



Smoothing parameter estimation

Simon N. Wood

University of Glasgow, UK

Overview

- 6 I will talk about estimating smoothing parameters (ridge parameters) λ , in penalized model fitting problems (generalized ridge regression problems) such as,

$$\text{minimize } \sum_{i=1}^n w_i (y_i - \mathbf{X}_i \boldsymbol{\beta})^2 + \sum_{j=1}^m \lambda_j \boldsymbol{\beta}^T \mathbf{S}_j \boldsymbol{\beta}$$

w.r.t. $\boldsymbol{\beta}$, or

$$\text{minimize } D(\boldsymbol{\mu}) + \sum_{j=1}^m \lambda_j \boldsymbol{\beta}^T \mathbf{S}_j \boldsymbol{\beta} \text{ w.r.t. } \boldsymbol{\beta},$$

where $\boldsymbol{\mu} = \mathbb{E}(\mathbf{y})$, and D is the deviance, $2(l(\mathbf{y}) - l(\boldsymbol{\mu}))$.

In particular, I will discuss efficient and stable λ estimation via criteria such as

- ⑥ GCV

$$\mathcal{V} = \frac{n \|\sqrt{\mathbf{W}}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})\|^2}{(n - \text{e.d.f.})^2} \quad \text{or} \quad \mathcal{V}_d = \frac{nD(\boldsymbol{\mu})}{(n - \text{e.d.f.})^2}$$

- ⑥ Mallows's statistic

$$C_p = \|\sqrt{\mathbf{W}}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})\|^2/n + 2\sigma^2 \times \text{e.d.f.}/n$$

- ⑥ or AIC

$$A = D(\boldsymbol{\mu}) + 2 \times \text{e.d.f.}$$

Typical context

- ⑥ A Generalized Additive Model has a structure like

$$g(\mathbb{E}[y_i]) = f_1(x_i) + f_2(z_i, v_i) + f_3(u_i) + \dots$$

where, usually, $y_i \sim$ some exponential family distribution.

- ⑥ We can choose bases to represent each f_j , e.g.

$$f_1(x) = \sum_{j=1} \alpha_j a_j(x)$$

- ⑥ Then the GAM becomes a GLM

$$g(\mathbb{E}[y_i]) = \mathbf{X}_i \boldsymbol{\beta}$$

Typical context II

- ⑥ If the GAM is estimated by likelihood maximization, then it will tend to *overfit*.
- ⑥ This tendency can be penalized during fitting. That is, minimize,

$$D(\boldsymbol{\mu}) + \sum \text{wiggleness of components}$$

rather than $D(\boldsymbol{\mu})$.

- ⑥ More formally, this becomes,

$$\text{minimize } D(\boldsymbol{\mu}) + \sum_{j=1}^m \lambda_j \boldsymbol{\beta}^T \mathbf{S}_j \boldsymbol{\beta} \text{ w.r.t. } \boldsymbol{\beta}$$

P-IRLS: penalized likelihood maximization

Given λ , and an initial guess at η ($= \mathbf{X}\beta$), iterate

- ⑥ $z_i = g'(\mu_i)(y_i - \mu_i) + \eta_i$ (note that $\mu_i = g^{-1}(\eta_i)$).
- ⑥ $W_{ii} = [V(\mu_i)g'(\mu_i)^2]^{-1}$.
- ⑥ Writing $\mathbf{S}_\lambda = \sum_j \lambda_j \mathbf{S}_j$,

$$\text{minimize } \|\sqrt{\mathbf{W}}(\mathbf{z} - \mathbf{X}\beta)\|^2 + \beta^\top \mathbf{S}_\lambda \beta \text{ w.r.t. } \beta$$

to find the next estimate of β .

