A review for change point detection methods

Yijin Zeng
Imperial College London

Contents

1 Introduction

Change point detection is the process of identifying points in a data sequence where the pattern of observations changes. The earliest work of change point detection dates back to the 1950s, when Page (1954) introduced control charts for detecting shifts in the mean of independent and identically distributed univariate Gaussian variables for industrial quality control. Since then, change point detection methods have been extensively studied and applied across a wide range of fields, including finance (Lavielle and Teyssiere, 2007), signal processing (Haynes et al., 2017), bioinformatics (Lio and Vannucci, 2000), and network traffic analysis (Lévy-Leduc and Roueff, 2009).

This report provides a concise review of various change point detection methods, focusing on their core methodologies and mathematical foundations. For each method, we also include references for readers interested in exploring the model's practical applications and theoretical properties in greater depth. This report is intended for readers who are already familiar with the basic concepts of change point detection, and wish to explore a broader range of methods in a short amount of time. For more gentle introduction of change point detection and change point detection methods, we refer to the work by (Aminikhanghahi and Cook, 2017; Truong et al., 2020). Section 5 provides a short introduction of these work.

This report divides change point detection methods into single change point methods and multiple change point methods, which can be found in Section 2, and Section 3, respectively. A short review for change point detection evaluation metric is provided in Section 4.

2 Single Change Point

Consider observations $y = (y_1, \cdots, y_n)$, where $y_i \in \mathbb{R}$ for each $i \in \{1, \ldots, n\}$. We assume y is a realization from the random sequence $Y = (Y_1, Y_2, \cdots, Y_n)$ such that $Y_1, \cdots, Y_\tau \overset{i.i.d}{\sim} F_1$ and $Y_{\tau+1}, \cdots, Y_n \overset{i.i.d}{\sim} F_2$, where F_1, F_2 are unknown cumulative density functions. Our goal is to decide whether $F_1 = F_2$, and if $F_1 \neq F_2$, to provide an estimation $\hat{\tau}$ of τ.

2.1 CUSUM method

The Cumulative Sum (CUSUM) algorithm is proposed by Page (1954) for sequence change point detection. The algorithm has been studied in later work (Healy, 1987), especially for normally distributed data (Duncan, 1974).

The CUSUM algorithm computes the statistic

$$S_m = \sum_{i=1}^{m} \log \frac{f_2(y_i)}{f_1(y_i)} - \min_{k \leq n} \sum_{i=1}^{k} \log \frac{f_2(y_i)}{f_1(y_i)} \tag{1}$$

where f_1, f_2 are the corresponding density functions of F_1, F_2, respectively. Equivalently, the equation (1) can be calculated recursively as

$$S_m = \max\left(S_{m-1} + \log \frac{f_2(y_i)}{f_1(y_i)}, 0\right),$$

which makes CUSUM suitable for online change point detection.

Suppose F_1 and F_2 are univariate normal distribution with mean values as μ_1 and $\mu_2 (\mu_2 > \mu_1)$ respectively with the same variance σ^2. In other words,

$$Y_i \sim \mathcal{N}(\mu_1, \sigma^2) \text{ for } i = 1, 2, \cdots, \tau,$$
$$Y_i \sim \mathcal{N}(\mu_2, \sigma^2) \text{ for } i = \tau + 1, \cdots, T.$$

The CUSUM statistic can be further written as

$$S_m = \max\left(S_{m-1} + y_m - (\mu_1 + \mu_2)/2, 0\right).$$

CUSUM detects change points when S_m is larger than a pre-defined threshold L, which L is often chosen to control the probability of making false detection. See Chapter 2 in Bodenham and Adams (2017) for more discussion about choosing L, and extensive simulations results.

2.2 Likelihood Ratio Test

Hinkley (1970) considers likelihood ratio test for change point detection (CPD). Suppose each Y_i follows a distribution F with probability density function f independently. The estimated change point $\hat{\tau}$ is the value of t that maximizes the likelihood function:

$$L(t) = \sum_{i=1}^{t} \log f(y_i, \theta_1) + \sum_{i=t+1}^{T} \log f(y_i, \theta_2).$$

For the case where θ_1 and θ_2 are unknown parameters in a normal distribution, the asymptotic distribution of $\hat{\tau}$ is given by Hinkley (1970).

