An optimization approach: estimating knots in the splines regression model

Alberto Rodrigues Ferreira[1]

[1] Institute of Mathematics and Statistics, University of São Paulo

Abstract: The regression splines model has received considerable attention in recent years. However, the splines regression model has a significant disadvantage: one of its main components, called knots, are usually chosen before the estimation process. We propose a new methodology that tries to solve this using an optimization approach.

Related publications:

–Sousa, A. R. D. S., Severino, M. T., & Leonardi, F. G. (2020). Model selection criteria for regression models with splines and the automatic localization of knots. *arXiv preprint arXiv:2006.02649*.

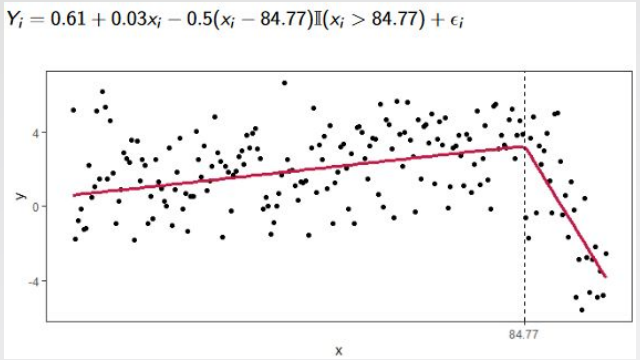
–Ruppert, D. (2002). Selecting the number of knots for penalized splines. *Journal of computational and graphical statistics*, *11*(4), 735-757.

3rd Conference on
Statistics and
Data Science
Salvador, Brazil (online)
October 28-30, 2021

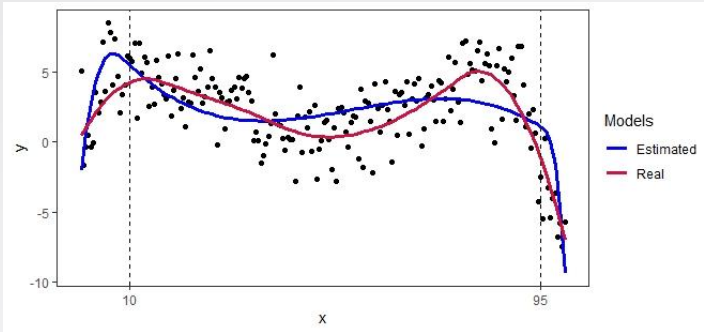
Problem and Usual Procedure

The studied model is given below:

$$Y_i = \beta_0 + \sum_{k=1}^{K} \beta_k x_i^k + \sum_{m=1}^{\alpha} \beta_{p+m}(x_i - t_m)^K \mathbb{I}(x_i > t_m) + \epsilon_i$$



However, the usual procedure is to choose knots before the estimation process.



Disadvantages

- Manual choice of the knots
- Increase in estimated standard errors
- Overfitting/Underfitting

Solution: *Consider location and number of knots as parameters.*

Proposed loss function

We propose a regularization method, calculated by the optimization algorithm called BFGS.

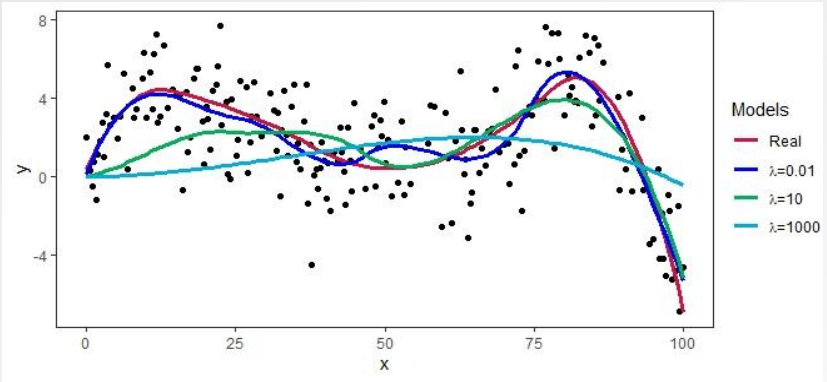
$$\frac{1}{n}\sum_{i=1}^{n}\left[y_i-\beta_0-\sum_{j=1}^{p}f_j(x_{ij})\right]^2+\sum_{j=1}^{p}\alpha_j\lambda_j\sum_{m=1}^{\alpha_j}\left|\beta_{(j,m+K)}\right|$$

The penalty deals directly with the balance between bias and variance in the model. Ideally, discarding irrelevant knots.

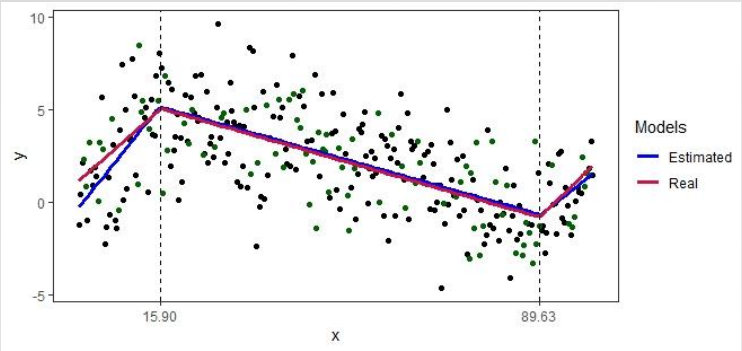
The larger the hyperparameter, the greater the penalty on knots.

The figure shows the estimated curves for three lambda values:

- λ̂=0.01
- λ̂=10
- λ̂=1000



Results and Conclusions



Our approach had several advantages over the usual procedure and estimated the knots appropriately.

Metrics	Real	Estimated
MSE	3.68	3.83
SD	4.55	4.62

Furthermore, we can note in this example that parameter estimates improved over the BFGS iterations.