

# **Numerically Efficient Penalized Spline Smoothing in Complex Models - Taking up the Cudgels for the Good Old Laplace Approximation**

Göran Kauermann  
Universität Bielefeld

based on joint work with  
Tatyana Krivobokova, Ciprian Crainiceanu and Theo Archontakis

München 12. October 2006

# Outline of Talk

- P-Spline Smoothing in a nutshell
- P-Spline Smoothing and Linear Mixed Models
- Bivariate Smoothing - *Term Structure Modeling*
- P-Spline Smoothing with Generalized Response
- Locally Adaptive Smoothing
- Discussion

## P Splines in a nutshell

Simple smoothing model  $y = \mu(x) + \varepsilon$  is fitted by replacing

$$\begin{aligned}\mu(x) &= \beta_0 + x\beta_x + x^2\beta_{xx} + \sum_{j=1}^k u_j(x - \tau_j)_+^2 \\ &= \mathbf{X}(x)\boldsymbol{\beta} + \mathbf{B}(x)\mathbf{u} = \mathbf{C}(x)\boldsymbol{\theta}\end{aligned}$$

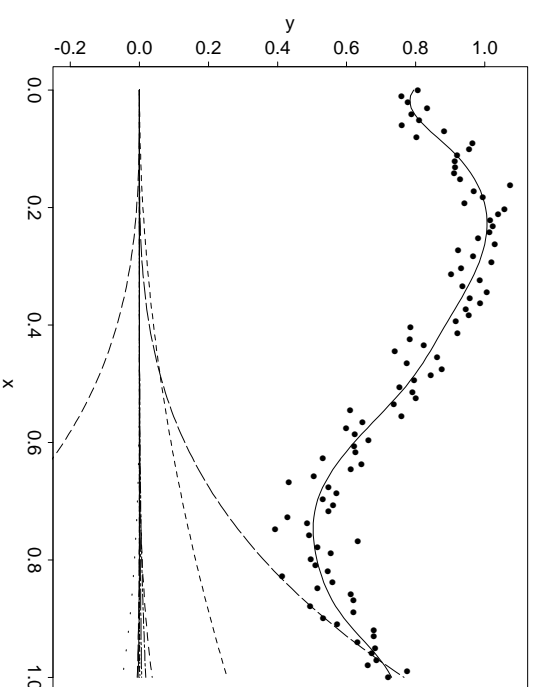
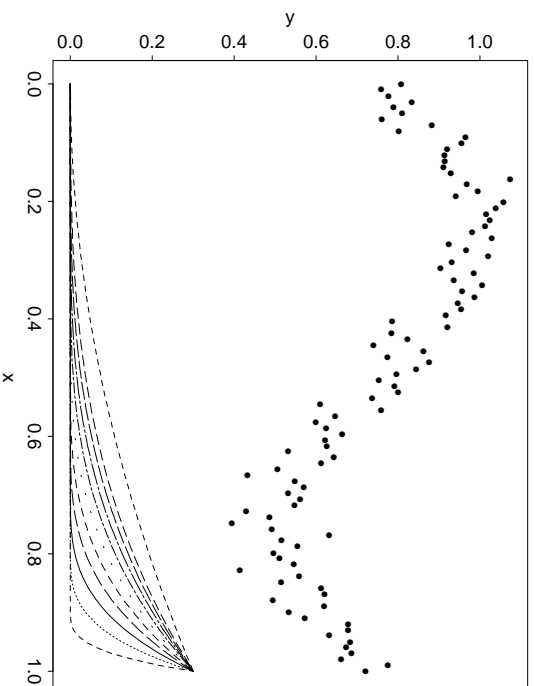
for knots  $\tau_1, \tau_2, \dots, \tau_k$  with  $k$  large. This yields the estimate

$$\hat{\boldsymbol{\mu}} = \mathbf{C}(\mathbf{C}^T\mathbf{C})^{-1}\mathbf{C}^T\mathbf{y}$$

