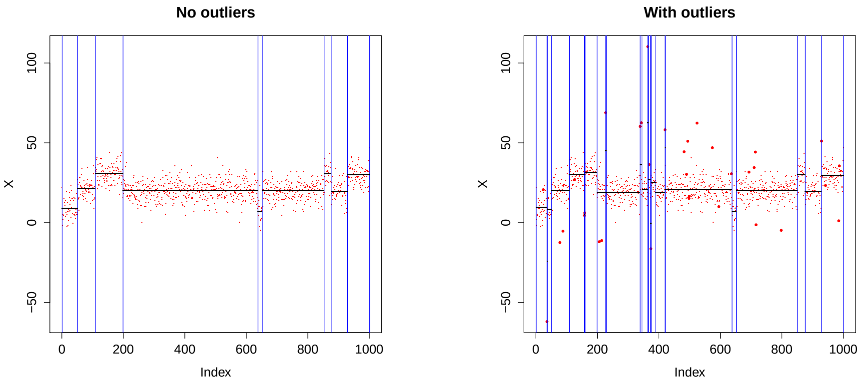
Detecting Outliers in Hidden Markov modeling through Relative Entropy: Applications to Change-Point Detection

V. Perduca, G. Nuel

IBC 2012, Kobe

Change-point detection, HMMs and outliers

- Given an heterogeneous sequence: find the segments in which the signal is homogeneous
- Here: Hidden Markov modeling
- Segmentation models are sensitive to the presence of outliers



Hidden Markov Models

- X_i observed variable
- S_i hidden variable, for $i = 1 \dots, n$

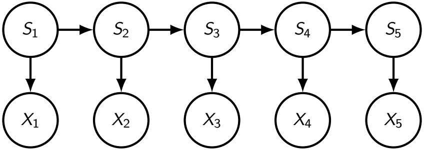
$$\begin{tikzcd}
S_1 \arrow[r] \arrow[d] & S_2 \arrow[r] \arrow[d] & S_3 \arrow[r] \arrow[d] & S_4 \arrow[r] \arrow[d] & S_5 \arrow[d] \\
X_1 & X_2 & X_3 & X_4 & X_5
\end{tikzcd}$$

Factorization of the joint probability distribution

$$\mathbb{P}(S_{1:n} = s_{1:n}, X_{1:n} = x_{1:n}) = \mathbb{P}(S_1 = s_1) \prod_{i=2}^{n} \mathbb{P}(S_i = s_i | S_{i-1} = s_{i-1}) \prod_{i=1}^{n} \mathbb{P}(X_i = x_i | S_i = s_i)$$

Hidden Markov Models

- X_i observed variable
- S_i hidden variable, for $i = 1 \dots, n$

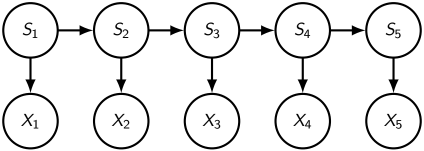


Factorization of the joint probability distribution

$$\mathbb{P}(S_{1:n}, X_{1:n}) = \mathbb{P}(S_1) \prod_{i=2}^{n} \mathbb{P}(S_i | S_{i-1}) \prod_{i=1}^{n} \mathbb{P}(X_i | S_i)$$

Hidden Markov Models

- X_i observed variable
- S_i hidden variable, for $i = 1 \ldots, n$



Factorization of the joint probability distribution

$$\mathbb{P}(S_{1:n}, X_{1:n}) = \mathbb{P}(S_1) \prod_{i=2}^{n} \mathbb{P}(S_i | S_{i-1}) \prod_{i=1}^{n} \mathbb{P}(X_i | S_i)$$

Example (Application to change point detection)

- Level-based: S_i = underlying level of observation X_i
- Segment-based: S_i = segment of X_i with $S_1 = 1$ and $S_n = \#$ segments

Inference in HMMs

If $\mathcal{E} = \{X_{1:n} = x_{1:n}\}$ observed, compute $\mathbb{P}(\mathcal{E}), \mathbb{P}(S_i|\mathcal{E}), \mathbb{P}(S_i|S_{i-1}, \mathcal{E})$...

