Mix local polynomial and spline truncated: the development of nonparametric regression model

To cite this article: Suparti et al 2018 J. Phys.: Conf. Ser. 1025 012102

View the article online for updates and enhancements.

Related content
- Modelling fourier regression for time series data: a case study: modelling inflation in foods sector in Indonesia
  Alan Prahutama, Suparti and Tiani Wahyu Utami
- Multicollinearity and Regression Analysis
  Jamal I. Daoud
- Application of Semiparametric Spline Regression Model in Analyzing Factors that Influence Population Density in Central Java
  Y D Sumantari, I Slamet and Sugiyanto
Mix local polynomial and spline truncated: the development of nonparametric regression model

Suparti¹, Alan Prahutama¹, Rukun Santoso¹
¹Department of Statistics, Faculty of Science and Mathematics, Diponegoro University
Jl. Prof. Soedharto, SH, Tembalang, Semarang 50275, Indonesia
E-mail: suparti702@gmail.com

Abstract. Regression analysis is a statistical method used to model response variables with predictor variables. The approach to regression analysis can be done by parametric, semiparametric and nonparametric. The nonparametric approach is more complex than parametric. Some nonparametric approaches include local and spline polynomial. Local polynomial is the development of the Taylor series which has an arbitrated fixed point, while the splines are pieces of the spline function. In both spline and local polynomial modeling requires a smoothing parameter determined by GCV (General Cross Validation) method. In this paper, we developed a method of combining local and spline polynomials. The parameter estimation uses the OLS (Ordinary Least Square) method, yielding an estimator \((X^TK_hX)^{-1}X^TK_hy\), where \(X\) is a predictor variable with combination of polynomial and spline function, whereas \(K_h\) is a kernel function used for the combination of local polynomial and spline function.

Keywords: Local Polynomial, Spline Truncated, Hybrid

1. Introduction
Regression analysis is a statistical method used for modeling between response variables and predictor variables. Regression approach can be done with three approaches, namely parametric approach, semiparametric, and nonparametric. The parametric approach is done when the curve shape is known, while the nonparametric approach is done when the curve shape is unknown. A semiparametric approach is done when some curves shape are known and some are unknown. The curve shape in question is the curve between the response variable and the predictor variable. The parametric approach is often done, due to the ease of estimation. However, the parametric approach is tied to assumptions such as normal distributed residuals, homocedasticity, nonautocorrelation, and among predictor variables do not occur multicollinearity [9]. So if a problem is modeled using a parametric approach, but the obtained model are not met the assumptions, of course the model will be biased to use.

Nowadays, the nonparametric approach is developed with several methods such as local polynomial, spline truncated, Fourier, Wavelet and others. Nonparametric approach does not require assumptions such as semiparametric approach. The focus of the nonparametric approach is modeling based on the shape of data pattern. So sometimes nonparametric approach has a better accuracy than parametric approach. However, the nonparametric approach is more complex than parametric [1]. The
development of spline such as adaptive spline for hierarchical data [11], cubic B-spline [12] and exponential pseudo spline [13].

One of nonparametric approach is local polynomial. The local polynomial approach is a nonparametric approach based on the degree of polynomial, kernel function and arbited fix point value [10]. In modeling, the local polynomial uses kernel function as its weighting by estimating the value of its smoothing parameter [2]. The estimation of smoothing parameter is based on General Cross Validation (GCV) calculation and also Cross Validation. In addition to local polynomial, another nonparametric approach method is Spline. Spline is a truncated polynomial based on the spline function, so spline has high flexibility. In spline modeling, firsts we determine the knot point using GCV [3].

The nonparametric approach is more complex than parametric. However, there is no guarantee that a more complex approach will result in a better model. This paper will discuss about the combined modeling between local polynomial and spline. The combining is based on data distribution, and combining is based on model and residual. On data based combining, the data pattern changes based on a period. Residual based combining means the model is estimated with local polynomial while the residual value is estimated with spline.

