REGULARIZED REGRESSION METHODS WITH SPATIAL BINARY AND MULTINOMIAL OUTCOMES

by

Rao Fu

A dissertation submitted in partial fulfillment of
the requirements for the degree of

Doctor of Philosophy

(Statistics)

at the

UNIVERSITY OF WISCONSIN–MADISON

2015

Date of final oral examination: 06/05/2015

The dissertation is approved by the following members of the Final Oral Committee:
Professor Jun Zhu, Statistics and Entomology
Associate Professor Sijian Wang, Statistics and Biostatistics and Medical Informatics
Professor Kam-Wah Tsui, Statistics
Assistant Professor Yingqi Zhao, Biostatistics and Medical Informatics
Associate Professor Menggang Yu, Biostatistics and Medical Informatics

Abstract

Categorical data analysis is common in such disciplines as landscape ecology and environmental history. A motivating example is a study conducted to assess the influence of social, economic, and historical factors on forest covers. In this thesis, I consider the statistical analysis of this type of problem in a spatial binary setting and a multinomial regression setting, and develop new methodology and theory for this purpose.

Autologistic regression models are proposed for relating spatial binary responses to spatial ownership characteristics. For big ecological data, a penalized estimation method is developed under pseudolikelihood and an approximation is derived for assessing the variation of pseudolikelihood estimates. A simulation study is conducted to evaluate the performance of the proposed method and algorithm, followed by a data example in a study of land cover in relation to land ownership characteristics.

Under the multinomial regression setting, I propose a group Lasso type of regularization method for multinomial regression models that can shrink some or all of the regression coefficients to zero simultaneously. Since the existing theorems cannot be directly applied to group Lasso for multinomial regression models, we establish a framework for selection consistency under suitable regularity conditions. Further, we devise an efficient algorithm to compute the group Lasso estimates. A simulation study shows that our method outperforms the traditional Lasso in terms of sum of the squared bias, Kullback-Leibler divergence, specificity, and correct model selection frequency. For illustration, our method is applied to evaluate the influence of past land ownership characteristics on land cover structure in northern Wisconsin, USA.

Acknowledgments

I would like to express the deepest appreciation to my PhD advisor Professor Jun Zhu, for her invaluable guidance and persistent help in my study. Her guidance has shaped me into a rigorous thinker, critical writer and skilled researcher. Also, I would like to thank my co-advisor Professor Sijian Wang for encouraging me during the stressful job hunting period and providing me with insightful comments and suggestions which lead to Chapter 3 of my dissertation.

I want to thank Professor Kam-Wah Tsui, Yingqi Zhao and Menggang Yu for kindly serving as my committee members. I wish to express my gratitude to Professor Kam-Wah Tsui for sharing his knowledge and stories with me during my PhD study. Also, a big thank you to Professor Yingqi Zhao and Menggang Yu who provided keen comments, questions, and recommendations that improved the final version of this document.

Finally, I am eternally grateful for the encouragement and support provided by my kind family and lovely girl friend Jie Guo. I especially thank my parents for devoting their energy and love to promote my happiness, and I dedicate this work to their honor.

Contents

List of Tables

List of Figures

Chapter 1

Introduction

Categorical data analysis is common in such disciplines as landscape ecology and environmental history. A motivating example is a study conducted to assess the influence of social, economic, and historical factors on forest covers. Researchers are particularly interested in quantifying relationships between land cover types and land ownership characteristics and identifying key factors. However, statistical methods for analyzing such data are limited. The purpose of this thesis is to fill some of this void by bringing together several strands of recent research in spatial statistics, developing new variable selection methods, and making these methodologies applicable for the analysis of spatial categorical data in practice. Although the specific application here concerns land cover types in relationship to land ownership characteristics in an intersecting area of landscape ecology and environmental history, the statistical methodologies developed can be adapted to analyze categorical data in general.

In Chapter 2, I focus on the spatial binary data on a lattice. Autologistic regression models are suitable for relating spatial binary responses in ecology to covariates such as environmental factors. For big ecological data, pseudolikelihood estimation is appealing due to its ease of computation, but at least two challenges remain. Although an important issue, it is unclear how model selection may be carried out under pseudolikelihood. In addition, for assessing the variation of pseudolikelihood estimates, parametric bootstrap using Monte Carlo simulation is often used but may be infeasible for very large data sizes. Here both these issues are addressed by developing a penalized pseudolikelihood estimation method and an approximation of the variance of the parameter estimates. Also, I develop a LARS-type algorithm for fast computation. A simulation study is conducted to evaluate the performance of the proposed method and algorithm, followed by a data example in a study of land cover in relation to land ownership characteristics. Extension of these models and methods to spatial-temporal binary data is further discussed. This chapter is published in Journal of Agricultural, Biological, and Ecological Statistics, and technical details for Chapter 2 is given in Appendix A.

In Chapter 3, I consider the problem in a multinomial regression setting and develop a group Lasso type of regularization method for multinomial regression models that can shrink some or all of the regression coefficients to zero simultaneously. Since the existing theorems cannot be directly applied to group Lasso for multinomial regression models, we establish a framework for selection consistency under suitable regularity conditions. Further, we devise an efficient algorithm to compute the group Lasso estimates. A simulation study shows that our method outperforms the traditional Lasso in terms of sum of the squared bias, Kullback-Leibler divergence, specificity, and correct model selection frequency. For illustration, our method is applied to evaluate the influence of past land ownership characteristics on land cover structure in northern Wisconsin, USA. In addition, alternative penalization methods and auto-multinomial Regression models are further discussed. And technical details and proof of theorem are given in Appendix B.

Chapter 2

On Estimation and Selection of Autologistic Regression Models via Penalized Pseudolikelihood

2.1 Introduction

Autologistic regression is suitable for modeling the relationship between a spatial binary response and covariates in environmental and ecological studies (see, e.g., Gumpertz et al. (1977); Huffer and Wu (1998); He et al. (2003)). For big ecological data, however, there are several challenging issues in terms of computation and methodology for statistical inference. This chapter aims to address some of these issues and in particular, we develop a new, computationally efficient method for selection of covariates and estimation of model parameters.

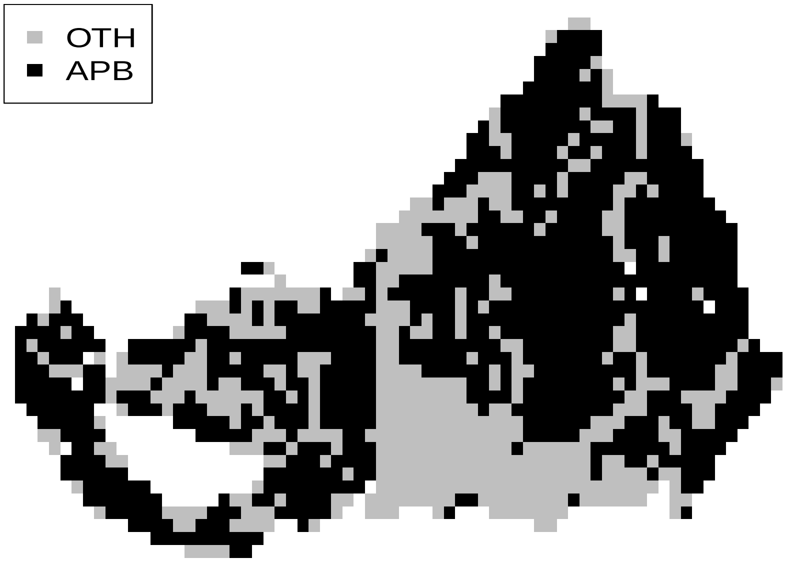

Figure 2.1: Map of the response variable in a study area of northern Wisconsin: Each of the 1,429 quarter sections is classified to be aspen-paper birch (APB) or other (OTH).

We will illustrate the method by a study in the intersection of landscape ecology and environmental history. An overarching goal of this study is to assess the influence of social, economic, and historical factors on landscape structure. The researchers are particularly interested in identifying and quantifying relationships between land cover types and land ownership characteristics. The study site is located in northern Wisconsin and is partitioned into 1,429 quarter sections (160 acre (65 ha) cells) (Figure 2.1). The spatial binary response is whether the dominant land cover type of a quarter section is aspen-paper birch or not, and the covariates are various measures of land ownership. As the study region expands in the future to the state or regional level, data could be composed of tens of thousands of quarter sections and thus require analytical tools that are computationally feasible.

Autologistic regression is such a data analytical tool. For an autologistic model (without regression), maximum pseudolikelihood (MPL) estimation is straightforward to implement and fast to compute (Besag, 1972, 1974). Maximum likelihood estimation is more challenging because of a normalizing constant that is analytically intractable in the likelihood function. The idea of Monte Carlo maximum likelihood (MCML) can be adopted to deal with the normalizing constant problem (Geyer, 1994). Various researchers have considered Bayesian estimation using innovations such as auxiliary variables (Moller et al., 2006), Monte Carlo approximations (Sun and Clayton, 2008), and other approximation methods (Friel et al., 2009). These methods tend to focus on autoregression but not as much on regression. For an autologistic model with regression, the idea of MPL can be applied (Gumpertz et al., 1977) as well as MCML (Huffer and Wu, 1998). More recently, Caragea and Kaiser (2009) suggested to center the autocovariates in the model, which allows more meaningful interpretation of the coefficients in the autologistic model. Here we will consider the centering suggested by Caragea and Kaiser (2009) and focus on MPL.

Compared with MCML, MPL is much faster to compute, which can be a great advantage for the analysis of big data (Hughes et al., 2011; Wang and Zheng, 2012; Y. and J., 2008). There are, however, at least two issues that appear to be unresolved. First, there appear to be no existing methods for the selection of covariates using

MPL. For autologistic regression, Zhu et al. (2008) applied MCML for regression and used an approximate information criterion for model comparison. When the number of candidate models is large, however, the usefulness of information criteria can be fairly limited due to a high computational cost. Second, the estimates of the variance of the MPL estimates are largely based on parametric bootstrap where the resamples are simulated using Markov chain Monte Carlo (MCMC) algorithms and can be time consuming as well (Gumpertz et al., 1977; Zhu et al., 2005). There clearly is a need to improve these aspects of the MPL method to make it more suitable for analyzing big ecological data.

Different approaches can be taken for variable selection in a standard linear regression assuming independent response variables. Regularization (or, penalized) methods for standard linear regression are becoming popular using, for example, least absolute shrinkage and selection operator (Lasso) (Tibshirani, 1996), adaptive Lasso (Zou, 2006), and least angle regression (LARS) algorithm (Efron et al., 2004). Penalized methods have also been developed for spatial data with continuous response variables (see, e.g., Huang et al. (2010), Zhu and Liu (2009), Zhu et al. (2010)) but not as much with discrete responses. An exception is Xue et al. (2012) who developed a penalized method for estimation of the Ising model, although the focus was on the spatial interactions not regression. In this chapter, we utilize the idea of adaptive Lasso and develop a penalized pseudolikelihood estimation method for selection of covariates. In addition, unlike the existing approach to estimate the variance of the parameter estimates by bootstrapping, we propose an analytical form based on asymptotic results.

In the remainder of this chapter, we will present the autologistic regression models in Section 3.2 and describe our method in Section 2.3. A simulation study is conducted in Section 3.6 to evaluate the performance of the proposed method, followed by a data example regarding land cover type in relation to land ownership characteristics in Section 3.7. Conclusion and discussion are given in Section 3.9.

2.2 Autologistic Regression Models

Autologistic Regression

For $i = 1, \ldots, n$, let Z_i denote the response variable at the ith site on a spatial lattice, such that $Z_i = 0$ or 1. Let $\boldsymbol{Z} = (Z_1, \ldots, Z_n)'$ denote the binary response variables at all n sites on this lattice and $\boldsymbol{Z}_{-i} = (Z_1, \ldots, Z_{i-1}, Z_i, \ldots, Z_n)'$ denote the vector that has all the response variables of $\boldsymbol{Z}$ except for Z_i. Further, consider a pre-specified spatial neighborhood structure. For example, the first-order neighborhood consists of up to four nearest neighbors on a regular square grid. Let $\mathcal{N}_i$ denote the set of indices of the neighbors of site i, and let $i' \in \mathcal{N}_i$ denote that site i' is a neighbor of site i.

To model the response variables $\boldsymbol{Z}$, we assume that the probability of the ith response, Z_i, conditional on $\boldsymbol{Z}_{-i}$ depends on only the responses in the neighborhood, $Z_{i'}$, where $i' \in \mathcal{N}_i$ (Gaetan and Guyon, 2010); that is,

$$p(Z_i|\boldsymbol{Z}_{-i}) \quad = \quad p(Z_i|Z_{i'} : i' \in \mathcal{N}_i). \tag{2.1}$$

Further, we assume that the conditional distribution $p(Z_i|Z_{i'} : i' \in \mathcal{N}_i)$ is Bernoulli with success probability π_i such that

$$\pi_i \;=\; p(Z_i = 1 | Z_{i'} : i' \in \mathcal{N}_i) \tag{2.2}$$

where π_i depends on $Z_{i'}$ for $i' \in \mathcal{N}_i$ via a logit link function

$$\text{logit}(\pi_i) \;=\; \log \frac{\pi_i}{1-\pi_i} = \boldsymbol{x}_i' \boldsymbol{\beta} + \sum_{i' \in \mathcal{N}_i} \eta_{ii'} Z_{i'}. \tag{2.3}$$

In (2.3), $\boldsymbol{x}_i$ denotes a $(p+1)$-dimensional vector of intercept 1 and p covariates at site i, and $\boldsymbol{\beta}$ denotes the corresponding vector of the regression coefficients. Also, $\eta_{ii'}$ for $i' \in \mathcal{N}_i$ denotes the autoregression coefficient between sites i and i' and the term $Z_{i'}$ can be thought of as an autocovariate. In the special case that all the autoregression coefficients are zero ($\eta_{ii'} = 0$), the model (2.1)–(2.3) reduces to a traditional logistic regression with independent responses.

We will restrict our attention to a constant autoregression coefficient $\eta_{ii'} = \eta$ for all $i' \in \mathcal{N}_i$ and the logit link becomes

$$\text{logit}(\pi_i) \;=\; \log \frac{\pi_i}{1-\pi_i} = \boldsymbol{x}_i' \boldsymbol{\beta} + \eta \sum_{i' \in \mathcal{N}_i} Z_{i'}. \tag{2.4}$$

However, this assumption may be relaxed to include different orders of neighborhood.

Centered Model

Caragea and Kaiser (2009) proposed to center the autocovariate around its expected value to achieve more meaningful interpretations for regression purposes. Here we consider this centered autologistic regression model by modifying (2.4) to

$$\mathrm{logit}(\pi_i) \quad = \quad \log \frac{\pi_i}{1-\pi_i} = \boldsymbol{x}_i'\boldsymbol{\beta} + \eta \sum_{i' \in \mathcal{N}_i} (Z_{i'} - \mu_{i'}), \tag{2.5}$$

where $\mu_{i'}$ is the expectation of $Z_{i'}$ assuming independence (i.e., $\eta = 0$) and thus

$$\mu_{i'} \quad = \quad \frac{\exp(\boldsymbol{x}_{i'}'\boldsymbol{\beta})}{1 + \exp(\boldsymbol{x}_{i'}'\boldsymbol{\beta})}.$$

The terms $\{\mu_i\}_{i=1}^n$ may be interpreted as the large-scale structure of the random field $\boldsymbol{Z}$ and with regression, they relate the response variables to covariates. The difference $Z_i - \mu_i$ represents the small-scale structure after adjusting for the large-scale structure. The effects of covariates are therefore captured in the regression coefficients $\boldsymbol{\beta}$ and the local spatial dependence in the autoregression coefficient η. In contrast, the uncentered autologistic regression model (2.3) does not permit such a clear interpretation (Caragea and Kaiser, 2009).

Alternative Coding

Although it is common to use 0 or 1 to code failure or success in the Bernoulli trial, in the uncentered model, the positive-valued autocovariates would artificially inflate the chance of success (Hughes et al., 2011). Here we consider an alternative coding is -1 for failure and $+1$ for success (Gaetan and Guyon, 2010), which can help address this issue with 0-1 coding. When $\eta > 0$, the conditional probability of a "presence" would increase when there are more $+1$ neighbors than -1 neighbors and would decrease when there are more -1 neighbors than $+1$ neighbors and vice versa for $\eta < 0$.

Here, we let $\tilde{Z}_i = 2Z_i - 1$ denote the binary response variable at site i under this alternative coding. Define $\tilde{\boldsymbol{Z}}$ and $\tilde{\boldsymbol{Z}}_{-i}$ as the counterparts of $\boldsymbol{Z}$ and $\boldsymbol{Z}_{-i}$. Let $\tilde{\pi}_i = P(\tilde{Z}_i = +1 | \tilde{Z}_{i'} : i' \in \mathcal{N}_i)$ denote the success probability. The logit link function in the uncentered model is

$$\text{logit}(\tilde{\pi}_i) = \log \frac{\tilde{\pi}_i}{1 - \tilde{\pi}_i} = \boldsymbol{x}_i' \tilde{\boldsymbol{\beta}} + \tilde{\eta} \sum_{i' \in \mathcal{N}_i} \tilde{Z}_{i'}, \tag{2.6}$$

but in the centered model is

$$\text{logit}(\tilde{\pi}_i) = \log \frac{\tilde{\pi}_i}{1 - \tilde{\pi}_i} = \boldsymbol{x}_i' \tilde{\boldsymbol{\beta}} + \tilde{\eta} \sum_{i' \in \mathcal{N}_i} (\tilde{Z}_{i'} - \tilde{\mu}_{i'}), \tag{2.7}$$

where it can be shown that the expectation of $\tilde{Z}_i$ assuming independence is

$$\tilde{\mu}_{i'} = \frac{\exp(2\boldsymbol{x}_{i'}' \tilde{\boldsymbol{\beta}}) - 1}{\exp(2\boldsymbol{x}_{i'}' \tilde{\boldsymbol{\beta}}) + 1}.$$

In fact, there is a one-to-one correspondence between the coefficients under the two different codings. For uncentered models, we have $\tilde{\eta} = \eta/4$ and $\boldsymbol{x}_i' \tilde{\boldsymbol{\beta}} = \boldsymbol{x}_i' \boldsymbol{\beta}/2 + |\mathcal{N}_i| \eta/4$, where $|\mathcal{N}_i|$ denotes the cardinality of $\mathcal{N}_i$ (i.e., the number of neighbors of the ith site). A similar connection can be made for centered models. See Appendix A in the supplementary materials for more details.

2.3 Estimation and Selection of Autologistic Regression Models

Maximum Pseudolikelihood

Let $\boldsymbol{\theta} = (\boldsymbol{\beta}', \eta)'$ denote the vector of all the parameters for either centered or uncentered models using the 0-1 coding of the response variable. Let $\tilde{\boldsymbol{\theta}} = (\tilde{\boldsymbol{\beta}}', \tilde{\eta})'$ be analogous to $\boldsymbol{\theta}$ but for the models using ±1 coding. We use maximum pseudolikelihood here for estimating the model parameters such that the pseudolikelihood function is the product of the full conditional probabilities at all sites (Cressie, 1993).

In an uncentered model, the full conditional distribution of Z_i is

$$p(Z_i = 1|\boldsymbol{Z}_{-i};\boldsymbol{\theta}) = p(Z_i = 1|Z_{i'} : i' \in \mathcal{N}_i;\boldsymbol{\theta}) = \frac{\exp(\boldsymbol{x}_i'\boldsymbol{\beta} + \eta\sum_{i'\in\mathcal{N}_i} Z_{i'})}{1+\exp(\boldsymbol{x}_i'\boldsymbol{\beta} + \eta\sum_{i'\in\mathcal{N}_i} Z_{i'})} \qquad (2.8)$$

and $p(Z_i = 0|\boldsymbol{Z}_{-i};\boldsymbol{\theta}) = 1/\{1+\exp(\boldsymbol{x}_i'\boldsymbol{\beta}+\eta\sum_{i'\in\mathcal{N}_i} Z_{i'})\}$. Thus, the log-pseudolikelihood function for the uncentered model is

$$\ell_{\mathrm{p}}^{\mathrm{u}}(\boldsymbol{\theta}|\boldsymbol{Z}) \quad = \quad \sum_{i=1}^{n} \log \frac{\exp\{Z_i(\boldsymbol{x}_i'\boldsymbol{\beta} + \eta\sum_{i'\in\mathcal{N}_i} Z_{i'})\}}{1+\exp(\boldsymbol{x}_i'\boldsymbol{\beta} + \eta\sum_{i'\in\mathcal{N}_i} Z_{i'})}. \qquad (2.9)$$

Similarly, in a centered model, the full conditional distribution of Z_i is

$$p(Z_i = 1|\boldsymbol{Z}_{-i};\boldsymbol{\theta}) = \frac{\exp\{\boldsymbol{x}_i'\boldsymbol{\beta} + \eta\sum_{i'\in\mathcal{N}_i}(Z_{i'} - \mu_{i'})\}}{1+\exp\{\boldsymbol{x}_i'\boldsymbol{\beta} + \eta\sum_{i'\in\mathcal{N}_i}(Z_{i'} - \mu_{i'})\}} \qquad (2.10)$$

and thus, the log-pseudolikelihood function is

$$\ell_{\mathrm{p}}^{\mathrm{c}}(\boldsymbol{\theta}|\boldsymbol{Z}) \quad = \quad \sum_{i=1}^{n} \log \frac{\exp[Z_i\{\boldsymbol{x}_i'\boldsymbol{\beta} + \eta \sum_{i' \in \mathcal{N}_i}(Z_{i'} - \mu_{i'})\}]}{1 + \exp\{\boldsymbol{x}_i'\boldsymbol{\beta} + \eta \sum_{i' \in \mathcal{N}_i}(Z_{i'} - \mu_{i'})\}}. \qquad (2.11)$$

Maximizing the log-pseudolikelihood function gives the maximum pseudolikelihood estimate (MPLE) of $\boldsymbol{\theta}$ and we denote them as $\widehat{\boldsymbol{\theta}}_{\mathrm{p}} = \mathrm{argmax}\{\ell_{\mathrm{p}}(\boldsymbol{\theta}|\boldsymbol{Z})\}$ for either $\ell_{\mathrm{p}}^{\mathrm{u}}(\boldsymbol{\theta}|\boldsymbol{Z})$ in (2.9) or $\ell_{\mathrm{p}}^{\mathrm{c}}(\boldsymbol{\theta}|\boldsymbol{Z})$ in (2.11).

Under the alternative ±1 coding of responses, the full conditional distribution of $\tilde{Z}_i$ in an uncentered model is

$$p(\tilde{Z}_i = 1|\tilde{\boldsymbol{Z}}_{-i}; \tilde{\boldsymbol{\theta}}) \quad = \quad \frac{\exp(\boldsymbol{x}_i'\tilde{\boldsymbol{\beta}} + \tilde{\eta} \sum_{i' \in \mathcal{N}_i} \tilde{Z}_{i'})}{2\cosh(\boldsymbol{x}_i'\tilde{\boldsymbol{\beta}} + \tilde{\eta} \sum_{i' \in \mathcal{N}_i} \tilde{Z}_{i'})}$$

and in a centered model is

$$p(\tilde{Z}_i = 1|\tilde{\boldsymbol{Z}}_{-i}; \tilde{\boldsymbol{\theta}}) \quad = \quad \frac{\exp\{\boldsymbol{x}_i'\tilde{\boldsymbol{\beta}} + \tilde{\eta} \sum_{i' \in \mathcal{N}_i}(\tilde{Z}_{i'} - \tilde{\mu}_{i'})\}}{2\cosh\{\boldsymbol{x}_i'\tilde{\boldsymbol{\beta}} + \tilde{\eta} \sum_{i' \in \mathcal{N}_i}(\tilde{Z}_{i'} - \tilde{\mu}_{i'})\}}.$$

The log-pseudolikelihood function for the uncentered model is

$$\ell_{\mathrm{p}}^{\mathrm{u}}(\tilde{\boldsymbol{\theta}}) \quad = \quad \sum_{i=1}^{n} \log \frac{\exp\{\tilde{Z}_i(\boldsymbol{x}_i'\tilde{\boldsymbol{\beta}} + \tilde{\eta} \sum_{i' \in \mathcal{N}_i} \tilde{Z}_{i'})\}}{2\cosh(\boldsymbol{x}_i'\tilde{\boldsymbol{\beta}} + \tilde{\eta} \sum_{i' \in \mathcal{N}_i} \tilde{Z}_{i'})} \qquad (2.12)$$

and for the centered model is

$$\ell_{\mathrm{p}}^{\mathrm{c}}(\tilde{\boldsymbol{\theta}}) \quad = \quad \sum_{i=1}^{n} \log \frac{\exp[\tilde{Z}_i\{\boldsymbol{x}_i'\tilde{\boldsymbol{\beta}} + \tilde{\eta} \sum_{i' \in \mathcal{N}_i}(\tilde{Z}_{i'} - \tilde{\mu}_{i'})\}]}{2\cosh\{\boldsymbol{x}_i'\tilde{\boldsymbol{\beta}} + \tilde{\eta} \sum_{i' \in \mathcal{N}_i}(\tilde{Z}_{i'} - \tilde{\mu}_{i'})\}}. \qquad (2.13)$$

We denote the MPLE of $\tilde{\boldsymbol{\theta}}$ by $\widehat{\tilde{\boldsymbol{\theta}}}_{\mathrm{p}} = \mathrm{argmax}\{\ell_{\mathrm{p}}(\tilde{\boldsymbol{\theta}}|\tilde{\boldsymbol{Z}})\}$ for either $\ell_{\mathrm{p}}^{\mathrm{u}}(\tilde{\boldsymbol{\theta}}|\tilde{\boldsymbol{Z}})$ in (2.12) or

$\ell_{\mathrm{p}}^{\mathrm{c}}(\tilde{\boldsymbol{\theta}}|\tilde{\boldsymbol{Z}})$ in (2.13).

The log-pseudolikelihood functions for the uncentered models $\ell_{\mathrm{p}}^{\mathrm{u}}(\boldsymbol{\theta}|\boldsymbol{Z})$ in (2.9) and $\ell_{\mathrm{p}}^{\mathrm{u}}(\tilde{\boldsymbol{\theta}})$ in (2.12) are concave, and thus a global maximum is achievable. Because of the additional complexity introduced by the centering term, a global maximum is not guaranteed for the centered models, and thus extra care will be needed in the optimization. We will discuss this further in the numerical examples in Sections 3.6 and 3.7.

Variable Selection

For an uncentered or centered model using the 0-1 coding of the response, we propose a penalized log-pseudolikelihood function

$$\ell_{\mathrm{pp}}(\boldsymbol{\theta}) = \ell_{\mathrm{p}}(\boldsymbol{\theta}) - n\sum_{j=1}^{p} \lambda_j |\beta_j| \tag{2.14}$$

where in the second term of (2.14), the regression coefficients are subject to an L_1-penalty function and λ_j is a regularization parameter for the jth regression coefficient β_j, where $j = 1, \ldots, p$. That is, no penalty is applied to the intercept β_0 nor the autoregression coefficient η. Maximizing (2.14) would enable variable selection and parameter estimation simultaneously, because the regression coefficients of the "less important" covariates will be shrunk to zero under the L_1-penalty and nonzero estimates of the other parameters will be given. Because the regularization parameter λ_j varies by j, the penalization is in the spirit of adaptive Lasso (Zou, 2006).

To maximize (2.14), we propose a one-step approximation. We first set the MPLE of $\boldsymbol{\theta}$ to be the initial value. That is, $\widehat{\boldsymbol{\theta}}^{(0)} = \widehat{\boldsymbol{\theta}}_{\mathrm{p}} \equiv (\widehat{\boldsymbol{\beta}}^{(0)\prime}, \widehat{\eta})'$. We approximate the penalized log-pseudolikelihood function (2.14) up to a constant by

$$\begin{aligned}\ell_{\mathrm{pp}}(\boldsymbol{\beta}) &= (\boldsymbol{\beta} - \widehat{\boldsymbol{\beta}}^{(0)})' \frac{\partial \ell_{\mathrm{p}}(\widehat{\boldsymbol{\theta}}^{(0)})}{\partial \boldsymbol{\beta}} - (1/2)(\boldsymbol{\beta} - \widehat{\boldsymbol{\beta}}^{(0)})' \boldsymbol{\mathcal{I}}(\widehat{\boldsymbol{\beta}}^{(0)})(\boldsymbol{\beta} - \widehat{\boldsymbol{\beta}}^{(0)}) \\ &\quad - n \sum_{j=1}^{p} \lambda_j |\beta_j|\end{aligned}$$

where $\boldsymbol{\mathcal{I}}(\boldsymbol{\beta}) = -\frac{\partial^2 \ell_{\mathrm{p}}(\boldsymbol{\theta})}{\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}'}$ is the negative of the second-order derivative of $\ell_{\mathrm{p}}(\boldsymbol{\theta})$ with respect to $\boldsymbol{\beta}$. We update $\widehat{\boldsymbol{\beta}}^{(0)}$ to be $\widehat{\boldsymbol{\beta}}^{(1)} = \arg\max_{\boldsymbol{\beta}} \{\ell_{\mathrm{pp}}(\boldsymbol{\beta})\}$. It can be shown that this solution can be attained equivalently by

$$\widehat{\boldsymbol{\beta}}^{(1)} = \arg\min_{\boldsymbol{\beta}} \left\{ (1/2)(\boldsymbol{y}^* - \boldsymbol{X}^* \boldsymbol{\beta}^*)'(\boldsymbol{y}^* - \boldsymbol{X}^* \boldsymbol{\beta}^*) + n \sum_{j=1}^{p} |\beta_j^*| \right\} \qquad (2.15)$$

where

$$\begin{aligned}\boldsymbol{y}^* &= (\boldsymbol{B}^{-1})' \left\{ \frac{\partial \ell_{\mathrm{p}}(\widehat{\boldsymbol{\theta}}^{(0)})}{\partial \boldsymbol{\beta}} + \boldsymbol{\mathcal{I}}(\widehat{\boldsymbol{\beta}}^{(0)}) \widehat{\boldsymbol{\beta}}^{(0)} \right\}, \boldsymbol{X}^* = \boldsymbol{B} \mathrm{diag}\{\lambda_j^{-1}\}_{j=1}^{p}, \\ \boldsymbol{\beta}^* &= \mathrm{diag}\{\lambda_j\}_{j=1}^{p} \boldsymbol{\beta},\ \boldsymbol{\mathcal{I}}(\widehat{\boldsymbol{\beta}}^{(0)}) = \boldsymbol{B}'\boldsymbol{B}.\end{aligned}$$

The minimization in (2.15) can be solved by a LARS algorithm and the computation is generally fast. We approximate the maximum penalized pseudolikelihood estimate (MPPLE), or $\widehat{\boldsymbol{\beta}}_{\mathrm{pp}}$, by $\widehat{\boldsymbol{\beta}}^{(1)}$ and let $\widehat{\boldsymbol{\theta}}_{\mathrm{pp}} = (\widehat{\boldsymbol{\beta}}'_{\mathrm{pp}}, \widehat{\eta})'$. In particular, some of the entries of $\widehat{\boldsymbol{\beta}}_{\mathrm{pp}}$ are zero (or nonzero), indicating that the corresponding covariates are selected out of (or kept in) the final model.

