

A Bayesian Approach for Penalized Splines with Hierarchical Penalty

Anne Wanjira Ndung'u¹, Samuel Musili Mwalili², Leo Odongo³

¹Pan African University, Nairobi, Kenya

²Department of Mathematics, Statistics and Actuarial Sciences, JKUAT University, Nairobi, Kenya

³Department of Mathematics and Actuarial Sciences, Kenyatta University, Nairobi, Kenya

Email: annewanjira@gmail.com, samuel.mwalili@gmail.com, leo_odongo.co.uk

How to cite this paper: Ndung'u, A.W., Mwalili, S.M. and Odongo, L. (2022) A Bayesian Approach for Penalized Splines with Hierarchical Penalty. *Open Journal of Statistics*, 12, 623-633.
<https://doi.org/10.4236/ojs.2022.125037>

Received: July 26, 2022

Accepted: October 11, 2022

Published: October 14, 2022

Copyright © 2022 by author(s) and Scientific Research Publishing Inc.

This work is licensed under the Creative Commons Attribution International License (CC BY 4.0).

<http://creativecommons.org/licenses/by/4.0/>



Open Access

Abstract

Penalized spline has largely been applied in many research studies not limited to disease modeling and epidemiology. However, due to spatial heterogeneity of the data because different smoothing parameter leads to different amount of smoothing in different regions the penalized spline has not been exclusively appropriate to fit the data. The study assessed the properties of penalized spline hierarchical model; the hierarchy penalty improves the fit as well as the accuracy of inference. The simulation demonstrates the potential benefits of using the hierarchical penalty, which is obtained by modelling the global smoothing parameter as another spline. The results showed that mixed model with penalized hierarchical penalty had a better fit than the mixed model without hierarchy this was demonstrated by the rapid convergence of the model posterior parameters and the smallest DIC value of the model. Therefore hierarchical model with fifteen sub-knots provides a better fit of the data.

Keywords

Penalized Splines, Mixed Model, Hierarchical Penalty

1. Introduction

Penalized splines have become a popular nonparametric smoothing technique [1] [2] using a global smoothing parameter in order to control the amount of smoothing in a function. Flexible methods that allow the smoothing parameter to vary over the domain as a penalty function of independent variables have been proposed [3] [4]. Adaptive smoothing method has also become increasingly important in statistics and it involves allowing the smoothing parameter, the bandwidth or the placement of knots to vary across the domain, adapting to the

change of roughness [5]-[14]. Modelling the smoothing parameter as a penalty function of independent variable can also be used to achieve adaptiveness. This involves formulating the adaptive smoothing as a minimization problem with a new penalty function in which the estimate has the same form as the smoothing spline and method developed for classical smoothing splines can be used. [4] derived the reproducing kernels for a generic penalty function and suggested modelling it by B-splines. [3] studied the solution of the penalized least square estimate in which the smoothing parameter is a varying function across the domain under the Reproducing Kernel Hilbert Space approach. [15] proposed to model the penalty function by a step function where the segmentation is data driven and estimate it by maximizing the generalized likelihood. A complexity penalty was added to the generalized likelihood in selecting the best step function from a collection of candidate. This approach was very computational expensive due to the large number of candidate models and proposed search algorithm and thus has a serious limitation. In this research, we aim at developing a Hierarchical penalty model using p-splines which will result in more adaptive smoothing. In penalized regression splines, [2] modelled the penalty function by a linear interpolation on the logarithmic scale, [16] modelled the penalty function from full Bayesian approach and used Markov chain Monte Carlo for computation, [17] developed a fast and simple algorithm for the Bayesian p-spline based on Laplace approximation for the marginal likelihood. To accommodate a flexible class of functions, a number of works have been constructed through the smoothing spline estimator of the non-parametric distribution function $m(\cdot)$. These include [18] [19] [20]. [4] [21] considered a mixed-effects model with a fixed smoothing parameter.

The study's primary objective was to assess the properties of the mixed model with penalized hierarchical penalty developed in [22] under Bayesian approach by simulation study and compared it with the mixed model without penalized hierarchical penalty. Therefore, the study's main contribution was to investigate the developed model fit performance when hierarchy penalty is added to mixed model.

