Changepoint Detection in Time Series Analysis with At Most Two Changepoints

Xinyuan Huang

Given a time series data with at most two unknown changepoints (AMTC), this algorithm provides a solution to find the changepoints in following procedures:

1. Suppose that the series has exactly two changepoints, the algorithm finds the most possible two changepoints and test if they are significant using null hypothesis.
2. If the two changepoints are not significant, the problem becomes a series with at most one changepoint (AMOC). The algorithm finds the most possible point and test the significance.
3. If no significant changepoint is detected, the series has no changepoint.

1. Two changepoints

Consider a linear series with two unknown changepoints, a regression model describing this scenario for data (X_t) over $t = 1, 2, \dots, n$ is

$$X_t = \mu + \beta t + \delta_t + \epsilon_t$$

where μ is the initial intercept, β is a linear trend, (ϵ_t) is a series of independent and identically distributed zero-mean error with variance σ^2. Assuming that the two changepoints occur at time c_1 and c_2, the intercept-shift factor (δ_t) obeys

$$\delta_t = \begin{cases} 0, & 1 \le t \le c_1, \\ \Delta_1, & c_1 < t \le c_2, \\ \Delta_1 + \Delta_2, & c_2 < t \le n. \end{cases}$$

and Δ's are the shift values after changepoints. We can show that the estimates that minimize the sum of squares

$$\sum_{\mathrm{t=1}}^{\mathrm{n}} \left\{ X_t - \left[\mu + \beta t + \Delta_1 1_{(t > c_1)} + \Delta_2 1_{(t > c_2)} \right] \right\}^2$$

are

$$\hat{\beta}(c_1, c_2) = \frac{\sum_{t=1}^{c_1} X_t \left[t - \frac{c_1 + 1}{2} \right] + \sum_{t=c_1+1}^{c_2} X_t \left[t - \frac{c_2(c_2+1) - c_1(c_1+1)}{2(c_2 - c_1)} \right] + \sum_{t=c_2+1}^{n} X_t \Big[t - \frac{n(n+1) - c_2(c_2 + 1}{2(n - c_2)}}{\frac{c_1(c_1+1)(c_1-1)}{12} + \frac{(c_2-c_1)(c_2-c_1+1)(c_2-c_1-1)}{12} + \frac{(n-c_2)(n-c_2+1)(n-c_2-1)}{12}}$$

$$\hat{\mu}(c_1, c_2) = \frac{1}{c_1} \sum_{t=1}^{c_1} \left[X_t - \hat{\beta}(c_1, c_2) t \right]$$

$$\widehat{\Delta_1}(c_1, c_2) = \frac{1}{c_2 - c_1} \sum_{t=c_1+1}^{c_2} \left[X_t - \hat{\beta}(c_1, c_2) t \right] - \hat{\mu}(c_1, c_2)$$

$$\widehat{\Delta_2}(c_1, c_2) = \frac{1}{n - c_2} \sum_{t=c_2+1}^{n} \left[X_t - \hat{\beta}(c_1, c_2) t \right] - \hat{\mu}(c_1, c_2) - \widehat{\Delta_1}(c_1, c_2)$$

We use genetic algorithm to discover the optimal solution [3, 5]. We generate some pairs of c_1 and c_2's as the initial solutions, the mutation and crossover operations can be defined as follow,
Mutation: add or subtract c_1 or c_2 by a random small amount;
Crossover: exchange the c_1 and c_2 from two selected solutions.

In my trial, this algorithm converges fast (typically in ~5 iterations) but may fall in sub-optimal solution. To achieve a better performance, we need a relatively high volume of initial choices.

We should always notice that the choices of the changepoints should not be too close to the boundaries or too close to each other.

After obtaining a solution for c_1 and c_2, we can conduct F-test. Suppose that there is no changepoint in the series, the regression model reduces to

$$X_t = \mu_{\text{Red}} + \beta_{\text{Red}} t + \epsilon_t$$

Let RSS_{Full} be the sum of squared residuals of the model with two changepoints, RSS_{red} be the sum of squared residuals of the reduced model, we can have the F-statistics

$$F_{c_1,c_2} = \frac{(\,RSS_{\text{Red}} - RSS_{\text{Full}})/2}{RSS_{\text{Full}}/(n-4)}$$

If changepoints exist at time c_1 and c_2, F_{c_1,c_2} should be large [4]. Thus, the test is concluded when

$$F_{\max} = \max_{1<c_1<c_2<n} F_{c_1,c_2}$$

is too large. According to Hinkley [2], $F_{\max}$ approximately follows $F_{4,n-4}$ considering c_1 and c_2 are two unknown parameters.

