European Journal of Statistics and Probability
Vol.5, No.1, pp.1-8, February 2017
Published by European Centre for Research Training and Development UK (www.eajournals.org)

SMOOTHING SPLINE OF ARMA OBSERVATIONS IN THE PRESENCE OF AUTOCORRELATION ERROR

[*1]Adams, S. O., [2]Ipinyomi, R.A. and [3]Yahaya, H.U
[*13]Department of Statistics, Faculty of Science,University of Abuja, P.M.B. 117, Abuja, FCT. Nigeria, West Africa.
[2]Department of Statistics, University of Ilorin, Kwara State, Nigeria

ABSTRACT: *Given a set of observations $x_1, \ldots x_n$, Spline Smoothing is of great importance in non-parametric regression because it is the fitting of smoothing function to filter out noise in an observation. Many methods of selecting smoothing parameters including; the Cross Validation (CV), Generalized Cross-Validation (GCV), Unbiased Risk (UBR) and Generalized Maximum Likelihood (GML) are developed under the assumption of independent observations. In this study, GML, GCV and UBR methods were extended to ARMA time series observations in the presence of autocorrelation at four levels, i.e. 0.1, 0.3, 0.5 and 0.8. Mean Bias, Mean Square Error and Variance were used to evaluate the performance of the three selection methods. Data were simulated to compare the performances of these three selection methods based on six sample sizes i.e. 20, 60, 200, 350, 500 and 750.GML method was computationally more effective and consistent than the UBR and GCV selection methods because it worked well for all samples sizes and at all levels of autocorrelation.*

KEYWORDS: Non-parametric regression, B-spline, P-Spline, Least Square Spline

INTRODUCTION

Spline Smoothing provides a powerful tool for estimating a nonparametric function, it is a one of the most popular methods used for the prediction of the nonparametric regression models and it is also a method used for fitting smooth curve to a set of noisy observations using a spline function. The role of this method is to estimate the nonparametric function that minimizes penalized least squares criterion. There are many spline smoothing selection methods, some of the methods are; Generalized Maximum Likelihood (GML), Cross Validation (CV), Generalized Cross Validation (GCV), Unbiased Risk (UBR) etc. This research work concentrates on the statistical aspects of nonparametric regression smoothing, two central problems were discussed i.e. the choice of smoothing parameter and the best smoothing method for time series observations with autocorrelation.

Aim and objectives of the study

The aim of this research work is to evaluate and compare the performance of three spline smoothing techniques in non-parametric regression model estimation i.e. the Generalized Maximum Likelihood (GML) and Generalized Cross Validation (GCV) and the Unbiased Risk (UBR). The intention is to make empirical comparison between these three selection methods in order to find out which one is most effective, efficient and consistent in estimating smoothing

ISSN 2055-0154(Print), ISSN 2055-0162(Online)

parameter. Very specifically, the objectives of this study include the following: Examine the effect of autocorrelation (AC) on the performance of the three selection methods i.e. GML, GCV and UBR. To compare the performance of the small, medium and large sample sizes based on the three selection methods. Identify the estimator that is most preferred in the smoothing of a time series data in the presence of autocorrelation error based on the bias, (MSE) and variance as the critical for measuring performance.

LITERATURE REVIEW

Many authors have studied modeling of time series data with spline smoothing using Generalized Cross Validation (GCV), Generalized Maximum Likelihood (GML) and Unbiased Risk (UBR) 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). (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. Yuedong et al (2000) described three methods: Generalized Maximum Likelihood (GML), Generalized Cross Validation (GCV) and leaving-out-one-pair cross validation (CV) to estimate the smoothing parameters, the weighting parameter and the correlation parameter simultaneously. Based on simulated data, they concluded that the GML method has smaller mean-square errors. (Wang, 2012) extended GML, GCV and UBR method to estimate smoothing parameter when data are correlated. (Dursin et al, 2013) recommended the parallel of Akaike's information criterion (GF_{AIC}) and generalized cross-validation (GCV) as being the best selection criteria. For large samples the GF_{AIC} method would seem to be more appropriate while for small samples they proposed the implementation of GCV criterion.

