Spline, and other basis function, regression

Mark Andrews
Psychology Department, Nottingham Trent University

✉ `mark.andrews@ntu.ac.uk`

Normal basis function regression

- Polynomial regression can be seen as a type of *basis function* regression.
- In general, in basis function regression where we have one predictor variable x, we model $f(x)$, which is the nonlinear function of x, as a linear sum of K simple functions of x known as basis functions

$$y_i \sim N(\mu_i, \sigma^2), \quad \mu_i = f(x_i) = \beta_0 + \sum_{k=1}^{K} \beta_k \phi_k(x_i), \quad \text{for } i \in 1 \ldots n.$$

- Here, $\phi_1(x_i), \phi_2(x_i) \ldots \phi_k(x_i) \ldots \phi_K(x_i)$ are simple deterministic functions of x_i.

Polynomial basis functions

- In polynomial regression, our basis functions are defined simply as follows:

$$\phi_k(x_i) \triangleq x_i^k.$$

Spline basis functions

- There are many different types of basis functions that are possible to use, but one particularly widely used class of basis functions are *spline* functions.
- The term *spline* is widely used in mathematics, engineering, and computer science and may refer to many different types of related functions, but in the present context, we are defining splines as piecewise polynomial functions that are designed in such a way that each piece or segment of the function joins to the next one without a discontinuity.
- As such, splines are smooth functions composed of multiple pieces, each of which is a polynomial.

Cubic b-splines

- There are many types of spline functions that can be used, but one of the most commonly used types is *cubic b-splines*.
- The *b* refers to *basis* and the *cubic* is the order of the polynomials that make up the pieces.
- Each cubic b-spline basis function is defined by 4 curve segments that join together smoothly.
- The breakpoints between the intervals on which these curves are defined are known as *knots*.
- If these knots are equally spaced apart, then we say that the spline is *uniform*.
- For basis function k, its knots can be stated as

$$t_0^k < t_1^k < t_2^k < t_3^k < t_4^k,$$

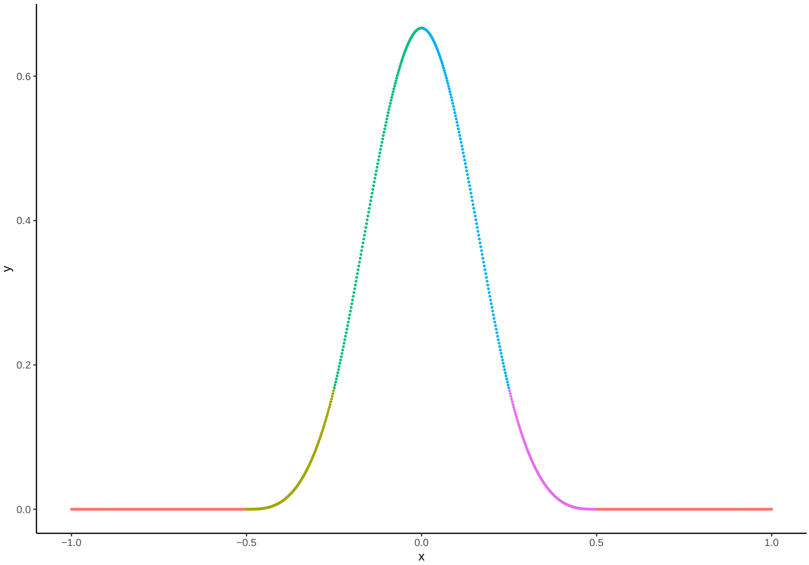
so that the curve segments are defined on the intervals $(t_0^k, t_1^k]$, $(t_1^k, t_2^k]$, $(t_2^k, t_3^k]$, (t_0^3, t_4^k).