2. Materials and Method
2.1 Parametric Regression
Parametric regression is a regression approach to determine the relationship pattern between response variables and predictor variables with regression curve shape is known. The equation of simple linear regression model as follows [1]

\[ y_i = \beta_0 + \beta_1 x_i + \epsilon_i, \quad i = 1,2,...,n. \]  

If written in matrix form, we obtain regression model:

\[ y = X\beta + \epsilon, \quad \text{where} \quad \epsilon \sim N(0,\sigma^2 I). \]

Estimation of regression coefficient \( \beta \) is done by using OLS (Ordinary Least Square). This method is done by minimizing \( \epsilon^T \epsilon \) against \( \beta \). The value of \( \epsilon^T \epsilon = (y - X\beta)^T (y - X\beta) \), derived against \( \beta \) so we obtain estimator

\[ \hat{\beta} = (X^T X)^{-1} X^T y. \]  

2.2. Local Polynomial
In the nonparametric regression approach where the curve is unknown, the regression function is assumed to be contained in a function space [4]. Nonparametric regression model have the form of:

\[ y_i = \eta(t_i) + \epsilon_i, i = 1,2,...,l \]

with \( y_i \) is response variable. The function \( \eta \) shape is unknown with \( t_i \) as a predictor variable and \( \epsilon_i \) is assumed to be \( N(0,\sigma^2) \) distributed and the curve \( \eta \) is assumed to be smooth and in a certain space. One approach to estimate \( \eta(t_i) \) is Local Polynomial method. The Local Polynomial Estimator is obtained by Taylor series containing polynomials of \( p \) degree. If \( \eta(t_i) \) is dragged by Taylor series with polynomial \( p \) degree then we obtain

\[ \eta(t_i) \approx \eta(t) + (t_i - t)\eta^{(1)}(t) + \cdots + (t_i - t)^p \eta^{(p)}(t) / p! \]  

where \( t_i \in [t-h,t+h] \).

If given \( \beta_r(t) = \frac{\eta^{(r)}(t)}{r!} \) with \( r = 0,1,2,...,p \) then Eq. (4) can be written as

\[ \eta(t_i) \approx \beta_0(t) + (t_i - t)\beta_1(t) + \cdots + (t_i - t)^p \beta_p(t). \]
Eq. 5 can be written as follows:
\[ \eta(t_i) \approx x_i^T \beta \] with
\[ x_i = \begin{bmatrix} 1 & (t_i - t)^1 & (t_i - t)^2 & \cdots & (t_i - t)^p \end{bmatrix}^T \] and \( \beta = \begin{bmatrix} \beta_0(t) & \beta_1(t) & \beta_2(t) & \cdots & \beta_p(t) \end{bmatrix}^T \).

Obtaining the \( \hat{\beta} \) estimator is done by minimizing the Weighted Least Square (WLS) criteria as follows:
\[ \sum_{i=1}^{n} (y_i - x_i^T \beta)^2 K_h(t_i - t). \tag{6} \]

\( K_h() = \frac{K(\cdot/h)}{h} \) where \( K \) is a kernel function and \( h \) is a bandwidth, so WLS criteria can be written as follows:
\[ (y - X\beta)^T K_h (y - X\beta) \tag{7} \]
where \( X_j = \begin{bmatrix} x_{j1}, x_{j2}, \ldots, x_{jn} \end{bmatrix}^T \); \( X = \begin{bmatrix} X_1^T, X_2^T, \ldots, X_m^T \end{bmatrix} \); and \( K_h = \text{diag}(K_1, \ldots, K_n) \) so we get estimation for \( \hat{\beta} \) as given:
\[ \hat{\beta} = (X^T K_h X)^{-1} X^T K_h y \tag{8} \]

One of estimation method on Local Polynomial is using WLS (Weighted Least Square) so we need weighting. One of the weighting used to obtain estimation is the Kernel Function (Eubank, 1988). Kernel function \( K \) with bandwidth \( h \) is defined as follows:
\[ K_h(x) = \frac{1}{h} K\left( \frac{x}{h} \right); \quad -\infty < x < \infty \quad \text{and} \quad h > 0. \]

The properties of kernel function are as follows [5]:
1. Uniform Kernel: \( K(x) = \frac{1}{2}; \quad I(|x| < 1) \)
2. Triangle Kernel: \( K(x) = (1 - |x|); \quad I(|x| < 1) \)
3. Eparichnikov Kernel: \( K(x) = \frac{3}{4}(1 - x^2); \quad I(|x| < 1) \)
4. Squares Kernel: \( K(x) = \frac{15}{16}(1 - x^2)^2; \quad I(|x| < 1) \)
5. Triweight Kernel: \( K(x) = \frac{35}{32}(1 - x^2)^3; \quad I(|x| < 1) \)
6. Cosinus Kernel: \( K(x) = \frac{\pi}{4} \cos \left( \frac{\pi x}{2} \right); \quad I(|x| < 1) \)
7. Gaussian Kernel: \( K(x) = \frac{1}{\sqrt{2\pi}} \exp \left( \frac{1}{2}(-x^2) \right) \)

