Statistical Methods on Variable Selection in Structured Longitudinal Data with Missing Information

by

Heejung Son

(Under the Direction of Ye Shen and Donglan Zhang)

Abstract

Despite substantial declines in cardiovascular disease (CVD) mortality across counties in the United States from 2009 to 2018, notable racial/ethnic, socioeconomic, and regional disparities persist. Health disparities in CVD mortality are closely linked to social determinants of health (SDOH), highlighting the need to address SDOH domains. Addressing these domains through targeted strategies is vital for reducing disparities and improving CVD outcomes. Challenges related to longitudinal data on SDOH include correlations of observations from the same subject and potential time-varying response patterns. Therefore, it is crucial to utilize statistical models that consider the within-subject correlation and the time-dependent effects of covariates. Models providing population-averaged effects or individual-specific estimates have been developed to address these challenges. Missing data often arise in longitudinal studies and are generally assumed to be missing at random when conditioned on relevant observed information. Modern longitudinal studies operate within a high-dimensional framework. Variable selection and regularization methods effectively address related challenges in SDOH, as they shrink coefficients to prevent overfitting and select variables within groups. The Exclusive Lasso manages grouped variables, ensuring at least one predictor from each predefined group is selected. Given the high-dimensional and longitudinal nature of county-level SDOH data, advanced clustering methods are necessary to reveal variations in the longitudinal relationship between SDOH domains and CVD mortality. Different subpopulations can demonstrate distinct behaviors over time, highlighting the necessity for clustering techniques to identify more homogeneous groups. In this dissertation, I developed a novel approach to integrate Exclusive Lasso into penalized weighted generalized estimating equations to facilitate domain-specific variable selection under missing at random. Furthermore, I propose a model-based clustering extension for high-dimensional longitudinal data, utilizing Exclusive Lasso to identify subpopulations of counties influenced by distinct covariates within each domain. Finally, to enhance this approach, I will employ the model-based clustering method using Exclusive Lasso to refine our understanding of county-level variations within each state. By integrating an additional algorithm that considers these variations, we can categorize counties based on their unique characteristics.

Index words: [Cardiovascular Disease, Exclusive Lasso, High-Dimensional Longitudinal Data, Missing Data, Model-Based Clustering, Penalized Weighted Generalized Estimating Equations, Social Determinant of Health, Variable Selection]

Statistical Methods on Variable Selection in Structured Longitudinal Data with Missing Information

by

Heejung Son

M.S., University of Georgia, 2019

A Dissertation Submitted to the Graduate Faculty of the
University of Georgia in Partial Fulfillment of the Requirements for the Degree.

Doctor of Philosophy

Athens, Georgia

2025

Statistical Methods on Variable Selection in Structured Longitudinal Data with Missing Information

by

Heejung Son

Major Professor: Ye Shen

Donglan Zhang

Committee: Zhuo Chen
Kevin K. Dobbin
Stephen L. Rathbun

Electronic Version Approved:

Ron Walcott
Dean of the Graduate School
The University of Georgia
May 2025

Acknowledgments

As I complete my educational journey, I want to acknowledge my advisors, Drs. Ye Shen and Donglan Zhang. Their guidance on research skills and learning strategies, based on my strengths, has motivated me to persist in my efforts. This work would not have been possible without their support. Especially their encouragement played a crucial role during the pandemic in helping me persevere and maintain a positive mindset toward my goals and assignments. Again, I am truly thankful for their mentorship. I would also like to acknowledge the committee members: Drs. Zhou Chen, Kevin K. Dobbin, and Stephen Lynn Rathbun, for their engagement with my work. Their transformative feedback helped me navigate challenges and think differently about my approach. Lastly, I sincerely appreciate my mom and brother for their support throughout this journey. Their encouragement has always inspired me to explore new challenges and experiences.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Background

Longitudinal data analysis is applied in various fields, including public health, clinical trials, and social science research. This approach allows for investigating how outcomes change over time and how these changes are associated with different covariates (Fitzmaurice et al., 2012; Hedeker & Gibbons, 1997). A key challenge in longitudinal settings is that observations on the same subject are correlated, and responses may exhibit time-varying structures. Accordingly, statistical models must account for this within-subject correlation and possible dynamic effects of covariates across different time points (Diggle, 2002; Liu & Deme, 2012). Generalized estimating equations (GEE) (Liang & Zeger, 1986) and linear or generalized linear mixed-effect models (LMMs) (Verbeke & Molenberghs, 2000) are popular modeling frameworks for longitudinal data. GEE targets marginal (population-averaged) inference and can be more robust to certain covariance mis-specifications than fully likelihood-based methods(Fitzmaurice et al., 2012; Hubbard et al., 2010). However, if the working correlation structure is incorrectly specified, it can result in a loss of efficiency or consistency (Fitzmaurice et al., 2012; Hubbard et al., 2010). In contrast, LMMs incorporate fixed and random effects, enabling the capture of variations at both the population and individual levels. This flexibility enables LMMs to effectively handle missing information, unbalanced data, and individual variations from a common pattern. (Laird & Ware, 1982; Verbeke & Molenberghs, 2000). Afterward, semiparametric mixed-effects models extend the LMMs, which can include an unknown function as an infinite dimensional parameter, providing flexibility to the model. Semiparametric approaches (Buckley & James, 1979; D. Y. Lin & Ying, 2001; Tsiatis, 2006) have been increasingly applied in recent longitudinal study designs to keep the balance between parametric and nonparametric methods and make concise inferences (Johnson et al., 2008; Wang et al., 2012).

