Random-Effects Spline Models

David Ruppert

School of OR & IE

Cornell University

The overheads and references for this lecture and related talks and papers are available at:

www.orie.cornell.edu/∼davidr

Two Messages

1. **Most nonparametric regression problems can be solved easily and effectively by reformulating them as**

   - *linear mixed models*, for **Gaussian responses**.
   - *generalized linear mixed models*, for **other types of responses**.

2. **Once a mixed model formulation is adopted, it is easy to extend nonparametric regression models** to *nested and crossed families of curves*, e. g., to model **longitudinal** data.

Outline

1. **Regression splines** are **linear models** but:
   - allow **essentially "nonparametric"** fitting.
   - are ideal nonparametric components in semiparametric models.
2. **High dimensionality** of the spline's parameter space can be accommodated by **penalized estimation** (**P-splines**).
3. Approach today:
   - **P-splines** are recast as **linear mixed models**.
   - The **spline fit** is the **BLUP** (or EBLUP).
   - The all-important **smoothing parameter** can be selected by **ML or REML**.

Outline, continued

4. **Today's ideas are not new,** but

    - They have been developed in the framework of **smoothing splines,** which causes **computational problems** because of the **excessive number of knots.**
    - They **haven't been well advertised** or well packaged.
    - Their **implications are just beginning to be appreciated.**

5. **Extensions** to **nested or crossed families of curves** are easy and will be important, e. g., to longitudinal data.

6. **Extensions** to other general **semiparametric** models are also **easy**.

Rodent Carcinogenicity Experiments

- From Parise, Ruppert, Ryan, and Wand (1999).
- **Animals are randomly assigned to one of several doses of a suspected carcinogen**, e. g., phenolphthalein in our example.
- **Outcomes are presence/absence of cancer**s of specific types, e. g., mammary tumors in our example.
- **Time of death is a covariate.**

Carcinogenicity Experiments — Controlling for Body Weight

- **Basic underlying fact: Lower body weight $\Rightarrow$ lower cancer incidences**.
  - **Many suspected carcinogens are toxic and cause lower body weight, thus masking their carcinogenicity**.
  - **A recent editoral in *Science* (Abelson, 1995) was highly critical of risk assessment agencies for not controlling for the effects of body weight.**
  - The **effects of body weight need not be linear**, even on the logit scale, e. g., there may be a threshold.
  - **Needed:** a **semiparametric model** including
    - **dose**
    - **time of death**
    - **body weight**

    and **with the effects of body weight modeled nonparametrically**.

Controlling for Body Weight with Historical Controls

- **Problem:** Body Weight and Dosage Are Confounded.
- **Solution:** Use historical controls.
  - Dosage is 0, since only controls are used.
  - **Body weight of historical controls has a large range** and includes that of all animals, controls and dosed, in the current experiment.
- **Random experiment-specifc intercepts controls for shifts in cancer rate over time.**

Historical Controls — Model

- $i =$ **experiment number**
  - $i = 0$ means current experiment
  - $i = 1, 2, \ldots$ means historical experiments.
- $y_{ij} =$ **indicator that** j **th rodent in** i **th experiment has mammary tumors**.
- **logit**$[(P(y_{ij} = 1)] = \alpha_0 + \alpha_1 t_{ij} + \gamma d_{ij} + s(w_{ij}) + u_i$**.**
- $t_{ij} =$ **time of death**
- $d_{ij} =$ **dosage**
- w_{ij} **= body weight at 12 months**
- s **is a spline**
  - So this is a **semiparametric** model.
- u_i **are IN(**$0, \tau^2$**).**
  - So this is a **mixed** model.
- We'll return to this model at the end.
  - By modeling the spline appropriately, the entire model becomes a *GLMM (generalized linear mixed model)*.

