Penalized Splines and Small Area Estimation

Jean Opsomer
Iowa State University

Joint work with
Jay Breidt, Colorado State University
Gerda Claeskens, Université Catholique de Louvain
Göran Kauermann, Universität Bielefeld
Giovanna Ranalli, Università di Perugia

July 23, 2004
Outline

1. Introduction
2. Nonparametric regression using penalized splines
3. Small area estimation
4. Nonparametric small area estimation
5. Northeastern lakes survey
6. Conclusion
1. Introduction: Lakes Survey

Ecological condition survey of Northeastern lakes conducted by U.S. Environmental Protection Agency

Data collected for 338 lakes
Lakes Survey (2)

Region includes 113 8-digit “Hydrologic Unit Codes” (HUC)

Goal: estimate mean lake Acid Neutralizing Capacity (ANC) for all HUCs
2. Nonparametric Regression Using Penalized Splines

Many nonparametric regression methods are available

- Kernel and local polynomial methods
- Splines
  - smoothing splines
  - regression splines
  - penalized splines (P-splines)
- Orthogonal decomposition (wavelet, Fourier series)

Penalized spline regression (Eilers and Marx, 1996) is simple, flexible and computationally attractive smoothing method
Definition of Penalized Splines Model

Regression model \( y_i = m(x_i) + \varepsilon_i \)

Function \( m(\cdot) \) is unknown but assumed well approximated by polynomial spline

\[
m_K(x) = \beta_0 + \beta_1 x + \ldots + \beta_p x^p + \sum_{k=1}^{K} \beta_{p+k} (x - \kappa_k)^p_+
\]

- \( p \): degree of spline (fixed)
- \( \kappa_1 < \ldots < \kappa_K \): set of \( K \) knots (fixed)
- \( \beta = (\beta_0, \ldots, \beta_{p+K}) \): vector of parameters (unknown)

(Ruppert, Wand and Carroll, 2003)
Polynomial Spline Basis Functions

$$(x - \kappa)_+^p \equiv \begin{cases} 
(x - \kappa)^p & \text{if } x - \kappa > 0 \\
0 & \text{if } x - \kappa \leq 0
\end{cases}$$

Other spline basis functions are possible (B-splines, radial splines)
Choosing $K$

If $K$ is sufficiently large, $m_K(\cdot)$ can approximate large class of functions

$\Rightarrow$ Rule of thumb: $K = \min(\#X/4, 35)$ (Ruppert, 2002)
Expressing Spline Model as Parametric Model

\[ m_K(x) = \beta_0 + \beta_1 x + \ldots + \beta_p x^p + \sum_{k=1}^{K} \beta_{p+k} (x - \kappa_k)_+^p \]

\[ \equiv \mathbf{x}^* \mathbf{\beta} \]

with

\[ \mathbf{x}^* = (1, x, \ldots, x^p, (x - \kappa_1)_+^p, \ldots, (x - \kappa_K)_+^p) \]
\[ \mathbf{\beta} = (\beta_0, \ldots, \beta_{p+K})^T \]
Fitting by Penalized Splines Regression

Minimize penalized sum of squares

$$\min_{\beta} \sum_{i=1}^{n} (y_i - m_K(x_i; \beta))^2 + \lambda \sum_{k=1}^{K} \beta_{p+k}^2$$

$$\Rightarrow \hat{m}_{K,\lambda}(x) = x^* \hat{\beta}_\lambda = x^* \left( X^{*T} X^* + \lambda A \right)^{-1} X^{*T} Y$$

$$\lambda = \text{smoothing penalty (fixed)}$$

$$A = \text{diag}\{0, \ldots, 0, 1, \ldots, 1\}$$

$$X^* = \text{design matrix (including spline terms)}$$

$\hat{\beta}_\lambda$ is ridge regression estimator, with ridge penalty on nonlinear (spline) terms of model
Fitting by Penalized Splines Regression (2)

λ protects against overfitting and determines smoothness of fit
Choosing the Penalty $\lambda$

