Choisir le nombre de segments dans des modèles de segmentation : un estimateur de vraisemblance pénalisée

A. Cleynen, E. Lebarbier

Harvard School of Public Health / AgroParisTech / INRA

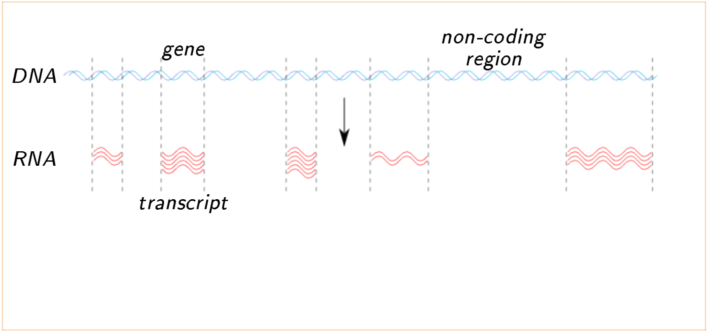
Journées MAS, 29 Aout 2014

Outline

1. Biological framework
2. Main Result
   - Framework
   - Scheme of the proof
   - Corrolary
3. Illustration
   - Segmentation procedure
   - Simulation study

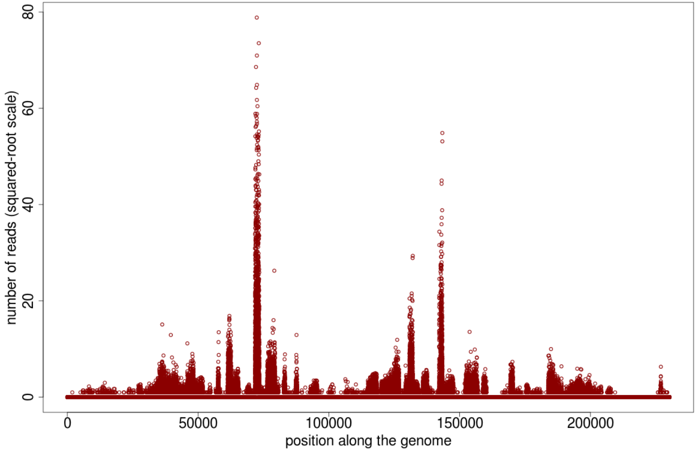
From DNA to RNA

Transcription step

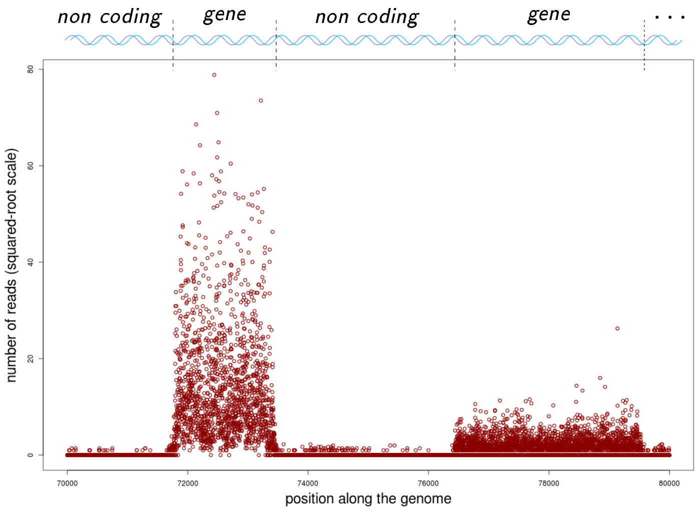


RNA-Seq data

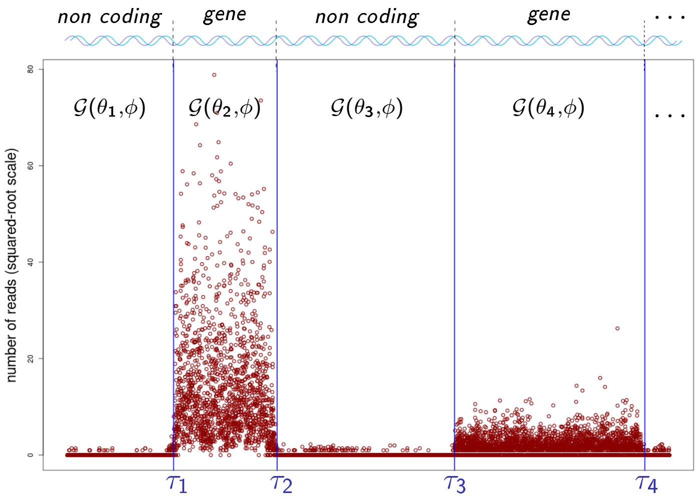
Mapping to the genome



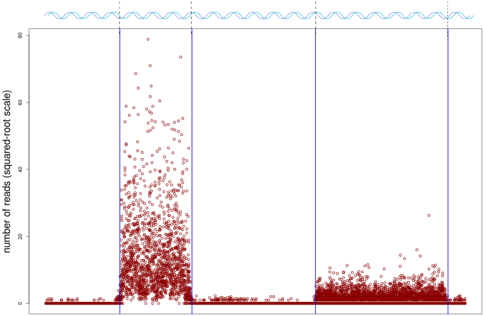
RNA-Seq data and annotation



RNA-Seq data and segmentation



Notations and model



- n length of profile
- m a partition of $[\![1, n]\!]$
- $|m|$ nb of segments of m
- J a segment of m
- $|J|$ length of segment J

$$\forall J \in m, \forall t \in J, Y_t \sim \mathcal{NB}(\theta_J, \phi) \quad [1]$$

[1] Cleynen *et. al.*, (2013)

Outline

1. Biological framework

2. Main Result
   - Framework
   - Scheme of the proof
   - Corrolary

3. Illustration
   - Segmentation procedure
   - Simulation study

Penalized log-likelihood framework

$$s(t) = \mathcal{NB}(p_t, \phi)$$

Collection of models $\mathcal{S}_m = \{s_m \mid \forall J \in m,\ \forall t \in J,\ s_m(t) = \mathcal{NB}(p_J, \phi)\}$.

Notations

- parameter ϕ known.
