Scientific Research Publishing

Open Journal of Statistics, 2019, 9, 657-663
https://www.scirp.org/journal/ojs
ISSN Online: 2161-7198
ISSN Print: 2161-718X

Hierarchical Penalized Mixed Model

A. W. Ndung'u[1], S. Mwalili[2], L. Odongo[3]

[1]Pan African University, Nairobi, Kenya
[2]Jomo Kenyatta University of Agriculture and Technology, Nairobi, Kenya
[3]Kenyatta University, Nairobi, Kenya
Email: nannewanjira@gmail.com

How to cite this paper: 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

Received: June 22, 2018
Accepted: November 25, 2019
Published: November 28, 2019

Abstract

Penalized spline has been a popular method for estimating an unknown function in the non-parametric regression due to their use of low-rank spline bases, which make computations tractable. However its performance is poor when estimating functions that are rapidly varying in some regions and are smooth in other regions. This is contributed by the use of a global smoothing parameter that provides a constant amount of smoothing across the function. In order to make this spline spatially adaptive we have introduced hierarchical penalized splines which are obtained by modelling the global smoothing parameter as another spline.

Keywords

Penalized Splines, Mixed Model, Smoothing Parameter

1. Introduction

Non parametric smoothing involves letting the data determine the amount of smoothing. Classical smoothing splines use a global smoothing parameter in order to control the amount of smoothing in a function. When homogeneity of the smoothness cannot be reasonably assumed across the whole domain of the function, a natural extension is to allow the smoothing parameter to vary over the domain as a penalty function of independent variable, adapting to the change of roughness [1] [2]. Adaptive smoothing has been an interesting topic 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 [3]-[10]. In penalized regression splines, [11] modeled the penalty function by a linear interpolation on the logarithmic scale, [12] modeled the penalty function from full Bayesian approach and used Markov chain Monte Carlo for computa-

tion, [13] developed a fast and simple algorithm for the Bayesian p-spline based on Laplace approximation for the marginal likelihood. Modeling 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. [1] derived the reproducing kernels for a generic penalty function and suggested modeling it by B-splines. [2] 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. [14] 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.

2. Modeling Approach

2.1. Penalized Splines

P-splines are low-order basis spline with a penalty to avoid under smoothing. They are typically not spatially adaptive and hence have trouble when functions are varying rapidly. Regression splines are approximations to functions typically using low-order number of basis function. These splines are subject to lack of smoothness and various strategies have been proposed to attain this smoothness. e.g Regression P-splines [15] achieves smoothness by penalizing the sum of squares or likelihood by a single penalty parameter. The penalty parameter and the fit using P-splines are easy to compute using mixed model technology; [16] [17] [18] and are not sensitive to knot parameter selection [11]. A penalized spline can be seen as a compromise between smoothing and regression spline and it combines the attractive attributes of regression and smoothing splines. They are basically regression spline in which the penalty is applied directly to the coefficients of the piecewise polynomial. Hence one can retain a large number of knots and constrain their effect using a penalty to avoid over fitting. The number of knots defining the spline function is larger than that justified by the data but smaller than the number of observations. Thus they are referred to as low-rank smoothers and this significantly reduces numerical effort. The level of over fitting is controlled by a roughness penalty over the curve. The most common choice is a penalty based on the integral of a squared derivative of a spline curve. To avoid the drawbacks in regression spline and optimize the fit we can choose a large number of knots e.g. min (n/4, 40) as suggested by [11] and prevent over fitting by penalizing the coefficients of splines. That is, one finds

$$\min_{\beta,d} \left\| Y - X\beta - Sd \right\|^2 \tag{1}$$

subject to $\left\| d \right\|^2 < a$ for non negative constant a. Where Y is the response variable, β and d are the fixed and random effects vectors, X and S are the design matrices associated with the fixed and random effects vectors. Using a Lagrange multiplier, this minimization can be written as

$$\min_{\beta,d} \left\| Y - X\beta - Sd \right\|^2 + \omega d^{\mathrm{T}} d = \min_{\beta,d} \left\| y - C\theta \right\|^2 + \omega \theta^{\mathrm{T}} D\theta \tag{2}$$

With $\theta = \left(\beta^{\mathrm{T}}, d^{\mathrm{T}}\right)^{\mathrm{T}}$, D is a block diag $0_{(p+1)\times(p+1)} I_K$ and $\omega \geq 0$.

