

Local Polynomial Estimation of Heteroscedasticity in a Multivariate Linear Regression Model and Its Applications in Economics

Liyun Su^{1*}, Yanyong Zhao¹, Tianshun Yan¹, Fenglan Li²

1 School of Mathematics and Statistics, Chongqing University of Technology, Chongqing, China, **2** Library, Chongqing University of Technology, Chongqing, China

Abstract

Multivariate local polynomial fitting is applied to the multivariate linear heteroscedastic regression model. Firstly, the local polynomial fitting is applied to estimate heteroscedastic function, then the coefficients of regression model are obtained by using generalized least squares method. One noteworthy feature of our approach is that we avoid the testing for heteroscedasticity by improving the traditional two-stage method. Due to non-parametric technique of local polynomial estimation, it is unnecessary to know the form of heteroscedastic function. Therefore, we can improve the estimation precision, when the heteroscedastic function is unknown. Furthermore, we verify that the regression coefficients are asymptotic normal based on numerical simulations and normal Q-Q plots of residuals. Finally, the simulation results and the local polynomial estimation of real data indicate that our approach is surely effective in finite-sample situations.

Citation: Su L, Zhao Y, Yan T, Li F (2012) Local Polynomial Estimation of Heteroscedasticity in a Multivariate Linear Regression Model and Its Applications in Economics. PLoS ONE 7(9): e43719. doi:10.1371/journal.pone.0043719

Editor: Timothy Ravasi, King Abdullah University of Science and Technology, Saudi Arabia

Received: May 2, 2012; **Accepted:** July 24, 2012; **Published:** September 17, 2012

Copyright: © 2012 Su et al. This is an open-access article distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

Funding: This work was supported by Chongqing CSTC foundations of China (CSTC2010BB2310, CSTC2011jjA40033), Chongqing CMEC foundations of China (KJ080614, KJ100810, KJ100818). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing Interests: The authors have declared that no competing interests exist.

* E-mail: cloudhopping@163.com

Introduction

The heteroscedasticity in classical linear regression model is defined by the variances of random items and which are not the same for different explanatory variables and observations [1,2]. Especially the heteroscedasticity is usually inevitable when we study the cross-sectional data [3–6]. When there is a heteroscedasticity in a linear regression model, estimations of parameters we obtained by ordinary least squares estimation (OLS) are still linear and unbiased [7–10]. However, the efficiency is bad [11,12]. This could lead to a incorrect statistical diagnosis for the parameters' significance test. Similarly, it unnecessarily enlarges the confidence interval when we estimate the parameter interval. Besides, the accuracy of predictive value may lower when we estimate with the regression model that we obtained. In order to solve the problem above, we can use generalized least squares estimation (GLS) when the covariance matrix of the random items is known. If it is unknown, we usually use two-stage least squares estimate. In other words, we first estimate variances of the residual error, and then the generalized least squares estimator is used to obtain the coefficients of the model by using the estimate of variances of the random items. However, the traditional estimation method is that we suppose the residual error variances as a certain parametric model. So estimations we obtained are always inaccurate. For a heteroscedasticity model which has only one explanatory variable, we have discussed in detail and given a rigorous proof in [13]. In this paper, we try to apply multivariate local polynomial fitting to random item variances as the first step, and then generalized least squares estimation is used to estimate the coefficients of the model. On the one hand, because of local

polynomial fitting's various nice statistical properties, the estimations obtained with this technology also have the same good statistical properties [14]. On the other hand, we exploit a heteroscedastic regression model rather than the artificial structure of heteroscedasticity. Then, we can directly get the heteroscedastic function based on the nonparametric technique, which shows the relationship between variance function of random items and explanatory variables from regression results [15]. Thus, it is unnecessary to test heteroscedasticity of the model. Particularly, the estimated value by multivariate local polynomial fitting are more accurate than that by the traditional method and univariate local polynomial fitting.

The rest of this paper is organized as follows. The multivariate local polynomial estimation adopted is explained in detail in Section Methods [16–18]: in its the first Subsection we study multivariate estimation with local polynomial fitting. The second subsection contains selections of parameters. Estimations of model coefficients and heteroscedastic function are presented in the third subsection Parametric estimations with multivariate local polynomial regression. In Sections Results and Discussion, we firstly give two models, do the simulations, then collect some real data, and use the local polynomial estimating the coefficients, respectively. Finally, Section Conclusions conclude.

Methods

Multivariate local polynomial fitting is an attractive method both from theoretical and practical point of view. It possesses a smaller mean squared error than that of classical kernel Nadaraya-Watson estimator which leads to an undesirable form of the bias

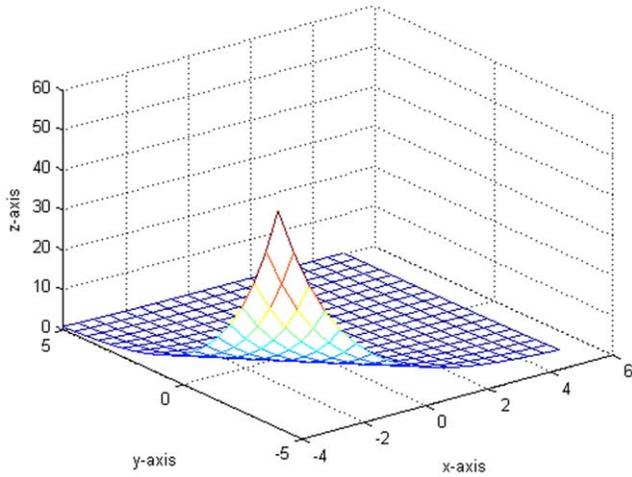


Figure 1. The fitted plot after 10000 replicates.
doi:10.1371/journal.pone.0043719.g001

and the Gasser-Muller estimator which has to pay a price in variance when dealing with a random design model. It also possesses other advantages [19,20]. In this Section, we briefly outline the idea of the extension of multivariate local polynomial fitting to multivariate linear regression.

Multivariate Estimator with Local Polynomial Fitting

We treat the m-dimension estimation problem where the measured data Y at the position $X = [x_1, x_2, \dots, x_m]^T$ is given by

$$Y = f(X) + \sigma(X)\varepsilon, \tag{1}$$

where $f(\cdot)$ is the regression function to be estimate, $E(\varepsilon) = 0$, $Var(\varepsilon) = I_n$ (I_n denoting the identity matrix of order n), and X and ε are independent. We always denote the conditional variance of Y given $X = X_T$ by $\sigma^2(X_T)$ and the marginal density of X , that is, the design density, by $f(\cdot)$.

While the specific form of $f(\cdot)$ may remain unspecified, if we assume that the $(p + 1)$ th derivative of $f(X)$ at the point X_T exists, then in order to estimate the value at this point, we can rely on a generic local expansion of the function about this point.

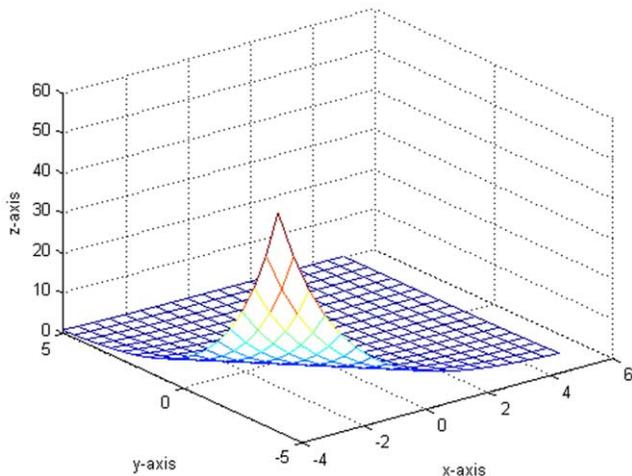


Figure 2. The original plot of the variance function.
doi:10.1371/journal.pone.0043719.g002

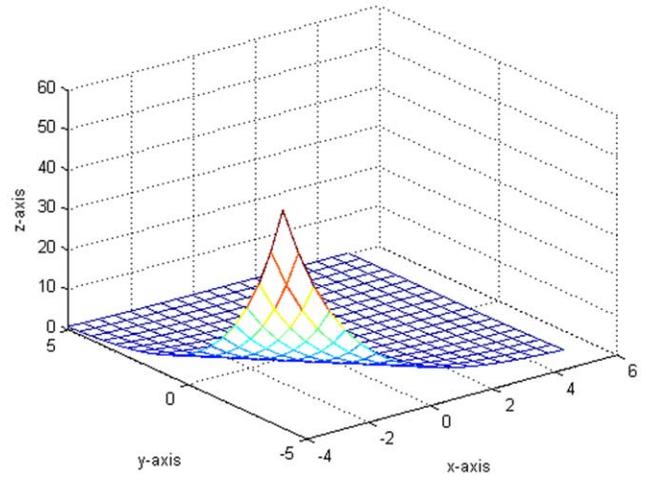


Figure 3. The plot of drawing fitted figure and original figure together.
doi:10.1371/journal.pone.0043719.g003

Specifically, for X in a neighborhood of X_T , a p -term Taylor expansion gives,

$$f(X) \approx \sum_{0 \leq |k| \leq p} \frac{1}{k!} D^{(k)} M(Z)|_{Z=X_T} (X - X_T)^k \tag{2}$$

where

$$k = (k_1, k_2, \dots, k_m),$$

$$k! = k_1! k_2! \dots k_m!,$$

Table 1. True value and fitted value at points.