Intuitively, $\hat{\tau}$ represents the "most likely" location of a change point, assuming one exists. However, we are often also interested in testing whether a change point exists at all. This is commonly approached by evaluating the following test statistic

$$S = 2\left(L(\hat{\tau}) - \sum_{i=1}^{n} \log f(y_i, \theta)\right).$$

If S exceeds a threshold c, we conclude that a change point is present at $\hat{\tau}$. Otherwise, we fail to reject the null hypothesis of no change point. The choice of c remains an open problem (Eckley et al., 2011; Killick and Eckley, 2014). One approach is to derive the asymptotic distribution of S under the null hypothesis when the underlying distribution is specified. Examples for normal, gamma, Poisson, and other cases are discussed by Chen and Gupta (2012).

2.3 Change Point Detection Based on T-test

The CUSUM method has the appealing property that, if the data are normally distributed and all parameters, pre- and post-change are known, it guarantees optimal performances (Moustakides, 1986), in the sense discussed by Lorden (1971). However, Hawkins et al. (2003) points out that the case where parameters are known is rare in practice, and considers an algorithm for normally distributed data in which all parameters are unknown.

Suppose the pre- and post-change distributions are Gaussian, but their parameters μ_1, μ_2, and σ are unknown:

$$Y_i \sim \mathcal{N}(\mu_1, \sigma^2),\ i = 1, 2, \ldots, \tau,$$
$$Y_i \sim \mathcal{N}(\mu_2, \sigma^2),\ i = \tau + 1, \ldots, T.$$

Following the notation by Hawkins et al. (2003), for any $n \geq j \geq 1$ define

$$\overline{Y}_{jn} = \frac{1}{j} \sum_{i=1}^{j} Y_i$$

as the sample mean of observations 1 to j, and

$$\overline{Y}^*_{jn} = \frac{1}{n-j} \sum_{i=j+1}^{n} Y_i$$

as the sample mean of observations $j+1$ to n. The residual sum of squares is given by

$$V_{jn} = \sum_{i=1}^{j} (Y_i - \overline{Y}_{jn})^2 + \sum_{i=j+1}^{n} (Y_i - \overline{Y}^*_{jn})^2.$$

For a given j, we can compute a two-sample t-statistic:

$$T_{jn} = \sqrt{\frac{j(n-j)}{n}} \frac{\overline{Y}_{jn} - \overline{Y}^*_{jn}}{\hat{\sigma}_{jn}},$$

where $\hat{\sigma}_{jn} = V_{jn}/(n-2)$. If $Y_1, \ldots, Y_n \overset{\text{i.i.d.}}{\sim} \mathcal{N}(\mu_1, \sigma)$, then T_{jn} follows a t-distribution with $n-2$ degrees of freedom. For $1 \leq j \leq n-1$, define

$$T_{\max,n} = \max_j T_{jn}, \text{ and } \tau_n = \arg\max_j T_{jn}.$$

Here, τ_n is the maximum likelihood estimator of the true change point τ. If $T_{\max,n}$ exceeds a decision threshold h_n, we conclude that $\mu_1 \neq \mu_2$ and estimate the change point as τ_n.

The main challenge following this approach is setting the threshold h_n to achieve a specified false alarm probability. One possible approach is to apply the Bonferroni inequality to choose h_n for a given false alarm probability. However, this bound is overly conservative when n is large, as noted by Hawkins et al. (2003): while false alarms can be controlled when the process is in control, the generalized likelihood ratio (GLR) statistic becomes very insensitive to real change points.

2.4 Non-parametric Change Point Detection Based On Rank

Hawkins and Deng (2010) consider a change point detection method without the distributional assumptions using rank. Following the notation by Hawkins and Deng (2010), define

$$D_{ij} = \operatorname{sgn}(Y_i - Y_j) = \begin{cases} 1, & \text{if } Y_i > Y_j, \\ 0, & \text{if } Y_i = Y_j, \\ -1, & \text{if } Y_i < Y_j, \end{cases}$$

$$U_{k,n} = \sum_{i=1}^{k} \sum_{j=k+1}^{n} D_{ij}, \quad 1 \leq k \leq n-1,$$

$$T_{k,n} = \frac{U_{k,n}}{\sqrt{k(n-k)(n+1)/3}}.$$

Consider the hypothesis test $H_0 : \theta = 0$ versus $H_\alpha : \theta \neq 0$. By relating $T_{k,n}$ to the Mann–Whitney statistic, we have

$$T_{k,n} \xrightarrow[n,\, n-k\to\infty]{d} \mathcal{N}(0,1)$$

under the null hypothesis. Similar to Hawkins et al. (2003), define

$$T_{\max,n} = \max_k T_{k,n}, \text{ and } \tau_n = \arg\max_k T_{k,n}.$$

If $T_{\max,n}$ exceeds a decision threshold h_n, we reject H_0 and estimate the change point as τ_n. This again raises the question about choosing h_n for a given false alarm probability. Due to the difficulty of specifying h_n theoretically, the authors suggest determine h_n by Monte Carlo simulations.

