Comparing the change-point location of independent profiles, and application to alternative splicing

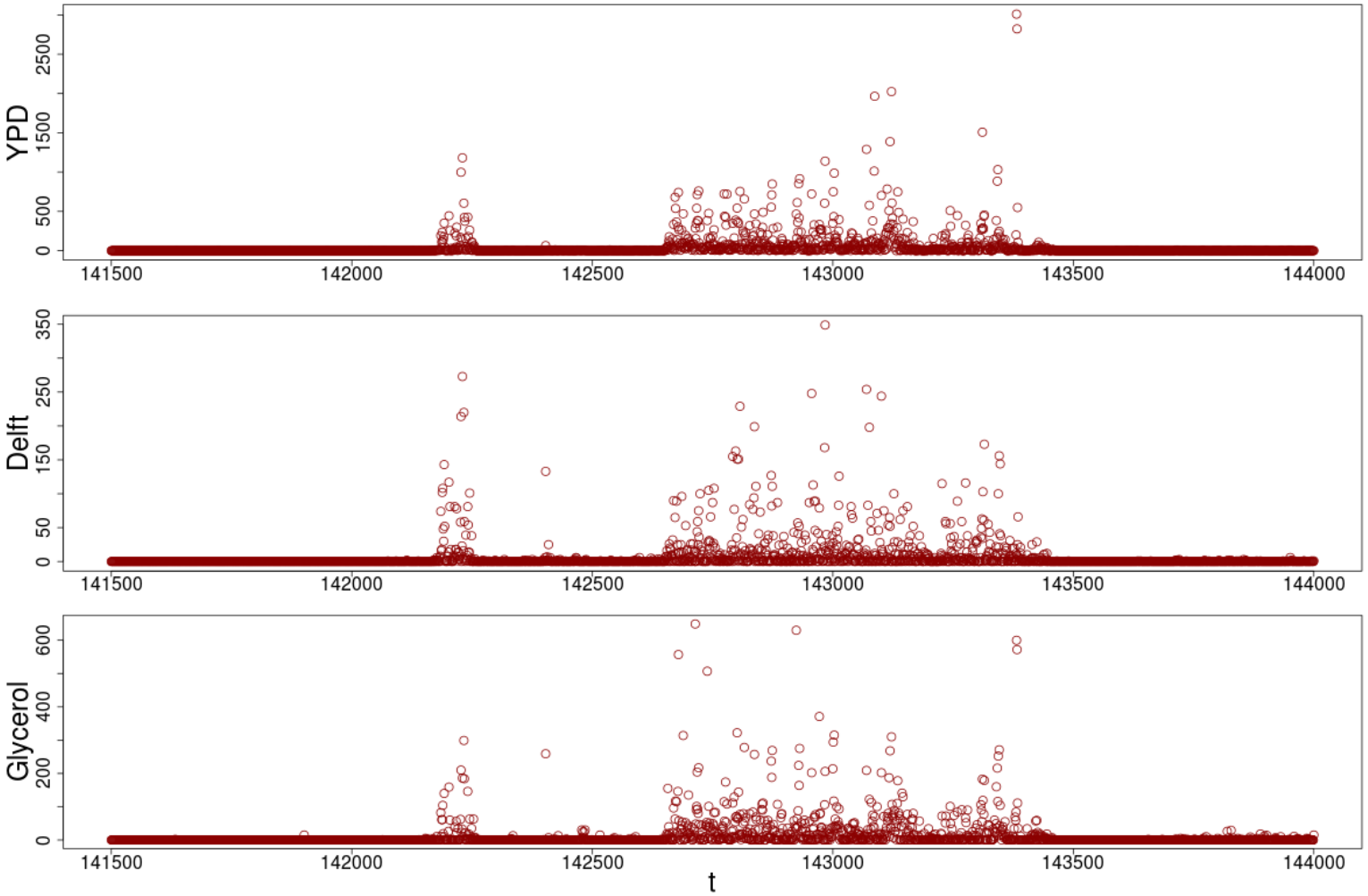
A. Cleynen, S. Robin

UMR 518 AgroParisTech / INRA / UC Berkeley

July 4th, 2013

Outline

Motivation


YPD
0
500
1500
2500
Delft
0
50
150
250
350
Glycerol
0
200
400
600
141500
142000
142500
143000
143500
144000
t

