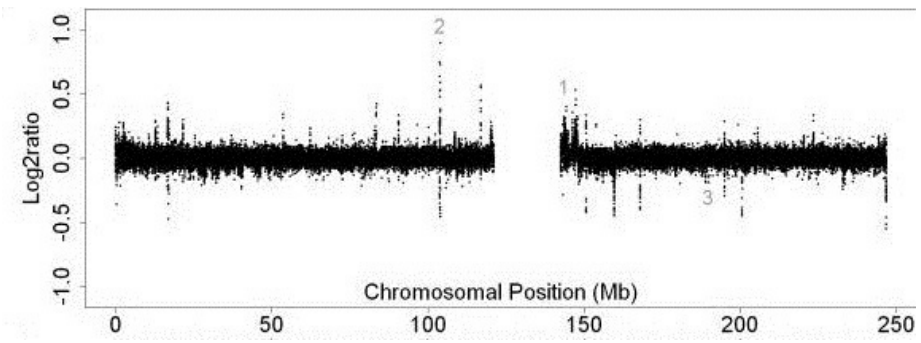
Fast estimation of posterior change-point probabilities for CNV data

The Minh Luong, Yves Rozenholc, Gregory Nuel, MAP5,
Université Paris Descartes

July 5, 2012

Introduction

- Change-point methods: applications in econometrics, engineering, network security, signal processing, music classification, bioinformatics
  - e.g. copy number variation (CNV), to identify regions where DNA mutations are related to disease susceptibility
- High-resolution data, 10's thousands of clones per chromosome
  - Array comparative genomic hybridization (aCGH)
  - Single nucleotide polymorphism (SNP) array



array CGH profile, source: Redon and Carter, Methods Mol Biol. 2009; 529: 37-49.

Examples of R packages for change-point analysis

Unsupervised hidden Markov model (HMM) approaches

- Willenbrock and Fridyland (2005) - aCGH package
- Marioni et al (2006) - snapCGH package

Non-HMM segmentation approaches

- Venkatraman and Olshen (2004) - DNAcopy package
- Hupé et al (2004) - GLAD package

Likelihood-based approaches - penalization criteria

- Picard et al (2005) - cghseg package

Change-point uncertainty (MCMC)

- Erdman et al (2008) - bcp package

Motivation

- Few exact non-MCMC methods for assessing uncertainty of change-point estimates
- Methods for finding exact posterior probabilities of change-points: $O(n^2)$ complexity
  - frequentist - Guédon (2007)
  - Bayesian - Rigaill (2011)
- High-resolution data in genomics technologies ($> 10,000$ observations per chromosome):
  - Smaller inter-segmental differences: characterize uncertainty
  - More data: need efficient estimates $O(n^2)$ not feasible
  - Next-generation sequencing: need methods adaptable to non-normal data

Segmentation approach to change-point detection

- Dataset: $X = (X_1, X_2, \ldots, X_n)$: real-valued observations.
- Hidden state space: $S = (S_1, S_2, \ldots, S_n)$: corresponding segment indices.
- Distribution: $\mathbb{P}(X_i | S_i = k, \theta_k) \sim g_{\theta_k}(\cdot)$: X_i belongs to segment k.
- Problem of interest: Find $\mathbb{P}(S_i | X; \theta) = ?$, when segments unknown given data

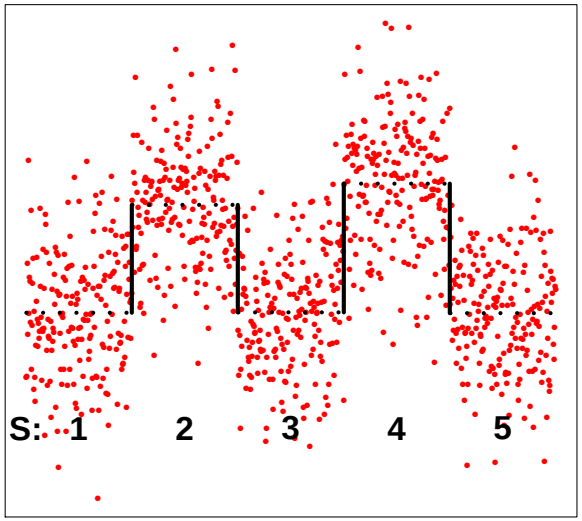


Figure: Segment-based change-point detection (K=5)

Constrained hidden Markov model for segmentation

Use of HMM algorithms to estimate posterior probabilities with **linear** complexity

- S: Markov chain over $\{1, 2, \ldots, K, K+1\}$, $\mathcal{M}_K$: set of possible S
- $\{S \in \mathcal{M}_K\}$: K states in n observations

Constraints on HMM correspond *exactly* to a segmentation change-point model.

