Multiple Change-Point Detection and Analysis of Chromosome Copy Number Variations

Heping Zhang

Yale School of Public Health

Joint work with Ning Hao, Yue S. Niu

presented @Tsinghua University

Outline

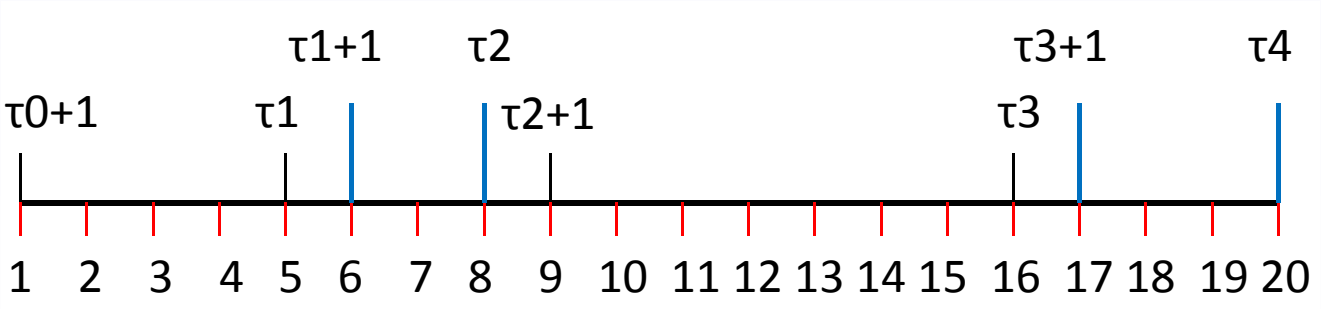
Outline

The Problem

Multiple Change-points Problem: $Y_1,...,Y_n$ are a sequence of independent random variables with $Y_j \sim F_j$. There are J change-points $0 = \tau_0 < \tau_1 < \cdots < \tau_J < \tau_{J+1} = n$ such that

- $F_{\tau_k+1} = F_{\tau_k+2} = \cdots = F_{\tau_{k+1}}$ for all $k = 0,...,J$;
- $F_{\tau_k} \neq F_{\tau_k+1}$ for all $k = 0,...,J$.

It is usually assumed that F_i belongs to a specified parametric family.



Highlights

Goal: Estimate the number and locations of the change-points.

Setting: n is large and $J \ll n$.

Feature: High dimensionality; Sparsity; Sequential Structure.

Tool: The Screening and Ranking Algorithm (SaRa).

Outline

Copy Number Variation

- DNA copy number: The number of copies of the DNA;
- Copy number variants (CNVs), i.e., gains or losses of segments of chromosomes, comprise an important class of genetic variation;
- CNVs: Inherited (present in parents) or de novo (absent in parents) mutation;
- CNVs: Associated with complex diseases.

