Simulation Study

We are mainly interested in the special case of single domain with common locations. The model supposes that the randomness comes both from covariates and functions. It is our interest to see how much the mixed-effects models can divide the randomness and how well it estimates the parameters in both covariates and functions.

4.1 Model comparison

The objective of this chapter is to evaluate the effectiveness of the proposed mixed-effects models. To this end, we compare our model with a competing method, the Generalized Additive Model proposed by [17]. Our evaluation is conducted on a C-shaped domain, with a surface test function obtained from Wood (2008), originally introduced by Ramsay (2002). The C-shaped function is unique in that the inner part of the domain is physically close, yet the evaluations contrast between the upper and lower parts. We generate different test functions for each unit by modifying the original test function. Figure 4.1 depicts the test functions used in this simulation, while Figure 4.2 illustrates the mesh used for discretizing our proposed method.

The competing model employs a soap film smoother basis function [20] in the C-shaped domain, which takes extra consideration on the boundary. The soap film smoother consists of two separate types of smoother; one for the boundary and one for the film itself. Moreover, the effective degrees of freedom for both smooths are automatically selected without requiring an explicit number of basis functions.

One difference between the competing model and the proposed model is the treatment of random intercepts. The competing model applies random intercept, while the proposed model absorbs it in the functional part. Since the randomness in the constant does not automatically transfer to the functional coefficients in the competing model, it is important to include random intercepts within the model.

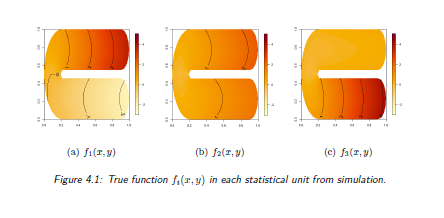
The proposed model is implemented using the R library fdaPDE [13], while the competing model is obtained from the R library mgcv [18]. Throughout the simulation, we refer to the results from each model accordingly. The usage of the R libraries fdaPDE and mgcv is detailed in Appendix B.

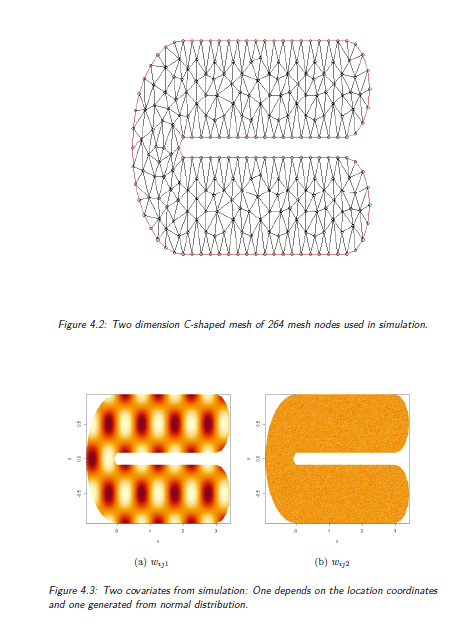
In this simulation, we generate 50 replicates with different realizations of errors e added as N(0, 5%\*range of functions). For each statistical unit i = 1 ~3, we observe n = 100 fixed locations pj for j = 1,... 100 lying on a C-shaped domain.

Formula:

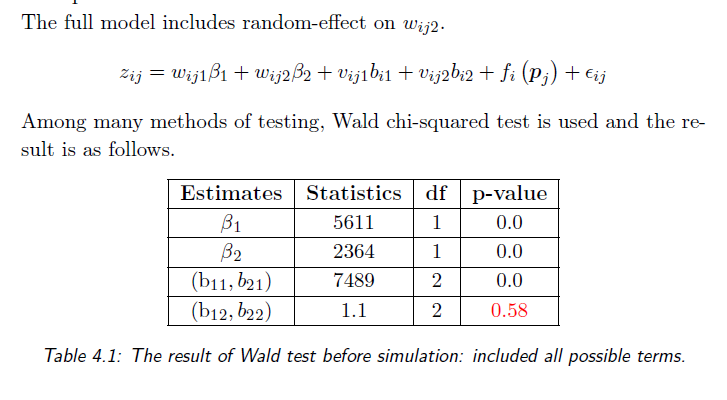
C:\Users\202207~1\AppData\Local\Temp\企业微信截图_16811934523702.png

