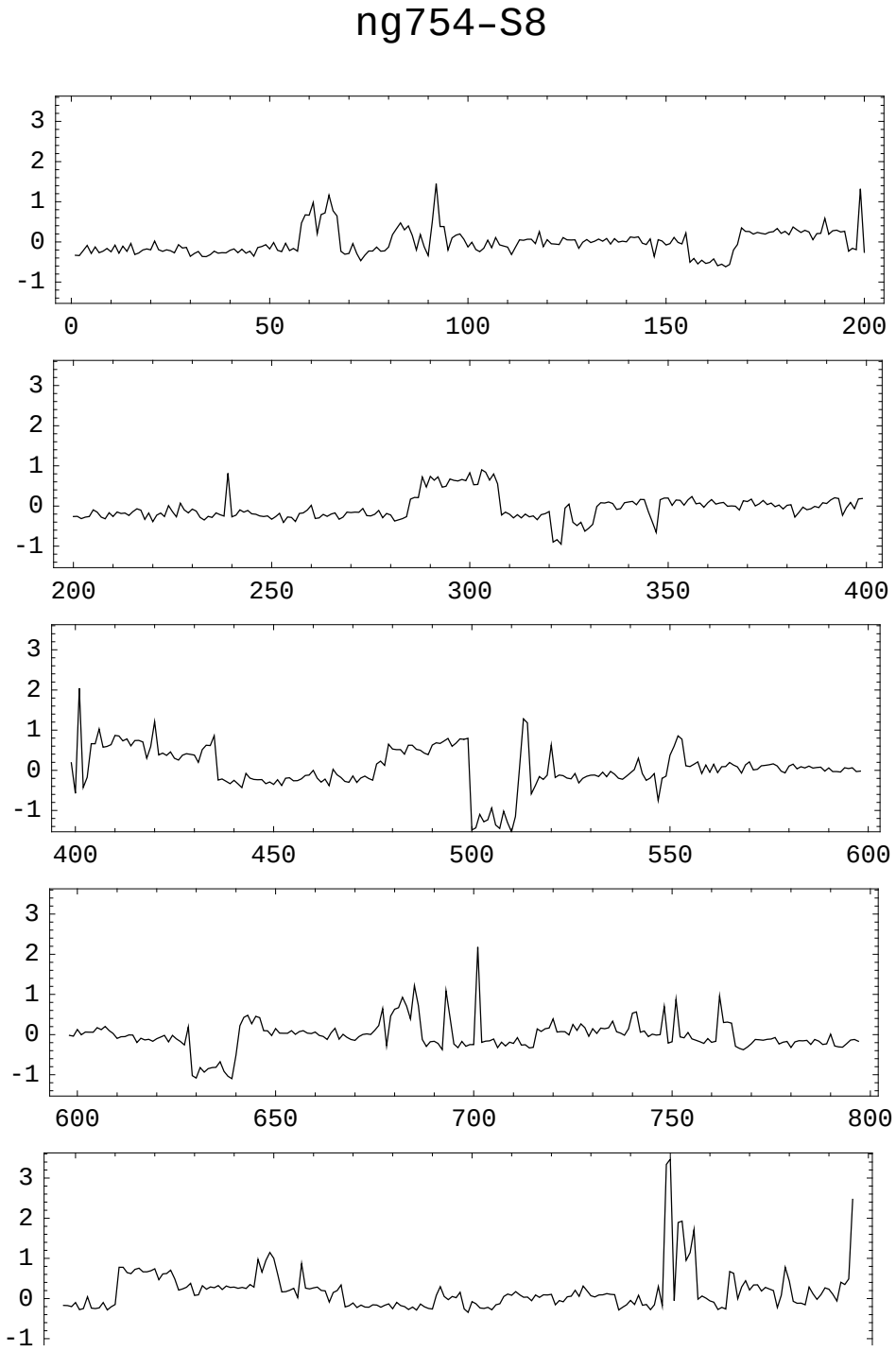
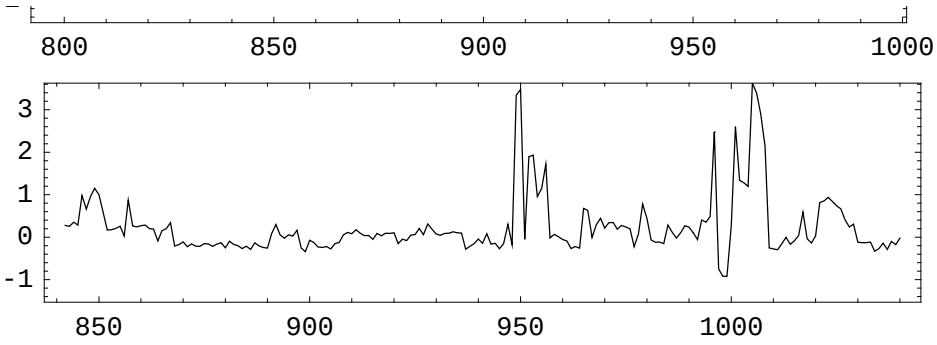
Change Point Detection using Wavelets

Introduction

The object is to detect jumps in array CGH data such as that shown in the plots below.



