Available online at www.sciencedirect.com

ScienceDirect

AASRI Procedia 6 (2014) 41 – 48

AASRI Procedia

www.elsevier.com/locate/procedia

2013 2nd AASRI Conference on Computational Intelligence and Bioinformatics

A New Fitting Scattered Data Method Based on the Criterion of Geometric Distance

Guowei Yang*, Jia Xu

College of Information Engineering, Nanchang Hangkong University, Nanchang 330063, China

Abstract

The traditional data fitting method based on least square method is not good for vector data fitting whose independent variable is random. So this paper proposes a new criterion of data fitting which is the least quadratic sum of geometrical distance, and brings forward the new fitting scattered data method based on the new criterion. At the same time the paper puts forward the optimization algorithm for the solution of the data fitting parameter. Simulation experiments show that the fitting precision of the new method is higher than the one of least square method for data fitting of vector, whose independent variable is random.

Peer-review under responsibility of Scientific Committee of American Applied Science Research Institute

Keywords: Data fitting, criterion of data fitting least square method, geometrical distance;

1. Introduction

In the experimental science, social sciences, behavioral science and the actual engineering fields, such as Computer-Aided Design, Manufacturing, virtual reality, medical imaging, the experiment, the survey or the test can frequently bring large numbers of data. In order to explain these data or according to these data to make the forecast, the judgment, provides the important basis to the policy-maker, we needs to carry on the

* Corresponding author. Tel.: +86-791-83953432; fax: +86-791-83953432.
E-mail address: ygw_ustb@163.com.

Peer-review under responsibility of Scientific Committee of American Applied Science Research Institute
doi:10.1016/j.aasri.2014.05.007

linear or nonlinear fitting modelling frequently to the survey data, seeks to the function (model) which can approximately reflect the change rules of the data. There are a long history and enrich results about data fitting and its application[1-3]. In linear data fitting, we frequently use the least square method in order to get data fitting parameter. The reason is that if the fitting model conforms to the Gauss -Markov hypothesis condition, least square method can obtain the fitting parameter with good statistical nature, like agonic, uniformity, smallest variance and so on. However, the actual test data infinitely varied, moreover the purpose of carrying on the data fitting to the test data are also different, and the precision request are also different, thus the data fitting result which use the least squares method to carry on the data fitting can not achieve the requested purpose. For example, there are some unusual data due to occasionally abnormal error or data probability distribution deviates normal distribution, if we use the regression analysis result of least square method, we will lose its good statistical property, one solution of this situation is to use the criterion function with steady performance[4-5]. Moreover, during linear fitting, compare to the Scatter-point actual distributed tendency, the straight line determined by least squares method in many fields all has a bit small slope phenomenon, especially the fluctuation of the sample Scatter-point is a little bigger, this phenomenon is more obvious. The reason is that, when we use ordinary least square method on parameter estimation, we have the following criterion: choose an equation, which the range difference sum of squares $\sum E_i^2 = \sum (y_i - \hat{y}_i)^2$ of the dependent variable between the observed value and the estimated value to be smallest, from all the possible linear equation. Its geometry significance is: To find a straight line, which longitudinal distance sum of squares between the scatter-point of the observed value and this straight line to be smallest, from all possible straight lines. Just because the goal is to make the longitudinal distance sum of squares between the scatter-point and this straight line to be smallest, rather than the geometrical distance (vertical distance) sum of squares between the scatter-point and this straight line to be smallest, which result in the estimated straight line has slope small tendency[6].

Lifts the background of least square method theory, we can discover the concealed default premise supposition when we use least square method on data fitting: because the dependent variable of the sample is influenced by the stochastic disturbing term, it is stochastic undulation, while the sample independent variable doesn't stochastic, that is scatter-point deviates linear equation, completely because scatter-point undulates in the fluctuation direction of dependent variable, rather than the combined effect in the fluctuation direction of all the dependent variable. Therefore, use least square method on data fitting directly without considering the premise supposition condition, and no wonder we will get the fitting equation with deviation. The different results are produced by different methods, however different methods are established on the foundation of different premise supposition. Therefore each method itself can not be said who is right in this question, just whose premise supposition more reasonable.

In practical life, there are generally two kinds of relationship of models, that is:

(1) The definite relation model, that is, in scatter-point (x, y), x and y has definite relation (that is: x is independent variable y is dependent variable or y is independent variable x is dependent variable)

(2) The independent variable is the random variable model, that is, in scatter-point (x, y), the x and y primary and secondary relations are fuzzy, not clear, the independent variable and the dependent variable do not always differentiate very clear. For example, the relations of human's height and weight, the degree of two variables are completely coordinated, not only have $y = f(x) + \mu$, but also have $x = g(y) + v$.