The paper is organized as follows: Section 2 describes the methods of the study; Section 3 presents the main results, simulation study results, and Section 4 presents conclusion and suggestions for further research.

2. Methods

In applied statistics, data normally plays an important role. However it can have different and complex shape that will result in parametric modeling being unfavorable. Non parametric modeling provides a solution to this in that the shape of the functional relationships is not determined before but is driven by the data. This means that it can adjust to capture unusual or unexpected features in the data. Penalized spline is a non parametric regression that is commonly used for scatter plot smoothing. Suppose we consider smoothing a scatter plot where the

data are denoted (x_i, y_i) for $i = 1, \dots, n$. The underlying trend would be a function m where $\beta(x)$ is the vector of known spline basis functions and β is the vector of unknown coefficients. P-spline is low order basis spline with a penalty to avoid under smoothing. Such splines are typically not spatially adaptive and hence have trouble when function is varying rapidly.

2.1. Bayesian Model

Bayesian approach allows the use of objective data or subjective opinion in specifying a prior distribution. Bayesian inference is the process of fitting a probability model to a set of data and summarizing the result by a probability distribution on the parameters of the model and on unobserved quantities as predictions for new observations. It is a process of using observed data to infer properties of the statistical distributions that generated that data. Given data (x_i, y_i) where x_i is univariate, our parametric model is defined by

$$Y_i = M(x_i) + \varepsilon_i$$

where $M(\cdot)$ is unknown function, the ε_i 's are independent conditional on x_i and normally distributed with mean zero and variance σ_ε^2 . To estimate $M(\cdot)$ we use regression P-splines. The basis function will be piecewise polynomial function whose highest order derivative take jumps at fixed "knots".

$$M(x) = \beta_0 + \beta_1 x + \dots + \beta_p x^q + \sum_{k=1}^r \rho_k (x - \tau_k)_+^q$$

where $\beta_0, \beta_1, \dots, \beta_p, \rho_1, \dots, \rho_r$ is a vector of regression coefficients and τ_1, \dots, τ_r are fixed knots.

To model the unknown function $M(\cdot)$ we can use regression spline of any degree but here the study used degree 2 for convenience so that,

$$M(x) = \beta_0 + \beta_1 x + \sum_{k=1}^r \rho_k (x - \tau_k)_+^q \quad (1)$$

r the number of knots is taken to be large but much less than the number of data points n . We use knots that are equally spaced sample quantiles of X . The number of knots are specified by the user although the choice is not crucial [2] a minimum number of knots is needed to capture the spatial variability in the data.

Equation (1) can be interpreted as a Bayesian linear model and rewritten as

$$Y = R_k P_k + \varepsilon_k \quad (2)$$

where $Y_{n \times 1} = (Y_1, \dots, Y_n)^T$, $R_k = (\beta_0, \beta_1, \rho_1, \dots, \rho_r)^T$ is $(r+2) \times 1$ vector of regression coefficients. $\varepsilon_k = (\varepsilon_1, \dots, \varepsilon_n)^T$ is $n \times 1$ error vector and the design matrix P_k is defined as

$$P_k = \begin{bmatrix} 1 & x_1 & (x_1 - \tau_1)_+ & \dots & (x_1 - \tau_r)_+ \\ 1 & x_2 & (x_2 - \tau_1)_+ & \dots & (x_2 - \tau_r)_+ \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_n & (x_n - \tau_1)_+ & \dots & (x_n - \tau_r)_+ \end{bmatrix}$$

Suppose $\varepsilon_1, \dots, \varepsilon_n$ are independent and identically distributed $N(0, \sigma_k^2)$. The terms involving (β_0, β_1) can be considered as fixed effects in the model. Normal prior on (β_0, β_1) with mean 0 and variance 100 was assumed. This effectively acts as a non informative uniform prior on the fixed effects. The random variables in the vector of spline coefficient are assumed a priori independent and normally distributed $\rho_k \sim N(0, \sigma_\rho^2(\tau_k))$ where $k = 1, 2, \dots, r$. $\sigma_\rho^2(\tau_k)$ is the smoothing parameter that shrinks all the coefficient of the spline hence a global smoothing parameter. The $\sigma_\rho^2(\tau_k)$ is all constant as a function of k that the smoothing is not spatially adaptive. In order to develop a spatially adaptive technique we need to model $\sigma_\rho^2(\tau_k)$ so as to capture the spatial heterogeneity of the data because different smoothing parameter leads to different amount of smoothing in different regions.

