## Flexible Regression

Session 3 - GAMs
Notes:https://warwick.ac.uk/fac/sci/statistics/ apts/students/resources/

Slides:www.stats.gla.ac.uk/~claire/APTS_FR_session_
3.pdf

Claire Miller \& Tereza Neocleous

## Session 1 - nonparametric regression summary

Slides:www.stats.gla.ac.uk/~claire/APTS_FR_session_1.pdf

$$
Y_{i}=f\left(x_{i}\right)+\varepsilon_{i}, \quad \varepsilon_{i} \sim \mathrm{~N}\left(0, \sigma^{2}\right)
$$

- Estimate $f()$ using a regression framework: $\hat{\mathbf{y}}=\mathbf{B} \hat{\boldsymbol{\beta}}$;
- Regression splines fit: $\sum_{i=1}^{n}\left(y_{i}-f\left(x_{i}\right)\right)^{2}$

$$
\hat{\boldsymbol{\beta}}=\left(\mathbf{B}^{\top} \mathbf{B}\right)^{-1} \mathbf{B}^{\top} \mathbf{y}
$$

- Penalised regression splines fit: $\sum_{i=1}^{n}\left(y_{i}-f\left(x_{i}\right)\right)^{2}+\lambda\|\mathbf{D} \boldsymbol{\beta}\|^{2}$

$$
\hat{\boldsymbol{\beta}}=\left(\mathbf{B}^{\top} \mathbf{B}+\lambda \mathbf{D}^{\top} \mathbf{D}\right)^{-1} \mathbf{B}^{\top} \mathbf{y}
$$

## Session 1 - nonparametric regression summary

$$
Y_{i}=f\left(x_{i}\right)+\varepsilon_{i}
$$

- Estimate $f()$ using a regression framework: $\hat{\mathbf{y}}=\mathbf{B} \hat{\boldsymbol{\beta}}$
- Regression splines fit: $\hat{\boldsymbol{\beta}}=\left(\mathbf{B}^{\top} \mathbf{B}\right)^{-1} \mathbf{B}^{\top} \mathbf{y}$
- Level of smoothing determined by number of basis functions (number of knots and degree (3))
- Penalised regression splines fit: $\hat{\boldsymbol{\beta}}=\left(\mathbf{B}^{\top} \mathbf{B}+\lambda \mathbf{D}^{\top} \mathbf{D}\right)^{-1} \mathbf{B}^{\top} \mathbf{y}$
- Level of smoothing determined by using 'too many' basis functions (number of knots and degree (3)) and smoothing through $\lambda$.


## Session 1 - nonparametric regression

library (mgcv)
model <- gam(Rc.age~s(Cal.age), data=radiocarbon)
model
plot(model, residuals=TRUE)


## What's in this session?

- How much to smooth?
- How to select smoothing parameters?
- Nonparametric regression in higher dimensions
- (Generalised) Additive Models


### 4.1 How much to smooth?

Fitted values can be expressed as:

$$
\hat{\mathbf{y}}=\hat{f}=\mathbf{S} \mathbf{y}
$$

Define: degrees of freedom for model:

$$
\mathrm{df}_{\mathrm{mod}}=\operatorname{tr}\{\mathbf{S}\} .
$$

### 4.1 How much to smooth?

Regression spline

$$
\mathbf{S}=\mathbf{B}\left(\mathbf{B}^{\top} \mathbf{B}\right)^{-1} \mathbf{B}^{\top}
$$

Penalised regression splines

$$
\mathbf{S}_{\lambda}=\mathbf{B}\left(\mathbf{B}^{\top} \mathbf{B}+\lambda \mathbf{D}^{\top} \mathbf{D}\right)^{-1} \mathbf{B}^{\top}
$$

Define effective degrees of freedom:

$$
\operatorname{edf}_{\bmod (\lambda)}=\operatorname{tr}\left(\mathbf{S}_{\lambda}\right)
$$

### 4.1 How much to smooth?



Figure: Radiocarbon data with fit from local linear regression with four different degrees of freedom

### 4.1 How much to smooth?



Figure: Radiocarbon data with fit from local linear regression with four different degrees of freedom

