EPSRC

Engineering and Physical Sciences
Research Council

UNIVERSITY OF BATH

Spatial confounding and spatial+ for non-linear covariate effects

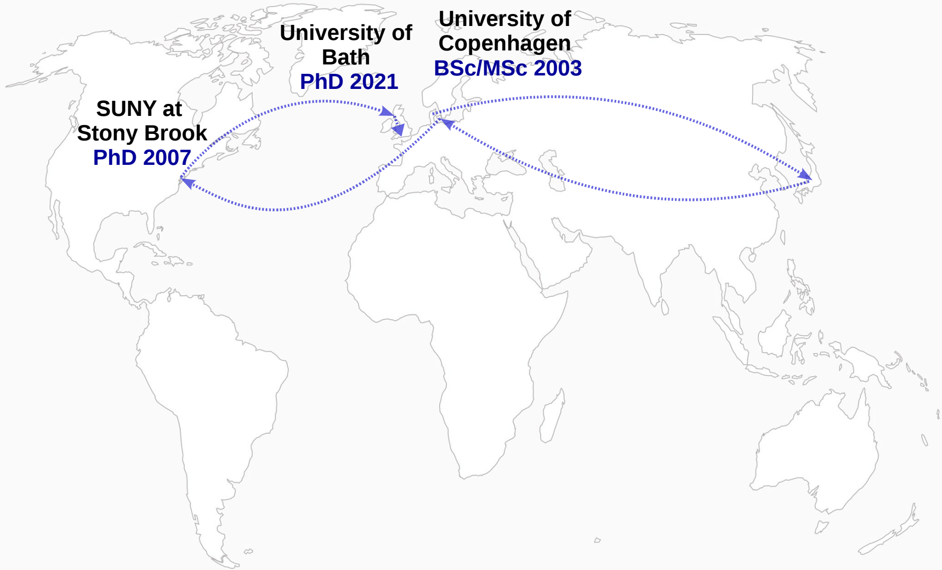
Dansk Statistisk Selskab todages møde

9 May 2023

Emiko Dupont, University of Bath

Nicole Augustin, University of Edinburgh

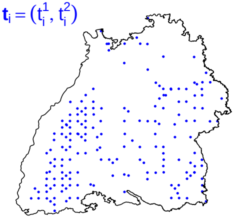
My background



Spatial confounding and spatial+ for non-linear covariate effects

1) Spatial confounding and spatial+ (linear effects)
2) Generalized Additive Models (GAMs)
3) Spatial confounding and GAMs

What is spatial confounding?



Response data: $\mathbf{y} = (y_1, \dots, y_n)^T$
Covariate data: $\mathbf{x} = (x_1, \dots, x_n)^T$

Data locations: $\mathbf{t}_1, \dots, \mathbf{t}_n \in \mathbb{R}^d$

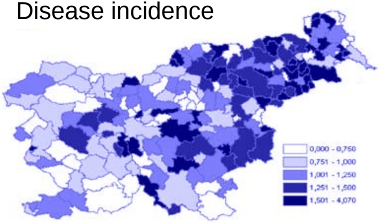
Spatial regression model:

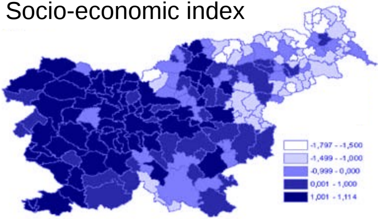
$$y_i = \beta x_i + \text{ spatial random effects } + \epsilon_i, \quad \epsilon_i \underset{\text{iid}}{\sim} N(0, \sigma^2)$$

Interference

Example: stomach cancer incidence in Slovenia

Reich et al., *Biometrics,* 2006


Disease incidence


Socio-economic index

Null model

$$y_i \quad = \quad \alpha \quad + \quad \beta\, x_i \quad + \quad \varepsilon_i$$

Disease incidence

Socio-economic index

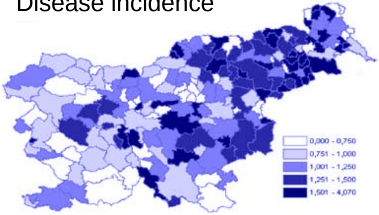
Random noise

→ Clear negative effect ($\hat{\beta} < 0$)

