2021

An Evaluation of Knot Placement Strategies for Spline Regression

William Klein

Claremont McKenna College

An Evaluation of Knot Placement Strategies for Spline Regression

submitted to
Professor Gillen

by
William Klein

for
Senior Thesis
Fall 2020
November 28, 2020

1 Acknowledgements

I would like to express my deepest gratitude to Professor Gillen for advising me through the writing of this thesis. I would also like to thank Professor Smith for her guidance and support.

Contents

2 Abstract

Regression splines have an established value for producing quality fit at a relatively low-degree polynomial. This paper explores the implications of adopting new methods for knot selection in tandem with established methodology from the current literature. Structural features of generated datasets, as well as residuals collected from sequential iterative models are used to augment the equidistant knot selection process. From analyzing a simulated dataset and an application onto the Racial Animus dataset, I find that a B-spline basis paired with equally-spaced knots remains the best choice when data are evenly distributed, even when structural features of a dataset are known and implementable. However, the residual-based knot selection outperforms both the equidistant knot placement and structural knot placement methods when data are irregularly distributed.

3 Literature Review

3.1 Introduction

The paper "Splines, Knots, and Penalties" (Eilers and Marx 2010) combines the concepts of spline regression, basis function selection, knot placement, and smoothness penalizations. In this section, I provide context from the relevant literature that allows me to expand upon the findings presented by Eilers and Marx. For the scope of this paper, a B-spline basis in conjunction with equally spaced knot selection is found to be the optimal foundation for expanding upon the existing literature. I expand upon these reasons below.

3.2 Spline and Penalized Spline Regression

In the context of regression analysis, spline modelling constitutes the separation of a data set through "knots" to facilitate fitting smooth functions onto sections of a larger data set. This technique is classified as nonparametric because it is not predicated on the fit to a given parameter (Perperoglou 2019). The number of knots selected affects the bias-variance tradeoff: as the number of knots increases, the model risks overfitting the data, but too few knots can produce a more restrictive function. The splines themselves are smooth up to the given polynomial order imposed on the model, d—i.e., they are continuous and differentiable. The implementation of regression splines allows us to write an unknown function f from the basis functions spanning a vector space, B_k, and the associated spline coefficients, β_k, where k represents the number of knots. This function is given in Equation 3.1.

$$f(X) = \sum_{k=1}^{K+d+1} \beta_k B_k(X) \tag{3.1}$$

In the case of penalized regression splines, a "roughness" penalty is applied to the selection of knots to bring more choice to the model beyond the selection of basis functions. As a result, the statistics used for penalized spline regression are balanced for both goodness-of-fit and smoothness. These conflicting ends necessitate some parameter to control the weighting towards either goal, which takes the form of the smoothing parameter λ (Gu 1993).

3.3 PB-Splines and PT-Splines

In their paper, "Splines, Knots, and Penalties," authors Eilers and Marx (2010) distinguish between two commonly used methods for spline regression: PB-splines and PT-splines. The first of these methods uses the B-spline basis. The B-spline basis is an extension of the commonly used cubic basis, which generates and fits a cubic polynomial between knots. An additional parametrization is then placed on the cubic splines, as given in Equation 3.2 (Perperoglou 2019). A difference penalty is also applied, based on the ℓ_1 norm (Eilers and Marx 2010).

$$\begin{aligned} \xi_1 \leq \dots \leq \xi_d \leq \xi_{d+1} < \xi_{d+2} < \dots < \xi_{d+K+1} \\ < \xi_{d+K+2} \leq \xi_{d+K+3} \leq \dots \leq \xi_{2d+K+2} \end{aligned} \tag{3.2}$$

In contrast, PT-splines use a truncated power series as a spline basis. In this case, a basic polynomial of degree d is chosen for the initial basis, and each successive function to the right of the Kth knot receives additional deviations. The PT-spline is taken from the basis function found in Equation 3.3 (Perperoglou 2019).

$$\begin{aligned}&B_1(x) = 1, B_2(x) = x, ..., B_{d+1}(x) = x^d,\\&B_{d+2}(x) = (x - \tau_1)^d, ..., B_{K+d+1} = (x - \tau_k)^d\end{aligned} \tag{3.3}$$

Unlike PB-splines, a ridge penalty is used for PT-splines via the ℓ_2 norm; however, as Eilers and Marx (2010) find, the two methods are effectively equivalent in their respective cases. As with PB-splines, the penalty parameter λ must necessarily take a positive value.

3.4 Knot Selection

Although polynomial degree and basis selection have a fairly low impact on model fit, the means by which knots are selected can yield a substantial effect on the model. In their paper, Eilers and Marx (2010) explore the difference between quantile placement and equidistant placement for knots. My paper builds on their work and explores the implications of incorporating new strategies for knot placement.

