The nonparametric regression model using Fourier series approximation and penalized least squares (PLS) (case on data poverty in East Java)

D R S Saputro¹, A Sukmayanti², P Widyaningsih³
¹, ², ³ Department of Mathematics, Faculty of Mathematics and Natural Sciences, Universitas Sebelas Maret

E-mail: dewiretnoss@staff.uns.ac.id

Abstract. Poverty is a problem all countries in the world, particularly developing countries, should encounter. The poverty level of a country can be a benchmark to see its social and economic conditions. Java is an island with the highest poor population density in Indonesia with the level of population density of 14.83 million people. East Java province had percentage of poverty of 11.85% in 2016 and it was included in hard core poverty (above 10%), indicating that East Java was on the highest level of poverty. Therefore, regression model is required to find out factors significantly influencing poverty. The regression model used is nonparametric regression since the scatter plot does not form certain plot and even tend to have repeated patterns of data on certain time interval. For that reason, Fourier series approximation is used in nonparametric regression. On such approximation, the model parameter can be estimated based on penalized least squares (PLS) and minimum generalized cross validation (GCV). The study investigated and applied the regression model on data of poverty in East Java. The study resulted in parameter estimation of 18 parameters and values of λ of nonparametric regression of Fourier series ranging from 0.05 to 0.1 on M=1 and GCV value of 5.83028 with coefficient of determination ($R^2$) of 83.28.

1. Introduction
Poverty is defined as the inability to fulfill minimum living standards ([1]). It is a problem all countries in the world, particularly developing countries, should encounter. The poverty level of a country can be a benchmark to see its social and economic conditions. For that reason, the poverty eradication becomes significant. If poverty is not immediately overcome, it will bring about the decline in the quality of human life and health ([2]). In 2016, the number of Indonesian people living in poverty reached 27.77 million, or about 12.33% of total population in Indonesia. Java is an island with the highest poor population density (level of population density of 14.83 million people (BPS—The Central Bureau of Statistics [2]) in Indonesia. According to the BPS [3], in 2016, the percentage of poverty in East Java was 11.85%, indicating that East Java had high poverty level (hard core poverty with percentage of above 10%). Thus, further analysis is required to find out factors influencing poverty in East Java.

A method used to find out factors influencing poverty is regression model. It is useful to recognize correlation patterns of response and predictor variables. There are three approximations in regression models: parametric, semi parametric, and nonparametric. Nonparametric regression is applied if the shape of the regression curve of the regression function is unknown ([4]). Several methods of
nonparametric approximation which have been developed include: spline, multivariate adaptive regression splines (MARS), Fourier, Wavelet, and Kernel series. The benefit of nonparametric regression approximation using Fourier series is its ability to encounter data with trigonometric distribution ([5]), in this case sin and cos. The Fourier series approximation in nonparametric regression is used if the shape of data patterns is unknown and the data patterns tend to be periodic; occur in fixed time interval ([6]). Fourier series is a trigonometric polynomial that has flexibility so that it can adjust itself effectively to the local nature of data.

Following are given several studies related to nonparametric regression with Fourier series, [7] comparing Fourier series and spline truncated in multivariable nonparametric regression models in cases of poverty in Papua estimated using least square optimization (LS). In addition, [8] applied a nonparametric regression with Fourier series on rainfall data with parameter estimates LS and [9] examined a combined estimator of Fourier and spline truncated series in multivariable nonparametric regression with estimation of penalized least squares (PLS). Meanwhile,[10] conducted a study of nonparametric regression models with Fourier series whose estimators were obtained through PLS.

Related to the 3 studies and the results of theoretical studies by the same author, [11], in this study a function approach was applied to the nonparametric regression model with Fourier-cosine series on poverty data in East Java. This is considering that in the initial research the patterns of distribution of poverty data in East Java were obtained which tended to be repeated (oscillation) data patterns.

2. Review Theory

2.1 Nonparametric Regression Model using Fourier Series

The Fourier series approximation in nonparametric regression is used if the shape of data patterns investigated is unknown and the data patterns tend to present repeated data (periodic) ([11]). Fourier series are trigonometric polynomials having flexibility to adapt to local characteristics of data ([7]).

Fourier series, the shape of regression curve \( f(x_i) \) is assumed to be unknown and random error \( \epsilon_i \) is independently distributed. According to [3], the regression curve can be approximated by the following function:

\[
    f(x_i) \approx g(x_i) = bx_i + \frac{1}{2}a_0 + \sum_{k=1}^{M} a_k \cos(kx_i)
\]

, where \( i = 1, 2, ..., n, k = 1, 2, ..., M \). \( g(x_i) \) represents regression function \( g \) of predictor variable \( x \) on the \( i^{th} \) observation, while \( b, a_0, a_k \) are model parameters. Function \( f(x_i) \) is the function of Fourier series approximation using trends.