Points	True value	Fitted value
(-0.5, -0.5)	1.64870	1.6345
(-0.5,0,0)	1.28400	1.27360
(-0.5,0,5)	1.00000	0.99139
(-0.5,1,0)	0.77880	0.77245
(0,0, -0.5)	1.28400	1.27360
(0,0,0,0)	1.00000	0.99230
(0,0,0,5)	0.77880	0.77245
(0,0,1,0)	0.60653	0.60186
(0.5, -0.5)	1.00000	0.99139
(0.5,0,0)	0.77880	0.77245
(0.5,0,5)	0.60653	0.60131
(0.5,1,0)	0.47237	0.46852
(1,0, -0.5)	0.77880	0.77245
(1,0,0,0)	0.60653	0.60186
(1,0,0,5)	0.47237	0.46852
(1,0,1,0)	0.36788	0.36505

doi:10.1371/journal.pone.0043719.t001

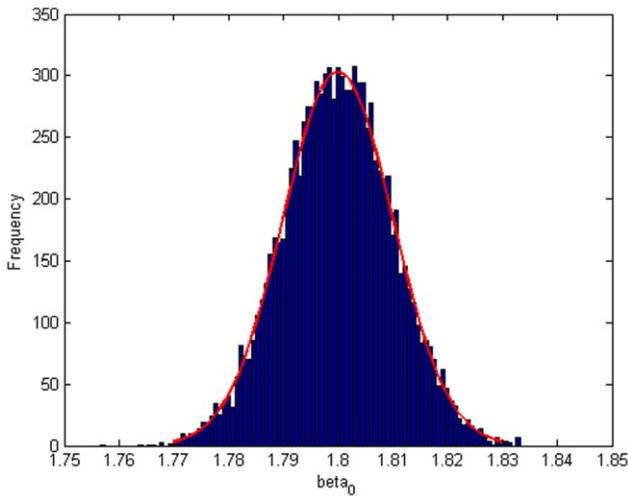


Figure 4. Histogram and asymptotic distribution for $\hat{\beta}_0$ of GLS when the variance function of the error term is $\sigma^2(x_1, x_2) = \exp(-(x_1 + x_2))$.
doi:10.1371/journal.pone.0043719.g004

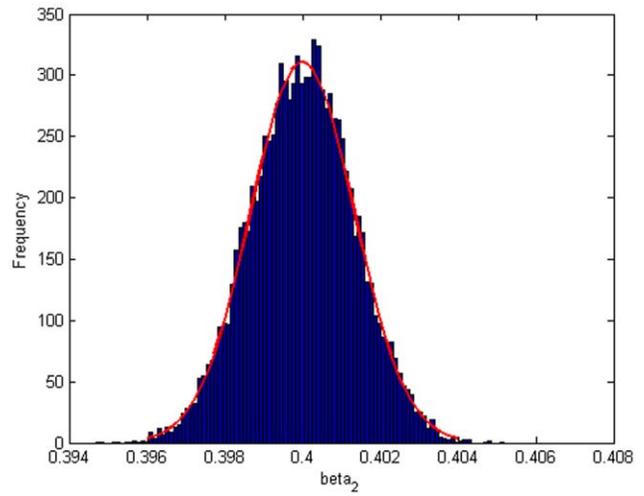


Figure 6. Histogram and asymptotic distribution for $\hat{\beta}_2$ of GLS when the variance function of the error term is $\sigma^2(x_1, x_2) = \exp(-(x_1 + x_2))$.
doi:10.1371/journal.pone.0043719.g006

$$|k| = \sum_{i=1}^m k_i,$$

$$\sum_{0 \leq |k| \leq p} = \sum_{j=0}^p \left(\sum_{k_1=0}^j \cdots \sum_{k_m=0}^j \right)_{j=k_1+k_2+\dots+k_m},$$

$$X^k = x_1^{k_1} x_2^{k_2} \cdots x_m^{k_m},$$

$$D^{(k)} M(Z) = \frac{\partial^k M(Z)}{\partial Z_1^{k_1} \cdots \partial Z_m^{k_m}}.$$

Given the series $\{X_i, Y_i\}_{i=1}^n$, this polynomial is fitted locally by a weighted least squares regression problem: minimize.

$$\sum_{i=1}^n \left\{ Y_i - \sum_{0 \leq |k| \leq p} \theta_k(X_T)(X_i - X_T) \right\}^2 K_M(X_i - X_T) \quad (3)$$

where M is a bandwidth matrix controlling the size of the local neighborhood and $K(\cdot)$ is the kernel function which penalizes both geometric and radiometric distances, where $K_M(\cdot)$ is defined as [21]

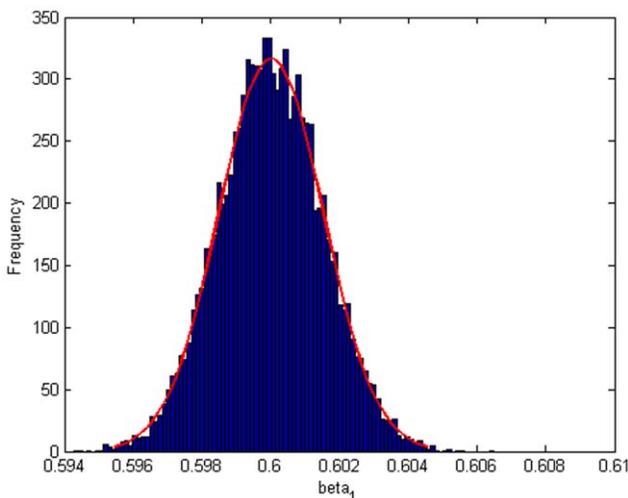


Figure 5. Histogram and asymptotic distribution for $\hat{\beta}_1$ of GLS when the variance function of the error term is $\sigma^2(x_1, x_2) = \exp(-(x_1 + x_2))$.
doi:10.1371/journal.pone.0043719.g005

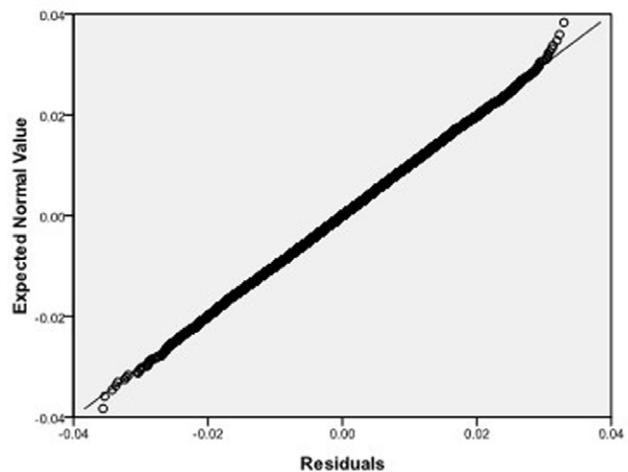


Figure 7. Normal Q-Q plot of residuals of $\hat{\beta}_0$ of GLS when the variance function of the error term is $\sigma^2(x_1, x_2) = \exp(-(x_1 + x_2))$.
doi:10.1371/journal.pone.0043719.g007

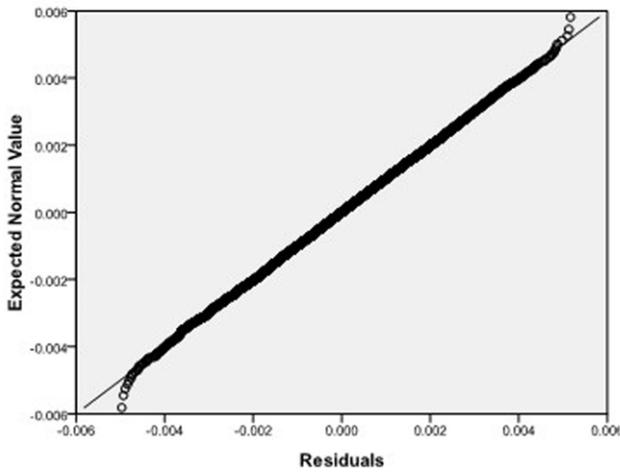


Figure 8. Normal Q-Q plot of residuals of $\hat{\beta}_1$ of GLS when the variance function of the error term is $\sigma^2(x_1, x_2) = \exp(-(x_1 + x_2))$.
doi:10.1371/journal.pone.0043719.g008

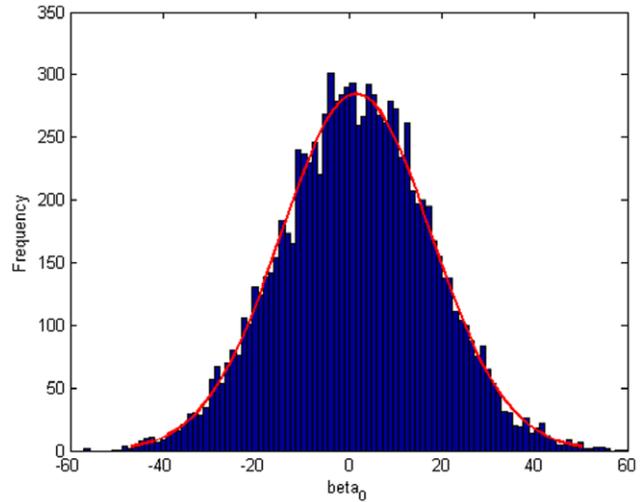


Figure 10. Histogram and asymptotic distribution for $\hat{\beta}_0$ of OLS when the variance function of the error term is $\sigma^2(x_1, x_2) = \exp(-(x_1 + x_2))$.
doi:10.1371/journal.pone.0043719.g010

$$K_M(X_i - X_T) = \frac{1}{|M|} K_M[M^{-1}(X_i - X_T)].$$

Denote by $\hat{\theta}_k(X)$ the solution to the least squares problem (3). It is clear from the Taylor expansion in (2) that $\hat{D}^{(k)}M(X_T) = k! \hat{\theta}_k(X_T)$ is an estimator for $D^{(k)}M(X_T)$.