In their analysis, Eilers and Marx (2010) primarily pair the use of equally spaced knots with PB-splines but take note of the potential for pairing PT-splines with equally spaced knots rather than the proposed quantile spacing. This aligns with the existing literature, which indicates that penalized splines are most often used with equally spaced knots, as quantile spacing tends to produce unevenly distributed knots which in turn demands the use of weights to the spline functions (Perperoglou 2019).

3.5 Adaptive Knot Placement

There are multiple different approaches for adaptive spline regression, which generates optimal knot placement for prioritizing model fit. Generally, the estimate for locally adaptive spline regression is presented in the form given in Equation 3.4, where TV indicates the total variation operator (Tibshirani and Wasserman 2013).

$$\hat{f} = \underset{f_k}{\arg\min} \sum_{i=1}^{n} (y_i f(x_i))^2 + \lambda TV(f^{(k)}) \tag{3.4}$$

A widely known form of adaptive spline placement is the Multivariate Adaptive Regression Splines (MARS), first presented by Jerome Friedman (1991). MARS functionally uses a two-step process to produce splines, wherein sets of basis functions are progressively added to the model in a forward stepwise procedure, while in the backward stepwise procedure, the least effective pair of basis functions is removed. These steps continue until the model has been grown and pared to a point of optimality (Friedman 1991). Unlike the semi-parametric approach to penalization by the PT-splines and PB-splines outlined by Eilers and Marx, MARS follows an entirely nonparametric approach.

Another recent paper published by Vivien Goepp, Olivier Bouaziz, and Grégory Nuel (2018) outlines the potential deviation from penalized spline generation towards the proposed "A-spline." Adaptive splines penalize using the ℓ_0 norm and produce a relatively low number of knots. In a key difference from other methods of selecting knots, A-spline methodology is an explicit advancement of the B-spline basis (Goepp 2018), and so it cannot be paired with other distinct methods of knot selection such as a quantile or equidistant approach, nor other types of basis functions.

3.6 Potential Contributions

Because of the incompatibility with adaptive spline regression and the selection of a B-spline basis or TPF-spline basis, expanding on Eilers and Marks' work leads to the consideration of a third method for knot placement. To this end, I argue that structural knot placement, using basic properties of the spline functions and hybridizing between free knot selection and equidistant knot placement, represents a viable area of furthering the existing literature. Fitting knots based on structural attributes of the data can be paired with of B-spline bases or TPF-spline bases and can be compared against evenly placed knot selection and quantile spaced knot selection as distinct methodologies, allowing for the creation of a 2x3 comparison matrix.

Based on the findings of Eilers and Marx, this paper focuses on the usage of the B-spline basis function in conjunction with different methods for knot-placement, as this method appears to yield the best model fit. In contrast to TPF-splines, B-splines also avoid sacrificing the numerical stability of the regression, which is particularly important for calculating derivatives from the estimated functions. As quantile-based knot placement was deemed computationally and analytically inferior to equally spaced knot placement, equal spacing will be used as the foundation for exploring alternative knot placement techniques. Future research could expand upon either the usage of a quantile-based knot selection process or the incorporation of an alternative basis, such as the TPF-spline basis used by Eilers and Marx.

4 Methodology

In some cases, structural features of a data set are known a priori and can be used to inform a regression or modeling approach. Alternatively, residuals can be taken from a basic spline regression and be used to inform the iterative placement of knots in the absence of prior information. These features can be internalized to a spline regression either by choosing a more complex basis or by using a B-spline basis with knot selection methodology that extends beyond equal spacing or quantile spacing. This section explores several different strategies for incorporating structural features of data while maintaining a B-spline basis selection.

4.1 Equally Spaced Knot Selection

For this analysis, three different methods were used to generate knots: equally spaced placement, residual based placement, and structural based placement. In all three cases, these knots were then used to fit a B-spline. In the equally spaced knots case, k knots were selected at even intervals throughout the data (for example, setting k to 10 resulted in knots being spaced at a distance of 0.2 apart), and a B-spline basis was applied for regression. Models were then evaluated for their goodness of fit as evaluated through the root mean squared error statistic.

