Quality Control Project : A review of different change point detection methods

Cyprien Ferraris

December 2019

1 Introduction

1.1 Presentation of the change point detection problem

The change point detection problem is a very important question for many fields. For example, quality control in industry, also often mentioned as SPC (statistical process control) [1], medical condition monitoring [2], speech recognition [3] and a lot of others, see for example [4].

Usually, the context is parametric, i.e. one associates an ensemble of probability distributions indexed by parameters of $\mathbb{R}^d$, $d \geq 1$. Then one observes data and wants to detect if there is a significant change in the parameters.

There is two ways of looking at this kind of problems. The first one called offline change detection. It is when one already has observed the entire sample and one wants to see if there is change in this sample. For example see [5].

The second way is when the data are observed one by one and one wants to detect as quickly as possible a change of distribution. To solve this problem, one of the main used technique is the CUSUM method [6] which as been proved to be exactly optimal under Lorden's optimality criterion, see for example [7] for the demonstration in the asymptotic case or Moustakides [8] for the demonstration that is exactly optimal under Lorden's criterion. Moreover it is asymptotic optimum under the Pollak's criterion [9] as the false alarm rates goes to zero[10].

The problem with the parametric cased mentioned over is the assumption that one knows the distribution, which is often no the case in practise.

That is why efficient nonparametric methods are very useful in practice.

1.2 Mathematical formulation of the Change point detection problem

Assume that one has an identically independent distributed (iid) sample $X_1, \dots, X_n$.

The idea is to test

- $H_0 : (X_1, \dots, X_n) \overset{iid}{\sim} F_0$
- $H_1 : \exists \tau \in \{1, \dots, n\}$ such that $(X_1, \dots, X_\tau) \overset{iid}{\sim} F_0$ and $(X_{\tau+1}, \dots, X_n) \overset{iid}{\sim} F_1$

Where $F_0 \neq F_1$. τ represent the unknown change point moment.

One will call this **problem (1)**.

It is also possible to assume that there is multiple change points in the data. In this case the problem (1) can be rewritten as

- $H_0 : (X_1, \dots, X_n) \overset{iid}{\sim} F_0$
- $H_1 : \exists 0 < \tau_1 < \cdots < \tau_m < n$ such that for all 0 >i >m $(X_{\tau_{i-1}}, \dots, X_{\tau_i}) \overset{iid}{\sim} F_{i-1}$ and $(X_{\tau_i}, \dots, X_{\tau_{i+1}}) \overset{iid}{\sim} F_i$

Where m is the number of change points (unknown), $\tau_0 = 0$, $\tau_m = n$. In this case one denotes by F the cumulative distribution function (c.d.f.) of the full data and by F_i the distribution of the i-th segment, i.e. of the data $x_{\tau_i}, \ldots, x_{\tau_{i+1}}$. One also assumes that $\forall i \in \{0, \ldots, m-1\}, F_i \neq F_{i+1}$

One calls this model **problem (2)**. In the next sections one would discuss around different methodology to deal with those problems. But first one will see different ways to judge and compare the performance of change point detection method.

1.3 Measure of performance for methods of change point detection

The first thing that can be expected of a change point detection method is **false alarm rate**. It corresponds to the probability of detecting that a change occur while in reality no change has occur. It represents the error of type I in test theory.

A second measure is the **average run length (ARL)**. Essentially, the ARL is the average number of points that must be plotted before a point indicates an out-of-control condition, so it is a measure of the ability to detect quickly the occurrence of a change point.

Another criterion is the **average detection delay (ADD)**.

Mathematically speaking, let's denotes S(t) a test statistic based on $x_1, \ldots, x_t$ to make a decision about the existence of one change point, and $T(\gamma) = inf\{t : S(t) > \gamma\}$ where γ corresponds to some threshold. The two previous measures of performance can be rewritten as

$$\begin{aligned} ARL(\gamma) &= E_\infty[T(\gamma)] \\ ADD(\gamma) &= \sup_\tau E_\tau[T(\gamma) - \tau | T(\gamma) \geq \tau] \end{aligned}$$

Where E_τ is the expectation assuming the change point is at τ and E_∞ denotes the expectation when there is no change at all.

