Optimizing Knot Placement in Regression Splines

Joseph Holey Supervisor: Michael O'Malley

6th September 2019

Introduction - 1

- In this work we simulated a function $q(x)$ given by:

$$\frac{\sin(x)}{x} + 0.5\cos(2x + \frac{\pi}{4}). \tag{1}$$

Introduction - 1

- In this work we simulated a function $q(x)$ given by:

$$\frac{\sin(x)}{x} + 0.5\cos(2x + \frac{\pi}{4}). \tag{1}$$

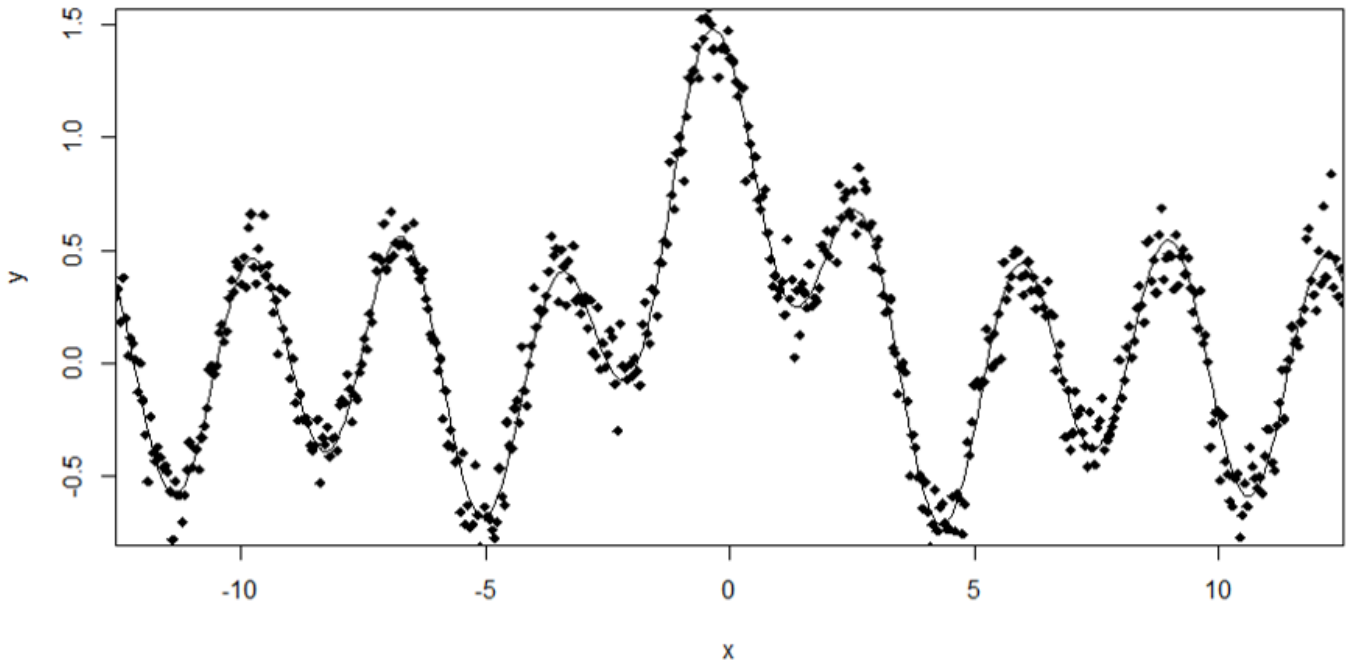
- We then added some random normally distributed noise with a mean of zero and a standard deviation of 0.1 to our grid of values.

Introduction - 1

- In this work we simulated a function $q(x)$ given by:

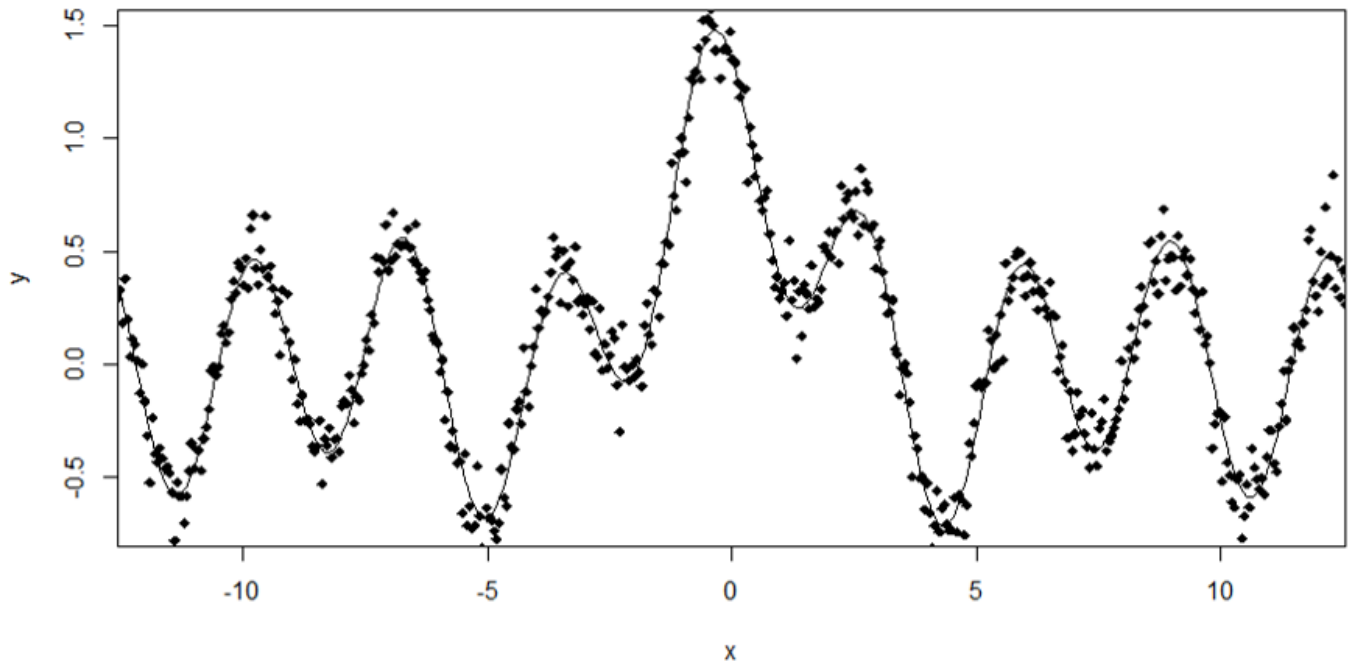
$$\frac{\sin(x)}{x} + 0.5\cos(2x + \frac{\pi}{4}). \tag{1}$$

- We then added some random normally distributed noise with a mean of zero and a standard deviation of 0.1 to our grid of values.



Introduction - 2

- Next, we used a variety of methods to obtain an estimate $\hat{f}(x)$ for the true function $q(x)$ based on simulated data. These methods required the use of the `mgcv` package and the `freeknotspline` package. We made use of bootstrapping in conjunction with both of these.



What is a Spline?

- Regressions splines are used to construct a model $\hat{f}(x)$ to fit a set of data. Many of them require knots to do this.

What is a Spline?

- Regressions splines are used to construct a model $\hat{f}(x)$ to fit a set of data. Many of them require knots to do this.
- There are many different types of spline but some of the most common are basis splines (B-splines) and penalised splines (P-Splines).

What is a Spline?

- Regressions splines are used to construct a model $\hat{f}(x)$ to fit a set of data. Many of them require knots to do this.
- There are many different types of spline but some of the most common are basis splines (B-splines) and penalised splines (P-Splines).

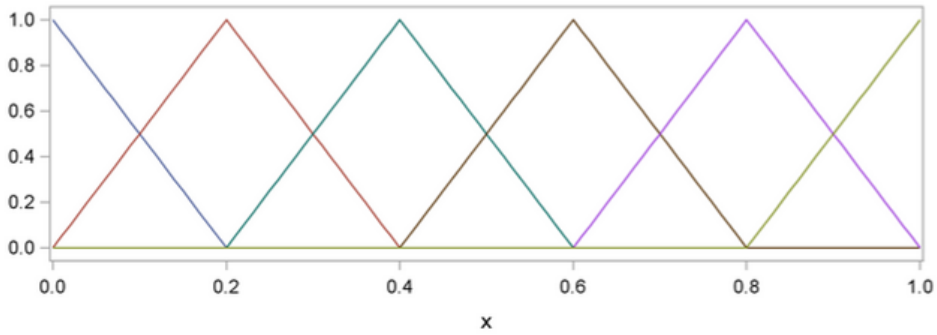


Figure: First degree B-spline basis functions

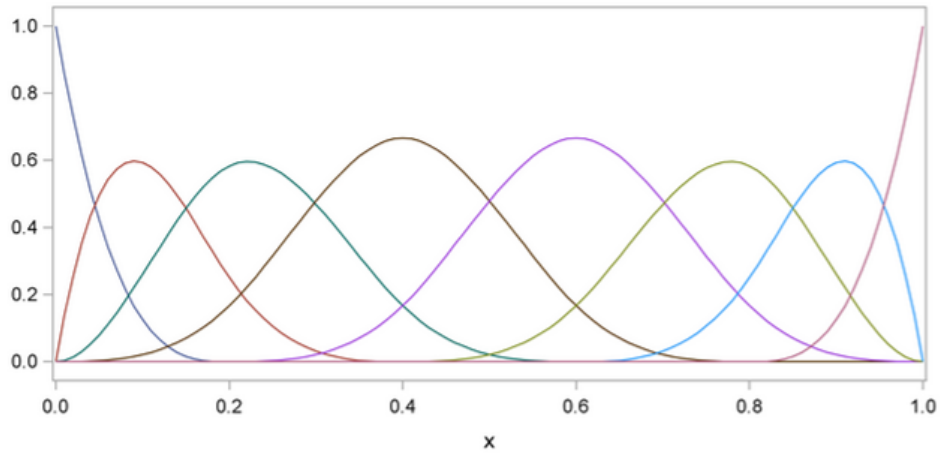


Figure: Third degree B-spline basis functions

P-Splines

- In this work we mainly used penalised splines (or P-splines). To fit this model we sought to minimise the following equation:

$$||\boldsymbol{y} - \mathbf{X}\boldsymbol{\beta}||^2 + \lambda \int_0^1 [f''(x)]^2 \, \mathrm{d}x \tag{2}$$

where $\mathbf{X}$ is our model matrix, $\boldsymbol{\beta}$ is a vector of unknown parameters that we are trying to find and λ is a tuning parameter.

