MONASH University

MONASH BUSINESS SCHOOL

ETC3580: Advanced Statistical Modelling

Week 11: Additive models

Outline

1 Penalized regression splines

2 Additive models

3 Generalized additive models

Recall cubic regression splines

$$y = f(x) + \varepsilon$$
$$f(x) = \beta_0 + \sum_{k=1}^{K+3} \beta_k \phi_k(x)$$
where $\phi_1(x), \ldots, \phi_{K+3}(x)$ is a family of spline functions.

Example:

- Knots: $\kappa_1 < \kappa_2 < \cdots < \kappa_K$.
- $\phi_1(x) = x$, $\phi_2(x) = x^2$, $\phi_3(x) = x^3$,
  $\phi_k(x) = (x - \kappa_{k-3})_+^3 \quad$ for $k = 4, \ldots, K+3$.
- Choice of knots can be difficult and arbitrary.

Penalized spline regression

Idea: Use many knots, but constrain their influence by

$$\sum_{k=4}^{K+3} \beta_k^2 < C.$$

Penalized spline regression

Idea: Use many knots, but constrain their influence by

$$\sum_{k=4}^{K+3} \beta_k^2 < C.$$

Let $D = \begin{bmatrix} \mathbf{0}_{4\times 4} & \mathbf{0}_{4\times K} \\ \mathbf{0}_{K\times 4} & \boldsymbol{I}_{K\times K} \end{bmatrix}$.

Then we want to minimize

$$\|\boldsymbol{y} - \boldsymbol{X}\beta\|^2 \qquad \text{subject to} \qquad \beta'\boldsymbol{D}\beta \leq C.$$

Penalized regression splines

A Lagrange multiplier argument shows that this is equivalent to minimizing

$$\|\boldsymbol{y} - \boldsymbol{X}\beta\|^2 + \lambda^2\beta'\boldsymbol{D}\beta$$

for some number $\lambda \geq 0$.

Solution: $\hat{\beta}_\lambda = (\boldsymbol{X}'\boldsymbol{X} + \lambda^2\boldsymbol{D})^{-1}\boldsymbol{X}'\boldsymbol{y}$.

- A type of ridge regression.

Mixed model representation

Split $\boldsymbol{X}$ matrix in two:

$$\boldsymbol{X} = \begin{bmatrix} 1 & x_1 & x_1^2 & x_1^3 \\ \vdots & \vdots & \vdots & \vdots \\ 1 & x_n & x_n^2 & x_n^3 \end{bmatrix} \text{ and } \boldsymbol{Z} = \begin{bmatrix} (x_1 - \kappa_1)_+^3 & \dots & (x_1 - \kappa_K)_+^3 \\ \vdots & \ddots & \vdots \\ (x_n - \kappa_1)_+^3 & \dots & (x_n - \kappa_K)_+^3 \end{bmatrix}$$

and let $\boldsymbol{\beta} = [\beta_0, \beta_1, \beta_2, \beta_3]'$ and $\boldsymbol{u} = [u_1, \dots, u_K]'$.

Then we want to minimize

$$\|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta} - \boldsymbol{Z}\boldsymbol{u}\|^2 + \lambda^2\|\boldsymbol{u}\|^2$$

This is equivalent to estimating the mixed model

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{Z}\boldsymbol{u} + \boldsymbol{\varepsilon}$$

where $u_i \sim \mathrm{N}(0, \sigma_u^2)$ and $\varepsilon_j \sim \mathrm{N}(0, \sigma_\varepsilon^2)$.

Mixed model representation

Advantages

- Automatic penalty selection: use REML.
- Easy to develop Bayesian version

Mixed model representation

Advantages

- Automatic penalty selection: use REML.
- Easy to develop Bayesian version

Formulas

Let $\lambda = \sigma_\varepsilon / \sigma_u$ and $\boldsymbol{V} = \mathrm{Cov}(\boldsymbol{y}) = \sigma_u^2 \boldsymbol{Z}\boldsymbol{Z}' + \sigma_\varepsilon^2 \boldsymbol{I}$.

Mixed model representation

Advantages

- Automatic penalty selection: use REML.
- Easy to develop Bayesian version

Formulas

Let $\lambda = \sigma_\varepsilon / \sigma_u$ and $\boldsymbol{V} = \text{Cov}(\boldsymbol{y}) = \sigma_u^2 \boldsymbol{Z}\boldsymbol{Z}' + \sigma_\varepsilon^2 \boldsymbol{I}$. Then

$$\hat{\beta} = (\boldsymbol{X}\boldsymbol{V}^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{V}^{-1}\boldsymbol{y}.$$

$$\hat{\boldsymbol{u}} = \sigma_u^2 \boldsymbol{Z}'\boldsymbol{V}^{-1}(\boldsymbol{y} - \boldsymbol{X}\hat{\beta}).$$

$\boldsymbol{V}$ estimated using profile log-likelihood methods.

Choice of knots

- Provided the set of knots is relatively dense with respect to the $\{x_j\}$, the result hardly changes.
- Choose enough knots to model structure, but not too many knots to cause computational problems.
- Ruppert, Wand and Carroll recommend:
  - $\max(n/4, 35)$ knots where n = number of unique observations.
  - $\kappa_j = \left(\frac{j+1}{K+1}\right)$th sample quantile of the unique $\{x_j\}$.
- `mgcv` package uses penalized regression splines by default.

