

Segmentation of paleoecological spatio- temporal count data

Vasko K.⁽¹⁾, Toivonen H.T.T.⁽²⁾, Korhola A.⁽³⁾

(1): CSC – Scientific Computing Ltd, Finland

(2): Department of Computer Science, University of Helsinki

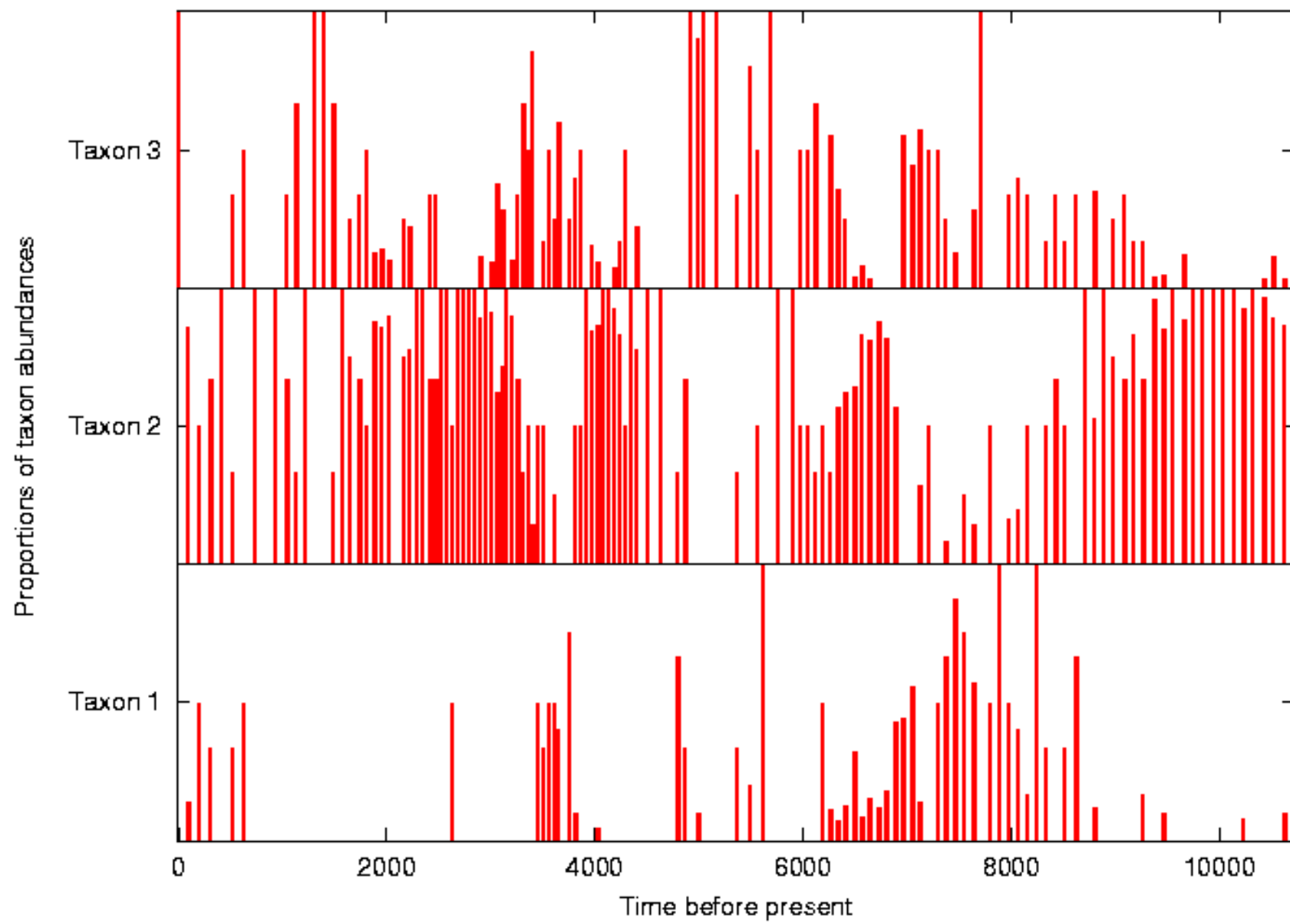
(3): Department of Biological and Environmental Sciences, University of Helsinki

