# Change: Detection, Estimation, Segmentation

**Abstract**

The maximum score statistic is used to detect and estimate changes in the level, slope, or other local feature of a sequence of observations, and to segment the sequence when there appear to be multiple changes. Control of false positive errors when observations are auto-correlated is achieved by using a first order autoregressive model. True changes in level or slope can lead to badly biased estimates of the autoregressive parameter and variance, which can result in a loss of power. Modifications of the natural estimators to deal with this difficulty are partially successful. Applications to temperature time series, atmospheric CO2 levels, COVID-19 incidence, excess deaths, copy number variations, and weather extremes illustrate the general theory.

This is joint research with Xiao Fang.

# A general formulation

Suppose that $dY_s = (\mu(s) + \sum \xi_j f(s - t_j))ds + \rho Y_s ds + \sigma dW_s$, where $dW$ is Gaussian white noise. Examples of the function $f(u)$ are (i) the indicator that $u > 0$, (ii) the positive part function $u^+$, (iii) the indicator of the interval $(0, 1]$ with unknown scale $\tau$, or (iv) a symmetric probability density function centered at $0$, also with unknown scale $\tau$. The process is observed for $s \in T$, which may be an interval of the real line or in some applications may be multi-dimensional. Initially we assume that $\sigma = 1$. Estimation of $\sigma$ and $\rho$ involve special problems that are discussed later.

The parameters of primary interest are $t, \xi$, which define the local signal. Let $\theta$ denote the nuisance parameters $\mu, \rho$. (Typically $\mu(s)$ is a parametric regression function.) Given $t$, the efficient score for testing $\xi = 0$ is

$$\frac{\partial \ell}{\partial \xi}(0, \hat{\theta}), \tag{1}$$

where $\hat{\theta}$ are maximum likelihood estimators of $\theta$ under the assumption that $\xi = 0$.

# Significance Thresholds

By standard likelihood theory (1 is asymptotically distributed as

$$\frac{\partial \ell}{\partial \xi} - I_{\xi,\theta} I_{\theta,\theta}^{-1} \frac{\partial \ell}{\partial \theta}, \tag{2}$$

where $I$ is the Fisher information matrix, partitioned according to the coordinates $\xi, \theta$, and all expressions are evaluated at $t$, $\xi = 0$ and true values of $\theta$. Hence (2) is of the form

$$\ell_\xi(t) - \Psi'(t) A \ell_\theta. \tag{3}$$

Here $\ell_\xi(t) = \partial \ell / \partial \xi$ is a Gaussian process with covariance function denoted by $G(s,t)$, while $\Psi(t)' = I_{\xi,\theta}$, $\ell_\theta = \partial \ell / \partial \theta$ is normally distributed with mean 0 and covariance matrix $I_{\theta,\theta}$, and $A = I_{\theta,\theta}^{-1}$.

Let $\Sigma(s,t) = G(s,t) - \Psi'(s) A \Psi(t)$ denote the covariance function of (3) under the hypothesis $\xi = 0$, and put

$$Z_t = [\Sigma(t,t)]^{-1/2} [\ell_\xi(t) - \Psi'(t) A \ell_\theta]. \tag{4}$$

This representaton provides an approximation for $\mathrm{P_0\{\max Z_t \geq b\}}$, which is independent of $\rho$.

