Modelling trend and detecting variation

Beyond the model

Christophe POUET

ECM

Semester 8, 2014

Outlines

Multiple regression

Model and observations:

$$Y_i = g(X_i) + \xi_i, \quad i = 1, \ldots, N \text{ or } T.$$

Assumption: there exists $\beta \in \mathbb{R}^{p+1}$ such that

$$g(x) = g_\beta(x).$$

The function $g_\beta(x)$ is linear in β.
Examples:

$$\begin{aligned} g_\beta(x) &= \beta_0 + \beta_1 x \\ g_\beta(x) &= \beta_0 + \beta_1 x + \beta_2 x^2 \\ g_\beta(x) &= \beta_0 + \beta_1 x + \ldots + \beta_p x^p \\ g_\beta(x) &= \beta_0 + \beta_1 \exp(x). \end{aligned}$$

Matrix model

Goal: estimate the parameter β.

$$Y = X\beta + \xi,$$

where

$$Y = \begin{pmatrix} Y_1 \\ \vdots \\ Y_N \end{pmatrix}, \quad \xi = \begin{pmatrix} \xi_1 \\ \vdots \\ \xi_N \end{pmatrix}$$

and X is the design matrix $N \times (p+1)$.

Examples

1. If $g_\beta(x) = \beta_0 + \beta_1 x$, we have

$$X = \begin{pmatrix} 1 & X_1 \\ 1 & X_2 \\ \vdots & \vdots \\ 1 & X_N \end{pmatrix}.$$

2. If $g_\beta(x) = \beta_0 + \beta_1 x + \beta_2 x^2$, we have

$$X = \begin{pmatrix} 1 & X_1 & X_1^2 \\ 1 & X_2 & X_2^2 \\ \vdots & \vdots & \vdots \\ 1 & X_N & X_N^2 \end{pmatrix}.$$

Examples

1. If $g_\beta(x) = \beta_0 + \beta_1 x + \ldots + \beta_p x^p$, we have

$$X = \begin{pmatrix} 1 & X_1 & \ldots & X_1^p \\ 1 & X_2 & \ldots & X_2^p \\ \vdots & \vdots & \vdots & \vdots \\ 1 & X_N & \ldots & X_N^p \end{pmatrix}.$$

2. If $g_\beta(x) = \beta_0 + \beta_1 \exp(x)$, we have

$$X = \begin{pmatrix} 1 & \exp(X_1) \\ 1 & \exp(X_2) \\ \vdots & \vdots \\ 1 & \exp(X_N) \end{pmatrix}.$$

Assumptions

Observation assumption: $p < N$
Noise assumptions: ξ is white noise, i.e.

- centered: $\mathbb{E}(\xi) = 0$
- homoskedastic: $\text{var}(\xi) = \sigma^2$
- independent (or at least uncorrelated)

Design assumption: matrix X is full rank.

Least square method

Definition

The least square estimator $\hat{\beta}$ is the solution of the minimization problem

$$\min_{\beta \in \mathbb{R}^{p+1}} \sum_{i=1}^{N} (Y_i - g_\beta(X_i))^2 = \min_{\beta \in \mathbb{R}^{p+1}} \|Y - X\beta\|_2^2 .$$

Proposition

The least square estimator is

$$\hat{\beta} = (X'X)^{-1}X'Y.$$

Properties

Proposition

The least square estimator is unbiased, i.e.

$$\mathbb{E}(\hat{\beta}) = \beta.$$

We have

$$\mathbb{E}(g_{\hat{\beta}}(x)) = g_{\beta}(x).$$

Its covariance matrix is

$$var(\hat{\beta}) = \sigma^2 (X'X)^{-1}.$$

Proposition

If the noise is gaussian, we have

$$\hat{\beta} \sim \mathcal{N}(\beta, \sigma^2 (X'X)^{-1}).$$

Tests

Proposition

Sums of squares identity:

$$\begin{aligned} \left\| Y - \bar{Y}_N \mathbb{I}_N \right\|^2 &= \left\| Y - X\hat{\beta} \right\|^2 + \left\| X\hat{\beta} - \bar{Y}_N \mathbb{I}_N \right\|^2 \\ \sum_{i=1}^{N} (Y_i - \bar{Y}_N)^2 &= \sum_{i=1}^{N} (Y_i - g_{\hat{\beta}}(X_i))^2 + \sum_{i=1}^{N} (g_{\hat{\beta}}(X_i) - \bar{Y}_N)^2 \end{aligned}$$

This reads: Corrected Sum of Squares (CSS) is the sum of Residual Sum of Squares (RSS) and Model Sum of Squares (MSS).
Cochran Theorem entails (if $\beta_j = 0, 1 \leq j \leq p$ and Gaussian noise)

$$\left\| Y - X\hat{\beta} \right\|^2 / \sigma^2 \sim \chi^2(N - p - 1), \quad \left\| X\hat{\beta} - \bar{Y}_N \mathbb{I}_N \right\|^2 / \sigma^2 \sim \chi^2(p),$$

and these two sums are independent.

Testing significativity

Correlation coefficient

$$R^2 = \frac{\text{MSS}}{\text{CSS}}.$$

Testing significativity of a model, i.e.

$$H_0 : \beta_1 = \ldots = \beta_p = 0.$$

Proposition

The test statistics is

$$F = \frac{MSS/p}{RSS/(N-p-1)}$$

and under H_0 its law is Fisher $\mathcal{F}(p, N-p-1)$.

Remark: this test is only for testing significativity of the whole model.

Testing submodels

Testing problem

$$H_0 : \beta_{m+1} = \beta_{m+2} = \ldots = \beta_p = 0.$$

The submodel is

$$g_\beta(x) = \beta_0 + \beta_1 x + \ldots + \beta_m x^m.$$

The test statistics is

$$F = \frac{(\text{MSS(model)} - \text{MSS(submodel)}/(p-m)}{\text{RSS(model)}/(N-p-1)}.$$

Example:

$$g_\beta(x) = \beta_0 + \beta_1 x + \beta_2 x^2 + \beta_3 x^3.$$

We want to test

$$H_0 : \beta_2 = \beta_3 = 0.$$

The alternative is $H_1 : \beta_2 \neq 0$ or $\beta_3 \neq 0$. The dimension of the model is 4 and the dimension of the submodel is 2.

Problems

- Multiple regression leads to a general model with fixed coefficients.
- Building an adequate model needs a lot of steps: estimating, testing,...
- Assume a parametric regression function.

Non parametric regression

We want to generalize the multiple regression model:

$$Y_i = g(X_i) + \xi_i, \quad i = 1, \ldots, N.$$

Noise assumptions are the same: white noise.
Design assumption: $0 \leq X_1 < \ldots < X_N \leq 1$.
But no parametric assumption on regression function $g(\cdot)$.