Contents

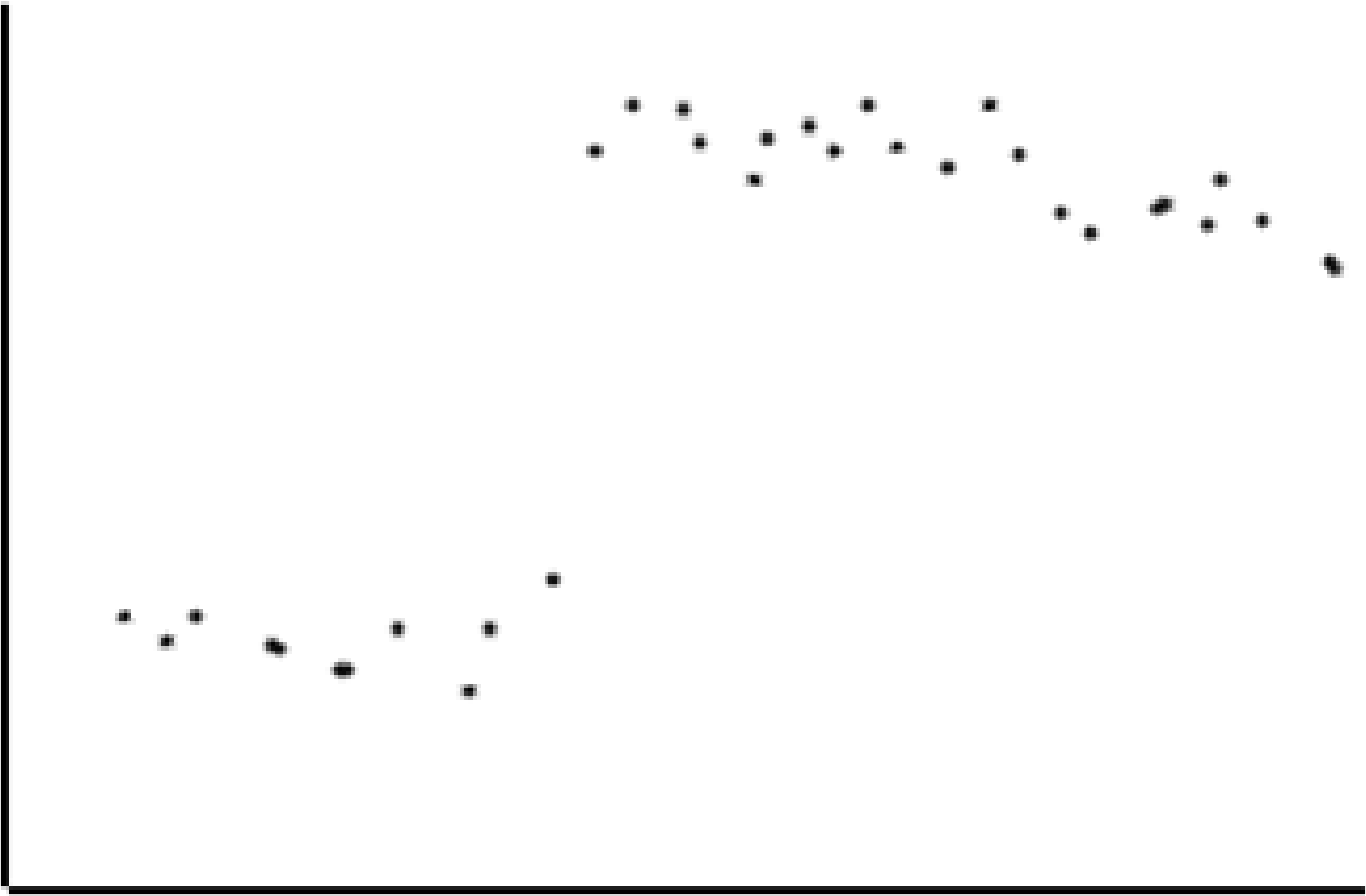
1. Background and motivation
2. Segmenting time series data
3. A Bayesian segmentation model for paleoecological time series
4. Demonstrations
5. Summary

Paleoecological spatio-temporal count data?

- Paleoecological count data = “Abundances of organisms”
- Paleoecological temporal count data = “A time series of paleoecological count data”
- Spatio-temporal = “There are several time series in different spatial locations”