2.3. Spline Truncated
The famous polynomial piece is spline [6]. Spline is defined as continuous and differentiable polynomial piece. Spline enables to adapt effectively toward data characteristics, so that optimal result can be obtained. The polynomial split points in spline are commonly called knot or knot point. The
knot point is a common fusion point that shows the behavioral changes of the spline function at different intervals. Spline of r order with knots on \( k_1, k_2, \ldots, k_p \) can be defined as a S function with the following form: \[ S(t) = \sum_{i=0}^{r-1} \theta_i t^i + \sum_{j=1}^{p} \delta_j (t-k_j)^{r-1} \] with \( \theta_i \) and \( \delta_j \) are real constants, in which:

\[ (t-k_j)^{r-1}, \text{ if } (t \geq k_j) \]
\[ 0, \text{ if } (t < k_j) \]

if \( r = 2 \), is called a linear spline, if \( r = 3 \), is called a quadratic spline, if \( r = 4 \), is called a cubic spline.

### 2.4. General Cross Validation

Bandwidth is controller between function and data so that the resulting function becomes smooth. Optimal bandwidth selection will obtain a good estimator of model. The selection of bandwidth that is too small will result in a very rough estimate curve (undersmoothing). Meanwhile, if the selection of bandwidth is too large, it will result in a very smooth estimation function curve (oversmoothing). Therefore optimal bandwidth selection is very important in nonparametric regression analysis [8]. One way to determine the optimal bandwidth is by using GCV (Generalized Cross Validation) method. The GCV function is given as follows:

\[ GCV(h) = \frac{MSE(h)}{\left( \frac{1}{n} tr[1 - A(h)] \right)^2} \]

with

\[ MSE(h) = \sum_{i=1}^{n} \left( y_i - \hat{y}_i \right)^2 \]

and \( A(h) \) is obtained from the relationship \( \hat{y} = A(h)y \).

The smallest GCV value will provide an optimal bandwidth \( h \) value.

### 3. Results and Discussion

Given response variable data \( (y_i) \) and predictor variable \( (x_i) \) amounting to \( m \) predictor variables. So it is modeled using regression model as follows:

\[ y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_{m} x_{im} + \varepsilon_i. \]  \[ \text{(10)} \]

The mix or combination approach is done by using local polynomial and spline truncated. The approach can be done with 2 different things. First, the combination approach is based on predictors and residuals. This means the predictor variables is approximated by using local polynomial while the residual is approached by spline, or vice versa. Second, the approach is based only on the predictor variables. This means for the \( j \)-th to \( v \)-th predictor variables use local polynomial, whereas the \( (v+1) \)-th to \( m \)-th predictor variables use spline, or vice versa.

#### 3.1. Predictors and residual based approach

Equation (9) can be written as follows:

\[ y_i = \eta(x_i) + \varepsilon_i, i=1,2,\ldots,l \]  \[ \text{(11)} \]

Where the value of \( \eta(x_i) \) is approached by using a local polynomial, while \( \varepsilon_i \) is approached by using a spline truncated, or vice versa. So if approached by using a local polynomial, then Eq. (10) becomes

\[ y_i = \beta_{01} + (x_{i1} - x)\beta_{11} + (x_{i1} - x)^2 \beta_{21} + \cdots + (x_{i1} - x)^p \beta_{p1} + \beta_{02} + (x_{i2} - x)\beta_{12} + (x_{i2} - x)^2 \beta_{22} \]
\[ + \cdots + (x_i\alpha - x)^p \beta_{p1} + \cdots + \beta_{lm} + (x_m\alpha - x) \beta_{lm} + (x_m\alpha - x)^2 \beta_{2m} + \cdots + (x_m\alpha - x)^p \beta_{pm} + \epsilon_i \]  

Eq. (11) is conversion of local polynomial for each predictor. So if predictor is approached by using local polynomial, each predictor will have an arbited fixed point value, polynomial order and their respective bandwidths. If the equation is written in matrix form \( \mathbf{y} = \mathbf{X}\mathbf{\beta} + \mathbf{\epsilon} \) with \( \mathbf{y} \) measures nx1, while \( \mathbf{\beta} \) measures (mp + 1)x1 while \( \mathbf{X} \) measures nx(mp + 1). Parameter estimation use WLS (Weighted Least Square) with kernel function as its weighting. Based on Eq. (11) we get the value of 

\[ \hat{\mathbf{\beta}} = (\mathbf{X}^T\mathbf{K_h}\mathbf{X})^{-1}\mathbf{X}^T\mathbf{K_h}\mathbf{y} \]  

with \( \mathbf{K_h} \) in nxn size.