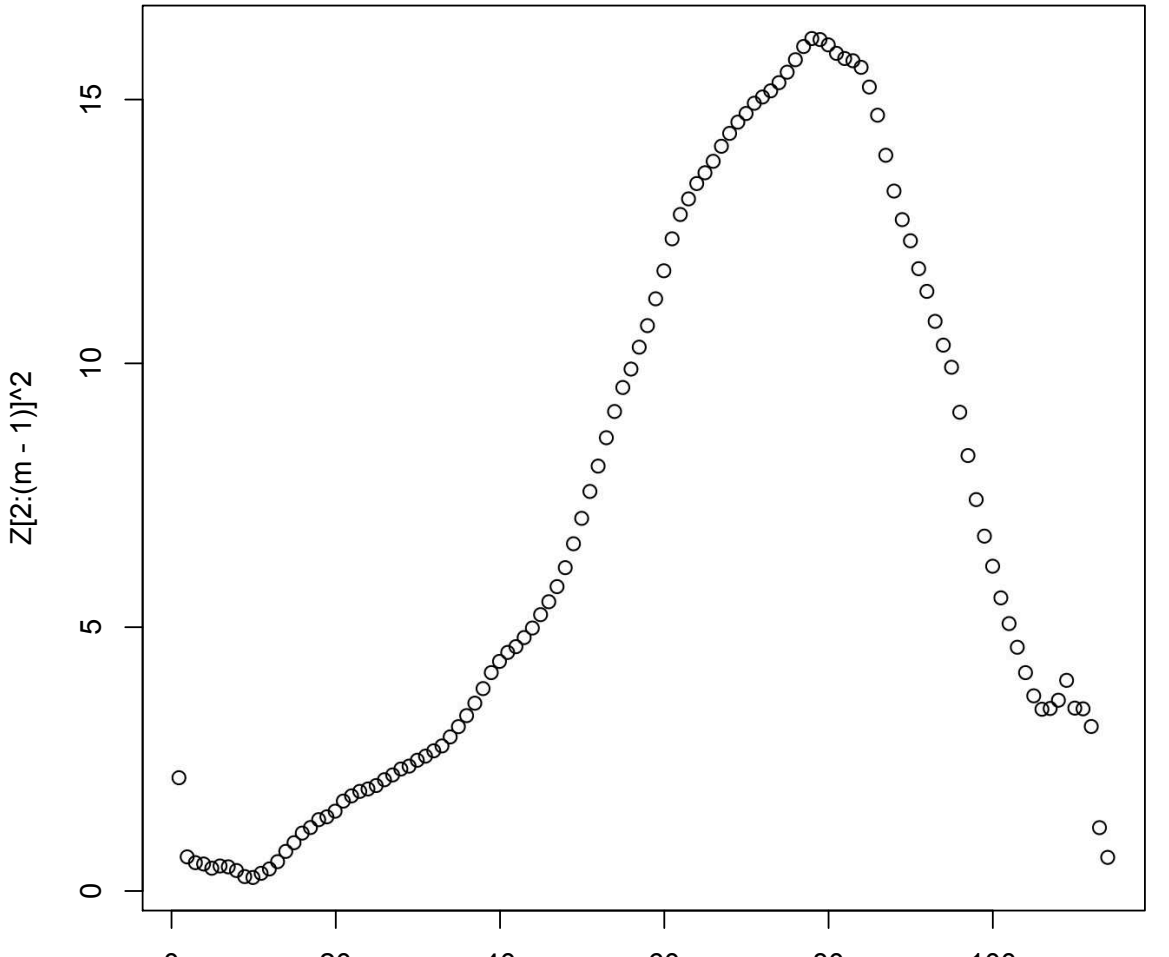
# Broken Line Regression

$\Sigma(s,t) = \mathrm{E}_0[\ell_\xi(s)\ell_\xi(t) - \Psi(s)'A\Psi(t)]$ is smooth and does not depend on nuisance parameters $\alpha, \beta, \rho$, so by Rice's formula

$$\mathrm{P}\{\max_{T_0<t<T_1} Z_t \geq b\} \sim (\varphi(b)/(2\pi)^{1/2}) \int_{T_0}^{T_1} [\mathrm{E}(\dot{Z}_t^2)]^{1/2} dt. \qquad (5)$$

For a numerical example, suppose $T = 116, b = 4.01$ (Annual average temperature of the Netherlands, 1901-2016). Then the approximation (5) with $T_0 = 1$ gives the value $0.0009$.

# Netherlands Temperature: 1901-2016



# Multidimensional Example

Consider the excess deaths in Germany, France, and Spain during the first 15 weeks of 2020. Assuming independence between weeks, analysis of each country separately indicates a slope increase after 8 weeks, but the numbers are small and the results somewhat unclear. Spain is not significant at the 0.05 level, France is, but not at the 0.01 level, and the p-value for Germany is 0.002. If we also assume independence between countries and use the norm of a three dimensional process, the p-value indicating a slope change after 8 weeks is 0.0001.

# Segmentation

Recall that $Z(t,T) = \{\ell_\xi(t,T) - \Psi'(t,T)A_T\ell_\theta(T)\}/\sigma(t,T)$. Let $Q = \mathrm{P}\{\max_{m_0 \le t < T \le m}\ Z(t,T) > b\}$.