Backward and Forward recursions (Baum-Welch algorithm)

Standard inference problems solved by *combining* the Forward and Backward quantities

- $F_i(S_i) := \mathbb{P}(S_i, X_{1:i} = x_{1:i})$
- $B_i(S_i) := \mathbb{P}(X_{i+1:n} = x_{i+1:n}|S_i)$,

which are computed recursively:

- $F_i(S_i) = \sum_{S_{i-1}} F_{i-1}(S_{i-1})\mathbb{P}(S_i|S_{i-1})\mathbb{P}(X_i|S_i)$
- $B_{i-1}(S_{i-1}) = \sum_{S_i} \mathbb{P}(S_i|S_{i-1})\mathbb{P}(X_i|S_i)B_i(S_i)$.

E.g. $\mathbb{P}(S_i, \mathcal{E}) = F_i(S_i)B_i(S_i)$

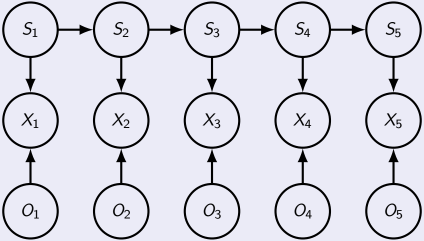
Parameter estimation

EM algorithm: Backward/Forward quantities provide explicit update formulas

Ad hoc model for outlier detection in HMMs

- X_i is an outlier if it is not generated by the underlying HMM
- $\Rightarrow$ extend the HMM with variables for the outliers status [Shah 2006]

Topology and conditional dependencies (homoscedastic Gaussian case)



- $O_i = 1$ iff X_i is outlier; $\mathbb{P}(O_i = 1) = \rho$
- $\mathbb{P}(S_1)$; $\mathbb{P}(S_i | S_{i-1})$: same as for underlying HMM
- $\mathbb{P}(X_i | S_i, O_i = 0) = \mathcal{N}(\mu_{S_i}, \sigma^2)$: same as for underlying HMM
- $\mathbb{P}(X_i | S_i, O_i = 1) = \mathcal{N}(\mu_{S_i}, \sigma^2) + \mathcal{N}(0, \delta^2)$

Inference in the ad hoc model

Inferring outlier posterior probabilities

$$\mathbb{P}(O_i = 1|\mathcal{E}) = \sum_{S_i} \left(\frac{\rho\mathbb{P}(X_i{=}x_i|S_i,O_i{=}1)}{\rho\mathbb{P}(X_i{=}x_i|S_i,O_i{=}1)+(1-\rho)\mathbb{P}(X_i{=}x_i|S_i,O_i{=}0)} \cdot \mathbb{P}(S_i|\mathcal{E}) \right)$$

where

$$\mathbb{P}(S_i|\mathcal{E}) = \frac{F_i(S_i)B_i(S_i)}{\sum_{S_i} F_i(S_i)B_i(S_i)}$$

$$F_i(S_i) = \sum_{S_{i-1}} F_{i-1}(S_i)\mathbb{P}(S_i|S_{i-1})\mathbb{P}(X_i|S_i)$$

Inference in the ad hoc model

Inferring outlier posterior probabilities

$$\mathbb{P}(O_i = 1|\mathcal{E}) = \sum_{S_i} \left(\frac{\rho\mathbb{P}(X_i{=}x_i|S_i,O_i{=}1)}{\rho\mathbb{P}(X_i{=}x_i|S_i,O_i{=}1)+(1-\rho)\mathbb{P}(X_i{=}x_i|S_i,O_i{=}0)} \cdot \mathbb{P}(S_i|\mathcal{E}) \right)$$

where

$$\mathbb{P}(S_i|\mathcal{E}) = \frac{F_i(S_i)B_i(S_i)}{\sum_{S_i} F_i(S_i)B_i(S_i)}$$

$$F_i(S_i) = \sum_{S_{i-1}} F_{i-1}(S_i)\mathbb{P}(S_i|S_{i-1}) \sum_{o=0,1} \mathbb{P}(X_i = x_i|S_i, O_i = o)\mathbb{P}(O_i = o)$$