MATERIAL AND METHODS

The general spline smoothing nonparametric regression model is given as;

$$y_i = f\left(\beta, x_i^{!}\right) + \varepsilon_i \text{ ----------equ}(1)$$

Where;

$\beta = (\beta_1, ..., \beta_p)$ is a vector of parameters to be estimated

$x_i = (x_1, ..., x_k)$ is a vector of predictors for the ith of n observations

The (error term) are assumed to be normally and independently distributed with mean 0 and constant variance σ^2

$f(\cdot)$ is a smooth, continuous function are to be estimated from the data

Generalized Maximum Likelihood (GML) selection method

A Bayesian model provides a general framework for the GML method and can be used to calculate the posterior confidence intervals of a spline estimate.

- The GML estimates and $\hat{\tau}$ of λ and τ are the minimizers of

ISSN 2055-0154(Print), ISSN 2055-0162(Online)

$$M(\lambda,\tau)=\frac{Z'\left(Q_2'B(\lambda,\tau)Q_2\right)^{-1}Z}{\left[\det\left(Q_2'B(\lambda,\tau)Q_2\right)^{-1}\right]^{1/(n-m)}}\text{----------}equ(2)$$

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

Generalized Cross-Validation (GCV) selection method

The generalized cross-validation (GCV) selection method is given as;

$$GCV=\frac{\frac{1}{n}\|(1-A\lambda)y\|^2}{\left[\frac{1}{n}Trace(1-A\lambda)\right]^2}\text{----------}equ\ (4)$$

Where;

n = sample size

$$\lambda = smoother\ parameter$$

$$A(\lambda)=X\left(X^TX+n\lambda I\right)^{-1}X^T\ i.e. the\ diagonal\ element\ of\ the\ smoother\ matrix$$

Unbiased Risk selection method

The UBR method has been successfully used to select smoothing parameters for spline estimates with non-Gaussian data (Gu 1992; Wahba, Wang, Gu, Klein, and Klein 1995). It can be developed by applying the Weighted Mean Square Errors (Wang, 1998). The Unbiased Risk is therefore given as;

$$V_k(\lambda,\tau)=\frac{\frac{1}{n}\left\|W^{\frac{k}{2}}(I-A)y\right\|^2}{\left[\frac{1}{n}tr\left(W^{k-1}(I-A)\right)\right]^2}\text{----------}equ(5)$$

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;

$$y_t=\frac{Sin\pi_t}{n}+\varepsilon_t\text{----------}equ(6)$$

t = 1 . . . 100

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

Experimental design and data generation

A Monte carlo experimental design was carried out in this research work to evaluate the performances of the three selection methods i.e. GML, GCV and UBR using program coded in R (version 3.2.3)

1. Sample Size(n) of 20, 100, 200, 350, 500 and 750 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 the simulation experiment.

ISSN 2055-0154(Print), ISSN 2055-0162(Online)

4. Data were generated for 1000 replications of each of the 60 combinations for cases and n's.
5. The Bias, Mean Squared Errors (MSE) and Variance were the criteria used for evaluation and comparison.

Criteria for comparison

The Evaluation and comparison of the three (3) estimations method were examined using the finite sampling properties of estimators which are; Mean Square Error (MSE), Mean Bias, and Variance; criteria.

$$(i)\ Mean\ Bias\left(\hat{\theta}_i\right)=\frac{1}{n}\sum_{j=1}^{n}\left(\hat{\theta}_{ij}-\theta_j\right)\text{----------}equ(7)$$

$$(ii)\ MSE\left(\hat{\theta}_i\right)=\frac{1}{n}\sum_{j=1}^{n}\left(\hat{\theta}_{ij}-\theta_i\right)^2\text{----------}equ(8)$$

$$(iii)\ Var\left(\hat{\theta}_i\right)=\frac{1}{n}\sum_{j=1}^{n}\left(\hat{\theta}_{ij}-\bar{\hat{\theta}}_i\right)^2\text{----------}equ(9)$$

