Change-point Detection

Summary

Change-point detection had its origins almost sixty years ago in the work of Page, Shiryayev, and Lorden, who focused on sequential detection of a change-point in an observed process. The process was typically a model for the measured quality of a continuous production process, and the change-point indicated a deterioration in quality that must be detected and corrected. Recently, motivation from a broad range of applications has lead to a variety of different problems. In this talk I will review this history with emphasis on applications in biology and common features of likelihood based approaches.

Page’s Problem (1954, 1955)

Suppose $X_1, X_2, \ldots, X_m, \ldots$ are independent observations. For $j \leq K$, they have the distribution F_0, while for $j > K$ they have the distribution F_1. The distributions F_i may be completely specified or may depend on unknown parameters. In the case of sequential observations we desire to detect the change-point K as soon as possible after it occurs, while rarely claiming a detection before it occurs. In the case of a fixed number m of observations, we would like to test the null hypothesis of no change, i.e., that $F_0 = F_1$. An essential feature of the problem is that K is undefined under the null hypothesis, which could also be specified as $K \geq m$.

Page's Solution, Barnard's (1959) Suggestion

For sequential detection Page (1954) suggested the stopping rule

$$N_0 = \min\{n : S_t - \min_{0 \leq k \leq t} S_k \geq b\},$$

where S_t is the tth cumulative sum (hence CUSUM) of "scores," $Z(X_i)$. An important special case is $Z(x) = \log f_1(x)/f_0(x)$, where $f_0(x)$ is the probability density function of the X's before the change-point and $f_1(x)$ the density after the change-point. For a fixed sample of size t, Page (1955) suggested the test statistic $S_t - \min_{0 \leq k \leq t} S_k$.

For the special case that f_0 is a normal distribution with mean 0 and variance σ^2, while f_1 is normal with mean δ, Barnard (1959) observed that $\log[f_1(x)/f_0(x)] = \delta x - \delta^2/2$ is maximized with respect to δ at $\hat{\delta} = x$, which suggests that in testing for a change in a normal mean from an initial value of 0, one should use

$$N = \min\{t : \max_{k \leq t} |S_t - S_k|/\sigma(t-k)^{1/2} \geq b\}.$$

Shiryayev's Method

Shiryayev (1963) considered the case of completely specified F_0 and F_1. He assumed that K is random and used optimal stopping theory to describe an exact solution to a well formulated Bayesian version of the problem, where detection of a change-point at time n incurs a loss equal to $\mathbf{1}\{K \geq n\} + C(n-K)^+$. Let $\ell_k(n) = \sum_{k+1}^{n} \log[f_1(X_j)/f_0(X_j)]$. Assume K has the distribution $\mathrm{P}\{K=n\} = q^{n-1}p$ and $p \to 0$. An approximation to Shiryayev's stopping rule is

$$N_1 = \min\{n : \sum_{k=0}^{n} \exp[\ell_k(n)] \geq B\} \tag{1}$$

Fundamental Properties

For both Page's and Shiryayev's stopping rules

$$\mathrm{E}_\infty(N_i) \sim C_i \exp(b) \tag{2}$$

and $\max_K \mathrm{E}_K(N_i - K|N_i > K) \sim b/\mathrm{E}_1 \log[f_1(X)/f_0(X)]$ as $b = \log(B) \to \infty$. Proving these results is relatively easy, but identification of $C_i > 1$ is more complex.

Davies' Contributions

Davies (1976, 1986 *Biometrika*) pointed out a class of problems having the essential feature of change-point detection—the existence of a nuisance parameter that is only present under the alternative hypothesis, although (unlike Page's problem) the log likelihood might be (piecewise) smooth. In fact, a large class of similar problems was originally discussed in companion papers by Hotelling and Weyl in 1939, who considered the model $Y_i = \alpha + \beta f_i(\theta) + e_i$, where f is nonlinear and θ can be multidimensional. For example, $f_i(\theta) = (x_i - \theta)^+$. Davies applied Rice's (1939) formula to evaluation of significance levels and enlarged the scope of potential applications.

