Change-Point Detection and Copy Number Variation

Benjamin Yakir

Department of Statistics
The Hebrew University

IMS, Singapore, 2009

Outline

Classical change-point detection

- At each monitoring period we observe a single observation.
- The distribution may shift over an unknown interval of time.

Off-line monitoring

- The entire sequence is given.
- Goals: testing for a change, estimating the change.

On-line monitoring

- The sequence is observed sequentially.
- Goals: quickest detection, avoiding false stops.

Example: SPRT

The problem

- $X_i \sim N(\mu, 1)$, $i = 1, 2, \ldots$.
- Test $H_0 : \mu = 0$ versus $H_1 : \mu = \mu_1$.
- Procedure $=$ a stopping time N.
- Reject $H_0 \Leftrightarrow \{N < \infty\}$.

Sequential Probability Ratio Test

- Compute $\ell(n) = \mu_1 \sum_{i=1}^{n} X_i - n\mu_1^2/2$, $n = 1, 2, \ldots$.
- $N_A = \inf\{n : \ell(n) \geq \log A\}$.

Significance level approximation

A sequential method

- Likelihood ratio transformation:

$$\mathbb{P}(N_A < \infty) = \mathbb{E}_1\left[e^{-\ell(N_A)}; N_A < \infty\right]$$
$$\approx (1/A)\mathbb{E}_1\left[e^{-(\ell(N_A)-\log A)}\right] .$$

- Overshoot: $\ell(N_A) - \log A \to U$, in distribution.
- Laplace transform: $\mathbb{E}_1[e^{-U}] = (\mu_1^2/2)\nu(\mu_1)$.
- The approximation: $\mathbb{P}(N_A < \infty) \approx (\mu_1^2/2)\nu(\mu_1)/A$.

Significance level approximation

A change-point method

- $\{N_A \le m\} = \{\max_{n \le m} \ell(n) \ge \log A\}$.
- Log-likelihood ratio: $\ell(n) \Leftrightarrow \mathbb{P}_n$.
- Likelihood ratio transformation:

$$\mathbb{P}(N_A \le m) = \frac{1}{A} \sum_{n=1}^{m} \mathbb{E}_n \left[\frac{e^{\ell(n) - \log A}}{\sum_j e^{\ell(j) - \ell(n)}} ; \max_j \ell(j) - \ell(n) + \ell(n) \ge \log A \right]$$

- Localization: $\approx \mathbb{E}_n \left[\frac{\max_j e^{\ell(j) - \ell(n)}}{\sum_j e^{\ell(j) - \ell(n)}} \right] \times \mathbb{P}_n(\ell(n) = \log A)$.
- The approximation: $\mathbb{P}(N_A < \infty) \approx \mathbb{E}[\mathcal{M}/\mathcal{S}]/A$.

The Change-point method

1. Express the event "Detection" as a maximum of likelihood ratios.
2. Transform the distribution: $\mathbb{P} \to \sum_n \mathbb{P}_n$.
3. Approximate each term in the sum:
   - Separate between the local effect and the global effect.
   - Evaluate the local limit of $\mathbb{E}[\mathcal{M}/\mathcal{S}]$.
4. Sum the approximations in order to obtain the final approximation.

Example: Scanning statistic in linkage

Affected sib-pairs

- Affected sib-pairs are collected and genotyped.
- The total identity by decent (IBD) in the sample is measured at each marker.
- One looks for loci with access IBD.

Large sample approximation

- The centered statistic Z_t is approximately standard normal.
- Haldane model of crossover: $\mathbb{C}\mathrm{ov}(Z_t, Z_s) = e^{-\beta|t-s|}$.
- Approximation $\mathbb{P}(\max_t Z_t \geq b) =?$

Example: Scanning statistic in linkage

Measure transformation

- $\mathbb{P}_t \Leftrightarrow \ell(t) = bZ_t - b^2/2$.
- Under $\mathbb{P}_t$: $Z_t \sim N(b, 1)$.
- The covariance structure is unchanged.

Approximation

- $\ell(s) - \ell(t) \sim N(-\beta b^2|s - t|, 2\beta b^2|s - t|)$.
- Asymptotically: A two-sided random walk with negative drift.
- $\mathbb{P}(\max_t Z_t \geq b) \approx \mathbb{E}[\mathcal{M}/\mathcal{S}] \times \#\{\text{markers}\} \times \phi(b)/b$.
- $\mathbb{E}[\mathcal{M}/\mathcal{S}] = \beta b^2\Delta \cdot \nu(b[2\beta\Delta]^{1/2})$.

Example: detecting DNA copy-number variation

DNA copy-number

- Deletions and insertions may produce variations in the copy-number.
- Zygotic copy-number variations are inheritable.

Detecting variation

- Microarray produce measurements on the DNA copy number at each loci. The single-to-noise ratio is low.
- Examine a sample in parallel $\Rightarrow$ increases the power of detection (Zhang et al., 2008).

The model

Hypothesis testing

- The sample: $i = 1, \dots n$.
- The loci: $j = 1, \dots J$.
- The observations: $X_{ij} \sim N(\mu_{ij}, 1)$.
- Test $H_0 : \mu_{ij} = 0$ versus $H_1 : \mu_{ij} = \mu_i \neq 0$,
  for some k and sub-interval $k - m \leq j \leq k + m$.

Scanning statistics

- $Z_i(k) = \sum_{j=k-m}^{k+m} X_{ij} / \sqrt{2m+1}$.
- $G_k = \sum_{i=1}^{n} g(Z_i(k))$.

Measure transformation

The significance level

- $\mathbb{P}(\max_k G_k \geq x) =?$.

The likelihood ratio

- Log-moment generating function: $\psi(\theta) = \log \mathbb{E} \exp\{\theta g(Z)\}$.
- Selecting θ: $\dot{\psi}(\theta) = x/n$.
- Log-Likelihood ratio: $\ell(k) = \theta G_k - n\psi(\theta)$.

Localization

The local process

$$\ell(l) - \ell(k) \approx \sum_{j=k+1}^{l} \hat{G}_j$$

where

$$\begin{aligned} \hat{G}_j &= \frac{\theta}{(2m+1)^{1/2}} \sum_{i=1}^{n} \dot{g}(Z_{i,k})[X_{i,j+m} - X_{i,j-m}] \\ &+ \frac{\theta/2}{2m+1} \sum_{i=1}^{n} \mathbb{E}_k^{\theta}\left[\ddot{g}(Z_{i,k})(X_{i,j+m} - X_{i,j-m})^2\right] . \end{aligned}$$

Approximation

The local term

$\mathbb{E}[\mathcal{M}/\mathcal{S}] = \mu(\theta)\nu\big([2\mu(\theta)]^{1/2}\big)$, where

$$\mu(\theta) = \frac{\theta^2 n}{2m+1} \int [\dot{g}(z)]^2 e^{\theta g(z) - \psi(\theta)} \phi(z) dz \; .$$

The probability

$$\mathbb{P}\Big(\max_{k \le J} G_k \ge x\Big) \sim J e^{-n\{\theta\dot{\psi}(\theta) - \psi(\theta)\}} \{2\pi n \theta^2 \ddot{\psi}(\theta)\}^{-1/2} \mathbb{E}[\mathcal{M}/\mathcal{S}] \; .$$

Bibliography

1. Nancy Zhang, David Siegmund, Hanlee Ji and Jun Li. (2008). Detecting simultaneous change-points in multiple sequences.
2. David Siegmund, Benjamin Yakir and Nancy Zhang. (2009). Tail approximations for maxima of random fields by likelihood ratio transformations.