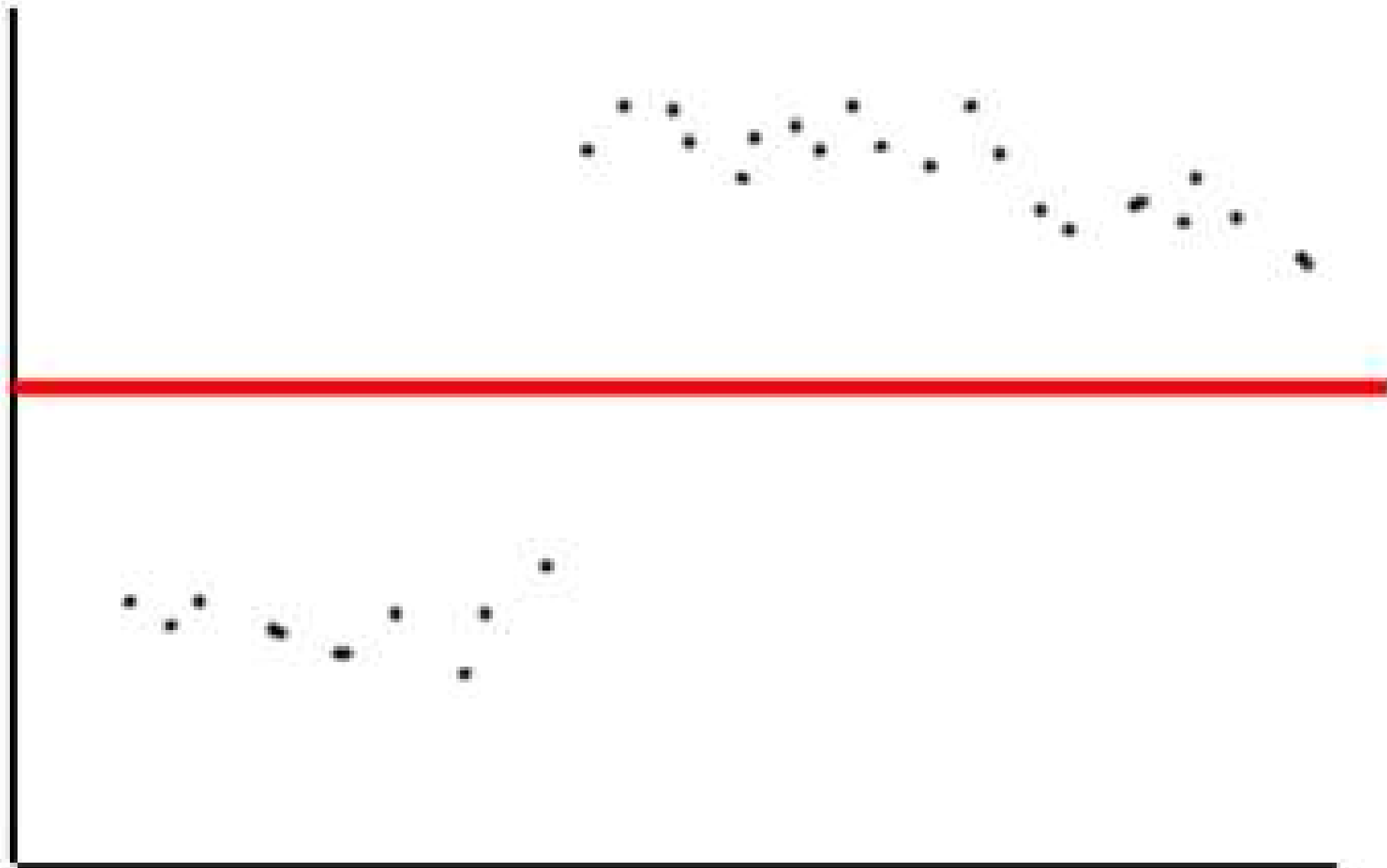


DOMAIN



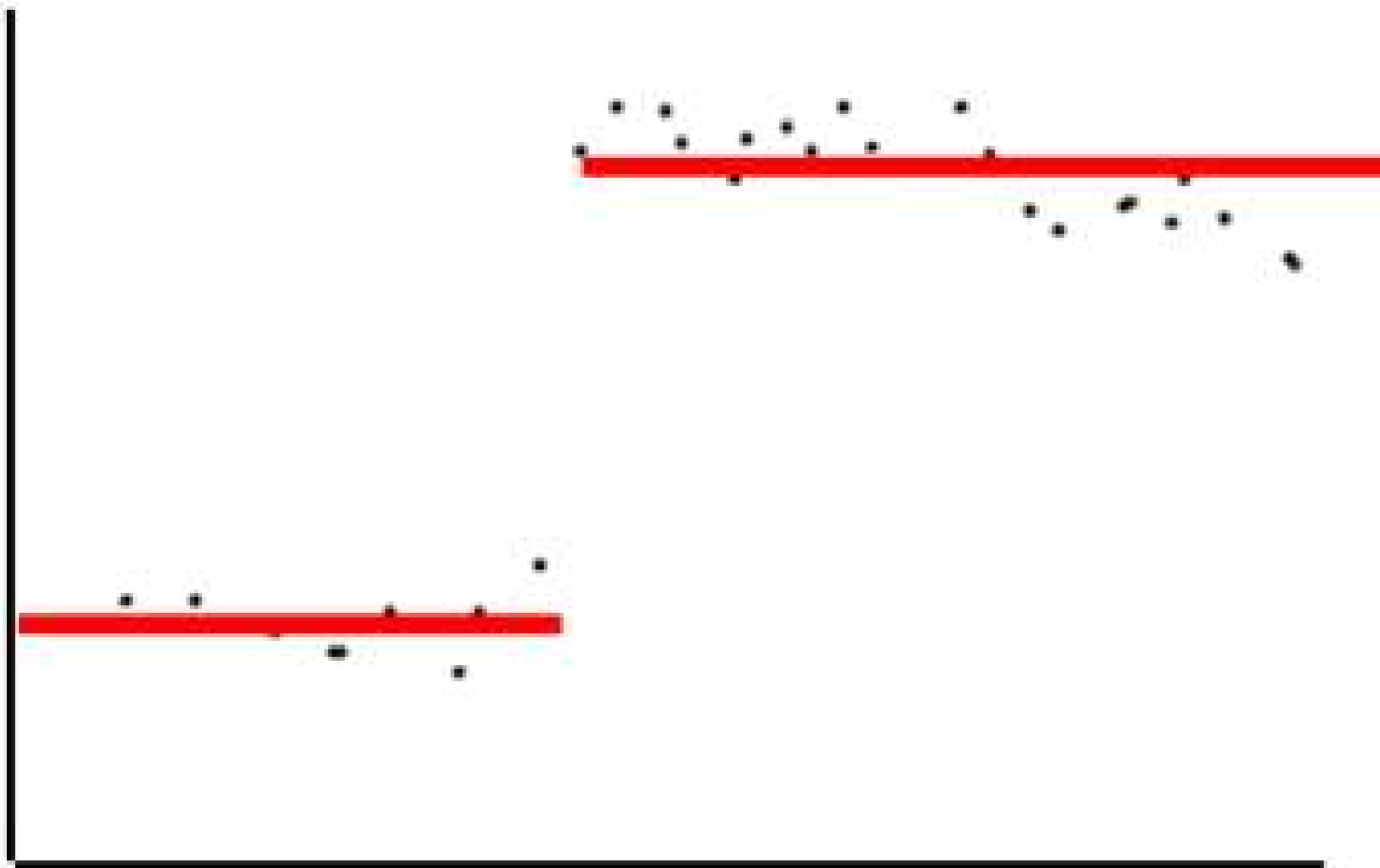
TIME

DOMAIN



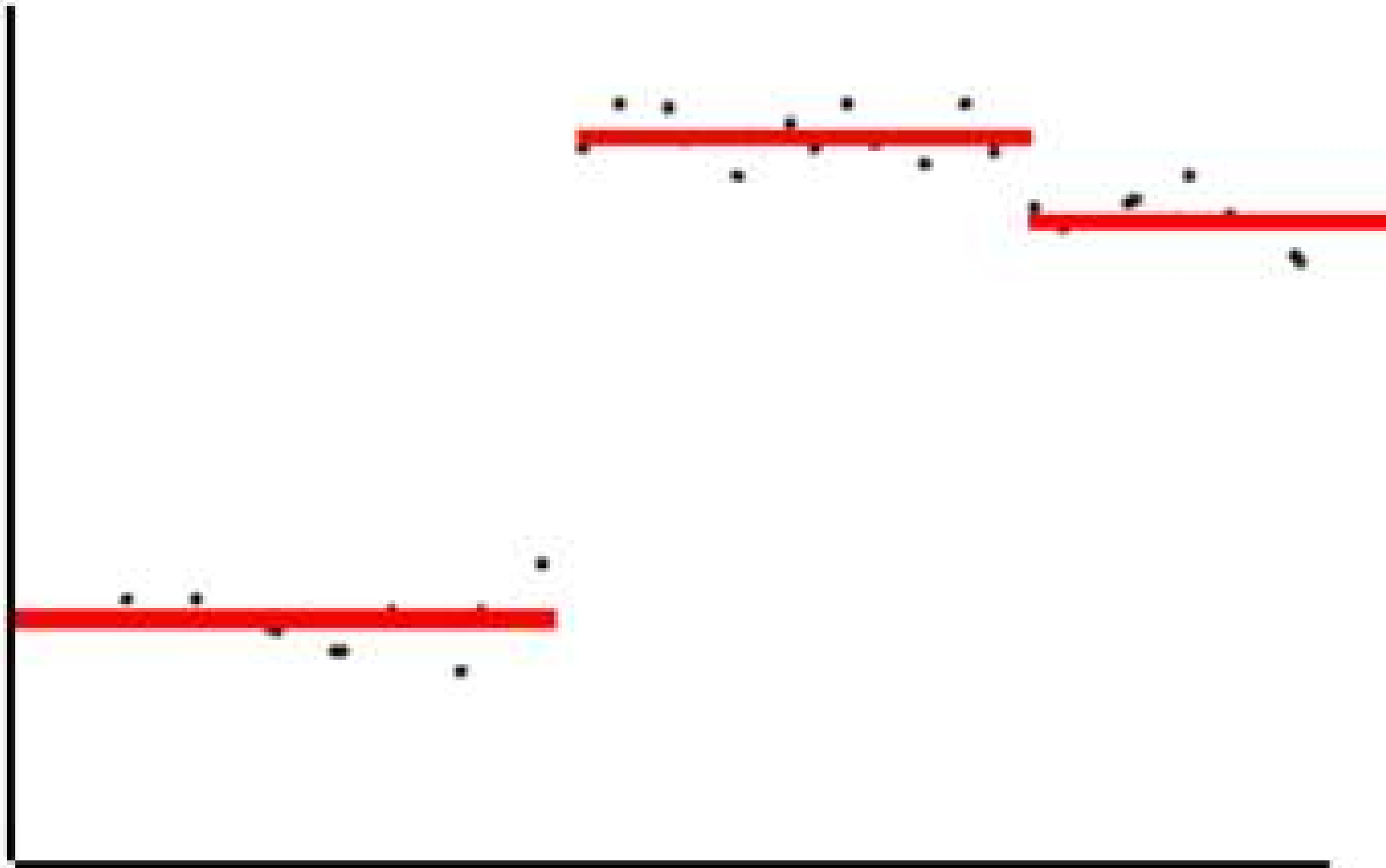
TIME

DOMAIN



TIME

DOMAIN



TIME

Segmenting paleoecological data

Ecological Motivation:

- Reduction of data set to manageable units
- Identification of zones of uniform environmental conditions
- Detection of significant periods of change
- Identification of local vs. regional (global) changes (multiple data sets)

Segmentation problem in a nutshell

- Find a partitioning of a time series to segments
 - Pattern language? Distribution of noise?
- For a fixed number segments it is easy to find a solution
 - Use the parameter values that maximize the score, e.g., the likelihood function.
- However, the number of segments is unknown
- *Ad hoc*-methods widely used by paleoecologists.
- Segmentation is unsupervised learning in nature, i.e., beauty is in the eye of the beholder
- We choose a Bayesian framework

A Bayesian approach to segmentation

- Let T be a set of time series, m the number of segments and $h(m)$ parameters as there is m segments
- Optimal choice is to choose the number of segments that maximize the posterior probability

$$\arg \max_m \Pr[m|T] = \arg \max_m \frac{\Pr[m]\Pr[T|m]}{\sum_{m'} \Pr[m']\Pr[T|m']}$$