In addition, for estimating the regularization parameters $\{\lambda_j\}_{j=1}^{p}$, we let $\lambda_j =$

$\lambda \log(n)/(n|\widehat{\beta}_j|)$ and compute a Bayesian information criterion (BIC) type criterion for determining λ. In particular, define

$$\mathrm{BIC}_{\mathrm{p}}(\lambda) = -2\ell_{\mathrm{p}}(\widehat{\boldsymbol{\theta}}_{\mathrm{pp}};\lambda) + e(\lambda)\log(n)$$

where $e(\lambda)$ denotes the number of nonzero estimates in $\widehat{\boldsymbol{\beta}}_{\mathrm{pp}}$. We estimate λ by $\widehat{\lambda} = \mathrm{argmin}_{\lambda}\{\mathrm{BIC}_{\mathrm{p}}(\lambda)\}$. Even though $\mathrm{BIC}_{\mathrm{p}}(\lambda)$ is not the actual BIC under maximum likelihood, it is in the same spirit, and our empirical study via simulation will demonstrate that the method works fairly well as a variable selection technique.

For models using ±1 as the response variables, the MPPLE of $\tilde{\boldsymbol{\theta}}$ can be obtained analogously by considering $\ell_{\mathrm{p}}(\tilde{\boldsymbol{\theta}})$ and applying the same penalty as in (2.14). The resulting maximum penalized pseudolikelihood estimates are denoted as $\widehat{\tilde{\boldsymbol{\theta}}}_{\mathrm{pp}} = (\widehat{\tilde{\boldsymbol{\beta}}}'_{\mathrm{pp}}, \widehat{\tilde{\eta}})'$.

Variance Estimation

Let $\boldsymbol{\mathcal{J}}(\boldsymbol{\theta}) = \sum_{i=1}^{n} \sum_{i' \in \mathcal{N}_i, i'=i} \frac{\partial \ell_{\mathrm{p}i}(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} \{\frac{\partial \ell_{\mathrm{p}i'}(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}\}'$, where $\ell_{\mathrm{p}i}$ is the log-pseudolikelihood of site i. By arguments similar to Comets and Janžura (1998), the following central limit theorem holds for the MPLE $\widehat{\boldsymbol{\theta}}_{\mathrm{p}}$:

$$\{\boldsymbol{\mathcal{J}}(\widehat{\boldsymbol{\theta}}_{\mathrm{p}})\}^{-1/2}\boldsymbol{\mathcal{I}}(\widehat{\boldsymbol{\theta}}_{\mathrm{p}})\{\widehat{\boldsymbol{\theta}}_{\mathrm{p}} - \boldsymbol{\theta}\} \to_d \ \mathcal{N}_p(0, \boldsymbol{I}_{p+2}).$$

Therefore, an estimate of the variance of $\widehat{\boldsymbol{\theta}}_{\mathrm{p}}$ is

$$\widehat{Var}(\widehat{\boldsymbol{\theta}}_{\mathrm{p}}) \approx \boldsymbol{\mathcal{I}}(\widehat{\boldsymbol{\theta}}_{\mathrm{p}})^{-1}\boldsymbol{\mathcal{J}}(\widehat{\boldsymbol{\theta}}_{\mathrm{p}})\boldsymbol{\mathcal{I}}(\widehat{\boldsymbol{\theta}}_{\mathrm{p}})^{-1}. \tag{2.16}$$

With the alternative ±1 coding, the variance estimation can be obtained in analogy to the case of 0-1 coding:

$$\widehat{Var}(\widehat{\tilde{\boldsymbol{\theta}}}_{\mathrm{p}}) \approx \boldsymbol{\mathcal{I}}(\widehat{\tilde{\boldsymbol{\theta}}}_{\mathrm{p}})^{-1}\boldsymbol{\mathcal{J}}(\widehat{\tilde{\boldsymbol{\theta}}}_{\mathrm{p}})\boldsymbol{\mathcal{I}}(\widehat{\tilde{\boldsymbol{\theta}}}_{\mathrm{p}})^{-1}. \tag{2.17}$$

For the MPPLE, we replace the MPLE $\widehat{\boldsymbol{\theta}}_{\mathrm{p}}$ (or $\widehat{\tilde{\boldsymbol{\theta}}}_{\mathrm{p}}$) in the variance formula (2.16) (or(2.17)) with the vector of non-zero entries of the MPPLE $\widehat{\boldsymbol{\theta}}_{\mathrm{p}p}$ (or $\widehat{\tilde{\boldsymbol{\theta}}}_{\mathrm{p}p}$). Thus, the operations involved in the variance estimation are of dimension up to $(p+2)\times(p+2)$ and generally manageable. More technical details can be found in Appendix B in the supplementary materials.

2.4 Simulation Study

Simulation Set-up

We conducted a simulation study to examine the finite-sample properties of the method developed in Sections 3.2–2.3. Consider an $m \times m$ square lattice, where $m = 15$ or 30, corresponding to sample sizes $n = 225$ or 900. For spatial dependence, the neighborhood structure is of the first order, and the autoregression coefficient η is either 0.3 or 0.7, corresponding to weaker or stronger spatial dependence.

Let $\boldsymbol{u}_j = (u_{j1}, \ldots, u_{jn})'$ denote the jth covariate vector such that $\{u_{ji} : i = 1, \ldots, n\}$ is a Gaussian random field with mean 0 and an exponential covariance function

$$Cov(u_{ji}, u_{ji'}) \quad = \quad \sigma^2 \exp(-|i - i'|/\tau),$$

where we let the variance parameter be $\sigma^2 = 1$ and the range parameter be $\tau = 0.1$. To obtain cross-covariate correlation, let $\boldsymbol{u}_i = (u_{1i}, \ldots, u_{pi})'$ and $\boldsymbol{x}_i = \boldsymbol{A}\boldsymbol{u}_i$ for site i, where $\boldsymbol{A}\boldsymbol{A'} = [\rho^{|j-j'|}]_{j,j'=1}^{p}$ and $\rho = 0.4$.

Let $p = 10$ be the number of covariates. The regression coefficients are set to be $\boldsymbol{\beta} = (1, \beta_1, 1, 1, \boldsymbol{0}_7')'$, that is, 3 out of 10 coefficients are non-zero, and the remaining 7 coefficients are zero. For the 0-1 coding, we adopt the notion of an average large-scale structure as the average of μ_i over all sites and covariates (Caragea and Kaiser, 2009). Let

$$\bar{\mu} = \frac{1}{n}\sum_{i=1}^{n} \mu_i = \frac{1}{n}\sum_{i=1}^{n} \frac{\exp(\boldsymbol{x}_i'\boldsymbol{\beta})}{1 + \exp(\boldsymbol{x}_i'\boldsymbol{\beta})}$$

The large-scale structure is considered to be weak when $\bar{\mu}$ is around 0.5 and strong otherwise. For the ±1 coding, we define the average large-scale structure analogously as

$$\bar{\tilde{\mu}} = \frac{1}{n}\sum_{i=1}^{n} \tilde{\mu}_i = \frac{1}{n}\sum_{i=1}^{n} \frac{\exp(2\boldsymbol{x}_i'\tilde{\boldsymbol{\beta}}) - 1}{\exp(2\boldsymbol{x}_i'\tilde{\boldsymbol{\beta}}) + 1}$$

In this case, the large-scale structure is considered to be weak if $\bar{\tilde{\mu}}$ is close to 0 but strong otherwise. Here, we let $\beta_1 = 1$ or 5, which corresponds to a stronger or weaker large-scale structure, respectively.

For each combination of m, β_1, and η, $S = 100$ samples were generated from the uncentered and centered model. In particular, the data were simulated by a Gibbs sampler using the full conditional probabilities in (2.8) and (2.10) (Wasserman, 2003). Since the initial value for the one-step approximation algorithm may not be in a strictly concave area of the likelihood function, the matrix $\boldsymbol{\mathcal{I}}(\cdot)$ is not always positive definite. This occurred occasionally to only the centered model, in which case we modified the $\boldsymbol{\mathcal{I}}(\cdot)$ matrix by adding a positive diagonal matrix (Nocedal and Wright, 2000). In addition, for comparison, we consider a nonconvex penalty function, the smoothly clipped absolute deviation (SCAD), and apply a one-step sparse estimation method (Zou and Li, 2008).

Simulation Results

Table 2.1: Average number of correctly identified non-zero and zero regression coefficients by adaptive Lasso (AL) or smoothly clipped absolute deviation (SCAD) when $\beta_1 = 5$ (weak large-scale structure) for uncentered and centered model, sample size $n = 225$ or 900, and antoregression coefficient $\eta = 0.3$ or 0.7.

	$\{\beta_j\}$	Number of non-zero estimates				Number of zero estimates			
Model	n	$\eta = 0.3$		$\eta = 0.7$		$\eta = 0.3$		$\eta = 0.7$	
		AL	SCAD	AL	SCAD	AL	SCAD	AL	SCAD
Uncentered	225	2.77	2.89	2.69	2.90	6.12	5.98	6.14	5.74
	900	3.00	3.00	3.00	3.00	6.81	5.95	6.88	6.19
Centered	225	2.75	2.85	2.67	2.87	6.21	4.87	6.17	4.74
	900	3.00	3.00	3.00	3.00	6.87	5.74	6.82	5.59

Tables 2.1–2.2 provide the results of variable selection for sample size $n = 225$ and 900 in terms of the average numbers of correctly identified zero and non-zero regression coefficients. The true number of non-zero and zero regression coefficients

Table 2.2: Average number of correctly identified non-zero and zero regression coefficients by adaptive Lasso (AL) or smoothly clipped absolute deviation (SCAD) when $\beta_1 = 1$ (strong large-scale structure) for uncentered and centered model, sample size $n = 225$ or 900, and antoregression coefficient $\eta = 0.3$ or 0.7.

	$\{\beta_j\}$	Number of non-zero estimates				Number of zero estimates			
Model	n	$\eta = 0.3$		$\eta = 0.7$		$\eta = 0.3$		$\eta = 0.7$	
		AL	SCAD	AL	SCAD	AL	SCAD	AL	SCAD
Uncentered	225	2.94	3.00	2.52	2.99	6.09	5.61	6.12	5.55
	900	3.00	3.00	3.00	3.00	6.88	5.69	6.95	6.43
Centered	225	2.92	3.00	2.90	2.99	6.31	5.03	6.16	4.95
	900	3.00	3.00	3.00	3.00	6.87	5.30	6.92	5.18

are 3 and 7, respectively. When the sample size n is larger, the number of correctly identified zero (or the non-zero) coefficients is closer to the truth. The number of correctly identified non-zero coefficients is closer to the truth compared with zero coefficients. When the sample size is smaller ($n = 225$), the average number of correctly identified zero (and non-zero) regression coefficients is largely closer to the truth when η is smaller corresponding to weaker spatial dependence. For the larger sample ($n = 900$), the effect of η is not as obvious. For either sample size, there is no apparent effect of β_1 (i.e., large-scale structure) on the average number of correctly identified zero or non-zero regression coefficients. Compared to the SCAD penalty, the adaptive Lasso identified the non-zero regression coefficients somewhat less frequently and the zero regression coefficients more frequently. Overall, selection by adaptive Lasso appears to perform somewhat better.

Figures 2.2–2.3 provide the results of parameter estimation for the regression coefficients and the autoregression coefficient for sample sizes $n = 225$ and 900. In general, for both uncentered and centered models, the bias and variance of the

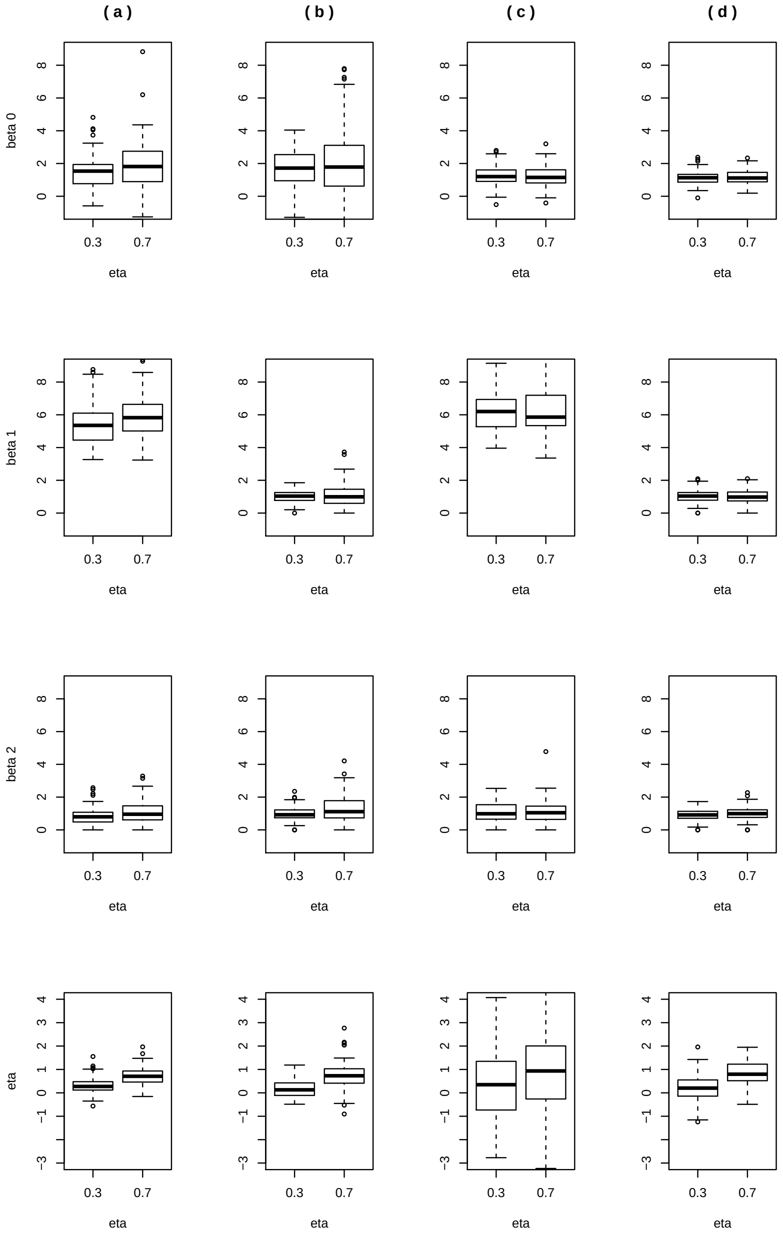

Figure 2.2: Box plot of the MPPLE $\widehat{\beta}_0$ (row 1), $\widehat{\beta}_1$ (row 2), $\widehat{\beta}_2$ (row 3), and $\widehat{\eta}$ (row 4) from the 100 simulations with sample size $n = 225$. Column (a): uncentered model with $\beta_1 = 5$ (weak large-scale structure); (b): uncentered model with $\beta_1 = 1$ (strong large-scale structure); (c): centered model with $\beta_1 = 5$ (weak large-scale structure); (d): centered model with $\beta_1 = 1$ (strong large-scale structure). The true values are $\beta_0 = 1$, $\beta_1 = 1$ or 5, $\beta_2 = 1$, $\beta_3 = 1$, and $\eta = 0.3$ or 0.7. The box plot of $\widehat{\beta}_3$ is similar to $\widehat{\beta}_2$ and omitted to save space.

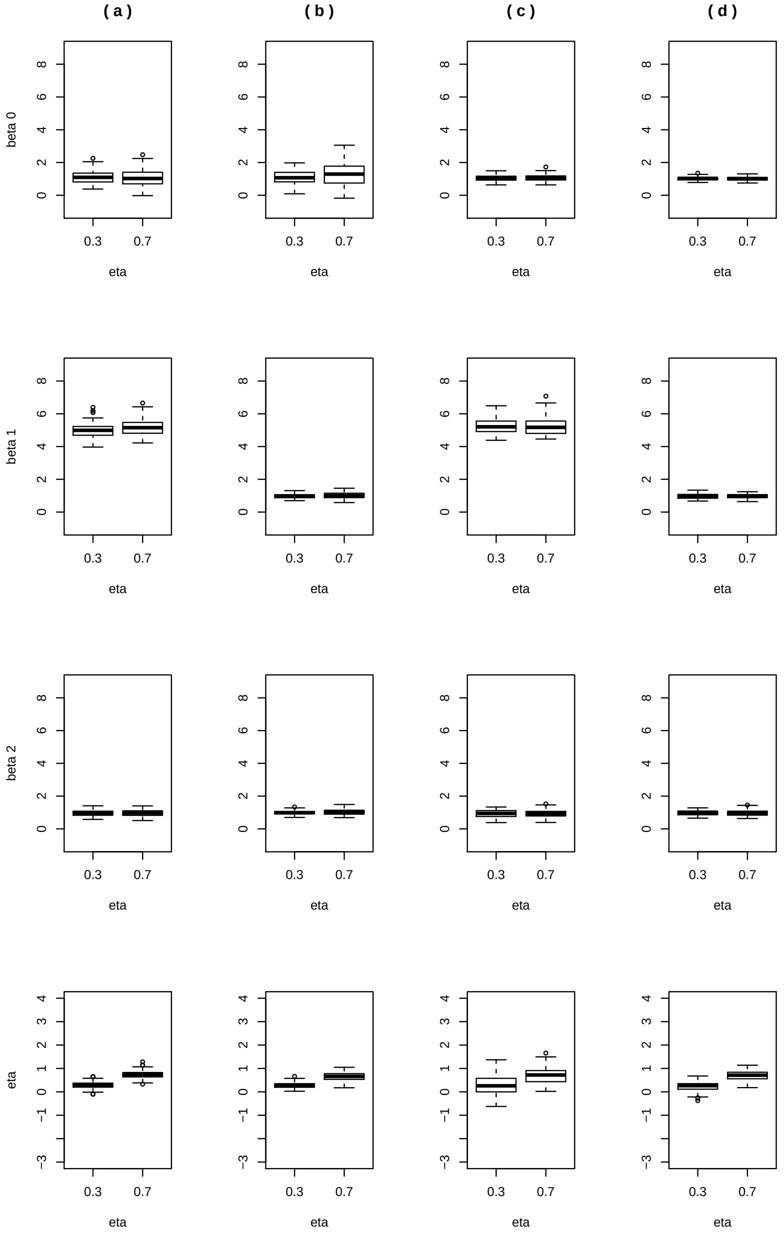

Figure 2.3: Box plot of the MPPLE $\widehat{\beta}_0$ (row 1), $\widehat{\beta}_1$ (row 2), $\widehat{\beta}_2$ (row 3), and $\widehat{\eta}$ (row 4) from the 100 simulations with sample size $n = 900$. Column (a): uncentered model with $\beta_1 = 5$ (weak large-scale structure); (b): uncentered model with $\beta_1 = 1$ (strong large-scale structure); (c): centered model with $\beta_1 = 5$ (weak large-scale structure); (d): centered model with $\beta_1 = 1$ (strong large-scale structure). The true values are $\beta_0 = 1$, $\beta_1 = 1$ or 5, $\beta_2 = 1$, $\beta_3 = 1$, and $\eta = 0.3$ or 0.7. The box plot of $\widehat{\beta}_3$ is similar to $\widehat{\beta}_2$ and omitted to save space.

MPPLE of $\{\beta_j\}$ and η are smaller as the sample size increases from 225 to 900, but are larger when the autoregression coefficient η is larger. Also, the bias and variance are larger when the large-scale structure is weaker. This is plausible, because when the large-scale structure is weaker, the small-scale structure induced by the spatial dependence is relatively stronger.

For each simulated data set, we also computed the standard error of the MPPLE of the nonzero coefficients $\{\beta_j : j = 0, \ldots, 3\}$ by taking the square root of the diagonal elements of the estimated variance matrix (2.16). Figures 2.4–2.5 give the box plots of these standard errors in each of the $S = 100$ simulated datasets, along with the nominal standard error of the MPPLE from these 100 simulations. The bias and variance of standard errors of all the MPPLEs are smaller when the sample size is larger but are larger when η is larger. In particular, the bias and variance of the standard error of the MPPLE of β_1 are larger when β_1 is larger corresponding to a weaker large-scale structure.

The selection and estimation results above are for the 0-1 coding of response variables, and those for the ±1 coding are similar and thus omitted to save space.

2.5 Data Example

In this section, we illustrate our method by a real data example in landscape ecology and environmental history. The study is aimed at assessing the influence of past land ownership characteristics on land cover structure in northern Wisconsin, USA. Landscape ecologists are interested in land ownership because associated variables can

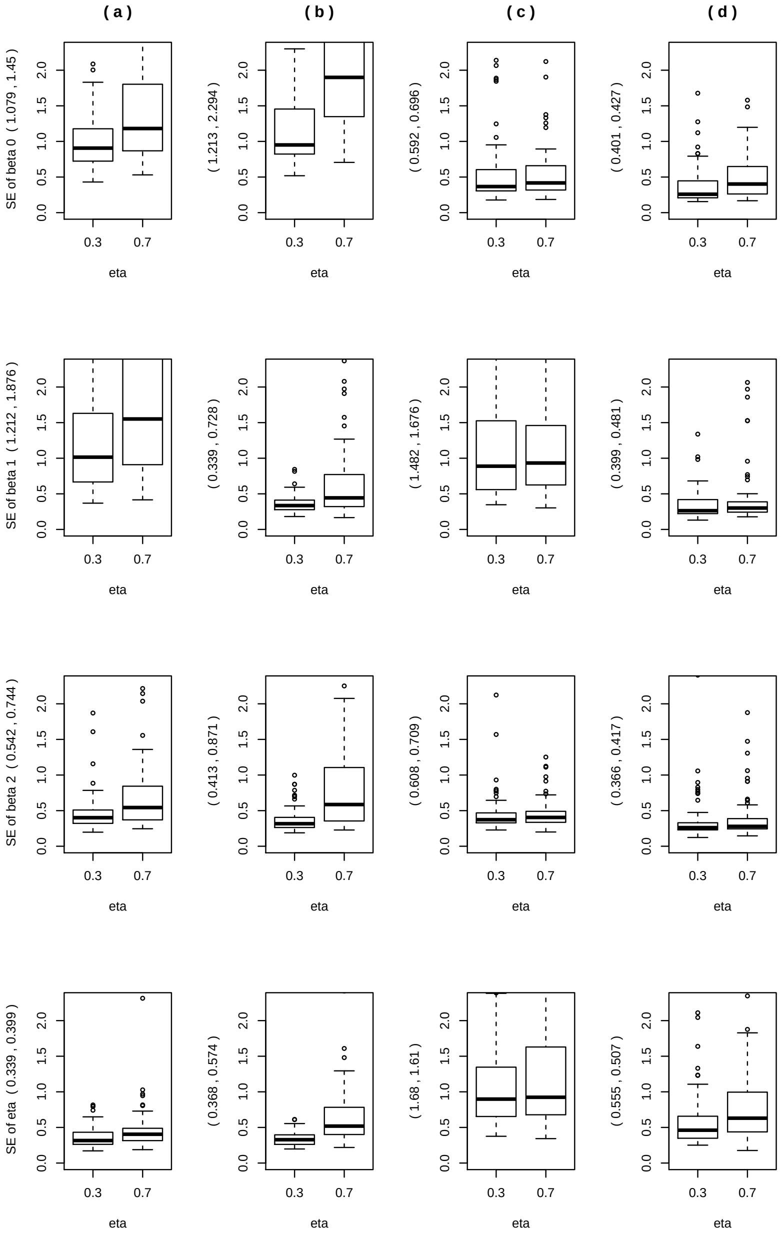

Figure 2.4: Box plot of the standard error of $\widehat{\beta}_0$ (row 1), $\widehat{\beta}_1$ (row 2), $\widehat{\beta}_2$ (row 3), and $\widehat{\eta}$ (row 4) from the 100 simulations with sample size $n = 225$. Column (a): uncentered model with $\beta_1 = 5$ (weak large-scale structure); (b): uncentered model with $\beta_1 = 1$ (strong large-scale structure); (c): centered model with $\beta_1 = 5$ (weak large-scale structure); (d): centered model with $\beta_1 = 1$ (strong large-scale structure). The nominal standard error of the MPPLE from the 100 simulations are given along the y-axis on the left. The box plot of the standard error of $\widehat{\beta}_3$ is similar to the standard error of $\widehat{\beta}_2$ and omitted to save space.

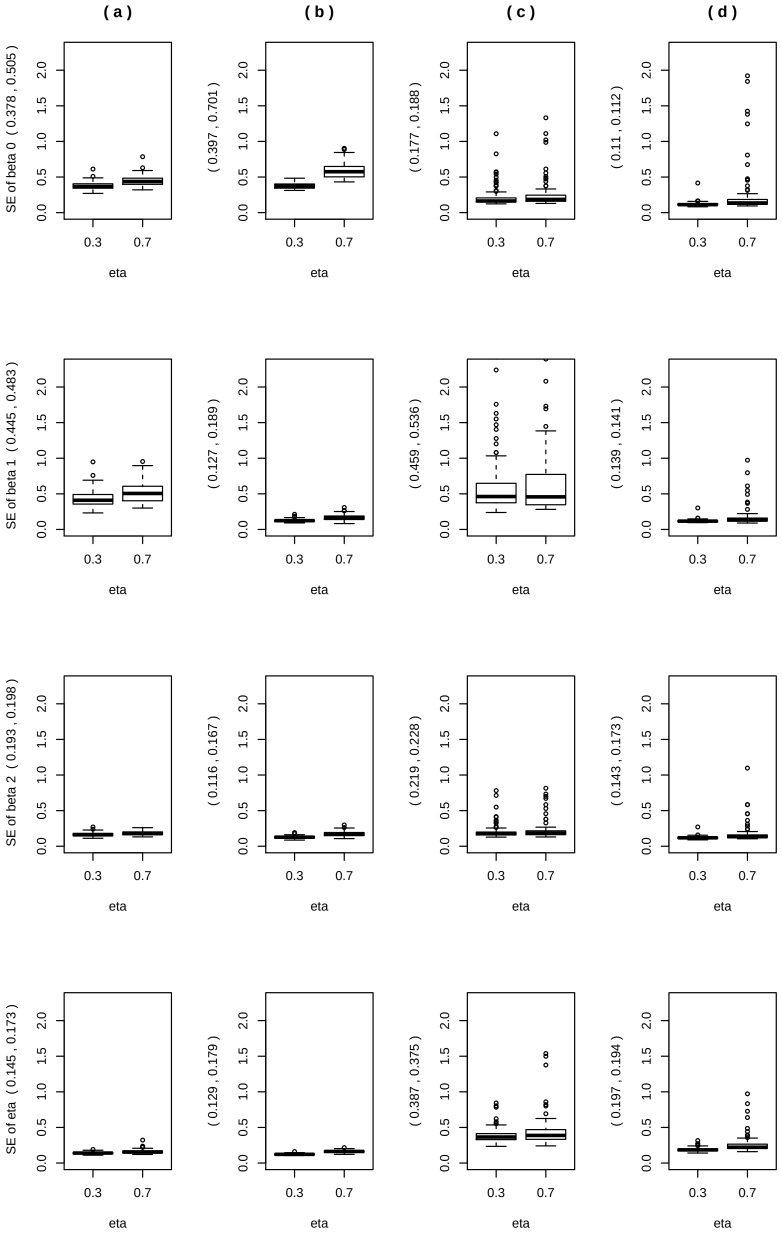

Figure 2.5: Box plot of the standard error of $\widehat{\beta}_0$ (row 1), $\widehat{\beta}_1$ (row 2), $\widehat{\beta}_2$ (row 3), and $\widehat{\eta}$ (row 4) from the 100 simulations with sample size $n = 900$. Column (a): uncentered model with $\beta_1 = 5$ (weak large-scale structure); (b): uncentered model with $\beta_1 = 1$ (strong large-scale structure); (c): centered model with $\beta_1 = 5$ (weak large-scale structure); (d): centered model with $\beta_1 = 1$ (strong large-scale structure). The nominal standard error of the MPPLE from the 100 simulations are given along the y-axis on the left. The box plot of the standard error of $\widehat{\beta}_3$ is similar to the standard error of $\widehat{\beta}_2$ and omitted to save space.

function as a proxy for underlying important, yet spatially imprecise social influences on landscape structure (Turner et al., 1996).

Data were derived from historical plat maps and the Wisconsin Land Economic Inventory, a land survey. The spatial unit of analysis is a quarter section (or, 1/36 township, 160 acres $\approx$ 65 ha) and there are 1429 units in the study area. The original data have multiple categories of land cover (Steen-Adams et al., 2011; Jin et al., 2013). Here, for illustration, we focus on a binary response variable indicating whether the land cover in a quarter section is aspen (*Populus* spp.)-paper birch (*Betula papryfera*) (APB), an early successional forest class, or not (Figure 2.1).

Forest composition can be associated with several land ownership characteristics, including ownership class, parcel size, and ownership size (Crow et al., 1999; Stanfield et al., 2002). The specific covariates of interest are reservation, number of parcels, average size of parcels within a quarter section, proportion of the largest parcel, total area, and average size of all parcels in (but not necessarily all contained within) a quarter section. More specifically, reservation (Reserv) indicates whether the quarter section is on an Indian reservation or not. The number of parcels (PolyNm) in a given quarter section is a measure of parcel density. Average size (PolyPr) is the average size of parcel polygons of a quarter section. Proportion of the largest parcel (MxPolyPr) reports the largest parcel polygon as a proportion of quarter section. Total area (TotOwn) shows the total property area (measured in acres) in this study area associated with owner of the largest parcel polygon in a given quarter section. Because the values of this covariate are right-skewed, we used a log transformation. Average size of all parcels (AvParcel) is the average size of all parcels that lay in whole or in part within a given quarter section (measured in acres). All the covariates except reservation were standardized to have mean 0 and variance 1 before being used in the regression procedure. Moreover, all pairwise interactions were included to investigate the potential for interaction between land ownership characteristics.

We considered the four combinations of uncentered or centered models and the 0-1 or ±1 coding of APB. For autologistic regression, a first-order neighborhood structure was assumed as in section 4. The MPPLE of the regression coefficients are given in Table 2.3, as well as the standard errors for those MPPLE that are non-zero. For the centered models, multiple initial values were tested to facilitate convergence of the computational algorithm. For the two centered models, the same sets of covariates were chosen, and for the two uncentered models, nearly the same sets were chosen. Thus, the coding of the response variable seems to have little effect on selection. However, the sets of covariates selected were very different depending on whether the model was centered or not. Reserv was selected in all four cases, whereas TotOwn and AvParcel were selected only in the uncentered models. For the uncentered models, several interactions were selected, whereas for the centered models, only one interaction was selected. Furthermore, the autoregression coefficient η was consistently estimated to be positive, indicating a positive spatial dependence. The magnitude of the estimates is about four times as large for the 0-1 coding than the ±1 coding. This is perhaps expected, as given in Appendix A, the analytical relation is $\tilde{\eta} = \eta/4$.

We focus our attention on the covariates selected in all four cases. We see that reservation (Reserv) is positively associated with APB. That is, whether a

Table 2.3: Maximum penalized pseudolikelihood estimates (MPPLE) of the regression coefficients and the autoregression coefficient η for uncentered or centered models and 0-1 or ±1 coding of the response variable. The standard errors of MPPLE of the regression coefficients and the autoregression coefficient are given in the parentheses.