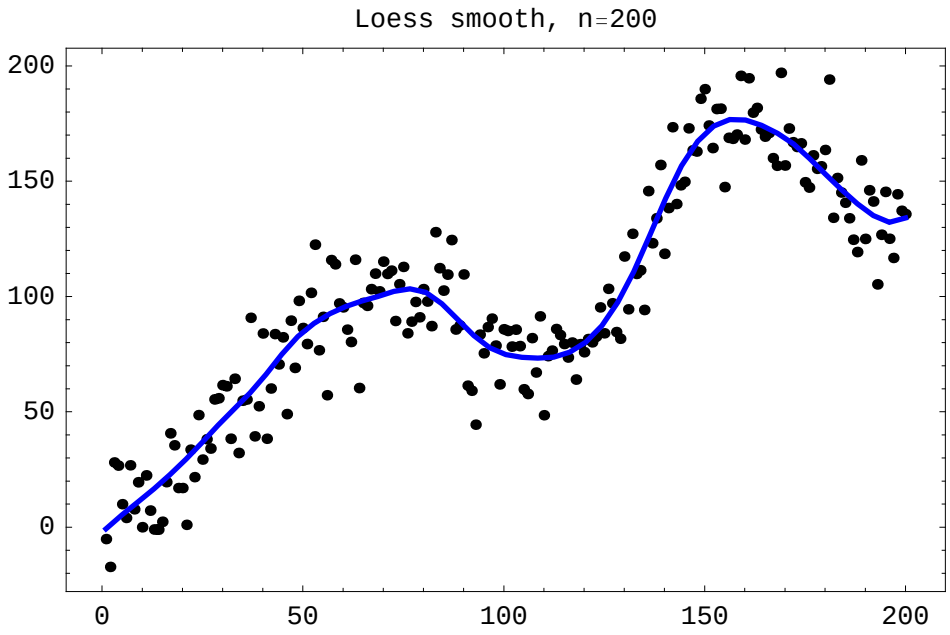


Given a sequence of values $z_0,\ z_1,\ \ldots,\ z_{N-1}$ assumed to be generated by

$z_t = f(t) + e_t$

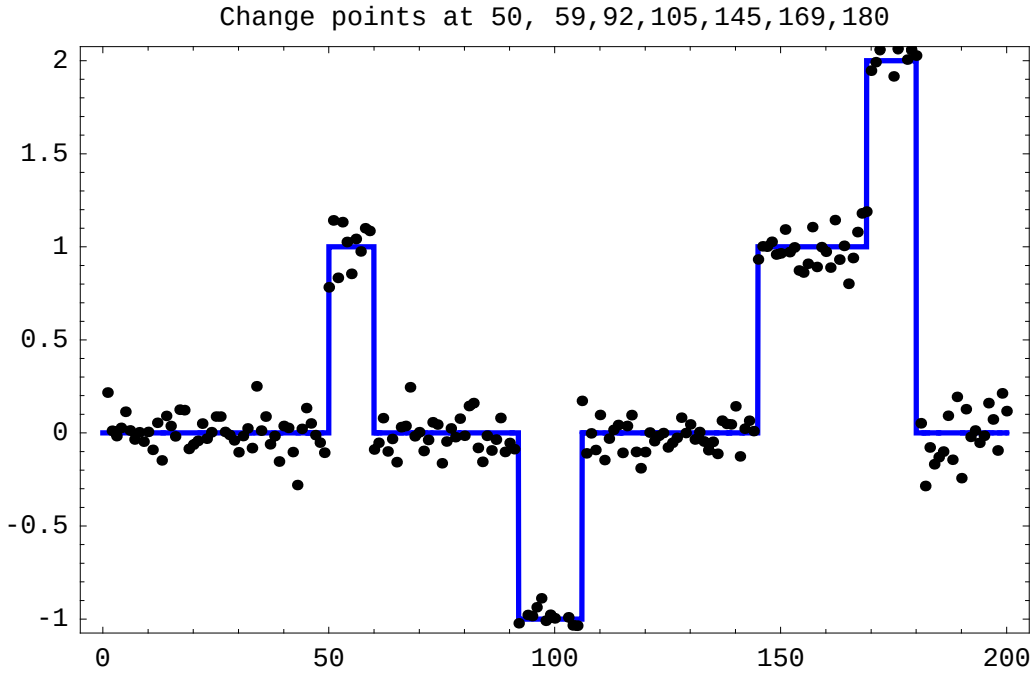
where $f(t)$ is function and $e_t \sim \text{NID}(0, \sigma^2)$. Two examples are shown below. In the first example, we have a smooth curve with random variation about it.

