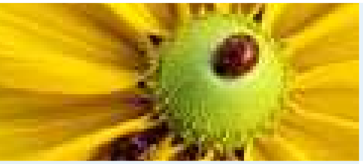
REML Estimation and Linear Mixed Models

5. Mixed model splines

Sue Welham

Rothamsted Research
Harpenden UK AL5 2JQ

December 14, 2008



Introduction

Splines have become a popular method for modelling non-linear trend within linear models, especially where the underlying form is unknown.

Several types of polynomial spline have been proposed within the mixed model context:

- cubic smoothing splines
- P-splines
- penalised splines

These splines are very closely related.

We will look at

- how splines fit into the mixed model framework
- connections between these different methods
- the practical use of splines for modelling non-linear curves in mixed models

There is also a wider class of L-splines, which we will not consider here.

Simple spline model

- Suppose we have n data values $\boldsymbol{y} = (y_1 \ldots y_n)'$
- and explanatory variable $\boldsymbol{x} = (x_1 \ldots x_n)'$ on range $[a, b]$ with $a < x_1 < x_2 < \ldots < x_n < b$
- We assume the data follows some unknown but smooth function $g(x)$ with error

Evaluated at the data points, $\boldsymbol{x}$,

$$\boldsymbol{y} = g(\boldsymbol{x}) + \boldsymbol{e}$$

with $\boldsymbol{e} \sim N(\boldsymbol{0}, \sigma^2 \boldsymbol{R})$.

Polynomial splines

A polynomial spline $g(x)$ of degree k (k=3 for cubic splines) with r knots $t_1 < t_2 < \ldots < t_r$ has the properties:

- on each interval $[t_j, t_{j+1}]$, $g(x)$ is a polynomial of degree k
- $g(x)$ and derivatives to order $k-1$ are continuous on $[a, b]$

Natural splines of odd degree $k = 2m - 1$ obey an additional constraint

$$g^{(j)}(t_1) = g^{(j)}(t_r) = 0 \quad \text{for} \quad j = m \ldots k$$

ie. natural splines have higher order derivatives constrained to zero outside the range of the knots.

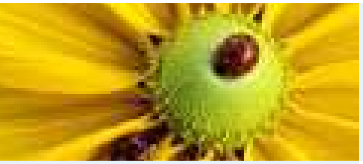
Natural cubic splines are piecewise cubic, continuous in the second derivative, and linear outside the range of the knots.

Polynomial splines

A polynomial spline $g(x)$ of degree k with r knots has:

- polynomial of kth degree on $(r+1)$ sections $\Rightarrow (k+1)(r+1)$ parameters
- $0 \ldots k-1$ derivatives continuous at r knots $\Rightarrow rk$ constraints
- $\Rightarrow r+k+1$ free parameters

Natural splines of degree k obey an additional $k+1$ constraints $\Rightarrow r$ free parameters.



Spline basis functions

A set of polynomial spline (degree k) basis functions for knot points $\boldsymbol{t} = (t_1 \dots t_r)'$ is a set of functions $\{s_i(x) \; ; \; i = 1 \dots r + k + 1\}$ such that

- each basis function $s_i(x)$ is itself a polynomial spline of degree k
- the set $\{s_i(x)\}$ spans the set of polynomial splines (degree k) for the knot points $\boldsymbol{t}$

Then any polynomial spline of degree k, $g(x)$, can be represented as

$$g(x) = \sum_{i=1}^{r+k+1} c_i s_i(x)$$

for suitably chosen coefficients c_i.

For natural splines, the basis functions should also be natural splines, then with only r basis functions required.

Regression splines

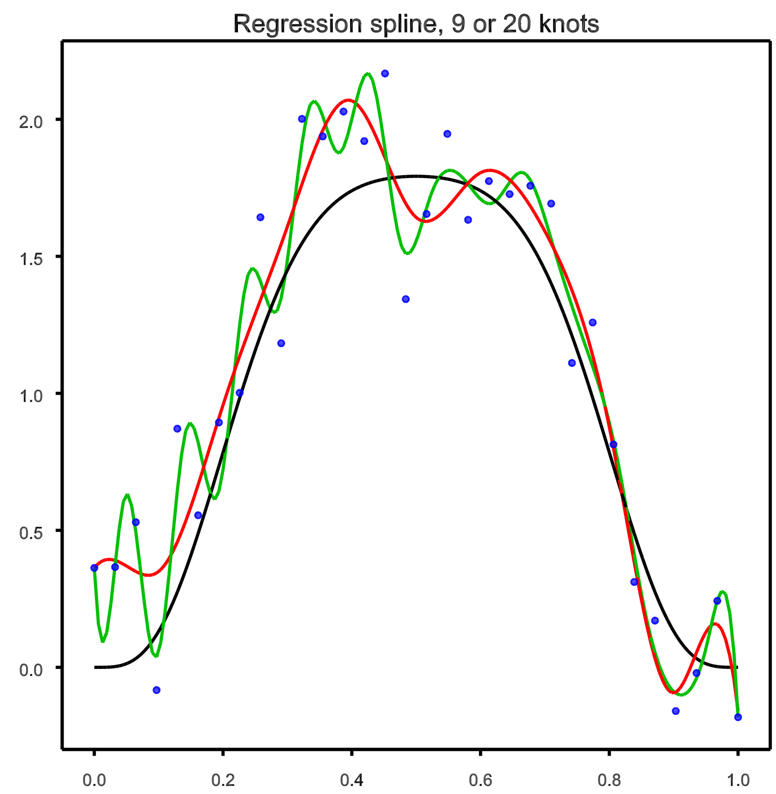
Regression splines are constructed as follows:

- a set of r knots is chosen, $r \ll n$
- a set of $r+k+1$ basis functions $s_i(x)$ is constructed
- the spline is fitted via regression using the model

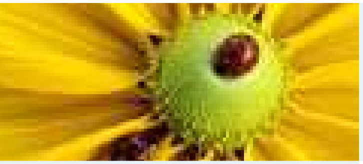
$$\boldsymbol{y} = g(\boldsymbol{x}) + \boldsymbol{e} = \sum_{i=1}^{r+k+1} c_i s_i(\boldsymbol{x}) + \boldsymbol{e}$$

- If the n (distinct) covariate values are used as knots, we get an exact fit and the fitted curve is the interpolation spline
- The smoothness of the curve is determined by the number and position of the knots: fewer knots $\Rightarrow$ smoother curve
- Choice of knots (and $\therefore$ smoothness of curve) subjective

Regression splines



- black line is function
- blue dots are data = function + normal error (32 points)
- green line = regression spline with 20 knots (equal spacing)
- red line = regression spline with 9 knots (equal spacing)



Smoothing splines

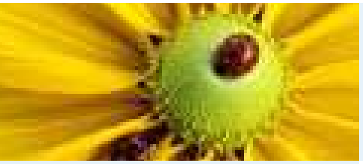
Smoothing splines of degree $k = 2m - 1$ ($k = 3 \Rightarrow m = 2$) are fitted to the model

$$\boldsymbol{y} = g(\boldsymbol{x}) + \boldsymbol{e}$$

by minimising a penalised sum of squares (PSS):

$$[\boldsymbol{y} - g(\boldsymbol{x})]' \boldsymbol{R}^{-1} [\boldsymbol{y} - g(\boldsymbol{x})] + \lambda \int_a^b [g^{(m)}(s)]^2 ds$$

- the first term (residual sum of squares) measures fidelity of the fitted curve to the data
- the second term (penalty) quantifies undesirable behaviour (roughness, ie. wigglyness) of the fitted curve
- in the PSS, the fit of the curve to the data competes against the roughness of the fitted curve
- λ ($\lambda > 0$) is a smoothing parameter that controls the balance between the two terms, hence smoothness of fitted curve



Cubic smoothing spline

It can be shown that, for a given value of λ, the curve that minimises the PSS

$$[\boldsymbol{y} - g(\boldsymbol{x})]' \boldsymbol{R}^{-1} [\boldsymbol{y} - g(\boldsymbol{x})] + \lambda \int_a^b [g''(s)]^2 ds$$

is a natural cubic spline with knots at the n (distinct) covariate values, $x_1 \ldots x_n$. This spline is called the cubic smoothing spline.

Here the penalty

- is the integrated squared second derivative of the fitted spline
- can be regarded as measuring accumulated deviations from straight line

Cubic smoothing spline - penalty

For a natural cubic spline, the penalty can be rewritten as

$$\int_a^b [g''(s)]^2 ds = \boldsymbol{\delta}' \boldsymbol{G}_s^{-1} \boldsymbol{\delta}$$

where

- $\boldsymbol{\delta} = (\delta_2 \dots \delta_{n-1})'$ holds values of second derivatives of fitted spline at internal knots $x_2 \dots x_{n-1}$
- $\boldsymbol{G}_s^{-1}$ is a tri-diagonal matrix with $n-2$ rows defined in terms of distances between successive knots, defined in terms of its inverse as

  $$[\boldsymbol{G}_s^{-1}]_{ij} = \begin{cases} \frac{1}{6}(x_i - x_{i-1}) = \frac{1}{6} h_{i-1} & j = i-1 \\ \frac{1}{3}(x_{i+1} - x_{i-1}) = \frac{1}{6}(h_i + h_{i-1}) & j = i \\ \frac{1}{6}(x_{i+1} - x_i) = \frac{1}{6} h_i & j = i+1 \\ 0 & \text{otherwise} \end{cases}$$

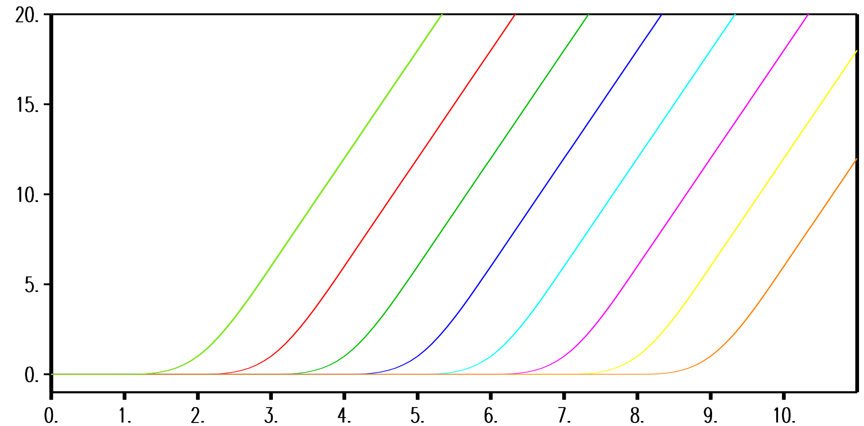
  where $h_j = x_{j+1} - x_j$.
- details in Green & Silverman (1994)

Cubic smoothing spline - basis

A natural cubic spline can be written as

$$g(x) = \tau_1 + \tau_2 x + \sum_{i=2}^{n-1} \delta_i P_i(x)$$

where $\{P_i(x), i = 2 \ldots n-1\}$ are basis functions for natural cubic splines, representing smooth deviations about linear trend.