Regression Spline Models

$$E(Y_i|X_i) = m(X_i;\boldsymbol{\beta}) := \beta_0 + \beta_1 X_i + \cdots + \beta_p X_i^p + \sum_{j=1}^{K} \beta_{p+j}(X_i - t_j)_+^p$$

- **Here** $t_1 < \cdots < t_K$ **are fixed "knots"**
- $u_+^p = u^p$ **if** $u > 0$ **and equals 0 otherwise.**
- $(x-t)_+^p$ **is called a truncated power function**.
- $m(x;\boldsymbol{\beta})$ **is a** p**th degree polynomial on each interval between knots and has** $p-1$ **continuous derivatives everywhere**.
- **Linear in** $\boldsymbol{\beta} = (\beta_0, \ldots, \beta_{p+K})^t \Rightarrow$ **a linear model**.
- **References:**
  - Ruppert and Carroll (2000)
  - Brumback, Ruppert, and Wand (1999)
  - Eilers and Marx (1996)

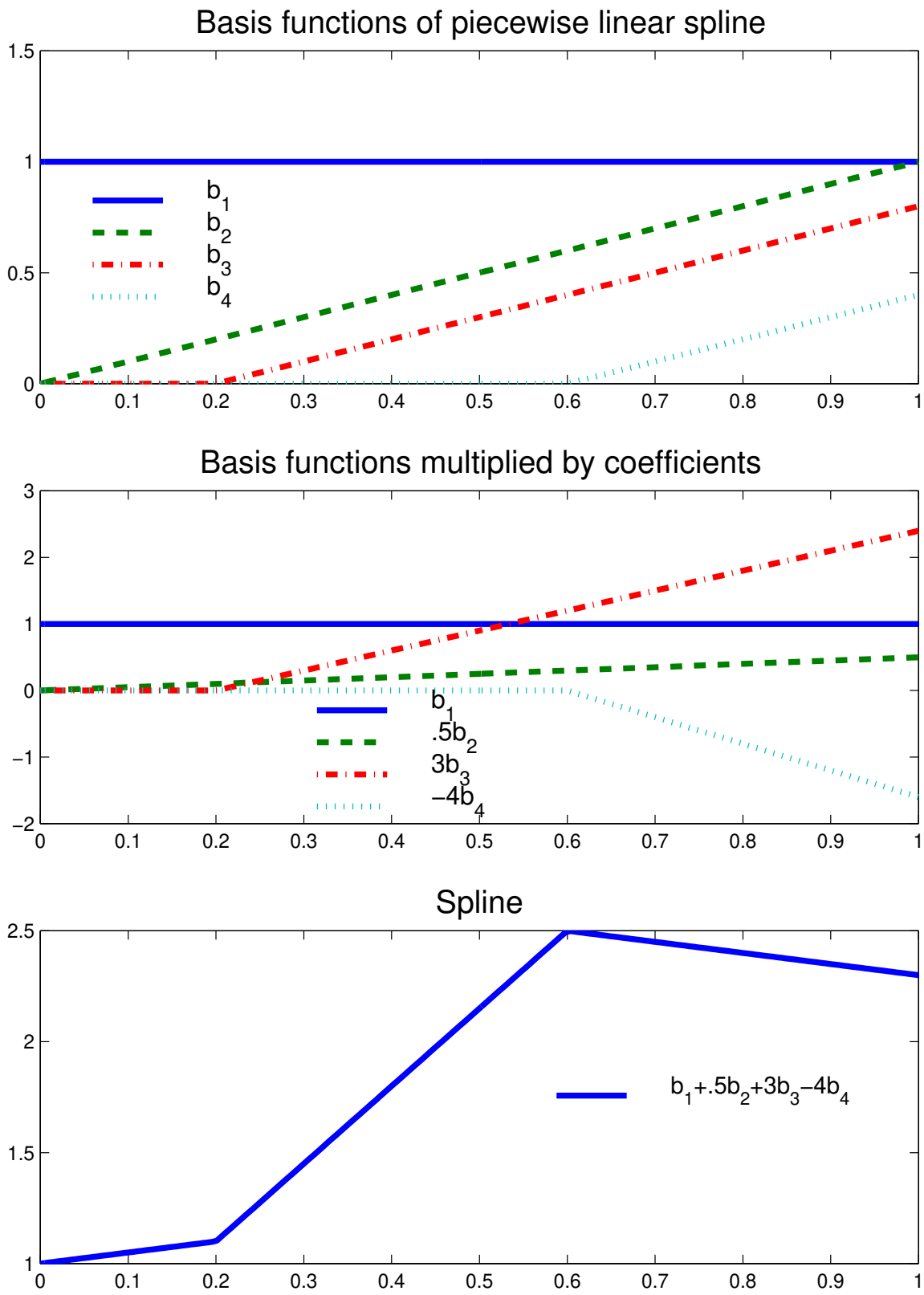


Figure 1: **Piecewise linear spline and basis functions.**

P-splines

- Let $\widehat{\boldsymbol{\beta}}$ minimize

$$\sum_{i=1}^{n}[Y_i - m(X_i;\boldsymbol{\beta})]^2 + \lambda \sum_{j=1}^{K} \beta_{p+j}^2$$