Computational/Analytic challenge

- From a computational point of view the hard part is the marginal likelihood ($h(m) = (c(m), p(m))$)

$$\begin{aligned}\Pr[T|m] &= \int \Pr[T, h(m)|m] dh(m) \\ &= \sum_{c(m) \in C_m} \int \Pr[c(m), p(m)] \Pr[T|m, c(m), p(m)] dp(m)\end{aligned}$$

- We need to integrate over all possible m-segmentations $c(m)$ and their parameter values $p(m)$

Approximating the marginal likelihood

- There is no efficient technique to integrate over all m -segmentations => an approximation is needed, e.g. BIC
- We propose an ML-MAP approximation

$$\text{ML-MAP} := \arg \max_m \frac{\Pr[m] \Pr[T|c^*(m), m]}{\sum_{m'} \Pr[m'] \Pr[T|c^*(m'), m']}$$

- where $c^*(m)$ are the maximum likelihood change points
- We need to integrate still over the segment specific parameter values $p(m)$

$$\Pr(T|c^*(m), m) = \int \Pr[T, p(m)|c^*(m), m] dp(m)$$

How to compute ML segment boundaries?

-
- Suppose the parameters of the segments are conditionally independent of each other => Dynamic programming can be applied (an extension of Bellman's idea 1961)
- DP algorithm has a quadratic time complexity w.r.t the number of data points => Heuristic approximations, e.g., top-down and its variants, may be needed

A multinomial segmentation model

- Suppose, for notational simplicity, that there is only one time series
- Let $T=\{y(1),y(2),\dots,y(n)\}$ be a paleoecological abundance time series s.t. $y(i)$ is a vector of organism abundances observed at time point i .
- Segment of T is $S=\{y(i),\dots,y(i+k)\}$, where $(i+k\leq n$ and $k\geq 0)$
- It is assumed segments are independent of each other given zone boundaries, i.e.
 - $Pr[S,S'|c(m)]=Pr[S|c(m)]Pr[S'|c(m)]$, where S and S' are disjoint segments.

A multinomial segmentation model: Likelihood

- m -segmentation of T is a set of m segments $S(1), S(2), \dots, S(m)$ s.t. they are non-overlapping and their union is T
- Due to the independence assumption
 - $Pr[T|c(m)] = Pr[S(1)|c(m)] * \dots * Pr[S(m)|c(m)]$
- **The concept of an environmental segment:** Probabilities of organism occurrences $q(i)$ are constants within a segment $S(i)$.
- If $S(i)$ is a segment of T then for all $y \in S(i)$
 - $y = (y(1), \dots, y(k)) \sim Mult(q(i), y(+))$
- $\Rightarrow Pr[S(i)|c(m)]$ is a product of multinomials having the same probability vector $q(i)$

A multinomial segmentation model: Priors

- We assume occurrence probabilities $q(i) = (q(i,1), \dots, q(i,k))$ at segment i are *a priori* Dirichlet distributed
 - $q(i) \sim \text{Dirichlet}(a(i,1), \dots, a(i,k))$
- We assume a uniform distribution over the different number of segments $\Rightarrow$ Bayes factors are equivalent to the posterior odds, i.e.
 - $P(T|k)/P(T|m) = [P(k)*P(T|k)]/[P(m)*P(T|m)]$
- If independence between organisms is assumed use a binomial or Poisson likelihood.

A multinomial segmentation model: Marginal likelihood

- The local marginal likelihood can be derived on the basis of known identities of Dirichlet integrals

$$\int \text{Pr}(S(i), p(i)) dp(i) = \frac{\Gamma(\sum_k a(i,k)) (\sum_k y(i,k))! \prod_k \Gamma(y(i,k) + a(i,k))}{\prod_k \Gamma(a(i,k)) \prod_k y(i,k)! \Gamma(\sum_k y(i,k) + a(i,k))}$$

- where $y(i,k)$ is the total sum of the occurrences of a taxon k at a segment i
- The total marginal likelihood is a product of the local marginal likelihoods