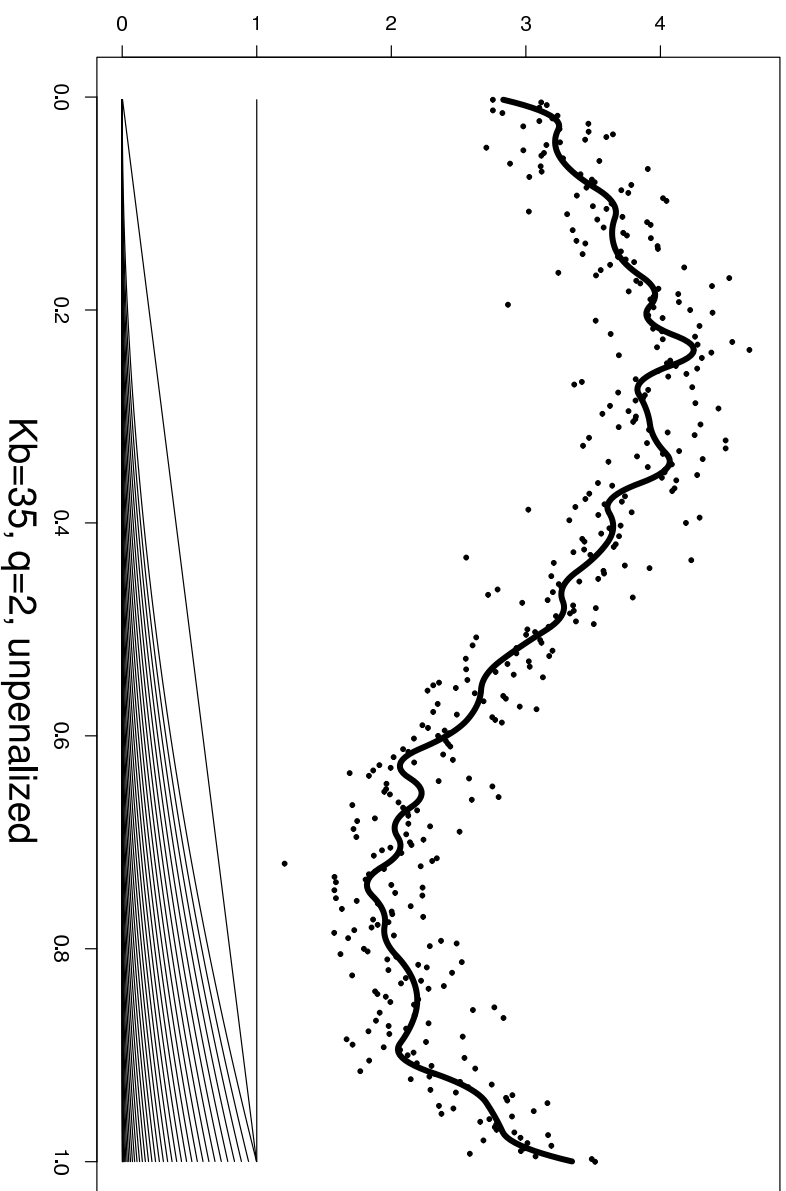
↑

$(k + q) \times (q + k)$ , with  $k$  “large, but not too large”

# Truncated Polynomial



# Need for Penalization



# P-Splines

O'Sullivan, 1986, Eilers & Marx, 1996, Ruppert, Wand & Carroll 2003.

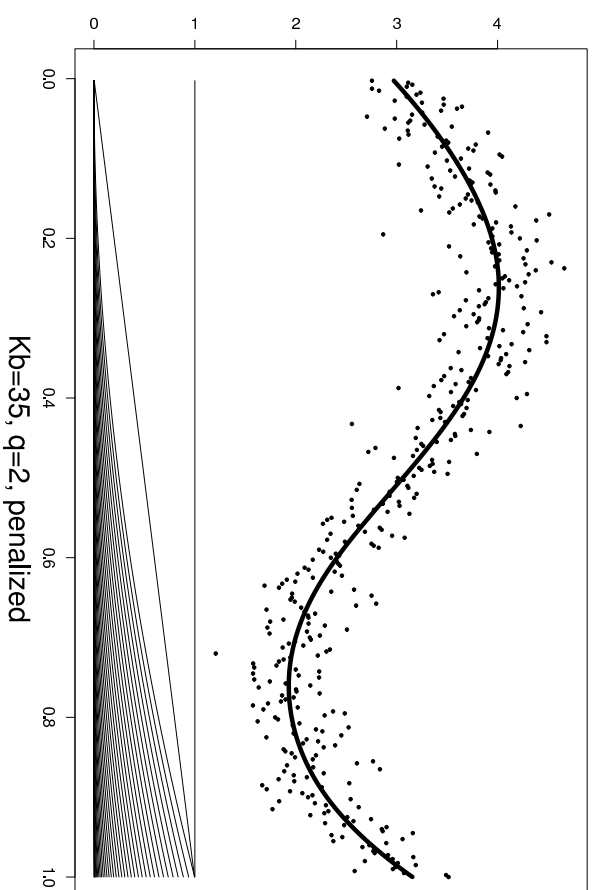
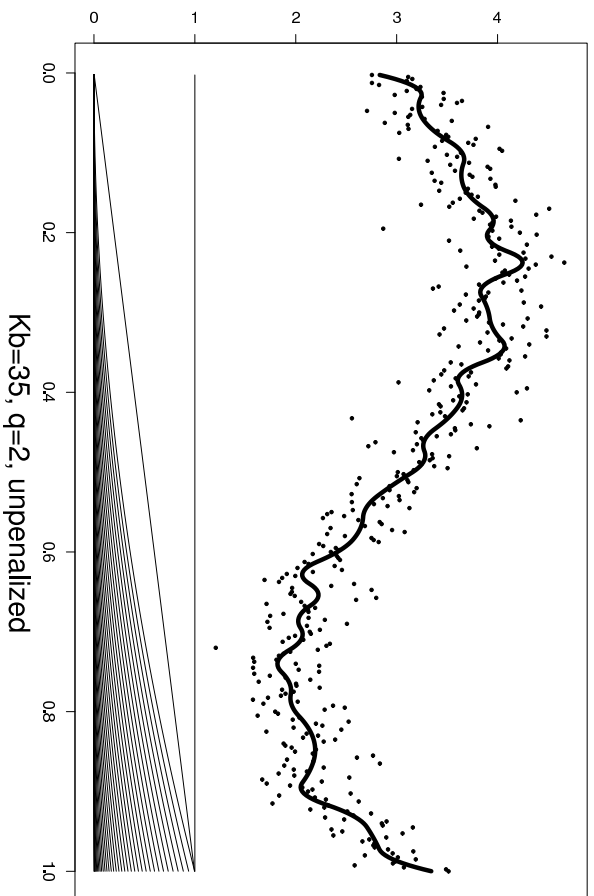
Recipe: take a rich dimensional Basis  $B(x)$ , that is take  $k$  large.