Cubic b-splines

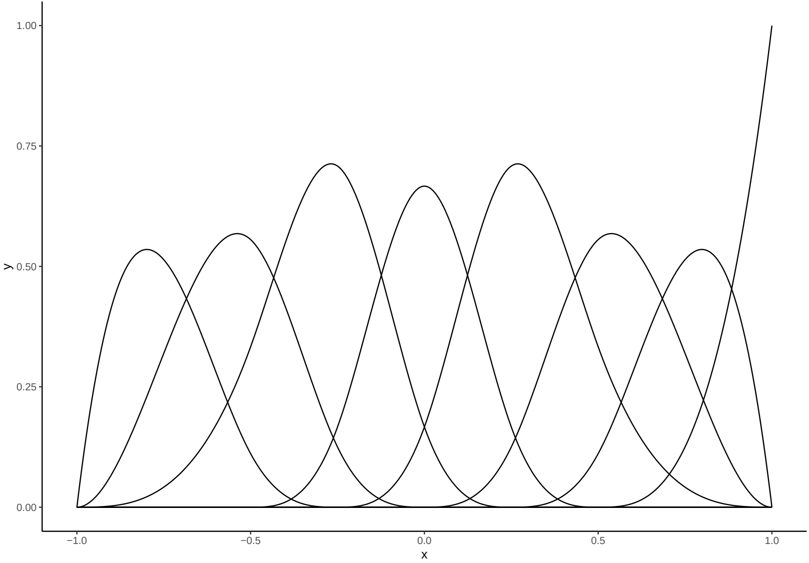
The cubic b-spline is then defined as follows: {

$$\phi_k(x_i) = \begin{cases} \frac{1}{6}u^3, & \text{if } x_i \in (t_0^k, t_1^k], \quad \text{with } u = (x_i - t_0^k)/(t_1^k - t_0^k) \\ \frac{1}{6}(1 + 3u + 3u^2 - 3u^3), & \text{if } x_i \in (t_1^k, t_2^k], \quad \text{with } u = (x_i - t_1^k)/(t_2^k - t_1^k) \\ \frac{1}{6}(4 - 6u^2 + 3u^3), & \text{if } x_i \in (t_2^k, t_3^k], \quad \text{with } u = (x_i - t_2^k)/(t_3^k - t_2^k) \\ \frac{1}{6}(1 - 3u + 3u^2 - u^3), & \text{if } x_i \in (t_3^k, t_4^k), \quad \text{with } u = (x_i - t_3^k)/(t_4^k - t_3^k) \\ 0 & \text{if } x_i < t_0^k \text{ or } x_i > t_4^k \end{cases}$$

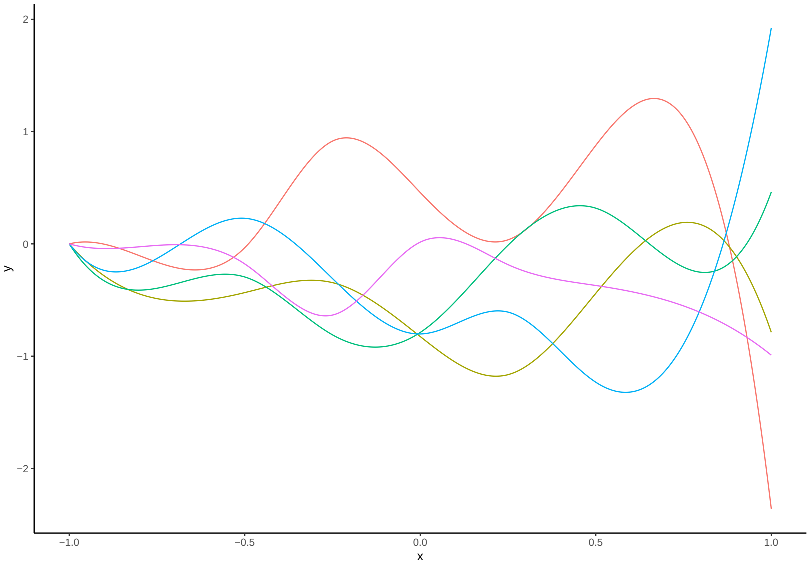
Cubic b-spline example



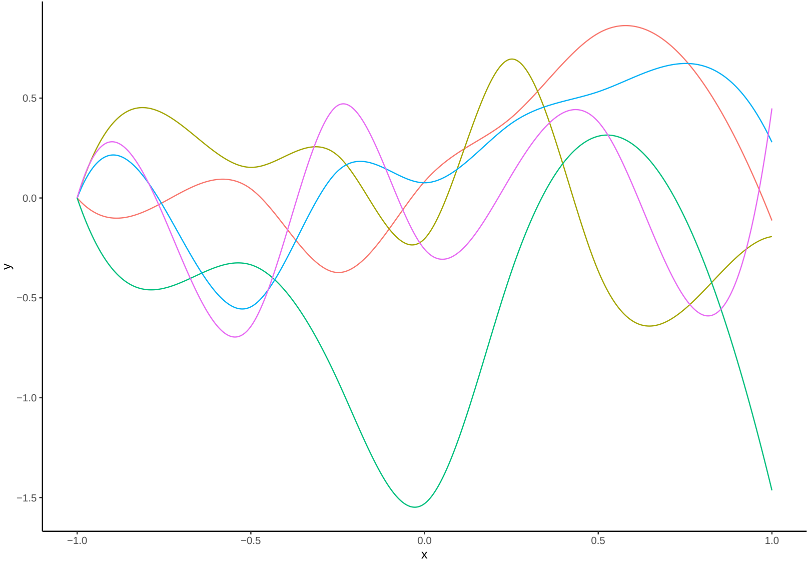
Uniformly spaced b-splines using `splines` *package*



