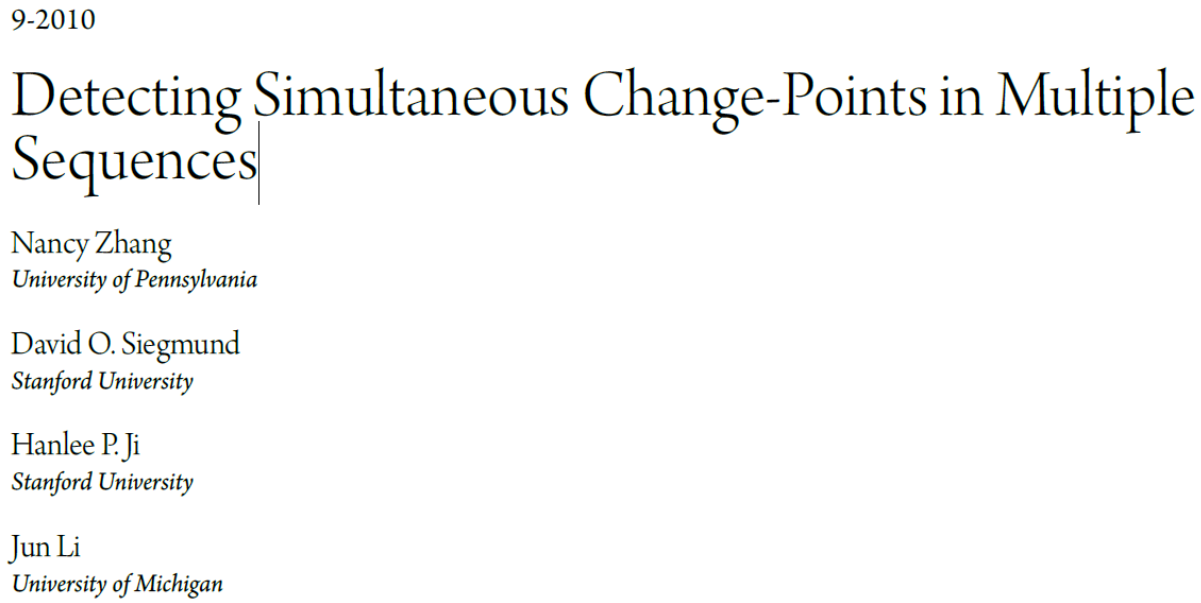
0. 論文



9-2010

Detecting Simultaneous Change-Points in Multiple Sequences

Nancy Zhang
University of Pennsylvania

David O. Siegmund
Stanford University

Hanlee P. Ji
Stanford University

Jun Li
University of Michigan

タイトル：Detecting Simultaneous Change-Points in Multiple Sequences
著者：

arXiv投稿日：
学会/ジャーナル：

1. どんなもの？

- 複数の1次元シーケンスに共通する変化点の検出を行う
  - シーケンスのほんの一部にしか起きないような変化も考慮する
  - シーケンス間のデータを結合する統計量の提案
  - 問題設定としてガウス分布の平均における一時的なシフトを変化とする
    - 各サンプルはノイズが独立して同一分布のガウス分布であることの仮定を置いている(σ_i^2を持つ)
  - DNAコピー数への適用を考えている

2. 先行研究

- アプリケーションとして一度の分析に1つのサンプルしか使えないことが多い
  - circular binary segmentationと呼ばれる手法が良く機能するらしいい
    - Olshen et al(2004), Venkatraman & Olshen(2007)
- 系列の選択はこの研究分野ではクロスサンプル解析と呼ばれる
  - 基本的にセグメンテーション後にこのような選択が行われる
  - Diskin et al. (2006), Wang et al. (2008), and Newton et al. (1998),Newton & Lee (2000)
  - 階層的隠れマルコフモデルもあるが，使用的な仮定が多すぎる気がする(異常区間の発生，異常の持続時間，振幅など)
- 全ゲノムでの分析もある
  - 複数ゲノムスキャンの同時解析のための統計的基準の提案もある
-

