Physics-informed spatial and functional data analysis over non-Euclidean domains

Laura Sangalli
MOX - Dept. of Mathematics, Politecnico di Milano

Discussant: Antonio Calcagnì
DPSS, University of Padova
GNCS, INdAM

Statistical methods and models for complex data

September, 21-23 2022

Synopsis

- Modeling of complicated spatio-temporal data by adding information regarding the (often complicated) spatial domain which might include concavities, holes, and complicated topologies

Synopsis

- Modeling of complicated spatio-temporal data by adding information regarding the (often complicated) spatial domain which might include concavities, holes, and complicated topologies

- SR-PDE: Spatial Regression + Partial Differential Equation

  > SR: effect of covariates $\mathbf{x}_i$ on the outcome variable $\mathbf{y}_i$ at the sample point $\mathbf{p}_i \in \mathcal{D} \subset \mathbb{R}^2$

  > PDE: spatial domain structure encapsulated in $f(\mathbf{p}_i)_{|\partial\mathcal{D}}$ with $\partial\mathcal{D}$ being the continuous boundaries of the domain structure $\mathcal{D}$

Synopsis

- Modeling of complicated spatio-temporal data by adding information regarding the (often complicated) spatial domain which might include concavities, holes, and complicated topologies
- SR-PDE: Spatial Regression + Partial Differential Equation
  - > SR: effect of covariates $\mathbf{x}_i$ on the outcome variable $\mathbf{y}_i$ at the sample point $\mathbf{p}_i \in \mathcal{D} \subset \mathbb{R}^2$
  - > PDE: spatial domain structure encapsulated in $f(\mathbf{p}_i)_{|\partial\mathcal{D}}$ with $\partial\mathcal{D}$ being the continuous boundaries of the domain structure $\mathcal{D}$
- The estimation problem is solved by minimizing the penalized criterion

$$\frac{1}{n}\sum_{i=1}^{n}\Big(y_i - \mathbf{x}_i\boldsymbol{\beta}^T - f(\mathbf{p}_i)\Big)^2 + \lambda\int_{\mathcal{D}}(\mathcal{L}f - u)^2\, d\mathbf{p}$$

Università degli Studi di Padova — 800 anni 1222·2022

Synopsis

$$\mathcal{H}(\mathbf{y}, \mathbf{X}, f) + \lambda \int_{\mathcal{D}} (\mathcal{L}f - u)^2 \; d\mathbf{p}$$

$$\downarrow$$

it models the spatial domain deterministically
(anisotropy + non-stationarity)

Synopsis

$$\mathcal{H}(\mathbf{y}, \mathbf{X}, f) + \lambda \int_{\mathcal{D}} (\mathcal{L}f - u)^2 \, d\mathbf{p}$$

$$\mathcal{L}f \stackrel{\text{def}}{=} -\text{div}(\mathbf{K}(\mathbf{p})\nabla f) + \nabla f \, \mathbf{b}(\mathbf{p}) + f \, c(\mathbf{p})$$

$\mathbf{K}(\mathbf{p})$ matrix governing **spatial anisotropy**

$\mathbf{b}(\mathbf{p}), c(\mathbf{p})$ vector and scalar for smoothing and shrinkage effects

UNIVERSITÀ DEGLI STUDI DI PADOVA

Synopsis

The minimization of

$$\mathcal{H}(\mathbf{y}, \mathbf{X}, f) + \lambda \int_{\mathcal{D}} (\mathcal{L}f - u)^2 \, d\mathbf{p}$$

requires **finite element methods**.

Proposal: using Delaunay triangulation on the bounded spatial domain coupled with the basis expansion of f with $\boldsymbol{\psi} = (\psi_1, \ldots, \psi_{N_{\mathcal{T}}})$ bases over the vertices $\boldsymbol{\xi} = (\xi_1, \ldots, \xi_{N_{\mathcal{T}}})$ of the triangulation.

Solution: $\hat{\mathbf{f}} = \frac{1}{n} \left(\boldsymbol{\Psi}^T \mathbf{Q} \boldsymbol{\Psi}/n + \lambda \mathbf{R}_1^T \mathbf{R}_0^{-1} \mathbf{R}_1 \right)^{-1} \boldsymbol{\Psi}^T \mathbf{Q} \mathbf{y}$

Synopsis

The minimization of

$$\mathcal{H}(\mathbf{y}, \mathbf{X}, f) + \lambda \int_{\mathcal{D}} (\mathcal{L}f - u)^2 \, d\mathbf{p}$$

requires **finite element methods**.

Proposal: using Delaunay triangulation on the bounded spatial domain coupled with the basis expansion of f with $\boldsymbol{\psi} = (\psi_1, \dots, \psi_{N_{\mathcal{T}}})$ bases over the vertices $\boldsymbol{\xi} = (\xi_1, \dots, \xi_{N_{\mathcal{T}}})$ of the triangulation.

Solution: $\hat{\mathbf{f}} = \frac{1}{n}(\boldsymbol{\Psi}^T \mathbf{Q} \boldsymbol{\Psi}/n + \lambda \mathbf{R}_1^T \mathbf{R}_0^{-1} \mathbf{R}_1)^{-1} \boldsymbol{\Psi}^T \mathbf{Q} \mathbf{y}$

$\downarrow$

discretization of the regularization term

it involves: $\boldsymbol{\Psi}, \mathbf{K}, \nabla f, \mathbf{b}, c$

Discussion

The estimator $\hat{\mathbf{f}}$ depends on the penalized term λ and, most importantly, on the assumptions about the spatial domain $\{\mathbf{K}, \mathbf{b}, c\}$ and the approximation $V_{\mathcal{T}}(\mathcal{D})$.

How could the parameters of the spatial domain approximation (i.e., $\mathbf{K}$, $\mathbf{b}$, c, $N_{\mathcal{T}}$) be chosen in advance (e.g., data-driven approach)? How do they affect the bias-variance tradeoff in the SR-PDE modeling?

Discussion

There are situations where spatial data show some levels of **spatial fuzziness** (e.g., inexact locations, vague boundaries or interiors).

A few examples: regions classified by spoken languages or dialects, marine surfaces at the intersection between two seas, brain network communities or, more generally, overlapped brain areas (e.g., do memory and learning processes share the same neuronal community?).

Discussion

The SR-PDE modeling approach might improve the analysis of situations involving fuzzy boundaries or interiors.

For instance:

- fuzzy data $(\tilde{\mathbf{p}}_1, \ldots, \tilde{\mathbf{p}}_n)$ might be modeled by representing each datum as a function $\xi_{\widetilde{\mathbf{p}}_i}(x_1, x_2)$: a fuzzy number can be represented as particular type of functional data itself. The parameter estimation can be performed by treating $V_{\mathcal{T}}$ as non-fuzzy (the same as before).

- fuzzy boundaries might be modeled via fuzzy equations directly. The parameter estimation can be performed by letting $\widetilde{V}_{\mathcal{T}}$ be fuzzy itself (e.g., via fuzzy differential equations).

antonio.calcagni@unipd.it

https://tinyurl.com/mtu3ttrm