Version of Record: https://www.sciencedirect.com/science/article/pii/S016794732200247X
Manuscript_25bbe6a94940ef16ddc63b28621e00a5

Spatial Heterogeneity Automatic Detection and Estimation

Xin Wang[a,*], Zhengyuan Zhu[b], Hao Helen Zhang[c]

[a]*Department of Statistics, Miami University, 311 Upham Hall, Oxford, 45056, OH, USA*
[b]*Department of Statistics, Iowa State University, Snedecor Hall, Ames, 50011, IA, USA*
[c]*Department of Mathematics, University of Arizona, 617 N. Santa Rita Ave, Tucson, 85721, AZ, USA*

Abstract

Spatial regression is widely used for modeling the relationship between a dependent variable and explanatory covariates. Oftentimes, the linear relationships vary across space, such that some covariates have location-specific effects on the response. One fundamental question is how to detect the systematic variation in the model and identify which locations share common regression coefficients and which do not. Only a correct model structure can assure unbiased estimation of coefficients and valid inferences. A new procedure is proposed, called Spatial Heterogeneity Automatic Detection and Estimation (SHADE), for automatically and simultaneously subgrouping and estimating covariate effects for spatial regression models. The SHADE employs a class of spatially-weighted fusion type penalty on all pairs of observations, with location-specific weight constructed using spatial information, to cluster coefficients into subgroups. Under certain regularity conditions, the SHADE is shown to be able to identify the true model structure with probability approaching one and estimate regression coefficients consistently. An alternating direction method of multiplier algorithm (ADMM) is developed to compute the SHADE. In numerical studies, the empirical performance of the SHADE is demonstrated by using different choices of weights and comparing their accuracy. The results suggest that spatial information can enhance subgroup structure analysis in challenging situations when the spatial variation among regression coefficients is small or the number of repeated measures is small. Finally, the SHADE is applied to find the relationship between a natural resource survey and a land cover data layer to identify spatially interpretable groups.

Keywords: Areal data, Structure Selection, Penalization, Repeated measures, Spatial heterogeneity, Subgroup analysis

*Corresponding author
Email addresses: wangx172@miamioh.edu (Xin Wang), zhuz@iastate.edu (Zhengyuan Zhu), hzhang@math.arizona.edu (Hao Helen Zhang)

1. Introduction

Spatial regression is commonly used to model the relationship between a response and explanatory variables. For complex problems, some covariates (we call them global covariates) may have constant effects across space, while other covariates (we call them local covariates) may have location-specific effects, i.e, their effects on the response variable vary across space. This has received wide attention in many fields, such as environmental sciences (Hu and Bradley, 2018), biology (Zhang and Lawson, 2011), social science (Bradley et al., 2018), economics (Brunsdon et al., 1996), and biostatistics (Xu et al., 2019).

A motivating example is about studying the relationship between two land-cover data sources. One is the National Resources Inventory (NRI, Nusser and Goebel 1997) survey conducted by the USDA Natural Resources Conservation Service (NRCS), the other one is the Cropland data layer (CDL, Han et al. 2012) produced by the USDA National Agricultural Statistics Service (NASS). An accurate estimate of local landcover information from NRI is essential for developing conservation policies and land management plans. However, direct estimates in small geographical areas such as at the county level may not be accurate due to small sample sizes. Auxiliary information such as CDL can be used to improve the small area estimator in NRI (Wang et al., 2018). Traditional regression models used in the small area estimation problems typically assume common regression coefficients over all domains, which may not be appropriate. For example, when we looked at the linear relationship between the NRI and CDL estimates of different types of land covers at the county level, the regression coefficients in the Mountain states are quite different from the west coast and the vast areas in the east. This is reflected in Figure 10 (a) in Section 5. One reason for that difference is due to the NRI survey's scope, which only includes non-federal land in the US. Another contributing reason is that CDL is created by training separate machine learning models at the state level using only ground observations from that state, which creates variations among states. A common regression assumption would be too simple to capture the regional differences and lead to biases in the estimators. This type of spatial heterogeneity is also known as structural instability. For linear models, this implies that the linear relationship changes geographically over space, and the linear regression coefficients may form subgroups. It is an important and challenging problem to identify the correct grouping structure of the regression coefficients, as only a correct model structure can lead to an unbiased estimation of the regression coefficients and their valid inference. In practice, ad hoc grouping of states as regions defined by tradition or for federal administrative purposes is sometimes used to address this issue. However, such grouping is not driven by the data in specific problems and may not be ap-

propriate or efficient. For example, the central region in Figure 10 (a) includes not only all the Mountain states, but also the North and South Dakotas which are not traditionally considered Mountain states. Figure 10 (b) suggests a further division of the east into sub-regions, which does not align with any well known administrative regions. One natural approach is to assume that regression coefficients of states nearby are more likely to belong to the same subgroup than states which are further apart, and use both the estimated regression coefficients and the spatial structure to guide the clustering of states into subregions, which is what we propose to do in this paper.

The problem of taking into account spatial dependence structure in linear regression has been studied for a long time in literature. Classical works include spatial expansion methods (Casetti, 1972; Casetti and Jones, 1987), which treat the spatially varying regression coefficients as a function of expansion variables, typically using longitude and latitude coordinates as location variables. One popular approach to accounting for spatial variations in the model is by introducing an additive spatial random effect for each location, as done for linear models by Cressie (2015) and generalized linear models by Diggle et al. (1998). Other classes of models in wide use are spatial varying coefficient models, including the geographically weighted regression (GWR; Brunsdon et al. 1996) and its extensions to generalized linear models (Nakaya et al., 2005) and the Cox model for survival analysis (Xue et al., 2020; Hu and Huffer, 2020). There have also been developments in the Bayesian framework, such as Gelfand et al. (2003).

The methods mentioned above typically assume that the regression parameters are smooth functions of location variables. This assumption is reasonable in certain practices, but may not be appropriate for applications where the covariate effects are constant over subregions defined by some unobserved hidden factors. In this work, we take a different perspective by grouping the covariate effects into spatially-interpretable subgroups or clusters. As in the motivating example, different clusters have different patterns, which can be used to build more flexible estimators to improve the original direct estimates. A majority of existing work in the literature on spatial cluster detection is based on hypothesis tests, including the scan statistic methods based on the likelihood ratio (Kulldorff and Nagarwalla, 1995; Jung et al., 2007; Cook et al., 2007) and the two-step spatial test methods under the GWR framework (Lee et al., 2017, 2020). Test-based methods are intuitive and useful in practice, but proper test statistics are often difficult to construct, and the tests may have low power when the number of locations is large. In addition, these methods handle the cluster detection problem and the model estimation separately, making it challenging to study the inferential properties of the final estimator. The main

purpose of this article is to fill this gap by developing a unified framework to detect clusters of regression coefficients, estimate them consistently, and make valid inferences.

In the context of non-spatial data analysis, a variety of clustering methods have been proposed to identify homogeneous groups for either observations or regression coefficients. Chi and Lange (2015) developed a method for the convex clustering problem through the alternating direction method of multiplier algorithm (ADMM) (Boyd et al., 2011) with pairwise $L_p(p \geq 1)$ penalty. Nonnegative weights are considered to reduce bias for pairwise penalties. Fan et al. (2018) considered a clustering problem with l_0 penalty on graphs. For clustering the regression coefficients, Ma and Huang (2017) and Ma et al. (2020a) proposed a concave fusion approach for estimating the group structure and estimating subgroup-specific effects, where smoothly clipped absolute deviation (SCAD) penalty (Fan and Li, 2001) and the minimax concave penalty (MCP) (Zhang, 2010) are considered. The pairwise penalty approach is also applied in partially linear models Liu and Lin (2019) and mixture models Im and Tan (2021).

For spatial analysis, some interesting work are recently proposed for grouping regression coefficients in the spatial regression. Hallac et al. (2015) considered a general network setting, and their focus was on optimization, where an ADMM based algorithm was proposed for network LASSO with a global convergence guarantee. They did not discuss statistical properties for any estimators. By contrast, our paper focuses on spatial regression and gives a comprehensive investigation of the topic, covering estimation framework, statistical theory, computation and tuning, as well as spatial applications. Both Ma et al. (2020a) and Hallac et al. (2015) considered on finding clusters based on the whole vector of regression coefficients. Ma et al. (2020b) proposed the Bayesian heterogeneity pursuit regression models to detect clusters in the covariate effects based on the Dirichlet process. Hu et al. (2020) proposed a Bayesian method for clustering coefficients with auxiliary covariates random effects, based on a mixture of finite mixtures (MFM). Li and Sang (2019) proposed a penalized approach based on the minimum spanning tree. Luo et al. (2021) generalized the work of Li and Sang (2019) using a Bayesian method and random spanning tree models, which was not based on penalty approaches. These two methods considered clusters on each covariate. In the area of spatial boundaries detection, Lu and Carlin (2005) and Lu et al. (2007) considered the areal boundary detection using a Bayesian hierarchical model based on the conditional autoregressive model (Banerjee et al., 2014). The boundaries were determined by the posterior distribution of the corresponding spatial process or spatial weights. These boundary detection methods focused on clustering of observations instead of regression coefficients.

The main difficulty in clustering spatial covariate effects is how to consistently estimate the number of clusters and cluster memberships, when properly considering spatial neighborhood information. To our knowledge, most existing methods do not rigorously prove the theoretical results and explore the choices of spatial weights. In this work, we fill the gap by proposing a new procedure, called Spatial Heterogeneity Automatic Detection and Estimation (SHADE), for automatically grouping and estimating local covariate effects simultaneously. The SHADE employs a class of spatially-weighted fusion type penalty on all pairs of observations, with location-specific weights adaptively constructed using geographical proximity of locations, and achieves spatial clustering consistency for spatial linear regression models. In theory, we show that the oracle estimator based on weighted least squares is a local minimizer of the objective function with probability approaching 1 under certain regular conditions, which indicates that the number of clusters can be estimated consistently. We believe that this result is the first of its kind in the contest of spatial data analysis. To implement the SHADE, an alternating direction method of multiplier (ADMM) algorithm is developed. To make the best choices of spatial weights and to understand their roles in practical applications, we consider different schemes to choose pairwise weights and compare them numerically and theoretically. Our numerical examples suggest that the number of clusters and the group structure can be recovered with high probability, and the spatial information can help in spatial clustering analysis when the minimal group difference is small or the number of repeated measures is small.

The article is organized as follows. In Section 2, we describe the Spatial Heterogeneity Automatic Detection and Estimation (SHADE) model and the corresponding ADMM algorithm. In Section 3, we establish the theoretical properties of the SHADE estimator. The simulation study is conducted in Section 4 under several scenarios to show the performances of the proposed estimator. The proposed method is applied to an NRI small area estimation problem to illustrate the use of SHADE in real-data applications in Section 5. Finally, some discussions are given in Section 6.

2. Methodology and Algorithm

2.1. Methodology: SHADE

Assume our spatial data consist of multiple measurements at each location or subject. Let y_{ih} be the hth response for the ith subject observed at location $\boldsymbol{s}_i$, where $i = 1, \ldots, n$, $h = 1, \ldots, n_i$. Based on their effects on the response variable, the covariates can be divided into two categories: "global" covariates which have common effects on the response across all the locations, and "local" covariates which

have location-specific effects on the response. To reflect this, let $\boldsymbol{z}_{ih}$ and $\boldsymbol{x}_{ih}$ be the corresponding covariate vectors with dimension q and p, respectively, where $\boldsymbol{z}_{ih}$'s are "global" covariates with common linear effects to the response across all the locations, while $\boldsymbol{x}_{ih}$'s are "local" covariates with location-specific linear effects on the response. We consider the following linear regression model

$$y_{ih} = \boldsymbol{z}_{ih}^T \boldsymbol{\eta} + \boldsymbol{x}_{ih}^T \boldsymbol{\beta}_i + \epsilon_{ih}, \tag{1}$$

where $\boldsymbol{\eta}$ represents the vector of common regression coefficients shared by global effects, $\boldsymbol{\beta}_i$'s are location-specific regression coefficients, and ϵ_{ih}'s are i.i.d random errors with $E(\epsilon_{ih}) = 0$ and $Var(\epsilon_{ih}) = \sigma^2$. Furthermore, some locations may have the same or similar location-specific effects, grouping locations with the same location-specific effects can help to achieve dimension reduction and improve model prediction accuracy. Assume the n location-specific effects belong to K mutually exclusive subgroups: the locations with a common $\boldsymbol{\beta}_i$ belong to the same group. Denote the corresponding partition of $\{1, \ldots, n\}$ as $\mathcal{G} = \{\mathcal{G}_1, \ldots, \mathcal{G}_K\}$, where the index set $\mathcal{G}_k$ contains all the locations belonging to the group k for $k = 1, \ldots, K$. For convenience, denote the regression coefficients associated with $\mathcal{G}_k$ as $\boldsymbol{\alpha}_k$. In practice, since neither K nor the partition $\mathcal{G}_k$'s are known, the goal is to use the observed data $\{(y_{ih}, \boldsymbol{z}_{ih}, \boldsymbol{x}_{ih})\}$ to construct the estimator $\hat{K}$ and the partition $\hat{\mathcal{G}} = \{\hat{\mathcal{G}}_1, \ldots, \hat{\mathcal{G}}_{\hat{K}}\}$, where $\hat{\mathcal{G}}_k = \{i : \hat{\boldsymbol{\beta}}_i = \hat{\boldsymbol{\alpha}}_k, 1 \leq i \leq n\}$.

To achieve this goal, we use the following optimization problem: minimize the weighted least squares objective function subject to a spatially-weighted pairwise penalty

$$Q_n(\boldsymbol{\eta}, \boldsymbol{\beta}; \lambda, \psi) = \frac{1}{2} \sum_{i=1}^{n} \frac{1}{n_i} \sum_{h=1}^{n_i} \left(y_{ih} - \boldsymbol{z}_{ih}^T \boldsymbol{\eta} - \boldsymbol{x}_{ih}^T \boldsymbol{\beta}_i \right)^2 + \sum_{1 \leq i < j \leq n} p_\gamma \left(\|\boldsymbol{\beta}_i - \boldsymbol{\beta}_j\|, c_{ij} \lambda \right), \tag{2}$$

where $\boldsymbol{\eta} = (\eta_1, \ldots, \eta_q)^T$, $\boldsymbol{\beta} = \left(\boldsymbol{\beta}_1^T, \ldots, \boldsymbol{\beta}_n^T\right)^T$, $\|\cdot\|$ denotes the Euclidean norm, $p_\gamma(\cdot, \lambda)$ is a penalty function imposed on all distinct pairs. In the penalty function, $\lambda \geq 0$ is a tuning parameter, $\gamma > 0$ is a built-in constant in the penalty function, and different weights c_{ij}'s are assigned to different pairs of locations $\boldsymbol{s}_i$ and $\boldsymbol{s}_j$ for any $1 \leq i < j \leq n$. One popular choice of penalty is the L_1 penalty (lasso) (Tibshirani, 1996) with the form $p_\gamma(t, \lambda) = \lambda |t|$. Since L_1 penalty tends to produce too many groups as shown in Ma and Huang (2017), we consider the SCAD penalty, which is defined as

$$p_\gamma(t, \lambda) = \lambda \int_0^{|t|} \min\{1, (\gamma - x/\lambda)_+ / (\gamma - 1)\} dx. \tag{3}$$

Here we treat γ as a fixed value as in Fan and Li (2001), Zhang (2010) and Ma et al. (2020a).

2.2. Choices of Spatial Weights

In (2), the values of weights c_{ij} are crucial, as they control the number of subgroups and grouping results. The pairs $\|\boldsymbol{\beta}_i - \boldsymbol{\beta}_j\|$ with larger weights $c_{ij}\lambda$ are shrunk together more than those pairs with smaller weights. For spatial data, reasonable choices of c_{ij} should take into account two factors: locations with closer $\boldsymbol{\beta}_j$ values are more likely grouped together and locations closer to each other are more likely to form a subgroup as they typically have similar trends. Since the true values of $\boldsymbol{\beta}$ are not available, we use their estimators $\tilde{\boldsymbol{\beta}}$ as the surrogates. For example, we can define the weights c_{ij} as

$$c_{ij} = \exp\left(-\psi \left\|\boldsymbol{s}_i - \boldsymbol{s}_j\right\| \cdot \left\|\tilde{\boldsymbol{\beta}}_i - \tilde{\boldsymbol{\beta}}_j\right\|\right),$$

where $\tilde{\boldsymbol{\beta}}_i$ is an initial estimate of $\boldsymbol{\beta}_i$, and ψ is a scale parameter to control the magnitudes of the weights. In areal data, we suggest three different ways of taking into account spatial information in the data to construct the weights.