3. コアアイディア

- 問題設定
  - y_{it}のようにi番目の系列のt番目の要素として観測されるとする
  - 帰無仮説？はすべての系列に変化がない，対立仮説はある部分系列集合$\mathcal{J}$のみに$\tau_1<t<\tau_2$のtに変化(平均が変化)するものとする
- 以下の値を最大にする始点sと終点tの探索を行う

$$Z^{(p)}(s,t) = \left[\sum_{i=1}^{N} w_p\{U_i(s,t)\}U_i^2(s,t) - N\mu_p\right] \Big/ \sigma_p N^{1/2}.$$

  -
  - $w_p(x) = \exp(x^2/2)/{r_p + \exp(x^2/2)}$
  - $U_i(s,t) = \hat{\sigma}^{-1}{S_t-S_s-(t-s)\bar{y}_T}/[(t-s){1-(t-s)/T}]^{1/2}$
  - S_t: tまでの平均
  - $\bar{y}_t = S_t/t$
  - $\hat{\sigma}^2$: 標本分散
  - $\mu_p = \mathrm{E}[w_p(U)U^2]$
  - $\sigma_p = \mathrm{var}(w_p(U)U^2]$
  - この式を標準化したものを考えている

$$\max_{s,t} \sum_{i=1}^{N} w_p[U_i(s,t)]U_i^2(s,t),$$

  -
- アルゴリズム

Algorithm 2.2. *Fix the global significance level α, parameter p, and a maximum window $T_0 < T$. We denote by $\boldsymbol{y}_{h:k}$ the matrix $\{y_{i,t} : \ 1 \le i \le N, \ \ h \le t \le k\}$.*

1. *Initialize $T_1 = 1$ and $T_2 = T$.*

---

2. *Compute*

$$Z_{max} = \max_{\substack{T_1 \le s < t \le T_2 \\ 1 \le t-s \le T_0}} \{Z^{(p)}(s,t)\}.$$

Let (s^*, t^*) be the maximizing interval.

3. *If the p-value of Z_{max}, as computed using the approximations in Section 2.4, is less than α, then for each $(u,v) \in \{(T_1, s^*-1), \ (s^*, t^*), \ (t^*+1, T_2)\}$, do:*

   (a) *Determine which samples carry the variation, as described below. If a sample carries the variation, let $\hat{\boldsymbol{y}}_{i,u:v} = \bar{\boldsymbol{y}}_{i,u:v}$, and for the other samples let $\hat{\boldsymbol{y}}_{i,u:v} = \bar{\boldsymbol{y}}_{i,T_1:T_2}$. Let $\boldsymbol{y}'_{u:v} = \boldsymbol{y}_{u:v} - \hat{\boldsymbol{y}}_{u:v}$.*

   (b) *Repeat steps 2-3 for $T_1 = u$, $T_2 = v$ and the newly normalized $\boldsymbol{y}'_{u:v}$.*

-
- 結局はバイナリセグメンテーション的な手法を取る
- 系列の選択はどの部分でできるのかが分からない
  - ヒューリスティックにやっている
  - 全てのサンプルに対して閾値処理

4. どうやって有効だと検証した？

- 人工データ実験
  - (おそらく)FPRの実験結果
    - bというパラメータを大きくすることでp値を小さくすることができる(なにが言いたいのかが分からない)
    - モンテカルロ近似と同様の結果が出せることを示したい
    - パラメータbによってp値の制御を行っているとみて良い
  - 検出力の実験
    - より大きなr_pの値を使うことによって検出力が上がることが期待できる
- 実データ実験

- CVNデータセット
  - これは良く研究されているもの
  - de novo検出に焦点を当てる
  - 単一の突然変異イベントに起因するもので，サンプル間で共有される
  - 大体1コピーの変化のみある
- このデータでは関係ないがアルゴリズム2.2は疎なクロスサンプルサマリーを得られるのに有用
  - これは系列の選択を行っているという認識でよい？
  - この選択を行うために3つの手法が考えられる
  - １．各サンプルでセグメント化して閾値異常の領域は変化点としてみなす
  - ２．各サンプルを個別に分割．サンプルを整列させて順列法を用いて選択する
  - ３．HMMベースの手法．すべてのサンプルが同時に変化する仮定を置いている．手法によっては全てのサンプルで同じ平均値を持つ必要がある
- 22番染色体に複雑な欠損があることが議論されている(Iafrate et al. 2004)
  - 全てのサンプルに変化があるわけでないのでアルゴリズム2.2を利用する
  - 適用した結果3つのコピー数レベル?を持つことが分かった