Inference in the ad hoc model

Inferring outlier posterior probabilities

$$\mathbb{P}(O_i = 1|\mathcal{E}) = \sum_{S_i} \left(\frac{\rho\mathbb{P}(X_i{=}x_i|S_i,O_i{=}1)}{\rho\mathbb{P}(X_i{=}x_i|S_i,O_i{=}1)+(1-\rho)\mathbb{P}(X_i{=}x_i|S_i,O_i{=}0)} \cdot \mathbb{P}(S_i|\mathcal{E}) \right)$$

where

$$\mathbb{P}(S_i|\mathcal{E}) = \frac{F_i(S_i)B_i(S_i)}{\sum_{S_i} F_i(S_i)B_i(S_i)}$$

$$F_i(S_i) = \sum_{S_{i-1}} F_{i-1}(S_i)\mathbb{P}(S_i|S_{i-1}) \sum_{o=0,1} \mathbb{P}(X_i = x_i|S_i, O_i = o)\mathbb{P}(O_i = o)$$

$$B_{i-1}(S_{i-1}) = \sum_{S_i} \mathbb{P}(S_i|S_{i-1}) \sum_{o=0,1} \mathbb{P}(X_i = x_i|S_i, O_i = o)\mathbb{P}(O_i = o)B_i(S_i)$$

EM algorithm for the ad hoc model

Parameter updates: 2 new parameters (ρ, δ^2)

- Transition parameters have same update formulas as in plain HMM
- $\rho = \frac{\sum_{i=1}^{n} \mathbb{P}(O_i{=}1|\mathcal{E})}{n}$
- $\mu_s, \sigma^2, \delta^2$ found as fixed points:

$$\begin{cases} \mu_s & = \frac{\sum_i x_i[\mathbb{P}(S_i{=}s,O_i{=}1|\mathcal{E})\sigma^2+\mathbb{P}(S_i{=}s,O_i{=}0|\mathcal{E})(\sigma^2+\delta^2)]}{\sum_i[\mathbb{P}(S_i{=}s,O_i{=}1|\mathcal{E})\sigma^2+\mathbb{P}(S_i{=}s,O_i{=}0|\mathcal{E})(\sigma^2+\delta^2)]} \\ \sigma^2 & = \frac{\sum_i\sum_s(x_i-\mu_s)^2\mathbb{P}(S_i{=}s,O_i{=}0|\mathcal{E})}{\sum_i\sum_s\mathbb{P}(S_i{=}s,O_i{=}0|\mathcal{E})} \\ \sigma^2+\delta^2 & = \frac{\sum_i\sum_s(x_i-\mu_s)^2\mathbb{P}(S_i{=}s,O_i{=}1|\mathcal{E}))}{\sum_i\sum_s\mathbb{P}(S_i{=}s,O_i{=}1|\mathcal{E})} \end{cases}$$

where $\mathbb{P}(S_i, O_i|\mathcal{E}) = \frac{\rho\mathbb{P}(X_i{=}x_i|S_i,O_i{=}1)}{\rho\mathbb{P}(X_i{=}x_i|S_i,O_i{=}1)+(1-\rho)\mathbb{P}(X_i{=}x_i|S_i,O_i{=}0)} \cdot \mathbb{P}(S_i|\mathcal{E})$