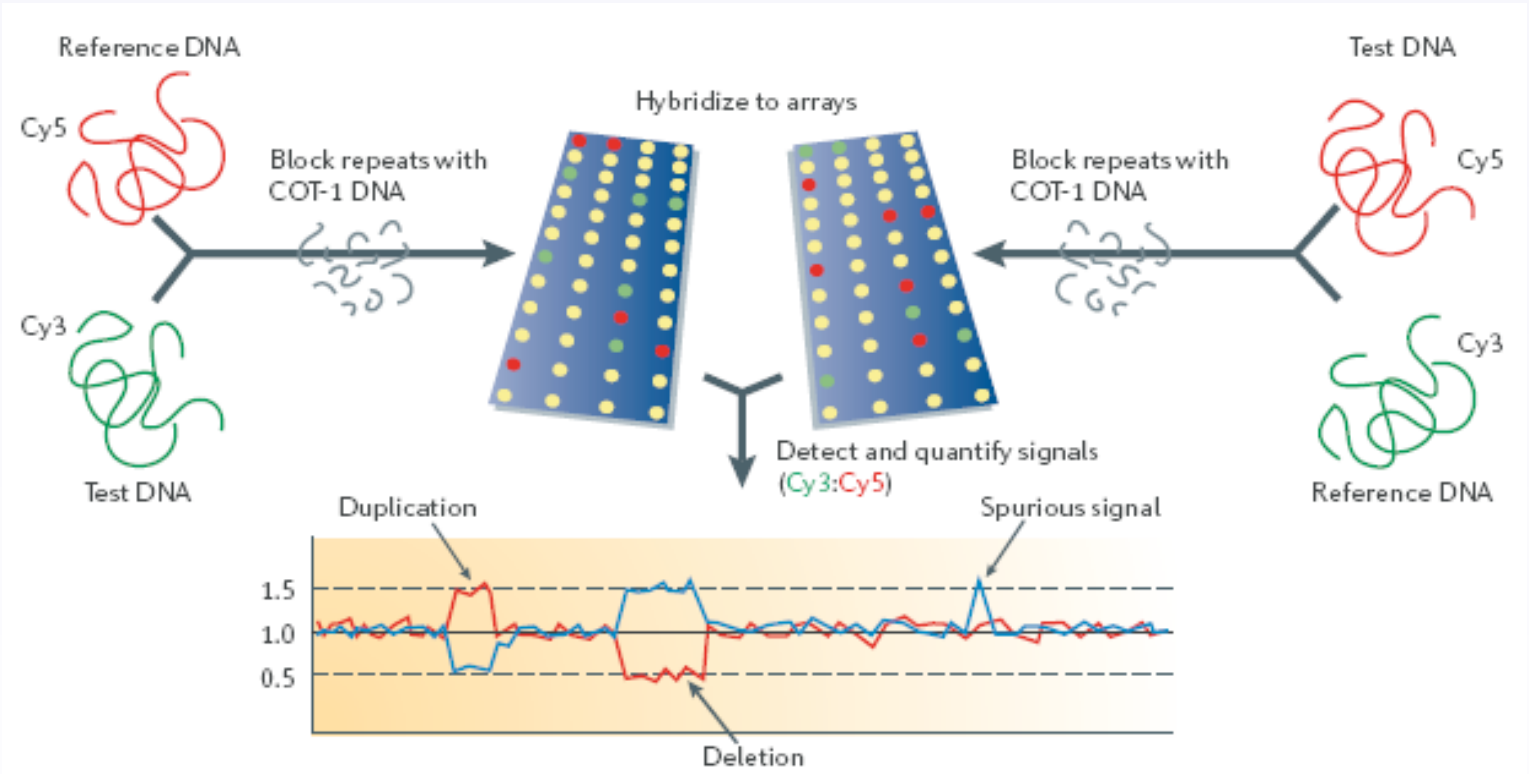
Copy Number Variations and Diseases

- Autism: MZ twins share the same deletion/duplication event, explaining why the concordance rate in MZ twins is high.
- Schizophrenia: Deletion in the 22q11.2 region from 17-21Mb to 3Mb was identified.
- Crohn's disease: The causal mutations were reported.

Platforms and data

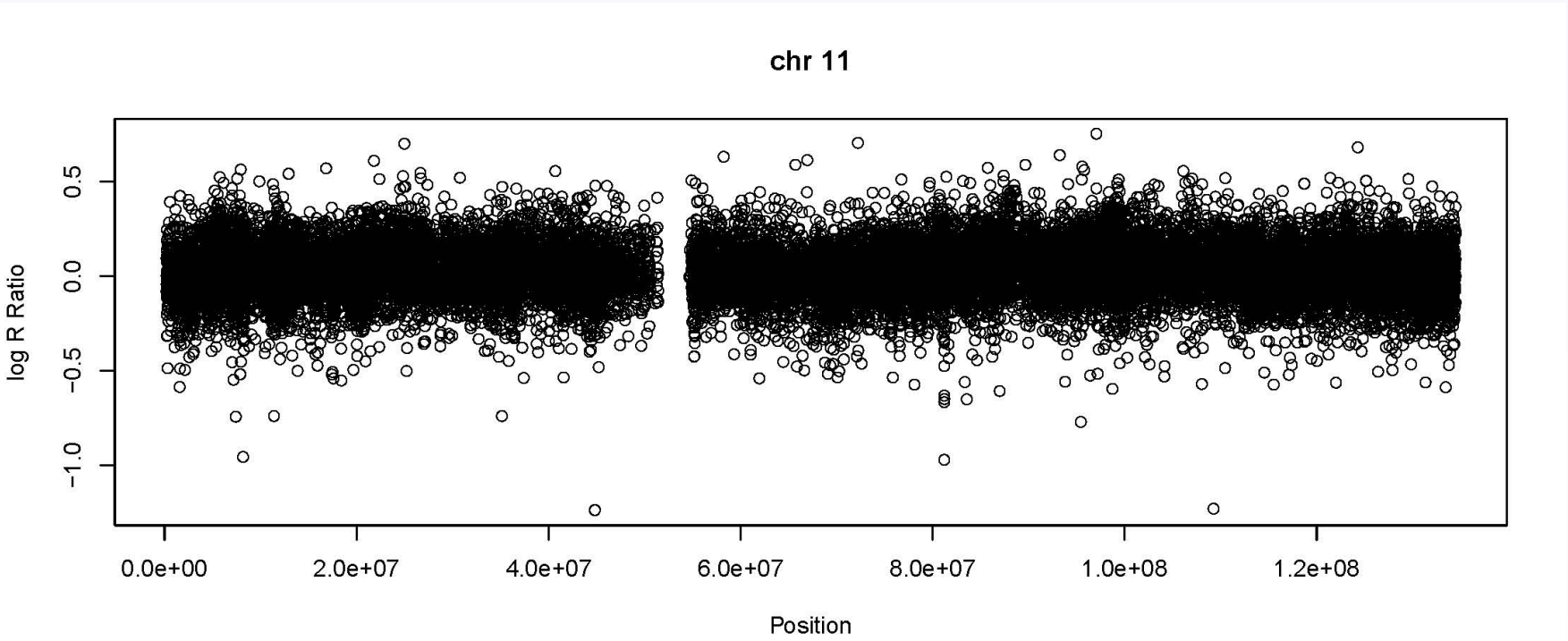
- Popular genome analysis platforms include array comparative genomic hybridization (aCGH) and SNP genotyping platforms.
- aCGH: data = $\log_2$ ratios of test and reference fluorescent intensities. Sample size $\approx$ a few thousands.
- SNP genotyping: data = "Log R Ratio"
  Total fluorescent intensity signals (alleles A and B) at each SNP. Sample size $\approx$ tens of thousands per chromosome, tens of thousands or millions along whole genome.
- Goal: identify segments of concentrated high or low log-ratios.

aCGH(Pinkel & Albertson 2001)



SNP genotyping data: a first look

Data: SNP genotyping data from illumina 500K platform.