Worsley's Model for Brain Scans

A random field to detect local activity in an fMRI scan can be modeled as

$$Z_{t,\sigma} = \int \sigma^{-1/4} f[(s-t)/\sigma] dX_s,$$

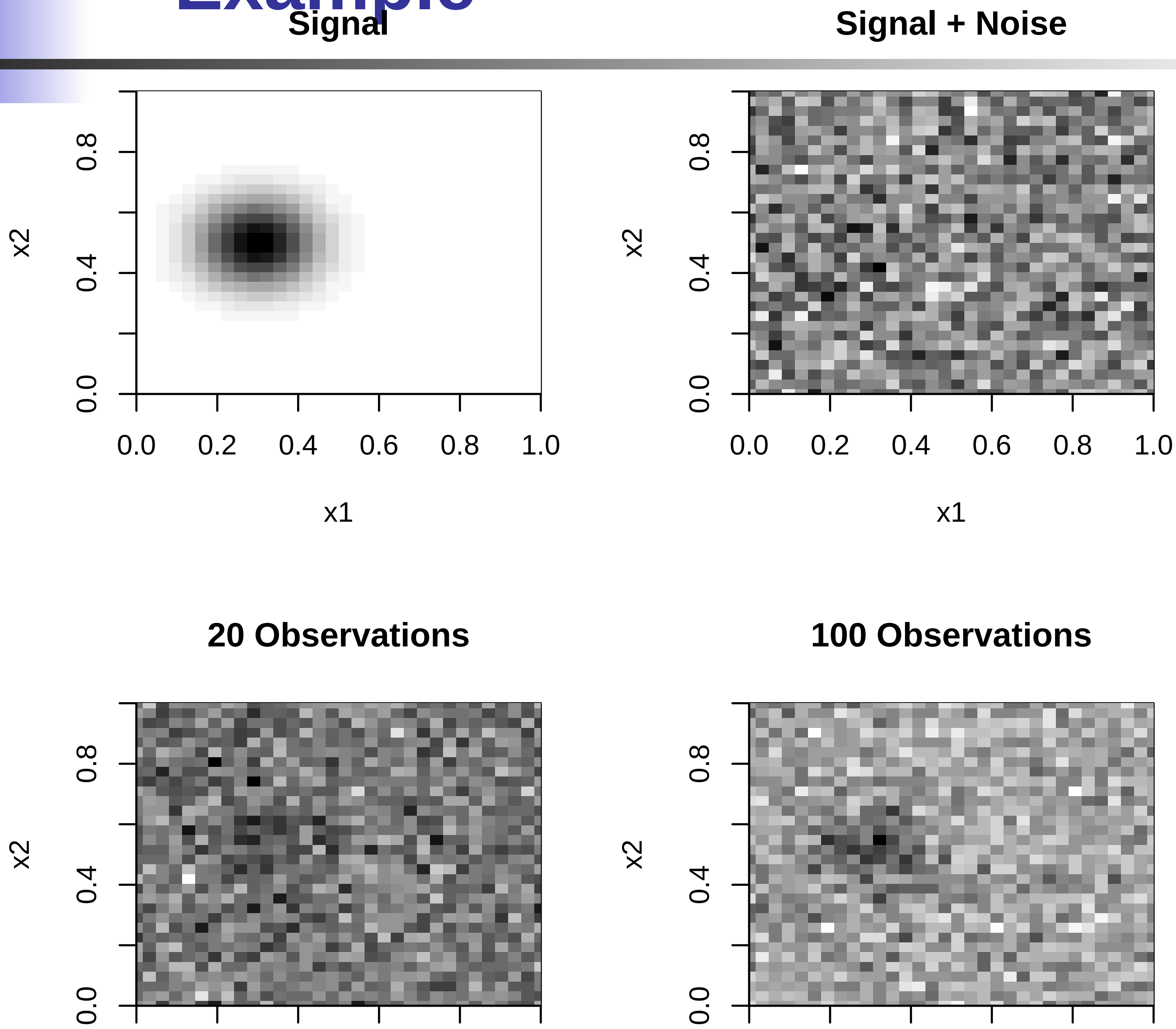
where f has been standardized so that $\int f^2(t)dt = 1$, and under the hypothesis of no signal at t, $dX_s = dW_s$ is Gaussian white noise. If there is a signal at τ, $dX_s = \sigma_0^{-1/4} \xi f_0[(s-\tau)/\sigma_0]ds + dW_s$. Ideally $f = f_0$; often f is taken to be a Gaussian kernel. The presence of a signal can be detected by $\max_{t,\sigma} Z_{t,\sigma}$. If f and f_0 are indicators of an interval, this would be a change-point problem. Smooth f_0 fits into the framework of Davies. Worsley applied geometric methods along the lines pioneered by Hotelling and Weyl and developed by Adler to estimate significance thresholds; and he applied these results in a series of papers.

Sequential Detection of a Structured Image

An image is a finite collection of pixels $\mathcal{J}$. At each time t and pixel location j is an observation $X_{t,j}$. Assume that all observations are Gaussian, independent, and have a known variance σ^2. Initially they also have mean 0.

The shape of a signal is given by a real valued function g, defined on $\mathcal{J}$; its intensity is a positive number ξ. When a signal is present, the expectation of the measurement $X_{t,j}$ is $\xi g(j)$; otherwise it is 0. The image at time t is the random field $X_t = \{X_{t,j} : j \in \mathcal{J}\}$. The log-likelihood ratio for testing a single image for the presence of a signal takes the form $\{\xi\langle g, X_t\rangle - \xi^2\langle g, g\rangle/2\}/\sigma^2$.