Nadaraya-Watson estimator

Definition

The Nadaraya-Watson estimator is given by

$$\hat{g}_h(x) = \frac{\sum_{j=1}^{N} Y_j K\left(\frac{x-X_j}{h}\right)}{\sum_{j=1}^{N} K\left(\frac{x-X_j}{h}\right)}$$

The function K is called a kernel and the parameter h is called the bandwidth.

Examples of kernels:

$$\begin{aligned} \text{Rectangular kernel:} \quad K_R(u) &= \frac{1}{2}\mathbb{I}_{[-1,\,1]}(x), \\ \text{Gaussian kernel:} \quad K_G(u) &= \frac{1}{\sqrt{2\pi}} \exp(-\frac{x^2}{2}). \end{aligned}$$

Priestley-Chao estimator, Gasser-Müller estimator

Definition

The Priestley-Chao estimator is

$$\hat{g}_h(x) = \frac{1}{h}\sum_{i=1}^{N}(X_i - X_{i-1})\, Y_i K(\frac{x - X_i}{h}).$$

The Gasser-Müller estimator is

$$\hat{g}_h(x) = \frac{1}{h}\sum_{i=1}^{N} Y_i \int_{s_{i-1}}^{s_i} K(\frac{x-u}{h})\, du,$$

where $s_i = (X_i + X_{i+1})/2$ for $i = 1, \ldots, N-1$, $s_0 = 0$ and $s_N = 1$.

All these estimators are weighted sums: $\hat{g}(x) = \sum_{i=1}^{N} w_{i,N}(x)\, Y_i$.
Same idea as moving average but a different goal.

Kernel and bandwidth

Each method has two parameters: the bandwidth and the kernel.
Assumptions for the kernel

$$\begin{aligned} \int K(u)\, du &= 1 \\ \int u\, K(u)\, du &= 0 \\ \int K^2(u)\, du &< +\infty \\ \int u^2\, K(u)\, du &< +\infty. \end{aligned}$$

Few facts

What is important?

- The kernel is not so important.
- The bandwidth really matters!

Bandwidth's effect:

- Small bandwidth means large variance and small bias.
- Large bandwidth means small variance and large bias.

Fourier series

Consider the orthogonal system

$$\mathcal{C} = \{1, \cos(\pi x), \cos(2\pi x), \ldots\}.$$

This system is complete for $\mathcal{C}[0,1]$.
Fourier coefficients:

$$\phi_j = \int_0^1 g(x)\ \cos(\pi j x)\ dx.$$

The truncated Fourier series is

$$g_m(x) = \phi_0 + 2\sum_{j=1}^{m} \phi_j \cos(\pi j x).$$

If m goes to ∞, then $g_m(x)$ converges to $g(x)$ in mean square.

Fourier series estimator

We estimate the Fourier coefficients:

$$\hat{\phi}_j = \sum_{i=1}^{N} Y_i \int_{s_{i-1}}^{s_i} \cos(\pi j u)\ du,$$

with $s_i = (X_i + X_{i+1})/2$ for $i = 1, \ldots, N-1$, $s_0 = 0$ and $s_N = 1$.

Definition

The truncated Fourier series estimator is

$$\hat{g}_m(x) = \hat{\phi}_0 + 2\sum_{j=1}^{m} \hat{\phi}_j \cos(\pi j x).$$

The truncated series estimator can be generalized to any orthogonal system.

Fourier series estimator as weighted estimator

One can write

$$\hat{g}_m(x) = \sum_{i=1}^{N} Y_i \int_{s_{i-1}}^{s_i} K_m(x,u)\, du.$$

The kernel is

$$K_m(x,u) = 1 + 2\sum_{j=1}^{m} \cos(\pi j u)\cos(\pi j x).$$

It can also be written

$$\begin{aligned} K_m(x,u) &= D_m(x-u) + D_m(x+u), \\ D_m(t) &= \frac{\sin((2m+1)\pi\frac{t}{2})}{2\sin(\pi\frac{t}{2})}. \end{aligned}$$

The function $D_m(t)$ is called the Dirichlet kernel.

Local polynomials

It is based on the idea that a polynomial of degree 1 should be fine if the function g has two continuous derivatives.
We consider the solution of the minimization problem

$$min_{\beta_0,\beta_1} \sum_{i=1}^{N} (Y_i - \beta_0 - \beta_1(X_i - x))^2 K\left(\frac{x - X_i}{h}\right).$$

Definition

The local polynomials estimator of order 1 is

$$\hat{g}(x) = \hat{\beta}_0(x).$$

Local polynomials

Proposition

We have

$$\hat{\beta}_0(x) = \frac{\sum_{i=1}^N w_{N,i}(x) Y_i}{\sum_{i=1}^N w_{N,i}(x)},$$

where

$$\begin{aligned} w_{N,i}(x) &= K\left(\frac{x - X_i}{h}\right)(m_{N,2}(x) - (x - X_i)m_{N,1}(x)), \\ m_{N,k}(x) &= \sum_{j=1}^N K\left(\frac{x - X_j}{h}\right)(x - X_j)^k, \quad k = 1, 2. \end{aligned}$$

Remarks on local polynomials

- Instead of a local polynomial of degree 1, we can consider a polynomial of degree 0. The local polynomial estimator is exactly the Nadaraya-Watson estimator.
- One can consider higher order local polynomials. The smoother the underlying function, the higher the degree of the local polynomial can be.

Smoothing splines

Consider the set $\mathcal{W}_2[0,\ 1]$ of continuously differentiable functions on $[0,\ 1]$ with a square integrable second derivative.
Criterion

$$R_\lambda(g) = \frac{1}{N}\sum_{i=1}^{N}(Y_i - g(X_i))^2 + \lambda \int_0^1 g^{(2)}(x)^2\ dx,$$

where λ is a positive constant.

- 1st part of $R_\lambda(g)$ measures the goodness-of-fit of g.
- 2nd part of $R_\lambda(g)$ measures te smoothness of g.
- Parameter λ is a trade-off between these two parts.

Smoothing splines

Definition

A spline is a piecewise polynomial such that it is smooth at points called knots.

The minimizer of $R_\lambda(g)$ is a spline such that

- knots : $X_1, \ldots, X_N$,
- cubic polynomial on $[X_{i-1},\ X_i]$,
- 2 continuous derivatives.

Wavelets

Goal: orthogonal series representation of $L^2(\mathbb{R})$.
Desirable features: ability to adapt to local features of curves, especially functions with jumps.
A wavelet function ψ entails an orthonormal system through dilatation and translation. We have

$$g(x) = \sum_{j,k=-\infty}^{+\infty} c_{j,k} 2^{j/2} \psi(2^j x - k),$$

where

$$c_{j,k} = 2^{j/2} \int_{-\infty}^{+\infty} g(x)\, \psi(2^j x - k)\, dx.$$

Example of a wavelet: Haar wavelet

$$\psi_H(x) = \mathbb{I}_{[0,\frac{1}{2}[} - \mathbb{I}_{[\frac{1}{2},1[}.$$

Wavelets have good data compression properties and good localization properties both in time and frequency.