Similar approaches based on the Kolmogorov–Smirnov and Cramér–von Mises tests are considered by Ross and Adams (2012).

3 Multiple Change Points

Similarly to the single change point problem, suppose we observe $y = (y_1, y_2, \ldots, y_n)$, where $y_i \in \mathbb{R}$ and y is a realization of the random sequence $Y = (Y_1, Y_2, \ldots, Y_n)$. The process Y is assumed to be piecewise i.i.d., with a set of change points $\tau = (\tau_1, \ldots, \tau_m)$, where each $\tau_i \in \mathbb{N}$, $\tau_0 = 1$, $\tau_{m+1} = n$, and $\tau_i < \tau_j$ if and only if $i < j$.

The change points partition Y into segments such that

$$Y_{\tau_j}, \ldots, Y_{\tau_{j+1}} \overset{\text{i.i.d.}}{\sim} F_j, \; j = 0, \ldots, m,$$
$$F_i \neq F_{i+1}, \; i = 0, \ldots, m-1.$$

For notation convenience, we denote the sequence $y_a, y_{a+1}, \ldots, y_b$ by $y_{a:b}$ in the following of the report.

3.1 Binary Segmentation

Binary segmentation (Scott and Knott, 1974) is a popular algorithm for multiple change point detection (Killick and Eckley, 2014) and is among the best-performing methods for practical change point detection problems (van den Burg and Williams, 2020).

Binary segmentation adapts a recursive segmentation approach: a single change point test statistic is first applied to the entire sequence, and if a change point is detected, the sequence is split into two segments. The procedure is then applied recursively to each segment until no further change points are found.

Specifically, Scott and Knott (1974) assumes that Y_i are independent and follow $\mathcal{N}(\mu_i, \sigma)$, and considers testing the null hypothesis H_0: $\mu_1 = \mu_2 = \cdots = \mu_n$ against the alternative that μ_i takes one of two possible values m_1 or m_2. To implement binary segmentation, we first find the partition B_0 that maximizes the between-groups sum of squares:

$$B_0 = \arg\max_t \left(\sum_{i=1}^{t} y_i - \sum_{i=1}^{n} y_i \right)^2 + \left(\sum_{i=t+1}^{n} y_i - \sum_{i=1}^{n} y_i \right)^2. \tag{2}$$

The maximum likelihood estimate of the variance is

$$\hat{\sigma}^2 = \frac{1}{n} \sum_{i=1}^{n} \left(y_i - \frac{1}{n} \sum_{j=1}^{n} y_j \right)^2.$$

Then, under H_0 (equal means in both groups), the test statistic

$$\lambda = \frac{\pi}{2(\pi - 2)} \frac{B_0}{\hat{\sigma}^2}$$

is asymptotically χ-distributed. If λ leads to the rejection of H_0, the procedure is repeated on the resulting two segments until no further rejections occur.

3.2 Wild Binary Segmentation

Wild Binary Segmentation (WBS) (Fryzlewicz, 2014) is a refinement of binary segmentation (Scott and Knott, 1974). WBS is motivated by the observation that binary segmentation is effective in the presence of a single change point, but can perform poorly when multiple change points occur in the sequence.

Unlike binary segmentation, WBS does not attempt to estimate change points using the whole sequence directly. Instead, Fryzlewicz (2014) proposes first randomly sampling subintervals $(Y_s, Y_{s+1}, \ldots, Y_e)$, where s and e are integers with $1 \leq s < e \leq n$. Then, compute the following CUSUM statistic for all subintervals:

$$Y_{s,e}^b = \sqrt{\frac{e-b}{n(b-s+1)}} \sum_{t=s}^{b} Y_t - \sqrt{\frac{b-s+1}{n(e-b)}} \sum_{t=b+1}^{e} Y_t.$$

The largest $|Y_{s,e}^b|$ among all subintervals is taken as the candidate change point for the entire sequence. A hypothesis test is then performed under the null hypothesis that the candidate is not a change point. If the null is rejected, the sequence is split at this point and the procedure is applied recursively until no further rejections occur.

