IOSR Journal of Mathematics (IOSR-JM)
e-ISSN: 2278-5728, p-ISSN: 2319-765X. Volume 13, Issue I Ver. V (Jan. - Feb. 2017), PP 41-44
www.iosrjournals.org

Smoothing Parameter Estimation of The Generalized Cross-Validation And Generalized Maximum Likehood

[*1]Adams, S. O., [2]Yahaya, H.U. [3]Nasiru O. M.
[123]Department of Statistics, University of Abuja, FCT, Nigeria

Abstract: *Spline Smoothing is a popular method of estimating the functions in a non-parametric regression model, its performance depends on the choice of smoothing parameter. The two most important methods for choosing the spline smoothing parameter are the Generalized Cross-Validation (GCV) and Generalized Maximum Likelihood (GML). A Monte Carlo Method using a program written in R, evaluated these two estimators to compare their performance. The Monte Carlo experiment was designed to see if the asymptotic results in the smooth case were evident in small, medium, and large sample sizes, the Mean Square error (MSE) criterion was used for the comparison. It was discovered that GML was better than GCV because it is stable and works well in all simulations and at all sample size and it does not overfit data when the sample size is small.*

Mathematics Subject Classification: *JSEM -39-94-119*
Keywords*: Nonparametric regression, Smoothing spline, Smoothing parameter, Selection criteria, Generalized cross validation, Generalized Maximum Likehood, Mean bias, Mean square error.*

I. Introduction

Spline Smoothing provides a powerful tool for estimating a nonparametric function (Eubank, 1988; Green and Silverman 1994; Hastie and Tibshirami 1990; Wahba 1990).
The traditional nonlinear regression model fits the model

$$y_i = f\left(\beta, x_i^!\right) + \varepsilon_i \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \quad (1)$$

where β = (β1, ..., βp)$^!$ is a vector of parameters to be estimated, and x_i = (x_1, ..., x_k) is a vector of predictors for the ith of n observations; the errors εi are assumed to be normally and independently distributed with mean 0 and constant variance σ^2. The function f(·), relating the average value of the response y to the predictors, is specified in advance, as it is in a linear regression model.

Generalized Cross-Validation (GCV) estimate method is the most attractive class of the Spline smoothing selection method, which is popular generally for choosing the complexity of statistical models. The basic principle of cross-validation is to leave the data points out one at a time and to choose the value of α under which the missing data points are best predicted by the remainder of the data. GML is a Bayesian model that provides a general framework for the spline smoothing selection methods, it can be used to calculate the posterior confidence intervals of a spline estimate.
In this research work, GML and GCV methods were extended to an ARMA time series observations in the presence of autocorrelation error

II. Literature Review

Many authors have studied modeling of time series data with spline smoothing using Generalized Cross Validation (GCV) and Generalized Maximum Likelihood (GML) method. Kernel Regression estimation using repeated measurement data (Hart, 1986), Regression with autocorrelated errors (Hurvich, 1990), Times series moving average error (Kohn and Wong, 1992), FMRI time series (John, Wahba, Xianhong, Erik and Nordheim, 2002), FMRI time series revisited (Worsley and Friston 1995) and (Diggle and Hutchinson, 1989) extended the GCV method to estimate the smoothing parameter and the autocorrelation parameters simultaneously. (Kohn, Ansley, and Wong 1992) represented a smoothing spline by a state-space model and extended the CV, GCV, and GML methods to an autoregressive moving average error sequence. (Hurvich and Zeger, 1990) used a frequency domain cross-validation method to estimate the smoothing parameter and more recently (Wang, 2012) extended GML, GCV and UBR method to estimate smoothing parameter when data are correlated. Almost all of these methods were developed for time series observations while some others require that the design points are equally spaced.

III. Research methodology

3.1.1 Generalized Maximum Likelihood (GML) estimate method

$$M(\lambda,\tau) = \frac{Z'(Q_2'B(\lambda,\tau)Q_2)^{-1}Z}{\left[\det(Q_2'B(\lambda,\tau)Q_2)^{-1}\right]^{1/(n-m)}}$$

$$= \frac{Y'W(1-A)Y}{\left[\det^{+}(W(1-A))\right]^{1/(n-m)}}$$

Where; $\det^{+}(I-A(\lambda))$ is the product of the n – m nonzero eigenvalues of [I – A(λ)].

3.1.2 Generalized Cross-Validation (GCV) estimate method

$$V(\lambda) = \frac{\frac{1}{n}\|(\nabla - A(\lambda))\|^2}{\left[\frac{1}{n}tr(\nabla - A(\lambda))\right]^2}$$

Where; n = sample size, $tr(\nabla - A(\lambda))$ is the trace of square of the nonzero eigenvalues of $(\nabla - A(\lambda))$

3.2 Source of data

The data used in this research work was simulated to evaluate the performances of the two selection methods i.e. GML and GCV. By using a program coded in R (version 3.2.3) data were generated with sample sizes; 20, 100, 200, 400 and 600. The number of replications was 1000 for each of the samples and each simulated data sets, the mean squared-errors (MSE) was used for evaluation and comparison.

