Two parts

- **Old:** overview of the book *Semiparametric Regression* by Ruppert, Wand, and Carroll (2003)
- **New:** asymptotics of penalized splines
Example 1 (courtesy of Rich Canfield, Nutrition, Cornell)

- blood lead and intelligence measured on children
- **Question:** how do low doses of lead affect IQ?
  - important since doses are decreasing with lead now out of gasoline
- several IQ measurements per child
  - so longitudinal
- nine “confounders”
  - e. g., maternal IQ
  - need to adjust for them
- **effect of lead appears nonlinear**
  - important conclusion
Dose-response curve

Thanks to Rich Canfield for data and estimates
Spinal bone mineral density example

Example II (in Ruppert, Wand, Carroll (2003), *Semiparametric Regression*)

- age and spinal bone mineral density measured on girls and young women
- several measurements on each subject
- increasing but nonlinear curves
Spinal bone mineral density data
What is needed to accommodate these examples

We need a model with

- potentially many variables
- possibility of nonlinear effects
- random subject-specific effects

The model should be one that can be fit with readily available software such as SAS, Splus, or R.
Underlying philosophy

1. minimalist statistics
   - keep it as simple as possible
2. build on classical parametric statistics
3. modular methodology
   - so we can add components to accommodate special features in data sets
Outline of the approach

- Start with linear mixed model
  - allows random subject-specific effects
  - fine for variables that enter linearly
- Expand the basis for those variables that have nonlinear effects
  - we will use a spline basis
  - treat the spline coefficients as random effects to induce empirical Bayes shrinkage = smoothing
- End result
  - linear mixed model from a software perspective, but nonlinear from a modeling perspective
Example: pig weights (random effects)

Example III (from Ruppert, Wand, and Carroll (2003))

(a)

(b)

(weight vs. number of weeks)

(weight vs. number of weeks)
Random intercept model

\[ Y_{ij} = (\beta_0 + b_{0i}) + \beta_1 \text{week}_j \]

- \( Y_{ij} \) = weight of \( i \)th pig at the \( j \)th week
- \( \beta_0 \) is the average intercept for pigs
- \( b_{0i} \) is an offset for \( i \)th pig
- So \( (\beta_0 + b_{0i}) \) is the intercept for the \( i \)th pig
Are random intercepts enough?

Example III

(a)

(b)

weight

number of weeks

weight

number of weeks

20 40 60 80

20 40 60 80
Random lines model

\[ Y_{ij} = (\beta_0 + b_{0i}) + (\beta_1 + b_{1i}) \text{ week}_j \]

- \( \beta_1 \) is the average slope
- \( b_{ii} \) is an adjustment to slope of the \( i \)th pig
- So \( (\beta_1 + b_{1i}) \) is the slope for the \( i \)th pig
- \( b_{0i} \) and \( b_{1i} \) seem positively correlated
  - makes sense: faster growing pigs should be larger at the start of data collection
General form of linear mixed model

Model is:

\[ Y_i = X_i^T \beta + Z_i^T b + \epsilon_i \]

- \( X_i = (X_{i1}, \ldots, X_{ip}) \) and \( Z_i = (Z_{i1}, \ldots, Z_{iq}) \) are vectors of predictor variables
- \( \beta = (\beta_1, \ldots, \beta_p) \) is a vector of fixed effects
- \( b = (b_1, \ldots, b_q) \) is a vector of random effects
  - \( b \sim MVN(0, \Sigma(\theta)) \)
  - \( \theta \) is a vector of variance components
Estimation in linear mixed models

- $\beta$ and $\theta$ are the parameter vectors
  - estimated by
    - ML (maximum likelihood), or
    - REML (maximum likelihood with degrees of freedom correction)

- $b$ is a vector of random variables
  - predicted by a BLUP (Best linear unbiased predictor)
  - BLUP is shrunk towards zero (mean of $b$)
  - amount of shrinkage depends on $\hat{\theta}$
Estimation in linear mixed models, cont.

