Generalized Additive Models 2

M1 Maths-IA

Yannig Goude [1]

8 février 2023

Sommaire

1 Interactions

2 Model selection
- Generalized likelihood ratio test
- Double penalty approach
- Shrinkage approach

Interaction terms : s and te

We previously saw that interaction can be fitted with `mgcv`, using `s` or `te` functions :

```
fit <- gam(formula = y ~ x1 + s(x2, x3))
fit2 <- gam(formula = y ~ x1 + te(x2, x3))
```

with `s`, a single penalty is optimized for the all interaction function whereas with `te` the penalty could be different for each marginals.

Interaction terms : ti

If the goal is interpretability, distinguish between interaction and single effect matter, we can than use the`ti` function :

```
fit3 <- gam(formula = y ~ x1 + ti(x2) + ti(x3) + ti(x2,x3))
```

`mgcv` compute in that case tensor product interactions from which the main effects have been excluded, under the assumption that they will be included separately.
This can be seen as functional ANOVA decomposition.

Interaction terms : anova

Testing the existance of interaction can be done with a ANOVA using the `anova` function :

```
fit1 <- gam(formula = y ~ x1 + ti(x2) + ti(x3))
fit2 <- gam(formula = y ~ x1 + ti(x2) + ti(x3) + ti(x2,x3))
anova(fit1, fit2, test = "Chisq")
```

Model comparison with ANOVA

Let say we have $X_, 1, ..., X_p$ explanatory variables, the question behind model comparison is wether we need to include a subset of covariate $X_{q+1}, ..., X_p$?
The test is then :

$$\begin{aligned} H_0 &: \beta_{q+1} = 0 = ... = \beta_p \\ H_1 &: H_0 \quad \text{is false} \end{aligned}$$

Let denote L_F and LR the likelihood of the full (p variables) and reduced (first q variables) models.
The generalized likelihood ratio test is base on the test statistic $W = -2 \log \Lambda_0$ with $\Lambda_0 = \frac{L_R}{L_F}$.

Model comparison with ANOVA

Under H_0, $\Lambda_0 \neq 1$ and $W \neq 0$. If H_0 is false $W >> 0$.
More precisely, under H_0, $W \sim \chi^2_{p-q}$, sowe reject H_0 at level α if $W > \chi^2_{p-q,\alpha}$.
Let denote L_S the likelihood of the saturated model,

$$\Lambda_0 = \frac{L_R}{L_S} / \frac{L_F}{L_S}$$

so that $W = (D_R - D_F)/\phi$ the difference of the scaled deviances.

mgcv basics

```
Analysis of Deviance Table

Model 1: Load ~ ti(Load.1, bs = "cr") + ti(Load.7, bs = "cr")
Model 2: Load ~ ti(Load.1, bs = "cr") + ti(Load.7, bs = "cr") + ti(Load.1,
    Load.7)
  Resid. Df Resid. Dev     Df  Deviance  Pr(>Chi)
1      2913 2.6308e+10
2      2905 2.5832e+10 8.0495 476547831 8.599e-09 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Double penalty approach

We saw that fitting a GAM corresponds to find the optimal λ and minimizing :

$$\|\boldsymbol{Y} - \boldsymbol{X}\beta\|^2 + \sum_j \lambda_j \beta^T S_j \beta$$

leading to the following estimator of β :

$$\widehat{\beta}_\lambda = (\boldsymbol{X}^T\boldsymbol{X} + \sum_j \lambda_j \boldsymbol{S_j})^{-1}\boldsymbol{X}^T\boldsymbol{Y}$$

Double penalty approach

This procedure is useful to regularize and reduce the complexity of an additive model but can not select/ remove a smooth term.

- One drawback of penalizing the second derivative of the smooth function is that it doesn't penalize functions in the penalty null space (e.g. linear functions).
- $\boldsymbol{S_j}$ could be decomposed as $\boldsymbol{U_j}\Lambda_j\boldsymbol{U_j}^T$ where $\boldsymbol{U_j}$ is an eigenvector matrix, Λ_j the corresponding diagonal matrix of the eigenvalues, containing 0 values due to the previous fact.
- [MW11] propose a shrinkage approach to shrink the functions to zero for high penalties.

Double penalty approach

Denoting $\boldsymbol{U_j^*}$ the matrix of eigenvectors corresponding to the zero eigenvalues of Λ_j and $S_j^* = \boldsymbol{U_j^*} \boldsymbol{U_j^*}^T$
They propose to add an extra penalty :

$$\lambda_j \beta^T \boldsymbol{S_j} \beta + \lambda_j^* \beta^T \boldsymbol{S_j^*} \beta$$

the first term penalizes function component in the range space, the 2nd term components in the null space. Both can be shrinked to 0 with high λs. It implies optimizing 2 *lambda* per function.

For cubic spline : the first term penalizes functions which deviate from linear and the 2nd term penalizes linear terms.

Implementation in `mgcv` with the `select=TRUE` argument.

Another proposal of [MW11] is to replace *bmS*$_j$ with

$$\boldsymbol{\tilde{S}_j} = \boldsymbol{U_j}\tilde{\Lambda}_j\boldsymbol{U_j^T}$$

replacing 0 eigen values in Λ_j by ε (small proportion of the smallest $\lambda_j > 0$).

This is equivalent to the double penalty approach with $\lambda_j = \varepsilon\lambda_j$

Implementation in `mgcv` with the `s(x, bs='cs')` or `te(x, z, bs='ts')`.

References I

Giampiero Marra and Simon N Wood. « Practical variable selection for generalized additive models ». In: **Computational Statistics & Data Analysis** 55.7 (2011), pp. 2372–2387.