Fast estimation of posterior change-point probabilities for CNV data

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

March 27, 2012

Introduction

- Change-point methods: applications in economics, engineering, bioinformatics
- Common application: copy number variation (CNV), identifies regions of DNA with gain or deletion that may be related to disease susceptibility
- High-resolution data, 10's thousands of clones per chromosome
  - Array comparative genomic hybridization (aCGH)
  - Single nucleotide polymorphism (SNP) array

Example: array CGH data with copy number variations

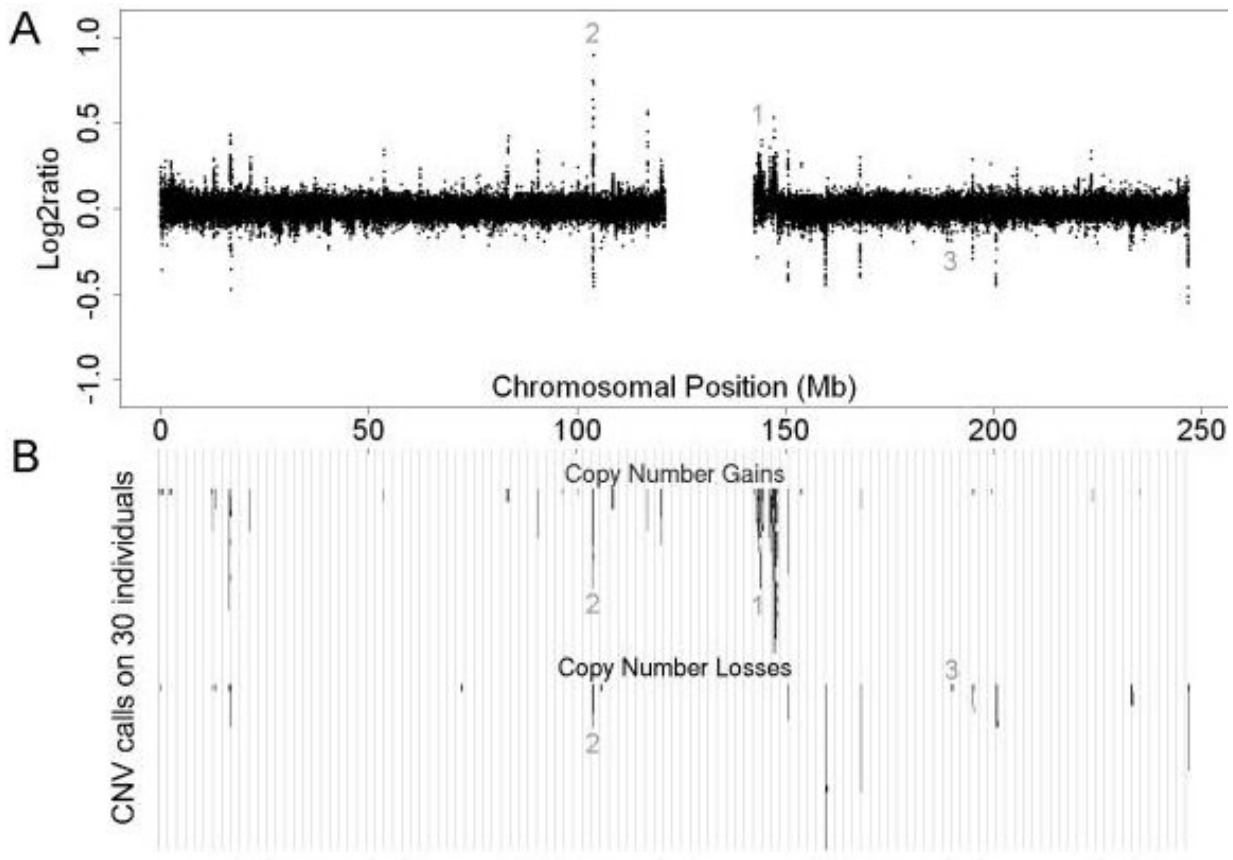

Figure: A: array CGH profile, B: identified DNA copy gains and losses, source: Redon and Carter, Methods Mol Biol. 2009; 529: 3749.