Wavelets in statistics

Estimate the wavelet coefficient:

$$\hat{c}_{j,k} = 2^{j/2} \frac{1}{N} \sum_{i=1}^{N} Y_i \psi(2^j X_i - k).$$

The estimate of the regression function

$$\hat{g}(x) = \sum_{j,k \in \mathcal{S}} \hat{c}_{j,k} 2^{j/2} \psi(2^j x - k).$$

Problem: the choice of the set $\mathcal{S}$.
Example of choice (Donoho and Johnstone):

$$\mathcal{S}_\lambda = \left\{ (j,k) : |\hat{c}_{j,k}| > \lambda \right\}.$$

Statistical properties of smoothers

Example of Gasser-Müller estimator

Design

Assumption on the design

$$X_i = F^{(-1)}\left(\frac{i-\frac{1}{2}}{N}\right),$$

where F is a cumulative distribution function with density f:

$$F(x) = \int_0^x f(u)\, du.$$

Moreover the density is Lipschitz continuous, i.e.

$$\exists L > 0 : |f(u) - f(v)| \leq L|u - v| \quad \forall u, v \in [0,\ 1].$$

Kernel

Assumptions on the kernel: it is supported on $[-1,\ 1]$, Lipschitz continuous and

$$\int_{-1}^{1} K(u)\,du = 1, \quad \int_{-1}^{1} uK(u)\,du = 0$$

$$J_K = \int_{-1}^{1} K^2(u)\,du, \quad \sigma_K^2 = \int_{-1}^{1} u^2 K(u)\,du.$$

Risk

Way of measuring the fit: pointwise mean square error

$$\mathbb{E}\left((\hat{g}_h(x) - g(x))^2\right) = \text{var}(\hat{g}_h(x)) + \left(\mathbb{E}(\hat{g}_h(x)) - g(x)\right)^2.$$

Theorem

The variance is

$$var(\hat{g}_h(x)) = \frac{\sigma^2}{Nh}\frac{1}{f(x)}J_K + \mathcal{O}(N^{-1}) + \mathcal{O}((Nh)^{-2}).$$

If g has two continuous derivatives in an interval containing x, the bias is

$$\mathbb{E}(\hat{g}_h(x)) - g(x) = \frac{h^2}{2}g^{(2)}(x)\sigma_K^2 + o(h^2) + \mathcal{O}(N^{-1}).$$

Choice of the bandwidth

Corollary

If $Nh \to +\infty$, $N \to +\infty$ and $h \to 0$, then

$$\begin{aligned}
\mathbb{E}\left((\hat{g}_h(x) - g(x))^2\right) &= E_1 + E_2 \\
E_1 &= \frac{\sigma^2}{Nh}\frac{1}{f(x)}J_K + \frac{h^4}{4}\left(g^{(2)}(x)\right)^2\sigma_K^4 \\
E_2 &= o(h^4) + \mathcal{O}(N^{-1}) + \mathcal{O}((Nh)^{-2}).
\end{aligned}$$

The optimal choice for the bandwidth is

$$h = \left(\frac{4\sigma^2 J_K}{f(x)(g^{(2)}(x))^2\sigma_K^4}\right)^{1/5} N^{-1/5}.$$

Problems: σ^2 and $g^{(2)}(x)$ are unknown!

Data driven choice of smoothing parameters

Cross Validation: build the model with one part of the sample and predict the other part with the model. We consider the special case: leave-one-out. The criterion is

$$CV(h) = \frac{1}{N}\sum_{i=1}^{N}(Y_i - \hat{g}_{h,i}(X_i))^2.$$

Find h which minimizes $CV(h)$.

Data driven choice of smoothing parameters

Plug-in rule: estimate the unknown parameters in the bandwidth, e.g. the Gasser-Müller estimator

$$\hat{\sigma}^2 = \frac{1}{2(N-1)} \sum_{i=2}^{N} (Y_i - Y_{i-1})^2, \quad \hat{g}_h^{(2)}(x).$$

Change point problem

Two aspects will be reviewed:

1. detection
2. estimation

We will consider both parametric and nonparametric approaches.

Definition

We observe a sample $\{Y_i, i = 1, \ldots, n\}$.

Definition

A change point is said to occur at time τ if the distributions of $\{Y_i, i = 1, \ldots, \tau\}$ and $\{Y_j, j = \tau + 1, \ldots, n\}$ are different.
If there is m change points, then the sample can be divided into $m + 1$ homogeneous intervals.

Examples

Let $1 < \tau < n$. We can observe

- mean change: for $\theta_0 \neq \theta_1$

$$Y_1, \ldots, Y_\tau \sim \mathcal{N}(\theta_0, \sigma^2)$$

and

$$Y_{\tau+1}, \ldots, Y_n \sim \mathcal{N}(\theta_1, \sigma^2),$$

- two-segmented regression:

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i, i = 1, \ldots, \tau$$

and

$$Y_j = \beta_0' + \beta_1' X_j + \varepsilon_j, j = \tau + 1, \ldots, n.$$

Applications

Applications of change point models are numerous

- industrial control: quality control, CUSUM chart
- computer science: bioinformatic, software reliability engineering, detection of malware within software
- network traffic analysis: inhomogeneous Poisson processes
- finance: stock prices
- climatology: detect regime shifts in climate
- ecology
- oceanography
- geology: seismic activity
- archeology

Problems to address

How many change points are needed to represent the data in an efficient way?
Can we estimate the parameters of the distribution in each interval?
What are the rates of estimation?
First, we will be interested in the single change point model.

Change in mean: estimation

David Hinkley (1970): *Inference about the change-point in a sequence of random variables*, Biometrika.