- This **penalized least-squares estimator** will be called a penalized spline, or **P-spline** (term due to Eilers and Marx (1996)).
- Other quadratic penalties on $\boldsymbol{\beta}$ are possible.
  - In particular, one can find $\boldsymbol{B}$ so that

$$\boldsymbol{\beta}^{\mathsf{T}}\boldsymbol{B}\boldsymbol{\beta} = \int_{\min(X_i)}^{\max(X_i)}[m''(x;\boldsymbol{\beta})]^2 dx.$$

- The choice of spline basis is unrestricted.
  - Truncated power functions are not needed, though we find them convenient.
  - B-splines could be used for numerical stability, but there are other ways to achieve numerical stability.
  - A natural cubic spline basis could be used, so that cubic smoothing splines are a special case of P-splines.

Mixed Linear Models

The linear mixed model is

$$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{Z}\boldsymbol{u} + \boldsymbol{\epsilon} = \begin{bmatrix} \boldsymbol{X} & \boldsymbol{Z} \end{bmatrix} \begin{bmatrix} \boldsymbol{\beta} \\ \boldsymbol{u} \end{bmatrix} + \boldsymbol{\epsilon}$$

where

- $\boldsymbol{X}$ and $\boldsymbol{Z}$ are known matrices.
- $\boldsymbol{\beta}$ **is a fixed parameter vector** (its components are called the fixed effects).
- $\boldsymbol{u}$ **is a random vector** (its components are called the random effects).
- $E(\boldsymbol{u}) = 0$, $E(\boldsymbol{\epsilon}) = 0$, and

$$\operatorname{Var} \begin{bmatrix} \boldsymbol{u} \\ \boldsymbol{\epsilon} \end{bmatrix} = \begin{bmatrix} \boldsymbol{G}(\boldsymbol{\theta}) & 0 \\ 0 & \boldsymbol{R}(\boldsymbol{\theta}) \end{bmatrix}.$$

- $\boldsymbol{\theta}$ is a vector of **"variance components."**
- Typically, $(\boldsymbol{u}, \boldsymbol{\epsilon})$ is jointly normal. Then $\boldsymbol{\beta}$ **and** $\boldsymbol{\theta}$ **can be estimated by ML** (Maximum Likelihood) or REML (Restricted or Residual ML).
- $\boldsymbol{u}$ **can be predicted by BLUP** (Best Linear Unbiased Prediction).

Generalized Linear Mixed Models

For a GLMM,

- **$\boldsymbol{X}$, $\boldsymbol{Z}$, $\boldsymbol{\beta}$, and $\boldsymbol{u}$ are as in the linear mixed model.**
- Y_i, X_i and Z_i are the i th components of $\boldsymbol{Y}$, $\boldsymbol{X}$, and $\boldsymbol{Z}$.
- The **distribution of** Y_i given (X_i, Z_i) is in an **exponential family**, e.g., is Bernoulli (or binomial) for logistic regression.
- $$\eta\{E(Y_i|X_i, Z_i)\} = X_i\boldsymbol{\beta} + Z_i\boldsymbol{u}$$
  for a known **"link" function** η**,** e.g., η is the logit function for logistic regression.

Mixed Model Formulation

Suppose that

1. $Y_i = \beta_0 + \beta_1 X_i + \cdots + \beta_p X_i^p + \Sigma_{j=1}^K u_j (X_i - t_j)_+^p + \epsilon_i$
2. $u_1, \ldots, u_K$ are random variables.
3. $E(u_j) = 0 \quad \forall j$.
4. Var$(u_j) = \sigma_u^2 \quad \forall\, j$.
5. **The u_j s are mutually independent.**
6. $\{\epsilon_i\}_{i=1}^n$ are IN$(0, \sigma^2)$.

5. may be the least tenable of these assumptions. However, it is a reasonable, simple starting point. Time series models for correlation, e. g., AR(1), could be used instead. Similarly, 6. could be weakened.