	0-1 coding		±1 coding	
Covariates	Uncentered	Centered	Uncentered	Centered
Reserv	2.90 (1.528)	1.53 (0.283)	1.67(0.383)	0.78 (0.061)
PolyNm	–	–	–	–
PolyPr	–	–	–	–
MxPolyPr	–	–	–	–
log(TotOwn)	0.68 (0.374)	–	0.35 (0.065)	–
AvParcl	−0.44 (0.356)	–	−0.23 (0.126)	–
Reserv×PolyNm	−0.38 (0.522)	–	−0.40 (0.087)	–
Reserv×PolyPr	−0.67 (0.319)	–	–	–
Reserv×MxPolyPr	–	–	–	–
Reserv×log(TotOwn)	–	–	–	–
Reserv×AvParcl	–	–	–	–
PolyNm×PolyPr	0.04 (0.067)	–	–	–
PolyNm×MxPolyPr	–	–	–	–
PolyNm×log(TotOwn)	0.03 (0.220)	–	–	–
PolyNm×AvParcl	0.35 (0.264)	–	0.10 (0.057)	–
PolyPr×MxPolyPr	0.83 (0.271)	0.51 (0.117)	0.40 (0.086)	0.27 (0.028)
PolyPr×log(TotOwn)	−0.65 (0.588)	–	−0.39 (0.144)	–
PolyPr×AvParcl	–	–	–	–
MxPolyPr×log(TotOwn)	–	–	–	–
MxPolyPr×AvParcl	–	–	−0.10 (0.091)	–
log(TotOwn)×AvParcl	–	–	–	–
η	1.22 (0.089)	1.39 (0.160)	0.33 (0.001)	0.35 (0.002)

site lies on the Indian reservation or not influences the likelihood that APB is the dominant land cover. This is plausible because sites on the reservation lay within a single administrative unit (the Indian Agency, later the Bureau of Indian Affairs), which historically has had authority over forest management decisions, regardless of ownership type. Indian Agency authority was especially pervasive in the early 20th century (for historical political reasons), which corresponds with the time period represented by the land cover data.

The interaction variable, PolyPr × MxPolyPr, was selected in all four cases, indicating a positive association with APB. This is again plausible. In this particular ecoregion, the land cover class of interest, APB, is strongly associated with a specific forest management practice, i.e., short-rotation, even-aged management. Large landowners are most likely to implement this practice: forest stands on some parcels could be harvested, while those on other parcels could be held for harvest. By distributing harvest rotations across a large land area, forest products could be continuously harvested. This management logic typically assumes large real estate holding, and functions less optimally on smaller ownerships.

The sets of covariates selected in the uncentered cases are complex. For example, the interaction between Reserv and PolyNm or PolyPr indicates that increasing parcel density lessens the effect of reservation on APB. However, interpretation of the regression coefficients may not be intuitive, given that the model is not centered.

2.6 Conclusion and Discussion

In this chapter, we have proposed maximum penalized pseudolikelihood estimation for simultaneous selection of covariates and estimation of parameters. Our simulation study has shown desirable large-sample properties in the sense that the performance of variable selection and parameter estimation improves as the sample size increases. In addition, a variance estimation based on the limiting distribution of the parameter estimates appears to work reasonably well. We have illustrated our method by a data example in the intersection of landscape ecology and environmental history.

Further, we simulated data on larger lattices and with larger numbers of covariates (p) to assess the feasibility of our method. For example, the computation times for maximum penalized pseudolikelihood estimation for $p = 100$ covariates when the lattice size was 60×60 (i.e., $n =$ 3,600) were 22 seconds and 2.3 minutes for the uncentered and centered models, respectively, using R (Team, 2011) on a 64 bit Linux operating system with 8 cores and 48-64 GB of RAM. When we increased the lattice size to 120×120 (i.e., $n =$ 14,400), the computing times were about 5.7 minutes and 49.7 minutes for the uncentered and centered models, respectively. The most time consuming part of the computation is the initial step that computes the MPLE and can be further improved using another programming language like C++.

For future research, it would be interesting to use some of the alternatives for estimation and computation such as Monte Carlo maximum likelihood. It would also be of interest to investigate the selection of the autocovariates and establish asymptotic properties of the model selection such as consistency of our BIC type criterion. A possible approach is to adopt the generalized information criterion framework in Zhang et al. (2010).

Other approaches are possible for modeling spatial binary response in relation to covariates. One possibility is spatial generalized linear mixed models, where the response variable is binary and linked to a regression on covariates and a spatial random effect as two additive components in a link function. Bayesian hierarchical modeling provides a general framework for drawing statistical inference about the regression coefficients and the spatial random effects, while maximum likelihood estimation is also possible (Banerjee et al., 2004; Diggle and Ribeiro, 2007; Rue et al., 2009). To the best of our knowledge, however, the issue of model selection has not always been addressed and is worth further investigation.

Spatial confounding occurs when a spatial covariate competes with residual spatial structure to account for variation in a response variable (Paciorek, 2010). We conducted a small simulation to investigate the effects of spatial confounding. The setup was similar to our simulation study in Section 3.6, but we included an unobserved confounding variable. The confounding variable was allowed to have spatial dependence ranges that were shorter, the same as, or longer than the observed covariates, and we considered two cases where the confounding variable was either more correlated with the covariates with zero coefficients (noise covariates) or more correlated with the covariates with nonzero coefficients (true covariates).

Some of the findings are summarized as follows. First, when the confounding variable was more correlated with a noise covariate, our method tended to detect spurious relationship with that noise covariate more frequently. Second, when the confounding variable was more correlated with a true covariate, the estimates of the coefficients of the true covariates tended to be biased, which is typical in regression studies when covariates related to the response are not included in the fit of a model. Third, as the dependence range of the confounding variable increased, the standard deviations of the coefficient estimates tended to increase for both true and noise covariates. An advantage of our method is that a large number of covariates can be considered and selected simultaneously, which should help to reduce the chance of confounding. However, if confounding is still an issue, further research will be needed to develop ways of alleviating confounding (see, e.g., Paciorek (2010), Hughes and

Haran (2013)).

Finally, we may consider spatial-temporal autologistic models as an extension of the spatial-only model above. For site $i = 1, \ldots, n$ and time $t = 1, \ldots, T$, let $Z_{i,t} = 0$ or 1 denote the response variable and define the set of indices of spatial-temporal neighbors as

$$\begin{aligned} \mathcal{N}_{i,t} &= \{(i', t') : i' \in \mathcal{N}_i, \max\{1, t - L\} \le t' \le \min\{T, t + L\}\} \\ &\cup \{(i', t') : i' = i, t' \neq t, \max\{1, t - L\} \le t' \le \min\{T, t + L\}\} \end{aligned}$$

where $\mathcal{N}_i$ is a pre-specified spatial neighborhood and L denotes the maximum time lag in the temporal neighborhood. That is, the spatial-temporal neighborhood contains the spatial neighbors of site i at time t and up to L time lags in the past and future and site i itself up to L time lags in the past and future. For a given time point t, let $\boldsymbol{Z}_t = (Z_{1,t}, \ldots, Z_{n,t})'$ denote the binary response variables at all n sites.

To model the response variables $\{Z_{i,t} : i = 1, \ldots, n,\ t = 1, \ldots, T\}$, we assume a random field such that

$$p(Z_{i,t} | Z_{i',t'} : (i', t') \neq (i, t)) \;=\; p(Z_{i,t} | Z_{i',t'} : (i', t') \in \mathcal{N}_{i,t}).$$

The conditional distribution $p(Z_{i,t} | Z_{i',t'} : (i', t') \in \mathcal{N}_{i,t})$ is Bernoulli with success probability $\pi_{i,t} = p(Z_{i,t} = 1 | Z_{i',t'} : (i', t') \in \mathcal{N}_{i,t})$.

The logit function of the success probability can be defined analogously to the spatial-only autologistic regression model. For an uncentered model, let

$$logit(\pi_{i,t}) = \log \frac{\pi_{i,t}}{1-\pi_{i,t}} \quad = \quad \boldsymbol{x}'_{i,t}\boldsymbol{\beta} + \sum_{(i',t') \in \mathcal{N}_{i,t}} \eta_{i,i',t,t'} Z_{i',t'}$$

and for a centered model, let

$$logit(\pi_{i,t}) = \log \frac{\pi_{i,t}}{1-\pi_{i,t}} \quad = \quad \boldsymbol{x}'_{i,t}\boldsymbol{\beta} + \sum_{(i',t') \in \mathcal{N}_{i,t}} \eta_{i,i',t,t'} (Z_{i',t'} - \mu_{i',t'})$$

where $\boldsymbol{x}_{i,t}$ denotes the covariate vector at site i and time t and $\mu_{i,t} = \exp(\boldsymbol{x}'_{i,t}\boldsymbol{\beta})/\{1+\exp(\boldsymbol{x}'_{i,t}\boldsymbol{\beta})\}$ is the expectation of $Z_{i,t}$ assuming independence.

Let $\boldsymbol{\theta} = (\boldsymbol{\beta}', \eta_{i,i',t,t'})'$ denote the vector of the regression coefficients $\boldsymbol{\beta}$ and autoregression coefficients $\{\eta_{i,i',t,t'}\}$. The log-pseudolikelihood function for the uncentered model is

$$\ell^{\mathrm{u}}_{\mathrm{p}(st)}(\boldsymbol{\theta}|\boldsymbol{Z}) \quad = \quad \sum_{t=1}^{T}\sum_{i=1}^{n} \log \frac{\exp\{Z_{i,t}(\boldsymbol{x}'_{i,t}\boldsymbol{\beta} + \sum_{(i',t') \in \mathcal{N}_{i,t}} \eta_{i,i',t,t'} Z_{i',t'})\}}{1+\exp(\boldsymbol{x}'_{i,t}\boldsymbol{\beta} + \sum_{(i',t') \in \mathcal{N}_{i,t}} \eta_{i,i',t,t'})}$$

and for the centered model is

$$\ell^{\mathrm{c}}_{\mathrm{p}(st)}(\boldsymbol{\theta}|\boldsymbol{Z}) \quad = \quad \sum_{t=1}^{T}\sum_{i=1}^{n} \log \frac{\exp[Z_{i,t}\{\boldsymbol{x}'_{i,t}\boldsymbol{\beta} + \sum_{(i',t') \in \mathcal{N}_{i,t}} \eta_{i,i',t,t'} (Z_{i',t'} - \mu_{i',t'})\}]}{1+\exp\{\boldsymbol{x}'_{i,t}\boldsymbol{\beta} + \sum_{(i',t') \in \mathcal{N}_{i,t}} \eta_{i,i',t,t'} (Z_{i',t'} - \mu_{i',t'})\}}$$

For simultaneous variable selection and parameter estimation, a penalized log-pseudolikelihood function similar to (2.14) can be defined:

$$\ell_{\mathrm{pp}(st)}(\boldsymbol{\theta}) = \ell_{\mathrm{p}(st)}(\boldsymbol{\theta}) - N \sum_{j=1}^{p} \lambda_j |\beta_j| \tag{2.18}$$

where $N = nT$. Maximization of (2.18) can be attained following the computation algorithm in Section 2.3 and similarly, the variance estimation of the MPPLE can be derived. In both cases, fast computation is attainable given the computational efficiency of adaptive Lasso and relatively low dimension of the variance matrix.

Chapter 3

Regularized Multinomial Regression with Application to Historical Ecology

3.1 Introduction

Categorical data analysis is common in such disciplines as landscape ecology and environmental history. A motivating example is a study conducted to assess the influence of social, economic, and historical factors on forest covers (Steen-Adams et al., 2011). Researchers are particularly interested in quantifying relationships between land cover types and land ownership characteristics and identifying key factors. In the past, Fu et al. (2013) collapsed the multiple categories in the response variable to be binary and developed a regularized method for variable selection in autologistic regression models, whereas Jin et al. (2013) developed Bayesian hierarchical modeling for multinomial response variable, but did not consider variable selection. Here, we consider the identification of key covariates in a multinomial regression setting and develop new methodology and theory for this purpose.

For linear models and univariate generalized linear models (GLMs), a variety of methods have been proposed to select covariates and estimate coefficients simultaneously under the assumption of sparse models. An earlier, highly prominent method is least absolute shrinkage and selection operator (Lasso) proposed by Tibshirani (1996). However, Lasso is not always consistent in terms of variable selection (Zou, 2006). Zhao and Yu (2004) derived a necessary and sufficient condition for the consistency of the Lasso. Fan and Li (2001) developed nonconcave penalized likelihood for variable selection and established its oracle properties. Such properties and selection consistency continue to hold in high dimension settings (Fan and Lv, 2011). Alternative regularized methods for variable selection include the adaptive Lasso (Zou, 2006), fused Lasso (Tibshirani et al., 2005), and elastic net (Zou and Hastie, 2005).

For multinomial regression models, the relation between the response and a given covariate is represented by several coefficients for different response categories, and thus, variable selection should generally select or remove all the coefficients within a group. However, the commonly used methods shrink the regression coefficients with Lasso-type penalties, and ignore the natural grouping of coefficients associated with individual covariates (Friedman et al., 2010). Thus, the existing approaches for common univariate GLMs are not directly applicable. In addition, when a covariate is a categorial variable, the Lasso-type methods are not quite satisfactory, as the selection depends on how the dummy variables are encoded and only individual dummy variables can be selected instead of the entire categorical covariate.

Here we consider group Lasso, as it overcomes the difficulties of Lasso by using an extension of the L_1 penalty which shrinks the coefficients toward zero with pre-specified groups (Yuan and Lin, 2006). Although regularized methods with group-wise penalties have been widely studied (Zou and Yuan, 2008; Zhao et al., 2009), there remain several challenges with the group Lasso for multinomial regression. First, the existing theorems for selection consistency cannot be directly applied to GLMs with multi-category responses (Zhao and Yu, 2004; Fan and Lv, 2011). Here we study a weak oracle property in multinomial regression models, such that our method can remove unimportant covariates and consistently estimate the effect of important covariates with large probability under suitable regularity conditions. Second, parameter estimation methods for group Lasso are only available for linear regression and logistic models (Yuan and Lin, 2006; Meier et al., 2008). Here, for multinomial regression, we develop a new, efficient algorithm to compute group Lasso estimates.

In the remainder of this chapter, we present the multinomial regression models in Section 3.2 and develop a group Lasso method in Section 3.3. In Section 3.4, we develop a computational algorithm for parameter estimation and variable selection. The weak oracle property is established in Section 3.5. A simulation study is conducted in Section 3.6 to evaluate the performance of the proposed method, followed by the land cover data example in Section 3.7. An alternative penalization procedure is discussed in Section 3.8. A conclusion and further discussion are given in Section 3.9.

The technical details are given in Appendixes as Web-based Supplementary materials.

3.2 Multinomial Regression Model

For $i = 1, \ldots, n$, let y_i denote a nominal categorical response with K (≥ 2) categories for the ith observation. Consider G (≥ 1) pre-specified groups of covariates, each of which contains s_g covariates, for $g = 1, \ldots, G$. In the data example, each continuous covariate is a group of its own, but for a categorical covariate with $H + 1$ levels, corresponding H dummy variables are considered as a group. Let $\boldsymbol{x}_{gj} = [x_{gj1}, \ldots, x_{gjn}]^T$ denote an n-dimensional vector of the jth covariate in the gth group, for $j = 1, \ldots, s_g$ and $g = 1, \ldots, G$. Let β_{k0} denote the intercept for the kth category of the response and let β_{kgj} denote the regression coefficient of the jth covariate in the gth group for the kth category of the response, for $k = 1, \ldots, K$. We model the nominal categorical response y_i by a multinomial distribution

$$p_{ik} = \Pr(y_i = k) = \exp(\theta_{ik}) \left\{ \sum_{l=1}^{K} \exp(\theta_{il}) \right\}^{-1}, \tag{3.1}$$

where $\theta_{ik} = \beta_{k0} + \sum_{g=1}^{G} \sum_{j=1}^{s_g} x_{gji} \beta_{kgj}$, for $i = 1, \ldots, n$ and $k = 1, \ldots, K$.

The parametrization in (3.1) is not identifiable, because for any constants $\{c_0, c_{gj} : j = 1, \ldots, s_g, g = 1, \ldots, G\}$, $\{\beta_{k0} + c_0, \beta_{kgj} + c_{gj} : k = 1, \ldots, K, j = 1, \ldots, s_g, g = 1, \ldots, G\}$ would give identical probabilities in (3.1). To ensure model identifiability, a commonly used approach is to choose one of the response categories as a baseline category. Suppose the Kth category is chosen as the reference. Then, $\beta_{K0} = 0$ and

$\beta_{Kgj} = 0$, for $j = 1, \ldots, s_g$ and $g = 1, \ldots, G$. Model (3.1) becomes

$$p_{ik} = \exp(\theta_{ik}) \left\{ 1 + \sum_{l=1}^{K-1} \exp(\theta_{il}) \right\}^{-1},$$

where $i = 1, \ldots, n$ and $k = 1, \ldots, K$. An alternative approach is a sum-zero constraint such that,

$$\sum_{k=1}^{K} \beta_{k0} = 0, \; \sum_{k=1}^{K} \beta_{kgj} = 0, \; \text{for } j = 1, \ldots, s_g \text{ and } g = 1, \ldots, G. \tag{3.2}$$

With the sum-zero constraint, the geometric mean of the logits, $(\prod_{l=1}^{K} p_{il})^{1/K}$, can be viewed as the reference category, since $\log\{p_{ik}(\prod_{l=1}^{K} p_{il})^{-1/K}\} = \theta_{ik}$. Here, the sum-zero constraint is preferred for ease of the methodology development in Section 3.3, although analogous results can be derived for the baseline category approach.

Let $\boldsymbol{\beta} = (\beta_{k0}, \beta_{kgj}, k = 1, \ldots, K, j = 1, \ldots, s_g, g = 1, \ldots, G)^T$ denote all of the regression coefficients and $\boldsymbol{y} = (y_1, \ldots, y_n)^T$ denote the vector of all the observed responses. The log-likelihood function for $\boldsymbol{\beta}$ is

$$\begin{aligned} \ell_n(\boldsymbol{\beta}; \boldsymbol{y}) &= \sum_{k=1}^{K} \left\{ \mathbb{1}(\boldsymbol{y} = k\mathbf{1})^T \left(\beta_{k0}\mathbf{1} + \sum_{g=1}^{G} \sum_{j=1}^{s_g} \boldsymbol{x}_{gj} \beta_{kgj} \right) \right\} \\ &\quad - \sum_{i=1}^{n} \log \left\{ \sum_{k=1}^{K} \exp(\theta_{ik}) \right\} \end{aligned} \tag{3.3}$$

where $\mathbb{1}(\cdot)$ denotes the indicator function and $\mathbf{1} = (1, \ldots, 1)^T$ is an n-dimensional vector of ones.

3.3 Group Lasso Method

For variable selection in multinomial regression, the effect of a continuous covariate is represented by K regression coefficients. To include or exclude any given continuous covariate, we regularize the K coefficients associated with the K response categories by a group-wise penalty. However for a categorical covariate with $H + 1$ levels represented by H binary (dummy) variables, there are a total of HK coefficients, which form a group of coefficients and are subject to a group-wise penalty.

Generally, the penalized log-likelihood function for group Lasso with G pre-specified groups is

$$Q_n(\boldsymbol{\beta}) = n^{-1}\ell_n(\boldsymbol{\beta}) - \lambda_n \sum_{g=1}^{G} \left(s_g \sum_{j=1}^{s_g} \sum_{k=1}^{K} \beta_{kgj}^2 \right)^{1/2} \tag{3.4}$$

where $\ell_n(\boldsymbol{\beta}) \equiv \ell_n(\boldsymbol{\beta}; \boldsymbol{y})$ is the log-likelihood function given in (3.3), the second term is a group-wise penalty, and λ_n is a non-negative regularization parameter. Here, the intercept β_{k0} is not penalized and $s_g^{1/2}$ is used to ensure that the penalty term is of the order of the number of covariates for each group (Yuan and Lin, 2006). With a proper choice of the regularization parameter λ_n, maximizing the penalized log-likelihood function (3.4) can provide model parameter estimation and variable selection simultaneously. We denote the penalized maximum likelihood estimate (PMLE) of $\boldsymbol{\beta}$ as $\hat{\boldsymbol{\beta}} = \arg\max_{\boldsymbol{\beta}}\{Q_n(\boldsymbol{\beta})\}$. It can be easily shown that $\hat{\boldsymbol{\beta}}$ automatically satisfies the sum-zero constrain (3.2) except the intercept. An outline of the proof is given in Appendix B. An advantage of using the sum-zero constraint is that the PMLE $\hat{\boldsymbol{\beta}}$ does not require choosing a baseline category. The group-wise penalty enables the coefficients from the same group to shrink to zero altogether. In addition, the selection of the dummy variables that represent a categorical covariate does not depend on how the dummy variable are encoded, as they are included or excluded as a group.

3.4 Computational Algorithm

We now develop an iterative algorithm to maximize (3.4), as a special case of the block coordinate descent algorithm (Nocedal and Wright, 2000). Let $\boldsymbol{\beta}_0 = [\beta_{10}, \ldots, \beta_{K0}]^T$ denote the vector of intercepts, let $\boldsymbol{\beta}_g = [\beta_{1g1}, \ldots, \beta_{1gs_g}, \ldots, \beta_{Kg1}, \ldots, \beta_{Kgs_g}]^T$ denote the vector of coefficients for the gth group, and let $\boldsymbol{\beta}_{-g}$ denote all of the parameters except the ones associated with the gth group. Within each iteration, we first fix $\{\boldsymbol{\beta}_g, g = 1, \ldots, G\}$ and maximize (3.4) with respect to $\boldsymbol{\beta}_0$. Next, we maximize (3.4) with respect to each $\boldsymbol{\beta}_g$ for $g = 1, \ldots, G$, while holding the other coefficients $\boldsymbol{\beta}_{-g}$ constant.

Define the log-likelihood function of $\boldsymbol{\beta}_g$ evaluated at $\hat{\boldsymbol{\beta}}_{-g}$ as

$$\begin{aligned}\ell_n^g(\boldsymbol{\beta}_g \mid \hat{\boldsymbol{\beta}}_{-g}) = \quad & \sum_{k=1}^{K} \left\{ \mathbb{1}(\boldsymbol{y} = k\mathbf{1})^T (\hat{\beta}_{k0}\mathbf{1} + \sum_{j=1}^{s_g} \boldsymbol{x}_{gj}\beta_{kgj} + \sum_{h \neq g} \sum_{j=1}^{s_h} \boldsymbol{x}_{hj}\hat{\beta}_{khj}) \right\} \\ & - \sum_{i=1}^{N} \log \left\{ \sum_{k=1}^{K} \exp(\hat{\beta}_{k0} + \sum_{j=1}^{s_g} x_{gji}\beta_{kgj} + \sum_{h \neq g} \sum_{j=1}^{s_h} x_{hji}\hat{\beta}_{khj}) \right\}.\end{aligned}$$

Let $\nabla f(\boldsymbol{\theta})$ denote the first-order derivative of a function f with respect to a vector $\boldsymbol{\theta}$ and $\| \cdot \|_2$ denote the Euclidean distance.

Lemma 3.1. *If* $\|\nabla \ell_n^g(\boldsymbol{\beta}_g \mid \hat{\boldsymbol{\beta}}_{-g})|_{\boldsymbol{\beta}_g=\mathbf{0}}\|_2 \leq n{s_g}^{1/2}\lambda_n$, *then* $\underset{\boldsymbol{\beta}_g}{\arg\max}\{n^{-1}\ell_n^g(\boldsymbol{\beta}_g \mid \hat{\boldsymbol{\beta}}_{-g}) - \lambda_n s_g^{1/2}\|\boldsymbol{\beta}_g\|_2\} = \mathbf{0}$.

Since (3.4) is not differentiable at $\boldsymbol{\beta}_g = \mathbf{0}$, Newton-Raphson type algorithms are not guaranteed to converge if the maximizer is at $\boldsymbol{\beta}_g = \mathbf{0}$. However, Lemma 1 provide us with a condition to check if the maximizer with respect to $\boldsymbol{\beta}_g$ is $\mathbf{0}$ prior to the maximization. If the condition in Lemma 1 does not hold, then we can proceed with a Newton-Raphson type algorithm, since the target function is differentiable around the maximizer. The proof of Lemma 1 is given in Appendix B, and the algorithm to maximize (3.4) is summarized as follow:

Algorithm 1:

1. For a given λ_n, let $\hat{\boldsymbol{\beta}}^{[0]} = (\hat{\beta}_{k0}^{[0]}, \hat{\beta}_{kgj}^{[0]}, k = 1, \ldots, K, j = 1, \ldots, s_g, g = 1, \ldots, G)^T$ denote the initial values of all of the coefficients.

2. Let $\hat{\boldsymbol{\beta}}^{[t-1]} = (\hat{\beta}_{k0}^{[t-1]}, \hat{\beta}_{kgj}^{[t-1]}, k = 1, \ldots, K, j = 1, \ldots, s_g, g = 1, \ldots, G)^T$ denote the values at the $(t-1)$th iteration, and $\hat{\boldsymbol{\beta}}_{-0}^{[t-1]}$ denote the parameters of $\hat{\boldsymbol{\beta}}^{[t-1]}$ except the intercepts. At the tth iteration, for

$$\begin{aligned}\ell_n^0(\boldsymbol{\beta}_0 \mid \hat{\boldsymbol{\beta}}_{-0}^{[t-1]}) \;&=\; \sum_{k=1}^{K}\left\{\mathbb{1}(\boldsymbol{y} = k\mathbf{1})^T(\beta_{k0}\mathbf{1} + \sum_{g=1}^{G}\sum_{j=1}^{s_g}\boldsymbol{x}_{gj}\hat{\beta}_{kgj}^{[t-1]})\right\} \\ &\quad -\sum_{i=1}^{N}\log\left\{\sum_{k=1}^{K}\exp(\beta_{k0} + \sum_{g=1}^{G}\sum_{j=1}^{s_g}x_{gji}\hat{\beta}_{kgj}^{[t-1]})\right\},\end{aligned}$$

estimate $\boldsymbol{\beta}_0$ by $\hat{\boldsymbol{\beta}}_0^{[t]} = \arg\max_{\boldsymbol{\beta}_0} \ell_n^0(\boldsymbol{\beta}_0 \mid \hat{\boldsymbol{\beta}}_{-0}^{[t-1]})$ and then update $\hat{\beta}_{k0}^{[t]}$ to $\hat{\beta}_{k0}^{[t]} - K^{-1}\sum_{k=1}^{K}\hat{\beta}_{k0}^{[t]}$, for $k = 1, \ldots, K$.

3. At the tth iteration, repeat for $g = 1, \ldots, G$:
   If $\|\nabla \ell_n^g(\boldsymbol{\beta}_g \mid \hat{\boldsymbol{\beta}}_1^{[t]}, \ldots, \hat{\boldsymbol{\beta}}_{g-1}^{[t]}, \hat{\boldsymbol{\beta}}_{g+1}^{[t-1]}, \ldots, \hat{\boldsymbol{\beta}}_G^{[t-1]})|_{\boldsymbol{\beta}_g = \mathbf{0}}\|_2 \leq n{s_g}^{1/2}\lambda_n$, then $\hat{\boldsymbol{\beta}}_g^{[t]} = \mathbf{0}$.
   Otherwise $\hat{\boldsymbol{\beta}}_g^{[t]} = \arg\max_{\boldsymbol{\beta}_g}\Big\{ n^{-1}\ell_n^g(\boldsymbol{\beta}_g \mid \hat{\boldsymbol{\beta}}_1^{[t]}, \ldots, \hat{\boldsymbol{\beta}}_{g-1}^{[t]}, \hat{\boldsymbol{\beta}}_{g+1}^{[t-1]}, \ldots, \hat{\boldsymbol{\beta}}_G^{[t-1]}) - \lambda_n s_g^{1/2}\|\boldsymbol{\beta}_g\|_2 \Big\}$.
4. Repeat steps 2 and 3 until a convergence criterion is met.

To further improve variable selection and parameter estimation, we use a penalized estimator for selection, followed by a refit of the model with the set of non-zero coefficients and without penalty (see, e.g., Efron et al. (2004)), and this refitting step is only applied in the simulation. In addition, for estimating the regularization parameter λ_n, we compute a type of Bayesian information criterion (BIC),

$$\mathrm{BIC}(\lambda_n) = -2\ell_n(\hat{\boldsymbol{\beta}}; \lambda_n) + e(\lambda_n)\log(n), \tag{3.5}$$

where $e(\lambda_n) = \sum_{g=1}^{G}\{\mathbb{1}(\hat{\boldsymbol{\beta}}_g^T\hat{\boldsymbol{\beta}}_g > 0) + (\hat{\boldsymbol{\beta}}_g^T\hat{\boldsymbol{\beta}}_g)(\tilde{\boldsymbol{\beta}}_g^T\tilde{\boldsymbol{\beta}}_g)^{-1}(s_gK - 1)\}$ approximates the degree of freedom and $\tilde{\boldsymbol{\beta}}_g$ is the least square estimates of $\boldsymbol{\beta}_g$ for $g = 1, \ldots, G$ (Yuan and Lin, 2006). The first part of $\mathrm{BIC}(\lambda_n)$ in (3.5) evaluates the goodness of fit of model using the penalized estimates, and the second part penalizes the estimates with a large degree of freedom. We estimate λ_n by $\hat{\lambda}_n = \arg\min_{\lambda_n}\{\mathrm{BIC}(\lambda_n)\}$.

3.5 Selection Consistency

In this section, we study the nonasymptotic weak oracle property of the PMLE $\hat{\boldsymbol{\beta}}$. That is, with large probability, $\hat{\boldsymbol{\beta}}$ identifies the sparse structure of the true parameter vector, and the non-zero component of $\hat{\boldsymbol{\beta}}$ are consistent at some rate.

For ease of presentation, we assume $\boldsymbol{\beta}_0 = \mathbf{0}$ and drop $\boldsymbol{\beta}_0$ from (3.3)-(3.4). Under the sum-zero constraint, $\beta_{Kgj} = -\sum_{k=1}^{K-1} \beta_{kgj}$ for $j = 1, \ldots, s_g, g = 1, \ldots, G$. Drop the coefficients for the Kth category from $\boldsymbol{\beta}$ and $\boldsymbol{\beta}_g$, then $\boldsymbol{\beta} = (\beta_{kgj}, k = 1, \ldots, K-1, j = 1, \ldots, s_g, g = 1, \ldots, G)^T$, and $\boldsymbol{\beta}_g = (\beta_{kgj}, k = 1, \ldots, K-1, j = 1, \ldots, s_g)^T$ for $g = 1, \ldots, G$. Plug the sum-zero constraint (3.2) into the penalized log-likelihood function (3.4):

$$\begin{aligned} Q_n(\boldsymbol{\beta}) = n^{-1} \sum_{k=1}^{K-1} & \left[\{\mathbb{1}(\boldsymbol{y} = k\mathbf{1}) - \mathbb{1}(\boldsymbol{y} = K\mathbf{1})\}^T \sum_{g=1}^{G} \sum_{j=1}^{s_g} \boldsymbol{x}_{gj} \beta_{kgj} \right] \\ -n^{-1} \sum_{i=1}^{n} & \log \left\{ \sum_{k=1}^{K-1} \exp(\theta_{ik}) + \exp(-\sum_{k=1}^{K-1} \theta_{ik}) \right\} \\ -\lambda_n \sum_{g=1}^{G} & \left[s_g \sum_{j=1}^{s_g} \left\{ \sum_{k=1}^{K-1} \beta_{kgj}^2 + \left(\sum_{k=1}^{K-1} \beta_{kgj} \right)^2 \right\} \right]^{1/2} \end{aligned} \tag{3.6}$$

where $\theta_{ik} = \sum_{g=1}^{G} \sum_{j=1}^{s_g} x_{gji} \beta_{kgj}$, $i = 1, \ldots, n, k = 1, \ldots, K-1$.

