UNIVERSITETET I OSLO

Matematisk Institutt

EXAM IN:	**STK 4444SP, special curriculum: Statistical Changepoint Methods**
FOR:	**Elisabeth Nesheim Hagen**
WITH:	**Nils Lid Hjort**
TIME FOR EXAM:	**15.v.–29.v.2019**

This is the exam project set for STK 4444SP, special curriculum on statistical changepoint methods, spring semester 2019. It is made available for the candidate as of *Wednesday 15 May 12:00*, and she should submit her written report by *Wednesday 29 May 13:00* (or earlier), electronically as a pdf, to Nils Lid Hjort. There will also be a *conversation* with the candidate, Hjort, and a colleague, with a blackboard nearby, on *Friday 31 May*; that meeting starts with the candidate presenting her project report, and might then touch both the report and other aspects of the special curriculum.

The candidate is required to work by herself, i.e. independently of others. Importantly, in her report the candidate should also include a one-page summary of the work carried out, and this should also contain a brief self-assessment of its quality. Kandidaten er hjertelig velkommen til å skrive på norsk, om hun vil, men hun kan også velge å skrive på engelsk.

This exam set contains three exercises and comprises five pages, including an Appendix.

Exercise 1

Consider independent observations $Y_1, \ldots, Y_n$ from the normal distribution with parameters (ξ, σ). For a relevant sequence of τ, say $c \le \tau \le d$, form from these the process

$$H_n(\tau) = \frac{\widehat{\xi}_L - \widehat{\xi}_R}{\{\widehat{\sigma}_L^2/\tau + \widehat{\sigma}_R^2/(n-\tau)\}^{1/2}} \quad \text{for } \tau = c, c+1, \ldots, d-1, d.$$

Here $\widehat{\xi}_L$ and $\widehat{\sigma}_L$ are the empirical mean and standard deviation for the left stretch of data $y_1, \ldots, y_L$, and correspondingly with $\widehat{\xi}_R$ and $\widehat{\sigma}_R$ for the right stretch $y_{\tau+1}, \ldots, y_n$. It is convenient to transform the time-scale via $s = \tau/n$, so that the cousin process $H_n^*(s) = H_n(\tau)$ is defined for $s \in [c/n, d/n]$, inside the unit interval.

(a) Show that for each s, $H_n^*(s)$ tends to the standard normal distribution.

(b) Working with the empirical partial-sum process

$$A_n(s) = \frac{1}{\sqrt{n}} \sum_{i \le [ns]} \frac{Y_i - \xi}{\sigma} \quad \text{for } s \in [0,1],$$

and its convergence to the Brownian motion process $W(s)$, show that

$$H_n^*(s) \to_d H^*(s) = (1-s)^{1/2} \frac{W(s)}{s^{1/2}} - s^{1/2} \frac{W(1) - W(s)}{(1-s)^{1/2}} \quad \text{for } s \in (0,1).$$

Here you may use Donskers's Theorem, see this exam set's Appendix, with natural consequences.

(c) Show also that the limit process above may be expressed as

$$H^*(s) = \frac{W^0(s)}{\sqrt{s(1-s)}} \quad \text{for } s \in (0,1),$$

where $W^0(s) = W(s) - sW(1)$ is a Brownian bridge, with variance $s(1-s)$ and covariance

$$\text{cov}\{W^0(s), W^0(t)\} = s(1-t) \quad \text{for } s \le t.$$

Find the variance of $H^*(s)$ and also the correlation between $H^*(s)$ and $H^*(t)$.

(d) These results may be used to construct tests for the null hypothesis H_0 that there has been no change in an ongoing process. Explain first that for a given subinterval $[c_0, d_0]$ of the unit interval,

$$M_n = \max_{nc_0 \le \tau \le nd_0} |H_n(\tau)| = \max_{c_0 \le s \le d_0} |H_n^*(s)|,$$

and then that under the null hypothesis,

$$M_n \to_d M = \max_{c_0 \le s \le d_0} \frac{|W^0(s)|}{\sqrt{s(1-s)}}.$$

(e) For $[c_0, d_0] = [0.20, 0.80]$, simulate a reasonably high number of realisations of M, perhaps using details from the Appendix, and record its 0.50, 0.90, 0.95 quantiles.