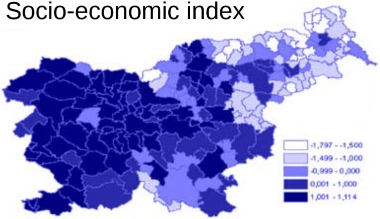
Example: stomach cancer incidence in Slovenia

Reich et al., *Biometrics,* 2006

Disease incidence



Socio-economic index



Null model

$$y_i = \alpha + \beta x_i + \varepsilon_i$$

→ Clear negative effect ($\hat{\beta} < 0$)

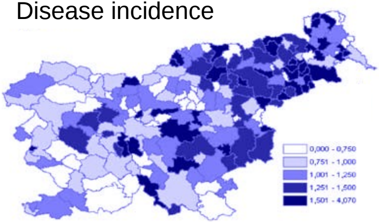
Spatial model

$$y_i = \alpha + \beta x_i + \underset{\text{Spatial effects}}{u_i} + \varepsilon_i$$

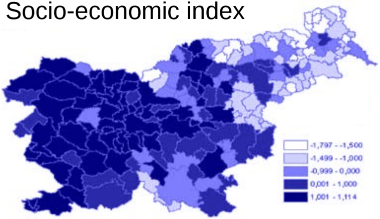
Example: stomach cancer incidence in Slovenia

Reich et al., *Biometrics,* 2006

Disease incidence



Socio-economic index



Null model

$$y_i = \alpha + \beta x_i + \varepsilon_i$$

→ Clear negative effect ($\hat{\beta} < 0$)

Spatial model

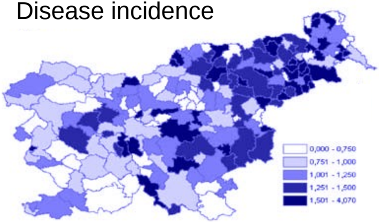
$$y_i = \alpha + \beta x_i + \underbrace{u_i}_{\text{Spatial effects}} + \varepsilon_i$$

→ No significant effect

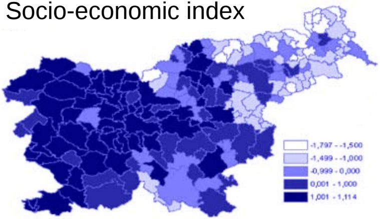
Example: stomach cancer incidence in Slovenia

Reich et al., *Biometrics,* 2006

Disease incidence



Socio-economic index



Null model

$$y_i = \alpha + \beta x_i + \varepsilon_i$$

→ Clear negative effect ($\hat{\beta} < 0$)

?

Spatial model

$$y_i = \alpha + \beta x_i + u_i + \varepsilon_i$$

→ No significant effect

Restricted spatial regression (RSR)

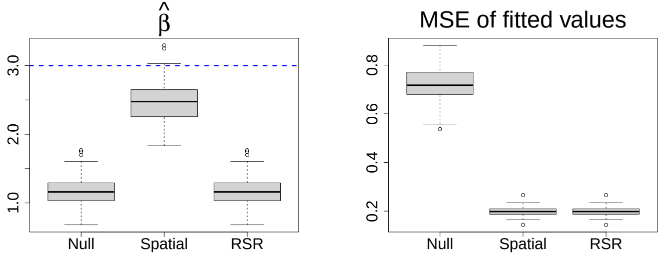
$$y_i = \alpha + \beta x_i + \tilde{u}_i + \varepsilon_i$$

→ Same $\hat{\beta}$ as linear model

Restricted spatial effects

Simulations

Data: $$\mathbf{y} = \underbrace{\beta\mathbf{x}}_{\beta = 3} + \mathbf{u} + \boldsymbol{\epsilon}^y \text{ where } \boldsymbol{\epsilon}^y \sim N(\mathbf{0}, \sigma_y^2\mathbf{I})$$



Null model/RSR

Data: $\mathbf{y} = \beta \mathbf{x} + \mathbf{u} + \epsilon^y, \quad \epsilon^y \sim N(\mathbf{0}, \sigma_y^2 \mathbf{I})$

Null model: $\mathbf{y} = \beta \mathbf{x} + \epsilon, \quad \epsilon \sim N(\mathbf{0}, \sigma^2 \mathbf{I})$