We consider a Gaussian sample: $Y_1, \ldots, Y_\tau \sim \mathcal{N}(\theta_0, \sigma^2)$ and $Y_{\tau+1}, \ldots, Y_n \sim \mathcal{N}(\theta_1, \sigma^2)$ with $\theta_0 \neq \theta_1$.
Goal: estimate τ.
Method: maximum likelihood

$$L(\tau) = \sum_{i=1}^{\tau} \ln\left(f(X_i, \theta_0)\right) + \sum_{i=\tau+1}^{n} \ln\left(f(X_i, \theta_1)\right).$$

Change in mean: estimation

We can also write:

$$L(\tau) = \sum_{i=1}^{\tau} U_i + \sum_{i=1}^{n} \ln\left(f(X_i, \theta_1)\right),$$

where

$$U_i = \ln\left(f(X_i, \theta_0)\right) - \ln\left(f(X_i, \theta_1)\right).$$

In the Gaussian case, we have

$$U_i = \frac{(\theta_0 - \theta_1)(X_i - \frac{1}{2}(\theta_0 + \theta_1))}{\sigma^2}.$$

First, assume θ_0 and θ_1 to be known.
The maximum likelihood estimator $\hat{\tau}$ is the value of τ which maximizes

$$\sigma^2 \sum_{i=1}^{\tau} U_i.$$

Change in mean: estimation

First, remark that the distribution of $\hat{\tau} - \tau$ is symmetric.
If τ and $N - \tau$ tend to ∞ when N goes to ∞, then the asymptotic distribution of $\hat{\tau} - \tau$ can be derived and approximated (Hinkley, 1970).
Let $\Delta = \frac{|\theta_0 - \theta_1|}{2\sigma}$.

k/Δ	0.5	1	1.5
0	0.280(0.640)	0.641(0.820)	0.857(0.928)
1	0.114(0.754)	0.113(0.933)	0.059(0.988)
2	0.067(0.821)	0.038(0.971)	0.01(0.997)
3	0.044(0.865)	0.015(0.987)	
4	0.031(0.896)	0.007(0.993)	
5	0.023(0.919)	0.003(0.997)	
6	0.017(0.935)		
7	0.013(0.948)		
8	0.010(0.956)		
9	0.008(0.973)		
10	0.006(0.973)		

Change in mean: estimation

In the case of θ_0 and θ_1 unknown, these parameters have to be estimated:

$$\begin{aligned} \hat{\theta}_{0\tau} &= \frac{1}{\tau} \sum_{i=1}^{\tau} Y_i, \\ \hat{\theta}_{1\tau} &= \frac{1}{n-\tau} \sum_{i=\tau+1}^{n} Y_i. \end{aligned}$$

The maximum likelihood estimator $\hat{\tau}$ is the value of τ which maximizes

$$Z_\tau^2 = \tau(n-\tau) \frac{\left(\hat{\theta}_{0\tau} - \hat{\theta}_{1\tau}\right)^2}{n}.$$

The asymptotic distribution of $\hat{\tau} - \tau$ is the same as in the case of θ_0 and θ_1 known.

Change in mean: test

We are interested in hypothesis testing such as

$$H_0 : \tau = \tau_0 \quad \text{versus} \quad H_1 : \tau > \tau_0,$$
$$H_0 : \tau = \tau_0 \quad \text{versus} \quad H_2 : \tau \neq \tau_0.$$

First assume that the parameters θ_0 and θ_1 are known. The likelihood ratio tests are

$$\Lambda_1 = \max_{\tau \geq \tau_0} \left(\sum_{i=1}^{\tau} U_i \right) - \sum_{i=1}^{\tau_0} U_i,$$
$$\Lambda_2 = L(\hat{\tau}) - L(\tau_0) = \sum_{i=1}^{\hat{\tau}} U_i - \sum_{i=1}^{\tau_0} U_i.$$

Change in mean: test

Assume that the sample is Gaussian. Then the likelihood ratio tests become

$$\Lambda_1 = \frac{(\theta_0 - \theta_1)\left(\max_{\tau \geq \tau_0} \sum_{i=1}^{\tau} (Y_i - \frac{1}{2}(\theta_0 + \theta_1)) - \sum_{i=1}^{\tau_0} (Y_i - \frac{1}{2}(\theta_0 + \theta_1))\right)}{\sigma^2}$$
$$\Lambda_2 = \frac{(\theta_0 - \theta_1)\left(\sum_{i=1}^{\hat{\tau}} (Y_i - \frac{1}{2}(\theta_0 + \theta_1)) - \sum_{i=1}^{\tau_0} (Y_i - \frac{1}{2}(\theta_0 + \theta_1))\right)}{\sigma^2}.$$

The rejection regions are of the form $\mathcal{W} = \{\Lambda > x\}$.

Change in mean: test

In the case of θ_0 and θ_1 unknown, we have

$$\begin{aligned} \Lambda_1 &= \frac{1}{2\sigma^2} \max_{\tau \geq \tau_0} \left(Z_\tau^2 - Z_{\tau_0}^2 \right) \\ \Lambda_2 &= \frac{1}{2\sigma^2} \left(Z_{\hat{\tau}}^2 - Z_{\tau_0}^2 \right), \\ \text{with} \quad Z_\tau^2 &= \frac{\tau(n-\tau)}{n} \left(\hat{\theta}_{0\tau} - \hat{\theta}_{1\tau} \right)^2. \end{aligned}$$

Remark: when the variance σ^2 is also unknown, it can be replaced by its maximum likelihood estimator under the relevant alternative.

Change in mean: test

Hinkley (1970) has tabulated the asymptotic distributions of the likelihood ratio tests. The 95% quantiles are given in the following table

Δ	Λ_1	Λ_2
0.5	2.42	3.09
1	1.79	2.53
1.5	0.62	1.59

Change in mean: detection

We are interested in the detection problem

$$\begin{aligned} H_0 &: \mu_1 = \ldots = \mu_n = \theta_0 \\ \text{versus} & \\ H_1 &: \mu_1 = \ldots = \mu_\tau = \theta_0 \neq \mu_{\tau+1} = \ldots = \mu_n = \theta_1, \end{aligned}$$

where τ is the unknown position of the change point.

Results based on the work by Gombay and Horvath (1990): *Asymptotic distributions of maximum likelihood tests for change in the mean*, Biometrika.

Change in mean: detection

Let g be a given function.
The test statistic is based on

$$Z_t = 2\left(tg(\hat{\theta}_{0,t}) + (n-t)g(\hat{\theta}_{1t}) - ng(\bar{Y}_n)\right),$$

and is

$$Z = \max_{1\leq t\leq n}\left(\frac{Z_t}{g^{(2)}(\theta)}\right).$$

Examples for function g:

$$g(u) = -\ln(u), \quad g(u) = \frac{1}{2}u^2.$$

In the Gaussian case, it is exactly the same statistic as in Hinkley (1970).

Change in mean: detection

Assumptions:

1. g has three continuous derivatives in a neighbourhood of θ, $g^{(2)}(\theta_0) \neq 0$
2. under the null hypothesis the random variables have a common finite variance and $\mathbb{E}\left(|Y_i|^{2+\delta}\right) < +\infty$ for some $\delta > 0$.

Theorem

Under H_0 and for all x, the limiting distribution of Z is given by

$$\lim_{n\to+\infty} \mathbb{P}\left(e^2(\ln n)\frac{Z}{\sigma^2} \leq (x + h(\ln n))^2\right) = \exp\left(-e^{-x}\right).$$

where

$$e(u) = \sqrt{2\ln(u)}, \quad h(u) = 2\ln(u) + \frac{1}{2}\ln\ln(u) - \frac{1}{2}\ln(\pi).$$

Change in mean: detection

Asymptotic results obtained by Gombay and Horvath (1990) were successfully applied to

- Annual volumes of discharge from the Nile River at Aswan between 1871 and 1970 (Cobb, 1978): according to historical records there was a shift in the flow levels starting from the year 1899. This shift in 1899 is attributed partly to the weather changes and partly to the start of construction work for a new dam at Aswan.
- Changes in variation of stock market returns (Hsu, 1979): weekly closing values of the Dow Jones Industrial Average from July 1, 1971 to August 2, 1974.
- Time intervals between coal mine explosions in which more than ten people were killed between 1851 and 1950 (Maguire et al, 1952; Jarrett, 1979).