Intuition

- Segment independently each profile in K segments $\rightarrow$ m
- Change-points τ_k correspond to transcript boundaries
- Compare change-point localization between profiles

Intuition

- Segment independently each profile in K segments $\rightarrow$ m
- Change-points τ_k correspond to transcript boundaries
- Compare change-point localization between profiles

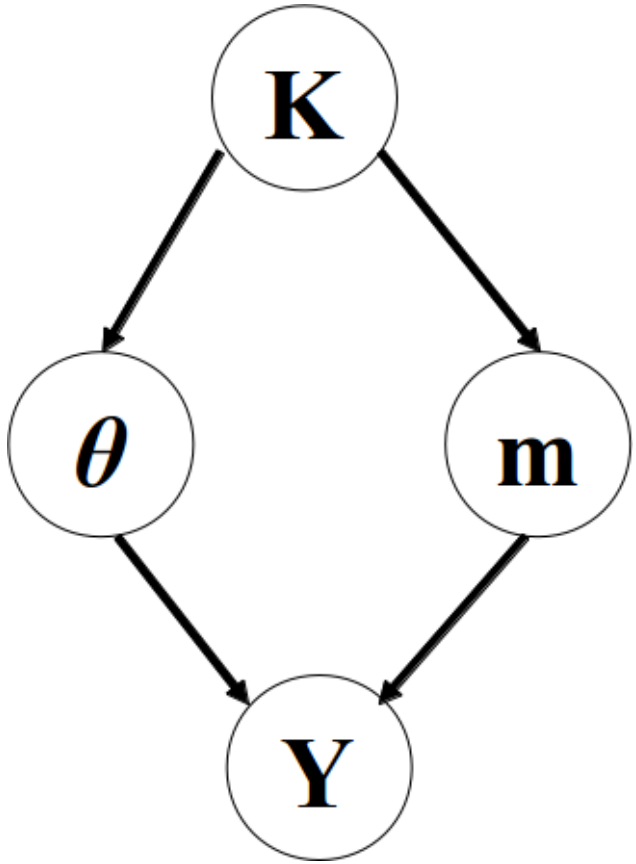
$\rightarrow$ requires ability to measure the uncertainty on the change-point location

Model from Rigaill *et al.* (2012)

- $K \sim P(K)$;
- $m|K \sim P(m|K)$;
- $\theta_J|K$ iid, $\theta_J|K \sim P(\theta_J|K)$;
- $Y = (Y_t)$ independent conditional on m and (θ_J) ,

$$(Y_t|m, \theta_J \in m, t \in J) \sim \mathcal{G}(\theta_J, \phi)$$

ϕ global, known parameter.



Note that typically, $P(m|K) = \mathcal{U}(\mathcal{M}_K)$

Computations

Key point: $\forall 1 \leq i < j \leq n+1$,

$$\begin{aligned}
[A]_{i,j} &= P(Y_{[\![i,j]\!]} | [\![i,j]\!]) \\
&= \int \mathbb{P}(Y_{[\![i,j]\!]} | \theta_{[\![i,j]\!]}, \phi) \mathbb{P}(\theta_{[\![i,j]\!]}) d\theta_{[\![i,j]\!]}
\end{aligned}$$

Computations

Key point: $\forall 1 \leq i < j \leq n+1$,

$$\begin{aligned}[A]_{i,j} &= P(Y_{[\![i,j]\!]}|[\![i,j]\!]) \\ &= \int \mathbb{P}(Y_{[\![i,j]\!]}|\theta_{[\![i,j]\!]},\phi)\mathbb{P}(\theta_{[\![i,j]\!]})d\theta_{[\![i,j]\!]}\end{aligned}$$

- $\mathbb{P}(Y|K) = C\ [A^K]_{1,n+1}$;
- $p_k(t;Y;K) = P(\tau_k = t|Y,K)$

$$p_k(t;Y;K) = \frac{\left[(A)^k\right]_{1,t}\left[(A)^{K-k}\right]_{t,n+1}}{\left[(A)^K\right]_{1,n+1}}.$$