Minimized the Penalized least square

$$(\mathbf{y} - \mathbf{C}\boldsymbol{\theta})^T (\mathbf{y} - \mathbf{B}\boldsymbol{\theta}) + \lambda \boldsymbol{\theta}^T \mathbf{D}\boldsymbol{\theta} \rightarrow \min$$

where  $\mathbf{D}$  is a penalty matrix chosen adequately.

# Unpenalized versus Penalized Spline



## Reformulation

We can always decompose the spline basis to

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{B}\mathbf{u} + \boldsymbol{\epsilon}$$

- $\mathbf{X}\boldsymbol{\beta}$  unpenalized parametric part,
- $\mathbf{B}\mathbf{u}$  penalized part.

Penalized least square

$$(\mathbf{y} - \mathbf{X}\boldsymbol{\beta} - \mathbf{B}\mathbf{u})^T (\mathbf{y} - \mathbf{X}\boldsymbol{\beta} - \mathbf{B}\mathbf{u}) + \lambda \mathbf{u}^T \mathbf{D}\mathbf{u} \rightarrow \min_{\mathbf{u}, \boldsymbol{\beta}}$$

## P-Spline fitting and Linear Mixed Models

P-Spline estimation is **equivalent** to posterior Bayes estimation in the following **Linear Mixed Model**

$$Y|u \sim N(\mathbf{X}\beta + \mathbf{B}u, \sigma_\epsilon^2 \mathbf{I}) \quad \text{and} \quad u \sim N(\mathbf{0}, \sigma_u^2 \mathbf{D}^{-1})$$

The Posterior Bayes estimate (or the BLUP) is

$$\begin{aligned} \check{u} &= (\mathbf{B}^T \mathbf{B} + \lambda \mathbf{D})^{-1} \mathbf{B}^T (\mathbf{y} - \mathbf{X} \hat{\beta}) \\ \hat{\beta} &= (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T (\mathbf{y} - \mathbf{B} \check{u}) \end{aligned}$$

which is **equivalent** to P-Spline smoothing, where  $\lambda = \sigma_\epsilon^2 / \sigma_u^2$ .

## Why linking P-Splines with Mixed Models

In the Mixed Model the smoothing parameter  $\lambda$  plays the role of the a priori variance and hence allows for maximum likelihood based estimation.

Some of the available results are:

- Comparison of ML estimates  $\hat{\lambda} = \hat{\sigma}_\epsilon^2 / \hat{\sigma}_u^2$  with cross validation based choices (Kauermann, 2005; Ruppert et al, 2003).
- Tests on smooth effects can be linked to tests on positive variances of random effects (Crainiceanu et al. 2005)
- Model selection based on the Mixed Model (Kauermann, Ormerod, Wand, 2006)

# Term Structure Modeling - An Example for Bivariate Smoothing

Daily prices  $P_{t,m}$  of a zero coupon bond at time  $t \in \{07/98 - 07/03\}$  with maturity  $m \in \{08/03 - 07/33\}$  (US Treasury STRIPS)

We have 126251 observations from 107 bonds.

