Nonparametric Methods for Longitudinal Data Using Regression Splines

Jin-Ting Zhang

National University of Singapore

NUS Summer School
June 17, 2009

stazjt@nus.edu.sg

OUTLINE

- Part I: Regression Spline Smoothing
- Part II: Regression Spline Smoothing for Longitudinal Data
- Summary

Regression Spline Smoothing

OUTLINE for Part I

- Review of Parametric Regression Modelling
- Nonparametric Regression Model
- Nonparametric Smoothing Methods
- Regression Spline Modelling

Regression Spline Smoothing

Review of parametric modelling

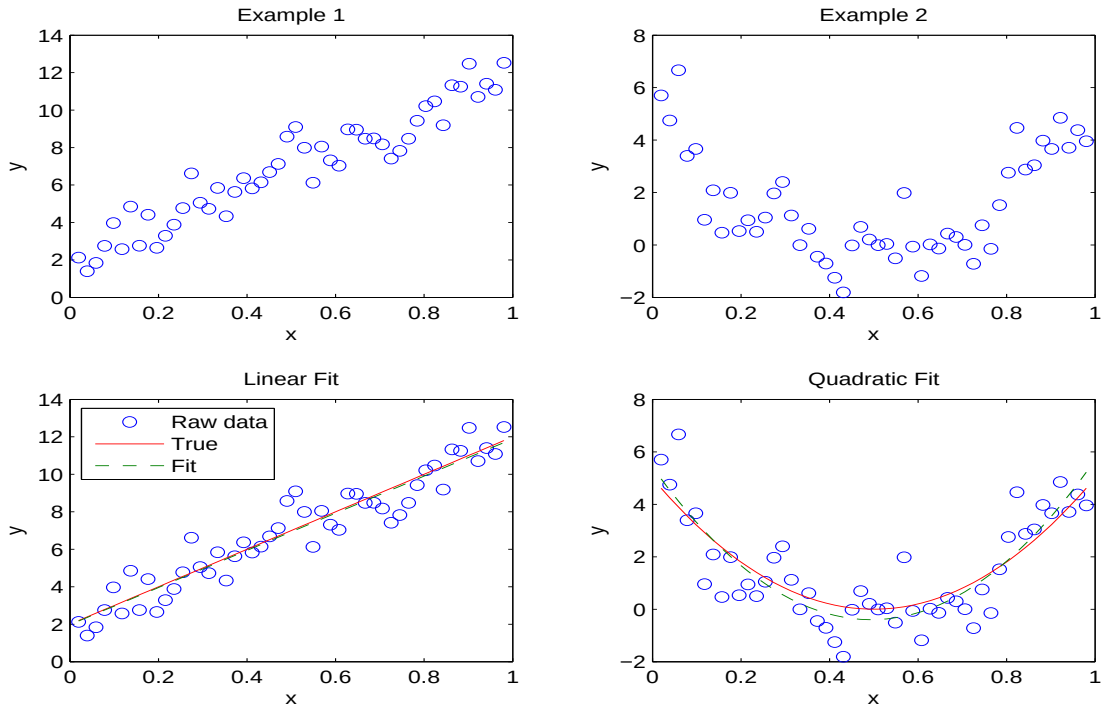


Figure 1: Examples for parametric regression

Regression Spline Smoothing

Polynomial Regression

- **Model** $y_i = \mathbf{\Phi}_0(x_i)^T \boldsymbol{\beta} + \epsilon_i, i = 1, 2, \cdots, n.$
- **Polynomial basis vector** $\mathbf{\Phi}_0(x) = [1, x, x^2, ..., x^k]^T$
- **Matrix form** $\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}.$
- **Easy to fit** $\hat{\boldsymbol{\beta}} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{y}.$

Regression Spline Smoothing

Advantages for Parametric Regression Modelling

- Easy to fit
- Methods for **estimation**, **hypothesis testing** and **prediction** well established

Drawbacks for Parametric Regression Modelling

- Parametric models applied, e.g., linear or quadratic, must be valid
- Invalid parametric models may lead to misleading results

Regression Spline Smoothing

The Motivating Data

The motorcycle data (Silverman 1985)

- Collected to study the crashed effects after the motorcycles hit by a stimulated impact
- **Dependent variable**: time after a stimulated impact with motorcycles
- **Response variable**: head acceleration of a PTMO (post mortem human test object), capturing the crash effects