A significant challenge in longitudinal studies is missing data, which often arises from dropout (when participants leave a study before completion) or intermittent missing observations, such as skipped visits or lost samples. Missing data mechanisms are typically classified into missing completely at random (MCAR), missing at random (MAR), and missing not at random (MNAR). MCAR indicates the probability of missingness is unrelated to any observed or unobserved data; in this ideal case, excluding incomplete obser-

vations only sacrifices power but does not bias estimates. MAR means the observed data can explain any systematic differences between missing and observed data. Formally, the probability that an observation is missing may depend on past observed outcomes or covariates but not on the unobserved value itself. This is a common assumption in longitudinal studies with dropouts. MNAR (also called informative missingness) occurs when the missingness depends on the unobserved data, which requires more complex modeling and is beyond the scope of most standard methods. In practice, dropout in longitudinal studies is often assumed MAR after conditioning on relevant observed information. It is crucial to address missingness because simply analyzing complete cases (ignoring subjects after they drop out) can lead to bias if the missingness is not MCAR. GEE can produce biased estimates unless the MCAR assumption holds, so weighted GEE (WGEE) methods were developed to address MAR mechanisms by incorporating inverse probability weighting (Fitzmaurice et al., 2012; Robins et al., 1995). When the dropout model is correctly specified, WGEE provides consistent estimates of the regression parameters. Notably, if there is no missing data (or if data are indeed MCAR), the weights are all equal, and WGEE reduces to the GEE, ensuring no efficiency is lost.

Beyond the need to address correlation and missingness, modern longitudinal studies frequently collect high-dimensional covariates. This high dimensionality can result from repeated measurements of numerous variables at each time point, which may be further complicated by additional time-varying factors and interaction effects (J. Fan et al., 2020). Traditional approaches often struggle to maintain statistical efficiency when the number of covariates grows relative to the sample size. To tackle these challenges, variable selection and regularization methods such as the Least Absolute Shrinkage and Selection Operator (Lasso) (Tibshirani, 1996) and smoothly clipped absolute deviation (SCAD) (J. Fan & Li, 2001) offer potential solutions by shrinking coefficients to control overfitting and select a parsimonious subset of predictors (Zou, 2006). However, Lasso may ignore the correlation between variables by selecting only one variable from a group of highly correlated variables (Friedman et al., 2010; Wang et al., 2012). Furthermore, the shrinkage effect in Lasso can lead to significant bias in parameter estimates, particularly when variables are highly correlated or organized in grouped structures (Friedman et al., 2010). Most existing literature on Lasso and SCAD has focused primarily on cross-sectional parametric models (Kowalski et al., 2018). Consequently, extensions of these methods have been developed for longitudinal contexts, including penalized GEE (W. J. Fu, 2003; Wang et al., 2012), penalized weighted GEE (Kowalski et al., 2018) and model selection in LMM (Arribas-Gil et al., 2015; Komárek & Komárková, 2013; Muller et al., 2013; Proust-Lima et al., 2015).

In addition to high dimensionality, large or complex longitudinal datasets often exhibit substantial heterogeneity. Different subgroups within the data may exhibit varying behaviors over time, prompting clustering techniques to identify more homogeneous subpopulations (Arribas-Gil et al., 2015; Proust-Lima et al., 2015). When both high dimensionality and heterogeneity are present, clustering methods can simultaneously address variable selection and the grouping of subjects (Komárek & Komárková, 2013; Yang & Wu, 2022). This combined approach enhances interpretability and predictive performance, which is particularly valuable in public health and clinical research, where identifying meaningful subgroups is crucial.

Building upon these motivations, we propose novel methods for variable selection in longitudinal analyses with missing information. For variable selection, we primarily incorporated the Exclusive Lasso to manage grouped variables, ensuring that at least one predictor from each predefined group is selected (Zhou et al., 2010). Unlike group Lasso (Yuan & Lin, 2006), which imposes an all-or-nothing selection of groups, Exclusive Lasso guarantees that each group contributes to the prediction of outcomes with at least one variable. This property is particularly valuable when domains or sets of covariates are thought to be relevant in principle based on prior knowledge, but the relative importance of variables within each domain in the relationship with the outcome remains uncertain. For instance, social determinants of health (SDOH) data often encompass multiple domains, including social context, economic stability, education, physical infrastructure, and healthcare context, but only a subset of specific variables within each domain may be crucial for a particular outcome (Son et al., 2023; Zhu & Xie, 2017).