Define $\boldsymbol{\theta} = (\theta_{ik}, i = 1, \ldots, n, k = 1, \ldots, K-1)^T$, and for $k = 1, \ldots, K-1$,

$$\boldsymbol{\mu}_k(\boldsymbol{\theta}) = \left(\left\{ \exp(\theta_{ik}) - \exp(-\sum_{k=1}^{K-1} \theta_{ik}) \right\} \left\{ \sum_{k=1}^{K-1} \exp(\theta_{ik}) + \exp(-\sum_{k=1}^{K-1} \theta_{ik}) \right\}^{-1}, i = 1, \ldots, n \right)^T$$

and $\nabla \ell_n^{kgj}(\hat{\boldsymbol{\beta}}) = \boldsymbol{x}_{gj}^T \{\mathbb{1}(\boldsymbol{y} = k\mathbf{1}) - \mathbb{1}(\boldsymbol{y} = K\mathbf{1}) - \boldsymbol{\mu}_k(\hat{\boldsymbol{\theta}})\}, j = 1, \ldots, s_g, g = 1, \ldots, G.$

Proposition 3.2. *A necessary and sufficient condition for $\hat{\boldsymbol{\beta}}$ to be a global maximizer* *of (3.6) is*

$$\nabla \ell_n^{kgj}(\hat{\boldsymbol{\beta}}) = n{s_g}^{1/2}\lambda_n \left\{\hat{\beta}_{kgj} + \left(\sum_{k=1}^{K-1}\hat{\beta}_{kgj}\right)\right\}\left[\sum_{j=1}^{s_g}\left\{\sum_{l=1}^{K-1}\hat{\beta}_{lgj}^2 + \left(\sum_{k=1}^{K-1}\hat{\beta}_{kgj}\right)^2\right\}\right]^{-1/2}$$

$$\textit{for } g \textit{ such that } \hat{\boldsymbol{\beta}}_g \neq \mathbf{0},\ j = 1, \ldots, s_g,\ k = 1, \ldots, K-1. \tag{3.7}$$

$$\sum_{j=1}^{s_g}\left[\sum_{k=1}^{K-1}\left\{\nabla \ell_n^{kgj}(\hat{\boldsymbol{\beta}})\right\}^2 + \left\{\sum_{k=1}^{K-1}\nabla \ell_n^{kgj}(\hat{\boldsymbol{\beta}})\right\}^2\right] \leq n^2 s_g \lambda_n^2,$$

$$\textit{for } g \textit{ such that } \hat{\boldsymbol{\beta}}_g = \mathbf{0}. \tag{3.8}$$

Proposition 3.2 is an extension of the Proposition 1 in Yuan and Lin (2006), which provides the Krush-Kuhn-Tucker (KKT) optimality condition for maximizing (3.6), and we use the sufficiency to prove the Theorem 3.3. The proof of Proposition 3.2 is given in Appendix B.

Next we consider the nonasymptotic weak oracle property of the PMLE $\hat{\boldsymbol{\beta}}$ in (3.4). Let $\boldsymbol{\beta}^* = (\beta_{kgj}^*, k = 1, \ldots, K-1, j = 1, \ldots, s_g, g = 1, \ldots, G)^T$ denote the true value of coefficients. Without loss of generality, we assume that the first M groups are non-zero groups (i.e., $\boldsymbol{\beta}_g^* \neq \mathbf{0}$ for $g = 1, \ldots, M$) and the remainder groups have zero coefficients (i.e., $\boldsymbol{\beta}_g^* = \mathbf{0}$ for $g = M+1, \ldots, G$). Accordingly, let $\boldsymbol{X}_I = [\boldsymbol{x}_{11}, \ldots, \boldsymbol{x}_{1s_1}, \ldots, \boldsymbol{x}_{M1}, \ldots, \boldsymbol{x}_{Ms_M}]$ be the design matrix of the covariates from the non-zero groups and $\boldsymbol{X}_{II} = [\boldsymbol{x}_{(M+1)1}, \ldots, \boldsymbol{x}_{M+1s_{M+1}}, \ldots, \boldsymbol{x}_{G1}, \ldots, \boldsymbol{x}_{Gs_G}]$ be the design matrix of the covariates from the zero groups, where $\boldsymbol{x}_{gj}$ is standardized so that $\|\boldsymbol{x}_{gj}\|_2 = n^{1/2}$ for $j = 1, \ldots, s_g$ and $g = 1, \ldots, G$. Further, define $\mathcal{N}_0 = \{\boldsymbol{\delta} \in \mathcal{R}^{(K-1)\sum_{g=1}^{M} s_g} : \|\boldsymbol{\delta} - \boldsymbol{\beta}_I^*\|_\infty \leq d, \boldsymbol{\beta}_I^* = [\boldsymbol{\beta}_1^{*T}, \ldots, \boldsymbol{\beta}_M^{*T}]^T\}$, where $d = (1/2)\min\{|\beta_{kgj}^*| :$

$\beta^*_{kgj} \neq 0,\ k = 1\dots,K-1, j = 1,\dots,s_g, g = 1\dots,M\}$ is half of the minimum non-zero coefficient.

Finally, define

$$\boldsymbol{A} = \begin{bmatrix} \boldsymbol{X}_I^T\boldsymbol{\Sigma}_{11}(\boldsymbol{\theta}^*)\boldsymbol{X}_I & \dots & \boldsymbol{X}_I^T\boldsymbol{\Sigma}_{1(K-1)}(\boldsymbol{\theta}^*)\boldsymbol{X}_I \\ \vdots & \ddots & \vdots \\ \boldsymbol{X}_I^T\boldsymbol{\Sigma}_{(K-1)1}(\boldsymbol{\theta}^*)\boldsymbol{X}_I & \dots & \boldsymbol{X}_I^T\boldsymbol{\Sigma}_{(K-1)(K-1)}(\boldsymbol{\theta}^*)\boldsymbol{X}_I \end{bmatrix},$$

$$\boldsymbol{B} = \begin{bmatrix} \boldsymbol{X}_{II}^T\boldsymbol{\Sigma}_{11}(\boldsymbol{\theta}^*)\boldsymbol{X}_I & \dots & \boldsymbol{X}_{II}^T\boldsymbol{\Sigma}_{1(K-1)}(\boldsymbol{\theta}^*)\boldsymbol{X}_I \\ \vdots & \ddots & \vdots \\ \boldsymbol{X}_{II}^T\boldsymbol{\Sigma}_{(K-1)1}(\boldsymbol{\theta}^*)\boldsymbol{X}_I & \dots & \boldsymbol{X}_{II}^T\boldsymbol{\Sigma}_{(K-1)(K-1)}(\boldsymbol{\theta}^*)\boldsymbol{X}_I \end{bmatrix},$$

and $\boldsymbol{C}_{lgj}(\boldsymbol{\delta}) = \left[\boldsymbol{c}_{kh} = \boldsymbol{X}_I^T \mathrm{diag}\{|\partial\{\partial\boldsymbol{\mu}_l(\boldsymbol{\theta})/\partial\boldsymbol{\theta}_k\mathbf{1}\}/\partial\boldsymbol{\theta}_h|_{\boldsymbol{\theta}=\boldsymbol{\Theta}(\boldsymbol{\delta})}|\}\boldsymbol{X}_I\right]_{k,h=1}^{K-1}$, where $\boldsymbol{\theta}_k = (\theta_{1k},\dots,\theta_{nk})^T$ for $k = 1,\dots,K-1$, $\boldsymbol{\Theta}(\boldsymbol{\delta}) = (\sum_{g=1}^M \sum_{j=1}^{s_g} x_{gji}\delta_{kgj}, i = 1,\dots,n, k = 1,\dots,K-1)^T$, and $\boldsymbol{\Sigma}_{kh}(\boldsymbol{\theta}) = \partial\boldsymbol{\mu}_k(\boldsymbol{\theta})/\partial\boldsymbol{\theta}_h$. Here, the derivative of a vector function with respect to a vector is the Jacobian matrix, and $\|\cdot\|_\infty$ is the L_∞-norm of a matrix.

We consider the following regularity conditions:

- (C1) $\|\boldsymbol{A}^{-1}\|_\infty = O(n^{-1})$.

- (C2) $\|\boldsymbol{B}\boldsymbol{A}^{-1}\|_\infty \leq 2^{-1}\max_{g=1}^M (s_g)^{-1/2}\{K(K-1)\}^{-1/2}$.

- (C3) $\max_{\boldsymbol{\delta}\in\mathcal{N}_0} \max_{g,j,l} \lambda_{max}\left[\boldsymbol{C}_{lgj}(\boldsymbol{\delta})\right] = O(n)$.

(C1) essentially requires that $\boldsymbol{A}$ is non-singular and there is a lower bound for its sup-norm. For a classical Gaussian linear regression model, Wainwright (2009) showed that $\|[\boldsymbol{X}_I^T\boldsymbol{X}_I]^{-1}\|_\infty = O_p(n^{-1})$ if the rows of $\boldsymbol{X}_I$ are i.i.d. Gaussian vectors with

$\|[E\boldsymbol{X}_I^T\boldsymbol{X}_I]^{-1}\|_\infty = O_p(n^{-1})$. (C2) is related to the strong irrepresentable condition of Lasso estimates. Zhao and Yu (2004) showed that the irrepresentable condition is sufficient and almost necessary for the Lasso to achieve model selection consistency. Here, (C2) essentially requires the correlation between the covariates from non-zero groups and zero groups cannot be too large. (C3) is a technical condition used in the proof, which holds automatically in classical Gaussian linear model (Fan and Lv, 2011).

Theorem 3.3. *Under condition (C1)-(C3), for fixed K and s_g for $g = 1, \dots, G$, if we choose λ_n satisfying $\lambda_n = o(n^{-1/2}\log n)$ and $\lambda_n^{-1}n^{-1/2}(\log n)^{1/2} \to 0$ as $n \to \infty$, then there exists an estimator $[\hat{\boldsymbol{\beta}}_I^T, \hat{\boldsymbol{\beta}}_{II}^T]^T$ for (3.6) that satisfies for sufficiently large n, with probability at least $1 - 2(K-1)(\sum_{g=1}^G s_g)n^{-1}$,*

$$\hat{\boldsymbol{\beta}}_{II} = \mathbf{0}, \tag{3.9}$$

$$\|\hat{\boldsymbol{\beta}}_I - \boldsymbol{\beta}_I^*\|_\infty = O(n^{-1/2}\log n), \tag{3.10}$$

$$sgn(\hat{\boldsymbol{\beta}}_I) = sgn(\boldsymbol{\beta}_I^*) \textit{ for only non-zero component of } \boldsymbol{\beta}_I^*, \tag{3.11}$$

where $\hat{\boldsymbol{\beta}}_I = [\hat{\boldsymbol{\beta}}_1^T, \dots, \hat{\boldsymbol{\beta}}_M^T]^T$, $\hat{\boldsymbol{\beta}}_{II} = [\hat{\boldsymbol{\beta}}_{M+1}^T, \dots, \hat{\boldsymbol{\beta}}_G^T]^T$, and $sgn(\hat{\beta}_{kgj}) = 1$ if $\hat{\beta}_{kgj} > 0$, $sgn(\hat{\beta}_{kgj}) = -1$ if $\hat{\beta}_{kgj} < 0$, $sgn(\hat{\beta}_{kgj}) = 0$ if $\hat{\beta}_{kgj} = 0$.

By (3.9) our proposed estimator identifies the zero groups with a large probability. The estimates of the covariates in the non-zero groups are consistent with a rate slower than $n^{1/2}$ by (3.10) and the sign of the non-zero coefficients are consistent by (3.11) with the large probability.

3.6 Simulation Study

Simulation Set-up

We conducted a simulation study to examine the finite-sample properties of the method developed in Section 3.3-3.4. We let $(n, G, K) = (900, 20, 4)$ be the sample size, number of groups, and number of response categories. Each covariate is a group of its own, and thus, $s_g = 1$ for $g = 1, \ldots, G$. The first 4 groups are the non-zero groups, with $\boldsymbol{\beta}_g^* = [1, 1, -1, -1]^T, g = 1, 2$, for the large-value groups and $\boldsymbol{\beta}_g^* = [0.3, 0.3, -0.3, -0.3]^T, g = 3, 4$, for the small-value groups. The remainder 16 groups are the zero groups. The intercepts are $\boldsymbol{\beta}_0 = (0.1, 0.1, -0.1, -0.1)^T$. The covariates $\boldsymbol{x}_{gj}$ are i.i.d. from standard normal distribution. To obtain cross-covariate correlations, we let the correlation between the non-zero groups (or zero groups) ρ_1 be 0.2. Then, we let the correlation between the non-zero groups and zero groups be ρ_2, where ρ_2 is {0, 0.1, 0.2, or 0.3}. For each ρ_2, 500 data sets were generated from the multinomial regression model (3.1).

For each simulated data set, the group Lasso method (GLasso) and group Lasso with a final refitting step (GLasso-r) were applied, as described in Section 3.3-3.4. For comparison, we performed the maximum likelihood estimation (MLE) with the model containing only the non-zero groups. Additionally,the standard Lasso method was used and the computation was carried out in the glmnet library in R (Team, 2011; Friedman et al., 2010). The regularization parameter was decided by the BIC criterion for GLasso, Glasso-r and Lasso.

Simulation Results

We evaluate the performance of the variable selection by three criteria. The first criterion is sensitivity, which is the proportion of correctly identified non-zero groups. The sensitivity to identify the first two non-zero groups are all equal to 1, and the sensitivity for the second two non-zero groups is given in Figure 3.1(a). The second criterion is specificity, which is the proportion of correctly identified zero groups (Figure 3.1(b)). The third criterion is correct model selection frequency (i.e., both of the non-zero and zero groups are correctly identified) among these 500 simulations (Figure 3.1(c)). When the magnitude of the coefficients is larger (group 1 and 2), all of the methods can identify them correctly. When the coefficients are smaller (group 3 and 4), the sensitivity for GLasso and Lasso are close 1, while GLasso-r is 20% smaller. The specificity for GLasso-r is close to 1, reduced from 0.9 to 0.8 for GLasso, and reduced from 0.7 to 0.2 for Lasso. As ρ_2 increases, the correct model selection frequency decreases from 0.5 to 0.4 for GLasso-r, decreases from 0.3 to 0 for GLasso, and is close to 0 for Lasso which suggests that Lasso cannot identify both the non-zero and zero groups simultaneously among these 500 simulations.

The group Lasso estimates are subject to the sum-zero constraint (3.2), while the Lasso estimates are subject to a median-zero constraint (Friedman et al., 2010). For a fair comparison in estimation accuracy among these methods, we converted all estimates under the sum-zero constraint (3.2). Figure 3.1(d) provides these scaled sum of the square of bias (SSB), $\sum_{k=1}^{K}\{(\hat{\beta}_{k0} - \hat{c}_0 - \beta^*_{k0})^2 + \sum_{g=1}^{G}\sum_{j=1}^{s_g}(\hat{\beta}_{kgj} - \hat{c}_{gj} - \beta^*_{kgj})^2\}$, where $\hat{c}_0 = \sum_{k=1}^{K}\hat{\beta}_{k0}/K, \hat{c}_{gj} = \sum_{k=1}^{K}\hat{\beta}_{kgj}/K, j = 1, \ldots, s_g, g = 1\ldots, G$. The MLE method has the smallest SSB, and the SSB for GLasso is quite close to the MLE.

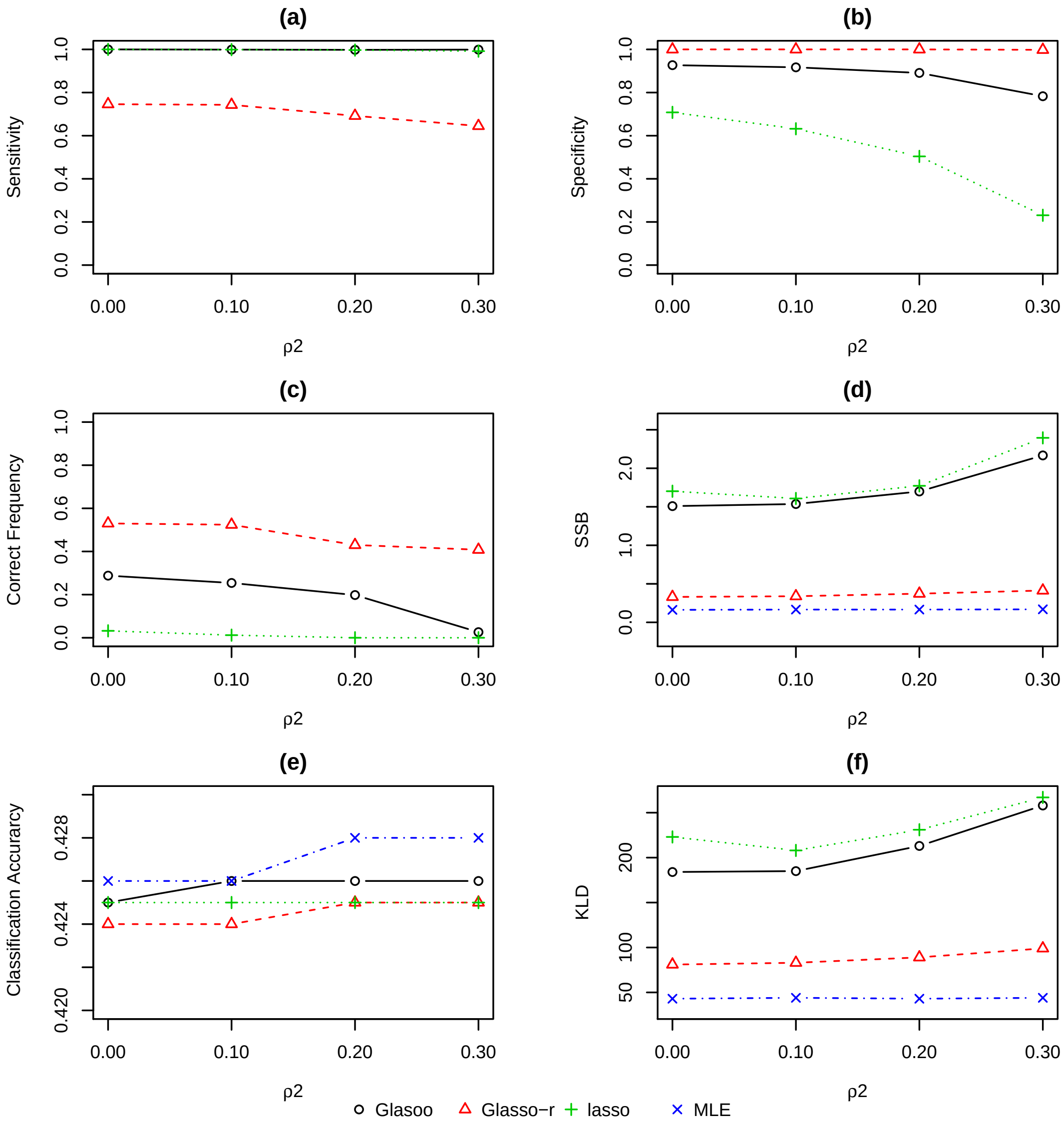

Figure 3.1: (a) provides the sensitivity for the second 2 important covariates. (b) provides the specificity. (c) provides the correct model selection frequency among 500 simulations. (d) provides sum of the square of bias (SSB). (e) provides the classification accuracy on the test data set. (f) provides the Kullback - Leibler divergence (KLD) on the test data set. Group Lasso (GLasso), group Lasso with refitting (GLasso-r), maximum likelihood (MLE) method only for the important covariates and Lasso are performed, and all of the reported values are the corresponding averaged values of 500 simulation results.

The SSB of GLasso and Lasso is 10 times more, and the SSB for Glasso is slightly smaller than Lasso.

Furthermore, based on a test data set with a sample size of 5000 for each simulation, we apply two criteria to compare the prediction performance. One is the classification accuracy on the test data set given in Figure 3.1(e) and the other is the Kullback-Leibler divergence (KLD), defined as: $\sum_{i=1}^{5000}\sum_{k=1}^{K}\{p_{ik}^{*}\log(p_{ik}^{*}/\hat{p}_{ik})\}$ given in Figure 3.1(f). The KLD is a measure of the difference between the actual probability p_{ik}^{*} and the predicted probability $\hat{p}_{ik}$ for $i = 1, \ldots, n, k = 1, \ldots, K$. While the values of classification accuracy for each method are quite similar, the KLD values are quite different with the general pattern of MLE < GLasso-r < GLasso < Lasso. These results suggest that GLasso-r gives the best prediction of the distribution among all the methods, even though the distributions have similar modes in most of the cases.

In addition, as the value of ρ_2 increases, the performance of GLasso in terms of parameter estimation and variable selection worsens. The poor performance is possibly because the weak oracle property in Theorem 2 is violated when ρ_2 increases.

Finally, we increased the number of covariates p from 20 to 40 and increased the percent of nonzero groups from 20% to 40%. The results under these scenarios are comparable with the description above and thus omitted. All of the reported values in Figure 3.1(a)-(f) are the corresponding averaged values of 500 simulation results. The standard error of these corresponding values are given in Table 3.1.

Table 3.1: Mean and standard error (se) of sensitivity (SEN), specificity (SPE), sum of the square of bias (SSB), classification accuracy (CLA) and Kullback - Leibler divergence (KLD) of the 500 simulations and the correct model selection frequency (COR) among the 500 simulations, by group Lasso (GLasso), group Lasso with refitting (GLasso-r), maximum likelihood (MLE) method only for the important covariates and Lasso, when ρ_1=0.2, and ρ_2 is from {0, 0.1, 0.2, 0.3}.

	$\rho_2 = 0.00$				$\rho_2 = 0.10$			
	GLasso	GLasso-r	Lasso	MLE	GLasso	GLasso-r	Lasso	MLE
SEN1 mean	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
SEN1 se	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
SEN2 mean	1.000	0.746	1.000	1.000	0.999	0.743	0.999	1.000
SEN2 se	0.000	0.013	0.000	0.000	0.001	0.013	0.001	0.000
SPE mean	0.922	1.000	0.708	1.000	0.917	1.000	0.632	1.000
SPE se	0.003	0.000	0.007	0.000	0.003	0.000	0.007	0.000
COR mean	0.288	0.530	0.032	1.000	0.254	0.524	0.012	1.000
COR se	0.020	0.022	0.008	0.000	0.019	0.022	0.005	0.000
SSB mean	1.509	0.330	1.702	0.162	1.536	0.339	1.608	0.167
SSB se	0.014	0.010	0.019	0.004	0.014	0.010	0.018	0.004
CLA mean	0.425	0.424	0.425	0.426	0.426	0.424	0.425	0.426
CLA se	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
KLD mean	184.422	81.216	223.516	43.416	185.378	83.956	208.483	44.762
KLD se	1.663	2.132	2.479	0.714	1.766	2.138	2.461	0.724

	$\rho_2 = 0.20$				$\rho_2 = 0.30$			
	GLasso	GLasso-r	Lasso	MLE	GLasso	GLasso-r	Lasso	MLE
SEN1 mean	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
SEN1 se	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
SEN2 mean	0.998	0.692	0.997	1.000	0.990	0.645	0.992	1.000
SEN2 se	0.001	0.013	0.002	0.000	0.003	0.016	0.003	0.000
SPE mean	0.892	1.000	0.504	1.000	0.783	0.998	0.231	1.000
SPE se	0.004	0.000	0.008	0.000	0.005	0.001	0.007	0.000
COR mean	0.198	0.430	0.000	1.000	0.026	0.408	0.000	1.000
COR se	0.018	0.022	0.000	0.000	0.007	0.022	0.000	0.000
SSB mean	1.699	0.372	1.772	0.166	2.167	0.413	2.395	0.169
SSB se	0.015	0.009	0.021	0.004	0.017	0.011	0.028	0.004
CLA mean	0.426	0.425	0.425	0.428	0.426	0.425	0.425	0.428
CLA se	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
KLD mean	213.545	89.995	231.087	43.555	258.021	99.845	267.572	44.673
KLD se	2.090	2.114	3.084	0.757	2.646	2.584	4.243	0.745

3.7 Data Example

We now return to the land cover data example. The study was aimed at assessing the influence of past land ownership characteristics on land cover structure in northern Wisconsin, USA. Landscape ecologists were interested in the effect of land ownership because it offers a spatially specific, quantifiable approach to assess the effect of important, yet often geographically amorphous social, economic, and historical factors on landscape structure (Turner et al., 1996).

Data were derived from historical plat maps and the Wisconsin Land Economic Inventory, a land survey. The spatial unit of analysis is a quarter section (or, 1/36 township, 160 acres $\approx$ 65 ha) and there are 1429 units in the study area (Figure 3.2). The original response variable is the land cover type with 7 categories, they are ag-grassland (AG), aspen (*Populus* spp.)-paper birch (*Betula papryfera*) (APB), lowland forest (LF), marsh-bog (MB), northern hardwood (NH), Others and Pine Species (PS) (Figure 3.2). Since the total observations for categories LF, MB, Others and PS is only about 3.2% of the whole data set, we combined these categories into one category, which results in a total of four categories: AG, APB, NH, Others. Both the original 7-category responses and the 4-category responses were used in the data analysis.

Forest composition can be associated with land ownership characteristics, such as ownership class, parcel size, and ownership size (Crow et al., 1999; Stanfield et al., 2002). The specific covariates of interest are: ownership, reservation, number of parcels, average size of parcels within a quarter section, proportion of the largest parcel, total area, and average size of all parcels in (but not necessarily all contained

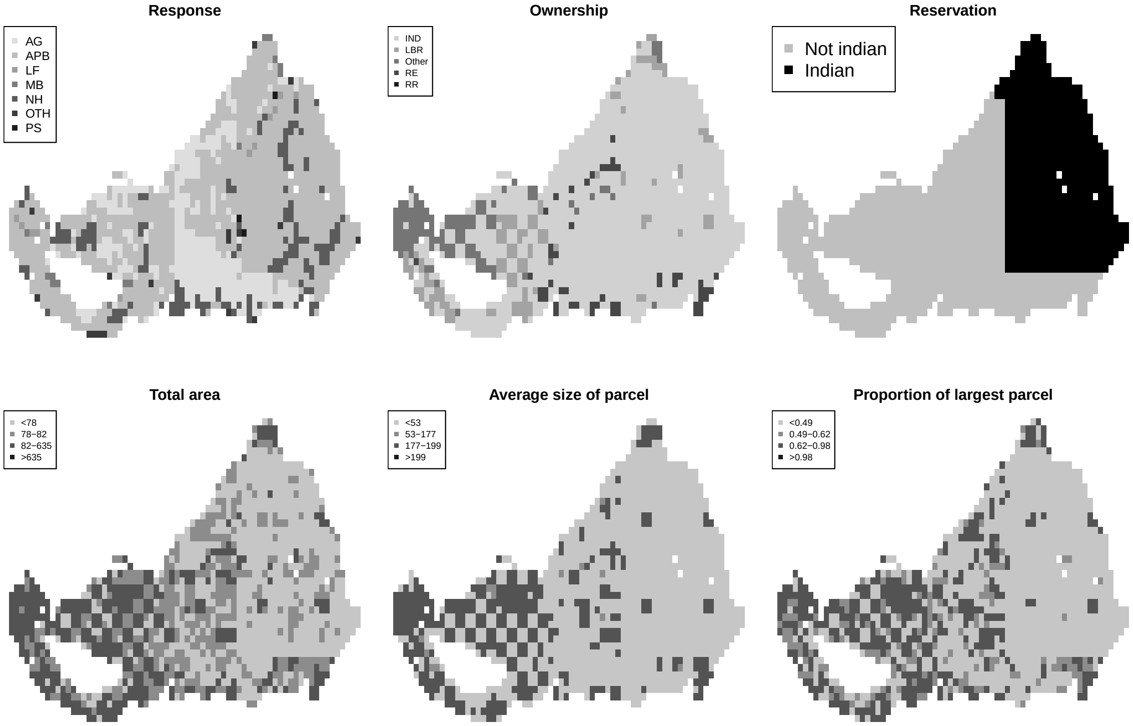

Figure 3.2: Map of the response variable, ownership, reservation, total area, average size of parcels and proportion of the largest parcel in a study area of northern Wisconsin which has 1,429 quarter sections has 7 categories. Response has 7 categories: ag-grassland (AG), aspen (*Populus* spp.)-paper birch (*Betula papryfera*) (APB), lowland forest (LF), marsh-bog (MB), northern hardwood (NH), Others and Pine Species (PS). Ownership has 5 levels: individual (IND), lumber company (LBR), real estate company (RE), Railroad (RR) and Others. Reservation has 2 levels: it's an Indian reservation or not. Total area, average size of parcels and proportion of the largest parcel are continuous covariates.

within) a quarter section. In particular, ownership (Own) is a categorical variable with 5 levels: individual (IND), lumber company (LBR), real estate company (RE), railroad (RR) and others. Four dummy variables are used to indicate the occurrence of IND (or IBR, RE, RR) against others. Further, reservation (Reserv) indicates whether the quarter section is on an Indian reservation or not. The number of parcels (PolyNm) in a given quarter section is a measure of parcel density. Average size (PolyPr) is the average size of parcel polygon, which is a measure of parcel size. Proportion of the largest parcel (MxPolyPr) reports the largest parcel polygon as a proportion of quarter section. Total area (TotOwn) shows the total property area (measured in acres) in this study area associated with owner of the largest parcel polygon in a given quarter section. Because the values of this covariate are skewed to the right, a log transformation was applied. Average size of all parcels (AvParcel) is the average size of all parcels that lay in whole or in part within a given quarter section (measured in acres). All the covariates except reservation and dummy variables are standardized to have mean 0 and variance 1. We applied the group Lasso estimation, where the grouping structure is such that reservation or each continuous covariate is a group of its own, and all of the dummy variables for Own are a group. Lasso was also applied for comparison.

The results are given in Table 3.2 and 3.3. Both the Glasso and Glasso-r methods selected Reserv, TotOwn and AvParcel. Glasso selected one more covariate PolyNm for the 7-category and 4-category cases, while Glasso selected Own in the 7-category case. In contrast, Lasso selected almost all of the covariates, possibly because Lasso have a poor performance in specificity compared to GLasso-r and Glasso as was

Table 3.2: Estimates by group Lasso (GLasso), group Lasso with refitting (GLasso-r) and Lasso for 7-category response: AG, APB, LF, MB, NH, Other and PS, the grouping structure for group Lasso is: each continuous covariate or Reserv is a group of its own, and all of the dummy variables are a group.

