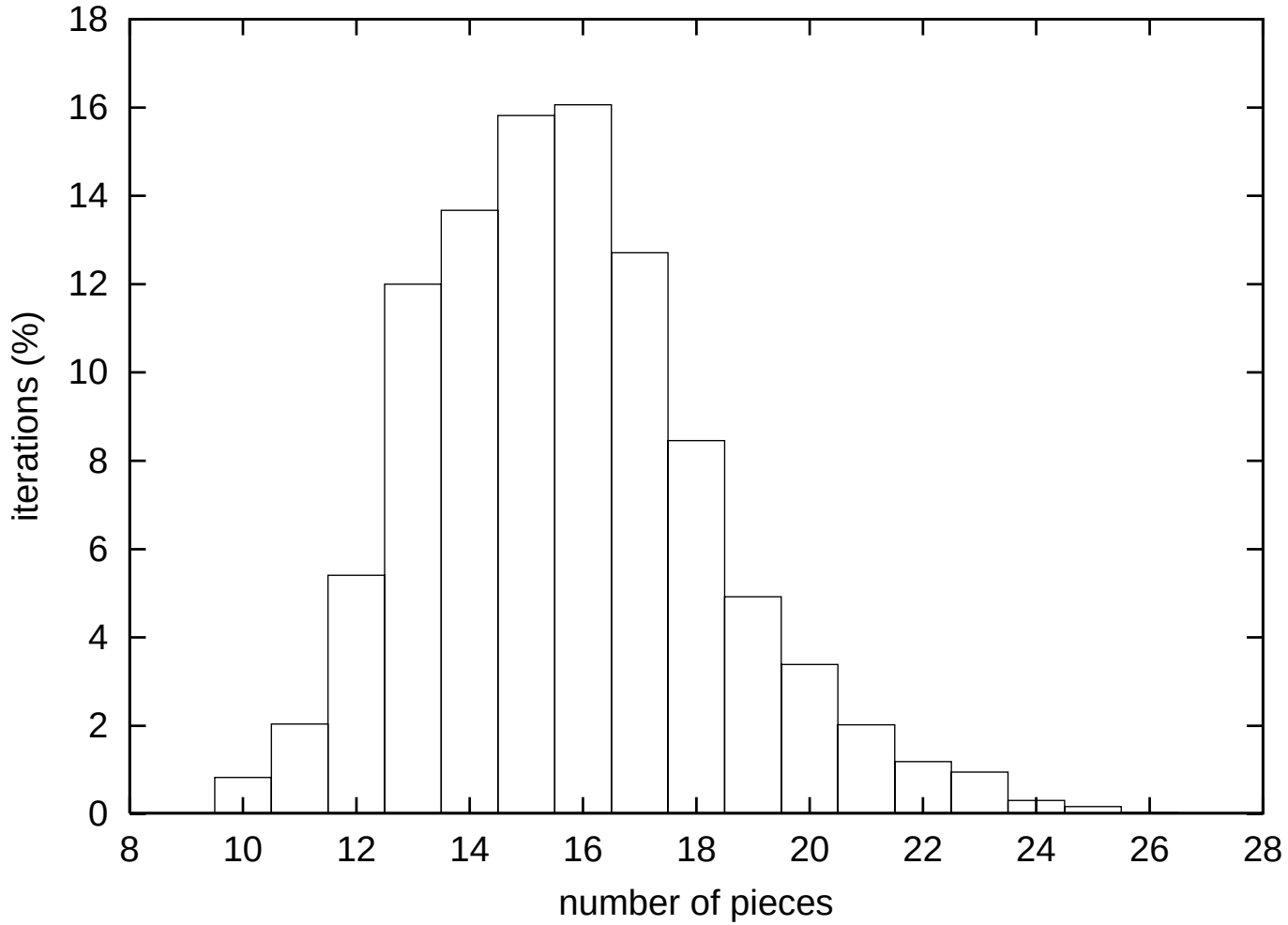
Example: telecom alarm data

- Telecom alarm data
- A single interesting event type has been selected
- 2371 events
- Geom(0.05) for the number of pieces
- Uniform(0.00001,1500) for the intensity
- Uniform(a,b) for the change points
- 510^6 iterations burn-in, actual run 10^6 iterations

