A penalized maximum likelihood estimator for the segmentation of RNA-Seq data

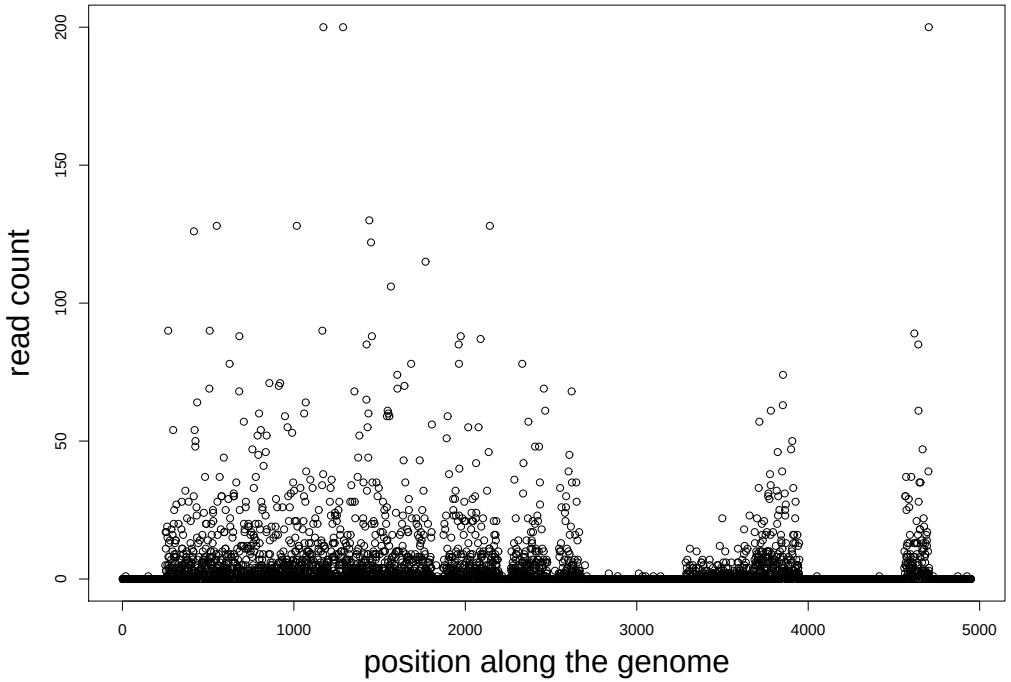
A. Cleynen, E. Lebarbier

UMR 518 AgroParisTech / INRA / UC Berkeley

Toulouse, May 27, 2013

Outline

Segmentation Model



- m a partition of $\{1, n\}$,
- J a segment of m,
- K the number of segments,

$$\forall t \in J, \ Y_t \sim \mathcal{NB}(p_J, \phi)$$

Penalized log-likelihood framework

$$s(t) = \mathcal{NB}(p_t, \phi) \quad \text{the true model}$$

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

Penalized log-likelihood framework

$$s(t) = \mathcal{NB}(p_t, \phi) \quad \text{the true model}$$

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

Minimal contrast estimator on partition m: $\hat{s}_m = \arg\min_{u \in \mathcal{S}_m} \gamma(u)$
Log-likelihood contrast $\gamma(u) = \sum_{t=1}^{n} -\phi \log p_t - Y_t \log(1 - p_t)$,
$\rightarrow$ defines collection $\mathcal{S} = \{(\hat{s}_m)_{m \in \mathcal{M}_n}\}$

Penalized log-likelihood framework

$$s(t) = \mathcal{NB}(p_t, \phi) \quad \text{the true model}$$

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

Minimal contrast estimator on partition m: $\hat{s}_m = \arg\min_{u \in \mathcal{S}_m} \gamma(u)$
Log-likelihood contrast $\gamma(u) = \sum_{t=1}^{n} -\phi \log p_t - Y_t \log(1 - p_t)$,
$\rightarrow$ defines collection $\mathcal{S} = \{(\hat{s}_m)_{m \in \mathcal{M}_n}\}$

$\rightarrow$ Choose best estimator $\hat{s}_{m(s)} = \arg\min_{u \in \mathcal{S}} \mathbf{E}[K(s, u)]$
$\rightarrow$ Requires the knowledge of s.