- Cross-Validation: minimize CV sum of squares with respect to $\lambda$
- Mixed model approach: treat spline parameters $\beta_{p+1}, \ldots, \beta_{p+K}$ as a random effect with common variance $\sigma^2_\beta$ and fit regression using Maximum Likelihood approach (MLE, REML)
P-spline: “Hybrid” Regression Method

• P-spline is a nonparametric regression method:
  – can fit very large classes of functions
  – adaptive to local features in the data
  – smoothness of function is determined by penalty parameter $\lambda$
P-spline: “Hybrid” Regression Method

- P-spline is a nonparametric regression method:
  - can fit very large classes of functions
  - adaptive to local features in the data
  - smoothness of function is determined by penalty parameter $\lambda$

- P-spline is a parametric regression method:
  - model can be written as $x^*\beta$
  - fitted by (global) least squares method
  - number of parameters $p + K$ puts upper bound on flexibility of model
P-Splines or Local Polynomials?

Advantages of P-Splines

• Closely related to parametric modelling
• Model is easy to extend to multivariate, additive, semiparametric cases
• Handles data sparseness easily, very fast to compute
• Fits “look” better
P-Splines or Local Polynomials?

Advantages of P-Splines

• Closely related to parametric modelling
• Model is easy to extend to multivariate, additive, semiparametric cases
• Handles data sparseness easily, very fast to compute
• Fits “look” better

Disadvantages of P-Splines

• Flexibility of model limited by number of parameters $p + K$
• No “true” asymptotic theory
Extending the model

- Semi-parametric regression

Model \( y_i = m(x_{1i}; \beta_1) + x_{2i}\beta_2 + \varepsilon_i \)

\[
\sum_{i=1}^{n} (y_i - m_K(x_{1i}; \beta_1) - x_{2i}\beta_2)^2 + \lambda \sum_{k=1}^{K} \beta_{1,p+k}^2
\]
Extending the model

- Semi-parametric regression

Model \( y_i = m(x_{1i}; \beta_1) + x_{2i}\beta_2 + \varepsilon_i \)

\[
\sum_{i=1}^{n} \left( y_i - m_K(x_{1i}; \beta_1) - x_{2i}\beta_2 \right)^2 + \lambda \sum_{k=1}^{K} \beta_{1,p+k}^2
\]

- Additive model

Model \( y_i = m_1(x_{1i}; \beta_1) + m_2(x_{2i}; \beta_2) + \varepsilon_i \)

\[
\sum_{i=1}^{n} \left( y_i - m_{1,K}(x_{1i}; \beta_1) - m_{2,K}(x_{2i}; \beta_2) \right)^2 + \lambda_1 \sum_{k=1}^{K} \beta_{1,p+k}^2 + \lambda_2 \sum_{k=1}^{K} \beta_{2,p+k}^2
\]

- Other...
Fits Often “Look” Better

Note: this is subjective...
Theory for P-spline Regression?

\[ \text{MSE} = \mathbb{E}\left( \hat{m}_{K,\lambda}(x) - m(x) \right)^2 \]

- Regression splines \((\lambda = 0, K \to \infty)\):
  \[ \text{MSE} = O \left( K^{-2p} + \frac{1}{nK^{-1}} \right) \]
  (Huang, 2001)

- Smoothing splines \((\lambda \to 0, K = n)\):
  \[ \text{MSE} = O \left( \lambda + \frac{1}{n\lambda^{1/2(p+1)}} \right) \]
  (Cox, 1983)

- Wand (1999): asymptotic approximation to P-spline MSE for \(K\) fixed and \(\lambda \to 0\)
Theory for P-spline Regression (2)

Hall and Opsomer (2004): white-noise model ($K = \infty$)

- Penalized least squares criterion

$$\min_{\beta(\cdot)} \int \left\{ y_t - \int \beta(s) \phi(t \mid s) \rho(s) \, ds \right\}^2 f(t) \, dt + \lambda \int \beta(t)^2 \, dt$$

with $\phi(t \mid s) = (t - s)^p_+$, and $\rho(\cdot)$ the density of the “knots”

- Estimator

$$\hat{m}(t) = \int \hat{\beta}(s) \phi(t \mid s) \rho(s) \, ds$$

- Mean squared error

$$\text{MSE} = O \left( \lambda + \frac{1}{n\lambda^{1/2(p+1)}} \right)$$
3. Small Area Estimation