In this dissertation, I developed a novel approach to integrate Exclusive Lasso into penalized weighted generalized estimating equations (PWGEE) to facilitate domain-specific variable selection under missing data mechanisms. The study utilizes the Social Determinants of Health database from the Agency for Healthcare Research and Quality (AHRQ), which comprises numerous county-level SDOH variables measured repeatedly over a decade across the United States (US) and is linked to county-level cardiovascular disease (CVD) mortality. In addition to missingness in certain covariates over time, the dataset encompasses five broad domains of predictors, each with multiple specific variables. The Exclusive Lasso penalty employed in PWGEE possesses the potential to retain at least one predictor from each domain while discarding variables that offer no or irrelevant information, thereby enhancing interpretability compared to group Lasso. Furthermore, CVD mortality exhibits variations across geographic regions and demographic groups, indicating inherent heterogeneity among counties. Consequently, I propose a model-based clustering extension for high-dimensional longitudinal data, utilizing Exclusive Lasso to identify subgroups of counties influenced by distinct covariates within each domain. Finally, to enhance this approach, I will employ a model-based clustering method using Exclusive Lasso to refine our understanding of county-level variations within each state. By integrating an additional algorithm that considers these variations, we can categorize counties based on their unique characteristics.

The remainder of this dissertation is organized as follows. In Chapter 2, we present a novel PWGEE method that integrates Exclusive Lasso to address missing at random and to perform domain-based selection of variables. Chapter 3 proposes a model-based clustering method that integrates Exclusive Lasso in high-dimensional longitudinal data to discover latent subgroups of US counties. In Chapter 4, we further investigate the properties of the proposed model-based clustering for the empirical study, focusing on reducing clustering variation to assist stakeholders or policymakers in making informed decisions. Final observations are provided in the Conclusion Chapter.

1.2 Motivation

1.2.1 The Longitudinal Social Determinants of Health Database

Our study is motivated by research on health disparities in social determinants of health (SDOH) and their impact on cardiovascular disease mortality. In 2020, the Agency for Healthcare Research and Quality (AHRQ) compiled and released the SDOH database to better understand the relationship between community-level factors, healthcare quality and delivery, and individual health to then address emerging health issues (for Healthcare Research & Quality, 2020). The AHRQ released its database on SDOH spanning from 2009 to 2018 (for Healthcare Research & Quality, 2020). The SDOH database is publicly available and compiles data from existing federal datasets and other public data sources (for Healthcare Research & Quality, 2020). Variables in the dataset are organized into five SDOH domains (Described in Table 1.1) (for Healthcare Research & Quality, 2020): 1) social context, such as age, race/ethnicity, and veteran status; 2) economic context, such as income and unemployment, 3) education, 4) physical infrastructure, such as housing, the built environment, and transportation, and 5) healthcare context, such as health insurance coverage and health care access (for Healthcare Research & Quality, 2020). Mortality data were sourced from the Interactive Atlas of Heart Disease and Stroke at the Centers for Disease Control and Prevention (CDC) (of Heart Disease & Stroke, n.d.), which were initially compiled from two data sources: 1) the National Vital Statistics System at the National Center for Health Statistics, and 2) the hospital discharge data from the Centers for Medicare and Medicaid Services' Medicare Provider Analysis and Review (MEDPAR) file (of Heart Disease & Stroke, n.d.). The SDOH data were linked to the corresponding mortality data at the county level.

Table 1.1: SDOH Domains and Topic Areas Represented in the SDOH Database

SDOH Domain	**SDOH Topic Area**
1. Social context	Demographics
	Living conditions
	Disability
	Immigration
	Socioeconomic disadvantage indices
	Segregation
2. Economic context	Income
	Employment
	Poverty
3. Education	Attainment
	School system
	Educational funding
	Literacy

Continued on next page

Table 1.1 – continued from previous page

SDOH Domain	SDOH Topic Area
	Numeracy
4. Physical infrastructure	Housing
	Transportation
	Migration
	Internet connectivity
	Environment
	Industry composition
	Social services
	Food access
	Access to exercise
	Crime
5. Healthcare context	Health insurance status
	Characteristics of health care providers
	Characteristics of health care facilities
	Distance to provider
	Utilization and costs
	Health behaviors
	Health outcomes
	health care quality