Change in variance

We observe a sample $\{Y_i, i = 1, \ldots, n\}$ and we assume

$$Y_i \sim \mathcal{N}(\mu, \sigma_i^2).$$

Goal: test if there is a change point for the variance.
This testing problem is

$$H_0 \quad : \quad \sigma_1^2 = \ldots = \sigma_n^2 = \sigma^2$$

versus

$$H_1 \quad : \quad \sigma_1^2 = \ldots = \sigma_{k_1}^2 \neq \sigma_{k_1+1}^2 = \ldots = \sigma_{k_2}^2 \neq \ldots \neq \sigma_{k_q+1}^2 = \ldots = \sigma_n^2,$$

where q is the unknown number of change points and
$1 \leq k_1 < k_2 < \ldots < k_q$.

Change in variance

We can iterate the testing procedure and assume $q = 1$. Therefore the testing problem is

$$H_0 \quad : \quad \sigma_1^2 = \ldots = \sigma_n^2 = \sigma^2$$

versus

$$H_1 \quad : \quad \sigma_1^2 = \ldots = \sigma_{k_1}^2 \neq \sigma_{k_1+1}^2 = \ldots = \ldots = \sigma_n^2,$$

where k_1 is the unknown position of the change point.

Change in variance

The test is based on the Schwarz Information Criterion (Schwarz, 1978), denoted SIC. It is of the form

$$-2\ln L(\hat{\theta}) + p\ln(n),$$

where $L(\hat{\theta})$ is the maximum likelihood function for the model, p is the number of free parameters and n is the sample size.
Here we have under H_0:

$$SIC(n) = n\ln(2\pi) + n\ln(\hat{\sigma}^2) + n + \ln(n),$$

where

$$\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \mu)^2.$$

Change in variance

Under the alternative H_1, we have

$$SIC(k) = n\ln(2\pi) + k\ln(\hat{\sigma}_1^2) + (n-k)\ln(\hat{\sigma}_n^2) + n + 2\ln(n),$$

where

$$\begin{aligned}\hat{\sigma}_1^2 &= \frac{1}{k}\sum_{i=1}^{k}(Y_i - \mu)^2,\\ \hat{\sigma}_n^2 &= \frac{1}{n-k}\sum_{i=k+1}^{n}(Y_i - \mu)^2.\end{aligned}$$

Remark: in order to compute the maximum likelihood estimators, we have to restrict to $2 \leq k_1 \leq n-1$.

Change in variance

The testing procedure is

- Do not reject (accept) H_0 if $SIC(n) \leq \min_{2\leq k\leq n-2} SIC(k) + c_\alpha$,
- Reject H_0 if $SIC(n) > \min_{2\leq k\leq n-2} SIC(k) + c_\alpha$.

The position of the change point is estimated by $\hat{k}_1$ such that

$$SIC(\hat{k}_1) = \min_{2\leq k\leq n-2} SIC(k).$$

Theorem

Let k_1 be the true position of the change point under the alternative. Then $\hat{k}_1$ is a strongly consistent estimator for k_1.

Change in variance

The threshold parameter c_α is a critical value such that

$$1 - \alpha = \mathbb{P}_{H_0}\left(SIC(n) < \min_{2\leq k\leq n-2} SIC(k) + c_\alpha\right).$$

Theorem

Under H_0 and for all $x \in \mathbb{R}$, we have

$$\lim_{n\to+\infty} \mathbb{P}\left(a_n\sqrt{\ln(n) - \min_{2\leq k\leq n-2}[SIC(k) - SIC(n)]} - b_n \leq x\right) = \exp\left(-2e^{-x}\right),$$

where

$$a_n = \sqrt{2\ln\ln(n)}, \quad b_n = 2\ln\ln(n) + \frac{1}{2}\ln\ln\ln(n) - \ln\Gamma\left(\frac{1}{2}\right).$$

Remark: the critical values have been tabulated (Chen and Gupta, J. Amer. Stat. Assoc. 1997).

Change in variance

We have to adjust to the unknown mean:

$$\begin{aligned} SIC(n) &= n\ln(2\pi) + n\ln(\hat{\sigma}^2) + n + 2\ln(n) \\ \hat{\sigma}^2 &= \frac{1}{n}\sum_{i=1}^{n}\left(Y_i - \bar{Y}_n\right)^2, \end{aligned}$$

and

$$\begin{aligned} \ln\left(L_1(\mu, \sigma_1^2, \sigma_n^2)\right) &= -\frac{n}{2}\ln(2\pi) - \frac{k}{2}\ln(\sigma_1^2) - \frac{n-k}{2}\ln(\sigma_n^2) \\ &\quad - \frac{\sum_{i=1}^{k}(Y_i-\mu)^2}{2\sigma_1^2} - \frac{\sum_{i=k+1}^{n}(Y_i-\mu)^2}{2\sigma_n^2}, \\ 0 &= \sigma_n^2\sum_{i=1}^{k}(Y_i-\mu)^2 + \sigma_1^2\sum_{i=k+1}^{n}(Y_i-\mu)^2, \end{aligned}$$

$$\sigma_1^2 = \frac{1}{k}\sum_{i=1}^{k}(Y_i-\mu)^2, \qquad \sigma_n^2 = \frac{1}{n-k}\sum_{i=k+1}^{n}(Y_i-\mu)^2.$$

Change in variance

Chen and Gupta (1997) successfully applied this testing procedure to the U.S. stock market return prices between 1971 and 1974, i.e.

$$R_t = \frac{P_{t+1} - P_t}{P_t}.$$

Striking results: they detect change points at the following dates

- March 19-23, 1973: Watergate
- July 19-August 8, 1971: several union strikes influenced the changes of the U.S. stock market
- November 15-December 12, 1971: increase of prices in some important industrial products

Multiple change points

Based on Scott and Knott (1974): *A cluster analysis method for grouping means in the analysis of variance*, Biometrics.