It is more convenient to work with matrix notation. For the weighted least squared problem, a matrix form can be depicted by

$$W^{1/2} \cdot Y = W^{1/2} \cdot X_x \cdot \theta(X_T) + \sigma(X)\varepsilon, \tag{4}$$

where

$$Y = \begin{pmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{pmatrix}, \quad \theta(X_T) = \begin{pmatrix} \theta_0(X_T) \\ \theta_1(X_T) \\ \vdots \\ \theta_p(X_T) \end{pmatrix}, \tag{5}$$

$$W = \begin{pmatrix} K_M(X_1 - X_T) & & & \\ & K_M(X_2 - X_T) & & \\ & & \ddots & \\ & & & K_M(X_n - X_T) \end{pmatrix}, \tag{6'}$$

and

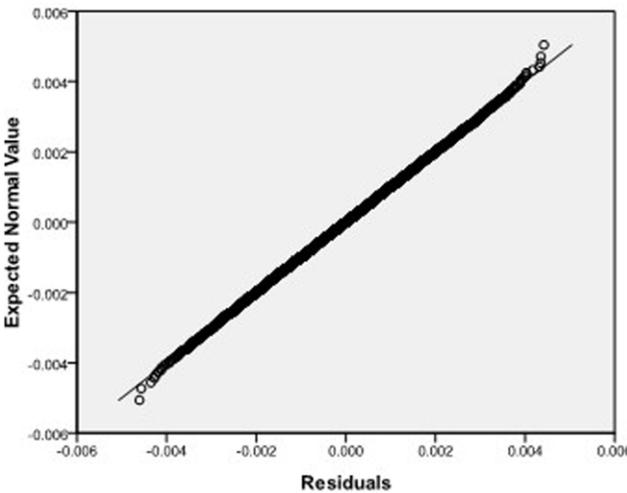


Figure 9. Normal Q-Q plot of residuals of $\hat{\beta}_2$ of GLS when the variance function of the error term is $\sigma^2(x_1, x_2) = \exp(-(x_1 + x_2))$.
doi:10.1371/journal.pone.0043719.g009

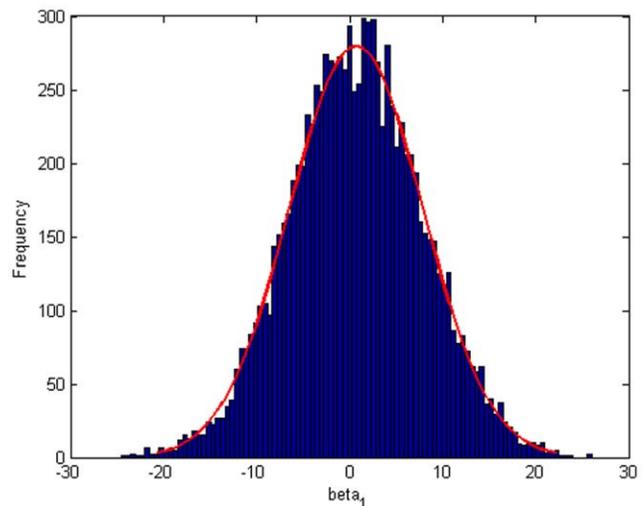


Figure 11. Histogram and asymptotic distribution for $\hat{\beta}_1$ of OLS when the variance function of the error term is $\sigma^2(x_1, x_2) = \exp(-(x_1 + x_2))$.
doi:10.1371/journal.pone.0043719.g011

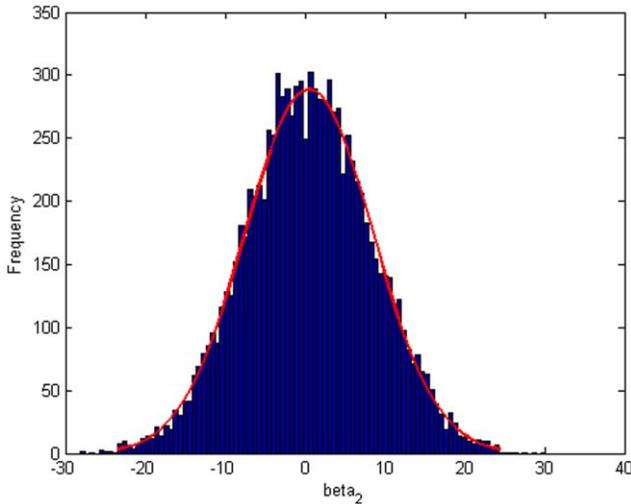


Figure 12. Histogram and asymptotic distribution for $\hat{\beta}_2$ of OLS when the variance function of the error term is $\sigma^2(x_1, x_2) = \exp(-(x_1 + x_2))$.
doi:10.1371/journal.pone.0043719.g012

$$X_x = \begin{pmatrix} 1 & (X_1 - X_T)^T & \text{vech}\{(X_1 - X_T)(X_1 - X_T)^T\} & \cdots \\ 1 & (X_2 - X_T)^T & \text{vech}\{(X_2 - X_T)(X_2 - X_T)^T\} & \cdots \\ \vdots & \vdots & \vdots & \vdots \\ 1 & (X_n - X_T)^T & \text{vech}\{(X_n - X_T)(X_n - X_T)^T\} & \cdots \end{pmatrix}. \quad (7)$$

We then have the least squared solution with multivariate local polynomial fitting [22].

$$\hat{\theta}(X_T) = (W^{1/2} X_x)^\dagger Y, \quad (8)$$

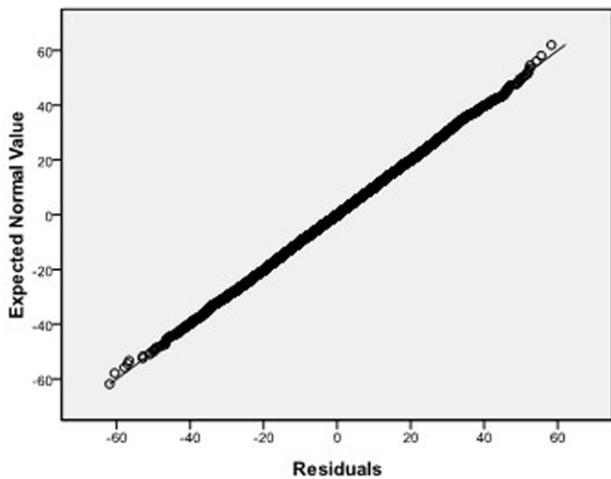


Figure 13. Normal Q-Q plot of residuals of $\hat{\beta}_0$ of OLS when the variance function of the error term is $\sigma^2(x_1, x_2) = \exp(-(x_1 + x_2))$.
doi:10.1371/journal.pone.0043719.g013

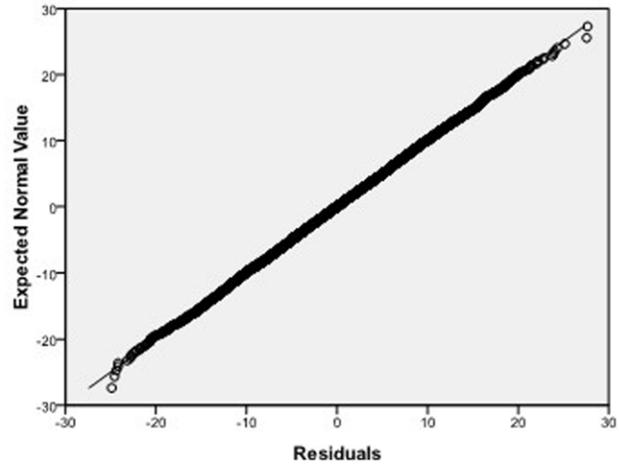


Figure 14. Normal Q-Q plot of residuals of $\hat{\beta}_1$ of OLS when the variance function of the error term is $\sigma^2(x_1, x_2) = \exp(-(x_1 + x_2))$.
doi:10.1371/journal.pone.0043719.g014

where (\dagger) denotes pseudo-inverse, or when $X_x^T W X_x$ is inverse, the estimation can be written by

$$\hat{\theta}(X_T) = (X_x^T W X_x)^{-1} X_x^T W Y. \quad (9)$$

Then, we can get the estimation $\hat{f}(X_T)$,

$$\hat{f}(X_T) = E_1 (X_x^T W X_x)^{-1} X_x^T W Y, \quad (10)$$

where E_1 is a column vector (the same size of a in (5)) with the first element equal to 1, and the rest equal to zero, that is, $E_1 = (1, 0, 0, \dots, 0)_{1 \times (p+1)}$.

Computing the \hat{a} will suffer from large computational cost. We can use the recursive least squared method to reduce the computation complexity, and it is very powerful especially in the local polynomial fitting problems. There are several important

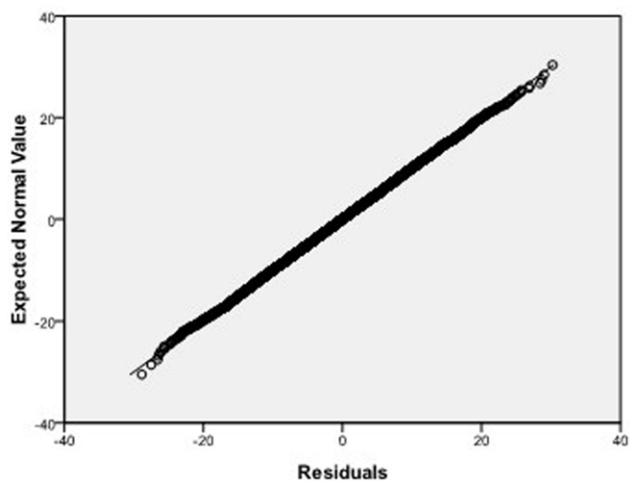


Figure 15. Normal Q-Q plot of residuals of $\hat{\beta}_2$ of OLS when the variance function of the error term is $\sigma^2(x_1, x_2) = \exp(-(x_1 + x_2))$.
doi:10.1371/journal.pone.0043719.g015

Table 2. Two-stage estimates of β_0 , β_1 , and β_2 of GLS and OLS.