4.2 Residual Based Knot Selection

Although moving away from equally spaced knot selection sacrifices some of the established numerical stability and quality of model fit, alternative techniques allow for the possibility of internalizing information into the model that might yield a better fit to the underlying function. In the case that structural features of the data are not known, the residuals of a given model fit contribute additional intuition about where knots can be placed to further optimize model fit and fully leverage the information available. For this approach, the following procedure was used:

1. Fit an initial regression using a B-spline basis and k / 2 equally spaced knots and recover the residuals
2. Select one new knot using normalized exponentiation of the squared residuals
3. Fit a new spline using the accumulated knots and recover the residuals
4. Repeat steps 2 and 3 until k total knots have been selected

4.3 Normalized Exponentiation Function

In order to assess the merits of using an alternative method for placing knots, an iterative knot placement method is used in conjunction with a normalized exponentiation, or "softmax," function to generate probabilities. This function is noted as Equation 4.1.

$$p_i = \frac{exp(\hat{u_i})}{\sum_j exp(\hat{u_j})} \tag{4.1}$$

First, an initial regression was performed using k / 2 equally spaced knots, and the squared residuals were calculated at each value along the x axis. Then, “probabilities” were generated at each x value by exponentiating the squared residuals and scaling against the sum of all exponentiated squared residuals. These probabilities could then be cumulatively summed at each x_i to create intervals such that the distance between x_i and x_i - 1 = p_i. In turn, these p_i were considered as disjoint events in the complete probability space from 0 to 1. A random number was then selected on the uniform distribution between 0 and 1 such that the corresponding event in the probability space, p_i, was mapped to a value x_i. This x value was taken as a new knot to be added to the existing k / 2 knots, and the procedure repeated until k knots had been chosen.

This process satisfied two goals: to help prevent over-clustering at regions where the residuals generated from the regression with k / 2 knots would otherwise be taken directly adjacent to one another, and to prioritize the relative extremity of the model's residuals so as to maximize the model fit.

4.4 Structurally Based Knot Selection

For the simulated data described in the following section, structural features of the dataset could easily be extracted from the underlying function. This allowed for a method of knot selection that could potentially be applied to datasets where discontinuities and derivatives are known a priori, despite this process being less generalizable than the novel residual based knot selection process.

The structurally-based knot approach followed a similar process to the residual approach, wherein the curvature index of the data was mapped onto a probability space from 0 to 1, and knots were selected at regions where the second derivative held the most extreme values. Because the curvature indices were known prior to the regression on the data, there was no need to iteratively regress 0.5 * k times as with the residual based knot selection method. However, a softmax function was employed to ensure that knots would not be clustered while still targeting points of structural extremity. Additional code was appended to the selection function in R to ensure that a single value of x would not be selected for two distinct knots.

5 Simulated Data Application

To analyze the goodness of fit of the residual, structural, and equally-spaced knot selection methods, simulated datasets were generated using four different functions. All of the datasets share the same primary four variables: an x and a y variable, constituting the values generated by the given function; a $y*$ variable indicating the "true" value of the function without any noise; and a "curvature index," which scaled the absolute value of the second derivative of the function across a specified positive range (either [0,1], [0,5], or [0,10]), so as to quantify a structural component of the underlying function in preparation for structural knot selection.

5.1 Fifth Degree Polynomial

The first data set used for this preliminary analysis was generated from the fifth degree polynomial function shown in Equation 5.1.

$$y = 5x^5 + 3x^4 - 10x^3 - 2x^2 + x \tag{5.1}$$

From this function, noise was generated using a standard deviation of 0.25, and points were created at intervals of 0.01 on the domain from -1 to 1, resulting in 201 evenly spaced data points. Summary statistics are given in Table 1.

Table 1: Fifth Degree Polynomial Function

Statistic	N	Mean	St. Dev.	Min	Pctl(25)	Pctl(75)	Max
X	201	0.000	0.582	−1	−0.5	0.5	1
Y	201	−0.065	1.872	−3.328	−0.971	0.542	5.162
Ystar	201	−0.061	1.838	−3	−0.9	0.3	5
Curvature.Index	201	1.109	0.891	0.017	0.554	1.424	5.000

5.2 Polynomial Function with Discontinuity

The second data set used was generated from a third degree polynomial function with an added discontinuity, which is shown in Equation 5.2.

$$y = \begin{cases} 4x^3 - 2x^2 + 1 & -0.53 \leq x \leq 0.03 \\ 4x^3 - 2x^2 - 1 & else \end{cases} \tag{5.2}$$

From this function, noise was generated using a standard deviation of 0.25, and points were created at intervals of 0.01 on the domain from -1 to 1, resulting in 201 evenly spaced data points. Summary statistics are given in Table 2.

Table 2: Polynomial Function with Discontinuity

Statistic	N	Mean	St. Dev.	Min	Pctl(25)	Pctl(75)	Max
X	201	0.000	0.582	−1	−0.5	0.5	1
Y	201	0.199	1.720	−3.636	−1.078	1.120	5.112
Ystar	201	0.221	1.682	−3	−1.0	1.1	5
Curvature.Index	201	1.950	1.569	0.011	0.903	2.651	10.000

5.3 Sine Function

The third data set used was generated from the sine function, as shown in Equation 5.3.

$$y = sin(x) \tag{5.3}$$

From this function, noise was generated using a standard deviation of 0.25, and points were created at intervals of 0.01 * 2π on the domain from -2π to 2π, resulting in 201 evenly spaced data points. Summary statistics are given in Table 3.