Regression Spline Smoothing

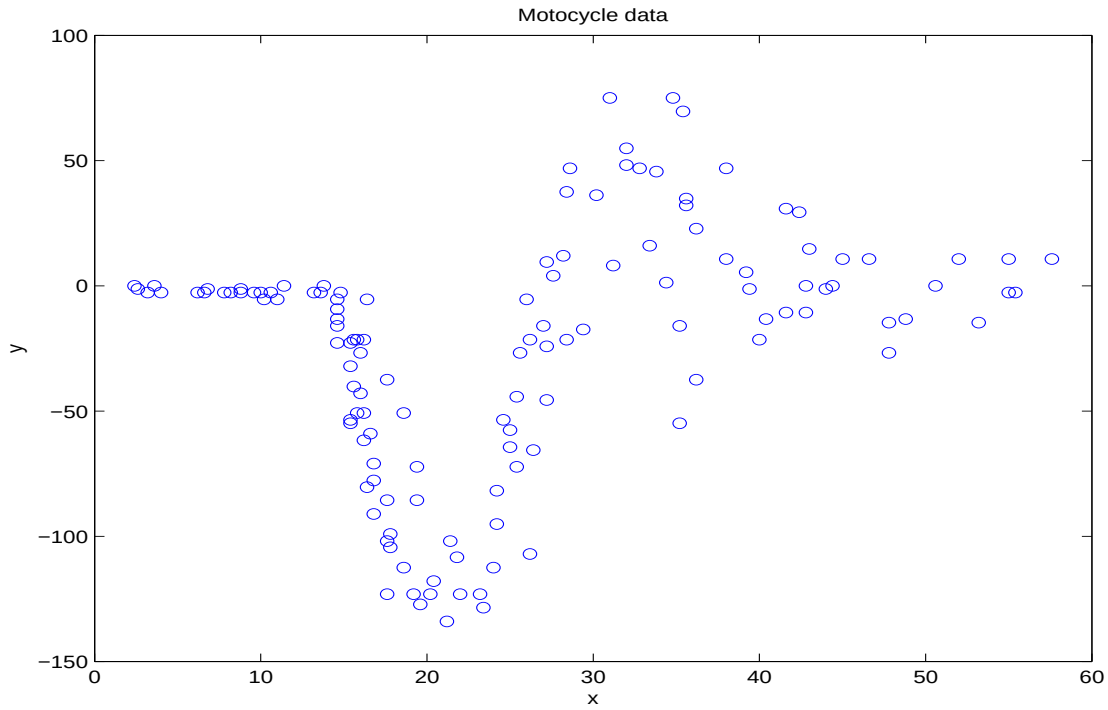


Figure 2: The motorcycle data

Regression Spline Smoothing

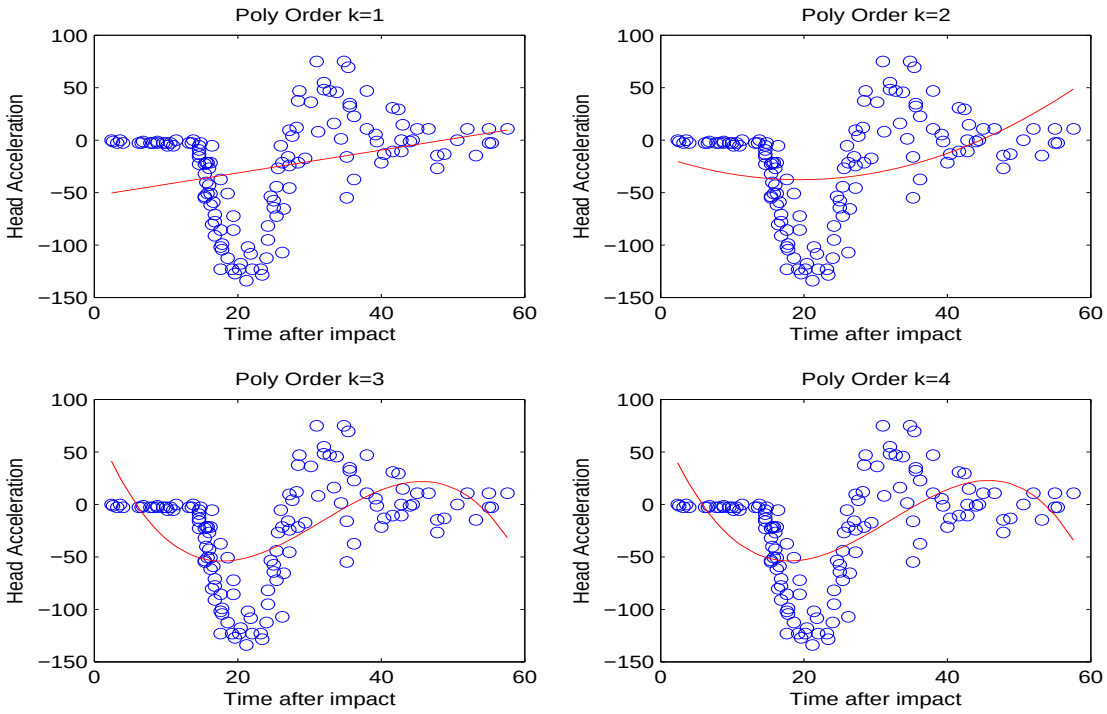


Figure 3: Polynomial fits for the motorcycle data

Regression Spline Smoothing

Nonparametric Regression Model

For a data set (x_i, y_i), $\quad i = 1, 2, \cdots, n,$

$$y_i = f(x_i) + \epsilon_i, \qquad i = 1, .., n,$$

- $f(\cdot)$ unknown but smooth, target for estimation
- $f(x) = \mathrm{E}(y_i | x_i = x)$, conditional expectation of the response
- $\mathrm{E}(\epsilon_i) = 0$ and $\mathrm{Var}(\epsilon_i) = \sigma^2$
- It reduces to a parametric model if $f(\cdot)$ is known except a parameter

Regression Spline Smoothing

Nonparametric Smoothing Methods

- **Local Polynomial Kernel Smoothing** (Wand and Jones 1995, Fan & Gijbels 1996)
- **Regression Splines** (Eubank 1988)
- **Smoothing Splines** (Wahba 1990, Green and Silverman 1994)
- **P-splines** (Ruppert, Wand and Carroll, 2003)