The study developed a Hierarchical model for $\sigma_\rho^2(\tau_k)$ where $\sigma_\rho^2(\cdot)$ is a function evaluated at the knots (τ_k) . By taking the functional form of $\sigma_\rho^2(\cdot)$ as another linear regression *i.e.*

$$\sigma_\rho^2(\tau_k) = \exp \left[\gamma_0 + \gamma_1 \tau^p + \sum_{j=1}^l h_j (\tau - \alpha_j)_+^q \right] \quad (3)$$

which is rewritten as

$$-\log \{ \sigma_\rho^2(\tau_k) \} = \gamma_0 + \gamma_1 \tau^p + \sum_{j=1}^l h_j (\tau - \alpha_j)_+^q \quad (4)$$

where $\alpha_1, \dots, \alpha_l$ are fixed knots. The number of sub knots l is user specified and typically less than r (the number of knots in the original spline). The knots $\{ \alpha_j \}_{j=1}^l$ are also taken to be equally spaced quantiles of τ . Rewriting Equation (4) as a Bayesian linear model we have

$$D = R_h P_h \quad (5)$$

where $D = [-\log \{ \sigma_\rho^2(\tau_1) \}, \dots, -\log \{ \sigma_\rho^2(\tau_r) \}]^T$, $P_h = (\gamma_0, \gamma_1, h_1, \dots, h_l)^T$ is an $(l+2) \times 1$ vector and R_h is the design matrix given by

$$P_h = \begin{bmatrix} 1 & \tau_1 & (\tau_1 - \alpha_1)_+ & \cdots & (\tau_1 - \alpha_l)_+ \\ 1 & \tau_2 & (\tau_2 - \alpha_1)_+ & \cdots & (\tau_2 - \alpha_l)_+ \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & \tau_r & (\tau_r - \alpha_1)_+ & \cdots & (\tau_r - \alpha_l)_+ \end{bmatrix}$$

The random variables in the Equation (4) are also assumed to be priori independent and normally distributed *i.e.* $h_j \sim N(0, \sigma_h^2)$ where $j = 1, \dots, l$ and the parameters (γ_0, γ_1) are also independent and normally distributed with mean zero and large variance.

2.2. MCMC Methods

MCMC has become a very important computational tool in Bayesian statistics, since it allows for Monte Carlo approximation of complex posterior distributions where analytical or numerical integration techniques are not applicable.

These methods are commonly used in exploration of joint posterior distribution. A Markov chain is basically a sequence of random variables generated from a transition distribution $q(\cdot)$ constructed such that the distribution of the next random variables to be generated depends only on the current state of the chain [23]. The aim of Bayesian analysis is to construct a Markov chain that converges to a stationary distribution which is the target posterior distribution. The Markov chain is then simulated until the point where the simulated draws resembled draws from the target distribution.

2.3. Model Diagnostics

The models were compared using the Deviance Information Criterion (DIC). The best fitting model is one with the smallest DIC given by $DIC = \bar{D}\theta + pD$, in which \bar{D} is the posterior mean of the deviance that measures the goodness of fit, and pD gives the effective number of parameters in the model which penalizes for complexity of the model. Lower values of \bar{D} indicate a better fit while small values of pD indicate a parsimonious model. To compare two models, it is observed how big the difference the Deviance Information Criterion (DIC) values of the models need to be so that one can declare that one model is better than the other one is not clearly cut. Studies have described that a difference of 3 in model DIC between two models is difficult to distinguish but a DIC difference greater than 3 can be loosely distinguished [24].

