Intro to Generalized additive models in R

Matteo Fasiolo (University of Bristol, UK)

matteo.fasiolo@bristol.ac.uk

Material available at:

https://github.com/mfasiolo/workshop_BOZEN19

These slides cover:

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 $\text{Dist}_m\{y|\theta_1(\mathbf{x}), \theta_2, \dots, \theta_q\}$, where $\theta_1, \dots, \theta_q$ are param.

We assume that $\theta_2, \dots, \theta_q$ do not depend on $\mathbf{x}$.

What is an additive model

Gaussian additive model:

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

where

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

and

$$\sigma^2 = \text{Var}(y),$$

f_j's can be fixed, random or smooth effects.

What is an additive model

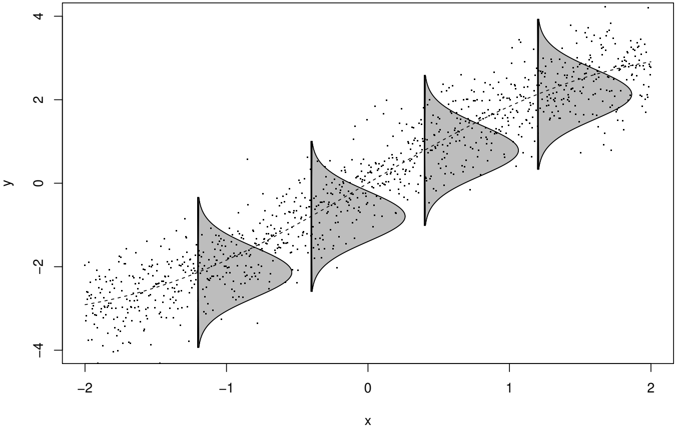


Figure: Gaussian model with variable mean.

In mgcv: `gam(y~s(x), family=gaussian)`.

What is an additive model

Generalized additive model (GAM) (Hastie and Tibshirani, 1990):

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

where

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

and g is the link function.

Poisson GAM:

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

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

Later we'll let all parameters be functions of **x**, eg:

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

These slides cover:

Introducing smooth effects

Consider additive model

$$\mathbb{E}(y|\mathbf{x}) = \mu(\mathbf{x}) = g^{-1}\Big\{f_1(\mathbf{x}) + f_2(\mathbf{x}) + f_3(\mathbf{x})\Big\},$$

where

- $f_1(\mathbf{x}) = \beta_0 + \beta_1 x_1 + \beta_2 x_1^2$
- $f_2(\mathbf{x}) = \begin{cases} 0 & \text{if } x_2 = \text{FALSE} \\ \beta_4 & \text{if } x_2 = \text{TRUE} \end{cases}$
- $f_3(\mathbf{x}) = f_3(x_3)$ is a non-linear smooth function.

Smooth effects built using spline bases

$$f_3(x_3) = \sum_{k=1}^{r} \beta_k b_k(x_3)$$

where β_k are unknown coeff and $b_k(x_3)$ are known spline basis functions.

Introducing smooth effects

Example: B-splines

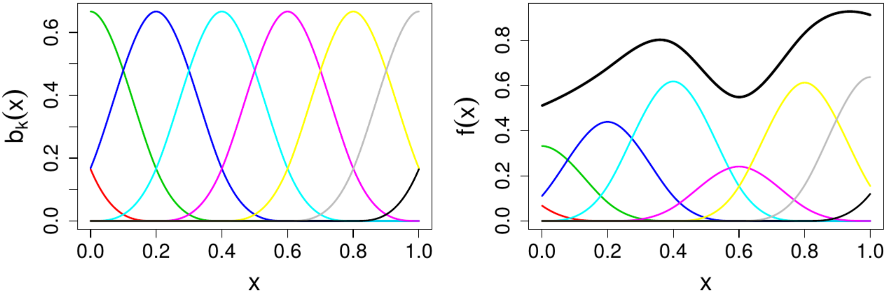
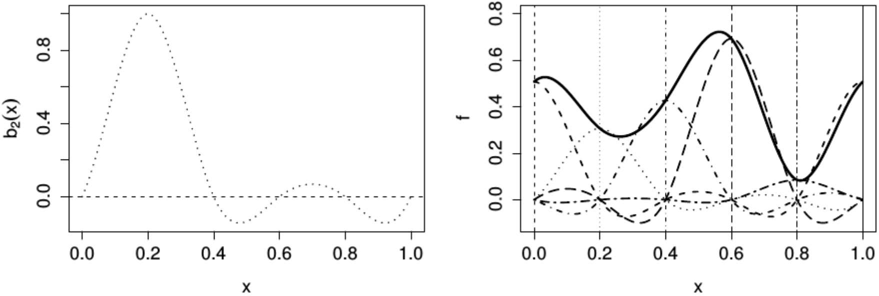