- $Y_J = \sum_{t \in J} Y_t$
- $\bar{Y}_J = Y_J / |J|$
- $\mathcal{M}_n$: a collection of partitions of $[\![1, n]\!]$,
- $E_t = \mathbf{E}(Y_t) = \phi \frac{1-p_t}{p_t}$,
- $E_J = \sum E_t,\ \bar{E}_J = E_J / |J|$,

Penalized log-likelihood framework (2)

Log-likelihood contrast $\gamma(u)$
Minimal contrast estimator $\hat{s}_m = \arg\min_{u \in \mathcal{S}_m} \gamma(u)$

$$\rightarrow \text{collection } \mathcal{S} = \{(\hat{s}_m)_{m \in \mathcal{M}_n}\}$$

Kullback-Leibler distance $K(s,u) = \mathbf{E}[\gamma(u) - \gamma(s)]$,
Projection $\bar{s}_m = \arg\min_{u \in \mathcal{S}_m} K(s,u)$ of s on $\mathcal{S}_m$:

Goal
Select estimator $\hat{s}_{m(s)}$ from collection $\mathcal{S} = \{(\hat{s}_m)_{m \in \mathcal{M}_n}\}$ which minimizes the Kullback-Leibler risk.
$\rightarrow$ Requires the knowledge of s.

Penalized log-likelihood framework (2)

Log-likelihood contrast $\gamma(u)$
Minimal contrast estimator $\hat{s}_m = \arg\min_{u \in \mathcal{S}_m} \gamma(u)$

$$\rightarrow \text{collection } \mathcal{S} = \{(\hat{s}_m)_{m \in \mathcal{M}_n}\}$$

Kullback-Leibler distance $K(s,u) = \mathbf{E}[\gamma(u) - \gamma(s)]$,
Projection $\bar{s}_m = \arg\min_{u \in \mathcal{S}_m} K(s,u)$ of s on $\mathcal{S}_m$:

Goal
Select estimator $\hat{s}_{m(s)}$ from collection $\mathcal{S} = \{(\hat{s}_m)_{m \in \mathcal{M}_n}\}$ which minimizes the Kullback-Leibler risk.
$\rightarrow$ Requires the knowledge of s.