P-Splines

- In this work we mainly used penalised splines (or P-splines). To fit this model we sought to minimise the following equation:

$$||\boldsymbol{y} - \mathbf{X}\boldsymbol{\beta}||^2 + \lambda \int_0^1 [f''(x)]^2 \, \mathrm{d}x \tag{2}$$

  where $\mathbf{X}$ is our model matrix, $\boldsymbol{\beta}$ is a vector of unknown parameters that we are trying to find and λ is a tuning parameter.

- For the purposes of computation we reformulated the integral term as:

$$\boldsymbol{\beta}^{\mathbf{T}}\mathbf{S}\boldsymbol{\beta} \tag{3}$$

  where $\mathbf{S}$ is a matrix of known coefficients.

K-fold Cross Validation

- Cross validation is an example of a resampling method.

K-fold Cross Validation

- Cross validation is an example of a resampling method.
- The primary use of cross validation is to select the best tuning parameters.

K-fold Cross Validation

- Cross validation is an example of a resampling method.
- The primary use of cross validation is to select the best tuning parameters.

Method

- For k-fold cross validation split the data set into k evenly sized sets of data.
- Train your model using the first $k-1$ data sets.
- Use the remaining subset as a test set.
- Repeat this process until each of k subsets has been used as the test set.

Motorcycle Data

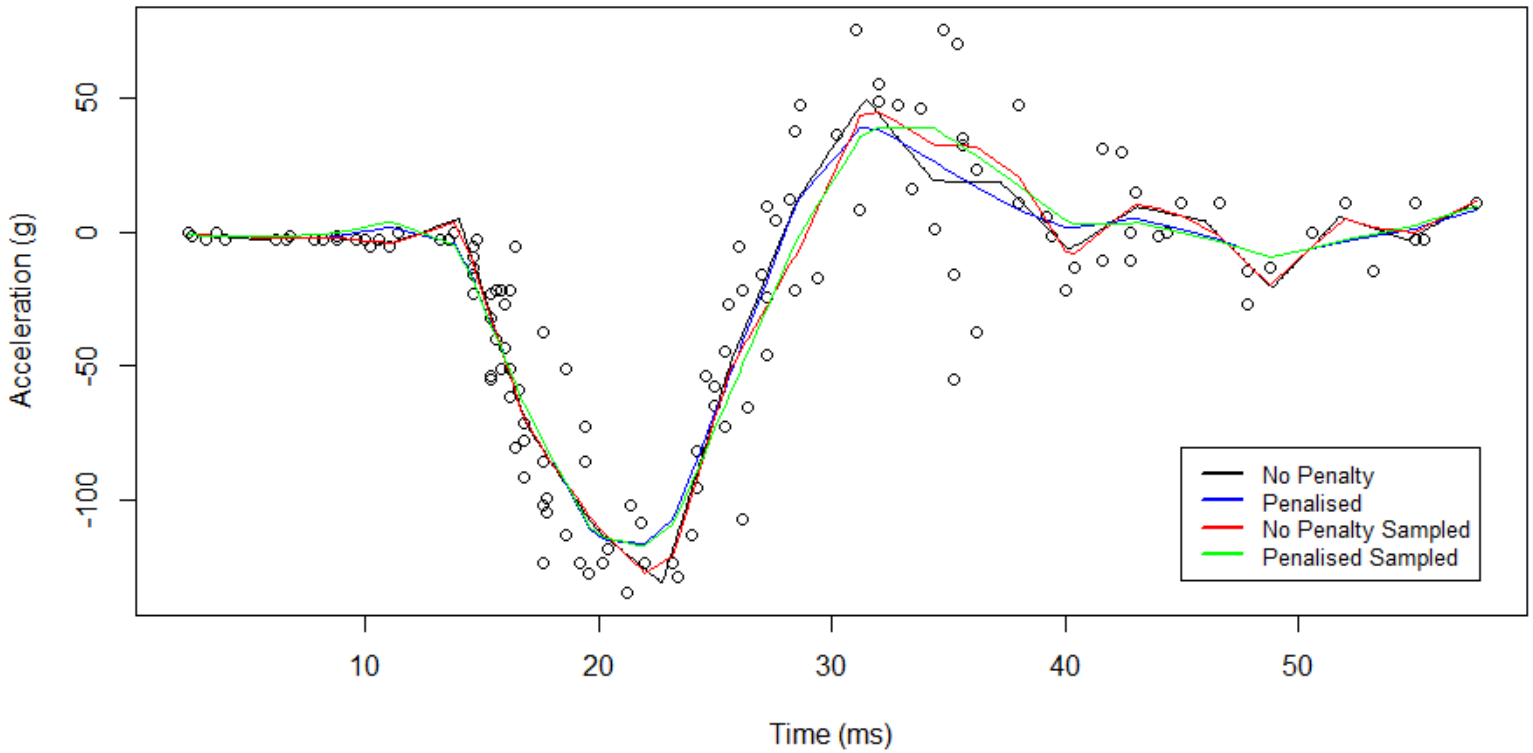


Figure: Data showing the acceleration that a motorcyclist undergoes in a crash with several different fits.