Finding a best segmentation for CNV

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

Estimate number of segments in aCGH data

- Picard et al (2005)
- Zhang and Siegmund (2007)

Posterior probabilities of change-point location

Methods to find posterior uncertainty of estimated change-point locations

- Asymptotic estimates: Bai (2003), Muggeo (2003)
- Bootstrapping methods: Hušková and Kirch (2008)
- Stochastic methods (Lai et al. 2008).
- Exact posterior distribution: Guédon (2008), Rigaill et al. (2011)

Motivation

- Relatively few algorithms for assessing uncertainty of change-point estimates
- Methods for finding posterior probabilities of change-points: $O(n^2)$ complexity
- Given high-resolution data in genomics technologies ($> 10,000$) observations per chromosome:
  - Smaller inter-segmental differences: need to characterize uncertainty
  - More data: need efficient estimates $O(n^2)$ not feasible

Segmentation approach to change-point model

- Dataset: $X = (X_1, X_2, \ldots, X_n)$ of real-valued observations
  - Segment indices: $S = (S_1, S_2, \ldots, S_n)$.
- Find best partitioning $S \in \mathcal{M}_K$ of the data into $K \geq 2$ non-overlapping intervals, assuming distribution is homogeneous within each interval.

$$\mathbb{P}(X|S;\theta) = \prod_{i=1}^{n} g\left(X_i;\theta_{S_i}\right) = \prod_{k=1}^{K} \prod_{i,S_i=k} g\left(X_i;\theta_k\right) \qquad (1)$$

  - $G(\cdot;\theta_k)$ is the parametric distribution (typically: Poisson or Gaussian) with parameter θ_k, $\theta = (\theta_1, \ldots, \theta_K)$ is the global parameter.
- $\mathbb{P}(S) = \mathbb{P}(S_1 = s_1) \prod_{i=2}^{n} \mathbb{P}(S_i = s_i | S_{i-1} = s_{i-1})$

Constrained hidden Markov model (HMM) model

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

- Permits use of HMM algorithms to estimate posterior probabilities with linear complexity

S: heterogeneous Markov chain over $\{1, 2, \ldots, K, K+1\}$

- $\{S \in \mathcal{M}_K\}$: K states in n observations
- $S_1 = 1, S_n = K$, junk state: $K+1$

Transitions: for all $2 \leqslant i \leqslant n$, and $1 \leqslant k \leqslant K$:

- $\mathbb{P}(S_i = k+1 | S_{i-1} = k) = \eta_k(i)$
- $\mathbb{P}(S_i = k | S_{i-1} = k) = 1 - \eta_k(i)$
- Allows for transitions of only 0 or $+1$, $S_i - S_{i-1} \in \{0, 1\}$