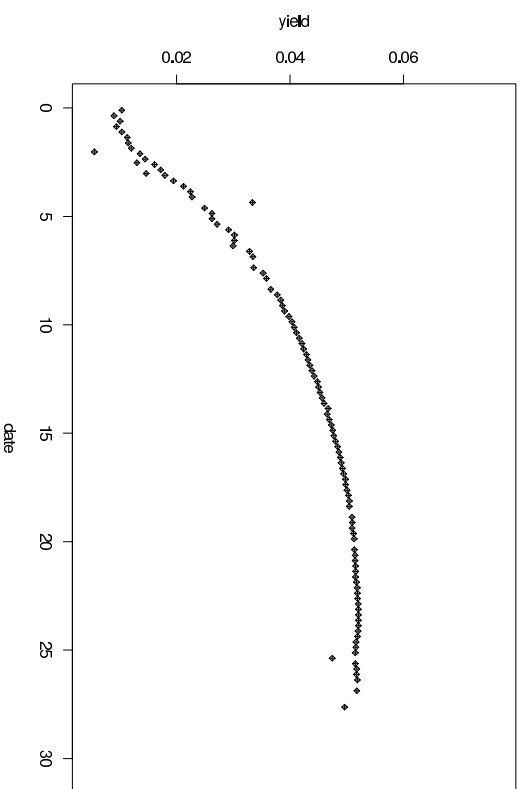
We model

$$y_{t,m} = -\log(P_{t,m})/m = \mu(t, m) + \epsilon_{t,m}$$

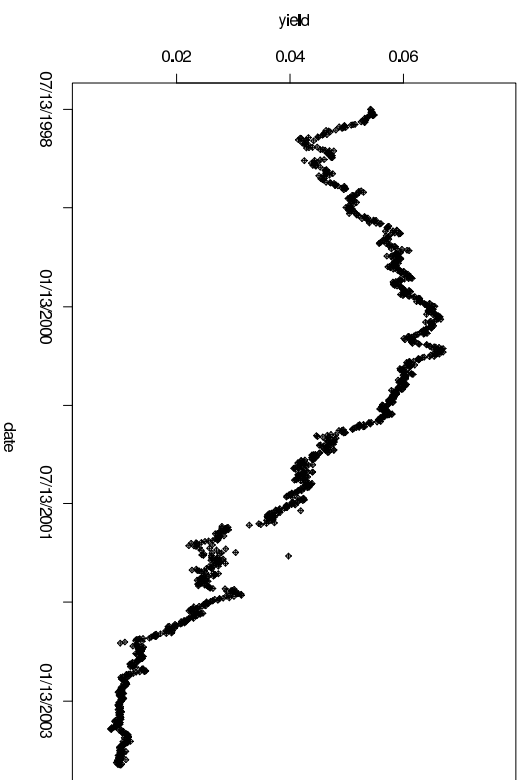
Note: Observations have a non-standard correlation structure (not focused in this presentation): correlation along single bonds but no correlation along time left to maturity.