The Bootstrap

- Like cross validation, the bootstrap is an example of a resampling method.

The Bootstrap

- Like cross validation, the bootstrap is an example of a resampling method.
- Its primary use is uncertainty estimation.

The Bootstrap

- Like cross validation, the bootstrap is an example of a resampling method.
- Its primary use is uncertainty estimation.
- If you had a set of n data points $(x_1, x_2, ..., x_n)$ then the bootstrap method would take a sample of these of size n with replacement so some of the data points may be repeated multiple times and others may not get selected at all.

The Bootstrap

- Like cross validation, the bootstrap is an example of a resampling method.
- Its primary use is uncertainty estimation.
- If you had a set of n data points $(x_1, x_2, ..., x_n)$ then the bootstrap method would take a sample of these of size n with replacement so some of the data points may be repeated multiple times and others may not get selected at all.
- Measuring the statistic of interest (e.g. the mean) for each of these bootstrapped samples then allows you to determine the variance of that statistic.

The Importance of Knot Positioning

- The positions of the knots for a spline can have a very large impact on how good a fit it provides to the data points.

The Importance of Knot Positioning

- The positions of the knots for a spline can have a very large impact on how good a fit it provides to the data points.
- A poor selection of the location of knots can lead to splines that are not even competitive with a simple polynomial regression.

The Importance of Knot Positioning

- The positions of the knots for a spline can have a very large impact on how good a fit it provides to the data points.
- A poor selection of the location of knots can lead to splines that are not even competitive with a simple polynomial regression.

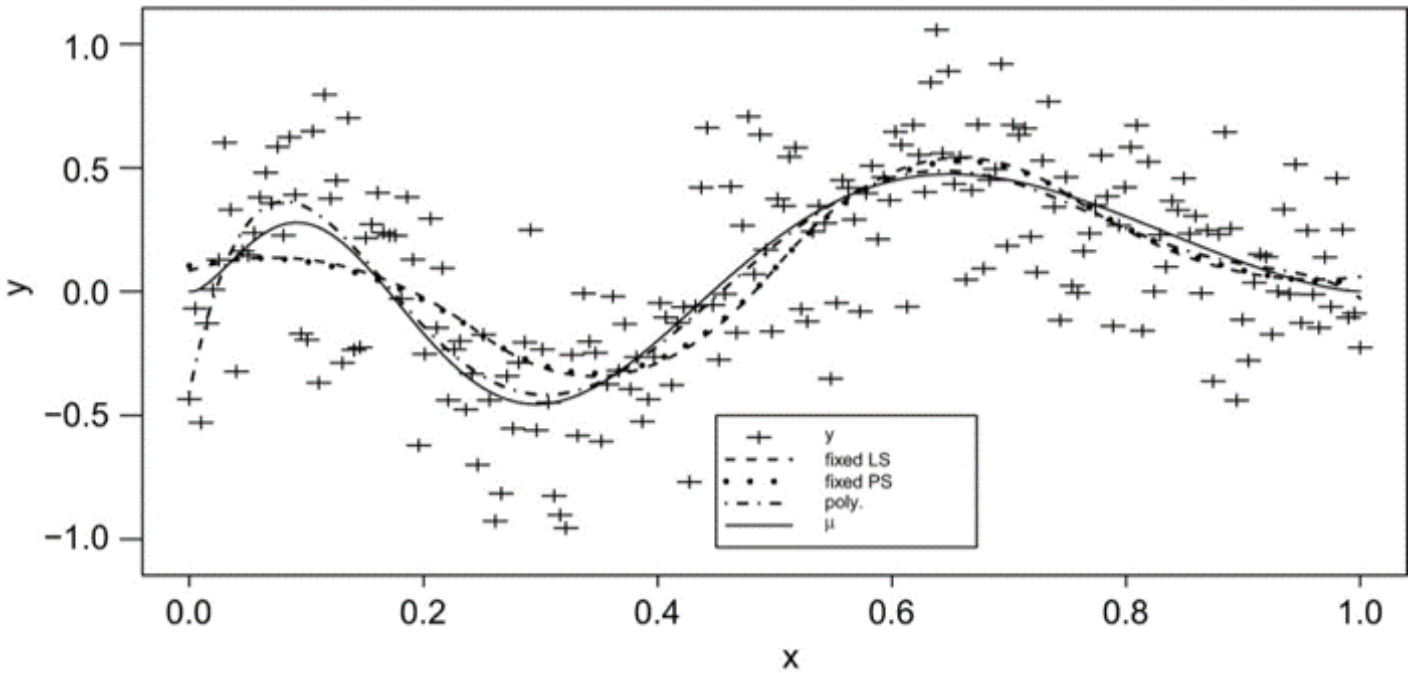
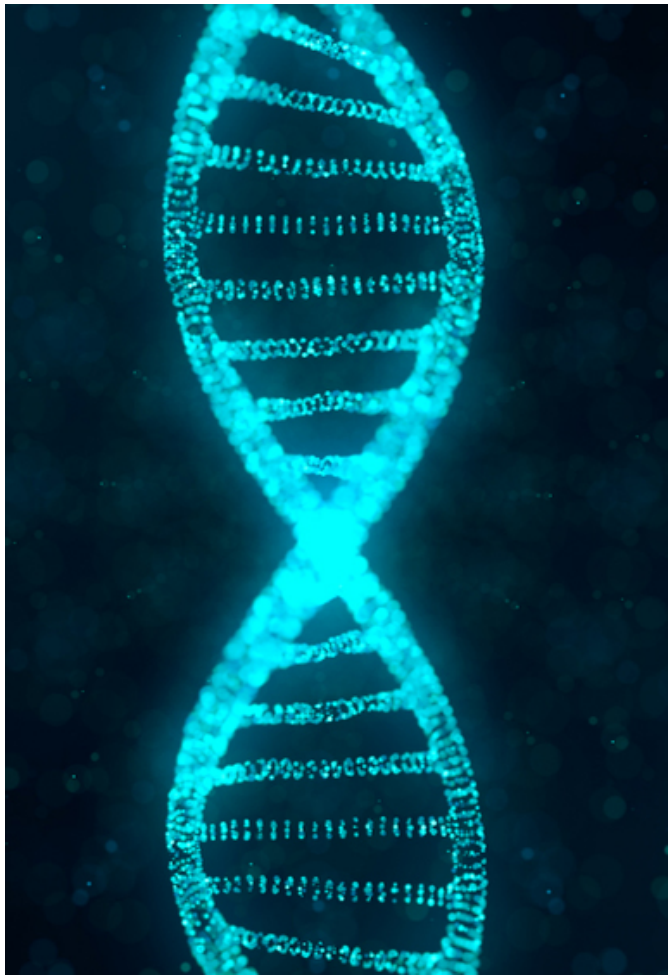