HMM example

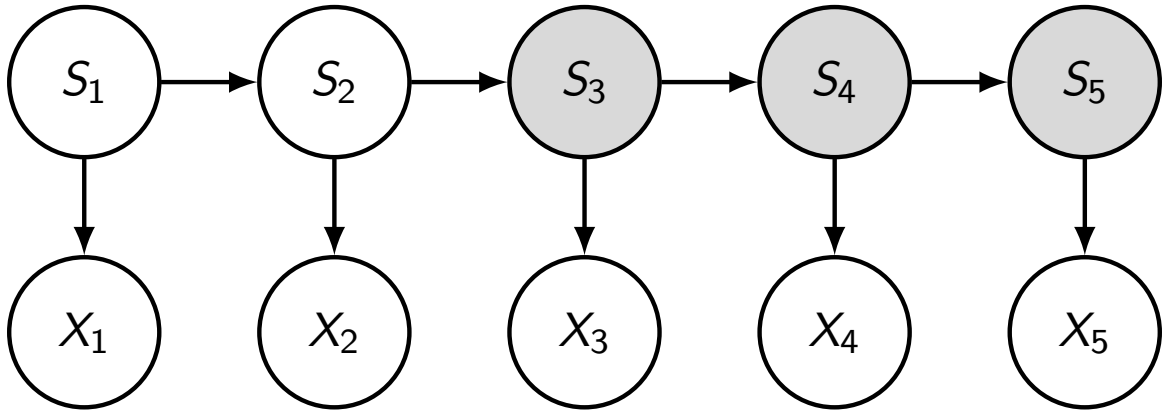

Figure: HMM topology. For $i = 1 \ldots 5$, S_i are the hidden states, and X_i are the observed states. White circles: $S_i = 1$, grey circles: $S_i = 2$

Example: if $n = 5$, $K = 2$, with change after $i = 2$ then:

- $S = (1, 1, 2, 2, 2)$
- $\mathbb{P}(S) = (1 - \eta_1(2))\eta_2(3)(1 - \eta_2(4))(1 - \eta_2(5))$.

Homogeneous constrained HMM

- Markov chain is homogeneous if $\eta_k(i) = \eta \in]0, 1[$ for all k,i
- Homogeneous HMM results in: $\mathbb{P}(S|S \in \mathcal{M}_K) = 1/|\mathcal{M}_K|$
  - constant and independent of S
  - $S = (1, 1, 2, 2, 2)$
  - For $\eta = 0.5$: $\mathbb{P}(S) = \eta(1-\eta)^3 = 0.5^4$.
- Can specify different $\eta_k(i)$ for heterogeneous HMM.

Forward-backward algorithm

Forward and backward quantities, for observation i and state k:
For $1 \leqslant i \leqslant n-1$:

$$F_i(k) = \mathbb{P}(X_{1:i} = x_{1:i}, S_i = k) \tag{2}$$
$$B_i(k) = \mathbb{P}(X_{i+1:n} = x_{i+1:n}, S_n = K | S_i = k) \tag{3}$$

Forward recursion:

$$F_1(k) = \begin{cases} G_{\theta_1}(x_1) & \text{if } k = 1 \\ 0 & \text{else} \end{cases} \tag{4}$$
$$F_i(k) = \left[F_{i-1}(k)(1 - \eta_k(i)) + \mathbf{1}_{k>1} F_{i-1}(k-1)\eta_k(i)\right] G_{\theta_k}(x_i) \tag{5}$$

Emission distribution of observed data $G_{\theta_k}(x_i)$

Forward-backward algorithm (cont)

Backward recursion:

$$B_{n-1}(k) = \begin{cases} G_{\theta_k}(x_n)\eta_K(x_n) & \text{if } k = K-1 \\ G_{\theta_k}(x_n)(1-\eta_K(x_n)) & \text{if } k = K \\ 0 & \text{else} \end{cases} \quad (6)$$

$$B_{i-1}(k) = (1-\eta_k(i))G_{\theta_k}(x_i)B_i(k) + \mathbf{1}_{k<K}\eta_{k+1}(i)G_{\theta_{k+1}}(x_i)B_i(k+1) \quad (7)$$