- **Example 1: Smooth Curve with Random Errors**



- **Example 2: Random points about a curve with jumps**

In the second example there are 7 points of discontinuity.



Discrete Wavelet Transformation

The partial DWT is a special orthonormal transformation:

$(z_0, \cdots, z_{N-1}) \longleftrightarrow (W_1, \cdots, W_J, V_j)$,

where W_j is a vector of length $N_j = N/2^j$ and V_J has the same length as W_J, N_J. For simplicity we have assumed that N is a multiple of 2^J. The vector W_j is called the vector of wavelet coefficients at level j and is associated with changes or differences on scale 2^{j-1}. The vector V_J, the scaling coefficients at level J, is associated with averages on scale 2^{J-1}.

We can write

$W = \mathcal{W} X$,

$$\mathcal{W} = \begin{pmatrix} \mathcal{W}_1 \\ \cdots \\ \mathcal{W}_J \\ \mathcal{V}_J \end{pmatrix}$$

$\mathcal{W}_j$ is $N_j \times N$, $j = 1, \cdots, J$ and $\mathcal{V}_J$ is $N_J \times N$.

In practice the DWT is computed using the pyramid algorithm which requires only $O(N)$ flops.

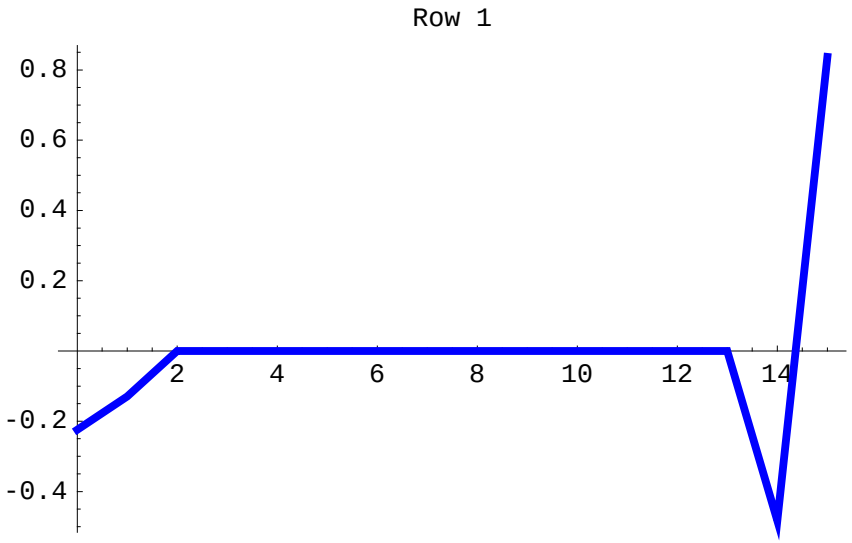
Daubechies D4 Wavelet Filter

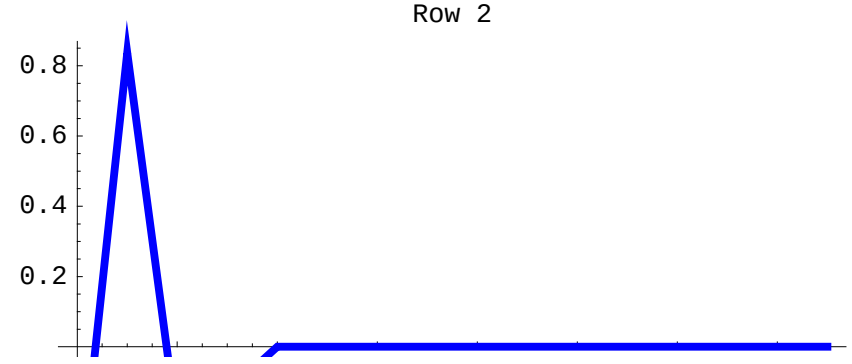
$$(a_0, a_1, a_2, a_3) = (\,-0.12941 \quad -0.224144 \quad 0.836516 \quad -0.482963\,)$$