The true parameter values are set to B1 = 3, B2 = 0.5, b11 = -5, b21 = 0, and b31 = 5. The mixed effects are applied to the first covariate wij1, so wij1 and vij1 are interchangeable. Covariates wij1 and wij2 are generated as wij1 = sin(2pipj1)cos(2pipj2) and wij2 = N(0, 4), respectively. Both covariates are the same across all statistical units. The first covariate depends on the location coordinates, while the second covariate is generated from a normal distribution. The two covariates are independent of each other. Figure 4.3 illustrates the spatially dependent covariate on the left and a realization of the random covariate on the right.





To evaluate the performance of the proposed mixed-effect model, we conduct a hypothesis test using the complete model that includes all possible terms. As the data generating process is already known, the results of the parameter inference should align with our expectations.



It is common practice to check the appropriate grid of gamma values before performing the simulation. The proposed model estimates coefficients given a fixed grid of gamma values, and the gamma value with the minimum generalized cross-validation (GCV) is chosen as the best gamma from the grid. The coefficients of the best gamma are then selected.

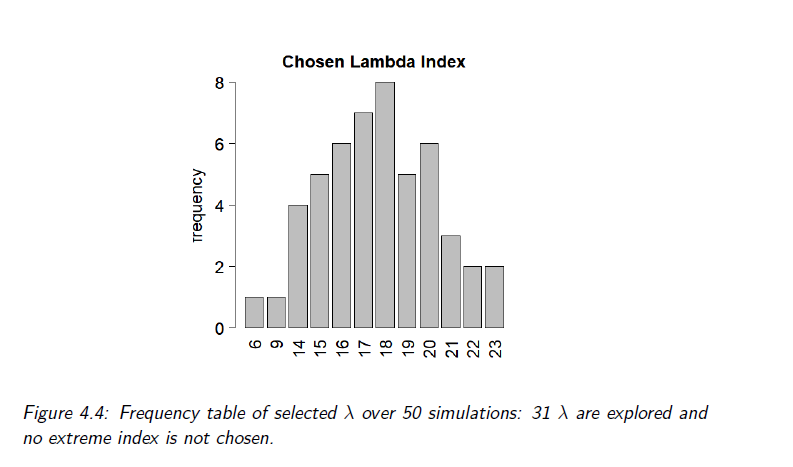
If the same gamma is chosen every time during the simulation with different repetitions of the errors, it means that the grid is not fine enough to be near the optimal gamma. On the other hand, if the grid is too fine, the method would be very slow. The grid of gamma should be big enough so that the boundary of the grid is not chosen. If the smallest or largest gamma is chosen, then the optimal gamma is smaller or bigger, so the gamma is not correctly selected.

In practice, the grid of gamma is tested through trial and error. After checking the selected gamma value, the grid should be restricted around that value. This helps to ensure that the gamma value is optimal and the simulation is efficient.

It is important to choose an appropriate grid of values for the regularization parameter gamma in the proposed model. This grid should be fine enough to be near the optimal value of gamma, but not too fine to avoid slowing down the method. Additionally, the grid should be big enough so that the boundary of the grid is not chosen. If the smallest or largest gamma value is chosen, then the optimal gamma might be smaller or larger, respectively, and the gamma value is not correctly selected.

During the simulation, it is important to check the chosen gamma value and restrict the grid around that value. If the same gamma value is chosen every time, it means that the grid is not fine enough to capture the optimal gamma value. In practice, the grid of gamma is tested using a trial and error approach. After selecting the best gamma value, the grid should be restricted around that value to ensure that the optimal gamma value is captured in subsequent simulations.

Once the best gamma value is selected, the proposed model estimates the coefficients using a fixed grid of gamma values. The gamma value with the minimum generalized cross-validation (GCV) score is chosen as the best gamma from the grid, and the coefficients of that best gamma value are selected.



It seems that the simulation results are consistent with the true parameters, as well as with the expectations from the hypothesis testing. Both fdaPDE and mgcv models provide good estimates for the fixed and mixed effects, as well as for the non-parametric part of the model. The frequency table of selected gama values also indicates that the grid of gama was appropriate for the simulation.

