

# Generalized additive models in R

Matteo Fasiolo (University of Bristol, UK)

Virgile Fritsch (Électricité de France R&D)

*matteo.fasiolo@bristol.ac.uk*

July 11, 2018

# Today's plan

## First session

- 1 Intro to Generalized Additive Models (GAMs) in R
- 2 Hands-on session

Coffee break

## Second session

- 1 Beyond mean modelling: GAMLSS and quantile GAMs
- 2 Hands-on session

# Intro to Generalized Additive Models (GAMs)

## Structure:

- 1 What is an additive model?
- 2 Introducing smooth effects
- 3 Big Data GAM methods
- 4 Diagnostics and model selection tools
- 5 GAM modelling using `mgcv` and `mgcViz`

# Structure of the talk

## Structure:

- 1 **What is an additive model?**
- 2 Introducing smooth effects
- 3 Big Data GAM methods
- 4 Diagnostics and model selection tools
- 5 GAM modelling using `mgcv` and `mgcViz`

# What is an additive model

Regression setting:

- $y$  is our response or dependent variable
- $\mathbf{x}$  is a vector of covariates or independent variables

In **distributional regression** we want a good model for  $\text{Dist}(y|\mathbf{x})$ .

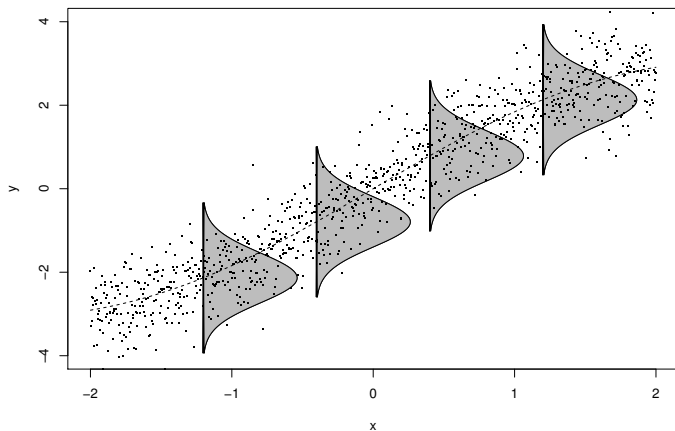
Model is  $\text{Dist}_m\{y|\theta_1(\mathbf{x}), \dots, \theta_q(\mathbf{x})\}$ , where  $\theta_1(\mathbf{x}), \dots, \theta_q(\mathbf{x})$  are param.

In a Gaussian model, the mean depends on the covariates

$$y|\mathbf{x} \sim N\{y|\mu = \theta(\mathbf{x}), \sigma^2\},$$

where  $\mu = \mathbb{E}(y|\mathbf{x})$  and  $\sigma^2 = \text{Var}(y)$ .

# What is an additive model



**Figure:** Gaussian model with variable mean.  
In mgcv: `gam(y~s(x), family=gaussian)`.

# What is an additive model

**Generalized** additive model (GAM):

$$y|\mathbf{x} \sim \text{Distr}\{y|\theta_1 = \mu(\mathbf{x}), \theta_2, \dots, \theta_p\},$$

where

$$\mathbb{E}(y|\mathbf{x}) = \mu(\mathbf{x}) = g^{-1}\left\{\sum_{j=1}^m f_j(\mathbf{x})\right\},$$

and  $g$  is the link function.

$f_j$ 's can be fixed, random or smooth effects with coefficients  $\beta$ .

Poisson GAM:

- $y|\mathbf{x} \sim \text{Pois}\{y|\mu(\mathbf{x})\}$
- $\mathbb{E}(y|\mathbf{x}) = \text{Var}(y|\mathbf{x}) = \exp\left\{\sum_{j=1}^m f_j(\mathbf{x})\right\}$
- $g = \log$  assures  $\mu(\mathbf{x}) > 0$

Here  $\mathbb{E}(y|\mathbf{x})$  and  $\text{Var}(y|\mathbf{x})$  is implied by model...

# What is an additive model

... or we can have extra parameters for scale and shape.