Computations

Key point: $\forall 1 \leq i < j \leq n+1$,

$$\begin{aligned}[A]_{i,j} &= P(Y_{[\![i,j[\![}|[\![i,j[\![) \\ &= \int \mathbb{P}(Y_{[\![i,j[\![}|\theta_{[\![i,j[\![}, \phi)\mathbb{P}(\theta_{[\![i,j[\![})d\theta_{[\![i,j[\![}\end{aligned}$$

- $\mathbb{P}(Y|K) = C\ [A^K]_{1,n+1}$;
- $p_k(t; Y; K) = P(\tau_k = t|Y, K)$

$$p_k(t; Y; K) = \frac{\left[(A)^k\right]_{1,t}\left[(A)^{K-k}\right]_{t,n+1}}{\left[(A)^K\right]_{1,n+1}}.$$

Provided we choose distribution $\mathcal{G}$ with conjugate prior, computations are done exactly and in quadratic time

Outline

Model for two profiles

- Series Y^1 and Y^2 with same length n
- Respective (known) number of segments K^1 and K^2.
- Change-points to compare: $\tau^1_{k_1}$ and $\tau^2_{k_2}$

Model for two profiles

- Series Y^1 and Y^2 with same length n
- Respective (known) number of segments K^1 and K^2.
- Change-points to compare: $\tau^1_{k_1}$ and $\tau^2_{k_2}$

- $\Delta = \tau^1_{k_1} - \tau^2_{k_2}$
- $\delta_{k_1,k_2}(d; K^1, K^2) = P(\Delta = d | Y^1, Y^2, K^1, K^2)$

Model for two profiles

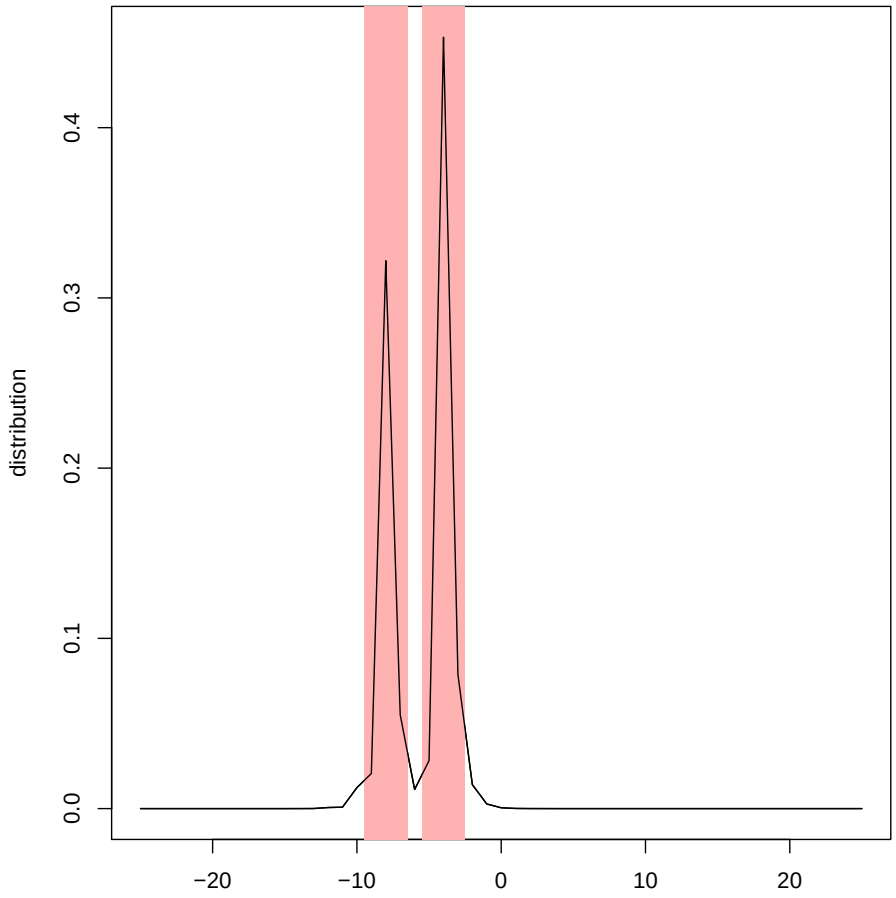
- Series Y^1 and Y^2 with same length n
- Respective (known) number of segments K^1 and K^2.
- Change-points to compare: $\tau^1_{k_1}$ and $\tau^2_{k_2}$