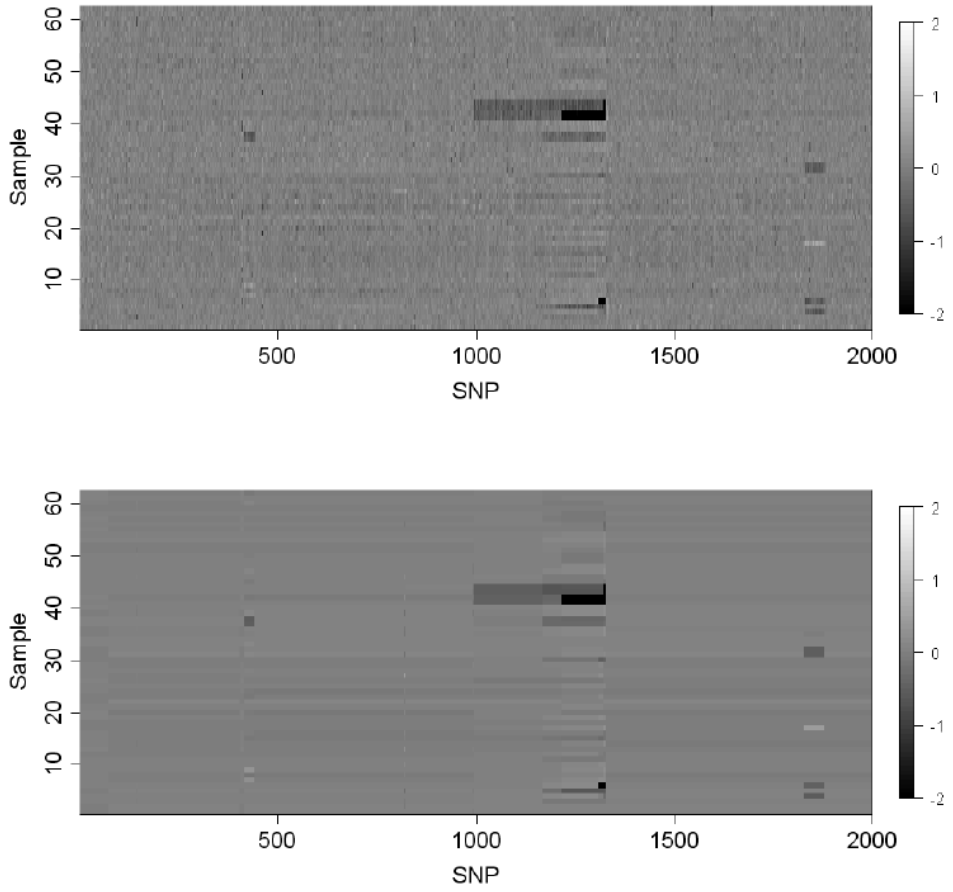


Figure 9. Example 2000 SNP region in cytoband 22q11 containing a complex CNV with nested deletions. Bottom panel shows segmentation given by Algorithm 2.2.

5. データセット

- CNV dataset (Fanciulli et al. 2007, Perry et al., 2007 etc)

6. 疑問点

- 得られた結果をどのように解釈すればよい？

    - 変化点検出というより異常領域の特定の重きを置いている気がする
    - 変化点検出に読み替えれない訳ではないが...
    - 実データ実験は遺伝子の研究をしている人じゃないと分かりにくい気がする

7. 次に読むべき論文は？

- CBS論文
    - Circular binary segmentation for the analysis of array-based dna copy number data.

キーワード

- scan統計量
- 変化点検出
- meta-analysis