				GLasso			
Covariates	AG	APB	LF	MB	NH	Other	PS
Intercept	3.077	1.923	-2.456	-0.395	1.283	-1.630	-1.802
Reserv	-3.113	0.704	0.834	0.386	0.208	-0.051	1.032
PolyNm	0.242	0.081	0.160	-0.079	-0.084	-0.178	-0.141
PolyPr	–	–	–	–	–	–	–
MxPolyPr	–	–	–	–	–	–	–
log(TotOwn)	-0.244	0.093	-0.031	0.040	0.032	0.267	-0.157
AvParcl	-0.001	0.001	0.003	–	–	-0.002	-0.001
IND	0.534	0.112	0.015	-0.937	0.126	0.090	0.061
LBR	-0.373	0.363	-0.171	0.096	0.186	-0.089	-0.012
RE	-0.007	-0.085	0.305	0.218	-0.271	-0.143	-0.018
RR	-0.220	0.038	-0.086	-0.222	0.531	-0.031	-0.010

				GLasso-r			
Covariates	AG	APB	LF	MB	NH	Other	PS
Intercept	5.169	2.425	-2.382	-2.595	1.474	-1.934	-2.157
Reserv	-3.750	0.218	0.808	0.087	-0.397	-0.662	3.696
PolyNm	–	–	–	–	–	–	–
PolyPr	–	–	–	–	–	–	–
MxPolyPr	–	–	–	–	–	–	–
log(TotOwn)	-0.317	0.174	0.097	0.369	0.093	0.374	-0.791
AvParcl	-0.002	–	0.002	-0.001	-0.001	-0.002	0.003
IND	–	–	–	–	–	–	–
LBR	–	–	–	–	–	–	–
RE	–	–	–	–	–	–	–
RR	–	–	–	–	–	–	–

				Lasso			
Covariates	AG	APB	LF	MB	NH	Other	PS
Intercept	4.212	2.419	-1.739	-1.429	0.879	-1.304	-3.038
Reserv	-2.774	0.366	–	–	–	–	–
PolyNm	–	–	–	–	–	–	–
PolyPr	-0.135	–	–	–	–	–	–
MxPolyPr	-0.603	–	–	–	–	–	–
log(TotOwn)	-0.330	–	–	–	–	–	–
AvParcl	–	–	–	–	–	–	–
IND	0.277	–	–	–	–	–	–
LBR	–	0.211	–	–	–	–	–
RE	–	–	–	–	–	–	–
RR	–	–	–	–	0.172	–	–

Table 3.3: Estimates by group Lasso (GLasso), group Lasso with refitting (GLasso-r) and Lasso for 4-category response: AG, APB, NH and other, the grouping structure for group Lasso is:each continuous covariate or Reserv is a group of its own, and all of the dummy variables are a group.

	GLasso			
Covariates	APB	AG	NH	Other
Intercept	0.342	2.347	-0.262	-2.427
Reserv	1.126	-2.401	0.538	0.736
PolyNm	0.025	0.134	-0.096	-0.063
PolyPr	–	–	–	–
MxPolyPr	–	–	–	–
log(TotOwn)	0.110	-0.298	0.028	0.160
AvParcl	0.001	-0.001	–	–
IND	–	–	–	–
LBR	–	–	–	–
RE	–	–	–	–
RR	–	–	–	–
	GLasso-r			
Covariates	APB	AG	NH	Other
Intercept	0.427	3.153	-0.508	-3.073
Reserv	1.180	-2.787	0.561	1.046
PolyNm	–	–	–	–
PolyPr	–	–	–	–
MxPolyPr	–	–	–	–
log(TotOwn)	0.113	-0.375	0.028	0.235
AvParcl	0.001	-0.001	–	–
IND	–	–	–	–
LBR	–	–	–	–
RE	–	–	–	–
RR	–	–	–	–
	Lasso			
Covariates	APB	AG	NH	Other
Intercept	0.617	2.414	-0.922	-2.109
Reserv	0.369	-2.772	–	–
PolyNm	–	–	–	–
PolyPr	–	-0.136	–	–
MxPolyPr	–	-0.605	–	–
log(TotOwn)	–	-0.331	–	–
AvParcl	–	–	–	–
IND	–	0.276	–	–
LBR	0.211	–	–	–
RE	–	–	–	–
RR	–	–	0.172	–

observed in the simulation study. Moreover, nearly the same sets of covariates were chosen using the 7-category response and the 4-category response. The selected covariates are comparable to the results in Fu et al. (2013). However, the results here are more informative, as we can compare the effect of a selected covariate across different response categories. For example, for Reserv using 4-category response, the difference between the coefficient for APB ($K = 1$) and AG ($K = 2$) is positive, and with all the other parameters held constant, $\log(p_{i1}/p_{i2})$ is a monotonically increasing function with respect to the difference. That is, a positive difference is associated with probability for the ith site to be APB, with a higher probability for a greater difference.

3.8 Alternative Penalization

Instead of using the log-likelihood function (3.3) with the sum-zero constraint (3.2), we can have a non-constraint log-likelihood function by plug in (3.2) to (3.3). For example, replacing $\beta_{K0} = -\sum_{k=1}^{K-1} \beta_{k0}$ and $\beta_{Kgj} = -\sum_{k=1}^{K-1} \beta_{kgj}$ for $j = 1, \dots, s_g$ and $g = 1, \dots, G$, then (3.3) becomes

$$\begin{aligned}\ell_n^a(\boldsymbol{\beta}; \boldsymbol{y}) &= \sum_{k=1}^{K-1} \left[\{\mathbb{1}(\boldsymbol{y} = k\mathbf{1}) - \mathbb{1}(\boldsymbol{y} = K\mathbf{1})\}^T \left(\beta_{k0}\mathbf{1} + \sum_{g=1}^{G}\sum_{j=1}^{s_g} \boldsymbol{x}_{gj}\beta_{kgj}\right)\right] - \\ & \sum_{i=1}^{n} \log\left\{ \sum_{k=1}^{K-1} \exp(\theta_{ik}) + \exp(-\sum_{k=1}^{K-1} \theta_{ik}) \right\}\end{aligned}$$

Since the coefficients for the Kth response category are replaced, the effect of each covariate is represented by $K - 1$ coefficients for K response categories. And the alternative penalized log-likelihood function is

$$Q_n^a(\boldsymbol{\beta}) = n^{-1}\ell_n^a(\boldsymbol{\beta}) - \lambda_n \sum_{g=1}^{G} \left(s_g \sum_{j=1}^{s_g} \sum_{k=1}^{K-1} \beta_{kgj}^2 \right)^{1/2} \tag{3.12}$$

With a proper choice of the regularization parameter λ_n, maximizing the penalized log-likelihood function (3.12) can also provide model parameter estimation and variable selection simultaneously. We denote the alternative penalized maximum likelihood estimate (APMLE) of $\boldsymbol{\beta}$ as $\hat{\boldsymbol{\beta}} = \arg\max_{\boldsymbol{\beta}}\{Q_n^a(\boldsymbol{\beta})\}$. In this example, since the coefficients for the Kth response category are replaced, only the coefficients for the 1th to $(K-1)$th response categories are penalized by the group-wise penalty. Thus, it's obviously to find if the coefficients for another response category are replaced, i.e. $\beta_{hgj} = -\sum_{k\neq h}\beta_{kgj}$ for $h \neq K, j = 1, \ldots, s_g, g = 1, \ldots, G$, then the penalized log-likelihood function (3.12) will be different, as the group-wise penalty penalized different set of coefficients. Thus the alternative penalization procedure is depend on which response category is replaced, and we preferred the penalization procedure described in Section 3.3.

For the alternative penalization procedure, we have the similar theoretical results and algorithm as in Section 3.4 to 3.5.

Lemma 3.4. *Let $\ell_n^{ag}(\boldsymbol{\beta}_g \mid \hat{\boldsymbol{\beta}}_{-g})$ be the log-likelihood function of $\boldsymbol{\beta}_g$ evaluated at $\hat{\boldsymbol{\beta}}_{-g}$,*

$$
\begin{aligned}
&\ell_n^{ag}(\boldsymbol{\beta}_g \mid \hat{\boldsymbol{\beta}}_{-g})\\
= \;&\sum_{k=1}^{K-1}\left[\left\{\mathbb{1}(\boldsymbol{y} = k\mathbf{1}) - \mathbb{1}(\boldsymbol{y} = K\mathbf{1})\right\}^T (\hat{\beta}_{k0}\mathbf{1} + \sum_{j=1}^{s_g} \boldsymbol{x}_{gj}\beta_{kgj} + \sum_{h\neq g}\sum_{j=1}^{s_h} \boldsymbol{x}_{hj}\hat{\beta}_{khj})\right]\\
&- \sum_{i=1}^{N} \log\Bigg[\exp\Bigg\{-\sum_{k=1}^{K-1}(\hat{\beta}_{k0} + \sum_{j=1}^{s_g} x_{gji}\beta_{kgj} + \sum_{h\neq g}\sum_{j=1}^{s_h} x_{hji}\hat{\beta}_{khj})\Bigg\}\\
&+ \sum_{k=1}^{K-1} \exp(\hat{\beta}_{k0} + \sum_{j=1}^{s_g} x_{gji}\beta_{kgj} + \sum_{h\neq g}\sum_{j=1}^{s_h} x_{hji}\hat{\beta}_{khj})\Bigg].
\end{aligned}
$$

If $\|\nabla \ell_n^{ag}(\boldsymbol{\beta}_g \mid \hat{\boldsymbol{\beta}}_{-g})|_{\boldsymbol{\beta}_g=\mathbf{0}}\|_2 \le n{s_g}^{1/2}\lambda_n$, *then* $\arg\max_{\boldsymbol{\beta}_g}\{n^{-1}\ell_n^{ag}(\boldsymbol{\beta}_g \mid \hat{\boldsymbol{\beta}}_{-g}) - \lambda_n s_g^{1/2}\|\boldsymbol{\beta}_g\|_2\} = \mathbf{0}$.

Algorithm 2:

1. For a given λ_n, let $\hat{\boldsymbol{\beta}}^{[0]} = (\hat{\beta}_{k0}^{[0]}, \hat{\beta}_{kgj}^{[0]}, k = 1, \ldots, K-1, j = 1, \ldots, s_g, g = 1, \ldots, G)^T$ denote initial values of all of the coefficients.

2. Let $\hat{\boldsymbol{\beta}}^{[t-1]} = (\hat{\beta}_{k0}^{[t-1]}, \hat{\beta}_{kgj}^{[t-1]}, k = 1, \ldots, K-1, j = 1, \ldots, s_g, g = 1, \ldots, G)^T$ denote the values at the $(t-1)$th iteration, and $\hat{\boldsymbol{\beta}}_{-0}^{[t-1]}$ denote the parameters of $\hat{\boldsymbol{\beta}}^{[t-1]}$ except the intercepts. At the tth iteration, for

$$
\begin{aligned}
&\ell_n^{0}(\boldsymbol{\beta}_0 \mid \hat{\boldsymbol{\beta}}_{-0}^{[t-1]}) = \sum_{k=1}^{K}\left\{\left\{\mathbb{1}(\boldsymbol{y} = k\mathbf{1}) - \mathbb{1}(\boldsymbol{y} = K\mathbf{1})\right\}^T (\beta_{k0}\mathbf{1} + \sum_{g=1}^{G}\sum_{j=1}^{s_g} \boldsymbol{x}_{gj}\hat{\beta}_{kgj}^{[t-1]})\right\}\\
&- \sum_{i=1}^{N} \log\Bigg[\exp\Bigg\{-\sum_{k=1}^{K-1}(\beta_{k0} + \sum_{g=1}^{G}\sum_{j=1}^{s_g} x_{gji}\hat{\beta}_{kgj}^{[t-1]})\Bigg\}\\
&+ \sum_{k=1}^{K-1} \exp(\beta_{k0} + \sum_{g=1}^{G}\sum_{j=1}^{s_g} x_{gji}\hat{\beta}_{kgj}^{[t-1]})\Bigg],
\end{aligned}
$$

estimate $\boldsymbol{\beta}_0$ by $\hat{\boldsymbol{\beta}}_0^{[t]} = \arg\max_{\boldsymbol{\beta}_0} \ell_n^0(\boldsymbol{\beta}_0 \mid \hat{\boldsymbol{\beta}}_{-0}^{[t-1]})$.

3. At the tth iteration, repeat for $g = 1, \ldots, G$:
   If $\|\nabla \ell_n^g(\boldsymbol{\beta}_g \mid \hat{\boldsymbol{\beta}}_1^{[t]}, \ldots, \hat{\boldsymbol{\beta}}_{g-1}^{[t]}, \hat{\boldsymbol{\beta}}_{g+1}^{[t-1]}, \ldots, \hat{\boldsymbol{\beta}}_G^{[t-1]})|_{\boldsymbol{\beta}_g=\mathbf{0}}\|_2 \leq n{s_g}^{1/2}\lambda_n$, then $\hat{\boldsymbol{\beta}}_g^{[t]} = \mathbf{0}$.
   Otherwise $\hat{\boldsymbol{\beta}}_g^{[t]} = \underset{\boldsymbol{\beta}_g}{\arg\max}\Big\{ n^{-1}\ell_n^g(\boldsymbol{\beta}_g \mid \hat{\boldsymbol{\beta}}_1^{[t]}, \ldots, \hat{\boldsymbol{\beta}}_{g-1}^{[t]}, \hat{\boldsymbol{\beta}}_{g+1}^{[t-1]}, \ldots, \hat{\boldsymbol{\beta}}_G^{[t-1]}) - \lambda_n s_g^{1/2}\|\boldsymbol{\beta}_g\|_2 \Big\}$.
4. Repeat steps 2 and 3 until a convergence criterion is met.

Proposition 3.5. *The necessary and sufficient condition for $\hat{\boldsymbol{\beta}}$ to be a global maximizer of (3.12) is*

$$\boldsymbol{x}_{gj}^T\{\mathbb{1}(\boldsymbol{y} = k\mathbf{1}) - \mathbb{1}(\boldsymbol{y} = K\mathbf{1}) - \boldsymbol{\mu}_k(\hat{\boldsymbol{\theta}})\} = n{s_g}^{1/2}\lambda_n\hat{\beta}_{kgj}\left\{\sum_{j=1}^{s_g}\sum_{k=1}^{K-1}\hat{\beta}_{kgj}^2\right\}^{-1/2}$$

$$\text{for } g \text{ such that } \hat{\boldsymbol{\beta}}_g \neq \mathbf{0},\ j = 1, \ldots, s_g,\ k = 1, \ldots, K-1. \tag{3.13}$$

$$\left\| \nabla \ell_n^g(\boldsymbol{\beta}_g \mid \hat{\boldsymbol{\beta}})|_{\boldsymbol{\beta}_g = \hat{\boldsymbol{\beta}}_g} \right\|_2 \leq n{s_g}^{1/2}\lambda_n \text{ for } g \text{ such that } \hat{\boldsymbol{\beta}}_g = \mathbf{0}. \tag{3.14}$$

We consider an alternative regularity conditions: (C4) $\|\boldsymbol{B}\boldsymbol{A}^{-1}\|_\infty < \max_{g=1}^M\{s_g(K-1)\}^{-1/2}$. This condition is also related to the irrepresentable condition (C2), but it's weaker than (C2), since the coefficients for the Kth response category are not penalized in (3.12).

Theorem 3.6. *Under condition (C1),(C3),(C4), for fixed K and s_g for $g = 1, \ldots, G$, if we choose λ_n satisfying $\lambda_n = o(n^{-1/2}\log n)$ and $\lambda_n^{-1}n^{-1/2}(\log n)^{1/2} \to 0$, then there exists an estimator $[\hat{\boldsymbol{\beta}}_I^T, \hat{\boldsymbol{\beta}}_{II}^T]^T$ for (3.12), that satisfies for sufficiently large n, with* *probability at least* $1 - 2(K-1)(\sum_{g=1}^{G} s_g)n^{-1}$,

$$\hat{\boldsymbol{\beta}}_{II} = \mathbf{0}, \tag{3.15}$$

$$\|\hat{\boldsymbol{\beta}}_I - \boldsymbol{\beta}_I^*\|_\infty = O(n^{-1/2}\log n), \tag{3.16}$$

$$sgn(\hat{\boldsymbol{\beta}}_I) = sgn(\boldsymbol{\beta}_I^*) \textit{ for only non-zero component of } \boldsymbol{\beta}_I^*. \tag{3.17}$$

where $\hat{\boldsymbol{\beta}}_I = [\hat{\boldsymbol{\beta}}_1^T, \ldots, \hat{\boldsymbol{\beta}}_M^T]^T$, $\hat{\boldsymbol{\beta}}_{II} = [\hat{\boldsymbol{\beta}}_{M+1}^T, \ldots, \hat{\boldsymbol{\beta}}_G^T]^T$, *and* $sgn(\hat{\beta}_{kgj}) = 1$ *if* $\hat{\beta}_{kgj} > 0$, $sgn(\hat{\beta}_{kgj}) = -1$ *if* $\hat{\beta}_{kgj} < 0$, $sgn(\hat{\beta}_{kgj}) = 0$ *if* $\hat{\beta}_{kgj} = 0$.

For simulation, with the same setting in Section 3.6, we have the following results. Under the alternative penalization procedure, the group Lasso method is denoted by AGLasso and group Lasso with a final refitting step is denoted by AGLasso-r. And from Table 3.4, the performance for penalization procedure described in Section 3.3 is slightly better than the alternative penalization procedure in terms of variable selection and parameter estimation. Moreover, the result for data analysis with the alternative penalization procedure is given in Table 3.5 and 3.6, and the results are comparable with result in Section 3.7.

Table 3.4: Mean and standard error (se) of sensitivity (SEN), specificity (SPE), sum of the square of bias (SSB), classification accuracy (CLA) and Kullback - Leibler divergence (KLD) of the 500 simulations and the correct model selection frequency (COR) among the 500 simulations, by group Lasso (GLasso), group Lasso with refitting (GLasso-r), alternative group Lasso (AGLasso), alternative group Lasso with refitting (AGLasso-r), when ρ_1=0.2, and ρ_2 is from {0, 0.1, 0.2, 0.3}.

	$\rho_2 = 0.00$				$\rho_2 = 0.10$			
	GLasso	GLasso-r	AGLasso	AGLasso-r	GLasso	GLasso-r	AGLasso	AGLasso-r
SEN1 mean	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
SEN1 se	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
SEN2 mean	1.000	0.746	1.000	0.914	0.999	0.743	1.000	0.913
SEN2 se	0.000	0.013	0.000	0.009	0.001	0.013	0.000	0.009
SPE mean	0.922	1.000	0.814	1.000	0.917	1.000	0.794	0.999
SPE se	0.003	0.000	0.005	0.000	0.003	0.000	0.005	0.000
COR mean	0.288	0.530	0.068	0.828	0.254	0.524	0.060	0.820
COR se	0.020	0.022	0.011	0.017	0.019	0.022	0.011	0.017
SSB mean	1.509	0.330	1.498	0.217	1.536	0.339	1.481	0.224
SSB se	0.014	0.010	0.015	0.007	0.014	0.010	0.017	0.007
CLA mean	0.425	0.424	0.425	0.425	0.426	0.424	0.426	0.426
CLA se	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
KLD mean	184.422	81.216	192.172	55.681	185.378	83.956	188.197	58.240
KLD se	1.663	2.132	1.893	1.494	1.766	2.138	2.094	1.552

	$\rho_2 = 0.20$				$\rho_2 = 0.30$			
	GLasso	GLasso-r	AGLasso	AGLasso-r	GLasso	GLasso-r	AGLasso	AGLasso-r
SEN1 mean	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
SEN1 se	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
SEN2 mean	0.998	0.692	0.999	0.841	0.990	0.645	0.998	0.712
SEN2 se	0.001	0.013	0.001	0.011	0.003	0.016	0.001	0.015
SPE mean	0.892	1.000	0.774	0.998	0.783	0.998	0.643	0.985
SPE se	0.004	0.000	0.006	0.001	0.005	0.001	0.006	0.001
COR mean	0.198	0.430	0.038	0.672	0.026	0.408	0.004	0.358
COR se	0.018	0.022	0.009	0.021	0.007	0.022	0.003	0.021
SSB mean	1.699	0.372	1.658	0.273	2.167	0.413	2.049	0.384
SSB se	0.015	0.009	0.017	0.009	0.017	0.011	0.016	0.011
CLA mean	0.426	0.425	0.426	0.426	0.426	0.425	0.426	0.425
CLA se	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
KLD mean	213.545	89.995	215.481	68.061	258.021	99.845	241.264	93.387
KLD se	2.090	2.114	2.421	1.875	2.646	2.584	2.531	2.340

Table 3.5: Estimates by group Lasso (AGLasso), group Lasso with refitting (AGLasso-r) under the alternative penalization method for 7-category response: AG, APB, LF, MB, NH, Other and PS, the grouping structure for group Lasso is: each continuous covariate or Reserv is a group of its own, and all of the dummy variables are a group.

				AGLasso			
Covariates	AG	APB	LF	MB	NH	Other	PS
Intercept	3.885	1.964	-2.005	-1.896	1.298	-1.568	-1.679
Reserv	-2.665	0.783	0.331	0.115	0.206	-0.171	1.401
PolyNm	0.155	0.045	0.067	-0.004	0.071	-0.027	-0.173
PolyPr	–	–	–	–	–	–	–
MxPolyPr	–	–	–	–	–	–	–
log(TotOwn)	-0.275	0.116	0.013	0.146	0.044	0.174	-0.219
AvParcl	-0.001	0.001	0.002	–	–	-0.001	-0.001
IND	–	–	–	–	–	–	–
LBR	–	–	–	–	–	–	–
RE	–	–	–	–	–	–	–
RR	–	–	–	–	–	–	–

				AGLasso-r			
Covariates	AG	APB	LF	MB	NH	Other	PS
Intercept	5.175	2.425	-2.404	-2.597	1.475	-1.933	-2.141
Reserv	-3.754	0.215	0.808	0.080	-0.401	-0.669	3.722
PolyNm	–	–	–	–	–	–	–
PolyPr	–	–	–	–	–	–	–
MxPolyPr	–	–	–	–	–	–	–
log(TotOwn)	-0.317	0.176	0.103	0.372	0.094	0.376	-0.804
AvParcl	-0.002	–	0.002	-0.001	-0.001	-0.002	0.004
IND	–	–	–	–	–	–	–
LBR	–	–	–	–	–	–	–
RE	–	–	–	–	–	–	–
RR	–	–	–	–	–	–	–

Table 3.6: Estimates by group Lasso (AGLasso), group Lasso with refitting (AGLasso-r) under the alternative penalization method for 4-category response: AG, APB, NH and other, the grouping structure for group Lasso is:each continuous covariate or Reserv is a group of its own, and all of the dummy variables are a group.

	AGLasso			
Covariates	APB	AG	NH	Other
Intercept	0.376	2.331	-0.240	-2.467
Reserv	1.052	-2.364	0.427	0.886
PolyNm	0.043	0.139	-0.058	-0.125
PolyPr	–	–	–	–
MxPolyPr	–	–	–	–
log(TotOwn)	0.096	-0.294	0.008	0.189
AvParcl	0.001	-0.001	–	–
IND	–	–	–	–
LBR	–	–	–	–
RE	–	–	–	–
RR	–	–	–	–

	AGLasso-r			
Covariates	APB	AG	NH	Other
Intercept	0.452	3.109	-0.484	-3.078
Reserv	1.176	-2.783	0.557	1.050
PolyNm	–	–	–	–
PolyPr	–	–	–	–
MxPolyPr	–	–	–	–
log(TotOwn)	0.108	-0.365	0.023	0.235
AvParcl	0.001	-0.001	–	–
IND	–	–	–	–
LBR	–	–	–	–
RE	–	–	–	–
RR	–	–	–	–

3.9 Discussion

In summary, we have developed a group Lasso type of regularization method for simultaneous selection of covariates and estimation of parameters for a multinomial regression model. By pre-specifying groups, our method not only selects the coefficients associated with individual covariates as a group, but also selects the coefficients of dummy variables associated a categorical covariates as a group. Selection consistency and weak oracle property are established under suitable regularity conditions and an efficient computational algorithm is developed for multinomial regression models. Our simulation study has shown that our methods have desirable advantages over the standard Lasso methods in both variable selection and parameter estimation under different scenarios. We have also illustrated our method by a data example in the intersection of landscape ecology and environmental history. For future research, it would be interesting to adopt some of the innovations in computation such as a closed-form solution for a block update as in Meier et al. (2008).

Auto-multinomial Regression

We may further consider spatial automultinomial models as an extension of the multinomial model with spatial correlation. Similar to Section 3.2, let $\boldsymbol{y} = (y_1, \ldots, y_n)^T$ denote the categorical response variables at all n sites on this lattice and $\boldsymbol{y}_{-i} = (y_1, \ldots, y_{i-1}, y_i, \ldots, y_n)^T$ denote the vector that has all the response variables of $\boldsymbol{y}$ except for y_i. Further, consider a pre-specified spatial neighborhood structure. For example, the first-order neighborhood consists of the four nearest neighbors on a regular grid. Let $\mathcal{N}_i$ denote the set of indices of the neighbors of site i and let $i' \in \mathcal{N}_i$ denote that site i' is a neighbor of site i. In addition, let η_k be the constant autoregression coefficient for the kth category of response and $\boldsymbol{\eta} = (\eta_1, \ldots, \eta_K)^T$.

To model the response variables $\boldsymbol{y}$, we assume the probability of the ith response, y_i, conditional on $\boldsymbol{y}_{-i}$ depends on only the responses in the neighborhood, $y_{i'}$, where $i' \in \mathcal{N}_i$. That is,

$$\Pr(y_i \mid \boldsymbol{y}_{-i}) \quad = \quad \Pr(y_i \mid y_{i'} : i' \in \mathcal{N}_i).$$

Further, we assume that the conditional distribution $\text{pr}(y_i \mid y_{i'} : i' \in \mathcal{N}_i)$ is multinomial distributed with probability π_{ik} for the kth category at the ith site,

$$\pi_{ik} \quad = \quad \Pr(y_i = k \mid y_{i'} : i' \in \mathcal{N}_i)$$

$$\pi_{ik} = \frac{\exp(\theta_{ik} + \eta_k \sum_{i' \in \mathcal{N}_i} \mathbb{1}\{y_{i'} = k\})}{\sum_{l=1}^{K} \exp(\theta_{il} + \eta_l \sum_{i' \in \mathcal{N}_i} \mathbb{1}\{y_{i'} = l\})} \tag{3.18}$$

Similarly, we model the conditional probability with a sum-zero constraint:

$$\sum_{k=1}^{K} \beta_{k0} = 0, \ \sum_{k=1}^{K} \beta_{kgj} = 0, j = 1, \ldots, s_g, g = 1, \ldots, G, \sum_{k=1}^{K} \eta_k = 0.$$

Caragea and Kaiser (2009) proposed to center the autocovariate around its expected value to achieve more meaningful interpretations for regression purposes.

Here we consider this centered multinomial model by modifying (3.18) to

$$\pi_{ik} = \frac{\exp(\theta_{ik} + \eta_k \sum_{i' \in \mathcal{N}_i} [\mathbb{1}\{y_{i'} = k\} - \mu_{i'k}])}{\sum_{l=1}^{K} \exp(\theta_{il} + \eta_l \sum_{i' \in \mathcal{N}_i} [\mathbb{1}\{y_{i'} = l\} - \mu_{i'l}])}$$
$$\mu_{i'k} = \frac{\exp(\theta_{i'k})}{\sum_{l=1}^{K} \exp(\theta_{i'l})}$$

Maximum pseudolikelihood method can be used for estimating the model parameters, where the pseudolikelihood function is the product of the full conditional probabilities at all sites (Cressie, 1993).

Fused Lasso

A multi-stage procedure was proposed for variable selection which utilizes group Lasso and fused Lasso. Group Lasso is considered for selecting categorical covariates and covariates for different response categories as described in Section 3.3, while fused Lasso (Tibshirani et al., 2005) is for identifying homogeneity of covariates by penalizing the pairwise differences of the corresponding coefficients. Such a multi-stage procedure ideally will produce sparse and interpretable models.

For example, we can fuse the coefficients for different response categories within each group. The fused type Lasso penalized function is:

$$Q_n^f(\boldsymbol{\beta}; \boldsymbol{y}) = n^{-1}\ell_n(\boldsymbol{\beta}) - \lambda_n \sum_{g=1}^{G} \left[\sum_{j=1}^{s_g} \sum_{k \neq h} (\beta_{kgj} - \beta_{hgj})^2 \right]^{1/2}$$

Maximizing the penalized log-likelihood function Q_n^f gives the maximum penalized estimates by fused type Lasso. The second term in Q_n^f is the fused-type penalty, it enables the L_2 difference of the estimates for different response categories shrink to zero, which helps to identify the homogeneity of the coefficients for different response categories.

Appendix A

Technical detials in Chapter 2

A.1 The relationship between 0-1 coding with ±1 coding

For an uncentered model, we have

$$\begin{aligned}
p(Z_i|Z_{i'}: i' \sim i) &= \frac{\exp\{Z_i(\boldsymbol{x}_i'\boldsymbol{\beta} + \eta\sum_{i'\sim i} Z_{i'})\}}{1+\exp(\boldsymbol{x}_i'\boldsymbol{\beta} + \eta\sum_{i'\sim i} Z_{i'})} \\
&= \frac{\exp[\{(\tilde{Z}_i+1)/2\}\{\boldsymbol{x}_i'\boldsymbol{\beta} + \eta\sum_{i'\sim i}(\tilde{Z}_{i'}+1)/2\}]}{1+\exp\{\boldsymbol{x}_i'\boldsymbol{\beta} + \eta\sum_{i'\sim i}(\tilde{Z}_{i'}+1)/2\}} \\
&= \frac{\exp[\tilde{Z}_i\{(\boldsymbol{x_i'}\boldsymbol{\beta}/2 + |\mathcal{N}_i|\eta/4) + (\eta/4)\sum_{i'\sim i}\tilde{Z}_{i'}\}]}{2\cosh\{(\boldsymbol{x_i'}\boldsymbol{\beta}/2 + |\mathcal{N}_i|\eta/4) + (\eta/4)\sum_{i'\sim i}\tilde{Z}_{i'}\}}
\end{aligned}$$

Thus, $\tilde{\eta} = \eta/4$ and $\boldsymbol{x}_i'\tilde{\boldsymbol{\beta}} = \boldsymbol{x}_i'\boldsymbol{\beta}/2 + |\mathcal{N}_i|\eta/4$.