We observe the means of k treatments: $Y_i \sim \mathcal{N}(\mu_i, \sigma^2), i = 1, \ldots, k$
We assume that there is an estimator of the common variance, denoted s^2 and independant of the means. Its distribution is such that $\frac{\nu s^2}{\sigma^2} \sim \chi^2_\nu$.
First we give another procedure for partitioning the treatments into two subgroups.
Let us consider the testing problem

$$H_0 : \mu_i = \mu \quad \text{versus} \quad H_1 : \mu_i = m_1 \text{ or } m_2 \text{ with } m_1 \neq m_2.$$

The test statistics is

$$\lambda = \frac{\pi}{2(\pi - 2)} \frac{B_0}{\hat{\sigma}_0^2},$$

where

$$\hat{\sigma}_0^2 \frac{\sum_{i=1}^k (Y_i - \bar{Y}_n)^2 + \nu s^2}{k + \nu}.$$

Multiple change points

Theorem

Under H_0, the distribution of the test statistic is asymptotically $\chi^2(\frac{k}{\pi - 2})$ (for k small the distribution has also been tabulated).

If H_0 is rejected then B_0 provides the maximum likelihood partition.

We should consider $2^{k-1} - 1$ partitions but according to Fisher it is sufficient to order the means and consider $k - 1$ partitions.

Multiple change points

A hierarchical splitting method is applied

1. find the best split according to the between groups sum of squares criterion, i.e. B_0
2. iterate on each subgroup
3. continue until the groups are found homogeneous according to the test statistic λ.

Problem: choice of the first-type error α

- α small: the splitting process will finish too early,
- α large: this leads to split homogeneous data.

Multiple change points

Examples taken from Scott and Knott.

Example 1: 7 varieties of barley were considered. The means were

$$49.6, 58.1, 61.0, 61.5, 67.6, 71.2, 71.3$$

An analysis of variance rejected the equality of the means. Therefore the splitting algorithm was applied

1. 1234567 was split between 1234 and 567 because of $\lambda = 17.73$,
2. 1234 could not be split into 1 and 234 because of $\lambda = 8.04$,
3. 567 could not be split into 5 and 67 because of $\lambda = 0.99$.

Multiple change points

Example 2: it came from Tukey (1949) and was about 6 varieties of potatoes. The means were

$$345, 405.2, 426.4, 477.8, 502.2, 601.8$$

The splitting algorithm was applied

1. 123456 was split between 123 and 456 because of $\lambda = 22.97$,
2. 123 was split into 1 and 23 because of $\lambda = 12.23$,
3. 23 could not be split into 2 and 3 because of $\lambda = 1.3$,
4. 456 was split into 45 and 6 because of $\lambda = 16.927$,
5. 45 could not be split into 4 and 5 because of $\lambda = 4.53$.

Multiple change points

Optimal segmentation algorithm due to Auger and Lawrence (1989).
Goal: optimally partition a sequence into Q contiguous segments neighbourhood based on the fit of a model.
Observation: $Y_1, \ldots, Y_n$
We denote $\tau_0 = 0 < \tau_1 < \ldots < \tau_{Q-1} < \tau_Q = n$ the change point positions. The vector of unknown parameters for the q-th segment neighbourhood is denoted θ_q.
The model of the relationship between $Y_i, \ldots, Y_j$ and θ_q is $F(Y_i, \ldots, Y_j, \theta_q)$.
we measure the fit of a model $F(\cdot)$ to the data by $C(F(Y_i, \ldots, Y_j, \theta_q))$.

Multiple change points

The global measure for the fit is

$$Z(Y, \theta, \tau, Q) = \min_{\theta,\tau} \sum_{q=1}^{Q} C(F(Y_{\tau_{q-1}+1}, \ldots, Y_{\tau_q}, \theta_q)).$$

Example: if $Y_i = \theta_q + \varepsilon_i$, then

$$Z(Y, \theta, \tau, Q) = \min_{\theta,\tau} \sum_{q=1}^{Q} \sum_{\tau_{q-1}+1}^{\tau_q} (Y_i - \theta_q)^2 .$$

We denote $c_{i,j}^q = Z(Y_i, \ldots, Y_j, \theta, \tau, q)$ the best partition of $Y_i, \ldots, Y_j$ into q segments neighbourhood.
Exhaustive enumeration is extremely computationally intensive.

Multiple change points

Auger and Lawrence proposed the following algorithm

1. Compute the measure of the fit for each segment
$$\forall i < j : c_{i,j}^1 \leftarrow C(F(Y_i, \ldots, Y_j, \theta_{ij})).$$
2. Compute optimal partition for $2, 3, \ldots, Q$ segments.
For $q = 2$ to Q do
for $j = 1$ to n do
$$c_{i,j}^q \leftarrow \min_{\tau} \left(c_{1,\tau}^{q-1} + c_{\tau+1,j}^1 \right)$$
end

Theorem

This algorithm computes $c_{1,n}^Q$ and is of order $\mathcal{O}(Qn^2)$.

Extension: this algorithm can be modified in order to identify a subset of P non-overlapping segments which do not necessarily splan the entire sequence.

Multiple change points

This is an introduction to P.E.L.T. algorithm developped by Killick, Fearnhead and Eckley (2011).

Observation: $Y_1, \ldots, Y_n$
We denote $\tau_0 = 0 < \tau_1 < \ldots < \tau_{Q-1} < \tau_Q = n$ the change point positions.
The model of the relationship between $Y_i, \ldots, Y_j$ and θ_q is $F(Y_i, \ldots, Y_j, \theta_q)$.
We measure the fit of a model $F(\cdot)$ to the data by $C(F(Y_i, \ldots, Y_j, \theta_q))$.

Multiple change points

We introduce a penalty term β which does not depend on τ or Q.
Goal: optimal partitioning with respect to

$$G(n) = \min_{\tau,\theta} \left(\sum_{q=1}^{m} \left[C(F(Y_{\tau_{q-1}+1}, \ldots, Y_{\tau_q}, \theta_q)) + \beta \right] \right).$$

Idea: recursive method for the optimal segmentation.
Conditionally on $\tau_m = \tau^\star$, we have

$$\begin{aligned} G(n) &= \min_{\tau^\star} \Big(\min_{\tau|\tau^\star} \left(\sum_{q=1}^{m-1} \left[C(F(Y_{\tau_{q-1}+1}, \ldots, Y_{\tau_q}, \theta_q)) + \beta \right] \right) \\ &\quad + C(F(Y_{\tau^\star+1}, \ldots, Y_n, \theta_q)) + \beta \Big) \\ &= \min_{\tau^\star} \left(G(\tau^\star) + C(F(Y_{\tau^\star+1}, \ldots, Y_n, \theta_q)) + \beta \right). \end{aligned}$$

Multiple change points

The algorithm by Killick, Fearnhead and Eckley is

- $G(0) = -\beta$ and $cp(0) = 0$.
- Iterate for $\tau^\star = 1, \ldots, n$
  1. Calculate
  $$G(\tau^\star) = \min_{0 \leq_t au \tau^\star} \left(G(\tau) + C(F(Y_{\tau^\star+1}, \ldots, Y_n, \theta_q)) + \beta \right)$$
  2. $$\tau' = \text{argmin}_{0 \leq \tau < \tau^\star} \left(G(\tau) + C(F(Y_{\tau^\star+1}, \ldots, Y_n, \theta_q)) + \beta \right).$$
  3. $cp(\tau^\star) = (cp(\tau'), \tau')$

This algorithm is $\mathcal{O}(n^2)$ which is faster than Auger and Lawrence. The reason is that the location and the number of change points are decided in one pass of the data.