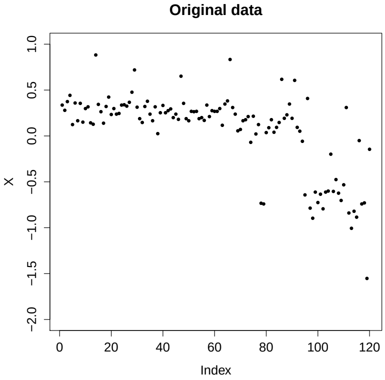
Initialization

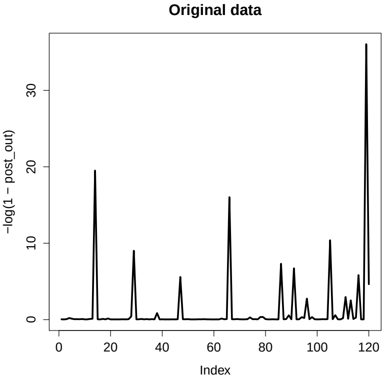
k-means algorithm and z-score

Application of the ad hoc model to real data

Ad hoc model performs well on data simulated by... ad hoc model. What about real data?

- CNV dataset from breast cancer cell line BT474 [Snijders 2001]
- Level-based model: S_i = level of observation X_i
- Parameters in the ad hoc model estimated with the EM algorithm

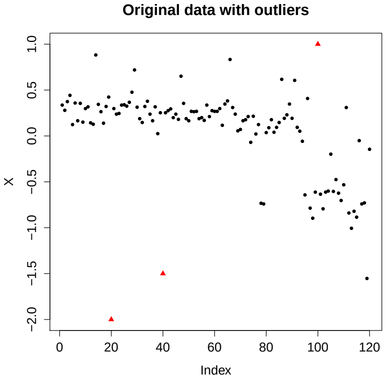


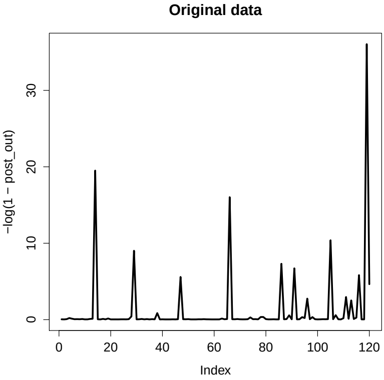


Application of the ad hoc model to real data

Ad hoc model performs well on data simulated by... ad hoc model. What about real data?

- CNV dataset from breast cancer cell line BT474 [Snijders 2001]
- Level-based model: $S_i =$ level of observation X_i
- Parameters in the ad hoc model estimated with the EM algorithm

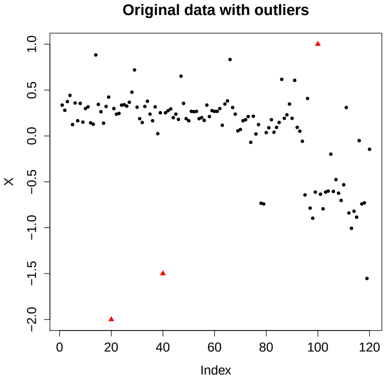


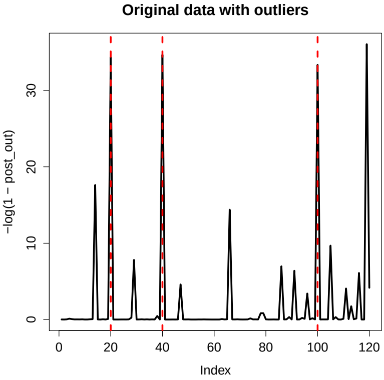


Application of the ad hoc model to real data

Ad hoc model performs well on data simulated by... ad hoc model. What about real data?