### 4.1 How much to smooth?



Figure: Radiocarbon data with fit from local linear regression with four different degrees of freedom

### 4.1 How much to smooth?



Figure: Radiocarbon data with fit from local linear regression with four different degrees of freedom

### 4.1 How much to smooth?

Error variance

$$
\begin{gathered}
\mathrm{RSS}=\sum\left\{y_{i}-\hat{f}\left(x_{i}\right)\right\}^{2} \\
\hat{\sigma}^{2}=\mathrm{RSS} / \mathrm{df}_{\mathrm{err}}
\end{gathered}
$$

$$
\mathrm{df}_{\mathrm{err}}=n-\operatorname{tr}(\mathbf{S}) \text { if } \mathbf{S}^{\top}=\mathbf{S} \text { and } \mathbf{S}^{2}=\mathbf{S}
$$

### 4.1 How much to smooth?

## Standard errors

$$
\operatorname{Var}\{\hat{f}\}=\operatorname{Var}\{\mathbf{S} \mathbf{y}\}=\mathbf{S S}^{\top} \sigma^{2}
$$

and so, by plugging in $\sqrt{\mathbf{S S}^{\top} \hat{\sigma}^{2}}{ }_{i i}$ the standard errors at each evaluation point are obtained.


### 4.2 Automatic methods for smoothing

- We can use the criteria (AIC, AICc, BIC, GCV, CV, ...) to automatically select smoothing parameters.
- General tendencies:
- AIC and cross-validation tend to overfit.
- BIC tends to underfit.
- For penalised regression spline models a mixed-model approach or a Bayesian approach for estimating / averaging over the smoothing parameter (to follow....).


## Selecting $\lambda$ by GCV - Radiocarbon dating



$\lambda=0.07$ selected as the smoothing parameter in a penalised regression fit.

### 4.2.1 Random effects interpretation

- We can interpret the penalised regression spline model (2.2) as a random effects model

$$
\begin{gathered}
\sum_{i=1}^{n}\left(y_{i}-f\left(x_{i}\right)\right)^{2}+\lambda\|\mathbf{D} \boldsymbol{\beta}\|^{2} \\
\|\mathbf{y}-\mathbf{B} \boldsymbol{\beta}\|^{2}+\lambda\|\mathbf{D} \boldsymbol{\beta}\|^{2}
\end{gathered}
$$

- We need to "split" $\boldsymbol{\beta}$ into an unpenalised fixed effect and a penalised random effect.
- Benefit: We can use mixed-model (REML) to estimate $\lambda=\frac{\sigma^{2}}{\tau^{2}}$.


### 4.2.1 Random effects interpretation

library (mgcv)
model <- gam(Rc.age~s(Cal.age), method="REML")


Comparison of automatic smoothing methods

| Method | GCV | REML | ML |
| :---: | :---: | :---: | :---: |
| edf | 7.56 | 7.44 | 7.42 |

### 4.2.2 Bayesian point-of-view

- Alternatively treat as a fully Bayesian model with priors on $\sigma^{2}$ and $\tau^{2}$ :

$$
\begin{aligned}
\mathbf{D} \boldsymbol{\beta} \mid \tau^{2} & \sim \mathrm{~N}\left(\mathbf{0}, \tau^{2} \mathbf{I}\right) \\
\mathbf{y} \mid \boldsymbol{\beta}, \sigma^{2} & \sim \mathrm{~N}\left(\mathbf{B} \boldsymbol{\beta}, \sigma^{2} \mathbf{I}\right) \\
\sigma^{2} & \sim \mathrm{IG}\left(a_{\sigma^{2}}, b_{\sigma^{2}}\right) \\
\tau^{2} & \sim \mathrm{IG}\left(a_{\tau^{2}}, b_{\tau^{2}}\right)
\end{aligned}
$$

- Inference can be done by a Gibbs sampler (BayesX)