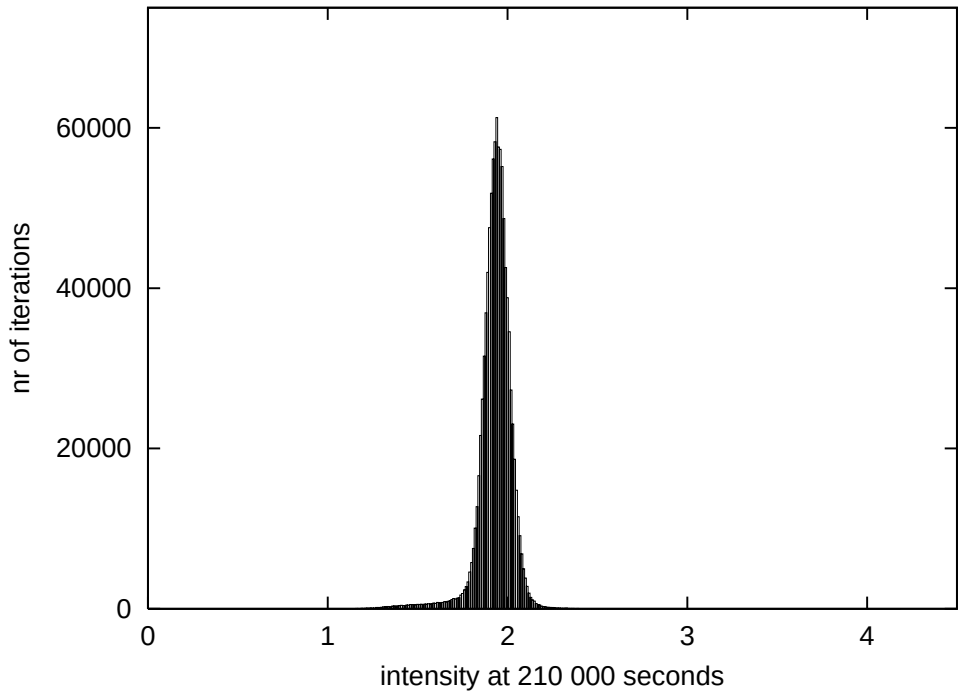
Posterior average and 99 % interval



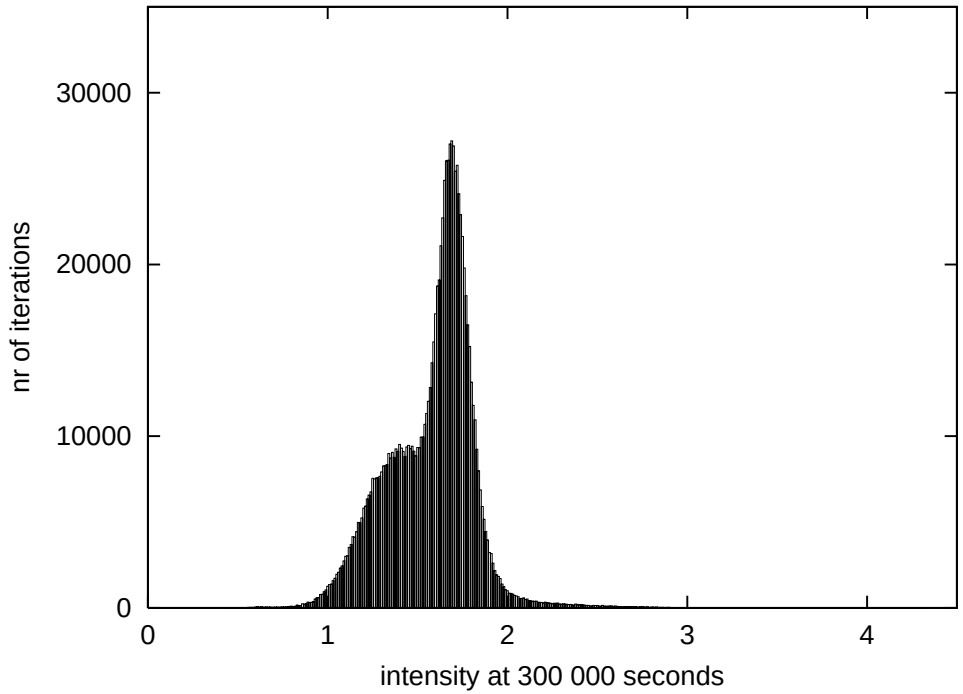
Marginal posterior distribution of the number of pieces