Regression Spline Smoothing

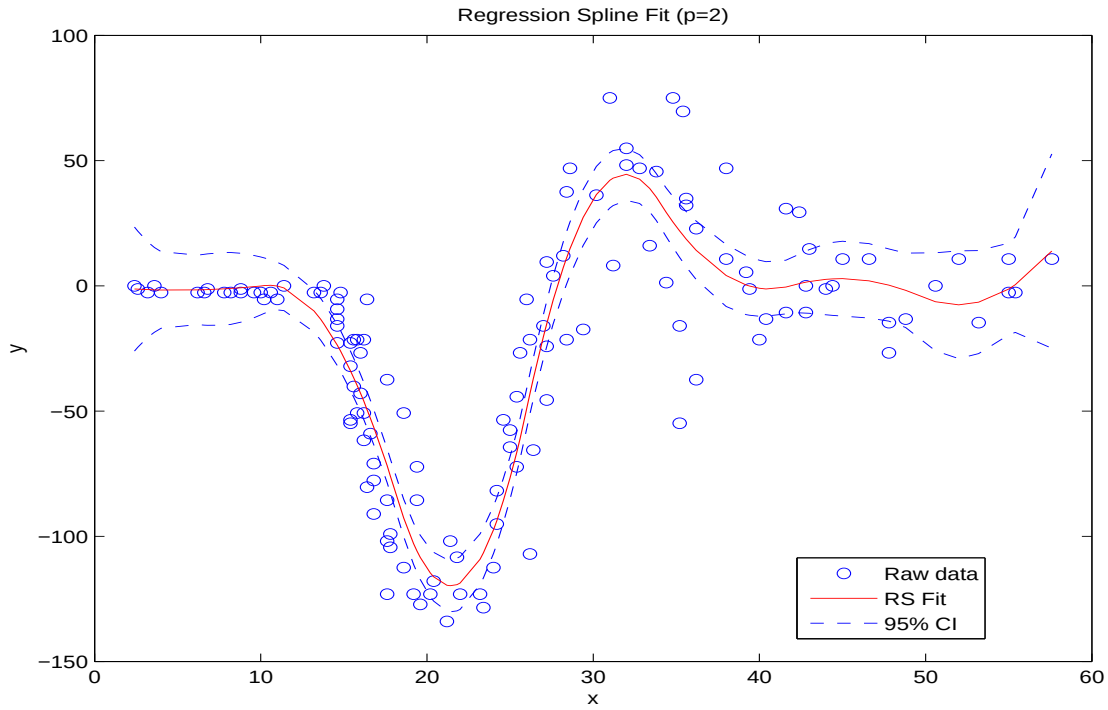


Figure 4: Regression spline fit for the motorcycle data

Regression Spline Smoothing

Regression Splines

- The simplest smoothing technique available
- A natural generalization of the polynomial regression
- Easy to understand and implement
- Widely used in data analysis

Regression Spline Smoothing

Motivation

- Polynomials are not flexible to model data in a big range
- They work well with a small range since within a small range, a Taylor's expansion up to some order is valid
- One can divide a big range, say $[a, b]$ into a few of small intervals:

$$[\tau_r, \tau_{r+1}), r = 0, 1, \cdots, K,$$

where **interior knots**: $\tau_r, r = 1, 2, \cdots, K$, **boundary knots**: a, b

$$a = \tau_0 < \tau_1 < \tau_2 < \cdots < \tau_K < \tau_{K+1} = b.$$

Regression Spline Smoothing

Definition

- **A regression spline** is a piecewise polynomial
- It is a polynomial of some order within any two neighboring knots τ_r and τ_{r+1} for $r = 0, 1, \cdots, K$
- The spline is jointed together at knots properly
- The spline allows discontinuous derivatives at the knots.

Regression Spline Smoothing

Truncated power basis (TPB)

- A regression spline can be constructed using a k-order TPB
- Given K interior knots $\tau_1, \tau_2, \cdots, \tau_K$, the k-order TPB is

$$1, x, \cdots, x^k, (x-\tau_1)_+^k, \cdots, (x-\tau_K)_+^k.$$

- The truncated power function $w_+^k = [w_+]^k$, $w_+ = \max(0, w)$.
- First $k+1$ basis functions are polynomials of order up to k
- Last K basis functions are truncated power basis functions of order k

Regression Spline Smoothing

Regressin splines using TPB:

$$f(x) = \sum_{s=0}^{k} \beta_s x^s + \sum_{r=1}^{K} \beta_{k+r} (x - \tau_r)_+^k.$$

- Within $[\tau_r, \tau_{r+1})$, $f(x)$ is a k-order polynomial:

$$f(x) = \sum_{s=0}^{k} \beta_s x^s + \sum_{l=1}^{r} \beta_{k+l} (x - \tau_l)^k,$$

- $f^{(k)}(x)$ jumps at τ_r with amount $\beta_{k+r} k!$, i.e.,

$$f^{(k)}(\tau_r+) - f^{(k)}(\tau_r-) = \beta_{k+r} k!.$$

Regression Spline Smoothing

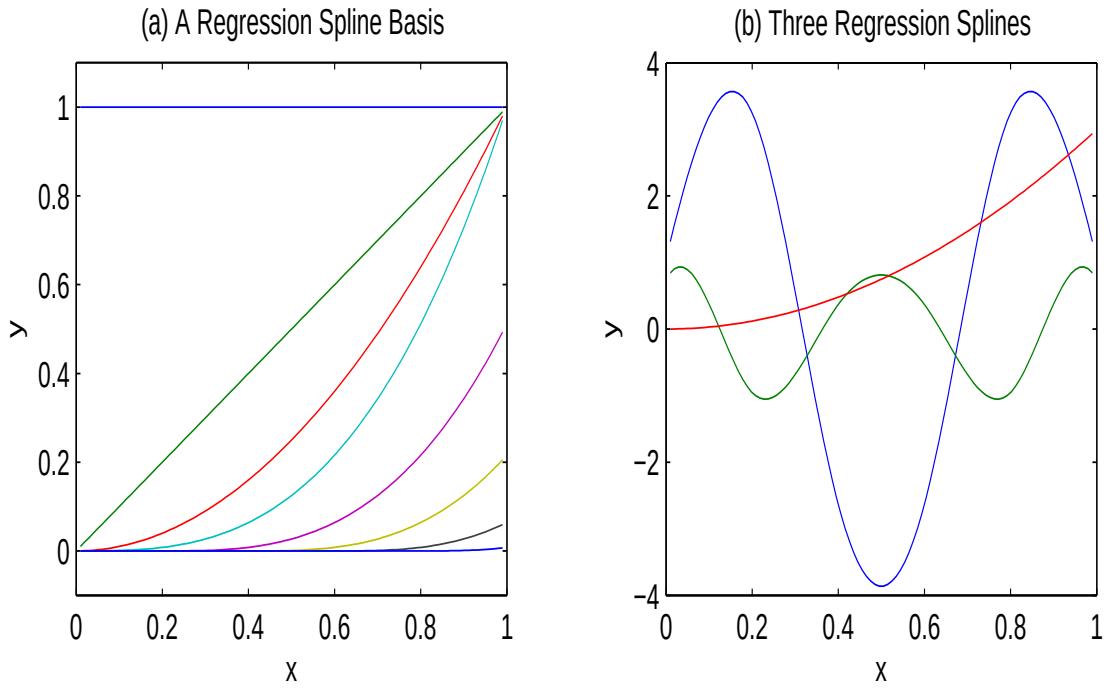


Figure 5: *(a) a cubic truncated power basis ($k = 3$) with interior knots .2, .4, .6, and .8; and (b) three cubic regression splines.*

Regression Spline Smoothing

Regression Spline Model

- Denote as the TPB as
  $\boldsymbol{\Phi}(x) = (1, x, \cdots, x^k, (x-\tau_1)_+^k, \cdots, (x-\tau_K)_+^k)^T$
- A regression spline can be written as $f(x) = \boldsymbol{\Phi}(x)^T \boldsymbol{\beta}$
- The model $y_i = f(x_i) + \epsilon_i, i = 1, \cdots, n$ becomes

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon},$$

  where $\boldsymbol{X} = (\boldsymbol{\Phi}(x_1), \cdots, \boldsymbol{\Phi}(x_n))^T$.