According to [3], multiple nonparametric regression model can be expressed:

\[
    y_{ij} = g_1(x_{1i}) + g_2(x_{2i}) + \cdots + g_j(x_{ji}) + \epsilon_i
\]

or

\[
    y_{ij} = \sum_{j=1}^{p} g_j(x_{ji}) + \epsilon_i \tag{1}
\]

, where \( i = 1, 2, ..., n, y_{ij} \) is the response variable of the \( j^{th} \) predictor variable \( x \) on the \( i^{th} \) observation, \( g_j(x_{ji}) \) denotes regression function \( g \) of the \( j^{th} \) predictor variable \( x \) on the \( i^{th} \) observation, and \( \epsilon_i \) is the residual on the \( i^{th} \) observation.

2.2 Parameter Estimation

The value of parameter estimation of nonparametric regression using Fourier series approximation is obtained by minimizing PLS, i.e. the criterion of approximation which combines goodness of fit and curve smoothness and both are controlled by a smoothing parameter \( (\lambda) \).
The regression function \( g \) continues from interval 0 to \( \pi \), and therefore the measurement of the curve fit towards data is:

\[
n^{-1} \sum_{i=1}^{n} \left( y_{ij} - g_j(x_{ij}) \right)^2
\]

where \( j = 1, 2, ..., p \) is the number of predictor variables, \( n \) is the number of observations, \( y_{ij} \) represents response variable of the \( j^{th} \) predictor variable on the \( i^{th} \) observation, and \( g_j(x_{ij}) \) is the regression function \( g \) of the \( j^{th} \) predictor variable on the \( i^{th} \) observation. The measurement of the curve smoothness is:

\[
\int_{0}^{\pi} \frac{2}{\pi} (g_j^2(x_{ij}))^2 \, dx_{ij}
\]

where \( g_j^2(x_{ij}) \) is the derivation of both regression functions \( g \).

Therefore, the approximation function for regression curve \( g \) obtained by minimizing PLS is obtained.

\[
\min_{g \in C([0,\pi])} \left( n^{-1} \sum_{i=1}^{n} (y_{ij} - g_j(x_{ij}))^2 + \lambda \int_{0}^{\pi} \frac{2}{\pi} (g_j^2(x_{ij}))^2 \, dx_{ij} \right) \tag{2}
\]

\( \lambda \) is the smoothing parameter, where \( \lambda > 0 \). The larger the value of \( \lambda \), the larger the resulted function will be. To minimize the value of PLS, the measurement of the smoothness of regression curve \( g \) should be determined first.

\[
\lambda \int_{0}^{\pi} \frac{2}{\pi} (g_j^2(x_{ij}))^2 \, dx_{ij} = \frac{2}{\pi} \lambda \int_{0}^{\pi} \left[ \frac{d^2}{dx_{ij}} \left( b x_{ij} + \frac{1}{2} a_0 + \sum_{k=1}^{M} a_k \cos(kx_{ij}) \right) \right]^2 \, dx_{ij}
\]

\[
\lambda \sum_{k=1}^{M} k^2 a_k^2.
\]

Such measurement is substituted to equation (2), and therefore the optimization of PLS is obtained.

\[
\min_{b, a_0, ..., a_m \in \mathbb{R}} \left( n^{-1} \sum_{i=1}^{n} (y_{ij} - b x_{ij} - \frac{1}{2} a_0 - \sum_{k=1}^{M} a_k \cos(kx_{ij}))^2 + \lambda \sum_{k=1}^{M} k^2 a_k^2 \right)
\]

It is expressed below in the form of matrix.

\[
\min_{b, a_0, ..., a_m \in \mathbb{R}} \left( n^{-1} (y' y - 2n^{-1} \beta' X' y + \beta'(n^{-1} X^{-1} X + \lambda D) \beta) \right)
\]

Suppose that \( n^{-1} (y' y - 2n^{-1} \beta' X' y + \beta'(n^{-1} X^{-1} X + \lambda D) \beta) = Q(\beta) \) and by deriving \( Q(\beta) \) in partial manner towards \( \beta \) the estimation value for \( \beta \) is obtained:

\[
\hat{\beta}(\lambda) = \frac{\partial Q(\beta)}{\partial \beta} = (n^{-1} X^{-1} X + \lambda D)^{-1} n^{-1} X' y \tag{3}
\]

and therefore the estimator for regression curve \( g \) is

\[
g_{ij}(x_{ij}) = \hat{\beta}(\lambda) x_{ij} + \frac{1}{2} \hat{a}_0(\lambda) + \sum_{k=1}^{M} \hat{a}_k \cos(kx_{ij}) \tag{4}
\]

or

\[
g_{ij}(x_{ij}) = X \hat{\beta}(\lambda) = S(\lambda) y
\]