2 Detecting a change in mean in a non parametric case

2.1 Modeling

In this paragraph, one assumes the following problem :

$X_1, \ldots, X_\tau, X_{\tau+1}, \ldots, X_n$ are independent, continuous R random variables. One wants to test

$$\begin{aligned} H_0 &: X_i \sim F_0(x), && for\ i = 1, \ldots, \tau \\ H_1 &: X_i \sim F_1(x) = F_0(x - \theta), && for\ i = \tau + 1, \ldots, n \end{aligned}$$

Where the parameter θ represents a shift in location occurring after the change point τ. Both θ and τ are assumed to be unknown. In this setting, testing whether the process has shifted corresponds to the hypothesis test

$$H_0 : \theta = 0 \ against \ H_1 : \theta \neq 0$$

To do so a way is to consider a U-statistic based on the Mann-Withney two-sample test. Let D_{ij} be the sign of $X_i - X_j$. The U-statistic $U_{k,n}$ is defined as a function of the D_{ij},

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

It can be shown [11] that the mean of this statistic is 0 and its variance depends on k.

So a way to delete this dependence is to look at the standardize statistic :

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

2.2 Hawkins et al Method [12]

From the asymptotic normality of Mann-Whitney statistics, $T_{k,n} \sim N(0,1)$, when k and n-k both go to ∞. Then a test statistic for the presence of change point and an estimate of its time of occurrence are given by :

$$\begin{aligned} T_{max,n} &= \max_{1\leq k\leq n-1} |T_{k,n}| \\ \hat{\tau}_T &= \operatorname*{argmax}_{1\leq k\leq n-1} |T_{k,n}| \end{aligned}$$

What one looks at his when the size of the data set is not fixed, but grows as long as the process is believed to still be in control. So one needs to define a sequence h_n such that for each n, one

- Computes $T_{max,n}$
- If $T_{max,n} \leq h_n$ then concludes that the process is in control and continue to the next reading
- If $T_{max,n} > h_n$ then concludes that the process has shifted and stop the process for diagnosis. Estimate the epoch of the shift by the maximizing k.

Then the question is to determine the sequence h_n. The ideal h_n is such that the probability of false alarm at any n is a constant α. Because then the run length follows a geometric distribution of parameter α, making the in-control average run length $1/\alpha$. So h_n is from now denoted by $h_{n,\alpha}$ to reflect this. The problem being that $h_{n,\alpha}$ is depending of the distribution of $T_{k,n}$ under the null hypothesis. However the exact distribution is very complicated because its depends on both k and n, so as soon as n is relatively big, it is impossible to compute it. So the idea is to process by simulation.

Hawkins et al underline the fact that for small α that are needed in SPC, the method as sense only after what they call "warm-up sequence", a sequence during which the data are not tested. They said that for a warm-up sequence of length 14, it is possible to choose any $\alpha > 1/3432$.

2.3 Zhou et al Method [13]

An other method based on this is the one of Zhou et al. They consider the same statistic $T_{k,n}$, but instead of directly looking at it they use an exponentially weighted moving average smooth version defined by

$$\begin{aligned} E_{k,n} &= (1-\lambda)E_{k,n-1} + \lambda T_{k,n} \\ W_{max,n} &= \max_{1\leq k\leq n-1} |E_{k,n}| \end{aligned}$$

In their paper, they suggest that $\lambda = 0.2$ is a good choice in general. A change point is declared if $W_{max,n}$ exceeds a control limit chosen to fix the false alarm rate at a constant α and again obtain by simulations.

2.4 Remarks on the methods above

First, a nice property in order to use those control charts is that the computation of the statistics of interest $T_{k,n}$ and $E_{k,n}$ scaled in both case linearly.

Hawkins et al perform a comparison study of the two methods. They concluded that the fact of having used a exponentially moving average filter leads to small benefits in case of a small shift, but on the other side it reduces its response for shift above 0.75σ.

2.5 A modification of CUSUM algorithm : comparison to Hawkins et al method

In their paper [14] Lau et al, consider a different approach more general that a change in mean. They consider the classical quick change detection model,

Let $X_1, \ldots, X_n$ be a random sequence of observations taking values in $\mathbb{R}$, such that

$$\begin{cases} X_t \sim F_0 & \text{iid for all t } \leq \nu, \\ X_t \sim F_1 & \text{iid for all t } \leq \nu \end{cases}$$

Where τ is an unknown deterministic change point and F_0, F_1 two probability distribution absolutely continuous with respect to (w.r.t.) the Lebesgue measure such that $F_0 \neq F_1$. Moreover they assume F_0 the pre-change distribution to be known while F_1 the post-change distribution is not.