Each basis function is cubic on across a range of 3 knots, and linear either side:

$$6P_j(x) = \begin{cases} 0 & x \leq t_{j-1} \ , \\ h_{j-1}^{-1}(x - t_{j-1})^3 & t_{j-1} < x \leq t_j \ , \\ h_{j-1}^{-1}(x - t_{j-1})^3 - (h_{j-1}^{-1} + h_j^{-1})(x - t_j)^3 & t_j < x \leq t_{j+1} \ , \\ (h_{j-1} + h_j)(3x - t_{j-1} - t_j - t_{j+1}) & t_{j+1} < x. \end{cases}$$

Cubic smoothing spline

In matrix form

$$g(\boldsymbol{x}) = \boldsymbol{X\tau} + \boldsymbol{P\delta}$$

where $\boldsymbol{X} = [\boldsymbol{1}\ \boldsymbol{x}]$, $\boldsymbol{\tau} = (\tau_1\ \tau_2)'$ and $\boldsymbol{P} = [P_2(\boldsymbol{x}) \ldots P_{n-1}(\boldsymbol{x})]$.

The PSS can now be rewritten as

$$[\boldsymbol{y} - \boldsymbol{X\tau} - \boldsymbol{P\delta}]'\boldsymbol{R}^{-1}[\boldsymbol{y} - \boldsymbol{X\tau} - \boldsymbol{P\delta}] + \lambda\boldsymbol{\delta}'\boldsymbol{G}_s^{-1}\boldsymbol{\delta}$$

Minimise with respect to $\boldsymbol{\tau}$ and $\boldsymbol{\delta}$: differentiate and set equal to zero gives equations

$$\begin{aligned} \boldsymbol{X}'\boldsymbol{R}^{-1}\boldsymbol{X\tau} + \quad & \boldsymbol{X}'\boldsymbol{R}^{-1}\boldsymbol{P\delta} &= \boldsymbol{X}'\boldsymbol{R}^{-1}\boldsymbol{y} \\ \boldsymbol{P}'\boldsymbol{R}^{-1}\boldsymbol{X\tau} + \quad & (\boldsymbol{P}'\boldsymbol{R}^{-1}\boldsymbol{P} + \lambda\boldsymbol{G}_s^{-1})\boldsymbol{\delta} &= \boldsymbol{P}'\boldsymbol{R}^{-1}\boldsymbol{y} \end{aligned}$$

This looks like the mixed model equations from a linear mixed model

$\longrightarrow$ mixed model smoothing splines described by Verbyla et al 1999, Brumback & Rice (1998), Wang (1998), Zhang et al (1998).

Mixed model cubic smoothing spline

Given the smoothing parameter, the cubic smoothing spline can be calculated as the BLUP

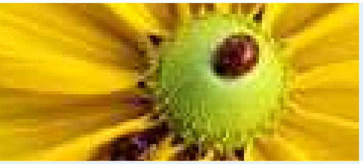
$$\tilde{g}(\boldsymbol{x}) = \boldsymbol{X}\hat{\boldsymbol{\tau}} + \boldsymbol{P}\tilde{\boldsymbol{\delta}}$$

from the mixed model

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\tau} + \boldsymbol{P}\boldsymbol{\delta} + \boldsymbol{e}$$

with $\mathrm{var}\,(\boldsymbol{\delta}) = \sigma_s^2 \boldsymbol{G}_s$, $\mathrm{var}\,(\boldsymbol{e}) = \sigma^2 \boldsymbol{R}$ and $\lambda = \sigma^2/\sigma_s^2 = \gamma_s^{-1}$.

- Note: fitting both $\boldsymbol{\tau}$ and $\boldsymbol{\delta}$ as fixed gives interpolation spline.
- Fitting deviations as random introduces smoothing via shrinkage of random effects.
- Amount of shrinkage depends on λ.
- $\lambda > 0 \Rightarrow \sigma_s^2 > 0$
- The smoothing parameter can also be estimated by REML as a variance parameter.



Transformation to independence

In practice, we make a transformation to independent random effects:

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\tau} + \boldsymbol{Z}_s\boldsymbol{u}_s + \boldsymbol{e}$$

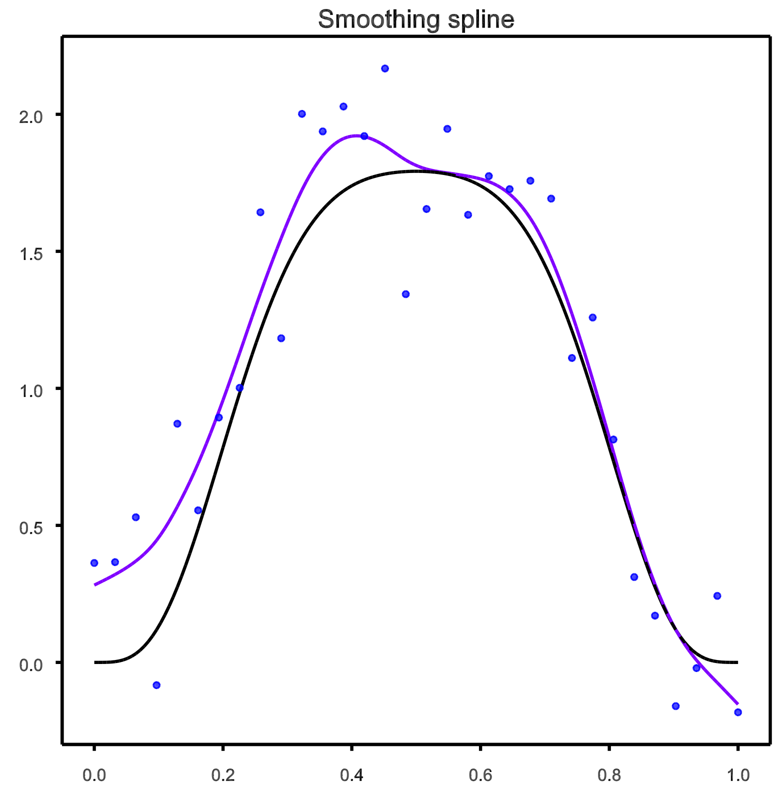
where $\boldsymbol{Z}_s = \boldsymbol{P}\boldsymbol{G}_s^{0.5}$, $\boldsymbol{u}_s = \boldsymbol{G}_s^{-0.5}\boldsymbol{\delta} \sim N(\boldsymbol{0}, \sigma_s^2\boldsymbol{I})$.

The random part of the spline can then be fitted as a simple variance component term.

In addition, we can transform so that $\boldsymbol{X}$ and $\boldsymbol{Z}_s$ are independent (to aid interpretation).

This model can then easily be fitted using most mixed model software.

Smoothing spline



- black line is function
- blue dots are data = function + normal error (32 points)
- purple line = smoothing spline fitted by REML

Mixed model cubic smoothing spline

- REML estimation selects smoothing parameter value that gives best fit to implicit variance model defined by $\boldsymbol{P}$ and $\boldsymbol{G}_s$
- Studies have shown that REML performs as well as GCV in terms of MSEP, although it yields smoother curves than GCV
- Advantage of REML approach: can build spline terms into complex mixed models that account for all sources of variation in the data, including correlated errors

But are we using this model just as a tool, or do we really believe it as a mixed model for the data?

- if a tool, need to bear this in mind when making inferences from fitted model
- might get better results when model is compatible with data

Inference

- In general mixed model, inference is made taking into account the influence of the population of random effects
- If population of random effects is not considered relevant then it is appropriate to exclude this variation from inferences and consider the model conditional on the random effects
- For mixed model spline, we are not interested in population of spline curves, we are interested only in the instance present in the data
- In the simple case, the mixed model smoothing spline is therefore considered as

$$\boldsymbol{y} \mid \boldsymbol{\tau}, \boldsymbol{\delta} \sim N(\,\boldsymbol{X\tau} + \boldsymbol{P\delta}, \sigma^2 \boldsymbol{R}\,)$$