Figure: A simple polynomial regression outperforming splines

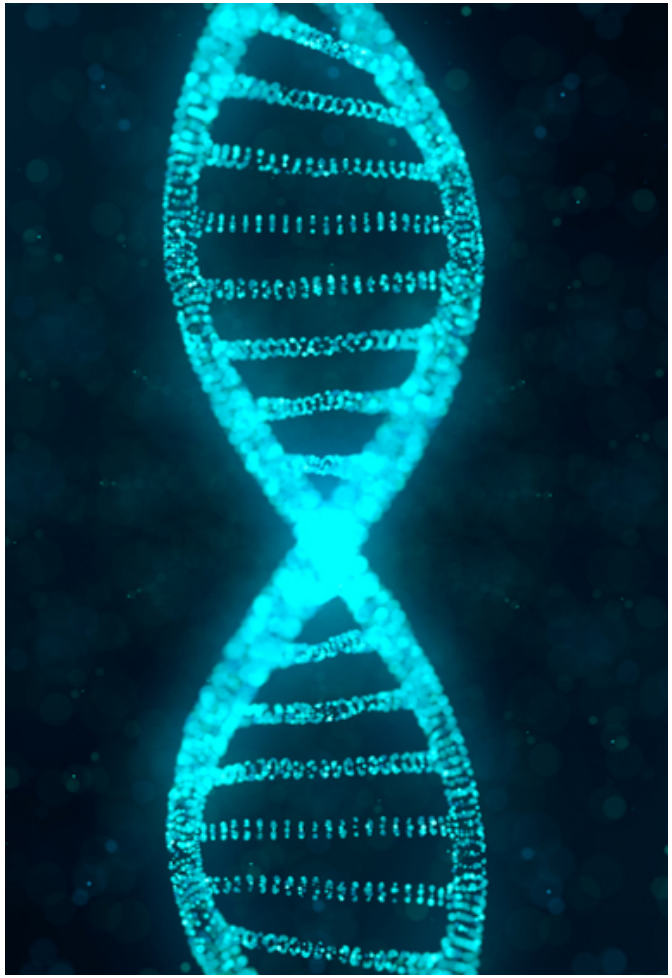
The Genetic Algorithm

- To improve our knot placement we considered the genetic algorithm.
- This algorithm was contained in the `freeknotsplines` package.
- The genetic algorithm begins by first randomly generating a large number of sets of k knots.
- There are 3 main steps to the algorithm - selection, crossover and mutation.
- These are iterated over multiple times until the best set of knots are found.