The u_j s are now random effects, so we have a *linear mixed model.*

Mixed Model Formulation, cont.

Now let

$$\boldsymbol{X} = \begin{bmatrix} 1 & X_1 & \cdots & X_1^p \\ 1 & X_2 & \cdots & X_2^p \\ \vdots & \vdots & \ddots & \vdots \\ 1 & X_n & \cdots & X_n^p \end{bmatrix},$$

$$\boldsymbol{Z} = \begin{bmatrix} (Z_1 - t_1)_+^p & \cdots & (Z_1 - t_K)_+^p \\ (Z_2 - t_1)_+^p & \cdots & (Z_2 - t_K)_+^p \\ \vdots & \ddots & \vdots \\ (Z_n - t_1)_+^p & \cdots & (Z_n - t_K)_+^p \end{bmatrix},$$

$$\boldsymbol{Y} = (Y_1, \ldots, Y_n)^\mathsf{T},$$

$$\boldsymbol{\beta} = (\beta_0, \ldots, \beta_p)^\mathsf{T}, \quad \boldsymbol{u} = (u_1, \ldots, u_K)^\mathsf{T},$$

and

$$\boldsymbol{\epsilon} = (\epsilon_1, \ldots, \epsilon_n)^\mathsf{T}.$$

Then the mixed model can be put in standard form:

$$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{Z}\boldsymbol{u} + \boldsymbol{\epsilon}.$$

Estimation

P-Splines — β and u are fixed

$$\begin{bmatrix} \boldsymbol{X}^\mathsf{T}\boldsymbol{X} & \boldsymbol{X}^\mathsf{T}\boldsymbol{Z} \\ \boldsymbol{Z}^\mathsf{T}\boldsymbol{X} & \boldsymbol{Z}^\mathsf{T}\boldsymbol{Z} + \lambda\boldsymbol{I} \end{bmatrix} \begin{pmatrix} \widehat{\boldsymbol{\beta}} \\ \widehat{\boldsymbol{u}} \end{pmatrix} = \begin{pmatrix} \boldsymbol{X}^\mathsf{T}\boldsymbol{Y} \\ \boldsymbol{Z}^\mathsf{T}\boldsymbol{Y} \end{pmatrix} \tag{1}$$

Typically, λ is chosen by GCV.

BLUPS - β fixed and u random

$\widehat{\beta}$ **(MLE)** and $\widehat{u}$ **(BLUP)** solve (1) with

$$\lambda = \frac{\sigma^2}{\sigma_u^2}.$$

This result is a special case of equation (1.2) of Robinson's (1991) review article on BLUP's. The original result is due to Henderson (1950) and (1) is sometimes called the "mixed models equations." Thus, P-splines are BLUPs. This fact was noted by Speed (1991) for smoothing splines.

λ could again be chosen by GCV, but **the mixed model formulation suggests ML or REML estimation of the two variance components and set λ equal to**