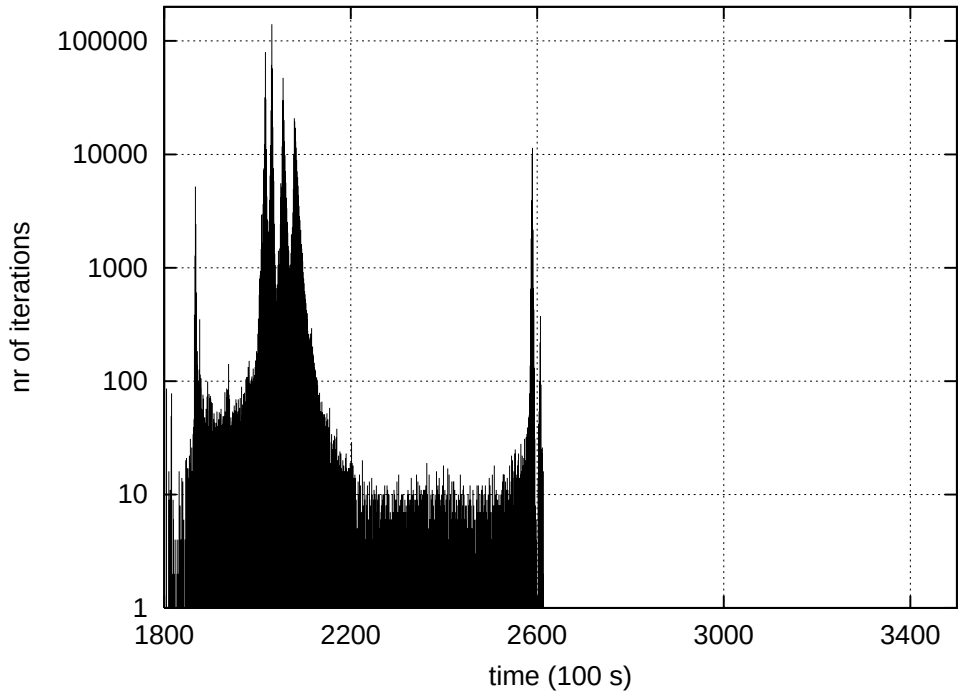
Intensity at 210,000



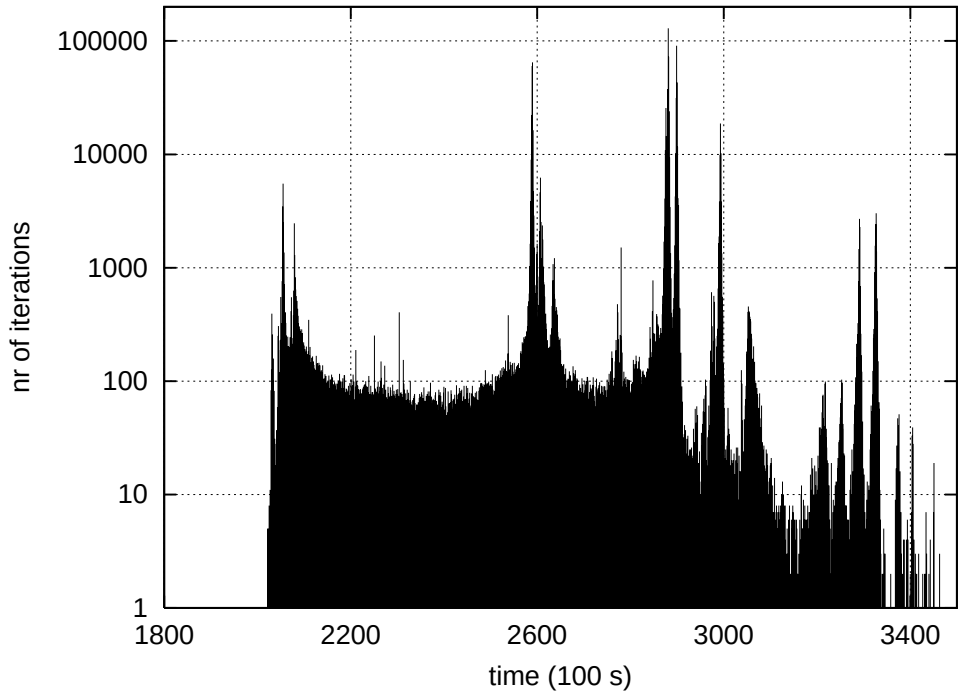
Intensity at 300,000



Distribution of the third change point



Distribution of the sixth change point