### 4.3 Nonparametric regression in higher dimensions

We want to develop a spline basis for a model of the form

$$
\mathbb{E}\left(Y_{i}\right)=f\left(x_{i 1}, x_{i 2}\right)
$$

### 4.3.2 Tensor-product splines

- We will use the following strategy.
- Place a basis on each dimension separately. $\rightsquigarrow$ Two bases

$$
\left(B_{1}^{(1)}\left(x_{1}\right), \ldots, B_{l_{1}+r-1}^{(1)}\right) \text { and }\left(B_{1}^{(2)}\left(x_{1}\right), \ldots, B_{l_{2}+r-1}^{(1)}\right)
$$

- Define bivariate-basis functions as

$$
B_{j k}\left(x_{1}, x_{2}\right)=B_{j}^{(1)}\left(x_{1}\right) \cdot B_{k}^{(2)}\left(x_{2}\right)
$$

for $j \in 1, \ldots, l_{1}+r-1$ and $k \in 1, \ldots, l_{2}+r-1$.
4.3.2 Tensor-product splines: basis degree 0

4.3.2 Tensor-product splines: basis degree 1


### 4.3.2 Tensor-product splines: basis degree 2


4.3.2 Tensor-product splines: basis degree 3


### 4.3.2 Tensor-product splines: entire basis



6 basis functions for each dimension
$\rightsquigarrow 36=6^{2}$ basis functions for the bivariate surface

### 4.3.2 Tensor-product splines: model fitting

- We will now use the basis expansion

$$
f\left(x_{i 1}, x_{i 2}\right)=\sum_{j=1}^{l_{1}+r-1} \beta_{j k} B_{j k}\left(x_{1}, x_{2}\right)
$$

- This corresponds to the design matrix
$\mathbf{B}=\left(\begin{array}{cccccc}B_{11}\left(x_{11}, x_{12}\right) & \ldots & B_{1, l_{2}+r-1}\left(x_{11}, x_{12}\right) & B_{21}\left(x_{11}, x_{12}\right) & \ldots & B_{l_{1}+r-1, l+2+r-1}\left(x_{11}, x_{12}\right) \\ \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ B_{11}\left(x_{n 1}, x_{n 2}\right) & \ldots & B_{1, l_{2}+r-1}\left(x_{n 1}, x_{n 1}\right) & B_{21}\left(x_{11}, x_{12}\right) & \ldots & B_{l_{1}+r-1, l+2+r-1}\left(x_{n 1}, x_{n 2}\right)\end{array}\right)$
- We apply univariate penalties to the "rows" and "columns" of the bivariate basis.


### 4.3.2 Tensor-product splines: Great Barrier Reef



### 4.3.2 Thin-plate splines - an alternative

Advantage: only one smoothing parameter is estimated (isotrophic smoothness assumption).

Thin-plate splines are the default in mgcv's function gam.
model <- gam(Score1~s(Latitude, Longitude), data=trawl) vis.gam(model, plot.type="contour")

### 4.3.2 Thin-plate splines: Great Barrier Reef

linear predictor


### 4.3.2 Thin plate splines (typo) page 88

In fact, we need to minimise the objective function

$$
\sum_{i=1}^{n}\left(y_{i}-f\left(x_{i 1}, x_{i 2}\right)\right)^{2}+\lambda \boldsymbol{\beta}^{\prime} \mathbf{R} \boldsymbol{\beta}
$$

subject to the constraints that
$\sum_{i=1}^{n} \beta_{2+i}=\sum_{i=1}^{n} x_{i 1} \beta_{2+i}=\sum_{i=1}^{n} x_{i 2} \beta_{2+i}=0$, where

$$
\mathbf{R}=\left(\begin{array}{cccccc}
0 & 0 & 0 & 0 & \cdots & 0 \\
0 & 0 & 0 & 0 & \cdots & 0 \\
0 & 0 & 0 & 0 \\
0 & 0 & 0 & \kappa\left(\left(x_{11}, x_{12}\right),\left(x_{11}, x_{12}\right)\right) & \cdots & \kappa\left(\left(x_{11}, x_{12}\right),\left(x_{n 1}, x_{n 2}\right)\right) \\
\vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\
0 & 0 & 0 & \kappa\left(\left(x_{n 11}, x_{n 2}\right),\left(x_{11}, x_{12}\right)\right) & \cdots & \kappa\left(\left(\left(x_{n 1} 1, x_{n 2}\right),\left(x_{n 1}, x_{n 2}\right)\right)\right.
\end{array}\right)
$$

### 4.4 Additive models

$$
Y_{i}=\beta_{0}+f_{1}\left(x_{1 i}\right)+\ldots+f_{p}\left(x_{p i}\right)+\varepsilon_{i}, \quad i=1, \ldots, n,
$$

where the $f_{i}$ are functions whose shapes are unrestricted, apart from an assumption of smoothness.