- Find best partitioning $S \in \mathcal{M}_K$ into K non-overlapping intervals, distribution homogeneous within each segment
- $S_1 = 1, S_n = K$, junk state: $K+1$
- Allow for transitions of only 0 or $+1$, $S_i - S_{i-1} \in \{0, 1\}$.
  - $\mathbb{P}(S_i = k+1 | S_{i-1} = k) = \eta_k(i)$
  - $\mathbb{P}(S_i = k | S_{i-1} = k) = 1 - \eta_k(i)$

Adapted forward-backward algorithm

Forward and backward quantities, for observation i and state k:

$$F_i(k) = \mathbb{P}(X_{1:i} = x_{1:i}, S_i = k)$$
$$B_i(k) = \mathbb{P}(X_{i+1:n} = x_{i+1:n}, S_n = K | S_i = k)$$

Initialization:

$$F_1(1) = g_{\theta_1}(x_1)$$
$$B_1(K-1) = \eta_K(x_n) g_{\theta_k}(x_n), B_1(K) = (1 - \eta_K(x_n)) g_{\theta_k}(x_n)$$

Recursion:

$$F_i(k) = \left[F_{i-1}(k)(1 - \eta_k(i)) + \mathbf{1}_{k>1} F_{i-1}(k-1)\eta_k(i)\right] g_{\theta_k}(x_i)$$
$$B_{i-1}(k) = (1 - \eta_k(i)) g_{\theta_k}(x_i) B_i(k) + \mathbf{1}_{k<K} \eta_{k+1}(i) g_{\theta_{k+1}}(x_i) B_i(k+1)$$

Posterior probabilities from forward-backward algorithm

Posterior probability of state k for observation i

$$\mathbb{P}(S_i = k | X_{1:n} = x_{1:n}) = \frac{F_i(k) B_i(k)}{F_1(1) B_1(1)}.$$

Posterior probability of obs i being the k^{th} change-point

$$\begin{aligned}\mathbb{P}(CP_k = i | X_{1:n} = x_{1:n}) &= \mathbb{P}(S_i = k, S_{i+1} = k+1 | X_{1:n} = x_{1:n}) \\ &= \frac{F_i(k) \eta_k(i) g_{\theta_{k+1}}(x_{k+1}) B_{i+1}(k+1)}{F_1(1) B_1(1)}\end{aligned}$$

Posterior transition probability from $k-1^{th}$ to k^{th} state

$$\mathbb{P}(S_i = k | S_{i-1} = k-1, X_{1:n} = x_{1:n}) = \frac{\eta_{k-1}(i-1) g_{\theta_k}(x_i) B_i(k)}{B_{i-1}(k-1)}.$$

R package: postCP

`http://cran.r-project.org/web/packages/postCP`

- Can be adapted to non-normal parametric distributions for data

Output includes:

- Confidence intervals around each change-point estimate
- Posterior probabilities of hidden state and change-point for each observation
- Obtain *a posteriori* most probable set of change-points (Viterbi algorithm)
- Sampling from original data set by generating random sets of change-points

R package: postCP - sample input, output

```
>postCP(data=LRR.PLP[chrom==10],seg=initseg,model=2,ci=0.90)
$cp.est
        est lo.0.9 hi.0.9
 [1,]   211    211    211
 [2,]   215    215    215
 [3,]   273    271    273
 [4,]   383    382    384
 [5,]   736    695    755
 [6,]  3091   3090   3091
 [7,]  3102   3101   3102
 [8,]  8308   8286   8417
 [9,]  8760   8703   8780
[10,] 12383  11931  12452
$bestcp
[1]  211  215  273  383  721 3091 3102 8308 8760 11943
```

Analysis of colorectal cancer, SNP array data

- $n = 14,241$ log-reference ratio (LRR) observations
- Used cbs algorithm in DNAcopy (Olshen), which found 10 change-points
- postCP took < 0.1 sec to estimate change-point probabilities