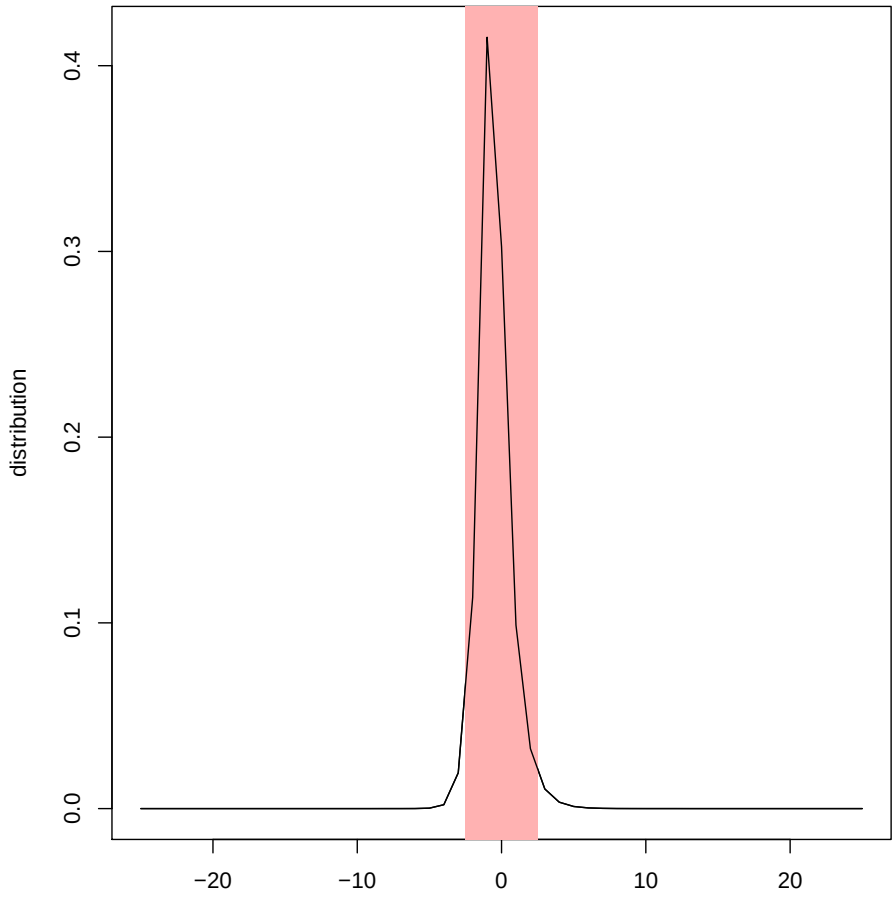
- $\Delta = \tau^1_{k_1} - \tau^2_{k_2}$
- $\delta_{k_1,k_2}(d; K^1, K^2) = P(\Delta = d | Y^1, Y^2, K^1, K^2)$

$$\delta_{k_1,k_2}(d; K^1, K^2) = \sum_t p_{k_1}(t; Y^1; K^1) p_{k_2}(t-d; Y^2; K^2).$$

Position of 0 with respect to the posterior distribution δ?

Examples


distribution
0.0
0.1
0.2
0.3
0.4
−20
−10
0
10
20


distribution
0.0
0.1
0.2
0.3
0.4
−20
−10
0
10
20

Outline

Framework of I profiles

- Series Y^ℓ for $1 \leq \ell \leq I$ $\rightarrow \mathbf{Y}$
- Partitions m^ℓ in K^ℓ segments $\rightarrow \mathbf{m}$ and $\mathbf{K}$
- Parameters θ^ℓ $\rightarrow \boldsymbol{\theta}$
- Change-points to compare: $\tau^\ell_{k_\ell}$

Framework of I profiles

- Series Y^ℓ for $1 \leq \ell \leq I$ $\rightarrow \mathbf{Y}$
- Partitions m^ℓ in K^ℓ segments $\rightarrow \mathbf{m}$ and $\mathbf{K}$
- Parameters θ^ℓ $\rightarrow \boldsymbol{\theta}$
- Change-points to compare: $\tau_{k_\ell}^\ell$