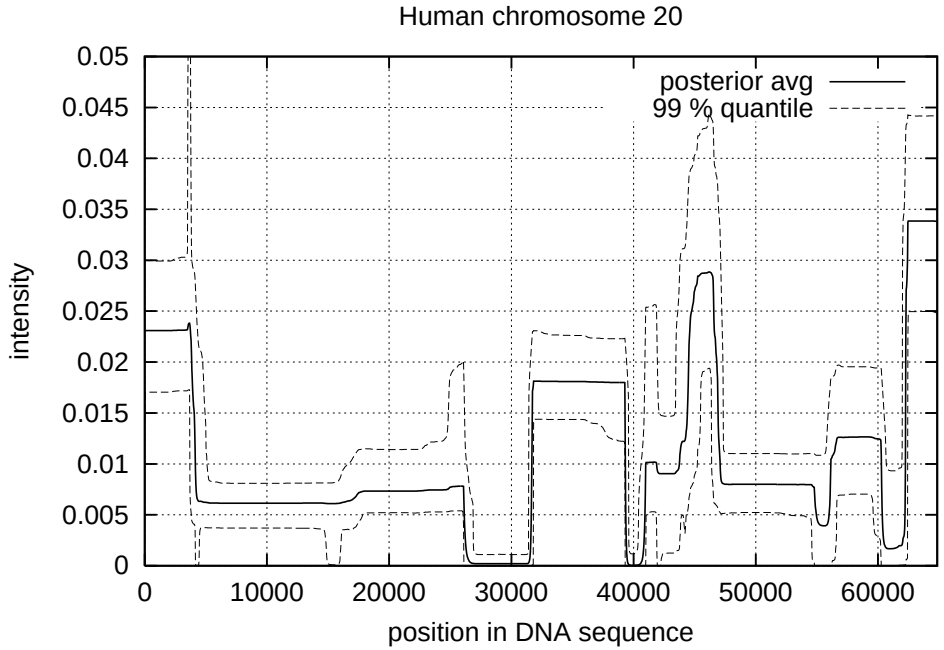
Convergence

Different types of diagnostics — converges, but not very fast

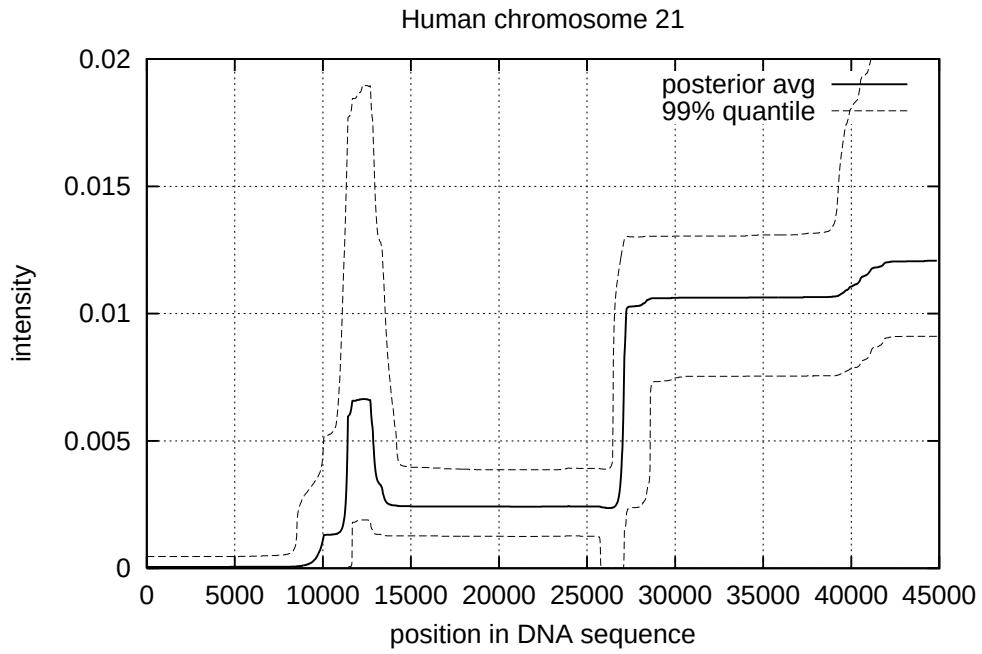
Example: human genome segmentation

- Data: starting points of ORFs in human genome