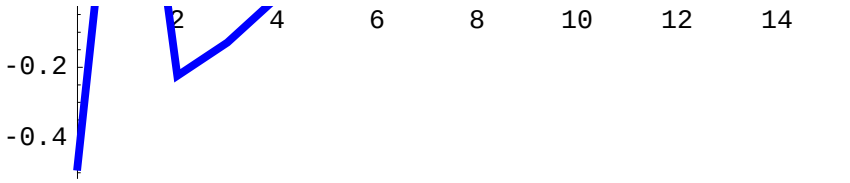
Taking $N = 16$,

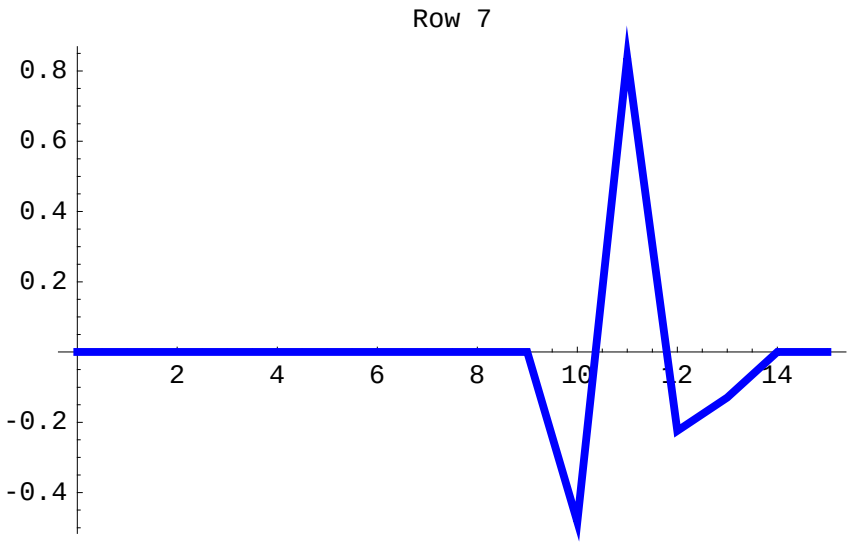
$$W_1 = \begin{pmatrix} a_1 & a_0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & a_3 & a_2 \\ a_3 & a_2 & a_1 & a_0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & a_3 & a_2 & a_1 & a_0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & a_3 & a_2 & a_1 & a_0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & a_3 & a_2 & a_1 & a_0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & a_3 & a_2 & a_1 & a_0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & a_3 & a_2 & a_1 & a_0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & a_3 & a_2 & a_1 & a_0 \end{pmatrix}$$

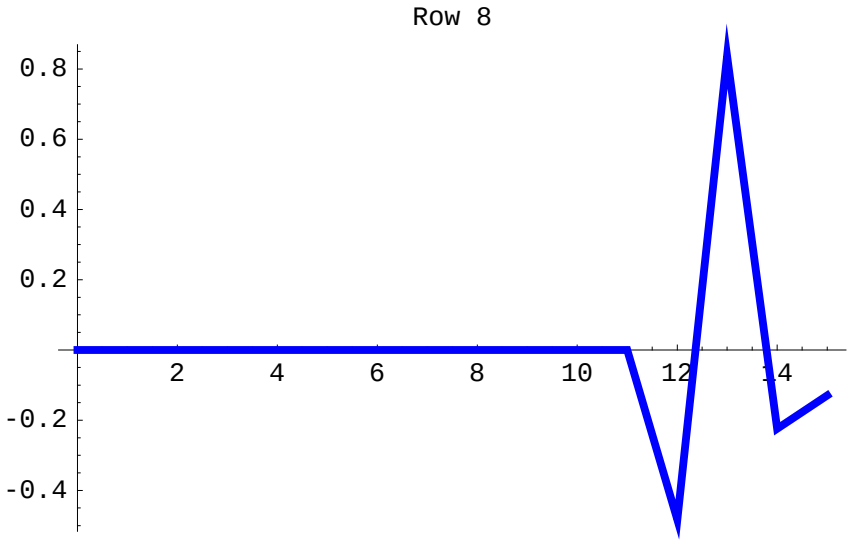
The rows of W_1 behave like approximately like first differences as can be seen from the diagram below which plots rows 1, 2, 7, 8.











■ Method of Wang (1995)

The underlying model or null hypothesis may be written $z_t = f(t) + a_t,\ t = 1, \cdots, n$ where $a_t \sim \text{NID}(0, \sigma^2)$ and $f(t)$ is a smooth function, ie. continuous and differentiable. The is a classic model in time series. Examples include the polynomial trend analysis (Fisher, 1921) and the lowess polynomial seasonal adjustment of Cleveland et al. (1990). Modern linear time series models such as the ARMA and its generalizations provide a

comprehensive and more realistic approach to building models which may be used for forecasting and intervention analysis.

Wang (1995) and Ogden and Parzen (1996) consider the problem of testing an abrupt change in the function $f(t)$. Ogden and Parzen (1996) approach the problem as one of estimating $f(t)$ when step functions are present and $f(t)$ is otherwise constant and their approach is again, like in MatLab, exploratory. Wang (1995) model is more general and focuses on detecting the change points at which a jump or sharp cusp occurs. A sharp cusp occurs at point t_0 if there exists a constant $K > 0$ such that

$$| f(t_0 + h) - f(t_0) | \geq K | h |^{\alpha}$$

for all h as $h \to 0$ and $0 \leq \alpha < 1$. When $\alpha = 0$, the function has a jump. Wang shows that, asymptotically with probability 1, all wavelet coefficients will have absolute value less than the universal threshold value $\sigma\sqrt{2\log(n)}$ provided there are no jumps or cusps. The asymptotics work best when $\alpha = 0$ which means in practice the wavelet coefficients are larger in absolute value. The unknown value of σ may be estimated robustly by the median absolute value of the wavelet coefficients at level 1 divided by 0.6745.