In the perspective of change-point comparison, we introduce the event:

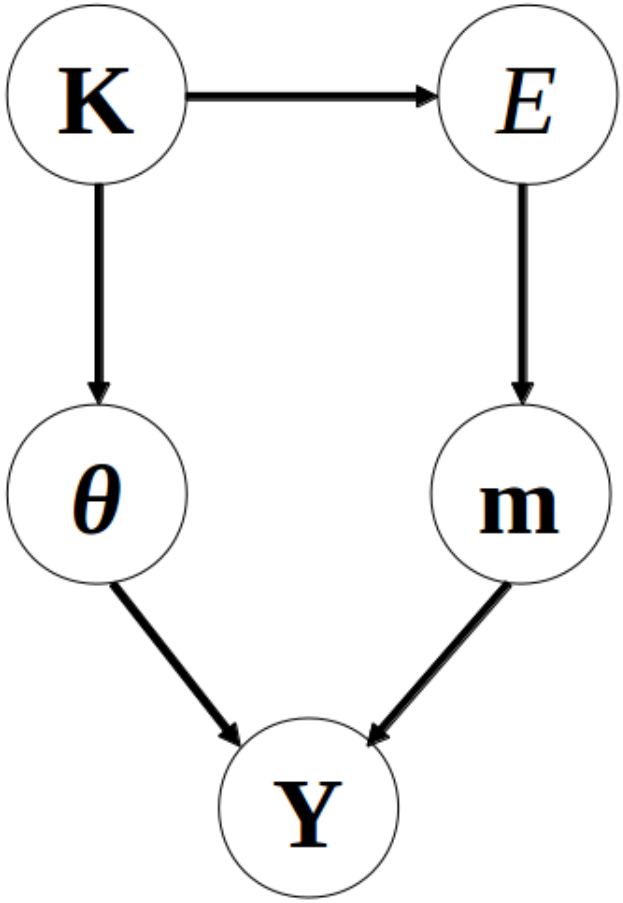
$$E_0 = \{\tau_{k_1}^1 = \cdots = \tau_{k_I}^I\}.$$

We further denote E_1 its complementary and define the random variable:

$$E = \mathbb{I}\{E_1\} = 1 - \mathbb{I}\{E_0\}.$$

Model for I profiles

- $E|\mathbf{K} \sim \mathcal{B}(1 - p_0(\mathbf{K}))$ where $p_0(\mathbf{K}) = P(E_0|\mathbf{K})$;
- $\mathbf{m}|\mathbf{K}, E \sim P(\mathbf{m}|\mathbf{K}, E)$;
- $\boldsymbol{\theta}|\mathbf{K} \sim P(\boldsymbol{\theta}|\mathbf{K})$;
- $\mathbf{Y}|\mathbf{m}, \boldsymbol{\theta} \sim P(\mathbf{Y}|\mathbf{m}, \boldsymbol{\theta})$



In this context, we choose

$$\begin{aligned} P(\mathbf{m}|\mathbf{K}, E_0) &= \mathcal{U}(\mathcal{M}_{n,\mathbf{K}} \cap E_0), \\ P(\mathbf{m}|\mathbf{K}, E_1) &= \mathcal{U}(\mathcal{M}_{n,\mathbf{K}} \cap E_1), \end{aligned}$$

where $\mathcal{M}_{n,\mathbf{K}} = \bigotimes_\ell \mathcal{M}_{n,K^\ell}$.