$$\hat{\lambda} = \frac{\widehat{\sigma}^2}{\widehat{\sigma}_u^2}$$

GCV versus REML

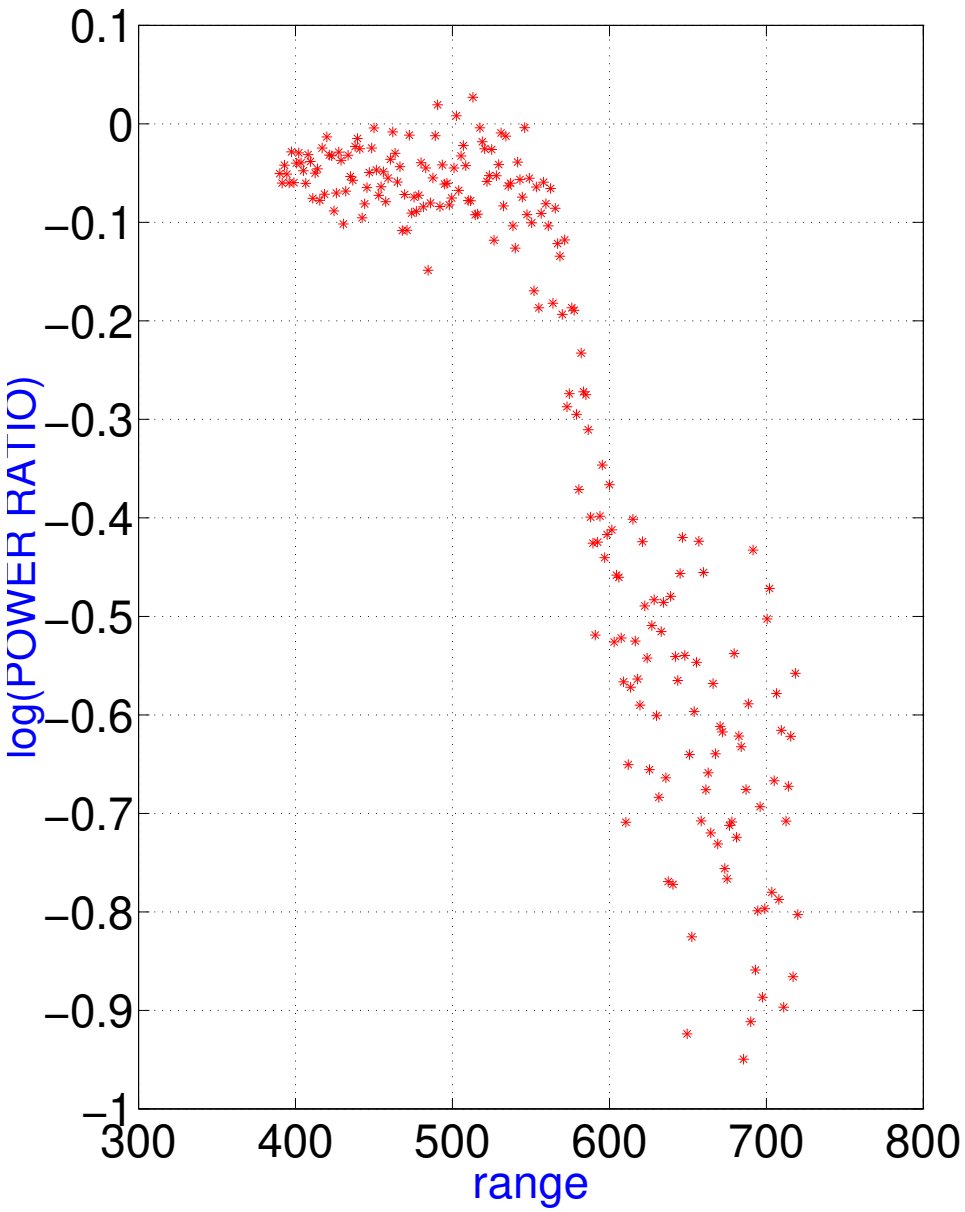
- Wahba (1985, *Annals*) introduced generalized maximum likelihood (GML) estimation for the smoothing parameter of smoothing splines. GML $=$ REML.
- Kohn, Ansley, and Tharm (1991, *JASA*) compared GCV to REML (which they called ML or marginal ML).
  - The methods were comparable, with REML having a slight edge.
- Yuedong Wang (1998, *JASA*) discusses REML and GCV with correlated errors. Correlation can be due to:
  - random effects.
  - correlated errors (the ϵ_i s).

  Time series models for the ϵ_i s fit naturally into the mixed model/REML framework.

  (See Altman (1990, *JASA*) and Hart (1991, *JRSS-B*) for earlier work on smoothing with correlated errors.)

LIDAR Data

Goal: detect mercury in an emissions plume. **Reference:** Ruppert (1997)