3. Main Results

Simulation Study

The model Equation (1) was fitted in Bayesian approach, and simulation was done using WinBUGS, four chains were run with a burn-in of 1000 iterations thinning to every 1 step (in order to reduce the auto-correlation of the chain) until we obtained a sample of 16,000. The mean function $m(\cdot)$ for the model (5) was simplified to

$$m(x, \theta) = \beta_0 + \beta_1 x + \sum_{k=1}^r \rho_k |x - \tau_k|^2 \quad (6)$$

where $\theta = (\beta_0, \beta_1, \rho_1, \dots, \rho_r)^\top$ is the vector of regression coefficients, and $\tau_1 < \tau_2 < \dots < \tau_r$ are fixed knots. The spline degree $q = 2$, number of knots as $p = 20$, sub-knots $t = 10, 15, 17$, sample sizes $N = 50, 100$ and 500 , $x \sim Unif(10, 100)$, $y \sim Unif(200, 300)$ and the following priors distribution were assumed $\beta_0, \beta_1 \sim N(0, 1 \times 10^{-6})$, $\rho_k \sim N(0, \tau_\rho^2)$, $\tau_k^2 \sim IG(0.01, 0.01)$. 4 chains were run for MCMC. To prevent over-fitting, then we minimize

$$\sum_{i=1}^n \{y_i - m(x_i, \theta)\}^2 + \frac{1}{\rho} \theta^\top P \theta \quad (7)$$

where ρ is the smoothing parameter and P is a known positive semi-definite penalty matrix.

$$P_k = \begin{bmatrix} 1 & x_1 & (x_1 - \tau_1)_+ & \cdots & (x_1 - \tau_r)_+ \\ 1 & x_2 & (x_2 - \tau_1)_+ & \cdots & (x_2 - \tau_r)_+ \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_n & (x_n - \tau_1)_+ & \cdots & (x_n - \tau_r)_+ \end{bmatrix}$$

The posterior distribution of the parameters for the model is given by

$$P_{post}(\beta, \rho, \tau^2 | y) = L(y | \beta, \rho, \tau^2) P_{prior}(\beta, \rho, \tau^2)$$

For each model run, 5000 Markov chain Monte Carlo (MCMC) iterations were ran, with the initial 1000 discarded to cater for the burn-in period and thereafter keeping every tenth sample value. The iterations left were used for assessing convergence of the MCMC and parameter estimation. We assessed MCMC convergence of all models parameters by checking trace plots of the MCMC output.

Table 1 presents model diagnostics for the fitted models having hierarchy penalty and without hierarchy penalty. The results showed that for Non-hierarchical model (2) with mean function (6), M_1 had a smallest Deviance Information Criterion (DIC) 493.216 compared to M_2 with DIC = 963.612 and M_3 with DIC = 4785.9 when the sample size taken was 50, however as the sample size increases the DIC become large. M_1 provide a better fit. The credible interval for the mean function $M(x)$ given the sample $N = 50$ was significantly large.

For the hierarchical model represented by the Equation (5), **Table 1** shows that the model with hierarchy penalty having 15 sub-knots had the lowest DIC = 196.209 compared to the model with 10 sub-knots (DIC = 202.321) and 17 sub-knots (DIC = 207.852), this indicate that the model with hierarchy penalty and 15 sub-knots provides a better fit for the data. The credible interval for the hierarchical regression function $M()$ was 243.5 - 289.0 (**Table 2**).

Table 1. DIC for models with and without the hierarchy penalty with sizes 50, 100, 500.

	Model Function without hierarchy Penalty			Model Function with Hierarchy Penalty		
	M_1	M_2	M_3	$M_1 (t = 10)$	$M_2 (t = 15)$	$M_3 (t = 17)$
p^D	7.628	7.628	9.494	7.805	7.607	6.981
\bar{D}	485.587	955.82	4776.410	194.516	188.602	200.871
DIC	493.216	963.612	4785.90	202.321	196.209	207.852

Table 2. 95% credible intervals for the model mean functions for models with and without the hierarchy penalty with sizes 50, 100, 500.