Outline

Model formulation: normal mean model

Normal mean model:

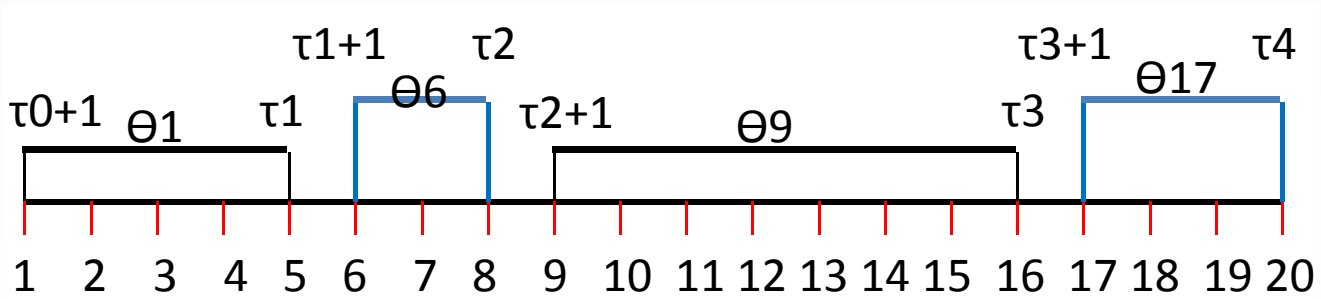
$$y_i = \theta_i + \varepsilon_i, \quad \varepsilon_i \overset{\text{iid}}{\sim} \mathcal{N}(0, \sigma^2), \quad i = 1, ..., n. \tag{1}$$

Moreover, we assume the mean vector $\boldsymbol{\theta} = (\theta_1, ..., \theta_n)^T$ is piecewise constant. In other words, we assume that

$$\theta_1 = \cdots = \theta_{\tau_1} \neq \theta_{\tau_1+1} = \cdots = \theta_{\tau_2} \neq \theta_{\tau_2+1} = \cdots$$

$$\cdots = \theta_{\tau_J} \neq \theta_{\tau_J+1} = \cdots = \theta_n,$$

where $\boldsymbol{\tau} = (\tau_1, ..., \tau_J)^T$ is the location vector of change-points.



Model formulation: existing work

Model (1) or more restrictive ones have been considered in Olshen et al.(2004); Huang et al.(2005); Zhang & Siegmund (2007); Tibshirani & Wang (2008); Jeng et al. (2010), among others.

Model formulation: regression model

Note that in model (1), the sparsity is encoded in the piecewise constant structure of $\boldsymbol{\theta}$. De-trend the θ's,

$$\beta_0 = \theta_1, \quad \beta_i = \theta_{i+1} - \theta_i; \quad i = 1, \ldots, n-1.$$

Model (1) is transformed to a sparse linear regression model,

$$y_i = \sum_{j=0}^{i-1} \beta_j + \varepsilon_i, \qquad i = 1, \ldots, n.$$

Model formulation: regression model

The model above can be rewritten as

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}, \tag{2}$$

where $\boldsymbol{\beta} = (\beta_0, \ldots, \beta_{n-1})^T$ is a sparse vector and the design matrix

$$\mathbf{X} = \begin{pmatrix} 1 & 0 & 0 & \cdots & 0 & 0 \\ 1 & 1 & 0 & \cdots & 0 & 0 \\ 1 & 1 & 1 & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & 1 & 1 & \cdots & 1 & 1 \end{pmatrix}.$$

Single change-point case

If we know in advance that there is at most one change-point in model (1), the problem becomes the following hypothesis testing problem

$$\begin{aligned} H_0 &: \theta_1 = \cdots = \theta_n, \quad \text{against} \\ H_1 &: \theta_1 = \cdots = \theta_j \neq \theta_{j+1} = \cdots = \theta_n \text{ for some } 1 \leq j < n. \end{aligned} \tag{3}$$

For simplicity, we assume $\sigma^2 = 1$. If j is fixed in H_1, we can calculate

$$-2\log\Lambda_j = (\bar{Y}_{j+} - \bar{Y}_{j-})^2/[1/j + 1/(n-j)], \tag{4}$$

where Λ_j is the likelihood ratio, $\bar{Y}_{j-} = \sum_{k=1}^{j} Y_k/j$ and $\bar{Y}_{j+} = \sum_{k=j+1}^{n} Y_k/(n-j)$.

Single change-point case

When j is unknown, it is natural to use

$$T_1 = \max_{1 \le j \le n-1} (-2 \log \Lambda_j)$$

as test statistic for problem (3).
Moreover, when the alternative is supported,

$$\hat{j} = \operatorname*{argmax}_{1 \le j \le n-1} (-2 \log \Lambda_j)$$

is the location estimator.

Single change-point case

Accuracy of $\hat{j}$

If H_1 is true, $j(n)/n \to 0$, $\delta(n) = \theta_{j+1}(n) - \theta_j(n) \to 0$, with $\lim_{n\to\infty} \frac{j(n)\delta^2}{\log\log n} = \infty$, then

$$\delta^2|\hat{j} - j| = O_P(1) \quad i.e. \quad \delta^2 \left|\frac{\hat{j}}{n} - \frac{j}{n}\right| = O_P\left(\frac{1}{n}\right).$$

If the change point is not too close to the end and the jump is not too small, we can detect the change point within a reasonable precision. The precision depends on the location and jump size.