Data contain 557 observations over 113 HUCs

Goal: produce estimates of ANC for all HUCs
HUC Sample Means

Problems: unreliable estimates, missing HUCs
HUC as “Small Areas”

Few sample observations available in most HUCs

- Average sample size/HUC: 4.9
- 64 HUCs contain less than 5 observations
- 27 out of 113 HUCs contain no sample observations

⇒ Modelling required to construct reliable HUC-level estimates
  - model combines overall trend for region with random effect for small areas
  - mixed model/prediction
  - called small area estimation in surveys statistics
Small Area Estimation as Mixed Model Regression

“Classical” small area estimation (Battese, Harter and Fuller, 1988):

- Population of interest $U$, divided into small areas $U_t$, $t = 1, \ldots, T$
- Variable of interest $y_i$ observed on sample, $i \in s$
- Auxiliary variable $x_i$ observed on sample, $i \in s$, with known small area means, $\bar{x}_t = \sum_{i \in U_t} x_i / N_t$
- Assume linear relationship between $y_i$ and $x_i$ in population, with random effect $u_t$ for small areas $U_t$, $t = 1, \ldots, T$
Small Area Estimation Regression Model

\[ y_i = x_i \beta + u_t + \varepsilon_i \quad i \in U_t \]

\[ = x_i \beta + d_i u + \varepsilon_i \]

\[ d_i = (d_{i1}, \ldots, d_{iT}) \quad d_{it} = \begin{cases} 1 & \text{if } i \in U_t \\ 0 & \text{otherwise} \end{cases} \]
Small Area Estimation Regression Model

\[ y_i = x_i \beta + u_t + \varepsilon_i \quad i \in U_t \]

\[ = x_i \beta + d_i u + \varepsilon_i \]

\[ d_i = (d_{i1}, \ldots, d_{iT}) \quad d_{it} = \begin{cases} 
1 & \text{if } i \in U_t \\
0 & \text{otherwise}
\end{cases} \]

\[ u = (u_1, \ldots, u_T) \sim \text{iid } \mathcal{F}_u(0, \sigma_u^2) \]

\[ \varepsilon_i \sim \mathcal{F}_\varepsilon(0, \sigma_\varepsilon^2) \]

In matrix form:

\[ Y = X\beta + Du + \varepsilon \]
Small Area Estimation: BLUP

- Small area estimation goal: predict
  \[ \bar{y}_t = \bar{x}_t \beta + u_t \quad t = 1, \ldots, T \]

- Assuming \( \sigma_u^2, \sigma_\varepsilon^2 \) known, Best Linear Unbiased Predictor (BLUP) of \( \bar{y}_t \) is
  \[ \hat{y}_t = \bar{x}_t \hat{\beta} + \hat{u}_t \]

  with
  \[ \hat{\beta} = (X'V^{-1}X)^{-1}X'V^{-1}Y \]
  \[ V = \text{Var}(Y) = \sigma_\varepsilon^2 I_n + \sigma_u^2 D'D \]
  \[ \hat{u} = \sigma_u^2 DV^{-1}(Y - X\hat{\beta}) \]

  (McCulloch and Searle, 2001)
BLUP as Ridge Regression Estimator

\[
\hat{y}_t = \bar{x}_t \hat{\beta} + \hat{u}_t
\]

with

\[
\begin{bmatrix}
\hat{\beta} \\
\hat{u}
\end{bmatrix} = \begin{bmatrix}
X'X & X'D \\
D'X & D'D + \frac{\sigma^2_{\varepsilon}}{\sigma^2_u}
\end{bmatrix}^{-1} \begin{bmatrix}
X'Y \\
D'Y
\end{bmatrix}
\]

\[
= \left( X^* X^* + \frac{\sigma^2_{\varepsilon}}{\sigma^2_u} A \right)^{-1} X^* Y
\]

where

\[
X^* = [XD]
\]
Small Area Estimation: EBLUP

When $\sigma^2_u, \sigma^2_\varepsilon$ unknown, Empirical BLUP (EBLUP) of $\bar{y}_t$ is found by Maximum Likelihood (ML) or Restricted Maximum Likelihood (REML)