A multinomial segmentation model: ML-MAP

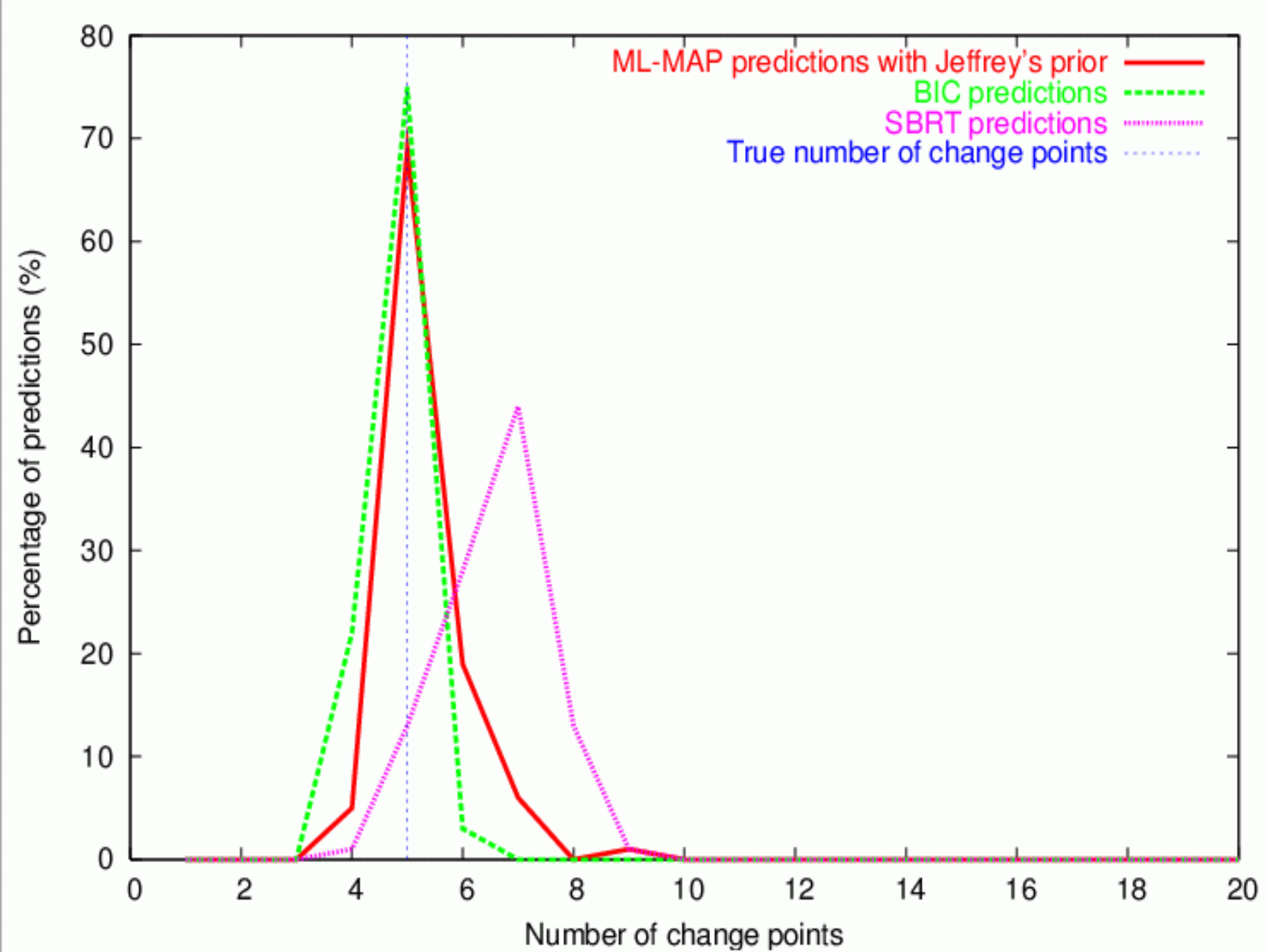
- Using previous identity ML-MAP is available in linear time and space given the ML change points $c^*(m)$ (space $\Omega(|T|)$ is required)

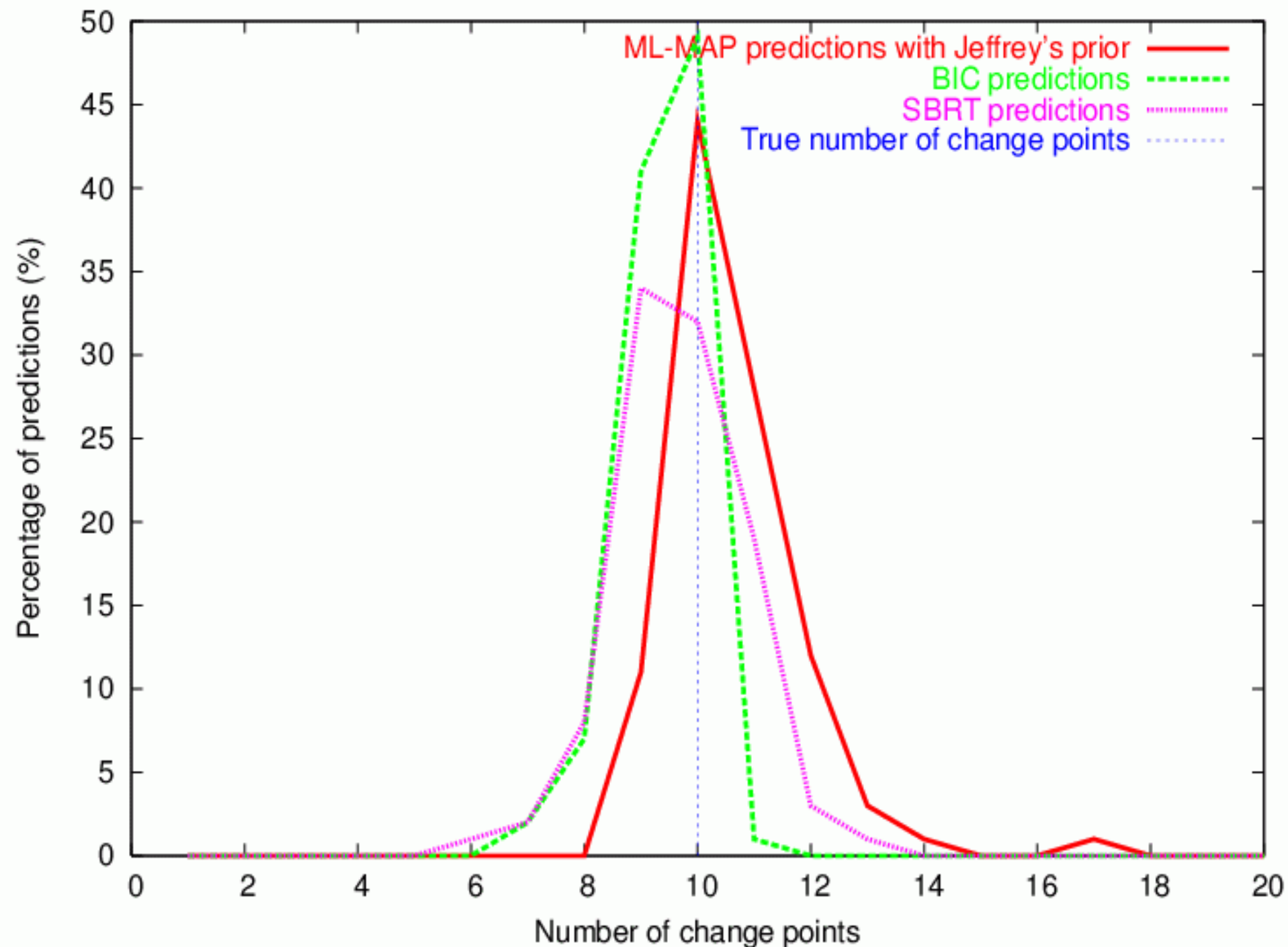
$$\text{ML-MAP} := \arg \max_m \frac{\Pr[m] \Pr[T|c^*(m), m]}{\sum_{m'} \Pr[m'] \Pr[T|c^*(m'), m']}$$