Penalized log-likelihood framework (2)

Penalized likelihood estimator $\hat{s}_{\hat{m}}$ such that:

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

Penalized log-likelihood framework (2)

Penalized likelihood estimator $\hat{s}_{\hat{m}}$ such that:

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

Oracle inequality

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

Our proposition

Assumptions:

- $\forall t, 0 < \rho_{min} \leq p_t \leq \rho_{max} < 1$ and
- $\forall J \in m_f, |J| \geq \Gamma(\log(n))^2$.

Let $\beta > 1/4$ and

$$pen(m) = \beta |m| \left(1 + 4\sqrt{1.1 + \log\left(\frac{n}{|m|}\right)}\right)^2, \quad \text{then}$$

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

Outline

Decomposition

As in [3], we write:

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

with $\bar{\gamma}(u) = \gamma(u) - \mathbf{E}[\gamma(u)]$ and $\bar{s}_m = \arg\min_{\mathcal{S}_m} K(s, u)$.

Decomposition

As in [3], we write:

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

with $\bar{\gamma}(u) = \gamma(u) - \mathbf{E}[\gamma(u)]$ and $\bar{s}_m = \arg\min_{\mathcal{S}_m} K(s, u)$.

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

Decomposition

As in [3], we write:

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

with $\bar{\gamma}(u) = \gamma(u) - \mathbf{E}[\gamma(u)]$ and $\bar{s}_m = \arg\min_{\mathcal{S}_m} K(s, u)$.

$\rightarrow$ control $\bar{\gamma}(\bar{s}_m) - \bar{\gamma}(\hat{s}_{m'})$ uniformly over $m' \in \mathcal{M}_n$.
Classic use of concentration inequalities on the supremum.

Decomposition

As in [3], we write:

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

with $\bar{\gamma}(u) = \gamma(u) - \mathbf{E}[\gamma(u)]$ and $\bar{s}_m = \arg\min_{\mathcal{S}_m} K(s, u)$.

$\rightarrow$ control $\bar{\gamma}(\bar{s}_m) - \bar{\gamma}(\hat{s}_{m'})$ uniformly over $m' \in \mathcal{M}_n$.
Classic use of concentration inequalities on the supremum.
Here we use instead the decomposition [4]:

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

Decomposition

As in [3], we write:

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

with $\bar{\gamma}(u) = \gamma(u) - \mathbf{E}[\gamma(u)]$ and $\bar{s}_m = \arg\min_{\mathcal{S}_m} K(s, u)$.

$\rightarrow$ control $\bar{\gamma}(\bar{s}_m) - \bar{\gamma}(\hat{s}_{m'})$ uniformly over $m' \in \mathcal{M}_n$.
Classic use of concentration inequalities on the supremum.
Here we use instead the decomposition [4]:

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

The χ^2 statistic

$(1) = (\bar{\gamma}(\bar{s}_{m'}) - \bar{\gamma}(\hat{s}_{m'})) \rightarrow$ introduce chi-square statistic :

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

$$Y_J = \sum_{t \in J} Y_t, \qquad E_J = \mathbf{E}[Y_J]$$

The χ^2 statistic

$(1) = (\bar{\gamma}(\bar{s}_{m'}) - \bar{\gamma}(\hat{s}_{m'})) \rightarrow$ introduce chi-square statistic :

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

$$Y_J = \sum_{t \in J} Y_t, \qquad E_J = \mathbf{E}[Y_J]$$

To apply Bernstein's inequality [6] we need:

- A space where Z_J is easily controled
- The control of Y_J around its expectation.

Applying Bernstein's inequality

We define

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

Applying Bernstein's inequality

We define

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

With a large deviation result established by Baraud and Birgé [2] we get

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

Applying Bernstein's inequality

We define

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

With a large deviation result established by Baraud and Birgé [2] we get

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

so that $\mathbf{P}(\Omega_m(\varepsilon)^C) \leq \dfrac{C(\phi, \Gamma, \rho_{min}, \varepsilon, a)}{n^a}$, with $a > 2$,

Applying Bernstein's inequality

We define

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

With a large deviation result established by Baraud and Birgé [2] we get

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

so that $\mathbf{P}(\Omega_m(\varepsilon)^C) \leq \dfrac{C(\phi, \Gamma, \rho_{min}, \varepsilon, a)}{n^a}$, with $a > 2$,

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

Segmentation procedure

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

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

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 $\gamma(s) = \sum_{t=1}^{n} -\phi \log p_t - Y_t \log(1 - p_t)$.