Posterior probability of E_0

$$P(E_0|\mathbf{Y},\mathbf{K}) = \frac{p_0(\mathbf{K})}{q_0(\mathbf{K})} Q(\mathbf{Y},E_0|\mathbf{K}) \Big/ \left[\frac{1-p_0(\mathbf{K})}{1-q_0(\mathbf{K})} Q(\mathbf{Y}|\mathbf{K}) + \frac{p_0(\mathbf{K})-q_0(\mathbf{K})}{q_0(\mathbf{K})[1-q_0(\mathbf{K})]} Q(\mathbf{Y},E_0|\mathbf{K})\right]$$

where

$$\begin{aligned} q_0(\mathbf{K}) &= \sum_t \prod_\ell \binom{t-2}{k_\ell-1}\binom{n-t}{K_\ell-k_\ell-1} \Big/ \binom{n-1}{K_\ell-1}, \\ Q(\mathbf{Y}|\mathbf{K}) &= \prod_\ell \left[(A_\ell)^{K_\ell}\right]_{1,n+1}, \\ \text{and } Q(\mathbf{Y},E_0|\mathbf{K}) &= \sum_t \prod_\ell \left[(A_\ell)^{k_\ell}\right]_{1,t} \left[(A_\ell)^{K_\ell-k_\ell}\right]_{t+1,n+1}. \end{aligned}$$

and A_ℓ stands for the matrix A defined before, corresponding to series ℓ.

$\rightarrow$ exact computation, and in quadratic time

Outline

Modeling RNA-Seq data

Using the negative binomial distribution:

$$\begin{cases} \theta_J & |K & \sim \mathcal{B}eta(a,b) \\ Y_t & |m,\theta_J & \sim \mathcal{NB}(\theta_J,\phi) \end{cases}$$

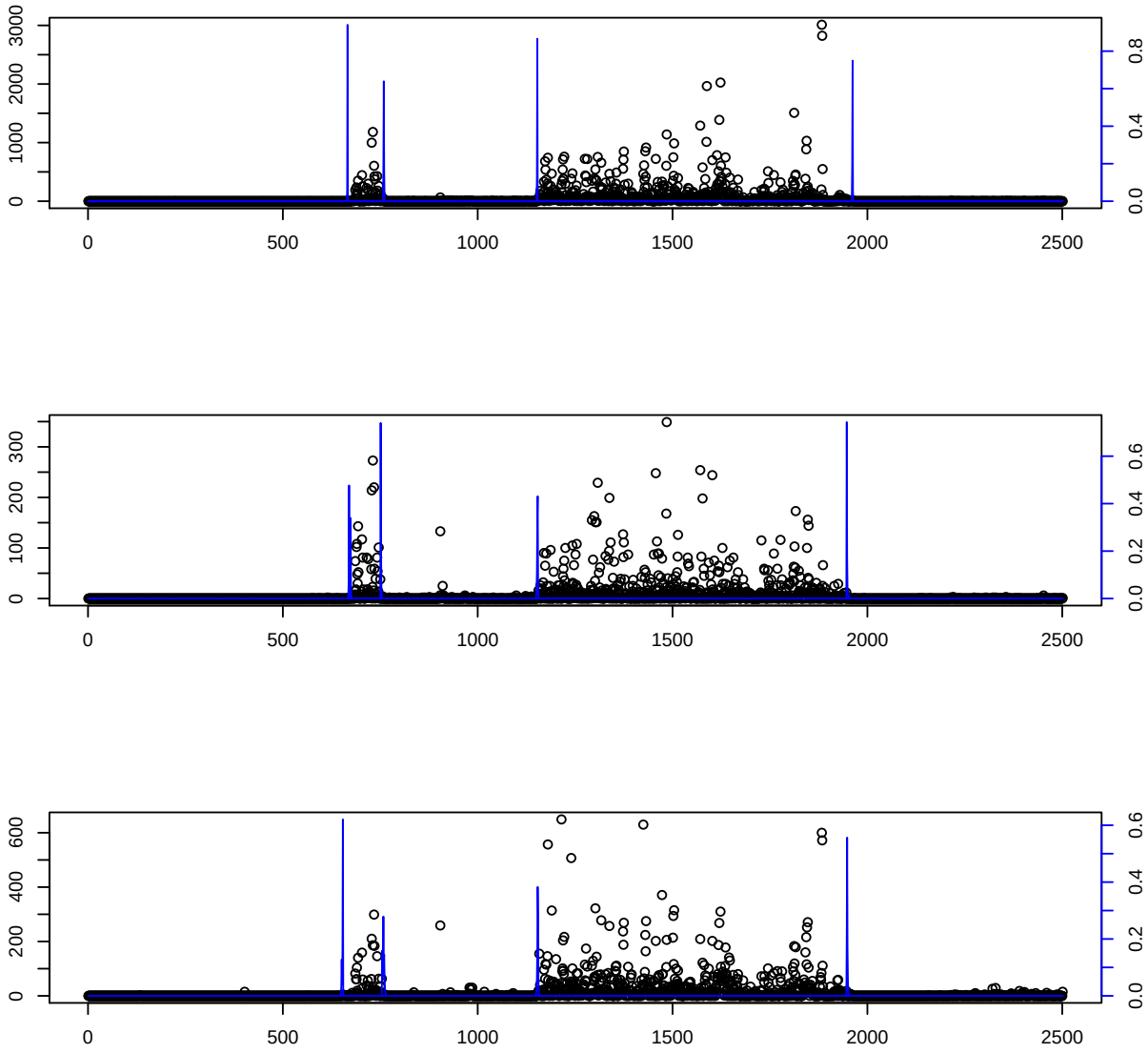
Estimation of ϕ (Inspired from Johnson, Kotz and Kemp):