We can have:

- More than one covariate;
- Smooth functions can be univariate, bivariate,......;
- Computational challenges can arise for higher dimensions.

Consider the case of only two covariates,

$$
Y_{i}=\beta_{0}+f_{1}\left(x_{1 i}\right)+f_{2}\left(x_{2 i}\right)+\varepsilon_{i}, \quad i=1, \ldots, n
$$

### 4.4 Additive models

A rearrangement of this as:

$$
y_{i}-\beta_{0}-f_{2}\left(x_{2 i}\right)=f_{1}\left(x_{1 i}\right)+\varepsilon_{i}
$$

suggests that an estimate of component $f_{1}$ can then be obtained by smoothing the residuals of the data after fitting $\hat{f}_{2}$,

$$
\hat{f}_{1}=S_{1}\left(\mathbf{y}-\overline{\mathbf{y}}-\hat{f}_{2}\right)
$$

and that, similarly, subsequent estimates of $f_{2}$ can be obtained.
$\rightsquigarrow$ the backfitting algorithm.

### 4.4 Additive models

If a spline basis is used, then the backfitting algorithm is not required as we have a form of linear model with a penalty term

$$
Y_{i}=\mathbf{B} \boldsymbol{\beta}+\varepsilon_{i}
$$

The model is fitted by choosing the vector of weights $\boldsymbol{\beta}$ to minimise

$$
(\mathbf{y}-\mathbf{B} \boldsymbol{\beta})^{\top}(\mathbf{y}-\mathbf{B} \boldsymbol{\beta})+\boldsymbol{\beta}^{\top} P \boldsymbol{\beta}
$$

where the penalty matrix $P$ is of block-diagonal form, constructed from the penalties from the individual model components, with the $j$ th component $\lambda_{j} \mathbf{D}_{j}^{\top} \mathbf{D}_{j}$, where $\mathbf{D}_{j}$ is a differencing matrix.

### 4.4 Additive models

This leads to the direct solution

$$
\hat{\boldsymbol{\beta}}=\left(\mathbf{B}^{\top} \mathbf{B}+P\right)^{-1} \mathbf{B}^{\top} \mathbf{y}
$$

Constraint for identifiability:

$$
\sum_{i=1}^{n} f_{j}\left(x_{i j}\right)=0
$$

for each component $j$.

All of the fitting methods above can be extended for more than 2 covariates (section 4.5).

### 4.4 Additive models - example

Two models fitted to the Reef data:

$$
Y_{i}=f\left(\text { lat }_{i}, \operatorname{long}_{i}\right)+\varepsilon_{i}
$$

$$
Y_{i}=\beta_{0}+f\left(\text { lat }_{i}\right)+f\left(\text { long }_{i}\right)+\varepsilon_{i}
$$

### 4.4 Additive models - example






### 4.6 Fitting GAMs

As illustrated previously, one way to fit (Generalised) Additive Models is to use the mgcv library in R.
$\operatorname{gam}\left(y^{\sim} s(x)+s(z)+s(t)\right)$

- bam
- plot(model)
- many options for different smoothers including cyclic, bs='cc'
- multiple family items for non-normal response distributions e.g. ziP - zero-inflated poisson
- the default basis functions can be altered, $\mathrm{s}(\mathrm{x}, \mathrm{k}=15)$
- basis dimension can be assessed, gam. check()


### 4.7 Inference - comparing additive models

One approach - approximate F-test:

$$
F=\frac{\left(\mathrm{RSS}_{2}-\mathrm{RSS}_{1}\right) /\left(\mathrm{df}_{2}-\mathrm{df}_{1}\right)}{\mathrm{RSS}_{1} / \mathrm{df}_{1}}
$$

RSS: $\sum_{i=1}^{n}\left(y_{i}-\hat{y}_{i}\right)^{2}, \mathrm{df}=$ degrees of freedom for error
No general expression for the distribution of this test statistic is available.