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 $\gamma(s) = \sum_{t=1}^{n} -\phi \log p_t - Y_t \log(1 - p_t)$.
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 Dynamic Programming algorithm.

Segmentation procedure

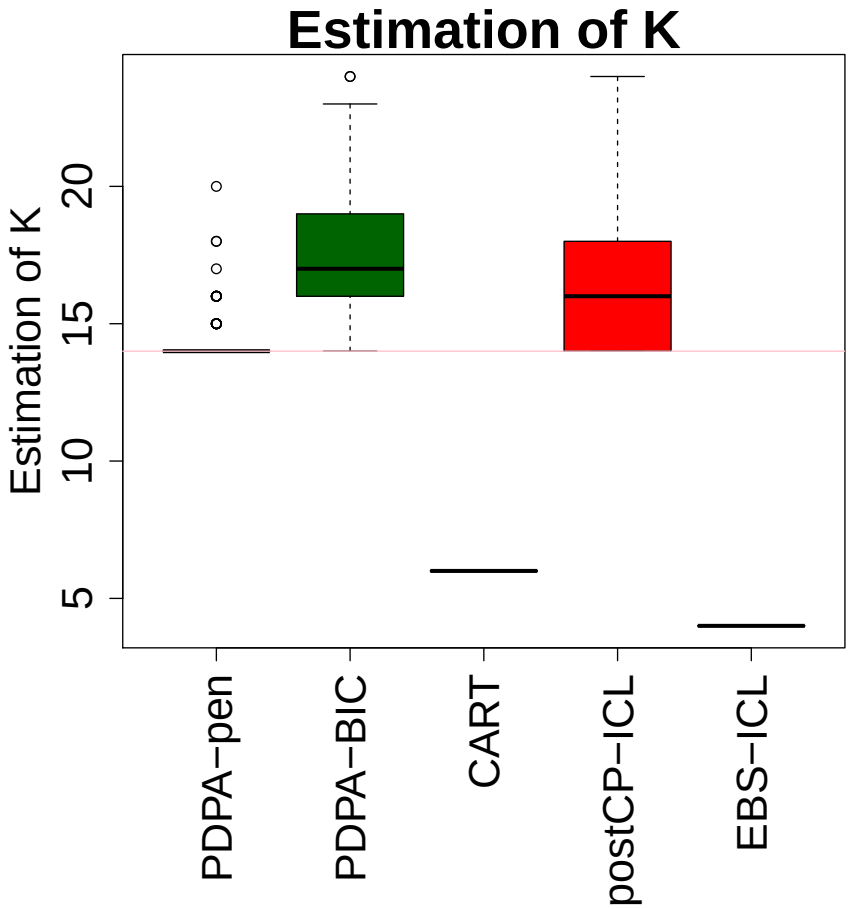
$$\hat{K} = \arg\min_k \left\{ \gamma(\hat{s}_k) + \beta k \left(1 + 4\sqrt{1.1 + \log \frac{n}{k}} \right)^2 \right\}$$