Posterior probability of state k for observation i

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

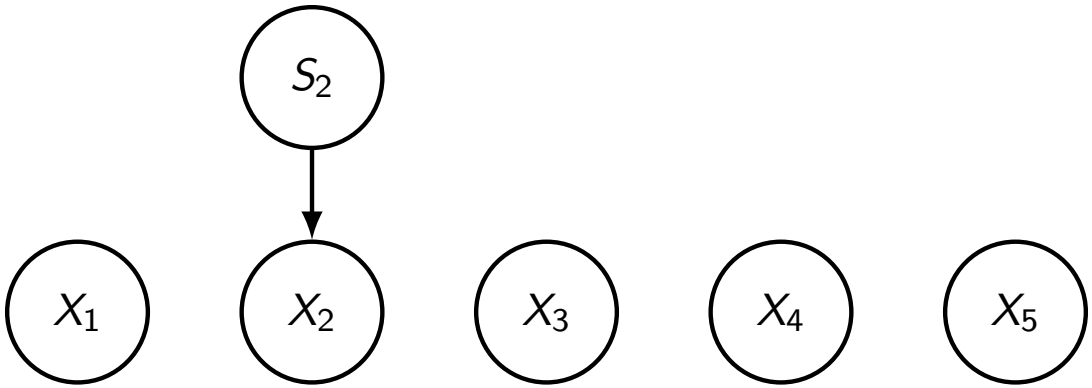

Figure: Posterior probability of observation 2 being in state 1, $\mathbb{P}(S_2 = 1 | X_{1:n} = x_{1:n})$

Posterior probability of k^{th} change-point occurring after observation i

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

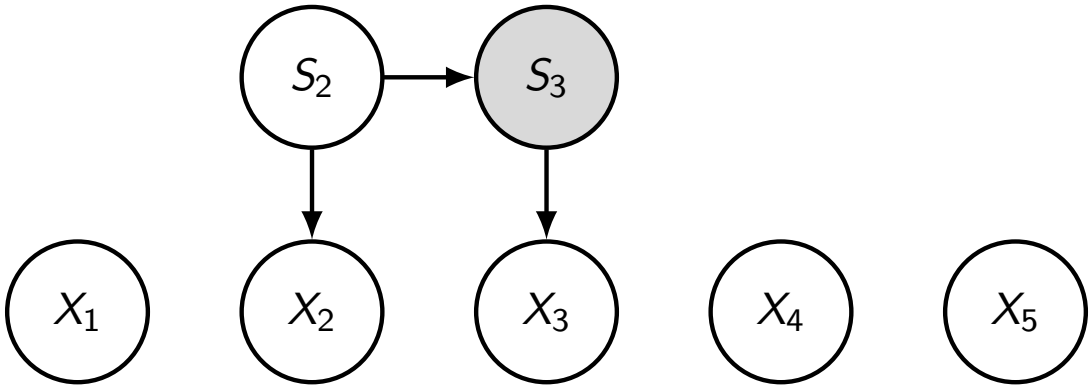

Figure: Posterior probability of change from state 1 to state 2 after observation 2, $\mathbb{P}(S_2 = 1, S_3 = 2 | X_{1:n} = x_{1:n})$

Best set of change-points (Viterbi algorithm)

Recursion (modified forward quantities):

$$
\begin{aligned}
V_1(1) &= G_{\theta_1}(x_1) \\
V_i(1) &= V_{i-1}(1)(1-\eta_k(x_i))G_{\theta_1}(x_i) && \text{if } i \geq 2 \\
V_i(k) &= \max\left\{V_{i-1}(k-1)\eta_k(x_i), V_{i-1}(k)(1-\eta_k(x_i))\right\} G_{\theta_k}(x_i) && \text{if } i,k \geq 2
\end{aligned}
$$

Obtain set of change-points with highest posterior probability by using path of indices k used to calculate $V_{i,k}$:

- $K-1^{th}$ change-point CP_{K-1}: largest i in $V_i(K-1)$ used to calculate $V_{i+1}(K)$
- k^{th} change-point CP_k: largest i in $V_i(k)$ used to calculate $V_{i+1}(k+1)$ (where $CP_k < CP_{k+1}$)

R package: postCP

Output includes:

- Confidence intervals around each initial change-point (either specified by user or found by Viterbi algorithm)
- Posterior probabilities of hidden state and change-point for each observation
- (Optional) Sampling from original data set by generating random sets of change-points

Analysis of colorectal cancer, SNP array data

- Used DNAcopy (Olshen), which found 10 change-points within $14,241$ observations
- postCP took < 0.1 sec to estimate change-point probabilities