# Term Structure

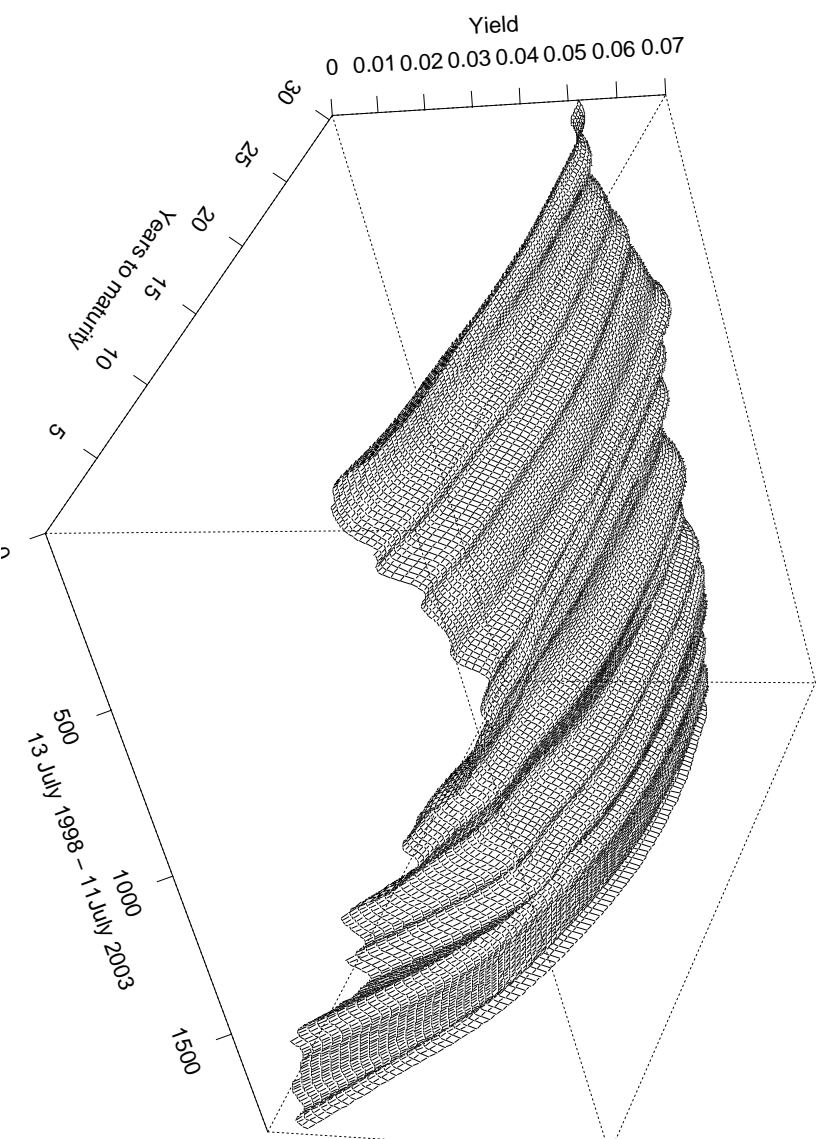
July 2003



6 year bond



# Bivariate Term Structure



## Generalized Smoothing Models

We extend the model to accommodate non-normal response. We assume

$$E(y|x) = h\{\eta(x)\}$$

with  $\eta(\cdot)$  as smooth function,  $h(\cdot)$  as link function and  $y \sim \exp\{y\vartheta - b(\vartheta)\}$ .

This is fitted by linking P-Splines to Generalized Linear Mixed Models:

$$\begin{aligned} E(y|x, u) &= h\{\mathbf{X}\boldsymbol{\beta} + \mathbf{B}u\} \\ u &\sim N(\mathbf{0}, \sigma_u^2 \mathbf{D}^{-1}). \end{aligned}$$

Note: The marginal likelihood  $\int f(\mathbf{y}|u)\phi(u)du$  is not analytical

## Calculating the Marginal Likelihood

- Laplace Approximation or similarly Penalized Quasi Likelihood (PQL) (Breslow & Clayton, 1993)
  - ⇒ fails asymptotically in “classical” GLMMs.
- Monte Carlo EM (Booth & Hobert, 1999)
  - ⇒ computer intensive
- Bayesian Modeling using MCMC
  - ⇒ even more computer intensive

## The Good Old Laplace Approximation

As simple way to approximate the marginal likelihood is using a Laplace approximation or a Penalized Quasi Likelihood (Breslow & Clayton, 1993), that is (with natural link)

$$l(\beta, \sigma_u^2) = -\frac{(k-1)}{2} \log(\sigma_u^{-2}) - \frac{1}{2} \log(|\mathbf{G}|) \\ + \mathbf{y}^T (\mathbf{X}\beta + \mathbf{Z}\hat{\mathbf{u}}) - 1_n^T b(\mathbf{X}\beta + \mathbf{Z}\hat{\mathbf{u}}) - \frac{k}{2} \hat{\mathbf{u}}^T \hat{\mathbf{u}} / \sigma_u^2$$

with  $\hat{\mathbf{u}}$  as maximizer (*penalized estimate*) and  $\mathbf{G} = \partial^2 l(\beta, u) / \partial \mathbf{u} \partial \mathbf{u}^T$ .

Note: Penalized Spline estimates are equivalent to Laplace estimates

## The dispute about Laplace approximation

It has been shown in Breslow & Lin (1995) and Shun & McCullagh (1995) that the Laplace approximation fails in standard GLMMs.

Asymptotic scenarios

- Classical Generalized Linear Mixed Model  
Number of independent individuals grows, while number of replicates per individual is limited.  
⇒ Laplace fails asymptotically
- P-Spline Model Scenario  
The number of (independent) splines is limited, while the support for each spline grows  
⇒ Laplace is justified asymptotically if the number of splines is fixed

## A Deeper Asymptotic Insight

So far we assumed that  $k$ , the number of splines is fixed (and limited).

⇒ This is not realistic.

We assume now that  $k$  depends on sample size  $n$ , that is

- $n \Rightarrow \infty$
- $k \Rightarrow \infty$ ,
- $|\tau_i - \tau_j| \Rightarrow 0, |x_i - x_j| \Rightarrow 0$

## Laplace Approximation

Question: If we consider spline coefficient  $u$  as random, but let its dimension grow, is the Laplace approximation still valid?

Taking  $\sigma_u^2 = 1/\lambda = O(n^{-\frac{1}{2q+3}})$  we can show that

$$\begin{aligned} l(\beta, \sigma_u^2) &= \int \exp\{l(\beta, u, \sigma_u^2)\} \phi(u, \frac{\sigma_u^2}{k}) du \\ &= l_{Laplace}(\beta, \sigma_u^2) \{1 + O(\varepsilon_0)\} \end{aligned}$$

where  $\varepsilon_0$  is the leading component in the Laplace approximation.

