Detection and Estimation of Local Signals

Summary

I will discuss a general framework for detection of local signals, primarily defined by change-points, in random sequences or random fields. Changes can occur continuously, e.g., a change in the slope of a regression line, or discontinuously, e.g., a jump in the level of a process. A motivating example of jump discontinuities is provided by copy number variation (CNV). I will focus on the simplest version of the problem: segmentation of independent normal observations according to changes in the mean. Results will be illustrated by simulations and applications. Confidence regions for the change-points and some difficulties associated with dependent observations will also be discussed.

Aspects of this research involve collaboration with Fang Xiao, Li Jian, Liu Yi, Nancy Zhang, Benjamin Yakir and Li (Charlie) Xia.

A general formulation

Suppose that $dY_s = (\mu + \xi f(s-t))ds + \rho Y_s + \sigma dW_s$. Here f is a positive kernel, for example, (i) the indicator that $s \geq 0$, (ii) the indicator of the interval $(0, \Delta]$, (iii) a symmetric probability density function centered at 0 with scale Δ, or (iv) the positive part function, s^+. The process is observed for $s \in T$, which may be an interval of the real line or in some applicationsmay be multi-dimensional. Initially we assume that σ is known and equals 1.

The parameters of primary interest are t, ξ, which define the local signal. Let β denote the nuisance parameters μ, ρ. Given t, the efficient score for testing $\xi = 0$ is

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

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

Significance Thresholds

By standard likelihood theory this is asymptotically distributed as

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

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

$$V_t - \Psi'(t) A\eta. \tag{3}$$

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

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

$$Z_t = [\sigma(t,t)]^{-1/2}[V_t - \Psi'(t)A\eta]. \tag{4}$$

We can use this representaton to approximate $\mathrm{P_0\{\max Z_t \geq b\}}$.

Examples

(1) For f equal to the indicator of the interval $(0, \infty]$, the test statistic when there is no change-point is distributed as

$$\max_{t_0<t<t_1} [W(t) - tW(T)/T]/[t(1 - t/T)]^{1/2}, \tag{5}$$

which does not depend on the nuisance parameters μ, ρ.

(2) For $\mu = \beta_0 + \beta_1(t - T/2)$ and $f(s) = \max(0, s)$, the covariance function again does not depend on the nuisance parameters, but now is differentiable, so a version of Rice's formula applies.

Binary Segmentation: Top Down (Vostrikova)

If we assume there is at most one change, the likelihood ratio test statistic is

$$\max_j |S_j - jS_m/m|/[j(1-j/m)]^{1/2}.$$

Use this statistic to test the hypothesis of no change-point against the alternative of “at least one change-point” by thresholding at level b. Then iterate. Problem: How do we choose the value of b if we do not know the number of change-points?

Segmentation Statistics

The log likelihood ratio statistic for a putative change-point in (i, k) is

$$\max_{i<j<k} |Z_{i,j,k}|$$

where

$$Z_{i,j,k} = [S_j - S_i - (j-i)(S_k - S_j)/(k-j)]/[(j-i)(1-(j-i)/(k-i))]^{1/2}. \quad (6)$$

Our basic segmentation statistic is $\max_{i<j<k} |Z_{i,j,k}|$, which generates a candidate set of change-points, by thresholding at a level b. We choose change-points from this candidate set by selecting those with the smallest value of $k - i$. An alternative possibility is to select those with the largest value of $|Z|$ and to impose a “no overlap” condition on i, k.

Niu and Zhang (2012) suggest a similar statistic, but with $k - j = j - i$.

p-values

A basic theoretical result is the approximation

$$\mathrm{P}\{\max_{i,j,k} |Z_{i,j,k}| \geq b\} \sim .25 b^5 \varphi(b)$$

$$\times \sum_{m_0}^{m_1} \sum_{m_0}^{m_1} \frac{(m-u-v)}{uv(u+v)} \nu[b(\frac{u}{v(u+v)})^{1/2}] \nu[b(\frac{v}{u(u+v)})^{1/2}] \nu[b(\frac{u+v}{uv})^{1/2}]. \tag{7}$$

where $\nu(x)$ is a special function defined, e.g., in Siegmund (1985) and easily computed numerically. For a simple numerical example, for $m = 500$ and $b = 4.83$, (2) gives 0.05. Simulated values based on 2000 repetitions yields 0.047.