(i) using both spatial and regression coefficients information:

$$c_{ij} = \exp\left(\psi \left(1 - a_{ij}\right) \cdot \left\|\tilde{\boldsymbol{\beta}}_i - \tilde{\boldsymbol{\beta}}_j\right\|\right), \tag{4}$$

where a_{ij} is the neighbor order between location $\boldsymbol{s}_i$ and location $\boldsymbol{s}_j$, which means that if i and j are neighbors, $a_{ij} = 1$. If i and j are not neighbors, but they have at least one same neighbor, $a_{ij} = 2$. Similarly, we can have all the neighborhood order for all subjects or locations.

(ii) using regression coefficients information only:

$$c_{ij} = \exp\left(-\psi \left\|\tilde{\boldsymbol{\beta}}_i - \tilde{\boldsymbol{\beta}}_j\right\|\right). \tag{5}$$

(iii) using spatial information only:

$$c_{ij} = \exp\left(\psi(1 - a_{ij})\right). \tag{6}$$

Weights in (4) and (5) both include the regression coefficients, which would depend on the accuracy of $\tilde{\boldsymbol{\beta}}_i$. If the number of repeated measures is not large, the values of $\tilde{\boldsymbol{\beta}}_i$ will not show the real relationship between different locations, which would lead to very bad weights. The phenomenon can be observed in the simulation study. The weights we use here are three special cases that use the information of

either regression coefficients or the spatial neighborhood orders. Definitely, there are other ways to construct weights, such as using distance to borrow spatial information. As long as the weights satisfy condition (C4) in Section 3, the theoretical results will hold under other conditions. For example, the weights in (5) will satisfy (C4) automatically if $\tilde{\boldsymbol{\beta}}_i$'s are consistent estimators. Besides condition (C4), there are no other conditions about the format of the weight function.

2.3. Algorithm for SHADE

In this section, we describe the ADMM algorithm to solve (2) in Section 2.1.

There are two tuning parameters, λ and ψ, in the proposed method. We choose them adaptively using some tuning procedures discussed at the end of this section. For now, we fix them and present the computational algorithm for solving (2). Denote the solution as

$$\left(\hat{\boldsymbol{\eta}}, \hat{\boldsymbol{\beta}}\right) = \underset{\boldsymbol{\eta}\in\mathbb{R}^q, \boldsymbol{\beta}\in\mathbb{R}^{np}}{\arg\min}\, Q_n\left(\boldsymbol{\eta}, \boldsymbol{\beta}, \lambda, \psi\right). \tag{7}$$

First, we introduce the slack variables for all the pairs (i, j) $\boldsymbol{\delta}_{ij} = \boldsymbol{\beta}_i - \boldsymbol{\beta}_j$, for $1 \le i < j \le n$. Then the problem is equivalent to minimizing the following objective function with regard to $(\boldsymbol{\eta}, \boldsymbol{\beta}, \boldsymbol{\delta})$,

$$\min_{\boldsymbol{\eta},\boldsymbol{\beta},\boldsymbol{\delta}} \; L_0\left(\boldsymbol{\eta}, \boldsymbol{\beta}, \boldsymbol{\delta}\right) = \frac{1}{2}\sum_{i=1}^{n}\frac{1}{n_i}\sum_{h=1}^{n_i}\left(y_{ih} - \boldsymbol{z}_{ih}^T\boldsymbol{\eta} - \boldsymbol{x}_{ih}^T\boldsymbol{\beta}_i\right)^2 + \sum_{1\le i<j\le n} p_\gamma\left(\left\|\boldsymbol{\delta}_{ij}\right\|, c_{ij}\lambda\right),$$
$$\text{subject to} \;\; \boldsymbol{\beta}_i - \boldsymbol{\beta}_j - \boldsymbol{\delta}_{ij} = \mathbf{0}, \;\; 1 \le i < j \le n,$$

where $\boldsymbol{\delta} = (\boldsymbol{\delta}_{ij}^T, 1 \le i < j \le n)^T$. To handle the equation constraints in the optimization problem, we introduce the augmented Lagrangian

$$L\left(\boldsymbol{\eta}, \boldsymbol{\beta}, \boldsymbol{\delta}, \boldsymbol{v}\right) = L_0\left(\boldsymbol{\eta}, \boldsymbol{\beta}, \boldsymbol{\delta}\right) + \sum_{i<j}\left\langle \boldsymbol{v}_{ij}, \boldsymbol{\beta}_i - \boldsymbol{\beta}_j - \boldsymbol{\delta}_{ij}\right\rangle + \frac{\vartheta}{2}\sum_{i<j}\left\|\boldsymbol{\beta}_i - \boldsymbol{\beta}_j - \boldsymbol{\delta}_{ij}\right\|^2,$$

where $\boldsymbol{v} = (\boldsymbol{v}_{ij}^T, 1 \le i < j \le n)^T$ are Lagrange multipliers and $\vartheta > 0$ is the penalty parameter.

To solve the problem, we use an iterative algorithm which updates $\boldsymbol{\beta}, \boldsymbol{\eta}, \boldsymbol{\delta}, \boldsymbol{v}$ sequentially, one at a time. At the $(m+1)$th iteration, given their current values $(\boldsymbol{\beta}^{(m)}, \boldsymbol{\eta}^{(m)}, \boldsymbol{\delta}^{(m)}, \boldsymbol{v}^{(m)})$, the updates of $\boldsymbol{\eta}, \boldsymbol{\beta}, \boldsymbol{\delta}, \boldsymbol{v}$ are

$$\begin{aligned}
\left(\boldsymbol{\eta}^{(m+1)}, \boldsymbol{\beta}^{(m+1)}\right) &= \underset{\boldsymbol{\eta},\boldsymbol{\beta}}{\arg\min}\, L\left(\boldsymbol{\eta}, \boldsymbol{\beta}, \boldsymbol{\delta}^{(m)}, \boldsymbol{v}^{(m)}\right), \\
\boldsymbol{\delta}^{(m+1)} &= \underset{\boldsymbol{\delta}}{\arg\min}\, L\left(\boldsymbol{\eta}^{(m+1)}, \boldsymbol{\beta}^{(m+1)}, \boldsymbol{\delta}, \boldsymbol{v}^{(m)}\right), \\
\boldsymbol{v}_{ij}^{(m+1)} &= \boldsymbol{v}_{ij}^{m} + \vartheta\left(\boldsymbol{\beta}_i^{(m+1)} - \boldsymbol{\beta}_j^{(m+1)} - \boldsymbol{\delta}_{ij}^{(m+1)}\right).
\end{aligned} \tag{8}$$

To update $\boldsymbol{\eta}$ and $\boldsymbol{\beta}$, we minimize the following objective function

$$f(\boldsymbol{\beta},\boldsymbol{\eta}) = \left\|\boldsymbol{\Omega}^{1/2}(\boldsymbol{y}-\boldsymbol{Z}\boldsymbol{\eta}-\boldsymbol{X}\boldsymbol{\beta})\right\|^2 + \vartheta\left\|\boldsymbol{A}\boldsymbol{\beta}-\boldsymbol{\delta}^{(m)}+\vartheta^{-1}\boldsymbol{v}^{(m)}\right\|^2,$$

where $\boldsymbol{y} = (y_{11},\ldots,y_{1n_1},\ldots,y_{n1},\ldots,y_{n,n_n})^T$, $\boldsymbol{Z} = (\boldsymbol{z}_{11},\ldots,\boldsymbol{z}_{1n_1},\ldots,\boldsymbol{z}_{n1},\ldots,\boldsymbol{z}_{n,n_n})^T$, $\boldsymbol{X} = \mathrm{diag}(\boldsymbol{X}_1,\ldots,\boldsymbol{X}_n)$ with $\boldsymbol{X}_i = (\boldsymbol{x}_{i1},\ldots,\boldsymbol{x}_{i,n_i})^T$, $\boldsymbol{\Omega} = \mathrm{diag}(1/n_1\boldsymbol{I}_{n_1},\ldots,1/n_n\boldsymbol{I}_{n_n})$ and $\boldsymbol{A} = \boldsymbol{D}\otimes\boldsymbol{I}_p$ with an $[n(n-1)/2]\times n$ matrix $\boldsymbol{D} = \{(\boldsymbol{e}_i-\boldsymbol{e}_j), i<j\}^T$, where $\boldsymbol{e}_i$ is an $n\times 1$ vector with ith element 1 and other elements 0. Then the solutions for $\boldsymbol{\beta}$ and $\boldsymbol{\eta}$ are

$$\boldsymbol{\beta}^{(m+1)} = \left(\boldsymbol{X}^T\boldsymbol{Q}_{Z,\Omega}\boldsymbol{X}+\vartheta\boldsymbol{A}^T\boldsymbol{A}\right)^{-1}\left[\boldsymbol{X}^T\boldsymbol{Q}_{Z,\Omega}\boldsymbol{y}+\vartheta\mathrm{vec}\left(\left(\boldsymbol{\Delta}^{(m)}-\vartheta^{-1}\boldsymbol{\Upsilon}^{(m)}\right)\boldsymbol{D}\right)\right], \tag{9}$$

$$\boldsymbol{\eta}^{(m+1)} = \left(\boldsymbol{Z}^T\boldsymbol{\Omega}\boldsymbol{Z}\right)^{-1}\boldsymbol{Z}^T\boldsymbol{\Omega}\left(\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta}^{(m+1)}\right), \tag{10}$$

where $\boldsymbol{\Delta}^{(m)} = \left(\boldsymbol{\delta}_{ij}^{(m)}, i<j\right)_{p\times n(n-1)/2}$, $\boldsymbol{\Upsilon}^{(m)} = \left(\boldsymbol{v}_{ij}^{(m)}, i<j\right)_{p\times n(n-1)/2}$ and

$$\boldsymbol{Q}_{Z,\Omega} = \boldsymbol{\Omega}-\boldsymbol{\Omega}\boldsymbol{Z}\left(\boldsymbol{Z}^T\boldsymbol{\Omega}\boldsymbol{Z}\right)^{-1}\boldsymbol{Z}^T\boldsymbol{\Omega}.$$

To update $\boldsymbol{\delta}_{ij}$'s componentwisely, it is equivalent to minimizing the following objective function

$$\frac{\vartheta}{2}\left\|\boldsymbol{\varsigma}_{ij}^{(m)}-\boldsymbol{\delta}_{ij}\right\|^2 + p_\gamma\left(\|\boldsymbol{\delta}_{ij}\|, c_{ij}\lambda\right),$$

where $\boldsymbol{\varsigma}_{ij}^{(m+1)} = \left(\boldsymbol{\beta}_i^{(m+1)}-\boldsymbol{\beta}_j^{(m+1)}\right)+\vartheta^{-1}\boldsymbol{v}_{ij}^{(m)}$. The solution based on SCAD penalty has a closed-form solution as

$$\boldsymbol{\delta}_{ij}^{(m+1)} = \begin{cases} S\left(\boldsymbol{\varsigma}_{ij}^{(m+1)}, \lambda c_{ij}/\vartheta\right) & \text{if } \left\|\boldsymbol{\varsigma}_{ij}^{(m+1)}\right\| \le \lambda c_{ij}+\lambda c_{ij}/\vartheta, \\ \frac{S\left(\boldsymbol{\varsigma}_{ij}^{(m+1)},\gamma\lambda c_{ij}/((\gamma-1)\vartheta)\right)}{1-1/((\gamma-1)\vartheta)} & \text{if } \lambda c_{ij}+\lambda c_{ij}/\vartheta < \left\|\boldsymbol{\varsigma}_{ij}^{(m+1)}\right\| \le \gamma\lambda c_{ij}, \\ \boldsymbol{\varsigma}_{ij}^{(m+1)} & \text{if } \left\|\boldsymbol{\varsigma}_{ij}^{(m+1)}\right\| > \gamma\lambda c_{ij}, \end{cases} \tag{11}$$

where $\gamma > c_{ij}+c_{ij}/\vartheta$ and $S(\boldsymbol{w},t) = (1-t/\|\boldsymbol{w}\|)_+\boldsymbol{w}$, and $(t)_+ = t$ if $t>0$, 0 otherwise.

In summary, the computational algorithm can be described as follows.

Algorithm: ADMM algorithm

Require: : Initialize $\boldsymbol{\beta}^{(0)}$, $\boldsymbol{\delta}^{(0)}$ and $\boldsymbol{v}^{(0)}$.

for $m = 0, 1, 2, \ldots$ **do**
  Update $\boldsymbol{\beta}$ by (9).
  Update $\boldsymbol{\eta}$ by (10).
  Update $\boldsymbol{\delta}$ by (11)
  Update $\boldsymbol{v}$ by (8).
  if convergence criterion is met **then**
    Stop and get the estimates
  else
    $m = m + 1$
  end if
end for

A good initial estimator of $\boldsymbol{\beta}$ will depend on many factors such as the number of covariates, the magnitude of difference between true parameters from different clusters, and signal to noise ratio. We can construct the initial values $\tilde{\boldsymbol{\beta}}^{(0)}$ by fitting a linear regression model $y_{ih} = \boldsymbol{z}_{ih}^T\boldsymbol{\eta} + \boldsymbol{x}_{ih}^T\boldsymbol{\beta}_i + \epsilon_{ih}$ for each $i = 1, \ldots, n$. Then, set $\boldsymbol{\delta}_{ij}^{(0)} = \boldsymbol{\beta}_i^{(0)} - \boldsymbol{\beta}_j^{(0)}$ and $\boldsymbol{v}^{(0)} = \mathbf{0}$. Alternatively, the initial values can be set using the procedure in Ma et al. (2020a). They used a ridge fusion criterion with a small tuning parameter value. The initial group structure was obtained by assigning objects into K^*(a given value) groups by ranking the estimated $\boldsymbol{\beta}_i$ based on the ridge fusion criterion.

If $\hat{\boldsymbol{\delta}}_{ij} = \mathbf{0}$, then the locations i and j belong to the same group. Thus, we can obtain the corresponding estimated partition $\hat{\mathcal{G}}$ and the estimated number of groups $\hat{K}(\lambda, \psi)$. For each group, the group-specific parameter vector is estimated as $\hat{\boldsymbol{\alpha}}_k = 1/|\hat{\mathcal{G}}_k| \sum_{i \in \hat{\mathcal{G}}_k} \hat{\boldsymbol{\beta}}_i$ for $k = 1, \ldots, \hat{K}$.

Remark 1. *If there are no global covariates, the model simplifies as* $y_{ih} = \boldsymbol{x}_{ih}^T\boldsymbol{\beta}_i + \epsilon_{ih}$. *The algorithm will be simplified, that is,* $\boldsymbol{Q}_{Z,\Omega}$ *will become* $\boldsymbol{\Omega}$. *The model we use in the application is the simplified model.*

Remark 2. *The convergence criterion used is the same as Ma and Huang (2017), which is based on the primal residual* $\boldsymbol{r}^{(m+1)} = \boldsymbol{A}\boldsymbol{\beta}^{(m+1)} - \boldsymbol{\delta}^{(m+1)}$. *The algorithm is stopped if* $\|\boldsymbol{r}^{(m+1)}\| < \varepsilon$, *where* ε *is a small positive number.*

Remark 3. *In (9), the computational cost of calculating the matrix inverse* $\left(\boldsymbol{X}^T\boldsymbol{Q}_{Z,\Omega}\boldsymbol{X} + \vartheta\boldsymbol{A}^T\boldsymbol{A}\right)^{-1}$ *is* $O(n^3)$ *if calculating the matrix inverse directly. However, based on the Sherman–Morrison–Woodbury (SMW) in Appendix, the matrix*

inverse can be calculated with less computational cost $O(n^2)$. The computational cost of updating pairwise differences is also $O(n^2)$. Thus, the computational cost of the whole algorithm is $O(n^2)$.