Note that at each iteration, $\sqrt{\mathbf{W}}\eta = \mathbf{A}_\lambda \sqrt{\mathbf{W}}\mathbf{z}$, where $\mathbf{A}_\lambda = \sqrt{\mathbf{W}}\mathbf{X}(\mathbf{X}^\top \mathbf{W}\mathbf{X} + \mathbf{S}_\lambda)^{-1} \mathbf{X}^\top \sqrt{\mathbf{W}}$. $\text{tr}(\mathbf{A}_\lambda)$ gives the effective degrees of freedom (e.d.f.) of the model.

Options for λ estimation

There are two alternative ways of using GCV, AIC, etc.

1. Estimate the optimal λ for the working linear model at each P-IRLS iteration, by e.g. minimizing

$$\mathcal{V} = n \|\sqrt{\mathbf{W}}(\mathbf{z} - \mathbf{X}\boldsymbol{\beta})\|^2 / [n - \text{tr}(\mathbf{A}_\lambda)]^2 \quad \text{w.r.t. } \lambda$$

2. Estimate λ as the minimizer of e.g.

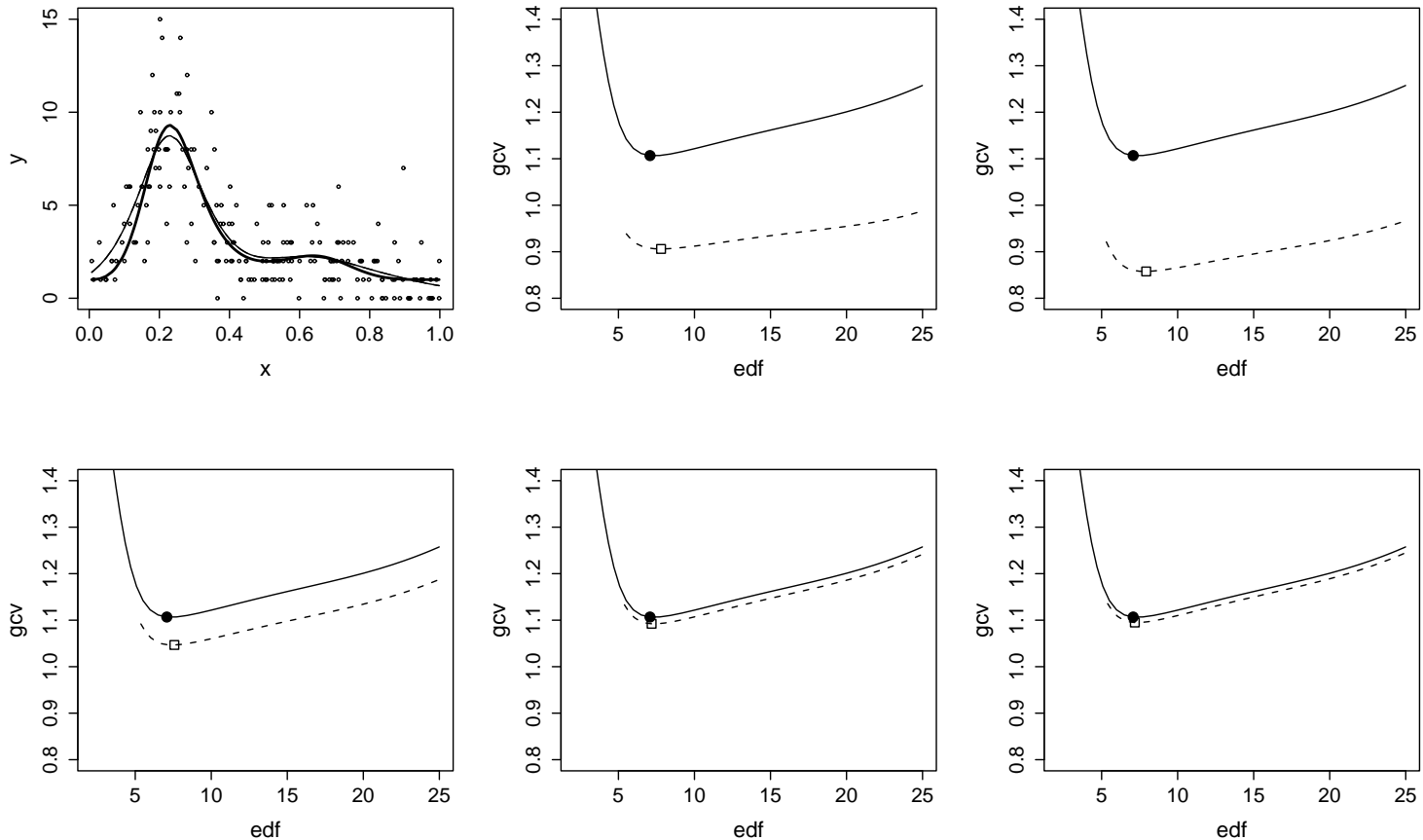
$$\mathcal{V}_d = nD(\hat{\boldsymbol{\mu}}) / [n - \text{tr}(\mathbf{A}_\lambda)]^2$$