Example


Signal
Signal + Noise
20 Observations
100 Observations
x1
x2
0.0
0.2
0.4
0.6
0.8
1.0

Log Likelihood Ratios

The possible locations and shapes of the signals are assumed to belong to a smooth parametric family $\{g_\tau : \tau \in T\}$, for a suitable compact set T. A signal is parameterized by $\theta = (\xi, \tau)$. The conditional log-likelihood for the sequence of images up to time n for the case that the signal first appeared at period $k+1$ is

$$\ell_k^\theta(n) = \sum_{t=k+1}^{n} \{\xi\langle g_\tau, X_t\rangle - \xi^2/2\}/\sigma^2 = \frac{\xi}{\sigma^2}\langle g_\tau, S_k(n)\rangle - (n-k)\frac{\xi^2}{2\sigma^2}\ , \tag{3}$$

where $S_k(n) = \sum_{i=k+1}^{n} X_i$ and g_τ has without loss of generality been standardized by $\sum_j g_\tau^2(j) = 1$.

Expected Delay

Subject to a constraint on the frequency of false positives, frequently specified by the condition

$$\mathrm{E}_{\infty}(N) \sim B$$

as $B \to \infty$, we want to minimize the expected delay until detection, which we measure by the expected Kullback-Leibler information accumulated until the detection occurs:

$$\mathrm{E}_{k}^{\theta}[\ell_{k}^{\theta}(N)|N > k] = \frac{\xi^2}{2\sigma^2}\mathrm{E}_{k}^{\theta}(N - k|N > k),$$

which we expect to be of order $\log(B)$.

Stopping Rules

Let ρ denote a prior distribution over Θ_ε and consider the (marginal) log-likelihood ratios

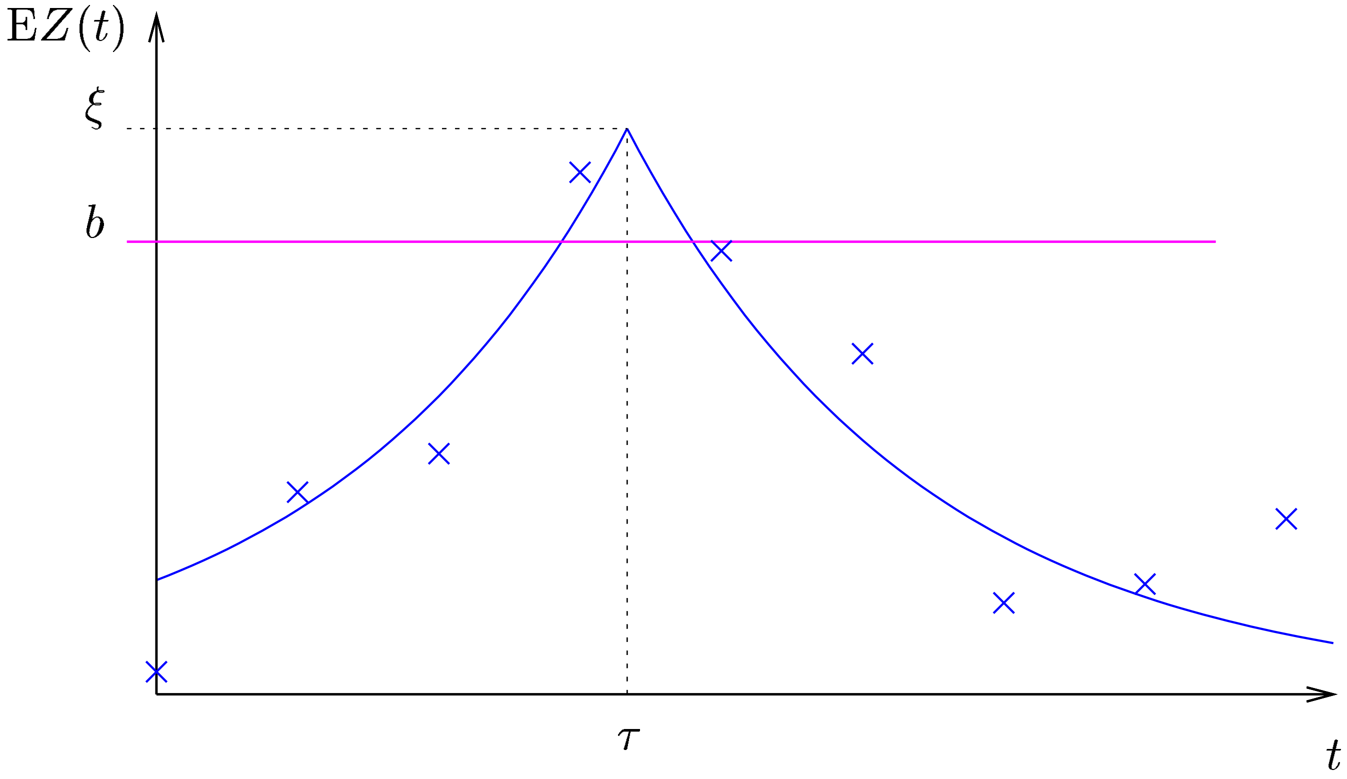
$$\ell_k(n) = \log \int_{\Theta_\varepsilon} \exp\{\ell_k^\theta(n)\}\rho(\theta)d\theta \tag{4}$$