Multiple change-point case: exhaustive search

Ignoring its computational complexity, an exhaustive search among all possibilities $0 \le J \le n-1$ and $0 < \tau_1 < \cdots < \tau_J < n$ can be applied. For any J and $\boldsymbol{\tau} = (\tau_1, \ldots, \tau_J)^T$, denote by $\hat{\sigma}^2_{J,\boldsymbol{\tau}}$ the MLE of the variance. Define $\hat{\sigma}^2_J = \min_{\boldsymbol{\tau}} \hat{\sigma}^2_{J,\boldsymbol{\tau}}$.
Yao (1988) showed that

$$\hat{J} = \underset{J}{\operatorname{argmin}}\left(\frac{n}{2}\log\hat{\sigma}^2_J + J\log n\right). \tag{5}$$

is consistent estimator for J^*–the true number of change points.
Yao & Au (1989) showed $\hat{\boldsymbol{\tau}} = \operatorname{argmin}\hat{\sigma}^2_{J^*,\boldsymbol{\tau}}$ is a consistent estimator for $\boldsymbol{\tau}^*$–the vector of the true change points.
Assumption: J is fixed and $\boldsymbol{\tau}/n \to \mathbf{t}$ as $n \to \infty$.

Multiple change-point case: binary segmentation

Binary Segmentation (BS) algorithm (Vostrikova 1981) is a method which applies the single change-point test recursively. The BS procedure can be summarized in the following steps.

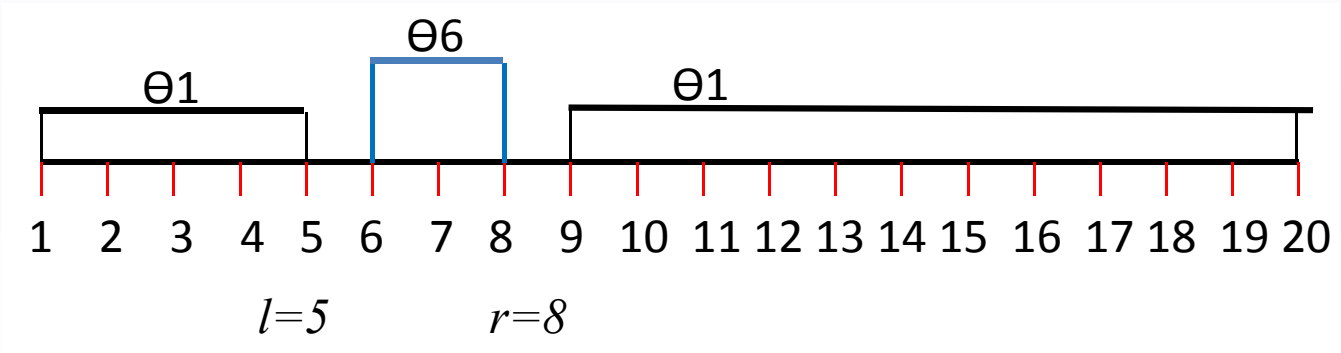
1. Test for no change-point versus one change point (3). If H_0 is not rejected, stop. Otherwise, there is a change-point $\hat{j}$.
2. Test the two segments before and after the change-point detected in step 1.
3. Step 3: Repeat the process until no further segments have change-points.

We see that this procedure is very similar to forward stepwise selection solving regression problem (2).

Multiple change-point case: binary segmentation

To make this algorithm more powerful in detecting short segments, Olshen et al. (2004) proposed Circular Binary Segmentation (CBS). The only difference is that CBS tests the epidemic alternative recursively over each segment.

$$
\begin{aligned}
H_0 \quad &: \quad \theta_1 = \cdots = \theta_n, \quad \text{against} \\
H_1 \quad &: \quad \theta_1 = \cdots = \theta_l = \theta_{r+1} = \cdots = \theta_n \neq \theta_{l+1} = \cdots = \theta_r \quad (6) \\
&\qquad \text{for some pair } l < r.
\end{aligned}
$$



Multiple change-point case: binary segmentation

Test statistic $T_2 = \max_{1 \le l < r \le n}(-2 \log \Lambda_{l,r})$,

$$-2 \log \Lambda_{l,r} = (\bar{Y}_I - \bar{Y}_O)^2 / [1/(r-l) + 1/(n-r+l)],$$

where

$$\bar{Y}_I = \sum_{k=l+1}^{r} Y_k/(r-l)$$

and

$$\bar{Y}_O = \sum_{k \le l \text{ or } k > r} Y_k/(n-r+l).$$

Multiple change-point case: ℓ_1 penalization

Huang et al. (2005) studied the following optimization problem

$$\text{minimize } ||\mathbf{y} - \boldsymbol{\theta}||^2 \qquad \text{subject to} \qquad \sum_j |\theta_j - \theta_{j+1}| \leq s. \tag{7}$$

After reparametrization $\beta_i = \theta_{i+1} - \theta_i$, the above optimization problem is equivalent to

$$\text{minimize } ||\mathbf{y} - \mathbf{X}\boldsymbol{\beta}||^2 \qquad \text{subject to} \qquad \sum_{j=1}^{n-1} |\beta_j| \leq s. \tag{8}$$

This is a special case of the fused lasso (Tibshirani & Wang 2008), which

$$\text{minimizes } ||\mathbf{y} - \boldsymbol{\theta}||^2 \text{ subject to } ||\boldsymbol{\theta}||_{\ell_1} \leq s_1, \quad \sum_j |\theta_j - \theta_{j+1}| \leq s_2.$$

Better Methods?