where $\hat{\boldsymbol{\mu}}$ and \mathbf{A}_λ are evaluated at P-IRLS convergence.

1. is known as *performance iteration* : it is fast, but can fail to converge. 2. ('outer iteration') is slower, but can be more reliable.

Comparison of approaches

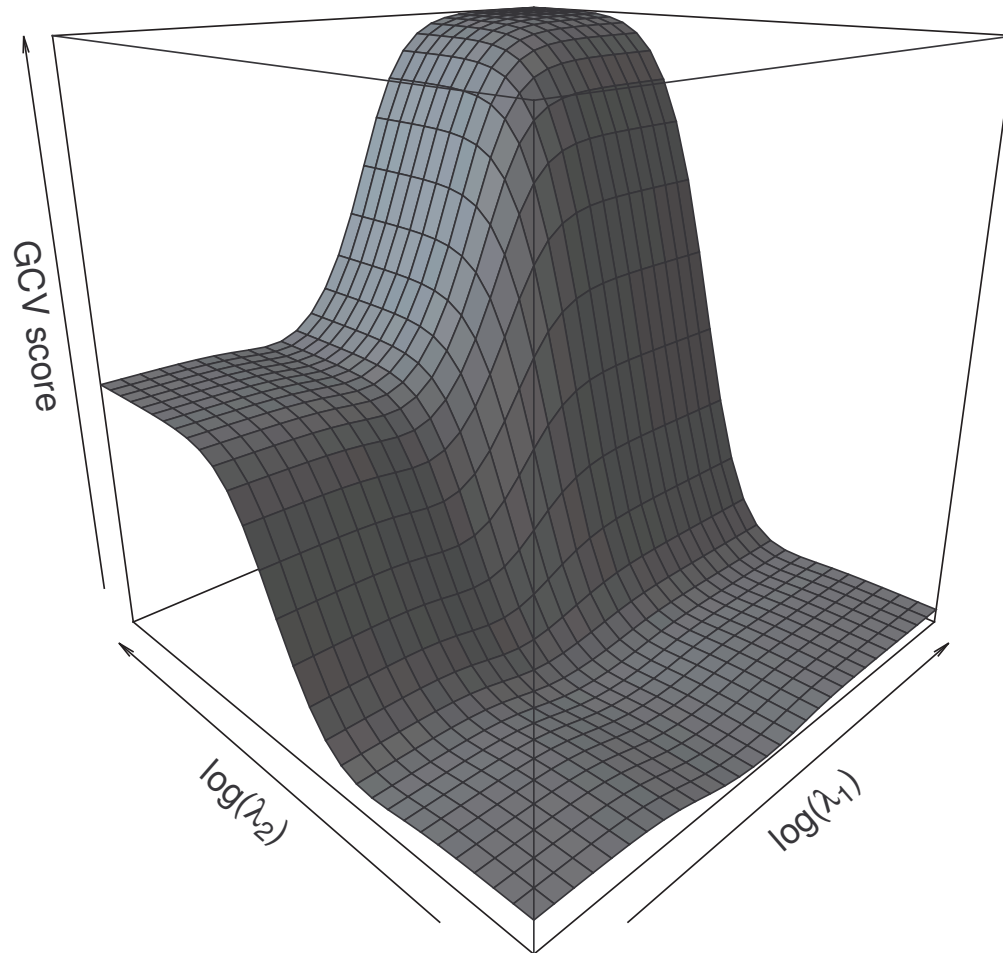
GCV for outer iteration (solid) and performance iteration (dashed). Poisson data and fits shown top right.