The SDOH data initially contained a total of 345 variables. However, it was noted that all variables included missing values, and some were not collected every year. In our previous study, the analytic sample was based on the following inclusion and exclusion criteria. First, variables that were measured repeatedly for 10 years were included. Second, variables with more than 60% missing values were excluded. Third, we eliminated redundant variables with the same or similar definitions (e.g., percentage of native-born residents and percentage of foreign-born residents) to avoid duplication. After applying these criteria, 78 variables were retained in the analytic sample, which encompassed 3,142 counties. Of these counties, 1,166 were classified as urban, while the remaining 1,976 were classified as rural. The rural-urban status of a county was determined according to the Urban-Rural Classification Scheme for Counties used by the National Center for Health Statistics (NCHS) in 2013 (Ingram & Franco, 2014). This Scheme categorized US counties into six groups based on population size, urbanization, and proximity to major cities (Ingram & Franco, 2014). Category 1, Large central metro, comprised counties in metropolitan statistical areas (MSA) with 1 million or more residents that either contained the entire population of the largest principal city, had their entire population within that city, or included at least 250,000 residents of any principal city. Category 2, Large fringe metro, included counties in the same large MSA that did not meet the central criteria, typically representing suburban areas. Category 3, Medium metro, consisted of counties in MSA with populations ranging from 250,000 to 999,999, while Category 4, Small metro, covered

counties in MSAs with populations under 250,000. On the rural side, Category 5, Micropolitan, identified counties in micropolitan statistical areas centered on urban clusters with 10,000 to 50,000 residents, and Category 6, Noncore, consisted of counties that did not qualify as micropolitan. Counties in Categories 1 through 4 were classified as urban, whereas those in Categories 5 and 6 were classified as rural, offering a nuanced framework for analyzing differences in infrastructure, economic opportunities, and public health resources (Ingram & Franco, 2014).

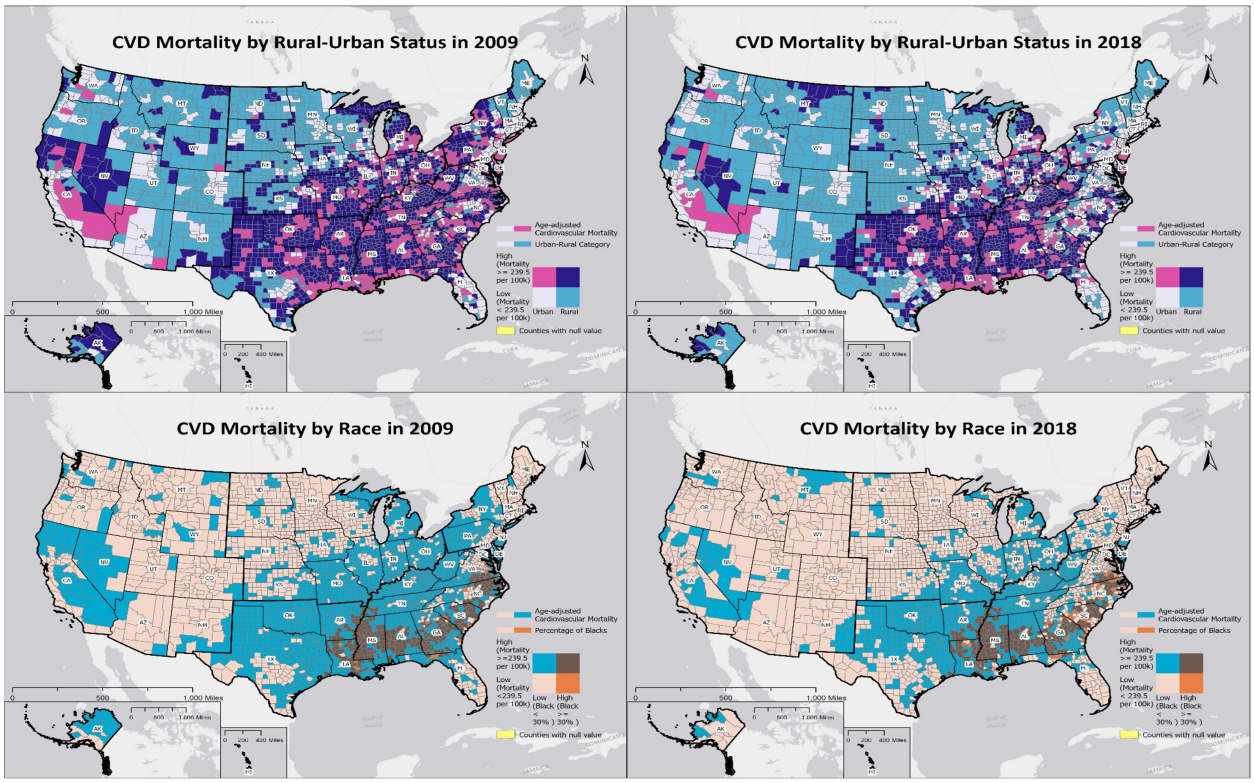


Figure 1.1: Geographic distribution of Age-adjusted Cardiovascular Disease Mortality across Counties in the United States, 2009 vs. 2018

Figure 1.1 shows the geographic distribution of age-standardized cardiovascular disease (CVD) mortality per 100,000 persons, stratified by race and rural-urban status of US counties in 2009 and 2018. The overall median CVD mortality rate (239.5; interquartile range [IQR]: 208.3 – 277.0) was used as the bivariate threshold, and a 30% cut-off value was used to define the racial composition of each county, measured by the percentage of Black residents (Explorer, n.d.). In 2018, the number of counties with a higher CVD mortality (>239.5) was smaller than those in 2009 in both rural and urban counties. Similarly, counties with a higher percentage of Black residents tend to show elevated CVD mortality rates compared to those with fewer Black residents.