- **Random intercepts example:**
  
  \[ Y_{ij} = (\beta_0 + b_{0i}) + \beta_1 \text{week}_j \]

- **High variability** among the intercepts \( \Rightarrow \) less shrinkage of \( b_{0i} \) towards 0
  - extreme case: intercepts are fixed effects

- **Low variability** among the intercepts \( \Rightarrow \) more shrinkage
  - extreme case: common intercept (another fixed effects model)
Comparison between fixed and random effects modeling

- fixed effects models allow only the two extremes:
  - no shrinkage
  - common intercept
- mixed effects modeling allows all possibilities between these extremes
Splines

- polynomials are excellent for local approximation of functions
- in practice, polynomials are relatively poor at global approximation
- a spline is made by joining polynomials together
  - takes advantage of polynomials strengths without inheriting their weaknesses
- splines have "maximal smoothness"
Piecewise linear spline model

“Positive part” notation:

\[ x_+ = x, \text{ if } x > 0 \] (1)
\[ = 0, \text{ if } x \leq 0 \] (2)

Linear spline:

\[ m(x) = \{ \beta_0 + \beta_1 x \} + \{ b_1 (x - \kappa_1)_+ + \cdots + b_K (x - \kappa_K)_+ \} \]

- \( \kappa_1, \ldots, \kappa_K \) are “knots”
- \( b_1, \ldots, b_K \) are the spline coefficients
Linear “plus” function with $\kappa = 1$
Linear spline

\[ m(x) = \beta_0 + \beta_1 x + b_1(x - \kappa_1)_+ + \cdots + b_K(x - \kappa_K)_+ \]

- slope jumps by \( b_k \) at \( \kappa_k, \ k = 1, \ldots, K \)
Fitting LIDAR data with plus functions
Generalization: higher degree splines

\[ m(x) = \beta_0 + \beta_1 x + \cdots + \beta_p x^p + b_1 (x - \kappa_1)_+^p + \cdots + b_K (x - \kappa_K)_+^p \]

- \( p \)th derivative jumps by \( p! b_k \) at \( \kappa_k \)
- first \( p - 1 \) derivatives are continuous
LIDAR data: ordinary Least Squares

- Raw Data
- 2 knots
- 3 knots
- 5 knots
- 10 knots
- 20 knots
- 50 knots
- 100 knots

Penalized Splines, Mixed Models, and Recent Large-Sample Results

David Ruppert

Outline

Semiparametric Regression

Introduction
Mixed linear models
Univariate splines
Back to examples
Summary

Asymptotic Theory

Framework and summary
0-degree splines
Linear Splines
Work in progress
Summary
Penalized least-squares

- Use matrix notation:

\[ m(X_i) = \beta_0 + \beta_1 X_i + \cdots + \beta_p X_i^p + b_1 (X_i - \kappa_1)_+^p + \cdots + b_K (X_i - \kappa_K)_+^p \]

\[ = X_i^T \beta_X + B^T (X_i) b \]

- Minimize

\[ \sum_{i=1}^{n} \left( Y_i - (X_i^T \beta_X + B^T (X_i) b) \right)^2 + \lambda b^T Db. \]
Penalized least-squares, cont.

- From previous slide: minimize
  \[
  \sum_{i=1}^{n} \left\{ Y_i - (X_i^T \beta X + B^T(X_i)b) \right\}^2 + \lambda b^TDb.
  \]

- \( \lambda b^TDb \) is a penalty that prevents overfitting
- \( D \) is a positive semidefinite matrix
  - so the penalty is non-negative
  - Example:
    \[
    D = I
    \]
- \( \lambda \) controls that amount of penalization
- the choice of \( \lambda \) is crucial
Penalized Least Squares

Penalized Splines, Mixed Models, and Recent Large-Sample Results

David Ruppert

Outline

Semiparametric Regression

Introduction

Mixed linear models

Univariate splines

Back to examples

Summary

Asymptotic Theory

Framework and summary

0-degree splines

Linear Splines

Work in progress

Summary
Selecting $\lambda$

To choose $\lambda$ use:

1. one of several model selection criteria:
   - cross-validation (CV)
   - generalized cross-validation (GCV)
   - AIC
   - $C_P$