Inference

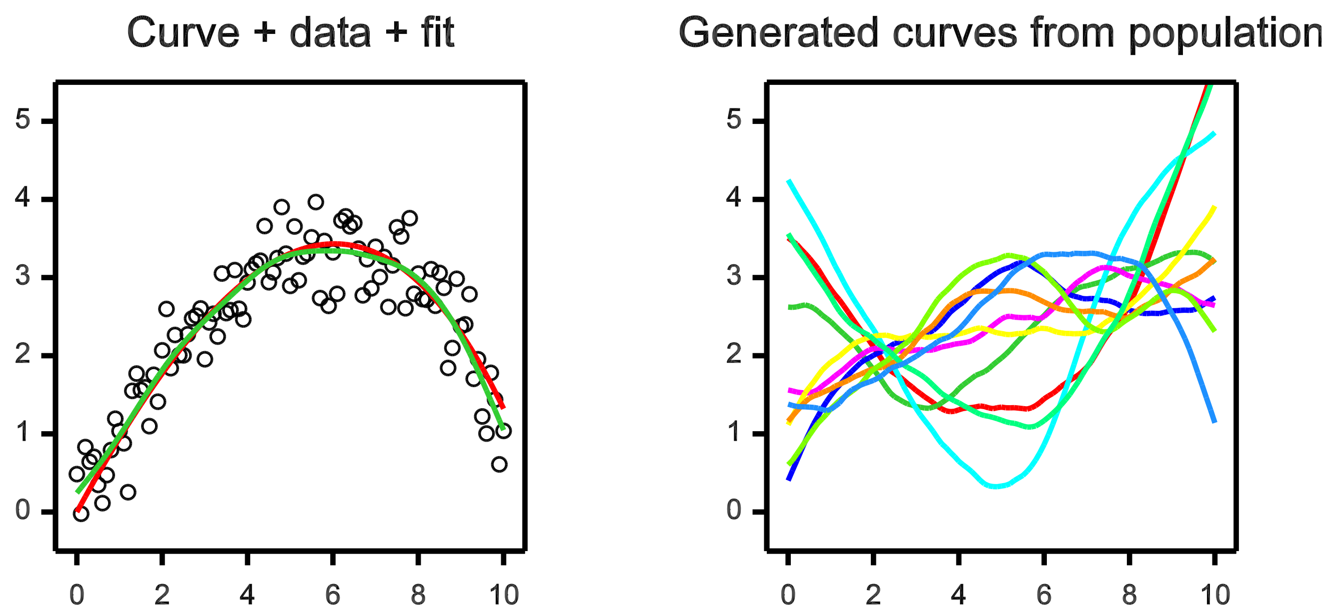


Figure 1: Fitted spline with population of splines

Inference

We are interested in conditional inference, ie.

$$\boldsymbol{y} \mid \boldsymbol{\tau}, \boldsymbol{\delta} \sim N(\ g(\boldsymbol{x}), \sigma^2 \boldsymbol{R}\ )$$

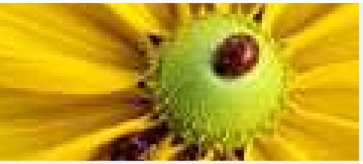
The conditional expectation takes the form

$$\begin{aligned} E(\ \tilde{\boldsymbol{y}} \mid \boldsymbol{\tau}, \boldsymbol{\delta}\ ) &= E(\ \boldsymbol{X}\hat{\boldsymbol{\tau}} + \boldsymbol{P}\tilde{\boldsymbol{\delta}} + \tilde{e} \mid \boldsymbol{\tau}, \boldsymbol{\delta}\ ) \\ &= \boldsymbol{X}\boldsymbol{\tau} + \boldsymbol{P}\boldsymbol{G}_s\boldsymbol{P}'\boldsymbol{P}_H E(\boldsymbol{y}|\boldsymbol{\tau}, \boldsymbol{\delta}) \\ &= \boldsymbol{X}\boldsymbol{\tau} + \boldsymbol{P}\boldsymbol{G}_s\boldsymbol{P}'\boldsymbol{P}_H\boldsymbol{P}\boldsymbol{\delta} \\ &\neq g(\boldsymbol{x}) \end{aligned}$$

where $\boldsymbol{P} = [P_2(\boldsymbol{x}) \ldots P_{n-1}(\boldsymbol{x})]$, $\boldsymbol{P}_H = \boldsymbol{H}^{-1} - \boldsymbol{H}^{-1}\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{H}^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}\boldsymbol{H}^{-1}$ and $\boldsymbol{P}_H\boldsymbol{X} = \boldsymbol{0}$.

The shrinkage that provides smoothing also introduces bias into the fitted function.

Bias is largest where shrinkage is largest, ie. where $|\delta_i|$ is large and hence where the absolute value of the second derivative, or curvature, is largest.



Inference

The conditional variance of the random effects takes the form:

$$\text{var}\left(\tilde{\boldsymbol{\delta}}|\boldsymbol{\delta}\right) = \sigma^2 \boldsymbol{G}_s \boldsymbol{P}' \boldsymbol{P}_H \boldsymbol{R} \boldsymbol{P}_H \boldsymbol{P} \boldsymbol{G}_s$$

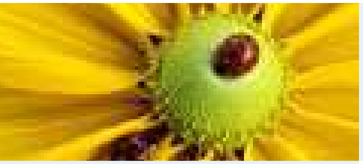
The conditional mean square error is calculated from the variance and the bias squared and takes the form

$$\begin{aligned} E\left((\tilde{\boldsymbol{\delta}}-\boldsymbol{\delta})(\tilde{\boldsymbol{\delta}}-\boldsymbol{\delta})'|\boldsymbol{\delta}\right) &= E\left((\tilde{\boldsymbol{\delta}}-\boldsymbol{\delta})|\boldsymbol{\delta}\right) E\left((\tilde{\boldsymbol{\delta}}-\boldsymbol{\delta})|\boldsymbol{\delta}\right)' + \text{var}\left(\tilde{\boldsymbol{\delta}} \mid \boldsymbol{\delta}\right) \\ &= (\boldsymbol{G}_s \boldsymbol{P}' \boldsymbol{P}_H \boldsymbol{P} - \boldsymbol{I}) \boldsymbol{\delta}\boldsymbol{\delta}' (\boldsymbol{P}' \boldsymbol{P}_H \boldsymbol{P} \boldsymbol{G}_s - \boldsymbol{I}) + \sigma^2 \boldsymbol{G}_s \boldsymbol{P}' \boldsymbol{P}_H \boldsymbol{R} \boldsymbol{P}_H \boldsymbol{P} \boldsymbol{G}_s \end{aligned}$$

Taking expectations over $\boldsymbol{\delta}$ and using $E(\boldsymbol{\delta}\boldsymbol{\delta}') = \sigma^2 \boldsymbol{G}_s$, we get

$$\begin{aligned} E(\tilde{\boldsymbol{\delta}}-\boldsymbol{\delta})^2 &= \sigma^2 (\boldsymbol{G}_s \boldsymbol{P}' \boldsymbol{P}_H \boldsymbol{P} - \boldsymbol{I}) \boldsymbol{G}_s (\boldsymbol{P}' \boldsymbol{P}_H \boldsymbol{P} \boldsymbol{G}_s - \boldsymbol{I}) + \sigma^2 \boldsymbol{G}_s \boldsymbol{P}' \boldsymbol{P}_H \boldsymbol{R} \boldsymbol{P}_H \boldsymbol{P} \boldsymbol{G}_s \\ &= \sigma^2 \left(\boldsymbol{G}_s \boldsymbol{P}' \boldsymbol{P}_H \boldsymbol{P} \boldsymbol{G}_s \boldsymbol{P}' \boldsymbol{P}_H \boldsymbol{P} \boldsymbol{G}_s - 2 \boldsymbol{G}_s \boldsymbol{P}' \boldsymbol{P}_H \boldsymbol{P} \boldsymbol{G}_s + \boldsymbol{G}_s \right. \\ &\quad \left. + \boldsymbol{G}_s \boldsymbol{P}' \boldsymbol{P}_H \boldsymbol{R} \boldsymbol{P}_H \boldsymbol{P} \boldsymbol{G}_s \right) \\ &= \sigma^2 \left(\boldsymbol{G}_s \boldsymbol{P}' \boldsymbol{P}_H \boldsymbol{H} \boldsymbol{P}_H \boldsymbol{P} \boldsymbol{G}_s - 2 \boldsymbol{G}_s \boldsymbol{P}' \boldsymbol{P}_H \boldsymbol{P} \boldsymbol{G}_s + \boldsymbol{G}_s \right) \\ &= \sigma^2 \left(\boldsymbol{G}_s - \boldsymbol{G}_s \boldsymbol{P}' \boldsymbol{P}_H \boldsymbol{P} \boldsymbol{G}_s \right) \end{aligned}$$

So the unconditional variance (readily available) can be used as an estimate of the average conditional Mean Squared Error.



Inference

Is the spline required?

- Test for curvature: REML likelihood ratio test of $H_0 : \sigma_s^2 = 0$
  - test on boundary of parameter space
  - does not obey conditions for 50:50 mixture of chi-squared distributions
  - use parametric bootstrap to get empirical distribution for test under null hypothesis

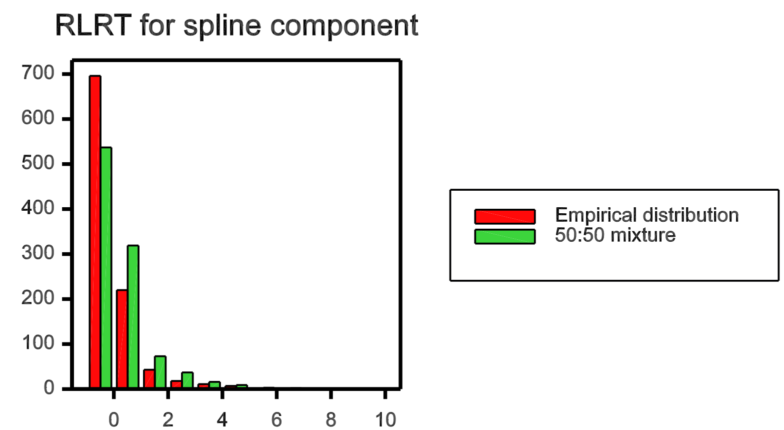
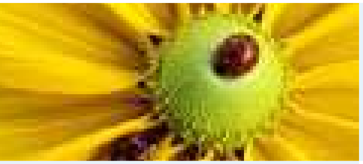


Figure 2: Bootstrap and theoretical distribution of RLRT

Empirical distribution: 67% zero values, 95th percentile = 1.74 (50:50 mixture 2.71)



More complex models

So far, we have considered the case of a single curve with distinct covariate values.