Regression Spline Smoothing

Regression Spline Smoother

- The LS estimator $\hat{\boldsymbol{\beta}} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{y}$
- The RS smoother $\hat{f}(x) = \boldsymbol{\Phi}(x)^T\hat{\boldsymbol{\beta}}$
- The fitted response vector $\hat{\mathbf{y}} = \boldsymbol{A}\boldsymbol{y}$ with the RS smoother matrix $\boldsymbol{A} = \boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T$
- $\boldsymbol{A}$ is a projection matrix satisfying $\boldsymbol{A}^T = \boldsymbol{A}, \boldsymbol{A}^2 = \boldsymbol{A}$ and $\mathrm{tr}(\boldsymbol{A}) = K + k + 1$. The trace of $\boldsymbol{A}$ measures the RS model complexity

Regression Spline Smoothing

Remarks for Nonparametric Regression

- For parametric regression, the model is fixed
- For nonparametric regression, the model is data-driven
- For the regression spline smoother, the model is specified by the TPB $\mathbf{\Phi}(x)$
- The TPB $\mathbf{\Phi}(x)$ is specified by the knot locations $\tau_1, \cdots, \tau_K$ and the number of knots, K
- Knot locating is important but the choice of the knot number is more important

Regression Spline Smoothing

Two Widely Used Knot Locating Methods

Equally Spaced Method Take K equally spaced points in the range of interest, say, $[a, b]$, as knots:

$$\tau_r = a + (b - a)r/(K + 1), r = 1, 2, \cdots, K.$$

- The method is independent of the design time points
- Employed when design time points are uniformly scattered

Regression Spline Smoothing

Equally Spaced Sample Quantiles Method Use equally spaced sample quantiles of the design time points $x_i,\ i = 1, 2, \cdots, n$ as knots:

$$\tau_r = x_{(1+[rn/(K+1)])}, r = 1, 2, \cdots, K,$$

- $x_{(1)}, \cdots, x_{(n)}$ the order statistics of the design time points.
- $[a]$ denotes the integer part of a.
- The method is design adaptive.
- More knots located where more design time points scattered.

Regression Spline Smoothing

Knot Number Selection

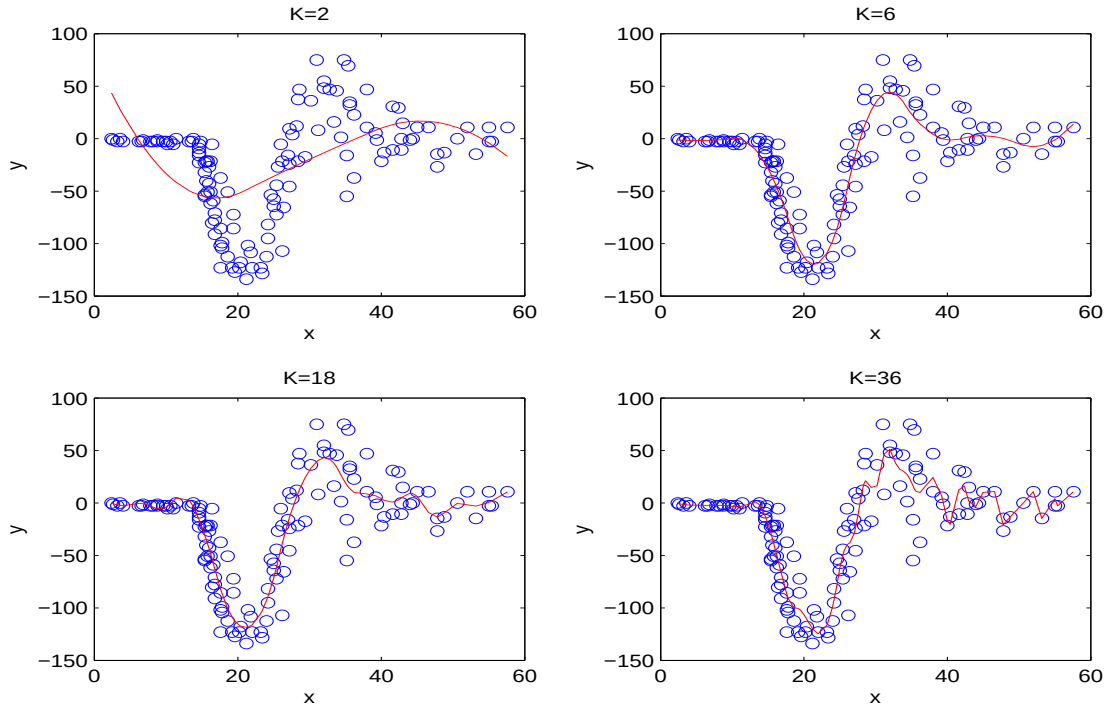


Figure 6: *RS fits to the motocycle data with different K.*

Regression Spline Smoothing

A Criterion for Knot Number Selection

Generalized Cross-Validation(GCV)

- This method was motified from cross-validation (CV) method
- Use SSE $= \boldsymbol{y}^T(\boldsymbol{I}_n - \boldsymbol{P}_X)\boldsymbol{y}$ to measure goodness of fit
- Use $\text{tr}(A) = K + k + 1$ to measure model complexity
- GCV $= \text{SSE}/(1 - \text{tr}(A)/n)^2$ trades off between goodness of fit and model complexity
- GCV often works well in choosing a good K