2. ML or REML in mixed model framework
   - convenient because one can add other random effects
   - also can use standard mixed model software
Return to spinal bone mineral density study

\[ \text{SBMD}_{i,j} = U_i + m(\text{age}_{i,j}) + \epsilon_{i,j}, \]

\[ i = 1, \ldots, m = 230, \quad j = i, \ldots, n_i. \]
Fixed effects

\[
X = \begin{bmatrix}
1 & \text{age}_{11} \\
\vdots & \vdots \\
1 & \text{age}_{1n_1} \\
\vdots & \vdots \\
1 & \text{age}_{m1} \\
\vdots & \vdots \\
1 & \text{age}_{mn_m}
\end{bmatrix}
\]
Random effects

$$Z = \begin{bmatrix} 1 & \cdots & 0 & (age_{11} - \kappa_1)^+ & \cdots & (age_{11} - \kappa_K)^+ \\ \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ 1 & \cdots & 0 & (age_{1n_1} - \kappa_1)^+ & \cdots & (age_{1n_1} - \kappa_K)^+ \\ \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ 0 & \cdots & 1 & (age_{m1} - \kappa_1)^+ & \cdots & (age_{m1} - \kappa_K)^+ \\ \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ 0 & \cdots & 1 & (age_{mn_m} - \kappa_1)^+ & \cdots & (age_{mn_m} - \kappa_K)^+ \end{bmatrix}$$
Random effects

\[
\mathbf{u} = \begin{bmatrix}
U_1 \\
\vdots \\
U_m \\
b_1 \\
\vdots \\
b_K
\end{bmatrix}
\]
Random effects

Variability bars on \( \hat{m} \) and estimated density of \( U_i \)
Modeling the blood lead and IQ data

For the $j$th measurements on the $i$th subject:

$$IQ_{ij} = b_i + m(lead_{ij}) + \beta_1 X_{ij}^1 + \cdots + \beta_L X_{ij}^L + \epsilon_{ij}$$

- $m(\cdot)$ is a spline
  - include the population average intercept
- $b_i$ is a random subject-specific intercept
  - $E(b_i) = 0$
  - model assumes parallel curves
- $X_{ij}^\ell$ is the value of the $\ell$th confounder, $\ell = 1, \ldots, L$
Summary (overview of semiparametric regression)

- **Semiparametric philosophy**
  - use nonparametric models where needed
  - but only where needed

- LMMs and GLMMs are fantastic tools, but (apparently) totally parametric

- By basis expansion, LMMs and GLMMs become semiparametric

- Low-rank splines eliminate computational bottlenecks

- Smoothing parameters can be estimated as ratios of variance components
Li and Ruppert (2008, *Biometrika*)

- $p$-degree spline model:

  $$f(x) = \sum_{k=1}^{K+p} b_k B_k(x), \quad x \in (0, 1)$$

- $p$th degree B-spline basis:

  $$\{B_k : k = 1, \ldots, K + p\}$$

- knots:

  $$\kappa_0 = 0 < \kappa_1 < \ldots < \kappa_K = 1$$
B-splines

0-degree B-splines

Linear B-splines

Quadratic B-splines
Outline of asymptotic theory

1. First: summary
2. Go through the case $p = 0$, $m = 1$, equally-spaced $x_i$ carefully
3. Then do $p = 0$ and $m = 2$
4. Discuss higher order cases and unequally-spaced data
Penalized spline estimators are approximately binned Nadaraya-Watson kernel estimators

Penalized splines are not design-adaptive in the sense of Fan (1992)

The order of the N-W kernel depends solely on $m$ (order of penalty)

- this was surprising to us
- order of kernel is $2m$ in the interior
- order is $m$ at boundaries
The spline degree $p$ does not affect the asymptotic distribution, but

- $p$ determines the type of binning and the minimum rate at which $K \to \infty$
- $p = 0 \implies$ usual binning
- $p = 1 \implies$ linear binning
- a higher value of $p$ means that less knots are needed (because there is less binning bias)
Penalized least-squares

Penalized least-squares minimizes

\[
\sum_{i=1}^{n} \left\{ y_i - \sum_{k=1}^{K+p} \widehat{b}_k B_i(x_i) \right\}^2 + \lambda \sum_{k=m+1}^{K+p} \{ \Delta^m(\widehat{b}_k) \}^2,
\]

\( \Delta b_k = b_k - b_{k-1} \) and \( \Delta^m = \Delta(\Delta^{m-1}) \)

