Hindawi Publishing Corporation
ISRN Applied Mathematics
Volume 2013, Article ID 346230, 4 pages
http://dx.doi.org/10.1155/2013/346230

Research Article

An Autocorrelation Term Method for Curve Fitting

Louis M. Houston

The Louisiana Accelerator Center, The University of Louisiana at Lafayette, Lafayette, LA 70504-4210, USA

Correspondence should be addressed to Louis M. Houston; houston@louisiana.edu

Received 7 May 2013; Accepted 8 July 2013

Academic Editors: K. Djidjeli, J. Kou, and M. Qatu

The least-squares method is the most popular method for fitting a polynomial curve to data. It is based on minimizing the total squared error between a polynomial model and the data. In this paper we develop a different approach that exploits the autocorrelation function. In particular, we use the nonzero lag autocorrelation terms to produce a system of quadratic equations that can be solved together with a linear equation derived from summing the data. There is a maximum of $2M$ solutions when the polynomial is of degree M. For the linear case, there are generally two solutions. Each solution is consistent with a total error of zero. Either visual examination or measurement of the total squared error is required to determine which solution fits the data. A comparison between the comparable autocorrelation term solution and linear least squares shows negligible difference.

1. Introduction

Curve fitting is the process of fitting a curve, in this case a polynomial, to a set of data points. There are different types of curve fitting, but we will discuss only the most popular method, the method of least squares [1].

It is assumed that the data consists of fluctuations about an ideal curve. These fluctuations create an error between the polynomial model and the actual data. After computing a total squared error, we can apply calculus to minimize the squared error with respect to the coefficients of the polynomial. This produces a set of linear equations called the normal equations. The coefficients are derived from solving the normal equations [2].

A different approach to the problem is as follows.

We propose that the data we are considering consists of a deterministic component and a random component with zero mean. The random component does not correlate with the deterministic component and it does not correlate with itself at nonzero lag. We want to extract only the portion of the data that is correlative at nonzero lag. This requires the computation of nonzero lag autocorrelation terms and produces a system of quadratic equations [3]. After substitution based on a linear equation derived from summing the data, the system of quadratic equations consists of M equations with M unknowns, where M is the degree of the polynomial. This results in a maximum of $2M$ solutions. Each solution is consistent with a total error between the polynomial and the data that is equal to zero. Either visual examination or measurement of the total squared error is required to determine which solution fits the data.

In this paper, we derive the autocorrelation term method and compare it to linear least squares.

2. The Autocorrelation Term Method

Consider a set of data (x_i, y_i) $(i = 1, 2, 3, \ldots, N)$. Characterize the y data as the sum of a polynomial function $y(x_i)$ and a discrete, zero-mean random variable [4] Z_i :

$$\begin{aligned} Z_i &: \Omega \rightarrow \mathbb{R}, \\ y_i &= y(x_i) + Z_i. \end{aligned} \tag{1}$$

Consequently, we can write

$$\begin{aligned} y_{i+m} y_i &= \left(y(x_{i+m}) + Z_{i+m}\right)\left(y(x_i) + Z_i\right), \\ &= y(x_{i+m})\, y(x_i) + Z_{i+m} y(x_i) + Z_i y(x_{i+m}) + Z_{i+m} Z_i, \\ & \qquad m \in \{1, 2, \ldots\}. \end{aligned} \tag{2}$$

Let $N \gg m$.

Summation yields

$$\sum_{i=1}^{N-m} y_{i+m} y_i = \sum_{i=1}^{N-m} y\left(x_{i+m}\right) y\left(x_i\right) \tag{3}$$

since the cross-correlation [5] and autocorrelation terms with Z_i must be zero. The polynomial function of degree M can be written as

$$\begin{aligned} y\left(x_i\right) &= a_0 + a_1 x_i + a_2 x_i^2 + \cdots + a_M x_i^M \\ &= \sum_{j=0}^{M} a_j x_i^j. \end{aligned} \tag{4}$$

Therefore, (3) becomes the system of equations:

$$\sum_{i=1}^{N-m} y_{i+m} y_i = \sum_{i=1}^{N-m} \left(\sum_{j=0}^{M} a_j x_{i+m}^j \right) \left(\sum_{j=0}^{M} a_j x_i^j \right), \quad m = 1, 2, \ldots, M. \tag{5}$$

Using (1), we can write the average

$$\begin{aligned} \left\langle y_i \right\rangle &= \left\langle y\left(x_i\right) + Z_i \right\rangle \\ &= \left\langle y\left(x_i\right) \right\rangle + \left\langle Z_i \right\rangle. \end{aligned} \tag{6}$$