BT474: Chromosomes 17 and 5

For chromosome 17, there are $m = 87$ observations, the standard deviation is 0.51; the 0.05 detection threshold is approximately 4.22. There is an increase in copy number at the 36th observation (17q11.2-12), with a change back to baseline just two observations later. There is a second increase at the 51st observation (17q21.3) and a return to the baseline at the 67th (17q23).

For chromosome 5 there are 99 observations and a standard deviation of 0.16. Change-points are detected at 25, 45, 51, 54, 65, 89, and 91.

Continuation of Example 1

Suppose there are N aligned sequences with aligned intervals $(s,t]$ where the ith interval shows a displacement from the baseline value of δ_i. It is assumed that any particular signal affects only a small fraction of the intervals, an assumption appropriate for detection of inherited copy number variations. The log likelihood function for a putative signal in $(s,t]$

$$\sum_1^N \log\{1 - p_0 + p_0 \exp[\delta_i \sum_{s+1}^t (Y_{u,i} - \delta_i/2)]\},$$

which when maximized with respect to δ_i becomes

$$\sum_1^N \log\{1 - p_0 + p_0 \exp[U_i^2(s,t)/2]\},$$

where $U_i(s,t) = [\sum_{s+1}^t Y_{u,i} - (t-s)\bar{Y}_i]/[(t-s)\{1-(t-s)/T\}]^{1/2}$. More generally, we consider statistics of the form

$$\sum_1^N f[U_i(s,t)]$$

for different functions f.

p-values for Example 1

Let $Z(s,t) = \sum_1^N f[U_i(s,t)]$. Let $\psi(\theta) = \log \mathrm{E}[\exp(\theta f(U)]$ and $\mathcal{I} = \theta\dot{\psi}(\theta) - \psi(\theta)$, where θ is chosen to satisfy $N\dot{\psi}(\theta) = b$. Put

$$\mu = .5N\mathrm{E}_\theta\{f'(U)U - f''(U)\} = .5N\theta\mathrm{E}_\theta\{f'(U)\}^2$$

and $\sigma^2 = N\mathrm{E}_\theta\{f'(U)\}^2$. Then

$$\mathrm{P}\left(\max_{T_0<t-s<T_1} Z(s,t) > b\right) \approx [2\pi N\ddot{\psi}(\theta)]^{-1/2}\theta\mu^2 e^{-N\mathcal{I}}$$

$$\times \int_{T_0/T}^{T_1/T} \frac{1}{u^2(1-u)}\nu^2\left[\frac{2\mu/\sigma T^{1/2}}{\{u(1-u)\}^{1/2}}\right] du,$$

where $\nu(x)$ is a special function defined, e.g., in Siegmund (1985) and easily computed numerically.

Example 2

$G(s,t) = \mathrm{E}_0[V_s V_t - \Psi(s)' A \Psi(t)]$ is smooth and does not depend on nuisance parameters α, β, ρ, so

$$\mathrm{P}\{\max_{T_0 < t < T_1} Z_t \geq b\} \sim (\varphi(b)/(2\pi)^{1/2}) \int_{T_0}^{T_1} [\mathrm{E}(\dot{Z}_t^2)]^{1/2} dt. \qquad (8)$$

For a numerical example, suppose $T = 136, b = 4.97$. (Singapore annual rainfall for 136 years.) Then the approximation (8) gives the value 4×10^{-6}.

Note also that $\mathrm{E}_{t,\xi} Z_t = \xi [G(t,t]^{1/2}$, as expected.

Confidence Regions for Example 2

Using the Kac-Slepian model process, we find that for a change-point τ, conditional on a large value Z_τ,

$$\max(Z_t^2 - Z_\tau^2) \approx \dot{Z}_\tau^2 / \mathrm{E}(\dot{Z}_\tau^2). \tag{9}$$

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

The result (9) can be used to give an approximation for the local power to detect a change at τ.