Estimate:

$$\begin{aligned} \hat{\beta} &= (\mathbf{x}^T \mathbf{x})^{-1} \mathbf{x}^T \mathbf{y} \\ &= (\mathbf{x}^T \mathbf{x})^{-1} \mathbf{x}^T (\beta \mathbf{x} + \mathbf{u} + \epsilon^y) \end{aligned}$$

$$\mathrm{E}(\hat{\beta}) = \beta + \underbrace{\mathrm{E}\left((\mathbf{x}^T \mathbf{x})^{-1} \mathbf{x}^T \mathbf{u}\right)}_{\text{Bias}}$$

Spatial+

Step 1: Fit a spatial model to the covariate

$$\mathbf{x} = \underbrace{\mathbf{u}^x}_{\substack{\text{spatial}\\\text{effects}}} + \quad \boldsymbol{\epsilon}^x, \quad \boldsymbol{\epsilon}^x \sim N(\mathbf{0}, \sigma_x^2\mathbf{I})$$

$$\Longrightarrow \qquad \mathbf{x} = \widehat{\mathbf{x}} + \mathbf{r}^x$$

Step 2: Replace $\mathbf{x}$ by $\mathbf{r}^x$ in the spatial model

$$\mathbf{y} = \beta\mathbf{r}^x + \mathbf{u} + \boldsymbol{\epsilon}, \quad \boldsymbol{\epsilon} \sim N(\mathbf{0}, \sigma^2\mathbf{I})$$

Spatial+

Step 1: Fit a spatial model to the covariate

$$\mathbf{x} = \underbrace{\mathbf{u}^x}_{\substack{\text{spatial}\\\text{effects}}} + \quad \boldsymbol{\epsilon}^x, \quad \boldsymbol{\epsilon}^x \sim N(\mathbf{0}, \sigma_x^2 \mathbf{I})$$

$$\implies \quad \mathbf{x} = \widehat{\mathbf{x}} + \mathbf{r}^x$$

Step 2: Replace $\mathbf{x}$ by $\mathbf{r}^x$ in the spatial model

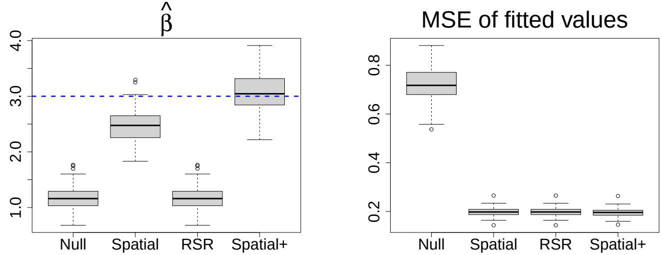
$$\mathbf{y} = \beta \mathbf{r}^x + \mathbf{u} + \boldsymbol{\epsilon}, \quad \boldsymbol{\epsilon} \sim N(\mathbf{0}, \sigma^2 \mathbf{I})$$

Idea

- Spatial dependence of $\mathbf{x}$ causes collinearity problems
- $\mathbf{x} = \widehat{\mathbf{x}} + \mathbf{r}^x \implies \beta\mathbf{x} = \beta\widehat{\mathbf{x}} + \beta\mathbf{r}^x$
- $\mathbf{r}^x$ is broadly independent of spatial location

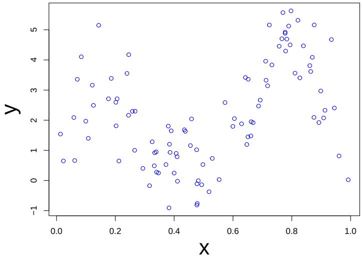
Simulations

Data: $$\mathbf{y} = \underbrace{\beta \mathbf{x}}_{\beta = 3} + \mathbf{u} + \boldsymbol{\epsilon}^y \text{ where } \boldsymbol{\epsilon}^y \sim N(\mathbf{0}, \sigma_y^2 \mathbf{I})$$



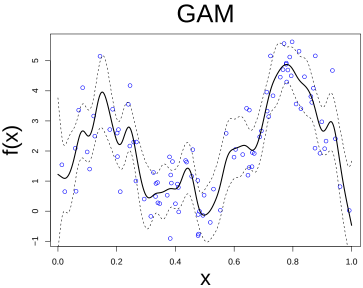
Generalized Additive Models (GAMs)