Table 3: Sine Function

Statistic	N	Mean	St. Dev.	Min	Pctl(25)	Pctl(75)	Max
X	201	0.000	3.655	−6.283	−3.142	3.142	6.283
Y	201	−0.014	0.716	−1.626	−0.612	0.569	1.409
Ystar	201	−0.000	0.707	−1	−0.7	0.7	1
Curvature.Index	201	0.633	0.311	0	0.4	0.9	1

5.4 Simulated Data Exponential

The fourth data set used was generated from the exponential function shown in Equation 5.4.

$$y = e^{(-(ln(x)-0.5)^2)} \tag{5.4}$$

From this function, noise was generated using a standard deviation of 0.25, and points were created at intervals of 0.02 on the domain from 0.01 to 4, resulting in 200 evenly spaced data points. Summary statistics are given in Table 4.

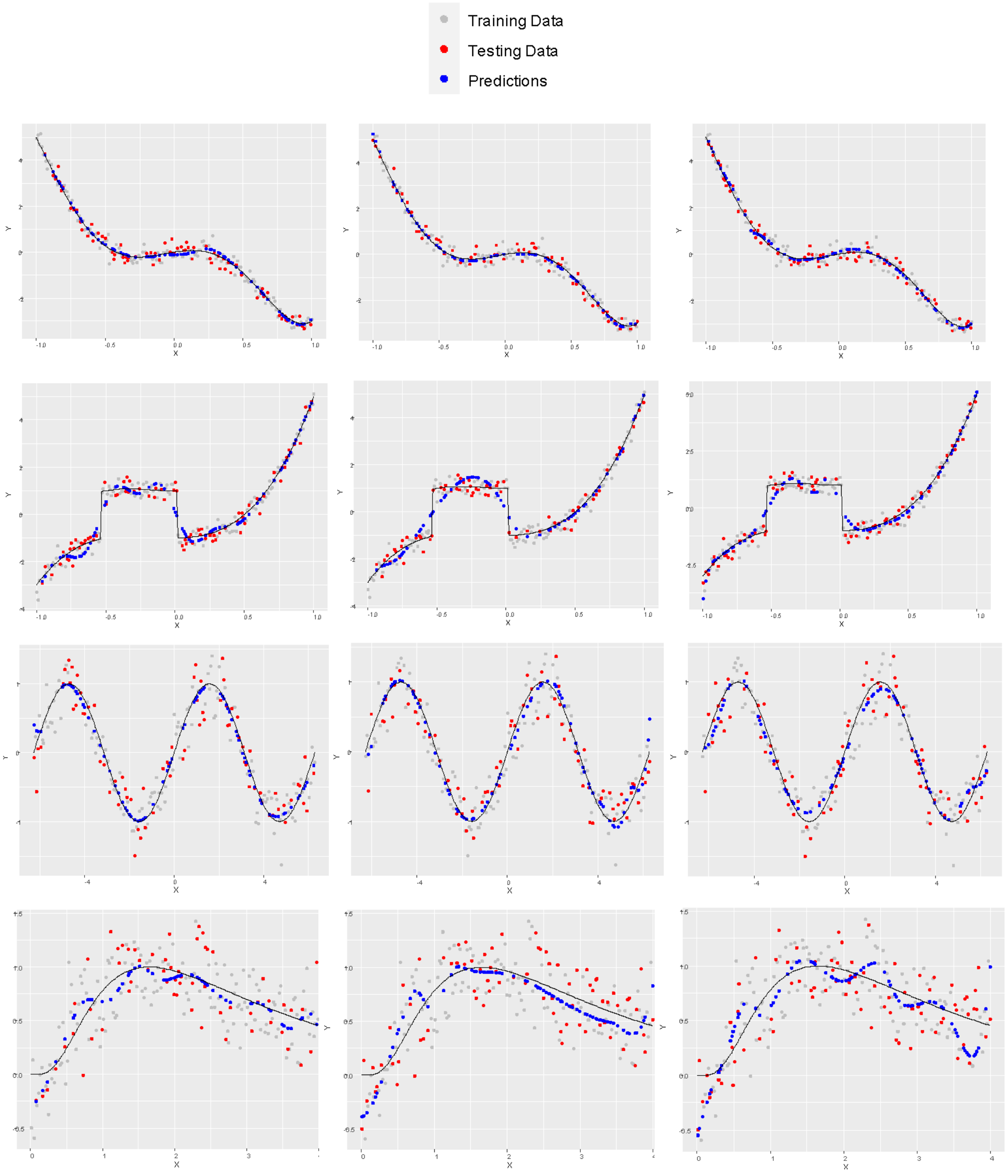
Table 4: Exponential Function

Statistic	N	Mean	St. Dev.	Min	Pctl(25)	Pctl(75)	Max
X	200	2.000	1.158	0.010	1.005	2.995	3.990
Y	200	0.644	0.392	−0.590	0.414	0.930	1.430
Ystar	200	0.664	0.293	0	0.5	0.9	1
Curvature.Index	200	0.180	0.233	0.00000	0.020	0.294	1.000

