An Improved Bayesian Method for DNA Copy Number Estimation

Paola M.V. Rancoita[1,2]
Marcus Hutter[3,4]
Francesco Bertoni[2]
Ivo Kwee[1,2]

[1]IDSIA, Manno, Switzerland,
[2]IOSI, Bellinzona, Switzerland,
[3]ANU, Canberra, Australia,
[4]NICTA, Canberra, Australia

Summary

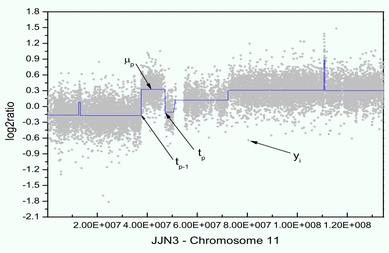
- Copy number data can be represented as a piecewise constant function
- We derive an improved version of Bayesian Piecewise Constant Regression (mBPCR)
- mBPCR is the best performing method for DNA copy number estimation

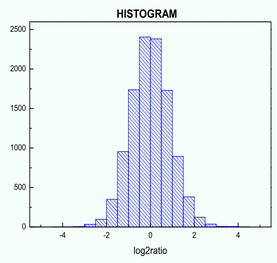
Copy number estimation

- Tumors are due to chromosomal aberrations affecting the DNA copy number in one or more regions of the genome.
- The Copy Number of a genomic region is the number of the copies of DNA in that region (in a healthy cell, the copy number is 2).
- Given a genomic region, we can divide it into subregions, where the CN (or better its $\log_2$ratio with respect to a normal reference sample) is constant.

  $\Rightarrow$ We can consider the $\log_2$ratio of the copy number as a piecewise constant function.

  $\Rightarrow$ We estimate the $\log_2$ratio copy number profile with a Bayesian piecewise constant regression (BPCR).
- The copy number is measured by microarray technology (such as a CGH array or SNP array), and the data can be very noisy due to both technical and biological reasons.





Hypotheses of BPCR model

We suppose that $\underline{Y}$ represents a noisy observation of a piecewise constant function with k_0 segments and boundaries $0 = t_0^0 < t_1^0 \cdots < t_{k_0-1}^0 < t_{k_0}^0 = n$ and that

$$Y_i|_{\mu_p,\,\sigma^2} \sim \mathcal{N}(\mu_p,\,\sigma^2) \qquad i = 1,\,\ldots,\,n$$

if Y_i belongs to the pth segment, for $p = 1,\,\ldots,\,k_0$.
Prior definition:

$$\mathrm{P}(K = k) = \frac{1}{k_{\max}} \qquad k \in \{1,\,\ldots,\,k_{\max}\} = I\!K$$

$$\mathrm{P}(\underline{T} = \underline{t}\,|\,K = k) = \frac{1}{\binom{n-1}{k-1}} \qquad \text{for each } \underline{t} \in \mathcal{T}_{k,n}$$

$$\underline{\mu}|_{\nu,\,\rho^2,\,K=k} \sim \mathcal{N}(\nu\underline{1},\,\rho^2\,I\!I)$$

Estimators of the original BPCR

Number of segment estimator:

$$\widehat{K}_{01} = \arg\max_{k \in I\!K} p(k\,|\,\underline{Y})$$

Boundary estimator:

$$\widehat{T}_{01,p} = \arg\max_{h \in \{p,\,\ldots,\,n-(\hat{k}-p)\}} \mathrm{P}(T_p = h\,|\,\underline{Y},\hat{k}) \qquad p = 1,\,\ldots,\,\hat{k}-1$$

Segment level estimators:

$$\widehat{\mu}_m = \mathrm{E}[\mu\,|\,\underline{Y},\widehat{\underline{t}},\hat{k}] \qquad m = 1,\,\ldots,\,\hat{k}$$

Hyperparameter estimators:

$$\widehat{\nu} = \frac{1}{n}\sum_{i=1}^{n} Y_i = \overline{Y},\;\; \widehat{\sigma}^2 = \frac{1}{2(n-1)}\sum_{i=1}^{n-1}(Y_{i+1}-Y_i)^2,\;\; \hat{\rho}^2 = \frac{1}{n}\sum_{i=1}^{n}(Y_i-\overline{Y})^2.$$

Practical and theoretical problems

- On simulated data with medium/high noise, BPCR did not perform best in comparison with the other existing methods.
- Since $\widehat{T}_{01}$ estimates each breakpoint separately, some breakpoints can be estimated with the same location loosing segments.