Obviously, least square method is not suitable for all linear fitting model parameter estimation. Therefore, to second kind model, we generally do not use least squares method criterion fitting. In order to solve the independent variable for the random variable model fitting, in recent years, many scientific and technology

workers carry on much research work on the least square fitting method and obtain fruitful efforts. For example, Principal Component Regression (PCR), Partial Least Squares (PLS), all kinds of Non-Linear Principal Component Regression (NLPCR), all kinds of Non-Linear Partial Least Squares (NLPLS), Neural Network method and the syntheses of these methods ect., which used in actual quite many in present project[4-9]. In this article, we do not use the method which is mentioned by the above literature when carry on scatter-point fitting, instead of using the minimum of geometrical distance sum of squares as new fitting standard, we propose a new data fitting method based on the new standard. In the method, including model transform, turns fitting parameter solution into "constrained optimization problem", and provides the fitting parameter solution algorithm. The simulation experiment indicated that, the fitting parameter solution algorithm is feasible and effective; in the data fitting which independent variable is the random variable vector data, we can get a higher fitting precision by using the new data fitting method instead of least square method.

2. A New Data Fitting Criterion and New Data Fitting Method

Question: Given the linear relation variable $X_1, X_2, \cdots, X_n$ have experiment or observation data (has error) in n dimension space: $(x_{i1}, x_{i2}, \cdots, x_{in})$, $i = 1, \cdots, N$ $(N \geq n)$, solve the actual linear relation formula of $X_1, X_2, \cdots, X_n$.

About this question, the traditional classical procedure is suppose X_1 is dependent variable, $X_2, \cdots, X_n$ is independent variable, establish $X_1, X_2, \cdots, X_n$ linear relation formula $X_1 = \beta_2 X_2 + \cdots + \beta_n X_n + \beta_n$ by least square method. The basic presupposition that establish $X_1, X_2, \cdots, X_n$ linear relation formula $X_1 = \beta_2 X_2 + \cdots + \beta_n X_n + \beta_n$ by least square method is: X_1 is dependent variable, $X_2, \cdots, X_n$ is independent variable, and $X_2, \cdots, X_n$ is fixed variable (means the experiment or observation data does not have error). It is proved that, under such presupposition, data fitting by least squares method is quite effective. However this presupposition is dissatisfy in many situations, sometimes independent variable $X_2, \cdots, X_n$ is random variable, sometimes $X_1, X_2, \cdots, X_n$ can not be distinguished which is dependent variable, which are independent variables.

While independent variable $X_2, \cdots, X_n$ is random variable, or $X_1, X_2, \cdots, X_n$ can not be distinguished which is dependent variable and which are independent variables, we naturally associate to their implicit function relations, that is, the variable $X_1, X_2, \cdots, X_n$ linear relations in n dimension space can be shown as follows,

$$a_1 X_1 + a_2 X_2 \cdots + a_n X_n + b = 0 \tag{1}$$

where $a_1, a_2 \cdots a_n$ are not all 0. Data vector $(x_{i1}, x_{i2}, \cdots, x_{in})$ $(i = 1, 2, \cdots, N)$ on the above or periphery of the hyperplane $a_1 X_1 + a_2 X_2 \cdots + a_n X_n + b = 0$.

Therefore, to solve the question which proposed above actually is to choose a “proper” hyperplane $a_1 X_1 + a_2 X_2 \cdots + a_n X_n + b = 0$ in n dimension space which can do vector data $(x_{i1}, x_{i2}, \cdots, x_{in})$ $(i = 1, 2, \cdots, N)$ fitting, where $a_1, a_2 \cdots a_n, b$ are some uncertain parameters. The visually and easy to excogitate hyperplane choice criterion is: the least quadratic sum of distance (or distance) from $(x_{i1}, x_{i2}, \cdots, x_{in})$ $(i = 1, 2, \cdots, N)$ to the hyperplane $a_1 X_1 + a_2 X_2 \cdots + a_n X_n + b = 0$.

Definition 1: Called the data point to the hyperplane distance

$$e_i = \frac{\left|a_1 x_{i1} + a_2 x_{i2} + \cdots + a_n x_{in} + b\right|}{\sqrt{a_1^2 + a_2^2 + \cdots + a_n^2}} \quad (i = 1, 2, \cdots, N) \tag{2}$$

is $(x_{1j}, x_{2j}, \cdots, x_{nj})$ to $a_1 X_1 + a_2 X_2 \cdots + a_n X_n + b = 0$ geometry distance.