5.5 Results

Model cross-validation with a 60:40 training/testing ratio was conducted for 1000 iterations for each technique of knot selection and spline regression. Based on preliminary analysis of RMSE reports and values of k, models were generated using 5, 10, or 20 knots. The RMSE and RMSE numerical standard error were gathered from each series of simulations and are listed in Table 5. Example iterations are shown at $k = 10$ for each dataset and method in Figure 1.

Figure 1:



Equal Spacing

Structural

Residual

Table 5: Model Outputs

Polynomial Function

	RMSE			RMSE NSE		
	k=5	k=10	k=20	k=5	k=10	k=20
Equal Space	0.05306	0.074088	0.132657	0.000975	0.000829	0.001666
Structural	0.138847	0.363466	3.798587	0.172062	0.109946	1.427978
Residual	0.065054	0.095911	0.174248	0.012305	0.005517	0.009443

Polynomial Function with Discontinuity

	RMSE			RMSE NSE		
	k=5	k=10	k=20	k=5	k=10	k=20
Equal Space	0.356342	0.271221	0.271648	0.001077	0.001244	0.004004
Structural	0.363125	0.36105	1.051922	0.006364	0.015684	0.385056
Residual	0.372629	0.414613	0.307968	0.00347	0.078814	0.00895

Sine Function

	RMSE			RMSE NSE		
	k=5	k=10	k=20	k=5	k=10	k=20
Equal Space	0.086872	0.096836	0.140028	0.00074	0.00083	0.00175
Structural	0.117827	0.112368	0.286989	0.007952	0.004357	0.050808
Residual	0.121186	0.141403	0.233538	0.004661	0.012426	0.015902

Exponential Function

	RMSE			RMSE NSE		
	k=5	k=10	k=20	k=5	k=10	k=20
Equal Space	0.084563	0.107876	0.153639	0.000723	0.000764	0.00143
Structural	0.104255	0.157902	0.199571	0.003873	0.020121	0.012499
Residual	0.11965	0.124476	0.212292	0.021571	0.002348	0.018105

As can be seen in Table 5, the equidistant placement of knots outperformed the other methods at all values of k, and across all four functions. For the polynomial function with a discontinuity, the sine function, and the exponential function, the curvature based approach tended to produce a slightly better fit than the residual knot selection method at low values of k; however, the residual based knot selection tended to outperform the structural model for these three functions as k increased beyond 10.

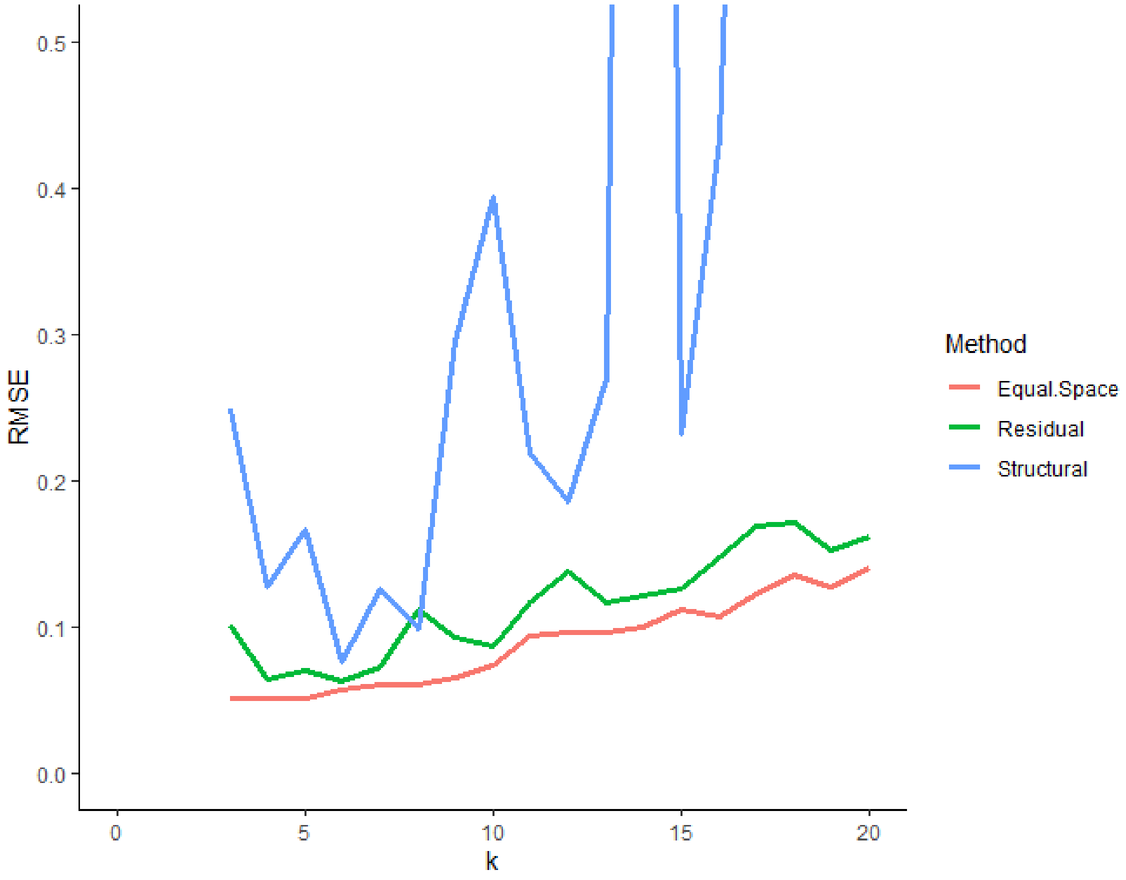
Notably, the worst performance across all methods and functions was seen at high values of k for the basic fifth-degree polynomial. The model particularly deteriorated at values of X close to -1 and 1, partly due to the high values for the second derivative clustered at the edge. This often resulted in several knots being placed on the border of the dataset and led to extreme predictions at the minimum and maximum values where data in the testing set was outside of the domain of the training set—ultimately resulting in a poor model fit.