The residual value is approached by using the spline truncated, so that in Eq. (10) there should be two functions as follows:

\[ y_i = \eta(x_i) + f(\epsilon_i) + \epsilon \]  

with \( f(\epsilon_i) \) residual value of local polynomial estimation. The curve \( f(\epsilon_i) \) will be approached using a spline truncated and \( \epsilon \) is residual value of the combined model.

\[ f(\epsilon_i) = \sum_{j=0}^{r-1} \theta_j \epsilon^j + \sum_{k=1}^{q} \delta_k (\epsilon - k_q)^{-1} \]  

with \( (\epsilon - k_q)^{-1} = \begin{cases} (\epsilon - k_q)^{-1}, & \text{if } \epsilon \geq k_j \\ 0, & \text{if } \epsilon < k_j \end{cases} \)

So the model becomes

\[ y_i = \eta(x_i) + \sum_{j=1}^{p} \theta_j \epsilon^j + \sum_{k=1}^{q} \delta_k (\epsilon_i - K_k)^p + \epsilon_i \]  

(14)

3.2. The Approach Based On All Predictor Values.

If given the regression model equation as follows:

\[ y_i = \beta_0 + \beta_1 x_i + \beta_2 x_2 + \cdots + \beta_p x_p + \epsilon_i \]  

(15)

So the Eq. (14) can be written

\[ y_i = \eta(x_i) + f(\pi_i) + \epsilon_i \]  

with \( f(\eta_i) \) will be approached by using local polynomial, while \( f(\pi_i) \) will be approached by using spline truncated. In equation (1) the constant value \( \beta_0 \) goes into local polynomial function, while in spline function there is no constant value. This can be done, or the constant value goes into spline function and local polynomial is not a problem. It is written in the following matrix form:

\[ \mathbf{y} = \mathbf{f}(\mathbf{\eta}) + \mathbf{f}(\mathbf{\pi}) + \mathbf{\epsilon} \]  

with

\[ \mathbf{y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}; \quad \text{If} \quad \mathbf{P}_1 = \begin{bmatrix} 1 & (x_{11}-x)^2 & \cdots & (x_{11}-x)^p \\ 1 & (x_{12}-x)^2 & \cdots & (x_{12}-x)^p \\ \vdots & \vdots & \ddots & \vdots \\ 1 & (x_{1n}-x)^2 & \cdots & (x_{1n}-x)^p \end{bmatrix}_{nx(p+1)} \]  

then

\[ f(\eta) = \begin{bmatrix} \mathbf{P}_1 & \mathbf{P}_2 & \cdots & \mathbf{P}_v \end{bmatrix} \]  

with \( v \) indicates the number of predictor variables approached by local polynomial.

If
\[ S_t = \begin{bmatrix} 1 & x_{(v+1)} & x_{(v+1)}^2 & \ldots & x_{(v+1)}^p & (x_{(v+1)} - K_1)^p & (x_{(v+1)} - K_2)^p & \ldots & (x_{(v+1)} - K_p)^p \\ 1 & x_{(v+2)} & x_{(v+2)}^2 & \ldots & x_{(v+2)}^p & (x_{(v+2)} - K_1)^p & (x_{(v+2)} - K_2)^p & \ldots & (x_{(v+2)} - K_p)^p \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \ldots & \vdots \\ 1 & x_{(v+n)} & x_{(v+n)}^2 & \ldots & x_{(v+n)}^p & (x_{(v+n)} - K_1)^p & (x_{(v+n)} - K_2)^p & \ldots & (x_{(v+n)} - K_p)^p \end{bmatrix} \]

with \( p \) is spline degree, while \( K \) is a knot point. Thus the \((v + 1)\)-th predictor to the \( m \)-th predictor is approached by using the spline truncated as follows:

\[
f(\pi) = \left[ S_{(v+1)} \quad S_{(v+2)} \quad \ldots \quad S_m \right]. \tag{16}\]

In combined modeling with all predictor variables, basically the determination of which should first be modeled using local polynomial or spline truncated depends on the researcher. For example, the 1st to \( v \)-th predictor variable form a truncated pattern, whereas the \((v + 1)\)-th to \( m \)-th predictor variable form a localization model, then the spline modeling followed by the local polynomial. The requirement of this method to run, predictor variables must be grouped in advance which forms the truncated and which forms the localization. As for estimation use WLS approach with kernel weighting.