- Pieces: Geom(α)
- Intensity: $\lambda_i \sim$ Gamma$(0.01, 0.75)$
- Fairly uninformative, converges faster

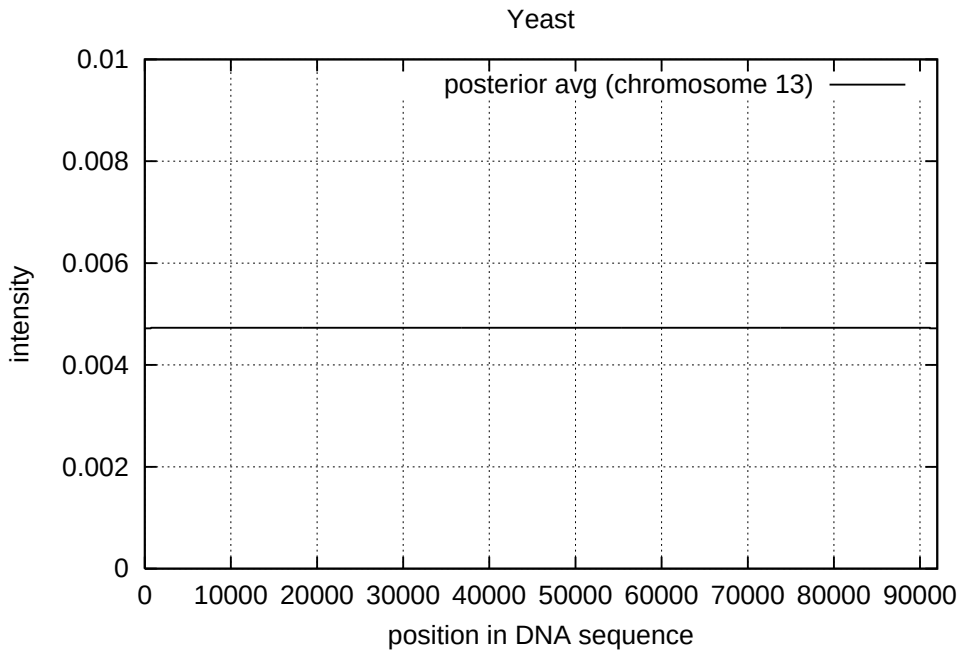
Posterior intensities of ORFs for chromosome 20