- Computational Complexity $O(n)$ or close to $O(n)$.
- Consistency: $P(\hat{J} = J^*) \to 1$; $\delta^2(\hat{\tau}_i - \tau_i^*) = O_P(1)$.
- Generalizability: Readily extendable to other settings.
- Nonasymptotic result, FDR control, etc.

Outline

The SaRa: the rationale

To determine whether a position is a change-point, it is enough to check observations in a neighborhood.
Suppose that the minimal distance between two change points is at least h. Consider the local hypothesis testing problem at position x:

$$\begin{aligned} H_0(x) &: F_{x+1-h} = \cdots = F_{x+h} \text{ vs} \\ H_1(x) &: F_{x+1-h} = \cdots = F_x \neq F_{x+1} = \cdots = F_{x+h}. \end{aligned} \tag{9}$$

The SaRa: the algorithm

Let $D(x)$ be a test statistic for (9), and $p(x)$ be the corresponding P-value. We may assume that a larger value of D is in favor of the alternative. The SaRa proceeds as follows.

1. Screening: calculate $D(x)$ (or $p(x)$) for each x.
2. Select all the local maximizers of $D(x)$ (or local minimizers of $p(x)$).
3. Ranking and thresholding: $D(x) > \lambda$ (or $p(x) < p^*$).

Here, we call x^* a local maximizer of $D(x)$ if

$$D(x^*) \geq D(x) \quad \text{for all} \quad x \in (x^* - h, x^* + h).$$

The SaRa estimator

$$\hat{\mathcal{J}}_{h,\lambda} = \{x | D(x) > \lambda \quad \& \quad x \text{ is a local max of } D(\cdot)\}.$$

$\hat{\boldsymbol{\tau}}$ is obtained by ordering elements in $\hat{\mathcal{J}}_{h,\lambda}$.

The SaRa for normal mean model

For the normal mean model, consider the local hypothesis testing problem at position x:

$$\begin{aligned} H_0(x) &: \theta_{x+1-h} = \cdots = \theta_{x+h} \text{ vs} \\ H_1(x) &: \theta_{x+1-h} = \cdots = \theta_x \neq \theta_{x+1} = \cdots = \theta_{x+h}. \end{aligned} \tag{10}$$

A reasonable test statistic is (Niu & Zhang 2010)

$$D_h(x) = \left| \left(\sum_{k=x-h+1}^{x} Y_k - \sum_{k=x+1}^{x+h} Y_k \right) / h \right|.$$

Computational complexity of the SaRa is $O(n)$, thanks to the recursion formula

$$D_h(x+1) = D_h(x) + (2Y_{x+1} - Y_{x-h+1} - Y_{x+h+1})/h.$$

The SaRa as "local correlation learning"

Let us revisit the high dimensional regression model

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}. \tag{2}$$

The correlation learning, e.g., Sure Independence Screening (Fan and Lv 2008), provides an approach to solving this regression problem. However, from the example below, we see that SIS may not work directly for (2). The SaRa is a localized version of the correlation learning and works well here.

Example: Assume $n = 300$, the true $\boldsymbol{\theta} = (-\mathbf{1}_{100}^T, \mathbf{0}_{100}^T, 2 \cdot \mathbf{1}_{100}^T)^T$, and $\sigma^2 = 0$. There are 2 change-points, 100 and 200.

The SaRa as "local correlation learning"