Penalized log-likelihood framework (2)

Log-likelihood contrast $\gamma(u)$
Minimal contrast estimator $\hat{s}_m = \arg\min_{u \in \mathcal{S}_m} \gamma(u)$

$$\rightarrow \text{collection } \mathcal{S} = \{(\hat{s}_m)_{m \in \mathcal{M}_n}\}$$

Kullback-Leibler distance $K(s,u) = \mathbf{E}[\gamma(u) - \gamma(s)]$,
Projection $\bar{s}_m = \arg\min_{u \in \mathcal{S}_m} K(s,u)$ of s on $\mathcal{S}_m$:

Goal

Select estimator $\hat{s}_{m(s)}$ from collection $\mathcal{S} = \{(\hat{s}_m)_{m \in \mathcal{M}_n}\}$ which minimizes the Kullback-Leibler risk.
$\rightarrow$ Requires the knowledge of s.

Penalized log-likelihood framework (3)

Penalized likelihood estimator $\hat{s}_{\hat{m}}$: Let $\mathcal{M}_n$ be a collection of partitions of $[\![1, n]\!]$. For a given nonnegative, increasing in the size of m penalty function pen: $\mathcal{M}_n \to \mathbb{R}_+$, we choose

$$\hat{m} = \arg \min_{m \in \mathcal{M}_n} \{\gamma(\hat{s}_m) + pen(m)\},$$

Oracle inequality

$$\mathbf{E}[K(s, \hat{s}_{\hat{m}})] \leq C \, \mathbf{E}[K(s, \hat{s}_{m(s)})],$$

Main Theorem

Let $\mathcal{M}_n$ be a collection of partitions constructed on a partition m_f such that there exist absolute positive constants ρ_{min}, ρ_{max} and Γ satisfying:

- $\forall t, \rho_{min} \leq p_t \leq \rho_{max}$ and
- $\forall J \in m_f, |J| \geq \Gamma\ (\log n)^2$.

Let $\beta > 1/2\rho_{min}$ and $(L_m)_{m\in\mathcal{M}_n}$ be some family of positive weights satisfying

$$\Sigma = \sum_{m\in\mathcal{M}_n} \exp(-L_m|m|) < +\infty.$$

If, $\forall m \in \mathcal{M}_n$, $pen(m) \geq \beta|m|\left(1+4\sqrt{L_m}\right)^2$, then

$$\mathbf{E}\left[h^2(s,\hat{s}_{\hat{m}})\right] \quad \leq \quad C_\beta \inf_{m\in\mathcal{M}_n}\{K(s,\bar{s}_m)+pen(m)\} + C(\phi,\Gamma,\rho_{min},\rho_{max},\beta,\Sigma),$$

with $C_\beta = \dfrac{(2\rho_{min}\beta)^{1/3}}{(2\rho_{min}\beta)^{1/3}-1}$.

Decomposition

$$\forall m \in \mathcal{M}_n, \quad \gamma(\hat{s}_{\hat{m}}) + pen(\hat{m}) \leq \gamma(\hat{s}_m) + pen(m) \leq \gamma(\bar{s}_m) + pen(m).$$

Then, with $\bar{\gamma}(u) = \gamma(u) - \mathbf{E}[\gamma(u)]$,

$$K(s, \hat{s}_{\hat{m}}) \leq K(s, \bar{s}_m) + \bar{\gamma}(\bar{s}_m) - \bar{\gamma}(\hat{s}_{\hat{m}}) - pen(\hat{m}) + pen(m).$$

$\rightarrow$ control $\bar{\gamma}(\bar{s}_m) - \bar{\gamma}(\hat{s}_{m'})$ uniformly over $m' \in \mathcal{M}_n$.
Use decomposition [2]:

$$\bar{\gamma}(\bar{s}_m) - \bar{\gamma}(\hat{s}_{m'}) = \underbrace{(\bar{\gamma}(\bar{s}_{m'}) - \bar{\gamma}(\hat{s}_{m'}))}_{(1)} + \underbrace{(\bar{\gamma}(s) - \bar{\gamma}(\bar{s}_{m'}))}_{(2)} + \underbrace{(\bar{\gamma}(\bar{s}_m) - \bar{\gamma}(s))}_{(3)}.$$

[2] Castellan, (2000)

Decomposition

$$\forall m \in \mathcal{M}_n, \quad \gamma(\hat{s}_{\hat{m}}) + pen(\hat{m}) \leq \gamma(\hat{s}_m) + pen(m) \leq \gamma(\bar{s}_m) + pen(m).$$