3.3 Pruned Exact Linear Time

Pruned Exact Linear Time (PELT) (Killick et al., 2012) builds upon the work by Jackson et al. (2005). Unlike binary segmentation and wild binary segmentation, which greedily search for multiple change points, Jackson et al. (2005); Killick et al. (2012) estimates all change points simultaneously by solving the optimization problem

$$\min_{\tau \subset \{1,\ldots,n\}} \left\{ \sum_{i=1}^{m+1} c\big(y_{(\tau_{i-1}+1):\tau_i}\big) \right\} + \beta f(m),$$

where $c(\cdot)$ is a segmentation cost function, e.g. variance of the sub-sequence, and $\beta f(m)$ is a penalty term for preventing overfitting.

Jackson et al. (2005) considers the case $\beta f(m) = \beta m$ and shows that

$$\begin{aligned} F(y_{1:n}) &= \min_{\tau \subset \{1,\dots,n\}} \left\{ \sum_{i=1}^{m+1} \left[c\big(y_{(\tau_{i-1}+1):\tau_i}\big) + \beta \right] \right\} \\ &= \min_{t} \left\{ F(y_{1:t}) + c\big(y_{(t+1):n}\big) + \beta \right\}. \end{aligned}$$

This formulation allows the problem to be solved recursively. However, the computational time for the recursive search is high, as noted by Killick et al. (2012), who then proposes an efficient pruning strategy for saving computational time. Specifically, if $c(\cdot)$ satisfies the condition that there exists a constant K such that for all $t < s < T$,

$$c\big(y_{(t+1):s}\big) + c\big(y_{(s+1):T}\big) + K \le c\big(y_{(t+1):T}\big),$$

then, if

$$F(y_{1:t}) + c\big(y_{(t+1):s}\big) + K \ge F(y_{1:s}),$$

at any future time $T > s$, t can never be the optimal last change point prior to T and can be pruned from consideration. This pruning process reduces computational cost while preserving the accuracy of the original algorithm by Jackson et al. (2005).

3.4 Non-parametric Change Point Detection

Zou et al. (2014) considers a nonparametric maximum likelihood approach by exploring the fact that for n univariate random variables $Y_1, Y_2, \cdots, Y_n \overset{i.i.d}{\sim} F$, the empirical cumulative distribution function $\hat{F}(\mu)$ of $y_1, \cdots, y_n$ satisfies

$$n\hat{F}(\mu) \sim \text{Binomial}\big(n, F(\mu)\big)$$

for any fixed μ.

Assume there are m change points, and denote the empirical cumulative distribution function of $y_{\hat{\tau}_i:\hat{\tau}_{i+1}}$ as $\hat{F}_{\hat{\tau}_i}^{\hat{\tau}_{i+1}}(\mu)$. For fixed μ, the log-likelihood can be written as

$$\mathcal{L}(\hat{\tau}_1, \cdots, \hat{\tau}_m) = \sum_{k=0}^{m} (\hat{\tau}_{k+1} - \hat{\tau}_k) \Big\{ \hat{F}_{\hat{\tau}_k}^{\hat{\tau}_{k+1}}(\mu) \log \hat{F}_{\hat{\tau}_k}^{\hat{\tau}_{k+1}}(\mu) + \big[1 - \hat{F}_{\hat{\tau}_k}^{\hat{\tau}_{k+1}}(\mu)\big] \log \big[1 - \hat{F}_{\hat{\tau}_k}^{\hat{\tau}_{k+1}}(\mu)\big] \Big\}.$$

For each fixed μ, the likelihood $\mathcal{L}(\hat{\tau}_1, \cdots, \hat{\tau}_m)$ is expected to increase as each $\hat{\tau}_i \to \tau_i$. Hence, one approach is to estimate the change points τ by maximizing the likelihood function $\mathcal{L}(\hat{\tau}_1, \cdots, \hat{\tau}_m)$ directly.