for $k \leq n$. The (window restricted) Shiryayev-Roberts stopping rule is defined by

$$N_{\rm SR}^w = \inf\{n : r^w(n) \geq \log B + \log \lambda_{\rm SR}\} \ ,$$

where $r^w(n) = \log \sum_{k=n-w}^n e^{\ell_k(n)}$. The constant $\lambda_{\rm SR}$ is a correction factor that produces the asymptotic false detection rate $\mathrm{E}_\infty(N_{\rm SR}^w) \sim B$ as $B \to \infty$. The asymptotic expected delay is proportional to $\log(B)$ and can be evaluated asymptotically up to to $o(1)$. A similar result can be obtained by maximizing over $\theta = (\xi, \tau)$ and/or k (In the spirit of Page and Lorden) instead of integrating.

A Typical Genetic Process



Example 1: Linkage/Association Analysis

A commonly used statistic to detect linkage in an experimental cross or in human genetics or to detect association in human genetics is $\max_t Z_t$, where for each t, Z_t has been standardized so that under the hypothesis that t is unlinked (not associated) it has a standard normal distribution. The maximum is taken over a set of markers, the cardinality of which can range from a few hundred to several hundreds of thousands. If τ denotes a locus contributing to the phenotype of interest, then $\xi = \mathrm{E}(\mathrm{Z}_\tau)$ is one measure of the genetic effect of that locus. Our goal is to detect genetic loci contributing to the trait, while modeling their interactions with each other and with the environment.

Example 2: DNA/Protein Sequence Analysis

A “score” is attached to each amino acid in a protein. On average the scores are negative. To test whether there is a “high scoring segment,” Dembo, Karlin, and Kawabata (1990) suggest a statistic of the form

$$M_T = \max_{0 \le t \le T} [S_t - \min_{0 \le k \le t} S_k] \tag{5}$$

and use earlier results of Iglehart to give a p-value for this statistic.

Dembo, Karlin and Zeitouni (1994) give an approximate p-value for DNA/Protein sequence alignment when gaps are not allowed. Siegmund and Yakir (2001, correction 2004) gave an approximation when there are affine gap costs with a large cost for a new gap.

Remark. $\mathrm{P}_0\{M_T \ge b\} = \mathrm{P}_0\{N_0(b) \le T\}$, where N_0 is the stopping rule suggested by Page.

Copy Number Variants (CNV) (Olshen and Venkatraman, 2004)

Assume that $y_t, 1 \leq t \leq T$ are independent and normally distributed with unit variance and with mean values equal to μ_i for $\tau_i < t \leq \tau_{i+1}$, where $0 = \tau_0 < \tau_1 < \ldots \tau_{m+1} = T$. The parameters m, τ_i for $1 \leq i \leq m$, and the μ_i are in general all unknown. To simplify the exposition we assume that $\mu_0 = 0$. Our goal is to decide on the correct value of m and the change-points $\tau_1, \ldots, \tau_m$.

For purposes of analytic analysis, we assume that $T >> 1$ and $1 << m << T$, while $\tau_{i+1} - \tau_i$ are uniformly sufficiently large that the μ_i can be estimated consistently.