	β_0	β_1	β_2
true value	1.8	0.6	0.4
GLS fitted value	1.79996	0.60014	0.39997
GLS relative error	0.0022%	0.233%	0.0075%
OLS fitted value	1.84603	0.57973	0.38452
OLS relative error	2.5572%	3.3783%	3.8700%

doi:10.1371/journal.pone.0043719.t002

issues about the bandwidth, the order of multivariate local polynomial function and the kernel function which have to be discussed. The three problems will be presented in following subsection.

Parameters Selections

For the multivariate local polynomial estimator, there are three important problems which have significant influence to the estimation accuracy and computational complexity [23].

First of all, there is the choice of the bandwidth matrix, which plays a rather crucial role. The bandwidth matrix M is taken to be a diagonal matrix. For simplification, the bandwidth matrix is designed into $M = hI_m$. Therefore, the most important thing is to find the bandwidth h [24,25]. In theory, there exists a optimal bandwidth h_{opt} in the meaning of mean integrated square error(MISE), fulfilling the equality

$$h_{opt} = \arg \min_h \int (f(X_T) - \hat{f}(X_T))^2 dX_T. \quad (11)$$

However, the theoretical bandwidth h_{opt} in formula (23) can not be directly calculated. Here, we propose a search method to select the bandwidth: Compare values of the objective function as the bandwidth h from small to large, and then find out the optimal bandwidth which minimize the objective function.

Suppose that $h_l = Kh_{min}$, where h_{min} is the minimum, K is coefficient of expansion. We search a bandwidth h to minimize the objective function in the interval $[h_{min}, h_{max}]$, where the objective

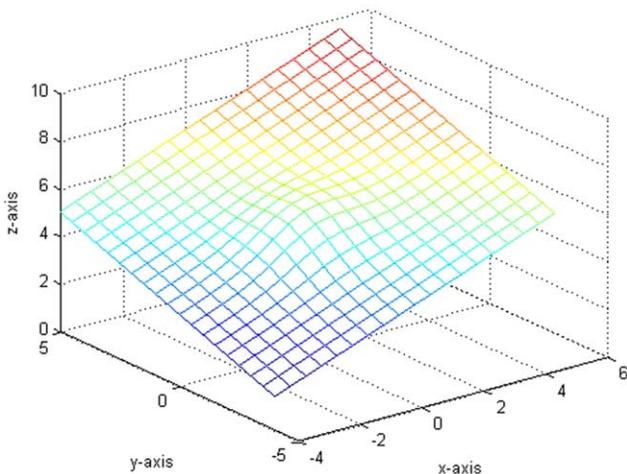


Figure 16. The fitted plot after 10000 replicates.
doi:10.1371/journal.pone.0043719.g016

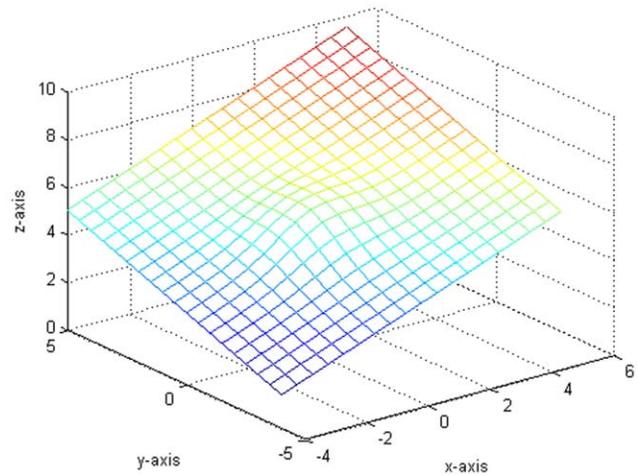


Figure 17. The original plot of the variance function.
doi:10.1371/journal.pone.0043719.g017

function refers to the prediction mean square error (MSE), denoted by $E_{MS}(h)$.

Firstly, we assume $h = h_{min}$, then increase h by efficient of expansion K and calculate value of objective function for each h . Stop down when $h > h_{max}$, and choose a bandwidth h which minimizes $E_{MS}(h)$ as the approximate optimal bandwidth. $E_{MS}(h)$ can be taken place by a estimation

$$e_{MS} = \frac{1}{n} \sum_{i=1}^n (Y_i - \hat{Y}_i)^2.$$

In this paper, we choose

$$h_{min} = \frac{2h_m}{n}, h_{max} = \frac{h_m}{2}, \quad (12)$$

where $h_m = \max \|X_i - X_j\|, K = 1.1$. Compared with other methods, this method is more convenient.

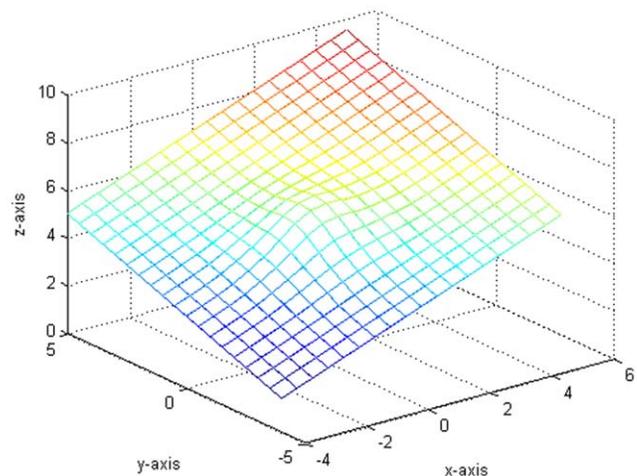


Figure 18. The plot of drawing fitted plot and original plot together.
doi:10.1371/journal.pone.0043719.g018

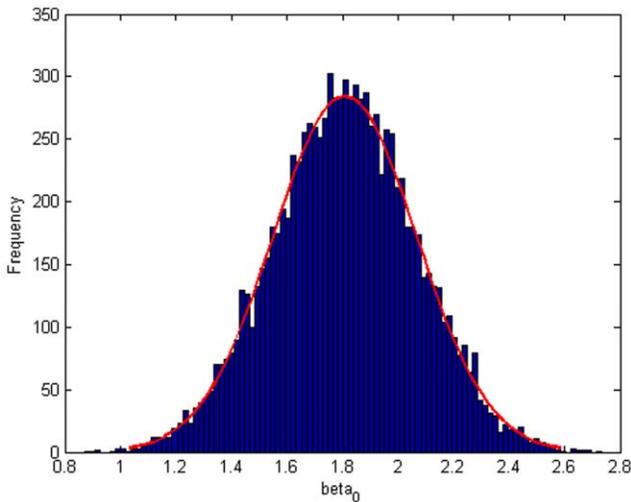


Figure 19. Histogram and asymptotic distribution for $\hat{\beta}_0$ of GLS when the variance function of the error term is $\sigma^2(x_1, x_2) = \exp(-(x_1^2 + x_2^2))$.
doi:10.1371/journal.pone.0043719.g019

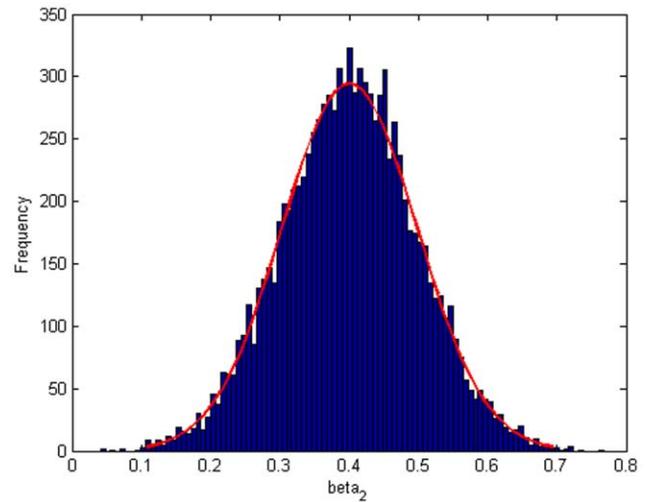


Figure 21. Histogram and asymptotic distribution for $\hat{\beta}_2$ of GLS when the variance function of the error term is $\sigma^2(x_1, x_2) = \exp(-(x_1^2 + x_2^2))$.
doi:10.1371/journal.pone.0043719.g021

In order to closer to the ideal optimal bandwidth, we search once again by narrowing the interval on the basis of the above searching process. Supposing j is the bandwidth which make $K^j h_{\min}$ optimal in the above searching process. Now, divide the small interval $[K^{j-1} h_{\min}, K^{j+1} h_{\min}]$ into n equal intervals. Supposing

$$h_i = h_{\min} K^{j-1} \left(1 + \frac{i}{n} (K^2 - 1) \right) \quad i = 1, \dots, n-1, \quad (13)$$

among these $n-1$ bandwidths, the approximate optimal bandwidth is the one that makes e_{MS} minimize. Obviously, this search method can quickly select the right bandwidth.