Definition 2: Define the evaluation function of data fitting is

$$J = \sum_{i=1}^{N} |e_i|^2 = \sum_{i=1}^{N} \left| \frac{a_1 x_{i1} + a_2 x_{i2} + \cdots + a_n x_{in} + b}{\sqrt{a_1^2 + a_2^2 + \cdots + a_n^2}} \right|^2 \quad (i = 1, 2, \cdots, N) \tag{3}$$

Call $\min J$ is geometrical distance criterion (new data fitting criterion), that is, the least quadratic sum of geometrical distance criterion. As follows, we will give the new data fitting method base on the least quadratic sum of geometrical distance criterion.

The new data fitting method base on the least quadratic sum of geometrical distance criterion mainly consists of the following two parts:

(a) Data fitting optimization model

(b) Data fitting optimization model solution

Data Fitting Optimization Model

Divide $\sqrt{a_1^2 + a_2^2 + \cdots + a_n^2}$ on each side of (1) equation synchronously, and suppose

$$\overline{a}_j = \frac{a_j}{\sqrt{a_1^2 + a_2^2 + \cdots + a_n^2}} \left(j = 1, 2, \cdots, n\right), \overline{b} = \frac{b}{\sqrt{a_1^2 + a_2^2 + \cdots + a_n^2}},$$

then (1) equation can be changed into:

$$\overline{a}_1 X_1 + \overline{a}_2 X_2 + \cdots + \overline{a}_n X_n + \overline{b} = 0 \tag{4}$$

and $\overline{a}_1^2 + \overline{a}_2^2 + \cdots + \overline{a}_n^2 = 1$. In a similar way, (3) equation can be changed into:

$$J = \sum_{i=1}^{N} |e_i|^2 = \sum_{i=1}^{N} \left| \frac{a_1 x_{i1} + a_2 x_{i2} + \cdots + a_n x_{in} + b}{\sqrt{a_1^2 + a_2^2 + \cdots + a_n^2}} \right|^2$$

$$= \sum_{i=1}^{N} \left| \overline{a}_1 x_{i1} + \overline{a}_2 x_{i2} + \cdots + \overline{a}_n x_{in} + \overline{b} \right|^2$$

Hence, data fitting model can be changed into the following optimization model:

$$\begin{aligned} &\min \sum_{i=1}^{N}\left(\overline{a}_1 x_{i1}+\overline{a}_2 x_{i2}+\cdots+\overline{a}_n x_{in}+\overline{b}\right)^2 \\ &s.t.\quad \overline{a}_1^2+\overline{a}_2^2+\cdots+\overline{a}_n^2-1=0 \qquad \left(i=1,2,\cdots,N\right) \end{aligned} \tag{5}$$

Where $\overline{a}_1,\overline{a}_2,\cdots,\overline{a}_n,\overline{b}$ are some uncertain parameters. $\left(x_{i1},x_{i2}\cdots x_{in}\right)$, $i=1,\cdots,N$ are N known data scatter-points.

Data Fitting Optimization Model Solution

By extremum solution theory, formula (5) can be transformed into Lagrange function:

$$L(\overline{a}_1,\overline{a}_2,\cdots,\overline{a}_n,\overline{b},\lambda)=\sum_{i=1}^{N}\left(\overline{a}_1 x_{i1}+\overline{a}_2 x_{i2}+\cdots+\overline{a}_n x_{in}+\overline{b}\right)^2+\lambda\left(\overline{a}_1^2+\overline{a}_2^2+\cdots+\overline{a}_n^2-1\right) \tag{6}$$

Consider the steady point $(\overline{a}_1,\overline{a}_2,\cdots,\overline{a}_n,\overline{b},\lambda)$. Take partial derivative to the above equation, we can obtain the following equation group.

$$F(\overline{a}_1,\overline{a}_2,\cdots,\overline{a}_n,\overline{b},\lambda)=0 \tag{7}$$

Simply written as: $F=0$, where $F=\left(F_1,F_2,\cdots,F_{n+1},F_{n+2}\right)^T$, that is

$$\begin{cases} \lambda\overline{a}_1+\sum_{i=1}^{N}x_{i1}\left(\overline{a}_1 x_{i1}+\overline{a}_2 x_{i2}+\cdots+\overline{a}_n x_{in}+\overline{b}\right)=0 \\ \qquad\qquad\cdots\cdots\cdots\cdots \\ \lambda\overline{a}_n+\sum_{i=1}^{N}x_{in}\left(\overline{a}_1 x_{i1}+\overline{a}_2 x_{i2}+\cdots+\overline{a}_n x_{in}+\overline{b}\right)=0 \\ \sum_{i=1}^{N}\left(\overline{a}_1 x_{i1}+\overline{a}_2 x_{i2}+\cdots+\overline{a}_n x_{in}+\overline{b}\right)=0 \\ \overline{a}_1^2+\overline{a}_2^2+\cdots+\overline{a}_n^2-1=0 \end{cases} \tag{8}$$