RESULT

Table 1: Bias result for the three selection methods of smoothing spline fitted with known first order correlations ( ) and standard deviation (σ = 0.3) for all sample size

N	Smoothing method	α = 0.1	α = 0.3	α = 0.5	α = 0.8	Mean
20	GML	0.577283	0.587909	0.588120	0.588482	0.585449
	GCV	0.014678	0.014256	0.013325	0.012839	0.013775
	UBR	0.566807	0.571591	0.573949	0.574351	0.571675
60	GML	0.494015	0.510958	0.526202	0.537691	0.517217
	GCV	0.084238	0.077236	0.072823	0.041794	0.069023
	UBR	0.508893	0.541234	0.545666	0.551108	0.536725
200	GML	0.300254	0.320644	0.310182	0.324417	0.313874
	GCV	0.271588	0.267005	0.257122	0.254669	0.262595
	UBR	0.354417	0.389064	0.390182	0.390254	0.380976
350	GML	0.124865	0.128390	0.130109	0.130895	0.128565
	GCV	0.3372093	0.358802	0.364319	0.344925	0.351314
	UBR	0.2423359	0.250660	0.266807	0.271587	0.257847
500	GML	0.013039	0.013172	0.013234	0.013476	0.0132302
	GCV	0.537662	0.528897	0.507521	0.459748	0.5084571
	UBR	0.233451	0.259953	0.269522	0.273292	0.2590545
750	GML	0.003633	0.0018117	0.001975	0.002047	0.0023667
	GCV	0.569467	0.553997	0.548991	0.511735	0.5460480
	UBR	0.302031	0.309301	0.311589	0.318134	0.3102635

The table above presents the Bias of the three estimators for all the sample sizes. It was discovered that for GML, GCV and UBR the mean bias increases as the scale of autocorrelation increases

ISSN 2055-0154(Print), ISSN 2055-0162(Online)

from 0.577283 when α = 0.1 to 0.5884820 when α = 0.8 i.e. It was also discovered that the bias decreases as the sample size increases; for n = 20 the bias decreased from 0.584449 to 0.517217 at n = 60. The main effect of autocorrelation was on GCV, it can be inferred that the performances of the GML and UBR is not affected by medium sample size nor autocorrelation. Thus, GML and UBR can be used as an appropriate spline smoothing method for medium sample size.

Table 2: MSE for the three selection methods of smoothing spline fitted with known first order correlations ( ) and standard deviation (σ = 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
	UBR	0.661073	0.709401	0.766523	0.841748	0.744686
60	GML	0.181779	0.189986	0.200859	0.213212	0.196459
	GCV	0.483469	0.453259	0.452603	0.44136	0.457673
	UBR	0.188423	0.208996	0.229928	0.257236	0.221146
200	GML	0.019526	0.020982	0.022083	0.0244210	0.016784
	GCV	0.548517	0.532719	0.511272	0.4928517	0.521340
	UBR	0.557992	0.581217	0.622587	0.645689	0.601871
350	GML	0.002498	0.003127	0.004058	0.0049996	0.003671
	GCV	0.554489	0.538209	0.523293	0.5016321	0.529406
	UBR	0.631964	0.638884	0.705325	0.7396382	0.678953
500	GML	0.000574	0.000724	0.000897	0.0010033	0.00080
	GCV	0.631338	0.667757	0.680748	0.7012048	0.670262
	UBR	0.729481	0.741138	0.753572	0.7734672	0.749415
750	GML	0.0000648	0.0001855	0.000197	0.0002034	0.000163
	GCV	0.7294816	0.7358454	0.775713	0.8074854	0.762131
	UBR	0.8074854	0.8274342	0.870173	0.9017735	0.851717

The table above presents the Mean Square Error (MSE) result of the three spline smoothing selection methods for small 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 = 60. 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, bias 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 60.