Another issue in multivariate local polynomial fitting is the choice of the order of the polynomial. For a given bandwidth h , a

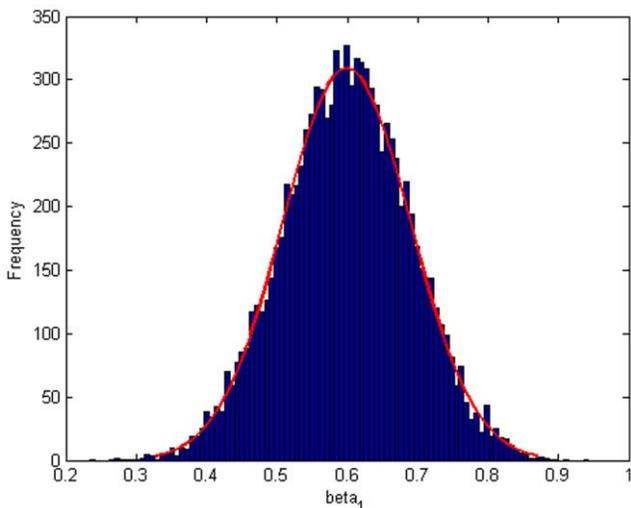


Figure 20. Histogram and asymptotic distribution for $\hat{\beta}_1$ of GLS when the variance function of the error term is $\sigma^2(x_1, x_2) = \exp(-(x_1^2 + x_2^2))$.
doi:10.1371/journal.pone.0043719.g020

large value of p would expectedly reduce the modeling bias, but would cause a large variance and a considerable computational cost. Since the bandwidth is used to control the modeling complexity, and due to the sparsity of local data in multi-dimensional space, a higher-order polynomial is rarely used. So we apply the local quadratic regression to fit the model (that is to say, $p=2$).

The third issue is the selection of the kernel function. In this paper, we choose the spherical Epanechnikov kernel as kernel function

$$K(x) = \{m(m+2)\Gamma(m/2)/(4\pi)^{m/2}\}(1 - \|x\|^2)_+, \quad (14)$$

where $x = (x_1, x_2, \dots, x_m)$, $\|x\| = (x_1^2 + x_2^2 + \dots + x_m^2)^{1/2}$, and $\Gamma(\cdot)$ represents Γ function. This is the optimal kernel function, see [14,19,20].

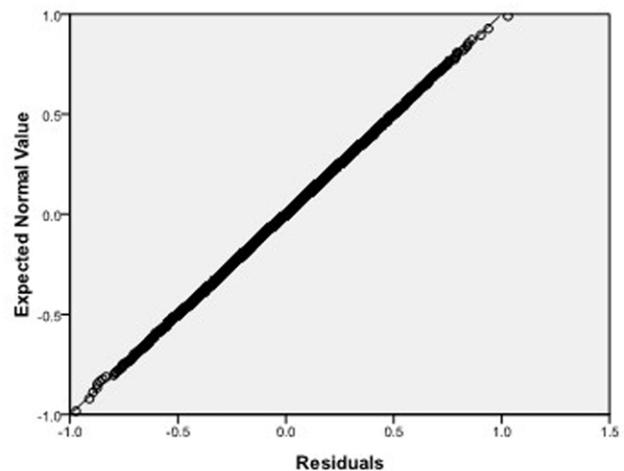


Figure 22. Normal Q-Q plot of residuals of $\hat{\beta}_0$ of GLS when the variance function of the error term is $\sigma^2(x_1, x_2) = \exp(-(x_1^2 + x_2^2))$.
doi:10.1371/journal.pone.0043719.g022

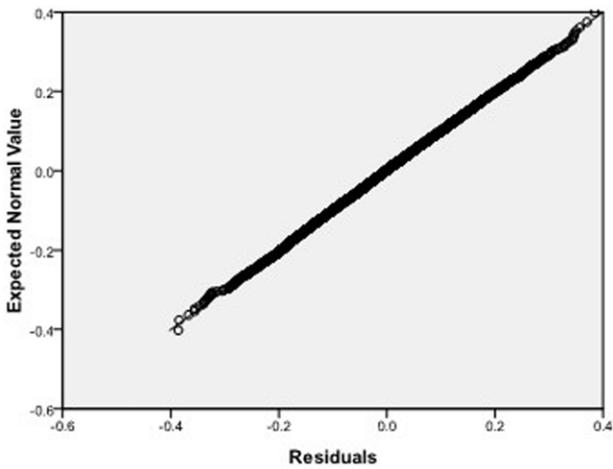


Figure 23. Normal Q-Q plot of residuals of $\hat{\beta}_1$ of GLS when the variance function of the error term is $\sigma^2(x_1, x_2) = \exp(-(x_1^2 + x_2^2))$. doi:10.1371/journal.pone.0043719.g023

Parametric Estimations with Multivariate Local Polynomial Regression

Let the dependent variable x and the explanatory variable y fulfill the following regression model \mathcal{L}° .

$$y_i = \beta_0 + \beta_1 x_{1i} + \dots + \beta_m x_{mi} + \sigma(x_1, \dots, x_m) \varepsilon_i, \quad i = 1, 2, \dots, n, \quad (15)$$

where y_i are the observations and x_{1i}, \dots, x_{mi} are independent variables. Denote

$$y = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix}, \quad \beta = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_m \end{pmatrix} \quad (16)$$

and

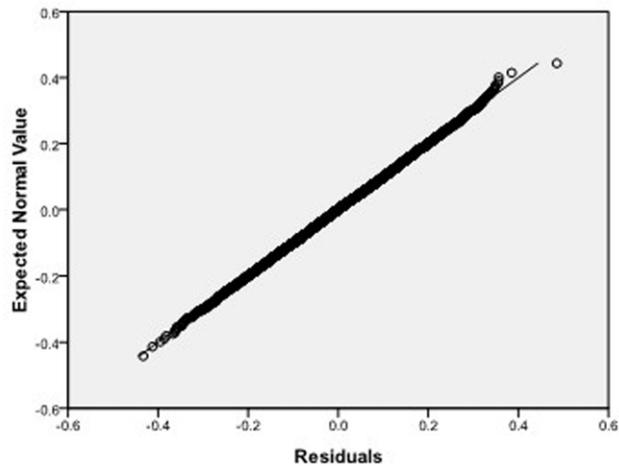


Figure 24. Normal Q-Q plot of residuals of $\hat{\beta}_2$ of GLS when the variance function of the error term is $\sigma^2(x_1, x_2) = \exp(-(x_1^2 + x_2^2))$. doi:10.1371/journal.pone.0043719.g024

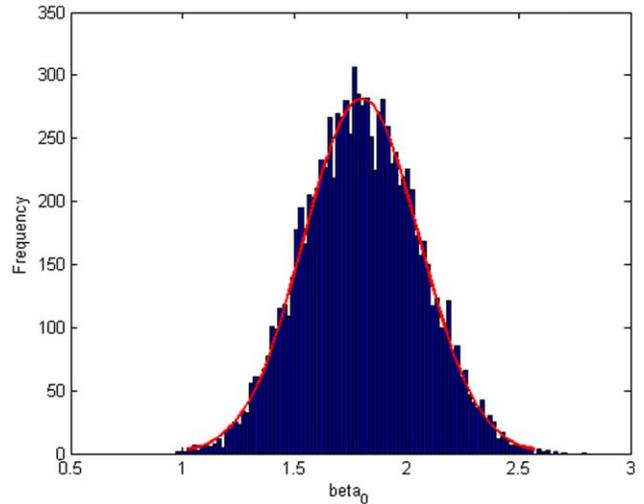


Figure 25. Histogram and asymptotic distribution for $\hat{\beta}_0$ of OLS when the variance function of the error term is $\sigma^2(x_1, x_2) = \exp(-(x_1^2 + x_2^2))$. doi:10.1371/journal.pone.0043719.g025

$$X_x = \begin{pmatrix} 1 & x_{11} & x_{21} & \dots & x_{m1} \\ 1 & x_{12} & x_{22} & \dots & x_{m2} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & x_{1n} & x_{2n} & \dots & x_{mn} \end{pmatrix}, \quad \varepsilon = \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{pmatrix}. \quad (17)$$

Therefore, the equation (15) can be abbreviated as

$$y = X_x \beta + \sigma(X) \varepsilon. \quad (18)$$

Suppose that

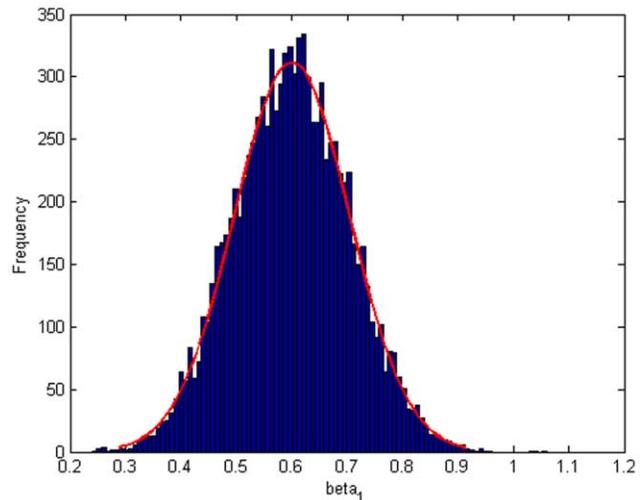


Figure 26. Histogram and asymptotic distribution for $\hat{\beta}_1$ of OLS when the variance function of the error term is $\sigma^2(x_1, x_2) = \exp(-(x_1^2 + x_2^2))$. doi:10.1371/journal.pone.0043719.g026

