Methodology Introduction

- NMCD Method (from a Statistical View)

2.1. *NMCD method.* Assume that $Z_1, \ldots, Z_n$ are independent and identically distributed from F_0, and let $\widehat{F}_n$ denote the empirical C.D.F. of the sample, then $n\widehat{F}_n(u) \sim \text{Binomial}(n, F_0(u))$. If we regard the sample as binary data with the probability of success $\widehat{F}_n(u)$, this leads to the nonparametric maximum log-likelihood

$$n\{\widehat{F}_n(u)\log(\widehat{F}_n(u)) + (1-\widehat{F}_n(u))\log(1-\widehat{F}_n(u))\}.$$

In the context of (1.1), we can write the joint log-likelihood for a candidate set of change-points $(\tau'_1 < \cdots < \tau'_L)$ as

$$\begin{aligned}\mathcal{L}_u(\tau'_1, \ldots, \tau'_L) = \sum_{k=0}^{L} (\tau'_{k+1} - \tau'_k)\{&\widehat{F}_{\tau'_k}^{\tau'_{k+1}}(u)\log(\widehat{F}_{\tau'_k}^{\tau'_{k+1}}(u)) \\ &+ (1-\widehat{F}_{\tau'_k}^{\tau'_{k+1}}(u))\log(1-\widehat{F}_{\tau'_k}^{\tau'_{k+1}}(u))\},\end{aligned} \tag{2.1}$$

where $\widehat{F}_{\tau'_k}^{\tau'_{k+1}}(u)$ is the empirical C.D.F. of the subsample $\{X_{\tau'_k}, \ldots, X_{\tau'_{k+1}-1}\}$ with $\tau'_0 = 1$ and $\tau'_{L+1} = n+1$. To estimate the change-points $1 < \tau'_1 < \cdots < \tau'_L \le n$, we can maximize (2.1) in an integrated form

$$R_n(\tau'_1, \ldots, \tau'_L) = \int_{-\infty}^{\infty} \mathcal{L}_u(\tau'_1, \ldots, \tau'_L)\, dw(u), \tag{2.2}$$

where $w(\cdot)$ is some positive weight function so that $R_n(\cdot)$ is finite, and the integral is used to combine all the information across u. The rationale of using (2.2) can be clearly seen from the behavior of its population counterpart. For simplicity, we assume that there exists only one change-point

τ_1, and let $\tau_1/n \to q_1 \in (0,1)$ and $\tau_1'/n \to \theta \in (0,1)$. Through differentiation with respect to θ, it can be verified that the limiting function of $\mathcal{L}_u(\tau_1')/n$,

$$\begin{aligned} Q_u(\theta) = \theta\{F_\theta^{(1)}(u)\log(F_\theta^{(1)}(u)) + (1-F_\theta^{(1)}(u))\log(1-F_\theta^{(1)}(u))\} \\ + (1-\theta)\{F_\theta^{(2)}(u)\log(F_\theta^{(2)}(u)) + (1-F_\theta^{(2)}(u))\log(1-F_\theta^{(2)}(u))\}, \end{aligned}$$

increases as θ approaches q_1 from both sides, where

$$F_\theta^{(1)}(u) = \frac{\min(q_1,\theta)F_1(u) + \max(\theta-q_1,0)F_2(u)}{\min(q_1,\theta)+\max(\theta-q_1,0)} \quad \text{and}$$

$$F_\theta^{(2)}(u) = \frac{\max(q_1-\theta,0)F_1(u) + \min(1-\theta,1-q_1)F_2(u)}{\max(q_1-\theta,0)+\min(1-\theta,1-q_1)},$$

are the limits of $\widehat{F}_1^{\tau_1'}(u)$ and $\widehat{F}_{\tau_1'}^{n+1}(u)$, respectively. This implies that the function $\int_{-\infty}^{\infty} Q_u(\theta)\,dw(u)$ attains its local maximum at the true location of the change-point, q_1.

If we take $dw(u) = \{\widehat{F}_n(u)(1-\widehat{F}_n(u))\}^{-1}\,d\widehat{F}_n(u)$, and also note that $\mathcal{L}_u$ is zero for $u \in (-\infty, X_{(1)})$ and $u \in (X_{(n)}, \infty)$ where $X_{(1)} < \cdots < X_{(n)}$ represent the order statistics, the objective function in (2.2) can be rewritten as

$$\begin{aligned} & R_n(\tau_1', \ldots, \tau_L') \\ & = \int_{X_{(1)}}^{X_{(n)}} \mathcal{L}_u(\tau_1', \ldots, \tau_L')\{\widehat{F}_n(u)(1-\widehat{F}_n(u))\}^{-1}\,d\widehat{F}_n(u) \\ & = n\sum_{k=0}^{L}\sum_{l=2}^{n-1}(\tau_{k+1}'-\tau_k')\frac{\widehat{F}_{kl}\log\widehat{F}_{kl} + (1-\widehat{F}_{kl})\log(1-\widehat{F}_{kl})}{l(n-l)}, \end{aligned} \tag{2.3}$$

where $\widehat{F}_{kl} = \widehat{F}_{\tau_k'}^{\tau_{k+1}'}(X_{(l)})$. As recommended by Zhang (2002), we take a common "continuity correction" by replacing $\widehat{F}_{kl}$ with $\widehat{F}_{kl} - 1/\{2(\tau_{k+1}'-\tau_k')\}$ for all k and l.