Where $i=1,\cdots,N$ are N known data scatter-points.

$$F_m=\lambda\overline{a}_m+\sum_{i=1}^{N}x_{im}\left(\overline{a}_1 x_{i1}+\overline{a}_2 x_{i2}+\cdots+\overline{a}_n x_{in}+\overline{b}\right)\left(m=1,2,\cdots,n\right)$$

$$F_{n+1}=\sum_{i=1}^{N}\left(\overline{a}_1 x_{i1}+\overline{a}_2 x_{i2}+\cdots+\overline{a}_n x_{in}+\overline{b}\right)$$

$$F_{n+2} = \overline{a}_1^2 + \overline{a}_2^2 + \cdots + \overline{a}_n^2 - 1$$

Solve the non-linear equations group of equation(7), the fitting parameter $\overline{a}_1, \overline{a}_2, \cdots, \overline{a}_n, \overline{b}$ and λ are obtained, farther solve the linear fitting equation $\overline{a}_1 X_1 + \overline{a}_2 X_2 \cdots + \overline{a}_n X_n + \overline{b} = 0$, as well as solve the equation $a_1 X_1 + a_2 X_2 \cdots + a_n X_n + b = 0$, thereby we can obtain the data fitting model.

3. Solution Algorithm of the Data Fitting Parameter

Iteration algorithm of fitting parameter will be given as follows.

For non-linear equation group $F(\overline{a}_1, \overline{a}_2, \cdots, \overline{a}_n, \overline{b}, \lambda) = 0$, $\overline{a}_1, \overline{a}_2, \cdots, \overline{a}_n, \overline{b}, \lambda$ are some uncertain parameters. Suppose $Y = (\overline{a}_1, \overline{a}_2, \cdots, \overline{a}_n, \overline{b}, \lambda)$, initial iteration point $Y^{(0)} = (\overline{a}_1^{(0)}, \overline{a}_2^{(0)}, \cdots, \overline{a}_n^{(0)}, \overline{b}^{(0)}, \lambda^{(0)})$, the iteration point after iterate k times is $Y^{(k)} = (\overline{a}_1^{(k)}, \overline{a}_2^{(k)}, \cdots, \overline{a}_n^{(k)}, \overline{b}^{(k)}, \lambda^{(k)})$, by Taylor formula, we can obtain

$$F\left(Y^{(k)}\right) \approx F(Y^{(k+1)}) + \nabla F(Y^{(k+1)})\left(Y^{(k)} - Y^{(k+1)}\right)$$

after rearrangement, we obtain the following equation:

$$\nabla F(Y^{(k+1)})\left(Y^{(k+1)} - Y^{(k)}\right) = F(Y^{(k+1)}) - F(Y^{(k)})$$

Definition 3: Call iteration algorithm

$$\begin{cases} Y^{(k+1)} = Y^{(k)} - \left(\nabla F\left(Y^{(k)}\right)\right)^{-1} F\left(Y^{(k)}\right) \\ \nabla F\left(Y^{(k+1)}\right)\left(Y^{(k+1)} - Y^{(k)}\right) = F(Y^{(k+1)}) - F(Y^{(k)}) \\ \nabla F\left(Y^{(k+1)}\right) = \nabla F\left(Y^{(k)}\right) + \Delta\left(\nabla F\left(Y^{(k)}\right)\right) \\ \qquad k = 0,1,\cdots \end{cases} \tag{9}$$

is the correction technique of solving non-linear equation group $F(\overline{a}_1, \overline{a}_2, \cdots, \overline{a}_n, \overline{b}, \lambda) = 0$ rank m.

By equation (9), we can solve $\overline{a}_1, \overline{a}_2, \cdots, \overline{a}_n, \overline{b}$, accordingly, can also determined the variable coefficient $a_1, a_2, \cdots, a_n, b$. As well as by equation (9), first solves $\overline{a}_1, \overline{a}_2, \cdots, \overline{a}_n$, $\overline{b}$, consequently, determines the variable coefficient $a_1, a_2, \cdots, a_n, b$.