However, the estimated change points would then depend on the choice of μ, which is hard to choose in practice. To overcome this problem, Zou et al. (2014) suggests choosing a weight function $w(\mu)$ and maximizing the integrated likelihood

$$\mathcal{R}(\hat{\tau}_1, \cdots, \hat{\tau}_m) = \int_{-\infty}^{\infty} \mathcal{L}(\hat{\tau}_1, \cdots, \hat{\tau}_m)\, dw(\mu),$$

where, in their implementation,

$$dw(\mu) = \big\{ \hat{F}_1^n(\mu)\big(1 - \hat{F}_1^n(\mu)\big) \big\}^{-1} d\hat{F}_1^n(\mu).$$

Since m is unknown, they further propose estimating it by minimizing

$$\mathcal{R}(\hat{\tau}_1, \cdots, \hat{\tau}_m) + m\, \xi_n, \tag{3}$$

where ξ_n denotes the Bayesian information criterion (Schwarz, 1978). Let $\overline{K}_n$ be the upper bound on the number of change points. Zou et al. (2014) show that if

$$\xi_n = \overline{K}_n^3 (\log K_n)^2 (\log n)^{2+c}$$

for any $c > 0$, then under mild conditions, the estimated number of change points converges in probability to the true number, and $\hat{\tau}$ is close to τ.

The optimization problem (3) is solved via dynamic programming with computational cost $O(mn^2)$. Haynes et al. (2017) further propose an approximation to (3) for reducing computational complexity.

3.5 Energy Change Point Detection

Energy change point detection (ECD) (Matteson and James, 2014) is a multivariate non-parametric change point detection algorithm, combining both the energy distance (Szekely et al., 2005), and the bisection method (Vostrikova, 1981).

The definition of the energy distance for two multivariate sequences $X^{n_1} = (X_1, \cdots, X_{n_1})$ and $Z^{n_2} = (Z_1, \cdots, Z_{n_2})$ is:

$$\epsilon(X^{n_1}, Z^{n_2}; \alpha) = \frac{1}{n_1 n_2} \sum_{j=1}^{n_1} \sum_{i=1}^{n_2} \|X_j - Z_i\|^\alpha - \frac{1}{2n_1^2} \sum_{j=1}^{n_1} \sum_{i=1}^{n_1} \|X_j - X_i\|^\alpha - \frac{1}{2n_2^2} \sum_{j=1}^{n_2} \sum_{i=1}^{n_2} \|Z_j - Z_i\|^\alpha. \tag{4}$$

It can be shown that the statistic

$$\frac{mn}{m+n} \epsilon(X^n, Z^m; \alpha)$$

converges in distribution to a non-degenerate random variable if and only if

$$(X_1, \ldots, X_{n_1}, Z_1, \ldots, Z_{n_2}) \overset{i.i.d}{\sim} F(x),$$

where $F(x)$ is a cumulative density function such that $\mathbb{E}|X|^\alpha < \infty$ ($\alpha < 2$). Otherwise,

$$\frac{mn}{m+n} \epsilon(X^n, Z^m; \alpha) \to \infty.$$

Denote $Y^\tau = (Y_1, \ldots, Y_\tau)$ and $Y_\tau^\kappa = (Y_{\tau+1}, \ldots, Y_\kappa)$. By modifying the bisection method, Matteson and James (2014) proposes to search a change point $\hat{\tau}$ by solving

$$\underset{\hat{\tau}, \hat{\kappa}}{\operatorname{argmin}} \, \epsilon\big(Z^{\hat{\tau}}, Z_{\hat{\tau}}^{\hat{\kappa}}\big).$$

Then, a two-sample hypothesis test between $Y_1, \ldots, Y_{\hat{\tau}}$ and $Y_{\hat{\tau}+1}, \ldots, Y_n$ is performed using the permutation method. If the null hypothesis is rejected at the specified significance level, $\hat{\tau}$ is considered a change point, and the sequence is split into two disjoint subsequences. This procedure will be repeated recursively on the split subsequences until no change point is detected.

A similar approach based on kernel is considered by Harchaoui et al. (2009a).

4 Change point detection evaluation metrics

Change point detection evaluation metrics can be roughly divided into clustering-based metrics and classification-based metrics (van den Burg and Williams, 2020). In this section, we consider the so-called covering metric (van den Burg and Williams, 2020) from the clustering category, and precision, recall from the classification category. For more evaluation metrics, we refer to the work by Aminikhanghahi and Cook (2017); Truong et al. (2020); van den Burg and Williams (2020).