Scaled Student's t GAM:

- $y|\mathbf{x} \sim \text{ScaledStud}\{y|\mu(\mathbf{x}), \sigma, \nu\}$
- $\mathbb{E}(y|\mathbf{x}) = \mu(\mathbf{x}) = \sum_{j=1}^m f_j(\mathbf{x})$
- $\sigma$  is scale parameter
- $\nu$  is shape parameter (degrees of freedom)
- $\text{Var}(y|\mathbf{x}) = \sigma^2 \frac{\nu}{\nu-2}$

Later we'll see models with multiple linear predictors, eg:

- $y|\mathbf{x} \sim N\{y|\mu(\mathbf{x}), \sigma(\mathbf{x})\}$



# Structure of the talk

## Structure:

- 1 What is an additive model?
- 2 **Introducing smooth effects**
- 3 Big Data GAM methods
- 4 Diagnostics and model selection tools
- 5 GAM modelling using `mgcv` and `mgcViz`

# Introducing smooth effects

Consider additive model

$$\mathbb{E}(y|\mathbf{x}) = \mu(\mathbf{x}) = g^{-1}\left\{f_1(\mathbf{x}) + f_2(\mathbf{x}) + \cdots\right\},$$

where

- $f_1(\mathbf{x}) = \beta_0 + \beta_1 x_1 + \beta_2 x_1^2$
- $f_2(\mathbf{x}) = f_2(x_2)$  is a non-linear smooth function

Smooth effects built using spine bases

$$f_2(x_2) = \sum_{k=1}^r \beta_k b_k(x_2)$$

where  $\beta_k$  are unknown coeff and  $b_k(x_2)$  are known spline basis functions.

NB: we call  $\sum_{j=1}^m f_j(\mathbf{x})$  **linear predictor** because it is linear in  $\beta$ .

# Introducing smooth effects

## B-splines:

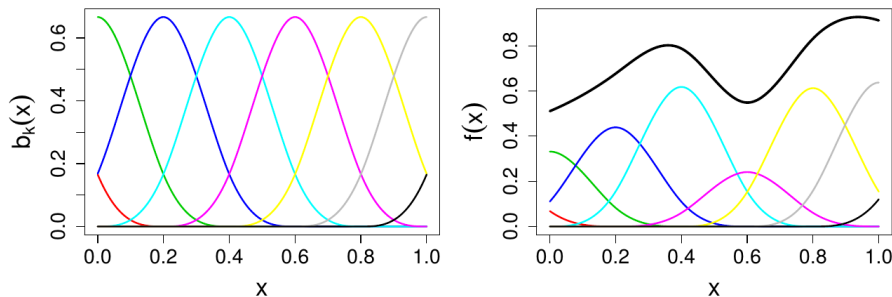
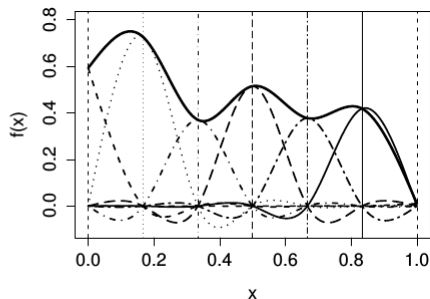
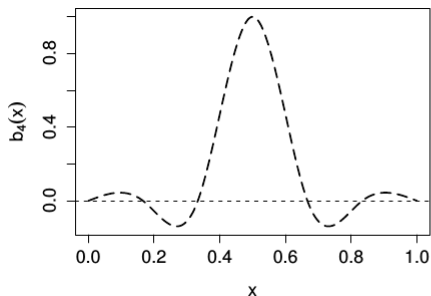


Figure: B-spline basis (left) and smooth (right).

# Types of smooths

`s(x, bs = "cr", k = 20)`

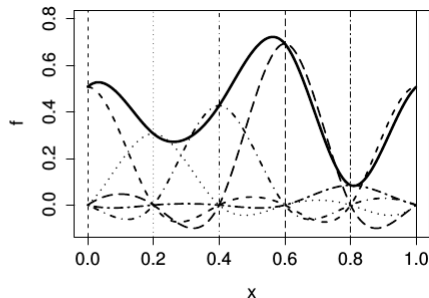
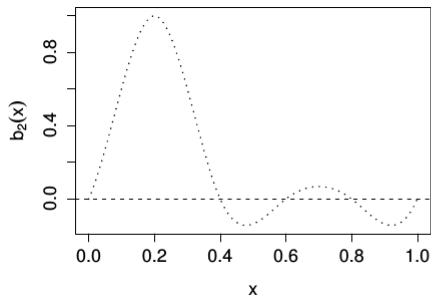