- ⑥ Criteria optimization is potentially expensive for $\dim(\boldsymbol{\lambda}) > 1$. Computational cost should scale linearly with number of data, n .
- ⑥ GAMs and similar models can be quite complex, leading to problems with numerical stability. Criteria optimization must be able to cope with these.
- ⑥ Co-linearity and concurvity can lead to problems with identifiability of $\boldsymbol{\lambda}$. This can make optimization awkward, and can cause performance iteration to fail altogether.

An example ν

GCV score for a 2 term GAM



Performance iteration

- ⑥ Need to minimize

$$\mathcal{V} = \|\mathbf{y} - \mathbf{A}_\lambda \mathbf{y}\|^2 / [n - \text{tr}(\mathbf{A}_\lambda)]^2$$

- ⑥ The potentially expensive part is getting $\text{tr}(\mathbf{A}_\lambda)$ and its derivatives w.r.t. λ , where $\mathbf{A}_\lambda = \mathbf{X}(\mathbf{X}^\top \mathbf{X} + \mathbf{S}_\lambda)^{-1} \mathbf{X}^\top$.
- ⑥ First form $\mathbf{X} = \mathbf{QR}$ (with pivoting). Most costly step, but only needed once. Note that $\mathbf{X}^\top \mathbf{X} = \mathbf{R}^\top \mathbf{R}$.
- ⑥ Now, truncating if necessary, form the SVD

$$\begin{bmatrix} \mathbf{R} \\ \sqrt{\mathbf{S}_\lambda} \end{bmatrix} = \mathbf{U} \mathbf{D} \mathbf{V}^\top = \begin{bmatrix} \mathbf{U}_1 \\ \mathbf{U}_2 \end{bmatrix} \mathbf{D} \mathbf{V}^\top$$

Performance iteration II

- From the SVD, $\mathbf{R}^T \mathbf{R} + \mathbf{S}_\lambda = \mathbf{V} \mathbf{D}^2 \mathbf{V}^T = \mathbf{X}^T \mathbf{X} + \mathbf{S}_\lambda$.
- So, $(\mathbf{X}^T \mathbf{X} + \mathbf{S}_\lambda)^{-1} = \mathbf{V} \mathbf{D}^{-2} \mathbf{V}^T$.
- Also, $\mathbf{X} = \mathbf{Q} \mathbf{R} = \mathbf{Q} \mathbf{U}_1 \mathbf{D} \mathbf{V}^T$, since, $\mathbf{R} = \mathbf{U}_1 \mathbf{D} \mathbf{V}^T$.
- Substitution into the expression for \mathbf{A}_λ yields:

$$\mathbf{A}_\lambda = \mathbf{Q} \mathbf{U}_1 \mathbf{U}_1^T \mathbf{Q}^T \Rightarrow \text{tr}(\mathbf{A}_\lambda) = \text{tr}(\mathbf{U}_1 \mathbf{U}_1^T).$$