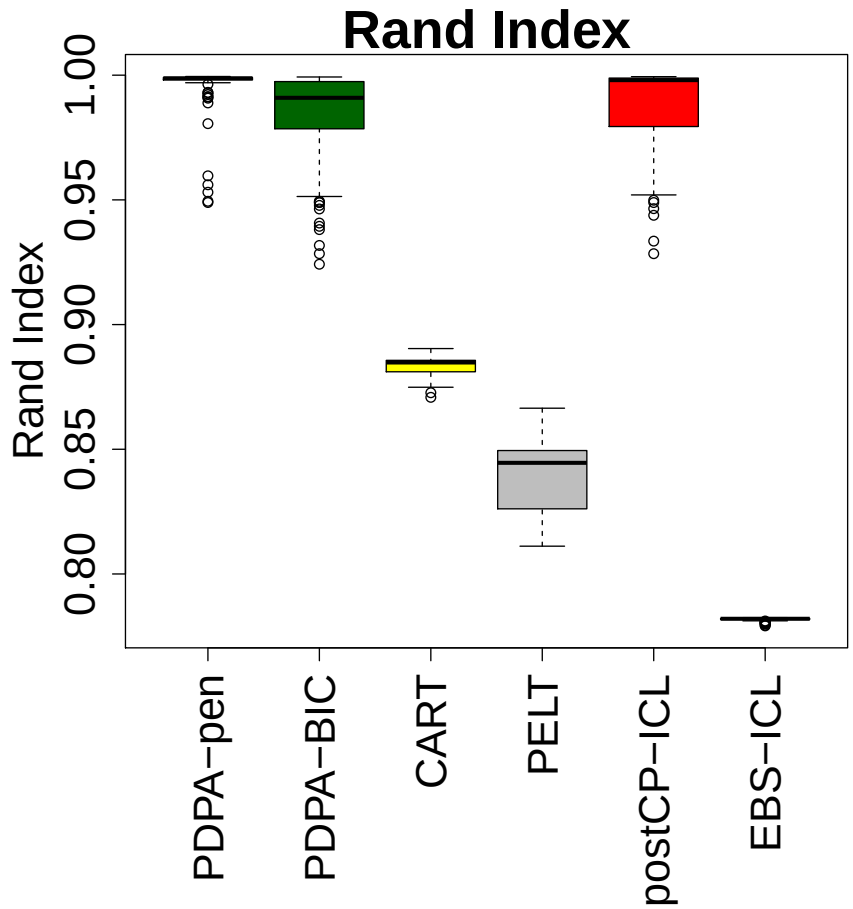
1. choose K_{max}: $k \in \{1, \ldots, K_{max}\}$.
2. define contrast $\gamma(s) = \sum_{t=1}^{n} -\phi \log p_t - Y_t \log(1 - p_t)$.
3. for $1 \leq k \leq K_{max}$, compute $\hat{s}_k = \arg\min_{m \in \mathcal{M}_k} \{\gamma(s_m)\}$. This can be done using the pruned Dynamic Programming algorithm.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

Outline

Short-signal analysis

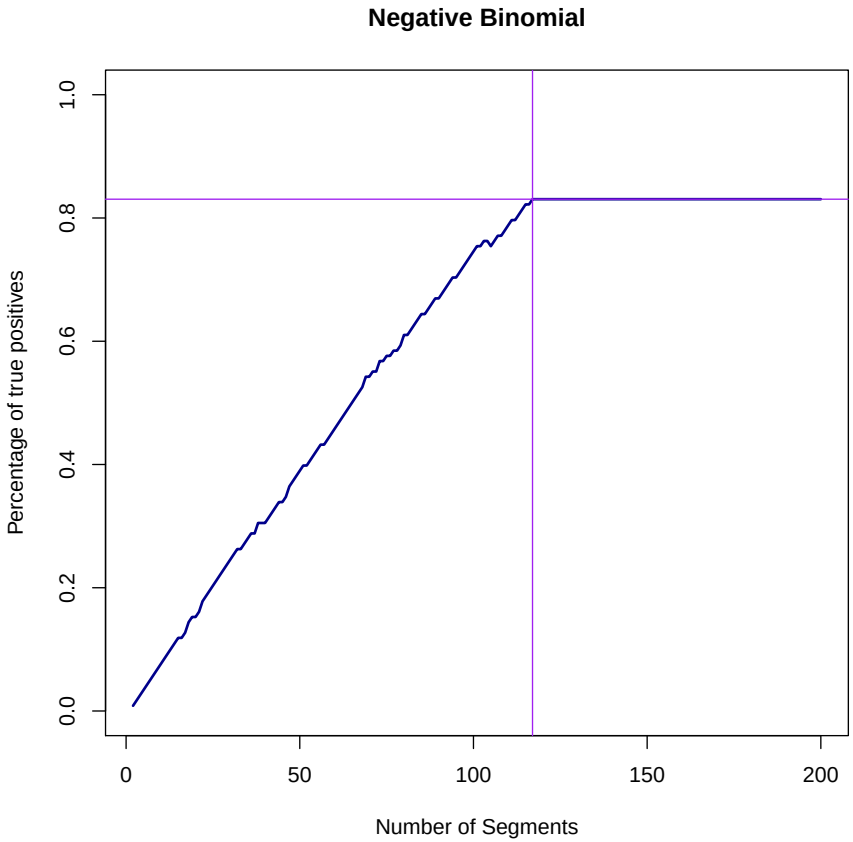
Resampling from a Drosophila gene with two isoforms
$n = 5000$; $K = 14$