Cubic regression splines are related to the optimal solution to

$$\sum_{i=1}^n \{y_i - f(x_i)\}^2 + \gamma \int f''(x)^2 dx.$$

# Types of smooths

`s(x, bs = "cc")`

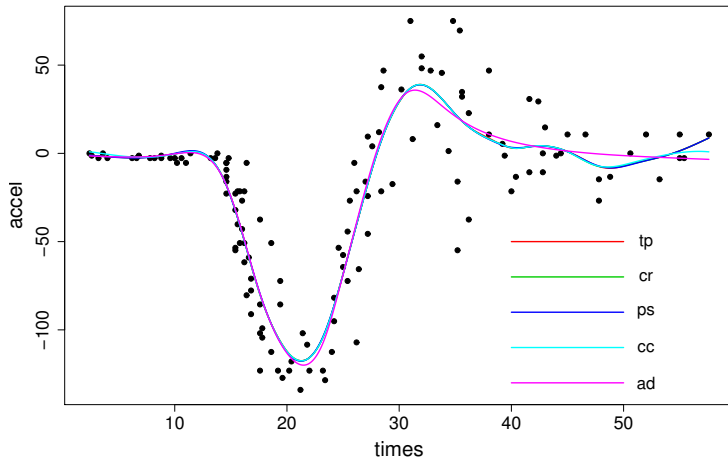


Cyclic cubic regression splines make so that

- $f(x_{min}) = f(x_{max})$
- $f'(x_{min}) = f'(x_{max})$

# Types of smooths

`s(x, bs = "ad")`



The wiggleness or smoothness of  $f(x)$  depends on  $x$ .

# Types of smooths

$s(x_1, x_2), s(x_1, x_2, x_3), \dots$

Based on thin plate regression splines basis.

Related to optimal solution to:

$$\sum_i \{y_i - f(x_i, z_i)\}^2 + \gamma \int f_{xx}^2 + 2f_{xz}^2 + f_{zz}^2 dx dz$$

A single smoothing parameter  $\gamma$ .

Isotropic: same smoothness along  $x_1, x_2, \dots$

# Types of smooths

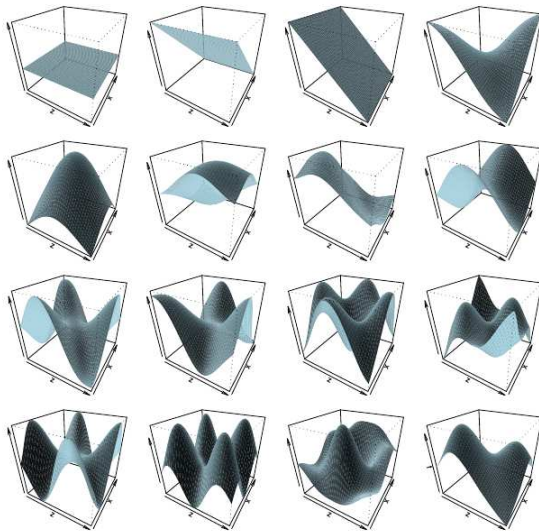


Figure: Rank 17 2D TPRS basis. Courtesy of Simon Wood.

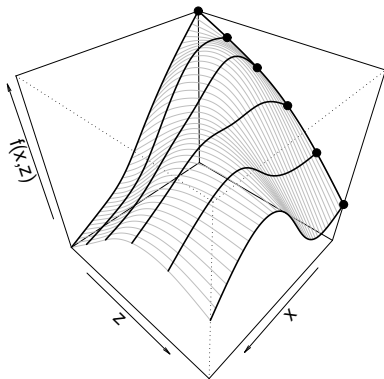
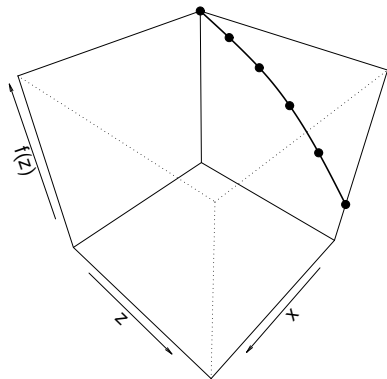


# Types of smooths

Isotropic effect of  $x_1, x_2$  are in same unit (e.g. Km).