We know that $\langle Z_i \rangle = 0$, so

$$\left\langle y_i \right\rangle = \left\langle y\left(x_i\right) \right\rangle \tag{7}$$

which implies

$$\sum y_i = \sum y\left(x_i\right) \tag{8}$$

or

$$\sum_{i=1}^{N} y_i = \sum_{i=1}^{N} \sum_{j=0}^{M} a_j x_i^j. \tag{9}$$

There are a maximum of $2M$ solutions to the system of equations created by (5) and (9) for the coefficients $\{a_0, a_1, \ldots, a_M\}$. Observe that the total error, E, is presumed to be zero:

$$E = \sum_{i=1}^{N} \left(\sum_{j=0}^{M} a_j x_i^j - y_i \right) = 0. \tag{10}$$

3. The Least-Squares Straight Line

Before examining the autocorrelation method, it is worth becoming familiar with the special case of fitting a straight line in the least-squares sense. We fit

$$y\left(x\right) = a_0 + a_1 x \tag{11}$$

to a set of data (x_i, y_i) $(i = 1, 2, 3, \ldots, N)$. We have two parameters a_0 and a_1 and want to minimize the function of two variables (a_0 and a_1):

$$F\left(a_0, a_1\right) = \sum_{i=1}^{N} \left[y\left(x_i\right) - y_i \right]^2. \tag{12}$$

This becomes

$$F\left(a_0, a_1\right) = \sum_{i=1}^{N} \left[a_0 + a_1 x_i - y_i \right]^2. \tag{13}$$

Apply the calculus method

$$\begin{aligned} \frac{\partial F\left(a_0, a_1\right)}{\partial a_0} &= 2 \sum \left(a_0 + a_1 x_i - y_i \right) = 0, \\ \frac{\partial F\left(a_0, a_1\right)}{\partial a_1} &= 2 \sum \left(a_0 + a_1 x_i - y_i \right) x_i = 0. \end{aligned} \tag{14}$$

Dropping factor 2 and rewriting, we get

$$\begin{aligned} a_0 N + a_1 \sum x_i &= \sum y_i, \\ a_0 \sum x_i + a_1 \sum x_i^2 &= \sum x_i y_i. \end{aligned} \tag{15}$$

Solving these equations for a_0 and a_1 yields

$$\begin{aligned} a_0 &= \frac{1}{N} \sum y_i - \frac{1}{N} \sum x_i \left(\frac{\sum x_i y_i - (1/N) \sum y_i \sum x_i}{\sum x_i^2 - \left(\left(\sum x_i \right)^2 / N \right)} \right), \\ a_1 &= \frac{\sum x_i y_i - (1/N) \sum y_i \sum x_i}{\sum x_i^2 - \left(\left(\sum x_i \right)^2 / N \right)}. \end{aligned} \tag{16}$$

4. The Autocorrelation Term Method for $M = 1$

For $M = 1$, system (5) becomes

$$\begin{aligned} \sum_{i=1}^{N-1} y_{i+1} y_i &= \sum_{i=1}^{N-1} \left(a_0 + a_1 x_{i+1} \right) \left(a_0 + a_1 x_i \right) \\ &= (N-1) a_0^2 + \left(\sum_{i=1}^{N-1} \left(x_{i+1} + x_i \right) \right) a_0 a_1 \\ &\quad + \left(\sum_{i=1}^{N-1} x_{i+1} x_i \right) a_1^2. \end{aligned} \tag{17}$$

Suppressing the summation limits, we can write

$$\begin{aligned} \sum y_{i+1} y_i &= (N-1) a_0^2 + \left(\sum \left(x_{i+1} + x_i \right) \right) a_0 a_1 \\ &\quad + \left(\sum x_{i+1} x_i \right) a_1^2. \end{aligned} \tag{18}$$

Equation (9) becomes

$$\sum y_i = N a_0 + a_1 \sum x_i. \tag{19}$$

Solving (19) for a_0 yields

$$a_0 = \frac{1}{N} \left(\sum y_i - a_1 \sum x_i \right). \tag{20}$$

Squaring yields

$$a_0^2 = \frac{1}{N^2}\left(\left(\sum y_i\right)^2 - 2a_1 \sum y_i \sum x_i + a_1^2\left(\sum x_i\right)^2\right). \quad (21)$$

Now substitute (20) and (21) into (18) and organize the terms:

$$\left[\frac{(N-1)}{N^2}\left(\sum x_i\right)^2 + \left(\sum x_{i+1}x_i\right) - \frac{1}{N}\sum (x_{i+1}+x_i)\left(\sum x_i\right)\right]a_1^2 + \left[\frac{1}{N}\sum (x_{i+1}+x_i)\left(\sum y_i\right) - \frac{2(N-1)}{N^2}\sum y_i \sum x_i\right] \times a_1 + \frac{(N-1)}{N^2}\left(\sum y_i\right)^2 - \sum y_{i+1}y_i = 0. \quad (22)$$

This is in the form of a quadratic equation that we can write as

$$Aa_1^2 + Ba_1 + C = 0, \quad (23)$$

where, with reference to (22), A is the precursor of a_1^2, B is the precursor of a_1, and C is the residual expression.