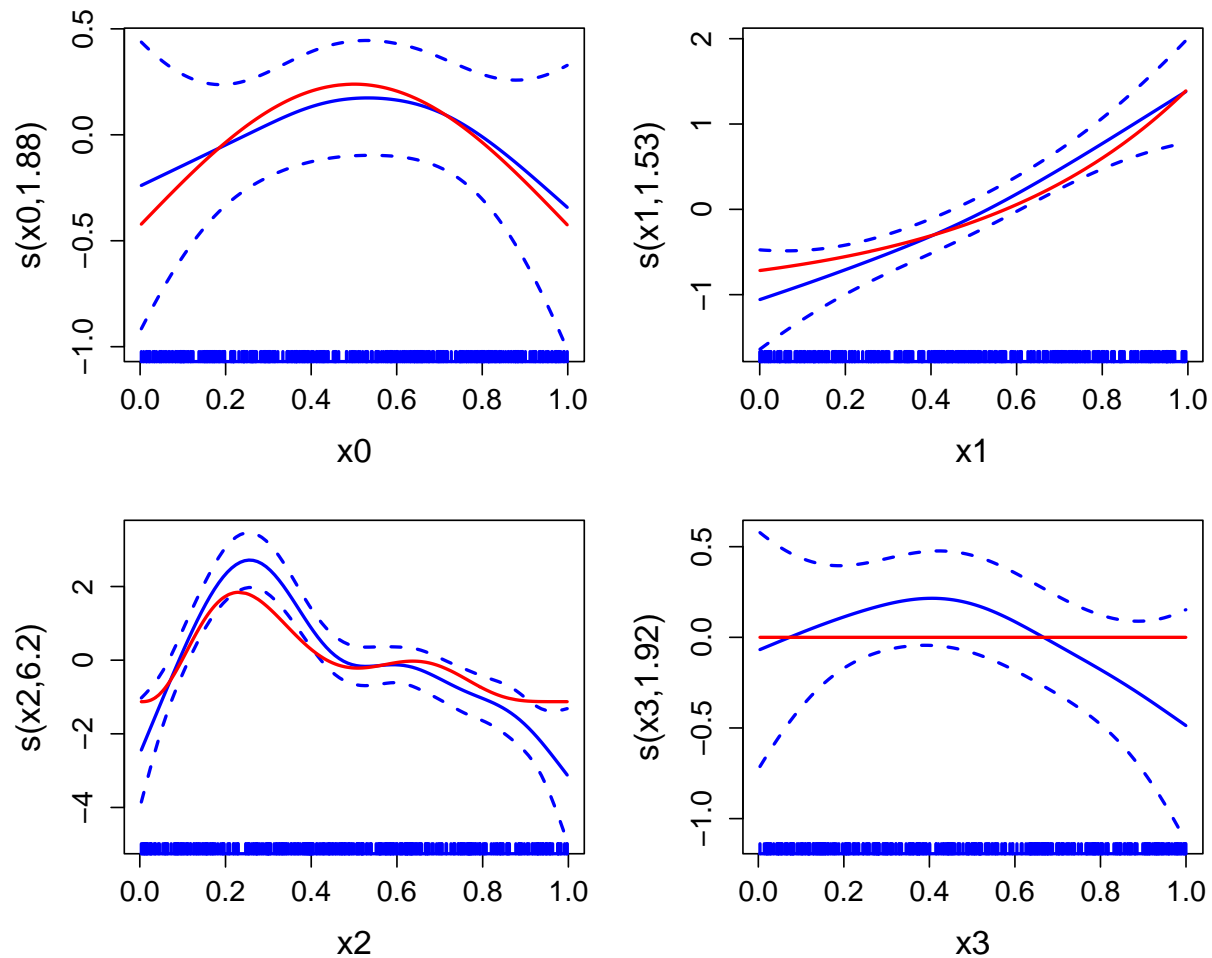
- Defining $\mathbf{y}_1 = \mathbf{U}_1^T \mathbf{Q} \mathbf{y}$ it is easy to show

$$\|\mathbf{y} - \mathbf{A}_\lambda \mathbf{y}\|^2 = \mathbf{y}^T \mathbf{y} - 2\mathbf{y}_1^T \mathbf{y}_1 + \mathbf{y}_1 \mathbf{U}_1^T \mathbf{U}_1 \mathbf{y}_1.$$

- Efficient and stable derivatives w.r.t. λ follow ...