Figure: B-spline basis (left) and smooth (right).

Types of smooths

Example: Cyclic cubic splines



Cyclic cubic regression splines make so that

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

Introducing smooth effects

Example: Thin plate regression splines (TPRS)

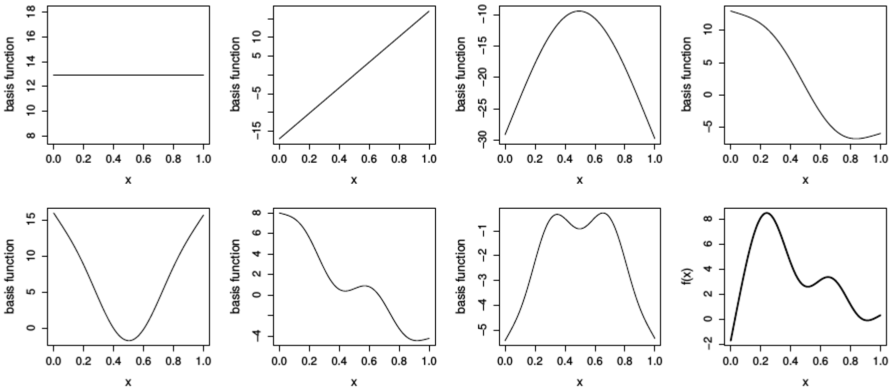


Figure: Rank 7 TPRS basis. Image from Wood (2017).

Introducing smooth effects

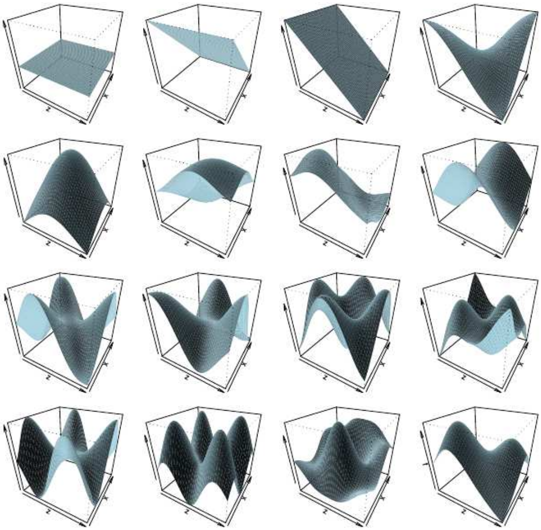


Figure: Rank 17 2D TPRS basis. Courtesy of Simon Wood.

Introducing smooth effects

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 $\boldsymbol{\beta}$ controls the wiggliness of the effects

GAM model fitting

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

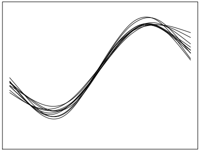
$$\hat{\boldsymbol{\beta}} = \underset{\boldsymbol{\beta}}{\text{argmax}}\ \text{PenLogLik}(\boldsymbol{\beta}|\boldsymbol{\gamma}) = \underset{\boldsymbol{\beta}}{\text{argmax}} \Big\{ \overbrace{\log p(\mathbf{y}|\boldsymbol{\beta})}^{\text{goodness of fit}} - \underbrace{\text{Pen}(\boldsymbol{\beta}|\boldsymbol{\gamma})}_{\text{penalize complexity}} \Big\}$$

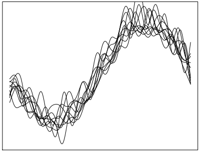
where:

- $\log p(\mathbf{y}|\boldsymbol{\beta}) = \sum_i \log p(y_i|\boldsymbol{\beta})$ is log-likelihood (i.i.d. case)
- $\text{Pen}(\boldsymbol{\beta}|\boldsymbol{\gamma})$ penalizes the complexity of the f_j's
- $\boldsymbol{\gamma} > 0$ smoothing parameters ($\uparrow \boldsymbol{\gamma}$ $\uparrow$smoothness)

GAM model fitting

`mgcv` uses a hierarchical fitting framework:



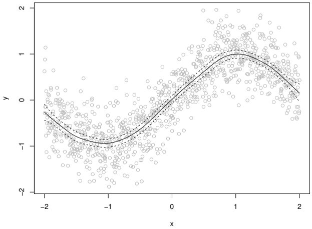


1. Select γ to determine smoothness

$$\hat{\gamma} = \underset{\gamma}{\operatorname{argmax}} \, \mathrm{LAML}(\gamma).$$

2. For fixed γ, estimate $\boldsymbol{\beta}$ to determine actual fit

$$\hat{\boldsymbol{\beta}} = \underset{\boldsymbol{\beta}}{\operatorname{argmax}} \, \mathrm{PenLogLik}(\boldsymbol{\beta}|\gamma).$$



These slides cover:

Introducing random effects

Suppose we have data on bone mineral density (*bmd*) as a function of *age*.

We have m subjects and n data pairs per subject

- subj 1: $\{bmd_{11}, age_{11}\}, \dots, \{bmd_{n1}, age_{n1}\}$
- subj j: $\{bmd_{1j}, age_{1j}\}, \dots, \{bmd_{nj}, age_{nj}\}$
- subj m: $\{bmd_{1m}, age_{1m}\}, \dots, \{bmd_{nm}, age_{nm}\}$

Standard linear model ignores individual differences

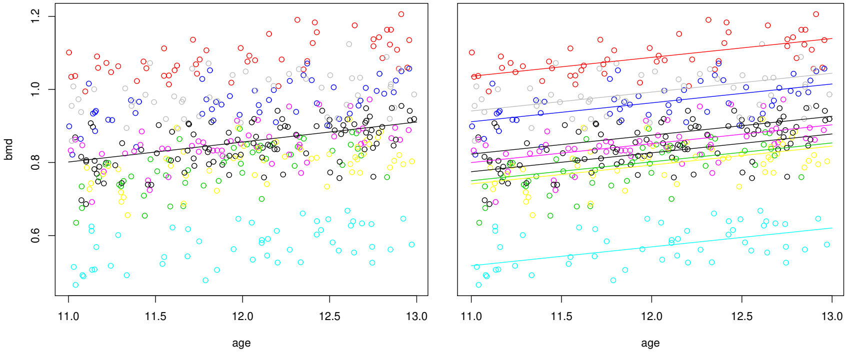
$$\mathbb{E}(bmd|age_{ij}) = \mu(age_{ij}) = \alpha + \beta age_{ij}.$$

We can include random intercept per subject

$$\mu(age_{ij}) = \alpha + \beta age_{ij} + \mathrm{a}_j,$$

where $\mathbf{a} = \{\mathrm{a}_1, \dots, \mathrm{a}_m\} \sim N(\mathbf{0}, \boldsymbol{\Sigma})$.

Introducing random effects



We can also include random slopes

$$\mu(age_{ij}) = \alpha + (\beta + b_j)age_{ij} + a_j,$$

where $\mathbf{a} \sim N(\mathbf{0}, \boldsymbol{\Sigma}_{\mathbf{a}})$ and $\mathbf{b} \sim N(\mathbf{0}, \boldsymbol{\Sigma}_{\boldsymbol{b}})$.

Introducing random effects

In `mgcv` random effect are specified as:

```
gam(bmd ~ 1 + s(subject, bs = "re") +
          age + s(age, subject, bs = "re"), ...)
```

In simplest case $\boldsymbol{\Sigma}_{\mathbf{a}} = \gamma_{\mathbf{a}}\mathbf{I}$ and $\boldsymbol{\Sigma}_{\mathbf{b}} = \gamma_{\mathbf{b}}\mathbf{I}$, that is

$$\boldsymbol{\Sigma}_{\mathbf{a}} = \begin{bmatrix} \gamma_{\mathbf{a}} & 0 & 0 & \dots & 0 \\ 0 & \gamma_{\mathbf{a}} & 0 & \dots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \dots & \gamma_{\mathbf{a}} \end{bmatrix}$$

Variances $\gamma_{\mathbf{a}}$ and $\gamma_{\mathbf{b}}$ must be estimated (later I'll explain how).