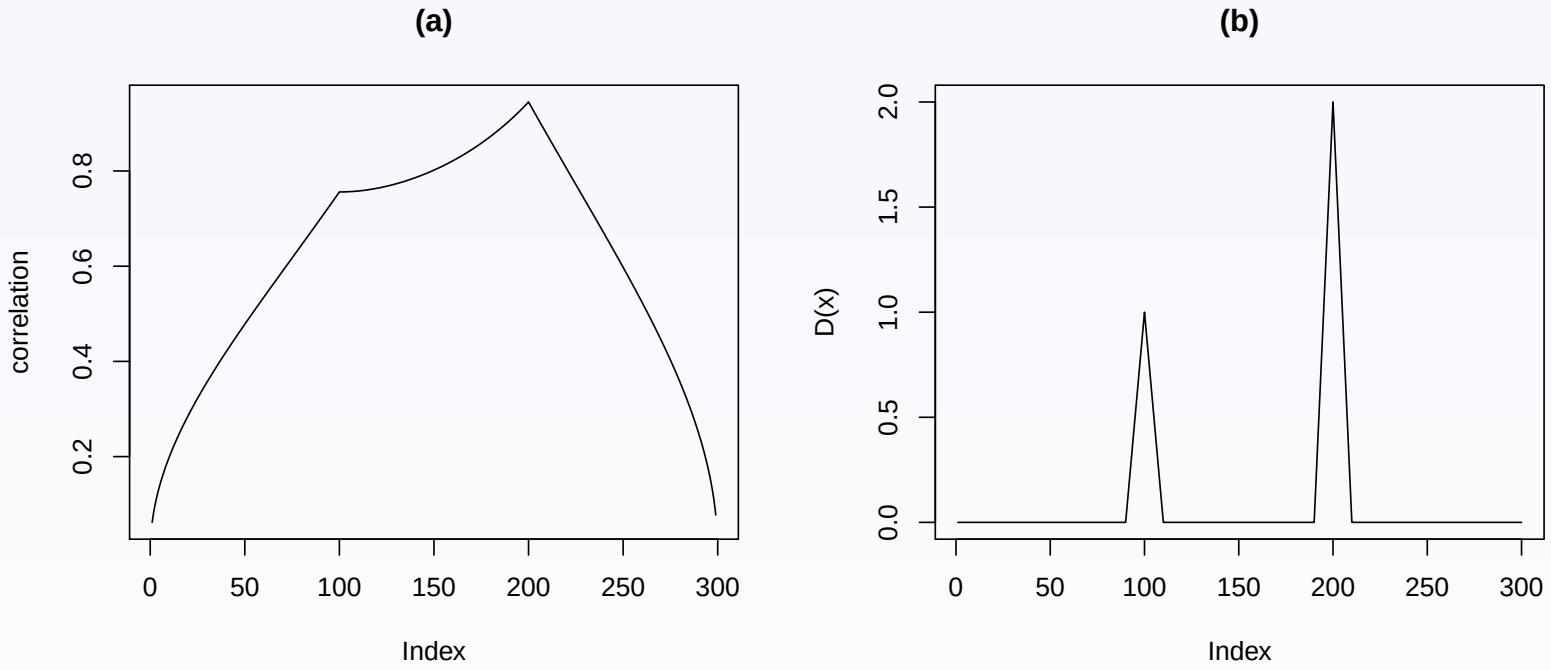


Figure: (a) Correlation between $\mathbf{y}$ and each $\mathbf{X}_i$; (b) Local statistic $D_{10}(\cdot)$.

$$D_{10}(100) = C \cdot \text{corr}(\mathbf{y}[91:110], \mathbf{X}_{100}[91:110]).$$

The SaRa: consistency

Asymptotic setting: Define

$$L = \min_{1\le j\le J+1}(\tau_j - \tau_{j-1}), \quad \delta = \min_{1\le j\le J}|\theta_{\tau_j+1} - \theta_{\tau_j}|,$$

where both J and $0 = \tau_0 < \tau_1 < \cdots < \tau_J < \tau_{J+1} = n$ depends on n.
We assume that

$$S^2 = \delta^2 L/\sigma^2 > 32\log n. \qquad (*)$$

The SaRa: consistency

Theorem 1

Under Assumption $(*)$, there exist $h = h(n)$ and $\lambda = \lambda(n)$ such that $\hat{\mathcal{J}} = \hat{\mathcal{J}}_{h,\lambda} = \{\hat{\tau}_1, \cdots, \hat{\tau}_{\hat{J}}\}$ satisfies

$$\lim_{n\to\infty} \mathbf{P}\left(\left\{\hat{J} = J\right\}\right) = 1;$$

$$\text{conditional on } \hat{J} = J, \qquad \delta^2(\hat{\tau}_i - \tau_i) = O_P(1).$$

In particular, taking $h = L/2$ and $\lambda = \delta/2$, we have

$$\mathbf{P}\left(\left\{\hat{J} = J\right\} \bigcap \bigcap_i \left\{|\hat{\tau}_i - \tau_i| < h\right\}\right) > 1 - 8S^{-1}\exp\{\log n - S^2/32\}.$$

The SaRa: FDR control

Multiple change-points problem can be considered as multiple testing problem. The tricky part is how to deal with "$H_0(x)$ vs $H_1(x)$" for those x's which are not change-points but close to change-points.

Define "$H_1(x)$ is discovered successfully" if a decision rule rejects $\hat{x}$ which is close to a true change-point x, say $\hat{x} \in [x + 1 - h, x + h]$.

Consider local minimal $p_{i_1}, ..., p_{i_N}$, which are nearly independent conditional on $i_1, ..., i_N$ are local mins of the P-value sequence. The conditional distribution p_{i_k}, depending only on h, can be approximated accurately and denoted by F_h. Any FDR control procedure can be applied to $F_h^{-1}(p_{i_1}), ..., F_h^{-1}(p_{i_N})$.

The SaRa: generalizability

The SaRa can be generalized to

- Heteroscedastic normal mean model.
- Mean shift model with non-Gaussian noise.
- Exponential family.
- Multivariate case.
- .....

Outline

Numerical Study I: Sure Coverage Property

Model: $Y_i = \theta_i + \varepsilon_i$, where $\theta_i = \delta \cdot I_{\{n/2 < i \le n/2+L\}}$.

- Fix jump size $\delta = 1$;
- Set $(n, L) = (400, 12), (3000, 16), (20000, 20)$ and $(160000, 24)$. $L \approx 2 \log n$.
- $\varepsilon_i \overset{\text{i.i.d.}}{\sim} N(0, \sigma^2)$ with $\sigma = 0.5, 0.25$. Correspondingly $S^2 \approx 8 \log n$ and $32 \log n$.
- Applying thresholding rule, $h = \frac{3}{4}L$, $\lambda = \frac{3}{4}\delta$.