Multiple change points

Improved efficiency can be obtained via "pruning", which means remove points which cannot be change points.

This is the P.E.L.T. method due to Killick, Fearnhead and Eckley.

Let K be a constant such that $\forall t < s < T$:

$$K + C(F(Y_{t+1}, \ldots, Y_s), \theta_{ts})) + C(F(Y_{s+1}, \ldots, Y_T, \theta_{sT})) < C(F(Y_{t+1}, \ldots, Y_T)).$$

Multiple change points

The algorithm is as follows

- Initialise: $F(0) = -\beta$, $cp(0) = 0$ and $R_1 = \{0\}$.
- Iterate for $\tau^\star = 1, \ldots, n$
  1. Calculate

$$G(\tau^\star) = \min_{\tau \in R_{\tau^\star}} \left(G(\tau) + C(F(Y_{\tau^\star+1}, \ldots, Y_n, \theta_{\tau+1,\tau^\star})) + \beta \right).$$

  2. $\tau^1 = \text{argmin}_{\tau \in R_{\tau^\star}} \left(G(\tau) + C(F(Y_{\tau^\star+1}, \ldots, Y_n, \theta_{\tau+1,\tau^\star})) + \beta \right)$.
  3. $cp(\tau^\star) = \{cp(\tau^1), \tau^1\}$.
  4.

$$\begin{aligned} R_{\tau^\star+1} \quad = \quad & \Big\{ \{0\} \cup \{\tau \in R_{\tau^\star}\} \cup \{\tau^\star\} : \\ & G(\tau) + C(F(Y_{\tau+1}, \ldots, Y_{\tau^\star}, \theta_{\tau+1,\tau^\star})) + K > G(\tau^\star) \Big\}. \end{aligned}$$

change points are in $cp(n)$.

Multiple change points

Applications to

- Dow Jones Industrial Index (daily data) from 1st October 1928 to 30th July 2010: change in variance studied, 82 change points found and 14 times faster than binary segmentation
- Human chromosome: 47 times faster than binary segmentation.

Regression and change points

This part is due to Tang and Wei (2004): *Detecting change points in a quadratic regression model*.
We consider quadratic regression :

$$Y_i = \beta_0 + \beta_1 X_i + \beta_2 X_i^2 + \varepsilon_i, \quad i = 1, \ldots, n.$$

If there are $s-1$ change points, the new model is

$$\begin{aligned} Y_i &= \beta_{01} + \beta_{11} X_i + \beta_{21} X_i^2 + \varepsilon_i, \quad i = 1, \ldots, \tau_1, \\ &\vdots \quad \vdots \quad \vdots \\ Y_i &= \beta_{0s} + \beta_{1s} X_i + \beta_{2s} X_i^2 + \varepsilon_i, \quad i = \tau_{s-1} + 1, \ldots, \tau_s = n. \end{aligned}$$

The number of change points s is assumed to be known but the locations of the change points are unknown

Regression and change points

For the model without change point, the Schwarz Information Criterion is

$$\begin{aligned} SIC(n) &= n \ln \left(\sum_{i=1}^{n} \left(Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i - \hat{\beta}_2 X_i^2 \right)^2 \right) \\ &\quad + n(1 + \ln(2\pi)) + (4 - n)\ln(n). \end{aligned}$$

The estimators are the usual ones

$$\begin{aligned} \hat{\beta} &= (X'X)^{-1} X'Y \\ \hat{\sigma}^2 &= \frac{1}{n}(Y - X\beta)'(Y - X\beta). \end{aligned}$$

Regression and change points

For the model with s change points, the Schwarz Information Criterion is

$$\begin{aligned} SIC(\tau_1, \ldots, \tau_{s-1}) &= n \ln \left(\sum_{j=1}^{s} \sum_{i=\tau_{j-1}+1}^{\tau_j} \left(Y_i - \hat{\beta}_{0j} - \hat{\beta}_{1j} X_i - \hat{\beta}_{2j} X_i^2 \right)^2 \right) \\ &\quad + n(1 + \ln(2\pi)) + (3s + 1 - n) \ln(n). \end{aligned}$$

The estimators are

$$\begin{aligned} \hat{\beta}_j &= (X_j' X_j)^{-1} X_j' Y_j, \\ \hat{\sigma}^2 &= \frac{1}{n} \sum_{j=1} s(Y_j - X_j \hat{\beta}_j)'(Y_j - X_j \beta_j). \end{aligned}$$

Select the model which minimizes Schwarz Information Criterion.

Nonparametric change point

This first part is due to Carlstein (1988).
We observe: $Y_1, \ldots, Y_n$.
The change point location is $[\theta n]$ and

$$Y_1, \ldots, Y_\tau \sim F, \quad Y_{\tau+1}, \ldots, Y_n \sim G.$$

Assumption:

$$\begin{aligned} \Lambda &= \{x \in \mathbb{R} : |F(x) - G(x)| > 0\} \\ &\quad \int_\Lambda dF(x) > 0 \quad \text{or} \quad \int_\Lambda dG(x) > 0. \end{aligned}$$

Nonparametric change point

Let $t \in T_n = \left\{ \frac{i}{n}, 1 \leq i \leq n-1 \right\}$.
We define

$$h_{0t}^n(x) = \frac{1}{nt} \sum_{i=1}^{nt} \mathbb{I}_{\{Y_i \leq x\}},$$

$$h_{1t}^n(x) = \frac{1}{n(1-t)} \sum_{i=nt+1}^{n} \mathbb{I}_{\{Y_i \leq x\}}.$$

We denote

$$d_{ni}^t = \left| h_{0t}^n(Y_i) - h_{1t}^n(Y_i) \right|, \quad 1 \leq i \leq n,$$

and

$$D_n(t) = \sqrt{t(1-t)} S_n \left(d_{n1}^t, \ldots, d_{nn}^t \right),$$

where $S_n(\ldots)$ is a mean-dominant norm.
Thes estimator is

$$\hat{\theta}_n = \text{argmax}_{t \in T_n} D_n(t).$$

Nonparametric change point

Definition

A function $S_n(\ldots)$ is to be a mean-dominant norm if

- symmetry: S_n is symmetric in its n arguments.
- homogeneity: $S_n(cy_n) = cS_n(y_n)$ whenever $c \geq 0$ and $y_n \geq 0$.
- triangle inequality: $S_n(y_n + z_n) \leq S_n(y_n) + S_n(z_n)$ whenever $y_n \geq 0$ and $z_n \geq 0$.
- identity: $S_n(1, \ldots, 1) = 1$
- monotonicity: $S_n(y_n) \geq S_n(z_n)$ whenever $y_n \geq z_n \geq 0$.
- mean-dominant: $S_n(y_n) \geq \frac{1}{n} \sum_{i=1}^n y_{in}$ whenever $y_n \geq 0$.