3.3 Equation used for generating values in simulation

A simulation study was conducted to evaluate and compare the performance of the two selection methods presented in previous sections. The model considered is

$$y_t = \frac{Sin\pi_i}{50} + \varepsilon_t \qquad t=1,...,n \;\; and \;\; i=1,...,50$$

Where the $\varepsilon's$ are generated by a first-order autoregressive process AR (1) with mean 0, standard deviation 0.3, and first-order correlation α ,and its 95% Bayesian confidence interval (Wahba, 1983 and Diggle, 1989)

3.4 Experimental design and data generation

The experimental plan applied in this research work was designed to

1. Sample Size(n) of 20, 100, 200, 400 and 600 were considered for the simulation.
2. The following autocorrelation levels were used for the correlations studied (RE) : 0.1, 0.3, 0.5 and 0.8
3. There are 4 x 3 x 5 = 60 combination setting in the design of my simulation experiment.
4. Data were generated for 500 replications of each of the 60 combinations for cases $\sigma's$ and n's.
5. Two RUNS were done for the simulations which were averaged at analysis stage.

Cubic splines (m = 2) were used to fit the mean function for all the sample sizes and σ = 0.3

3.5 Criteria for comparism

Comparison was made to test the performance of these methods in estimating the functions in the presence of autocorrelation error. The simulation study was performed according to R programming code, it was used to estimate all the model parameters, the criteria, the effect of autocorrelation on the estimated parameters and the performances of the two estimation methods i.e. Generalized Maximum Likelihood (GML) and Generalized Crossed Validation (GCV).

Two different criterion functions were used i.e. GML (M) and GCV (V) thus i conducted simulation the functions, M and V. The Evaluation and comparison of the two (2) estimations method were examined using the Mean Square Error (MSE) criterion.

This criterion is written mathematically as

$$MSE\left(\hat{\theta}_i\right) = \frac{1}{n}\sum_{j=1}^{n}\left(\hat{\theta}_{ij} - \theta_i\right)^2$$

IV. Result

Table 1: MSE for the three selection methods of smoothing spline fitted with known first order correlations (α) and standard deviation ($\sigma = 0.3$) for all sample size

N	Smoothing method	α = 0.1	α = 0.3	α = 0.5	α = 0.8	Mean
20	GML	0.485742	0.500942	0.513907	0.594412	0.523751
	GCV	0.487632	0.475321	0.333052	0.117645	0.353413
100	GML	0.181779	0.189986	0.200859	0.213212	0.196459
	GCV	0.483469	0.453259	0.452603	0.44136	0.457673
200	GML	0.019526	0.020982	0.022083	0.0244210	0.016784
	GCV	0.548517	0.532719	0.511272	0.4928517	0.521340
400	GML	0.002498	0.003127	0.004058	0.0049996	0.003671
	GCV	0.554489	0.538209	0.523293	0.5016321	0.529406
600	GML	0.000574	0.000724	0.000897	0.0010033	0.00080
	GCV	0.631338	0.667757	0.680748	0.7012048	0.670262

The table above presents the Mean Square Error (MSE) result of the two spline smoothing selection methods for the sample sizes. It was discovered that for GML, MSE increases as the scale of autocorrelation increases from less (α = 0.1) i.e.0.485742 to high autocorrelation level (α = 0.8) i.e. 0.594412. It was also discovered that the MSE decreases as the sample size increases; for n = 20 MSE decreased from 0.523751 to 0.196459 at n = 100.

For GCV: It was observed that; as the degree of autocorrelation increases the MSE deceases just like the case in bias, i.e. for α = 0.1, MSE is 0.487632 and for α = 0.8, MSE reduces to 0.117645. It was also discovered that the MSE increase as the sample size increases; for n = 20, MSE increased from 0.353413 to 0.457673 when the sample size increased to 100.

The Mean Square Error (MSE) result of the two spline smoothing selection methods for medium sample sizes above shows that, for GML; MSE increases as the scale of autocorrelation increases from less (α = 0.1) i.e.0.019526 to high autocorrelation level (α = 0.8) i.e. 0.022083. It was also discovered that the MSE decreases as the sample size increases; for n = 200, MSE decreased from 0.016784 to 0.003671 at n = 400. For GCV: It was observed that; as the degree of autocorrelation increases the MSE deceases just like the case in bias, i.e. for α = 0.1, MSE is 0.548517 and for α = 0.8, it reduces to 0.4928517. It was also discovered that the MSE increase as the sample size increases; for example, when n = 200, MSE increased from 0.521340 to 0.529406 when the sample size increased to 400.