5.6 Analysis

Several possible explanations exist for why the equal spacing method outperformed the structural and residual knot generation methods. In addition to the clustering of knots at the edges of the domain of the generated functions, it is possible that too many knots were generated following the alternative approaches. All three methods almost exclusively produced the greatest fit at $k = 5$, suggesting that only the first few knots generated from the curvature index or residuals might be of value towards generating a model. To this end, generating only a few structurally or residually based knots might be a superior means of supplementing the equally-spaced knot approach rather than relying on the new approaches for selecting half of the knots used for regressing.

Ultimately, these findings reaffirm the consensus of the existing literature, as the equally spaced knots produced the best model fit when paired with a B-spline basis. In particular, when half of the knots used for spline regression are generated based on the curvature of the underlying function, increasing k to high values rapidly deteriorates the model fit. Figure 2 demonstrates this phenomenon for the fifth-degree polynomial function, where the x-axis indicates the value of k and the y-axis reflects the average RMSE over 100 models.

Figure 2:



Future research could internalize the generation of the curvature index to the iterative knot selection process, such that after the initial regression using equally spaced splines, the derivatives of the piecewise functions could be evaluated and used to generate knots for further iterations. Special attention could also be given to the generation of knots at the borders of the training set such that knots will not clusters in areas that will ultimately yield a poor model fit to the testing data, and for varying the proportion of knots generated using the alternative methods.

6 Demonstrated Application

6.1 Racial Animus Data

To demonstrate how this methodology can be applied to a practical dataset, I used Professor Seth Stephens-Davidowitz' "Racial Animus" observations (Chae 2015). Stephens-Davidowitz measured the percent of Google search queries that included racially charged language for a given geographical area and used an algorithm to create the "Animus" variable. Animus is scaled between 0 and 250, with larger values indicating higher racial animus. The variable "ObamaKerry" is calculated as the percentage of the popular vote won by Barack Obama in the 2008 presidential election when subtracting the percentage won by John Kerry in 2004. The ObamaKerry and Animus variables are relatively correlated but share a nonlinear relationship, making the variables ideal candidates for this applied methodology.

Also included in these data sets is the "BachPlus" variable, which indicates the percentage of individuals in a given area with a bachelor's degree or higher level of education. Under the assumption that education levels are highly determinative of responsiveness of voting behavior to racial animus, the BachPlus variable contains structural information about the data. Due to this relationship, BachPlus is used similarly to the curvature index of the simulated data as an input into the softmax function to calculate where knots should be placed under the structural knot selection method. Summary statistics for these three variables are given in Table 6.