We need to select two tuning parameters, λ and ψ, in the SHADE algorithm. In this paper, we use the modified Bayes Information Criterion (BIC) (Wang et al., 2007) to determine the best tuning parameters adaptively from the data. In particular, we have

$$\mathrm{BIC}\,(\lambda,\psi) = \log\left[\frac{1}{n}\sum_{i=1}^{n}\frac{1}{n_i}\left(y_{ih} - \boldsymbol{z}_{ih}^T\hat{\boldsymbol{\eta}}(\lambda,\psi) - \boldsymbol{x}_{ih}^T\hat{\boldsymbol{\beta}}_i(\lambda,\psi)\right)^2\right] + C_n\frac{\log n}{n}\left(\hat{K}(\lambda,\psi)p + q\right), \tag{12}$$

where C_n is a positive number which can depend on n. Here $C_n = c_0\log(\log(np+q))$ with $c_0 = 0.2$ is used. For tuning parameter ψ, we select the best value from some candidate values, such as 0.1, 0.5, 1, 3. For each given ψ, we use the warm start and continuation strategy to select the tuning parameter λ. A grid of λ is predefined within $[\lambda_{\min}, \lambda_{\max}]$. For each λ, the initial values are the estimated values from the previous estimation. Denote the selected tuning parameters as $\hat{\lambda}$ and $\hat{\psi}$. Correspondingly, the estimated group number is $\hat{K}(\hat{\lambda},\hat{\psi})$, and the estimated regression coefficients are $\hat{\boldsymbol{\beta}}$ and $\hat{\boldsymbol{\eta}}$.

3. Theoretical Properties of SHADE

In this section, we study the theoretical properties of the proposed SHADE estimator. Assume $\mathcal{G}_k$'s are the true partition of location-specific regression coefficients. Let $|\mathcal{G}_k|$ be the number of subjects in group $\mathcal{G}_k$ for $k = 1,\ldots,K$, $|\mathcal{G}_{\min}|$ and $|\mathcal{G}_{\max}|$ be the minimum and maximum group sizes, respectively. Let $\widetilde{\boldsymbol{W}}$ be an $n\times K$ matrix with element w_{ik} and $w_{ik} = 1$ if $i\in\mathcal{G}_k$, $w_{ik} = 0$, otherwise. Denote $\boldsymbol{W} = \widetilde{\boldsymbol{W}}\otimes\boldsymbol{I}_p$, an $np\times Kp$ matrix, and $\boldsymbol{U} = (\boldsymbol{Z},\boldsymbol{X}\boldsymbol{W})$. Define $\mathcal{M}_{\mathcal{G}} = \{\boldsymbol{\beta}\in\mathbb{R}^{np} : \boldsymbol{\beta}_i = \boldsymbol{\beta}_j,\ \text{for } i,j\in\mathcal{G}_k,\ 1\le k\le K\}$. Using these notations, we can then express $\boldsymbol{\beta}$ as $\boldsymbol{\beta} = \boldsymbol{W}\boldsymbol{\alpha}$ if $\boldsymbol{\beta}\in\mathcal{M}_{\mathcal{G}}$, where $\boldsymbol{\alpha} = \left(\boldsymbol{\alpha}_1^T,\ldots,\boldsymbol{\alpha}_K^T\right)^T$. For any positive numbers, x_n and y_n, $x_n \gg y_n$ means that $x_n^{-1}y_n = o(1)$. Define the scaled penalty function as

$$\rho_\gamma(t) = \lambda^{-1}p_\gamma(t,\lambda). \tag{13}$$

Below are our assumptions.

(C1) The function $\rho_\gamma(t)$ is symmetric, non-decreasing, and concave on $[0,\infty)$. It is constant for $t\ge a\lambda$ for some constant $a>0$, and $\rho_\gamma(0) = 0$. Also, $\rho'(t)$ exists and is continuous except for a finite number values of t and $\rho'(0+) = 1$.

(C2) There exist finite positive constants $M_1, M_2, M_3 > 0$ such that $|x_{ih,l}| \leq M_1$, $|z_{ih,l}| \leq M_1$ for $j = 1, \ldots, n_i$ and $i = 1, \ldots, n$ and $M_2 \leq \max_i n_i / \min_i n_i \leq M_3$. Also, assume that $\lambda_{\min}\left(\boldsymbol{U}^T \boldsymbol{\Omega} \boldsymbol{U}\right) \geq C_1 |\mathcal{G}_{\min}|$, $\lambda_{\max}\left(\boldsymbol{U}^T \boldsymbol{\Omega} \boldsymbol{U}\right) \leq C_1' n$ for some constants $0 < C_1 < \infty$ and $0 < C_1' < \infty$, where $\lambda_{\min}$ and $\lambda_{\max}$ are the corresponding minimum and maximum eigenvalues respectively. In addition, assume that $\sup_{i,h} \|\boldsymbol{x}_{ih}\| \leq C_2 \sqrt{p}$ and $\sup_{i,h} \|\boldsymbol{z}_{ih}\| \leq C_3 \sqrt{q}$ for some constants $0 < C_2 < \infty$ and $0 < C_3 < \infty$.

(C3) The random error vector $\boldsymbol{\epsilon} = (\epsilon_{11}, \ldots, \epsilon_{1n_1}, \epsilon_{21}, \ldots, \epsilon_{2n_2}, \ldots, \epsilon_{n1}, \ldots, \epsilon_{nn_n})^T$ has sub-Gaussian tails such that $P\left(\left|\boldsymbol{a}^T \boldsymbol{\epsilon}\right| > \|\boldsymbol{a}\| x\right) \leq 2 \exp\left(-c_1 x^2\right)$ for any vector $\boldsymbol{a} \in \mathbb{R}^m$ and $x > 0$, where $0 < c_1 < \infty$ and $m = \sum_{i=1}^n n_i$.

(C4) The pairwise weights c_{ij}'s are bounded away from zero if i and j are in the same group.

(C5) Below is the condition for the minimum group size,

$$|\mathcal{G}_{\min}| \gg (q + Kp)^{1/2} \max\left(\sqrt{\frac{n}{\min_i n_i} \log n}, (q + Kp)^{1/2}\right).$$

Conditions (C1) and (C3) are commonly used in high-dimensional penalized regression problems. Condition (C2) is similar to the condition mentioned in Ma et al. (2020a). It also includes the bounded conditions for covariates, which are used in Huang et al. (2004). In general, if the weights functions are not zeros defined on a finite support, the c_{ij}'s will satisfy condition (C4). This mild condition can guarantee the consistency results. Different choices of weights that satisfy this condition will not have any effect on the consistency results. However, different c_{ij}'s may have different finite sample performances. We tried different c_{ij} in the simulation results to compare the performances.

First, we establish the properties of the oracle estimator, which is defined as the weighted least squares estimator, assuming that the underlying group structure is known. Specifically, the oracle estimator of $(\boldsymbol{\eta}, \boldsymbol{\alpha})$ is

$$\begin{aligned}(\hat{\boldsymbol{\eta}}^{or}, \hat{\boldsymbol{\alpha}}^{or}) &= \underset{\boldsymbol{\eta} \in \mathbb{R}^q, \boldsymbol{\alpha} \in \mathbb{R}^{Kp}}{\arg\min} \frac{1}{2} \left\|\boldsymbol{\Omega}^{1/2} (\boldsymbol{y} - \boldsymbol{Z}\boldsymbol{\eta} - \boldsymbol{X}\boldsymbol{W}\boldsymbol{\alpha})\right\|^2 \\ &= \left(\boldsymbol{U}^T \boldsymbol{\Omega} \boldsymbol{U}\right)^{-1} \boldsymbol{U}^T \boldsymbol{\Omega} \boldsymbol{y}. \end{aligned} \tag{14}$$

Then, the corresponding oracle estimator of $\boldsymbol{\beta}$ is $\hat{\boldsymbol{\beta}}^{or} = \boldsymbol{W}\hat{\boldsymbol{\alpha}}^{or}$. Let $\boldsymbol{\alpha}_k^0$ be the true coefficient vector for group k, $k = 1, \ldots, K$ and $\boldsymbol{\alpha}^0 = ((\boldsymbol{\alpha}_1^0)^T, \ldots, (\boldsymbol{\alpha}_K^0)^T)^T$, and let $\boldsymbol{\eta}^0$

be the true common coefficient vector. The following theorem shows the properties of the oracle estimator.

Theorem 1. *Under conditions (C1)-(C3) and (C5), $q = o(n)$ and $Kp = o(n)$, we have with probability at least $1 - 2(q + Kp)n^{-1}$,*

$$\left\| \begin{pmatrix} \hat{\boldsymbol{\eta}}^{or} - \boldsymbol{\eta}^0 \\ \hat{\boldsymbol{\alpha}}^{or} - \boldsymbol{\alpha}^0 \end{pmatrix} \right\| \leq \phi_n,$$

and

$$\left\| \hat{\boldsymbol{\beta}}^{or} - \boldsymbol{\beta}^0 \right\| \leq \sqrt{|\mathcal{G}_{\max}|}\phi_n; \sup_i \left\| \hat{\boldsymbol{\beta}}_i^{or} - \boldsymbol{\beta}_i^0 \right\| \leq \phi_n,$$

where

$$\phi_n = c_1^{-1/2} C_1^{-1} M_1 \sqrt{q + Kp}\, |\mathcal{G}_{\min}|^{-1} \sqrt{\frac{n}{\min n_i} \log n}.$$

Furthermore, for any vector $\boldsymbol{a}_n \in \mathbb{R}^{q+Kp}$, we have as $n \to \infty$

$$\sigma_n(\boldsymbol{a}_n)^{-1} \boldsymbol{a}_n^T \left(\left(\hat{\boldsymbol{\eta}}^{or} - \boldsymbol{\eta}^0 \right)^T, \left(\hat{\boldsymbol{\alpha}}^{or} - \boldsymbol{\alpha}^0 \right)^T \right)^T \xrightarrow{d} N(0, 1), \tag{15}$$

where

$$\sigma_n(\boldsymbol{a}_n) = \sigma \left[\boldsymbol{a}_n^T \left(\boldsymbol{U}^T \boldsymbol{\Omega} \boldsymbol{U} \right)^{-1} \boldsymbol{U}^T \boldsymbol{\Omega} \boldsymbol{\Omega} \boldsymbol{U} \left(\boldsymbol{U}^T \boldsymbol{\Omega} \boldsymbol{U} \right)^{-1} \boldsymbol{a}_n \right]^{1/2}. \tag{16}$$

Remark 4. *There are no specific assumptions about n_i. If $\min n_i \ll \frac{n}{q+Kp} \log n$, or $\min n_i = O\left(\frac{n}{q+Kp} \log n\right)$, (C5) becomes $|\mathcal{G}_{\min}| \gg (q + Kp)^{1/2} \sqrt{\frac{n}{\min n_i} \log n}$. If $\min n_i \gg \frac{n}{q+Kp} \log n$, (C5) becomes $|\mathcal{G}_{\min}| \gg q + Kp$. In this case, if q, p and K are fixed values, the condition (C5) becomes $1/|\mathcal{G}_{\min}| = o(1)$.*

Remark 5. *The model considered in Ma et al. (2020a) is a special case of the proposed model, and their condition is a special case of condition (C5) used here, that is when $n_i = 1$.*

Remark 6. *If let $|\mathcal{G}_{\min}| = \delta n/K$ for some constant $0 < \delta \leq 1$, then*

$$\phi_n = c_1^{-1/2} C_1^{-1} M_1 \delta^{-1} K \sqrt{q + Kp} \sqrt{\log n / (n \min n_i)}.$$

Moreover, if q, p and K are fixed values, then $\phi_n = C^* \sqrt{\log n / (n \min n_i)}$ for some constant $0 < C^* < \infty$.

Next, we study the properties of our proposed estimator. Let

$$b_n = \min_{i\in\mathcal{G}_k, j\in\mathcal{G}_{k'}} \left\|\boldsymbol{\beta}_i^0 - \boldsymbol{\beta}_j^0\right\| = \min_{k\neq k'} \left\|\boldsymbol{\alpha}_k^0 - \boldsymbol{\alpha}_{k'}^0\right\| \tag{17}$$

be the minimal difference among different groups.

Theorem 2. *Suppose the conditions of Theorem 1 hold and (C4) holds. If $b_n > a\lambda$ and $\lambda \gg \phi_n$ for some constant $a > 0$, then there exists a local minimizer $\left(\hat{\boldsymbol{\eta}}(\lambda,\psi)^T, \hat{\boldsymbol{\beta}}(\lambda,\psi)^T\right)^T$ of the objective function $Q_n(\boldsymbol{\eta},\boldsymbol{\beta})$ given in (2) such that*

$$P\left(\left(\hat{\boldsymbol{\eta}}(\lambda,\psi)^T, \hat{\boldsymbol{\beta}}(\lambda,\psi)^T\right)^T = \left((\hat{\boldsymbol{\eta}}^{or})^T, (\hat{\boldsymbol{\beta}}^{or})^T\right)^T\right) \to 1. \tag{18}$$

Remark 7. *Theorem 2 implies that true group structure can be recovered with probability approaching 1. It also implies that the estimated number of groups $\hat{K}$ satisfies $P\left(\hat{K}(\lambda,\psi) = K\right) \to 1$.*

Let $\hat{\boldsymbol{\alpha}}(\lambda,\psi)$ be the distinct group vectors of $\hat{\boldsymbol{\beta}}(\lambda,\psi)$. According to Theorem 1 and Theorem 2, we have the following result.

Corollary 1. *Suppose the conditions in Theorem 2 hold, for any vector $\boldsymbol{a}_n \in \mathbb{R}^{q+Kp}$, we have as $n \to \infty$*

$$\sigma_n(\boldsymbol{a}_n)^{-1}\boldsymbol{a}_n^T\left(\left(\hat{\boldsymbol{\eta}}(\lambda,\psi) - \boldsymbol{\eta}^0\right)^T, \left(\hat{\boldsymbol{\alpha}}(\lambda,\psi) - \boldsymbol{\alpha}^0\right)^T\right)^T \xrightarrow{d} N(0,1). \tag{19}$$

Remark 8. *The variance parameter σ^2 can be estimated by*

$$\sigma^2 = \frac{1}{m - q - \hat{K}p}\sum_{i=1}^{n}\sum_{h=1}^{n_i}\left(y_{ih} - \boldsymbol{z}_{ih}^T\hat{\boldsymbol{\eta}} - \boldsymbol{x}_{ih}^T\hat{\boldsymbol{\beta}}_i\right)^2 \tag{20}$$

The algorithm can be implemented through the package **Spgr** found in https://github.com/wangx23/Spgr.

4. Simulation Studies

In this section, we evaluate and compare the performance of the proposed SHADE estimator with different weight choices: equal weights $c_{ij} = 1$ (denoted as "equal"), weights defined in (4) (denoted as "reg-sp"), weights defined in (5) (denoted by "reg"), and weights defined in (6) (denoted by "sp").

The simulations are carried out as follows. Let $\boldsymbol{z}_{ih} = (z_{ih,1}, \ldots, z_{ih,5})^T$ with $z_{ih,1} = 1$ and $(z_{ih,2}, \ldots, z_{ih,5})^T$ are generated using a multivariate normal distribution

with mean 0, variance 1, and pairwise correlation $\rho = 0.3$. Define $\boldsymbol{x}_{ih} = (x_{ih,1}, x_{ih,2})^T$, where $x_{ih,1}$ is simulated from a standard normal distribution and $x_{ih,2}$ is simulated from a centered and standardized binomial $(n, 0.7)$. Let $\boldsymbol{\eta} = (\eta_1, \ldots, \eta_5)^T$, where η_k's are simulated from Uniform $[1, 2]$ and the standard deviation of the error term is $\sigma = 0.5$. We set $\vartheta = 1$ and $\gamma = 3$ and use the SCAD penalty function. The tuning parameters are chosen by the modified BIC defined by (12). We run the simulations under several scenarios. The results are based on 100 simulations.