The solutions are the roots

$$a_1 = \frac{-B \pm \sqrt{B^2 - 4AC}}{2A}. \quad (24)$$

a_0 is found using (20).

5. The Autocorrelation Term Method for Higher Values of M

The autocorrelation term method for curve fitting with an Mth degree polynomial requires the solution of one linear equation and M quadratic equations with $M + 1$ variables which reduces to M quadratic equations with M variables. This can produce $2M$ solutions [6]. For example, consider the $M = 2$ case. System (5) can be written as

$$\sum y_{i+1}y_i = \sum \left(a_0 + a_1x_{i+1} + a_2x_{i+1}^2\right)\left(a_0 + a_1x_i + a_2x_i^2\right),$$
$$\sum y_{i+2}y_i = \sum \left(a_0 + a_1x_{i+2} + a_2x_{i+2}^2\right)\left(a_0 + a_1x_i + a_2x_i^2\right). \quad (25)$$

Equation (9) can be written as

$$\sum y_i = \sum \left(a_0 + a_1x_i + a_2x_i^2\right). \quad (26)$$

From (26) we can write

$$a_0 = \frac{1}{N}\sum y_i - \frac{a_1}{N}\sum x_i - \frac{a_2}{N}\sum x_i^2. \quad (27)$$

Substitute (27) into (25):

$$\sum y_{i+1}y_i = \sum \left(\frac{1}{N}\sum y_i + \left(x_{i+1} - \frac{1}{N}\sum x_i\right)a_1 + \left(x_{i+1}^2 - \frac{1}{N}\sum x_i^2\right)a_2\right) * \left(\frac{1}{N}\sum y_i + \left(x_i - \frac{1}{N}\sum x_i\right)a_1 + \left(x_i^2 - \frac{1}{N}\sum x_i^2\right)a_2\right),$$
$$\sum y_{i+2}y_i = \sum \left(\frac{1}{N}\sum y_i + \left(x_{i+2} - \frac{1}{N}\sum x_i\right)a_1 + \left(x_{i+2}^2 - \frac{1}{N}\sum x_i^2\right)a_2\right) * \left(\frac{1}{N}\sum y_i + \left(x_i - \frac{1}{N}\sum x_i\right)a_1 + \left(x_i^2 - \frac{1}{N}\sum x_i^2\right)a_2\right). \quad (28)$$

Equations (28) are two quadratic equations in a_1 and a_2. This can produce four solutions. In general, these types of systems can be solved on a Grobner basis using Buchberger's algorithm [7].

6. A Comparison of the Autocorrelation Term Method to Linear Least Squares

We have constructed synthetic data consisting of the sum of a line and a random variable. The line has an intercept of fifteen and a slope of five. The random variable is Gaussian with a zero mean. The amplitude of the random variable is twenty. There are a total of one hundred equally spaced data points. Figure 1 shows the data, the least squares fit, and the autocorrelation term fit for $a_1 = (-B + \sqrt{B^2 - 4AC})/2A$. Figure 2 shows the data, the least squares fit, and the autocorrelation term fit for $a_1 = (-B - \sqrt{B^2 - 4AC})/2A$. In this example, we find that the quadratic roots are negatives of one another. This occurs because $x_i = i - 1$ which implies that $B = 0$. This follows from (22) which for $B = 0$ implies

$$\sum (x_{i+1} + x_i) = \frac{2(N-1)}{N}\sum x_i. \quad (29)$$

This becomes

$$\sum (2i - 1) = \frac{2(N-1)}{N}\sum (i - 1) \quad (30)$$

which reduces to

$$\sum_{i=1}^{N-1} (2i - 1) = (N - 1)^2. \quad (31)$$

7. Conclusions

We have derived a method for computing polynomial curve fits to data based on terms from the autocorrelation function. The method produces M quadratic equations with M