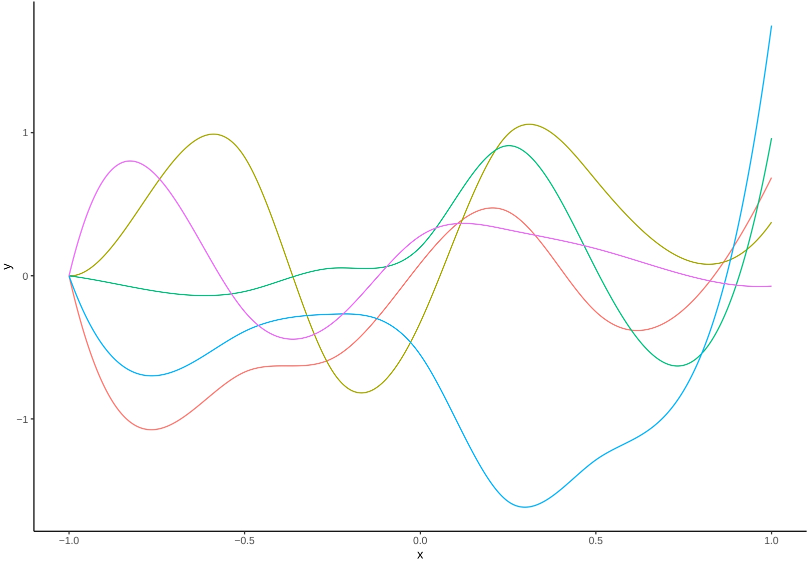
Weighted sums of b-splines: Example 1



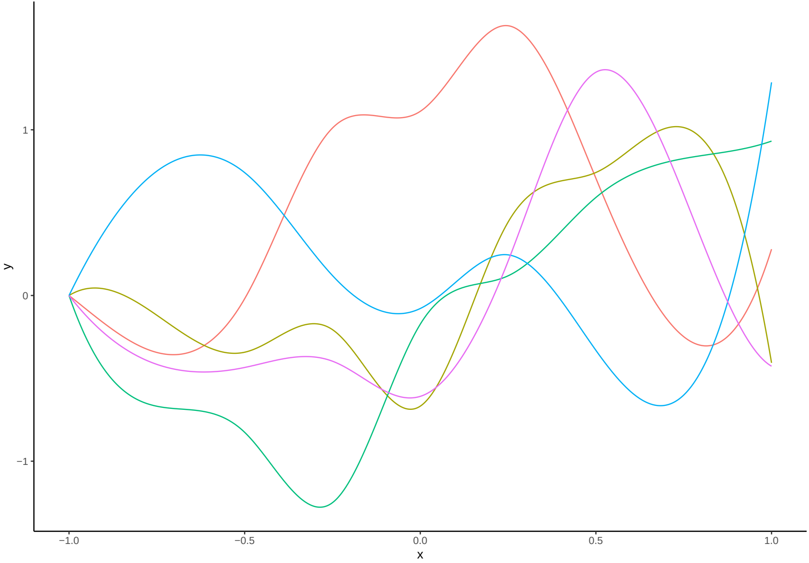
Weighted sums of b-splines: Example 2



Weighted sums of b-splines: Example 3



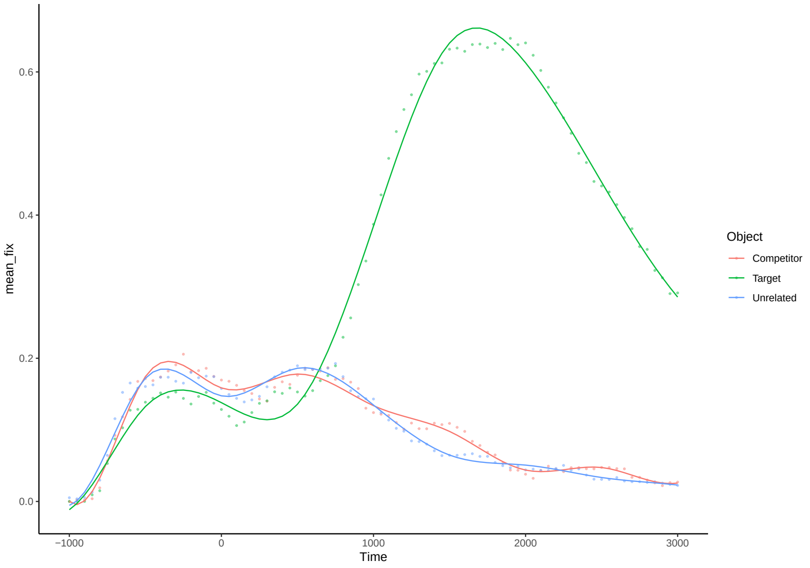
Weighted sums of b-splines: Example 4



Eyetracking data

```r
library(splines)
knots <- seq(-500, 2500, by = 500)
M_bs <- lm(mean_fix ~ bs(Time, knots = knots)*Object,
           data=eyefix_df_avg)
```

Model predictions



Radial basis functions

- An alternative, though related, class of basis functions to spline basis functions are *radial basis functions* (RBF).