The resulting estimate is given by

$$\hat{y} = Z\left(Z^{\mathrm{T}} Z + \omega D\right)^{-1} Z^{\mathrm{T}} y \tag{3}$$

The smoothness of this estimate varies continuously as a function of a global smoothing parameter ω. The larger the value of ω the more the fit shrinks towards polynomial fit while small values of ω result in an over fitted estimate. Penalized spline can be seen as a generalization of the spline smoothing with more flexible choice of bases, penalties and knots. One chooses the spline basis based on sufficiently large number of knot and penalizes unnecessary structure. This spline possesses a number of good properties: It shows no boundary effect as many kernels smoother do. *i.e.* the spreading of a fitted curve as density outside of the domain of the data generally accompanied by bending towards zero, it is a straight forward extension of (generalized) linear regression models, conserve moments (means, variances) of the data *i.e.* Given a linear p spline with degree $q+1$ and a penalty of order $q+1$ or higher

$$\sum_{j=1}^{k} x^q y_j = \sum_{j=1}^{k} x^q \hat{y}_j \tag{4}$$

For all values of the smoothing parameter ω where $\hat{y}_j$ the fitted values are. This property is very useful in density smoothing where mean and variance of the estimated density are the same as mean and the variance of the data for any amount of smoothing. It also has polynomial curve fit as its limits. That is, for a penalty of order q and large values of the smoothing parameter ω, the fitted function will approach a polynomial of degree $q-1$, if the degree of the p-spline is equal or higher than q. Also the computations, including those of cross validation are relatively cheap and can easily be incorporated into standard software [15].

2.2. Mixed Models

Mixed model are regression model with both the fixed effects and random effects. They correspond to a hierarchy of levels with the repeated, correlated measurement occurring among all the lower level units for each particular upper level. The standard linear mixed model has the form

$$Y = X\beta + Sd + \varepsilon \tag{5}$$

where Y is a vector of observed responses, β is an unknown vector of fixed ef-

fects, d is an unknown vector of random effects or subject specific, with mean zero and variance W, X and S are design matrices associated with a vector of fixed effects β and a vector of random effects d respectively and ε is a vector of residual error term with zero mean and covariance matrix P. The dimensions of the design matrices X and S must conform to the lengths of the observation vector Y and the number of fixed and random effects respectively. It is generally assumed that the elements of d are uncorrelated with the elements of ε in which case the covariance matrix of the random effects and residual error term is a block diagonal

$$var\begin{pmatrix} d \\ \varepsilon \end{pmatrix} = \begin{bmatrix} W & 0 \\ 0 & P \end{bmatrix} \tag{6}$$

The matrices S and W will themselves be block diagonal if the data arise from a hierarchical structure, where a fixed number of random effects common to observations within a single higher-level unit are assumed to vary across the units for a given level of the hierarchy. Typically the vectors of residual errors are taken to independent and identically distributed and thus $P = \sigma_\varepsilon^2 I$ where σ_ε^2 is the residual variance. The covariance matrix W of the random effects vector d is often assumed to have a structure that depends on a series of unknown variance component parameters that need to be estimated in addition to the residual variance σ_ε^2 and the vector of fixed effects β.

