# REAL-TIME TREND DETECTION USING SEGMENTAL LINEAR REGRESSION

WILLIAM J. LONG

## 1. Introduction

The challenge when monitoring one or more sources of recuring data is the early detection of trends providing evidence of important changes in the state of the system. Noise in the observed data adds uncertainty to detection and means that statistical methods are necessary to have early and accurate detection. This paper will develop a method for detecting trends as the data arrives by fitting line segments using linear regression.

## 2. Methods

The essential idea is to fit a regression line to the data and compare that to the alternative of fitting two lines. When the two line explanation of the data is better by some criterian, it will be accepted as the explanation of the data and the second line becomes the running hypothesis until it too is broken into two lines. The criterian can use the variation in the data unexplained by each hypothesis as well as other information about the system including other monitoring data streams and external influences.

The problem then is to compare two explanations for the data: that the data is best explained by $y_i = a_1(x_i - x_0) + y_0$, where $(x_0, y_0)$ is the last accepted value or by that value for $x_i \leq x_1$ and by $y_i = a_2(x_i - x_1) + y_1$ for $x_i > x_1$ with $y_1 = a_1(x_1 - x_0) + y_0$. The values of $a_1$ and $a_2$ are to be determined by linear regression and the value of $x_1$ is to be determined by search among the available $x$ values in the data.

There are variations on this theme that we will consider. In the beginning, when the first points are arriving, there is no $(x_0, y_0)$ unless knowledge of the system provides an appropraite value. Thus, the initial problem includes determining $y_0$ by linear regression.

The equations we are describing correspond to a system for which changes in the slope of the monitored parameter reflect the changes in the system. There are other possible behaviors that can still be reflected in segmentally linear descriptions. The simplest is a single outlier. That is, one $y$ value was generated by a different noise process from the rest. Such behavior can be handled by leaving the point out of the regression computation and using a comparison criterion that accounts for the process generating the outlier in accordance with the system. The second possible behavior is a shift in the $y$ values at some point. This would mean that the best explanation for the second segment of the data is $y_i = a_1(x_i - x_0) + y_0 + b_1$ so that at $x_1$ there is a jump in $y_1$ of $b_1$ from $a_1(x_1 - x_0) + y_0$ to $a_1(x_1 - x_0) + y_0 + b_1$. The final alternative is that there is both a shift and a change of slope at a point, with the corresponding equation.

For the change of slope explanation, we may also want to consider changes that take place between the available data points. This can be handled using the mechanisms developed for a shift and change of slope.

2.1. **Derivation of Equations for Change of Slope.** The essential criteria for the computational mechanisms we will derive is that they can be applied efficiently with each new data point. Thus, we need a mechanism for preserving as much intermediate state as possible to make the computations for each new point incremental.

First, we will derive the equations for connected line segments with a change of slope. Assume an initial point at $(0, 0)$ and the break at $x_1$ with $n_0$ points from 0 to $x_1$ and $n_1$ points from $x_1$ to the most recent value. To maintain these assumptions as we move from segment to segment we will translate the axes appropriately.

Under these assumptions the equations for the line segments become $y = a_1 x_1$ if $x_i \leq x_1$ and otherwise $y = a_2(x_i - x_1) + a_1 x_1$.

The problem is to determine $a_1$ and $a_2$. The sum of the squares of the difference between the $y$ values and the linear estimates of the values, $\hat{y}$, is minimized when the following relations are true:

$$\frac{\delta}{\delta a_1} \sum_{x_i} (y_i - \hat{y})^2 = 0 \quad \frac{\delta}{\delta a_2} \sum_{x_i} (y_i - \hat{y})^2 = 0$$

Therefore:

$$\frac{\delta}{\delta a_1} \Big( \sum_{x_i \leq x_1} (y_i - a_1 x_i)^2 + \sum_{x_i > x_1} (y_i - a_2 x_i + a_2 x_1 - a_1 x_1)^2 \Big) = 0 \tag{2.1}$$

and:

$$\frac{\delta}{\delta a_2} \Big( \sum_{x_i \leq x_1} (y_i - a_1 x_i)^2 + \sum_{x_i > x_1} (y_i - a_2 x_i + a_2 x_1 - a_1 x_1)^2 \Big) = 0 \tag{2.2}$$

Solving 2.1:

$$-2\sum_{x_i\le x_1}((y_i-a_1x_i)x_i)-2x_1\sum_{x_i>x_1}(y_i-a_2x_i+a_2x_1-a_1x_1)=0$$

$$\sum_{x_i\le x_1}(y_ix_i-a_1x_i^2)+x_1\sum_{x_i>x_1}(y_i-a_2x_i)+n_1a_2x_1^2-n_1a_1x_1^2=0$$