Repeated covariate values are easily dealt with:

- if we have n data values and $r < n$ distinct covariate values $\boldsymbol{t} \subset \boldsymbol{x}$
- use set of distinct values $\boldsymbol{t}$ as knots to form r basis functions
- apply basis functions to full covariate $\boldsymbol{x}$, *i.e.*

$$g(\boldsymbol{x}) = \tau_1 + \tau_2 \boldsymbol{x} + \sum_{i=1}^{r} \delta_i P_i(\boldsymbol{x})$$

- this effectively gives additional weight to repeated covariate values and still achieves minimum of PSS

More complex models

Similarly, splines can be included as part of more complex model formulae:

- denote random part of spline function ($\boldsymbol{P\delta}$) as spl(x)
- simple spline model = constant + lin(x) + spl(x)
- might introduce different splines for different treatments

$$\text{constant} + \text{trt} + \text{lin(x)} + \text{trt.lin(x)} + \text{spl(x)} + \text{trt.spl(x)}$$

  with trt.lin(x) fitted as fixed, trt.spl(x) fitted as random

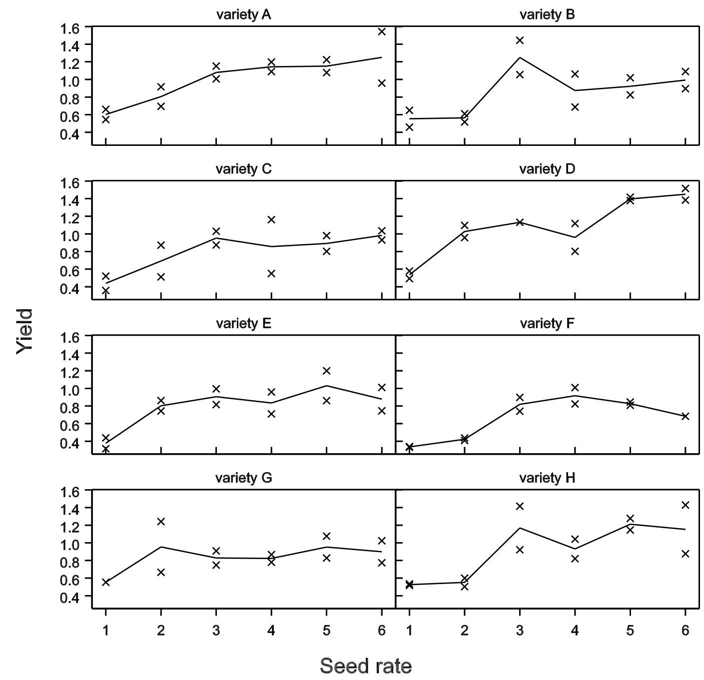
- can simplify model using RLRT

Can also consider interaction of spline with another covariate, z.spl(x), to get varying coefficient models

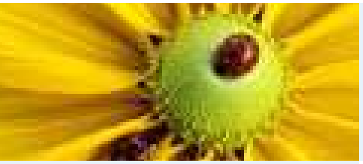
- allows coefficient of z to change as a smooth function of x
- this gives an alternative method of response surface modelling

Example: field trial

Field experiment to find optimal seeding rate (6 rates) for 8 varieties of lupin



- Completely randomized design (laid out on 16 x 6 grid)
- Theory suggests exponential model, but cannot fit this to all varieties
- Use splines to model profiles across seed rate and establish whether common model adequate



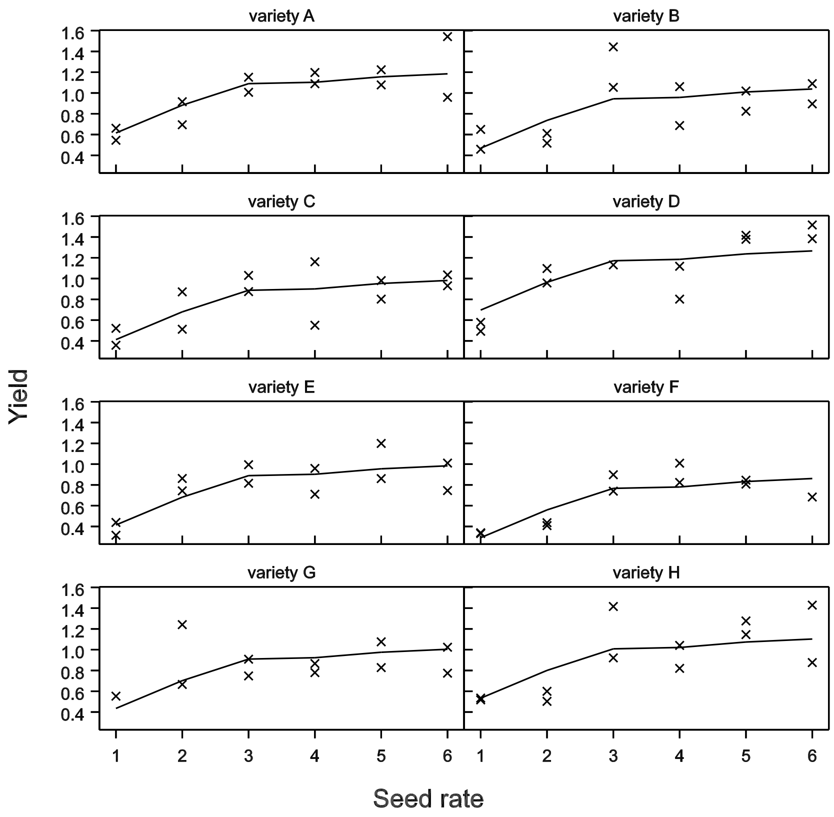
Example: model

- fixed terms = const + variety + lin(seedrate) + variety.lin(seedrate)
- spline random component = spl(seedrate) + variety.spl(seedrate)
- structural random terms = residual error (spatial model)

```
Estimated variance components

Random term                 component       s.e.
Spline(seedrate)              0.01214     0.01490
variety.Spline(seedrate)      0.00000       bound

Residual variance model

Term                    Model(order)  Parameter    Estimate      s.e.
row.column              Identity      Sigma2         0.0338   0.00562

Tests for fixed effects

Sequentially adding terms to fixed model

Fixed term         Wald statistic  n.d.f.   F statistic  d.d.f.    F pr
variety                     41.24       7          5.89    72.5  <0.001
seedrate                    86.90       1         86.90    72.5  <0.001
variety.seedrate             8.32       7          1.19    72.5   0.321
```

Example: fitted model = common spline

- fixed terms = const + variety + lin(seedrate)
- spline random component = spl(seedrate)
- structural random terms = residual error



Example: conclusions

Spline model is useful in this context

- can investigate differences between varieties without specifying form of curve
- can add other fixed or random effects, eg. to model spatial trend

In general, spline mixed models useful to model non-linear trend within complex mixed models, especially where

- form of non-linear trend does not fit standard curve
- form of non-linear trend may differ between treatments
- correlated error structure present, eg. longitudinal data
- contribution of nuisance variables needs to be removed

Reduced knot mixed model spline - O-spline

- Advantage of cubic smoothing spline: no need to use subjective choice of knots
- But, design matrix for spline basis functions is dense, with $\#$columns $=$ number of distinct covariate values - 2
- For large data sets, models may be slow to fit
- Use reduced knot version of cubic smoothing spline:
  - for specified knots $\boldsymbol{t} = (t_1 \ldots t_r)'$ form basis functions $P_2 \ldots P_{r-1}$
  - fit model with same fixed effects, but basis functions, random effects $\boldsymbol{\delta}$ and variance matrix $\boldsymbol{G}_s$ defined with respect to new knots
- Number of columns in random design matrix $= r - 2$
- Choice of knots subjective (and affects fitted spline)
- Fitted spline no longer minimises PSS
- Compromise solution: retain as many knots as possible

P-splines

Introduced by Eilers & Marx (1996) as simple and computationally efficient alternative to smoothing splines.

P-splines

- use a B-spline basis
- reduced knot set, knots usually equally spaced
- discrete penalty defined using differencing operators
- model

$$g(x) = \sum_{i=1}^{r+k+1} a_i B_i(x) \quad \text{or} \quad g(\boldsymbol{x}) = \boldsymbol{B}\boldsymbol{a}$$

- penalised sum of squares

$$[\boldsymbol{y} - \boldsymbol{B}\boldsymbol{a}]'\boldsymbol{R}^{-1}[\boldsymbol{y} - \boldsymbol{B}\boldsymbol{a}] + \lambda \boldsymbol{a}'\boldsymbol{\Delta}_d'\boldsymbol{\Delta}_d\boldsymbol{a}$$

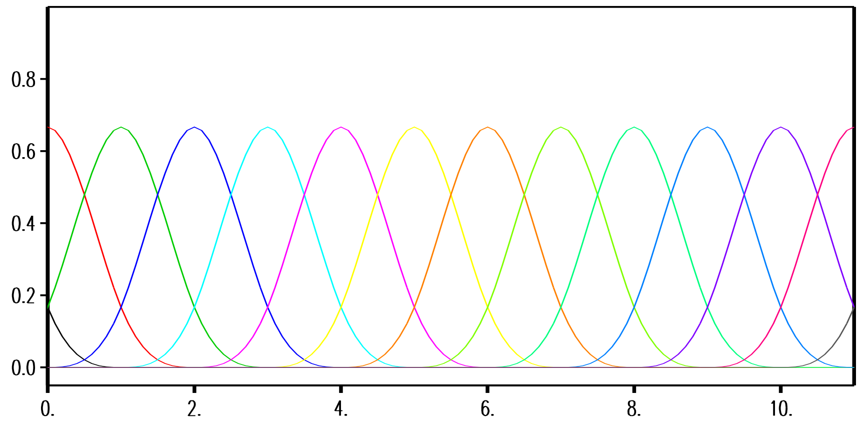
- where $\boldsymbol{\Delta}_d$ is dth order differencing matrix

P-splines - rationale

B-spline basis has many desirable properties

- individual B-spline basis functions have compact support $\Rightarrow$ matrix $\boldsymbol{B}$ has sparse form
- $\sum_i B_i(x) = 1 \; \forall x$ and the form of the curve follows the pattern in the coefficients
- hence restricting differences in adjacent coefficients compels fitted spline to be smooth

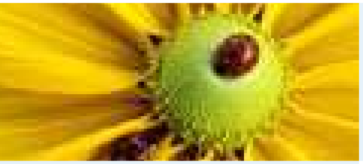
B-spline basis functions for knots $1 \ldots 10$



P-splines

User has choice of:

- degree of B-splines (k)
- number and position of knots $= t_1 \ldots t_r$
- order of differencing in the penalty (d)