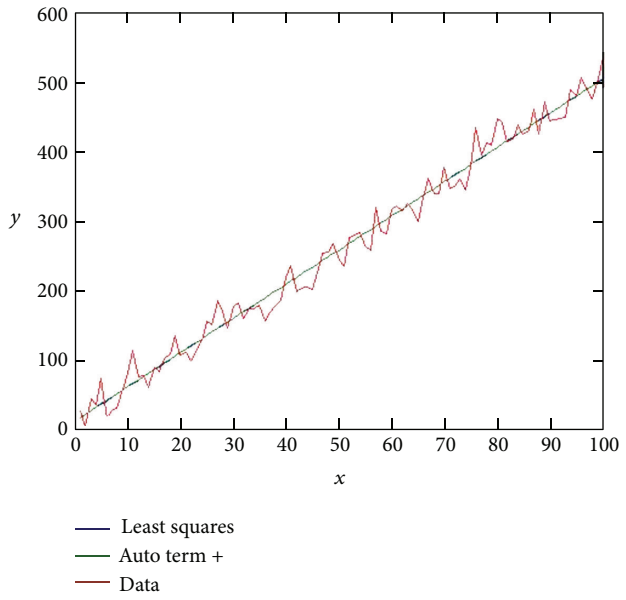

Figure 1: An overlay of the data, the least-squares fit, and the autocorrelation term fit for $a_1 = (-B + \sqrt{B^2 - 4AC})/2A$.

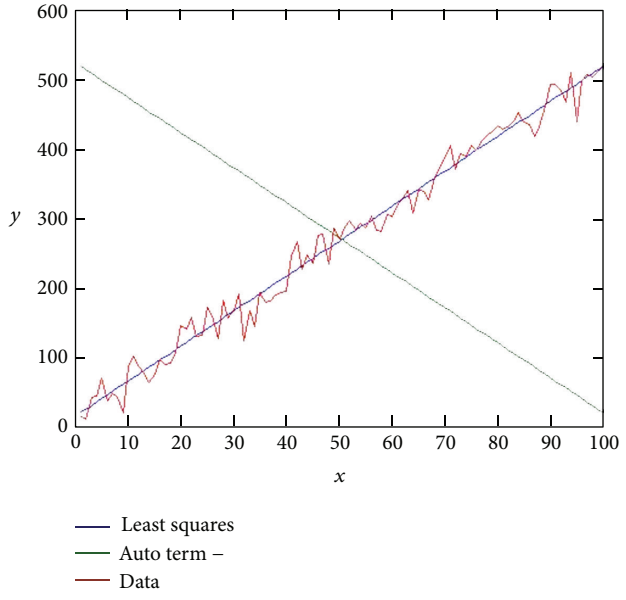

Figure 2: An overlay of the data, the least-squares fit, and the autocorrelation term fit for $a_1 = (-B - \sqrt{B^2 - 4AC})/2A$.

variables for a polynomial of degree M. The solutions of this system maximally produce $2M$ curves. Each solution is consistent with a total error between the polynomial and the data that is equal to zero. The proper curve can be selected with either visual examination or measurement of the total squared error. We tested this method with a linear curve fit and compared it to linear least squares. There is negligible difference between the comparable solutions.

Acknowledgment

Discussions with Gwen Houston and Dominique Luecken-hoff are greatly appreciated.

References

[1] S. D. Christian, E. H. Lane, and F. Garland, "Linear least-squares analysis: a caveat and a solution," *Journal of Chemical Education*, vol. 51, no. 7, p. 475, 1974.

[2] R. J. O'Dowd, "The Wiener-Levinson algorithm and ill-conditioned normal equations," *Geophysical Journal International*, vol. 106, no. 2, pp. 399–406, 1991.

[3] Y. Kinoshita, T. Asakura, and M. Suzuki, "Autocorrelation of Gaussian-Beam fluctuation caused by a random medium," *Journal of the Optical Society of America*, vol. 58, no. 8, pp. 1040–1047, 1968.

[4] K. Wong and S. Redman, "The recovery of a random variable from a noisy record with application to the study of fluctuations in synaptic potentials," *Journal of Neuroscience Methods*, vol. 2, no. 4, pp. 389–409, 1980.

[5] M. E. Keating, F. Bonnier, and H. J. Byrne, "Spectral cross-correlation as a supervised approach for the analysis of complex Raman datasets: the case of nanoparticles in biological cells," *Analyst*, vol. 137, pp. 5792–5802, 2012.

[6] A. Denis, "A discussion of the cases when two quadratic equations involving two variables can be solved by the method of quadratics," *The American Mathematical Monthly*, vol. 10, no. 8-9, pp. 192–199, 1903.

[7] B. Buchberger, "An algorithm for finding the basis elements of the residue class ring of a zero dimensional polynomial ideal," *Journal of Symbolic Computation*, vol. 41, no. 3-4, pp. 475–511, 2006.