The universal estimators of the fixed and random effects are the best linear unbiased estimators (BLUE) $\hat{\beta}$ of β and the best linear unbiased predictors (BLUP) $\hat{d}$ of d. This can be recovered as the solution to the mixed model equation,

$$\begin{bmatrix} X^{\mathrm{T}}P^{-1}X & X^{\mathrm{T}}P^{-1}Y \\ R^{\mathrm{T}}P^{-1}X & Y^{\mathrm{T}}P^{-1}S + F^{-1} \end{bmatrix} \begin{pmatrix} \hat{\beta} \\ \hat{d} \end{pmatrix} = \begin{bmatrix} X^{\mathrm{T}}P^{-1}Y \\ S^{\mathrm{T}}P^{-1}Y \end{bmatrix} \tag{7}$$

A mixed model is of the form,

$$Y = X\beta + Sd + \varepsilon$$

Assuming that d and ε are multivariate normal;

$$\begin{bmatrix} d \\ \varepsilon \end{bmatrix} \sim N\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} W & 0 \\ 0 & P \end{bmatrix} \right) \tag{8}$$

and taking $H = var(Y)$. Then $Y \sim N(X\beta, H)$ Which result into a pdf

$$f(y;\beta,H) = \frac{1}{\sqrt{2\pi H}} \exp\left\{ -\frac{1}{2H}(y - X\beta)^2 \right\}$$

The likelihood function becomes,

$$L(y;\beta,H) = \prod_{i=1}^{n} f(y;\beta,H) = (2\pi H)^{-\frac{n}{2}} \exp\left\{ -\frac{1}{2H} \sum_{i=1}^{n} (y - X\beta)^2 \right\} \tag{9}$$

The log likelihood becomes,

$$l(y;\beta,H) = -\frac{n}{2}\log 2\pi - \frac{n}{2}\log H + \exp\left[-\frac{1}{2}(y - X\beta)^{\mathrm{T}} H^{-1} (y - X\beta)^{\mathrm{T}} \right]$$

$$l(y;\beta,H) = -\frac{1}{2}\left\{\log 2\pi + \log H + (y - X\beta)^{\mathrm{T}} H^{-1}(y - X\beta)^{\mathrm{T}}\right\} \tag{10}$$

where $H = var(Y) = SWS^{\mathrm{T}} + P$. Assuming that the parameters defining the covariance matrices W and P are known, the MLE $\hat{\beta}$ of β is

$$\hat{\beta} = \left(X^{\mathrm{T}} H^{-1} X\right)^{-1} X^{\mathrm{T}} H^{-1} y \tag{11}$$

which although not obvious algebraically must also satisfy the mixed model equation given earlier. Since one of the ways in which these equations can be derived is directly from multivariate normality assumption. Typically W and P will not be known and can be estimated by substituting the expression for $\hat{\beta}$ back into $l(\beta;W;P)$ and maximizing the result over the parameters defining W and P. Once estimates for W and P have been determined, we can return to the mixed model equations and determined the BLUP $\hat{d}$ of random effects vector d as the vector that minimizes the expected mean squared error of prediction.

$$E\left\{\left(\hat{d} - d\right)^{\mathrm{T}}\left(\hat{d} - d\right)\right\} \tag{12}$$

The BLUP of d can be expressed as the posterior expectation of the random effects given the data $\hat{d} = E(d/Y)$ which can be solved explicitly under the normality assumption to yield,

$$\hat{d} = WS^{\mathrm{T}} H^{-1}\left(Y - X\hat{\beta}\right) \tag{13}$$

2.3. Hierarchical Penalized Mixed Model

Assuming

$$Y \sim N\left(m(X_i);\sigma_i^2\right), i = 1,2,\cdots,n \tag{14}$$

where $m(X_i)$ is modelled as truncated polynomial spline

$$m(X) = \beta_0 + \beta_1 X^1 + \cdots + \beta_p X^q + \sum_{k=1}^{r} \rho_k (X - \tau_k)_+^q \tag{15}$$

where $\tau_1, \tau_2, \cdots, \tau_r$ are the knots covering the range of x's and

$$(X - \tau_k)_+^q = \begin{cases} (X - \tau_k)_+^q, & (X - \tau_k)_+^q > 0 \\ 0, & \text{otherwise} \end{cases}$$

The knots are placed over the range of x's and the dimension of r is chosen generously. In penalized spline the approach is to put a penalty on the coefficient of ρ_k. The standard approach is to minimize sum of squares and the quadratic penalty $\omega\rho^{\mathrm{T}} D\rho$, where ω is the penalty parameter and D is the penalty square matrix. In truncated polynomial D is an identity matrix and the penalty is $\omega\rho^{\mathrm{T}}\rho$. In B spline basis the penalty is constructed using the difference between neighboring spline coefficients [15]. An important feature of penalized spline is its links to linear mixed model. Due to this link we assume

$$\rho \sim N\left(0, \sigma_\rho^2 D^{-1}\right)$$

where ρ is a vector of spline coefficients, $\sigma_\rho^2 = \sigma_\varepsilon^2/\omega$ and D^{-1} is a generalized inverse of D.