P-splines as mixed models

- For k=3, d=2, P-spline approximates O-spline with same knots
- In the general case, let

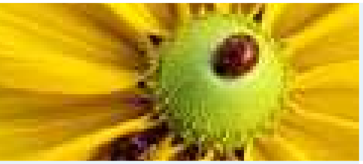
$$\boldsymbol{I} - \boldsymbol{\Delta}_d'(\boldsymbol{\Delta}_d\boldsymbol{\Delta}_d')^{-1}\boldsymbol{\Delta}_d \;=\; \boldsymbol{L}\boldsymbol{L}' \;,$$

  for full column-rank matrix $\boldsymbol{L}$

- Mixed model form of P-spline uses

$$\begin{aligned} \boldsymbol{y} &= \boldsymbol{B}\boldsymbol{L}\boldsymbol{L}'\boldsymbol{\alpha} + \boldsymbol{B}\boldsymbol{\Delta}_d'(\boldsymbol{\Delta}_d\boldsymbol{\Delta}_d')^{-1}\boldsymbol{\Delta}_d\boldsymbol{\alpha} + \boldsymbol{e} \\ &= \boldsymbol{X}_b\boldsymbol{\tau}_b + \boldsymbol{Z}_b\boldsymbol{u}_b + \boldsymbol{e}. \end{aligned}$$

- $\boldsymbol{X}_b = \boldsymbol{B}\boldsymbol{L}$ is a polynomial of degree d, $\boldsymbol{\tau}_b = \boldsymbol{L}'\boldsymbol{\alpha}$, $\boldsymbol{u}_b = \boldsymbol{\Delta}_d\boldsymbol{\alpha}$
- For d=2, fixed part of model represents linear trend, random part represents deviations about linear trend (as for O-spline)
- Sparse representation lost within mixed model context



Cubic P-splines

Comparison with O-spline:

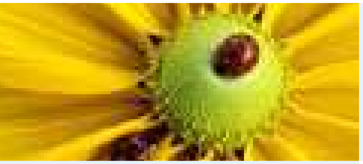
- both have choice of knots

P-splines:

- has choice of differencing order in penalty
- differencing order determines fixed terms in model
- non-natural basis (this can be modified)

O-splines:

- penalty determined by degree of spline
- optimisation property for full-knot basis (cubic smoothing spline)
- fixed terms determined by degree of spline basis
- can relax naturalness constraints



Penalised splines

Described in Semi-parametric Regression by Ruppert, Wand & Carroll (2003)

Penalised splines use:

- truncated power function basis, $\{T_j(x) = (x - t_j)_+^k \ ; \ j = 1 \ldots r\}$
- reduced knot set
- ad-hoc penalty to induce smoothing

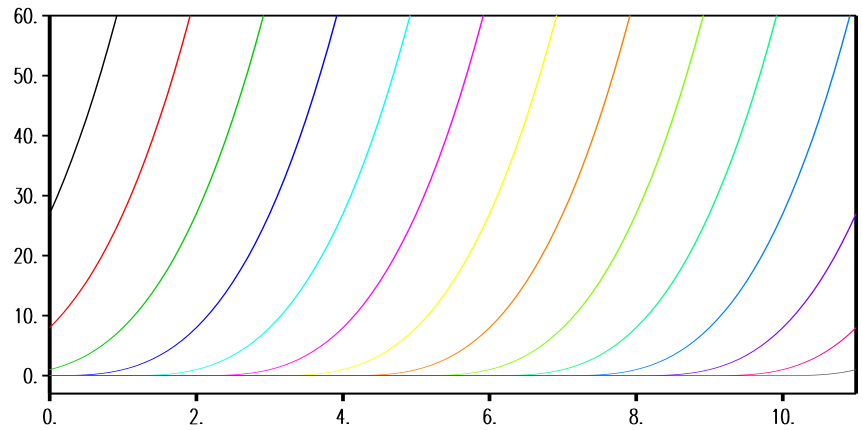
Model

$$g(x) \; = \; \sum_{j=0}^{k} \tau_{T,j} \, x^j + \sum_{j=1}^{r} \beta_j T_j(x) \quad \text{or} \quad g(\boldsymbol{x}) = \boldsymbol{X}_T \boldsymbol{\tau}_T + \boldsymbol{T}\boldsymbol{\beta}$$

where $\boldsymbol{X}_T = [\mathbf{1} \ \boldsymbol{x} \ , \ldots, \ \boldsymbol{x}^k]$, $\boldsymbol{\tau}_T = (\tau_{T,0}, \ldots, \tau_{T,k})'$, $\boldsymbol{T} = [\, T_1(\boldsymbol{x}), \ldots, T_r(\boldsymbol{x}) \,]$, $\boldsymbol{\beta} = (\beta_1 \ldots \beta_r)'$

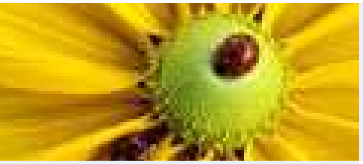
Penalised splines

Truncated power function (TPF) basis for knots $1 \ldots 10$ (excluding constant and linear)



TPF basis has properties

- easy to write down and calculate
- easy to add/remove knots from basis
- poor numerical properties of $\boldsymbol{T}$ (badly conditioned)



Penalised splines as mixed models

- Penalised sum of squares

$$[\,\boldsymbol{y} - \boldsymbol{X}_T\boldsymbol{\tau}_T - \boldsymbol{T}\boldsymbol{\beta}\,]'\boldsymbol{R}^{-1}[\,\boldsymbol{y} - \boldsymbol{X}_T\boldsymbol{\tau}_T - \boldsymbol{T}\boldsymbol{\beta}\,] \;+\; \lambda\boldsymbol{\beta}'\boldsymbol{\beta}$$

- ad-hoc penalty used to introduce shrinkage/smoothing
- User choice:
  - degree of basis
  - knot positions
- Mixed model formulation

$$\boldsymbol{y} + \boldsymbol{X}_T\boldsymbol{\tau}_T + \boldsymbol{T}\boldsymbol{\beta} + \boldsymbol{e}$$

with $\mathrm{var}\,(\boldsymbol{\beta}) = \sigma_s^2\boldsymbol{I}$, $\mathrm{var}\,(\boldsymbol{e}) = \sigma^2\boldsymbol{I}$.

Comparison of mixed model splines

O-spline	P-spline	Penalised spline
(Minimises general PSS)	No optimality	No optimality
Choice of k	Choice of k	Choice of k
-	Choice of d	-
$k \Rightarrow$ fixed terms	$d \Rightarrow$ fixed terms	$k \Rightarrow$ fixed terms
$k \Rightarrow$ penalty	$d \Rightarrow$ penalty	Identity penalty
(Choice of knots)	Choice of knots	Choice of knots
(Natural)	(Not natural)	(Not natural)

- what differences do these choices make?
- how can we compare the different penalties?

Connections between bases

Blurring over details (see Welham et al 2007):

- k-degree B-spline basis can be constructed as $(k+1)$th differenced TPF basis
- *i.e.* cubic B-spline basis is 4th difference of cubic TPF basis
- cubic spline basis functions can be constructed as 2nd differences of cubic TPF basis

Hence for equally-spaced knots, spacing $= h$, we have equivalent representations of cubic spline with

$$\boldsymbol{\beta} = \boldsymbol{\Delta}_2 \boldsymbol{\delta}/6h \ ; \quad \boldsymbol{\delta} = \boldsymbol{\Delta}_2 \boldsymbol{a}/h^2 \ ; \quad \boldsymbol{\beta} = \boldsymbol{\Delta}_4 \boldsymbol{a}/6h^3$$

Connections between penalties

Given the connections between bases, we can relate penalties:

- For O-spline and cubic P-spline with 2nd order differencing (recommended default)

  $$\boldsymbol{a}'\boldsymbol{\Delta}_2'\boldsymbol{\Delta}_2\boldsymbol{a} \propto \boldsymbol{\delta}'\boldsymbol{\delta} \approx \boldsymbol{\delta}'\boldsymbol{G}_s^{-1}\boldsymbol{\delta}$$

  so second difference penalty can be regarded as rough approximation to cubic smoothing spline penalty

- For cubic penalised spline

  $$\boldsymbol{\beta}'\boldsymbol{\beta} \propto \boldsymbol{a}'\boldsymbol{\Delta}_4'\boldsymbol{\Delta}_4\boldsymbol{a}$$

  so cubic penalised spline is equivalent to cubic P-spline with d=4

- Linear penalised spline (recommended default) is equivalent to linear P-spline with d=2.

- P-splines (or smoothing/O-splines) preferred (by me) because penalty is explicit

User choices that affect fitted spline

Degree of spline basis:

- for given penalty and reasonable number of knots, degree of basis makes little difference
- basis should be appropriate to aims of analysis
- eg. to estimate derivatives of curve, require differentiable function
- linear splines are continuous but not differentiable

User choices that affect fitted spline

Number of knots

- choice of both number and position of knots
- reducing number of knots reduces flexibility of fitted spline
- usually aim to have sufficient knots to allow flexible spline without incurring excessive computational load
- large regression spline literature on knot selection
- Ruppert et al (2003) advocate

$$r = \min\left\{\frac{n}{4}, 35\right\}$$

  with knots at quantiles of covariate
- Eilers & Marx strongly advocate equally spaced knots
- can produce examples for each set where it performs poorly, usually where gaps exist in distribution of covariate

User choices that affect fitted spline

Penalty matrix

- consider in P-spline context: penalty depends on order of differencing
- increased order of differencing $\Rightarrow$ more constrained/smoother spline
- in practice, selection of smoothing parameter by REML to optimise fit to data often (but not always) leads to similar fitted splines
- difficulties occur in comparing models with different penalties, as this also implies different fixed effects
- unresolved issue
- first step: understand implications of different penalties

Impact of penalty on fitted curve

Does order of penalty make any real difference?