	Model Function without hierarchy Penalty			Model Function with Hierarchy Penalty		
	M_1	M_2	M_3	$M_1 (t = 10)$	$M_2 (t = 15)$	$M_3 (t = 17)$
CI	246.5 - 268.5	250.2 - 268.9	248.2 - 254.8	237.6 - 294.4	243.5 - 289.0	227.7 - 281.2

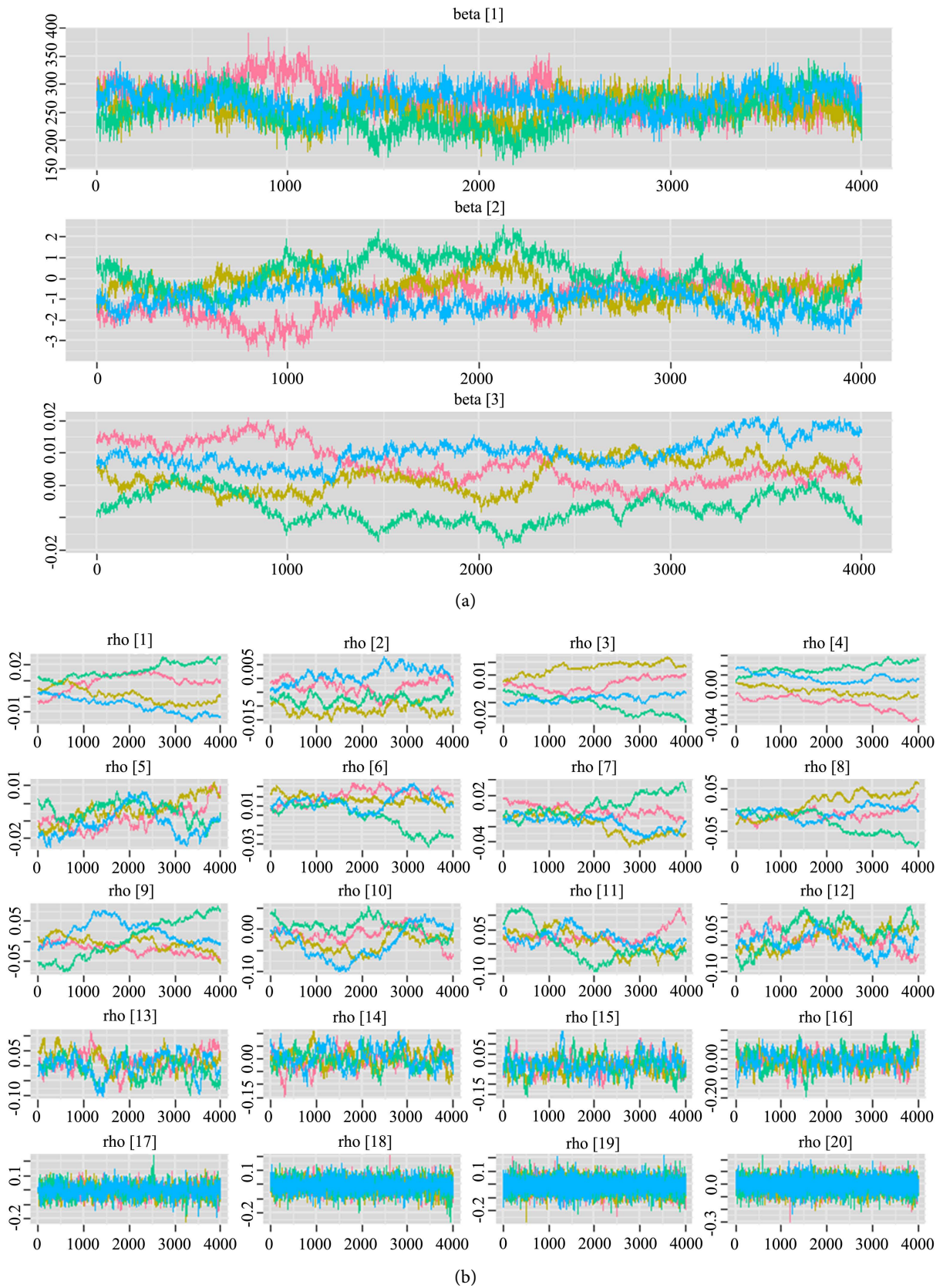


Figure 1. Traceplots plots for model parameters. (a) Beta parameter in Non-Hierarchical model; (b) Rho parameter in Non-Hierarchical model.