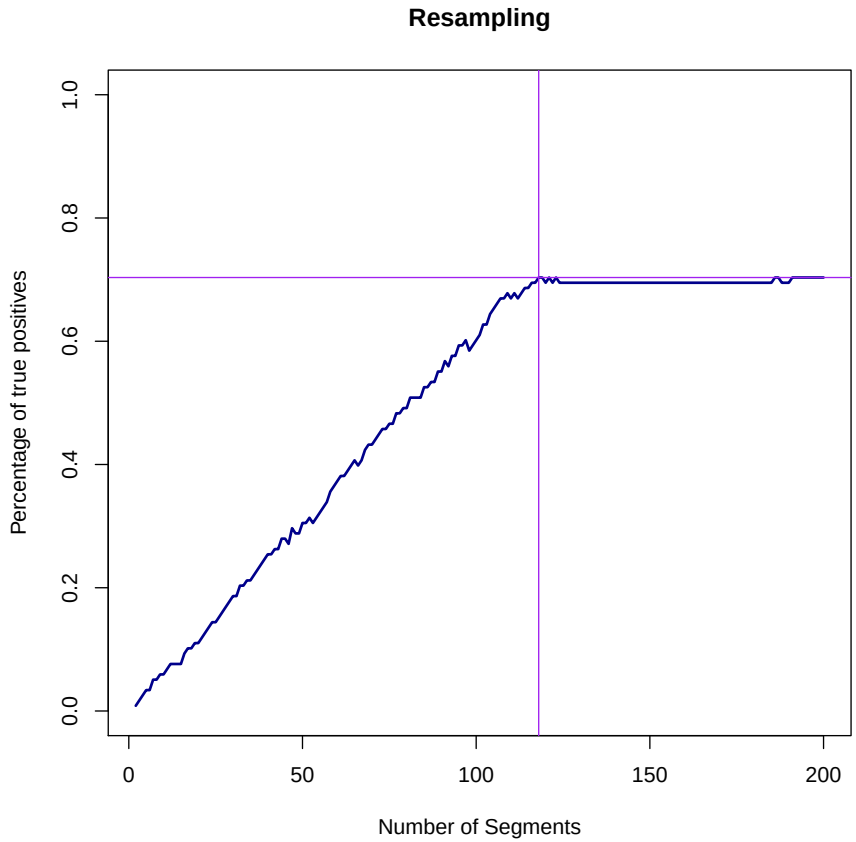




Long-signal analysis

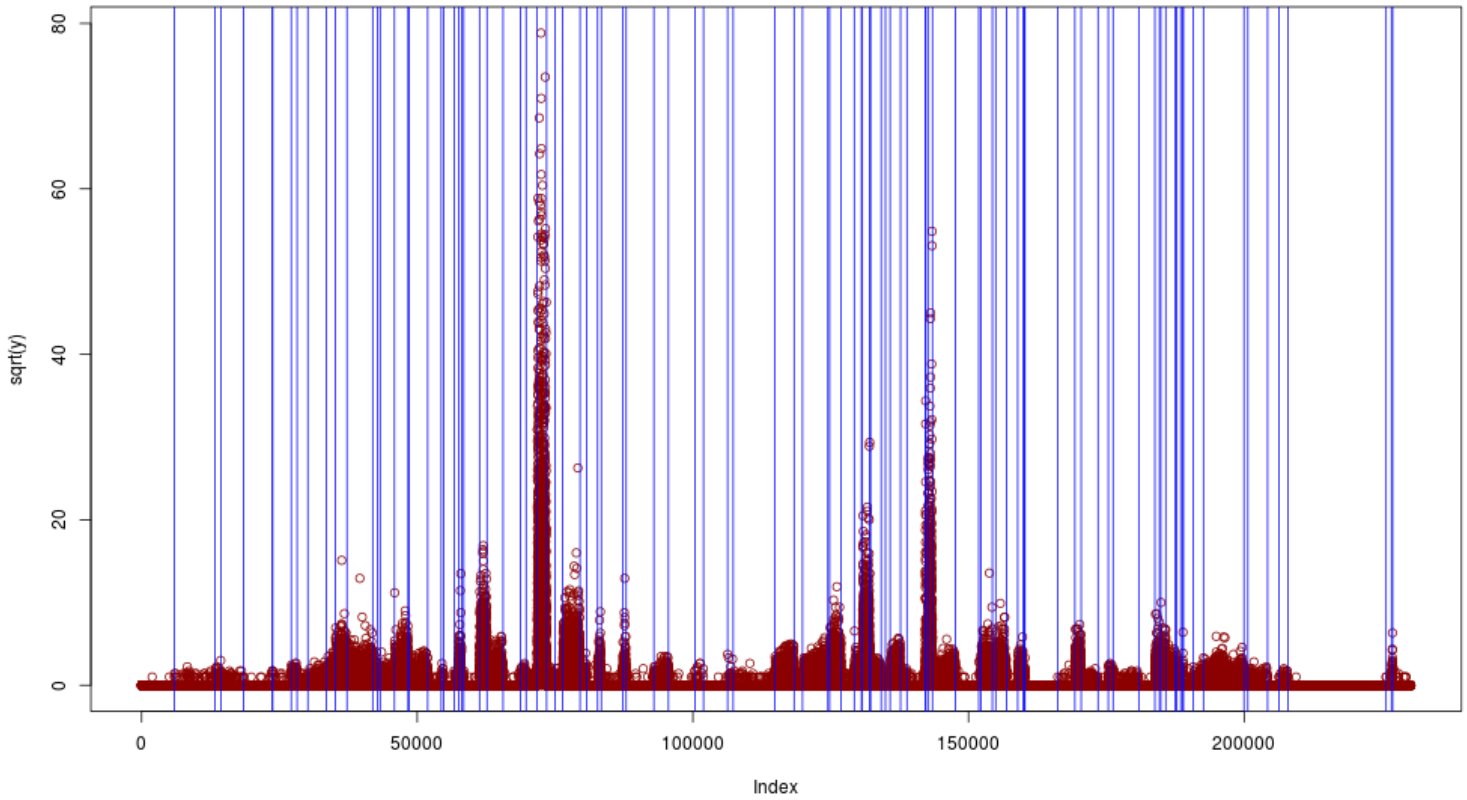
Two simulation studies: $n = 230000$; $K = 118$
S1= simulation from negative binomial with 4 levels
S2= simulation by resampling real RNA-seq data pooled in 4 categories





Application to real data

Real RNA-Seq data: chromosome 1 from yeast genome.
$\hat{K} = 103$



Conclusion

- Proposition of an estimator for the overdispersion ϕ:
  $\rightarrow$ inspired from [5], based on moment estimator
- Assuming ϕ is known:
  - Effective procedure for the segmentation of RNA-Seq data,
    $\rightarrow$ complexity in $\mathcal{O}(n \log(n))$
  - Theoretical guarantees,
    $\rightarrow$ oracle inequality
  - Excellent results in practice
- Complete procedure available in R package `Segmentor3IsBack` on the CRAN.

Thank you!

References I

S. Arlot and P. Massart.

Data-driven calibration of penalties for least-squares regression.
J. Mach. Learn. Res., 10:245–279 (electronic), 2009.

Y. Baraud and L. Birgé.

Estimating the intensity of a random measure by histogram type estimators.
Probab. Theory Related Fields, 143(1-2):239–284, 2009.

L. Birgé and P. Massart.

Gaussian model selection.
J. Eur. Math. Soc. (JEMS), 3(3):203–268, 2001.

G. Castellan.

Modified Akaike's criterion for histogram density estimation.
C. R. Acad. Sci., Paris, Sér. I, Math. 330, 8:729–732, 2000.

N. Johnson, A.W. Kemp, and S. Kotz.

Univariate discrete distributions.
John Wiley & Sons, Inc., 2005.

P. Massart.

Concentration inequalities and model selection, volume 1896 of *Lecture Notes in Mathematics*.
Springer, Berlin, 2007.
Lectures from the 33rd Summer School on Probability Theory held in Saint-Flour, July 6–23, 2003, With a foreword by Jean Picard.