The also consider that F_1 is in $D(F_0, N)$ (definition below), the set of distributions distinguishable w.r.t. (F_0, N), where N is an integer.

Definition : Let $I_1 = (-\infty, z_1], I_2 = (z_1, z_2], \ldots, I_N = (z_{N-1}, \infty)$, be a set of N intervals such that for each $i \in \{1, \ldots, N\}$, one has $\int_{I_i} dpi(x) = \frac{1}{N}$. A probability distribution μ absolutely continuous to a probability distribution π is distinguishable from π w.r.t. N if there exists $i \in \{1, \ldots, N\}$ such that $\int_{I_i} d\pi(x) \neq \int_{I_i} d\mu(x)$. The set $D(\pi, N)$ is the set of all probability distributions μ distinguishable from π for the given N.

For lighten the notation, in this part, one denotes $\pi = F_0$ and $\mu = F_1$

Since $\mu \in D(\pi, N)$, it is possible to define the quantities

$$\mu_N(k) = \int_{I_k} d\mu(x), \;\; \pi_N(k) = \int_{I_k} d\pi(x) = \frac{1}{N}$$

One also uses the notations

$$\mu_N(x) = \mu_N(i), \;\; \pi_N(x) = \pi_N(i)$$

Where i is the unique integer such that $x \in I_i$.

If μ_N and π_N are both known, comparing the log-likelihood ratio of $\{\nu \leq t\}$ against $\{\nu > t\}$given the observations $x_1, \ldots, x_n$ leads to the Page's CUSUM test :

$$\tau(\gamma) = inf\{t : S(t) > \gamma\}$$

Where

$$\begin{aligned} S(t) &= log \frac{\underset{1 \leq k \leq t+1}{max} \prod_{i=1}^{k-1} \pi_N(x_i) \prod_{i=k}^{t} \mu_N(x_i)}{\prod_{i=1}^{t} \pi_N(x_i)} \\ &= \underset{1 \leq k \leq t+1}{max} \sum_{i=k}^{t} log \frac{\mu_N(x_i)}{\pi_N(x_i)} \end{aligned}$$

Because here μ_N is not known, it can be replaced by its maximum likelihood estimator

$$\mu_N^{k:t}(i) = \frac{|\{x_r : k \leq r \leq t \; and \; x_r \in I_i\}|}{t - (k-1)}$$

Where $|E|$ denotes the cardinality of the set E.

Then they said that

$$S(t) \approx \underset{1 \leq k \leq t+1}{max} \sum_{i=k}^{t} log \frac{\mu_N^{k:t}(x_i)}{\pi_N(x_i)}$$

To prevent over-fitting on the observations $x_k, \ldots, x_t$ on the maximum likelihood estimator of $\mu_N^{k:t}$, they propose not to include the current observation x_t in the estimation of μ and they introduce a fixed positive constant R to prevent that $\mu_N^{k:t-1}(x_t) = 0$, when x_t is the first observation occurring in the interval. For those reasons they propose as estimator

$$\hat{\mu}_N^{k:t-1}(i) = \begin{cases} \frac{R + |\{x_r : k \leq r \leq t-1 \; and \; x_r \in I_i\}|}{N x R + t - (k-1)} & if \; k \leq t-1 \\ \frac{1}{N} & otherwise \end{cases}$$

And then the statistic becomes

$$S(t) \approx \max_{1\leq k\leq t+1} \sum_{i=k}^{t} log \frac{\hat{\mu}_N^{k:t-1}(x_i)}{\pi_N(x_i)}$$

Finally, they propose a **Binned Generalized CUSUM** statistic $\tilde{S}$ and the associated test τ as follows

$$\begin{aligned} S(t) &= \max_{\lambda_{t-1}\leq k\leq t+1} \sum_{i=k}^{t} log \frac{\mu_N^{\lambda_{t-1}:i-1}(x_i)}{\pi_N(x_i)} \\ \lambda_{t-1} &= max \left\{ \underset{\lambda_{t-2}\leq k\leq t}{argmax} \sum_{i=k}^{t-1} log \frac{\mu_N^{\lambda_{t-2}:i-1}(x_i)}{\pi_N(x_i)} \right\} \\ \tau(\gamma) &= inf\{t : S(t) > \gamma\} \end{aligned}$$

2.6 Empirical comparison with the Hawkins method

In their paper, Lau et al perform a comparison based on ADD of the method with the one of Hawkins et al in term of mean shift detection and they obtain as results that their method perform better whatever the size of the shift. Moreover they perform an equivalent study to detect a change in variance of a normal distribution. As it can be expected, the Hawkins method does not perform well since it is not is aim, but the method proposed by Lau et al work also well in this case.