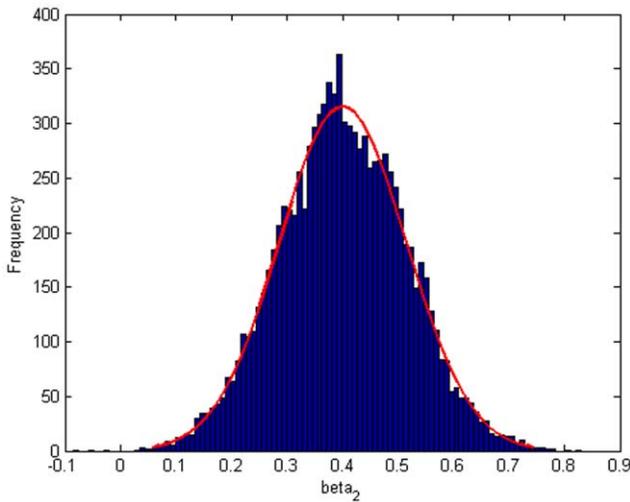


Figure 27. Histogram and asymptotic distribution for $\hat{\beta}_2$ of OLS when the variance function of the error term is $\sigma^2(x_1, x_2) = \exp(-(x_1^2 + x_2^2))$.
doi:10.1371/journal.pone.0043719.g027

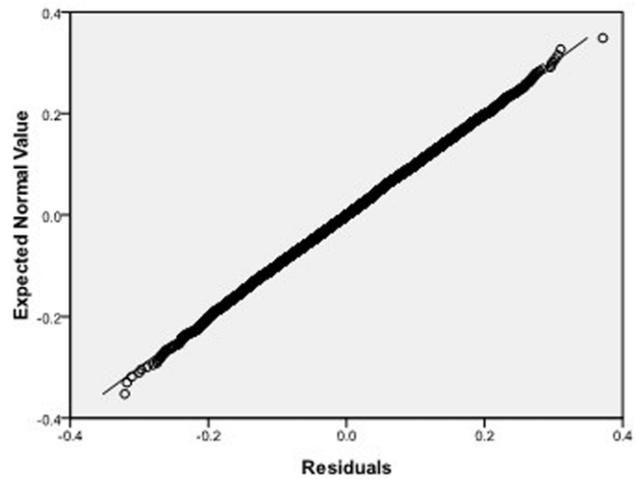


Figure 29. Normal Q-Q plot of residuals of $\hat{\beta}_1$ of OLS when the variance function of the error term is $\sigma^2(x_1, x_2) = \exp(-(x_1^2 + x_2^2))$.
doi:10.1371/journal.pone.0043719.g029

$$1) E(\sigma(X)\varepsilon) = (E(\sigma(X)\varepsilon_1), \dots, E(\sigma(X)\varepsilon_n))^T = (0, \dots, 0)^T.$$

$$2) \Sigma = E((\sigma(X)\varepsilon)(\sigma(X)\varepsilon)^T) = \text{diag}(\sigma_1^2, \dots, \sigma_n^2),$$

where $\sigma_i^2 = \text{var}((\sigma(x_i)\varepsilon_i)), i = 1, 2, \dots, n$.

$\sigma_1^2, \sigma_2^2, \dots, \sigma_n^2$ are all not equal, that is, there is a heteroscedasticity in model (18). Therefore, the WLS of β is.

$$\beta = (X_x^T \Sigma^{-1} X_x)^{-1} X_x^T \Sigma^{-1} y, \tag{19}$$

or it can be written as

$$\beta = \left(\sum_{i=1}^n \frac{1}{\sigma_i^2} x_i x_i^T \right)^{-1} \left(\sum_{i=1}^n \frac{1}{\sigma_i^2} x_i y_i \right), \tag{20}$$

where $x_i = (1, x_{1i}, x_{2i}, \dots, x_{mi})^T, i = 1, 2, \dots, n$. If covariance matrix Σ of ε is known, formula (20) can be estimated. Equation (20) is considered as the weighted least squares estimate of β and it possesses nice qualities. However, in many practical situations, the form of Σ is unknown. Therefore, the so-called two-stage method of estimation is used to solve the heteroscedasticity problem. It can be depicted as follows: first, apply multivariate local polynomial fitting to get the estimate $\hat{\sigma}_i^2$ of σ_i^2 , and then we can obtain the estimate $\hat{\beta}$ of β by using equation (20). The estimator follows that

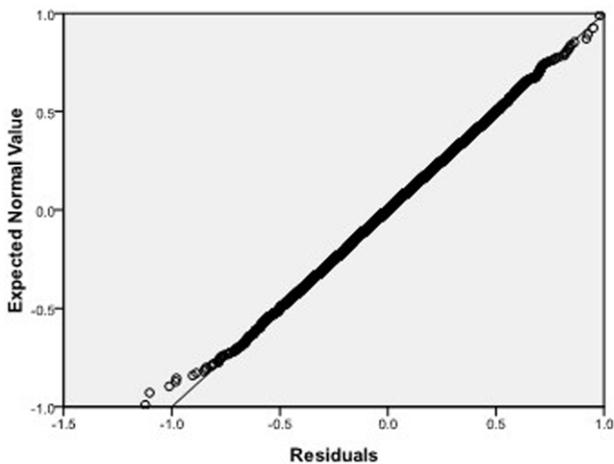


Figure 28. Normal Q-Q plot of residuals of $\hat{\beta}_0$ of OLS when the variance function of the error term is $\sigma^2(x_1, x_2) = \exp(-(x_1^2 + x_2^2))$.
doi:10.1371/journal.pone.0043719.g028

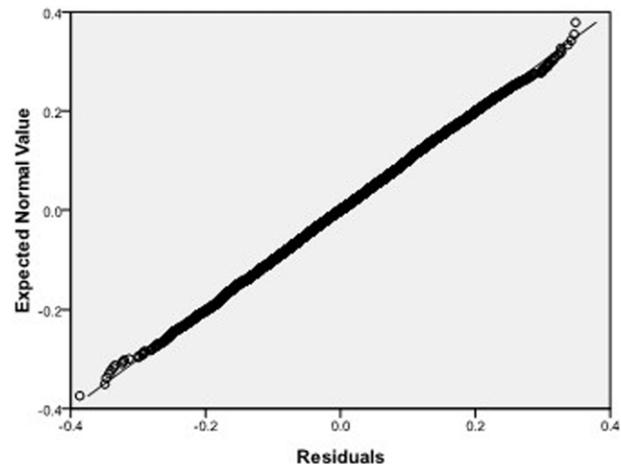


Figure 30. Normal Q-Q plot of residuals of $\hat{\beta}_2$ of OLS when the variance function of the error term is $\sigma^2(x_1, x_2) = \exp(-(x_1^2 + x_2^2))$.
doi:10.1371/journal.pone.0043719.g030

Table 3. Two-stage estimates for β_0 , β_1 , and β_2 of GLS and OLS.

true value	1.8	0.6	0.4
GLS fitted value	1.79996	0.60004	0.39992
GLS relative error	0.0022%	0.0067%	0.02%
OLS fitted value	1.7992	0.6045	0.3917
OLS relative error	0.044%	0.75%	2.075%

doi:10.1371/journal.pone.0043719.t003

$$\hat{\beta} = \left(\sum_{i=1}^n \frac{1}{\hat{\sigma}_i^2} x_i x_i^T \right)^{-1} \left(\sum_{i=1}^n \frac{1}{\hat{\sigma}_i^2} x_i y_i \right). \quad (21)$$

Because of $E(\sigma(X)\varepsilon_i|X) = \sigma_i^2$, we construct the following regression model in order to estimate σ_i^2 ,

$$(\sigma(X)\varepsilon_i)^2 = \sigma_i^2 + u_i, \quad E u_i = 0, \quad i = 1, 2, \dots, n, \quad (22)$$

where u_i is the difference between $(\sigma(X)\varepsilon_i)^2$ and its expectation. Suppose that $b = (X_x^T X_x)^{-1} X_x^T y$ is the OLS of model (18). Although the ordinary least squares estimate b is ineffective, it is still consistent. Therefore, the corresponding residuals hold that

$$\begin{aligned} e_i^2 &= (x_i^T \beta + (\sigma(X)\varepsilon_i) - x_i^T b)^2 \\ &= (\sigma(X)\varepsilon_i)^2 + [x_i^T (\beta - b)]^2 \\ &\quad + 2(\sigma(X)\varepsilon_i)x_i^T (\beta - b) \approx (\sigma(X)\varepsilon_i)^2. \end{aligned} \quad (23)$$

Consequently, we can approximately get

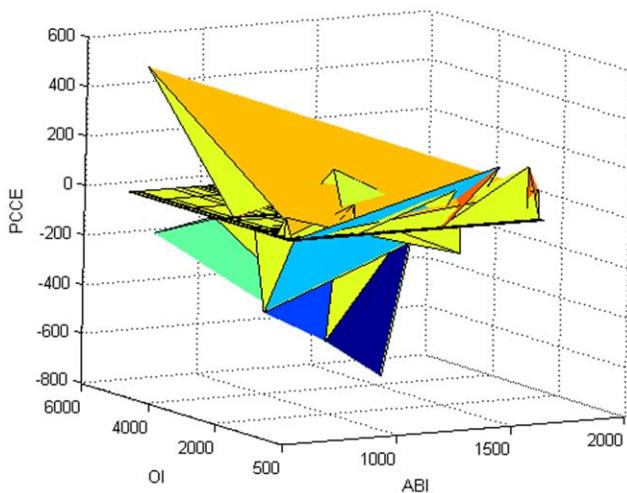


Figure 31. Variance squares.
doi:10.1371/journal.pone.0043719.g031

Table 4. The data of PCCE, ABI and OI for regions in China in 2009.