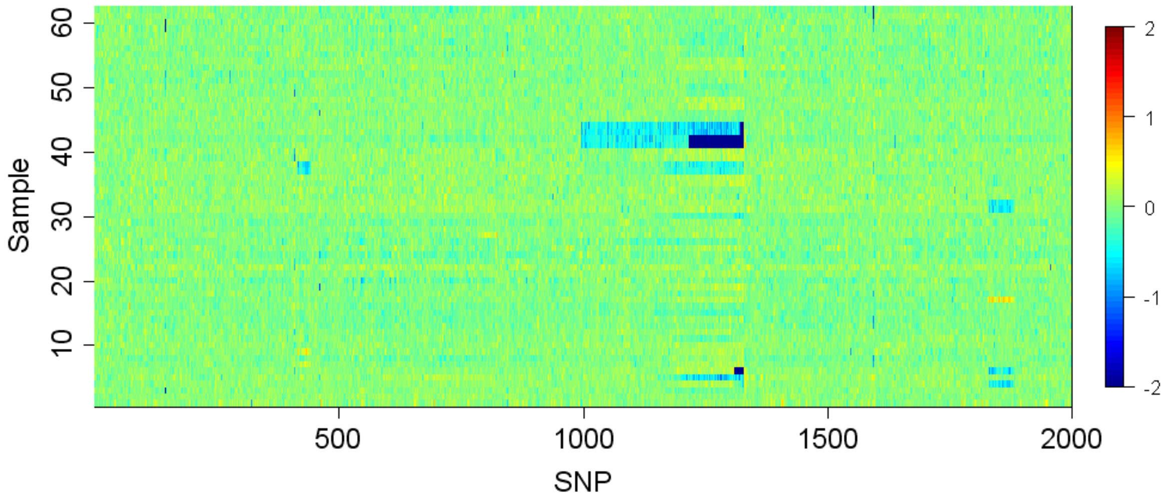
Variations

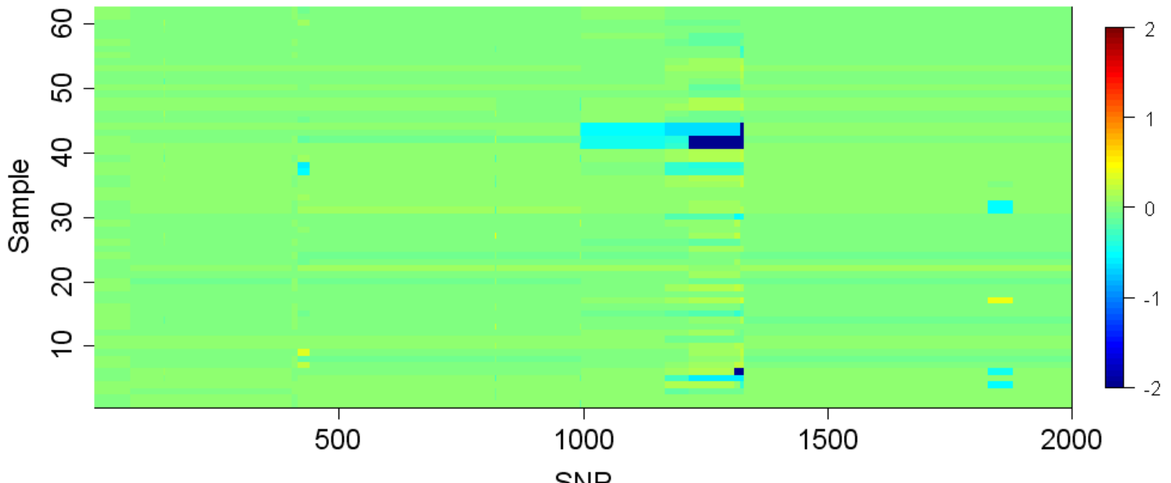
(1) There is a “baseline” value, say 0, such that $\mu_i = 0$ for even values of i.

(2) There are N sequences of observations $y_{nt}, 1 \leq t \leq T$, subsets J_i of which have a changed mean value in the interval $(\tau_i, \tau_{i+1}]$.

(3) In a gene mapping experiment there are m genetic loci contributing to the trait at genomic locations $\tau_1, \ldots \tau_m$. They may have pairwise (and higher way) interactions between them and/or with certain environmental variables. We want to determine m, the genomic locations, and the QTL effects (main effects and/or interactions).

Chromosome 22 CNV data





Test Statistic for a Single Sequence (Olshen and Venkatraman)

For a single sequence, $\{y_1, \ldots, y_T\}$, let $S_t = y_1 + \ldots + y_t$, $\bar{y}_t = S_t/t$, and $\hat{\sigma}^2 = T^{-1}\sum_1^T (y_t - \bar{y}_T)^2$. Consider

$$\max_{s,t} U^2(s,t),$$

where

$$U(s,t) = \hat{\sigma}^{-1}\{S_t - S_s - (t-s)\bar{y}_T\}/[(t-s)\{1-(t-s)/T\}]^{1/2},$$

and the max is taken over $0 \leq s < t \leq T, \;\; T_0 \leq t-s \leq T_1$. Here $T_0 < T_1$ are assumed lower and upper bounds on the length of a variant interval. Often $T_1 << T$, but it can be the entire interval. Within broad limits the exact choice of T_1 is unimportant.

Changepoint Model for Aligned CNV

Let the observed data be a two dimensional array $\{y_{nt} : 1 \leq n \leq N,\ 1 \leq t \leq T\}$, where $y_{n,t}$ is the data point for the n-th profile at location t. We assume that for each n, the random variables $\{y_{it} :\ 1 \leq t \leq T\}$ are mutually independent and normally distributed with mean values μ_{nt} and variances σ_i^2. Under the null hypothesis, for each profile i, the random variables $y_{n,t}$ are identically distributed with "baseline" mean value μ_n. Under the alternative hypothesis there exist values $1 \leq \tau_1 < \tau_2 \leq T$ and a set of profiles $\mathcal{J} \subset \{1, \ldots, N\}$, such that for $n \in \mathcal{J}$, $\mu_{nt} = \mu_{n0} + \delta_n I_{\{\tau_1 < t \leq \tau_2\}}$, where the δ_n are non-zero constants. Under the alternative detection and estimation of the change-points τ_1 and τ_2 are the primary interest; a secondary interest is determining the subset $\mathcal{J}$. We refer to $(\tau_1, \tau_2]$ as a variant interval. In the applications we discuss there are usually multiple variant intervals within the data, defined by different τ_1 and τ_2, and $\mathcal{J}$.