Examples:

$$S_n(y_n) = \frac{1}{n} \sum_{i=1}^{n} y_{in}, S_n(y_n) = \sqrt{\frac{1}{n} \sum_{i=1}^{n} y_{in}^2}, S_n(y_n) = \sup_{1 \leq i \leq n} y_{in}.$$

Nonparametric change point

We assume that the observation are all distincts.

Theorem

Let $\{S_n, , n \geq 1\}$ be a sequence of mean-dominant norms.
We have

$$\forall \varepsilon > 0, \quad \forall n \geq n(\varepsilon) : \mathbb{P}\left(|\hat{\theta}_n - \theta| > \varepsilon\right) \leq c_1 n \exp\left(-c_2 \varepsilon^2 n\right),$$

where c_1 and c_2 are constants.

Nonparametric change point

Applications:

- Nile River data
- Lindisfarne scribes data: text divided into 13 sections. It was assumed that each section was written by a scribe alone and sections written by a scribe were consecutive. Each scribe is characterized by his propensity to use one of two possible grammatical variants: either "s" or "δ" ending in the present indicative 3rd person singular.
  Let m_j be the number relevant words in the jth section. We have $n = \sum_{j=1}^{13} m_j$. We denote X_i the indicator of "δ" ending in the ith word. The model is $\mathcal{B}(p)$. Then $D_n(t)$ is proportional to

$$\sqrt{t(1t-)}\left|\frac{1}{t}\sum_{i=1}^{[nt]} X_i - \frac{1}{1-t}\sum_{i=[nt]+1}^{n} X_i\right|,$$

  but it is maximized only at points $t_k = \frac{1}{n}\sum_{j=1}^{k} m_j, \quad 1 \leq k \leq 12.$

Nonparametric change point

Here are Lindisfarne scribes data

k	1	2	3	4	5	6	
m_k	21	57	101	131	183	228	
$\sum_{i=1}^{n_k} X_i$	9	19	32	38	62	73	
$C_n(\frac{n_k}{n})$	18.5	15.2	17.4	12.9	34.9	34	
k	276	333	381	403	423	444	464
m_k	82	93	100	193	196	1110	114
$C_n(\frac{n_k}{n})$	28.9	24.8	16.7	11.8	7.3	4.5	

It was found that there was a change point at the end of section 5.

Nonparametric change point

This part is due to Kan, Koo and Park (2000).
We are interested in the nonparametric regression model:

$$Y_i = g(X_i) + \varepsilon_i, \quad \mathbb{E}(\varepsilon_i) = 0, \quad \text{var}(\varepsilon_i) = \sigma^2 < +\infty.$$

We assume that the regression function has a jump at τ and is smooth otherwise.
We can write

$$g(t) = f(t)\Delta \mathbb{I}_{[\tau,1]}(t),$$

where

$$\Delta = g(t^+) - g(t^-).$$

Nonparametric change point

We recall the Gasser-Müller estimator

$$\hat{g}(t) = \frac{1}{h}\sum_{i=1}^{n} Y_i \int_{s_{i-1}}^{s_i} K\left(\frac{t-u}{j}\right)\, du,$$

where $s_i = (X_i + X_{i-1})/2, \quad i = 1, \ldots, n-1$, $s_0 = 0$ and $s_n = 1$. The kernel is supported on $[-1,\ 1]$ and is of order k.
We also define

$$\begin{aligned}\hat{g}_+(t) &= \frac{1}{h}\sum_{i=1}^{n} Y_i \int_{s_{i-1}}^{s_i} K_+\left(\frac{t-u}{h}\right)\, du \\ \hat{g}_-(t) &= \frac{1}{h}\sum_{i=1}^{n} Y_i \int_{s_{i-1}}^{s_i} K_-\left(\frac{t-u}{h}\right)\, du,\end{aligned}$$

where the kernels K_- and K_+ are of order k and supported respectively on $[0,\ 1]$ and $[-1,\ 0]$.

Nonparametric change point

Let $Q \subset [0, 1]$ be a closed interval such that $\tau \in Q$.
The position of the change point is estimated by

$$\hat{\tau} = \inf\{\rho \in Q : \hat{\Delta}(\rho) = \sup_{x \in Q} |\hat{\Delta}(x)|\},$$

where $\hat{\Delta}(t) = \hat{g}_+(t) - \hat{g}_-(t)$.

Proposition

For the bandwidth $h \sim n^{-\frac{2k-1}{2k+1}}$, we have

$$|\hat{\tau} - \tau| = \mathcal{O}_P\left(n^{-\frac{2k}{2k+1}}\right).$$

For the bandwidth $h \sim n^{-\frac{1}{2k+3}}$, we have

$$|\hat{\Delta} - \Delta| = \mathcal{O}_P\left(n^{-\frac{k+1}{2k+3}}\right).$$

Nonparametric change point

Now let $Y_i^\star = Y_i - \hat{\Delta}\mathbb{I}_{[\hat{\tau},1]}(X_i)$.
The estimator of the regression function is defined as

$$\hat{g}^\star(t) = \frac{1}{h}\sum_{i=1}^{n} Y_i^\star \int_{s_{i-1}}^{s_i} K^\star\left(\frac{t-u}{h}\right)\, du + \hat{\Delta}\mathbb{I}_{[\hat{\tau},1]}(t),$$

where

$$K^\star(t-uh) = \begin{cases} K_+\left(\frac{t-u}{h}, q\right), & 0 \le t \le h, q = \frac{t}{h}, \\ K\left(\frac{t-u}{h}\right), & h \le t \le 1h, \\ K_-\left(\frac{t-u}{h}, q\right), & 1-h \le t \le 1, q = \frac{1-t}{h}. \end{cases}$$

The kernels $K_+(\cdot, q)$ and $K_-(\cdot, q)$ are kernels with general asymmetric support $[-1, q]$ and $[-q, 1]$

Nonparametric change point

Theorem

If all the kernels are of order K then

$$\sqrt{nh}\,(\hat{g}^\star(t) - g(t)) \xrightarrow{\mathcal{L}} \mathcal{N}\left(f^{(k)}(t)B_k, \sigma^2 V\right),$$

where

$$B_k = \frac{1}{k!}\int x^k K^\star(x)\, dx, \quad V = \int (K^\star(x))^2\, dx.$$

Examples of asymmetric kernels

$$K_-(x) = 6(1+x)(1+2x), \quad K_+(x) = 6(1-x)(1-2x).$$