- \( m = 1 \) ⇒ constant functions are unpenalized
- \( m = 2 \) ⇒ linear functions are unpenalized
\begin{align*}
p = 0, \ m = 1
\end{align*}

Assume:

- \(x_1 = 1/n, x_2 = 2/n, \ldots, x_n = 1\)
- \(\kappa_0 = 0, \kappa_1 = 1/K, \kappa_2 = 2/K, \ldots, \kappa_K = 1\)
- \(B_k(x) = I\{\kappa_{k-1} < x \leq \kappa_k\}, 1 \leq k \leq K\) (\(k\)th bin indicator)
- assume that \(n/K := M\) is an integer
- then \(X^T X = MI_K\) where \(I_K\)
Assume further:

- $m = 1$

Then

$$D^T D = \begin{pmatrix}
1 & -1 & 0 & \cdots & 0 & 0 \\
-1 & 2 & -1 & \cdots & 0 & 0 \\
0 & -1 & 2 & \cdots & 0 & 0 \\
\vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\
0 & 0 & 0 & \cdots & 2 & -1 \\
0 & 0 & 0 & \cdots & -1 & 1
\end{pmatrix},$$
\( p = 0, \ m = 1, \ \text{PLS estimator} \)

**The Penalized LS estimator solves:**

\[
\Lambda \hat{b} = z = \bar{y}/(1 + 2\lambda) \quad (\bar{y} = \text{bin averages})
\]

where

\[
\Lambda = \begin{pmatrix}
\theta & \eta & 0 & 0 & \cdots & 0 & 0 \\
\eta & 1 & \eta & 0 & \cdots & 0 & 0 \\
0 & \eta & 1 & \eta & \cdots & 0 & 0 \\
0 & 0 & \eta & 1 & \cdots & 0 & 0 \\
\vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\
0 & 0 & 0 & 0 & \cdots & 1 & \eta \\
0 & 0 & 0 & 0 & \cdots & \eta & \theta
\end{pmatrix}, \quad \eta = -\frac{\lambda}{1 + 2\lambda}
\]
Let $\rho \in (0, 1)$ be a root of

$$\eta + \rho + \eta \rho^2 = 0.$$ 

Then

$$\rho = \frac{1 - \sqrt{1 - 4\eta^2}}{-2\eta} = \frac{1 + 2\lambda - \sqrt{1 + 4\lambda}}{2\lambda}.$$ 

Define

$$T_i = (\rho^{i-1}, \rho^{i-2}, \ldots, \rho, 1, \rho, \rho^2, \ldots, \rho^{K-i})^T.$$ 

$T_i$ is orthogonal to all columns of $\Lambda$ except the first, last, and $i$th (so $T_i$ is the $i$th row of $\Lambda$, except for a geometrically convergent error)
Finite-sample kernel defined by:

\[ \hat{f}(x) = \sum_{j=1}^{K} H(x, \bar{x}_j) \bar{y}_i \]

\[ \frac{T_i^T}{1 + 2\lambda} = \frac{(\rho^{i-1}, \rho^{i-2}, \ldots, \rho, 1, \rho, \rho^2, \ldots, \rho^{K-i})}{1 + 2\lambda} \]

is the finite-sample kernel (ignoring asymptotically negligible boundary effects).
Three kernels corresponding to first-order penalty

- finite-sample kernel is linear combination of three kernels
  - double exponential kernel centered at \( x \)
  - boundary kernels are \( \exp(-x) \) and \( \exp(x) \)
- weights for the boundary kernels are asymptotically negligible in interior

\( x \) is an “estimation point” (here fixed at 0.4)
Finite-sample kernels, first-order penalty

![Graph showing finite-sample kernels, first-order penalty](image)
Connection with smoothing splines

We get the same equivalent kernels (Silverman, 1985) as for smoothing splines with a penalty on the first derivative.
Finding $\hat{b}_i$ – interior case

- Suppose $i/K \to x \in (0, 1)$ (non-boundary case)
- After some algebra:
  \[
  \hat{b}_i \sim \frac{\sum_{j=1}^{K} \rho|i-j| \bar{y}_j}{\sum_{j=1}^{K} \rho|i-j|}.
  \]
- Note that
  \[
  \hat{f}(x) = \hat{b}_i
  \]
  for $x$ in the $i$th bin
Equivalence to N-W kernel estimator