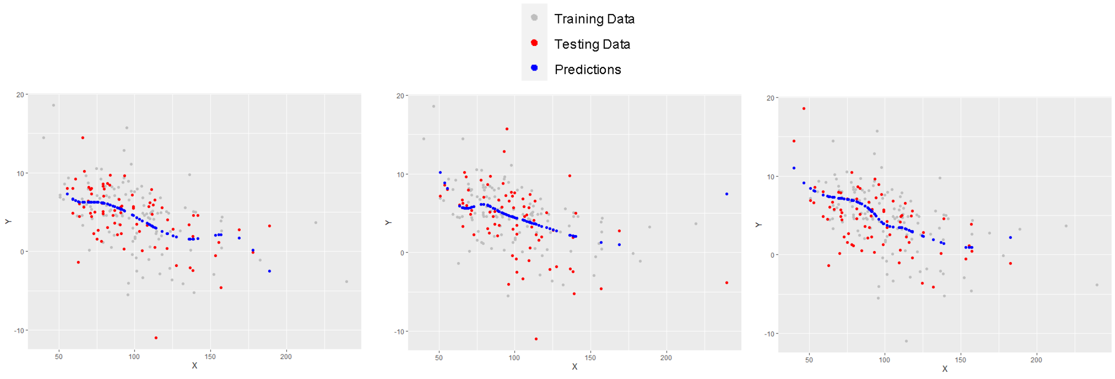
Table 6: Racial Animus

Statistic	N	Mean	St. Dev.	Min	Pctl(25)	Pctl(75)	Max
Animus	196	99.073	30.599	39.799	78.875	113.164	239.492
ObamaKerry	196	4.511	4.030	−10.975	2.448	7.056	18.602
BachPlus	196	23.415	5.747	12.491	18.782	26.761	42.534

6.2 Results

Model cross-validation with a 60:40 training/testing ratio was conducted for 1000 iterations for each technique of knot selection and spline regression. Models were fitted using different values of k and sample RMSE statistics were generated to evaluate model fit. Due to the presence of outliers on the domain of Animus, models were generated twice over the dataset: once using the complete data, and once using only points where Animus was between 40 and 200. Model results are given in Table 7, and visualizations for these methods over the complete data set are shown in Figure 3.

Figure 3:



Equal Spacing

Structural

Residual

Table 7: Racial Animus Application

Complete Dataset

	RMSE			RMSE NSE		
	k=5	k=10	k=20	k=5	k=10	k=20
Equal Space	17790.35	1958577	4630451	13154.85	1387927	1548361
Structural	388.419	5299.983	7.13E+08	282.3573	3117.584	4.3E+08
Residual	4.880723	2612.166	3.09E+09	0.151593	1448.408	1.72E+09

Filtered Dataset

	RMSE			RMSE NSE		
	k=5	k=10	k=20	k=5	k=10	k=20
Equal Space	4.635144	171.883	406790.3	0.130496	54.20774	124123.2
Structural	180.9906	1413.328	45837.93	110.4831	966.0656	17463.37
Residual	3.969452	11.02572	20690.83	0.084653	4.299016	10461.88

6.3 Analysis

The results given in Table 7 reveal the benefit of the residual knot placement technique. Although all three methods again produced the best model fit at lower values of k, the residual method outperformed the equidistant knot placement models. Unlike in the simulated data, the "X" variable of the Racial Animus dataset is irregularly distributed across the x-axis, resulting in regions on the domain where data are relatively thin. In these cases, the residual model proportionately places fewer knots relative to the equidistant knot placement model, ultimately minimizing the root mean squared error of the regression.

These model outputs also reinforce the prior explanation for the structural knot placement method underperforming. The most extreme values for the BachPlus variable are concentrated at the edges of the data, resulting in a higher density of knots in areas that do not benefit the model fit.

7 Conclusion

Based on the combined findings from the simulated application and Racial Animus application of these methodologies, the structural method should not be used in any case. Instead, either the equidistant or residual knot spacing methods should be used at low values of k depending on whether data is regularly or irregularly distributed. The equidistant method is particularly sensitive to the presence of outliers in a data set, which substantially reduce the model's goodness of fit.