- CNV dataset from breast cancer cell line BT474 [Snijders 2001]
- Level-based model: S_i = level of observation X_i
- Parameters in the ad hoc model estimated with the EM algorithm





Outlier detection through relative entropy

Intuition

- If $X_i = x_i$ is an outlier than it must have a strong influence on $\mathbb{P}(S_{1:n}|\mathcal{E}) = \mathbb{P}(S_{1:n}|X_{1:n} = x_{1:n})$
- As a consequence, $\mathbb{P}(S_{1:n}|X_{1:n} = x_{1:n})$ must differ significantly from $\mathbb{P}(S_{1:n}|X_{-i} = x_{-i})$
- We can try to use the relative entropy

$$K_i := \sum_{S_{1:n}} \mathbb{P}(S_{1:n}|X_{-i} = x_{-i}) \log \frac{\mathbb{P}(S_{1:n}|X_{-i} = x_{-i})}{\mathbb{P}(S_{1:n}|X_{1:n} = x_{1:n})}$$

for outlier detection: *"the higher K_i the more likely $X_i = x_i$ is an outlier"*

- Technical problem: *how to compute K_i?*

Computing K_i

- By using naively back/forw recursions, the complexity for computing K_i for a given i is $O(n)$ $\Rightarrow$ the overall complexity is $O(n^2)$

Linear time algorithm for computing K_i for all $i = 1, \dots, n$

$$K_i = \sum_{S_i} \mathbb{P}(S_i | X_{-i} = x_{-i}) \log \frac{\mathbb{P}(S_i | X_{-i} = x_{-i})}{\mathbb{P}(S_i | X_{1:n} = x_{1:n})},$$

with

$$\mathbb{P}(S_i | X_{-i} = x_{-i}) = \frac{F_i^*(S_i) B_i(S_i)}{\sum_{S_{i-1}} F_i^*(S_{i-1}) B_i(S_{i-1})}$$

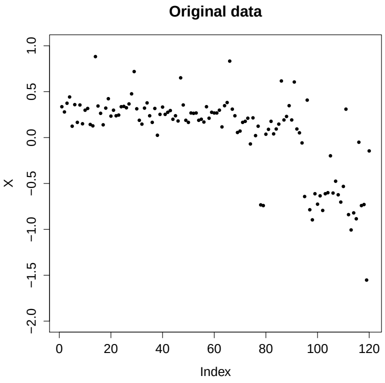
where B and F are the standard back/forw quantities for HMMs and

$$F_i^*(S_i) = \sum_{S_{i-1}} F_{i-1}(S_{i-1}) \mathbb{P}(S_i | S_{i-1}).$$

$\Rightarrow$ the overall complexity is $O(n)$

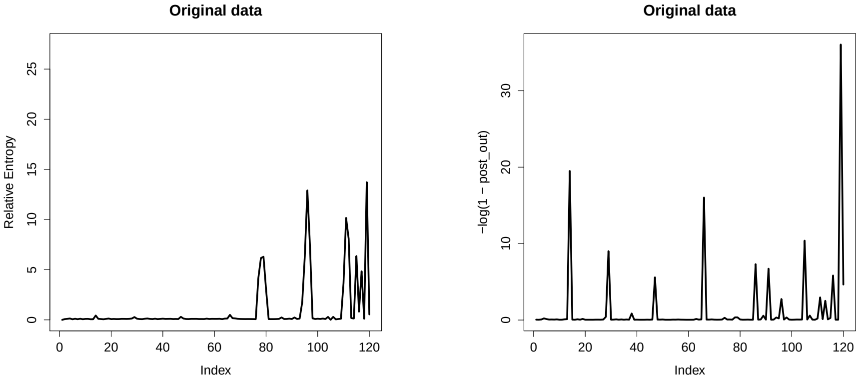
Application of the relative entropy based method to real data

- Same CNV dataset as before
- Parameters in the underlying HMM estimated with the EM algorithm



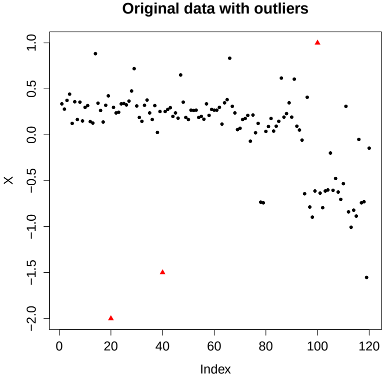
Application of the relative entropy based method to real data