3 Other methods of change points detection

The first two methods introduced are only working in order to detect a change in mean and the last one assume the pre-change distribution to be known. So those method cannot be applied in all situations.

A classical method to do fully nonparametric change point detection is based on log likelihood ratio has the last presented here are very used. But in nonparametric case, one first needs to estimate the density in order to compute the statistic. So there is classical methods to do it.

3.1 A c.d.f. based method

Maybe the most intuitive approach in a nonparametric approach is to look at the empirical cumulative distribution function. That is what is done in by Zou et al[15].

Assume that one has data $x_1, \ldots, x_n \in \mathbb{R}$. The aim is to find a segmentation following the notations of the problem (1).

One denotes $\hat{F}_i$ the estimation of the c.d.f. of the i-th segment defines by

$$\hat{F}_i(t) = \frac{1}{\tau_i - \tau_{i-1}} \sum_{j=\tau_{i-1}+1}^{\tau_i} \left(\mathbb{1}_{x_j<t} + 0.5 \times \mathbb{1}_{x_j=t} \right)$$

If one has n data points that are iid with c.d.f. F(t), then for a fixed value of t, $n\hat{F}(t) \sim Binomial(n, F(t))$. Thus the log-likelihood of F(t) is given by

$$n[\hat{F}(t)log(F(t)) + (1 - \hat{F}(t))log(1 - F(t))]$$

Which is maximized by the empirical c.d.f., so minus the maximum value of this log-likelihood can be used as a segment cost function in a way that for the segment i, one has cost function define has

$$-L_{np}(x_{\tau_{i-1}+1:\tau_i}|t) = (\tau_i - \tau_{i-1})[\hat{F}_i(t)log\hat{F}_i(t) + (1 - \hat{F}_i(t))log(1 - \hat{F}_i(t))]$$

One deduces a cost of a segmentation as the sum of the segment costs. Thus to segment the data with m change points one minimise

$$-\sum_{i=1}^{m+1} L_{np}(x_{\tau_{i-1}+1:\tau_i}|t)$$

To overcome the problem of evaluated the c.d.f. only for a fixed t, Zou et al considered the following cost function for a segment u:v,

$$\int_{-\infty}^{\infty} -L_{np}(x_{u:v}|t)dw(t)$$

With as weight, $dw(t) = [F(t)(1-F(t))]^{-1}dF(t)$. In practice, the c.d.f. of the full data is unknown so it is approximate. Finally, the objective function for a given m is given by

$$Q_{NMCD}(\tau_{1:m}|x_{1:n}) = -n\sum_{i=1}^{m+1}\sum_{t=1}^{n}(\tau_i - \tau_{i-1}) \times \frac{\hat{F}_i(t)log\hat{F}_i(t) + (1-\hat{F}_i(t))log(1-\hat{F}_i(t))}{(t-0.5)(n-t+0.5)}$$

As in practice m is not known, Zou et al suggest estimating m by using the Bayesian Information criterion (BIC)[16]. Finally, they minimize

$$BIC = \min_{m|\tau_1,\ldots,\tau_m} [Q_{NMCD}(\tau_{1:m}|x_{1:n}) + m\xi]$$

Where ξ_n is a sequence going to infinity.

As described here the method as a complexity in $\mathcal{O}(Mn^2+n^3)$, with M the maximum number of possible change points. For a quickest version of this algorithm see the work of Haynes et al [17].

3.2 A kernel method

Another classical statistical method for nonparametric estimation is estimation by Kernel. So there are also widespread method in change points detection.

A notable advantage of this method over the previous ones is that it works also in the multivariate case. Unlike the previous method, Kernel change points (KCP) detection methods often use other metrics to assess there is a change point in the data.

One of them is the running maximum partition strategy. Consider the case of one change point detection, i.e. **problem (1)**. The idea is to select the partition of a segment of observations which leads to a maximum heterogeneity between the two segments. Denoting $\Delta_{\tau,n}$ the heterogeneity between the segments $\{X_1,\ldots,X_\tau\}$ and $\{X_{\tau+1},\ldots,X_n\}$, the aim is to compute $\underset{1<\tau<n}{max}\,\Delta_{\tau,n}$ to test for the presence of a change point and provides an estimator of the true change point instant.