To evaluate the subgrouping performance of the proposed method, we report the estimated group number $\hat{K}$, adjusted Rand index (ARI) (Rand, 1971; Hubert and Arabie, 1985; Vinh et al., 2010), and the root mean square error (RMSE) for estimating $\boldsymbol{\beta}$. For the estimated $\hat{K}$ over 100 simulations, we report its average (denoted by "mean"), standard error in the parenthesis, and the occurrence percentage of $\hat{K} = K$ (denoted by "per"). The quantity ARI is used to measure the degree of agreement between two partitions, taking a value between 0 and 1: the larger ARI value, the more agreement. We report the average ARI across 100 simulations along with the standard error in the parentheses. To evaluate the estimation accuracy of $\boldsymbol{\beta}$, we also report the average RMSE

$$\sqrt{\frac{1}{n}\sum_{i=1}^{n} \|\hat{\boldsymbol{\beta}}_i - \boldsymbol{\beta}_i\|^2}. \tag{21}$$

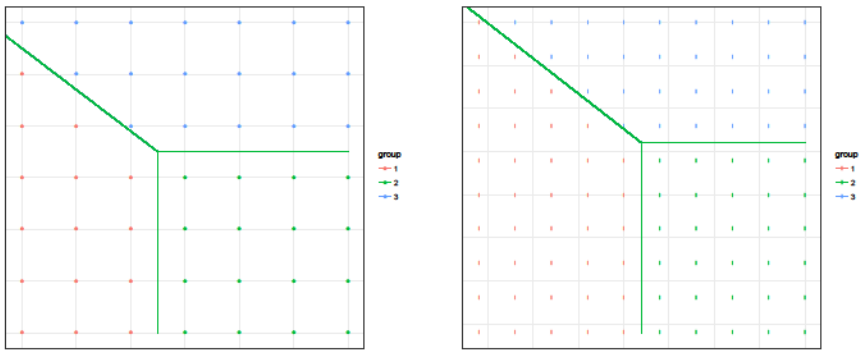
4.1. Balanced group

We assume that there are $K = 3$ true groups $\mathcal{G}_1, \mathcal{G}_2$ and $\mathcal{G}_3$. Consider the two spatial settings, for which the group parameters are respectively given by:

Setting 1: $\boldsymbol{\beta}_i = (1, 1)^T$ if $i \in \mathcal{G}_1$; $\boldsymbol{\beta}_i = (1.5, 1.5)^T$ if $i \in \mathcal{G}_2$; $\boldsymbol{\beta}_i = (2, 2)^T$ if $i \in \mathcal{G}_3$.

Setting 2: $\boldsymbol{\beta}_i = (1, 1)^T$ if $i \in \mathcal{G}_1$; $\boldsymbol{\beta}_i = (1.25, 1.25)^T$ if $i \in \mathcal{G}_2$; $\boldsymbol{\beta}_i = (1.5, 1.5)^T$ if $i \in \mathcal{G}_3$.

Under each setting, we simulate the data on two sizes of regular lattice, a 7×7 grid (left) and a 10×10 grid (right), as shown in Figure 1. Furthermore, for the 7×7 grid with $n_i = 10$, we use a 10-fold cross validation to select the tuning parameters. The repeated measures of location i are divided into 10 parts; the jth part of each location is combined as the validation data set, and the remaining observations form the training data set. The spatial weights (6) are considered. The results are labeled as "cv" in all the tables. Note that "reg_sp" and "reg" were not computed for the 10×10 grid.



(a) 7×7 grid (b) 10×10 grid

Figure 1: Two spatial settings in the simulation studies.

Results for Setting 1: Tables 1, 2 and 3 show the estimated number of groups and ARI. Figures 2 and 3 plot the RMSE of the estimates obtained using different weight choices. After estimating the group structure, one can also estimate parameters $\boldsymbol{\eta}$ and $\boldsymbol{\beta}$ again by assuming that the group information is known; the results are denoted as "refit". Based on the numerical results, we make the following observations.

First, we summarize the results for the 7×7 grid. In all the considered scenarios, the spatially weighted penalty outperforms the non-weighted penalty ("equal"). The upper panels in Tables 1 and 2, and the left plot in Figure 2 suggest that, if the number of repeated measurements is relatively small (say, $n_i = 10$), the weights "reg_sp" and "sp" perform similarly and they are the best in terms of estimating K, recovering the true subgroup structure (large ARI), and estimating regression coefficients (small RMSE); the weights "equal" and "reg" are much worse. The lower panels of Tables 1 and 2 and the right plot in Figure 2 show that when the number of repeated measurements gets larger (say, $n_i = 30$), all the methods improve and there is not much difference among them. Cross validation works well in terms of ARI and RMSE, but it tends to over-estimate the number of groups K. This is because that cross validation focuses more on the prediction accuracy; the coefficient estimates of some groups are close to the true coefficients, but they are not shrank together. In addition, refitting the model does not appear to further improve the accuracy of estimating $\boldsymbol{\beta}$.

Table 1: Summary of the estimate $\hat{K}$ for Setting 1 under the 7×7 grid.

		equal	reg_sp	reg	sp	cv
$n_i = 10$	mean	3.34(0.054)	3.15(0.039)	3.33(0.051)	3.13(0.034)	3.82(0.13)
	per	0.69	0.86	0.69	0.87	0.56
$n_i = 30$	mean	3.00(0)	3.00(0)	3.00(0)	3.00(0)	
	per	1.00	1.00	1.00	1.00	

Table 2: Average ARI for Setting 1 under the 7×7 grid

	equal	reg_sp	reg	sp	cv
$n_i = 10$	0.80(0.011)	0.92(0.008)	0.82(0.01)	0.92(0.007)	0.95(0.007)
$n_i = 30$	0.998(0.001)	0.999(0.0006)	0.998(0.001)	0.999(0.0006)	

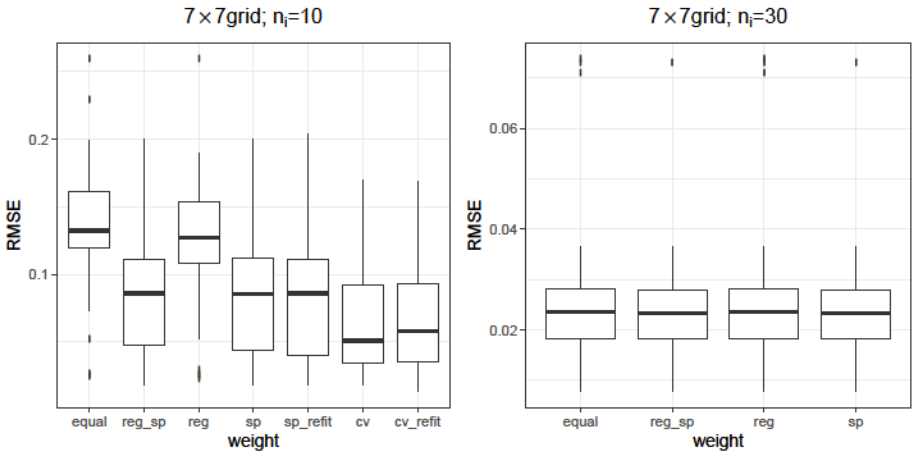


Figure 2: RMSE for Setting 1 under the 7×7 grid

Next, we summarize the results for the 10×10 grid. In this case, we consider equal weights and spatial weights only. Again, the spatially-weighted penalty outperforms the non-weighted penalty ("equal"). Table 3 and Figure 3 suggest that if the number of repeated measurements is relatively small (say, $n_i = 10$), "sp" performs much better in terms of grouping and estimating regression coefficients than "equal"; for a larger number of repeated measurements (say, $n_i = 30$), they perform similarly.

Table 3: Summary of $\hat{K}$ and average ARI for Setting 1 under the 10×10 grid.

		$\hat{K}$		ARI	
		equal	sp	equal	sp
$n_i = 10$	mean	3.59(0.073)	3.37(0.065)	0.70(0.009)	0.97(0.003)
	per	0.53	0.71	-	-
$n_i = 30$	mean	3(0)	3(0)	0.996(0.001)	1.00(0)
	per	1.00	1.00	-	-

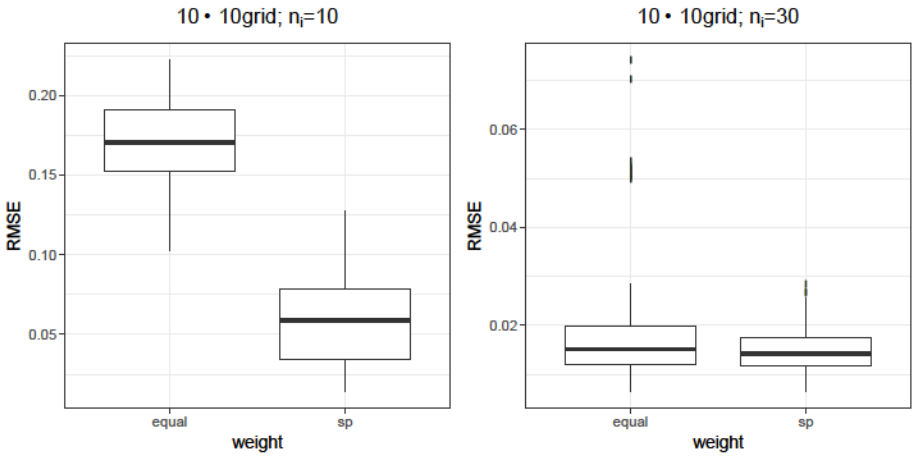


Figure 3: RMSE for Setting 1 under the 10×10 grid

Results for Setting 2: In this setting, the group difference becomes smaller. Tables 4, 5 and Figure 4 summarize the results for the 7×7 grid. For both values of n_i, the weights "sp" performs best in terms of estimating the number of groups ($\hat{K}$), recovering the true group structure (ARI), and estimating regression coefficients. In contrast to Setting 1, when the difference among groups becomes smaller, even with $n_i = 30$, the model with the spatial weight is superior to other models.

Table 4: Summary of $\hat{K}$ for Setting 2 under the 7×7 grid

		equal	reg_sp	reg	sp
$n_i = 10$	mean	3.25(0.119)	3.01(0.093)	3.14(0.107)	2.88(0.067)
	per	0.34	0.45	0.33	0.60
$n_i = 30$	mean	2.70(0.046)	2.90(0.030)	2.76(0.043)	2.95(0.022)
	per	0.70	0.90	0.76	0.95

Table 5: Average ARI for Setting 2 under the 7×7 grid

	equal	reg_sp	reg	sp
$n_i = 10$	0.32(0.011)	0.50(0.023)	0.33(0.01)	0.61(0.026)
$n_i = 30$	0.72(0.018)	0.86(0.015)	0.75(0.017)	0.90(0.012)

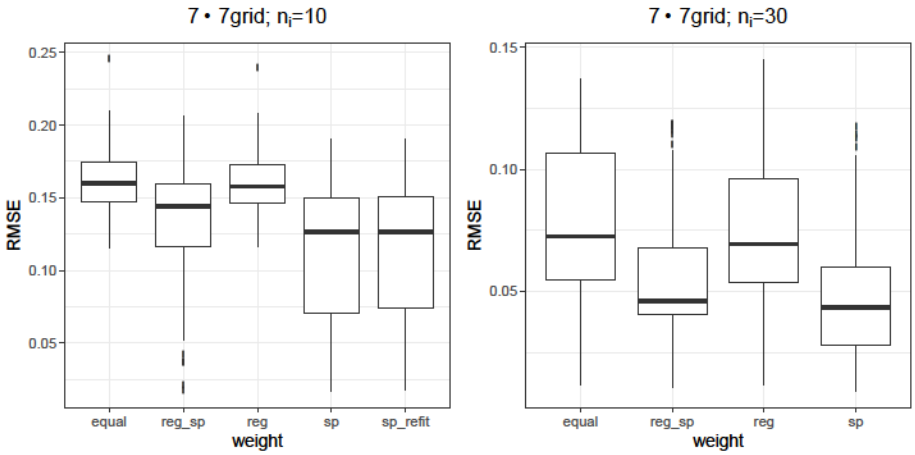


Figure 4: RMSE for setting 2 under the 7×7 grid

Table 6 and Figure 5 show the results for the 10×10 grid. Again, we only compare "equal" weights and "sp" weights. The results suggest similar conclusions to those for the 7×7 grid: the model with the spatial weight is superior even with a large number of repeated measurements ($n_i = 30$) by producing larger ARI and smaller RMSE.

Table 6: Summary of $\hat{K}$ and average ARI for Setting 2 under the 10×10 grid

		$\hat{K}$		ARI	
		equal	sp	equal	sp
$n_i = 10$	mean	3.82(0.146)	3.35(0.078)	0.32(0.009)	0.81(0.022)
	per	0.32	0.620	-	-
$n_i = 30$	mean	3.10(0.060)	3.00(0.0)	0.79(0.012)	0.94(0.005)
	per	0.64	1.0	-	-

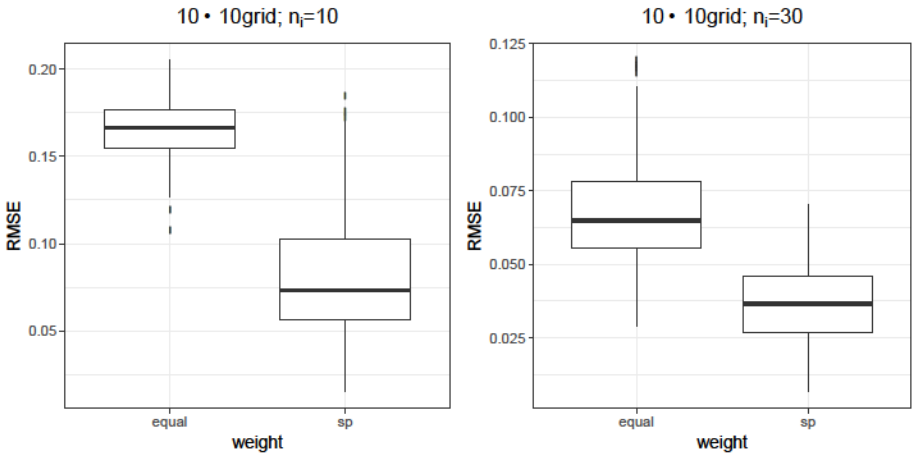


Figure 5: RMSE for Setting 2 under the 10×10 grid

4.2. Unbalanced group setting

Here we consider an unbalanced group setting as shown in Figure 6. In this setting, there are four groups, denoted as $\mathcal{G}_1, \mathcal{G}_2, \mathcal{G}_3$ and $\mathcal{G}_4$, and two groups have 9 subjects and the other two groups have 41 subjects. The group parameters are $\boldsymbol{\beta}_i = (1,1)^T$ if $i \in \mathcal{G}_1$, $\boldsymbol{\beta}_i = (1.5,1.5)^T$ if $i \in \mathcal{G}_2$, $\boldsymbol{\beta}_i = (2,2)^T$ if $i \in \mathcal{G}_3$ and $\boldsymbol{\beta}_i = (2.5,2.5)^T$ if $i \in \mathcal{G}_4$.

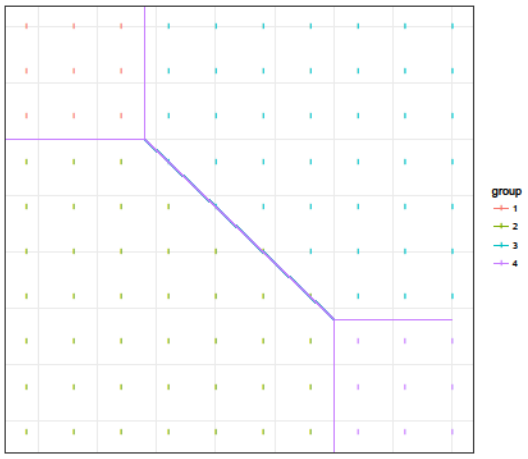


Figure 6: Unbalanced group setting

Table 7 and Figure 7 show the summaries of $\hat{K}$, ARI and RMSE for $\boldsymbol{\beta}$ when the number of repeated measurements is $n_i = 10$. In general, "reg_sp" and "sp" perform better than the other two types of weights. In particular, "sp" performs a slightly better than "reg_sp". The results are consistent with those under balanced cases. We expect that when the group difference becomes smaller, "sp" would still perform better than other weights even when the number of repeated measurements is large.