Introducing random effects

Using the gam function in mgcv we can create and estimate simple random effects:

```
gam(bmd ~ 1 + s(subject, bs = "re") +
          age + s(age, subject, bs = "re"), ...)
```

For more complex random effects (e.g. introducing correlation between random slopes and intercepts) we can use:

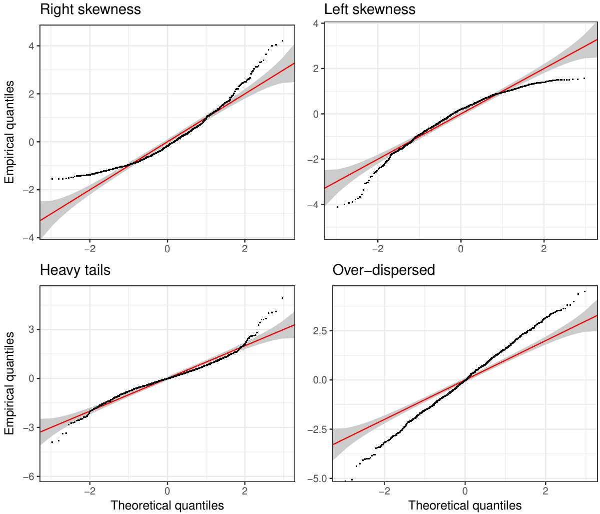
- gamm function estimates generalized additive mixed models using nlme
- gamm4 package based on lme4 package

Here we will use only simple effects.

These slides cover:

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

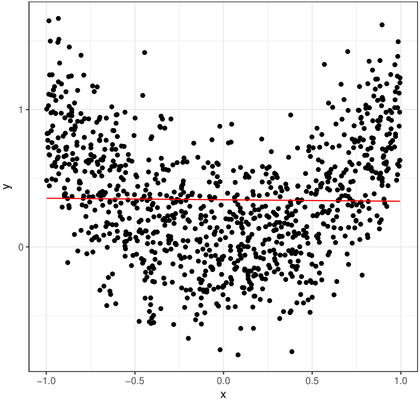
QQ-plots

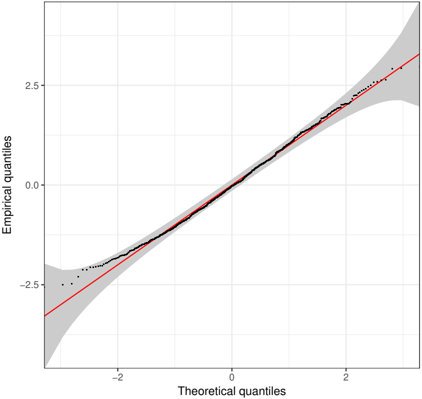


Diagnostics and model selection tools

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

Recall structure of smooth effects:

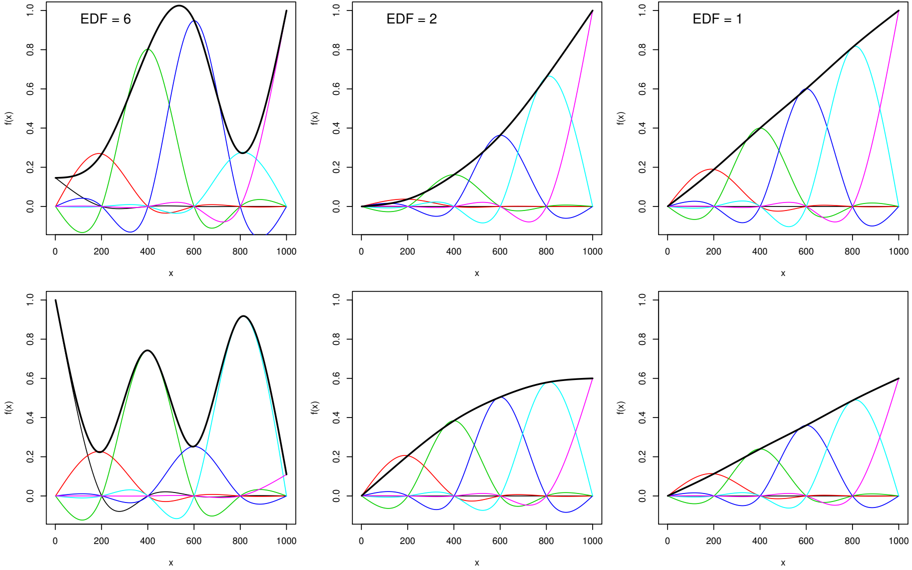
$$f(\mathbf{x}) = \sum_{j=1}^{k} \beta_j b_j(\mathbf{x}).$$