areas	PCCE	ABI	OI
xinjiang	1350.23	1300.10	410.30
neimenggu	1554.60	1497.80	480.50
jilin	1661.70	1634.60	547.60
heilongjiang	1604.50	1684.10	596.20
xizang	1123.71	589.60	614.40
yunnan	1336.25	889.40	644.30
guizhou	1098.39	764.00	647.80
qinghai	1334.45	803.80	753.50
hainan	1357.43	1386.70	839.80
guangxi	1550.62	1068.80	875.60
shanxi	1331.03	614.80	876.00
gansu	1127.37	621.60	887.00
ningxia	1388.79	859.60	963.40
hubei	1649.20	1352.00	1000.10
anhui	1412.40	1013.10	1006.90
henan	1375.60	1083.80	1014.10
sichuan	1497.52	919.30	1067.60
chongqing	1475.16	883.20	1088.00
jiangxi	1720.00	1027.80	1203.80
liaoning	1786.30	1254.30	1303.60
shanxi	1221.60	609.80	1346.20
hunan	1990.30	908.20	1391.30
shandong	1905.00	1293.00	1511.60
hebei	1429.80	928.80	1674.80
fujian	2503.10	1053.00	2327.70
guangdong	2703.36	1242.90	2526.90
jiangsu	2374.70	1177.60	2607.20
tianjin	2050.90	1314.60	2633.10
zhejiang	3479.20	985.80	3596.60
beijing	3552.10	579.10	4446.40
shanghai	4753.20	652.50	5218.40

doi:10.1371/journal.pone.0043719.t004

$$e_i^2 = \sigma_i^2 + v_i, \quad E(v_i) = 0, \quad i = 1, 2, \dots, n. \quad (24)$$

It can be taken as a regression model, in which the variance function is regression function and the squared residuals e_i^2 are dependent variables. In order to estimate this model, parameter estimation method would usually be taken in some articles. In other words, they suppose $\sigma_i^2 = f(c^T, x_{1i}, \dots, x_{mi})$, where the form of f is known and $c = (c_0, c_1, \dots, c_n)^T$ are the parameters to be estimated. Note that what we usually discuss about more and more detail are $\sigma_i^2 = \sigma^2(c^T, x_{1i}, \dots, x_{mi})$, $\sigma_i^2 = \sigma^2 \exp(c^T, x_{1i}, \dots, x_{mi})$ and so on [26]. However, the discussion of these models requires that the analyst have a better understanding for the background in practical problems. As an example, variance of corporate profits is often in direct proportion with family income. Since the variance function must be non-negative, a non-parametric method is proposed to fit σ_i^2 . This method can be depicted as follows. Then,

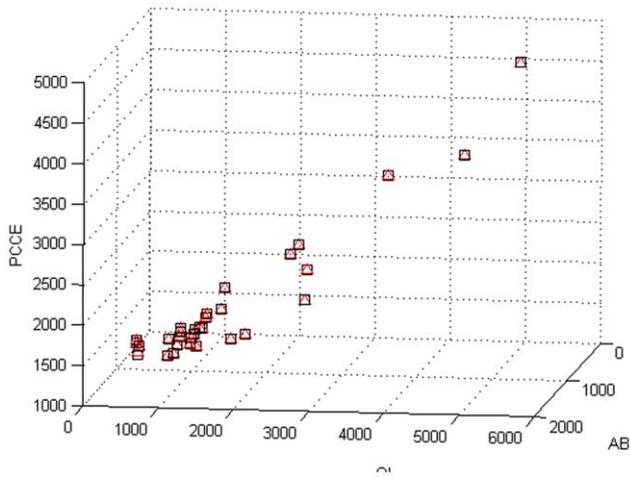


Figure 32. Scatter of original and fitting data.
doi:10.1371/journal.pone.0043719.g032

the p -order local polynomial estimate of the variances function $\sigma^2(x_{1i}, \dots, x_{mi})$ is according to formula (2). Using the least squares method for the data around the local window, we can estimate the local intercept via minimizing

$$\sum_{i=1}^n \left\{ Y_i - \sum_{0 \leq |k| \leq p} \alpha_k(X_T)(X_i - X_T) \right\}^2 K_M(X_i - X_T). \quad (25)$$

Furthermore, the solution vector can be written as

$$\alpha(X_T) = (X_{xx}^T W_x X_{xx})^{-1} X_{xx}^T W_x e^2, \quad (26)$$

where the weighted matrix

$$W_x = \text{diag}\{K_M(X_1 - X_T), \dots, K_M(X_n - X_T)\},$$

the the design matrix

$$X_{xx} = \begin{pmatrix} 1 & (x_1 - x)^T & \text{vech}\{(x_1 - x)(x_1 - x)^T\} & \dots \\ 1 & (x_2 - x)^T & \text{vech}\{(x_2 - x)(x_2 - x)^T\} & \dots \\ \vdots & \vdots & \ddots & \vdots \\ 1 & (x_n - x)^T & \text{vech}\{(x_n - x)(x_n - x)^T\} & \dots \end{pmatrix}, \quad (27)$$

and $e^2 = (e_1^2, e_2^2, \dots, e_n^2)^T$. Consequently, the estimated variance function is $\hat{\sigma}^2(X_{1i}, \dots, X_{mi}) = E_1 \alpha(X_T)$. Finally, we can get two-stage estimate $\hat{\beta}$ of β by substituting estimate $\hat{\sigma}_i^2$ of σ_i^2 into equation (21).

Results and Discussion

Simulations and Analysis

In this section, we give the following model to discuss the qualities of $\hat{\beta}$ under the limited sample. Considering the practical background which is applied to economics, we assume the variance function of the following two forms.

Model 1. Denote the linear model by

$$y_i = 1.8 + 0.6x_{1i} + 0.4x_{2i} + \sigma(x_{1i}, x_{2i})\varepsilon_i, \quad i = 1, 2, \dots, n \quad (28)$$

where x_{1i} and x_{2i} ($i = 1, 2, \dots, n$) are independent variables. Here, we assume the variance function of the error term $\sigma^2(x_1, x_2) = \exp(-(x_1^2 + x_2^2))$.

Step 1: Firstly, obtain the estimation of two coefficients in model (28) with the ordinary least squares estimation. Second, calculate the squares of the residuals e^2 . Third, do the local polynomial regression on the model (24) Figure 1 is the result after 10000 replicates, where $n = 20$, $p = 2$, and the kernel function is given by.

$$K_M(x_1, x_2) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{x_1^2}{2}\right) \cdot \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{x_2^2}{2}\right) \quad (29)$$

In the kernel function, the ranges of x_1 and the same with x_2 are $[-4, 5]$, and the optimal bandwidth matrix is $M = \begin{pmatrix} 0.67 & 0 \\ 0 & 0.67 \end{pmatrix}$. The criteria of selecting M is according to MISE, which can be given by

$$MISE \equiv E \int_{-4}^5 \{\sigma(X_T) - \hat{\sigma}(X_T)\}^2 dX_T.$$

That is $MISE = 0.00179$.

Minimizing the asymptotic MISE with respect to the bandwidth parameter h results in a bandwidth, called the asymptotically optimal bandwidth or simply optimal bandwidth. So we can obtain the optimal bandwidth matrix M . Figure 2 shows the plot which generates from the variance function.

$$\sigma^2(x_1, x_2) = \exp(-(x_1 + x_2)).$$

Figure 3 indicates the variance of the error term function and the fitting functions drawn in the same situation. Besides, we compare some fitted value with the true value in Table 1. It is worth pointing out that our fitting is very close to the real value from Figures 1, 2, 3 and Table 1.

Step 2: We substitute $\hat{\sigma}_i^2$ which obtained from step 1 into model (20), then we will get $\hat{\beta}$. Figures 4, 5, 6 depict the histograms and asymptotic distributions of $\hat{\beta}_0$, $\hat{\beta}_1$, and $\hat{\beta}_2$, respectively, which we do 10000 replicates with GLS and choose $n = 20$. It is easy to see from Figures 4, 5, 6 that the estimated distributions of parameters subject to normal asymptotically, which the proofs are in [27–30]. Further, we give normal Q-Q plots of residuals of estimated parameters in Figures 7, 8, 9. As illustrated in Figures 10, 11, 12, the histograms and asymptotic distributions of $\hat{\beta}_0$, $\hat{\beta}_1$, and $\hat{\beta}_2$ can be get by using OLS, respectively. Similarly, the normal Q-Q plots of residuals are depicted in Figures 13, 14, 15. However, the results are obviously unsatisfying in Figures 10, 11, 12, 13, 14, 15. Besides, the fitted value and true value of GLS and ones of OLS are listed in Table 2. Therefore, we can conclude that our fitting is very perfect. Here, we define the relative error as

$Re = \frac{|\hat{\xi} - \xi|}{\xi} \times 100\%$, where ξ is a true value, and $\hat{\xi}$ is a fitted value. So the GLS estimates of parameters are much better than OLS according to Table 2.