Table 7: Summary of $\hat{K}$ and average ARI for the unbalanced setting with $n_i = 10$

		equal	reg_sp	reg	sp
$\hat{K}$	mean	4.58(0.093)	4.23(0.049)	5.17(0.011)	4.35(0.059)
	per	0.570	0.800	0.300	0.710
ARI	mean	0.62(0.010)	0.94(0.061)	0.67(0.009)	0.96(0.004)

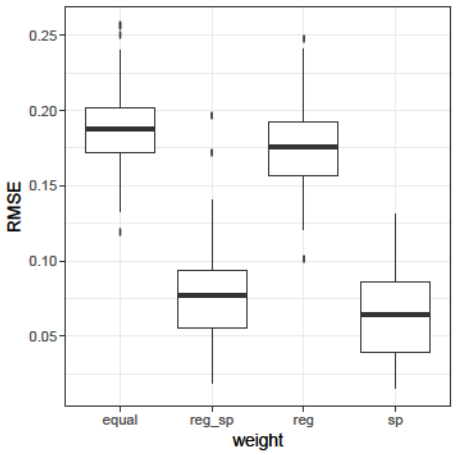


Figure 7: RMSE for unbalanced setting

4.3. Random group setting

We consider a setting without specified location group information. For each location, it has equal probability to three groups. Table 8 shows the summary of $\hat{K}$ and ARI for Setting 1 under the grid 7×7 with 10 repeated measures. Table 9 shows the summary of $\hat{K}$ and ARI for Setting 2 under the grid 7×7 with 30 repeated measures. Figure 8 shows the RMSE results for both cases. We can see that different weights have similar performances. The results suggest that even without prior information on the existence of spatial groups, "sp" weights can still produce comparable results to equal weights.

Table 8: Summary of $\hat{K}$ and average ARI for Setting 1 under the 7×7 grid with $n_i = 10$

		equal	reg_sp	reg	sp
$\hat{K}$	mean	3.42(0.064)	3.45(0.063)	3.40(0.059)	3.45(0.063)
	per	0.66	0.62	0.65	0.62
ARI	mean	0.78(0.011)	0.82(0.010)	0.81(0.010)	0.82(0.011)

Table 9: Summary of $\hat{K}$ and average ARI for Setting 2 under the 7×7 grid with $n_i = 30$

		equal	reg_sp	reg	sp
$\hat{K}$	mean	2.77(0.045)	2.77(0.045)	2.83(0.040)	2.73(0.047)
	per	0.75	0.75	0.81	0.71
ARI	mean	0.74(0.015)	0.76(0.016)	0.77(0.014)	0.74(0.017)

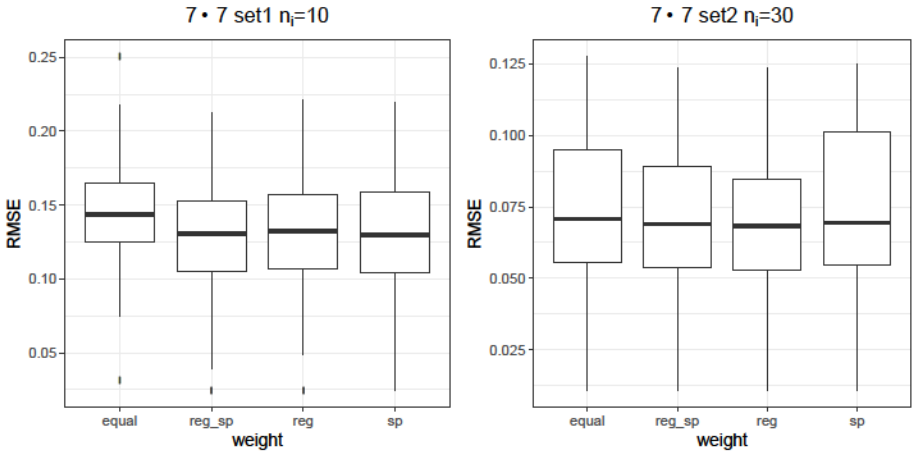


Figure 8: RMSE for random groups under the 7×7 grid

4.4. Computation time

In this section, we illustrate the computation time by using the algorithm implemented in Spgr. We consider four different values of n, which are 10×10, 15, 20×20 and 25×25 grid. The number of groups is 3, and the true group parameters follow Setting 2. The number of repeated measures is 30. We used 37 λ values and 4 ψ values when evaluating the computation time. The computation time is recorded based on an iMac with Processor 4.2 Hz Quad-Core Intel Core i7 and Memory 16GB. Figure 9 shows the results based on 100 simulations for equal weights and spatial weights, respectively. The y-axis is about the computation time in minutes. When n is 100, the computation time is less than 1 minute for spatial weights. When n is 400, the computation time is about 15 minutes for spatial weights and about 5 minutes for equal weights. When the number of locations is 625, the computation time can be about 15 minutes for equal weights and 70 minutes for spatial weights. A longer computation time of using spatial weights is because of an extra tuning parameter ψ.

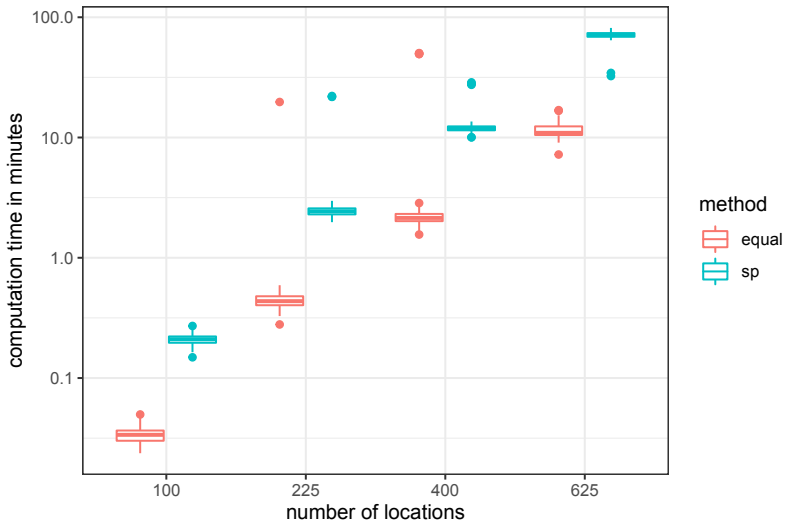


Figure 9: Computation time

5. Application

In this section, we apply our SHADE method to the modeling of the National Resources Inventory survey (NRI) data [1] for the purpose of small area estimation. The NRI survey is one of the largest longitudinal natural resource surveys in the U.S. The national and state level estimates of the status and change of landcover use and soil erosion have been used by numerous federal, state, and local agencies in the past few decades. In recent years, there is an increasing demand for NRI to provide various county level estimates. These include estimates of different land covers, such as cropland, pasture land, urban and forest. Due to the limitation of sample size, the uncertainty of the NRI direct county level estimates are usually too large for local stakeholders to make policy decisions. To make the county level estimates more useful, it is necessary to include some auxiliary information and an appropriate model to reduce the uncertainty of the estimates. One such set of auxiliary covariates is Cropland Data Layer (CDL), which is based on classification of satellite image pixels into several mutually exclusive and exhaustive land cover categories. In this section, we model the relationship between the NRI forest proportion and the CDL forest proportion among 48 states. In NRI, forests belonging to federal land, such as national

[1] https://www.nrcs.usda.gov/wps/portal/nrcs/main/national/technical/nra/nri/

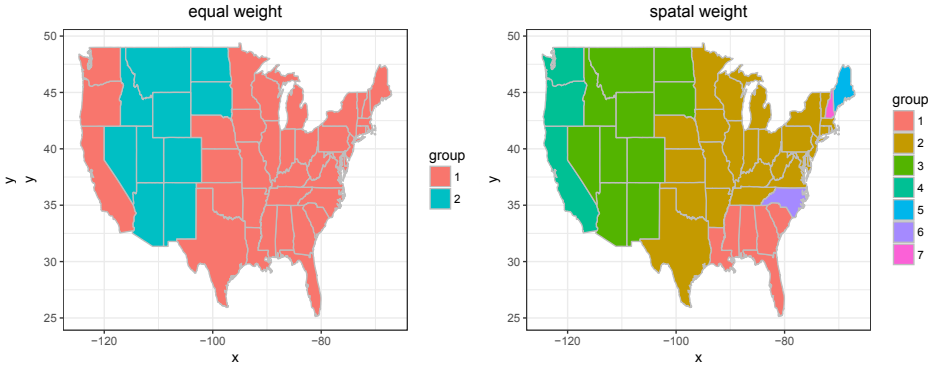
parks, are not included in the forest category. For states with more forest federal land, NRI estimates can be substantially smaller than CDL estimates. Therefore, different states could have different relationships between these two proportions.

The model we consider is,

$$y_{ih} = \beta_{0,i} + \beta_{1,i} x_{ih} + \epsilon_{ih} \tag{22}$$

where y_{ih} is the NRI forest proportion of the hth county in the ith state, x_{ih} is the corresponding CDL forest proportion of the hth county in the ith state, and $\beta_{0,i}$ and $\beta_{1,i}$ are the unknown coefficients. Both x and y are standardized. Instead of using the estimated linear regression coefficients as initial values directly, we use five sets of initial values which are simulated from a multivariate normal distribution with estimated coefficients as the mean vector and estimated covariance matrix as the covariance matrix. The models with the smallest modified BIC values are selected for equal weights and spatial weights, respectively.

In Figure 10, we display the estimated groups based on 2011 NRI data sets. The left figure shows the estimated groups based on equal weights, and the right one is for the estimated groups based on spatial weights in (6). We find that the two different weights give different estimated groups. Tables 10 and 11 are the corresponding estimates of regression coefficients in different groups.



(a) Estimated groups based on equal weights (b) Estimated groups based on spatial weights

Figure 10: Estimated groups for both equal weights and spatial weights.

Table 10: Estimated coefficients of different groups for equal weight

group	1	2
β_0	-0.029(0.006)	0.003(0.008)
β_1	0.885(0.011)	0.241(0.026)

Table 11: Estimated coefficients of different groups for spatial weights

group	1	2	3	4	5	6	7
β_0	-0.041(0.016)	-0.032(0.006)	0.003(0.007)	0.023(0.015)	-0.108(0.293)	0.275(0.038)	0.376 (0.309)
β_1	1.018(0.028)	0.867(0.012)	0.241(0.024)	0.608(0.033)	1.148 (0.377)	0.332(0.064)	0.341(0.384)

When considering equal weights, λ is the only tuning parameter in the algorithm. By changing the value of λ, we can have a different number of groups. We consider changing the λ value in the algorithm based on equal weights such that the number of groups is the same as what we have selected based on the spatial weights, that is, 7 groups. Figure 11 shows the group structure with 7 groups based on equal weights. In both Figure 11 and the left figure of Figure 10, "WA", "OR" and "CA" are not separated from the majority group (the group with the largest group size) when considering equal weights. These three states are in group 4, which are separated from the majority group (group 2) when considering spatial weights, which is more reasonable and intuitive based on the estimates of regression coefficients as shown in Table 11. One possible explanation of this result is that these three states have more national parks than those states in group 2.

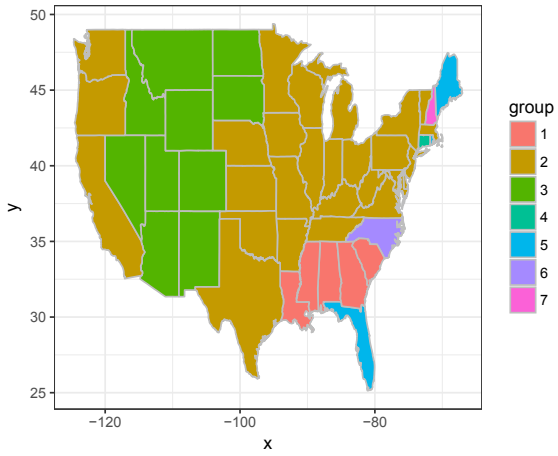


Figure 11: Estimated groups by changing the tuning parameter λ with equal weights.

Alternatively, we also implement K-means clustering based on the initial estimates to identify similar behaviors among the states. Figure 12 shows the maps based on 2-means clustering and 7-means clustering, respectively. The 2-cluster map is almost the same as the map based on equal weights. However, the 7-cluster map is not interpretable compared to the result based on spatial weights. This suggests

that the proposed procedure can produce more interpretable subgroup structures than K-means clustering methods.

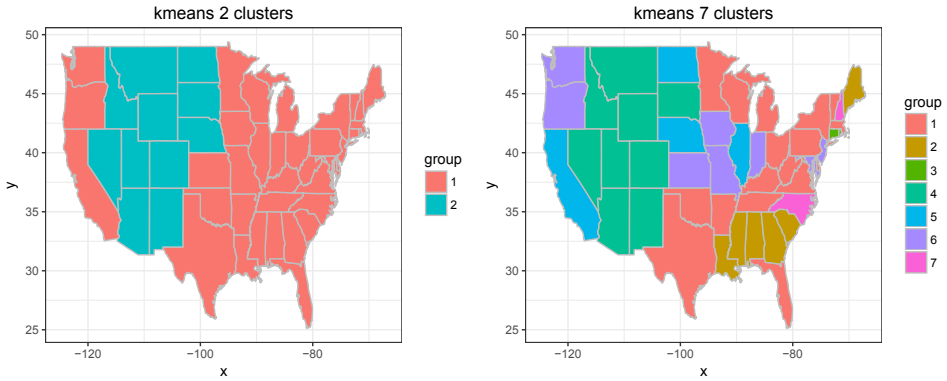


Figure 12: Group clustering results based on K-means.

6. Discussion

In this article, we considered the problem of spatial clustering of local covariate effects and develop a general framework called Spatial Heterogeneity Automatic Detection and Estimation (SHADE) for spatial areal data with repeated measures. In spatial data, since locations near each other usually have similar patterns, we proposed to take into account spatial information in the pairwise penalty, where closer locations are assigned with larger weights to encourage stronger shrinkage. In the simulation study, we used several examples to investigate and compare the performance of the procedure using different weights. We found that spatial information helps improve the accuracy of grouping, especially when the minimal group difference is small or the number of repeated measures is small. We also established theoretical properties of the proposed estimator in terms of its consistency in estimating the number of groups.

In the real data example, we have treated states as locations and counties as repeated measures. Alternatively, one can treat counties as individual units, since one state could have counties with two different features. Then, the algorithm will involve a matrix inverse with dimension more than 3000, which will require a higher computational burden. A further study is needed to compare these two models for the application.

The proposed method does not consider the spatial dependence in the regression error when constructing the objective function. The basic idea of this algorithm can

be extended to a general spatial clustering setup with consideration of the spatially dependent error. More specifically, the weighted least squared term in the objective function needs to be replaced by a generalized least squares term, which includes an estimated covariance matrix. The new algorithm should have two iterative steps. The first step is to update regression coefficients to find clusters and the second step is to update covariance parameters. More simulation studies are needed to explore the performance of the two-step algorithm. Moreover, the theoretical properties need to be established to support the new algorithm. Both theoretical and computational aspects of such extension are nontrivial and will be considered in a follow up work.

Acknowledgement

This research was supported in part by the Natural Resources Conservation Service of the U.S. Department of Agriculture and was supported in part by National Science Foundation grant NSF CCF-1740858.

Appendices

A. Proof of Theorem 1

In this section, we prove Theorem 1. When proving the central limit theorem (CLT) we use the technique in Huang et al. (2004).

The oracle estimator is defined in (14), which has the following form

$$\begin{pmatrix} \hat{\boldsymbol{\eta}}^{or} \\ \hat{\boldsymbol{\alpha}}^{or} \end{pmatrix} = \left(\boldsymbol{U}^T\boldsymbol{\Omega}\boldsymbol{U}\right)^{-1}\boldsymbol{U}^T\boldsymbol{\Omega}\boldsymbol{y}.$$

Thus, we have

$$\begin{pmatrix} \hat{\boldsymbol{\eta}}^{or} - \boldsymbol{\eta}^0 \\ \hat{\boldsymbol{\alpha}}^{or} - \boldsymbol{\alpha}^0 \end{pmatrix} = \left(\boldsymbol{U}^T\boldsymbol{\Omega}\boldsymbol{U}\right)^{-1}\boldsymbol{U}^T\boldsymbol{\Omega}\boldsymbol{\epsilon},$$

where $\boldsymbol{\epsilon} = (\boldsymbol{\epsilon}_i^T, \ldots, \boldsymbol{\epsilon}_n^T)^T$ with $\boldsymbol{\epsilon}_i = (\epsilon_{i1}, \ldots, \epsilon_{i,n_i})^T$. Therefore,

$$\left\| \begin{pmatrix} \hat{\boldsymbol{\eta}}^{or} - \boldsymbol{\eta}^0 \\ \hat{\boldsymbol{\alpha}}^{or} - \boldsymbol{\alpha}^0 \end{pmatrix} \right\| \leq \left\| \left(\boldsymbol{U}^T\boldsymbol{\Omega}\boldsymbol{U}\right)^{-1} \right\|_2 \left\| \boldsymbol{U}^T\boldsymbol{\Omega}\boldsymbol{\epsilon} \right\|, \tag{23}$$

where $\|\cdot\|_2$ is matrix norm, which is defined as, for a matrix $\boldsymbol{A}$, $\|\boldsymbol{A}\|_2 = \sup_{\|\boldsymbol{x}\|=1} \|\boldsymbol{A}\boldsymbol{x}\|$.