Then, with $\bar{\gamma}(u) = \gamma(u) - \mathbf{E}[\gamma(u)]$,

$$K(s, \hat{s}_{\hat{m}}) \leq K(s, \bar{s}_m) + \bar{\gamma}(\bar{s}_m) - \bar{\gamma}(\hat{s}_{\hat{m}}) - pen(\hat{m}) + pen(m).$$

$\rightarrow$ control $\bar{\gamma}(\bar{s}_m) - \bar{\gamma}(\hat{s}_{m'})$ uniformly over $m' \in \mathcal{M}_n$.
Use decomposition [2]:

$$\bar{\gamma}(\bar{s}_m) - \bar{\gamma}(\hat{s}_{m'}) = \underbrace{(\bar{\gamma}(\bar{s}_{m'}) - \bar{\gamma}(\hat{s}_{m'}))}_{(1)} + \underbrace{(\bar{\gamma}(s) - \bar{\gamma}(\bar{s}_{m'}))}_{(2)} + \underbrace{(\bar{\gamma}(\bar{s}_m) - \bar{\gamma}(s))}_{(3)}.$$

[2] Castellan, (2000)

Control of the first term

$$\chi^2_m = \sum_{J \in m} Z_J, \text{ with } Z_J = \frac{[Y_J - E_J]^2}{E_J}.$$

To apply Bernstein's inequality [3] we need a space of large probability where the Z_Js are easily controled.

Lemma Let $\Omega_m(\varepsilon) = \bigcap_{J \in m} \left\{ \left| \frac{Y_J}{E_J} - 1 \right| \leq \varepsilon \right\}$. Then

$$\mathbf{P}(\Omega_m(\varepsilon)^C) \leq \frac{2}{n^a}, \text{ with } a > 2,$$

and $\mathbf{P}\left[\chi^2_m \mathbf{1}_{\Omega_m} \geq \frac{2}{\rho_{min}} \left(|m| + 8(1+\varepsilon)\sqrt{x|m|} + 4(1+\varepsilon)x \right) \right] \leq e^{-x}$

[3] Massart (2007)

Proof of lemma

$$\begin{aligned}\log \mathbf{E}\left(e^{z(Y_t - E_t)}\right) &= \frac{z^2}{2}\sum_{k\geq 0}\frac{2\kappa_{k+2}}{(k+2)!}z^k \text{ for } z \leq -\log(1-p_t) \\ &\leq E_t\frac{z^2}{2}\frac{2}{p_t}\sum_{k\geq 0}\left(\frac{z}{p_t}\right)^k\end{aligned}$$

where the κ_k are the cumulants of the negative binomial distribution.

$$\log \mathbf{E}\left[e^{z(Y_t - E_t)}\right] \leq E_t\frac{z^2}{2}\frac{2}{\rho_{min}}\frac{1}{1-\frac{z}{\rho_{min}}} \quad \text{and} \quad \log \mathbf{E}\left[e^{-z(Y_t - E_t)}\right] \leq E_t\frac{z^2}{2}\frac{2}{\rho_{min}}$$

Using a large deviation result [4] we obtain

$$P\left[|Y_J - E_J| \geq x\right] \leq 2e^{-\frac{\rho_{min}x^2}{4(E_J + x)}}$$

[4] Baraud and Birgé (2009)

Proof of lemma

$$\log \mathbf{E}\left(e^{z(Y_t - E_t)}\right) = \frac{z^2}{2}\sum_{k\geq 0}\frac{2\kappa_{k+2}}{(k+2)!}z^k \text{ for } z \leq -\log(1-p_t)$$
$$\leq E_t \frac{z^2}{2}\frac{2}{p_t}\sum_{k\geq 0}\left(\frac{z}{p_t}\right)^k$$

where the κ_k are the cumulants of the negative binomial distribution.

$$\log \mathbf{E}\left[e^{z(Y_t - E_t)}\right] \leq E_t \frac{z^2}{2}\frac{2}{\rho_{min}}\frac{1}{1-\frac{z}{\rho_{min}}} \quad \text{and} \quad \log \mathbf{E}\left[e^{-z(Y_t - E_t)}\right] \leq E_t \frac{z^2}{2}\frac{2}{\rho_{min}}$$

Using a large deviation result [4] we obtain

$$P\left[|Y_J - E_J| \geq x\right] \leq 2e^{-\frac{\rho_{min}x^2}{4(E_J+x)}}$$

[4] Baraud and Birgé (2009)

Risk of a model

Proposition

Under same assumptions:

$$K(s, \bar{s}_m) - \frac{C_1(\phi, \Gamma, \rho_{min}, \rho_{max}, \varepsilon, a)}{n^{a/2-\alpha}} + C_2(\varepsilon)|m| \leq \mathbf{E}[K(s, \hat{s}_m)],$$

where $\alpha < 1$ and $C_2(\varepsilon) = \rho_{min}^2 \frac{(1-\varepsilon)^2}{(1+\varepsilon)^4}$.