2. One changepoint

If we fail to reject the null hypothesis (i.e. the two changepoints case does not hold), the algorithm will detect if there exists one changepoint. According to Lund [1], the regression model becomes $X_t = \mu + \beta t + \delta_t + \epsilon_t$ where

$$\delta_t = \begin{cases} 0, 1 \le t \le c, \\ \Delta, c < t \le n. \end{cases}$$

and the OLS problem becomes minimizing

$$\sum_{t=1}^{n} \left\{X_t - \left[\mu + \beta t + \Delta 1_{(t>c)}\right]\right\}^2$$

We can infer that

$$\hat{\beta}(c) = \frac{\sum_{t=1}^{c} X_t \left[t - \frac{c+1}{2}\right] + \sum_{t=c+1}^{n} X_t \left[t - \frac{n(n+1) - c(c+1)}{2(n-c)}\right]}{\frac{c(c+1)(c-1)}{12} + \frac{(n-c)(n-c+1)(n-c-1)}{12}}$$

$$\hat{\mu}(c) = \frac{1}{c} \sum_{t=1}^{c} \left[X_t - \hat{\beta}(c)t\right]$$

$$\widehat{\Delta}(c) = \frac{1}{n-c} \sum_{t=c+1}^{n} \left[X_t - \hat{\beta}(c)t\right] - \hat{\mu}(c)$$

A changepoint is suggested at time c when $\widehat{\Delta}(c)$ is far from zero. A t-statistics can be used that

$$T(c)=\frac{\widehat{\Delta}(c)}{\widehat{\mathrm{Var}}\left[\widehat{\Delta}(c)\right]^{1/2}}$$

where

$$\widehat{\mathrm{Var}}\left[\widehat{\Delta}(c)\right]=\frac{\hat{\sigma}^2}{c\left(1-\frac{c}{n}\right)\left[1-3\frac{c(n-c)}{n^2-1}\right]}$$

Let

$$\widetilde{D}_h^*=\max_{h\le(c/n)\le1-h}T(c)^2$$

where h is used to prevent the potential changepoint from getting too close to the boundaries. $\widetilde{D}_h^*$ follows certain distribution given by Lund [1] and $\mathrm{Prob}\left(\widetilde{D}_h^*\ge 20.114\right)=99.9\%$ when using $h=0.05$. We can reject the null hypothesis when $\widetilde{D}_h^*$ is large enough so that a changepoint is detected at $c=\underset{h\le(c/n)\le1-h}{\arg\max}\;T(c)^2$

3. Prospective

Generally, if N changepoints are given at time $c_1, c_2, \dots, c_N$, let $c_0=0, c_{N+1}=n$, denote $\boldsymbol{c}=(c_1, c_2, \dots, c_N)$, we have

$$\hat{\beta}(\boldsymbol{c})=\frac{\sum_{i=0}^{N}\left\{\sum_{t=c_i+1}^{c_{i+1}}X_t\left[t-\frac{c_{i+1}(c_{i+1}-1)-c_i(c_i-1)}{2(c_{i+1}-c_i)}\right]\right\}}{\sum_{i=0}^{N}\frac{(c_{i+1}-c_i)(c_{i+1}-c_i+1)(c_{i+1}-c_i-1)}{12}}$$

$$\hat{\mu}(\boldsymbol{c})=\frac{1}{c_1}\sum_{t=1}^{c_1}\left[X_t-\hat{\beta}(\boldsymbol{c})t\right]$$

$$\widehat{\Delta}_i(\boldsymbol{c})=\frac{1}{c_{i+1}-c_i}\sum_{t=c_i+1}^{c_{i+1}}\left[X_t-\hat{\beta}(\boldsymbol{c})t\right]-\hat{\mu}(\boldsymbol{c})-\sum_{j=1}^{i-1}\widehat{\Delta}_j(\boldsymbol{c})\qquad i=1,2,\dots,N$$

However, hypothesis testing is extremely hard to infer.

References

[1] Gallagher, Colin, Robert Lund, and Michael Robbins. "Changepoint detection in climate time series with long-term trends." Journal of Climate 26.14 (2013): 4994-5006.

[2] Hinkley, David V. "Inference about the intersection in two-phase regression." Biometrika 56.3 (1969): 495-504.

[3] Li, Shanghong, and Robert Lund. "Multiple changepoint detection via genetic algorithms." Journal of Climate 25.2 (2012): 674-686.

[4] Lund, Robert, and Jaxk Reeves. "Detection of undocumented changepoints: A revision of the two-phase regression model." Journal of Climate 15.17 (2002): 2547-2554.

[5] Segaran, Toby. Programming collective intelligence: building smart web 2.0 applications. " O'Reilly Media, Inc.", 2007. APA