$$\hat{y}_t = \bar{x}_t\hat{\beta} + \hat{u}_t$$

with

$$\hat{\beta} = (X'\hat{V}^{-1}X)^{-1}X'\hat{V}^{-1}Y$$

$$\hat{V} = \hat{\sigma}_\varepsilon^2 I_n + \hat{\sigma}_u^2 D'D$$

$$\hat{u} = \hat{\sigma}_u^2 D\hat{V}^{-1}(Y - X\hat{\beta})$$

(e.g. Searle, Casella and McCulloch, 1992)
Inference for Small Area Estimation

Target: prediction MSE = $\mathbb{E}(\hat{y}_t - \bar{y}_t)^2$

Asymptotic approximation of EBLUP available under linear mixed model specification

Can be estimated consistently
4. Nonparametric Small Area Estimation

- more flexible fixed component in model can improve prediction
- predicting for “empty” (no data) HUCs relies exclusively on fixed component
4. Nonparametric Small Area Estimation

- more flexible fixed component in model can improve prediction
- predicting for “empty” (no data) HUCs relies exclusively on fixed component

P-splines are ideally suited for small area estimation

- close relationship between “classical” small area estimation models and P-splines
- availability of existing software
- ability to evaluate need for nonlinearity in model and significance of small area effects
P-splines as Random Effects

\[ y_i = m_K(x_i) + \varepsilon_i \]
\[ = x_i^\ast \beta + \varepsilon_i \]
P-splines as Random Effects

\[ y_i = m_K(x_i) + \varepsilon_i \]
\[ = x^*_i \beta + \varepsilon_i \]
\[ = x^F_i \beta^F + z_i \gamma + \varepsilon_i \]

\[ x^F_i \beta^F \equiv \beta_0 + x_i \beta_1 + \ldots + x_i^p \beta_p \quad \text{(parametric, fixed component)} \]

\[ z_i \gamma = z_{1i} \gamma_1 + \ldots + z_{Ki} \gamma_K \]
\[ \equiv (x_i - \kappa_1)^p \beta_{p+1} + \ldots + (x_i - \kappa_K)^p \beta_{p+K} \]
P-splines as Random Effects

\[ y_i = m_K(x_i) + \varepsilon_i \]
\[ = x_i^* \beta + \varepsilon_i \]
\[ = x_i^F \beta^F + z_i \gamma + \varepsilon_i \]

\[ x_i^F \beta^F \equiv \beta_0 + x_i \beta_1 + \ldots + x_i^p \beta_p \quad \text{(parametric, fixed component)} \]

\[ z_i \gamma = z_1 \gamma_1 + \ldots + z_K \gamma_K \]
\[ \equiv (x_i - \kappa_1)^p \beta_{p+1} + \ldots + (x_i - \kappa_K)^p \beta_{p+K} \]
\[ \quad \text{(deviations from parametric, treated as random effect)} \]

\[ \gamma = (\gamma_1, \ldots, \gamma_K) \sim \text{iid } F_\gamma(0, \sigma^2_\gamma) \]
\[ \varepsilon_i \sim F_\varepsilon(0, \sigma^2_\varepsilon) \]
P-splines Estimator as BLUP

Assuming $\sigma^2_\gamma, \sigma^2_\varepsilon$ known, BLUP/BLUE for is solution to

$$\min_{\beta^F, \gamma} \sum_{i=1}^{n} (y_i - x_i^F \beta^F + z_i \gamma)^2 + \frac{\sigma^2_\varepsilon}{\sigma^2_\gamma} \sum_{k=1}^{K} \gamma_k^2$$

(Henderson et al., 1959)

$$\Rightarrow \begin{bmatrix} \hat{\beta}^F \\ \hat{\gamma} \end{bmatrix} = \left( X^{*\prime} X^* + \frac{\sigma^2_\varepsilon}{\sigma^2_\gamma} A \right)^{-1} X^{*\prime} Y$$

with

$$A = \text{diag}\{0, \ldots, 0, 1, \ldots, 1\}$$

$$X^* = [x^F \ z]$$

$$\begin{bmatrix} \hat{\beta}^F \\ \hat{\gamma} \end{bmatrix} = \text{P-splines (ridge) regression estimator } \hat{\beta}_\lambda \text{ with } \lambda = \frac{\sigma^2_\varepsilon}{\sigma^2_\gamma}$$
P-splines Estimator as EBLUP