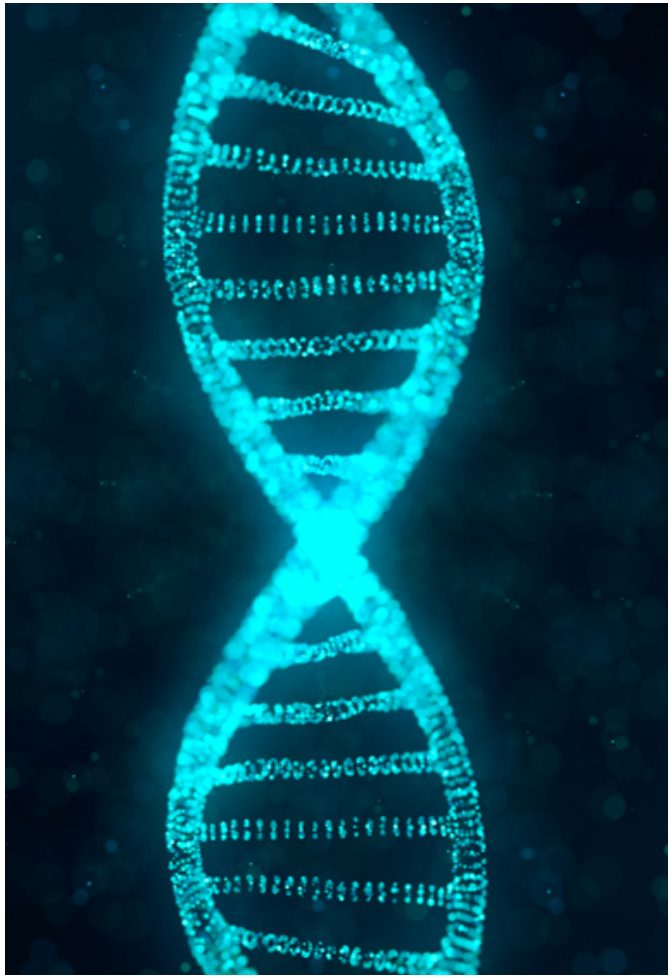
The Genetic Algorithm - Selection

- For selection we choose some measure such as the residual sum of squares.
- We then choose to keep the sets of knots which have a value of this measure below some threshold value and we discard the other sets of knots.



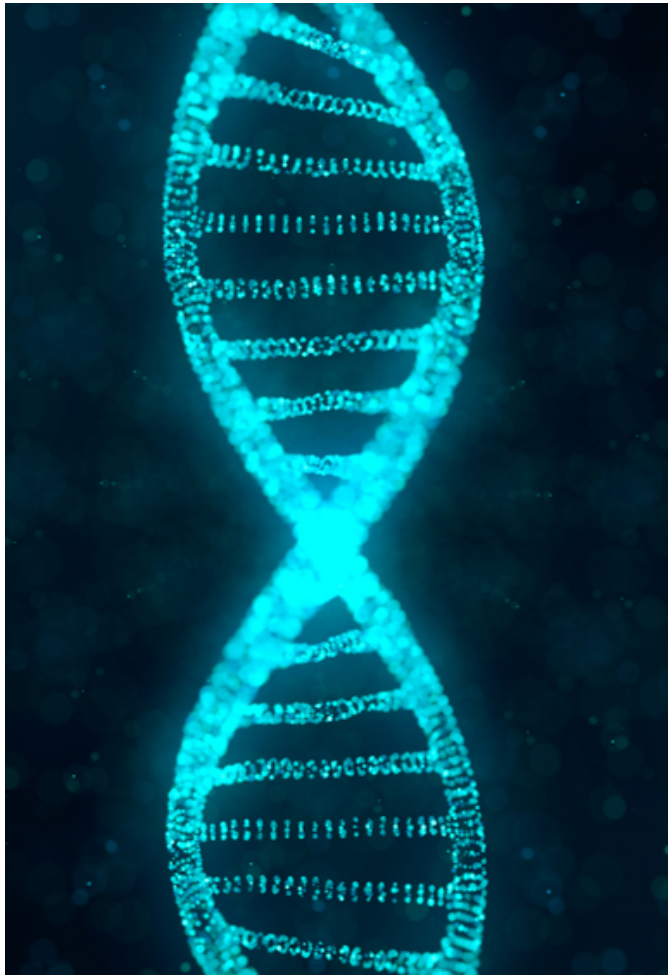
The Genetic Algorithm - Crossover

- In the crossover step an integer, ℓ is randomly chosen between 1 and k and two parent sets are also chosen at random.
- A child set is then produced consisting of the first $\ell - 1$ knots of one of the parent sets and the last $k - \ell + 1$ knots of the other parent set.



The Genetic Algorithm - Mutation

- Mutation consists of randomly choosing one of the surviving knot sets and then selecting one of the knots ξ_ℓ.
- This knot is then replaced with one randomly selected in $(\xi_{\ell-1}, \xi_{\ell+1})$.



Methods

- We used two main methods in this work.

Methods

- We used two main methods in this work.
- Firstly we used bootstrapping with the `mgcv` package to find our fits and confidence intervals.

Methods

- We used two main methods in this work.
- Firstly we used bootstrapping with the `mgcv` package to find our fits and confidence intervals.
- In the second method we ran the genetic algorithm to find the best set of knots for our data set. We then reused these knots to fit to all the bootstrapped data sets.