- In these basis functions, the value the function takes is defined by the distance of the input value from a fixed center.
- As an example, one of the most commonly used RBF models is the *Gaussian* or *squared exponential* RBF defined as follows.

$$\phi(x) = e^{-\frac{|x-\mu|^2}{2\sigma^2}}.$$

Radial basis function

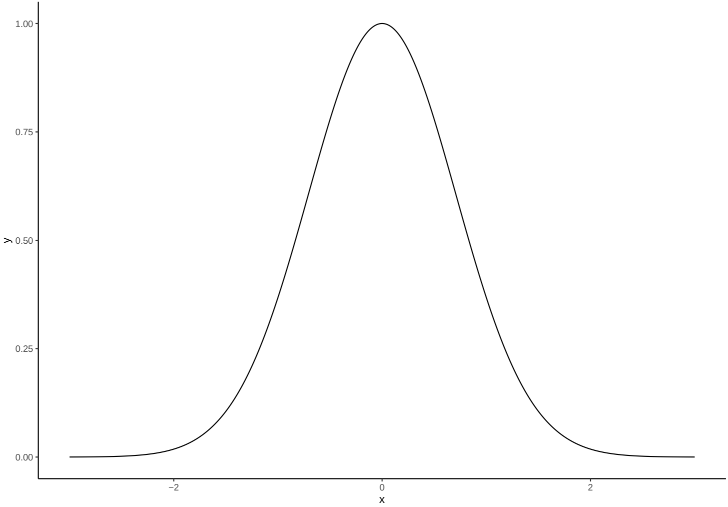


Figure 1: A Gaussian radial basis function (RBF) is essentially an unnormalized Normal distribution. In this figure, we display a Gaussian RBF that is centered at $\mu = 0$ and has a width parameter $\sigma = 1.0$.

Weighted sums of RBFs

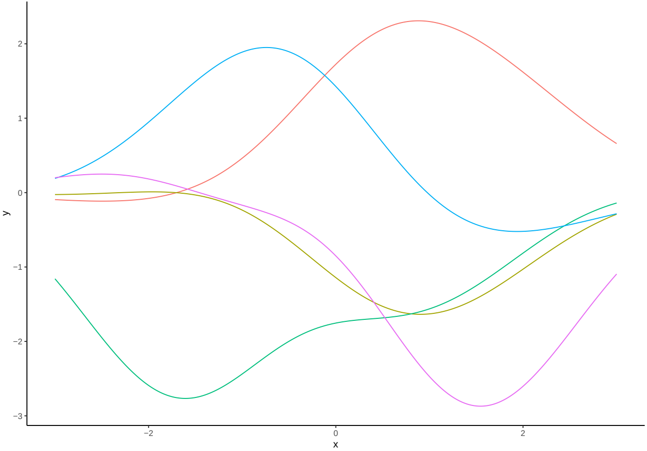


Figure 2: Examples of random sums of Gaussian RBF, with $\sigma = 1$.