$\Rightarrow$ Definition of other estimators

$\Downarrow$

- Change error that the Bayesian estimators minimizes $\left(\widehat{\Theta}_{Bayesian} = \arg\min_{\theta'} \mathrm{E}[\text{error}(\theta,\,\theta')\,|\,\underline{Y}]\right)$
- Change estimator of the variance of the segment levels ($\hat{\rho}^2$), which is very biased

New ρ^2 estimator & comparison

If Y_i and Y_j belong to the same segment,

$$\mathrm{Cov}(Y_i,\,Y_j|\nu,\,\rho^2,\,\sigma^2) = \rho^2 \qquad i \neq j$$

$$\Rightarrow \widehat{\rho}_1^2 := \frac{1}{n}\left|\sum_{i=1}^{n}(Y_i-\overline{Y})(Y_{i+1}-\overline{Y})\right|$$

Results on dataset sampled from our prior distributions:

	$\sigma^2=0.1$ $\rho^2=0.5$	$\sigma^2=0.3$ $\rho^2=0.05$	$\sigma^2=0.5$ $\rho^2=0.02$	$\sigma^2=0.5$ $\rho^2=0.05$	$\sigma^2=0.5$ $\rho^2=0.1$	$\sigma^2=0.7$ $\rho^2=0.5$	$\sigma^2=1$ $\rho^2=0.05$	$\sigma^2=1.2$ $\rho^2=0.5$
$\mathrm{MSE}_{\hat{\rho}^2}$	0.0633	0.0871	0.2508	0.2404	0.2426	0.4271	0.9921	1.3254
$\mathrm{MSE}_{\hat{\rho}_1^2}$	0.068	0.0009	0.0008	0.0014	0.0047	0.0593	0.0024	0.0623

$\Rightarrow$ It seems better to use $\hat{\rho}_1^2$ when $\sigma^2 > \rho^2$, while in the other case we can use also $\hat{\rho}^2$.

New k estimators & comparison

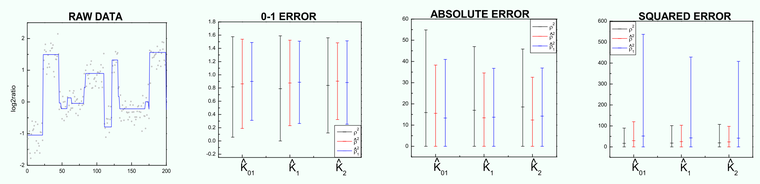
$$\widehat{K} = \arg\min_{k'} \mathrm{E}[\text{error}(k_0,\,k')\,|\,\underline{Y}]$$

$$\text{0-1 error} = 1 - \delta_{k_0-k'} \;\Rightarrow\; \widehat{K}_{01}$$

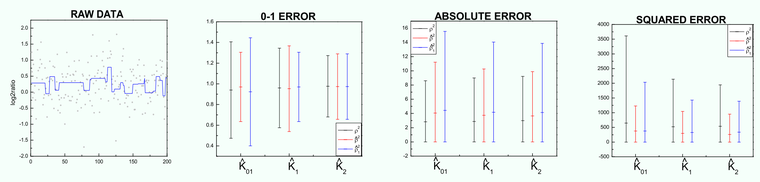
$$\text{absolute error} = \left|k_0 - k'\right| \;\Rightarrow\; \widehat{K}_1$$

$$\text{square error} = \left(k_0 - k'\right)^2 \;\Rightarrow\; \widehat{K}_2$$

Error confidence intervals on dataset with $\sigma^2 = 0.1$ and $\rho^2 = 0.5$:



Error confidence intervals on dataset with $\sigma^2 = 0.3$ and $\rho^2 = 0.05$:



$\Rightarrow$ $\widehat{K}_2$ always has small(est) error

New $\underline{t}$ estimators & comparison

$$\widehat{T}_{01,p} = \arg\max_{t'_p} \mathrm{P}(t'_p\,|\,\underline{Y},\,\hat{k})$$
$$\approx \arg\min_{t'_p} \mathrm{E}[1 - \delta_{t_p^0 - t'_p}\,|\,\underline{Y},\,k_0] \;\text{(0-1 error per component)}$$

$$\Rightarrow \widehat{\underline{T}}_{joint} = \arg\max_{\underline{t}'} \mathrm{P}(\underline{t}'\,|\,\underline{Y},\,\hat{k})$$
$$\approx \arg\min_{\underline{t}'} \mathrm{E}[1 - \delta_{\underline{t}^0 - \underline{t}'}\,|\,\underline{Y},\,k_0] \;\text{(total 0-1 error)}$$

The latter takes into account the dependency among the breakpoints, but the total 0-1 error cannot be defined for $\hat{k} \neq k_0$.

$\Rightarrow$ Definition of another error: the **binary error**

$\Downarrow$

The boundary vectors are mapped into vectors with the same length on which we define a suitable error:

1. Definition of the binary transformation for the boundary vectors

$$\underline{t} \mapsto \underline{\tau} \;\text{ such that }\; \tau_i = \begin{cases} 1 & \text{if } \exists p \text{ such that } t_p = i \\ 0 & \text{otherwise} \end{cases}$$

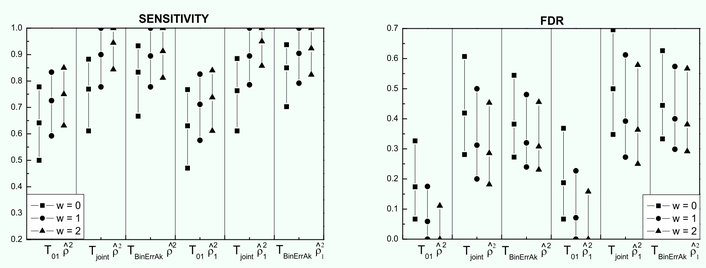
2. Definition of the binary error on the transformed vectors

$$\text{binary error} = k_0 - 1 - \langle \underline{\tau}^0,\,\underline{\tau}\rangle = k_0 - 1 - \sum_{i=1}^{n-1} \tau_i^0 \tau_i$$

3. Then, $\widehat{\underline{T}}_{\mathrm{BinErrAk}}$ is the inverse image of

$$\widehat{\underline{T}}_{\mathrm{BinErrAk}} = \arg\max_{\underline{\tau}' \in \mathcal{T}_{\hat{k},n}} \mathrm{E}\left[\sum_{i=1}^{n-1} \tau_i \tau_i' \,\middle|\, \underline{Y}\right].$$

Results on dataset *Simulated Chromosomes* (i.e. simulated chromosomic profiles used in Willenbrock and Fridlyand (2005), *Bioinformatics*, **21** 4084-4091):



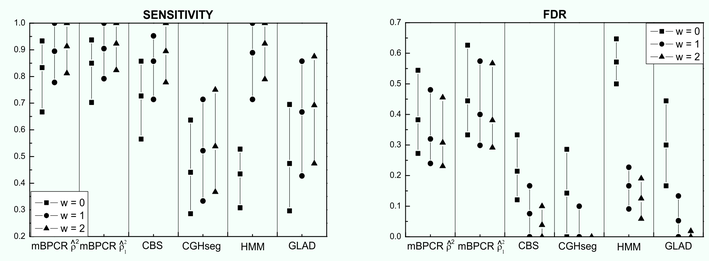
(if $w = i$, we look if $\hat{t}_p \in [t_q^0 - i,\,t_q^0 + i]$, for all $p = 1,\,\ldots,\,\hat{k}$ and $q = 1,\,\ldots,\,k_0$)