Let $\beta(t,T) = \{\mathrm{E}[\ell_\xi \partial \ell_\xi/\partial \mathrm{T}) - .5\partial\sigma^2(\mathrm{t},\mathrm{T})/\partial \mathrm{T}\}/\sigma^2(\mathrm{t},\mathrm{T})$ and $\lambda_t = \mathrm{E}[(\dot{\mathrm{Z}}_{\mathrm{t},\mathrm{T}})^2]$. Then

$$Q \approx (2/\pi)^{1/2} b\varphi(b) \sum_{m_0}^{m_1} \int_{m_0 \le t < T-1} [\lambda_t \beta(t,T)\nu[b(2\beta(t,T)]^{1/2}]^{1/2} dt.$$

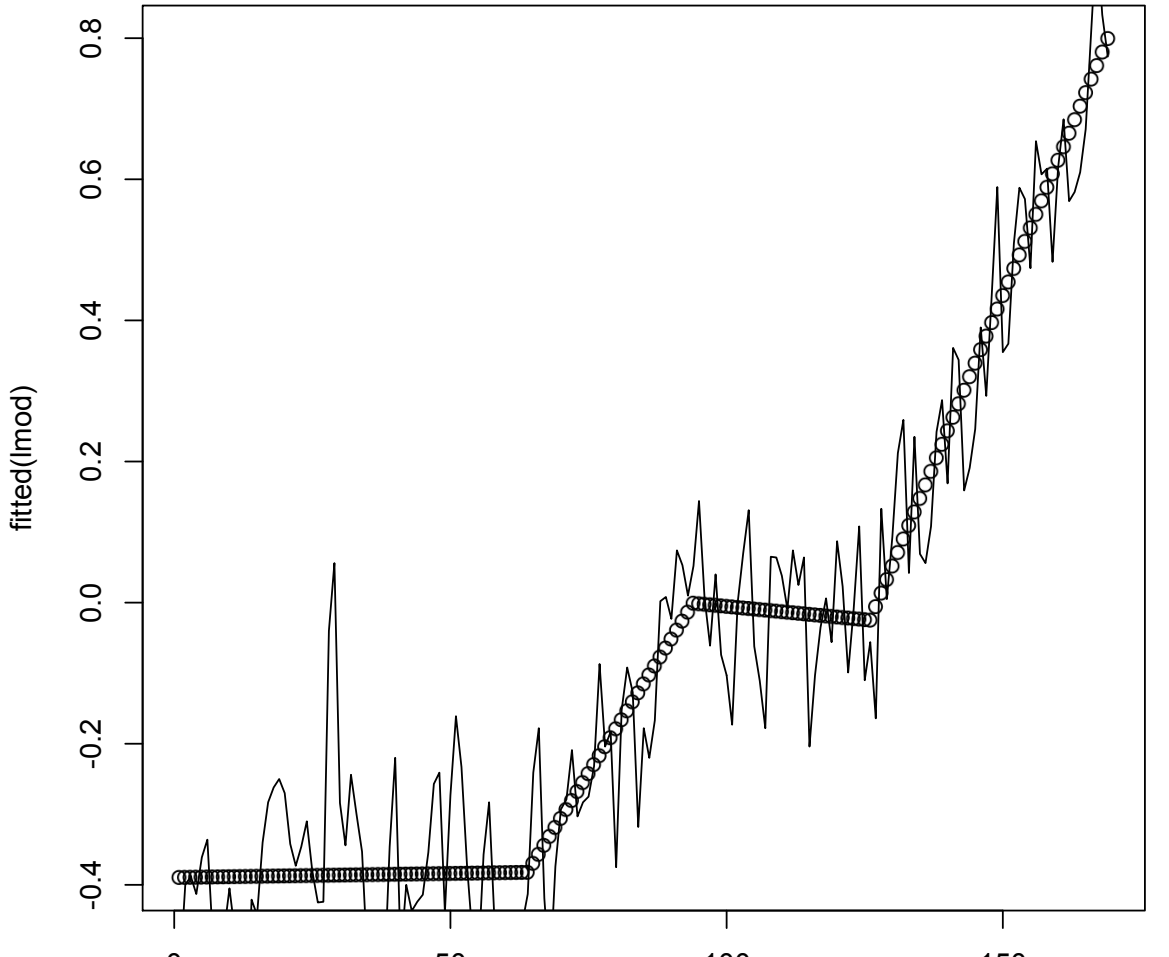
We can use this approximation for *pseudo-sequential segmentation* OR for sequential detection of a slope change.