Outline

1 Penalized regression splines

2 Additive models

3 Generalized additive models

Additive models

Avoid curse of dimensionality by assuming additive surface:

$$y = \beta_0 + \sum_{j=1}^{p} f_j(x_j) + \varepsilon$$

where $\varepsilon \sim N(0, \sigma^2)$.

- Restricts complexity but a much richer class of surfaces than parametric models.
- Need to estimate p one-dimensional functions instead of one p-dimensional function.
- Usually set each f_j to have zero mean.
- Some f_j may be linear.

Additive models

- Up to p different bandwidths to select.
- Generalization of multiple regression model

$$y = \beta_0 + \sum_{j=1}^{p} \beta_j x_j + \varepsilon$$

  which is also additive in its predictors.
- Estimated functions, f_j, are analogues of coefficients in linear regression.
- Interpretation easy with additive structure.

Additive models

- Categorical predictors: fit constant for each level as for linear models.
- Allow interaction between two continuous variables x_j and x_k by fitting a bivariate surface $f_{j,k}(x_j, x_k)$.
- Allow interaction betwen factor x_j and continuous x_k by fitting separate functions $f_{j,k}(x_k)$ for each level of x_j.

Additive models in R

- `gam` package: more smoothing approaches, uses a backfitting algorithm for estimation.
- `mgcv` package: simplest approach, with automated smoothing selection and wider functionality.
- `gss` package: smoothing splines only

Estimation

Back-fitting-algorithm (Hastie and Tibshirani, 1990)

1. Set $\beta_0 = \bar{y}$.
2. Set $f_j(x) = \hat{\beta}_j x$ where $\hat{\beta}_j$ is OLS estimate.
3. For $j = 1, \ldots, p, 1, \ldots, p, 1, \ldots, p, \ldots$

$$f_j(x) = S(x_j, y - \beta_0 - \sum_{i \neq j} f_i(x_i))$$

where $S(x, u)$ means univariate smooth of u on x.

Iterate step 3 until convergence.

Estimation

Back-fitting-algorithm (Hastie and Tibshirani, 1990)

1. Set $\beta_0 = \bar{y}$.
2. Set $f_j(x) = \hat{\beta}_j x$ where $\hat{\beta}_j$ is OLS estimate.
3. For $j = 1, \dots, p, 1, \dots, p, 1, \dots, p, \dots$

$$f_j(x) = S(x_j, y - \beta_0 - \sum_{i \neq j} f_i(x_i))$$

   where $S(x, u)$ means univariate smooth of u on x.

Iterate step 3 until convergence.

- S could be *any* univariate smoother.
- $y - \beta_0 - \sum_{i \neq j} f_i(x_i)$ is a "partial residual"

Estimation

Regression splines

No need for iterative back-fitting as the model can be written as a linear model.

Penalized regression splines

No need for iterative back-fitting as the model can be written as a linear mixed-effects model.

Inference for Additive Models

Each fitted function can be written as a linear smoother $\hat{\boldsymbol{f}}_j = \boldsymbol{S}_j \boldsymbol{y}$ for some $n \times n$ matrix $\boldsymbol{S}_j$.

$\hat{\boldsymbol{f}}(\boldsymbol{x})$ is a linear smoother. Denote smoothing matrix as $\boldsymbol{S}$:

$$\hat{\boldsymbol{f}}(\boldsymbol{x}) = \boldsymbol{S}\boldsymbol{y} = \beta_0 \mathbf{1} + \sum_{j=1}^{p} \boldsymbol{S}_j \boldsymbol{y}$$

where $\mathbf{1} = [1, 1, \ldots, 1]^T$. Then $\boldsymbol{S} = \sum_{j=0}^{p} \boldsymbol{S}_j$ where $\boldsymbol{S}_0$ is such that $\boldsymbol{S}_0 \boldsymbol{y} = \beta_0 \mathbf{1}$.

Thus all inference results for linear smoothers may be applied to additive model.

Outline

1 Penalized regression splines

2 Additive models

3 Generalized additive models

Generalised additive models

Generalized Linear Model (GLM)

- Distribution of y
- Link function g
- $\text{E}(y \mid x_1, \dots, x_p) = \mu$ where $g(\mu) = \beta_0 + \sum_{j=1}^{p} \beta_j x_j$.

Generalised additive models

Generalized Linear Model (GLM)

- Distribution of y
- Link function g
- $\mathrm{E}(y \mid x_1, \dots, x_p) = \mu$ where $g(\mu) = \beta_0 + \sum_{j=1}^{p} \beta_j x_j$.

Generalised Additive Model (GAM)

- Distribution of y
- Link function g
- $\mathrm{E}(y \mid x_1, \dots, x_p) = \mu$ where $g(\mu) = \beta_0 + \sum_{j=1}^{p} f_j(x_j)$.

Generalised additive models

Examples:

- Y binary and $g(\mu) = \log[\mu(1-\mu)]$. This is a logistic additive model.
- Y normal and $g(\mu) = \mu$. This is a standard additive model.

Estimation

Hastie and Tibshirani describe method for fitting GAMs using a method known as “local scoring” which is an extension of the Fisher scoring procedure.