Now, for a centered model, we have

$$\begin{aligned}
& p(Z_i|Z_{i'} : i' \sim i) \\
= \; & \frac{\exp[Z_i\{\boldsymbol{x}_i'\boldsymbol{\beta} + \eta \sum_{i'\sim i}(Z_{i'} - \mu_{i'})\}]}{1 + \exp\{\boldsymbol{x}_i'\boldsymbol{\beta} + \eta \sum_{i'\sim i}(Z_{i'} - \mu_{i'})\}} \\
= \; & \frac{\exp[\{(\tilde{Z}_i + 1)/2\}\{\boldsymbol{x}_i'\boldsymbol{\beta} + \eta \sum_{i'\sim i}(\tilde{Z}_{i'}/2 + 1/2 - \mu_{i'})\}]}{1 + \exp\{\boldsymbol{x}_i'\boldsymbol{\beta} + \eta \sum_{i'\sim i}(\tilde{Z}_{i'}/2 + 1/2 - \mu_{i'})\}} \\
= \; & \frac{\exp[\tilde{Z}_i\{\boldsymbol{x}_{\boldsymbol{i}}'\boldsymbol{\beta}/2 + |\mathcal{N}_i|\eta/4 - \eta \sum_{i'\sim i}\mu_{i'}/2 + \eta \sum_{i'\sim i}\tilde{\mu}_{i'}/4 + (\eta/4)\sum_{i'\sim i}(\tilde{Z}_{i'} - \tilde{\mu}_{i'})\}]}{2\cosh\{\boldsymbol{x}_{\boldsymbol{i}}'\boldsymbol{\beta}/2 + |\mathcal{N}_i|\eta/4 - \eta \sum_{i'\sim i}\mu_{i'}/2 + \eta \sum_{i'\sim i}\tilde{\mu}_{i'}/4 + (\eta/4)\sum_{i'\sim i}(\tilde{Z}_{i'} - \tilde{\mu}_{i'})\}}
\end{aligned}$$

Thus, $\tilde{\eta} = \eta/4$ and $\boldsymbol{x}_i'\tilde{\boldsymbol{\beta}} = \boldsymbol{x}_{\boldsymbol{i}}'\boldsymbol{\beta}/2 + |\mathcal{N}_i|\eta/4 - \eta \sum_{i'\sim i}\mu_{i'}/2 + \eta \sum_{i'\sim i}\tilde{\mu}_{i'}/4$

A.2 Formulas for $\widehat{Var}(\widehat{\tilde{\boldsymbol{\theta}}}_{\mathrm{p}})$

0-1 Coding

For the uncentered model and the 0-1 coding, the $\boldsymbol{\mathcal{J}}(\boldsymbol{\theta})$ and $\boldsymbol{\mathcal{I}}(\boldsymbol{\theta})$ are

$$\begin{aligned}
\boldsymbol{\mathcal{J}}(\boldsymbol{\theta}) \; = \; & \sum_{i=1}^{n} \sum_{i'\sim i, i'=i} \left[\begin{array}{c} \boldsymbol{x}_i \left\{ Z_i - \frac{\exp(\boldsymbol{x}_i'\boldsymbol{\beta} + \eta \sum_{i'\sim i} Z_{i'})}{1+\exp(\boldsymbol{x}_i'\boldsymbol{\beta} + \eta \sum_{i'\sim i} Z_{i'})} \right\} \\ (\sum_{i'\sim i} Z_{i'}) \left\{ Z_i - \frac{\exp(\boldsymbol{x}_i'\boldsymbol{\beta} + \eta \sum_{i'\sim i} Z_{i'})}{1+\exp(\boldsymbol{x}_i'\boldsymbol{\beta} + \eta \sum_{i'\sim i} Z_{i'})} \right\} \end{array} \right] \\
& \left[\begin{array}{c} \boldsymbol{x}_{i'} \left\{ Z_{i'} - \frac{\exp(\boldsymbol{x}_{i'}'\boldsymbol{\beta} + \eta \sum_{i''\sim i'} Z_{i''})}{1+\exp(\boldsymbol{x}_{i'}'\boldsymbol{\beta} + \eta \sum_{i''\sim i'} Z_{i''})} \right\} \\ (\sum_{i''\sim i'} Z_{i''}) \left\{ Z_{i'} - \frac{\exp(\boldsymbol{x}_{i'}'\boldsymbol{\beta} + \eta \sum_{i''\sim i'} Z_{i''})}{1+\exp(\boldsymbol{x}_{i'}'\boldsymbol{\beta} + \eta \sum_{i''\sim i'} Z_{i''})} \right\} \end{array} \right]'.
\end{aligned}$$

and

$$\boldsymbol{\mathcal{I}}(\boldsymbol{\theta}) \quad = \quad \begin{pmatrix} \boldsymbol{\mathcal{I}}(\boldsymbol{\beta}) & -\frac{\partial^2 \ell_{\mathrm{p}}(\boldsymbol{\theta})}{\partial \boldsymbol{\beta} \partial \eta} \\ -\frac{\partial^2 \ell_{\mathrm{p}}(\boldsymbol{\theta})}{\partial \eta \partial \boldsymbol{\beta}'} & -\frac{\partial^2 \ell_{\mathrm{p}}(\boldsymbol{\theta})}{\partial \eta^2} \end{pmatrix}$$

where

$$\begin{aligned} \boldsymbol{\mathcal{I}}(\boldsymbol{\beta}) \quad &= \quad \sum_{i=1}^{n} \boldsymbol{x}_i \boldsymbol{x}_i' \frac{\exp(\boldsymbol{x}_i'\boldsymbol{\beta} + \eta \sum_{i' \sim i} Z_{i'})}{\{1 + \exp(\boldsymbol{x}_i'\boldsymbol{\beta} + \eta \sum_{i' \sim i} Z_{i'})\}^2} \\ -\frac{\partial^2 \ell_{\mathrm{p}}(\boldsymbol{\theta})}{\partial \eta^2} \quad &= \quad \sum_{i=1}^{n} \frac{(\sum_{i' \sim i} Z_{i'})^2 \exp(\boldsymbol{x}_i'\boldsymbol{\beta} + \eta \sum_{i' \sim i} Z_{i'})}{\{1 + \exp(\boldsymbol{x}_i'\boldsymbol{\beta} + \eta \sum_{i' \sim i} Z_{i'})\}^2} \\ -\frac{\partial^2 \ell_{\mathrm{p}}(\boldsymbol{\theta})}{\partial \boldsymbol{\beta} \partial \eta} \quad &= \quad \sum_{i=1}^{n} \boldsymbol{x}_i \frac{(\sum_{i' \sim i} Z_{i'}) \exp(\boldsymbol{x}_i'\boldsymbol{\beta} + \eta \sum_{i' \sim i} Z_{i'})}{\{1 + \exp(\boldsymbol{x}_i'\boldsymbol{\beta} + \eta \sum_{i' \sim i} Z_{i'})\}^2} \end{aligned}$$

For the centered model,

$$\boldsymbol{\mathcal{J}}(\boldsymbol{\beta}) = \sum_{i=1}^{n} \sum_{i' \sim i, i' = i} \left[\begin{array}{c} \left\{ \boldsymbol{x}_i - \eta \sum_{i' \sim i} \frac{\exp(\boldsymbol{x}_{i'}'\boldsymbol{\beta})}{(1+\exp(\boldsymbol{x}_{i'}'\boldsymbol{\beta}))^2} \boldsymbol{x}_{i'} \right\} \left\{ Z_i - \frac{\exp(\boldsymbol{x}_i'\boldsymbol{\beta}+\eta \sum_{i' \sim i}(Z_{i'}-\mu_{i'}))}{1+\exp(\boldsymbol{x}_i'\boldsymbol{\beta}+\eta \sum_{i' \sim i}(Z_{i'}-\mu_{i'}))} \right\} \\ \{\sum_{i' \sim i}(Z_{i'} - \mu_{i'})\} \left\{ Z_i - \frac{\exp(\boldsymbol{x}_i'\boldsymbol{\beta}+\eta \sum_{i' \sim i}(Z_{i'}-\mu_{i'}))}{1+\exp(\boldsymbol{x}_i'\boldsymbol{\beta}+\eta \sum_{i' \sim i}(Z_{i'}-\mu_{i'}))} \right\} \end{array} \right]$$

$$\left[\begin{array}{c} \left\{ \boldsymbol{x}_{i'} - \eta \sum_{i'' \sim i'} \frac{\exp(\boldsymbol{x}_{i''}'\boldsymbol{\beta})}{(1+\exp(\boldsymbol{x}_{i''}'\boldsymbol{\beta}))^2} \boldsymbol{x}_{i''} \right\} \left\{ Z_{i'} - \frac{\exp(\boldsymbol{x}_{i'}'\boldsymbol{\beta}+\eta \sum_{i'' \sim i'}(Z_{i''}-\mu_{i''}))}{1+\exp(\boldsymbol{x}_{i'}'\boldsymbol{\beta}+\eta \sum_{i'' \sim i'}(Z_{i''}-\mu_{i''}))} \right\} \\ \{\sum_{i'' \sim i'}(Z_{i''} - \mu_{i''})\} \left\{ Z_{i'} - \frac{\exp(\boldsymbol{x}_{i'}'\boldsymbol{\beta}+\eta \sum_{i'' \sim i'}(Z_{i''}-\mu_{i''}))}{1+\exp(\boldsymbol{x}_{i'}'\boldsymbol{\beta}+\eta \sum_{i'' \sim i'}(Z_{i''}-\mu_{i''}))} \right\} \end{array} \right]' .$$

and

$$\boldsymbol{\mathcal{I}}(\boldsymbol{\theta}) \quad = \quad \begin{pmatrix} \boldsymbol{\mathcal{I}}(\boldsymbol{\beta}) & -\frac{\partial^2 \ell_{\mathrm{p}}(\boldsymbol{\theta})}{\partial \boldsymbol{\beta} \partial \eta} \\ -\frac{\partial^2 \ell_{\mathrm{p}}(\boldsymbol{\theta})}{\partial \eta \partial \boldsymbol{\beta}'} & -\frac{\partial^2 \ell_{\mathrm{p}}(\boldsymbol{\theta})}{\partial \eta^2} \end{pmatrix}$$

where

$$
\begin{aligned}
&\boldsymbol{\mathcal{I}}(\boldsymbol{\beta}) = \sum_{i=1}^{n} \left[\boldsymbol{x}_i - \eta \sum_{i' \sim i} \frac{\exp(\boldsymbol{x}_{i'}'\boldsymbol{\beta})}{\{1+\exp(\boldsymbol{x}_{i'}'\boldsymbol{\beta})\}^2} \boldsymbol{x}_{i'} \right] \left[\boldsymbol{x}_i - \eta \sum_{i' \sim i} \frac{\exp(\boldsymbol{x}_{i'}'\boldsymbol{\beta})}{\{1+\exp(\boldsymbol{x}_{i'}'\boldsymbol{\beta})\}^2} \boldsymbol{x}_{i'} \right]' \\
&\qquad\qquad\qquad \frac{\exp(\boldsymbol{x}_i'\boldsymbol{\beta} + \eta \sum_{i' \sim i}(Z_{i'} - \mu_{i'}))}{\{1+\exp(\boldsymbol{x}_i'\boldsymbol{\beta} + \eta \sum_{i' \sim i}(Z_{i'} - \mu_{i'}))\}^2} \\
&+ \sum_{i=1}^{n} \sum_{i' \sim i} \eta \left\{ Z_i - \frac{\exp(\boldsymbol{x}_i'\boldsymbol{\beta} + \eta \sum_{i'' \sim i}(Z_{i''} - \mu_{i''}))}{1+\exp(\boldsymbol{x}_i'\boldsymbol{\beta} + \eta \sum_{i'' \sim i}(Z_{i''} - \mu_{i''}))} \right\} \\
&\frac{\exp(\boldsymbol{x}_{i'}'\boldsymbol{\beta}) - \{\exp(\boldsymbol{x}_{i'}'\boldsymbol{\beta})\}^2}{\{1+\exp(\boldsymbol{x}_{i'}'\boldsymbol{\beta})\}^3} \boldsymbol{x}_{i'} \boldsymbol{x}_{i'}'
\end{aligned}
$$

$$
-\frac{\partial^2 \ell_{\mathrm{p}}(\boldsymbol{\theta})}{\partial \eta^2} = \frac{(\sum_{i' \sim i}(Z_{i'} - \mu_{i'}))^2 \exp(\boldsymbol{x}_i'\boldsymbol{\beta} + \eta \sum_{i' \sim i}(Z_{i'} - \mu_{i'}))}{\{1+\exp(\boldsymbol{x}_i'\boldsymbol{\beta} + \eta \sum_{i' \sim i}(Z_{i'} - \mu_{i'}))\}^2}
$$

$$
\begin{aligned}
&-\frac{\partial^2 \ell_{\mathrm{p}}(\boldsymbol{\theta})}{\partial \boldsymbol{\beta} \partial \eta} \\
= \;& \sum_{i=1}^{n} \left[\boldsymbol{x}_i - \eta \sum_{i' \sim i} \frac{\exp(\boldsymbol{x}_{i'}'\boldsymbol{\beta})}{\{1+\exp(\boldsymbol{x}_{i'}'\boldsymbol{\beta})\}^2} \boldsymbol{x}_{i'} \right] \frac{(\sum_{i' \sim i}(Z_{i'} - \mu_{i'})) \exp(\boldsymbol{x}_i'\boldsymbol{\beta} + \eta \sum_{i' \sim i}(Z_{i'} - \mu_{i'}))}{\{1+\exp(\boldsymbol{x}_i'\boldsymbol{\beta} + \eta \sum_{i' \sim i}(Z_{i'} - \mu_{i'}))\}^2} \\
&+ \sum_{i=1}^{n} \sum_{i' \sim i} \left\{ Z_i - \frac{\exp(\boldsymbol{x}_i'\boldsymbol{\beta} + \eta \sum_{i'' \sim i}(Z_{i''} - \mu_{i''}))}{1+\exp(\boldsymbol{x}_i'\boldsymbol{\beta} + \eta \sum_{i'' \sim i}(Z_{i''} - \mu_{i''}))} \right\} \frac{\exp(\boldsymbol{x}_{i'}'\boldsymbol{\beta})}{\{1+\exp(\boldsymbol{x}_{i'}'\boldsymbol{\beta})\}^2} \boldsymbol{x}_{i'}
\end{aligned}
$$

±1 **Coding**

For the uncentered model and the ±1 coding, we have

$$
\begin{aligned}
\boldsymbol{\mathcal{J}}(\tilde{\boldsymbol{\theta}}) \;=\; \sum_{i=1}^{n} \sum_{i'\sim i, i'=i} & \left[\begin{array}{c} \boldsymbol{x}_i \left\{ \tilde{Z}_i - \frac{\sinh(\boldsymbol{x}_i'\tilde{\boldsymbol{\beta}}+\tilde{\eta}\sum_{i'\sim i}\tilde{Z}_{i'})}{\cosh(\boldsymbol{x}_i'\tilde{\boldsymbol{\beta}}+\tilde{\eta}\sum_{i'\sim i}\tilde{Z}_{i'})} \right\} \\ (\sum_{i'\sim i}\tilde{Z}_{i'}) \left\{ \tilde{Z}_i - \frac{\sinh(\boldsymbol{x}_i'\tilde{\boldsymbol{\beta}}+\tilde{\eta}\sum_{i'\sim i}\tilde{Z}_{i'})}{\cosh(\boldsymbol{x}_i'\tilde{\boldsymbol{\beta}}+\tilde{\eta}\sum_{i'\sim i}\tilde{Z}_{i'})} \right\} \end{array} \right] \\
& \left[\begin{array}{c} \boldsymbol{x}_{i'} \left\{ \tilde{Z}_{i'} - \frac{\sinh(\boldsymbol{x}_{i'}'\tilde{\boldsymbol{\beta}}+\tilde{\eta}\sum_{i''\sim i'}\tilde{Z}_{i''})}{\cosh(\boldsymbol{x}_{i'}'\tilde{\boldsymbol{\beta}}+\tilde{\eta}\sum_{i''\sim i'}\tilde{Z}_{i''})} \right\} \\ (\sum_{i''\sim i'}\tilde{Z}_{i''}) \left\{ \tilde{Z}_{i'} - \frac{\sinh(\boldsymbol{x}_{i'}'\tilde{\boldsymbol{\beta}}+\tilde{\eta}\sum_{i''\sim i'}\tilde{Z}_{i''})}{\cosh(\boldsymbol{x}_{i'}'\tilde{\boldsymbol{\beta}}+\tilde{\eta}\sum_{i''\sim i'}\tilde{Z}_{i''})} \right\} \end{array} \right]' .
\end{aligned}
$$

and

$$
\boldsymbol{\mathcal{I}}(\tilde{\boldsymbol{\theta}}) \;=\; \begin{pmatrix} \boldsymbol{\mathcal{I}}(\tilde{\boldsymbol{\beta}}) & -\frac{\partial^2 \ell_{\mathrm{p}}(\tilde{\boldsymbol{\theta}})}{\partial\tilde{\boldsymbol{\beta}}\partial\tilde{\eta}} \\ -\frac{\partial^2 \ell_{\mathrm{p}}(\tilde{\boldsymbol{\theta}})}{\partial\tilde{\eta}\partial\tilde{\boldsymbol{\beta}}'} & -\frac{\partial^2 \ell_{\mathrm{p}}(\tilde{\boldsymbol{\theta}})}{\partial\tilde{\eta}^2} \end{pmatrix}
$$

where

$$
\begin{aligned}
\boldsymbol{\mathcal{I}}(\tilde{\boldsymbol{\beta}}) &= \sum_{i=1}^{n} \boldsymbol{x}_i \boldsymbol{x}_i' \left[1 - \left\{ \frac{\sinh(\boldsymbol{x}_i'\tilde{\boldsymbol{\beta}} + \tilde{\eta}\sum_{i'\sim i}\tilde{Z}_{i'})}{\cosh(\boldsymbol{x}_i'\tilde{\boldsymbol{\beta}} + \tilde{\eta}\sum_{i'\sim i}\tilde{Z}_{i'})} \right\}^2 \right] \\
-\frac{\partial^2 \ell_{\mathrm{p}}(\tilde{\boldsymbol{\theta}})}{\partial\tilde{\eta}^2} &= \sum_{i=1}^{n} \left\{ \frac{\sum_{i'\sim i}\tilde{Z}_{i'}}{\cosh(\boldsymbol{x}_i'\tilde{\boldsymbol{\beta}} + \tilde{\eta}\sum_{i'\sim i}\tilde{Z}_{i'})} \right\}^2 \\
-\frac{\partial^2 \ell_{\mathrm{p}}(\tilde{\boldsymbol{\theta}})}{\partial\tilde{\boldsymbol{\beta}}\partial\tilde{\eta}} &= \sum_{i=1}^{n} \frac{\boldsymbol{x_i}\sum_{i'\sim i}\tilde{Z}_{i'}}{\{\cosh(\boldsymbol{x}_i'\tilde{\boldsymbol{\beta}} + \tilde{\eta}\sum_{i'\sim i}\tilde{Z}_{i'})\}^2}
\end{aligned}
$$

For the centered model,

$$
\begin{aligned}
&\boldsymbol{\mathcal{J}}(\tilde{\boldsymbol{\beta}})\\
= \;& \sum_{i=1}^{n} \sum_{i' \sim i, i' = i} \left[\begin{array}{c} \left\{ \boldsymbol{x}_i - \tilde{\eta} \sum_{i' \sim i} (\cosh(\boldsymbol{x}_{i'}' \tilde{\boldsymbol{\beta}}))^{-2} \boldsymbol{x}_{i'} \right\} \left\{ \tilde{Z}_i - \frac{\sinh(\boldsymbol{x}_i' \tilde{\boldsymbol{\beta}} + \tilde{\eta} \sum_{i' \sim i} (\tilde{Z}_{i'} - \tilde{\mu}_{i'}))}{\cosh(\boldsymbol{x}_i' \tilde{\boldsymbol{\beta}} + \tilde{\eta} \sum_{i' \sim i} (\tilde{Z}_{i'} - \tilde{\mu}_{i'}))} \right\} \\ \left\{ \sum_{i' \sim i} (\tilde{Z}_{i'} - \tilde{\mu}_{i'}) \right\} \left\{ \tilde{Z}_i - \frac{\sinh(\boldsymbol{x}_i' \tilde{\boldsymbol{\beta}} + \tilde{\eta} \sum_{i' \sim i} (\tilde{Z}_{i'} - \tilde{\mu}_{i'}))}{\cosh(\boldsymbol{x}_i' \tilde{\boldsymbol{\beta}} + \tilde{\eta} \sum_{i' \sim i} (\tilde{Z}_{i'} - \tilde{\mu}_{i'}))} \right\} \end{array} \right] \\
&\qquad \left[\begin{array}{c} \left\{ \boldsymbol{x}_{i'} - \tilde{\eta} \sum_{i'' \sim i'} (\cosh(\boldsymbol{x}_{i''}' \tilde{\boldsymbol{\beta}}))^{-2} \boldsymbol{x}_{i''} \right\} \left\{ \tilde{Z}_{i'} - \frac{\sinh(\boldsymbol{x}_{i'}' \tilde{\boldsymbol{\beta}} + \tilde{\eta} \sum_{i'' \sim i'} (\tilde{Z}_{i''} - \tilde{\mu}_{i''}))}{\cosh(\boldsymbol{x}_{i'}' \tilde{\boldsymbol{\beta}} + \tilde{\eta} \sum_{i'' \sim i'} (\tilde{Z}_{i''} - \tilde{\mu}_{i''}))} \right\} \\ \left\{ \sum_{i'' \sim i'} (\tilde{Z}_{i''} - \tilde{\mu}_{i''}) \right\} \left\{ \tilde{Z}_{i'} - \frac{\sinh(\boldsymbol{x}_{i'}' \tilde{\boldsymbol{\beta}} + \tilde{\eta} \sum_{i'' \sim i'} (\tilde{Z}_{i''} - \tilde{\mu}_{i''}))}{\cosh(\boldsymbol{x}_{i'}' \tilde{\boldsymbol{\beta}} + \tilde{\eta} \sum_{i'' \sim i'} (\tilde{Z}_{i''} - \tilde{\mu}_{i''}))} \right\} \end{array} \right]'.
\end{aligned}
$$

and

$$
\boldsymbol{\mathcal{I}}(\tilde{\boldsymbol{\theta}}) \;=\; \begin{pmatrix} \boldsymbol{\mathcal{I}}(\tilde{\boldsymbol{\beta}}) & -\frac{\partial^2 \ell_{\mathrm{p}}(\tilde{\boldsymbol{\theta}})}{\partial \tilde{\boldsymbol{\beta}} \partial \tilde{\eta}} \\ -\frac{\partial^2 \ell_{\mathrm{p}}(\tilde{\boldsymbol{\theta}})}{\partial \tilde{\eta} \partial \tilde{\boldsymbol{\beta}}'} & -\frac{\partial^2 \ell_{\mathrm{p}}(\tilde{\boldsymbol{\theta}})}{\partial \tilde{\eta}^2} \end{pmatrix}
$$

where

$$
\begin{aligned}
\boldsymbol{\mathcal{I}}(\tilde{\boldsymbol{\beta}}) = & \sum_{i=1}^{n} \left[\boldsymbol{x}_i - \tilde{\eta} \sum_{i' \sim i} \frac{1}{\{\cosh(\boldsymbol{x}_{i'}' \tilde{\boldsymbol{\beta}})\}^2} \boldsymbol{x}_{i'} \right] \left[\boldsymbol{x}_i - \tilde{\eta} \sum_{i' \sim i} \frac{1}{\{\cosh(\boldsymbol{x}_{i'}' \tilde{\boldsymbol{\beta}})\}^2} \boldsymbol{x}_{i'} \right]' \\
& \cosh[\boldsymbol{x}_i' \tilde{\boldsymbol{\beta}} + \tilde{\eta} \sum_{i' \sim i} (\tilde{Z}_{i'} - \tilde{\mu}_{i'})]^{-2} \\
& - \sum_{i=1}^{n} \sum_{i' \sim i} 2\tilde{\eta} \left\{ \tilde{Z}_i - \frac{\sinh(\boldsymbol{x}_i' \tilde{\boldsymbol{\beta}} + \tilde{\eta} \sum_{i'' \sim i} (\tilde{Z}_{i''} - \tilde{\mu}_{i''}))}{\cosh(\boldsymbol{x}_i' \tilde{\boldsymbol{\beta}} + \tilde{\eta} \sum_{i'' \sim i} (\tilde{Z}_{i''} - \tilde{\mu}_{i''}))} \right\} \frac{\sinh(\boldsymbol{x}_{i'}' \tilde{\boldsymbol{\beta}})}{\{\cosh(\boldsymbol{x}_{i'}' \tilde{\boldsymbol{\beta}}\}^3} \boldsymbol{x}_{i'} \boldsymbol{x}_{i'}'
\end{aligned}
$$

$$
-\frac{\partial^2 \ell_{\mathrm{p}}(\tilde{\boldsymbol{\theta}})}{\partial \tilde{\eta}^2} \;=\; \sum_{i=1}^{n} \frac{\{\sum_{i' \sim i} (\tilde{Z}_{i'} - \tilde{\mu}_{i'})\}^2}{\cosh(\boldsymbol{x}_i' \tilde{\boldsymbol{\beta}} + \tilde{\eta} \sum_{i' \sim i} (\tilde{Z}_{i'} - \tilde{\mu}_{i'}))}
$$

$$\begin{aligned} & -\frac{\partial^2 \ell_{\mathrm{p}}(\tilde{\boldsymbol{\theta}})}{\partial \tilde{\boldsymbol{\beta}} \partial \tilde{\eta}} \\ = \ & \sum_{i=1}^{n} \left\{ \sum_{i' \sim i} (\cosh(\boldsymbol{x}'_{i'} \tilde{\boldsymbol{\beta}}))^{-2} \boldsymbol{x}_{i'} \right\} \left\{ \tilde{Z}_i - \frac{\sinh(\boldsymbol{x}'_i \tilde{\boldsymbol{\beta}} + \tilde{\eta} \sum_{i' \sim i} (\tilde{Z}_{i'} - \tilde{\mu}_{i'}))}{\cosh(\boldsymbol{x}'_i \tilde{\boldsymbol{\beta}} + \tilde{\eta} \sum_{i' \sim i} (\tilde{Z}_{i'} - \tilde{\mu}_{i'}))} \right\} \\ & + \sum_{i=1}^{n} \left\{ \sum_{i' \sim i} (\tilde{Z}_{i'} - \tilde{\mu}_{i'}) \right\} \frac{\boldsymbol{x}_i - \tilde{\eta} \sum_{i' \sim i} (\cosh(\boldsymbol{x}'_{i'} \tilde{\boldsymbol{\beta}}))^{-2} \boldsymbol{x}_{i'}}{\{\cosh(\boldsymbol{x}'_i \tilde{\boldsymbol{\beta}} + \tilde{\eta} \sum_{i' \sim i} (\tilde{Z}_{i'} - \tilde{\mu}_{i'}))\}^2} \end{aligned}$$

Appendix B

Proof for Chapter 3

B.1 The group Lasso estimates satisfied the sum-zero constrain

Proof. Suppose $\hat{\boldsymbol{\beta}} = \{\hat{\beta}_{0k}, \hat{\beta}_{kgj}, k = 1, \ldots, K, j = 1, \ldots, s_g, g = 1, \ldots, G\}$ is the maximizer of (3.4) but not satisfied the constraint (3.2). Let $\hat{c}_0 = \sum_{k=1}^{K} \hat{\beta}_{k0}/K$, $\hat{c}_{gj} = \sum_{k=1}^{K} \hat{\beta}_{kgj}/K, j = 1, \ldots, s_g, g = 1, \ldots, G$. Then let $\hat{\boldsymbol{\beta}}_c = \{\hat{\beta}_{k0} - \hat{c}_0, \hat{\beta}_{kgj} - \hat{c}_{gj}, j = 1, \ldots, s_g, g = 1, \ldots, G, k = 1, \ldots, K\}$. $\ell(\hat{\boldsymbol{\beta}}) = \ell(\hat{\boldsymbol{\beta}}_c)$. However $\sum_{k=1}^{K} \hat{\beta}_{kgj}^2 > \sum_{k=1}^{K} (\hat{\beta}_{kgj} - \hat{c}_{gj})^2$ for $j = 1 \ldots, s_g, g = 1, \ldots, G$, then $\hat{\boldsymbol{\beta}}$ cannot be the maximizer of (3.4). □

B.2 Lemma 3.1 Proof

Proof.

$$\begin{aligned}\ell_n^g(\boldsymbol{\beta}_g \mid \hat{\boldsymbol{\beta}}_{-g}) = \quad & \sum_{k=1}^{K}\left\{\mathbb{1}(\boldsymbol{y} = k\mathbf{1})^T(\hat{\beta}_{k0}\mathbf{1} + \sum_{j=1}^{s_g}\boldsymbol{x}_{gj}\beta_{kgj} + \sum_{h\neq g}\sum_{j=1}^{s_h}\boldsymbol{x}_{hj}\hat{\beta}_{khj})\right\} \\ & -\sum_{i=1}^{N}\log\{\sum_{k=1}^{K}\exp(\hat{\beta}_{k0} + \sum_{j=1}^{s_g}x_{gji}\beta_{kgj} + \sum_{h\neq g}\sum_{j=1}^{s_h}x_{hji}\hat{\beta}_{khj})\}.\end{aligned}$$

By the standard argument on the Taylor expansion of the likelihood function $\ell_n^g(\boldsymbol{\beta}_g \mid \hat{\boldsymbol{\beta}})$ at $\mathbf{0}$, we have

$$\ell_n^g(\boldsymbol{\beta}_g \mid \hat{\boldsymbol{\beta}}_{-g}) = \ell_n^g(\mathbf{0} \mid \hat{\boldsymbol{\beta}}_{-g}) + \nabla\ell_n^g(\mathbf{0} \mid \hat{\boldsymbol{\beta}}_{-g})\boldsymbol{\beta}_g + 2^{-1}\boldsymbol{\beta}_g\nabla^2\ell_n^g(\mathbf{0} \mid \hat{\boldsymbol{\beta}}_{-g})\boldsymbol{\beta}_g(1 + o(1))$$

If $\|\nabla\ell_n^g(\boldsymbol{\beta}_g \mid \hat{\boldsymbol{\beta}}_{-g})|_{\boldsymbol{\beta}_g=\mathbf{0}}\|_2 \leq n{s_g}^{1/2}\lambda_n$, then $\nabla\ell_n^g(\mathbf{0} \mid \hat{\boldsymbol{\beta}}_{-g})\boldsymbol{\beta}_g - n\lambda_n s_g^{1/2}\|\boldsymbol{\beta}_g\|_2 \leq 0$. Since the log-likelihood function is concave (Davidson and G.J., 2003), then $\boldsymbol{\beta}_g\nabla^2\ell_n^g(\mathbf{0} \mid \hat{\boldsymbol{\beta}}_{-g})\boldsymbol{\beta}_g \leq 0$, and clearly, $\boldsymbol{\beta}_g = \mathbf{0}$ is the maximizer. □

B.3 Proposition 3.2 Proof

Proof. For ease of presentation, assume $s_g = 1$ for $g = 1, \ldots, G$. When $s_g > 1$, the proof still holds with a similar procedure. Here, $\boldsymbol{\beta} = (\beta_{kgj}, k = 1, \ldots, K-1, j =$

$1, g = 1, \ldots, G)^T$, and

$$\ell_n(\boldsymbol{\beta}; \boldsymbol{y}) = \sum_{k=1}^{K-1} \left[\{\mathbb{1}(\boldsymbol{y} = k\mathbf{1}) - \mathbb{1}(\boldsymbol{y} = K\mathbf{1})\}^T (\sum_{g=1}^{G} \sum_{j=1}^{s_g} \boldsymbol{x}_{gj} \beta_{kgj}) \right] -$$
$$\sum_{i=1}^{n} \log \left\{ \sum_{k=1}^{K-1} \exp(\theta_{ik}) + \exp(-\sum_{k=1}^{K-1} \theta_{ik}) \right\}$$