Response data: $\mathbf{y} = (y_1, \ldots, y_n)$
Covariate data: $\mathbf{x} = (x_1, \ldots, x_n)$

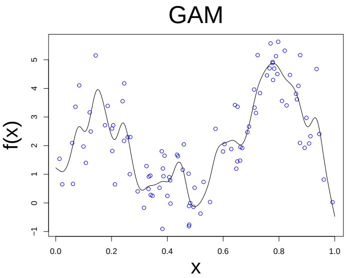


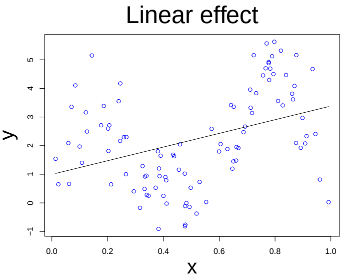
Generalized Additive Models (GAMs)

Response data: $\mathbf{y} = (y_1, \ldots, y_n)$
Covariate data: $\mathbf{x} = (x_1, \ldots, x_n)$



Generalized Additive Models (GAMs)


GAM


Linear effect

GAM: $y_i = f(x_i) + \epsilon_i$

Penalised ML: $\hat{f}$ minimises

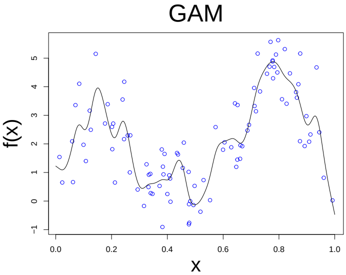
$$\underbrace{\sum_{i=1}^{n}(y_i - f(x_i))^2}_{\text{distance to data}} + \underbrace{n\lambda \int |f''(x)|^2 dx}_{\substack{\text{smoothing}\\\text{penalty}\\(\lambda>0)}}$$

Linear effect: $y_i = \alpha + \beta x_i + \epsilon_i$

ML/LLS: $\hat{\alpha}, \hat{\beta}$ minimise

$$\underbrace{\sum_{i=1}^{n}(y_i - \alpha - \beta x_i)^2}_{\text{distance to data}}$$

Generalized Additive Models (GAMs)

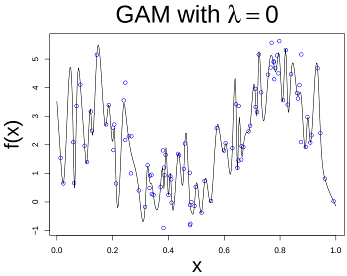


GAM: $y_i = f(x_i) + \epsilon_i$

Penalised ML: $\hat{f}$ minimises

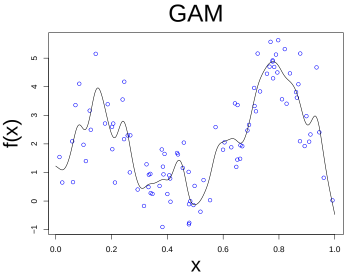
$$\underbrace{\sum_{i=1}^{n}(y_i - f(x_i))^2}_{\text{distance to data}} + \underbrace{n\lambda\int|f''(x)|^2dx}_{\substack{\text{smoothing}\\ \text{penalty}\\ (\lambda>0)}}$$

Smoothing penalty:



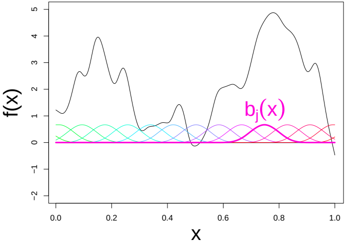
$\lambda > 0$ estimated

Generalized Additive Models (GAMs)



Basis expansion:

$$f(x) = \sum_{j=1}^{p} \beta_j \underbrace{b_j(x)}_{\text{basis functions}}$$

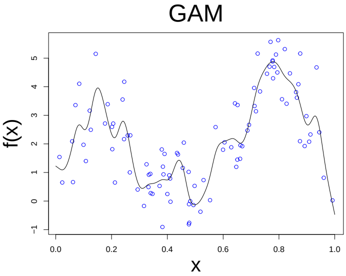


GAM: $y_i = f(x_i) + \epsilon_i$

Penalised ML: $\hat{f}$ minimises

$$\underbrace{\sum_{i=1}^{n}(y_i - f(x_i))^2}_{\text{distance to data}} + \underbrace{n\lambda \int |f''(x)|^2 dx}_{\text{smoothing penalty } (\lambda>0)}$$