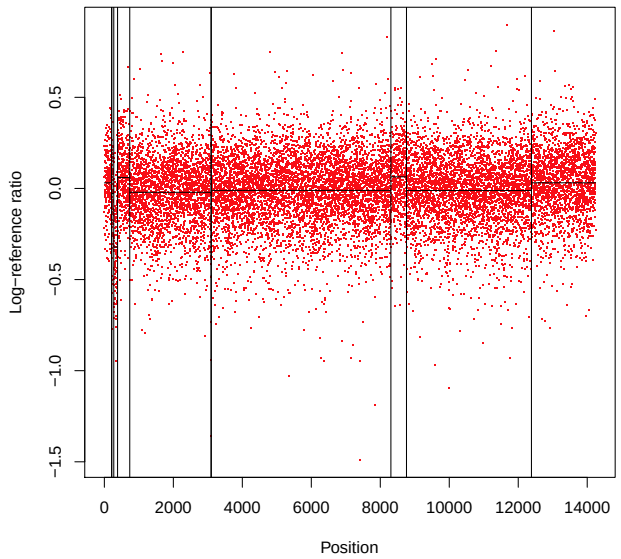

Figure: SNP array data with 11 segments

R package: postCP

```
>postCP(data=LRR.PLP[chrom==10],seg=initseg,model=2,
   ci=0.95)$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
```

Posterior-change point probabilities for first 5 change-points in SNP array data, $n = 14,241$, (Laurent-Puig) by postCP

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

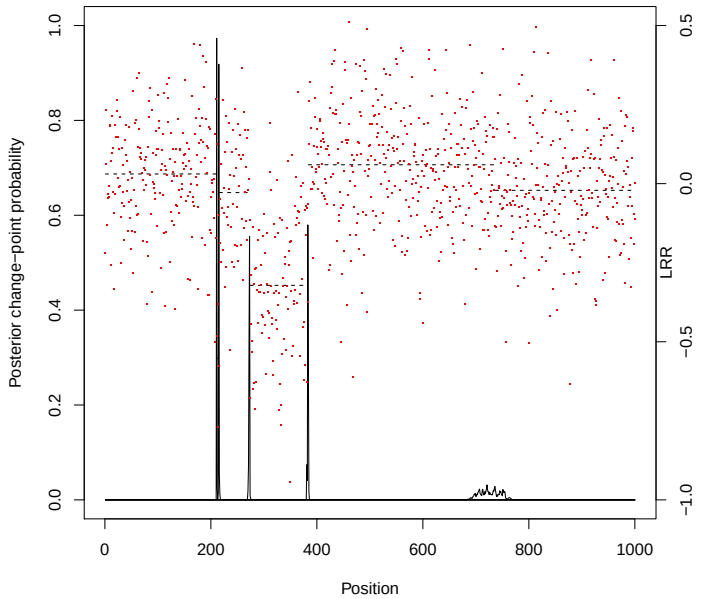

Figure: postCP: first 5 change-points

Posterior-change point probabilities for 10th change-point in SNP array data, $n = 14,241$, (Laurent-Puig) by postCP

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

- Irregular nature of posterior change-point probability from discreteness of locations.