Approximate guidance can be given by referring $F$ to an $F$ distribution $\left(\left(\mathrm{df}_{2}-\mathrm{df}_{1}\right), \mathrm{df}_{1}\right)$.

### 4.8 Example - Mackerel eggs

A multi-country survey of mackerel eggs in the Eastern Atlantic:

$$
\log \left(\text { density }_{i}\right)=\beta_{0}+f_{1}(\text { depth })+f_{2}(\text { temp })+f_{34}\left(\text { lat }_{i}, \operatorname{long}_{i}\right)+\varepsilon_{i}
$$

$$
\varepsilon_{i} \sim N\left(0, \sigma^{2}\right)
$$

$$
\begin{aligned}
\text { model1 <- gam(log(Density) } & \sim s(l o g(\text { mack.depth) ) } \\
+ & \mathrm{s}(\text { Temperature }) \\
& +\mathrm{s}(\text { mack.lat, mack.long) })
\end{aligned}
$$

### 4.8 Example - Mackerel eggs





Figure: Depth (left), Temperature (middle), and spatial location (right longitude ( $y$-axis), latitude ( $x$-axis))

### 4.8 Example - Mackerel eggs

Approximate significance of smooth terms:

|  | edf | Ref.df | F | p-value |
| :--- | ---: | ---: | ---: | ---: |
| s(log(mack.depth)) | 2.815 | 3.538 | 18.055 | $9.55 \mathrm{e}-12$ |
| s(Temperature) | 2.316 | 2.904 | 3.872 | 0.0147 |
| s(mack.lat,mack.long) | 20.197 | 24.788 | 5.060 | $1.03 \mathrm{e}-12$ |

### 4.8.2 Correlation in GAMs

The random effects framework introduced earlier can also be used in order to incorporate, and account for, correlation in GAMs.

## (Example 4.5)

- Daily river flow data were collected for a Scottish river between 1997 and 2001.
- It was of interest to investigate the long-term trend and any cyclical patterns in the data.


### 4.8.2 Correlation in GAMs

Flow data:



### 4.8.2 Correlation in GAMs

$$
\begin{aligned}
\log \left(\text { (low }_{i}\right)=\beta_{0}+s\left(\text { Year }_{i}\right)+s\left({\text { Day of } \left.\text { Year }_{i}\right)}\right) & \varepsilon_{i} \\
\varepsilon_{i} & \sim N\left(0, \sigma^{2}\right)
\end{aligned}
$$




### 4.8.2 Correlation in GAMs

## ACF/PACF of residuals:




### 4.8.2 Correlation in GAMs

## Incoporating correlated errors:

Take, $\varepsilon \sim N\left(0, V \sigma^{2}\right)$ for a correlation matrix $V$.
Therefore, here we will fit:

$$
\varepsilon_{i}=\phi \varepsilon_{i-1}+\epsilon_{i}
$$

with $\epsilon_{i} \sim N\left(0, \sigma^{2}\right)$.

Fitting in R:
$\operatorname{gamm}(\log ($ Flow $) \sim s($ Year, $b s=" c r ")+s($ doy, $b s=" c c "), ~ c o r r e l a t i o n=c o r A R 1(f o r m=\sim 1)) ~$

### 4.8.2 Correlation in GAMs

Fitted models after incoporating correlated errors:



### 4.8.2 Correlation in GAMs

ACF/PACF of residuals after incorporating correlated errors:



### 4.8.2 Correlation in GAMs

Fitted models:


### 4.8.3 Bayesian additive models

A fully Bayesian approach can be used extending the ideas in section 4.2.2, including priors for the unknown hyperparameter $\lambda$.

The R2BayesX package can be used to experiment with this approach.

Reef data example, Fig 4.20:
model2 <- bayesx(Score1 ~ sx(Longitude) + sx(Latitude))

## Summary

## What have we covered?

- How much to smooth?
- How to select smoothing parameters?
- random effect and fully Bayesian implementations
- Nonparametric regression in higher dimensions
- (Generalised) Additive Models

Tea time......