We rewrite the maximization of (3.4) in the form

$$\begin{aligned} &\min_{\boldsymbol{\beta}, \boldsymbol{v} \in \mathcal{R}^G} -n^{-1} \ell_n(\boldsymbol{\beta}) + \lambda_n \sum_{g=1}^{G} s_g^{1/2} v_g, \\ &\text{s.t. } \|M \boldsymbol{\beta}_g\|_2 \le v_g, \ g = 1, \ldots, G. \end{aligned} \tag{B.1}$$

where $\boldsymbol{\beta}_g = (\beta_{1g1}, \ldots, \beta_{(K-1)g1})^T$ and

$$M = \begin{bmatrix} 1 & & \\ & \ddots & \\ & & 1 \\ -1 & \ldots & -1 \end{bmatrix}_{K \times (K-1)} .$$

Then we consider the Lagrangian with dual variable $\{\boldsymbol{\zeta}_g, \gamma_g\} \in R^{s_g(K-1)} \times R$ (Boyd and Vandenberghe, 2003):

$$L(\boldsymbol{\beta}, \boldsymbol{v}, \boldsymbol{\zeta}, \boldsymbol{\gamma}) = -n^{-1} \ell_n(\boldsymbol{\beta}) + \lambda_n \sum_{g=1}^{G} s_g^{1/2} v_g - \sum_{g=1}^{G} ((M\boldsymbol{\beta}_g)^T M \boldsymbol{\zeta}_g + v_g \gamma_g),$$
$$\text{s.t. } \|M \boldsymbol{\zeta}_g\|_2 \le \gamma_g, \ g = 1, \ldots, G.$$

For $k = 1, \dots, K-1$, define
$\boldsymbol{\mu}_k(\boldsymbol{\theta}) = \Bigg(\Big\{ \exp(\theta_{ik}) - \exp(-\sum_{k=1}^{K-1} \theta_{ik}) \Big\} \Big\{ \sum_{k=1}^{K-1} \exp(\theta_{ik}) + \exp(-\sum_{k=1}^{K-1} \theta_{ik}) \Big\}^{-1}$
$, i = 1, \dots, n \Bigg)^T$. The KKT conditions are: for $g = 1, \dots, G$,

$$\begin{bmatrix} -n^{-1}\boldsymbol{x}_{g1}^T\{\mathbb{1}(\boldsymbol{y}=1) - \mathbb{1}(\boldsymbol{y}=K) - \boldsymbol{\mu}_1(\boldsymbol{\theta})\} \\ \vdots \\ -n^{-1}\boldsymbol{x}_{g1}^T\{\mathbb{1}(\boldsymbol{y}=K-1)\mathbb{1}(\boldsymbol{y}=K) - \boldsymbol{\mu}_{K-1}(\boldsymbol{\theta})\} \end{bmatrix} = M^T M \boldsymbol{\zeta}_g, \tag{B.2}$$

$$\lambda_n s_g^{1/2} = \gamma_g, \tag{B.3}$$

$$\|M\boldsymbol{\beta}_g\|_2 \leq v_g, \tag{B.4}$$

$$\|M\boldsymbol{\zeta}_g\|_2 \leq \gamma_g, \tag{B.5}$$

$$(M\boldsymbol{\beta}_g)^T M\boldsymbol{\zeta}_g + v_g\gamma_g = 0. \tag{B.6}$$

where (B.2) and (B.3) are for stationarity, (B.4) is for primary feasibility, (B.5) is for dual feasibility and (B.6) is for complimentary slackness. If $\boldsymbol{\beta}_g = \mathbf{0}$, then we just need $\|M\boldsymbol{\zeta}_g\|_2 \leq \lambda_n s_g^{1/2}$. Denote the left hand side of (B.2) is $\nabla \ell_n^g$, then $(M^TM)^{-1}\nabla \ell_n^g = \boldsymbol{\zeta}_g$.

$$\|M\boldsymbol{\zeta}_g\|_2 = (\boldsymbol{\zeta}_g^T M^T M \boldsymbol{\zeta}_g)^{1/2} = \{(\nabla \ell_n^g)^T M^T M \nabla \ell_n^g\}^{1/2} = \|M\nabla \ell_n^g\|_2$$

Then it reduce to and (3.8). If $\boldsymbol{\beta}_g \neq \mathbf{0}$, we can have $M\boldsymbol{\zeta}_g = -M\boldsymbol{\beta}_g \lambda_n s_g^{1/2}(\|M\boldsymbol{\beta}_g\|_2)^{-1}$, plug into (B.2) then

$$\begin{bmatrix} -n^{-1}\boldsymbol{x}_{g1}^T\{\mathbb{1}(\boldsymbol{y}=1)-\mathbb{1}(\boldsymbol{y}=K)-\boldsymbol{\mu}_1(\boldsymbol{\theta})\} \\ \vdots \\ -n^{-1}\boldsymbol{x}_{g1}^T\{\mathbb{1}(\boldsymbol{y}=K-1)\mathbb{1}(\boldsymbol{y}=K)-\boldsymbol{\mu}_{K-1}(\boldsymbol{\theta})\} \end{bmatrix} = -M^TM\boldsymbol{\beta}_g\lambda_n s_g^{1/2}(\|M\boldsymbol{\beta}_g\|_2)^{-1}.$$

Note that M^TM is a $K-1\times K-1$ matrix with diagonal element 2, off-diagonal element 1, this can reduce to (3.7). Since the dual and primal problems are strictly feasible and the $\ell_n(\boldsymbol{\beta})$ is concave, the KKT condition is a sufficient and necessary condition for primary and dual optimal (Boyd and Vandenberghe, 2003). When $s_g > 1$, the KKT condition can be easily proved by updating M to M', where M' would be a block diagonal matrix, where there are s_g diagonal blocks and each block is M. □

B.4 Theorem 3.3 Proof

First, we need to show the existence of such λ_n. It's easy to prove by providing a special example is $\lambda_n = n^{-1/2}(\log n)^{3/5}$. Second, we prove the Proposition B.1 which is used to prove Theorem 3.3.

Proposition B.1. *For any $\epsilon \in (0,\infty)$, $\boldsymbol{a} \in \mathcal{R}^n$ and $k = 1,\dots,K-1$, we have*

$$Pr(|\boldsymbol{a}^T\mathbb{1}(\boldsymbol{y}=k)-\mathbb{1}(\boldsymbol{y}=K)-\boldsymbol{a}^T\mu_k(\boldsymbol{\theta}^*)| > \epsilon) \leq 2\exp\left[-\frac{2\epsilon^2}{||\boldsymbol{a}||_2^2}\right]$$

Proof. since $E(\mathbb{1}(\boldsymbol{y} = k) - \mathbb{1}(\boldsymbol{y} = K)) = \mu_k(\boldsymbol{\theta}^*)$, and y_is' are independent, then by Hoeffding inequality, it follows. □

Then, we follow the technique in the proof of Theorem 2 in Fan and Lv (2011).

Proof. First, let

$$\begin{aligned}
&\boldsymbol{\xi}_{Ik} = \boldsymbol{X}_I^T\{\mathbb{1}(\boldsymbol{y} = k) - \mathbb{1}(\boldsymbol{y} = K)\} - \boldsymbol{X}_I^T\boldsymbol{\mu}_k(\boldsymbol{\theta}^*)\\
&\boldsymbol{\xi}_{IIk} = \boldsymbol{X}_{II}^T\{\mathbb{1}(\boldsymbol{y} = k) - \mathbb{1}(\boldsymbol{y} = K)\} - \boldsymbol{X}_{II}^T\boldsymbol{\mu}_k(\boldsymbol{\theta}^*)\\
&\boldsymbol{\xi}_I = [\boldsymbol{\xi}_{I1}^T, \ldots, \boldsymbol{\xi}_{IK-1}^T]^T,\ \boldsymbol{\xi}_{II} = [\boldsymbol{\xi}_{II1}^T, \ldots, \boldsymbol{\xi}_{IIK-1}^T]^T.
\end{aligned}$$

then define:

$$E_1 = \{\|\boldsymbol{\xi}_I\|_\infty \le 2^{-1/2}(n\log n)^{1/2}\},\ E_2 = \{\|\boldsymbol{\xi}_{II}\|_\infty \le 2^{-1/2}(n\log n)^{1/2}\}.$$

then by general Bonferroni's inequality and Proposition B.1:

$$\begin{aligned}
\Pr(E_1 \cap E_2) &\ge 1 - \sum_{k=1}^{K-1}\sum_{g=1}^{G}\sum_{j=1}^{s_g}\Pr(|\xi_{kgj}| > 2^{-1/2}(n\log n)^{1/2})\\
&\ge 1 - 2(K-1)(\sum_{g=1}^{G} s_g)n^{-1}.
\end{aligned}$$

where $\xi_{kgj} = \boldsymbol{x}_{gj}^T\{\mathbb{1}(\boldsymbol{y} = k) - \mathbb{1}(\boldsymbol{y} = K) - \boldsymbol{\mu}_k(\boldsymbol{\theta}^*)\}$.

Under $E_1 \cap E_2$, we will show that there exists $\hat{\boldsymbol{\beta}}$ satisfying the (3.7) and (3.8). We break the prove into two steps.

Step 1: Existence of a solution to (3.7):

Let $\boldsymbol{\delta}_k = [\delta_{k11}, \dots, \delta_{k1s_1}, \dots, \delta_{kM1}, \dots, \delta_{kMs_M}]^T$, $\boldsymbol{\delta} = [\boldsymbol{\delta}_1^T, \dots, \boldsymbol{\delta}_{K-1}^T]^T$, and $\boldsymbol{\beta}_{Ik} = [\beta_{k11}, \dots, \beta_{k1s_1}, \dots, \beta_{kM1}, \dots, \beta_{kMs_M}]^T$, $\boldsymbol{\beta}_I = [\boldsymbol{\beta}_{I1}^T, \dots, \boldsymbol{\beta}_{I(K-1)}^T]^T$. we prove for sufficiently large n, (3.7) have solutions inside the hypercube:

$$\mathcal{N} = \{\boldsymbol{\delta} : \|\boldsymbol{\delta} - \boldsymbol{\beta}_I^*\|_\infty = n^{-1/2} \log n\}.$$

For $\boldsymbol{\delta} \in \mathcal{N}$, since $d \geq n^{-1/2} \log n$ for sufficient large n, we have

$$\begin{aligned} \min |\delta_{kgj}| &\geq \min |\beta_{kgj}^*| - d = d \text{ for } \beta_{kgj}^* \neq 0, \\ \operatorname{sgn}(\delta_{kgj}) &= \operatorname{sgn}(\beta_{kgj}^*) \text{ for } \beta_{kgj}^* \neq 0. \end{aligned}$$

where $g \in \{1, \dots, M\}, k \in \{1, \dots, K-1\}, j \in \{1, \dots, s_g\}$.

Let $\eta_{kgj} = n\lambda_n {s_g}^{1/2}(\delta_{kgj} + \sum_{k=1}^{K-1} \delta_{kgj})[\sum_{j=1}^{s_g}\{\sum_{k=1}^{K-1} \delta_{kgj}^2 + (\sum_{k=1}^{K-1} \delta_{kgj})^2\}]^{-1/2}$, $\boldsymbol{\eta}_k = [\eta_{k11}, \dots, \eta_{k1s_1} \dots, \eta_{kM1}, \dots, \eta_{kMs_M}]^T$, and $\boldsymbol{\eta} = [\boldsymbol{\eta}_1^T, \dots, \boldsymbol{\eta}_{K-1}^T]^T$. Then $\|\boldsymbol{\eta}\|_\infty \leq 2n\lambda_n\sqrt{s_*}$, $\sqrt{s_*} = max_{g=1}^M \sqrt{s_g}$, which along with the definition of E_1 entails: $\|\boldsymbol{\xi}_I - \boldsymbol{\eta}\|_\infty \leq 2^{-1/2}\sqrt{n \log n} + 2n\lambda_n\sqrt{s_*}$.

Define vector-valued functions:

$$\boldsymbol{\gamma}_I(\boldsymbol{\delta}) = \begin{bmatrix} \boldsymbol{\gamma}_{I1}(\boldsymbol{\delta}) \\ \vdots \\ \boldsymbol{\gamma}_{I(K-1)}(\boldsymbol{\delta}) \end{bmatrix},$$

where $\boldsymbol{\gamma}_{Ik}(\boldsymbol{\delta}) = \boldsymbol{X}_I^T(\{\exp(\theta_{ik}) - \exp(-\sum_{l=1}^{K-1} \theta_{il})\}\{\exp(\sum_{l=1}^{K-1} \theta_{il}) + \sum_{k=1}^{K-1} \exp(\theta_{ik})\}^{-1}, i = 1, \dots, n)^T$, where $\theta_{ik} = \sum_{g=1}^M \sum_{j=1}^{s_g} x_{gji}\delta_{kgj}$ for $k = 1, \dots, K-1$.

Then define:

$$\boldsymbol{\Psi}(\boldsymbol{\delta}) \;=\; \boldsymbol{\gamma}_I(\boldsymbol{\delta}) - \boldsymbol{\gamma}_I(\boldsymbol{\beta}_I^*) - (\boldsymbol{\xi}_I - \boldsymbol{\eta}).$$

Note that (3.7) is equivalent to $\boldsymbol{\Psi}(\boldsymbol{\delta}) = \mathbf{0}$. We need to show the latter has a solution inside $\mathcal{N}$. We represent $\boldsymbol{\gamma}_I(\boldsymbol{\delta})$ by using a second order Taylor expansion for around $\boldsymbol{\beta}_I^*$:

$$\begin{aligned}
\boldsymbol{\gamma}_I(\boldsymbol{\delta}) &= \boldsymbol{\gamma}_I(\boldsymbol{\beta}_I^*) + \boldsymbol{A}(\boldsymbol{\delta} - \boldsymbol{\beta}_I^*) + \boldsymbol{r}, \\
\boldsymbol{r} &= [\boldsymbol{r}_1^T, \ldots, \boldsymbol{r}_{K-1}^T]^T \\
\boldsymbol{r}_k &= [r_{k11}, \ldots, r_{k1s_1}, \ldots, r_{kM1}, \ldots, r_{kMs_M}]^T \\
r_{kgj} &= \frac{1}{2}(\boldsymbol{\delta} - \boldsymbol{\beta}_I^*)^T \nabla^2 \gamma_{kgj}(\breve{\boldsymbol{\delta}})(\boldsymbol{\delta} - \boldsymbol{\beta}_I^*)
\end{aligned}$$

where $\breve{\boldsymbol{\delta}}_k$ is a vector lying on the line segment joining $\boldsymbol{\delta}_k$ and $\boldsymbol{\beta}_{Ik}^*$, and

$$\boldsymbol{A} = \begin{bmatrix} \boldsymbol{X}_I^T \boldsymbol{\Sigma}_{11}(\boldsymbol{\theta}^*) \boldsymbol{X}_I & \ldots & \boldsymbol{X}_I^T \boldsymbol{\Sigma}_{1(K-1)}(\boldsymbol{\theta}^*) \boldsymbol{X}_I \\ \vdots & \ddots & \vdots \\ \boldsymbol{X}_I^T \boldsymbol{\Sigma}_{(K-1)1}(\boldsymbol{\theta}^*) \boldsymbol{X}_I & \ldots & \boldsymbol{X}_I^T \boldsymbol{\Sigma}_{(K-1)(K-1)}(\boldsymbol{\theta}^*) \boldsymbol{X}_I \end{bmatrix},$$

$$\begin{aligned}
&\boldsymbol{\Sigma}_{kk}(\boldsymbol{\theta}) = \partial \boldsymbol{\mu}_k(\boldsymbol{\theta})/\partial \boldsymbol{\theta}_k \\
&= \operatorname{diag}\Bigg\{ \{\exp(\theta_{ik}) + \exp(-\sum_{k=1}^{K-1} \theta_{ik})\}\{\sum_{k=1}^{K-1} \exp(\theta_{ik}) + \exp(-\sum_{k=1}^{K-1} \theta_{ik})\} \\
&-\{\exp(\theta_{ih}) - \exp(-\sum_{k=1}^{K-1} \theta_{ik})\}^2\{\sum_{k=1}^{K-1} \exp(\theta_{ik}) + \exp(-\sum_{k=1}^{K-1} \theta_{ik})\}^{-2}, i = 1, \ldots, n\Bigg\}, \\
&\boldsymbol{\Sigma}_{kh}(\boldsymbol{\theta}) = \partial \boldsymbol{\mu}_k(\boldsymbol{\theta})/\partial \boldsymbol{\theta}_h = \\
&\operatorname{diag}\Bigg\{ - \{\exp(\theta_{ik}) - \exp(-\sum_{k=1}^{K-1} \theta_{ik})\}\{\exp(\theta_{ih}) - \exp(-\sum_{k=1}^{K-1} \theta_{ik})\} \\
&\{\sum_{k=1}^{K-1} \exp(\theta_{ik}) + \exp(-\sum_{k=1}^{K-1} \theta_{ik})\}^{-2}, i = 1, \ldots, n\Bigg\},\ \text{for }, k \neq h \in \{1, \ldots, K-1\}.
\end{aligned}$$

and $\boldsymbol{X}_I = [\boldsymbol{x}_{11} \ldots, \boldsymbol{x}_{1s_1}, \ldots, \boldsymbol{x}_{M1} \ldots, \boldsymbol{x}_{Ms_M}]$, $\boldsymbol{X}_{II} = [\boldsymbol{x}_{(M+1)1} \ldots, \boldsymbol{x}_{(M+1)s_{M+1}}, \ldots, \boldsymbol{x}_{G1} \ldots, \boldsymbol{x}_{Gs_G}]$, $\boldsymbol{\theta}_k = [\theta_{1k}, \ldots, \theta_{nk}]^T$. And here, the derivative of a vector function with respect to a vector is known as the Jacobian matrix. By condition (C3),

$$\begin{aligned}
\|\boldsymbol{r}\|_\infty &\leq \max_{\boldsymbol{\delta}_0 \in \mathcal{N}} \max_{l,g=1,\ldots,M,j} \frac{1}{2} \lambda_{\max}[\boldsymbol{C}_{lgj}(\boldsymbol{\delta}_0)] \|(\boldsymbol{\delta} - \boldsymbol{\beta}_I^*)\|_2^2 \\
&= O[(K-1)\sum_{g=1}^{M} s_g (\log n)^2] \qquad \text{(B.7)}
\end{aligned}$$

where

$$\boldsymbol{C}_{lgj}(\boldsymbol{\delta}_0) = \big[\boldsymbol{c}_{kh} = \boldsymbol{X}_I^T \operatorname{diag}\{|\partial\{\partial \boldsymbol{\mu}_l(\boldsymbol{\theta})/\partial \boldsymbol{\theta}_k \mathbf{1}\}/\partial \boldsymbol{\theta}_h|_{\boldsymbol{\theta}=\boldsymbol{\Theta}(\boldsymbol{\delta})} \boldsymbol{x}_{gj}|\} \boldsymbol{X}_I\big]_{k,h=1}^{K-1},$$

and $\boldsymbol{\Theta}(\boldsymbol{\delta}) = (\sum_{g=1}^{M} \sum_{j}^{s_g} x_{gji} \delta_{kgj} : i = 1, \ldots, n, k = 1, \ldots, K-1)^T$, $\boldsymbol{\theta} = \{\theta_{ik}, i = 1, \ldots, n, k = 1, \ldots, K-1\}, \boldsymbol{\theta}_k = [\theta_{1k}, \ldots, \theta_{nk}]^T$, and the L_∞ norm of a matrix is the maximum of the L_1 norm of each row. Let

$$\begin{aligned}\bar{\boldsymbol{\Psi}}(\boldsymbol{\delta}) &= \boldsymbol{A}^{-1}\boldsymbol{\Psi}(\boldsymbol{\delta}) = \boldsymbol{\delta} - \boldsymbol{\beta}_I^* - \boldsymbol{u} \\ \boldsymbol{u} &= -\boldsymbol{A}^{-1}(\boldsymbol{\xi}_I - \boldsymbol{\eta} - \boldsymbol{r})\end{aligned}$$

then by condition (C1), we have:

$$\begin{aligned}\|\boldsymbol{u}\|_\infty &\leq \|\boldsymbol{A}^{-1}\|_\infty(\|\boldsymbol{\xi}_I - \boldsymbol{\eta}\|_\infty + \|\boldsymbol{r}\|_\infty) \\ &= O[n^{-1}(2^{-1/2}\sqrt{n\log n} + 2\sqrt{s_*}n\lambda_n + (K-1)\sum_{g=1}^{M} s_g(\log n)^2)] \\ &= O(2^{-1/2}n^{-1/2}\sqrt{\log n} + 2\sqrt{s_*}\lambda_n + n^{-1}(K-1)\sum_{g=1}^{M} s_g(\log n)^2) = o(n^{-1/2}\log n)\end{aligned}$$

For sufficiently large n, if the jth component of $\boldsymbol{\delta} - \boldsymbol{\beta}_I^*$, $(\boldsymbol{\delta} - \boldsymbol{\beta}_I^*)_{(j)} = n^{-1/2}\log n$, then we have

$$\bar{\boldsymbol{\Psi}}_{(j)}(\boldsymbol{\delta}) \geq n^{-1/2}\log n - \|\boldsymbol{u}\|_\infty \geq 0$$

and if $(\boldsymbol{\delta} - \boldsymbol{\beta}_I^*)_{(j)} = -n^{-1/2}\log n$

$$\bar{\boldsymbol{\Psi}}_{(j)}(\boldsymbol{\delta}) \leq -n^{-1/2}\log n + \|\boldsymbol{u}\|_\infty \leq 0$$

where $\bar{\boldsymbol{\Psi}}_{(j)}(\boldsymbol{\delta})$ is the jth component of $\bar{\boldsymbol{\Psi}}(\boldsymbol{\delta})$. Then by Miranda's existence theorem, $\bar{\boldsymbol{\Psi}}(\boldsymbol{\delta}) = 0$ has a solution in $\mathcal{N}$, so does $\boldsymbol{\Psi}(\boldsymbol{\delta}) = 0$.

Step 2: (Verification of (3.8)): Let $\hat{\boldsymbol{\beta}}$ be the estimates with $\hat{\boldsymbol{\beta}}_I$ is the solution from the first step, and $\hat{\boldsymbol{\beta}}_{II} = \mathbf{0}$. And we can find a necessary condition for

$$\sum_{j=1}^{s_g} \left[\sum_{k=1}^{K-1} \left\{ \nabla \ell_n^{kgj}(\hat{\boldsymbol{\beta}}) \right\}^2 + \left\{ \sum_{k=1}^{K-1} \nabla \ell_n^{kgj}(\hat{\boldsymbol{\beta}}) \right\}^2 \right] \leq n^2 s_g \lambda_n^2$$

is

$$\left\| \begin{bmatrix} \boldsymbol{X}_{II}^T\{\mathbb{1}(\boldsymbol{y}=1) - \mathbb{1}(\boldsymbol{y}=K) - \boldsymbol{\mu}_1(\hat{\boldsymbol{\beta}})\} \\ \vdots \\ \boldsymbol{X}_{II}^T\{\mathbb{1}(\boldsymbol{y}=K-1) - \mathbb{1}(\boldsymbol{y}=K) - \boldsymbol{\mu}_1(\hat{\boldsymbol{\beta}})\} \end{bmatrix} \right\|_\infty \leq n\lambda_n / \sqrt{K(K-1)}$$

Then let:

$$\boldsymbol{\gamma}_{II}(\boldsymbol{\delta}) = \begin{bmatrix} \boldsymbol{\gamma}_{II1}(\boldsymbol{\delta}) \\ \vdots \\ \boldsymbol{\gamma}_{II(K-1)}(\boldsymbol{\delta}) \end{bmatrix}, \boldsymbol{\gamma}_{IIk}(\boldsymbol{\delta}) \quad = \quad \boldsymbol{X}_{II}^T \boldsymbol{\mu}_k(\boldsymbol{\delta}).$$

Then define:

$$\boldsymbol{z} = (n\lambda_n)^{-1}\{\boldsymbol{\xi}_{II} - [\boldsymbol{\gamma}_{II}(\hat{\boldsymbol{\beta}}_I) - \boldsymbol{\gamma}_{II}(\boldsymbol{\beta}_I^*)]\}$$

By the rate of λ_n, $\|(n\lambda_n)^{-1}\boldsymbol{\xi}_{II}\|_\infty = O(2^{-1/2}n^{-1/2}\lambda_n^{-1}\sqrt{\log n}) = o(1)$.

We represent $\boldsymbol{\gamma}_{II}(\hat{\boldsymbol{\beta}}_I)$ by using a second order Taylor expansion around $\boldsymbol{\beta}_I^*$:

$$\begin{aligned}
\boldsymbol{\gamma}_{II}(\hat{\boldsymbol{\beta}}_I) &= \boldsymbol{\gamma}_{II}(\boldsymbol{\beta}_I^*) + \boldsymbol{B}(\hat{\boldsymbol{\beta}}_I - \boldsymbol{\beta}_I^*) + \boldsymbol{w} \\
\boldsymbol{w} &= [\boldsymbol{w}_1^T, \ldots, \boldsymbol{w}_K^T]^T \\
\boldsymbol{w}_k &= [w_{k(M+1)1}, \ldots, w_{k(M+1)s_(M+1)}, \ldots, w_{kG1}, \ldots, w_{kGs_G}]^T \\
w_{kgj} &= \frac{1}{2}(\hat{\boldsymbol{\beta}}_I - \boldsymbol{\beta}_I^*)^T \nabla^2 \gamma_{kgj}(\breve{\boldsymbol{\delta}})(\hat{\boldsymbol{\beta}}_I - \boldsymbol{\beta}_I^*)
\end{aligned}$$

where

$$\boldsymbol{B} = \begin{bmatrix} \boldsymbol{b}_{11} & \ldots & \boldsymbol{b}_{1(K-1)} \\ \vdots & \ddots & \vdots \\ \boldsymbol{b}_{(K-1)1} & \ldots & \boldsymbol{b}_{(K-1)(K-1)} \end{bmatrix}, \boldsymbol{b}_{kh} = \boldsymbol{X}_{II}^T \boldsymbol{\Sigma}_{kh}(\boldsymbol{\theta}^*)\boldsymbol{X}_I.$$

$\boldsymbol{w}$ can be bounded by similar argument in(B.7), and from the first step of the proof, we already have $\hat{\boldsymbol{\beta}}_I - \boldsymbol{\beta}_I^* = \boldsymbol{A}^{-1}(\boldsymbol{\xi}_I - \boldsymbol{\eta} - \boldsymbol{r})$, by condition (C2) we have:

$$\begin{aligned}
&\|\boldsymbol{z}\|_\infty \le o(1) + (n\lambda_n)^{-1}\|\boldsymbol{\gamma}_{II}(\hat{\boldsymbol{\beta}}_I) - \boldsymbol{\gamma}_{II}(\boldsymbol{\beta}_I^*)\|_\infty \\
\le\ & o(1) + (n\lambda_n)^{-1}\|\boldsymbol{B}\boldsymbol{A}^{-1}\|_\infty(\|\boldsymbol{\xi}_I - \boldsymbol{\eta}\|_\infty + \|\boldsymbol{r}\|_\infty) + (n\lambda_n)^{-1}\|\boldsymbol{w}\|_\infty \\
\le\ & o(1) + 1/\sqrt{K(K-1)} + (n\lambda_n)^{-1}O(\sqrt{n\log n} + s(\log n)^2) \\
=\ & o(1) + 1/\sqrt{K(K-1)} + O(\lambda_n^{-1}n^{-1/2}\sqrt{\log n} + \lambda_n^{-1}n^{-1}(\log n)^2) \\
=\ & 1/\sqrt{K(K-1)} + o(1)
\end{aligned}$$

Therefore, by Proposition 3.2, we have shown that there is a $\hat{\boldsymbol{\beta}}$ with (3.9), (3.10), (3.11) under event $E_1 \cap E_2$. This completes the proof. □

B.5 Lemma 3.4 Proof

Proof.

$$\begin{aligned}\ell_n^g(\boldsymbol{\beta}_g \mid \hat{\boldsymbol{\beta}}_{-g}) = \quad & \sum_{k=1}^{K}\left\{\mathbb{1}(\boldsymbol{y}=k\mathbf{1})^T(\hat{\beta}_{k0}\mathbf{1}+\sum_{j=1}^{s_g}\boldsymbol{x}_{gj}\beta_{kgj}+\sum_{h\neq g}\sum_{j=1}^{s_h}\boldsymbol{x}_{hj}\hat{\beta}_{khj})\right\}\\ & -\sum_{i=1}^{N}\log\{\sum_{k=1}^{K}\exp(\hat{\beta}_{k0}+\sum_{j=1}^{s_g}x_{gji}\beta_{kgj}+\sum_{h\neq g}\sum_{j=1}^{s_h}x_{hji}\hat{\beta}_{khj})\}.\end{aligned}$$

By the standard argument on the Taylor expansion of the likelihood function $\ell_n^g(\boldsymbol{\beta}_g \mid \hat{\boldsymbol{\beta}})$ at $\mathbf{0}$, we have

$$\ell_n^g(\boldsymbol{\beta}_g \mid \hat{\boldsymbol{\beta}}_{-g}) = \ell_n^g(\mathbf{0} \mid \hat{\boldsymbol{\beta}}_{-g}) + \nabla\ell_n^g(\mathbf{0} \mid \hat{\boldsymbol{\beta}}_{-g})\boldsymbol{\beta}_g + 2^{-1}\boldsymbol{\beta}_g\nabla^2\ell_n^g(\mathbf{0} \mid \hat{\boldsymbol{\beta}}_{-g})\boldsymbol{\beta}_g(1+o(1))$$

If $\|\nabla\ell_n^g(\boldsymbol{\beta}_g \mid \hat{\boldsymbol{\beta}}_{-g})|_{\boldsymbol{\beta}_g=\mathbf{0}}\|_2 \leq n{s_g}^{1/2}\lambda_n$, then $\nabla\ell_n^g(\mathbf{0} \mid \hat{\boldsymbol{\beta}}_{-g})\boldsymbol{\beta}_g - n\lambda_n s_g^{1/2}\|\boldsymbol{\beta}_g\|_2 \leq 0$. Since the log-likelihood function is concave, then $\boldsymbol{\beta}_g\nabla^2\ell_n^g(\mathbf{0} \mid \hat{\boldsymbol{\beta}}_{-g})\boldsymbol{\beta}_g \leq 0$, and clearly, $\boldsymbol{\beta}_g = \mathbf{0}$ is the maximizer. □

B.6 Proposition 3.5 Proof

Proof.

$$\begin{aligned}\ell_n(\boldsymbol{\beta};\boldsymbol{y}) = \sum_{k=1}^{K-1}\left[\{\mathbb{1}(\boldsymbol{y}=k\mathbf{1})-\mathbb{1}(\boldsymbol{y}=K\mathbf{1})\}^T(\sum_{g=1}^{G}\sum_{j=1}^{s_g}\boldsymbol{x}_{gj}\beta_{kgj})\right] - \\ \sum_{i=1}^{n}\log\left\{\sum_{k=1}^{K-1}\exp(\theta_{ik})+\exp(-\sum_{k=1}^{K-1}\theta_{ik})\right\}\end{aligned}$$

We rewrite the maximization of (3.4) in the form

$$\min_{\boldsymbol{\beta},\boldsymbol{v}\in\mathcal{R}^G} -n^{-1}\ell_n(\boldsymbol{\beta}) + \lambda_n \sum_{g=1}^{G} s_g^{1/2} v_g,$$
$$\text{s.t. } \|\boldsymbol{\beta}_g\|_2 \le v_g,\ g = 1,\dots,G.$$

Then we consider the Lagrangian with dual variable $\{\boldsymbol{\zeta}_g, \gamma_g\} \in R^{s_g(K-1)} \times R$ (Boyd and Vandenberghe, 2003):

$$L(\boldsymbol{\beta},\boldsymbol{v},\boldsymbol{\zeta},\boldsymbol{\gamma}) = -n^{-1}\ell_n(\boldsymbol{\beta}) + \lambda_n \sum_{g=1}^{G} s_g^{1/2} v_g - \sum_{g=1}^{G}(\boldsymbol{\beta}_g^T\boldsymbol{\zeta}_g + v_g\gamma_g),$$
$$\text{s.t. } \|\boldsymbol{\zeta}_g\|_2 \le \gamma_g,\ g = 1,\dots,G.$$

The KKT conditions are: for $g = 1,\dots,G$.

$$-n^{-1}\boldsymbol{x}_{gj}^T\{\mathbb{1}(\boldsymbol{y} = k\mathbf{1}) - \boldsymbol{\mu}_k(\boldsymbol{\theta})\} = \zeta_{gj}, k = 1,\dots,K-1, j = 1,\dots,s_g, \tag{B.8}$$

$$\lambda_n s_g^{1/2} = \gamma_g, \tag{B.9}$$

$$\|\boldsymbol{\beta}_g\|_2 \le v_g, \tag{B.10}$$

$$\|\boldsymbol{\zeta}_g\|_2 \le \gamma_g, \tag{B.11}$$

$$\boldsymbol{\beta}_g^T\boldsymbol{\zeta}_g + v_g\gamma_g = 0. \tag{B.12}$$

where (B.8) and (B.9) are for stationarity, (B.10) is for primary feasibility, (B.11) is for dual feasibility and (B.12) is for complimentary slackness. If $\boldsymbol{\beta}_g = \mathbf{0}$, then we just need $\|\boldsymbol{\zeta}_g\|_2 \le \lambda_n s_g^{1/2}$. If $\boldsymbol{\beta}_g \ne \mathbf{0}$, we can have $\zeta_{gj} = -\beta_{gj}\lambda_n s_g^{1/2}(\|\boldsymbol{\beta}_g\|_2)^{-1}$. These KKT conditions reduce to (3.13) and (3.14) by plug in (B.8). Since the dual and primal problems are strictly feasible, and the $\ell_n(\boldsymbol{\beta})$ is concave. Then KKT condition is a sufficient and necessary condition for primary and dual optimal (Boyd and Vandenberghe, 2003). □

B.7 Theorem 3.6 Proof

We follow the technique in the proof of Theorem 3.3.