where $\boldsymbol{\beta}$ shrunk toward smoothness by penalty.

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

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

Diagnostics and model selection tools



Diagnostics and model selection tools

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

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

Checking whether k is too low:

1. look at conditional residuals checks
2. look at output of `gam.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

Model selection

General criterion is approximate Akaike Information Criterion (AIC):

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

where τ 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
```

These slides cover:

1. What is an additive model?
2. What is an additive model?
3. Introducing smooth effects
4. Introducing random effects
5. Diagnostics and model selection tools
6. GAM modelling in mgcv

GAMs in mgcv

The exercises will be based on the `mgcv` package for GAM modelling.

`mgcv` is a recommended R package, included in R by default.

It contains methods for:

- creating GAM models
- fitting them
- visualizing and summarizing model output

GAMs in mgcv

There are alternatives to `mgcv`, such as:

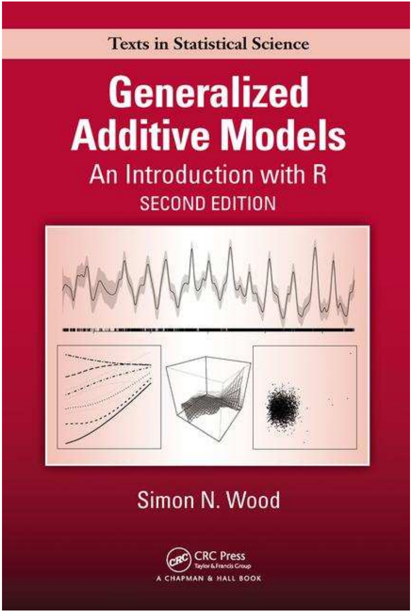
- `mboost` (Hothorn et al., 2010)
- `gamlss` (Rigby and Stasinopoulos, 2005)
- `brms` (Bürkner et al., 2017)
- `BayesX` (Brezger et al., 2003)
- `INLA` (Rue et al., 2009)

Each offers much flexibility (e.g. smooth effects types and distributions).

Strong points of `mgcv`'s methods:

1. little tuning needed (automatic smoothing parameters selection)
2. efficient and stable numerical implementation

Further reading



References I

Brezger, A., T. Kneib, and S. Lang (2003). Bayesx: Analysing bayesian structured additive regression models. Technical report, Discussion paper//Sonderforschungsbereich 386 der Ludwig-Maximilians .

Bürkner, P. C. et al. (2017). brms: An r package for bayesian multilevel models using stan. *Journal of Statistical Software 80*(1), 1–28.

Hastie, T. and R. Tibshirani (1990). *Generalized Additive Models*, Volume 43. CRC Press.

Hothorn, T., P. Bühlmann, T. Kneib, M. Schmid, and B. Hofner (2010). Model-based boosting 2.0. *The Journal of Machine Learning Research 11*, 2109–2113.

Rigby, R. A. and D. M. Stasinopoulos (2005). Generalized additive models for location, scale and shape. *Journal of the Royal Statistical Society: Series C (Applied Statistics) 54*(3), 507–554.

Rue, H., S. Martino, and N. Chopin (2009). Approximate bayesian inference for latent gaussian models by using integrated nested laplace approximations. *Journal of the royal statistical society: Series b (statistical methodology) 71*(2), 319–392.

Wood, S. (2017). *Generalized additive models: an introduction with R*. CRC press.