This strategy is considered by Bach et al [18]. To measure the heterogeneity, they choose the kernel Fisher discriminant ratio (KFDR) define as below.

First define the empirical means elements and covariance operators as

$$\begin{aligned}\hat{\mu}_{i:j} &= \frac{1}{j-i+1}\sum_{l=i}^{j} k(X_l,\cdot)\\ \hat{\Sigma}_{i:j} &= \frac{1}{j-i+1}\sum_{l=i}^{j}[k(X_l,\cdot)-\hat{\mu}_{i:j}]\otimes[k(X_l,\cdot)-\hat{\mu}_{i:j}]\end{aligned}$$

Where k is a bounded kernel and the reproducing kernel Hilbert space associated with k is dense in L^2. And the KFDR is defined as :

$$KFDR_{k,n,\gamma}(X_1,\ldots,X_n) = \frac{k(n-k)}{n}\left\|\left(\frac{k}{n}\hat{\Sigma}_{1:k}+\frac{n-k}{n}\hat{\Sigma}_{k+1:n}+\gamma I\right)^{-1/2}(\hat{\mu}_{k+1:n}-\hat{\mu}_{1:k}\right\|^2$$

Finally they define their statistic as :

$$T_{n:\gamma} = \max_{a_n<\tau<b_n} \frac{KFDR_{\tau,n,\gamma} - d_{1,\tau,n,\gamma}(\hat{\Sigma}^W_{\tau,n})}{\sqrt{2}d_{2,\tau,n,\gamma}(\hat{\Sigma}^W_{\tau,n})}$$

Where $n\hat{\Sigma}^W_{k,n} = k\hat{\Sigma}_{1:k} + (n-k)\hat{\Sigma}_{k+1:n}$ and $d_{1,k,n,\gamma}(\hat{\Sigma}^W_{k,n} = Tr\left\{(\hat{\Sigma}^W_{k,n} + \gamma I)^{-1}\hat{\Sigma}^W_{k,n}\right\}$, $d_{2,k,n,\gamma}(\hat{\Sigma}^W_{k,n}) = Tr\left\{(\hat{\Sigma}^W_{k,n} + \gamma I)^{-2}(\hat{\Sigma}^W_{k,n})^2\right\}$. Moreover a_n and b_n are such that $a_n > 1$, $b_n < n$ to avoid problems in the neighborhood of the interval boundaries.

The notable property of their method is the following :

The test : $\max_{a_n<\tau<b_n} T_{n,\gamma}(\tau) \geq t_{1-\alpha}(\Sigma,\gamma)$ has a false alarm rate of α as n tends to infinity.

Where $t_{1-\alpha}(\Sigma,\gamma)$ is the quantile $1-\alpha$ of the limit distribution of $\max_{a_n<\tau<b_n} T_{n,\gamma}(\tau)$ under H_0 and can be obtained by Monte Carlo simulation.

Moreover they prove that the power of their test tends to 1 as n goes to infinity.

There exists many more other approaches to change point detection with kernels [19, 20].

An interesting method is the one proposed by Garreau and Arlot [21] based on model selection with a penalized kernel empirical criterion, and for which they give non asymptotic results on the ability of the method to detect the true number of change points.

3.3 A divergence based method

It is possible to perform direct ratio estimation without a step of estimating the density in a non parametric way [22]. They proposed a method based on f-divergence also called $\phi-divergence$ [23, 24]. The advantage of this kind of method relative to the previous ones is that as the kernel methods they can be used for multivariate data and they suffer less of the problems induced by high-dimensional data.

So let's consider **problem (1)**.

In their work Liu et al consider dissimilarity measure of the form

$$D(F_0\|F_1) + D(F_1\|F_0)$$

With $D(F\|F')$ of the form

$$D(F\|F') = \int f'(x)\phi(\frac{f(x)}{f'(x)})dx$$

Where ϕ is a convex function such that $\phi(1) = 0$ and f,f' are the density functions (assumed to be strictly positive) of F and F' respectively.

The $\phi-divergence$ includes various divergences such as the Kullback-Leibler (KL) divergence by $\phi(t) = t\ log(t)$ or the χ^2 divergence with $\phi(t) = (t-1)^2$.