Numerical Study I: Sure Coverage Property

Table: The estimated model sizes $\hat{J}$ and Sure Coverage Probabilities (SCP) of SaRa. Column 3 lists the distribution and mean value of the estimated number of change-points. Column 4 and 5 list SCPs of two change-points as well as mean distance between estimated change-point locations and true locations. The results are based on 1000 replications.

		number of change-points				change-point 1	change-point
(n, L)	σ	$\hat{J} = 2$	< 2	> 2	Mean	SCP (Mean)	SCP (Mean
(400,12)	0.5	63.5%	11.7%	24.8%	2.175	91.3% (0.756)	91.3% (0.71
	0.25	98.2%	1.8%	0.0%	1.980	98.9% (0.129)	99.1% (0.11
(3000,16)	0.5	60.3%	8.3%	31.4%	2.306	92.8% (0.814)	93.4% (0.77
	0.25	98.1%	1.9%	0.0%	1.980	99.3% (0.118)	98.7% (0.12
(20000,20)	0.5	60.2%	6.3%	33.5%	2.343	94.3% (0.862)	94.8% (0.84
	0.25	99.3%	0.7%	0.0%	1.993	99.5% (0.139)	99.8% (0.10
(160000,24)	0.5	49.5%	5.0%	45.5%	2.599	95.8% (0.877)	95.0% (1.01
	0.25	99.5%	0.5%	0.0%	1.995	99.8% (0.096)	99.7% (0.14

Numerical Study II: FDR control

Still consider model (1).

- We set $n = 30000$, $\sigma = 1$, $J = 50$.
- We drew 50 change-points uniformly among $\{x \in \mathbb{N} : x < 20000\}$, producing $\boldsymbol{\tau} = (430, 570, \cdots, 19750)^T$.
- $L = \min(\tau_{j+1} - \tau_j) = 15$.
- $\theta_i = 0$ when $\tau_{2j-1} \leq i \leq \tau_{2j}$; $\theta_i = 1.5$ or 3 otherwise.

We tried the SaRa with $h = 10$, 20, 30 and the threshold chosen by Benjamini Hochberg procedure with target FDR $q = 0.05, 0.10, 0.15$.
$\hat{\tau}_k$ is "falsely discovered" if there is no τ_j such that $|\hat{\tau}_k - \tau_j| < 10$. Otherwise, $\hat{\tau}_k$ is a "true positive".

Numerical Study II: FDR control

Table: The average estimated number of change-points $\hat{J}$, true positives (TP) and false discovery proportion (FDP). The results are based on 100 replications.

	q=0.05			q=0.10			q=0.15		
(δ, h)	$\hat{J}$	TP	FDP	$\hat{J}$	TP	FDP	$\hat{J}$	TP	FD
$(1.5, 10)$	3.70	3.52	0.4%	20.86	19.13	7.6%	27.69	23.64	13.6
$(1.5, 20)$	45.73	43.60	4.5%	50.71	45.60	9.9%	54.62	46.56	14.5
$(1.5, 30)$	50.58	47.13	6.7%	53.80	47.38	11.7%	56.74	47.46	16.1
$(3, 10)$	51.50	49.92	3.0%	53.68	49.97	6.7%	57.04	49.98	12.1
$(3, 20)$	50.38	49.07	2.5%	52.82	49.07	7.0%	55.00	49.07	10.6
$(3, 30)$	50.77	48.65	4.1%	53.00	48.65	8.0%	55.49	48.65	12.1

Numerical Study III: CNV detection

Data: SNP genotyping data from illumina 550K platform.
(father.txt included in PennCNV package)
Y = Log R Ratios of Chr 11, $n = 27272$.