## Results

One finds

$$\varepsilon_0 = O \left\{ n^{-\frac{2q}{2q+3}} \left( 1 + n^{-\frac{1}{2q+3}} \sigma_u^{-2} \right)^{-3} \right\},$$

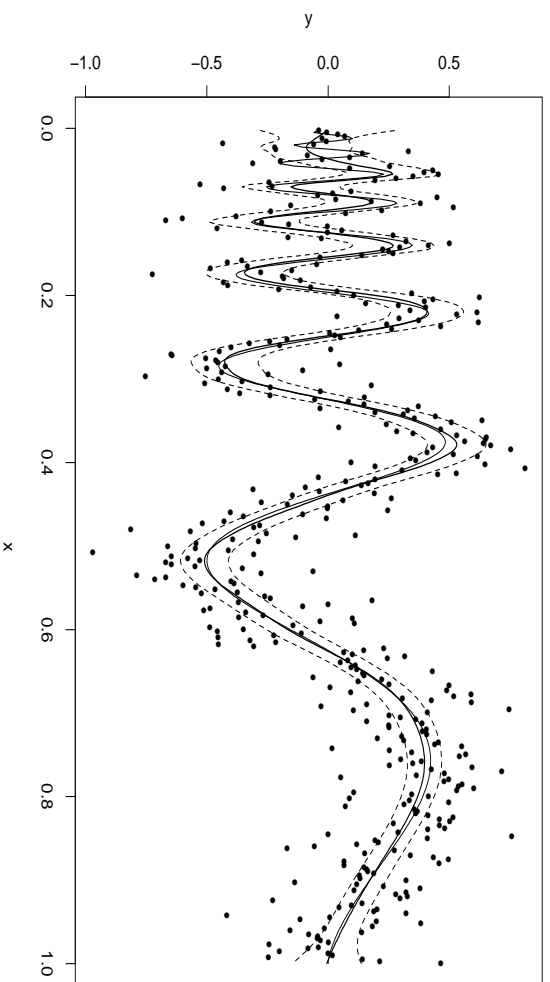
With  $k = O(n^{\frac{1}{2q+3}})$  and  $\sigma_u^2 = O(n^{-\frac{1}{2q+3}})$  we get

$$\varepsilon_0 = O \left\{ n^{-\frac{2q+1}{2q+3}} \right\},$$

**Good News:** Laplace approximation is justified asymptotically, even for growing spline dimension.

# Local Adaptive Smoothing

Taking advantage of the Laplace approximation once more.



# The Idea of Local Adaptive Smoothing

We assume that

$$y_i \sim N(\mu(x_i), \sigma_\epsilon^2), \quad i = 1, \dots, n,$$

where  $\mu(\cdot)$  is of locally varying complexity.

This means that fitting  $\mu(x)$  with a global smoothing parameter is not recommendable.

We therefore allow  $\lambda$  to depend on  $x$ , that is  $\lambda(x)$ .

In the Mixed Model scenario this means, we allow  $\sigma_u^2 = \sigma_u^2(x)$ .

# Hierarchical Modeling

We extend the Mixed Model to

$$Y|\mathbf{u} \sim N(\mathbf{X}\boldsymbol{\beta} + \mathbf{B}\mathbf{u}, \sigma_\epsilon^2 \mathbf{I}) \quad \text{and} \quad \mathbf{u} \sim N(\mathbf{0}, \sigma_u(\mathbf{x})^2 \mathbf{D}^{-1})$$

where  $\sigma_u^2(\mathbf{x})$  is a smooth function in  $x$ .

The idea is now to estimate  $\sigma_u^2(x)$  by penalized splines, that is we assume

$$\sigma_u^2(x) = \gamma_0 + x\gamma_p + \dots + x^p\gamma_p + \sum_{j=1}^{k_c} (x - \tau_j^{(c)})_+^p c_j$$

where  $\tau_j^{(c)}$  are an additional layer of knots covering the range of  $x$ .

For estimation we impose a penalty of coefficients  $c_j$ .