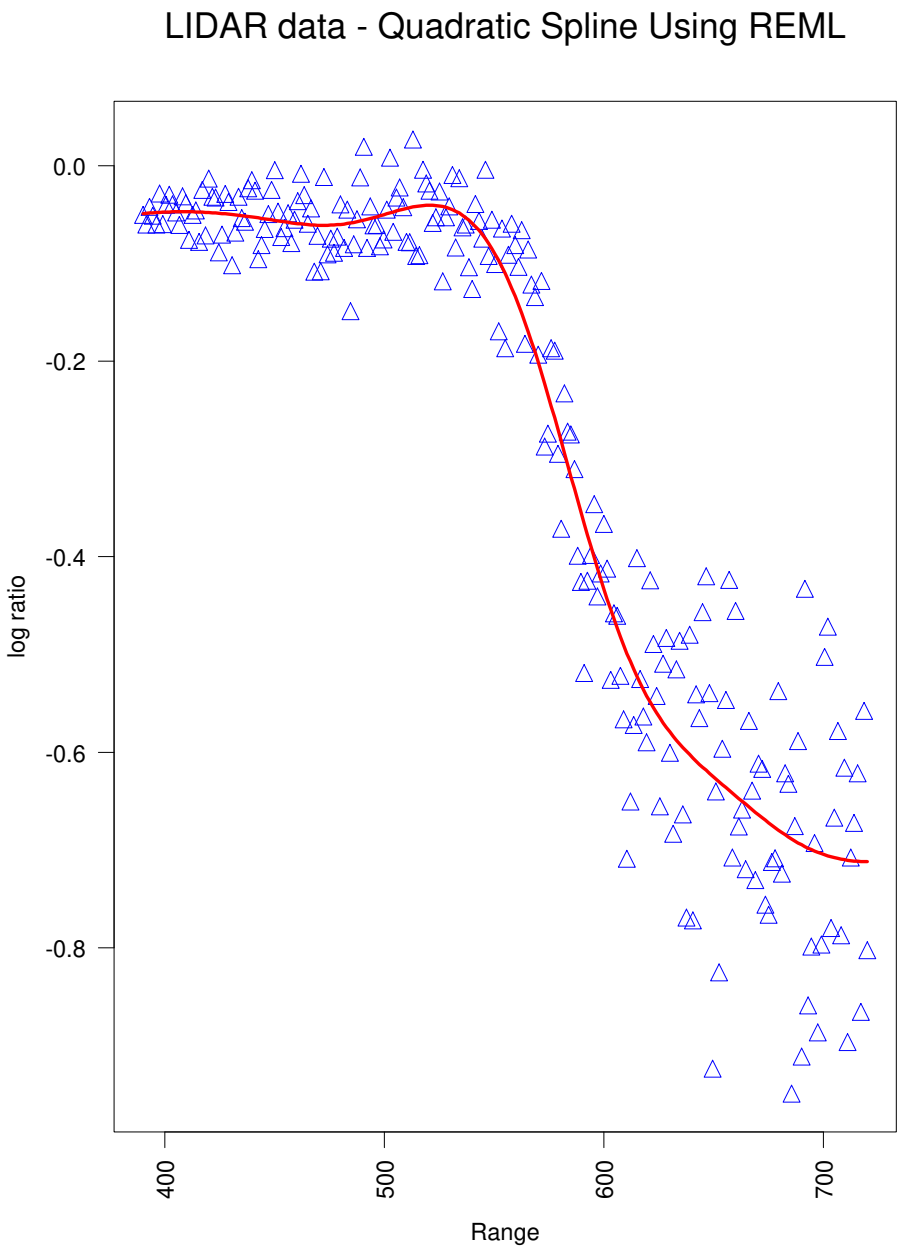


LIDAR example — LME Program

```
######### lid.lme.S ##########
# For fitting a P-spline to the LIDAR data
# using lme(), with REML smoothing parameter choice.

num.knots <- 30
data <- read.table("lidar.dat",header=T)
x <- data$range
y <- data$logratio
n <- length(x)
knots <- seq(min(x),max(x),length=(num.knots+2))
     [-c(1,(num.knots+2))]
          # Equally spaced knots
Zmat <- matrix(0,n,num.knots)
for (j in 1:num.knots)
     Zmat[,j] <- ((x-knots[j])*(x>knots[j]))^2
Xmat <- cbind(x,x^2)
data.fr <- cbind(y,Xmat,Zmat)
lidar.fit <- lme(y~Xmat,
     random = ~ Zmat -1 ,
     cluster = ~ rep(1,n),
     data.fr,
     re.structure="identity",
     est.method="RML")
plot(data$range, lidar.fit$fitted$cluster,
     type="l",
     ylim=c(min(data$logratio),
     max(data$logratio)),
     xlab="Range",ylab="log ratio",
     main="LIDAR data - Quadratic Spline Using REML")
points(data$range,data$logratio,pch=1)
########## End of lid.lme.S  ##########
```