Moreover, our previous study (Son et al., 2023) findings identified 17 key SDOH associated with county-level CVD mortality, including rural-urban status, racial composition, median household income, food insecurity, and housing instability. Although there was an overall decrease in CVD mortality by 1.08

deaths per 100,000 people each year from 2009 to 2018, rural counties and those with higher percentages of Black residents consistently experienced higher CVD mortality rates than urban counties and those with lower percentages of Black residents. The rural-urban CVD mortality gap did not change significantly over the past decade, whereas the association between the percentage of Black residents and CVD mortality showed a significant diminishing trend over time.

Nevertheless, several analytical gaps remain unresolved in this evaluation. First, the analysis did not address the missing value issues during the variable selection process, which could influence the robustness of the findings. Second, the analysis did not select variables within each domain, limiting interpretability regarding domain-specific influences on CVD mortality. Third, the analysis did not consider clustering across geographic regions, which may bias results by ignoring ecological correlations (Robinson, 2009). Lastly, the analysis did not account for the state as a natural grouping or clustering factor, potentially obscuring important variations related to state-specific policies or environmental conditions.

Adverse social and environmental conditions, such as barriers to accessing healthcare, unsafe living environments, inadequate education, and unequal employment opportunities, which are defined as SDOH, are associated with poor health outcomes through racial disparities. (Churchwell et al., 2020; Virani et al., 2021). SDOH can influence disparities in CVD outcomes in various ways. Therefore, it is important to understand the comprehensive SDOH framework and how the burden of racial disparities contributes to the risk of CVD (Javed et al., 2022). A nuanced understanding of which SDOH indicators within specific domains may be more important in determining CVD mortality over time is needed to identify promising approaches to address disparities in CVD outcomes. Despite a decrease in CVD mortality over a decade, the decline may not be uniform across different geographic areas. This is evident in the variation of mortality rates in certain counties differing compared to national levels (Case & Deaton, 2015). Thus, addressing varying CVD mortality rates at the county, state, or other geographic levels based on SDOH is important for informing local and national health policies. In our research, addressing missing data for high-dimensional SDOH indicators presents a challenge when fitting longitudinal models and interpreting their association with mortality within the SDOH framework. Under the assumption that the missing data is MAR, and given the large number of predictors and repeated measurements, we aim to select the variables within the stratified framework. Additionally, we aim to identify variables within each cluster to describe distinct trends in CVD mortality based on the SDOH framework across counties or counties within the state.

Chapter 2

Penalized Weighted Generalized Estimating Equations via Exclusive Lasso Penalty

2.1 Introduction

Longitudinal data analysis is a widely used technique in public health, clinical trials, and social science research. Using data collected from repeatedly measuring individuals over time, this analysis enables the identification of temporal changes in responses and the elucidation of intricate relationships between these changes and covariates. Semi-parametric approaches to longitudinal study designs have increasingly been employed to balance parametric and non-parametric methods and make concise inferences. The generalized estimating equations (GEE) (Liang & Zeger, 1986) are popular approach for population-averaged effects and substitute for the likelihood-based generalized linear mixed model, which is more susceptible to consistency loss when specifying variance structures (Breslow & Clayton, 1993; Diggle, 2002; Fitzmaurice et al., 2012; Stokes et al., 2012). Subsequently, weighted generalized estimating equations (WGEE), along with the technique of inverse missing probability weighting (IPW) adjusting for bias, were proposed to analyze longitudinal data that incorporate missing values (Robins et al., 1995). In the GEE and WGEE approaches, valid inference for various classes of data distributions is assigned by specifying the conditional mean of responses given the independent covariates. However, unless the data are missing completely at random (MCAR), the GEE approach may introduce bias in parametric estimates (Fitzmaurice et al., 2012; Hardin & Hilbe, 2003; Preisser et al., 2002; Robins et al., 1995). In addition, the WGEE provide consistent estimates under the missing at random (MAR) mechanism, where the probability of missing responses is independent of current and future responses, given the observed past responses and covariates, particularly when a monotone missing data pattern or dropouts occur (Fitzmaurice et al., 2012; Mallinckrodt, 2013; O'Kelly & Ratitch, 2014). Consequently, the WGEE facilitates the adoption of a practical approach for longitudinal study designs that exhibit missing data patterns.

Our research is motivated by addressing health disparities related to the social determinants of health (SDOH) and cardiovascular disease (CVD) mortality. Our dataset, the SDOH database from the Agency for Healthcare Research and Quality (AHRQ), included variables that have been repeatedly measured across the United States (US) at the county level. It encompasses a substantial number of multiple measurements on the same subject over time and comprises five domains of variables. Furthermore, besides the SDOH data, most explanatory variables are subject to some missingness over the course of the study period. Consequently, model selection presents a significant challenge that arises in a wide range of disciplines.