Regression Spline Smoothing

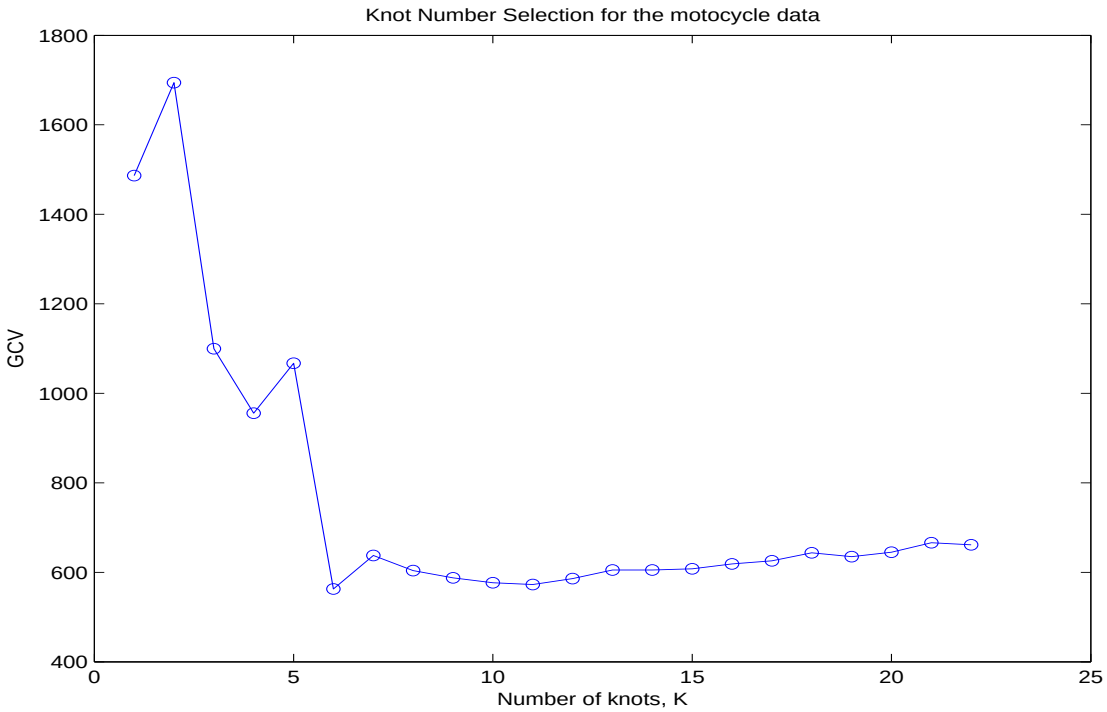


Figure 7: *GCV for knot number selection for the motocycle data.*

Regression Spline Smoothing

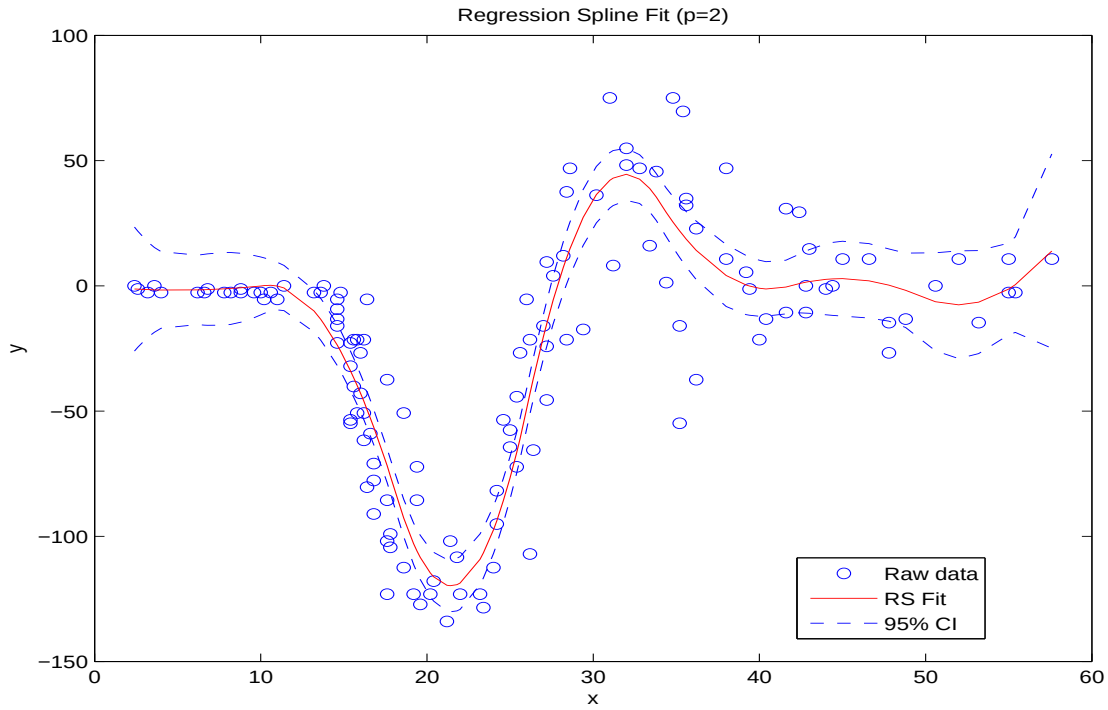


Figure 8: Regression spline fit for the motorcycle data

Regression Spline Smoothing for Longitudinal Data

OUTLINE for Part II

- Review of LME Modelling
- Motivating Longitudinal Data
- Nonparametric Mixed-effects (NPME) Model
- Fitting the NPME Model Using Regression Splines
- Extensions to other Nonparametric/Semiparametric ME Models

Regression Spline Smoothing for Longitudinal Data

Review of the LME Model:

$$\begin{aligned} y_{ij} &= x_{ij}^T\boldsymbol{\beta} + z_{ij}^T\boldsymbol{b}_i + \epsilon_{ij}, \quad j = 1, 2, \cdots, n_i;\ i = 1, 2, \cdots, n, \\ \boldsymbol{b}_i &\sim N(0, \boldsymbol{D}), \quad \boldsymbol{\epsilon}_i \sim N(0, \sigma^2 \boldsymbol{I}_{n_i}) \end{aligned}$$

- x_{ij}, z_{ij}: fixed-effects (FE) and random-effects(RE) covariates,
- $\boldsymbol{\beta}, \boldsymbol{b}_i$: FE and RE vectors, modeling population and individual features respectively
- ϵ_{ij}: measurements errors
- $\boldsymbol{D}$ and σ^2: variance components
- Mixed-effects modelling allows to pull information across subjects to estimate $\boldsymbol{\beta}$ and $\boldsymbol{b}_i$