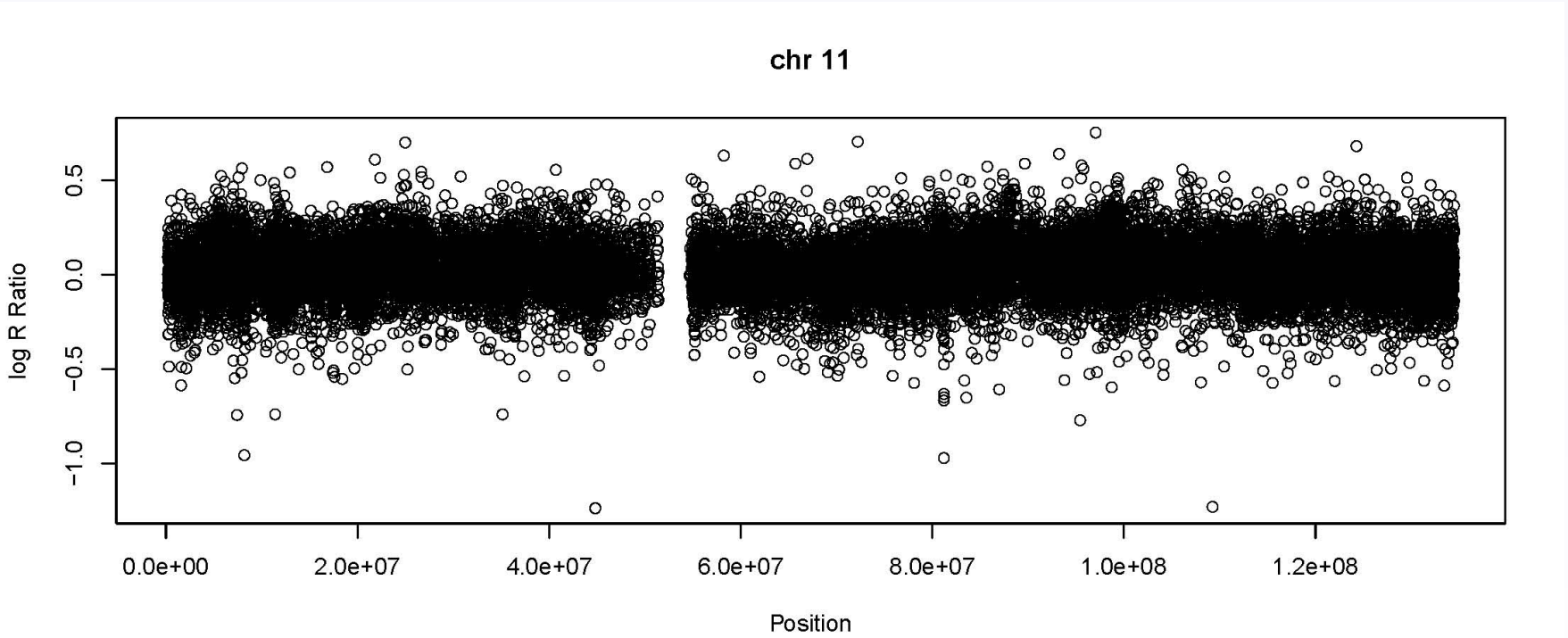


Figure: Log R Ratio of Chromosome 11.

Numerical Study III: CNV detection

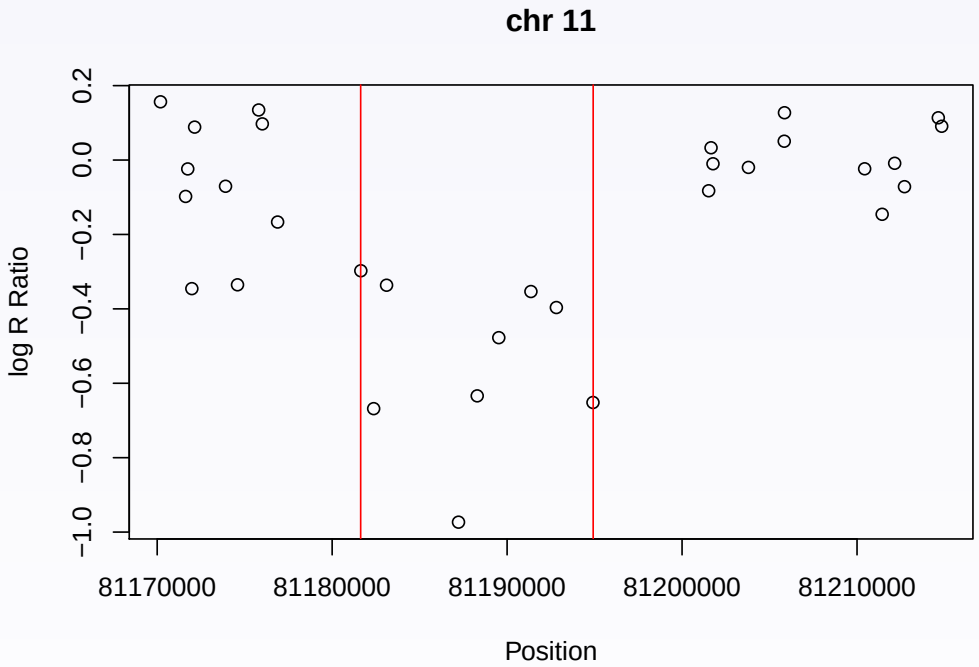


Figure: CNV in Chromosome 11.

Outline

Better Methods? Answer= The SaRa.

- Computational Complexity $O(n)$ or close to $O(n)$.
- Consistency: $P(\hat{J} = J^*) \to 1$; $\delta^2(\hat{\tau}_i - \tau_i^*) = O_P(1)$.
- Extensibility: Extensible to more general setting.
- nonasymptotic result, FDR control, etc.

Reference

Huang, T., Wu, B., Lizardi, P. & Zhao, H. (2005). Detection of DNA copy number alterations using penalized least squares regression. Bioinformatics 21, 3811-3817.

Jeng, X. J., Cai, T. T. & Li, H. (2010). Optimal sparse segment identification with application in copy number variation analysis. Journal of the American Statistical Association 105, 1056-1066.

Niu, Y. S. & Zhang, H. (2010). The Screening and Ranking Algorithm to Detect DNA Copy Number Variations. manuscript.

Olshen, A. B., Venkatraman, E. S., Lucito, R. & Wigler, M. (2004). Circular binary segmentation for the analysis of array-based DNA copy number data. Biostatistics 5, 557-72.

Tibshirani, R. & Wang, P. (2008). Spatial smoothing and hot spot detection for cgh data using the fused lasso. Biostatistics 9, 18-29.

Yao, Y.-C. (1988). Estimating the number of change-points via schwarz criterion. Statistics & Probability Letters 6, 181-189.

Yao, Y.-C. & Au, S. T. (1989). Least-squares estimation of a step function. Sankhya: The Indian Journal of Statistics, Series A 51, 370-381.

Zhang, N. R. & Siegmund, D. O. (2007). A modified bayes information criterion with applications to the analysis of comparative genomic hybridization data. Biometrics 63, 22-32.

The End

Thank you!