- Let h equal 15;
- for each sliding window L of size h, compute moment estimator of ϕ_L using

$$\phi_L = \frac{\mathbf{E}^2(X^L)}{\mathbf{V}(X^L) - \mathbf{E}(X^L)}$$

- if $\hat{\phi} = median\{\widehat{\phi_L}\} < 0$, $h \leftarrow h * 2$, else keep $\hat{\phi}$

Illustration



Illustration

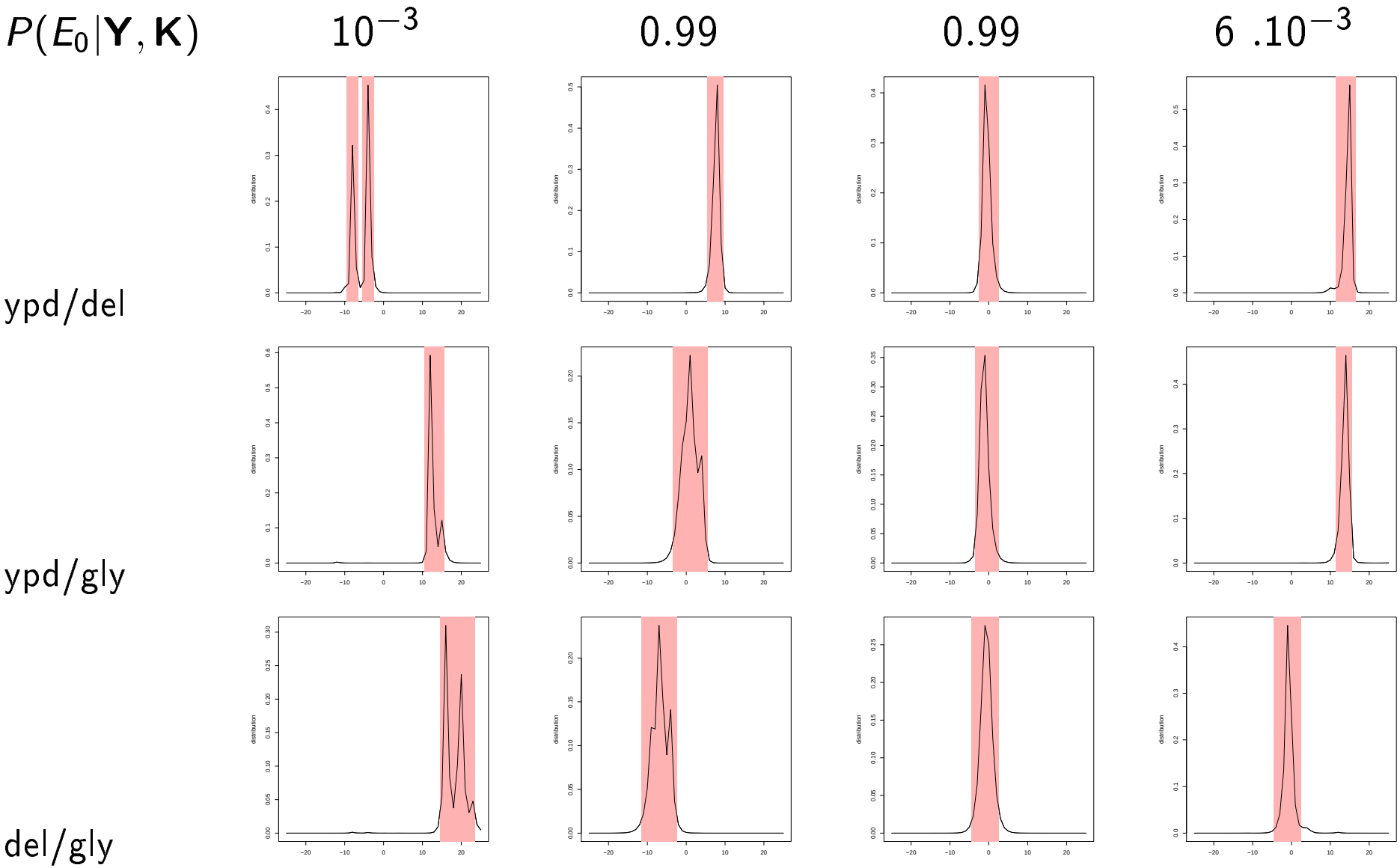
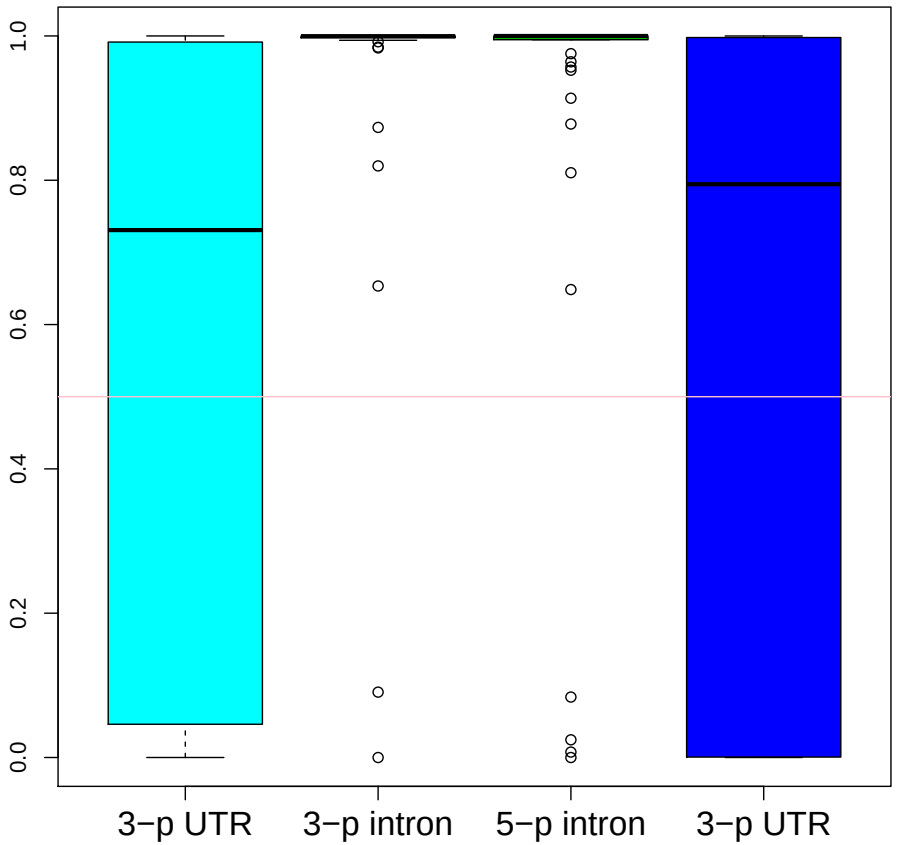


Illustration: 50 genes with 2 exons

Distribution of the posterior probability of event *Eo*
($p_0(K) = 1/2$)



Conclusion

Two exact approaches for the comparison of change-point location

- Posterior distribution of the shift in two profiles
- Posterior probability of the equality of change-point location for *I* profiles

$\rightarrow$ evidence of alternative splicing in the UTRs

$\rightarrow$ evidence of alternative isoforms

Available in R package `EBS` on the CRAN

Reference: arxiv

Thank you...