Overall, the simulation study provides evidence that the proposed method can be used to estimate the parameters of a mixed-effects model with non-parametric covariate functions, and that it can provide accurate estimates of the fixed and mixed effects as well as the non-parametric functions. However, it is important to note that the simulation study was conducted under specific assumptions and conditions, and the performance of the method may vary in different settings.

4.2

It is important to note that missing data can have a significant impact on the performance of statistical models. In the study described, the RMSE increased with the presence of missing data in the response variable. However, it is encouraging to see that the model was able to recover a considerable amount of information even in the presence of missing data.

Furthermore, the study also highlighted the impact of randomness in covariates on the performance of the model. When there was a high level of randomness in the covariates, the model was less accurate. However, as the level of randomness decreased, the model's accuracy improved.

Overall, the results suggest that mixed-effect models can be a useful tool for handling missing data in real-world applications. However, it is important to carefully consider the level of randomness in the covariates and to choose an appropriate model based on the specific characteristics of the data.

5 Application

5.1 Brain fMRI

Functional Magnetic Resonance Imaging (fMRI) is an imaging technique developed to detect regional, time-varying changes in brain metabolism, which can arise from task-induced cognitive state changes or unregulated processes in the resting brain. Since its inception in 1990, fMRI has become widely used in the cognitive neurosciences, clinical psychiatry and psychology, and presurgical planning, with a large number of studies available in the literature. The popularity of fMRI can be attributed to its widespread availability, non-invasive nature, relatively low cost, and good spatial resolution, as it can be performed on a clinical 1.5T scanner without requiring the injection of a radioisotope or other pharmacologic agent. Moreover, fMRI is increasingly being used as a biomarker for disease, for monitoring therapy, or for studying pharmacological efficacy. Therefore, it is important to critically evaluate the contrast mechanisms, strengths and limitations, and evolutionary trends of fMRI. The use of fMRI has significantly advanced the scientific investigation of the human mind. [refer toOverview of Functional Magnetic Resonance Imaging https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3073717/ ]

In this study, we aim to construct mixed-effect models using clinical data from multiple patients. Due to the large amount of time-series data in fMRI (1200 rows per brain location), we instead focus on constructing Functional Connectivity (FC) maps. FC maps are constructed by computing pairwise correlations between the fMRI time-series of all brain locations and the mean time-series of a Region of Interest (ROI) on the cortical surface, followed by Fisher transformation. This approach allows us to identify the areas of the brain that are functionally connected to the ROI. In this study, we use the precuneus as the ROI, which has been previously used in similar studies. We also examine the relationship between the FC map and the cortical thickness of the cerebral cortex, as the cortex is known to stimulate neural activity. The thin cerebral cortex is represented as a 2D surface embedded in a 3D space, and a specific mesh is used to analyze the spatial dataset using the finite element method.

5.1.1 fMRI Data

The previous study by [12] utilized Principal Component Analysis (PCA) to analyze fMRI data distributed over curved domains, using 69 statistical units for fPCA analysis. Our current objective is to extend this analysis by performing mixed-effect models in spatial regression, incorporating as many statistical units as possible.

The dataset used in this study comprises six variables, with observations limited to the left hemisphere. Each statistical unit is identified by a patient index, while cortical thickness at each brain location and functional connectivity (FC) map between the region of interest and each brain location are represented by the variables 'thickness' and 'connectivity', respectively. To represent the location coordinates, the three-dimensional coordinates x, y, and z are utilized.

The Conte69 brain atlas contains a large number of locations (32492) in the brain, which correspond to evaluations of the FC map and cortical thickness in each patient. Due to computational limitations, it is necessary to limit the number of patients included in the analysis. To address this issue, two different meshes are used, as shown in Figure 5.5: a full mesh with 32492 nodes and a simplified mesh with 10000 nodes. The latter allows for the inclusion of more patients while keeping the system manageable.

A mesh simplification strategy is designed specifically for spatial regression analysis over the cortical surface of the brain. It is important to note that the original locations must be projected onto the simplified mesh. Although the locations are changed, the corresponding observations remain valid. The mesh simplification algorithm takes into account various criteria, such as preserving the topology and shape of the original mesh, to ensure that the data is still suitable for analysis.