In this approach a single parameter σ_ρ^2 is used to shrink all the coefficients of spline and this can be a limitation especially if the underlying function is locally varying, *i.e.* it fails to completely capture the features of functions that exhibit strong heterogeneity. One way to avoid this is to allow the coefficients $\rho_1,\cdots,\rho_r$ to have prior variances $\rho_k \sim N\left(0,\sigma_\rho^2\{\tau_k\}\right)$ and assume that the shrinkage variance process $\sigma_\rho^2\{\tau_k\}$ is a smooth function modeled as a log-penalized spline

$$\sigma_\rho^2\{\tau_k\} = \exp\left[\gamma_0 + \gamma_1^{\tau^1} + \cdots + \gamma_p^{\tau^p} + \sum_{j=1}^{l} h_j \left(X-\alpha_j\right)_+^q\right] \tag{16}$$

where $\alpha_1,\alpha_2,\cdots,\alpha_l$ is a second layer of knots covering the range of $\tau_1,\tau_2,\cdots,\tau_r$. l is practically less than r. The hierarchical penalized smoothing model is completed by the shrinkage assumption $h_j \sim N\left(0,\sigma_h^2\right), j=1,2,\cdots,l$ and σ_h^2 is constant.

Thus our hierarchical smoothing model can be written as

$$Y \mid \rho,h = X_\rho\beta + S_\rho\rho + \varepsilon, \varepsilon \sim N\left(0,\sigma_\varepsilon^2 I_n\right)$$

$$\rho \mid h \sim N\left(0,\Sigma_\rho\right)$$

$$\Sigma_\rho = diag\left\{\exp\left(X_h^{\mathrm{T}} y + Z_h C\right)\right\}$$

$$C \sim N\left(0,\sigma_l^2 I_n\right)$$

where:

$$y = \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix}, X_\rho = \begin{pmatrix} \begin{bmatrix} 1 & \cdots & x_1^q \\ \vdots & \ddots & \vdots \\ 1 & \cdots & x_n^q \end{bmatrix} \end{pmatrix}, S_\rho = \begin{pmatrix} \begin{bmatrix} \left(x_1-\tau_1\right)_+^q & \cdots & \left(x_1-\tau_r\right)_+^q \\ \vdots & \ddots & \vdots \\ \left(x_n-\tau_1\right)_+^q & \cdots & \left(x_n-\tau_r\right)_+^q \end{bmatrix} \end{pmatrix}$$

$$X_h = \begin{pmatrix} \begin{bmatrix} 1 & \cdots & \tau_1^p \\ \vdots & \ddots & \vdots \\ 1 & \cdots & \tau_l^p \end{bmatrix} \end{pmatrix}, Z_h = \begin{pmatrix} \begin{bmatrix} \left(\tau_1-\alpha_1\right)_+^p & \cdots & \left(\tau_1-\alpha_l\right)_+^p \\ \vdots & \ddots & \vdots \\ \left(\tau_r-\alpha_1\right)_+^p & \cdots & \left(\tau_r-\alpha_l\right)_+^p \end{bmatrix} \end{pmatrix}$$

$$\beta = \left(\beta_0,\cdots,\beta_p\right)^{\mathrm{T}}, \rho = \left(\rho_0,\cdots,\rho_r\right)^{\mathrm{T}}$$

3. Results and Conclusion

Penalized splines are very common in parametric regression but they have one major drawback in that they are not spatially adaptive. This is due to the use of a global smoothing parameter across the whole heterogeneous function. In this research we aimed at coming up with a spatially adaptive penalized spline by introducing hierarchical splines. This was achieved by modeling the global smoothing parameter ω that is normally used in classical smoothing as another spline.