Performance iteration example

Fit from 400 binary observations of 3 term additive truth.



Outer iteration

- ⑥ Minimization of $\mathcal{V}_d = nD(\hat{\boldsymbol{\mu}})/[n - \text{tr}(\mathbf{A}_\lambda)]^2$ is more tedious.
- ⑥ Possible to base minimization on finite differenced derivatives, but this is slow and unreliable for *difficult* models.
- ⑥ Easy to get

$$\frac{\partial D}{\partial \boldsymbol{\beta}} \quad \text{and} \quad \frac{\partial \text{tr}(\mathbf{A}_\lambda)}{\partial \boldsymbol{\beta}}, \quad \text{but} \quad \frac{\partial \boldsymbol{\beta}}{\partial \boldsymbol{\lambda}}$$

has to be iterated along side P-IRLS scheme.

Outer iteration II

- ⑥ Propagation of derivatives through the fitting of the working linear model involves a similar approach to \mathcal{V} evaluation in *performance iteration*.
- ⑥ But some of the other steps are less convenient. e.g. the following term cannot be avoided

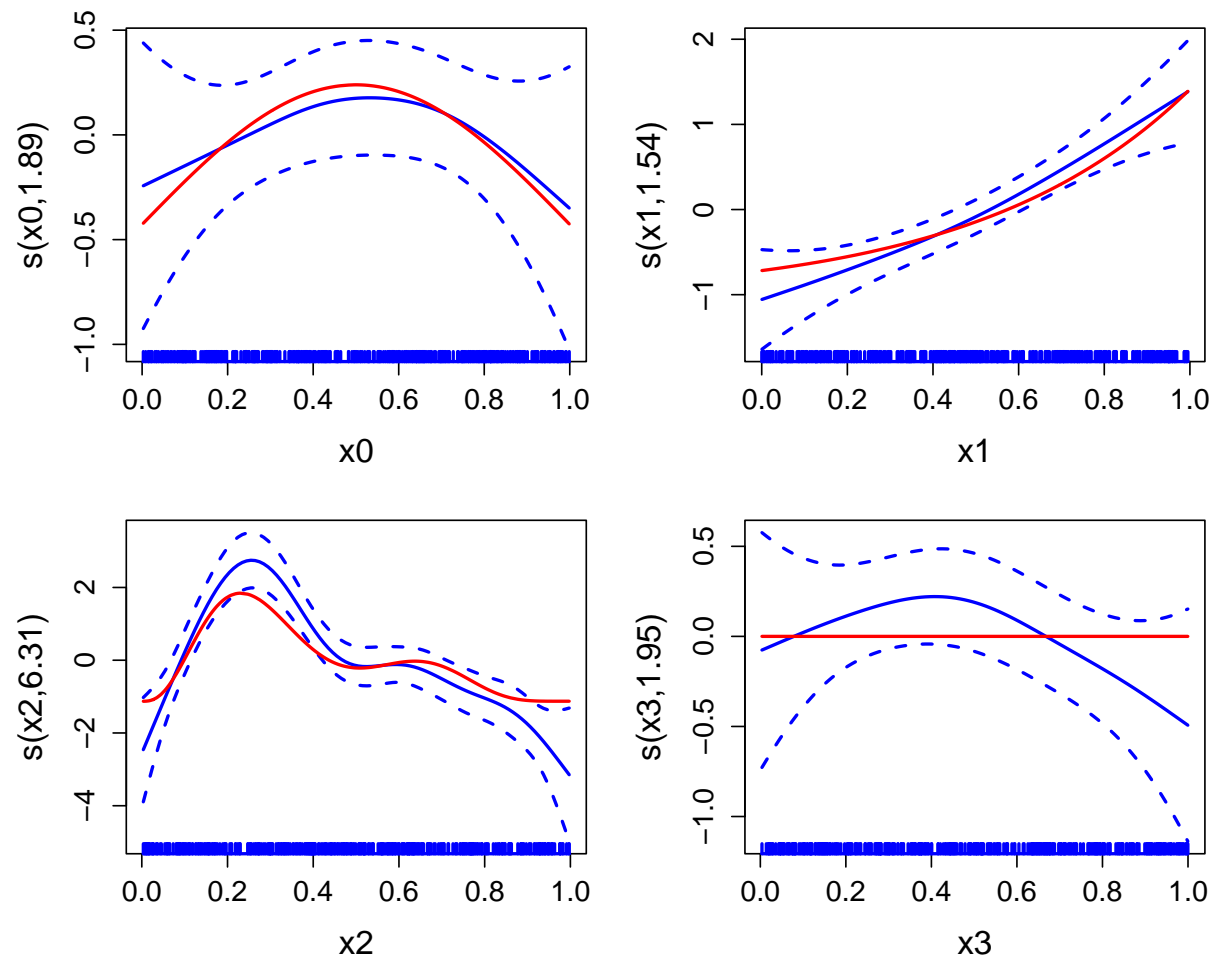
$$\frac{\partial w_i}{\partial \rho_k} = -\frac{1}{2} w_i^3 \left[\frac{\partial V}{\partial \mu_i} \frac{\partial \mu_i}{\partial \eta_i} + 2V(\mu_i) g''(\mu_i) \right] \frac{\partial \eta_i}{\partial \rho_k}$$