Model 2. In formula (28), we assume the error term function is

$$\sigma^2(x_1, x_2) = 5 + 0.5x_1 + 0.4x_2 + \exp(-(x_1^2 + x_2^2)/2),$$

and the optimal bandwidth matrix is $M = \begin{pmatrix} 0.95 & 0 \\ 0 & 0.95 \end{pmatrix}$, where $MISE = 0.01748$. The process of parameter's estimation and calculation is the same with Model 1. Also we choose $p = 2$, the ranges of x_1 and the same with x_2 are $[-4, 5]$. Besides, the kernel function is still defined as equation (29). We do a total of 10000 replicates. Figure 16 shows the estimated variance function $\hat{\sigma}^2(x_{1i}, \dots, x_{mi})$, where the selection of optimal bandwidth matrix is similar with that in step 1. Figure 17 shows the plot which generates from the variance function

$$\sigma^2(x_1, x_2) = 5 + 0.5x_1 + 0.4x_2 + \exp(-(x_1^2 + x_2^2)/2).$$

Figure 18 indicates the plot of drawing fitted plot and original plot together. According to Figures 16, 17, 18, we can easily reach the conclusion that our fitting is nice and close to the true value. The histograms and asymptotic distributions of $\hat{\beta}_0$, $\hat{\beta}_1$ and $\hat{\beta}_2$ are depicted in Figures 19, 20 and 21. Besides, Figures 22, 23 and 24 contain the normal Q-Q plots of residuals, respectively. It is worth pointing out that the asymptotic distributions of parameters are symmetric. As shown in Figures 25, 26 and 27, the histograms and asymptotic distributions of $\hat{\beta}_0$, $\hat{\beta}_1$, and $\hat{\beta}_2$ of OLS are not symmetrical, but irregular. There are larger fluctuations in normal Q-Q plots of residuals in Figures 28, 29 and 30 than those in Figures 22, 23 and 24. So it is not what we want to reach. Besides, we list fitted value and true value of β_0 , β_1 , and β_2 of GLS and ones of OLS in Table 3. In order to obtain a visual effect, we calculate the relative errors and do a comparison. It is easy to see that the GLS estimates of parameters are much better than OLS from Table 3.

Application

As an example of pure cross-sectional data with potential for heteroscedasticity, consider the data given in Table 4, which gives data on per capita consumption expenditure (PCCE), agricultural business income (ABI) and other income (OI) for 31 different regions in China in 2009. The cross-sectional data presented in this table are quite heterogenous in a regression of PCCE on ABI and OI, so heteroscedasticity is likely.

We hope to understand the relationship of PCCE, ABI and OI, then the regression function is as follows:

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \sigma(x_{1i}, x_{2i}) \varepsilon_i.$$

References

- Galtchouk L, Pergamenschikov S (2009) Adaptive asymptotically efficient estimation in heteroscedastic nonparametric regression. *J Korean Stat Soc* 38: 305–322.
- Galtchouk L, Pergamenschikov S (2009) Sharp non-asymptotic oracle inequalities for nonparametric heteroscedastic regression models. *J Nonparametr Stat* 21: 1–18.
- Amemiya T (1977) A note on a heteroscedastic model. *J Econom* 6: 365–370.
- Chen P, Suter D (2007) A bilinear approach to the parameter estimation of a general heteroscedastic linear system, with application to conic fitting. *J Math Imaging Vis* 28: 191–208.
- Ma JH, Liu LX (2008) Multivariate nonlinear analysis and prediction of Shanghai stock market. *Discrete Dyn Nat Soc* 2008: 526734.
- Liu H, Zhang ZQ, Zhao Q (2009) The volatility of the index of shanghai stock market research based on ARCH and its extended forms. *Discrete Dyn Nat Soc* 2009: 743685.

Firstly, the ordinary least squares estimators can be obtained, say, $\hat{\beta}_{O0} = 356.4174$, $\hat{\beta}_{O1} = 0.4282$, $\hat{\beta}_{O2} = 0.6931$. Then, inspect that if there is a heteroscedasticity. According to formula (23), the residual squares can be considered as variances. So we need only to fit residual squares, seeing Figure 31, where residual squares are variance squares. Obviously, there are biases among the residual squares fitted. Therefore, it can be concluded that there is a heteroscedasticity. Furthermore, according to the two-stage estimate, the generalized least squares estimation for β_{G0} , β_{G1} , and β_{G2} can be estimated after 10000 replicates provided with the bandwidth matrix $M = \begin{pmatrix} 0.48 & 0 \\ 0 & 0.48 \end{pmatrix}$, the order $p = 2$, and $MISE = 0.047369$. The estimations are $\hat{\beta}_{G0} = 361.7432$, $\hat{\beta}_{G1} = 0.4165$, and $\hat{\beta}_{G2} = 0.7087$. Finally, the regression equation of GLS can easily be written as follow:

$$\hat{y}_i = 361.7432 + 0.4165x_{1i} + 0.7087x_{2i}. \quad (30)$$

The scatter of original and fitting data is also given, see Figure 32.

Conclusions

In this paper \mathcal{L} - we presented a new method for estimation of multivariate linear heteroscedastic regression model based on multivariate local polynomial estimation with non-parametric technique. The proposed scheme firstly adopted the local polynomial fitting to estimate heteroscedastic function, then the coefficients of regression model are obtained based on generalized least squares method. Our approach avoided the test of heteroscedasticity for the multivariate linear model. Due to non-parametric technique of local polynomial estimation, if the heteroscedastic function is unknown, the precision of estimation was improved. Furthermore, the asymptotic normality of parameters was verified by the results of numerical simulations normal Q-Q plots. Finally, the simulation results and local polynomial estimation of real data really indicated that our approach was effective in finite-sample situations, which did not need to assume the form of heteroscedastic function. The presented algorithm could be easily used to heteroscedastic regression model in some practical problems.

Author Contributions

Conceived and designed the experiments: LYS YYZ. Performed the experiments: LYS YYZ TSY. Analyzed the data: LYS FLL. Contributed reagents/materials/analysis tools: LYS YYZ. Wrote the paper: LYS YYZ TSY. Collected the data: TSY FLL. Applied the real data: LYS TSY FLL.

7. Cai L, Hayes AF (2008) A new test of linear hypotheses in OLS regression under heteroscedasticity of unknown form. *J Educ Behav Stat* 33: 21–40.
8. Hayes AF, Cai L (2007) Using heteroskedasticity-consistent standard error estimators in OLS regression: An introduction and software implementation. *Behav Res Methods* 39: 709–722.
9. Lima VMC, Souza TC, Cribari-Neto F, Fernandes GB (2009) Heteroskedasticity-robust inference in linear regressions. *Commun Stat Simul Comput* 39: 194–206.
10. Almeida AM, Castel-Branco MM, Falcão AC (2002) Linear regression for calibration lines revisited: weighting schemes for bioanalytical methods. *J Chromatogr B* 774: 215–222.
11. Härdle W, Werwatz A, Müller M, Sperlich S (2004) *Nonparametric and semiparametric models*. New York: Springer. 85–88.
12. Fan JQ (1993) Local linear regression smoothers and their minimax efficiency. *Ann Stat* 21: 196–216.
13. Su LY, Zhao YY, Yan TT (2012) Two-stage method based on local polynomial fitting for a linear heteroscedastic regression model and its application in economics. *Discrete Dyn Nat Soc* 2012: 696927.
14. Fan JQ (1992) Design-adaptive nonparametric regression. *J Am Stat Assoc* 87: 998–1004.
15. Godfrey LG (2006) Tests for regression models with heteroskedasticity of unknown form. *Comput Stat Data Anal* 50: 2715–2733.
16. Su LY (2010) Prediction of multivariate chaotic time series with local polynomial fitting. *Comput Math Appl* 59: 737–744.
17. Su LY (2011) Multivariate local polynomial estimation with application to Shenzhen component index. *Discrete Dyn Nat Soc* 2011: 930958.
18. Su LY, Ma YJ, Li JJ (2012) Application of local polynomial estimation in suppressing strong chaotic noise. *Chinese Physics B*. 21: 020508.
19. Fan JQ, Yao QW (2003) *Nonlinear time series: nonparametric and parametric methods*. New York: Springer. 313–317.
20. Fan JQ, Zhang WY (1999) Statistical estimation in varying coefficient models. *Ann Stat* 27: 1491–1518.
21. Farsiu S, Robinson MD, Elad M, Milanfar P (2004) Fast and robust multiframe super-resolution. *IEEE Trans Image Process* 13: 1327–1344.
22. Ruppert D, Wand MP (1994) Multivariate locally weighted least squares regression. *Ann Stat* 22: 1346–1370.
23. Su LY, Li FL (2010) Deconvolution of defocused image with multivariate local polynomial regression and iterative Wiener filtering in DWT domain. *Math Probl Eng* 2010: 605241.
24. Fan JQ, Gijbels I (1996) *Local polynomial modelling and its applications*. London: Chapman and Hall. 263–276.
25. Fan JQ, Gijbels I (1995) Data-driven bandwidth selection in local polynomial fitting: variable bandwidth and spatial adaptation. *J R Stat Soc Series B Stat Methodol* 57: 371–394.
26. Rutemiller HC, Bowers DA (1968) Estimation in a heteroscedastic regression model. *J Am Stat Assoc* 63: 552–557.
27. White H (1980) A heteroskedasticity-consistent covariance matrix estimator and a direct test for heteroskedasticity. *Econometrica* 48: 817–838.
28. Cribari-Neto F (2004) Asymptotic inference under heteroskedasticity of unknown form. *Comput Stat Data Anal* 45: 215–233.
29. Cribari-Neto F, Souza TC, Vasconcellos KLP (2007) Inference under heteroskedasticity and leveraged data. *Commun Stat Theory Methods* 36: 1877–1888.
30. Greene WH (1993) *Econometric analysis*. New York: Macmillan. 362–366.