Generalized Additive Models (GAMs)



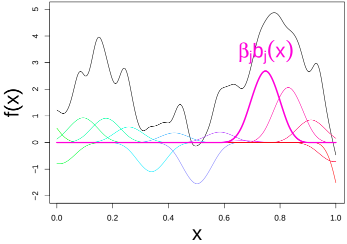
Basis expansion:

$$f(x) = \sum_{j=1}^{p} \beta_j \underbrace{b_j(x)}_{\substack{\text{basis}\\\text{functions}}}$$

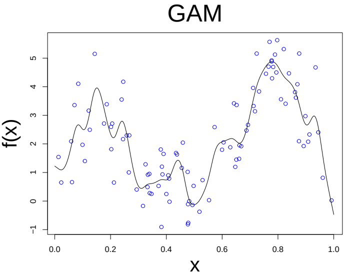
GAM: $y_i = f(x_i) + \epsilon_i$

Penalised ML: $\hat{f}$ minimises

$$\underbrace{\sum_{i=1}^{n}(y_i - f(x_i))^2}_{\text{distance to data}} + \underbrace{n\lambda \int |f''(x)|^2 dx}_{\substack{\text{smoothing}\\\text{penalty}\\(\lambda>0)}}$$



Generalized Additive Models (GAMs)



GAM: $y_i = f(x_i) + \epsilon_i$

Penalised ML: $\hat{f}$ minimises

$$\underbrace{\sum_{i=1}^{n} (y_i - f(x_i))^2}_{\text{distance to data}} + \underbrace{n\lambda \int |f''(x)|^2 dx}_{\text{smoothing penalty } (\lambda>0)}$$

Basis expansion:

$$f(x) = \sum_{j=1}^{p} \beta_j \underbrace{b_j(x)}_{\text{basis functions}}$$

Penalised LLS:

$\widehat{\boldsymbol{\beta}} = (\hat{\beta}_1, \ldots, \hat{\beta}_p)$ minimises

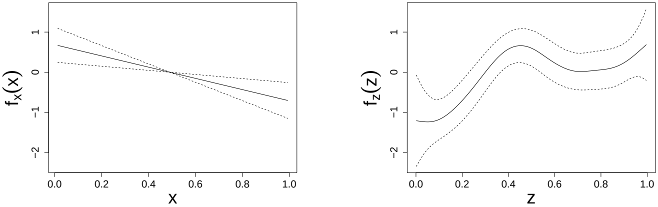
$$\|\mathbf{y} - \mathbf{B}\boldsymbol{\beta}\|^2 + \lambda \boldsymbol{\beta}^T \mathbf{S} \boldsymbol{\beta}$$

Model matrix $\mathbf{B} = \left[\mathbf{b}_1 \mid \cdots \mid \mathbf{b}_p\right]$
Penalty matrix $\mathbf{S}$

Generalized Additive Models (GAMs)

Response data: $\mathbf{y} = (y_1, \ldots, y_n)$
Covariate data: $\mathbf{x} = (x_1, \ldots, x_n)$, $\mathbf{z} = (z_1, \ldots, z_n)$

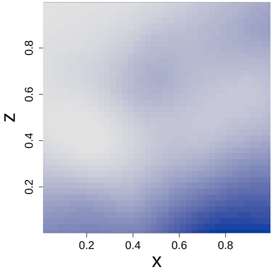
GAM: $$y_i = f_x(x_i) + f_z(z_i) + \epsilon_i, \quad \epsilon_i \underset{\text{iid}}{\sim} N(0, \sigma^2)$$



Generalized Additive Models (GAMs)

Response data: $\mathbf{y} = (y_1, \ldots, y_n)$
Covariate data: $\mathbf{x} = (x_1, \ldots, x_n)$, $\mathbf{z} = (z_1, \ldots, z_n)$

GAM: $y_i = f(x_i, z_i) + \epsilon_i, \quad \epsilon_i \underset{\text{iid}}{\sim} N(0, \sigma^2)$