The Lasso and smoothly clipped absolute deviation penalty (SCAD) class of methods has been popular for variable selection in regression analysis and regularization to deal with high-dimensional data in clinical and public health research (J. Fan & Li, 2001; Tibshirani, 1996; Zou, 2006). SCAD modifies the Lasso framework by using a piecewise penalty that reduces bias for larger coefficients while still promoting sparsity (J. Fan & Li, 2001; J. Fan & Lv, 2011; Tibshirani, 1996; Zou, 2006). In current studies, penalized generalized estimating equations (PGEE) were derived from applying a framework to the longitudinal analysis in which regularized penalty methods were used (W. J. Fu, 2003; Wang et al., 2012). In particular, Wang et al., 2012 developed SCAD penalized GEE to analyze longitudinal data with a high dimensional framework (large n, diverging p) for simultaneous variable selection and estimation (Blommaert et al., 2014; Wang et al., 2012). While PGEE and related penalized methods significantly enhance GEE by addressing high-dimensional covariate selection, they retain GEE's limitations with regard to missing data. In particular, these methods typically assume the same missingness conditions as standard GEE, which is formally derived under the assumption that data are MCAR. If the longitudinal outcome or covariates are missing in a systematic manner (e.g., missing at random dropout), a penalized GEE without correction can yield biased estimates and suboptimal variable selection (Blommaert et al., 2014; Wang et al., 2012). Recently, Kowalski et al., 2018 developed penalized weighted generalized estimating equations (PWGEE), which integrates the SCAD class of variable selection with weighted GEE under the MAR missing mechanism (Kowalski et al., 2018). However, this penalized WGEE method, along with penalties such as Lasso and SCAD, cannot effectively address variable selection while considering between-group and within-group correlations in a predefined group structure (Kowalski et al., 2018). This limitation highlights the need for improved methods that account for these correlations.

Thus, another critical consideration in longitudinal data analysis is the potential grouping of covariates and the desire for grouped variable selection. Often, covariates can be naturally divided into groups, such as SDOH domains, multiple measurements related to the same risk factor, gene expression probes corresponding to the same pathway, or dummy variables representing categories of a factor. Group Lasso (Yuan & Lin, 2006) selects or discards covariates in predefined groups using an L_2 norm within each group and an L_1 penalty across groups, forcing all-or-nothing selection at the group level. However, this can be restrictive if only a subset of variables in a group is relevant. The Exclusive Lasso (Zhou et al., 2010) encourages within-group sparsity while ensuring no group is neglected. It uses a combination of L_1 and L_2 norms (sometimes called an $L_{1,2}$ penalty in reverse order) to penalize having multiple non-zero coefficients in the same group, effectively selecting at least one variable from each group but not necessarily all of them.

For each group of related features, the Exclusive Lasso tends to pick the single most predictive feature and set the rest to zero rather than selecting the entire group or potentially dropping the whole group. This within- and between-grouped variable selection approach is relevant for longitudinal data with covariates grouped by meaningful criteria (e.g., multiple time-specific measurements of the same variable from each domain). It ensures each important covariate domain contributes to the model, improving interpretability and preventing the omission of important factors due to redundancy. Recent studies have explored the Exclusive Lasso for structured feature selection, showing its advantages in selecting representative features from each domain (Campbell & Allen, 2017; Kong et al., 2017; Zhao et al., 2018).

Given the gaps in the existing methods, we propose a penalized weighted GEE with an Exclusive Lasso to tackle missing data and perform variable selection based on predefined domains in longitudinal analysis. This approach integrates inverse probability weighting for missing data and an Exclusive Lasso penalty for covariates within predefined domains. Doing so achieves consistent estimates and structured variable selection, selecting a representative subset from each domain rather than individual covariates. This extension of the penalized GEE framework accommodates incomplete data and emphasizes group-level sparsity. The Exclusive Lasso ensures that at least one covariate from each relevant group is retained in the model. This might be useful in longitudinal settings, such as selecting the most informative time point for a risk factor measured multiple times since it can incorporate the most relevant time-specific data into predictive models to improve their accuracy (Chen et al., 2018). The proposed penalized WGEE addresses two key challenges in longitudinal data analysis: robustness to missing data through weighting and within- and between-grouped variable selection in complex covariate structures through the Exclusive Lasso. By leveraging the strength of all available data and encouraging a parsimonious yet domain-conscious model, it enhances estimation accuracy and interpretability in longitudinal studies of CVD mortality with complex covariate structures and incomplete observations in SDOH.

2.2 Methods

In this section, we use the composite penalty to extend the WGEE with Exclusive Lasso variable selection (PWGEE-eLasso), which performs l_1 norm within the group by applying separate l_2 penalties to each group. This method can also be extrapolated to other multidimensional social determinant data.