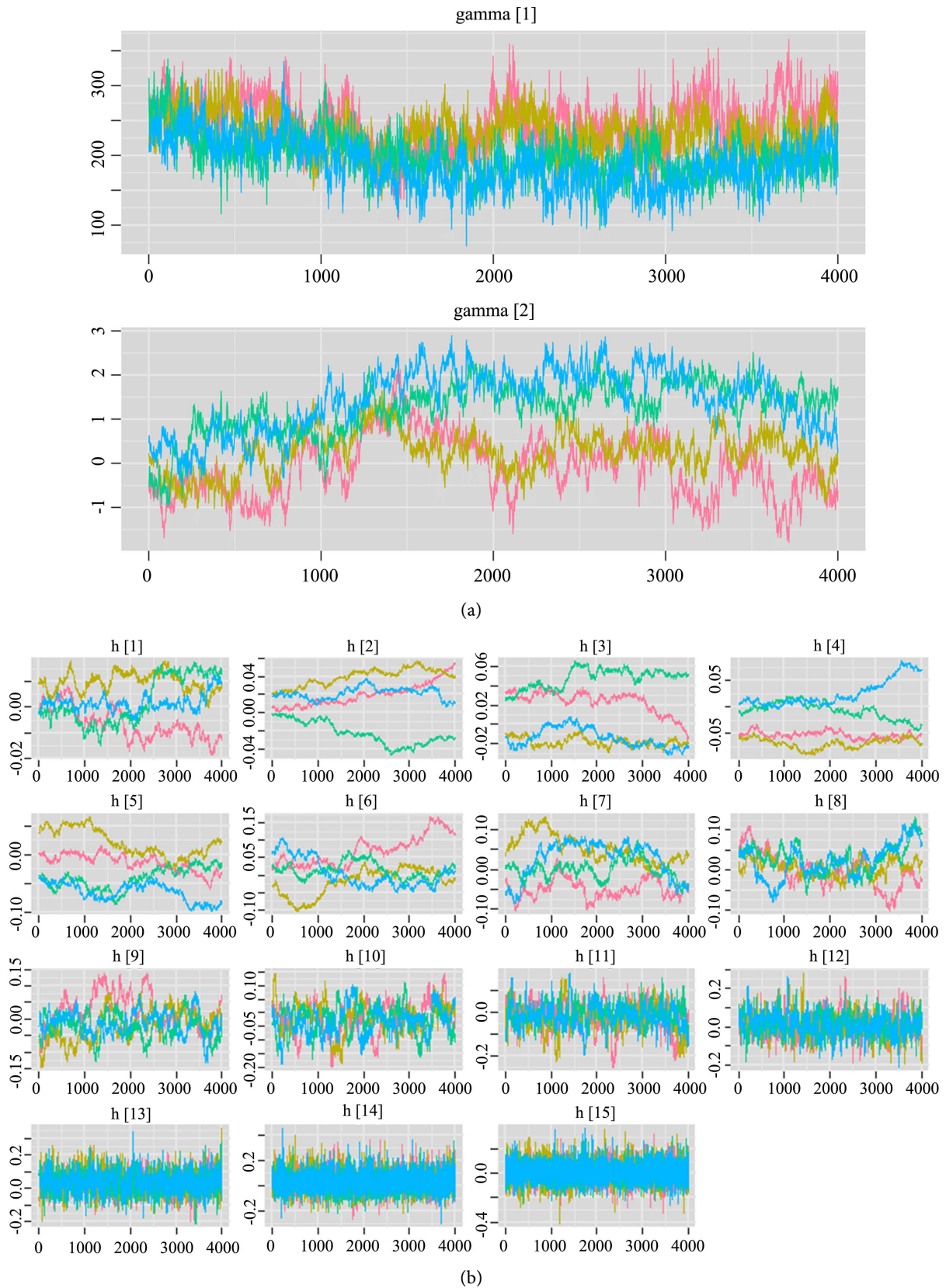


Figure 2. Traceplots plots for model parameters. (a) Gamma parameter in Hierarchical model; (b) h parameter in Hierarchical model.

Convergence of the Parameters

In this section we checked for the convergence of the model parameters by examining their respective trace plots. The parameters ρ and β of the model without hierarchy penalty showed an acceptable convergence to the stationary distribution both with graphical tests shown by trace plots in **Figure 1**.

Figure 2 shows that the parameters γ and h_k of the model with hierarchy penalty also had an acceptable convergence to the stationary distribution both with graphical tests.