Weighted sums of RBFs

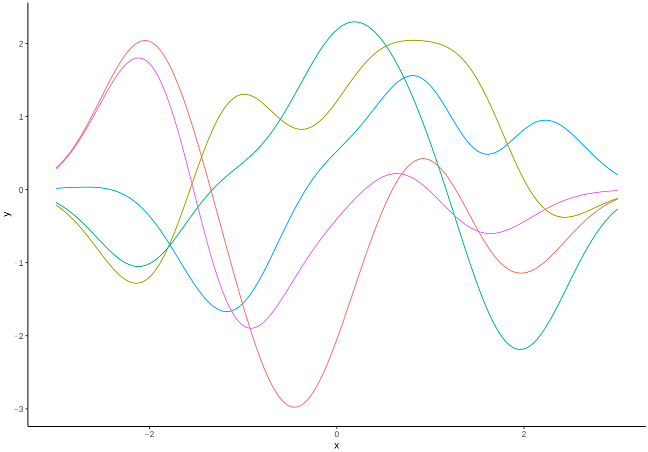


Figure 3: Examples of random sums of Gaussian RBF, with $\sigma = 0.5$.

RBF regression

- We can perform a RBF regression using `lm` similarly to how we used `lm` with `poly` or `splines:bs`.
- To do so, we will create a custom `rbf` function that returns the values of set of Gaussian RBF functions defined at specified centres and with a common width parameter.

```r
rbf <- function(x, centres, sigma = 1.0){
  map(centres,
      ~exp(-(x-.)^2/(2*sigma^2))
  ) %>% do.call(cbind, .)
}
```

RBF regression

- We may then use this `rbf` function inside `lm` by choosing the location of the centres, which we set to be at every 250ms beween -1000 and 3000 ms, and the width parameter, which we set to be 500.

```r
centres <- seq(-1000, 3000, by = 250)
M <-lm(mean_fix ~ rbf(Time, centres, sigma = 500)*Object,
       data=eyefix_df_avg)
```

RBF regression

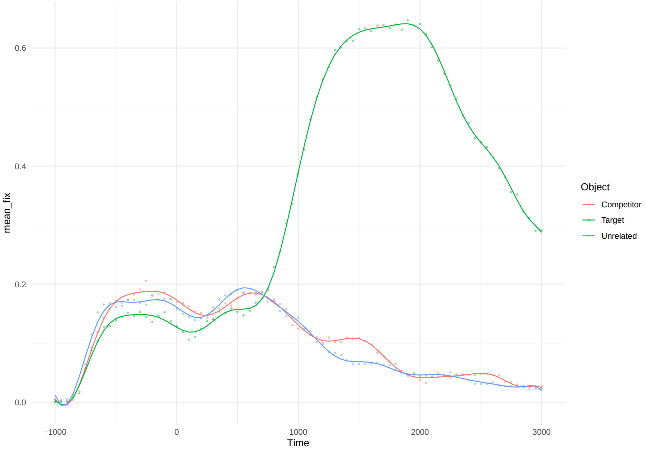


Figure 4: The fit of a Gaussian RBF, with centres at every 250ms and $\sigma = 500$, to the average eye fixation rates to each `Object` category.

Choosing basis function parameters

- A persistent and major issue in basis function regression is choosing between or evaluating the different parameters of the basis functions.
- In the case of cubic b-splines, for example, this would primarily concern the choice of the number and location of the knots.
- Other basis functions, as we will see, have other parameters whose values must also be chosen.
- Although this issue can in principle be treated as just another type of parametric inference, i.e., where the basis function parameters are inferred along with the standard regression coefficients and the standard deviation of the outcome variable, doing so can often be technically very difficult.
- As a result, more commonly, this issue is treated as a model evaluation issue.

GSSvocab problem

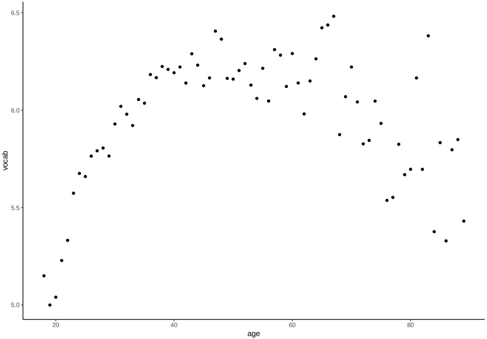


Figure 5: Average score on a vocabulary test for each year of age in a sequences of years from 18 to 89.

Using AIC to select knots

- Let us now fit a sequence of cubic b-spline regression model to this data, where we vary the number of knots from a minimum of 3 to 30.

```
df_seq <- seq(3, 30) %>% set_names(.,.)

M_gssvocab <- map(df_seq,
                  ~lm(vocab ~ ns(age, df = .),
                      data = gssvocab)
)

aic_results <- map_dbl(M_gssvocab, aic_c) %>%
  enframe(name = 'df', value = 'aic')
```