Multiple Sequences

In some problems (e.g., for inherited copy number variations), we expect a relatively small set of the sequences to contain any particular variant interval. The log likelihood function for a putative CNV in $(s, t]$

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

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 $U_i(s,t) = [\sum_{s+1}^{t} y_{v,i}]/(t-s)^{1/2}$. For smooth f we consider statistics of the form

$$\sum_{1}^{N} f[U_i(s,t)].$$

BIC: The Classical Formulation

Assume a family of “regular” parametric models indexed by $k = 1, \ldots K$ with log likelihood functions $\ell_k(\theta_k)$ for $\theta_k \in \Theta_k$. The posterior probability of the kth model when there is a smooth prior density $\pi_k(\theta_k)$ is proportional to $\int_{\Theta_k} \exp[\ell_k(\theta_k)]\pi_k(\theta_k)d\theta_k$. By using the expansion $\ell_k(\theta_k) \approx \ell_k(\hat{\theta}_k) + (\theta_k - \hat{\theta}_k)'[-\ddot{\ell}_k(\hat{\theta}_k)](\theta_k - \hat{\theta}_k)$, one sees that the logarithm of the posterior probability of the kth model, up to terms that are bounded as the sample size n becomes infinite, is the Bayes Information Criterion:

$$\mathrm{BIC} = \max_k \{\ell_k(\hat{\theta}_k) - 2^{-1} d_k \log(n)\},$$

where d_k is the dimension of the parameter space Θ_k. Note that terms we have neglected could be important if the number of models, $K = K_n$ is not bounded.

BIC: Single Sequence

Let X_τ by a vector with ith coordinate $X_{\tau,i} = (S_T/T - S_{\tau_i})$ and covariance matrix. Σ_τ. The likelihood function can be shown to equal

$$\ell(\tau, \delta) = \exp\{\delta' X_\tau - \delta' \Sigma_\tau \delta/2\}, \qquad (6)$$

where $\delta_i = \mu_i - \mu_{i-1}$ for $i = 1, \ldots, m$.

If the τ_i are uniformly distributed on $(0, T]$ and m remains bounded as $T \to \infty$, then (Zhang and Siegmund, 2006)

$$\mathrm{BIC1} = \ell(\hat{\tau}, \hat{\delta}) - .5 \sum_{i=1}^{m+1} \log(\hat{\tau}_i - \hat{\tau}_{i-1}) - (m - 1/2) \log T.$$

Prior Distributions

The classical BIC procedure does not depend on the specific prior distribution, although it would if the expansion of the Bayes factor were carried out to terms of constant order of magnitude, or if the number K of models were allowed to increase indefinitely with the sample size n. See Berger, Ghosh and Mykhopadhyay (2003). Modifying slightly the prior suggested by B, G, and M (2003), we assume δ, conditional on ω, is multivariate normally distributed with mean 0 and covariance matrix (a) $\omega^{-1}\Sigma_{\tau}^{-1}$, or (b) $\omega^{-1}I$, where ω has a smooth density that is positive and bounded on $[0, \infty)$.

The τ_i are assumed to be order statistics from a uniform distribution on $[0, T]$.

BIC

$$\mathrm{BIC2} = 2^{-1} X'_{\hat{\tau}} \Sigma_{\hat{\tau}}^{-1} X_{\hat{\tau}} - 2^{-1} m \log(X'_{\hat{\tau}} \Sigma_{\hat{\tau}}^{-1} X_{\hat{\tau}}/m) - m/2$$

$$+ \log(m!) - m \log(T) - \sum_{1}^{m} \log[\hat{\delta}_i^2]$$

$$+mE\{\log \int_{-\infty}^{\infty} \exp[W_t - |t|/2] dt\} - mE\{\max_t (W_t - |t|/2)\},$$

where $\delta_i = \mu_{i+1} - \mu_i$.

$$\mathrm{BIC3} = 2^{-1} X'_{\hat{\tau}} \Sigma_{\hat{\tau}}^{-1} X_{\hat{\tau}} - 2^{-1} m \log(X'_{\hat{\tau}} X_{\hat{\tau}}/m) - m/2$$

$$-2^{-1} \sum \log(\hat{\tau}_{i+1} - \hat{\tau}_i) + \log(m!) - m \log(T) - \sum_{1}^{m} \log[\hat{\delta}_i^2]$$

$$+mE\{\log \int_{-\infty}^{\infty} \exp[W_t - |t|/2] dt\} - mE\{\max_t (W_t - |t|/2)\}.$$

Large BIC for Multiple Sequences I

Assume that each $X_{\tau,i}$ is an N-dimensional vector with covariance matrix W. Then X_τ is an mN dimensional vector with covariance matrix$W \otimes \Sigma$. For simplicity we assume that W is the identity matrix and Σ is diagonal with entries $(\tau_{i+1} - \tau_i)^{-1}$, as above; but with minor notational changes one can handle the more general case. Assume that in the ith interval, $\tau_{i+1} - \tau_i$ there is a subset $J_i \subset \{1, \ldots, N\}$ where $\delta_{i,j} = \mu_{i,j} - \mu_{i-1,j} \neq 0.$