■ Practical Implementation Details

We follow the notation in the book Percival and Walden (2000) which is also used in the S-Plus, R and MatLab software. The principal differences are that level 1 refers highest time domain resolution and filter width rather than half-width is used in the naming of the Daubechies wavelets. So for example, Wang's $D(1)$ corresponds to $D(2)$ which is also equivalent to the Haar wavelet. Another difference is that Wang pads the real data so that n is a power of 2 because he uses Mallat's algorithm which is extremely efficient. However other algorithms which are only slight less efficient when n is not a power of 2 are available and will be used.

For detection of changepoints, Wang (1996) recommends examining plots of the absolute empirical wavelets at various levels j and finding those values which exceed the threshold line and are larger than others. Daubechies wavelets $D(k)$, $k = 2, 4, ..., 20$ The level j should be chosed as small as possible in order to obtain the highest for the changepoint time.

■ Cusum Test

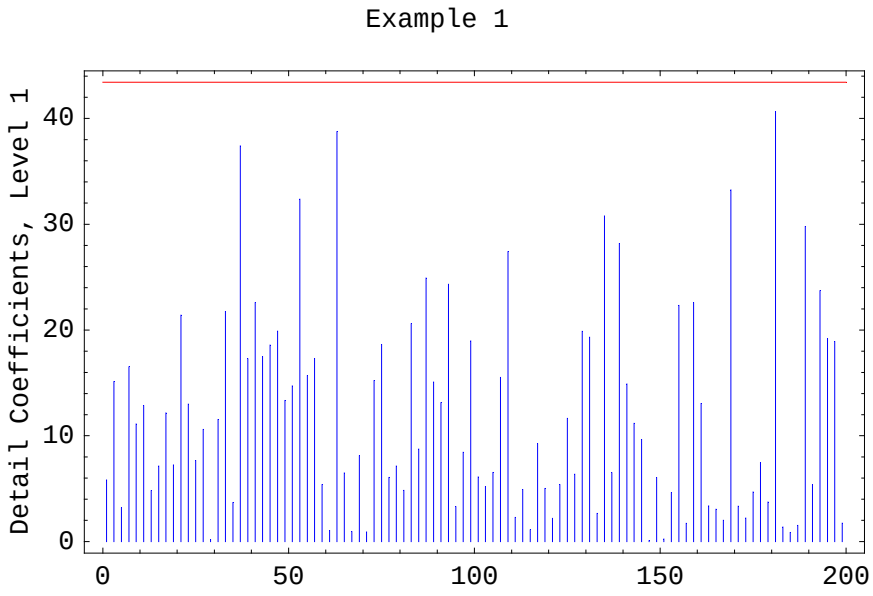
In changepoint problems it has often been found that the cusum provides an effective method of detecting changepoints (Page, 1955; Barnard, 1959; Lombard 1988)

The power of Wang's test can be improved by working with the cusum,

$$y_t = \sum_{s=1}^{t} z_s$$

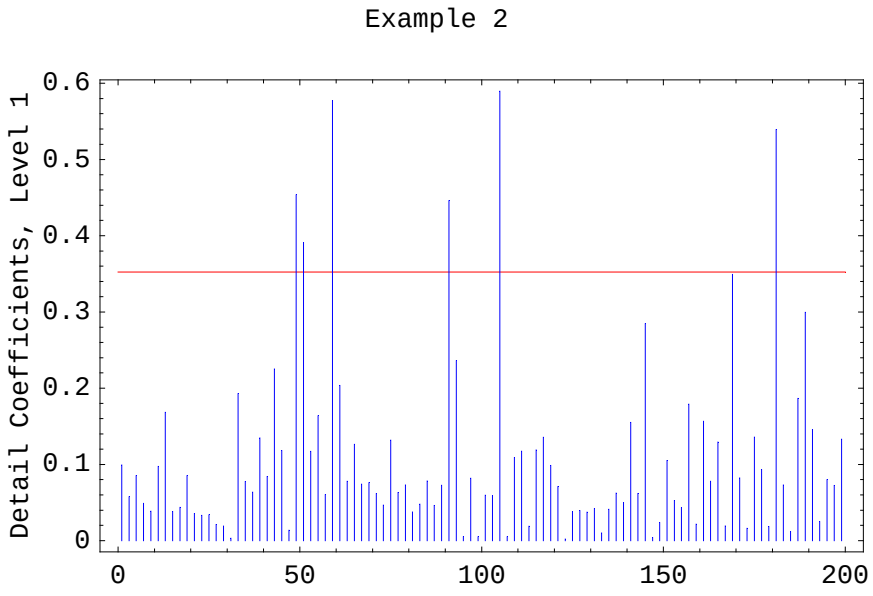
■ Illustrative Examples 1

In this example, no discontinuity is detected.