\[
y = A(\eta, \pi)\omega(\beta, \delta, \theta) + \varepsilon \tag{17}\]

with \( A(\eta, \pi) \) is a predictor composition matrix with local polynomial and spline truncated approaches and \( \omega(\beta, \delta, \theta) \) is parameter of local polynomial and spline truncated. Therefore the estimation with WLS obtains:

\[
\hat{\omega}(\beta, \delta, \theta) = (A^T(\eta, \pi)K_{bh}(\eta, \pi)A(\eta, \pi))^{-1}A^T(\eta, \pi)K_{bh}(\eta, \pi)y \tag{18}\]

where \( K_{bh} \) is kernel function matrix as a weighting. The kernel weighting is done not only on the local polynomial approach but also on the spline truncated function. However, if it is the approach of original spline truncated function, then the kernel weighting can be value of one.

While to obtain the optimum bandwidth value with GCV approach requires \( A(h) \) value, where \( A(h) = (A^T(\eta, \pi)K_{bh}(\eta, \pi)A(\eta, \pi))^{-1}A^T(\eta, \pi)K_{bh}(\eta, \pi) \). MSE value is obtained from MSE value of combined models between local polynomial and spline truncated.

4. Conclusion

Approximation using local polynomial only requires one parameter with kernel weighting. As for the spline truncated requires two parameters, but it needs knot points as weighting. The combined modeling between local polynomial and spline truncated yields a fairly complex model. However, it does not guarantee that the combined model is better than the simple model. Parameter estimation of combined model using WLS method, yields estimator formula that is similar to local polynomial parameter estimator or parameter estimator of spline truncated only.

Acknowledgments

This research supported by Ministry of Research Technology and Higher Education Indonesia as a founder in applied of superior research for University in 2017 (first year). Thanks to laboratory of Statistics Department Diponegoro University for supporting this research.

References

[1] Eubank R L, Huang C, Maldonado YM, Wang N, Wang, S, and Buchanan RJ 2004 Smoothing Spline Estimation in Varying-Coefficient Models *J. R. Statist. Soc.* 66 Part 3 653-667.

[2] Bellhouse DR and Stafford JE 2001 Local Polynomial regression in Complex Surveys *Statistics Canada, Catalogue: Survey Methodology* Vol 27 No 2 197-203.
[3] Bryson KMO and Ko M 2004 Exploring The Relationship between Information Technology Investment and Firm Performance Using Regression Spline Analysis Information and Management 42 1-13.

[4] Delaigle A, Fan J, and Carroll RJ 2009 A Design Adaptive Local Polynomial Estimator for the errors in Variable Problem, Journal of the American Association Vol. 104 No 485 348-359.

[5] Welsh AH and Yee TY 2005 Local Regression for Vector Responses Journal of Statistical Planning and Inference 136 3007-3031.

[6] Dette H, and Melas VB 2010 A Note on All-Bias Designs with Applications in Spline Regression Models Journal of Statistical Planning and Inference 140 2037-2045.

[7] Lestari B, Budiantara IN, Sunaryo S, and Mashuri M 2010 Spline Estimator in Multi-Response Nonparametric Regression Model with Unequal Correlation of Errors. Journal of Mathematics and Statistics Vol 6 No 3 327-332.

[8] Wu H and Zhang JT 2006 Nonparametric Regression Methods for Longitudinal Data Analysis A John-Wiley and Sons Inc. Publication New Jersey.

[9] Liu S and Li G, 2015, Varying-coefficient mean-covariance regression analysis for longitudinal data Journal of Statistical planning and inference 160 89-106.

[10] Kong D, Bondell HD and Wu Y 2015 Domain selection for varying coefecient model via local Polynomial regression Computational Statistics and Data Analysis 83 236-250.

[11] Bracco C, Giannell C, and Sestini A 2017 Adaptive scattereddata fitting by extention of local approximations to hierarchical splines Computer aided geometric design 52-53 90-105.

[12] Qiao B, Chen X, Xue X, Luo X and Liu R 2015 The application of Cubic B-spline collocation method in impact force identification Mechanical Systems and signal Processing 64-65 413-427.

[13] Conti C, Gemignani L and Romani L 2016 Exponential pseudo-spline: Looking beyond exponential B-spline Journal of Mathematical Analysis and Applications 439 32-56.