4. Conclusion and Suggestions

This study utilizes Bayesian technique to model the penalized spline smoothing regression model with hierarchical penalty. We build on existing model by Eilers [25] and Rupert [2]. The model in this study is based on the penalized regression spline, in a semi-parametric modeling paradigm. Many situations arise where the relationship between the function of the response variable and covariates is non-linear. Due to the high complexity and intractability of this non-parametric model, the maximum likelihood approaches are not viable in the estimation of these models. Consequently, the Bayesian inference, utilizing MCMC techniques is highly favoured. Based on simulation results, penalized spline smoothing with hierarchical penalty provides a better fit compared to penalized spline smoothing without hierarchy; this was shown by the rapid convergence of the model posterior parameters and the lowest DIC value of the model. Therefore from the simulation results, the study concluded that the hierarchical model with fifteen sub-knots provides a better fit of the data.

Suggestions

Although the proposed model provides a better fit compared to the penalized spline without hierarchy penalty, the mean function density becomes rough as the sample size increases therefore the study proposes further research on the nature of an appropriate smoothing parameter for the model can be done.

Acknowledgements

Sincere thanks to my supervisors Prof. Samuel Mwalili and Prof. Leo Odongo for their professional guidance and performance, and special thanks to my husband and parents for their moral support and rare attitude of high quality.

Conflicts of Interest

The authors declare no conflicts of interest.

References

- [1] Eilers, P.H.C. and Marx, B.D. (1996) Flexible Smoothing with B -Splines and Penalties. *Statistical Science*, **11**, 89-121. <https://doi.org/10.1214/ss/1038425655>
- [2] Ruppert, D. (2002) Selecting the Number of Knots for Penalized Splines. *Journal of Computational and Graphical Statistics*, **11**, 735-757.

- <https://doi.org/10.1198/106186002853>
- [3] Pinheiro, J.C. and Bates, D.M. (1995) Approximations to the Loglikelihood Function in the Nonlinear Mixed-Effects Model. *Journal of Computational and Graphical Statistics*, **4**, 12-35. <https://doi.org/10.1080/10618600.1995.10474663>
- [4] Wahba, G., Wang, Y.D., Gu, C., Klein, R., Klein, B., et al. (1995) Smoothing spline Anova for Exponential Families, with Application to the Wisconsin Epidemiological Study of Diabetic Retinopathy: The 1994 Neyman Memorial Lecture. *The Annals of Statistics*, **23**, 1865-1895. <https://doi.org/10.1214/aos/1034713638>
- [5] Fan, J.Q. and Gijbels, I. (1995) Data-Driven Bandwidth Selection in Local Polynomial Fitting: Variable Bandwidth and Spatial Adaptation. *Journal of the Royal Statistical Society: Series B (Methodological)*, **57**, 371-394. <https://doi.org/10.1111/j.2517-6161.1995.tb02034.x>
- [6] Silverman, B.W. (1984) Spline Smoothing: The Equivalent Variable Kernel Method. *The Annals of Statistics*, **12**, 898-916. <https://doi.org/10.1214/aos/1176346710>
- [7] Brockmann, H. and Klein, T. (2004) Love and Death in Germany: The Marital Biography and Its Effect on Mortality. *Journal of Marriage and Family*, **66**, 567-581. <https://doi.org/10.1111/j.0022-2445.2004.00038.x>
- [8] Donoho, D.L. and Johnstone, J.M. (1994) Ideal Spatial Adaptation by Wavelet Shrinkage. *Biometrika*, **81**, 425-455. <https://doi.org/10.1093/biomet/81.3.425>
- [9] Faucett, C.L. and Thomas, D.C. (1996) Simultaneously Modelling Censored Survival Data and Repeatedly Measured Covariates: A Gibbs Sampling Approach. *Statistics in Medicine*, **15**, 1663-1685. [https://doi.org/10.1002/\(SICI\)1097-0258\(19960815\)15:15<1663::AID-SIM294>3.0.CO;2-1](https://doi.org/10.1002/(SICI)1097-0258(19960815)15:15<1663::AID-SIM294>3.0.CO;2-1)
- [10] Luo, Z. and Wahba, G. (1997) Hybrid Adaptive Splines. *Journal of the American Statistical Association*, **92**, 107-116. <https://doi.org/10.1080/01621459.1997.10473607>
- [11] DiMatteo, I., Genovese, C.R. and Kass, R.E. (2001) Bayesian Curve-Fitting with Free-Knot Splines. *Biometrika*, **88**, 1055-1071. <https://doi.org/10.1093/biomet/88.4.1055>
- [12] Zhou, S.G. and Shen, X.T. (2001) Spatially Adaptive Regression Splines and Accurate Knot Selection Schemes. *American Statistical Association*, **96**, 247-259. <https://doi.org/10.1198/016214501750332820>
- [13] Wood, S.A., Jiang, W.X. and Tanner, M. (2002) Bayesian Mixture of Splines for Spatially Adaptive Nonparametric Regression. *Biometrika*, **89**, 513-528. <https://doi.org/10.1093/biomet/89.3.513>
- [14] Miyata, S. and Shen, X.T. (2003) Adaptive Free-Knot Splines. *Computational and Graphical Statistics*, **12**, 197-213. <https://doi.org/10.1198/1061860031284>
- [15] Liu, Z.Y. and Guo, W.S. (2010) Data driven adaptive spline smoothing. *Statistica Sinica*, **20**, 1143-1163.
- [16] Baladandayuthapani, V., Mallick, B.K. and Carroll, R.J. (2005) Spatially Adaptive Bayesian Penalized Regression Splines (P-Splines). *Journal of Computational and Graphical Statistics*, **14**, 378-394. <https://doi.org/10.1198/106186005X47345>
- [17] Krivobokova, T., Crainiceanu, C.M. and Kauermann, G. (2008) Fast Adaptive Penalized Splines. *Journal of Computational and Graphical Statistics*, **17**, 1-20. <https://doi.org/10.1198/106186008X287328>
- [18] Cox, D.R. (1972) Regression Models and Life-Tables. *Journal of the Royal Statistical Society: Series B (Methodological)*, **34**, 187-202.