Oracle inequality

$\mathbf{E}\left[h^2(s, \hat{s}_{\hat{m}})\right] \leq C \log(n) \inf_{m \in \mathcal{M}_n} \left\{\mathbf{E}[K(s, \hat{s}_m)]\right\} + C(\phi, \Gamma, \rho_{min}, \rho_{max}, \beta, \Sigma).$

Risk of a model

Proposition

Under same assumptions:

$$K(s, \bar{s}_m) - \frac{C_1(\phi, \Gamma, \rho_{min}, \rho_{max}, \varepsilon, a)}{n^{a/2-\alpha}} + C_2(\varepsilon)|m| \leq \mathbf{E}[K(s, \hat{s}_m)],$$

where $\alpha < 1$ and $C_2(\varepsilon) = \rho_{min}^2 \frac{(1-\varepsilon)^2}{(1+\varepsilon)^4}$.

Oracle inequality

$\mathbf{E}\left[h^2(s, \hat{s}_{\hat{m}})\right] \leq C \log(n) \inf_{m \in \mathcal{M}_n} \{\mathbf{E}[K(s, \hat{s}_m)]\} + C(\phi, \Gamma, \rho_{min}, \rho_{max}, \beta, \Sigma).$

Outline

1 Biological framework

2 Main Result
- Framework
- Scheme of the proof
- Corrolary

3 Illustration
- Segmentation procedure
- Simulation study

Segmentation procedure

$$\hat{K} = \arg\min_k \left\{ \gamma(\hat{s}_k) + \beta k \left(1 + 4\sqrt{1.1 + \log \frac{n}{k}} \right)^2 \right\}$$

[5] Arlot and Massart (2009)

Segmentation procedure

$$\hat{K} = \arg\min_{k} \left\{ \gamma(\hat{s}_k) + \beta k \left(1 + 4\sqrt{1.1 + \log \frac{n}{k}} \right)^2 \right\}$$

1. choose K_{max}: $k \in \{1, \ldots, K_{max}\}$.

[5] Arlot and Massart (2009)

Segmentation procedure

$$\hat{K} = \arg\min_k \left\{ \gamma(\hat{s}_k) + \beta k \left(1 + 4\sqrt{1.1 + \log \frac{n}{k}} \right)^2 \right\}$$

1. choose K_{max}: $k \in \{1, \ldots, K_{max}\}$.
2. define contrast $\gamma(s) = -\sum_{t=1}^{n} \phi \log \theta_t + Y_t \log(1 - \theta_t)$.

[5] Arlot and Massart (2009)

Segmentation procedure

$$\hat{K} = \arg\min_{k} \left\{ \gamma(\hat{s}_k) + \beta k \left(1 + 4\sqrt{1.1 + \log \frac{n}{k}} \right)^2 \right\}$$

1. choose K_{max}: $k \in \{1, \dots, K_{max}\}$.
2. define contrast $\gamma(s) = -\sum_{t=1}^{n} \phi \log \theta_t + Y_t \log(1 - \theta_t)$.
3. for $1 \leq k \leq K_{max}$, compute $\hat{s}_k = \arg\min_{m \in \mathcal{M}_k} \{\gamma(s_m)\}$.
   This can be done using the pruned DPA.

[5] Arlot and Massart (2009)

Segmentation procedure

$$\hat{K} = \arg\min_k \left\{ \gamma(\hat{s}_k) + \beta k \left(1 + 4\sqrt{1.1 + \log\frac{n}{k}} \right)^2 \right\}$$

1. choose K_{max}: $k \in \{1, \dots, K_{max}\}$.
2. define contrast $\gamma(s) = -\sum_{t=1}^{n} \phi \log\theta_t + Y_t \log(1-\theta_t)$.
3. for $1 \leq k \leq K_{max}$, compute $\hat{s}_k = \arg\min_{m \in \mathcal{M}_k} \{\gamma(s_m)\}$.
   This can be done using the pruned DPA.
4. tune β using the slope heuristic [5]