Simulation study

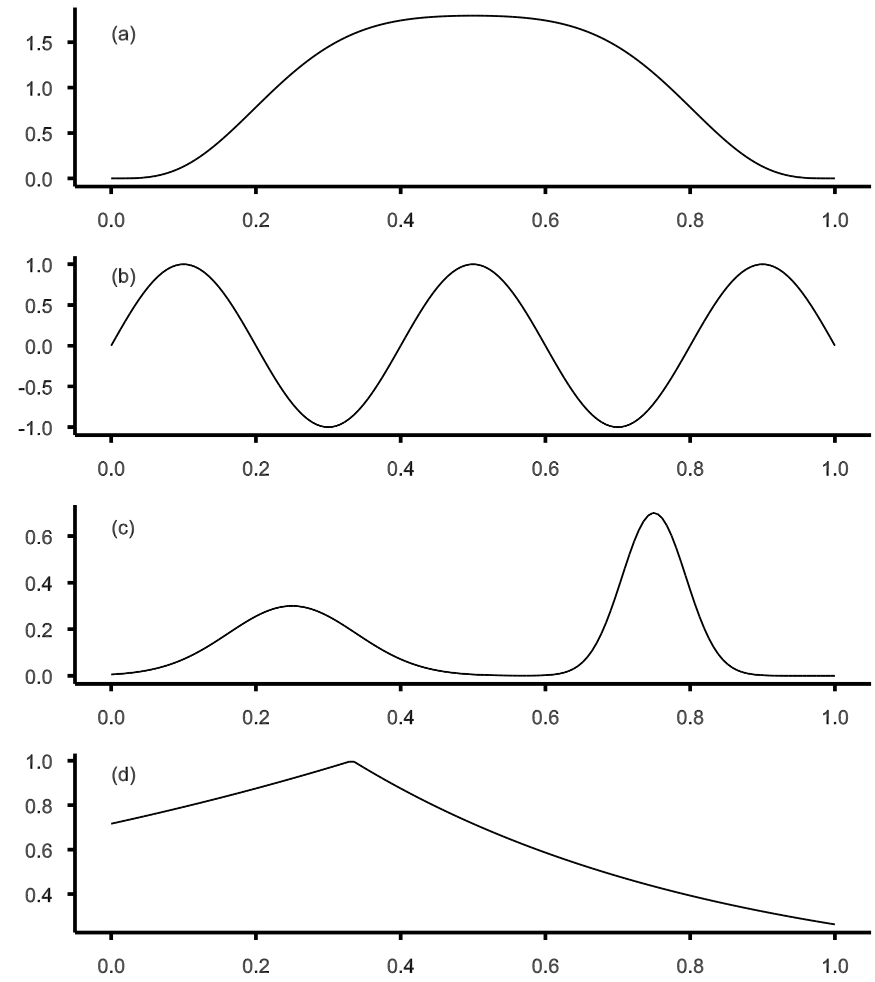
- four known functions on [0,1]
- data generated as function + error with normal errors at 32 or 128 equally-spaced points
- four different error variances, defined with as 0.5, 0.25, 0.1 or 0.05 $\times$ range of function

Compare splines, all with full knot set

- NCS = natural cubic spline
- cubic P-spline with d=1,2,3,4
- linear penalised spline

Evaluate models in terms of average mean squared error of prediction (MSEP) across 500 data sets

Known functions used for simulation study with different properties



Performance for function (a)

Average MSEP as % of minimum (se)

σ^2	d=1	d=2	d=3	d=4
0.5	127 (3.6)	100 (2.9)	104 (2.7)	117 (3.1)
0.25	127 (2.9)	100 (2.2)	126 (2.2)	112 (2.5)
0.1	142 (2.6)	100 (2.0)	112 (2.7)	111 (2.0)
0.05	203 (3.1)	111 (2.2)	100 (2.1)	110 (2.2)

100 x Average ratio estimated σ^2: actual σ^2 (% zero σ_s^2)

σ^2	d=1	d=2	d=3	d=4
0.5	91 (1)	97 (0)	99 (64)	98 (49)
0.25	85 (0)	99 (0)	104 (46)	101 (17)
0.1	72 (0)	97 (0)	104 (1)	106 (0)
0.05	53 (0)	91 (0)	97 (0)	104 (0)

- d=2 gives best MSEP for large σ^2
- d=3 gives best MSEP for small σ^2
- d=1 over-fits (allocates noise as signal)
- d=4 over-smooths (allocates signal as noise)

Performance for function (b)

Average MSEP as % of minimum (se)

σ^2	d=1	d=2	d=3	d=4
0.5	100 (1.7)	119 (1.5)	120 (1.5)	112 (1.9)
0.25	119 (2.9)	110 (4.3)	105 (4.5)	100 (3.0)
0.1	183 (2.7)	116 (2.0)	100 (1.9)	101 (1.9)
0.05	250 (2.7)	134 (2.1)	106 (1.8)	100 (1.8)

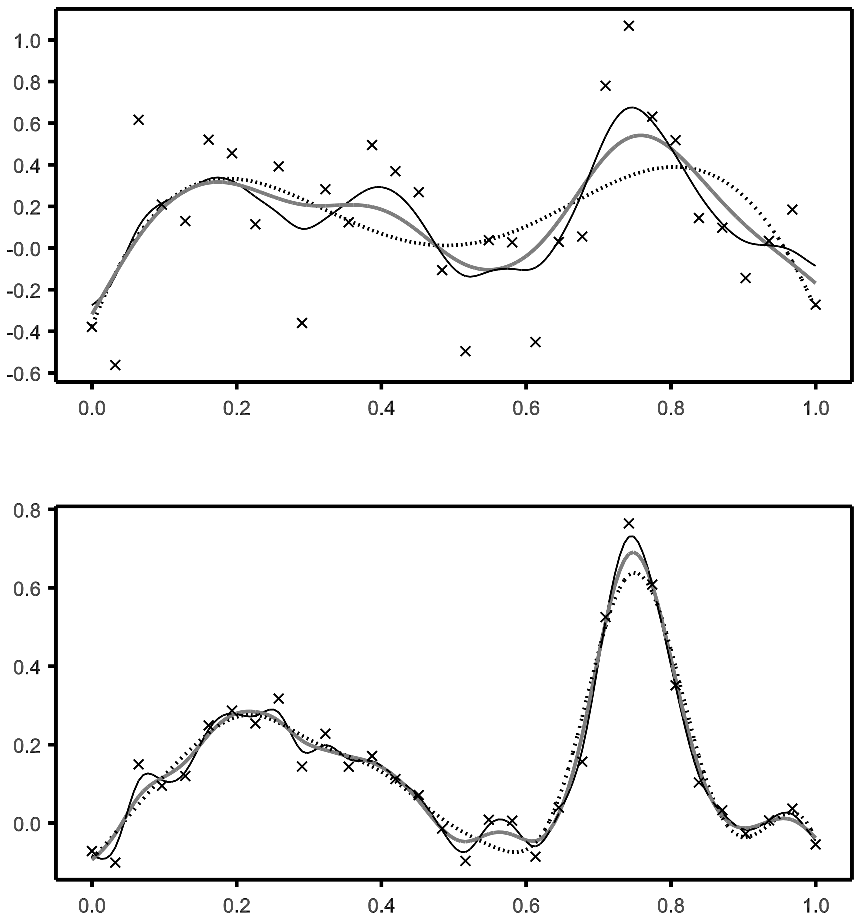
100 x Average ratio estimated σ^2: actual σ^2 (% zero σ_s^2)

σ^2	d=1	d=2	d=3	d=4
0.5	110 (39)	126 (38)	126 (62)	117 (47)
0.25	85 (1)	97 (1)	101 (4)	102 (1)
0.1	53 (0)	83 (0)	93 (1)	97 (0)
0.05	13 (0)	73 (0)	90 (0)	95 (0)

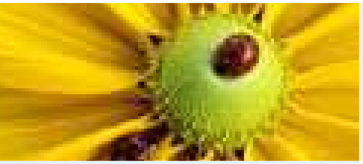
For large σ^2, pattern lost in noise for all splines (many zero spline variance components)

- d=3,4 give best MSEP
- d=1 over-fits (allocates noise as signal)
- d=2 slightly over-fits

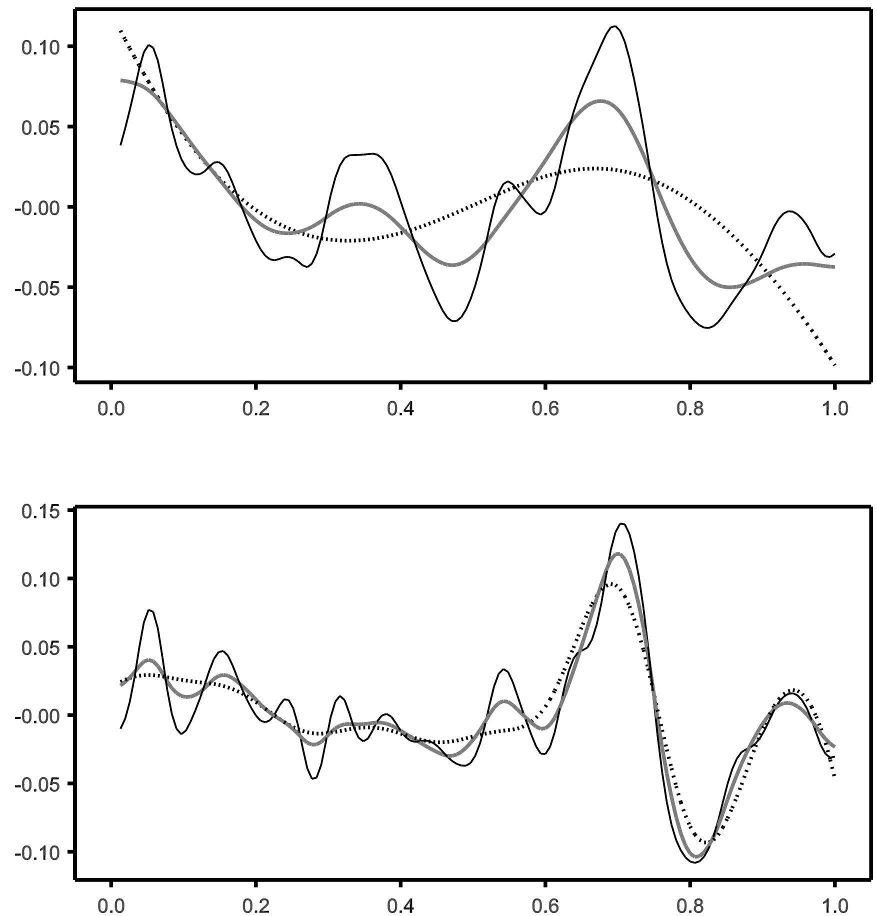
Performance for function (c)



Simulated data generated from function (c) with $\sigma^2 = 0.5$ (top) or 0.1 (bottom) with fitted splines from cubic P-spline mixed models with $d = 1$ (thin black line), $d = 2$ (thick grey line) or $d = 4$ (dotted line).