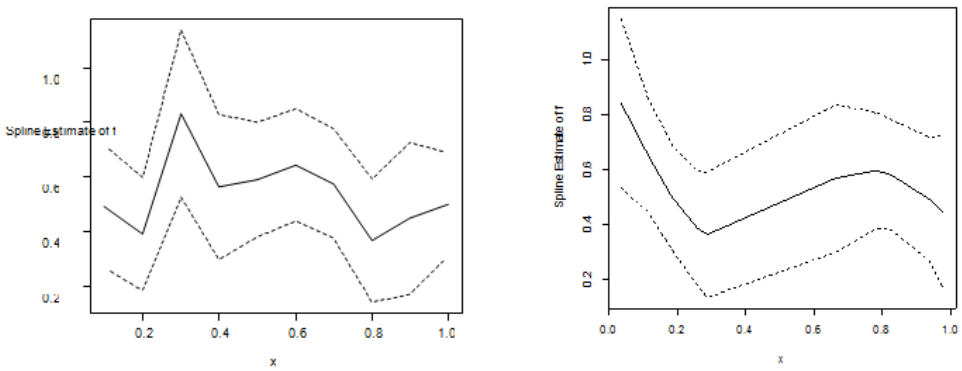

Figure 1: These are plots of the GML and GCV Spline smoothing selection method, the solid curve is the estimates corresponding to the MSE of the simulated study while the two dotted lines are the 95 Bayesian Confidence intervals.

V. Result Discussion

The plots and results presented above indicated that; GML showed signs of efficiency and consistency for all sample sizes i.e. small (n = 20 and 100), medium (n = 200 and 400) and large (n = 600)
The result also showed that, GCV is an inconsistent estimator and it's not an appropriate spline smoothing selection method for all sample sizes when a time series observation possesses an autocorrelation error.

5.1 Effect of autocorrelation under the three criteria

The main effect of autocorrelation was on GCV because their performance was affected by the presence of autocorrelation of all degree i.e. less (0.1), moderate (0.3 and 0.5) and strong (0.8). The result also showed that the performance of the GML was not affected by the presence of autocorrelation, because it worked well at all levels of autocorrelation i.e. (0.1 – 0.8)

It is obvious that the performance of the estimators were affected by autocorrelation under the different sample size. GCV's performance was affected by the presence of autocorrelation at all sample sizes (n = 20 to 600).It reveals that GML estimator is preferred for as a spline smoothing selection method for time series observation at all the levels of autocorrelation.

VI. Conclusion

In all, GML provided better estimates and proved to be most preferred than GCV as a spline smoothing selection method in terms MSE criterion. GML method is computationally more effective and consistence than the GCV selection methods because it worked well for all samples sizes and for all degrees of autocorrelation. GML is most preferred out of the three estimators and is therefore recommended as the best spline smoothing selection method for all sample sizes in the presence of autocorrelation error and for a monte-carlo experiment.

VII. Recommendation

The research work has revealed that GML selection method for spline smoothing is the most preferred estimator in the presence of autocorrelation error based on MSE criterion under the five level of sample sizes considered and the four autocorrelation levels. It can therefore be recommended that when the assumption of mutually independent is abandoned in favour of an autocorrelation error, the most preferred estimator to use is GML.

References

[1]. Diggle, P.J. and Hutchinson, M.F. (1989). On spline smoothing with autocorrelated errors, *Australian Journal of Statistics*, 31: 166 –182.
[2]. Eubank, R. L. (1988). Spline Smoothing and Nonparametric Regression, New *York: Marcel, Dekker, Inc., New York , Basel.*
[3]. Green, P.J. and Silverman, B.W. (1994). Nonparametric regression and generalized linear Models, A roughness penalty approach. *Chapman & Hall, London.*
[4]. Hart, J.D. (1986). Kernel Regression estimation using repeated measurement data, *Journal of American Statistical Association*, 81(396):1080 – 1088.
[5]. Hastie, T. and Tibshirani, R. (1990). Generalized Additive Models, *Chapman and Hall.* London.
[6]. Hurvich, C. M., and Zeger, S. L. (1990). A Frequency Domain Selection Criterion for Regression with Autocorrelated Errors, *Journal of the American Statistical Association, 85*: 705 – 714.
[7]. John, D.C., Wahba G. and Brown M.B. (2000). Spline smoothing for fMRI time series by Generalized Cross-Validation, *NeuroImage*, 18(4): 950 – 961.
[8]. Kohn, R., Ansley, C. F., and Wong, C. (1992), 'Nonparametric Spline Regression with Autoregressive Moving Average Errors,' Biometrika, 79:44 – 50.
[9]. Wahba, G. (1983), 'Bayesian Confidence intervals for the cross-validated smoothing Spline', Journal of Royal. Statistical Society Service. B. 45:133-150.
[10]. Worsley, K.J., and Friston, K.J. (1995). 'Analysis of fMRI time-series revisited again', NeuroImage, 2:173-181.
[11]. Yanrong W. (2012). 'Smoothing Spline Models with Correlated Random Errors. Journal of the American Statistical Association, 93:441, 341–348.