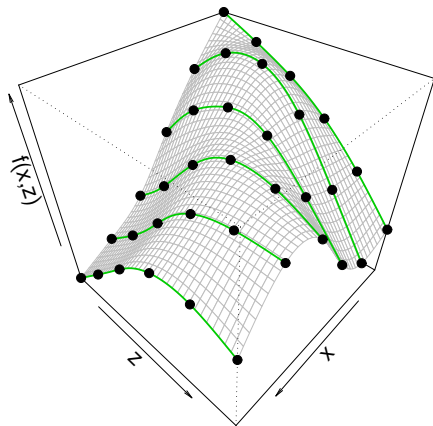
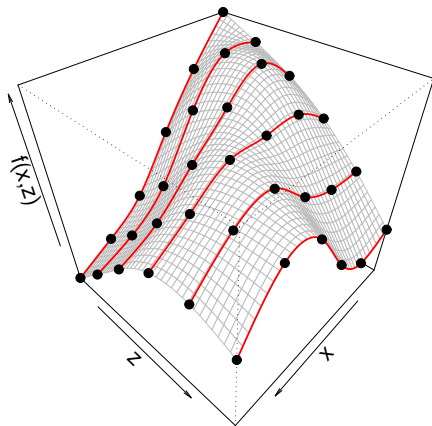
If different units better use tensor product smooths  $\text{te}(x_1, x_2)$ .

Construction: make a spline  $f_z(z)$  a function of  $x$  by letting its coefficients vary smoothly with  $x$



# Types of smooths

- x-penalty: average wiggleness of red curves
- z-penalty: average wiggleness of green curves



# Types of smooths

Can use (almost) any kind of marginal:

- `te(x1, x2, x3)` product of 3 cubic regression splines bases
- `te(x1, x2, bs = c("cc", "cr"), k = c(10, 6))`
- `te(L0, LA, t, d=c(2,1), k=c(20,10), bs=c("tp", "cc"))`

Basis of `te` contains functions of the form  $f(x_1)$  and  $f(x_2)$ .

To fit  $f(x_1) + f(x_2) + f(x_1, x_2)$  separately use:

```
y ~ ti(x1) + ti(x2) + ti(x1, x2)
```

# Types of smooths

## By-factor smooths

Approach (1) is  $s(x, \text{by} = \text{subject})$ , which means

- $\mu(x) = f_1(x) + \dots$  if subject = 1
- $\mu(x) = f_2(x) + \dots$  if subject = 2
- ...

Approach (2) is  $s(x, \text{subject}, \text{bs} = "fs")$ , which means

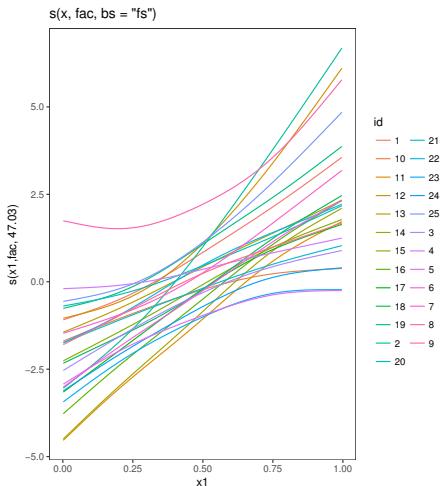
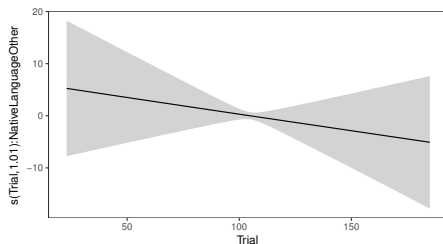
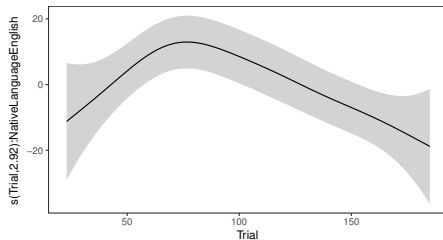
- $\mu(x) = b_1 + f_1(x) + \dots$  if subject = 1
- $\mu(x) = b_2 + f_2(x) + \dots$  if subject = 2
- ...

where  $b_1, b_2, \dots \sim N(0, \gamma_{\mathbf{b}} \mathbf{I})$  are random effects.