- <https://doi.org/10.1111/j.2517-6161.1972.tb00899.x>
- [19] Chen, Q.X., Wu, H.Y., Ware, L.B. and Koyama, T. (2014) A Bayesian Approach for the Cox Proportional Hazards Model with Covariates Subject to Detection Limit. *International Journal of Statistics in Medical Research*, **3**, 32-43. <https://doi.org/10.6000/1929-6029.2014.03.01.5>
- [20] Liu, L., Ma, J.Z. and O'Quigley, J. (2008) Joint Analysis of Multi-Level Repeated Measures Data and Survival: An Application to the End Stage Renal Disease (ESRD) Data. *Statistics in Medicine*, **27**, 5679-5691. <https://doi.org/10.1002/sim.3392>
- [21] Cantoni, E. and Hastie, T. (2002) Degrees-of-Freedom Tests for Smoothing Splines. *Biometrika*, **89**, 251-263. <https://doi.org/10.1093/biomet/89.2.251>
- [22] Ndung'u, A.W., Mwalili, S. and Odongo, L. (2019) Hierarchical Penalized Mixed Model. *Open Journal of Statistics*, **9**, 657-663. <https://doi.org/10.4236/ojs.2019.96042>
- [23] Chib, S. and Greenberg, E. (1995) Hierarchical Analysis of Sur Models with Extensions to Correlated Serial Errors and Timevarying Parameter Models. *Journal of Econometrics*, **68**, 339-360. [https://doi.org/10.1016/0304-4076\(94\)01653-H](https://doi.org/10.1016/0304-4076(94)01653-H)
- [24] Kazembe, L.N., Chirwa, T.F., Simbeye, J.S. and Namangale, J.J. (2008) Applications of Bayesian Approach in Modelling Risk of Malaria-Related Hospital Mortality. *BMC Medical Research Methodology*, **8**, Article No. 6. <https://doi.org/10.1186/1471-2288-8-6>
- [25] Eilers, P.H.C. and Marx, B.D. (2004) Splines, Knots, and Penalties. *WIREs Computational Statistics*, **2**, 637-353. <https://doi.org/10.1002/wics.125>