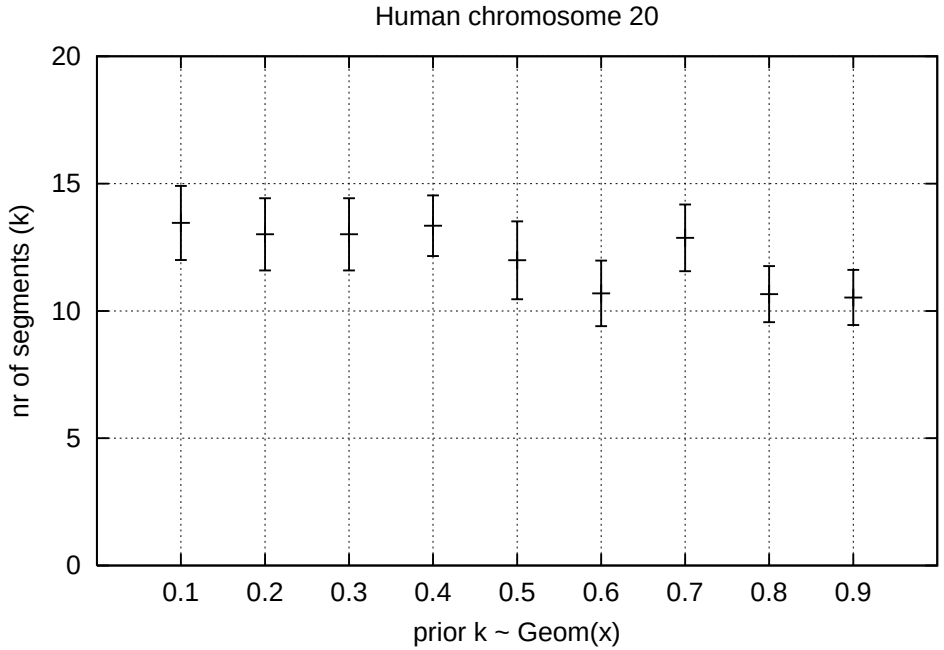
Posterior intensities of ORFs for chromosome 21



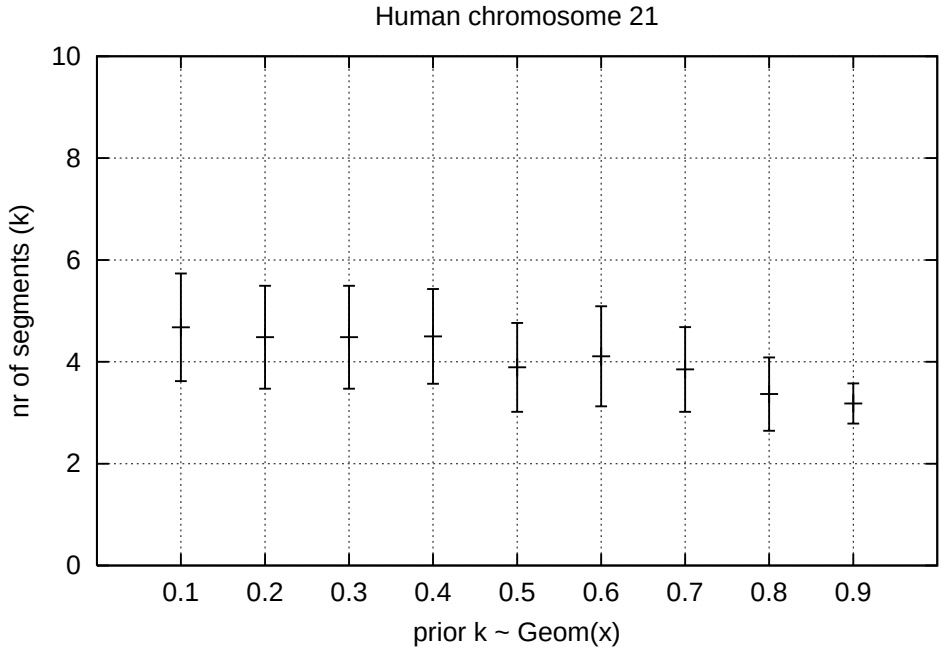
Posterior intensities of ORFs for yeast chr 13