Performance for function (c)



Second differences of fitted mixed model splines. Fitted splines generated from cubic P-spline mixed models with $d = 1$ (thin black line), $d = 2$ (thick grey line) or $d = 4$ (dotted line).

Performance for function (c)

Average MSEP as % of minimum (se)

σ^2	d=1	d=2	d=3	d=4
0.5	100 (1.3)	111 (1.3)	116 (1.4)	116 (1.6)
0.25	100 (2.7)	132 (2.9)	153 (2.5)	150 (2.1)
0.1	101 (1.5)	100 (1.7)	155 (3.5)	255 (4.0)
0.05	140 (1.6)	100 (1.4)	110 (1.8)	181 (7.7)

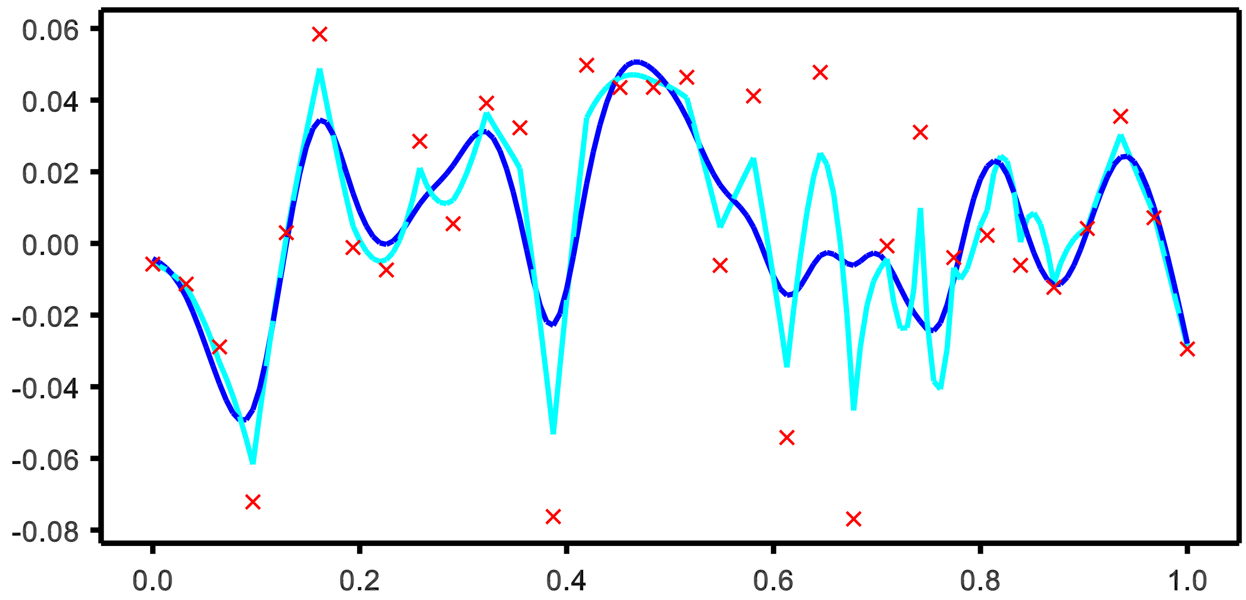
100 x Average ratio estimated σ^2: actual σ^2 (% zero σ_s^2)

σ^2	d=1	d=2	d=3	d=4
0.5	109 (49)	115 (51)	115 (50)	113 (27)
0.25	104 (10)	133 (14)	147 (15)	146 (3)
0.1	68 (0)	109 (0)	158 (0)	234 (0)
0.05	25 (0)	83 (0)	117 (0)	185 (0)

For large σ^2, pattern lost in noise for all splines (many zero spline variance components)

- d=1,2 gives best MSEP
- d=3,4 over-smooths (allocates signal as noise)

Performance for function (c)



Generated errors (×) with $\sigma^2 = 0.05$ with deviations of fitted splines from underlying function f_C. Splines were fitted with knots at data points as a cubic P-spline mixed model with $d = 2$ (—) or linear P-spline with $d = 2$ (ie. linear penalised spline —).

Linear splines have continuous function but derivatives are not continuous.

Performance for function (d)

Average MSEP as % of minimum (se)

σ^2	d=1	d=2	d=3	d=4
0.5	101 (3.1)	100 (2.6)	101 (2.3)	111 (3.2)
0.25	103 (2.7)	100 (2.7)	102 (2.4)	103 (2.5)
0.1	130 (2.7)	100 (1.7)	123 (2.1)	139 (2.4)
0.05	146 (2.5)	100 (1.8)	120 (2.2)	136 (2.5)

100 x Average ratio estimated σ^2: actual σ^2 (% zero σ_s^2)

σ^2	d=1	d=2	d=3	d=4
0.5	93 (1)	98 (30)	98 (42)	97 (66)
0.25	90 (0)	99 (2)	100 (11)	99 (61)
0.1	81 (0)	98 (0)	106 (0)	108 (37)
0.05	71 (0)	98 (0)	110 (0)	116 (2)

- for large σ^2, all splines similar as discontinuity masked by noise
- as noise decreases, discontinuity becomes apparent and fitted better by small d
- d=2 fits uniformly better: balances good fit of smooth trend with adaptation to discontinuity

Impact of the penalty

Overall summary:

- $d = 1$ fits detail well, but smooth trend badly
- $d = 4$ fits detail badly, smooth trend well
- As signal:noise ratio increases, smooth trend becomes more apparent and optimal order of d tends to increase
- Hence, optimal d depends on characteristics of curve and signal:noise ratio
- As d increases, splines are more likely to allocate high frequency pattern as noise
  - for small d, there is a tendency to over-fit.
  - for large d, there is a tendency to over-smooth.
  - both ameliorated by replication enabling independent estimate of error

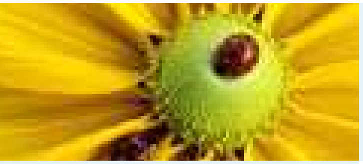
Which penalty should you use?

For exploratory analysis:

- purpose is to explore shape of underlying function
- use spline with small d, eg. cubic spline, cubic P-spline with d=2

For modelling curves:

- purpose is to get reliable description of underlying function
- more problematic: no objective method to choose between penalties
- requires further work
- wider class of L-splines that uses different penalties, useful for periodic data, should also be considered (Welham *et al*, 2006)
- we are looking at diagnostics to identify which spline is most consistent with data



A unified form of mixed model spline

Use B-spline basis, because this seems the most natural scale.

For spline basis $\{Bi(x); i = 1q\}$, we use model

$$\boldsymbol{y} = \sum_{i=1}^{q} \alpha_i B_i(\boldsymbol{x}) + \boldsymbol{e} = \boldsymbol{B\alpha} + \boldsymbol{e}$$

with penalised sum of squares

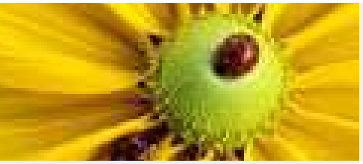
$$(\boldsymbol{y} - \boldsymbol{B\alpha})' \boldsymbol{R}^{-1} (\boldsymbol{y} - \boldsymbol{B\alpha}) + \lambda \boldsymbol{\alpha}' \boldsymbol{G}^{-1} \boldsymbol{\alpha}$$

where G is a semi-positive definite symmetric penalty matrix (eg. $\boldsymbol{\Delta}'_d \boldsymbol{\Delta}$ or an evaluated integral).

We can extend this model to impose c constraints of the form $\boldsymbol{C}'\boldsymbol{\alpha} = \boldsymbol{0}$ (eg. naturalness or periodicity) via $\boldsymbol{\alpha} = \boldsymbol{Sv}$ where $\boldsymbol{C}'\boldsymbol{S} = \boldsymbol{0}$.

The penalised sum of squares (PSS) then becomes

$$(\boldsymbol{y} - \boldsymbol{BSv})' \boldsymbol{R}^{-1} (\boldsymbol{y} - \boldsymbol{BSv}) + \lambda \boldsymbol{v}' \boldsymbol{S}' \boldsymbol{G}^{-1} \boldsymbol{Sv}$$



A unified form of mixed model spline

This general model can be fitted via eigen-value decomposition:

$$\boldsymbol{S}\boldsymbol{G}^{-1}\boldsymbol{S} = \boldsymbol{U}'\boldsymbol{D}\boldsymbol{U}$$

for some $(q-c) \times (q-c-s)$ matrix $\boldsymbol{U}$ with $\boldsymbol{U}'\boldsymbol{U} = \boldsymbol{I}_{q-c-s}$ and $\boldsymbol{D} = \text{diag}\{d_1 \ldots d_{q-c-s}\}$ with $d_i > 0$ for $i = 1 \ldots q-c-s$.

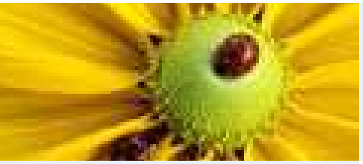
Then we can write

$$\boldsymbol{B}\boldsymbol{S}\boldsymbol{v} = \boldsymbol{B}\boldsymbol{S}\boldsymbol{L}\boldsymbol{\tau} + \boldsymbol{B}\boldsymbol{S}\boldsymbol{U}\boldsymbol{D}^{0.5}\boldsymbol{u}$$

We can use this form as a mixed model with

- $\boldsymbol{\tau}$ is an $s \times 1$ vector of fixed effects and design matrix $\boldsymbol{B}\boldsymbol{S}\boldsymbol{L}$
- $\boldsymbol{u}$ is a $(q-c-s) \times 1$ vector of random effects with design matrix $\boldsymbol{B}\boldsymbol{S}\boldsymbol{U}\boldsymbol{D}^{0.5}$ and variance $\sigma_s^2 \boldsymbol{I}_{q-c-s}$ for $\lambda = \sigma^2/\sigma_s^2$

It is straightforward to show that BLUPs from this mixed model minimise the PSS.