ISSN 2055-0154(Print), ISSN 2055-0162(Online)

Table 3: Variance result for the three selection methods of smoothing spline fitted with known first order correlations ( ) and standard deviation (σ = 0.3) for all sample size

N	Smoothing method	α = 0.1	α = 0.3	α = 0.5	α = 0.8	Mean
20	GML	0.672943	0.701755	0.818362	0.860062	0.763281
	GCV	0.580553	0.742472	0.971773	1.163785	0.864646
	UBR	0.873948	1.151315	1.538486	1.828795	1.348136
60	GML	0.421131	0.555184	0.868063	1.066889	0.727817
	GCV	0.700402	0.789726	0.824556	0.868568	0.795813
	UBR	0.399591	0.666253	0.857896	1.114419	0.759541
200	GML	0.419038	0.467496	0.517099	0.556339	0.489993
	GCV	0.581751	0.614757	0.647253	0.671889	0.628913
	UBR	0.768259	0.822937	0.863854	0.900431	0.838872
350	GML	0.190483	0.200868	0.224191	0.252616	0.21704
	GCV	0.404532	0.438309	0.452731	0.474389	0.44249
	UBR	0.427856	0.474889	0.494637	0.514333	0.477929
500	GML	0.015243	0.024416	0.043959	0.051107	0.033681
	GCV	0.007154	0.031891	0.054221	0.079384	0.043163
	UBR	0.025983	0.046689	0.074342	0.095972	0.060747
750	GML	0.009125	0.020359	0.024552	0.035482	0.022384
	GCV	0.190943	0.214454	0.247782	0.276889	0.232515
	UBR	0.019102	0.025005	0.044671	0.051345	0.035031

The variance of the three spline smoothing selection methods for small sample sizes presented above shows that, for GML; The variance increases as the scale of autocorrelation increases from less (α = 0.1) i.e.0.672943 to high autocorrelation level (α = 0.8) i.e. 0.860062. It was also discovered that as the sample size increases the variance decreases, for example when n = 20, variance decreased from 0.672943 to 0.431131 when sample size increased to 60. It was observed that; as the degree of autocorrelation increases, variance also increased, i.e. for α = 0.1, variance was 0.580553 and for α = 0.8, it increased to 1.163785.

ISSN 2055-0154(Print), ISSN 2055-0162(Online)

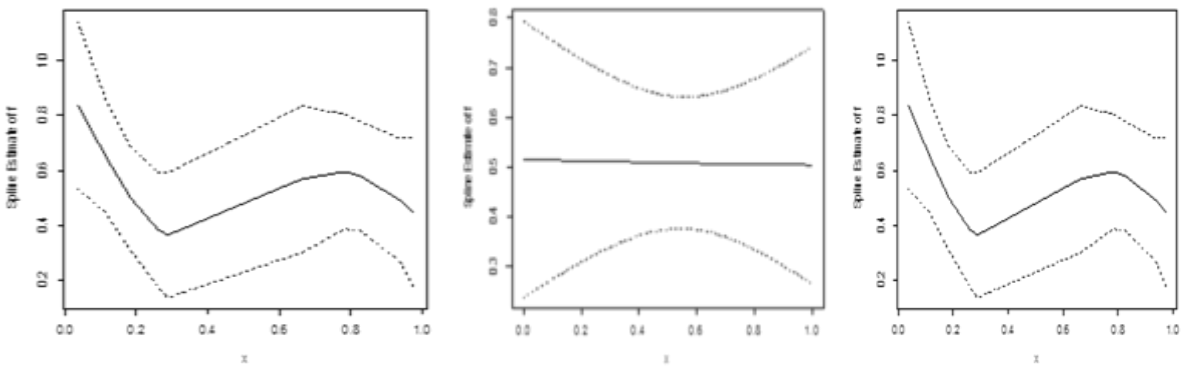

Figure 1: The plot of the GML, GCV and UBR, the solid curve is the estimates corresponding to the MSE of the simulated study while the two dotted lines are the 95 Bayesian Confidence interval.