If $\sigma_\gamma^2, \sigma_\varepsilon^2$ are unknown, estimates can be obtained by ML/REML

$$\hat{\beta}_{\lambda} = \left[ \hat{\beta}_{F}, \hat{\gamma} \right] = \left( X^* X^* + \frac{\hat{\sigma}_\varepsilon^2}{\hat{\sigma}_\gamma^2} D \right)^{-1} X^* Y$$

$\Rightarrow \hat{\beta}_{\lambda}$ is *Empirical BLUP (EBLUP)* for $\beta$
P-splines Estimator as EBLUP

If $\sigma^2_\gamma, \sigma^2_\varepsilon$ are unknown, estimates can be obtained by ML/REML

$$\hat{\beta}_\lambda = \begin{bmatrix} \hat{\beta}^F \\ \hat{\gamma} \end{bmatrix} = \left( X^*'X^* + \frac{\hat{\sigma}^2_\varepsilon}{\hat{\sigma}^2_\gamma} D \right)^{-1} X^*'Y$$

$\Rightarrow \hat{\beta}_\lambda$ is *Empirical BLUP (EBLUP)* for $\beta$

- Smoothing penalty $\lambda = \frac{\hat{\sigma}^2_\varepsilon}{\hat{\sigma}^2_\gamma}$ is determined by data
- Automatically adjusts $\lambda$ to “patterns” in data
  - small deviations from parametric shape $\rightarrow \hat{\sigma}^2_\gamma$ small $\rightarrow$ more smoothing
  - data exhibit significant deviations from parametric shape $\rightarrow \hat{\sigma}^2_\gamma$ large $\rightarrow$ less smoothing

(Wand, 2003)
Nonparametric Small Area Model

Combine both random effects models

\[ y_i = m_K(x_i) + d_i u + \varepsilon_i \]
\[ = x_i F \beta^F + z_i \gamma + d_i u + \varepsilon_i \]

Variance components

\[ \gamma \sim \text{iid } F_{\gamma}(0,\sigma_\gamma^2) \]
\[ u \sim \text{iid } F_{u}(0,\sigma_u^2) \]
\[ \varepsilon_i \sim F_{\varepsilon}(0,\sigma_\varepsilon^2) \]
Nonparametric Small Area Model

Combine both random effects models

\[ y_i = m_K(x_i) + d_iu + \varepsilon_i \]
\[ = x_i^F \beta^F + z_i\gamma + d_iu + \varepsilon_i \]

Variance components

\[ \gamma \sim \text{iid } F_{\gamma}(0, \sigma_\gamma^2) \]
\[ u \sim \text{iid } F_u(0, \sigma_u^2) \]
\[ \varepsilon_i \sim F_\varepsilon(0, \sigma_\varepsilon^2) \]

EBLUP can be computed by (RE)ML, and

\[ \bar{y}_t = \bar{x}_i^F \hat{\beta}^F + \bar{z}_t\hat{\gamma} + \hat{u}_t \]
Inference for Nonparametric Small Area Estimation

• What is right target?
  1. full prediction MSE: $\mathbb{E}((\hat{y}_t - \bar{y}_t)^2$
  2. full ridge regression: $\mathbb{E}((\hat{y}_t - \bar{y}_t)|\gamma, u)^2$
  3. prediction MSE conditional on spline: $\mathbb{E}((\hat{y}_t - \bar{y}_t)|\gamma)^2$
Inference for Nonparametric Small Area Estimation

• What is right target?
  1. full prediction MSE: $\mathbb{E}(\hat{y}_t - \bar{y}_t)^2$
  2. full ridge regression: $\mathbb{E}(\hat{y}_t - \bar{y}_t|\gamma, \mathbf{u})^2$
  3. prediction MSE conditional on spline: $\mathbb{E}(\hat{y}_t - \bar{y}_t|\gamma)^2$

• No clear winner:
  1. spline mean function is fixed, not random
  2. small areas too numerous to treat as fixed
  3. complicated (?)