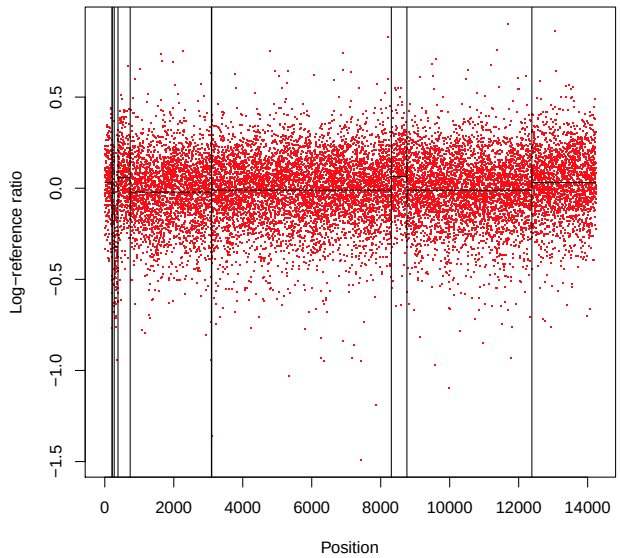
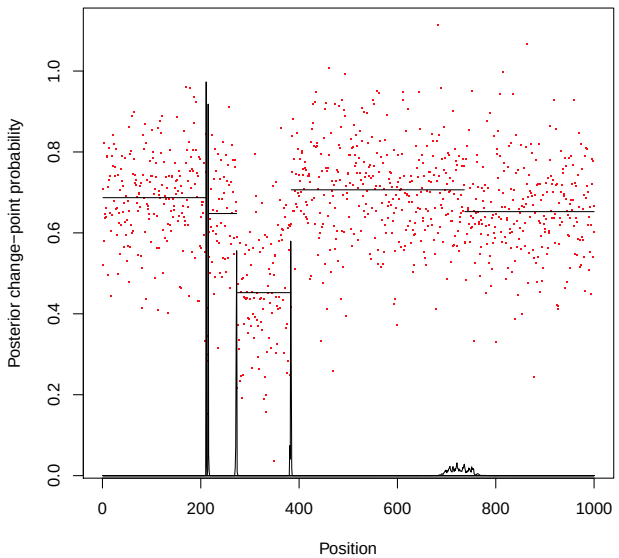


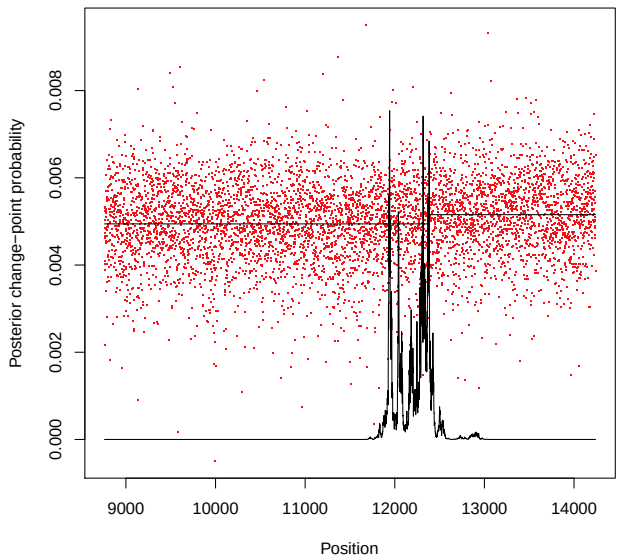
Figure: SNP array data with 11 segments, from Dr. Pierre Laurent-Puig, INSERM S775, Paris Descartes

Posterior-change point probabilities in SNP array data

CP	Est	post Prob	Δ Mean	width 0.9 CI
1	211	0.973	-0.582	1
2	215	0.918	0.523	1
3	273	0.556	-0.293	3
4	383	0.580	0.381	3
5	736	0.028	-0.081	61

CP	Est	Post Prob	Δ Mean	width 0.9 CI
10	12383	0.006	0.042	522





Change-point location estimates for Snijders breast cancer aCGH data (2001)

- $n = 120$ log-reference ratio (LRR) observations
- Initial change-points from modified greedy K-means algorithm.
- Less conservative intervals found by postCP
  - Frequentist: fixed parameters

Comparison vs Bayesian (Rigaill, 2011)

CP	Δ	est	postCP 95%CI	Bayes 95%CI
		Three segments		
1	-0.22	68	66-76	64-78
2	-0.71	96	96-96	92-97
		Four segments		
1	-0.34	68	66-76	66-78
2	-0.20	80	79-85	78-97
3	-0.80	96	96-96	91-112

Simulation - comparison of mean square error in posterior mean estimates of normally distributed data

Alternating means between θ_0 and θ_1. MSE=$mean((\hat{\theta} - \theta_{true})^2)$