In (1) each  $f_j$  has its own smoothing parameter.

In (2) all  $f_j$ 's have the same smoothing parameter.

# Types of smooths



In general

$$f(\mathbf{x}) = \sum_{k=1}^r \beta_k b_k(\mathbf{x}).$$

To determine complexity of  $f(\mathbf{x})$ :

- the basis rank  $r$  is large enough for sufficient flexibility
- a complexity penalty on  $\beta$  controls the wiggleness of the effects

# GAM model fitting

$\hat{\beta}$  is the maximizer of **penalized** log-likelihood

$$\hat{\beta} = \operatorname{argmax}_{\beta} \operatorname{PenLogLik}(\beta|\gamma) = \operatorname{argmax}_{\beta} \left\{ \overbrace{L_y(\beta)}^{\text{goodness of fit}} - \underbrace{\operatorname{Pen}(\beta|\gamma)}_{\text{penalize complexity}} \right\}$$

where:

- $L_y(\beta) = \sum_i \log p(y_i|\beta)$  is log-likelihood
- $\operatorname{Pen}(\beta|\gamma)$  penalizes the complexity of the  $f_j$ 's
- $\gamma > 0$  smoothing parameters ( $\uparrow \gamma \uparrow$  smoothness)

Concrete example  $\mu(\mathbf{x}) = f(x_1) + g(x_2, x_3)$ :

$$\operatorname{Pen}(\beta|\gamma) \approx \gamma_1 \int f_{x_1 x_1}^2 dx_1 + \gamma_2 \int g_{x_2 x_2}^2 + g_{x_3 x_3}^2 + 2g_{x_2 x_3}^2 dx_2 dx_3$$

# GAM model fitting

We use a hierarchical fitting framework:

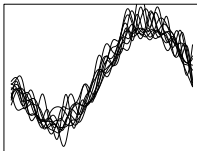
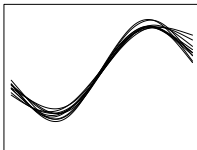
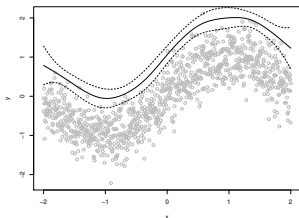
- 1 Select  $\gamma$  determine smoothness

$$\hat{\gamma} = \operatorname{argmax}_{\gamma} \text{LAML}(\gamma)$$

where  $\text{LAML}(\gamma) \approx p(y|\gamma) = \int p(y, \beta|\gamma) d\beta$ .

- 2 For fixed  $\gamma$ , estimate  $\beta$  to determine actual fit

$$\hat{\beta} = \operatorname{argmax}_{\beta} \text{PenLogLik}(\beta|\gamma).$$





# Structure of the talk

## Structure:

- 1 What is an additive model?
- 2 Introducing smooth effects
- 3 **Big Data GAM methods**
- 4 Diagnostics and model selection tools
- 5 GAM modelling using `mgcv` and `mgcViz`

# Big Data methods

Recall the GAM model structure

$$\mu(\mathbf{x}) = \mathbb{E}(y|\mathbf{x}) = g^{-1}\left\{\sum_{j=1}^m f_j(\mathbf{x})\right\}$$

Here  $\mu(\mathbf{x}_i)$  can be written as  $g^{-1}(\mathbf{X}_i\boldsymbol{\beta})$ , where  $\mathbf{X}_i$  row of

$$\mathbf{X} = \begin{bmatrix} A_{11} & A_{12} & \cdots & b_{11}(x_{11}) & b_{12}(x_{11}) & \cdots & b_{21}(x_{21}) & \cdots \\ A_{21} & A_{22} & \cdots & b_{11}(x_{12}) & b_{12}(x_{12}) & \cdots & b_{21}(x_{22}) & \cdots \\ \cdot & \cdot & \cdots & \cdot & \cdot & \cdots & \cdot & \cdots \\ \cdot & \cdot & \cdots & \cdot & \cdot & \cdots & \cdot & \cdots \end{bmatrix}$$

with  $n$  rows and

$$d = p + k_1 + \cdots + k_j + \cdots + k_m$$

columns.