Conflicts of Interest

The authors declare no conflicts of interest regarding the publication of this paper.

References

[1] Grace Wahba, Y.W. (1995) Smoothing Spline ANOVA for Exponential Families, with Application to the Wisconsin Epidemiological Study of Diabetic Retinopathy. *Annals of Statistics*, **23**, 1865-1895. https://doi.org/10.1214/aos/1034713638

[2] Pintore, A., Speckman, P. and Holmes, C.C. (2006) Spatially Adaptive Smoothing Splines. *Biometrika*, **93**, 113-125. https://doi.org/10.1093/biomet/93.1.113

[3] Muller, H.-G. and Stadtmuller, U. (1987) Estimation of Heteroscedasticity in Regression Analysis. *Annals of Statistics*, **15**, 610-635. https://doi.org/10.1214/aos/1176350364

[4] Silver, B.W. (1985) Some Aspects of the Spline Smoothing Approach to Nonparametric Regression Curve Fitting. *Journal of the Royal Statistical Society*, *Series B*, **47**, 1-52.

[5] Brockmann, M., Gasser, T. and Herrmann, E. (1993) Locally Adaptive Bandwidth Choice for Kernel Regression Estimators. *Journal of the American Statistical Association*, **88**, 1302-1309. https://doi.org/10.1080/01621459.1993.10476411

[6] Donoho, D.L. and Johnstone, I.M. (1994) Ideal Spatial Adaptation by Wavelet Shrinkage. *Biometrica*, **81**, 425-455. https://doi.org/10.1093/biomet/81.3.425

[7] Fan, I.J. (1995) Data-Driven Bandwidth Selection in Local Polynomial Fitting: Variable Bandwidth and Spatial Adaptation. *Journal of the Royal Statistics Society B*, **53**, 371-394. https://doi.org/10.1111/j.2517-6161.1995.tb02034.x

[8] Luo, G.Z. (1997) Further Results on Non-Paratemtric Linear Regression Model in Survival Analysis. American Statistical Association, Alexandria, 107-116.

[9] Dimatteo, I., Genovese, C.R. and Kass, R.E. (2001) Bayesian Curve-Fitting with Free Knots Splines. *Biometrika*, **88**, 1055-1071. https://doi.org/10.1093/biomet/88.4.1055

[10] Wood, S., Jiang, W. and Tanner, M. (2002) Bayesian Mixture of Splines for Spatially Adaptive Non Parametric Regression. *Biometrika*, **83**, 513-528. https://doi.org/10.1093/biomet/89.3.513

[11] 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

[12] 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

[13] Krivobokova, T., Crainiceanu, C.M. and Kauermann, G. (2008) Fast Adaptive Penalized Splines. *Journal of Computational and Graphical Statistics*, **17**, 1-20.

[14] Liu, W.Z. (2010) Data Driven Adaptive Spline Smoothing. *Statistica Silica*, **20**, 1143-1163.

[15] Eilers, B.P. (1996) Flexible Smoothing with b-Splines and Penalties. *Statistical Science*, **11**, 89-102. https://doi.org/10.1214/ss/1038425655

[16] Robinson, G.K. (1991) That Blup Is a Good Thing the Estimation of Random Effects. *Statistical Science*, **6**, 15-51. https://doi.org/10.1214/ss/1177011926

[17] Coull, B.A., Ruppert, D. and Wand, M.P. (2001) Simple Incorporation of Interactions into Additive Models. *Biometric*, **57**, 539-545. https://doi.org/10.1111/j.0006-341X.2001.00539.x

[18] Chiang, C.-T., Rice, J.A. and Wu, C.O. (2001) Smoothing Spline Estimation for Varying Coefficient Models with Repeatedly Measured Dependent Variables. *Journal of the American Statistical Association*, **96**, 605-619. https://doi.org/10.1198/016214501753168280