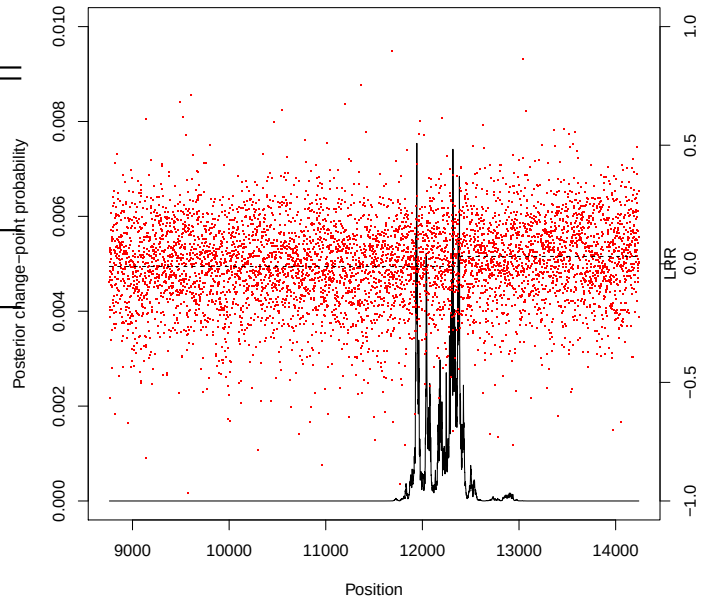

Figure: postCP: 10th change-point

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

- Initial change-points from modified greedy K-means algorithm.
- Less conservative intervals found by postCP
  - Frequentist approach: fixed parameter values

CP	Δ	est	postCP	Bayes
		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

Posterior-change point probabilities for aCGH data, $n = 120$, (Snijders et al, 2001) by postCP

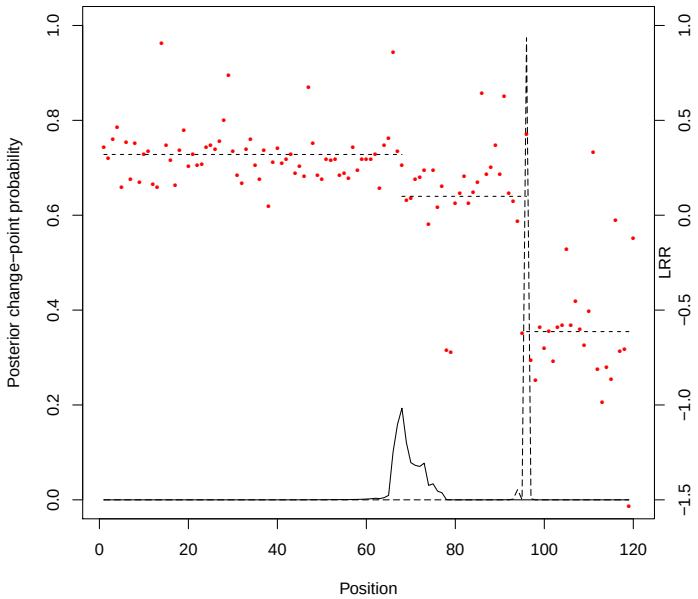

Figure: postCP: 3 segments

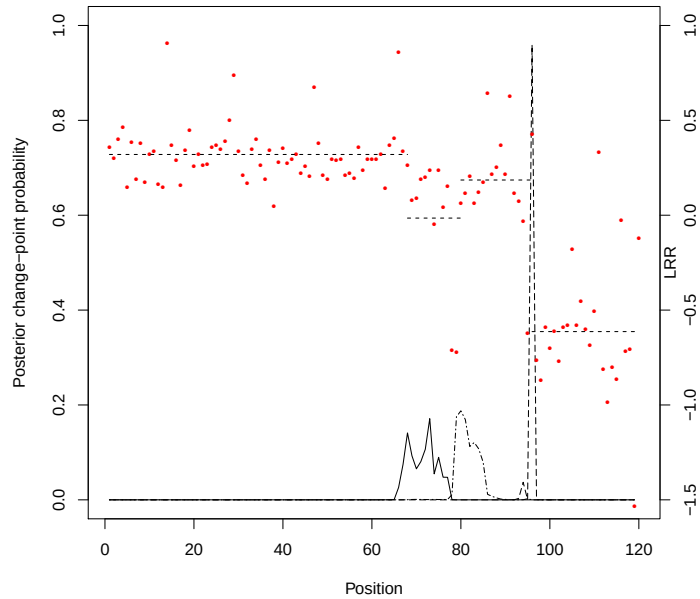

Figure: postCP: 4 segments

Misclassified initial change-points, normal data, one true change-point at $i = 500$