- Same CNV dataset as before
- Parameters in the underlying HMM estimated with the EM algorithm



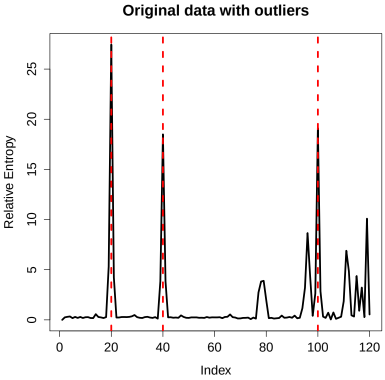
Application of the relative entropy based method to real data

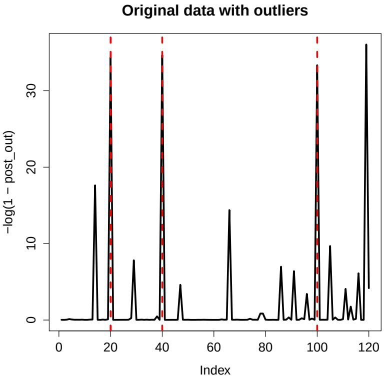
- Same CNV dataset as before
- Parameters in the underlying HMM estimated with the EM algorithm



Application of the relative entropy based method to real data

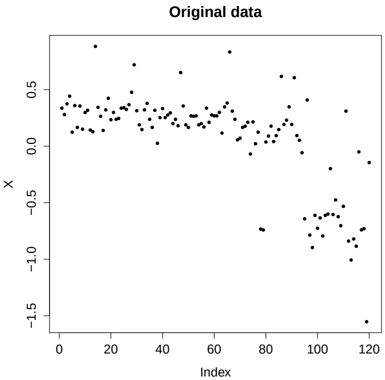
- Same CNV dataset as before
- Parameters in the underlying HMM estimated with the EM algorithm





Comparison on real CNV dataset

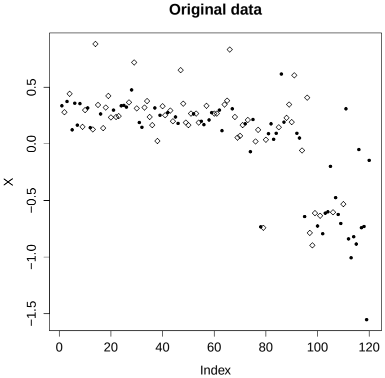
- Original data: $n = 120$ observations
- H0: random samples of $n/2$ observations from the original dataset
- H1: $\rho \times n/2$ outliers added with $\mathcal{N}(0, \delta^2)$ $(\rho = 0.05, \delta = 6)$



- Parameters estimated
- Global statistics for ad hoc model: $T = \max_{i=1,\dots,n} \mathbb{P}(O_i = 1|\mathcal{E})$
- Global statistics for relative entropy method: $S = \max_{i=1,\dots,n} K_i$

Comparison on real CNV dataset

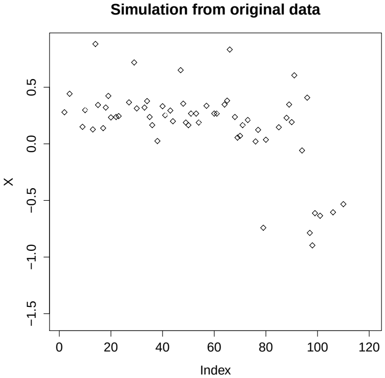
- Original data: $n = 120$ observations
- H0: random samples of $n/2$ observations from the original dataset
- H1: $\rho \times n/2$ outliers added with $\mathcal{N}(0, \delta^2)$ ($\rho = 0.05, \delta = 6$)