A unified form of mixed model spline

This general model allows easy transformation between bases.

Suppose we have two bases

- represented by matrices $\boldsymbol{T\beta}$ and $\boldsymbol{B\alpha}$
- with $\boldsymbol{T} = \boldsymbol{BA}$ for some full rank matrix $\boldsymbol{A}$
- then $\boldsymbol{\alpha} = \boldsymbol{A\beta}$

We can then rewrite the PSS as

$$(\boldsymbol{y} - \boldsymbol{BA\beta})'\boldsymbol{R}^{-1}(\boldsymbol{y} - \boldsymbol{BA\beta}) + \lambda\boldsymbol{\beta}'\boldsymbol{A}'\boldsymbol{G}^{-1}\boldsymbol{A\beta}$$

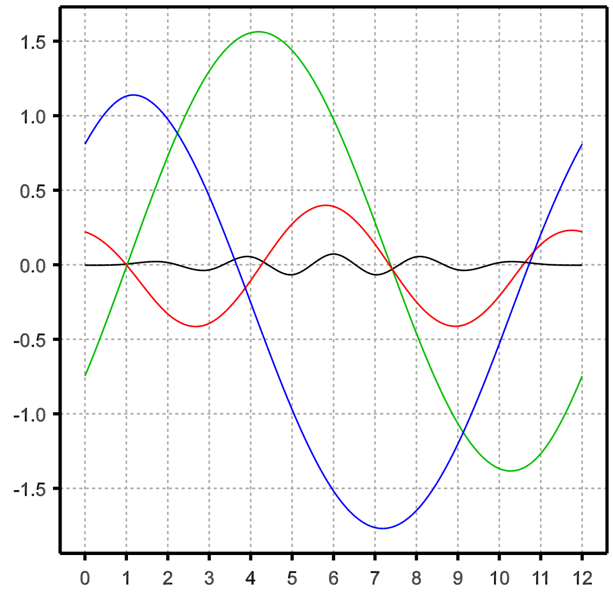
and use the resulting mixed model.

Example

Example: cubic smoothing spline with additional periodic constraints.

- suppose data periodic on range $[0, \omega]$
- use full TPF basis with penalty $\int_0^\omega [g''(s)]^2 ds$ expressed as $\boldsymbol{\beta}' \boldsymbol{G}^{-1} \boldsymbol{\beta}$
- impose constraints $g^{(j)}(0) = g^{(j)}(\omega)$ for $j = 0, 1, 2$.

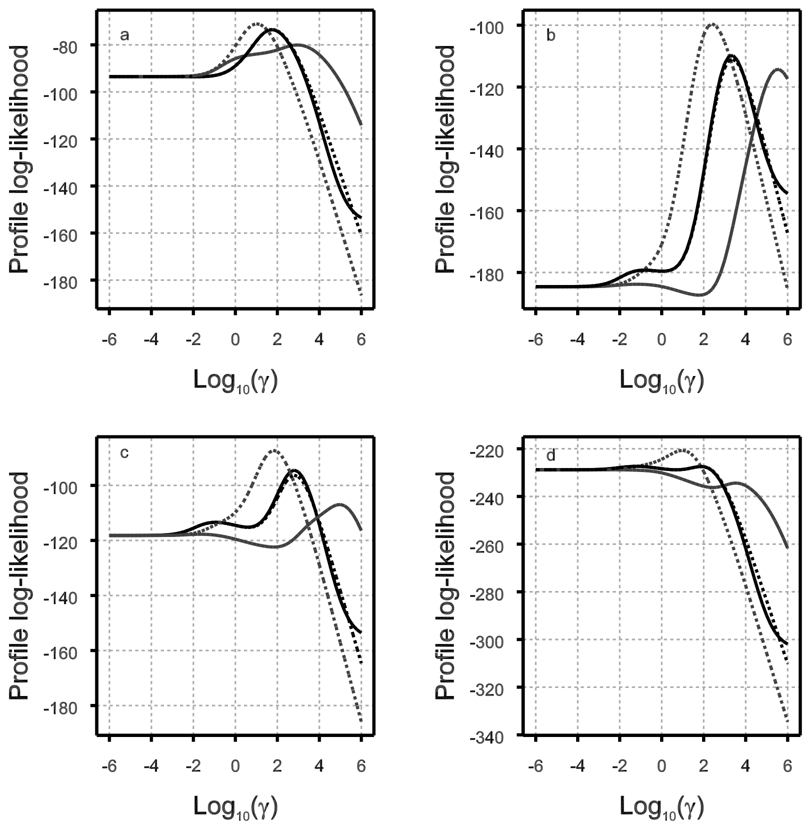
Derived basis has only 1 fixed term: constant, with constrained basis functions:



An issue

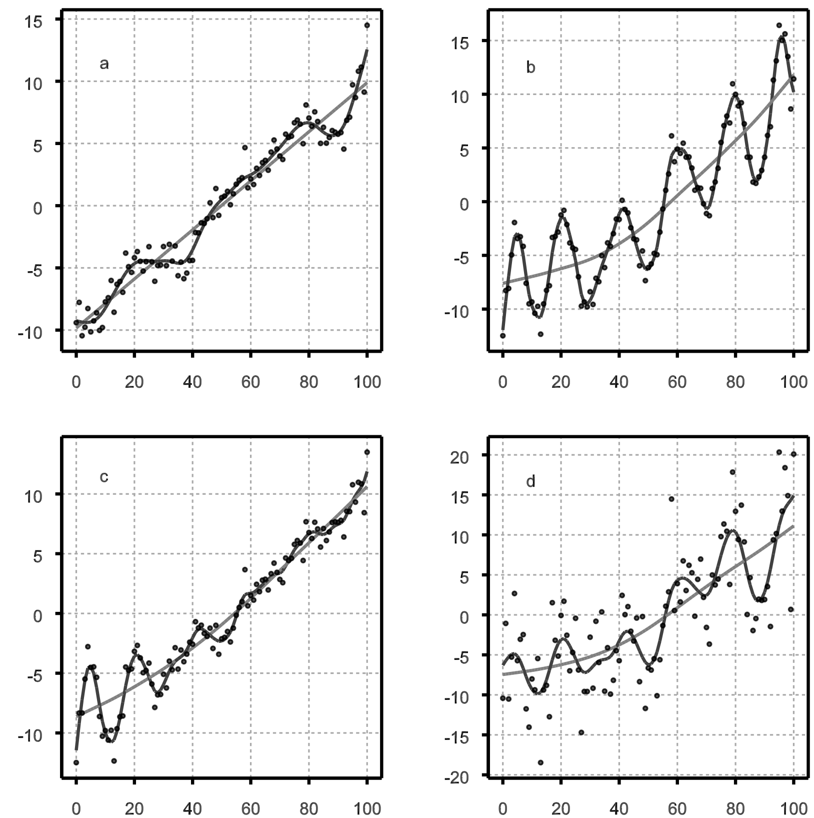
For some data sets, log-likelihood function is bimodal:

- several fitted splines may be obtained from standard algorithms
- fitted spline dependent on initial value of smoothing parameter



An issue

Introduction
Simple spline model
Polynomial splines
Spline basis
Regression splines
Smoothing splines
Cubic SS
Mixed model CSS
Inference
More complex models
Example
Reduced knot spline
P-splines
Penalised splines
Comparison
Connections
User choices
Penalty
Unified model
An issue
References



- This feature common to smoothing & penalised splines
- Usually indicates incompatibility between data & spline model - cubic spline model assumes non-linear trend dominated by low frequency components
- Problem not flagged in software packages

References

Brumback, B. A., and Rice, J. A. (1998). Smoothing spline models for the analysis of nested and crossed samples of curves (with discussion). *Journal of American Statistical Association* **93,** 961–994.
Eilers, P. H. C., and Marx, B. D. (1996). Flexible smoothing with B-splines and penalties. *Statistical Science* **11,** 89–121.
Green P.J. and Silverman B.W. (1994). *Nonparametric Regression and Generalized Linear Models*. Chapman and Hall, London.
Hastie, T. J., Tibshirani, R. J., and Friedman, J. (2001). *The Elements of Statistical Learning*. New York: Springer-Verlag.
Ruppert, D., Wand, M. P., and Carroll, R. J. (2003). *Semiparametric regression*. Cambridge: Cambridge University Press.
Verbyla, A. P., Cullis, B. R., Kenward, M. G., and Welham, S. J. (1999). The analysis of designed experiments and longitudinal data using smoothing splines. *Applied Statistics* **48,** 269–311.
Welham, S. J., Cullis B. R., Kenward, M. G., and Thompson, R. (2006). The analysis of longitudinal data using mixed model L-splines. *Biometrics* **62,** 392–401.
Welham, S. J., Cullis B. R., Kenward, M. G. and Thompson, R. (2007). A comparison of mixed model splines. *Australian and New Zealand Journal of Statistics* **49,** 1–23.
Welham, S. J., and Thompson, R. (2008). A note on bimodality in the log-likelihood function for penalized spline mixed models. *Computational Statistics and Data Analysis* **in press**.
Wood, S.N. (2006). *Generalized Additive Models: An Introduction with R*. Chapman & Hall/CRC, Boca Raton.
Wood, S. N. (2006). Low-rank scale-invariant tensor product smooths for generalized additive mixed models. *Biometrics* **62,** 1025–1036.
Zhang, D. W., and Lin, X. H. (2003). Hypothesis testing in semiparametric additive mixed models. *Biostatistics* **4,** 57–74.
Zhao, Y., Staudenmayer, J., Coull, B. A., and Wand M. P. (2006). General design Bayesian generalized linear mixed models. *Statistical Science* **21,** 35–51.