To make these equations correspond to the way they will be computed, we will make the following substitutions:

$$\begin{array}{rlrlrl} s_{xy0}= & \sum_{x_i\le x_1} y_ix_i & s_{xx0}= & \sum_{x_i\le x_1} x_i^2 & s_{yy0}= & \sum_{x_i\le x_1} y_i^2 \\ s_{x0}= & \sum_{x_i\le x_1} x_i & s_{y0}= & \sum_{x_i\le x_1} y_i & s_{xy1}= & \sum_{x_i>x_1} y_ix_i \\ s_{xx1}= & \sum_{x_i>x_1} x_i^2 & s_{yy1}= & \sum_{x_i>x_1} y_i^2 & s_{x1}= & \sum_{x_i>x_1} x_i \\ s_{y1}= & \sum_{x_i>x_1} y_i & & & & \end{array}$$

Then:

$$s_{xy0}-a_1s_{xx0}+x_1s_{y1}-a_2x_1s_{x1}+n_1a_2x_1^2-n_1a_1x_1^2=0$$

Arranging this in terms of $a_1$ and $a_2$:

$$a_1(s_{xx0}+n_1x_1^2)+a_2x_1(s_{x1}-n_1x_1)=s_{xy0}+x_1s_{y1} \tag{2.3}$$

Solving 2.2:

$$-2\sum_{x_i>x_1}((y_i-a_2x_i+a_2x_1-a_1x_1)(x_i-x_1))=0$$

$$\sum_{x_i>x_1}(x_iy_i-a_2x_i^2+2a_2x_1x_i-a_1x_1x_i-a_2x_1^2+a_1x_1^2)=0$$

Making the same substitution and arranging the terms:

$$a_1x_1(s_{x1}-n_1x_1)+a_2(s_{xx1}-x_1(2s_{x1}-n_1x_1))=s_{xy1}-x_1s_{y1} \tag{2.4}$$

Equations 2.3 and 2.4 are now simultaneous equations in $a_1$ and $a_2$. If $a_1b_1+a_2c_1=d_1$ and $a_1b_2+a_2c_2=d_2$ then $a_1=\frac{c_2d_1-c_1d_2}{b_1c_2-b_2c_1}$ and $a_2=\frac{b_1d_2-b_2d_1}{b_1c_2-b_2c_1}$.

Thus, the denominator is:

$$(s_{xx0}+n_1x_1^2)(s_{xx1}-2x_1s_{x1}+n_1x_1^2)-x_1^2(s_{x1}-n_1x_1)^2$$

The numerator of $a_1$ is:

$$(s_{xx1}-2x_1s_{x1}+n_1x_1^2)(s_{xy0}+x_1s_{y1})-x_1(s_{x1}-n_1x_1)(s_{xy1}-x_1s_{y1})$$

The numerator of $a_2$ is:

$$(s_{xx0} + n_1 x_1^2)(s_{xy1} - x_1 s_{y1}) - (x_1(s_{x1} - n_1 x_1))(s_{xy0} + x_1 s_{y1})$$

This provides us with the equations necessary to compute $a_1$ and $a_2$ in terms of values that can be accumulated incrementally. That is, $s_{xy1}(i) = s_{xy1}(i-1) + x(i)y(i)$ and so forth.

The basis for deciding whether the slope actually changes is the unexplained variation in the $y$ values. This is:

$$\sum_{x_i} (y_i - \hat{y})^2$$

Substituting in the equations for $\hat{y}$ we have:

$$\sum_{x_i \leq x_1} (y_i - a_1 x_i)^2 + \sum_{x_i > x_1} (y_i - a_2 x_i + a_2 x_1 - a_1 x_1)^2 \tag{2.5}$$

Multiplying out and substituting, letting $q = x_1(a_2 - a_1)$:

$$s_{yy0} - 2a_1 s_{xy0} + a_1^2 s_{xx0} + s_{yy1} - 2a_2 s_{xy1} + 2q s_{y1} + a_2^2 s_{xx1} - 2q a_2 s_{x1} + n_1 q^2$$

By using the equalities in 2.3 and 2.4 this can be simplified to:

$$s_{yy0} + s_{yy1} - a_1 s_{xy0} - a_2 s_{xy1} + (a_2 - a_1)(x_1 s_{y1}) \tag{2.6}$$

These then are the computations that need to be made with each new data point.