$f(x, z)$

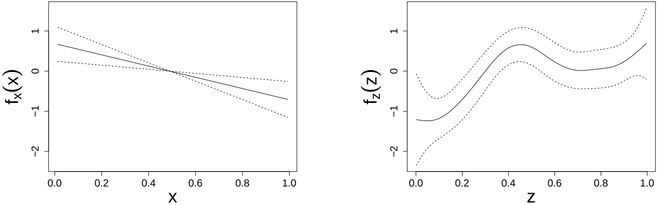
Generalized Additive Models (GAMs)

Response data: $\mathbf{y} = (y_1, \dots, y_n)$
Covariate data: $\mathbf{x} = (x_1, \dots, x_n)$, $\mathbf{z} = (z_1, \dots, z_n)$

GAM:

$$y_i \sim \mathrm{EF}(\mu_i, \phi)$$

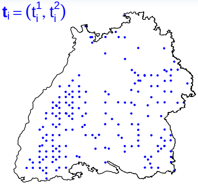
$$g(\mu_i) = f_x(x_i) + f_z(z_i)$$



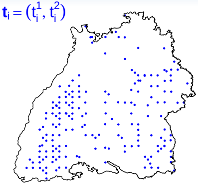
Spatial confounding and GAMs

Response data: $\mathbf{y} = (y_1, \ldots, y_n)^T$
Covariate data: $\mathbf{x} = (x_1, \ldots, x_n)^T$

Data locations: $\mathbf{t}_1, \ldots, \mathbf{t}_n \in \mathbb{R}^d$



Spatial confounding and GAMs

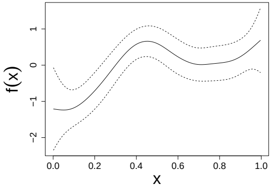


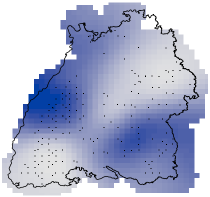
Response data: $\mathbf{y} = (y_1, \ldots, y_n)^T$
Covariate data: $\mathbf{x} = (x_1, \ldots, x_n)^T$

Data locations: $\mathbf{t}_1, \ldots, \mathbf{t}_n \in \mathbb{R}^d$

Null GAM: $y_i = f(x_i) + \epsilon_i, \quad \epsilon_i \underset{\text{iid}}{\sim} N(0, \sigma^2)$

Spatial GAM: $y_i = f(x_i) + \underbrace{f_{\text{sp}}(\mathbf{t}_i)}_{\substack{\text{spatial}\\ \text{effects}}} + \epsilon_i, \quad \epsilon_i \underset{\text{iid}}{\sim} N(0, \sigma^2)$

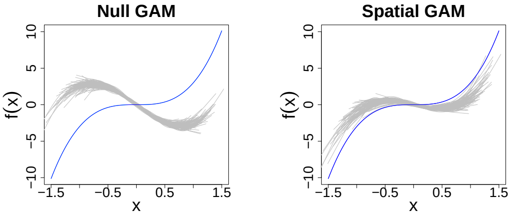


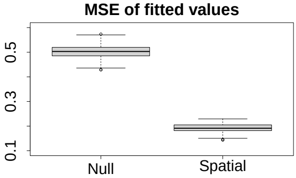


$f_{sp}(\mathbf{t})$

Spatial confounding and GAMs

Data:
$$y_i = \underbrace{f(x_i)}_{f(x) = 3x^3} + f_{\text{sp}}(\mathbf{t}_i) + \epsilon_i^y \text{ where } \epsilon_i^y \sim N(0, \sigma_y^2)$$





Spatial+

Spatial model: $\mathbf{y} = \beta\mathbf{x} + \mathbf{u} + \boldsymbol{\epsilon}, \quad \boldsymbol{\epsilon} \sim N(\mathbf{0}, \sigma^2\mathbf{I})$

Spatial+: $\mathbf{y} = \beta\mathbf{r}^x + \mathbf{u} + \boldsymbol{\epsilon}, \quad \boldsymbol{\epsilon} \sim N(\mathbf{0}, \sigma^2\mathbf{I})$