BIC for Multiple Sequences II

With an appropriate multivariate normal prior for δ (conditional on the change-points and the sets J_i) and a Bernoulli prior for the J_i,

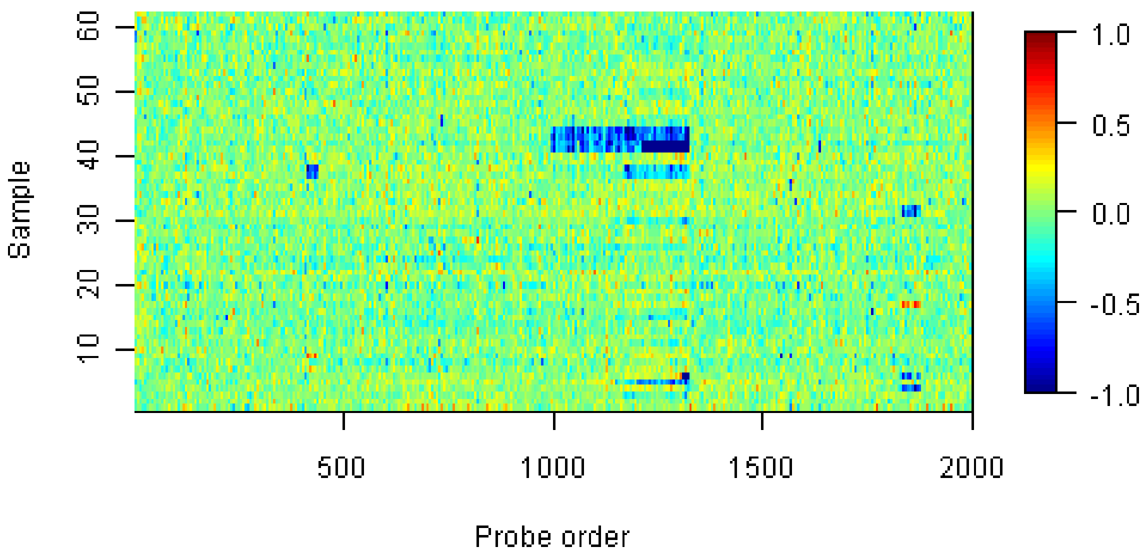
$$\mathrm{BIC2} = \ell(\hat{\tau}, \hat{\delta}) - 0.5M \log\{\frac{\ell(\hat{\tau}, \hat{\delta})}{M}\} - .5M - \sum_{i=1}^{m} \log[\sum_{j \in J_i} \hat{\delta}_{i,j}^2]$$

$$+ \log m! - m \log(T) + \sum_i \log \int \exp[W_{i,t} - |t|/2] dt - \sum_i \max_t (W_{i,t} - |t|/2)$$

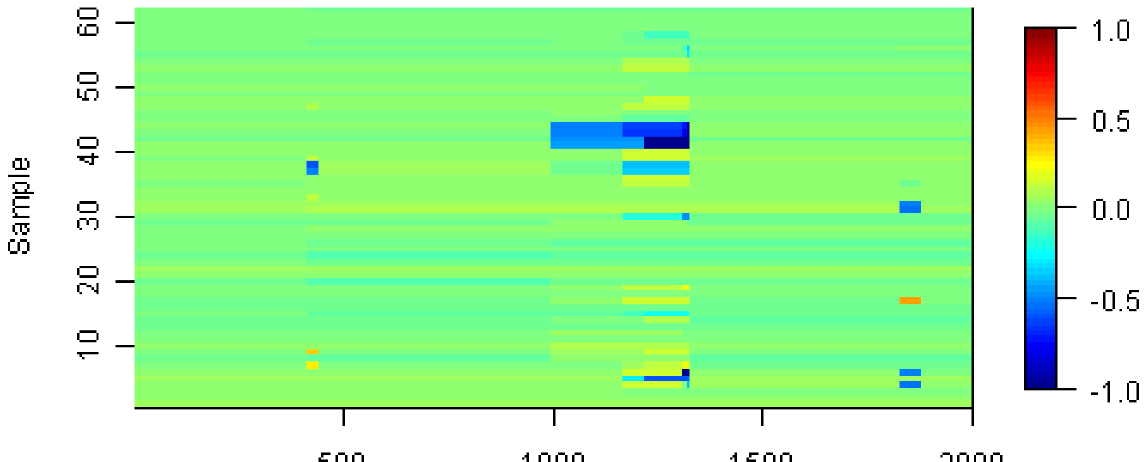
$$+ mN[\hat{p} \log(\hat{p}) + (1 - \hat{p}) \log(1 - \hat{p})],$$

where $M = \sum_i |J_i|$ and $\hat{p} = M/mN$.