Effect of priors on 20



Effect of priors on 21



Extensions

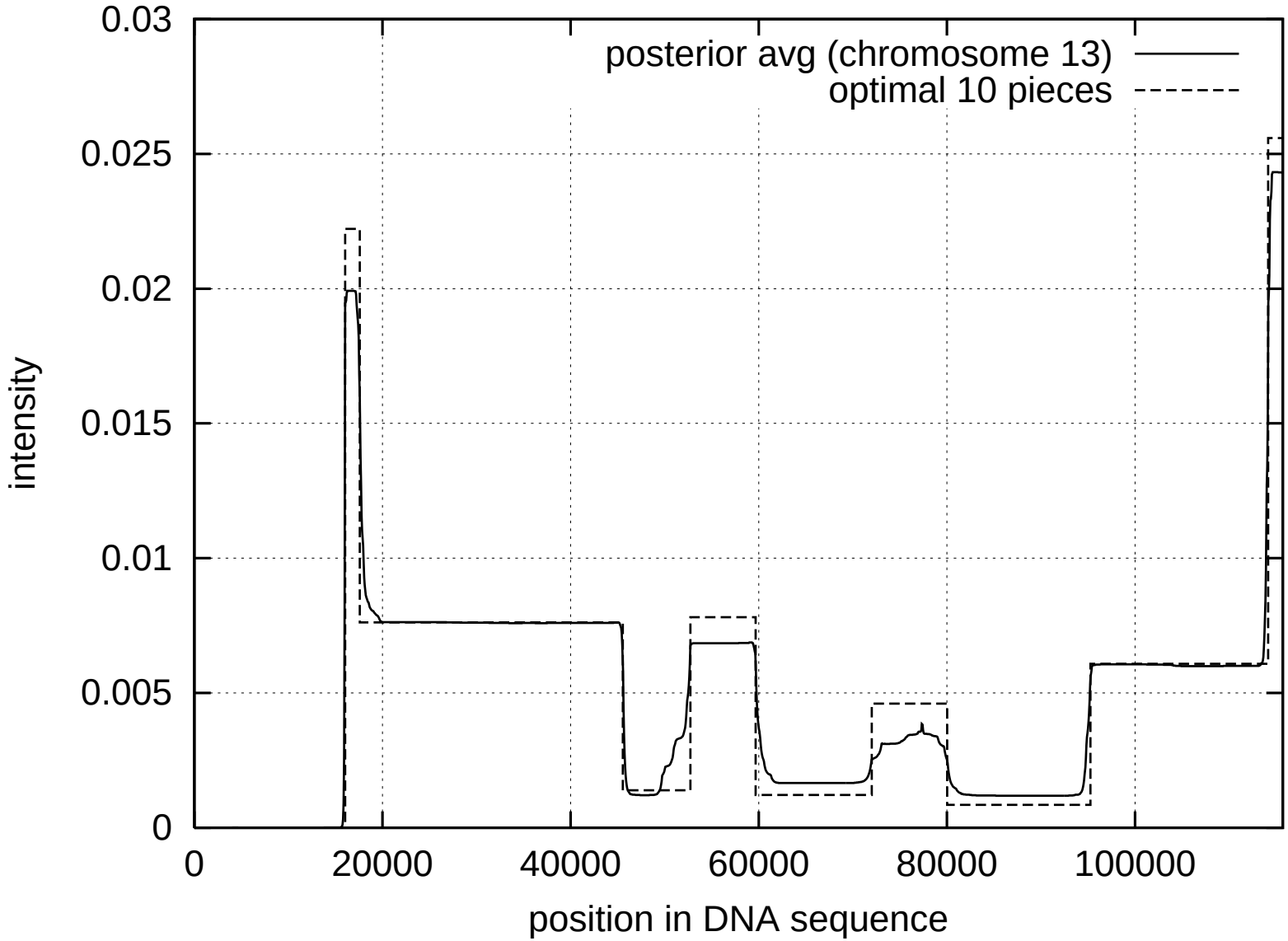
$\lambda_A(t)$ piecewise constant

$\lambda_B(t) = \beta\lambda_A(t)$

can the same intensity be used to describe both A and B

etc.

MCMC and dynamic programming



MCMC and dynamic programming