# Big Data methods

Bottom line:  $\mathbf{X}$  can get very big, which causes problems:

- storing  $\mathbf{X}$  takes too much memory
- computing things involving  $\mathbf{X}$  (e.g.  $\mathbf{X}^T \mathbf{X}$ ) takes time

Solution implemented in `mgcv::bam` function:

- do not create  $\mathbf{X}$  but only sub-blocks:

$$\mathbf{X} = \begin{bmatrix} \mathbf{X}_{11} & \mathbf{X}_{12} \\ \mathbf{X}_{21} & \mathbf{X}_{22} \end{bmatrix}$$

do not store them either, but create them when needed;

- any computation involving  $\mathbf{X}$  is based on the blocks;
- use parallelization when possible;

Further acceleration and memory savings by discretization.

Instead of having  $n$  unique rows of **X** discretize to  $b \ll n$  rows.

In mgcv:

```
fit <- bam(y ~ s(x),  
          discrete = TRUE,  
          nthreads = 2,  
          ...)
```

# Structure of the talk

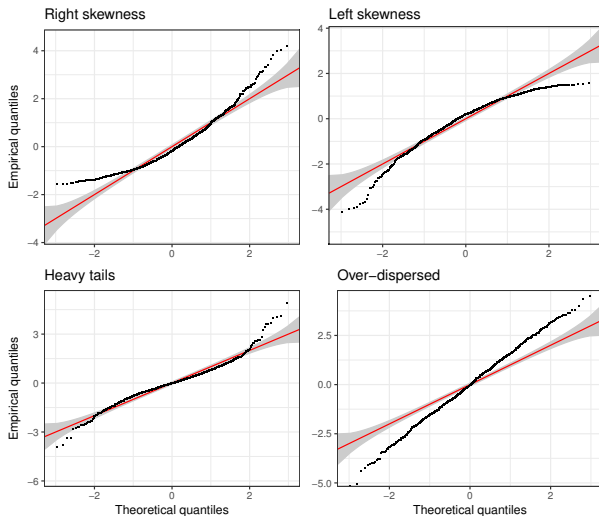
## Structure:

- 1 What is an additive model?
- 2 Introducing smooth effects
- 3 Introducing random effects
- 4 **Diagnostics and model selection tools**
- 5 GAM modelling using `mgcv` and `mgcViz`

# Diagnostics and model selection tools

In first hands-on session we'll use few basic diagnostics.

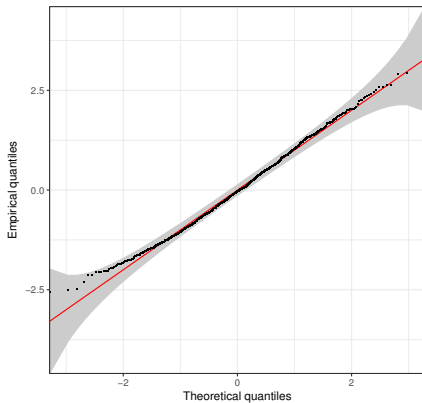
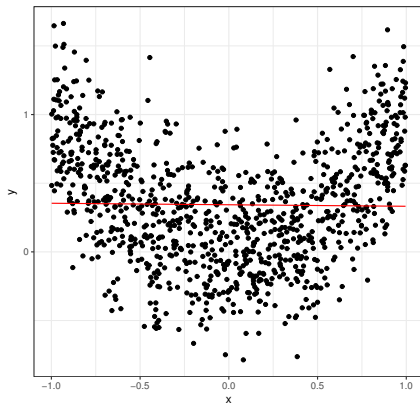
## QQ-plots