4. Simulation Analysis

Suppose the point on the straight line $10x - y + 1 = 0$, the observation data (has error) as follows:

	$i = 1$	$i = 2$	$i = 3$	$i = 4$	$i = 5$	$i = 6$	$i = 7$	$i = 8$	$i = 9$
x	-2	-2	-0.5	-1	0.5	0	1.5	1	2.5
y	-19	-14	-9	-4	1	6	11	16	21

Determine the fitting straight line $ax + by + c = 0$ by the above discrete point, where a, b, c are some unknown parameter. Then solve the equation (8) with the data fitting parameter solution algorithm.

Following Table 1 and Fig. 1 respectively is under the perfect condition, under the least square criterion, the straight line equation fitting parameter and the simulation figure based on the geometrical distance criterion.

Table 1. Fitting parameter under different criterion

	perfect condition	the least square criterion	based on the geometrical distance criterion
a	10	8.4211	-0.99377232
b	-1	-1	0.11142971
c	1	1	-0.1115

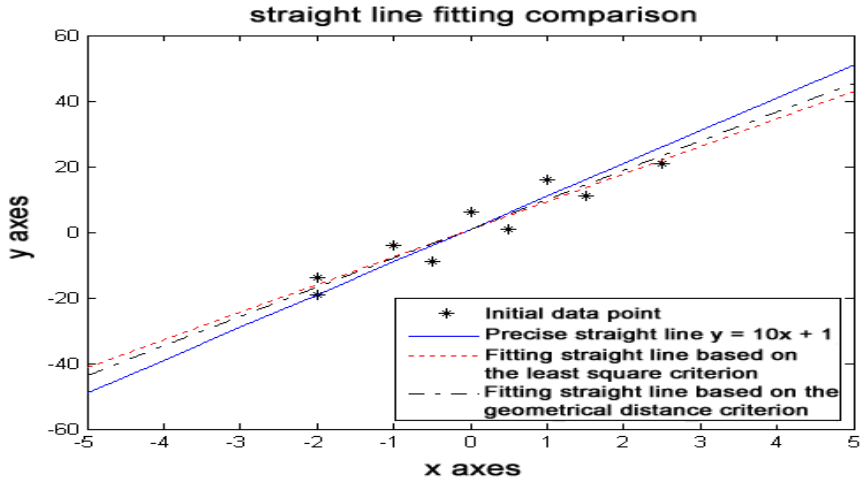

Fig. 1. Simulation result under different fitting standard

By the figure we can see, the discrete point uniform distributes on the different sides of three straight lines. It is obvious that the fitting effect by the least square criterion and based on the geometrical distance criterion is all good. However, reference to the ideal straight line we can know that the fitting straight line under new criterion obviously better than the least square fitting straight line, the fitting straight line under new criterion located between the ideal straight line and the least square straight line, the fitting precision is higher than the least square straight line.

Acknowledgements

The authors would like to thank the editors and the anonymous reviewers for their valuable comments and constructive suggestions. This research is supported by the National Natural Science Foundation of China (No. 61272077, 61202319), the Natural Science Foundation of Jiangxi Province(No. 20114BAB201034) and the Scientific and Technological Project of Jiangxi Province(No. 20133BBE50022).

References

[1] Lin Honghua. Data processing of dynamic measurement. Beijing: Beijing Institute of Technology Press;1952.
[2] Lin Hongwei. Adaptive data fitting by the progressive-iterative approximation. Computer Aided Geometric Design. 2012; 2(7): 463-473.

[3] Lapo Governi, Rocco Furferi, Matteo Palai, Yary Volpe. 3D geometry reconstruction from orthographic views: A method based on 3D image processing and data fitting. Computers in Industry, In Press, online 19 March 2013.
[4] E. Vassiliou, I.C. Demetriou. An adaptive algorithm for least squares piecewise monotonic data fitting. Computational Statistics & Data Analysis. 2005; 49(2):591-609.
[5] G. Casciola, L. Romani. A Newton-type method for constrained least-squares data-fitting with easy-to-control rational curves. Journal of Computational and Applied Mathematics. 2009; 223:672-692.
[6] Huang Minjie, Ye Hao, Wang Guizeng. The regression analysis method summary based on projection, Control theory and application. 2001; 8(18):1-6.
[7] Alexandru Mihai Bica. Fitting data using optimal Hermite type cubic interpolating splines. Applied Mathematics Letters. 2012; 25(12):2047-2051.
[8] Philipp Reinecke, Tilman Krauß, Katinka Wolter. Cluster-based fitting of phase-type distributions to empirical data. Computers & Mathematics with Applications. 2012; 64(12);3840-3851.
[9] Akemi Gálvez, Andrés Iglesias, Andreina Avila. Immunological-based Approach for Accurate Fitting of 3D Noisy Data Points with Bézier Surfaces. Procedia Computer Science. 2013; 18:50-59.