- Parameters estimated
- Global statistics for ad hoc model: $T = \max_{i=1,\ldots,n} \mathbb{P}(O_i = 1|\mathcal{E})$
- Global statistics for relative entropy method: $S = \max_{i=1,\ldots,n} K_i$

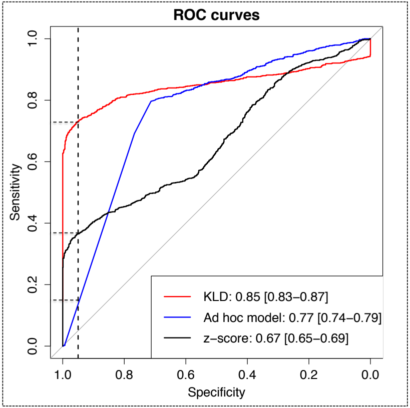
Comparison on real CNV dataset

- Original data: $n = 120$ observations
- H0: random samples of $n/2$ observations from the original dataset
- H1: $\rho \times n/2$ outliers added with $\mathcal{N}(0, \delta^2)$ $(\rho = 0.05, \delta = 6)$



- Parameters estimated
- Global statistics for ad hoc model: $T = \max_{i=1,\dots,n} \mathbb{P}(O_i = 1|\mathcal{E})$
- Global statistics for relative entropy method: $S = \max_{i=1,\dots,n} K_i$

Results of comparison on real data



Discussion

- When data is not generated accordingly to the ad hoc model, the method based on relative entropy is more performant
- The method based on z-score is not satisfactory

Final word

Conclusions

- Ad hoc model:
  - + Outlier explicit modeling, convenient for simulating
  - − Intricate EM algorithm
  - − Very sensitive, false positives
- Method based on relative entropy:
  - + Model free
  - + Parameter estimation simple to implement and fast
  - + Robust

Perspectives

- Comparison with standard outliers detection methods (e.g. LOF)
- Local statistics for outlier detection based on relative entropy
- Application to biological data

References

J. Fridlyand et al. *Hidden Markov models approach to the analysis of array CGH data*; Journal of multivariate analysis, 2004.

A. Olshen et al. *Circular binary segmentation for the analysis of array-based DNA copy number data*; Biostatistics, 2004.

J. Bilmes. *A gentle tutorial of the EM algorithm and its application to parameter estimation for Gaussian mixture and hidden Markov models*; International Computer Science Institute, 1998.

S.S. Shah et al. *Integrating copy number polymorphisms into array CGH analysis using a robust HMM*; Bionformatics, 2006.

A. Snijders et al. *Assembly of microarrays for genome-wide measurement of DNA copy number by CGH*; Nature Genetics, 2001.

Appendix

Initialization of the EM algorithm for the ad hoc model

Using z-score

- Cluster the observations with the k-means algorithm
- μ_{S_i} = mean of all the observations within the same cluster
- $\sigma^2 = \frac{\sum_{i=1}^{n}(X_i - \mu_{S_i})^2}{n}$
- Compute for each observation its z-score:

$$z_i = \frac{X_i - \mu_{S_i}}{\sigma} \sim \mathcal{N}(0,1)$$

- If $|z_i| > 1.96$, then $O_i = 1$
- Compute ρ and δ^2

Relative entropy

Definition

Given two probability distributions p and q their relative entropy (Kullback-Leibler divergence) is

$$D_{KL}(p||q) := \int_z p(z) \log \frac{p(z)}{q(z)} dz = \mathbb{E}_p[\log p] - \mathbb{E}_p[\log q]$$