The covering metric measures the overlap between segments of the sequence $y = (y_1, \ldots, y_n)$ induced by the true change point set τ and the estimated set $\hat{\tau}$. To be more formal, let us denote the segments induced by the true change points as

$$\mathcal{A}_j = \{y_{\tau_j}, \ldots, y_{\tau_{j+1}}\}, \text{ for } j = 0, \ldots, m,$$

and denote

$$\mathcal{A} = \{\mathcal{A}_0, \ldots, \mathcal{A}_m\}.$$

Similarly, let us denote the segments induced by the estimated change points as

$$\hat{\mathcal{A}}_j = \{y_{\hat{\tau}_j}, \ldots, y_{\hat{\tau}_{j+1}}\}, \text{ for any } j = 0, \ldots, \hat{m},$$

and denote

$$\hat{\mathcal{A}} = \{\hat{\mathcal{A}}_0, \ldots, \hat{\mathcal{A}}_{\hat{m}}\}.$$

For any segments $\mathcal{A}_j$ and $\hat{\mathcal{A}}_j$, define the Jaccard index as

$$J(\mathcal{A}_j, \hat{\mathcal{A}}_j) = \frac{|\mathcal{A}_j \cap \hat{\mathcal{A}}_j|}{|\mathcal{A}_j \cup \hat{\mathcal{A}}_j|}.$$

The covering metric is then defined by van den Burg and Williams (2020) as :

$$\mathcal{C}(\mathcal{A}, \hat{\mathcal{A}}) = \frac{1}{n} \sum_{\mathcal{A}_j \in \mathcal{A}} |\mathcal{A}_j| \max_{\hat{\mathcal{A}}_j \in \hat{\mathcal{A}}} J(\mathcal{A}_j, \hat{\mathcal{A}}_j).$$

Note that $\mathcal{C}(\mathcal{A}, \hat{\mathcal{A}}) \to 1$ if and only if $\hat{\mathcal{A}} \to \mathcal{A}$. In other words, the larger the covering metric, the better the estimated change points.

Precision and recall measure the correspondence between estimated and true change point locations. The true positive is defined as the number of estimated change points that are close enough to a true change point. Formally,

$$\mathrm{Tp}(\tau, \hat{\tau}) = \{\tau_i \in \tau \mid \exists\, \hat{\tau}_j \in \hat{\tau} \text{ s.t. } |\tau_i - \hat{\tau}_j| < M\},$$

where $M \in \mathbb{N}$ is tolerance parameter. The precision and recall are then defined as

$$\mathrm{Pre}(\tau, \hat{\tau}) = \frac{|\mathrm{Tp}(\tau, \hat{\tau})|}{|\tau|}, \text{ and } \mathrm{Rec}(\tau, \hat{\tau}) = \frac{|\mathrm{Tp}(\tau, \hat{\tau})|}{|\hat{\tau}|},$$

respectively. The estimated change points are evaluated using both $\mathrm{Pre}(\tau, \hat{\tau})$ and $\mathrm{Rec}(\tau, \hat{\tau})$. A good segmentation would have metrics are close to 1. While only one metric is close to 1 could imply over-segmentation or under-segmentation results.

5 Other review papers

As mentioned earlier, this report provides only a brief review of change point detection methods, focusing on their core methodologies and mathematical foundations. More comprehensive review papers are available in the literature. This section introduces two excellent review papers by Truong et al. (2020) and Aminikhanghahi and Cook (2017).

Truong et al. (2020) reviews general offline change point detection methods, and decompose all the methods into three components, namely cost functions, penalty terms, and search methods. Specifically, the cost of a subsequence $y_{a:b}$, with $1 \le a \le b \le n$, is denoted by $c(y_{a:b})$, where $c(\cdot)$ is a cost function. The cost function is designed to be sensitive to the presence of change points in the subsequence: if the subsequence does not contain change points, the cost should be small; otherwise, it should be large. The penalty terms are designed to prevent over-segmentation by penalizing the number of change points detected. Combining both the cost function and the penalty term, the change point detection problem can be formulated into the following minimization problem:

$$\sum_{i=1}^{m+1} c\big(y_{(\hat{\tau}_{i-1}+1):\hat{\tau}_i}\big) + \beta f(m), \tag{5}$$

where $\hat{\tau}_i$ are the estimated change points, and $\beta f(m)$ is a penalty term to prevent over-segmentation. The choice of the cost function $c(\cdot)$ is often based on the characteristics of the change points to be detected, e.g. detecting changes in mean, or variance could require different cost functions. Common options include the negative log-likelihood (Chen and Gupta, 2012), squared quadratic loss, nonparametric log-likelihood cost (Zou et al., 2014), rank-based cost, and kernel-based cost (Harchaoui and Cappé, 2007). Common penalty terms include linear penalties (Killick et al., 2012), ℓ_1-type penalties (Harchaoui and Lévy-Leduc, 2010), and more complex forms (Zhang and Siegmund, 2007).