## Hierarchical Mixed Model

Formulating the penalty on coefficient  $\mathbf{c}$  as a *priori* normality leads to the Hierarchical Mixed Model

$$\mathbf{y} | \mathbf{u}, \mathbf{c} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{u} + \boldsymbol{\epsilon}, \quad \boldsymbol{\epsilon} \sim N(0, \sigma_{\boldsymbol{\epsilon}}^2 \mathbf{I}_n),$$

$$\mathbf{u} | \mathbf{c} \sim N(0, \Sigma_{\mathbf{u}}), \quad \Sigma_{\mathbf{u}} = \text{diag}[\exp(\mathbf{X}_c \boldsymbol{\gamma} + \mathbf{Z}_c \mathbf{c})],$$

$$\mathbf{c} \sim N(0, \sigma_c^2 \mathbf{I}_{K_c}).$$

See also Crainiceanu, Ruppert & Carroll (2005)

## Marginal Likelihood

The marginal likelihood is not analytic and results to

$$\begin{aligned} L(\boldsymbol{\beta}, \boldsymbol{\gamma}, \sigma_\epsilon^2, \sigma_c^2) &= f(\mathbf{y}; \boldsymbol{\beta}, \boldsymbol{\gamma}, \sigma_\epsilon^2, \sigma_c^2) \\ &= (2\pi)^{-\frac{(n+k_c)}{2}} \sigma_\epsilon^{-n} \sigma_c^{-k_c} \int_{R^{k_c}} \exp[-g(\mathbf{c})] d\mathbf{c}, \end{aligned}$$

with

$$g(\mathbf{c}) = \frac{1}{2} \log |V_\epsilon| + \frac{\mathbf{c}^T \mathbf{c}}{2\sigma_c^2} + \frac{(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^T V_\epsilon^{-1} (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})}{2\sigma_c^2}$$

and  $V_\epsilon = I_n + \mathbf{Z}\Sigma_u\mathbf{Z}^T/\sigma_\epsilon^2$ .

## Laplace Approximation in this Setting ?

Why should we apply Laplace approximation?

- MCMC methods are time consuming
- The hierarchical stochastic structure assumed for  $\sigma_u^2(x)$  is a working model only anyway. Hence, why using exact solutions for an approximate model?
- We will see that Laplace is fast and satisfactory.

## Laplace Approximation in this Setting !

The marginal likelihood can be approximated by

$$\begin{aligned} -2l(\boldsymbol{\beta}, \gamma, \sigma_\epsilon^2, \sigma_c^2) &\approx n \log \sigma_\epsilon^2 + k_c \log \sigma_c^2 + \log |V_\epsilon(\hat{\mathbf{c}})| + \log |I_{cc}(\hat{\mathbf{c}})| \\ &+ \hat{\mathbf{c}}^T \hat{\mathbf{c}} / \sigma_c^2 + (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^T V_\epsilon^{-1}(\hat{\mathbf{c}}) (\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) / \sigma_\epsilon^2, \end{aligned}$$

where  $\hat{\mathbf{c}}$  is the solution to

$$\frac{\partial g(\hat{\mathbf{c}})}{\partial c_i} = \frac{1}{2} \text{tr} \left( V_\epsilon^{-1} \frac{\partial V_\epsilon}{\partial c_i} \right) + \frac{c_i}{\sigma_c^2} - \frac{1}{2\sigma_c^2} (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^T V_\epsilon^{-1} \frac{\partial V_\epsilon}{\partial c_i} V_\epsilon^{-1} (\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) = 0$$

and

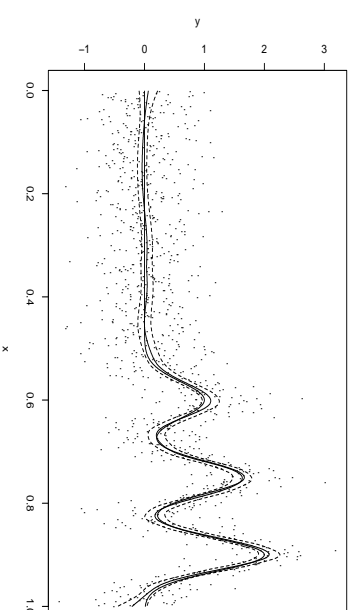
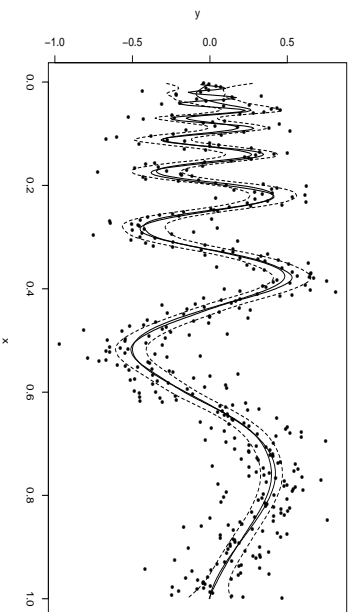
$$(I_{cc}(\mathbf{c}))_{ij} = \frac{\delta_{ij}}{\sigma_c^2} + \frac{1}{2} \text{tr} \left( V_\epsilon^{-1} \frac{\partial V_\epsilon}{\partial c_i} V_\epsilon^{-1} \frac{\partial V_\epsilon}{\partial c_j} \right)$$

## Practical implementation

Estimates for  $u$  and  $c$  are easily available by using a Backfitting strategy.