where $\rho_k = \log(\lambda_k)$. Note that $\partial V / \partial \mu_i$ and $g''(\mu_i)$ are not needed for the P-IRLS itself...

- ⑥ Given first derivatives, optimization by quasi-Newton is straightforward.

Outer iteration example

Fit from 400 binary observations of 3 term additive truth.



Example: Death & Air Pollution

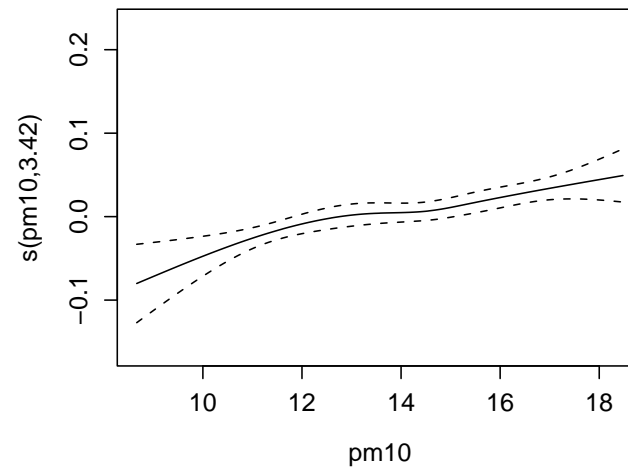
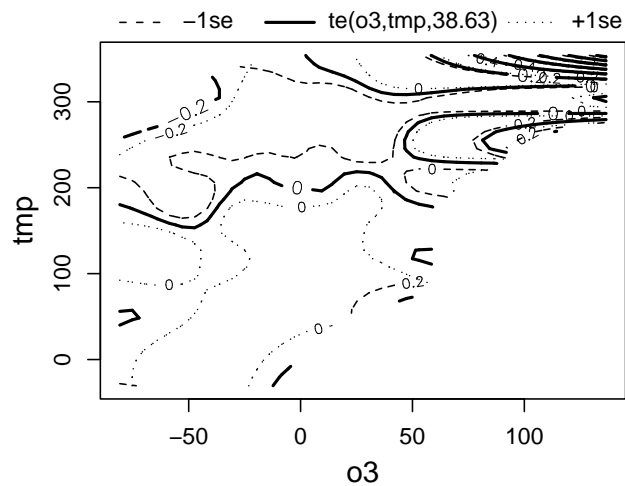
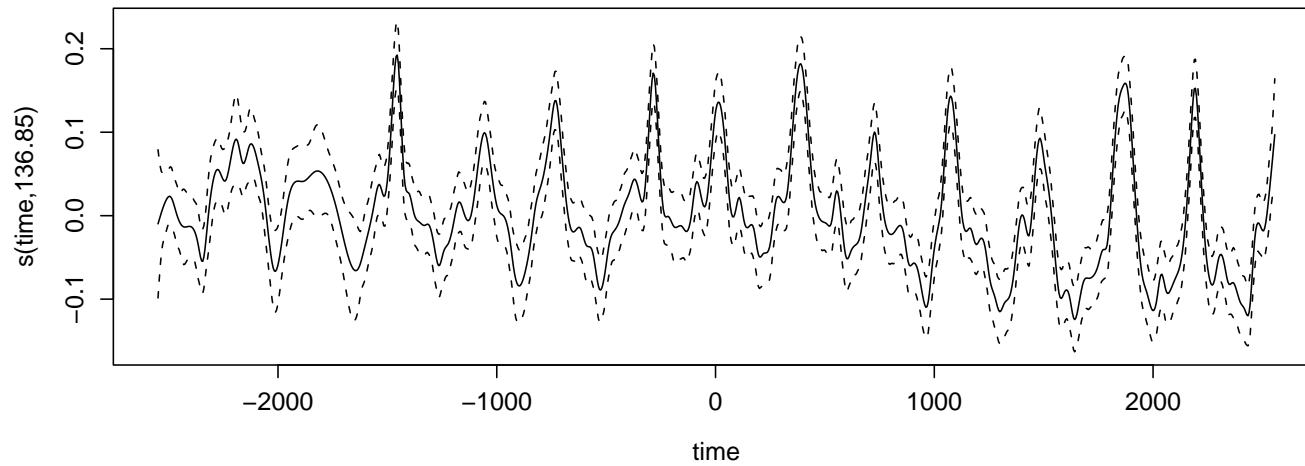
- ⑥ Response data is daily `death` rate in Chicago over 13+ years.
- ⑥ Interest is in the relationship with pollution related predictors (all summed over 4 days leading to death): ozone, `o3`; temperature `tmp`; particulate matter, `pm10`.
- ⑥ Also need to account for `time` variation in non-pollution related death rate.
- ⑥ Model is $\text{death}_i \sim \text{Poisson}$,

$$\log(\mathbb{E}[\text{death}_i]) = f_1(\text{time}_i) + f_2(\text{o3}_i, \text{tmp}_i) + f_3(\text{pm10}_i)$$

Death & Air Pollution II

- ⑥ The smooth ozone - temperature interaction is represented using a tensor product smooth (with 2 penalties), while the other smooth terms are represented using penalized cubic regression splines.
- ⑥ Performance iteration fails for this model, if particulate matter is included to make a 3-way interaction, but outer iteration has no problem.
- ⑥ Performance iteration problems probably relate to the fact that all predictors are themselves strongly time dependent (concurvity).
- ⑥ Fit suggests that high temperatures with high ozone is a bad thing.

Death & Air Pollution fits



Extensions

- ⑥ Quadratic approximations to any regular likelihood can be written as a pseudo-model and data, and are hence susceptible to smoothing parameter estimation by performance iteration, at least.
- ⑥ Non-linear models estimable by Gauss-Newton are particularly easy to deal with.
- ⑥ REML estimation of smoothing parameters involves similar calculations to those presented here.

Further information

- ⑥ The methods discussed here are available in R package `mgcv`.
- ⑥ Start R, type `library(mgcv)` and `?gam` for information, including references.
- ⑥ To get R, go to <http://cran.r-project.org>.
- ⑥ Alternatively, see <http://www.stats.gla.ac.uk/~simon> for references etc.
- ⑥ The air pollution data are from Roger D. Peng's `NMMAPSdata` R library.