Properties

- $D_{KL}(p||q)$ is a non-symmetric measure of the distance between p and q: it measures the extra number of bits required for encoding events sampled from p using a code based on q
- Monte Carlo estimation: if $x_1, \ldots, x_N$ is a sample of p, then

$$D_{KL}(p||q) \approx \frac{\sum_{j=1}^{N} \log p(x_j) - \log q(x_j)}{N}$$

- For common (e.g. normal) distributions: exact closed forms of D_{KL}

Parameters for simulations

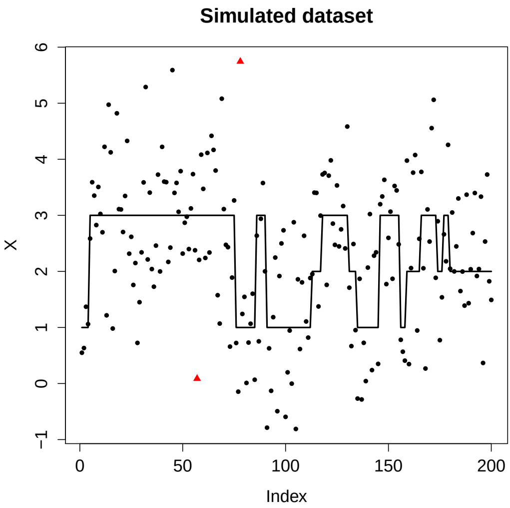
Homoscedastic Gaussian ad hoc model

- $\mathbb{P}(O_i = 1) = \rho$
- $\mathbb{P}(X_i | S_i, O_i = 0) = \mathcal{N}(\mu_{S_i}, \sigma^2)$
- $\mathbb{P}(X_i | S_i, O_i = 1) = \mathcal{N}(\mu_{S_i}, \sigma^2) + \mathcal{N}(0, \delta^2)$

Parameters

- $S_i \in \{1, 2, 3\}$
- $\mathbb{P}(S_1 = 1) = 1$
- $\mathbb{P}(S_i | S_{i-1}) = \begin{pmatrix} 1-\eta & \eta/2 & \eta/2 \\ \eta/2 & 1-\eta & \eta/2 \\ \eta/2 & \eta/2 & 1-\eta \end{pmatrix}$ with $\eta = 0.05$; i=2,...,n
- $\mu_1 = 1$, $\mu_2 = 2$, $\mu_3 = 3$
- $\sigma = 1$
- $\rho = 0.05$
- δ `in seq(from=0.00,to=4.50,by=0.50)`

Example of simulation under H1



Validation of the ad hoc model on a toy example

- Simulations done with the homoscedastic Gaussian ad hoc model:
  - H0: no outlier ($\delta = 0$)
  - H1: presence of outliers ($\delta \neq 0$)
- Global statistics: $T = \max_{i=1,\dots,n} \mathbb{P}(O_i = 1|\mathcal{E})$
- $\mathbb{P}(O_i|\mathcal{E})$ computed using the true parameters

δ	$AUC(T)$
0.00	0.52 [0.48,0.55]
0.50	0.49 [0.46,0.53]
1.00	0.56 [0.52,0.59]
1.50	0.74 [0.71,0.78]
2.00	0.87 [0.85,0.89]
2.50	0.93 [0.91,0.95]
3.00	0.97 [0.95,0.98]
3.50	0.99 [0.98,0.99]
4.00	0.99 [0.99,1.00]
4.50	0.99 [0.99,1.00]

Validation of the relative entropy method on the toy example and comparison

- Global statistics: $S = \max_{i=1,\ldots,n} K_i$

δ	$AUC(S)$	$AUC(T)$
0.00	0.50 [0.47,0.54]	0.52 [0.48,0.55]
0.50	0.50 [0.46,0.53]	0.49 [0.46,0.53]
1.00	0.52 [0.49,0.56]	0.56 [0.52,0.59]
1.50	0.55 [0.52,0.59]	0.74 [0.71,0.78]
2.00	0.69 [0.66,0.73]	0.87 [0.85,0.89]
2.50	0.76 [0.73,0.79]	0.93 [0.91,0.95]
3.00	0.84 [0.81,0.87]	0.97 [0.95,0.98]
3.50	0.87 [0.85,0.90]	0.99 [0.98,0.99]
4.00	0.92 [0.91,0.94]	0.99 [0.99,1.00]
4.50	0.96 [0.95,0.97]	0.99 [0.99,1.00]