LIDAR Data — LME Fit



Testing Parmetric Null Hypotheses

- There is a large literature on testing **parametric null** hypotheses versus **nonparametric alternatives**.
- Hastie and Tibshirani's (1990) book on GAMs discusses approximate F-tests
- Hart's (1997) book discusses other tests.
- In the mixed model framework, **standard likelihood ratio tests** can be used.
- Example:
  - **H$_0$:** $m(x)$ **is linear.**
  - **H$_1$:** $m(x)$ **is nonlinear.**
  - To test, use the model:
    $m(x) = \beta_0 + \beta_1 x + \Sigma_{j=1}^{K} u_j (x - t_j)_+$ with $\{u_j\}_{j=1}^{K}$ distributed IN$(0, \sigma_u^2)$.
    Then the hypotheses are:
    - **H$_0$:** $\sigma_U^2 = 0$**.**
    - **H$_1$:** $\sigma_U^2 > 0$**.**
- **Null distribution doesn't depend on distributional assumptions about the u_j's**, i. e., the normality and independence assumptions.
- However, asymptotic **null distribution is not "standard"** since the null value of the **parameter vector is on the boundary** (see Self and Liang (1987))

Nested Models

- Suppose that there are I **subjects, each with a subject-specific regression curves**. The curves may be similar. We want to shrink towards a common "population curve" but we also want to estimate differences between the curves.
- **Data:** $\{(X_{ij}, Y_{ij})\}_{j=1}^{J_i},\ i = 1, \ldots, I$
- **Model:** $Y_{ij} = m_i(X_{ij}) + \epsilon_{ij}$
  - m_i is the subject-specific regression curve.
  - $m_i(x) = \Sigma_{k=0}^{p}(\beta_k + b_{ik})\, x^k + \Sigma_{k=1}^{K}(u_k + w_{ik})\, (x - t_k)_+^p$
  - $\{u_k\}_{k=1}^{K}$ are IN$(0, \sigma_u^2)$
  - $\{b_{ik}\}$ is N$(0, \sigma_{b,k}^2)$.
  - $\{w_{ik}\}_{i=1\ k=1}^{I\ \ K}$ are distributed IN$(0, \sigma_w^2)$.
  - $m(x) = \Sigma_{k=0}^{p}\, \beta_k\, x^k + \Sigma_{k=1}^{K}\, u_k\, (x - t_k)_+^p$ is the population curve
  - In some applications one might set $\sigma_w^2 = 0$.
    - Also setting $\sigma_{b,k}^2 = 0$ for $k > 1$ (or even for $k > 0$) might be contemplated.
- **References:** Altman and Casella (1995, *JASA*), Brumback and Rice (1998, *JASA*), Wang (1998, *JRSS-B*).

Phenolphthalein Experiment

Analysis	$\widehat{\gamma}$	Std Err	p-value
Current experiment only			
1. Original model	-0.076	0.021	< 0.001
2. Include weight as spline	-0.042	**0.030**	0.156
With historical controls			
3. Random intercept, no weight term	-0.046	0.156	0.003
4. Proposed model	-0.023	**0.015**	0.144

Note: γ is the effect of dosage. Use of historical controls cuts the std err of $\hat{\gamma}$ in half.

All models were logistic mixed models and were fit using **`GLMMIX`** in **`SAS`**.

Response:

- Indicator of mammary tumors

Fixed effects:

- intercept
- dosage (linear)
- time of death (linear)
- quadratic in weight (analyses 2. and 4. only)

Random effects:

- experiment number (analyses 3. and 4. only)
- truncated quadratics in weight (analyses 2. and 4. only)

Summary

- **P-Splines, which include smoothing splines, can be formulated as linear mixed models.**
- **Smoothing parameters can be estimated by**
  - **ML**
  - or **REML**.
- **Nonparametric regression can be added to a graduate course on linear models without much difficulty.**
  - If the mixed model is already in the course, then little new material is required.
- **Nonparametric modeling can be extended to non-Gaussian, e.g., binary, responses via the GLMM.**
- **Extension to nested curves and even more general models is possible.**
  - MLE provides an effective framework for estimating the smoothing parameters as ratios of variances.
- **The ideas presented in this talk**
  - Were discovered by the smoothing spline community.
  - Have a long history.
  - Deserve much wider appreciation.