(f) To get a picture of how well such a test may work, consider a setup where there actually is a changepoint, with $Y_1, \ldots, Y_\tau$ from $\text{N}(0,1)$ and $Y_{\tau+1}, \ldots, Y_n$ from $\text{N}(\delta, 1)$, with $n = 100$ and $\tau = 50$. Study the test which rejects the null hypothesis of no change when $M_n \ge m_0$, with m_0 the 0.95 quantile in the limit distribution M, found in the previous point. Then use simulation to approximate the power function

$$\pi_n(\delta) = P_\delta\{M_n \ge m_0\},$$

for a reasonable grid of δ values. Make a plot of this power function, and comment on your findings.

Exercise 2

In addition to testing the null hypothesis of constancy, the H_n plot of the previous exercise can potentially be used to estimate the position of a changepoint, is such is present. Consider for concreteness the model where the Y_i are taken to come from a $\text{N}(\xi, 1)$, with known standard deviation 1, with $Y_i \sim \text{N}(\xi_1, 1)$ for $i = 1, \ldots, \tau$ and $Y_i \sim \text{N}(\xi_2, 1)$ for $i = \tau + 1, \ldots, n$. Thus there are three unknown parameters in this setup; the changepoint τ and the two levels, to the left and to the right. Two estimators to consider are

$$\widehat{\tau} = \text{argmax}|H_n|, \quad \text{the position where } |H_n(\tau)| \text{ is maximal,}$$

and the maximum likelihood estimator, i.e.

$$\tau^* = \text{argmax}\{\ell_{n,\text{prof}}(\tau)\},$$

where

$$\ell_{n,\text{prof}}(\tau) = \max\{\ell_n(\tau, \xi_1, \xi_2)\text{: all } \xi_1, \xi_2\}$$

is the profiled log-likelihood function, from $\ell_n(\tau, \xi_1, \xi_2)$ which uses ξ_1 from 1 to τ and ξ_2 from $\tau + 1$ to n.

Implement algorithms for finding $\widehat{\tau}$ and τ^* for a given set of data. Work with the simple concrete case of $n = 100$ and $\tau_{\text{true}} = 50$. Conduct and report on a simple simulation experiment to compare the performance of the two estimators.

Exercise 3

The dataset `cow-bigwars-data` is being sent separately to the candidate. It consists of the four columns

$$(i, x_i, z_i, y_i) \quad \text{for } i = 1, \ldots, 51,$$

for the $n = 51$ horribly great interstate wars our world has experienced since 1823, where the number z of battle-deaths tragically has exceeded the threshold $z_0 = 7000$. Here i is the index, x_i is the point in time when the war broke out (with month and day being transformed to decimals, so that the start of the Korean War is at 1950.825, etc.), z_i the battle-deaths count, and

$$y_i = \log(z_i/z_0) \quad \text{for } i = 1, \ldots, n.$$

I have excerpted these data from the Correlates-of-War database.

I've chosen the threshold z_0 so high that z_i above this level can be seen as coming from a so-called heavy-tailed power-law distribution – read Steven Pinker's *Statistical Sightings of Better Angels* (2011) and *Enlightenment Now* (2018) for discussion of this statistical concept. The heavy-tailed-ness property can be defined and worked with in several ways. One version is as follows, in probabilistic terms: for a certain positive heavy-tail parameter θ, it holds that

$$P\{Z \geq z\} = \left(\frac{z_0}{z}\right)^{\theta} \quad \text{for } z \geq z_0.$$

The θ may be called the heavy-tail power parameter, and for low or moderate parameter values even very large outcomes z are not infrequent. The density of Z becomes proportional to $1/z^{\theta+1}$, going much more slowly to zero than for most typical distributions.

(a) Assume Z follows the distribution above. Show that $Y = \log(Z/z_0)$ must follow the exponential distribution, with the familiar density $\theta \exp(-\theta y)$ for $y \geq 0$.

(b) Consider therefore the observed $y_1, \ldots, y_n$, with the aim of checking whether the underlying θ parameter has remained constant in time, or perhaps not. Construct an appropriate *monitoring plot*, following the recipes of Hjort and Koning, 'Tests for constancy of model parameters over time' (*Journal of Nonparametric Statistics*, 2002). Comment on what you might learn from this.