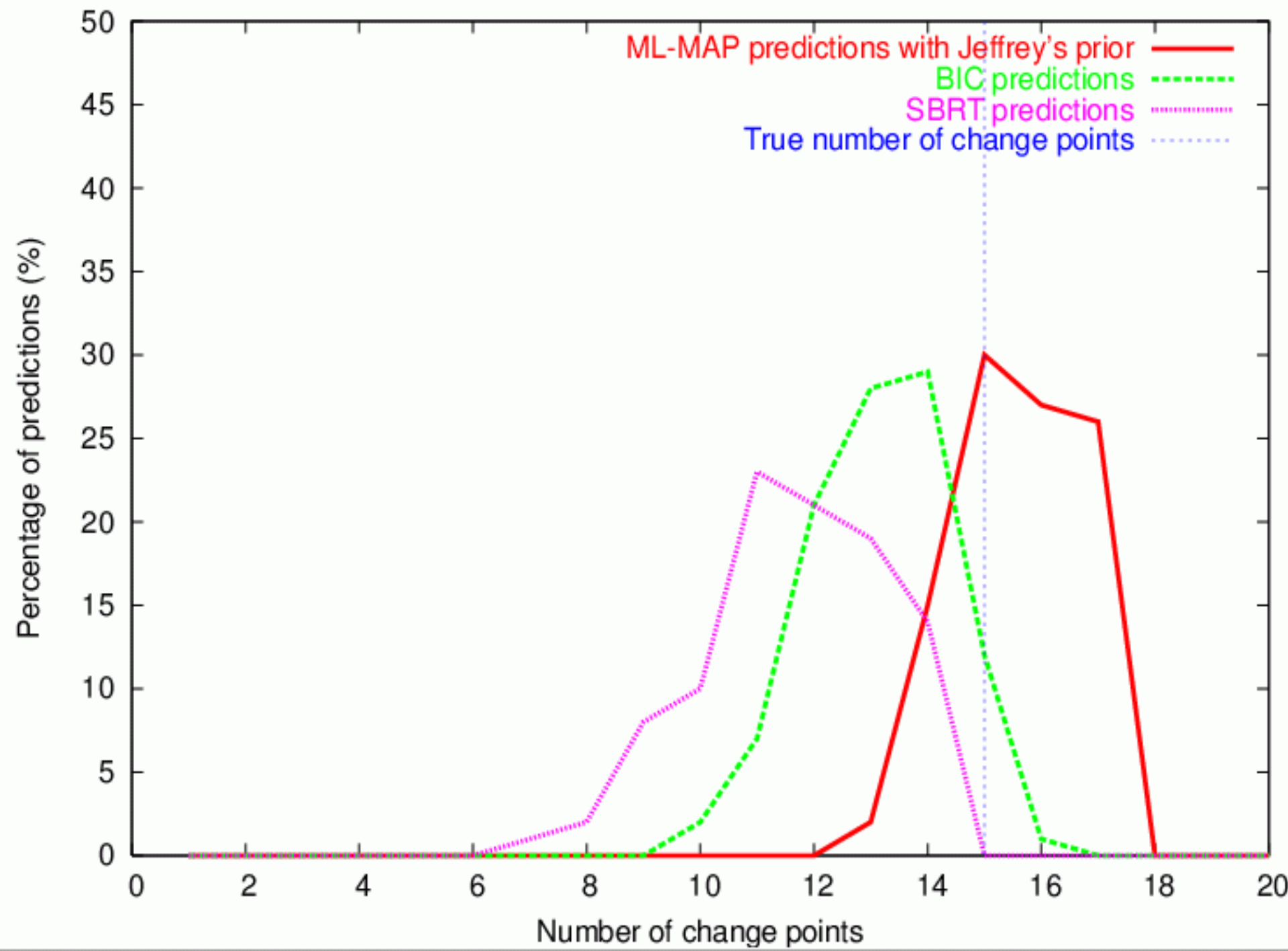
- Obs! It is straightforward to extend the model to take into account a set of time series s.t. they have different time domains, different organisms *etc.*
- => Multiple time series analysis is useful in identification of local vs. regional (global) environmental changes