The comparison is to $y_i = a_1 x_i$, since we make the same translation of the initial point to $(0,0)$. The minimization then produces a partial derivative that looks like the first part of equation 2.1, which with the substitutions is (let $x_1$ be beyond the last point, so the sums include all of the values):

$$s_{xy0} - a_1 s_{xx0} = 0$$

$$a_1 = \frac{s_{xy0}}{s_{xx0}} \tag{2.7}$$

The corresponding unexplained variation is the first part of equation 2.5, which with multiplication and substitution is $s_{yy0} - 2a_1 s_{xy0} + a_1^2 s_{xx0}$. Substituting from 2.7 simplifies this to:

$$s_{yy0} - a_1 s_{xy0} \tag{2.8}$$

Thus, the values in equations 2.6 and 2.8 need to be compared to decide whether a new segment should be added. This needs to be done for each point beyond $x_0$. To do this we need to keep vectors of the various sums for each of the points. The easiest way to do this is to keep two vectors, one with lists of the sums from $x_0$ to each possible $x_1$ and one with lists of the sums from the $x_1$'s to the most recent point. These can be updated for a new point as follows: Add a new sum to each of the lists in the first vector. Add the appropriate new value onto each item in each list of the second vector and add the value as a new item on each list. When a change in slope is accepted, the reference point ($x_0$) moves forward to the point of change and the vectors are recomputed.

2.2. **Starting the Segments.** Initially, there is no value for $y_0$. Thus, this needs to be determined by regression as well. The regression line with $x_0 = 0$ is $y_i = a_1 x_i + y_0$ and the problem is to find $a_1$ and $y_0$. Taking the partial derivative with respect to $a_1$ we get:

$$\frac{\delta}{\delta a_1} \sum_{x_i} (y_i - a_1 x_i - y_0)^2 = 0$$

$$\sum_{x_i} (x_i(y_i - a_1 x_i - y_0)) = 0$$

$$s_{xy0} - a_1 s_{xx0} - s_{x0} y_0 = 0$$

Taking the partial derivative with respect to $y_0$ we get:

$$\frac{\delta}{\delta y_0} \sum_{x_i} (y_i - a_1 x_i - y_0)^2 = 0$$

$$\sum_{x_i} (y_i - a_1 x_i - y_0) = 0$$

$$s_{y0} - a_1 s_{x0} - n_0 y_0 = 0$$

Solving simultaneously as above, the denominator is $n_0 s_{xx0} - s_{x0}^2$. The numerator for $a_1$ is $n_0 s_{xy0} - s_{x0} s_{y0}$ and the numerator for $y_0$ is $s_{xx0} s_{y0} - s_{x0} s_{xy0}$. The unexplained variance can be computed from:

$$\begin{aligned}
&\sum_{x_i} (y_i - a_1 x_i - y_0)^2 \\
&= s_{yy0} - 2a_1 s_{xy0} - 2y_0 s_{y0} + a_1^2 s_{xx0} + 2a_1 y_0 s_{x0} + y_0^2 \\
&= s_{yy0} - a_1 s_{xy0} - y_0 s_{y0}
\end{aligned}$$

To determine when the initial segment has a change in direction we need to compare this to $y_i = a_1 x_i + y_0$ for $x_i \leq x_1$ and $y_i = a_2(x_i - x_1) + a_1 x_1 + y_0$ for $x_i < x_1$.

Thus we need to solve for the three unknowns using the partial derivatives.

$$\frac{\delta}{\delta a_1}\Big(\sum_{x_i \leq x_1} (y_i - a_1 x_i - y_0)^2 + \sum_{x_i > x_1} (y_i - a_2 x_i + a_2 x_1 - a_1 x_1 - y_0)^2\Big) = 0$$

$$s_{xy0} - a_1 s_{xx0} - y_0 s_{x0} + x_1 s_{y1} - a_2 x_1 s_{x1} + n_1 a_2 x_1^2 - n_1 a_1 x_1^2 - y_0 n_1 x_1 = 0$$

$$a_1(s_{xx0} + n_1 x_1^2) + a_2 x_1(s_{x1} - n_1 x_1) + y_0(s_{x0} + n_1 x_1) = s_{xy0} + x_1 s_{y1}$$

With respect to $a_2$:

$$\begin{aligned} s_{xy1} - a_2 s_{xx1} + 2a_2 x_1 s_{x1} - a_1 x_1 s_{x1} - x_1 s_{y1} \\ -n_1 a_2 x_1^2 + n_1 a_1 x_1^2 - y_0 s_{x1} + n_1 y_0 x_1 = 0 \end{aligned}$$

$$\begin{aligned} a_1(x_1^2 n_1 - x_1 s_{x1}) - a_2(2x_1 s_{x1} - n_1 x_1^2 - s_{xx1}) \\ +y_0(n_1 x_1 - s_{x1}) = x_1 s_{y1} - s_{xy1} \end{aligned}$$