- After some more algebra
  \[ \rho |i-j| \sim \exp \left\{ -\frac{|x_i - x_j|}{hn^{-1/5}} \right\} \]

- Thus, \( \hat{f}_n \) is asymptotically equivalent to the Nadaraya-Watson estimator with
  - double exponential kernel \( H(x) = (1/2) \exp(-|x|) \)
  - bandwidth \( hn^{-1/5} \)
Nadaraya-Watson kernel estimators

Model:

\[ y_i = f(x_i) + \epsilon_i \]

Nadaraya-Watson estimator:

\[
\hat{f}(x) = \frac{\sum_{i=1}^{n} H\{ (x_i - x)/h_n \} y_i}{\sum_{i=1}^{n} H\{ (x_i - x)/h_n \}}
\]

- \( H(\cdot) \) is called the kernel function
- \( h_n \) is the bandwidth
Binned Nadaraya-Watson kernel estimators

Binned Nadaraya-Watson estimator:

- range of the $x_i$ divided into $K$ subintervals (bins)
- $\overline{x_j}$ is average of $x_i$ in $i$th bin
- $\overline{y_j}$ is average of $y_i$ such that $x_i$ is in the $i$th bin

$$
\hat{f}(x) = \frac{\sum_{j=1}^{K} H\{(\overline{x_j} - x)/h_n\}\overline{y_j}}{\sum_{j=1}^{K} H\{(\overline{x_i} - x)/h_n\}}
$$
Thus, \( \hat{f}_n \) is asymptotically equivalent to a binned Nadaraya-Watson estimator with

- double exponential kernel \( H(x) = \frac{1}{2} \exp(-|x|) \)
- bandwidth \( hn^{-1/5} \)

Binning bias is negligible if \( K = Cn^\gamma \) for \( \gamma > 2/5 \) and \( C > 0 \)

“negligible” means \( o(n^{-2/5}) \)
Selecting $\lambda$ to achieve desired bandwidth

To get bandwidth $hn^{-1/5}$ we need $\lambda$ chosen as

$$\lambda \sim \{(Cn^\gamma)(hn^{-1/5})\}^2 = (\# \text{ knots} \times \text{bandwidth})^2$$
Asymptotic Distribution

For $x \in (0, 1)$, as $n \to \infty$ we have

$$n^{2/5} \{ \hat{f}_n(x) - f(x) \} \Rightarrow N \{ B(x), V(x) \}$$

where

- $B(x) = h^2 f^{(2)}(x)$
- $V(x) = 4^{-1} h^{-1} \sigma^2(x)$
Some folklore

- **Folklore:** The number of knots is not important, provided that it is large enough.
  - **Confirmation:**
    \[ K \sim C n^\gamma \text{ with } C > 0 \text{ and } \gamma > 2/5. \] (3)

- **Folklore:** The value of the penalty parameter is crucial.
  - **Confirmation:**
    \[ \lambda \sim C^2 h^2 n^{2\gamma - 2/5} = (\# \text{ knots} \times \text{ bandwidth})^2 \] (4)
    for some \( h > 0 \).

- **Folklore:** Modeling bias is small.
  - **Confirmation:** Modeling bias does not appear in asymptotic bias provided (3) and (4) hold.
Order of a kernel and bias

Moments: \( k \)-th moment is \( \int x^k H(x) \, dx \)

Order of kernel: A kernel is of \( k \)-th order if the first non-zero moment is the \( k \)-th

- Non-negative kernel: order is at most 2

Bias: \( \text{bias} = O\{(\text{bandwidth})^k\} \)

Variance:
\[
\text{variance} = O\left(\frac{1}{n \times \text{bandwidth}}\right)
\]

and
\[
\text{optimal RMSE} = O\left(n^{-k/(2k+1)}\right)
\]
2nd order-penalty gives 4th order kernel (in interior)

Now let \( m = 2 \) (2nd order difference penalty)