For solving the minimization problem as formulated in (5), different search methods are available, such as PELT (Killick et al., 2012), binary segmentation (Scott and Knott, 1974), and dynamic programming approaches (Bai and Perron, 2003).

Aminikhanghahi and Cook (2017) considers change point detection from a machine learning perspective. In the supervised setting, change point detection can be formulated as either binary or multiclass classification. In the unsupervised setting, methods are broadly divided into six categories:

likelihood ratio approaches (Basseville et al., 1993; Kawahara and Sugiyama, 2012), subspace modelling (Liu et al., 2013), probabilistic methods (Adams and MacKay, 2007; Saatçi et al., 2010), kernel-based methods (Harchaoui et al., 2009b), graph-based models (Chen and Zhang, 2015), and clustering methods (Rakthanmanon et al., 2011; Zakaria et al., 2012).

There is also a large body of Bayesian change point detecion methods (Adams and MacKay, 2007; Fearnhead and Liu, 2007; Tartakovsky and Moustakides, 2010). These typically infer the next change point conditional on previously estimated change points and are widely regarded as producing state-of-the-art results (van den Burg and Williams, 2020). Killick (2011) compares different change point detection approaches based on frequentist and Bayesian.

References

Adams, R. P. and MacKay, D. J. (2007). Bayesian online changepoint detection. *arXiv preprint arXiv:0710.3742*.

Aminikhanghahi, S. and Cook, D. J. (2017). A survey of methods for time series change point detection. *Knowledge and information systems*, 51(2):339–367.

Bai, J. and Perron, P. (2003). Computation and analysis of multiple structural change models. *Journal of applied econometrics*, 18(1):1–22.

Basseville, M., Nikiforov, I. V., et al. (1993). *Detection of abrupt changes: theory and application*, volume 104. prentice Hall Englewood Cliffs.

Bodenham, D. A. and Adams, N. M. (2017). Continuous monitoring for changepoints in data streams using adaptive estimation. *Statistics and Computing*, 27(5):1257–1270.

Chen, H. and Zhang, N. (2015). Graph-based change-point detection. *The Annals of Statistics*, 43(1):139–176.

Chen, J. and Gupta, A. K. (2012). Parametric statistical change point analysis: with applications to genetics, medicine, and finance.

Duncan, A. J. (1974). Quality control and industrial statistics.

Eckley, I. A., Fearnhead, P., and Killick, R. (2011). Analysis of changepoint models. *Bayesian time series models*, pages 205–224.

Fearnhead, P. and Liu, Z. (2007). On-line inference for multiple changepoint problems. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 69(4):589–605.

Fryzlewicz, P. (2014). Wild binary segmentation for multiple change-point detection. *The Annals of Statistics*, 42(6):2243–2281.

Harchaoui, Z. and Cappé, O. (2007). Retrospective mutiple change-point estimation with kernels. In *2007 IEEE/SP 14th Workshop on Statistical Signal Processing*, pages 768–772. IEEE.

Harchaoui, Z. and Lévy-Leduc, C. (2010). Multiple change-point estimation with a total variation penalty. *Journal of the American Statistical Association*, 105(492):1480–1493.

Harchaoui, Z., Moulines, E., and Bach, F. R. (2009a). Kernel change-point analysis. In *Advances in neural information processing systems*, pages 609–616.

Harchaoui, Z., Vallet, F., Lung-Yut-Fong, A., and Cappé, O. (2009b). A regularized kernel-based approach to unsupervised audio segmentation. In *2009 IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 1665–1668. IEEE.

Hawkins, D. M. and Deng, Q. (2010). A nonparametric change-point control chart. *Journal of Quality Technology*, 42(2):165–173.

Hawkins, D. M., Qiu, P., and Kang, C. W. (2003). The changepoint model for statistical process control. *Journal of quality technology*, 35(4):355–366.

Haynes, K., Fearnhead, P., and Eckley, I. A. (2017). A computationally efficient nonparametric approach for changepoint detection. *Statistics and Computing*, 27(5):1293–1305.

Healy, J. D. (1987). A note on multivariate cusum procedures. *Technometrics*, 29(4):409–412.

Hinkley, D. V. (1970). Inference about the change-point in a sequence of random variables.