A similar result can be obtained for $\max_{T_0<t<T_1} Z(T_0,t,T_1)$, which facilitates searching for change-points over all possible background intervals (or a random selection of background intervals), $(T_0,T_1)$.

# NH Anomalies:1850-2019

The Northern Hemisphere average annual temperature anomalies from the Berkeley Earth web site are a relatively simple and interesting example, since a plot of the data suggests there may be multiple slope changes in the 20th century. For $m = 170$, $b = 3.77$, $m_0 = 5$, and $\rho$ set equal to 0.3, the pseudo-sequential method detects slope changes in the 64th, 94th, and 126th years At the conventional level of 0.05, the method using all possible backgrounds detects essentially the same three change-points. A multiple regression model with these three changes assumed to be true produces $R^2 = 0.92$ and $\rho = 0.33$

# NH Anomalies: 1880-2019



# A Top Down Procedure to Detect Slope Changes in Pairs

Consider the statistic $\max_{s<t-h} U_{s,t}$, where

$$U_{s,t} = (V_s, V_t)\Sigma_{s,t}^{-1}(V_s, V_t)'. \tag{6}$$

and $h$ is a parameter that represents a minimum distance between changes (usually taken to be 5 or 10). An appropriate threshold may be determined from the approximation

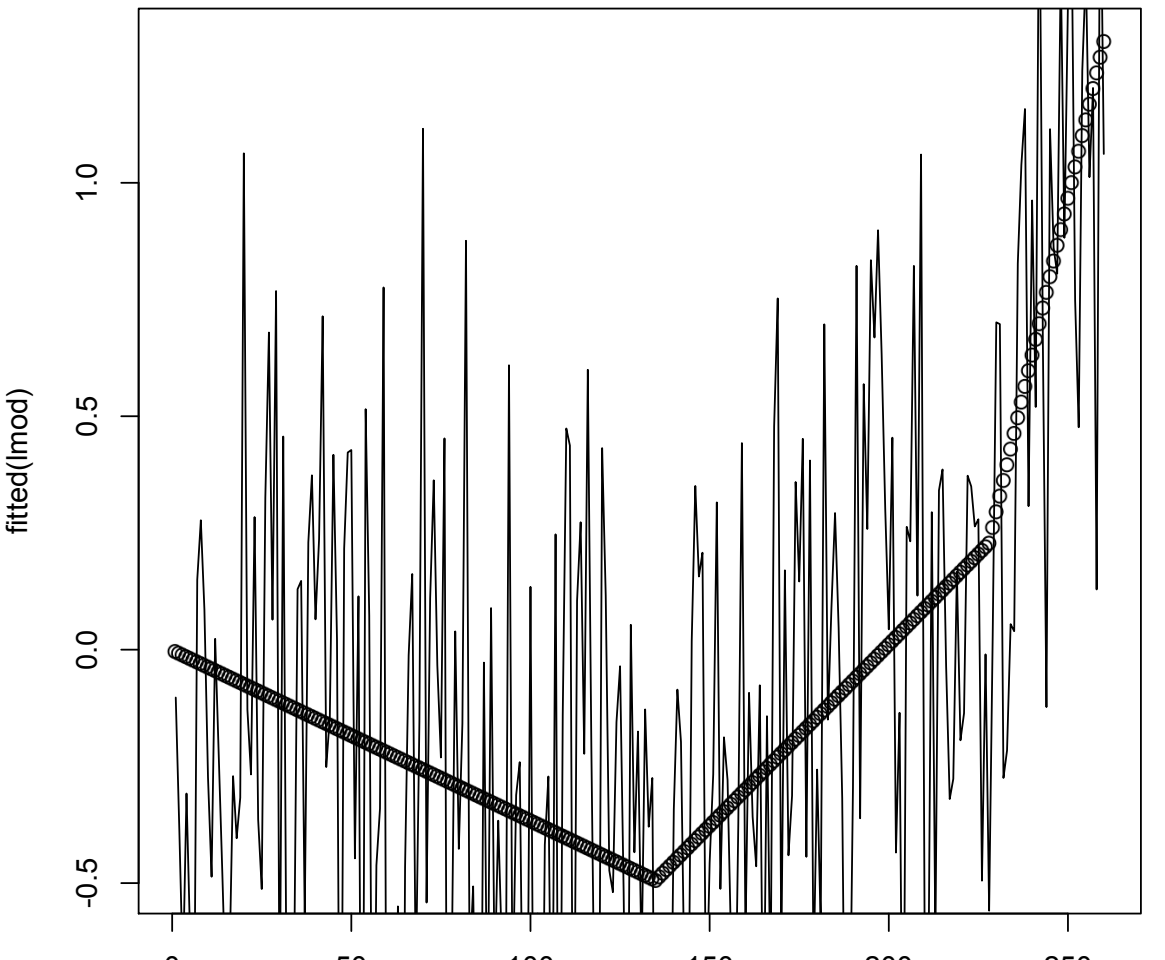
$$\mathcal{P}\{\max_{s<t-h} U_{s,t} > b\} \sim [2b\varphi(b)/(2\pi)] \int_{s<t-h} \det[\mathrm{E}(\dot{\mathrm{U}}\dot{\mathrm{U}}')]^{1/2} \mathrm{dsdt}. \tag{7}$$