We know that

$$\begin{aligned} P\left(\left\| \boldsymbol{U}^T\boldsymbol{\Omega}\boldsymbol{\epsilon} \right\|_\infty > C\sqrt{\frac{n}{\min n_i}\log n} \right) \leq\ & P\left(\left\| (\boldsymbol{X}\boldsymbol{W})^T\boldsymbol{\Omega}\boldsymbol{\epsilon} \right\|_\infty > C\sqrt{\frac{n}{\min n_i}\log n} \right) \\ & + P\left(\left\| \boldsymbol{Z}^T\boldsymbol{\Omega}\boldsymbol{\epsilon} \right\|_\infty > C\sqrt{\frac{n}{\min n_i}\log n} \right), \end{aligned} \tag{24}$$

where C is a finite positive constant and $\|\cdot\|_\infty$ is defined as, for a vector $\boldsymbol{x} \in \mathbb{R}^m$, $\|\boldsymbol{x}\|_\infty = \max_{1\le i\le m} x_i$. By condition (C2), we have

$$\sqrt{\sum_{i=1}^{n}\sum_{h=1}^{n_i}\frac{x_{ih,l}^2}{n_i^2}1\left\{i\in\mathcal{G}_k\right\}} \le M_1\sqrt{\sum_{i=1}^{n}\frac{1}{n_i}\left\{i\in\mathcal{G}_k\right\}} \le M_1\sqrt{\sum_{i=1}^{n}\frac{1}{n_i}} \le M_1\sqrt{\frac{n}{\min n_i}}.$$

Since

$$\left\|(\boldsymbol{XW})^T\boldsymbol{\Omega\epsilon}\right\|_\infty = \sup_{k,l}\left|\sum_{i=1}^{n}\frac{1}{n_i}\sum_{h=1}^{n_i}x_{ih,l}\epsilon_{ih}1\left\{i\in\mathcal{G}_k\right\}\right|,$$

from condition (C3), it follows that

$$\begin{aligned}
&P\left(\left\|(\boldsymbol{XW})^T\boldsymbol{\Omega\epsilon}\right\|_\infty > C\sqrt{\frac{n}{\min n_i}\log n}\right)\\
\le&\sum_{l=1}^{p}\sum_{k=1}^{K}P\left(\left|\sum_{i=1}^{n}\sum_{j=1}^{n_i}\frac{1}{n_i}x_{ih,l}\epsilon_{ih}1\left\{i\in\mathcal{G}_k\right\}\right| > C\sqrt{\frac{n}{\min n_i}\log n}\right)\\
=&\sum_{l=1}^{p}\sum_{k=1}^{K}P\left(\left|\sum_{i=1}^{n}\sum_{h=1}^{n_i}\frac{1}{n_i}x_{ih,l}\epsilon_{ih}1\left\{i\in\mathcal{G}_k\right\}\right| > \frac{\sqrt{\sum_{i=1}^{n}\sum_{h=1}^{n_i}\frac{x_{ih,l}^2}{n_i^2}1\left\{i\in\mathcal{G}_k\right\}}}{\sqrt{\sum_{i=1}^{n}\sum_{h=1}^{n_i}\frac{x_{ih,l}^2}{n_i^2}1\left\{i\in\mathcal{G}_k\right\}}}C\sqrt{\frac{n}{\min n_i}\log n}\right)\\
\le&\sum_{l=1}^{p}\sum_{k=1}^{K}P\left(\left|\sum_{i=1}^{n}\sum_{h=1}^{n_i}\frac{1}{n_i}x_{ih,l}\epsilon_{ih}1\left\{i\in\mathcal{G}_k\right\}\right| > \sqrt{\sum_{i=1}^{n}\sum_{h=1}^{n_i}\frac{x_{ih,l}^2}{n_i^2}1\left\{i\in\mathcal{G}_k\right\}}\frac{C}{M_1}\sqrt{\log n}\right)\\
\le&2Kp\exp\left(-c_1\frac{C^2}{M_1^2}\log n\right) = 2Kpn^{-c_1C^2/M_1^2}.
\end{aligned}$$

Similarly, $\left|\sum_{i=1}^{n}\sum_{h=1}^{n_i}\frac{z_{ih,l}^2}{n_i^2}\right| \le M_1^2\sum_{i=1}^{n}1/n_i \le M_1^2\frac{n}{\min n_i}$. Again, by condition (C3), we have

$$\begin{aligned}
&P\left(\left\|\boldsymbol{Z}^T\boldsymbol{\Omega\epsilon}\right\|_\infty > C\sqrt{\frac{n}{\min n_i}\log n}\right)\\
\le&\sum_{l=1}^{q}P\left(\left|\sum_{i=1}^{n}\sum_{h=1}^{n_i}\frac{1}{n_i}z_{ih,l}\epsilon_{ih}\right| > C\sqrt{\frac{n}{\min n_i}\log n}\right)\\
\le&\sum_{l=1}^{q}P\left(\left|\sum_{i=1}^{n}\sum_{h=1}^{n_i}\frac{1}{n_i}z_{ih,l}\epsilon_{ih}\right| > \sqrt{\sum_{i=1}^{n}\sum_{h=1}^{n_i}\frac{z_{ih,l}^2}{n_i^2}}\frac{C}{M_1}\sqrt{\log n}\right)\\
\le&2q\exp\left(-c_1\frac{C^2}{M_1^2}\log n\right) = 2qn^{-c_1C^2/M_1^2}.
\end{aligned}$$

Thus, (24) can be bounded by

$$P\left(\left\|\boldsymbol{U}^T\boldsymbol{\Omega}\boldsymbol{\epsilon}\right\|_\infty > C\sqrt{\frac{n}{\min n_i}\log n}\right) \le 2\left(Kp+q\right)n^{-c_1C^2/M_1^2}.$$

Since $\left\|\boldsymbol{U}^T\boldsymbol{\Omega}\boldsymbol{\epsilon}\right\| \le \sqrt{q+Kp}\left\|\boldsymbol{U}^T\boldsymbol{\Omega}\boldsymbol{\epsilon}\right\|_\infty$,

$$P\left(\left\|\boldsymbol{U}^T\boldsymbol{\Omega}\boldsymbol{\epsilon}\right\| > C\sqrt{q+Kp}\sqrt{\frac{n}{\min n_i}\log n}\right) \le 2\left(Kp+q\right)n^{-c_1C^2/M_1^2}.$$

Let $C = c_1^{-1/2}M_1$, thus

$$P\left(\left\|\boldsymbol{U}^T\boldsymbol{\Omega}\boldsymbol{\epsilon}\right\| > C\sqrt{q+Kp}\sqrt{\frac{n}{\min n_i}\log n}\right) \le 2\left(Kp+q\right)n^{-1}. \tag{25}$$

Also, according to condition (C2), we have

$$\left\|\left(\boldsymbol{U}^T\boldsymbol{\Omega}\boldsymbol{U}\right)^{-1}\right\|_2 \le C_1^{-1}\left|\mathcal{G}_{\min}\right|^{-1}. \tag{26}$$

Combining (23), (25) and (26), with probability at least $1-2\left(Kp+q\right)n^{-1}$, we have

$$\left\|\begin{pmatrix}\hat{\boldsymbol{\eta}}^{or}-\boldsymbol{\eta}^0\\ \hat{\boldsymbol{\alpha}}^{or}-\boldsymbol{\alpha}^0\end{pmatrix}\right\| \le CC_1^{-1}\sqrt{q+Kp}\left|\mathcal{G}_{\min}\right|^{-1}\sqrt{\frac{n}{\min n_i}\log n}.$$

Let

$$\phi_n = c_1^{-1/2}C_1^{-1}M_1\sqrt{q+Kp}\left|\mathcal{G}_{\min}\right|^{-1}\sqrt{\frac{n}{\min n_i}\log n}.$$

Furthermore,

$$\begin{aligned}\left\|\hat{\boldsymbol{\beta}}^{or}-\boldsymbol{\beta}^0\right\|^2 &= \sum_{k=1}^K\sum_{i\in\mathcal{G}_k}\left\|\hat{\boldsymbol{\alpha}}_k^{or}-\boldsymbol{\alpha}_k^0\right\|^2 \le \left|\mathcal{G}_{\max}\right|\sum_{k=1}^K\left\|\hat{\boldsymbol{\alpha}}_k^{or}-\boldsymbol{\alpha}_k^0\right\|^2\\ &= \left|\mathcal{G}_{\max}\right|\left\|\hat{\boldsymbol{\alpha}}^{or}-\boldsymbol{\alpha}^0\right\|^2 \le \left|\mathcal{G}_{\max}\right|\phi_n^2,\end{aligned}$$

and

$$\sup_i\left\|\hat{\boldsymbol{\beta}}_i^{or}-\boldsymbol{\beta}_i^0\right\| = \sup_k\left\|\hat{\boldsymbol{\alpha}}_k^{or}-\boldsymbol{\alpha}_k^0\right\| \le \left\|\hat{\boldsymbol{\alpha}}^{or}-\boldsymbol{\alpha}^0\right\| \le \phi_n.$$

Next, we consider the central limit theorem. Let $\boldsymbol{U} = \left(\boldsymbol{U}_1^T,\ldots,\boldsymbol{U}_n^T\right)^T$ with $\boldsymbol{U}_i = \left(\boldsymbol{U}_{i1},\ldots,\boldsymbol{U}_{i,n_i}\right)^T$ for $i=1,\ldots,n$. Consider

$$\boldsymbol{a}_n^T\left(\left(\hat{\boldsymbol{\eta}}^{or}-\boldsymbol{\eta}^0\right)^T,\left(\hat{\boldsymbol{\alpha}}^{or}-\hat{\boldsymbol{\alpha}}^0\right)^T\right)^T = \sum_{i=1}^n\boldsymbol{a}_n^T\left(\sum_{i=1}^n\boldsymbol{U}_i^T\boldsymbol{\Omega}_i\boldsymbol{U}_i\right)^{-1}\boldsymbol{U}_i^T\boldsymbol{\Omega}_i\boldsymbol{\epsilon}_i,$$

where $\boldsymbol{\Omega}_i = 1/n_i \boldsymbol{I}_{n_i}$. By the assumption of $\boldsymbol{\epsilon}_i$ in the model (1), we have

$$E\left[\boldsymbol{a}_n^T \left(\left(\hat{\boldsymbol{\eta}}^{or} - \boldsymbol{\eta}^0\right)^T, \left(\hat{\boldsymbol{\alpha}}^{or} - \hat{\boldsymbol{\alpha}}^0\right)^T \right)^T \right] = 0.$$

The variance of $\boldsymbol{a}_n^T \left(\left(\hat{\boldsymbol{\eta}}^{or} - \boldsymbol{\eta}^0\right)^T, \left(\hat{\boldsymbol{\alpha}}^{or} - \hat{\boldsymbol{\alpha}}^0\right)^T \right)^T$ can be written as

$$\begin{aligned}
&Var\left\{ \boldsymbol{a}_n^T \left(\left(\hat{\boldsymbol{\eta}}^{or} - \boldsymbol{\eta}^0\right)^T, \left(\hat{\boldsymbol{\alpha}}^{or} - \hat{\boldsymbol{\alpha}}^0\right)^T \right)^T \right\} \\
=&\sigma^2 \left[\boldsymbol{a}_n^T \left(\boldsymbol{U}^T \boldsymbol{\Omega} \boldsymbol{U}\right)^{-1} \boldsymbol{U}^T \boldsymbol{\Omega} \boldsymbol{\Omega} \boldsymbol{U} \left(\boldsymbol{U}^T \boldsymbol{\Omega} \boldsymbol{U}\right)^{-1} \boldsymbol{a}_n \right] \\
=&\sigma^2 \left[\boldsymbol{a}_n^T \left(\boldsymbol{U}^T \boldsymbol{\Omega} \boldsymbol{U}\right)^{-1} \sum_{i=1}^{n} \boldsymbol{U}_i^T \boldsymbol{\Omega}_i \boldsymbol{\Omega}_i \boldsymbol{U}_i \left(\boldsymbol{U}^T \boldsymbol{\Omega} \boldsymbol{U}\right)^{-1} \boldsymbol{a}_n \right].
\end{aligned}$$

We use the technique of Huang et al. (2004) in the proof of their Theorem 3. That is,
$\boldsymbol{a}_n^T \left(\left(\hat{\boldsymbol{\eta}}^{or} - \boldsymbol{\eta}^0\right)^T, \left(\hat{\boldsymbol{\alpha}}^{or} - \hat{\boldsymbol{\alpha}}^0\right)^T \right)^T$ can be written as $\sum_{i=1}^n a_i \xi_i$ with

$$a_i^2 = \boldsymbol{a}_n^T \left(\boldsymbol{U}^T \boldsymbol{\Omega} \boldsymbol{U}\right)^{-1} \boldsymbol{U}_i^T \boldsymbol{\Omega}_i \boldsymbol{\Omega}_i \boldsymbol{U}_i \left(\boldsymbol{U}^T \boldsymbol{\Omega} \boldsymbol{U}\right)^{-1} \boldsymbol{a_n},$$

where ξ_i's are independent with mean zero and variance one. If

$$\frac{\max_i a_i^2}{\sum_{i=1}^n a_i^2} \to 0,$$

then $\sum_{i=1}^n a_i \xi_i / \sqrt{\sum_{i=1}^n a_i^2}$ is asymptotically $N(0,1)$.

For any $\boldsymbol{\lambda} = (\lambda_1, \ldots, \lambda_{q+Kp})^T$, we have

$$\begin{aligned}
\boldsymbol{\lambda}^T \boldsymbol{U}_i^T \boldsymbol{\Omega}_i \boldsymbol{\Omega}_i \boldsymbol{U}_i \boldsymbol{\lambda} &= \frac{1}{n_i^2} \boldsymbol{\lambda}^T \boldsymbol{U}_i^T \boldsymbol{U}_i \boldsymbol{\lambda} = \frac{1}{n_i^2} \sum_{h=1}^{n_i} \boldsymbol{\lambda}^T \boldsymbol{U}_{ih} \boldsymbol{U}_{ih}^T \boldsymbol{\lambda} \\
&= \frac{1}{n_i^2} \sum_{h=1}^{n_i} \left(\sum_{l=1}^{q+Kp} U_{ih,l} \lambda_l \right)^2 \\
&\le \frac{1}{n_i^2} \sum_{h=1}^{n_i} \left(\sum_{l=1}^{q+Kp} U_{ih,l}^2 \right) \left(\sum_{l=1}^{q+Kp} \lambda_l^2 \right) \le \frac{M_1^2}{n_i} (q + Kp) \|\boldsymbol{\lambda}\|^2 .
\end{aligned}$$

$$\begin{aligned}
\boldsymbol{\lambda}^T \left(\sum_{i=1}^{n} \boldsymbol{U}_i^T \boldsymbol{\Omega}_i \boldsymbol{\Omega}_i \boldsymbol{U}_i \right) \boldsymbol{\lambda} &\ge \frac{1}{\max_i n_i} \boldsymbol{\lambda}^T \left(\sum_{i=1}^{n} \boldsymbol{U}_i^T \boldsymbol{\Omega}_i \boldsymbol{U}_i \right) \boldsymbol{\lambda} \ge \frac{1}{\max_i n_i} \boldsymbol{\lambda}^T \boldsymbol{U}^T \boldsymbol{\Omega} \boldsymbol{U} \boldsymbol{\lambda} \\
&\ge \frac{1}{\max_i n_i} C_1 |\mathcal{G}_{\min}| \|\boldsymbol{\lambda}\|^2 ,
\end{aligned}$$

where the last inequality is by condition (C2). So,

$$\frac{\max_i \boldsymbol{\lambda}^T \boldsymbol{U}_i^T \boldsymbol{\Omega}_i \boldsymbol{\Omega}_i \boldsymbol{U}_i \boldsymbol{\lambda}}{\boldsymbol{\lambda}^T \left(\sum_{i=1}^n \boldsymbol{U}_i^T \boldsymbol{\Omega}_i \boldsymbol{\Omega}_i \boldsymbol{U}_i\right) \boldsymbol{\lambda}} \leq \left(\max_i n_i\right) \left(\max_i \frac{1}{n_i}\right) M_1^2 C_1^{-1} \left|\mathcal{G}_{\min}\right|^{-1} (q + Kp)$$
$$= M_1^2 C_1^{-1} \frac{\max_i n_i}{\min_i n_i} \left|\mathcal{G}_{\min}\right|^{-1} (q + Kp) \to 0, \tag{27}$$

by assumption.