Regression Spline Smoothing for Longitudinal Data

The LME Solution: Given the variance components $\boldsymbol{D}$ and $\boldsymbol{R}$,

$$\begin{aligned} \hat{\boldsymbol{\beta}} &= (\boldsymbol{X}^T\boldsymbol{V}^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{V}^{-1}\boldsymbol{y}, \\ \hat{\mathbf{b}} &= \boldsymbol{D}\boldsymbol{Z}^T\boldsymbol{V}^{-1}(\boldsymbol{y}-\boldsymbol{X}\hat{\boldsymbol{\beta}}), \end{aligned}$$

with $\boldsymbol{X}, \boldsymbol{Z}$ properly defined,

- $\hat{\boldsymbol{\beta}}$ is the generalized least squares estimator of $\boldsymbol{\beta}$
- $\boldsymbol{V}$ and $\boldsymbol{R}$ are estimated using EM-algorithm (Vonesh and Chinchilli 1996)
- Existing software for solving LME models, e.g., *lme* in Splus and *Proc Mixed* in SAS.

Regression Spline Smoothing for Longitudinal Data

Advantages for LME Modelling

- Easy to fit
- Information across subjects and within subjects are used
- Methods for fitting the LME models well established

Drawbacks for LME Modelling

- Strong assumption about the model form needed
- Invalid parametric models may lead to misleading results

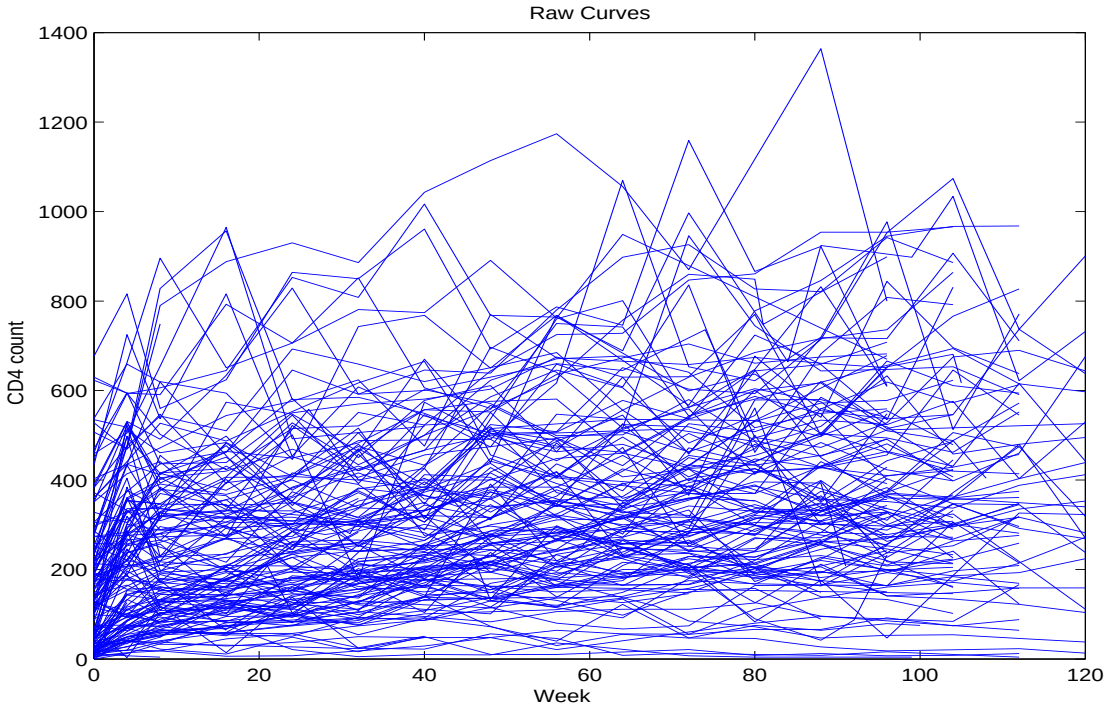
Regression Spline Smoothing for Longitudinal Data

Motivating Longitudinal Data(Park and Wu 2004)

- Collected in an AIDS clinical study conducted by the AIDS Clinical Trials Group (ACTG), called ACTG 388 data.
- The study randomized 517 HIV-1 infected patients in three antiviral treatments. The data from one of the three treatments used
- 166 patients treated with highly active antiviral therapy (HAART) for 120 weeks during which CD4 cell counts were monitored at weeks 4, 8, and every 8 weeks thereafter (up to 120 weeks)
- **Response variable**: CD4 cell counts as an important marker for assessing immunologic response of an antiviral regimen.
- **Covariate**: time after antiviral treatments

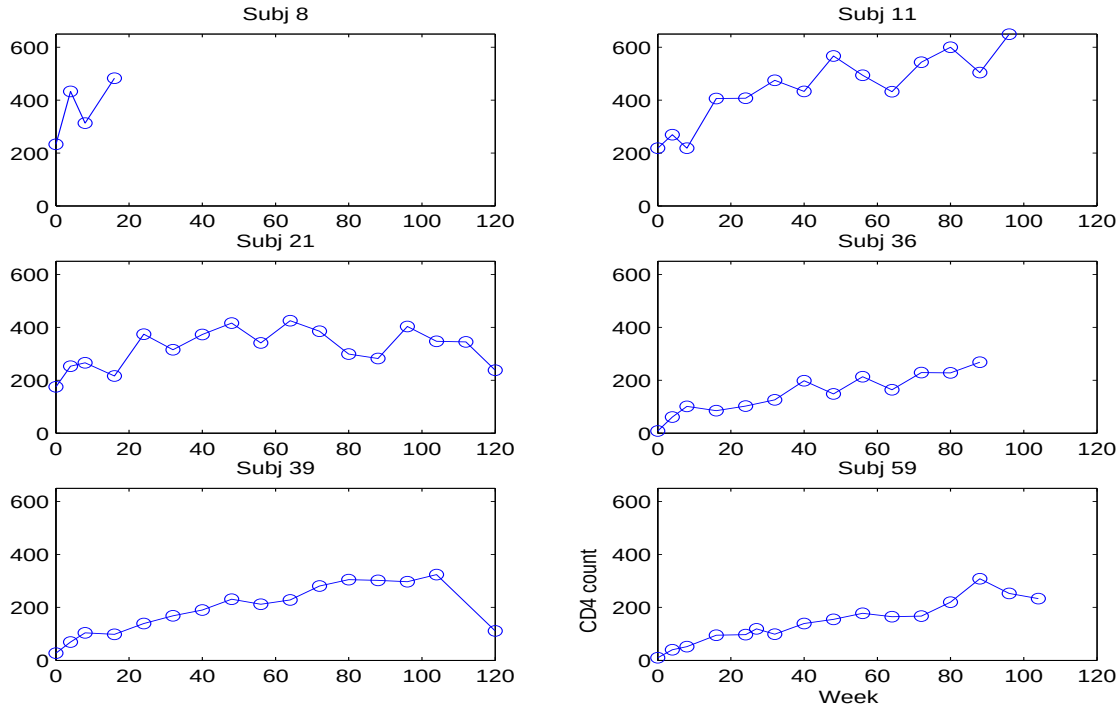
Regression Spline Smoothing for Longitudinal Data