- Taking  $\sigma_u^2(x)$  as fix we get an estimate for  $u$  with simple penalized least squares.
- Taking  $\hat{u}$  as fixed, we get an estimate for  $\hat{c}$  by penalized weighted least square
- Taking  $\hat{u}$  and  $\hat{c}$  as given we get an update for the remaining (hyper) parameters.
- We iterate until convergence.

# Simulation



---

3.3e-3	<i>Krivobokova et al (2006)</i>	4.7e-3
2.7e-3	Ruppert et. al. (2000)	6.1e-3
2.6e-3	Baladandayuthapani et al, (2005)	6.5e-3

---

## **Extensions**

The idea can be extended to

- Generalized Response Models
- Spatial Smoothing
- or both

## Data Example

Absenteeism of workers in a German company. We consider data of

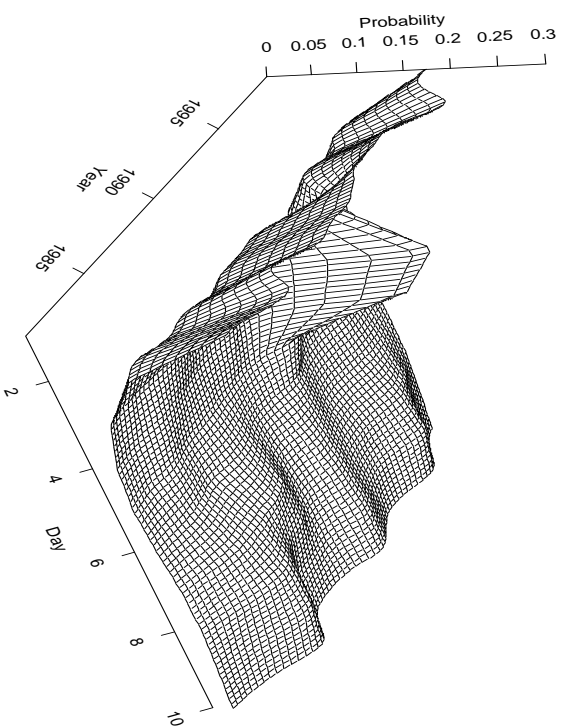
- 370 employees
- for calendar time  $c \in \{1981 - 1998\}$
- with about 27.000 absenteeism spells

We focus on the duration of absenteeism  $d$  (in days) using a discrete hazard model

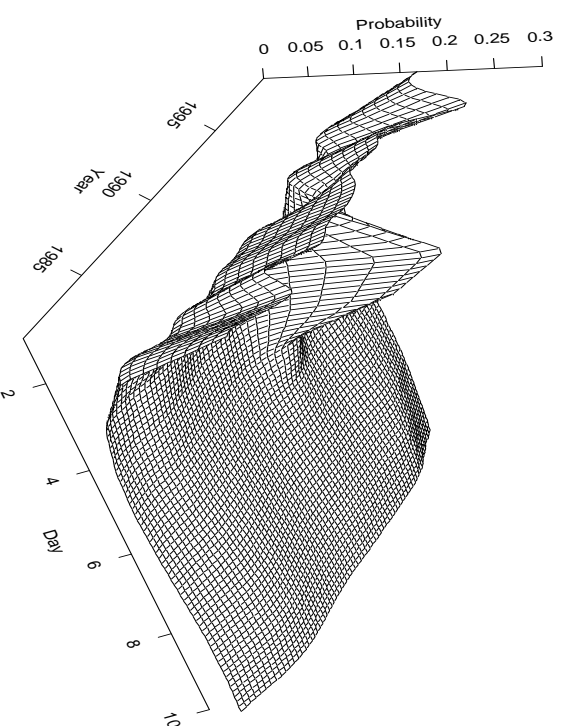
$$\text{logit}\{P(d = t | d \geq t, c)\} = \mu(t, c)$$

# Example (cont)

Non Adaptive



Adaptive



## **AdaptFit - Fitting in Practice**

The R package “SemiPar” written by Matt Wand has been extended to accommodate locally and spatially adaptive smoothing. the routines build the new R package AdaptFit available from the CRAN server.

The routines include the calculation of confidence intervals (not shown in this talk).

## Discussion

- P-spline smoothing is a powerful and numerically handy smoothing method.
- The link to Generalized Linear Mixed Models is worthwhile.
- Laplace approximation is asymptotically and practically justified.
- Laplace approximation works in more complex models as well.

Paper and Preprints available from [www.wiwi.uni-bielefeld.de/~statistics](http://www.wiwi.uni-bielefeld.de/~statistics)