By (27), we have that $\max_i a_i^2 / \sum_{i=1}^n a_i^2 \to 0$, so (15) exists.

B. *Proof of Theorem 2*

In this section, we prove Theorem 2. As in Ma et al. (2020a) and Ma and Huang (2017), we define $T : \mathcal{M}_{\mathcal{G}} \to \mathbb{R}^{Kp}$ to be the mapping that $T(\boldsymbol{\beta}) = \boldsymbol{\alpha}$ and $T^* : \mathbb{R}^{np} \to \mathbb{R}^{Kp}$ to be the mapping that $T^*(\boldsymbol{\beta}) = \left(\left|\mathcal{G}_k\right|^{-1} \sum_{i \in \mathcal{G}_k} \boldsymbol{\beta}_i^T,\ k = 1, \ldots, K\right)^T$.

Consider the following neighborhood of $(\boldsymbol{\eta}^0, \boldsymbol{\beta}^0)$,

$$\Theta = \left\{\boldsymbol{\eta} \in \mathbb{R}^q, \boldsymbol{\beta} \in \mathbb{R}^{np} : \left\|\boldsymbol{\eta} - \boldsymbol{\eta}^0\right\| \leq \phi_n, \sup_i \left\|\boldsymbol{\beta}_i - \boldsymbol{\beta}_i^0\right\| \leq \phi_n\right\}.$$

According to Theorem 1, there exists an event E_1 where $\|\boldsymbol{\eta} - \boldsymbol{\eta}^0\| \leq \phi_n$ and $\sup_i \|\boldsymbol{\beta}_i - \boldsymbol{\beta}_i^0\| \leq \phi_n$ such that $P(E_1) \geq 1 - 2(q + Kp) n^{-1}$.

Recall that the objective function to minimize is given in (2), which has the following form

$$Q_n(\boldsymbol{\eta}, \boldsymbol{\beta}; \lambda, \psi) = \frac{1}{2} \sum_{i=1}^n \frac{1}{n_i} \sum_{h=1}^{n_i} \left(y_{ih} - \boldsymbol{z}_{ih}^T \boldsymbol{\eta} - \boldsymbol{x}_{ih}^T \boldsymbol{\beta}_i\right)^2 + \sum_{1 \leq i < j \leq n} p_\gamma\left(\left\|\boldsymbol{\beta}_i - \boldsymbol{\beta}_j\right\|, c_{ij}\lambda\right). \tag{28}$$

Here we show that $\left((\hat{\boldsymbol{\eta}}^{or})^T, (\hat{\boldsymbol{\beta}}^{or})^T\right)^T$ is a strict local minimizer of the above objective function with probability approaching 1 by two steps as in Ma et al. (2020a). The first step is to show that in event E_1, $Q_n(\boldsymbol{\eta}, \boldsymbol{\beta}^*) > Q_n(\hat{\boldsymbol{\eta}}^{or}, \hat{\boldsymbol{\beta}}^{or})$ for any $(\boldsymbol{\eta}^T, \boldsymbol{\beta}^T)^T \in \Theta$ and $(\boldsymbol{\eta}^T, \boldsymbol{\beta}^{*T})^T \neq ((\hat{\boldsymbol{\eta}}^{or})^T, (\hat{\boldsymbol{\beta}}^{or})^T)^T$, where $\boldsymbol{\beta}^* = T^{-1}(T^*(\boldsymbol{\beta}))$ and $\boldsymbol{\beta} \in \mathbb{R}^{np}$. The proof of this step is almost the same as the first step in Ma et al. (2020a) , which is omitted here.

Here we show the second step, that is, there exists an event E_2 such that $P(E_2) \geq 1 - 2n^{-1}$. In the event $E_1 \cap E_2$, there is a neighborhood Θ_n of $\left((\hat{\boldsymbol{\eta}}^{or})^T, \left(\hat{\boldsymbol{\beta}}^{or}\right)^T\right)^T$, such that $Q_n(\boldsymbol{\eta}, \boldsymbol{\beta}) \geq Q_n(\boldsymbol{\eta}, \boldsymbol{\beta}^*)$ for any $\left(\boldsymbol{\eta}^T, \boldsymbol{\beta}^T\right)^T \in \Theta_n \cap \Theta$ for sufficiently large n.

Let $\Theta_n = \left\{\boldsymbol{\beta}_i : \sup_i \left\|\boldsymbol{\beta}_i - \hat{\boldsymbol{\beta}}_i^{or}\right\| \leq t_n\right\}$, where t_n is a positive sequence with $t_n = o(1)$. By Taylor's expansion, for $\left(\boldsymbol{\eta}^T, \boldsymbol{\beta}^T\right)^T \in \Theta_n \cap \Theta$,

$$Q_n\left(\boldsymbol{\eta}, \boldsymbol{\beta}\right) - Q_n\left(\boldsymbol{\eta}, \boldsymbol{\beta}^*\right) = \Gamma_1 + \Gamma_2, \tag{29}$$

where

$$\begin{aligned}
\Gamma_1 &= -\left(\boldsymbol{y} - \boldsymbol{Z}\boldsymbol{\eta} - \boldsymbol{X}\boldsymbol{\beta}^m\right)^T \boldsymbol{\Omega}\boldsymbol{X}\left(\boldsymbol{\beta} - \boldsymbol{\beta}^*\right), \\
\Gamma_2 &= \sum_{i=1}^{n} \frac{\partial\left[\lambda \sum_{l<j} c_{lj}\rho_\gamma\left(\left\|\boldsymbol{\beta}_l^m - \boldsymbol{\beta}_j^m\right\|\right)\right]}{\partial \boldsymbol{\beta}_i^T}\left(\boldsymbol{\beta}_i - \boldsymbol{\beta}_i^*\right),
\end{aligned}$$

with $\boldsymbol{\beta}^m = \alpha\boldsymbol{\beta} + (1-\alpha)\boldsymbol{\beta}^*$ for some constant $\alpha \in (0,1)$.

We have Γ_2 as follows,

$$\Gamma_2 = \lambda \sum_{i<j} c_{ij}\rho_\gamma'\left(\left\|\boldsymbol{\beta}_i^m - \boldsymbol{\beta}_j^m\right\|\right)\left\|\boldsymbol{\beta}_i^m - \boldsymbol{\beta}_j^m\right\|^{-1}\left(\boldsymbol{\beta}_i^m - \boldsymbol{\beta}_j^m\right)^T\left\{\left(\boldsymbol{\beta}_i - \boldsymbol{\beta}_i^*\right) - \left(\boldsymbol{\beta}_j - \boldsymbol{\beta}_j^*\right)\right\}.$$

For $i, j \in \mathcal{G}_k$, $\boldsymbol{\beta}_i^* = \boldsymbol{\beta}_j^*$ and $\boldsymbol{\beta}_i^m - \boldsymbol{\beta}_j^m = \alpha\left(\boldsymbol{\beta}_i - \boldsymbol{\beta}_j\right)$, then

$$\begin{aligned}
\Gamma_2 = {} & \lambda \sum_{k=1}^{K} \sum_{\{i,j\in\mathcal{G}_k, i<j\}} c_{ij}\rho_\gamma'\left(\left\|\boldsymbol{\beta}_i^m - \boldsymbol{\beta}_j^m\right\|\right)\left\|\boldsymbol{\beta}_i^m - \boldsymbol{\beta}_j^m\right\|^{-1}\left(\boldsymbol{\beta}_i^m - \boldsymbol{\beta}_j^m\right)^T\left(\boldsymbol{\beta}_i - \boldsymbol{\beta}_j\right) \\
& + \lambda \sum_{k=1}^{K} \sum_{\{i\in\mathcal{G}_k, j\in\mathcal{G}_{k'}\}} c_{ij}\rho_\gamma'\left(\left\|\boldsymbol{\beta}_i^m - \boldsymbol{\beta}_j^m\right\|\right)\left\|\boldsymbol{\beta}_i^m - \boldsymbol{\beta}_j^m\right\|^{-1}\left(\boldsymbol{\beta}_i^m - \boldsymbol{\beta}_j^m\right)^T\left\{\left(\boldsymbol{\beta}_i - \boldsymbol{\beta}_i^*\right) - \left(\boldsymbol{\beta}_j - \boldsymbol{\beta}_j^*\right)\right\}.
\end{aligned}$$

Since $\sup_i \left\|\boldsymbol{\beta}_i^m - \boldsymbol{\beta}_i^0\right\| \leq \phi_n$, for $k \neq k'$, $i \in \mathcal{G}_k, j \in \mathcal{G}_{k'}$,

$$\left\|\boldsymbol{\beta}_i^m - \boldsymbol{\beta}_j^m\right\| \geq \min_{i\in\mathcal{G}_k, j\in\mathcal{G}_{k'}}\left\|\boldsymbol{\beta}_i^0 - \boldsymbol{\beta}_j^0\right\| - 2\max_i\left\|\boldsymbol{\beta}_i^m - \boldsymbol{\beta}_i^0\right\| \geq b_n - 2\phi_n > a\lambda.$$

Thus, $\rho_\gamma'(\left\|\boldsymbol{\beta}_i^m - \boldsymbol{\beta}_j^m\right\|) = 0$ by assumption (C1). Therefore,

$$\Gamma_2 = \lambda \sum_{i=1}^{K} \sum_{\{i,j\in\mathcal{G}_k, i<j\}} c_{ij}\rho_\gamma'\left(\left\|\boldsymbol{\beta}_i^m - \boldsymbol{\beta}_j^m\right\|\right)\left\|\boldsymbol{\beta}_i - \boldsymbol{\beta}_j\right\|. \tag{30}$$

Also, for $i, j \in \mathcal{G}_k$, $\sup_i \left\|\boldsymbol{\beta}_i^m - \boldsymbol{\beta}_j^m\right\| \leq 4t_n$, so $\rho_\gamma'\left(\left\|\boldsymbol{\beta}_i^m - \boldsymbol{\beta}_j^m\right\|\right) \geq \rho'\left(4t_n\right)$ by assumption (C1). Thus, we have

$$\Gamma_2 \geq \sum_{k=1}^{K} \sum_{\{i,j\in\mathcal{G}_k, i<j\}} \lambda c_{ij}\rho_\gamma'\left(4t_n\right)\left\|\boldsymbol{\beta}_i - \boldsymbol{\beta}_j\right\|.$$

Let $\boldsymbol{Q} = \left(\boldsymbol{Q}_1^T, \ldots, \boldsymbol{Q}_n^T\right)^T = \left[\left(\boldsymbol{y} - \boldsymbol{Z}\boldsymbol{\eta} - \boldsymbol{X}\boldsymbol{\beta}^m\right)^T \boldsymbol{\Omega}\boldsymbol{X}\right]^T$ with

$$\boldsymbol{Q}_i = \frac{1}{n_i} \sum_{h=1}^{n_i} \left(y_{ih} - \boldsymbol{z}_{ih}^T \boldsymbol{\eta} - \boldsymbol{x}_{ih}^T \boldsymbol{\beta}_i^m\right) \boldsymbol{x}_{ih}.$$

We have,

$$\begin{aligned}
\Gamma_1 &= -\left(\boldsymbol{y} - \boldsymbol{Z}\boldsymbol{\eta} - \boldsymbol{X}\boldsymbol{\beta}^m\right)^T \boldsymbol{\Omega}\boldsymbol{X} \left(\boldsymbol{\beta} - \boldsymbol{\beta}^*\right) \\
&= -\boldsymbol{Q}^T \left(\boldsymbol{\beta} - \boldsymbol{\beta}^*\right) \\
&= -\sum_{k=1}^{K} \sum_{\{i,j \in \mathcal{G}_k, i<j\}} \frac{\left(\boldsymbol{Q}_i - \boldsymbol{Q}_j\right)^T \left(\boldsymbol{\beta}_i - \boldsymbol{\beta}_j\right)}{|\mathcal{G}_k|}.
\end{aligned} \tag{31}$$

Moreover,

$$\boldsymbol{Q}_i = \frac{1}{n_i} \sum_{h=1}^{n_i} \left(\epsilon_{ih} + \boldsymbol{z}_{ih}^T \left(\boldsymbol{\eta}^0 - \boldsymbol{\eta}\right) + \boldsymbol{x}_{ih}^T \left(\boldsymbol{\beta}_i^0 - \boldsymbol{\beta}_i^m\right)\right) \boldsymbol{x}_{ih},$$

so

$$\begin{aligned}
\sup_i \|\boldsymbol{Q}_i\| &\leq \sup_{i,h} \|\boldsymbol{x}_{ih}\| \left(\|\boldsymbol{\xi}\|_\infty + \sup_{i,h} \|\boldsymbol{z}_{ih}\| \left\|\boldsymbol{\eta}^0 - \boldsymbol{\eta}\right\| + \sup_{i,h} \|\boldsymbol{x}_{ih}\| \left\|\boldsymbol{\beta}_i^0 - \boldsymbol{\beta}_i^m\right\|\right) \\
&\leq C_2\sqrt{p}\left(\|\boldsymbol{\xi}\|_\infty + C_3\sqrt{q}\phi_n + C_2\sqrt{p}\phi_n\right),
\end{aligned}$$

where $\boldsymbol{\xi} = (\xi_1, \ldots, \xi_n)^T$ with $\xi_i = \frac{1}{n_i}\sum_{h=1}^{n} \epsilon_{ih}$. According to Condition (C3),

$$\begin{aligned}
P\left(\|\boldsymbol{\xi}\|_\infty > \sqrt{2c_1^{-1}}\sqrt{\log n / \min n_i}\right) &\leq \sum_{i=1}^{n} P\left(|\xi_i| > \sqrt{2c_1^{-1}}\sqrt{\log n / \min n_i}\right) \\
&= \sum_{i=1}^{n} P\left(\left|\frac{1}{n_i}\sum_{j=1}^{n_i} \epsilon_{ij}\right| > \sqrt{2c_1^{-1}}\sqrt{\log n / \min n_i}\right) \\
&\leq \sum_{i=1}^{n} P\left(\left|\frac{1}{n_i}\sum_{j=1}^{n_i} \epsilon_{ij}\right| > \sqrt{2c_1^{-1}}\sqrt{\log n / n_i}\right) \\
&\leq 2\sum_{i=1}^{n} \exp\left\{-c_1 2c_1^{-1} \log n\right\} \leq \frac{2}{n}.
\end{aligned}$$

Thus, there exists an event E_2 such that $P(E_2) \geq 1 - 2n^{-1}$ and

$$\sup_i \|\boldsymbol{Q}_i\| \leq C_2\sqrt{p}\left(\sqrt{2c_1^{-1}}\sqrt{\log n / \min_i n_i} + C_3\sqrt{q}\phi_n + C_2\sqrt{p}\phi_n\right).$$

Thus,

$$\begin{aligned}
&\left|\frac{\left(\boldsymbol{Q}_i-\boldsymbol{Q}_j\right)^T\left(\boldsymbol{\beta}_i-\boldsymbol{\beta}_j\right)}{|\mathcal{G}_k|}\right| \\
\leq & 2\left|\mathcal{G}_{\min}\right|^{-1}\sup_i\left\|\boldsymbol{Q}_i\right\|\left\|\boldsymbol{\beta}_i-\boldsymbol{\beta}_j\right\| \\
\leq & 2C_2\left|\mathcal{G}_{\min}\right|^{-1}\sqrt{p}\left(\sqrt{2c_1^{-1}}\sqrt{\log n/\min_i n_i}+C_3\sqrt{q}\phi_n+C_2\sqrt{p}\phi_n\right)\left\|\boldsymbol{\beta}_i-\boldsymbol{\beta}_j\right\|. \qquad (32)
\end{aligned}$$

Combining (30), (31) and (32), (29) follows that

$$\begin{aligned}
&Q_n\left(\boldsymbol{\eta},\boldsymbol{\beta}\right)-Q_n\left(\boldsymbol{\eta},\boldsymbol{\beta}^*\right) \\
\geq & \sum_{k=1}^K\sum_{\{i,j\in\mathcal{G}_k,i<j\}}\left\{\lambda c_{ij}\rho'\left(4t_n\right)-2C_2\left|\mathcal{G}_{\min}\right|^{-1}\sqrt{p}\left(\sqrt{2c_1^{-1}}\sqrt{\frac{\log n}{\min_i n_i}}+C_3\sqrt{q}\phi_n+C_2\sqrt{p}\phi_n\right)\right\}\left\|\boldsymbol{\beta}_i-\boldsymbol{\beta}_j\right\|.
\end{aligned}$$

As $t_n=o\left(1\right)$, $\rho'\left(4t_n\right)\to 1$. Since $\left|\mathcal{G}_{\min}\right|\gg\left(q+Kp\right)^{1/2}\max\left(\sqrt{\frac{n}{\min_i n_i}\log n},\left(q+Kp\right)^{1/2}\right)$, $p=o(n)$ and $q=o(n)$, then $\left|\mathcal{G}_{\min}\right|^{-1}p=o\left(1\right)$ and $\left|\mathcal{G}_{\min}\right|^{-1}\sqrt{pq}=o\left(1\right)$. Thus, $\lambda\gg\left|\mathcal{G}_{\min}\right|^{-1}\sqrt{p}\sqrt{\frac{\log n}{\min n_i}}$, $\lambda\gg\left|\mathcal{G}_{\min}\right|^{-1}\sqrt{pq}\phi_n$ and $\lambda\gg\left|\mathcal{G}_{\min}\right|^{-1}p\phi_n$. Therefore, $Q_n\left(\boldsymbol{\eta},\boldsymbol{\beta}\right)-Q_n\left(\boldsymbol{\eta},\boldsymbol{\beta}^*\right)\geq 0$ for sufficiently large n by the assumption (C4) that c_{ij}'s are bounded if i and j are in the same group.