8 Appendix I: R Code

```
equalSplineFit <- function(df,k,fulldf=df){
  knots <- equalknots(c(min(df$X),max(df$X)),k)
  model <- lm (Y ~ bs(X, knots = knots,
              Boundary.knots = c(min(fulldf$X),max(fulldf$X))),
              data = df)
  predictions <- model %>%
    predict(df)
  df <- df %>%
    add_predictions(model)
  model
}


strucSplineFit <- function(df,k,fulldf=df){
  knots <- equalknots(c(min(df$X),max(df$X)),k%/%2)
  inds = c()
  while (length(knots) < k){
    df <- df %>%
      mutate(origind = row_number())
    if (!length(inds)==0){
      df1 <- df [-inds,]
    }
    else{df1<-df}
    denom = sum(exp(df1$Curvature.Index))
    df1 <- df1 %>%
      mutate(Pvals = exp(Curvature.Index)/denom) %>%
      mutate(Cvals = cumsum(Pvals))
    rn = runif(1,0,1)
    for (i in 1:length(df1$X)){
      if (i == 1){
        if (rn < df1$Cvals[i]){
          index = i
        }
      }
      else if (rn < df1$Cvals[i] & rn >= df1$Cvals[i-1]){
        index = i
      }
    }
    ind = df1$origind[index]
    inds <- c(inds, ind)
    newknot = df$X[ind]
    knots <- c(knots, newknot)
  }
  model <- lm (Y ~ bs(X, knots = knots), data = df)
  df <- df %>% add_predictions(model=model)
  model
}
```

```
residSplineFit <- function(df,k,fulldf=df){
  knots <- equalknots(c(min(df$X),max(df$X)),k%/%2)
  inds = c()
  while (length(knots) < k){
    model <- lm (Y ~ bs(X, knots = knots), data = df)
    df <- df %>% add_residuals(model=model) %>%
      mutate(uhatsq = resid^2, origind = row_number())
    if (!length(inds)==0){
      df1 <- df [-inds,]
    }
    else{df1<-df}
    denom = sum(exp(df1$uhatsq))
    df1 <- df1 %>%
      mutate(Pvals = exp(uhatsq)/denom) %>%
      mutate(Cvals = cumsum(Pvals))
    rn = runif(1,0,1)
    for (i in 1:length(df1$X)){
      if (i == 1){
        if (rn < df1$Cvals[i]){
          index = i
        }
      }
      else if (rn < df1$Cvals[i] & rn >= df1$Cvals[i-1]){
        index = i
      }
    }
    ind = df1$origind[index]
    inds <- c(inds, ind)
    newknot = df$X[ind]
    knots <- c(knots, newknot)
    }
  model <- lm (Y ~ bs(X, knots = knots), data = df)
  df <- df %>% add_predictions(model=model)
  model
}
```

9 Appendix II: Sample RMSE Statistics for Simulated Data Sets

Table 8: Model Outputs

Polynomial Function

	RMSE			RMSE NSE		
	k=5	k=10	k=20	k=5	k=10	k=20
Equal Space	0.252101	0.261458	0.272785	0.000533	0.000686	0.00123
Structural	0.319492	0.532675	3.922286	0.014932	0.109395	1.427476
Residual	0.257363	0.277179	0.3154	0.001376	0.004968	0.008793

Polynomial Function with Discontinuity

	RMSE			RMSE NSE		
	k=5	k=10	k=20	k=5	k=10	k=20
Equal Space	0.450699	0.389712	0.38091	0.001128	0.00105	0.00377
Structural	0.458122	0.470025	1.153672	0.006227	0.015253	0.384834
Residual	0.464759	0.517483	0.415557	0.003231	0.078759	0.008509

Sine Function

	RMSE			RMSE NSE		
	k=5	k=10	k=20	k=5	k=10	k=20
Equal Space	0.260221	0.264732	0.290236	0.000519	0.000616	0.001194
Structural	0.281409	0.277879	0.426513	0.007683	0.003905	0.050408
Residual	0.282384	0.303956	0.37125	0.004119	0.011954	0.015198

Exponential Function

	RMSE			RMSE NSE		
	k=5	k=10	k=20	k=5	k=10	k=20
Equal Space	0.254392	0.262608	0.277108	0.000448	0.000582	0.001113
Structural	0.267755	0.303832	0.320606	0.003335	0.019748	0.012118
Residual	0.284524	0.268743	0.332782	0.021297	0.001853	0.017723

References

Chae DH, Clouston S, Hatzenbuehler ML, Kramer MR, Cooper HLF, Wilson SM, et al. (2015). Association between an Internet-Based Measure of Area Racism and Black Mortality, *PLoS ONE* 10(4): e0122963. doi:10.1371/journal.pone.0122963

Chib, Siddhartha, and Edward Greenberg. (2014). Nonparametric Bayes Analysis of the Sharp and Fuzzy Regression Discontinuity Designs." *Mimeo.*

Eilers, Paul, and Brian D. Marx. Splines, Knots, and Penalties. (2010). *WIREs Comp Stat* 2: 637-653.

Friedman, Jerome H. Multivariate Adaptive Regression Splines. (1991). *The Annals of Statistics* 19, no. 1: 1-67.

Goepp, Vivien, Olivier Bouaziz, and Grégory Nuel. (2018). Spline Regression with Automatic Knot Selection. *arXiv* preprint arXiv:1808.01770.

Gu, Chong. (1993). Smoothing Spline Density Estimation: A Dimensionless Automatic Algorithm. *Journal of the American Statistical Association* 88, no. 422: 495-504.

Perperoglou, Aris, Willi Sauerbrei, Michal Abrahamowicz, and Matthias Schmid. (2019). A Review of Spline Function Procedures in R. *BMC Medical Research Methodology* 19, no. 1: 46.

Tibshirani, Ryan, and Larry Wasserman. (2013). Nonparametric Regression. *Statistical Machine Learning*, Spring.