2.2.1 WGEE

At the beginning of this section, we define some underlying notations. For $i = 1, ..., n$ and $t = 1, ..., m$, we consider n as subjects (i.e., counties in our data) and m assessment times in our longitudinal study. Let Y_{it} indicate a response for subject i at time t, and let $\mathbf{X}^*_{it}$ denote a p-dimensional vector that includes the SDOH variables from the ith subject at time t $(1 \leq i \leq n, 1 \leq t \leq m)$. $\mathbf{X}_{it} = (1, \mathbf{X}^*_{it})$ have a $(1 + p)$ column vector, where the first component of the column vector is the constant one and the remaining p components are the explanatory variables. As is well known, WGEE comprises two modules: 1) the main module, which estimates the association between primary outcomes and its predictors, and 2) the

missing data module under MAR assumption. We denote that WGEE in the main module assumes the relationship between y_{it} and $\mathbf{X_{it}}$ via the conditional mean of Y_{it} given $\mathbf{X_{it}}$:

$$E(Y_{it}|\mathbf{X}_{it}) = \mathbf{g}(\mathbf{X}_{it}^T\boldsymbol{\beta}) = \mathbf{g}_{it}, \quad 1 \leq i \leq n, 1 \leq t \leq m, \tag{2.1}$$

where $g(\cdot)$ is a link function (depending on the type of Y_{it}) and $\boldsymbol{\beta} = (\beta_0, \beta_1, ..., \beta_p)^T$ is a vector of parameters with β_0 denoting the intercept term. The link function depends on the type of y_{it}, but the identity link function will be used in our study since the study outcome of primary interest is assumed to follow the Gaussian distribution. Under the MAR missing mechanism, WGEE handles missing data to perform valid inferences about $\boldsymbol{\beta}$. We assume that Y_{it} and $\mathbf{X_{it}}$ are either observed or missing together at time t (Robins et al., 1995). Suppose that missing data indicators R_{it} in the missing data module are defined by:

$$R_{it} = \begin{cases} 1 & \text{if } (Y_{it}, \mathbf{X_{it}}) \text{is observed} \\ 0 & \text{otherwise} \end{cases}. \tag{2.2}$$

Under MAR, the probability of missing data process satisfies:

$$\begin{aligned} &\pi_{it} = Pr(R_{it} = 1|H_{it}), \quad 2 \leq t \leq m, \\ &H_{it} = \{\mathbf{X_{it-}}, \mathbf{Y_{it-}}; 2 \leq t \leq m\}, \\ &\mathbf{X_{it-}} = ((X_{i1}^*)^T, ..., (X_{i(t-1)}^*)^T)^T, \quad \mathbf{Y_{it-}} = (Y_{i1}, ..., Y_{i(t-1)})^T, \end{aligned} \tag{2.3}$$

where H_{it} denotes the observed information prior to time $t (2 \leq t \leq m, 1 \leq i \leq n)$. In addition to $\mathbf{X_{it-}}$ and $\mathbf{Y_{it-}}$, we can incorporate other variables observed before time t. Although these ancillary variables are not relevant for modeling $\mathbf{Y_{it-}}$ in the primary module, they may be useful for modeling R_{it}. For simplicity, we assume $\mathbf{X_{it}}$ contains all this information. $R_{i1} = 1$ for all observations $1 \leq i \leq n$, i.e., we assume no missing data are observed at baseline $t = 1$. A monotone missing data pattern, such as dropouts, is assumed to estimate π_{it} and construct weights to facilitate inference about $\boldsymbol{\beta}$ in the main module. Under the monotone missing data pattern, H_{it} is observed if $H_{i(t-1)}$ is. This monotone missing data pattern is a reasonable assumption when dropouts are the major source of missing data, as study dropouts in longitudinal studies typically follow this type of missing data pattern. Thus, we let $p_{it} = Pr(R_{it} = 1|R_{i(t-1)} = 1, H_{it})$.

Consider the following WGEE:

$$\mathbf{U}_n(\boldsymbol{\beta}; \boldsymbol{\xi}) = \sum_{i=1}^{n} \mathbf{U}_i = \sum_{i=1}^{n} D_i V_i^{-1} W_i (Y_i - \mathbf{g}_i) = \mathbf{0}, \tag{2.4}$$

where

$$\begin{aligned} &W_{it} = \tfrac{R_{it}}{\pi_{it}}, \quad W_i = \text{diag}_t(W_{it}), \\ &D_i = \tfrac{\partial \mathbf{g}_i}{\partial \boldsymbol{\beta}}, \quad \mathbf{g}_i = (g_{i1}, ..., g_{im})^T, \quad V_i = A_i^{1/2} R(\boldsymbol{\alpha}) A_i^{1/2}, \\ &A_{it} = Var(Y_{it}|\mathbf{X_{it}}), \quad A_i = \text{diag}_t(A_{it}), \end{aligned} \tag{2.5}$$