			SSQ	MAD	accuracy
$\widehat{K}_2$	$\hat{\rho}^2$	$\widehat{\mathrm{T}}_{01}$	14.23	0.00877	0.889
		$\widehat{\mathrm{T}}_{joint}$	2.22	0.00840	0.904
		$\widehat{\mathrm{T}}_{BinErrAk}$	1.70	0.00733	0.936
	$\hat{\rho}_1^2$	$\widehat{\mathrm{T}}_{01}$	9.74	0.00952	0.881
		$\widehat{\mathrm{T}}_{joint}$	2.67	0.00970	0.882
		$\widehat{\mathrm{T}}_{BinErrAk}$	1.85	0.00781	0.929

$\Rightarrow$ $\widehat{\mathrm{T}}_{BinErrAk}$ performs best

Comparisons with existing methods

Results on dataset *Simulated Chromosomes*:



	SSQ	MAD	accuracy
mBPCR $\hat{\rho}^2$	1.70	0.00733	0.936
mBPCR $\hat{\rho}_1^2$	1.85	0.00781	0.929
CBS	1.56	0.00705	0.953
CGHseg	5.42	0.00795	0.925
HMM	4.47	0.00350	0.993
GLAD	4.15	0.00846	0.939

Results on real data: gene regions

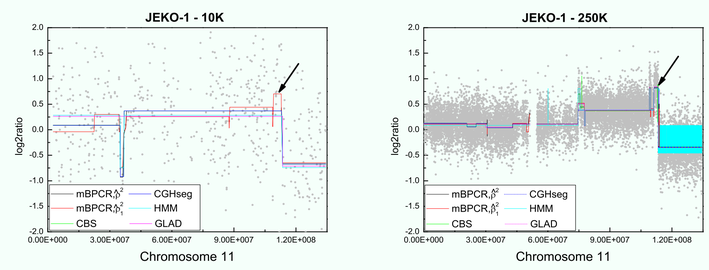
JEKO-1 10K

Estimated copy numbers

gene region	FISH CN	mBPCR $\hat{\rho}^2$	mBPCR $\hat{\rho}_1^2$	CBS	CGHseg	HMM	GLAD
BCL6	3/2	2.97	2.99	2.97	2.90	2.92	2.92
C-MYC	ampl	12.11	9.35	10.27	10.27	13.95	9.82
CCND1	2	*4.08*	*3.77*	*4.08*	*4.08*	*3.84*	*3.79*
BIRC3	4/5	4.08	4.29	4.08	4.08	3.84	3.79
ATM	4	4.08	4.29	4.08	4.08	3.84	3.79
D13S319	4	3.72	3.59	3.57	3.72	3.62	3.58
LAMP1	4	*3.41*	3.82	*3.41*	*3.41*	3.62	*2.49*
TP53	2/3	2.81	3.00	2.83	2.50	*3.52*	2.93
BCL2	4	3.63	3.62	*3.48*	3.64	*3.42*	*3.42*
MALT1	4	3.63	3.62	*3.48*	3.64	*3.42*	*3.42*

= cn wrongly identified

Example of estimated profile:



On the 10K data, mBPCR $\hat{\rho}_1^2$ is the only method which identifies the amplification after position 105Mb.

Conclusions

- mBPCR is the best method in detecting the true position of the breakpoints also on very noisy data
- On real data, mBPCR performed best (the second best method is CBS).

Remarks:

- mBPCR has a high FDR in the detection of the breakpoints, but this usually does not affect the estimation

  $\Rightarrow$ we obtained a reduction of the FDR using prior of K proportional to $1/k^2$ and its MAP estimator
- If the real data contains only few aberrations the variance of the level ρ^2 cannot be well estimated

  $\Rightarrow$ we can use ρ^2 estimated with $\hat{\rho}_1^2$ on one or more cell line genomes which contains several aberrations
- The computational complexity is $O(k_{max}n^2)$, but the computation can be parallelized by chromosome.