Jackson, B., Scargle, J. D., Barnes, D., Arabhi, S., Alt, A., Gioumousis, P., Gwin, E., Sangtrakulcharoen, P., Tan, L., and Tsai, T. T. (2005). An algorithm for optimal partitioning of data on an interval. *IEEE Signal Processing Letters*, 12(2):105–108.

Kawahara, Y. and Sugiyama, M. (2012). Sequential change-point detection based on direct density-ratio estimation. *Statistical Analysis and Data Mining: The ASA Data Science Journal*, 5(2):114–127.

Killick, R. (2011). *Analysis of changepoint models.*

Killick, R. and Eckley, I. (2014). changepoint: An r package for changepoint analysis. *Journal of statistical software*, 58(3):1–19.

Killick, R., Fearnhead, P., and Eckley, I. A. (2012). Optimal detection of changepoints with a linear computational cost. *Journal of the American Statistical Association*, 107(500):1590–1598.

Lavielle, M. and Teyssiere, G. (2007). Adaptive detection of multiple change-points in asset price volatility. In *Long memory in economics*, pages 129–156. Springer.

Lévy-Leduc, C. and Roueff, F. (2009). Detection and localization of change-points in high-dimensional network traffic data. *The Annals of Applied Statistics*, pages 637–662.

Lio, P. and Vannucci, M. (2000). Wavelet change-point prediction of transmembrane proteins. *Bioinformatics*, 16(4):376–382.

Liu, S., Yamada, M., Collier, N., and Sugiyama, M. (2013). Change-point detection in time-series data by relative density-ratio estimation. *Neural Networks*, 43:72–83.

Lorden, G. (1971). Procedures for reacting to a change in distribution. *The Annals of Mathematical Statistics*, pages 1897–1908.

Matteson, D. S. and James, N. A. (2014). A nonparametric approach for multiple change point analysis of multivariate data. *Journal of the American Statistical Association*, 109(505):334–345.

Moustakides, G. V. (1986). Optimal stopping times for detecting changes in distributions. *the Annals of Statistics*, 14(4):1379–1387.

Page, E. S. (1954). Continuous inspection schemes. *Biometrika*, 41(1/2):100–115.

Rakthanmanon, T., Keogh, E. J., Lonardi, S., and Evans, S. (2011). Time series epenthesis: Clustering time series streams requires ignoring some data. In *2011 IEEE 11th International Conference on Data Mining*, pages 547–556. IEEE.

Ross, G. J. and Adams, N. M. (2012). Two nonparametric control charts for detecting arbitrary distribution changes. *Journal of Quality Technology*, 44(2):102–116.

Saatçi, Y., Turner, R. D., and Rasmussen, C. E. (2010). Gaussian process change point models. In *ICML*.

Schwarz, G. (1978). Estimating the dimension of a model. *The annals of statistics*, pages 461–464.

Scott, A. J. and Knott, M. (1974). A cluster analysis method for grouping means in the analysis of variance. *Biometrics*, pages 507–512.

Szekely, G. J., Rizzo, M. L., et al. (2005). Hierarchical clustering via joint between-within distances: Extending ward's minimum variance method. *Journal of classification*, 22(2):151–184.

Tartakovsky, A. G. and Moustakides, G. V. (2010). State-of-the-art in bayesian changepoint detection. *Sequential Analysis*, 29(2):125–145.

Truong, C., Oudre, L., and Vayatis, N. (2020). Selective review of offline change point detection methods. *Signal Processing*, 167:107299.

van den Burg, G. J. and Williams, C. K. (2020). An evaluation of change point detection algorithms. *arXiv preprint arXiv:2003.06222*.

Vostrikova, L. Y. (1981). Detecting "disorder" in multidimensional random processes. In *Doklady akademii nauk*, volume 259, pages 270–274. Russian Academy of Sciences.

Zakaria, J., Mueen, A., and Keogh, E. (2012). Clustering time series using unsupervised-shapelets. In *2012 IEEE 12th International Conference on Data Mining*, pages 785–794. IEEE.

Zhang, N. R. and Siegmund, D. O. (2007). A modified bayes information criterion with applications to the analysis of comparative genomic hybridization data. *Biometrics*, 63(1):22–32.

Zou, C., Yin, G., Feng, L., and Wang, Z. (2014). Nonparametric maximum likelihood approach to multiple change-point problems. *The Annals of Statistics*, 42(3):970–1002.