- Assume:
  - \( K \sim Cn^{\gamma} \) with \( C > 0 \) and \( \gamma > 4/9 \)
  - \( \lambda \sim 4C^4h^4n^{4\gamma-4/9} \sim 4(Khn^{-1/9})^4. \)

Then for any \( x \in (0, 1) \), when \( n \to \infty \), we have

\[
n^{4/9}\{ \hat{f}_n(x) - f(x) \} \Rightarrow N\{B_1(x), V_1(x)\},
\]

where

- \( B_1(x) = (1/24)h^4f^{(4)}(x) \int x^4 T(x) \, dx \)
- \( V_1(x) = h^{-1} \{ \int T^2(x) \, dx \} \sigma^2(x) \)
  - \( T(x) \) is a fourth order kernel
Mathematical approach

Main technical device uses roots of the polynomial

\[ w(\xi) = \lambda(1-4\xi+6\xi^2-4\xi^3+\xi^4) + \xi^2 = \lambda(1-\xi)^4 + \xi^2, \quad \lambda > 0 \]

- No real roots and no roots of modulus one
- Roots are: \( r, \text{conj}(r), r^{-1}, \text{conj}(r)^{-1} \) (all distinct)
- Use the conjugate pair with modulus less than one
The asymptotic kernel is a linear combination of

\[ \exp(-|x|) \cos(x) \quad \text{and} \quad \exp(-|x|) \sin(|x|) \]

Same equivalent kernel (Silverman, 1985) as for smoothing splines with a penalty on the second derivative.
Finite-sample kernels, second-order penalty
Linear splines need less knots

Assume $m = 1$ (1st-order difference penalty).

- If $p = 1$ (linear), then require
  \[ K \sim Cn^\gamma \text{ with } C > 0 \text{ and } \gamma > 1/5 \]

- When $p$ was 0 (piecewise constant), we required
  \[ \gamma > 2/5 \]

- Otherwise, results are the same as for 0-degree and linear splines

A similar result holds for $m = 2$. 
Conjectures

- **Conjecture**: For $x$ in the interior:

  \[ P\text{-spline} \sim \text{N-W estimator with an } 2m\text{-order kernel} \]

  - Recall: $m$ is order of difference penalty
  - Kernel order independent of $p = \text{degree of spline}$
  - Shown to hold for $m = 1, 2$ and $p = 0, 1$

  - $p = 1$ requires less knots than $p = 0$
    - What happens for $p > 1$?
    - **Conjecture**: Still less knots are needed

- Conjectures are nearly proved: Li, Apanosovich, Ruppert (2009)
Unequally spaced $X$

- Assume $G(x_t) = t/n$ for a smooth $G$ with $g = G'$
- Fit a spline to $(Y_t, u_t)$ with regression function $f \circ G^{-1}$
  - evaluate this estimate at $G(x)$ to estimate $f(x)$
- Equally spaced knots for $(Y_t, u_t)$ implies knots at sample quantiles for $(Y_t, x_t)$
- Asymptotic bias is
  \[
  h^2 (f \circ G^{-1})^{(2)} \{ G(x) \} = \frac{h^2}{g^2(x)} \left\{ f^{(2)}(x) - \frac{f'(x)g'(x)}{g(x)} \right\}
  \]
- Nadaraya-Watson bias is
  \[
  h^2 \left\{ f^{(2)}(x) + \frac{2f'(x)g'(x)}{g(x)} \right\}
  \]
We use only one of two potential smoothing parameters

Both \( K \) and \( \lambda \) are potential smoothing parameters

- In our asymptotic theory, only \( \lambda \) plays the role of a smoothing parameter
- Could develop a theory where only \( K \) plays this role
  - would be similar to regression spline (\( \lambda = 0 \)) theory
- One could also choose \( K \) and \( \lambda \) so that both have a non-negligible effect
- Our theory mimics actual practice
Summary (asymptotics)

- P-spline estimators \( \approx \) binned N-W kernel estimators
- The number of knots unimportant if above a minimum
- Degree of spline
  - determines minimum convergence rate for number of knots
  - does not affect rate of convergence
- Order of penalty determines
  - order of equivalent kernel
  - convergence rate of estimator
- \( m \)th order penalty \( \iff \) smoothing spline with penalty on \( m \)th difference
Thanks for your attention