[5] Arlot and Massart (2009)

Segmentation procedure

$$\hat{K} = \arg\min_k \left\{ \gamma(\hat{s}_k) + \beta k \left(1 + 4\sqrt{1.1 + \log\frac{n}{k}}\right)^2 \right\}$$

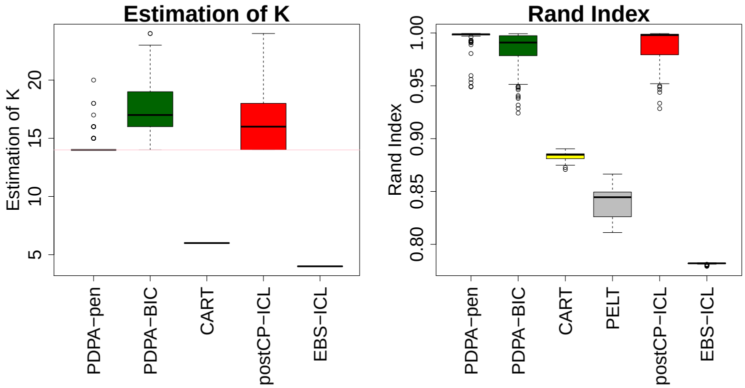
1. choose K_{max}: $k \in \{1, \ldots, K_{max}\}$.
2. define contrast $\gamma(s) = -\sum_{t=1}^{n} \phi \log\theta_t + Y_t \log(1-\theta_t)$.
3. for $1 \leq k \leq K_{max}$, compute $\hat{s}_k = \arg\min_{m\in\mathcal{M}_k}\{\gamma(s_m)\}$.
   This can be done using the pruned DPA.
4. tune β using the slope heuristic [5]

$\rightarrow$ implemented in R package `Segmentor3IsBack`

[5] Arlot and Massart (2009)

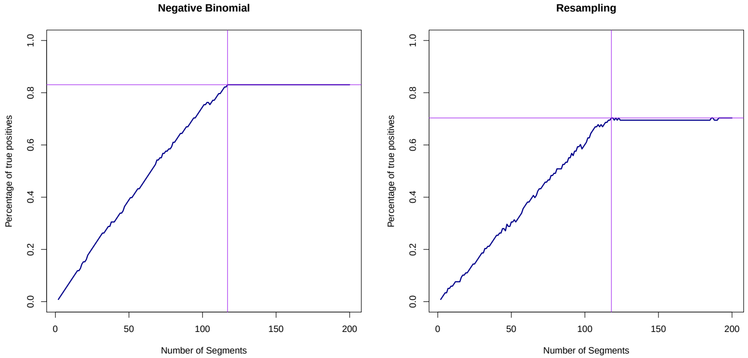
Short-signal analysis

Toy example, resampling from RNA-Seq data
$n = 5000$; $K = 14$



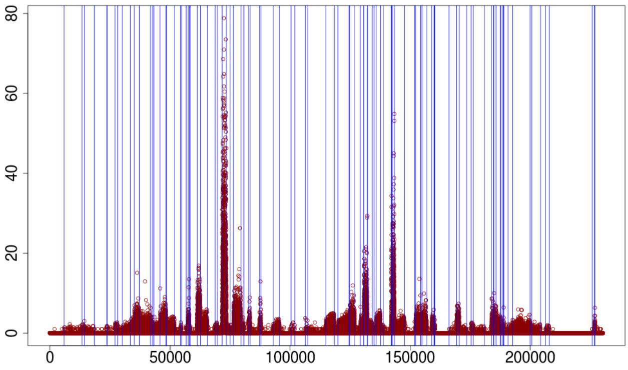
Long-signal analysis

Two simulation studies: $n = 230000$; $K = 118$
S1= simulation from negative binomial with 4 levels
S2= simulation by resampling real RNA-seq data



Segmentation of a yeast chromosome

$$K_{max} = 1000, \quad K_{annot} = 137, \quad \hat{K} = 201$$



Conclusion

- Theoretic framework for the choice of the number of segments
- Implemented in an operational package for long datasets
- Extension to distributions from the exponential family

Segmentation of the Poisson and Negative Binomial Rate Models: a Penalized Estimator

Alice Cleynen and Emilie Lebarbier

To appear in Esaim : Proba & Stats

Thank you for your attention!