where \( S(\lambda) = X(n^{-1} X^{-1} X + \lambda D)^{-1} n^{-1} X' \) is \( n \times n \) matrix ([12]).
2.3 The Selection of Optimum Smoothing Parameter (λ)
The thing that should be a concern in nonparametric regression approximation using Fourier series is to determine the optimum value of parameter λ to obtain a good regression curve. The determination of the optimum value of λ using GCV method can be seen from the value of λ which generates minimum GCV value. According to [9], the general formula to calculate GCV value is:

\[ GCV(\lambda) = \frac{\text{RKS}(\lambda)}{(n^{-1}\text{trace}(I - S(\lambda)))^2} \]  

where \( \text{RKS}(\lambda) = n^{-1}y'(I - S(\lambda))(I - S(\lambda))y \) and \( S(\lambda) = X(n^{-1}X^{-1}X + \lambda D)^{-1}n^{-1}X' \).

3. Research Method
This research is an applied research based on data and research steps below.

3.1 Research Data
The data used in the study include secondary data of poverty in 29 regencies and 9 cities in East Java in the year of 2016 obtained from publication of the BPS. Response and predictor variables used in the study cover the percentage of poor population as the response variable \( Y \), human development index \( X_1 \), average years of schooling \( X_2 \), literacy rate \( X_3 \), life expectancy rate \( X_4 \), regency/city-based per capita expenditure \( X_5 \), and school life expectancy rate \( X_6 \).

3.2. Research Procedure
The steps taken in the research are to plot the data and analysis, determine the optimum parameter λ based on the minimum GCV value through equation (3) and the model (4) and equation (5). Determination of the optimum λ and GCV are with several experiments. Next, determining the parameter estimation of the nonparametric regression model with Fourier and PLS series approaches, determining the regression model based on the optimum parameter value λ, and determining the size of the model improvement by determining the coefficient of determination \( R^2 \) and root mean square error (MSE). Each of these sizes is written as

\[ R^2 = \frac{\sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} \] and \[ RMSE = \sqrt{\frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{n}} \]

with \( \bar{y} \) is the average observation data for the percentage of poverty, \( y_i \) is the observation data for the percentage of poverty in the regency, and \( \hat{y}_i \) is the estimated data on the percentage of poverty in the regency and \( i = 1, 2, 3, ..., n \).

4. Results and Discussion
This part discusses the modeling of poverty in East Java using Fourier series approximation and PLS.

4.1. The Modeling of Poverty in East Java Using Fourier Series Approximation.
The first procedure in determining nonparametric regression model using Fourier series approximation was determining the optimum value of parameter λ obtained from minimum GCV value. Seven trials were carried out to determine minimum GCV value, and therefore optimum parameter λ was obtained. The values of parameter λ used include positive real numbers with different values in each trial. The value of parameter M used ranges from 1 to 6 since the greater the value of parameter M, the more damping the resulted curve of Fourier series will be. Out of the seven trials, the minimum GCV value was selected and the number of simple parameter was determined. Based on these experiments the results obtained are shown in Table 1.
Fourier series are shown in Table 2. The parameter estimates of the nonparametric regression model with the full Fourier series approach based on GCV values for 7 range \( \lambda \) and optimum M oscillation parameters are as shown in Table 1.

Based on Table 1, the minimum GCV is selected at the value \( \lambda \) with intervals of 0.005-0.1 and \( M = 1 \) with 18 parameters. The parameter estimates of the nonparametric regression model with the Fourier series are shown in Table 2.

The regression model with the full Fourier series approach based on the model (4) on poverty data in East Java is written as

\[
\hat{Y} = \hat{\beta}_1X_1 + \hat{\beta}_0 - \hat{\beta}_c_1cosX_1 + \hat{\beta}_2X_2 + \hat{\beta}_0 - \hat{\beta}_c_2cosX_2 - \hat{\beta}_3X_3 + \hat{\beta}_0 - \hat{\beta}_c_3cosX_3 + \hat{\beta}_4X_4 + \hat{\beta}_0 - \hat{\beta}_c_4cosX_4 - \hat{\beta}_5X_5 + \hat{\beta}_0 - \hat{\beta}_c_5cosX_5 - \hat{\beta}_6X_6 + \hat{\beta}_0 - \hat{\beta}_c_6cosX_6.
\]

Next, parameter estimation was conducted to model (6) using PLS. The nonparametric regression model using Fourier series approximation on poverty in East Java is expressed:

\[
\hat{Y} = -0.23397X_1 + 11.06751 - 0.01159cosX_1 + 1.11055X_2 + 11.06751 - 0.03825cosX_2 - 0.72614cosX_3 + 11.06751 - 0.00356cosX_4 - 0.31621cosX_5 + 11.06751 - 0.23397cosX_6.
\]

The interpretation of the model is that if the HDI value increases by 1 percent then the percentage of poverty will decrease by 0.23397 and if the average value of school length increases by 1 year, the
percentage of poverty decreases by 1.11055. In addition, if the literacy rate increases by 1 percent, the percentage of poverty decreases by 0.72614 and if the life expectancy increases by 1 year, the percentage of poverty decreases by 0.565584. Furthermore, if the expenditure per capita increases by 1 unit then the percentage of poverty decreases by 0.31621 and if the old school expectation rate increases by 1 year then the percentage of poverty decreases 0.34111.