# Diagnostics and model selection tools

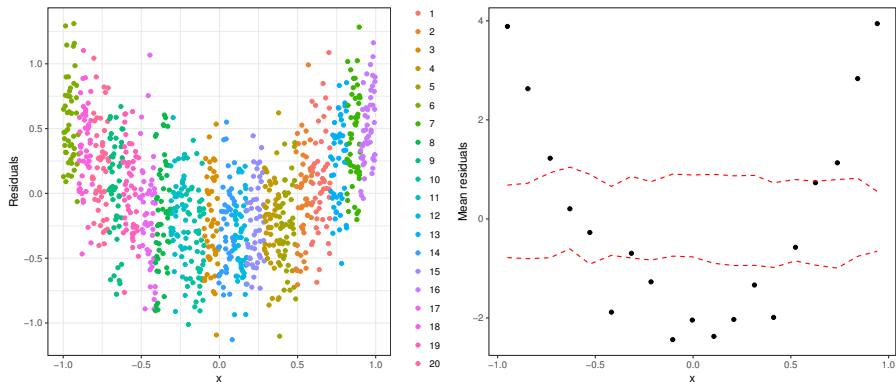
Useful for choosing model  $\text{Dist}_m(y|\mathbf{x})$  (e.g. Poisson vs Neg. Binom.)

Less useful for finding omitted variables and non-linearities.



# Diagnostics and model selection tools

**Conditional residuals checks** are more helpful here.





# Diagnostics and model selection tools

Recall structure of smooth effects:

$$f(\mathbf{x}) = \sum_{j=1}^k \beta_j b_j(\mathbf{x}).$$

where  $\beta$  shrunk toward zero by smoothness penalty.

Effective number of parameters we are using is  $< k$ .

Approximation is **Effective Degrees of Freedom** (EDF)  $< k$ .

By default  $k = 10$  but this is arbitrary.

Exact choice of  $k$  not important, but it must not be too low.

# Diagnostics and model selection tools

Checking whether  $k$  is too low:

- 1 look at conditional residuals checks
- 2 look at output of `check(fit)`:

##		$k'$	edf	k-index	p-value
##	s(wM)	9.00	8.60	0.91	<2e-16 ***
##	s(wM_s95)	9.00	8.13	1.02	0.76
##	s(Posan)	8.00	2.66	1.04	0.97

- 3 increase  $k$  and see if a **model selection criterion** improves

# Diagnostics and model selection tools

## Model selection

General criterion is approximate Akaike Information Criterion (AIC):

$$\text{AIC} = \underbrace{-2 \log p(\mathbf{y}|\hat{\beta})}_{\text{goodness of fit}} + \underbrace{2\tau}_{\text{model complexity}}$$

where  $\tau$  is EDF.

If  $\text{AIC}_{m1} < \text{AIC}_{m2}$  choose model 1.

To select which effects to include we can also look at p-values:

```
summary(fit)
```

##	Estimate	Std. Error	t value	Pr(> t )	
## (Intercept)	267.2004	75.4197	3.543	0.000405	***
## Fl	6.2854	1.0457	6.010	2.20e-09	***
## loc2	79.8459	80.4130	0.993	0.320858	
## loc3	-71.2728	86.1725	-0.827	0.408284	

# Structure of the talk

## Structure:

- 1 What is an additive model?
- 2 Introducing smooth effects
- 3 Introducing random effects
- 4 Diagnostics and model selection tools
- 5 **GAM modelling using mgcv and mgcViz**

# GAMs in mgcv and mgcViz

`mgcv` is the recommended R package for fitting GAMs.

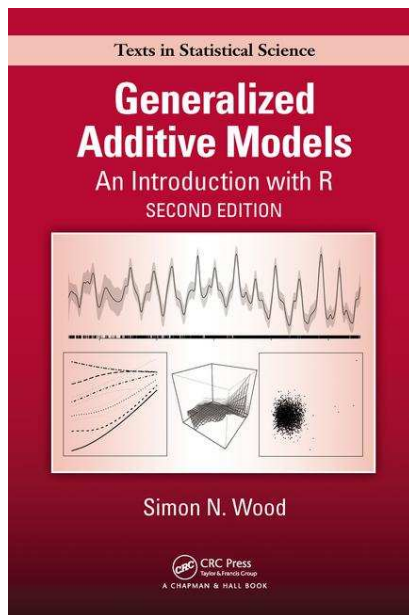
Today we'll work with `mgcViz`'s interface.

`mgcViz` extends `mgcv`'s tools for:

- plotting the estimated effects
- doing visual model checking

But most of the computation is done by `mgcv`.

# Further reading



# References I

Fasiolo, M., Y. Goude, R. Nedellec, and S. N. Wood (2017). Fast calibrated additive quantile regression. *arXiv preprint arXiv:1707.03307*.