• Asymptotic approximation can be derived for all 3 under mixed model specification
Inference for Nonparametric Small Area Estimation (2)

• Inference about variance components

1. $H_0 : \sigma^2_\gamma = 0$ versus $H_a : \sigma^2_\gamma > 0$
2. $H_0 : \sigma^2_u = 0$ versus $H_a : \sigma^2_u > 0$
3. $H_0 : \sigma^2_\gamma = \sigma^2_u = 0$ versus $H_a : \sigma^2_\gamma > 0$ or $\sigma^2_u > 0$

• Existing results:
  – asymptotic distribution of likelihood ratio for parameter on boundary: Self and Liang (1987)

⇒ neither one applies here...

• Develop parametric bootstrap approach
Spatial Smoothing using P-splines

- NE Lakes auxiliary variable is location: requires bivariate (spatial) smoothing
- Low-rank radial basis (≈ thin-plate spline)

\[ z = \left[ C(x_i - \kappa_k) \right]_{1 \leq i \leq n} \left[ C(\kappa_k - \kappa_{k'}) \right]_{1 \leq k, k' \leq K}^{-1/2} \]

with \( C(r) = ||r||^2 \log ||r|| \) (Ruppert et al. 2003)

- Mixed model

\[ y_i = x_i^F \beta^F + z_i \gamma + \varepsilon_i \]

\[ \gamma = (\gamma_1, \ldots, \gamma_K) \sim \text{iid } \mathcal{F}_\gamma(0, \sigma_\gamma^2) \]

\[ \varepsilon_i \sim \mathcal{F}_\varepsilon(0, \sigma_\varepsilon^2) \]

- Knot selection: regular spatial grid, space-filling algorithm
5. Small Area Estimation for Lakes Survey

- 557 measurements on 338 lakes
- Dependent variable
  - ANC | Acid Neutralizing Capacity
- Independent variables
  - Fixed effects
    - INT | intercept
    - ELEV | elevation
  - Random effects
    - TPS | spatial thin-plate spline for $K = 81$
    - HUC | 113 small areas
Knot Locations for Spatial Spline

Algorithm: funfits() in R/S-Plus
Full Model: Model Fit

- Estimates
  - Fixed effects
    
    | \( \hat{\beta}^F \) |
    |---------------------|
    | INT 586             |
    | ELEV -0.74          |

  - Random effects
    
    | \( \hat{\sigma} \) |
    |---------------------|
    | TPS 79              |
    | HUC 420             |
    | Error 173           |

- Correlation between model predictions and ANC: 0.96
Full Model: HUC Predictions

Correlation between HUC means and model predictions: 0.97
Are both random effects needed?

- AIC model selection criterion

<table>
<thead>
<tr>
<th></th>
<th>HUC</th>
</tr>
</thead>
<tbody>
<tr>
<td>yes</td>
<td>7755</td>
</tr>
<tr>
<td>no</td>
<td>7968</td>
</tr>
<tr>
<td>TPS</td>
<td>7933</td>
</tr>
<tr>
<td>no</td>
<td>8497</td>
</tr>
</tbody>
</table>

- Correlation between ANC and model prediction

<table>
<thead>
<tr>
<th></th>
<th>HUC</th>
</tr>
</thead>
<tbody>
<tr>
<td>yes</td>
<td>0.96</td>
</tr>
<tr>
<td>no</td>
<td>0.90</td>
</tr>
<tr>
<td>TPS</td>
<td>0.90</td>
</tr>
<tr>
<td>no</td>
<td>0.19</td>
</tr>
</tbody>
</table>
Spline model provides better predictions for “empty” HUCs
Small Area Random Effect in Spatial Model?

HUC effect results in predictions that are closer to observed data.
6. Conclusions

- P-spline regression is promising and flexible new tool in smoothing applications
  - full theoretical development still lacking
- Mixed model formulation allows easy incorporation into existing small area estimation techniques
- To do:
  - tests for significance of random effects

Contact information:
- jopsomer@iastate.edu
- http://www.public.iastate.edu/~jopsomer/home.html