		Mean square error		
θ_0	θ_1	cbs	cbs+postCP	bcp
		n=500, K=7		
1.0	1.50	0.058	0.055	0.045
	2.00	0.068	0.055	0.052
	2.50	0.050	0.039	0.047
	3.00	0.047	0.037	0.043
	3.50	0.042	0.034	0.039
		n=10000, K=40		
1.0	1.50	0.021	0.018	0.015
	2.00	0.018	0.014	0.015
	2.50	0.017	0.013	0.014
	3.00	0.015	0.012	0.014
	3.50	0.014	0.011	0.017

cbs: Venkatraman (2007), cbs+postCP, bcp: Erdman (2008)

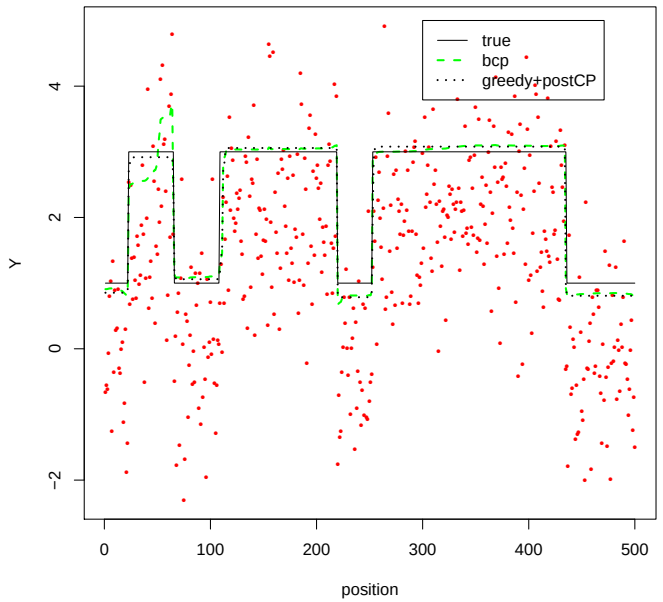


Figure: Posterior mean estimates

Summary

- Combine with effective method which obtains initial estimates of distribution of change-points (e.g. cbs, greedy algorithms)
  - Less conservative confidence intervals than those from exact formulae (Rigaill, 2011), postCP uses frequentist framework
  - With larger intersegmental differences, comparable loss to Bayesian methods (Erdman, 2008)
- Estimates of change-point probabilities in linear time $O(Kn)$
  - 10 change-points in $>$ 14000 SNPs: $<$ 0.1 second,$\sim$ 100 change-points in 200000 observations: $\sim$ 10 seconds
- Methods are easily adapted to non-normal data, such as those from next-generation sequencing (negative Binomial)

Further applications

- Posterior probabilities and means may be used to calculate model selection criteria
- Constrained HMM may be adapted to alternate change-point models
- Sampling methods can account for parameter uncertainty by Sequential Monte Carlo (SMC) methods
- Can use posterior estimates to detect simultaneous change-points across multiple samples
- Segment multiple outcomes at same time (LRR and BAF in CNV)


Erdman C, Emerson J (2008) *A fast bayesian change point analysis for the segmentation of microarray data.* Bioinformatics 24(19):2143–2148.


Guédon Y (2007) *Exploring the state sequence space for hidden Markov and semi- Markov chains.* Computational Statistics & Data Analysis 51(5):23792409


Hupé, Stransky N, Thiery J, Radvanyi F, Barillot E (2004) *Analysis of array CGH data: from signal ratio to gain and loss of DNA regions.* Bioinformatics 20(18):3413–3422.


Marioni J, Thorne N, Tavaré S (2006) *BioHMM: a heterogeneous hidden Markov model for segmenting array CGH data.* Bioinformatics 22(9):1144–1146.


Picard F, Robin S, Lavielle M, Vaisse C, Daudin J (2005) *A statistical approach for array CGH data analysis.* BMC Bioinformatics 6(1):27.


Rabiner L (1989) *A tutorial on hidden Markov models and selected applications in speech recognition.* In: Proceedings of the IEEE, IEEE, vol 77, pp 257–286.


Rigaill G, Lebarbier E, Robin S (2011) *Exact posterior distributions and model selection criteria for multiple change-point detection problems.* Statistics and Computing pp 113


Snijders A, Nowak N, Segraves R, Blackwood S, Brown N, Conroy J, Hamilton G, Hindle A, Huey B, Kimura K, et al (2001) *Assembly of microarrays for genome-wide measurement of DNA copy number by CGH.* Nature Genetics 29:263–264.


Venkatraman E, Olshen A (2007) *A faster circular binary segmentation algorithm for the analysis of array CGH data.* Bioinformatics 23(6):657–663


Willenbrock H, Fridlyand J (2005) *A comparison study: applying segmentation to array CGH data for downstream analyses.* Bioinformatics 21(22):4084–4091.