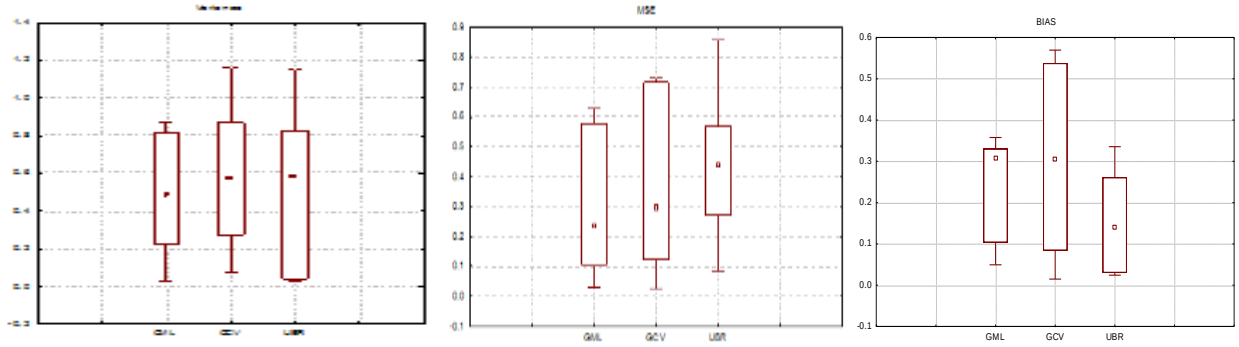

Figure 2: The boxplots of GML, GCV and UBR spline smoothing selection method for the three comparison criteria

Figure 2, presents the boxplot from left to right are: estimates of GML, estimates GCV, and UBR. From these plots we can see that the GML and UBR estimates have small MSE, bias and variances. The GML estimate of have smallest MSE, UBR estimates has an average bias, MSE and variance while GCV estimates had the larger MSE, biases and variances. From the boxplot and plots of the estimated functions, it can be concluded that two out of the three methods estimate the smoothing parameters and the functions very well.

CONCLUSION

In all, GML and UBR provided better estimates and proved to be most preferred than GCV as a spline smoothing selection method in terms of the three criteria. GML method is computationally more effective and consistence than the UBR and GCV selection methods because it worked well for all samples sizes and for all degrees of autocorrelation.

ISSN 2055-0154(Print), ISSN 2055-0162(Online)

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.

REFERENCE

Diggle, P.J. and Hutchinson, M.F. (1989). On spline smoothing with autocorrelated errors, *Australian Journal of Statistics*, 31: 166 –182.

Eubank, R. L. (1988). Spline Smoothing and Nonparametric Regression, New *York: Marcel, Dekker, Inc., New York , Basel.*

Green, P.J. and Silverman, B.W. (1994). Nonparametric regression and generalized linear Models, A roughness penalty approach. *Chapman & Hall, London.*

Hart, J.D. (1986). Kernel Regression estimation using repeated measurement data, *Journal of American Statistical Association*, 81(396):1080 – 1088.

Hastie, T. and Tibshirani, R. (1990). Generalized Additive Models, *Chapman and Hall.* London.

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.

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.

Kohn, R., Ansley, C. F., and Wong, C. (1992), 'Nonparametric Spline Regression with Autoregressive Moving Average Errors,' Biometrika, 79:44 – 50.

Wahba, G. (1983), 'Bayesian Confidence intervals for the cross-validated smoothing Spline', Journal of Royal. Statistical Society Service. B. 45:133-150.

Worsley, K.J., and Friston, K.J. (1995). 'Analysis of fMRI time-series revisited again', NeuroImage, 2:173-181.

Yanrong W. (2012). 'Smoothing Spline Models with Correlated Random Errors. Journal of the American Statistical Association, 93:441, 341–348.

ISSN 2055-0154(Print), ISSN 2055-0162(Online)