Chromosome 22

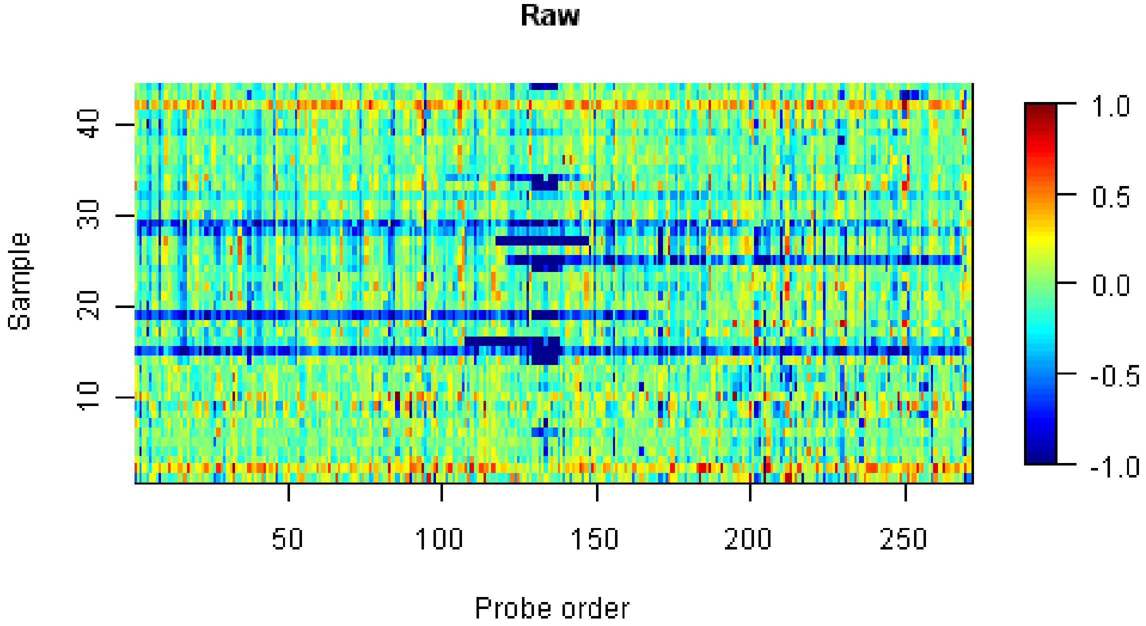


Raw

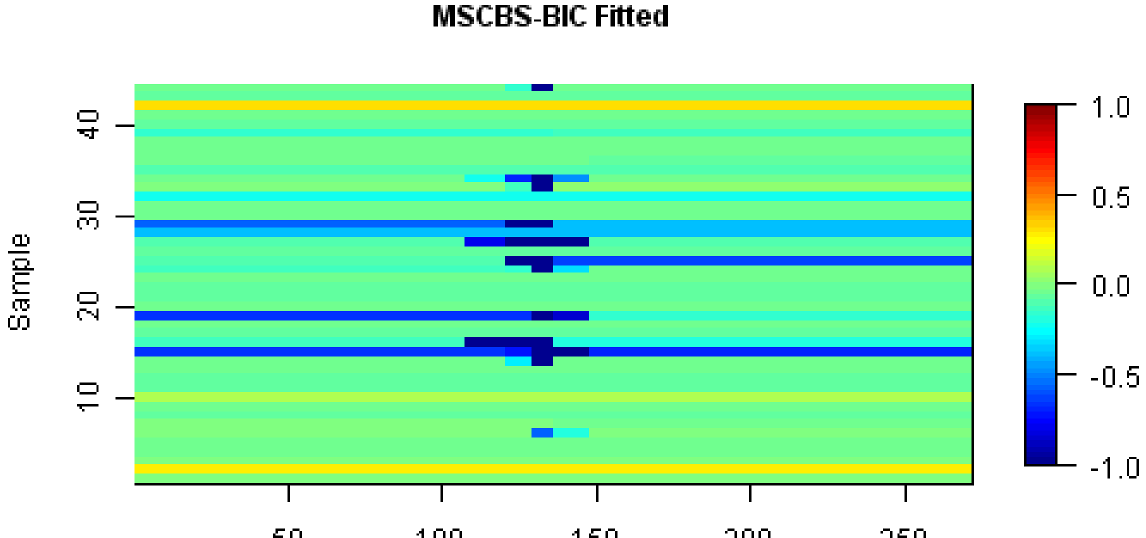


MSCBS BIC Fitted

Leukemia Example



Raw



MSCBS-BIC Fitted

References

Siegmund, D. (2013). Change-points: from sequential detection to biology and back (with discussion). *Sequential Analysis* **32** 2-14, 43-46.

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.

Siegmund, D. and Yakir, B. (2008). Detecting the emergence of a signal in a noisy image *Statistics and Its Interfaces* **1**, 3-12.

Zhang, N.R. and Siegmund, D.O. (2007) A Modified Bayes Information Criterion with Applications to the Analysis of Comparative Genomic Hybridization Data, *Biometrics* **63**, 22-32.

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