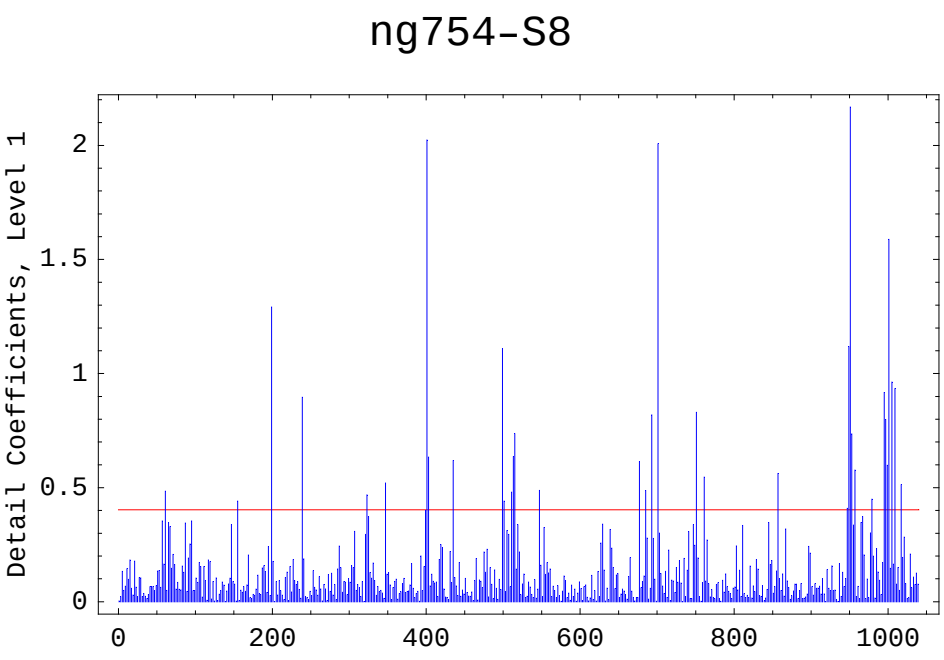
■ Illustrative Examples 2

In this example, 6 points are clearly detected and the 7th is very close to the boundary.