Experiments

- We carried out experiments with both simulated and real data
- Why simulated data?
- We generated 100 random instances of the following time series
 - Dimension = 50 such that 5, 10 and 15 segments
 - Length of the time series were 150 data points (a typical length in paleoecological studies)
 - Change points uniformly distributed
 - Top-down approximation of ML
 - Probabilities were generated as
 - $p[j] := X[j]/[X[1]+...+X[k]]$, where $X[i] \sim \text{unif}(0,1)$

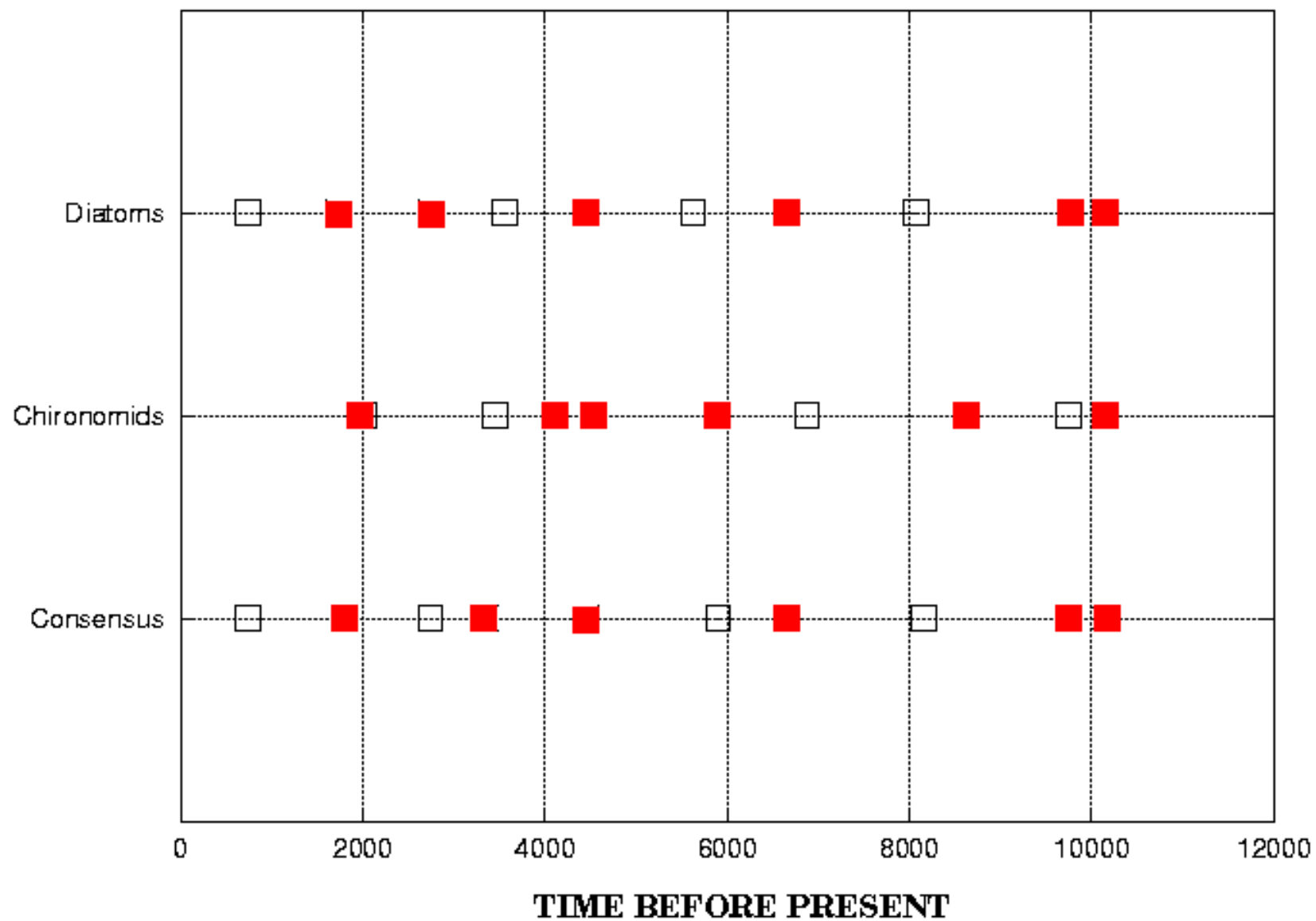






Tsuolbmajavri: Multinomial likelihood

C
O
R
E



Summary

- We represented a probabilistic approach to paleoecological segmentation task and a multinomial zonation model.
- ML-MAP is more reliable than BIC when the number segments is high w.r.t to the length of time series
- Obs! Full details (and much more) are available in
 - “Computational methods and models for paleoecology” by Kari Vasko, PhD Thesis (Chapter 7). The thesis is available in pdf-format at web
 - <http://ethesis.helsinki.fi/julkaisut/mat/tieto/vk/vasko>
 - A hard copy is available by request