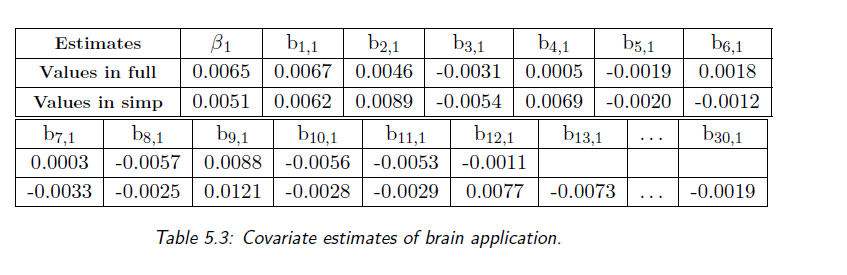
We consider a sample of m = 12 patients for the full mesh study and m = 30 for the simplified mesh study. The response variable zij represents the functional connectivity between each brain vertex and the region of interest. Cortical thickness is included as both a fixed-effect and random-effect covariate, and we aim to estimate the coefficients B1 and bi1 associated with it.

C:\Users\202207~1\AppData\Local\Temp\企业微信截图_16811979641294.png

The proportion of missing data (NA) in this application is reported to be 0.08, which is considerably lower than the proportion observed in the birds application. This suggests that the results obtained from this application would be more reliable. Appendix C provides a detailed procedure for downloading the fMRI/cortical thickness dataset, the brain atlas, and simplifying the brain mesh. The procedure requires the use of R, Matlab, AWS credential, and installation of various packages in both R and Matlab. Additionally, Appendix C.4 contains information on the patient index and corresponding ID.

5.1.2 Result

As in the previous study, gamma is chosen by minimizing the generalized cross-validation (GCV) criterion over a suitable grid. To obtain function estimates, six patients from the full mesh dataset are selected that exhibit distinctive aspects of connectivity.



Figures 5.6 and 5.7 display the estimated functions of the brain data in full mesh. The color scale in these maps ranges from red (high values) to yellow (low values). It is worth noting that these maps use the same scale for all 12 patients. The region of interest, precuneus, is situated in the upper left area of the brain, as illustrated in Figure 5.4. Given that the connectivity of this region with itself should be high, this area is colored in red in each estimated function. Although patients share the same red pattern in the precuneus, they exhibit different colors in distinct regions on Figure 5.6. While some patients exhibit significant contrast in certain areas of the brain, others lack distinctive characteristics. A few patients possess a similar pattern to that seen in Figure 5.7.

The function estimates shown in Figures 5.6 and 5.7 exhibit noticeable variations in color between different patients, indicating that there may be significant differences in brain connectivity patterns across individuals. Further analysis using additional patient-specific information could potentially provide deeper insights into these variations.

6 future work

The present implementation of the mixed-effect models in Spatial Regression is limited by its high computational cost. To make the model applicable to larger datasets, an iterative procedure algorithm should be developed in the future. In the brain application, only the estimation is performed, and inference is currently unfeasible due to memory constraints. An iterative procedure algorithm should be implemented to enable inference in this massive application, which would avoid constructing a large system and make the model flexible to add patients. While the focus has been on models sharing the same domain and location points, future work should include multiple domains with different location points based on the methodology presented in Chapter 2. This would make it possible to apply extra data with different domains and location points for each statistical unit.

The current models have a single smoothing parameter, gamma, for each statistical unit, but it is possible to develop models with different smoothing parameters for each unit, allowing for consideration of individual functional wiggliness. Additionally, if problem-specific information about the spatial field is known, such as from physics, mechanics, or chemistry, the models can be extended to include different penalty terms with general PDEs Lf = u, as demonstrated in [2] and [3]. These extensions can also incorporate anisotropy or non-stationarity, as in [4].

The methodology can also be extended to spatio-temporal data, as demonstrated in [1] and [5]. However, this extension would significantly increase computational costs and may not be feasible from an implementation standpoint. Finally, other discretization techniques, such as Non-Uniform Rational B-Splines (NURBS), can be adopted, as demonstrated in [16], which explores the use of isogeometric analysis with high smoothness.