Nonparametric Small-area Estimation Using Penalized Splines

Jean Opsomer

Iowa State University

Joint work with
Jay Breidt, Colorado State University
Gerda Claeskens, Katholieke Universiteit Leuven,
Giovanna Ranalli, Università di Perugia

August 8, 2004
Outline

1. Introduction
2. Nonparametric regression using penalized splines
3. Nonparametric small area estimator
4. Northeastern Lakes Survey
1. Introduction

Ecological condition survey of Northeastern lakes conducted by U.S. EPA

557 observations for whole region
Small Area Problem

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

Goal: improve estimates of HUC mean Acid Neutralizing Capacity (ANC)
2. Nonparametric Regression Using Penalized Splines

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

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

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

- \( p \) : 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

\[(x - \kappa)^p_+ \equiv \begin{cases} 
(x - \kappa)^p & \text{if } x - \kappa > 0 \\
0 & \text{if } x - \kappa \leq 0
\end{cases}\]
Fitting by Penalized Splines Regression

Minimize penalized sum of squares

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

$$
\Rightarrow \hat{\beta}_\lambda = (X^*^T X^* + \lambda D)^{-1} X^*^T Y
$$

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

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

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

$\hat{\beta}_\lambda$ 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
P-splines as Random Effects

\[ y_i = m(x_i; \beta) + \varepsilon_i \]
\[ = x_i \beta^F + z_i \gamma + \varepsilon_i \]

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

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

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

Assuming $\sigma^2_\gamma, \sigma^2_\varepsilon$ known, Best Linear Unbiased Estimator/Predictor (BLUE/BLUP) is solution to

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

(Henderson et al., 1959)

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

with

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

$$
X^* = [\mathbf{x}^F \ z]
$$

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

If $\sigma_\gamma^2, \sigma_\varepsilon^2$ are unknown, estimates can be obtained by Maximum Likelihood (ML) or Restricted Maximum Likelihood (REML)

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

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

If $\sigma^2$, $\sigma^2_\gamma$ are unknown, estimates can be obtained by Maximum Likelihood (ML) or Restricted Maximum Likelihood (REML)

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

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

- Smoothing penalty $\hat{\lambda} = \frac{\hat{\sigma}^2_\varepsilon}{\hat{\sigma}^2_\gamma}$ is determined by data
- Penalty choice and model fit performed at same time
- Readily implemented in PROC MIXED or lme()
3. Nonparametric Small Area Estimation

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

- \( T \) small areas, with \( u_t \) the random effect for small area \( t = 1, \ldots, T \)

- Model

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

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

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

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

- Target \( \bar{Y}_t = \bar{X}_t\beta + u_t \) estimated by (E)BLUP \( \bar{y}_t = \bar{X}_t\hat{\beta} + \hat{u}_t \)
Nonparametric Small Area Model

Combine both random effects models

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

Variance components

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

Combine both random effects models

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

Variance components

\[ \gamma \sim \text{iid } \mathcal{F}_\gamma(0, \sigma^2_\gamma) \]
\[ u \sim \text{iid } \mathcal{F}_u(0, \sigma^2_u) \]
\[ \varepsilon_i \sim \mathcal{F}_\varepsilon(0, \sigma^2_\varepsilon) \]

EBLUP easily computed by REML

\[ \bar{y}_t = \bar{X}_t^F \hat{\beta}^F + \bar{Z}_t \hat{\gamma} + \hat{u}_t \]
4. Northeastern Lakes Survey Results

- Dependent variable
  - ANC Acid Neutralizing Capacity

- Independent variables
  - Fixed effects
    - INT intercept
    - ELEV elevation
  - Random effects
    - TPS thin-plate (spatial) spline for $K = 81$
    - HUC 113 small areas
Full Model: Model Fit

- Estimates
  - Fixed effects
    - INT: $\hat{\beta}^F = 586$
    - ELEV: $\hat{\beta}^F = -0.74$
  - Random effects
    - TPS: $\hat{\sigma} = 79$
    - HUC: $\hat{\sigma} = 420$
    - Error: $\hat{\sigma} = 173$

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

-106.4 - 300.0
300.1 - 600.0
600.1 - 900.0
900.1 - 1200.0
1200.1 - 2727.0
Are both random effects needed?

Correlation between ANC and model prediction

<table>
<thead>
<tr>
<th></th>
<th>HUC</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>yes</td>
<td>no</td>
<td></td>
</tr>
<tr>
<td>TPS</td>
<td>yes</td>
<td>0.96</td>
<td>0.90</td>
</tr>
<tr>
<td></td>
<td>no</td>
<td>0.90</td>
<td>0.19</td>
</tr>
</tbody>
</table>
Spline in Small Area Model?

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.
5. Conclusions

- Nonparametric methods can improve small area estimation
  - better overall model fit
  - improved prediction for small areas without observations
- To do:
  - inference (prediction MSE)
  - test for significance of random effects

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