With respect to $y_0$:

$$s_{y0} - a_1 s_{x0} - y_0 + s_{y1} - a_2 s_{x1} + n_1 a_2 x_1 - n_1 a_1 x_1 - y_0 n_1 = 0$$

$$a_1(s_{x0} + n_1 x_1) + a_2(s_{x1} - n_1 x_1) + y_0(1 + n_1) = s_{y0} + s_{y1}$$

These simultaneous equations can then be solved. The denominator is:

$$\begin{aligned} den \;=\; & (s_{x1} - n_1 x_1)^2(2x_1 s_{x0} - s_{xx0} - n_1 x_1^2) \\ & + ((s_{x0} + n_1 x_1)^2 - (s_{xx0} + x_1^2 n_1)(n_1 + n_0))(2s_{x1} x_1 - s_{xx1} - x_1^2 n_1) \end{aligned}$$

The numerators are:

$$\begin{aligned}
a_1 &= (x_1 s_{y0} - s_{xy0})(s_{x1} - n_1 x_1)^2 + (s_{x0} - n_0 x_1)(s_{x1} - n_1 x_1)(s_{xy1} - x_1 s_{y1}) \\
&\quad + ((s_{x0} + n_1 x_1)(s_{y0} + s_{y1}) - (s_{xy0} + x_1 s_{y1})(n_1 + n_0))(2x_1 s_{x1} - s_{xx1} - x_1^2 n_1) \\
a_2 &= ((s_{xx0} + x_1^2 n_1)(s_{y0} + s_{y1}) - (x_1 s_{y0} + 2x_1 s_{y1} + s_{xy0})(s_{x0} + n_1 x_1)) \\
&\quad - (s_{xy0} + x_1 s_{y1}) x_1 (n_1 + n_0))(s_{x1} - n_1 x_1) \\
&\quad - ((s_{x0} + n_1 x_1)^2 - (s_{xx0} + x_1^2 n_1)(n_1 + n_0))(s_{xy1} - x_1 s_{y1}) \\
y_0 &= (s_{xy0} - x_1 s_{y0}) x_1 (s_{x1} - n_1 x_1)^2 - (s_{xx0} - x_1 s_{x0})(s_{xy1} - x_1 s_{y1})(s_{x1} - x_1 n_1) \\
&\quad + ((s_{xy0} + x_1 s_{y1})(s_{x0} + n_1 x_1) - (s_{xx0} + x_1^2 n_1)(s_{y0} + s_{y1})) \\
&\quad (2x_1 s_{x1} - s_{xx1} - x_1^2 n_1)
\end{aligned}$$

The unexplained variance is computed similarly.

$$((a_2 - a_1)x_1 - y_0)s_{y1} - s_{y0} y_0 + s_{yy0} + s_{yy1} - a_2 s_{xy1} - a_1 s_{xy0}$$

2.3. **Alternate Hypotheses.** If instead there is just a shift, that would be modelled as $y_i = a_1 x_i$ for $x_i \leq x_1$ and $y_i = a_1 x_i + y_0$ for $x_i > x_1$.

$$a_1 = \frac{s_{y1} s_{x1} - s_{xy0} - s_{xy1}}{s_{x1}^2 - s_{xx0} - s_{xx1}},$$

$$y_0 = \frac{s_{x1}(s_{xy0} + s_{xy1}) - s_{xx0} s_{y1} - s_{y1} s_{xx1}}{s_{x1}^2 - s_{xx0} - s_{xx1}}$$

The unexplained variance is:

$$s_{yy0} + s_{yy1} - a_1(s_{xy0} + s_{xy1}) - y_0 s_{y1}$$

If there is both a shift and a change, that would be modeled as $y_i = a_1 x_i$ for $x_i \leq x_1$ and $y_i = a_2 x_i + y_0$ for $x_i > x_1$.

$$a_2 = \frac{s_{xy1} - s_{y1} s_{x1}}{s_{xx1} - s_{x1}^2}, y_0 = \frac{s_{y1} s_{xx1} - s_{x1} s_{xy1}}{s_{xx1} - s_{x1}^2}, a_1 = \frac{s_{xy0}}{s_{xx0}}$$

Unexplained:

$$(s_{xx1} - s_{x1}^2) a_2^2 + s_{yy0} + s_{yy1} + a_1^2 s_{xx0} - 2a_1 s_{xy0} - s_{y1}^2$$

2.4. **Reasoning About Hypotheses.** Given a set of data there are several hypotheses that might explain it. The problem is how to decide among the hypotheses. Each hypothesis has a degree of fit as expressed by the unexplained variance and a probability for that hypothesis based on a model of the domain. Thus, the probability of a random noise value may be set at p1 and a the probability a change in the slope may be p2 while other changes are inconsistent with the underlying behavior of the system.

MIT-CSAIL
*E-mail address*: wjl@mit.edu