Based on the results of the partial test of the model parameters, the parameter that has a significant effect on the model is $\hat{\beta}_2$. This indicates that the factors that significantly affected poverty in East Java in 2016 were the average school year ($X_2$). However, other factors that are not significant can be considered in the model.

4.2. Suitability Model.
As stated in the research method, the suitability of the model is shown by 2 measurements, namely the value of $R^2$ and RMSE. The $R^2$ value obtained is 0.8328 or around 83.28% and the RMSE value is 0.1653. This value indicates that the nonparametric regression model with the Fourier series approach is suitable because it is close to 1 or close to 100% and with a small RMSE. In addition, the plot and interpolation of real data in the percentage of poverty with the estimated data of the constructed model shown in Figure 1 appear to have almost the same pattern.

![Figure 1. Interpolation of real percentages of poverty data with estimated model data.](image-url)

5. Conclusion
In reference to the results and discussion, nonparametric regression model using Fourier series approximation is obtained on $M = 1$ and the resulted optimum values of $\lambda$ optimum are $\lambda_1 = 0.1$, $\lambda_2 = 0.1$, $\lambda_3 = 0.1$, $\lambda_4 = 0.05$, $\lambda_5 = 0.1$, $\lambda_6 = 0.1$ with GCV value of 5.83028.

The nonparametric regression model using Fourier series on data of poverty in East Java is expressed below:

$$\hat{Y} = -0.23397 X_1 + 11.06751 - 0.01159 \cos X_1 + 1.11055 X_2 + 11.06751 - 0.03825 \cos X_2$$
$$-0.72614 X_3 + 11.06751 - 0.00356 \cos X_3 + 0.56558 X_4 + 11.06751 - 0.25080 \cos X_4$$
$$-0.31621 X_5 + 11.06751 - 0.02274 \cos X_5 - 0.34111 X_6 + 11.06751 - 0.01606 \cos X_6.$$  

with the coefficient of determination (R2) obtained is 0.832778 or around 83.2778% with RMSE 0.1653. Based on these 2 statistical measures, the model obtained can be used to predict the percentage of poverty in East Java in 2017 and the following year.
References

[1] Kuncoro M 2006 *Ekonomika pembangunan, teori, masalah, dan kebijakan* (4 ed, Yogyakarta: UPP AMP YKPN)

[2] Groce N 2011 Poverty and disability a critical review of the literature in Low and Middle Income Countries, *UCL Paper Series: 16*

[3] BPS [Badan Pusat Statistik] *Persentase penduduk miskin 2016*, Semarang: BPS—The Central Bureau of Statistics of East Java Province

[4] Budiantara I N 2000 *Optimasi dan proyeksi dalam regresi nonparametrik spline*, Bulletin of Mathematics and Natural Science of Universitas Gajah Mada 10 35-44

[5] Eubank RL 1998 *Spline smoothing and nonparametric Regression* (2 ed, New York : Marcel Dekker)

[6] Sukmayanti A, Saputro D R S, and Widyaningsih P 2017 Pendekatan model regresi nonparametrik dengan deret Fourier, Proceeding of National Seminar on Mathematics, Department of Mathematics and Natural Science UNS 107-112

[7] Asrini L J and Budiantara I N 2014 Fourier series semiparametric regression models (Case study: The production of lowland rice irrigation in Central Java), *ARPN Journal of Engineering and Apllied Sciences*, 9 1501-1506

[8] Sudiarso I W, Budiantara I N, Suhartono, and Purnami S W 2015 Combined estimator Fourier series and spline truncated in multivariabel nonparametric regression *Applied Mathematical Science*, No.100 9 4997-5010

[9] Tripena A 2009 Pemilihan parameter penghalus pada estimator deret Fourier dalam regresi nonparametrik, *Jurnal Ilmiah Matematika dan Pendidikan Matematika* Universitas Jenderal Soedirman Purwokerto No.2 1 (2009) 2-16

[10] Wu H and Zhang J T 2006 *Nonparametric regression methods for longitudinal data analysis* (New Jersey: John-Wiley and Sons)

[11] Bilodeau M 1992 Fourier smoother and additive models *The Canadian Journal of Statistic* No.3 20 257-9