Estimation of σ^2 and ρ

When there are signals in the form of change-points, the usual estimators of σ^2 and of ρ 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, a reasonable estimator of σ^2 is $\sum(Y_t - Y_{t-1})^2/(2T)$. Other possibilities remain to be explored.

Confidence Regions for Example 1

Given the number of change-points M, the likelihood ratio statistic for the locations and size of jumps of the change points is

$$\ell(\hat{\tau}_i, \hat{\mu}_j) - \ell(\tau_i, \mu_j),$$

where $1 \leq i \leq m$ and $0 \leq j \leq M$. In the special case $M = 1$, with some approximations, this can be simplified to

$$(S_\tau - \tau\mu_0)^2/(2\tau) + (S_m - S_\tau - (m-\tau)\mu_1)^2/[2(m-\tau)]$$

$$+ \max_i [\hat{\delta}_\tau (S_{\tau+i} - S_\tau) - \hat{\delta} i/2],$$

where $\hat{\delta}$ is the maximum likelihood estimate of $\mu_0 - \mu_1$.

Probability Approximation

Let $Z_i(\delta) = \delta[S_{\tau+i} - S_\tau - \delta i/2]$. Then $P\{\max_i Z_i(\hat{\delta}_\tau) \geq x|\hat{\delta}_\tau)\}$

$$\approx 2\nu(\hat{\delta}) \exp(-x) - [\nu(\hat{\delta})]^2 \exp(-2x).$$

If we replace $\hat{\delta}_\tau$ by δ (consistency), we have expressed the probability on the preceding slide as the distribution of the sum of three (almost) INDEPENDENT random variables with known distributions. Generalization to the case of M change-points is straightforward.

Numerical Examples.

Table 1: Likelihood ratio based joint confidence intervals. $\hat{p}$ is the simulated probability that the parameters t_1 and t_2 are rejected when the true parameter values are τ_1 and τ_2. Nominal confidence level is 0.05. Simulations are based on 1000 (500) repetitions in the first four (last 12) rows.

δ_1	δ_2	b	$\tau_1,\ \tau_2$	$t_1,\ t_2$	$\hat{p}$
2.13	1.33	6.4	9, 33	9, 33	0.049
2.5	4.0	5.35	87, 104	87, 104	0.051
0.65	2.5	6.65	138, 225	138, 225	0.047
1.73	2.13	6.23	57, 66	57, 66	0.049
2.13	1.33	6.4	9, 33	7, 33	0.59
2.13	1.33	6.4	9, 33	11, 33	0.58
2.13	1.33	6.4	9, 33	9, 29	0.47
2.13	1.33	6.4	9, 33	9, 37	0.44
0.65	2.5	6.65	138, 225	138, 227	0.75
0.65	2.5	6.65	138, 225	138, 223	0.73

References

Frick, K, Munk, A. and Sieling, H. (2014). Multiscale change-point inference, *JRSS* B.

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.

Fang, Xiao, Li, Jian, and Siegmund D. (2016). Segmentation and Estimation of Change-point Models, *Arkiv*.

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

Li Charlie Xia, Sukolsak Sakshuwong, Erik Hopmans, John Bell, Sue Grimes, David O. Siegmund, Hanlee P. Ji, Nancy R. Zhang (2016) A genome-wide approach for detecting novel insertion-deletion variants of mid-range size, *Nucleic Acids Research*.

Niu, S. Y. and Zhang, H. (2012). The screening and ranking algorithm to detect DNA copy number variations. *Ann. Appl. Statist.* **6**, 1306-1326.

Zhang, N. R., Siegmund, D. O.,Yakir, B. (2011). Detectiing simultaneous variant intervals in aligned sequences *Ann. Appl. Statist.* **5** 645-668.

Zhang, N., Yakir, B., Xia, L., Siegmund, D. (2016). Scan statistics on Poisson random fields with applications in genomics. *Ann. Appl. Statist.*