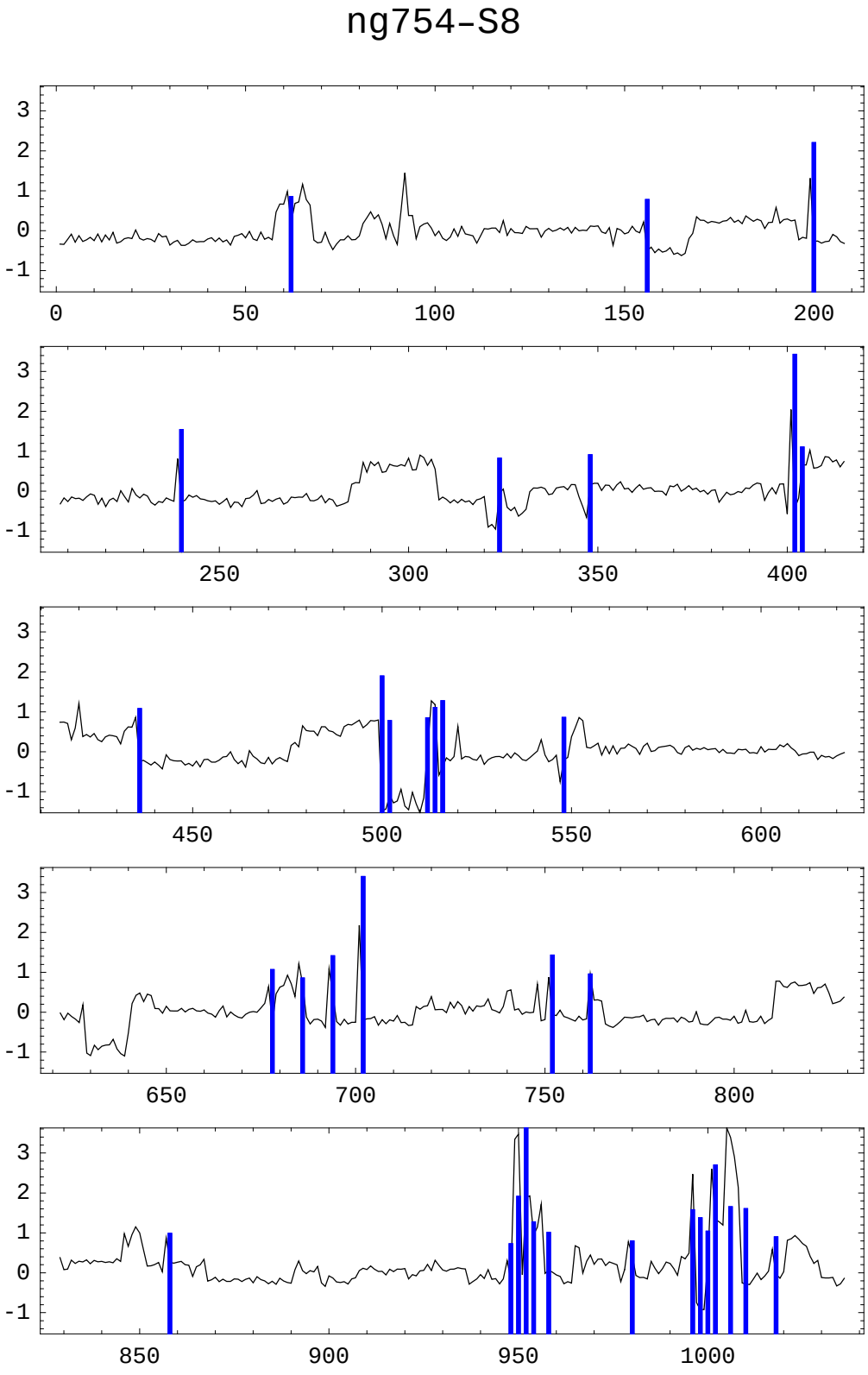


Application to NG754-S8 Data

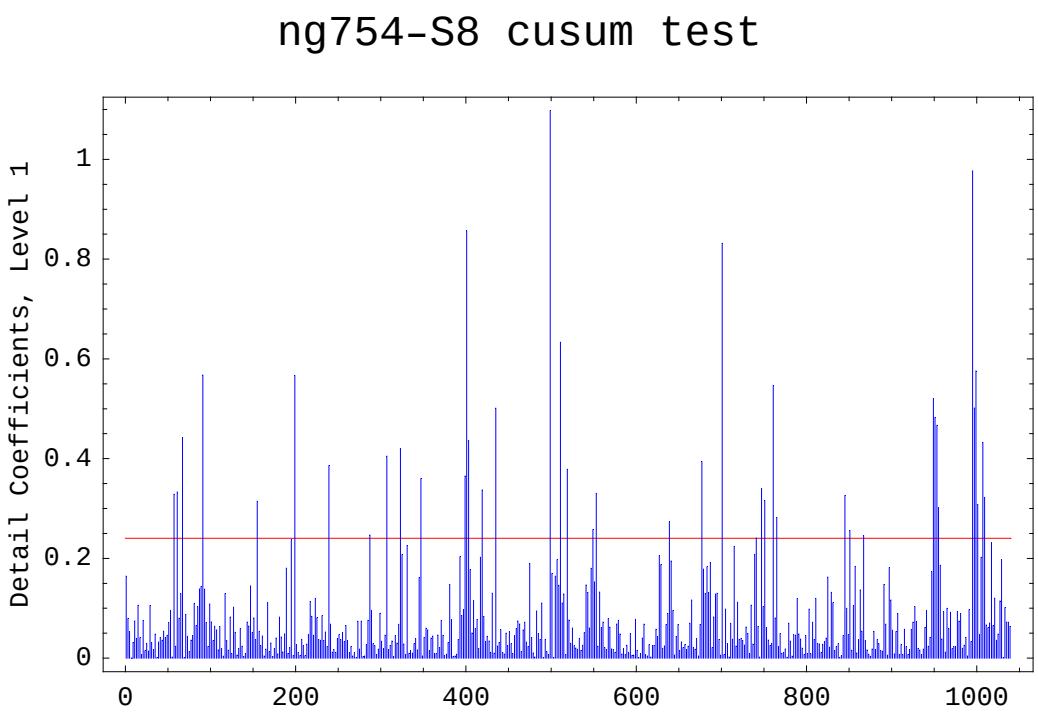
Applying Wang's test, quite a few change points are detected as shown below.