As said in introduction, because the densities are unknown a naive approach is two estimate them and plug the estimation in the definition of the $\phi-divergence$ but because density estimation is an hard problem, it is not reliable in practice. So a method based on direct density ratio estimation is explored. As said in its name, the idea is to compute directly the density ratio without looking for the two densities.

Their exist various method of directly estimating the density ratio. One can cite as examples the KL importance estimation procedure (KLIEP) [25], the unconstrained least-squares importance fitting (uLSIF) [26] and the relative uLSIF (RuLSIF) [27].

One will focus on the **KLIEP** method [25].

In this case, the density ratio $\frac{F_0}{F_1}$ is modeled by the kernel model:

$$g(X;\theta) = \sum_{l=1}^{n} \theta_l k(X, X_l)$$

Where θ represent the parameters to be learned from the data samples, and k is a kernel basis function. In practice, the Gaussian kernel is used $k(x,y) = exp\left(-\frac{\|x-y\|^2}{2\sigma^2}\right)$, where $\sigma > 0$ is determined by cross-validation.

The parameters θ in the model $g(X,\theta)$ are determined so that it minimize

$$\begin{aligned} KL(F_0\|F_1) &= \int f_0(x) log\left(\frac{f_0(x)}{f_1(x)g(x;\theta)}\right) dx \\ &= \int f_0(x) log\left(\frac{f_0(x)}{f_1(x)}\right) dx - \int f_0(x) log(g(x;\theta)) dx \end{aligned}$$

As the first term does not depends on θ, it does not impact the estimation and finally the **KLIEP** optimization problem is given by

$$\begin{array}{l} \underset{\theta}{max}\ \frac{1}{n}\sum_{i=1} log\left(\sum_{l=1}^{n}\theta_l k(X_i, X_l)\right) \\ s.t. \frac{1}{n}\sum_{j=1}^{n}\sum_{l=1}^{n}\theta_l k(x_j, x_l) = 1\ and\ \theta_1, \ldots, \theta_n \geq 0 \end{array}$$

The constraint are due to the fact that $g(X,\theta)F_1(X)$ should be a probability density function and that the density ratio should be positive.

Because it is a convex optimization problem, the global optimal solution $\hat{\theta}$ can be obtained by gradient-projection iteration [28], and a density-ratio estimator is given by

$$\hat{g}(X) = \sum_{l=1} \hat{\theta}_l k(X, X_l)$$

A remarkable result about this method is that it achieve the optimal non-parametric convergence rate[29].

And an approximation of the KL divergence is given by

$$\hat{KL} = \frac{1}{n}\sum_{i=1}^{n} log\hat{g}(X_i)$$

Then this statistic is compared to some threshold to assess if there is a change point or not.

However there is at my knowledge no theoretical results about it but this method as shown working well in practice [22]. Even there is another method uLISF and RuLISF that have the same performance but are more robust and quicker. Moreover an other idea to fasten the method and make it more robust is to consider an α-ratio density instead of the classical ratio density [30].

4 Conclusion

Various methods are presented here with different approaches to the quickest change point detection problem in a nonparametric setting. This review is not intended to be exhaustive but just to present different points of view. Moreover to be complete this study would have needed experimental study to properly compare the performance of the methods presented here both in terms of ability to quickly detect a change and in terms of complexity.

References

[1] Douglas M. Hawkins, Peihua Qiu, and Chang Wook Kang. The changepoint model for statistical process control. *Journal of Quality Technology*, 35(4):355–366, 2003.

[2] Ping Yang, Guy Dumont, and Mark Ansermino. Adaptive change detection in heart rate trend monitoring in anesthetized children. *IEEE transactions on bio-medical engineering*, 53:2211–9, 12 2006.

[3] M. Chowdhury, Sid Ahmed Selouani, and D. O'Shaughnessy. Bayesian on-line spectral change point detection: A soft computing approach for on-line asr. *International Journal of Speech Technology*, 15, 03 2012.

[4] Samaneh Aminikhanghahi and Diane Cook. A survey of methods for time series change point detection. *Knowledge and Information Systems*, 51, 09 2016.

[5] Charles Truong, Laurent Oudre, and Nicolas Vayatis. A review of change point detection methods. 01 2018.

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

[7] Gary Lorden. Procedures for reacting to a change in distribution. *The Annals of Mathematical Statistics*, 42, 12 1971.