The ACTG388 Data



Regression Spline Smoothing for Longitudinal Data

Six Selected Subjects



Regression Spline Smoothing for Longitudinal Data

Nonparametric Mixed-effects (NPME) Model

$$\begin{aligned} y_{ij} &= \eta(t_{ij}) + v_i(t_{ij}) + \epsilon_{ij}, \quad j = 1, 2, \cdots, n_i;\ i = 1, 2, \cdots, n, \\ v_i(t) &\sim \mathrm{GP}(0, \gamma), \quad \boldsymbol{\epsilon}_i = [\epsilon_{i1}, \cdots, \epsilon_{in_i}]^T \sim N(0, \sigma^2 \boldsymbol{I}_{n_i}), \end{aligned}$$

where

- ϵ_{ij}: j-th measurement error of i-th subject
- $\eta(t)$: smooth FE function, modelling the popular feature
- $v_i(t)$: smooth RE function, modelling the individual feature

Aims: Estimate $\eta(t)$, $\gamma(s,t)$ and σ^2

Regression Spline Smoothing for Longitudinal Data

Fitting the NPME model Using Regression Splines

- Approximating $\eta(t)$ by a regression spline $\boldsymbol{\Phi}(t)^T\boldsymbol{\beta}$
- Approximating $v_i(t)$ by regression splines $\boldsymbol{\Psi}(t)^T\boldsymbol{b}_i$
- $\boldsymbol{\Phi}(t)$: order k TPB with K interior knots
- $\boldsymbol{\Psi}(t)$: order k_v TPB with K_v interior knots
- $\boldsymbol{\beta}$ and $\boldsymbol{b}_i$ are the associated coefficient vectors

Regression Spline Smoothing for Longitudinal Data

The Approximation LME Model

$$
\begin{aligned}
y_{ij} &= x_{ij}^T \boldsymbol{\beta} + z_{ij}^T \boldsymbol{b}_i + \epsilon_{ij}, \quad j = 1, 2, \cdots, n_i; \; i = 1, 2, \cdots, n, \\
\boldsymbol{b}_i &\sim N(0, \boldsymbol{D}), \quad \boldsymbol{\epsilon}_i \sim N(0, \sigma^2 \boldsymbol{I}_{n_i})
\end{aligned}
$$

- $x_{ij} = \boldsymbol{\Phi}(t_{ij})$ and $z_{ij} = \boldsymbol{\Psi}(t_{ij})$
- For given $\boldsymbol{\Phi}(t)$ and $\boldsymbol{\Psi}(t)$, the approximation LME model can be fitted easily using existing software

Regression Spline Smoothing for Longitudinal Data

Remarks for NPME Modelling

- For parametric ME modelling, the model is fixed
- For NPME modelling, the model is data-driven
- For the regression spline-based approximation LME model, the model is specified by the TPBs $\mathbf{\Phi}(t)$ and $\mathbf{\Psi}(t)$
- The knots of $\mathbf{\Phi}(t)$ and $\mathbf{\Psi}(t)$ can be located using the two methods described in Part I
- Choosing the knot numbers, K and K_v to tradeoff the goodness of fit and model complexity

Regression Spline Smoothing for Longitudinal Data

Model Complexity and Goodness of Fit

- **FE and RE Predictions**: $\hat{\mathbf{y}} = \boldsymbol{X}\hat{\boldsymbol{\beta}} = \boldsymbol{A}\boldsymbol{y}$ and $\hat{\mathbf{v}} = \boldsymbol{Z}\hat{\mathbf{b}} = \boldsymbol{A}_v\boldsymbol{y}$
- **FE and RE Smoother Matrices**: $\boldsymbol{A}$ and $\boldsymbol{A}_v$
- **FE and RE Model Complexity**: $\mathrm{df} = \mathrm{tr}(\boldsymbol{A})$ and $\mathrm{df}_v = \mathrm{tr}(\boldsymbol{A}_v)$
- **Goodness of fit**: Loglik (Log-likelihood)