(c) We may also work with a relative of the H_n process from Exercise 1, but now utilising the extra information on exponentiality. Recall that if Y is exponential with parameter θ, then its mean and variance are equal to $1/\theta$ and $1/\theta^2$. Consider now the process

$$H_n(\tau) = \frac{\bar{y}_L - \bar{y}_R}{\{\bar{y}_L^2/\tau + \bar{y}_R^2/(n-\tau)\}^{1/2}} \quad \text{for } \tau = c, c+1, \ldots, d-1, d,$$

with $\bar{y}_L$ and $\bar{y}_R$ being the means of the left and right stretches of data, as with Exercise 1. Compute and display this plot, say for $\tau = 3, \ldots, n-3$.

(d) Try to show, perhaps via variations of arguments used in Exercise 1, that the time-scaled process $H_n^*(s) = H_n(\tau)$, with $s = \tau/n$, converges in distribution to the same limit process $H(s)$ as worked with there.

(e) Show that if the null hypothesis of no change holds, then

$$K_n = \max_{c_0 n \le \tau \le d_0 n} H_n(\tau) \to_d K = \max_{c_0 \le s \le d_0} H(s).$$

(f) Choose a reasonable time interval $[c, d]$ for τ, so that c corresponds to about 1925 and d to about 1975, and transform to $c_0 = c/n$ and $d_0 = d/n$. What is the observed value of K_n? Simulate the distribution of the limit variable K, and give the upper 0.05 value. What is the p-value for the test?

(g) Use the H_n plot, or perhaps other methods, to estimate the position of a changepoint, assuming there is one. Attempt also to form a confidence curve for this changepoint. Discuss what you find.

Appendix: Donsker's Theorem

The following is one way of spelling out Donsker's Theorem (from 1951). It says that if $V_1, V_2, \ldots$ is a sequence of independent observations from the same distribution, with mean 0 and variance 1, then the empirical partial-sum process

$$A_n(s) = \frac{1}{\sqrt{n}} \sum_{i \le [ns]} V_i \quad \text{for } s \in [0,1]$$

converges in distribution to $W = \{W(s): s \in [0,1]\}$, the Brownian motion process. Here $[ns]$ is the 'integer-value' of $[ns]$, so that $[17.01] = 17$, $[16.99] = 16$, &c. The W is a zero-mean Gaussian process with independent increments, with $W(t) - W(s) \sim \mathrm{N}(0, t-s)$. The process convergence in question is taking place in the space $D[0,1]$ of all right-continuous functions $x: [0,1] \to \mathbb{R}$ with left-hand limits, with the so-called Skorokhod topology; see e.g. Billingsley's book *Convergence of Probability Measures* (1968) for the required mathematical details.

Among the crucial points here is that if $g: D[0,1] \to \mathbb{R}$ is a continuous function, then $g(A_n) \to_d g(W)$. So, for example, with $g(x) = \max\{|x(s)|: s \in [c_0, d_0]\}$, with $[c_0, d_0]$ a subinterval of the unit interval, we have

$$D_n = \max_{c_0 \le s \le d_0} |A_n(s)| \to_d D = \max_{c_0 \le s \le d_0} |W(s)|,$$

and so on. Similar remarks apply to cases where the limit distribution of a suitable empirical process is not W itself, but a relative, such as with Exercise 1. If $Z_n \to_d Z$, with Z_n an empirical process with a limit process Z, then again $g(Z_n) \to_d g(Z)$ for each continuous function $g(z)$.

Distributions such as for D above may be simulated. This is a simple little code for that purpose, for suitably high values of `m` and `sim`.

```
sval <- (1:mm)/mm
cc0 <- 0.12
dd0 <- 0.88
ok <- 1*(sval >= cc0)*(sval <= dd0)
sshort <- sval[ok==1]

keep <- 0*(1:sim)
for (ss in 1:sim)
{
vv <- rnorm(mm)
A <- cumsum(vv)/sqrt(mm)
Ashort <- A[round(cc0*mm):round(dd0*mm)]
# matplot(sshort,Ashort,type="l")
keep[ss] <- max(abs(Ashort))
}
```