Proof. First, let

$$\begin{aligned}
&\boldsymbol{\xi}_{Ik} = \boldsymbol{X}_I^T \mathbb{1}(\boldsymbol{y} = k\mathbf{1}) - \boldsymbol{X}_I^T \boldsymbol{\mu}_k(\boldsymbol{\theta}^*)\\
&\boldsymbol{\xi}_{IIk} = \boldsymbol{X}_{II}^T \mathbb{1}(\boldsymbol{y} = k\mathbf{1}) - \boldsymbol{X}_{II}^T \boldsymbol{\mu}_k(\boldsymbol{\theta}^*)\\
&\boldsymbol{\xi}_I = [\boldsymbol{\xi}_{I1}^T, \ldots, \boldsymbol{\xi}_{I(K-1)}^T]^T,\ \boldsymbol{\xi}_{II} = [\boldsymbol{\xi}_{II1}^T, \ldots, \boldsymbol{\xi}_{II(K-1)}^T]^T.
\end{aligned}$$

then define:

$$E_1 = \{\|\boldsymbol{\xi}_I\|_\infty \leq 2^{-1/2}(n\log n)^{1/2}\},\ E_2 = \{\|\boldsymbol{\xi}_{II}\|_\infty \leq 2^{-1/2}(n\log n)^{1/2}\}.$$

then by general Bonferroni's inequality and Proposition B.1:

$$\begin{aligned}
\Pr(E_1 \cap E_2) &\geq 1 - \sum_{k=1}^{K-1}\sum_{g=1}^{G}\sum_{j=1}^{s_g} \Pr(|\xi_{kgj}| > 2^{-1/2}(n\log n)^{1/2}\\
&\geq = 1 - 2(K-1)(\sum_{g=1}^{G} s_g)n^{-1}.
\end{aligned}$$

where $\xi_{kgj} = \boldsymbol{x}_{gj}^T\{\mathbb{1}(\boldsymbol{y} = k\mathbf{1}) - \boldsymbol{\mu}_k(\boldsymbol{\theta}^*)\}$.

Under $E_1 \cap E_2$, we will show that there exists $\hat{\boldsymbol{\beta}}$ satisfying the (3.13) and (3.14). We break the prove into two steps.

Step 1: Existence of a solution to (3.13):

Let $\boldsymbol{\delta}_k = [\delta_{k11}, \dots, \delta_{k1s_1}, \dots, \delta_{kM1}, \dots, \delta_{kMs_M}]^T$, $\boldsymbol{\delta} = [\boldsymbol{\delta}_1^T, \dots, \boldsymbol{\delta}_{K-1}^T]^T$, and $\boldsymbol{\beta}_{Ik} = [\beta_{k11}, \dots, \beta_{k1s_1}, \dots, \beta_{kM1}, \dots, \beta_{kMs_M}]^T$, $\boldsymbol{\beta}_I = [\boldsymbol{\beta}_{I1}^T, \dots, \boldsymbol{\beta}_{IK-1}^T]^T$. we prove for sufficiently large n, (3.13) have solutions inside the hypercube:

$$\mathcal{N} = \{\boldsymbol{\delta} : \|\boldsymbol{\delta} - \boldsymbol{\beta}_I^*\|_\infty = n^{-1/2} \log n\}.$$

For $\boldsymbol{\delta} \in \mathcal{N}$, since $d \geq n^{-1/2} \log n$ for sufficient large n, we have

$$\begin{aligned} \min|\delta_{kgj}| &\geq \min|\beta_{kgj}^*| - d = d \text{ for } \beta_{kgj}^* \neq 0, \\ sgn(\delta_{kgj}) &= sgn(\beta_{kgj}^*) \text{ for } \beta_{kgj}^* \neq 0. \end{aligned}$$

where $g \in \{1, \dots, M\}, k \in \{1, \dots, K-1\}, j \in \{1, \dots, s_g\}$.

Let $\eta_{kgj} = n\lambda_n {s_g}^{1/2} \delta_{kgj} (\sum_{k=1}^{K-1} \sum_{j=1}^{s_g} \delta_{kgj}^2)^{-1/2}$, $\boldsymbol{\eta}_k = [\eta_{k11}, \dots, \eta_{k1s_1} \dots, \eta_{kM1}, \dots, \eta_{kMs_M}]^T$, and $\boldsymbol{\eta} = [\boldsymbol{\eta}_1^T, \dots, \boldsymbol{\eta}_{K-1}^T]^T$. Then $\|\boldsymbol{\eta}\|_\infty \leq n\lambda_n \sqrt{s_*}$, $\sqrt{s_*} = \max_{g=1}^S \sqrt{s_g}$, which along with the definition of E_1 entails: $\|\boldsymbol{\xi}_I - \boldsymbol{\eta}\|_\infty \leq 2^{-1/2} \sqrt{n \log n} + n\lambda_n \sqrt{s_*}$.

Define vector-valued functions:

$$\boldsymbol{\gamma}_I(\boldsymbol{\delta}) = \begin{bmatrix} \boldsymbol{\gamma}_{I1}(\boldsymbol{\delta}) \\ \vdots \\ \boldsymbol{\gamma}_{I(K-1)}(\boldsymbol{\delta}) \end{bmatrix},$$

where $\boldsymbol{\gamma}_{Ik}(\boldsymbol{\delta}) = \boldsymbol{X}_I^T(\{\exp(\theta_{ik}) - \exp(-\sum_{l=1}^{K-1} \theta_{il})\}\{\exp(\sum_{l=1}^{K-1} \theta_{il}) + \sum_{k=1}^{K-1} \exp(\theta_{ik})\}^{-1}, i =$

$1, \ldots, n)^T$, where $\theta_{ik} = \sum_{g=1}^{M} \sum_{j=1}^{s_g} x_{gji} \delta_{kgj}$ for $k = 1, \ldots, K-1$.

Then define:

$$\boldsymbol{\Psi}(\boldsymbol{\delta}) = \boldsymbol{\gamma}_I(\boldsymbol{\delta}) - \boldsymbol{\gamma}_I(\boldsymbol{\beta}_I^*) - (\boldsymbol{\xi}_I - \boldsymbol{\eta}).$$

Note that (3.13) is equivalent to $\boldsymbol{\Psi}(\boldsymbol{\delta}) = \mathbf{0}$. We need to show the latter has a solution inside $\mathcal{N}$. We represent $\boldsymbol{\gamma}_I(\boldsymbol{\delta})$ by using a second order Taylor expansion for around $\boldsymbol{\beta}_I^*$:

$$\begin{aligned}
\boldsymbol{\gamma}_I(\boldsymbol{\delta}) &= \boldsymbol{\gamma}_I(\boldsymbol{\beta}_I^*) + \boldsymbol{A}(\boldsymbol{\delta} - \boldsymbol{\beta}_I^*) + \boldsymbol{r}, \\
\boldsymbol{r} &= [\boldsymbol{r}_1^T, \ldots, \boldsymbol{r}_{K-1}^T]^T \\
\boldsymbol{r}_k &= [r_{k11}, \ldots, r_{k1s_1}, \ldots, r_{kM1}, \ldots, r_{kMs_M}]^T \\
r_{kgj} &= \frac{1}{2}(\boldsymbol{\delta} - \boldsymbol{\beta}_I^*)^T \nabla^2 \gamma_{kgj}(\breve{\boldsymbol{\delta}})(\boldsymbol{\delta} - \boldsymbol{\beta}_I^*)
\end{aligned}$$

where $\breve{\boldsymbol{\delta}}_k$ is a vector lying on the line segment joining $\boldsymbol{\delta}_k$ and $\boldsymbol{\beta}_{Ik}^*$, and

$$\boldsymbol{A} = \begin{bmatrix} \boldsymbol{X}_I^T \boldsymbol{\Sigma}_{11}(\boldsymbol{\theta}^*) \boldsymbol{X}_I & \ldots & \boldsymbol{X}_I^T \boldsymbol{\Sigma}_{1(K-1)}(\boldsymbol{\theta}^*) \boldsymbol{X}_I \\ \vdots & \ddots & \vdots \\ \boldsymbol{X}_I^T \boldsymbol{\Sigma}_{(K-1)1}(\boldsymbol{\theta}^*) \boldsymbol{X}_I & \ldots & \boldsymbol{X}_I^T \boldsymbol{\Sigma}_{(K-1)(K-1)}(\boldsymbol{\theta}^*) \boldsymbol{X}_I \end{bmatrix}$$

$$\boldsymbol{\Sigma}_{kk}(\boldsymbol{\theta}) = \partial \boldsymbol{\mu}_k(\boldsymbol{\theta}) / \partial \boldsymbol{\theta}_k, \text{ for } k = 1, \ldots, K-1,$$

$$\boldsymbol{\Sigma}_{kh}(\boldsymbol{\theta}) = \partial \boldsymbol{\mu}_k(\boldsymbol{\theta}) / \partial \boldsymbol{\theta}_h, \text{ for } k \neq h, k, h = 1, \ldots, K-1.$$

and $\boldsymbol{X}_I = [\boldsymbol{x}_{11} \ldots, \boldsymbol{x}_{1s_1}, \ldots, \boldsymbol{x}_{M1} \ldots, \boldsymbol{x}_{Ms_M}]$, $\boldsymbol{X}_{II} = [\boldsymbol{x}_{(M+1)1} \ldots, \boldsymbol{x}_{(M+1)s_{M+1}}, \ldots, \boldsymbol{x}_{G1} \ldots, \boldsymbol{x}_{Gs_G}]$, $\boldsymbol{\theta}_k = [\theta_{1k}, \ldots, \theta_{nk}]^T$. And here, the derivative of a vector function with respect to a vector is known as the Jacobian matrix. By condition (C3),

$$\begin{aligned} \|\boldsymbol{r}\|_\infty &\leq \max_{\boldsymbol{\delta}_0 \in \mathcal{N}} \max_{l,g=1,\ldots,M,j} \frac{1}{2} \lambda_{\max}[\boldsymbol{C}_{lgj}(\boldsymbol{\delta}_0)] \|(\boldsymbol{\delta} - \boldsymbol{\beta}_I^*)\|_2^2 \\ &= O[(K-1) \sum_{g=1}^{M} s_g (\log n)^2] \end{aligned} \quad \text{(B.13)}$$

where

$$\boldsymbol{C}_{lgj}(\boldsymbol{\delta}_0) = \left[\boldsymbol{c}_{kh} = \boldsymbol{X}_I^T \mathrm{diag}\{|\partial\{\partial \boldsymbol{\mu}_l(\boldsymbol{\theta})/\partial \boldsymbol{\theta}_k \mathbf{1}\}/\partial \boldsymbol{\theta}_h|_{\boldsymbol{\theta}=\boldsymbol{\Theta}(\boldsymbol{\delta})} \boldsymbol{x}_{gj}|\} \boldsymbol{X}_I\right]_{k,h=1}^{K-1},$$

and $\boldsymbol{\Theta}(\boldsymbol{\delta}) = (\sum_{g=1}^M \sum_j^{s_g} x_{gji}\delta_{kgj} : i = 1, \ldots, n, k = 1, \ldots, K-1)^T$, $\boldsymbol{\theta} = \{\theta_{ik}, i = 1, \ldots, n, k = 1, \ldots, K-1\}, \boldsymbol{\theta}_k = [\theta_{1k}, \ldots, \theta_{nk}]^T$, and the L_∞ norm of a matrix is the maximum of the L_1 norm of each row. Let

$$\begin{aligned} \bar{\boldsymbol{\Psi}}(\boldsymbol{\delta}) &= \boldsymbol{A}^{-1} \boldsymbol{\Psi}(\boldsymbol{\delta}) = \boldsymbol{\delta} - \boldsymbol{\beta}_I^* - \boldsymbol{u} \\ \boldsymbol{u} &= -\boldsymbol{A}^{-1}(\boldsymbol{\xi}_I - \boldsymbol{\eta} - \boldsymbol{r}) \end{aligned}$$

then by condition (C1), we have:

$$\begin{aligned}
\|\boldsymbol{u}\|_\infty &\leq \|\boldsymbol{A}^{-1}\|_\infty(\|\boldsymbol{\xi}_I - \boldsymbol{\eta}\|_\infty + \|\boldsymbol{r}\|_\infty) \\
&= O[n^{-1}(2^{-1/2}\sqrt{n\log n} + \sqrt{s_*}n\lambda_n + (K-1)\sum_{g=1}^{M} s_g(\log n)^2)] \\
&= O(2^{-1/2}n^{-1/2}\sqrt{\log n} + \sqrt{s_*}\lambda_n + n^{-1}(K-1)\sum_{g=1}^{M} s_g(\log n)^2) = o(n^{-1/2}\log n)
\end{aligned}$$

For sufficiently large n, if the jth component of $\boldsymbol{\delta} - \boldsymbol{\beta}_I^*$, $(\boldsymbol{\delta} - \boldsymbol{\beta}_I^*)_{(j)} = n^{-1/2}\log n$, then we have

$$\bar{\boldsymbol{\Psi}}_{(j)}(\boldsymbol{\delta}) \geq n^{-1/2}\log n - \|\boldsymbol{u}\|_\infty \geq 0$$

and if $(\boldsymbol{\delta} - \boldsymbol{\beta}_I^*)_{(j)} = -n^{-1/2}\log n$

$$\bar{\boldsymbol{\Psi}}_{(j)}(\boldsymbol{\delta}) \leq -n^{-1/2}\log n + \|\boldsymbol{u}\|_\infty \leq 0$$

where $\bar{\boldsymbol{\Psi}}_{(j)}(\boldsymbol{\delta})$ is the jth component of $\bar{\boldsymbol{\Psi}}(\boldsymbol{\delta})$. Then by Miranda's existence theorem, $\bar{\boldsymbol{\Psi}}(\boldsymbol{\delta}) = 0$ has a solution in $\mathcal{N}$, so does $\boldsymbol{\Psi}(\boldsymbol{\delta}) = 0$.

Step 2: (Verification of (3.8)): Let $\hat{\boldsymbol{\beta}}$ be the estimates with $\hat{\boldsymbol{\beta}}_I$ is the solution from the first step, and $\hat{\boldsymbol{\beta}}_{II} = \boldsymbol{0}$. And we can find a necessary condition for

$$\left\| \begin{bmatrix} \boldsymbol{X}_g^T \mathbb{1}(\boldsymbol{y} = 1\boldsymbol{1}) \\ \vdots \\ \boldsymbol{X}_g^T \mathbb{1}(\boldsymbol{y} = K-1\boldsymbol{1}) \end{bmatrix} - \begin{bmatrix} \boldsymbol{X}_g^T \boldsymbol{\mu}_1(\hat{\boldsymbol{\beta}}) \\ \vdots \\ \boldsymbol{X}_g^T \boldsymbol{\mu}_{K-1}(\hat{\boldsymbol{\beta}}) \end{bmatrix} \right\|_2 < n{s_g}^{1/2}\lambda_n$$

is

$$\left\| \begin{bmatrix} \boldsymbol{X}_{II}^T \mathbb{1}(\boldsymbol{y} = 1\mathbf{1}) \\ \vdots \\ \boldsymbol{X}_{II}^T \mathbb{1}(\boldsymbol{y} = (K-1)\mathbf{1}) \end{bmatrix} - \begin{bmatrix} \boldsymbol{X}_{II}^T \boldsymbol{\mu}_1(\hat{\boldsymbol{\beta}}) \\ \vdots \\ \boldsymbol{X}_{II}^T \boldsymbol{\mu}_{K-1}(\hat{\boldsymbol{\beta}}) \end{bmatrix} \right\|_\infty < n\lambda_n/\sqrt{K-1}$$

Then let:

$$\boldsymbol{\gamma}_{II}(\boldsymbol{\delta}) = \begin{bmatrix} \boldsymbol{\gamma}_{II1}(\boldsymbol{\delta}) \\ \vdots \\ \boldsymbol{\gamma}_{II(K-1)}(\boldsymbol{\delta}) \end{bmatrix}, \boldsymbol{\gamma}_{IIk}(\boldsymbol{\delta}) \quad = \quad \boldsymbol{X}_{II}^T \boldsymbol{\mu}_k(\boldsymbol{\delta}).$$

Then define:

$$\boldsymbol{z} = (n\lambda_n)^{-1}\{\boldsymbol{\xi}_{II} - [\boldsymbol{\gamma}_{II}(\hat{\boldsymbol{\beta}}_I) - \boldsymbol{\gamma}_{II}(\boldsymbol{\beta}_I^*)]\}$$

By the rate of λ_n, $\|(n\lambda_n)^{-1}\boldsymbol{\xi}_{II}\|_\infty = O(2^{-1/2}n^{-1/2}\lambda_n^{-1}\sqrt{\log n}) = o(1)$.

We represent $\boldsymbol{\gamma}_{II}(\hat{\boldsymbol{\beta}}_I)$ by using a second order Taylor expansion around $\boldsymbol{\beta}_I^*$:

$$\begin{aligned} \boldsymbol{\gamma}_{II}(\hat{\boldsymbol{\beta}}_I) &= \boldsymbol{\gamma}_{II}(\boldsymbol{\beta}_I^*) + \boldsymbol{B}(\hat{\boldsymbol{\beta}}_I - \boldsymbol{\beta}_I^*) + \boldsymbol{w} \\ \boldsymbol{w} &= [\boldsymbol{w}_1^T, \ldots, \boldsymbol{w}_{K-1}^T]^T \\ \boldsymbol{w}_k &= [w_{k(M+1)1}, \ldots, w_{k(M+1)s_(M+1)}, \ldots, w_{kG1}, \ldots, w_{kGs_G}]^T \\ w_{kgj} &= \frac{1}{2}(\hat{\boldsymbol{\beta}}_I - \boldsymbol{\beta}_I^*)^T \nabla^2 \gamma_{kgj}(\breve{\boldsymbol{\delta}})(\hat{\boldsymbol{\beta}}_I - \boldsymbol{\beta}_I^*) \end{aligned}$$

where

$$\boldsymbol{B} = \begin{bmatrix} \boldsymbol{b}_{11} & \dots & \boldsymbol{b}_{1(K-1)} \\ \vdots & \ddots & \vdots \\ \boldsymbol{b}_{(K-1)1} & \dots & \boldsymbol{b}_{(K-1)(K-1)} \end{bmatrix}, \boldsymbol{b}_{kh} = \boldsymbol{X}_{II}^T \boldsymbol{\Sigma}_{kh}(\boldsymbol{\theta}^*)\boldsymbol{X}_I.$$

$\boldsymbol{w}$ can be bounded by similar argument in(B.13), and from the first step of the proof, we already have $\hat{\boldsymbol{\beta}}_I - \boldsymbol{\beta}_I^* = \boldsymbol{A}^{-1}(\boldsymbol{\xi}_I - \boldsymbol{\eta} - \boldsymbol{r})$, by condition (C4) we have:

$$\begin{aligned} & \|\boldsymbol{z}\|_\infty \le o(1) + (n\lambda_n)^{-1}\|\boldsymbol{\gamma}_{II}(\hat{\boldsymbol{\beta}}_I) - \boldsymbol{\gamma}_{II}(\boldsymbol{\beta}_I^*)\|_\infty \\ \le \quad & o(1) + (n\lambda_n)^{-1}\|\boldsymbol{B}\boldsymbol{A}^{-1}\|_\infty(\|\boldsymbol{\xi}_I - \boldsymbol{\eta}\|_\infty + \|\boldsymbol{r}\|_\infty) + (n\lambda_n)^{-1}\|\boldsymbol{w}\|_\infty \\ < \quad & o(1) + 1/\sqrt{K-1} + (n\lambda_n)^{-1}O(\sqrt{n\log n} + s(\log n)^2) \\ = \quad & o(1) + 1/\sqrt{K-1} + O(\lambda_n^{-1}n^{-1/2}\sqrt{\log n} + \lambda_n^{-1}n^{-1}(\log n)^2) \\ = \quad & 1/\sqrt{K-1} + o(1) \end{aligned}$$

Therefore, by Proposition 3.5, we have shown that there is a $\hat{\boldsymbol{\beta}}$ with (3.15), (3.16), (3.17) under event $E_1 \cap E_2$. This completes the proof. □

References

Banerjee, S., B.P. Carlin, and A.E. Gelfand. 2004. *Hierarchical modeling and analysis for spatial data.* Chapman and Hall: Boca Raton.

Besag, J. 1972. Nearest-neighbour systems and the auto-logistic model for binary data. *Journal of the Royal Statistical Society Series B* 34:75–83.

Besag, J., J.esag. 1974. Spatial interaction and the statistical analysis of lattice systems (with discussion). *Journal of the Royal Statistical Society Series B* 36: 192–236.

Boyd, S., and L. Vandenberghe. 2003. *Convex optimization.* Cambridge Univ. Press.

Caragea, P.C., and M.S. Kaiser. 2009. Autologistic models with interpretable parameters. *Journal of Agricultural, Biological, and Environmental Statistics* 14: 281–300.

Comets, F., and M. Janžura. 1998. A central limit theorem for conditionally centred random fields with an application to markov fields. *Journal of Applied Probability* 35:608–621.

Cressie, N. 1993. *Statistics for spatial data, rev ed.* Wiley: New York.

Crow, T.R., G.E. Host, and D.J. Mladenoff. 1999. Ownership and ecosystem as sources of spatial heterogeneity in a forested landscape, wisconsin usa. *Landscape Ecology* 14:449–463.

Davidson, R., and MacKinnon G.J. 2003. *Econometric theory and methods.* Oxford University Press, New York.

Diggle, P.J., and P.J. Ribeiro. 2007. *Model-based geostatistics.* New York: Springer.

Efron, B., T. Hastie, I. Johnstone, and R. Tibshirani. 2004. Least angle regression (with discussion). *Annals of Statistics* 32:407–499.

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

Fan, J., and J. Lv. 2011. Nonconcave penalized likelihood with np-dimensionality. *IEEE Transactions on Information Theory* 57(8).

Friedman, J., T Hastie, and R. Tibshirani. 2010. Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software* 44(1).

Friel, N., A.N. Pettitt, R. Reeves, and E. Wit. 2009. Bayesian inference in hidden Markov random fields for binary data defined on large lattices. *Journal of Computational and Graphical Statistics* 18:243–261.

Fu, R., A.L. Thurman, T. Chu, M.M. Steen-Adams, and J. Zhu. 2013. On estimation and selection of autologistic regression models via penalized pseudolikelihood. *Journal of Agricultural, Biological, and Ecological Statistics* 3:429–449.

Gaetan, C., and X. Guyon. 2010. *Spatial statistics and modeling.* New York: Springer.

Geyer, C. J. 1994. On the convergence of Monte Carlo maximum likelihood calculations. *Journal of the Royal Society of Statistics Series B* 56:261–274.

Gumpertz, M.L., J.M. Graham, and J.B. Ristaino. 1977. Autologistic model of spatial pattern of *Phytophthora* epidemic in bell pepper: Effects of soil variables on disease presence. *Journal of Agricultural, Biological, and Environmental Statistics* 2: 131–156.

He, F., J. Zhou, and H. Zhu. 2003. Autologistic regression model for the distribution of vegetation. *Journal of Agricultural, Biological and Environmental Statistics* 8: 205–222.

Huang, H.-C., N.-J. Hsu, D.M. Theobald, and F.J. Breidt. 2010. Spatial LASSO with applications to GIS model selection. *Journal of Computational and Graphical Statistics* 19:963–983.

Huffer, F.W., and H. Wu. 1998. Markov chain Monte Carlo for autologistic regression models with application to the distribution of plant species. *Biometrics* 54:509–524.

Hughes, J., and M. Haran. 2013. Dimension reduction and alleviation of confounding for spatial generalized linear mixed models. *Journal of the Royal Statistical Society Series B* 75:139–159.

Hughes, J., M. Haran, and P.C. Caragea. 2011. Autologistic models for binary data on a lattice. *Environmetrics* 22:857–871.

Jin, C., J. Zhu, M.M. Steen-Adams, S.R. Sain, and R.E. Gangnon. 2013. Spatial multinomial regression models for nominal categorical data: a study of land cover in northern Wisconsin, USA. *Environmetrics* 24:98–108.

Meier, L., S Geer, and P. Buhlmann. 2008. The group lasso for logistic regression. *Journal of the Royal Statistical Society: Series B* 70:53–71.

Moller, J., A.N. Pettitt, R. Reeves, and K.K. Berthelsen. 2006. An efficient Markov chain Monte Carlo method for distributions with intractable normalising constants. *Biometrika* 93:451–458.

Nocedal, J., and S.J. Wright. 2000. *Numerical optimization*. 2nd ed. New York: Springer.

Paciorek, C.J. 2010. The importance of scale for spatial-confounding bias and precision of spatial regression estimators. *Statistical Science* 25:107–125.

Rue, H., S. Martino, and N. Chopin. 2009. Approximate bayesian inference for latent gaussian models by using integrated nested laplace. *Journal of Royal Statistical Society* 71:319–392.

Stanfield, B.J., J.C. Bliss, and T.A. Spies. 2002. Land ownership and landscape structure: A spatial analysis of sixty-six oregon (usa) coast range watersheds. *Landscape Ecology* 17:685–697.

Steen-Adams, M.M., D.J. Mladenoff, N.E. Langston, F. Liu, and J. Zhu. 2011. Influence of biophysical factors and differences in ojibwe reservation versus euro-american social histories on forest landscape change in northern Wisconsin, USA. *Landscape Ecology* 26:1165–1178.

Sun, L., and M.K. Clayton. 2008. Bayesian analysis of cross-classified spatial data with autocorrelation. *Biometrics* 64:74–84.

Team, R Development Core. 2011. *R: A language and environment for statistical computing*. Vienna, Austria ISBN 3-900051-07-0 http://www.R-project.org/.

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

Tibshirani, R., S Rosset, J. Zhu, and K. Knight. 2005. Sparsity and smoothness via the fused lasso. *Journal of the Royal Statistical Society: Series B* 67:91–108.

Turner, M.G., D.N. Wear, and R.O. Flamm. 1996. Land ownership and land-cover change in the Southern Appalachian Highlands and the Olympic Peninsula. *Ecological Applications* 6:1150–1172.

Wainwright, M. L. 2009. Sharp thresholds for high-dimensional and noisy recovery of sparsity using l_1-constrained quadratic programming. *IEEE Transactions on Information Theory* 55:2183–2202.

Wang, Z., and Y. Zheng. 2012. Analysis of binary data via a centered spatial-temporal autologistic regression model. *Environmental and Ecological Statistics* 20: 37–57.

Wasserman, L. 2003. *All of statistics: A concise course in statistical inference*. New York: Springer.

Xue, L., H. Zou, and T. Cai. 2012. Nonconcave penalized composite conditional likelihood estimation of sparse Ising models. *Annals of Statistics* 40:1403–1429.

Y., Zheng, and Zhu J. 2008. Markov chain Monte Carlo for spatial-temporal autologistic regression model. *Journal of Computational and Graphical Statistics* 17: 123–127.

Yuan, M., and Y. Lin. 2006. Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society: Series B* 68:49–67.

Zhang, Y., R. Li, and C.-L. Tsai. 2010. Regularization parameter selections via generalized information criterion. *Journal of the American Statistical Association* 105:312–323.

Zhao, P., G. Rocha, , and B. Yu. 2009. The composite absolute penalties family for grouped and hierarchical variable selection. *The Annals of Statistics* 37:3468–3497.

Zhao, P., and B. Yu. 2004. On model selection consistency of lasso. *Journal of Machine Learning Research* 7:2541–2567.

Zhu, J., H.-C. Huang, and P.E. Reyes. 2010. On selection of spatial linear models for lattice data. *Journal of Royal Statistics Society Series B* 72:389–402.

Zhu, J., H.-C. Huang, and J.-P. Wu. 2005. Modeling spatial-temporal binary data using Markov random fields. *Journal of Agricultural, Biological, and Environmental Statistics* 10:212–225.

Zhu, J., Y. Zheng, A.L. Carroll, and B.H. Aukema. 2008. Autologistic regression analysis of spatial-temporal binary data via Monte Carlo maximum likelihood. *Journal of Agricultural, Biological, and Environmental Statistics* 13:84–98.

Zhu, Z., and Y. Liu. 2009. Estimating spatial covariance using penalized likelihood with weighted L_1 penalty. *Journal of Nonparametric Statistics* 21:925–942.

Zou, H. 2006. The adaptive LASSO and its oracle properties. *Journal of the American Statistical Association* 101:1418–1429.

Zou, H., and T. Hastie. 2005. Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B* 67:301–320.

Zou, H., and R. Li. 2008. One-step sparse estimates in nonconcave penalized likelihood models. *Annals of Statistics* 36:1509–1533.

Zou, H., and M. Yuan. 2008. The F_∞ norm support vector machine. *Statistica Sinica* 18:379–398.