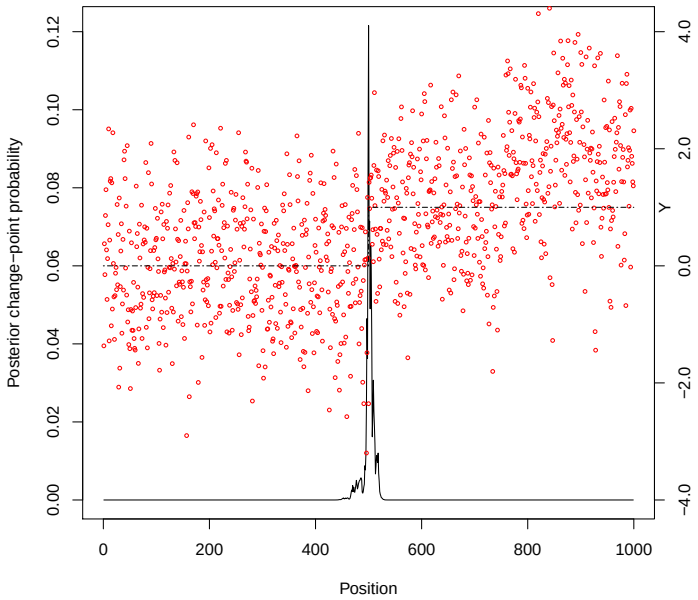

Figure: Misplaced initial change-point $i = 100$

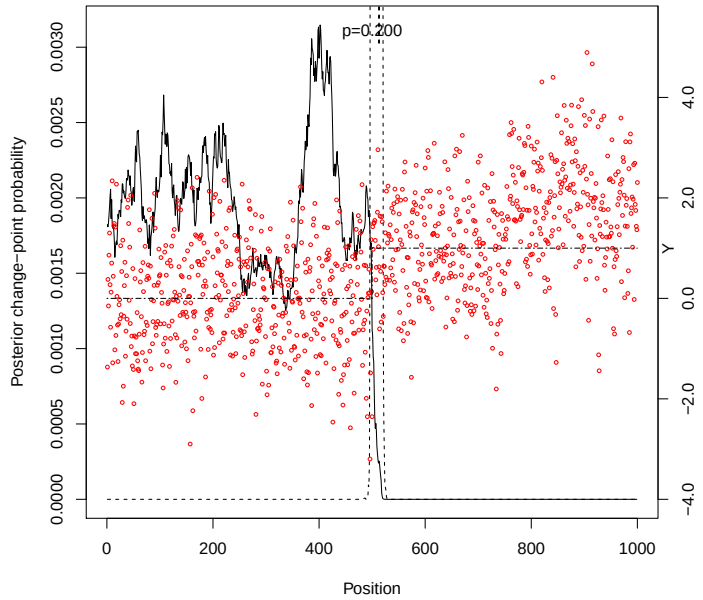

Figure: Extra initial change-point $i = 100$ and $i = 500$

Summary

- Estimates of change-point probabilities in linear time $O(Kn)$
- Calculations for 10 change-points in > 14000 SNPs took < 0.1 second, took ~ 10 seconds for ~ 100 change-points 200000 observations
- Less conservative confidence intervals than those from exact formulae (Rigaill, 2011), postCP uses frequentist framework
- Probability estimates may be inaccurate if change-point locations misspecified

Practical applications

- Useful when combined with effective method to obtain initial estimates of distribution of change-points
- Efficient calculations, feasible for high-throughput data such as CNV data from SNP arrays
- Overlapping confidence intervals across multiple samples may yield useful information

Future work

- Combination in *R* package with dynamic programming algorithm for detecting change-points from Rozenholc (2011)
- Model selection through criteria (BIC, ICL) using posterior probabilities from forward-backward algorithm
  - ICL uses entropy: $\sum_S \mathbb{P}(S|X, K, \hat{\theta}_K) \log \mathbb{P}(S|X, K, \hat{\theta}_K)$
- Specification of priors, E-M algorithm
- Segmentation of multiple outcomes (LRR and BAF in CNV)