To determine L in the MCP, we observe that $Q_u(\theta)$ is a convex function with respect to θ, and thus

$$\max_{\tau_1'<\cdots<\tau_L'} R_n(\tau_1', \ldots, \tau_L') \le \max_{\tau_1'<\cdots<\tau_{L+1}'} R_n(\tau_1', \ldots, \tau_{L+1}'),$$

which means that the maximum log-likelihood $\max_{\tau_1'<\cdots<\tau_L'} R_n(\tau_1', \ldots, \tau_L')$ is a nondecreasing function in L. Hence, we can use Schwarz's Bayesian information criterion (BIC) to strike a balance between the likelihood and the number of change-points by incorporating a penalty for large L. More specifically, we identify the value of L by minimizing

$$\text{BIC}_L = -\max_{\tau_1'<\cdots<\tau_L'} R_n(\tau_1', \ldots, \tau_L') + L\zeta_n \tag{2.4}$$

and ζ_n is a proper sequence going to infinity. Yao (1988) used the BIC with $\zeta_n = \log n$ to select the number of change-points and showed its consistency in the least-squares framework. However, the traditional BIC tends to select a model with some spurious change-points. Detailed discussions on the choice of ζ_n and other tuning parameters are given in Section 3.2.

- Implementation of NMCD Method

3.1. *Algorithm.* One important property of the proposed maximum likelihood approach is that (2.3) is separable. The optimum for splitting cases $1, \ldots, n$ into L segments conceptually consists of first finding the rightmost change-point $\hat{\tau}_L$, and then finding the remaining change-points from the fact that they constitute the optimum for splitting cases $1, \ldots, \hat{\tau}_L$ into $L-1$ segments. This separability is called Bellman's "principle of optimality" [Bellman and Dreyfus (1962)]. Thus, (2.3) can be maximized via the DP algorithm and fitting such a nonparametric MCP model is straightforward and fast. The total computational complexity is $O(Ln^2)$ for a given L; see Hawkins (2001) and Bai and Perron (2003) for the pseudo-codes of the DP. Hawkins (2001) suggested using the DP on a grid of $m \ll n$ values. Harchaoui and Lévy-Leduc (2010) proposed using a LASSO-type penalized estimator to achieve a reduced version of the least-squares method. Niu and Zhang (2012) developed a screening and ranking algorithm to detect DNA copy number variations in the MCP framework.

Due to the DP's computational complexity in n^2, an optimal segmentation of a very long sequence could be computationally intensive; for example, DNA sequences nowadays are often extremely long [Fearnhead and Vasileiou (2009)]. To alleviate the computational burden, we introduce a preliminary screening step which can exclude most of the irrelevant points and, as a consequence, the NMCD is implemented in a much lower-dimensional space.

1. The Screening Procedure

Screening algorithm.

(i) Choose an appropriate integer n_I which is the length of each subsequence of the data, and take the estimated change-point set $\mathcal{O} = \varnothing$.

(ii) Initialize $\gamma_i = 0$ for $i = 1, \ldots, n$; and for $i = n_I, \ldots, n - n_I$, update γ_i to be the Cramér–von Mises two-sample test statistic for the samples $\{X_{i-n_I+1}, \ldots, X_i\}$ and $\{X_{i+1}, \ldots, X_{i+n_I}\}$.

(iii) For $i = n_I, \ldots, n - n_I$, define $k = \arg\max_{i-n_I<j\le i+n_I} \gamma_j$. If $k = i$, update $\mathcal{O} = \mathcal{O} \cup \{i\}$.

Intuitively speaking, this screening step finds the most influential points that have the largest *local* jump sizes quantified by the Cramér–von Mises statistic, and thus helps to avoid including too many candidate points around the true change-point. As a result, we can obtain a candidate change-point set, $\mathcal{O}$, of which the cardinality, $|\mathcal{O}|$, is usually much smaller than n. Finally, we run the NMCD procedure within the set $\mathcal{O}$ using the DP algorithm to find the solution of

$$\arg\max_{\tau_1'<\cdots<\tau_L'\in\mathcal{O}} R_n(\tau_1', \ldots, \tau_L').$$

Apparently, the screening procedure is fast because it mainly requires calculating $n - 2n_I + 1$ Cramér–von Mises statistics. In contrast, Lee (1996) used a thresholding step to determine the number of change-points. The main difference between Lee (1996) and Niu and Zhang (2012) lies in the choice of the local test statistic; the former uses some seminorm of empirical distribution functions and the latter is based on the two-sample mean difference.

2. Select the proper n_I

In the screening procedure, the choice of n_I needs to balance the computation and underfitting. By Proposition 1, $n_I \in (\log n, \lambda_n^{1/2})$, while λ_n is typically unknown. In practice, we recommend to choose $n_I = \lceil (\log n)^{3/2}/2 \rceil$,

3. Implementing the Algorithm by

$$\arg\max_{\tau_1'<\cdots<\tau_L'\in\mathcal{O}} R_n(\tau_1', \ldots, \tau_L')$$