This search can be iterated. If we assume that in early iterations, only true positives are detected, then the iterative process does not create a multiple comparisons issue, since the probability is roughly linear in the length of the segment searched.

# Examples: Detecting Slope Changes in Pairs

For the Berkeley Land-Ocean data, the method of detecting changes in pairs gives essentially the same results as the other methods. In a few examples it seems to be the method of choice. For example, the Berkeley Earth web site gives average annual temperture anomalies for individual European countries, beginning in 1753. The pseudo-sequential and all possible backgrounds method often detect one slope increase, in either about 1880 or about one hundred years later. The method to detect paired changes often detects two. An example is provided by the annual anomalies of Switzerland, where positive slope changes can be detected in the 135th and the 228th years. The pseudo-sequential method detects only the first of these changes, although it detects both if it is run backwards.

# Swiss Anomalies: 1753-2012

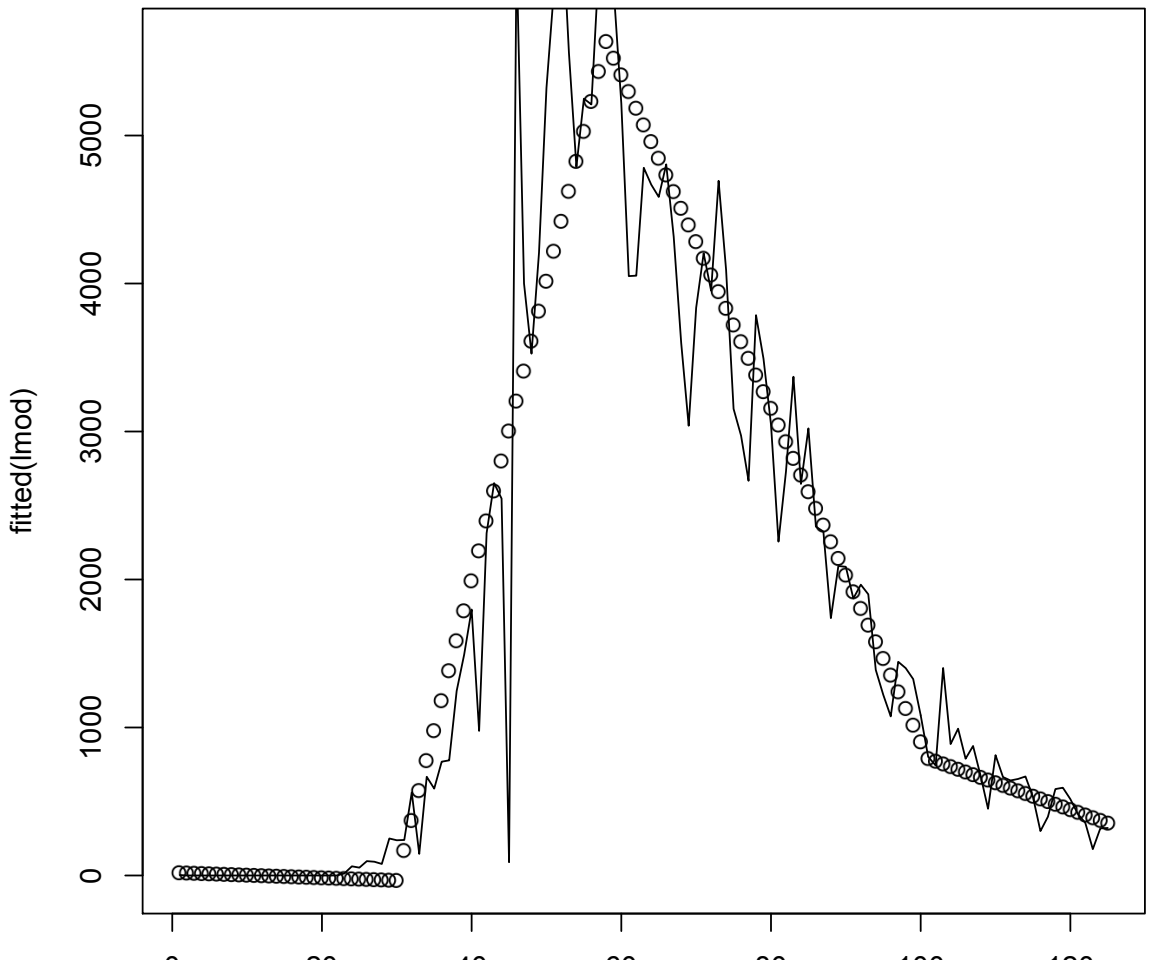


# COVID-19

Our broken line model can be useful in tracking the incidence of COVID-19. We consider here Italy for $T = 124$ days after the first case appeared on 31.01.20 The pseudo-sequential method puts the first slope increase at 20 days, a large slope decrease at 63, and another much smaller increase at 101 days. The method using all possible backgrounds misses the first slope change, although this is fairly inconsequential for the overall fit, since the slope at 0 compensates. The best result comes from the method designed to detect two changes at a time, which puts changes at 36, 58, and 101 days. It also suggests that there may be a relatively small slope increase at 24 days.

We noted above that the pseudo-sequential method could be used as a legitimate sequential method. If applied to the China COVID-19 data, which reported its first cases on 31.12.19, calibrated to allow one expected false positive in 100 years, with $\rho = 0.4$, it detects a change after the 27th day, which it estimates to have occurred on the 22nd.

# COVID-19 in Italy



# Confidence Regions for Broken Line Regression

Using the Kac-Slepian model process, or equivalently as an application of LeCam large sample theory, we find that for a putative change-point $t$,