Results - 1

- We assessed the quality of the fits by considering the mean squared error given by:

$$\frac{1}{n}\sum_{i=1}^{n}(\hat{f}(x_i) - q(x_i))^2, \tag{4}$$

where the x_i are our grid points and $q(x)$ is the true function.

Results - 1

- We assessed the quality of the fits by considering the mean squared error given by:

$$\frac{1}{n}\sum_{i=1}^{n}(\hat{f}(x_i) - q(x_i))^2, \tag{4}$$

where the x_i are our grid points and $q(x)$ is the true function.

Method	Mean Squared Error
Fixed Knots	5.896×10^{-4}
Genetic algorithm	5.762×10^{-4}

Table: This table shows the Mean Squared Error for the two methods between the truth and the fit.

Results - 2

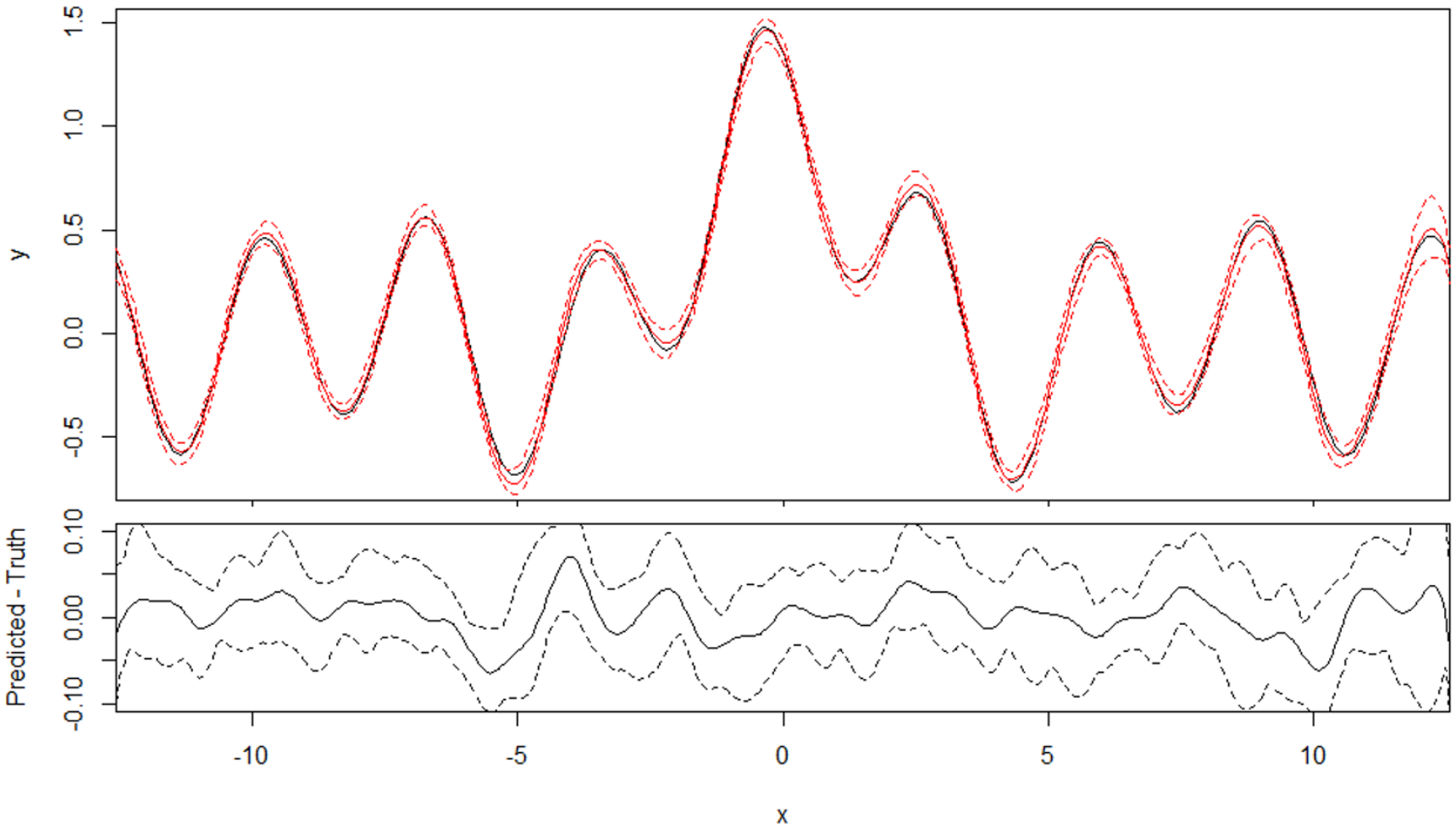


Figure: Fit resulting from using the `mgcv` package along with the bootstrap.