Idea

- $\mathbf{x} = \widehat{\mathbf{x}} + \mathbf{r}^x \implies \beta\mathbf{x} = \beta\widehat{\mathbf{x}} + \beta\mathbf{r}^x$

Spatial GAM: $y_i = f(x_i) + f_{\text{sp}}(\mathbf{t}_i) + \epsilon_i, \quad \epsilon_i \sim N(0, \sigma^2)$

Issue

- $\mathbf{x} = \widehat{\mathbf{x}} + \mathbf{r}^x$ but $f(x_i) \xcancel{=} f(\hat{x}_i) + f(r_i^x)$

Spatial+

Spatial model: $\mathbf{y} = \beta \mathbf{x} + \mathbf{u} + \boldsymbol{\epsilon}, \quad \boldsymbol{\epsilon} \sim N(\mathbf{0}, \sigma^2 \mathbf{I})$

Spatial+: $\mathbf{y} = \beta \mathbf{r}^x + \mathbf{u} + \boldsymbol{\epsilon}, \quad \boldsymbol{\epsilon} \sim N(\mathbf{0}, \sigma^2 \mathbf{I})$

Idea

- $\mathbf{x} = \widehat{\mathbf{x}} + \mathbf{r}^x \implies \beta \mathbf{x} = \beta \widehat{\mathbf{x}} + \beta \mathbf{r}^x$

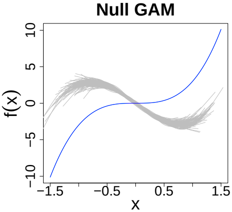
Spatial GAM: $y_i = f(x_i) + f_{\text{sp}}(\mathbf{t}_i) + \epsilon_i, \quad \epsilon_i \sim N(0, \sigma^2)$

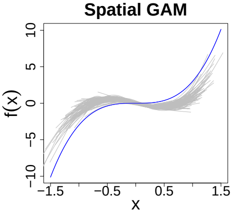
Idea

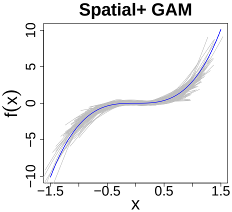
- Use basis expansion $f(x) = \sum_{j=1}^{p} \beta_j b_j(x)$
  - Model matrix $\mathbf{B} = \left[\mathbf{b}_1 \mid \cdots \mid \mathbf{b}_p\right]$
  - $\mathbf{b}_j = \hat{\mathbf{b}}_j + \mathbf{r}_j \implies \beta_j \mathbf{b}_j = \beta_j \hat{\mathbf{b}}_j + \beta_j \mathbf{r}_j$
  - Replace $\mathbf{b}_j$ by $\mathbf{r}_j$ in $\mathbf{B}$

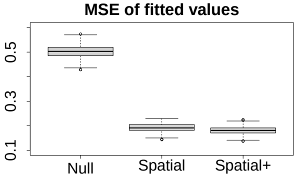
Spatial+

Data: $$y_i = \underbrace{f(x_i)}_{f(x) = 3x^3} + f_{\text{sp}}(\mathbf{t}_i) + \epsilon_i^y \text{ where } \epsilon_i^y \sim N(0, \sigma_y^2)$$


Null GAM


Spatial GAM


Spatial+ GAM


MSE of fitted values

References

- E. Dupont, and N. Augustin, *Spatial confounding and spatial+ for non-linear covariate effects.* (Under review)

- E. Dupont, N. Augustin, and S. Wood, *Spatial+: a novel approach to spatial confounding*, Biometrics (2022), 1– 12, https://doi.org/10.1111/biom.13656.

  Discussions:

  - I. Marques, and T. Kneib, doi.org/10.1111/biom.13650.
  - B. Reich, S. Yang, and Y. Guan, doi.org/10.1111/biom.13651.
  - A. Schmidt, doi.org/10.1111/biom.13654.
  - G. Papadogeorgou, doi.org/10.1111/biom.13655.

  Rejoinder:

  - E. Dupont, N. Augustin, and S. Wood, doi.org/10.1111/biom.13653.

- B. J. Reich,J. S. Hodges and V. Zadnik, *Effects of residual smoothing on the posterior of the fixed effects in disease-mapping models*, Biometrics, 62 (2006), pp. 1197–1206.

- Wood, S. N., *Generalized additive models: an introduction with R*, CRC press (2017).