$$\max(Z_u^2 - Z_t^2) \approx \dot{Z}_t^2 / \mathrm{E}(\dot{Z}_t^2). \tag{8}$$

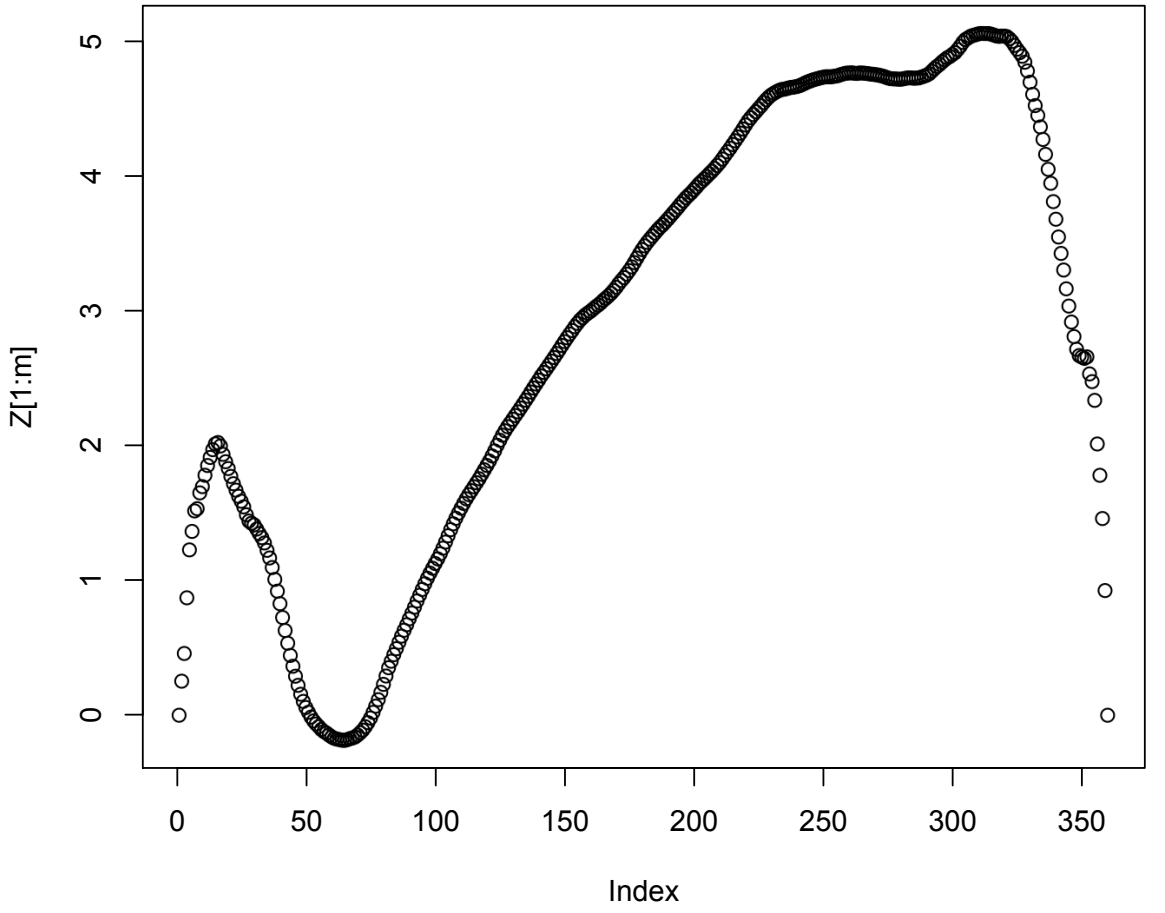
Hence a 0.9 confidence region for $t$ is the set of all $t$ such that $Z_t^2 \geq \max_u Z_u^2 - \chi_1^2(.9)$. Note that this is exactly what "regular" likelihood theory would suggest for a likelihood ratio statistic with one degree of freedom.

The result (8) can be used to give an approximation for the local power to detect a change at $\tau$.

# Example: Central England Temperature since 1760

Consider the annual average tempertures in central England since 1760. (See next slide for a plot of the statistic to detect at least one change, since the series beginning in 1659.) A change is detected about 1980, but the plot suggests that the temperature may have begun to increase about 100 years earlier. A 95% confidence region extends only 10 years the the right of 1980, but almost 100 years to the left.

# Central England Annual Temperature Since 1659



# Estimation of $\sigma^2$ and $\rho$

When there are signals in the form of change-points, the usual estimators of $\sigma^2$ and of $\rho$ can be very badly biased. Using them can lead to a serious loss of power. If there is a known segment of the data without local signals, these parameters can be estimated from that part of the data. If we assume the observations are independent, for some problems a reasonable estimator of $\sigma^2$ is $\sum(Y_t - Y_{t-1})^2/(2T)$. An estimator of $\sigma^2$ that also removes the effect of linear drift would be to use second differences, although this reduces the effective sample size. Many of our examples involve time series, where serial correlation must be regarded a possibility.

# References


Fang, Xiao, Li, Jian, and Siegmund D. (2020). Segmentation and estimation of change-point models: false positive control and confidence regions, to appear in *Ann. Statist.*.

Fang, Xiao and Siegmund D. (2020) Detection and estimation of local signals, *Arkiv*

Fryzlewicz, P. (2014). Wild binary segmentation for multiple change-pont detection. *Ann. Statist.* **42**, 2243-2281.

Olshen, A. B., Venkatraman, E. S., Lucito, R. and Wigler, M. (2004). Circular binary segmentation for the analysis of array-based DNA copy number data, *Biostatistics*. **5**, 557-572.

Zhang, N. R., Siegmund, D. O., Ji, Hanlee, and Li, Jun (2010). Detecting simultaneous change-points in multiple sequences, *Biometrika* **97** 631-646.