Results - 3

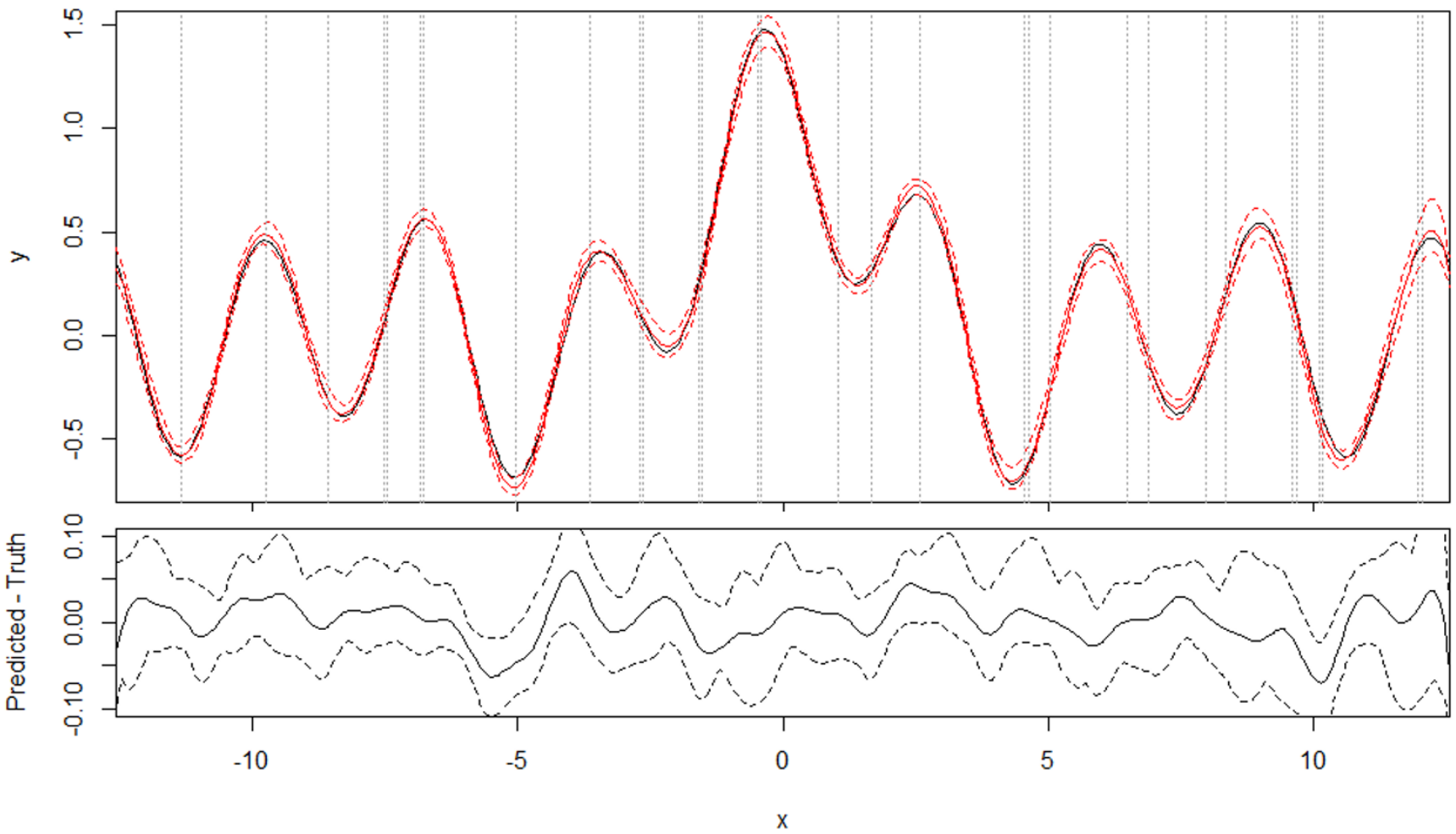
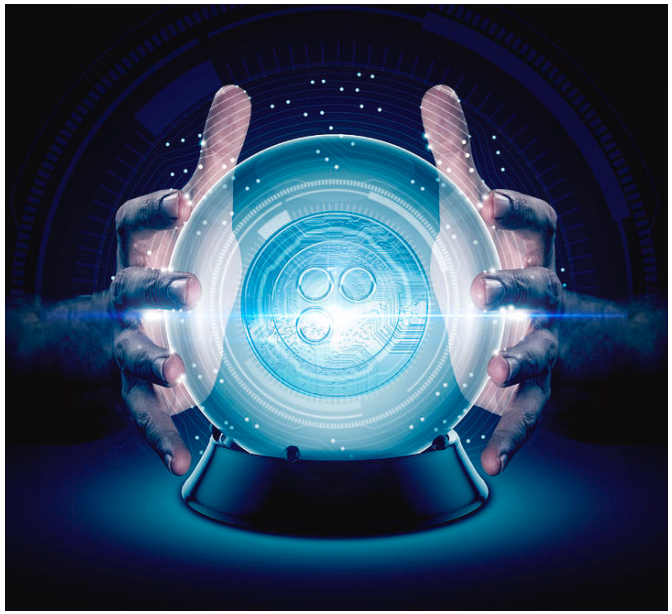


Figure: Fit resulting from using the genetic algorithm and the `mgcv` package along with the bootstrap.

Future Work

- Looking into the future I would like to try running the genetic algorithm on each of the bootstrapped data sets to find the optimum set of knots for each of them.
- I would also like to investigate whether the genetic algorithm could be competitive with the standard method using fewer knots.



Any Questions?