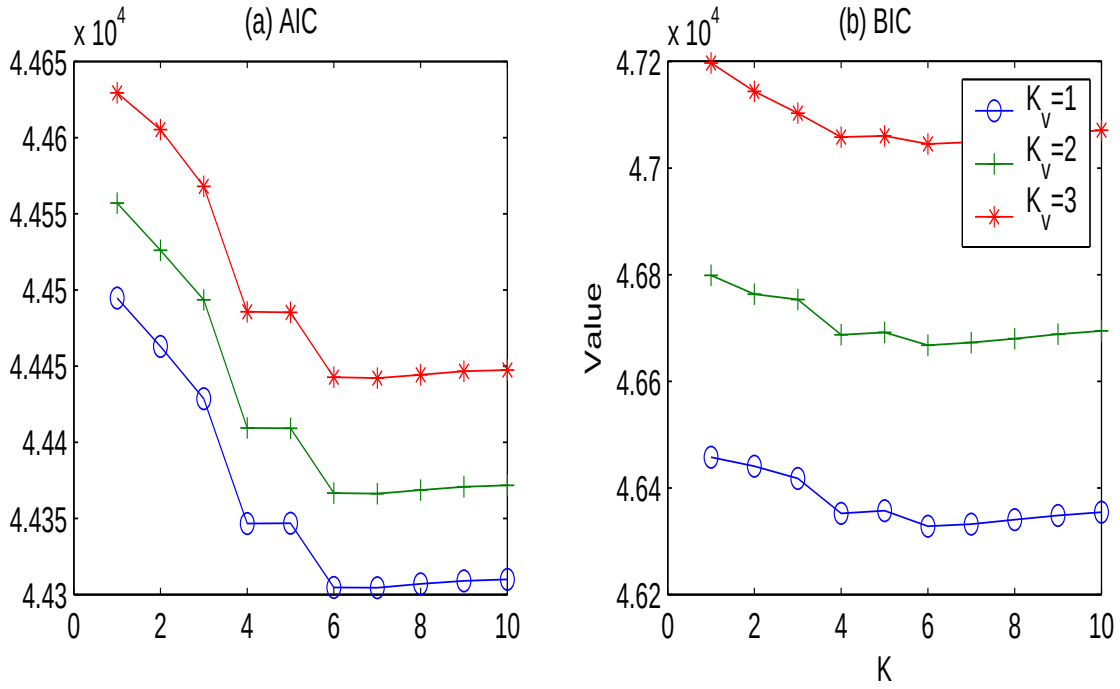
Regression Spline Smoothing for Longitudinal Data

Criteria for Knot Number Selection: AIC and BIC

- **A Criterion** should trade off "Goodness of Fit" and "Model Complexity"
- For the regression spline-based NPME modelling, one may define
  - $\text{AIC}(K, K_v) = -2\text{Loglik} + 2(\text{df} + \text{df}_v + 1)$,
  - $\text{BIC}(K, K_v) = -2\text{Loglik} + log(N)(\text{df} + \text{df}_v + 1)$.

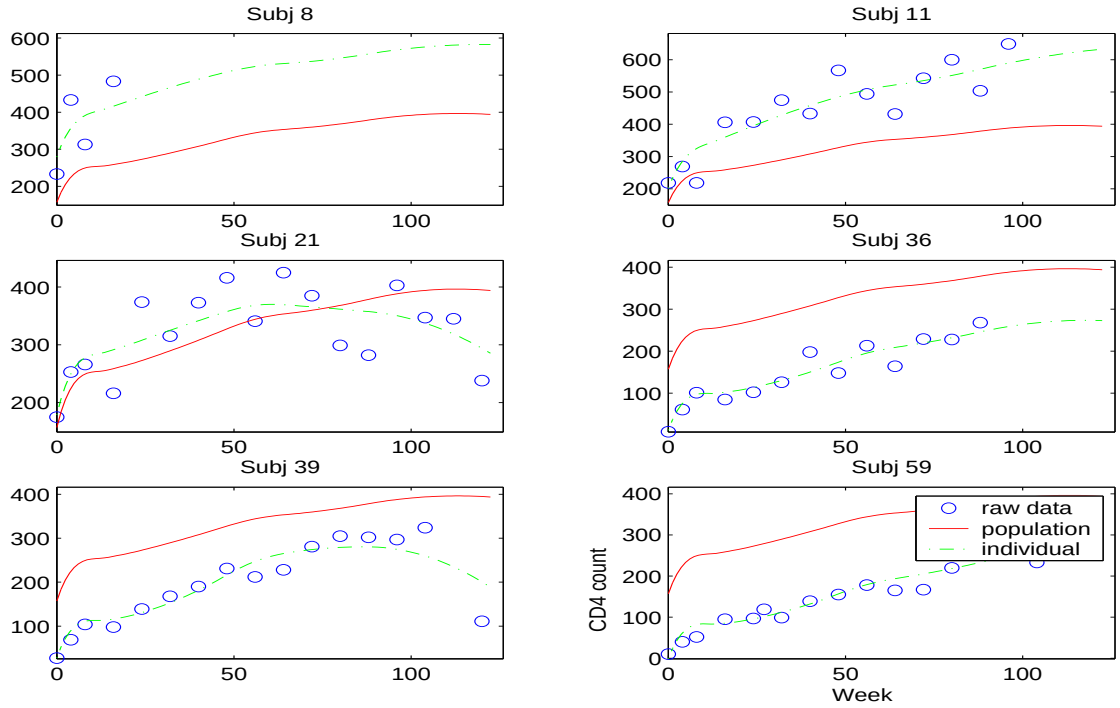
Regression Spline Smoothing for Longitudinal Data

Applications to the ACTG 388 Data



Regression Spline Smoothing for Longitudinal Data

Plots of six individual fits



Regression Spline Smoothing for Longitudinal Data

Extensions to other Non/Semiparametric ME Models

- **Semiparametric ME Model**:

$$\begin{aligned} y_{ij} &= \eta(t_{ij}) + \boldsymbol{z}_{ij}^T \boldsymbol{b}_i + \epsilon_{ij}, \\ y_{ij} &= \boldsymbol{x}_{ij}^T \boldsymbol{\beta} + v_i(t_{ij}) + \epsilon_{ij}. \end{aligned}$$

- **Varying-coefficients ME Model**:

$$y_{ij} = \boldsymbol{x}_{ij}^T \boldsymbol{\beta}(t_{ij}) + v_i(t_{ij}) + \epsilon_{ij}$$

- **Random-coefficient Model**: $y_{ij} = \boldsymbol{x}_{ij}^T \{\eta(t_{ij}) + v_i(t_{ij})\} + \epsilon_{ij}$,
- **Generalized Nonparametric ME Model**:
  $y_{ij} = \phi\{\eta(t_{ij}) + v_i(t_{ij})\} + \epsilon_{ij}$, where $\phi(\cdot)$ is known.

.

Summary

- In Part I, we show how to fit a single curve using regression splines
- In Part II, we show how to fit a group of curves using regression splines
- Regression splines can be used to fit other non/semiparametric models
- Other smoothing methods can also be applied; see Wu and Zhang (2006) for details

That's all!

THANK YOU VERY MUCH