[8] George V. Moustakides. Optimal stopping times for detecting changes in distributions. *The Annals of Statistics*, 14(4):1379–1387, 1986.

[9] Moshe Pollak. Optimal detection of a change in distribution. *The Annals of Statistics*, 13, 03 1985.

[10] Tze Leung Lai. Information bounds and quick detection of parameter changes in stochastic systems. *IEEE Trans. Inf. Theor.*, 44(7):2917–2929, September 2006.

[11] W.J. Conover. *Practical Nonparametric Statistical*. John Wiley Sons Inc., New York, 1999.

[12] Douglas M. Hawkins and Qiqi Deng. A nonparametric change-point control chart. *Journal of Quality Technology*, 42(2):165–173, 2010.

[13] Chunguang Zhou, Changliang Zou, Yujuan Zhang, and Zhaojun Wang. Nonparametric control chart based on change-point model. *Statistical Papers*, 50:13–28, 02 2009.

[14] T. S. Lau, W. P. Tay, and V. V. Veeravalli. Quickest change detection with unknown post-change distribution. In *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 3924–3928, March 2017.

[15] Changliang Zou, Guosheng Yin, Long Feng, and Zhaojun Wang. Nonparametric maximum likelihood approach to multiple change-point problems. 2014.

[16] Gideon Schwarz. Estimating the dimension of a model. *The Annals of Statistics*, 6(2):461–464, 1978.

[17] Kaylea Haynes, Paul Fearnhead, and Idris A. Eckley. A computationally efficient nonparametric approach for changepoint detection. *Statistics and Computing*, 27(5):1293–1305, Sep 2017.

[18] Zaïd Harchaoui, Eric Moulines, and Francis R. Bach. Kernel change-point analysis. In D. Koller, D. Schuurmans, Y. Bengio, and L. Bottou, editors, *Advances in Neural Information Processing Systems 21*, pages 609–616. Curran Associates, Inc., 2009.

[19] C. Truong, L. Oudre, and N. Vayatis. Supervised kernel change point detection with partial annotations. In *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 3147–3151, May 2019.

[20] Shuang Li, Yao Xie, Hanjun Dai, and Le Song. M-statistic for kernel change-point detection. In C. Cortes, N. D. Lawrence, D. D. Lee, M. Sugiyama, and R. Garnett, editors, *Advances in Neural Information Processing Systems 28*, pages 3366–3374. Curran Associates, Inc., 2015.

[21] Damien Garreau and Sylvain Arlot. Consistent change-point detection with kernels. 2016.

[22] Song Liu, Makoto Yamada, Nigel Collier, and Masashi Sugiyama. Change-point detection in time-series data by relative density-ratio estimation. *Neural Networks*, 43:72 – 83, 2013.

[23] I. Csiszár. Eine informationstheoretische ungleichung und ihre anwendung auf den beweis der ergodizitat von markoffschen ketten. *Magyar. Tud. Akad Mat. Kutato Int. Kozl.*, pages 85–108, 1963.

[24] Noel Cressie and Timothy R. C. Read. Multinomial goodness-of-fit tests. *Journal of the Royal Statistical Society. Series B (Methodological)*, 46(3):440–464, 1984.

[25] Masashi Sugiyama, Shinichi Nakajima, and Hisashi Kashima. Direct importance estimation with model selection and its application to covariate shift adaptation. 01 2007.

[26] Takafumi Kanamori, Shohei Hido, and Masashi Sugiyama. A least-squares approach to direct importance estimation. *Journal of Machine Learning Research*, 10:1391–1445, 07 2009.

[27] Makoto Yamada, Taiji Suzuki, Takafumi Kanamori, Hirotaka Hachiya, and Masashi Sugiyama. Relative density-ratio estimation for robust distribution comparison. *Neural computation*, 25:1324–70, 05 2013.

[28] J. B. Rosen. The gradient projection method for nonlinear programming. part i. linear constraints. *Journal of the Society for Industrial and Applied Mathematics*, 8(1):181–217, 1960.

[29] X. Nguyen, M. J. Wainwright, and M. I. Jordan. Estimating divergence functionals and the likelihood ratio by convex risk minimization. *IEEE Transactions on Information Theory*, 56(11):5847–5861, Nov 2010.

[30] Anis Hamadouche, Abdelmalek Kouadri, and Azzedine Bakdi. A modified kullback divergence for direct fault detection in large scale systems. *Journal of Process Control*, 59:28 – 36, 2017.