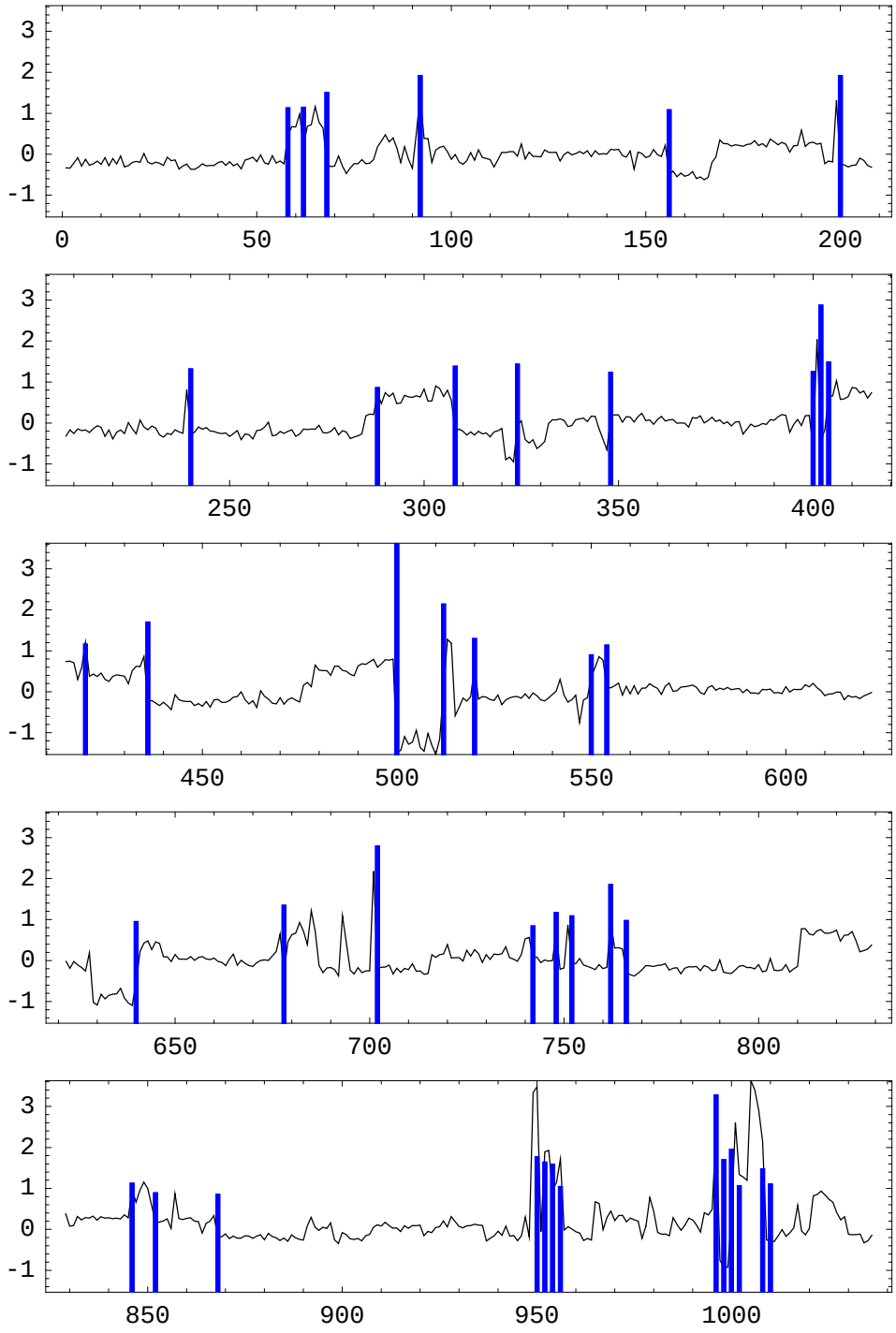
In the figure below we plot the data and a scaled version of the absolute values of those wavelet coefficients which exceed the threshold.



Next we apply the Wang test to the cusum.



As shown in the plot below the cusum approach does a better job of detecting changepoints.


ng754-S8 cusum test
3
2
1
0
-1
0
50
100
150
200
250
300
350
400
450
500
550
600
650
700
750
800
850
900
950
1000

References

Barnard, G.A.(1959). Control charts and stochastic processes. *Journal of the Royal Statistical Society B* 21, 239-270.

R. B. Cleveland, W. S. Cleveland, J.E. McRae, and I. Terpenning (1990) STL: A Seasonal-Trend Decomposition Procedure Based on Loess. *Journal of Official Statistics*, **6**, 3–73.

Fisher, R.A. (1921). Studies in crop variation. I. An examination of the yield of dressed grain from Broadbalk. *J. Agric. Sci.*, 11: 107-135.

Lai, T.L., Liu, H. and Xing, H. (2005). Autoregressive models with piecewise constant volatility and regression parameters. *Statistica Sinica* 15, 279-301.

Lombard, F. (1988). Detecting Change Points by Fourier Analysis. *Technometrics* 30, 305-310.

Ogden, Todd and Parzen, Emanuel (1976). Data dependent wavelet thresholding in nonparametric regression with change-point applications. *Computational Statistics & Data Analysis*, 22, 53-70.

Page, E.S. (1955). A test for change in a parameter occuring at an unknown point. *Biometrika* 42, 523-527.

Percival, D.B. & Walden, A.T. (2000). *Wavelet Methods for Time Series Analysis.* Cambridge University Press.

Wang, Yazhen (1995). Jump and Sharp Cusp Detection by Wavelets. *Biometrika*, 82, 385-397.

▼ Code