Therefore, combining the two steps, we will have that $Q_n\left(\boldsymbol{\eta},\boldsymbol{\beta}\right)>Q_n(\hat{\boldsymbol{\eta}}^{or},\hat{\boldsymbol{\beta}}^{or})$ for any $\left(\boldsymbol{\eta}^T,\boldsymbol{\beta}^T\right)^T\in\Theta_n\cap\Theta$ and $(\boldsymbol{\eta}^T,\boldsymbol{\beta}^T)^T\neq((\hat{\boldsymbol{\eta}}^{or})^T,(\hat{\boldsymbol{\beta}}^{or})^T)^T$. This shows that $((\hat{\boldsymbol{\eta}}^{or})^T,(\hat{\boldsymbol{\beta}}^{or})^T)^T$ is a strict local minimizer of the objective function (2) on $E_1\cap E_2$ with probability at least $1-2(K+p+1)n^{-1}$ for sufficiently large n.

C. Sherman–Morrison–Woodbury formula

Consider $\left(\boldsymbol{X}^T\boldsymbol{Q}_{Z,\Omega}\boldsymbol{X}+\nu\boldsymbol{A}^T\boldsymbol{A}\right)^{-1}$. It is known that $\boldsymbol{A}^T\boldsymbol{A}=n\boldsymbol{I}_{np}-\left(\mathbf{1}_n\otimes\boldsymbol{I}_p\right)\left(\mathbf{1}_n\otimes\boldsymbol{I}_p\right)^T$. Let $\boldsymbol{X}^*=\boldsymbol{\Omega}^{1/2}\boldsymbol{X}$ and $\boldsymbol{Z}^*=\boldsymbol{\Omega}^{1/2}\boldsymbol{Z}$, then the target matrix becomes

$$\begin{aligned}
&\left(\boldsymbol{X}^{*T}\boldsymbol{Q}_{Z^*}\boldsymbol{X}^*+\nu\boldsymbol{A}^T\boldsymbol{A}\right)^{-1} \\
=&\left(\boldsymbol{X}^{*T}\boldsymbol{X}^*+\nu n\boldsymbol{I}_{np}-\boldsymbol{X}^{*T}\boldsymbol{Z}^*\left(\boldsymbol{Z}^{*T}\boldsymbol{Z}^*\right)^{-1}\boldsymbol{Z}^{*T}\boldsymbol{X}^*-\nu\left(\mathbf{1}_n\otimes\boldsymbol{I}_p\right)\left(\mathbf{1}_n\otimes\boldsymbol{I}_p\right)^T\right)^{-1}.
\end{aligned}$$

Let $\boldsymbol{A}_1=\boldsymbol{X}^{*T}\boldsymbol{X}^*+\nu n\boldsymbol{I}_{np}-\boldsymbol{X}^{*T}\boldsymbol{Z}^*\left(\boldsymbol{Z}^{*T}\boldsymbol{Z}^*\right)^{-1}\boldsymbol{Z}^{*T}\boldsymbol{X}^*$, $\boldsymbol{B}=\mathbf{1}_n\otimes\boldsymbol{I}_p$, $\boldsymbol{C}=\nu\boldsymbol{I}_p$ and $\boldsymbol{D}=\boldsymbol{B}^T$, then based on Sherman–Morrison–Woodbury formula the orignal inverse

can be written as

$$(\boldsymbol{A}_1 - \boldsymbol{BCD})^{-1} = \boldsymbol{A}_1^{-1} + \boldsymbol{A}_1^{-1}\boldsymbol{B}\left(\frac{1}{\nu}\boldsymbol{I}_p - \boldsymbol{B}^T\boldsymbol{A}_1^{-1}\boldsymbol{B}\right)^{-1}\boldsymbol{B}^T\boldsymbol{A}_1^{-1}.$$

Use Sherman–Morrison–Woodbury formula again to calculate $\boldsymbol{A}_1^{-1}$, that is

$$\boldsymbol{A}_1^{-1} = \left(\boldsymbol{X}^{*T}\boldsymbol{X}^* + \nu n\boldsymbol{I}_{np} - \boldsymbol{X}^{*T}\boldsymbol{Z}^*\left(\boldsymbol{Z}^{*T}\boldsymbol{Z}^*\right)^{-1}\boldsymbol{Z}^{*T}\boldsymbol{X}^*\right)^{-1},$$

which can be written as

$$\begin{aligned}\boldsymbol{A}_1^{-1} &= \left(\boldsymbol{X}^{*T}\boldsymbol{X}^* + \nu n\boldsymbol{I}_{np}\right)^{-1} \\ &+ \left(\boldsymbol{X}^{*T}\boldsymbol{X}^* + \nu n\boldsymbol{I}_{np}\right)^{-1}\boldsymbol{X}^{*T}\boldsymbol{Z}^*\left[\boldsymbol{Z}^{*T}\boldsymbol{Z}^* - \boldsymbol{Z}^{*T}\boldsymbol{X}^*\left(\boldsymbol{X}^{*T}\boldsymbol{X}^* + \nu n\boldsymbol{I}_{np}\right)^{-1}\boldsymbol{X}^{*T}\boldsymbol{Z}^*\right]^{-1}. \\ &\cdot \boldsymbol{Z}^{*T}\boldsymbol{X}^*\left(\boldsymbol{X}^{*T}\boldsymbol{X}^* + \nu n\boldsymbol{I}_{np}\right)^{-1}\end{aligned}$$

Also it is known that,

$$A_{11}^{-1} = \left(\boldsymbol{X}^{*T}\boldsymbol{X}^* + \nu n\boldsymbol{I}_{np}\right)^{-1} = \begin{pmatrix} \left(\boldsymbol{x}_1^*\boldsymbol{x}_1^{*T} + \nu n\boldsymbol{I}_p\right)^{-1} & \cdots & \boldsymbol{0} \\ \vdots & \ddots & \vdots \\ \boldsymbol{0} & \cdots & \left(\boldsymbol{x}_n^*\boldsymbol{x}_n^{*T} + \nu n\boldsymbol{I}_p\right)^{-1} \end{pmatrix},$$

and

$$\begin{aligned}\boldsymbol{B}^T\boldsymbol{A}_1^{-1}\boldsymbol{B} &= \sum_{i=1}^n\left(\boldsymbol{x}_i^*\boldsymbol{x}_i^{*T} + \nu n\boldsymbol{I}_p\right)^{-1} \\ &+ \left(\sum_{i=1}^n\left(\boldsymbol{x}_i^*\boldsymbol{x}_i^{*T} + \nu n\boldsymbol{I}_p\right)^{-1}\boldsymbol{x}_i^*\boldsymbol{z}_i^{*T}\right)\left[\boldsymbol{Z}^{*T}\boldsymbol{Z}^* - \boldsymbol{Z}^{*T}\boldsymbol{X}^*\left(\boldsymbol{X}^{*T}\boldsymbol{X}^* + \nu n\boldsymbol{I}_{np}\right)^{-1}\boldsymbol{X}^{*T}\boldsymbol{Z}^*\right]^{-1} \\ &\cdot \left(\sum_{i=1}^n\left(\boldsymbol{x}_i^*\boldsymbol{x}_i^{*T} + \nu n\boldsymbol{I}_p\right)^{-1}\boldsymbol{x}_i^*\boldsymbol{z}_i^{*T}\right)^T.\end{aligned}$$

References

Banerjee, S., Carlin, B. P., and Gelfand, A. E. (2014). *Hierarchical modeling and analysis for spatial data*. Crc Press.

Boyd, S., Parikh, N., Chu, E., Peleato, B., and Eckstein, J. (2011). Distributed optimization and statistical learning via the alternating direction method of multipliers. *Foundations and Trends in Machine Learning*, 3(1):1–122.

Bradley, J. R., Holan, S. H., and Wikle, C. K. (2018). Computationally efficient multivariate spatio-temporal models for high-dimensional count-valued data (with discussion). *Bayesian Analysis*, 13(1):253–310.

Brunsdon, C., Fotheringham, A. S., and Charlton, M. E. (1996). Geographically weighted regression: a method for exploring spatial nonstationarity. *Geographical analysis*, 28(4):281–298.

Casetti, E. (1972). Generating models by the expansion method: applications to geographical research. *Geographical analysis*, 4(1):81–91.

Casetti, E. and Jones, J. P. (1987). Spatial aspects of the productivity slowdown: an analysis of us manufacturing data. *Annals of the Association of American Geographers*, 77(1):76–88.

Chi, E. C. and Lange, K. (2015). Splitting methods for convex clustering. *Journal of Computational and Graphical Statistics*, 24(4):994–1013.

Cook, A. J., Gold, D. R., and Li, Y. (2007). Spatial cluster detection for censored outcome data. *Biometrics*, 63(2):540–549.

Cressie, N. (2015). *Statistics for spatial data*. John Wiley & Sons.

Diggle, P. J., Tawn, J. A., and Moyeed, R. A. (1998). Model-based geostatistics. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 47(3):299–350.

Fan, J. and Li, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American statistical Association*, 96(456):1348–1360.

Fan, Z., Guan, L., et al. (2018). Approximate l_0-penalized estimation of piecewise-constant signals on graphs. *The Annals of Statistics*, 46(6B):3217–3245.

Gelfand, A. E., Kim, H.-J., Sirmans, C., and Banerjee, S. (2003). Spatial modeling with spatially varying coefficient processes. *Journal of the American Statistical Association*, 98(462):387–396.

Hallac, D., Leskovec, J., and Boyd, S. (2015). Network lasso: Clustering and optimization in large graphs. In *Proceedings of the 21th ACM SIGKDD international conference on knowledge discovery and data mining*, pages 387–396.

Han, W., Yang, Z., Di, L., and Mueller, R. (2012). Cropscape: A Web service based application for exploring and disseminating US conterminous geospatial cropland data products for decision support. *Computers and Electronics in Agriculture*, 84:111–123.

Hu, G. and Bradley, J. (2018). A bayesian spatial–temporal model with latent multivariate log-gamma random effects with application to earthquake magnitudes. *Stat*, 7(1):e179.

Hu, G. and Huffer, F. (2020). Modified Kaplan–Meier estimator and Nelson–Aalen estimator with geographical weighting for survival data. *Geographical Analysis*, 52(1):28–48.

Hu, G., Xue, Y., and Ma, Z. (2020). Bayesian clustered coefficients regression with auxiliary covariates assistant random effects. *arXiv preprint arXiv:2004.12022*.

Huang, J. Z., Wu, C. O., and Zhou, L. (2004). Polynomial spline estimation and inference for varying coefficient models with longitudinal data. *Statistica Sinica*, 14(3):763–788.

Hubert, L. and Arabie, P. (1985). Comparing partitions. *Journal of classification*, 2(1):193–218.

Im, Y. and Tan, A. (2021). Bayesian subgroup analysis in regression using mixture models. *Computational Statistics & Data Analysis*, 162:107252.

Jung, I., Kulldorff, M., and Klassen, A. C. (2007). A spatial scan statistic for ordinal data. *Statistics in medicine*, 26(7):1594–1607.

Kulldorff, M. and Nagarwalla, N. (1995). Spatial disease clusters: detection and inference. *Statistics in medicine*, 14(8):799–810.

Lee, J., Gangnon, R. E., and Zhu, J. (2017). Cluster detection of spatial regression coefficients. *Statistics in medicine*, 36(7):1118–1133.

Lee, J., Sun, Y., and Chang, H. H. (2020). Spatial cluster detection of regression coefficients in a mixed-effects model. *Environmetrics*, 31(2):e2578.

Li, F. and Sang, H. (2019). Spatial homogeneity pursuit of regression coefficients for large datasets. *Journal of the American Statistical Association*, 114(527):1050–1062.

Liu, L. and Lin, L. (2019). Subgroup analysis for heterogeneous additive partially linear models and its application to car sales data. *Computational Statistics & Data Analysis*, 138:239–259.

Lu, H. and Carlin, B. P. (2005). Bayesian areal wombling for geographical boundary analysis. *Geographical Analysis*, 37(3):265–285.

Lu, H., Reilly, C. S., Banerjee, S., and Carlin, B. P. (2007). Bayesian areal wombling via adjacency modeling. *Environmental and Ecological Statistics*, 14(4):433–452.

Luo, Z., Sang, H., and Mallick, B. (2021). A bayesian contiguous partitioning method for learning clustered latent variables. *Journal of machine learning research*, 22.

Ma, S. and Huang, J. (2017). A concave pairwise fusion approach to subgroup analysis. *Journal of the American Statistical Association*, 112(517):410–423.

Ma, S., Huang, J., Zhang, Z., and Liu, M. (2020a). Exploration of heterogeneous treatment effects via concave fusion. *The international journal of biostatistics*, 16(1).

Ma, Z., Xue, Y., and Hu, G. (2020b). Heterogeneous regression models for clusters of spatial dependent data. *Spatial Economic Analysis*, pages 1–17.

Nakaya, T., Fotheringham, A. S., Brunsdon, C., and Charlton, M. (2005). Geographically weighted Poisson regression for disease association mapping. *Statistics in medicine*, 24(17):2695–2717.

Nusser, S. M. and Goebel, J. J. (1997). The National Resources Inventory: a long-term multi-resource monitoring programme. *Environmental and Ecological Statistics*, 4(3):181–204.

Rand, W. M. (1971). Objective criteria for the evaluation of clustering methods. *Journal of the American Statistical association*, 66(336):846–850.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, 58(1):267–288.

Vinh, N. X., Epps, J., and Bailey, J. (2010). Information theoretic measures for clusterings comparison: variants, properties, normalization and correction for chance. *Journal of Machine Learning Research*, 11:2837–2854.

Wang, H., Li, R., and Tsai, C.-L. (2007). Tuning parameter selectors for the smoothly clipped absolute deviation method. *Biometrika*, 94(3):553–568.

Wang, X., Berg, E., Zhu, Z., Sun, D., and Demuth, G. (2018). Small area estimation of proportions with constraint for National Resources Inventory survey. *Journal of Agricultural, Biological and Environmental Statistics*, 23(4):509–528.

Xu, Z., Bradley, J. R., and Sinha, D. (2019). Latent multivariate log-gamma models for high-dimensional multi-type responses with application to daily fine particulate matter and mortality counts. *arXiv preprint arXiv:1909.02528*.

Xue, Y., Schifano, E. D., and Hu, G. (2020). Geographically weighted Cox regression for prostate cancer survival data in louisiana. *Geographical Analysis*, 52(4):570–587.

Zhang, C.-H. (2010). Nearly unbiased variable selection under minimax concave penalty. *The Annals of statistics*, 38(2):894–942.

Zhang, J. and Lawson, A. B. (2011). Bayesian parametric accelerated failure time spatial model and its application to prostate cancer. *Journal of applied statistics*, 38(3):591–603.