where $\boldsymbol{\xi}$ is denoted as parameter estimates in the missing data module, $\text{diag}_t(W_{it})$ denotes a $m \times m$ diagonal matrix with W_{it} on the ith elements, a working correlation matrix $R(\boldsymbol{\alpha})$ is parameterized by $\boldsymbol{\alpha}$ and A_i is defined by a $m \times m$ diagonal matrix with A_{it} as the tth diagonal elements, $\text{diag}_t(A_{it})$ (Robins et al., 1995). For inference about $\boldsymbol{\beta}$, we will apply a popular two-step procedure (G. Lin et al., 2015). This procedure first estimates $\boldsymbol{\xi}$ with logistic regression and then solves for $\boldsymbol{\beta}$ in Equation 2.2 by substituting such estimates in lieu of $\boldsymbol{\xi}$ in the weight function. Mild regularity conditions in WGEE refer to standard theoretical assumptions that include the smoothness and differentiability of estimating functions, boundedness of covariates and weights, correct specification of the regression model mean structure, nonsingularity and positive definiteness of variance-covariance matrices, and proper handling of missing data mechanisms (Kowalski & Tu, 2008; Robins et al., 1995). These conditions ensure the consistency and asymptotic normality of parameter estimates, $\hat{\boldsymbol{\beta}}$, and are typically met in practical longitudinal analyses. While the main module is of primary interest, we must also take into account the inference regarding $\boldsymbol{\xi}$ in the missing data module. This is important because the inference concerning $\boldsymbol{\beta}$ depends on a set of WGEE that involves the $\pi_{it}(\boldsymbol{\xi})'s$ (Robins et al., 1995). Using logistic regression, following the one-step transition probability of observing the subject at t given observing this subject at $t-1$ is modeled:

$$\begin{aligned} \text{logit}(p_{it}(\boldsymbol{\xi}_t)) &= \text{logit}[Pr(r_{it}=1|r_{i(t-1)}=1, H_{it})] \\ &= \xi_0 + \boldsymbol{\xi}_{xt}^T(\mathbf{X}_{it-}) + \boldsymbol{\xi}_{yt}^T(\mathbf{y}_{it-}), \quad 2 \le t \le m, \end{aligned} \tag{2.6}$$

where $\boldsymbol{\xi}_t = (\xi_{0t}, \boldsymbol{\xi}_{xt}^T, \boldsymbol{\xi}_{yt}^T)^T$. Since

$$\pi_{it}(\boldsymbol{\xi}) = p_{it}Pr(r_{i(t-1)}=1|H_{i(t-1)}) = \prod_{s=2}^{t} p_{is}(\boldsymbol{\xi}_s), \quad 2 \le t \le m, 1 \le i \le n,$$

where $\boldsymbol{\xi} = (\boldsymbol{\xi}_2^T, ..., \boldsymbol{\xi}_m^T)^T$, the weight function π_{it} in Equation 2.4 can be estimated by the missing data module in Equation 2.6. Thus, when the number of covariates is large, estimation in models 2.4 and 2.6 becomes challenging. In such instances, it is imperative to employ variable selection techniques, as they simultaneously reduce dimensionality and estimate parameters, effectively enhancing model interpretability and predictive performance (Zou & Hastie, 2005).

2.2.2 Variable Selection

The Exclusive Lasso function will be applied to estimate $\boldsymbol{\xi}$ for the missing data modules. In the main module, Exclusive Lasso will be further described for the variable selection for our study. Penalty functions in Exclusive Lasso adaptively shrink parameter estimates to select variables, similar to Lasso and SCAD. However, the Exclusive Lasso penalty is composed of the l_1 norm within groups and the l_2 norm between them (Campbell & Allen, 2017). That composite penalty conducts selection within the group by applying separate Lasso penalties to each group (Campbell & Allen, 2017). Thus, the Exclusive Lasso always selects one non-zero variable in each group. The Exclusive Lasso assumes two structural conditions for the parameters in Equation 2.7 (Campbell & Allen, 2017; Zhou et al., 2010):

Assumption (1): There exists a collection of non-overlapping predefined groups given, G, such that $\bigcup_{g \in G} g = (1, ..., p)$, and any pair of groups $g, g' \in G$ satisfies $\bigcap_{g,g'} = \emptyset$.

Assumption (2): The support set S of the true parameter β^* intersects each group, such that for all $g \in G$ we have $S \bigcap g \neq \emptyset$ and $\beta_i^* \neq 0$ for all $j \in S$.

The penalty function of Exclusive Lasso is captured by the equation:

$$p_\lambda(\hat{\beta}) = \text{argmin} \frac{1}{2}(\mathbf{y} - \mathbf{X}\beta)^2 + \frac{\lambda}{2} \sum_{g \in G} \left(\sum_{j \in g} |\beta_j| \right)^2 \tag{2.7}$$

where λ is a regularization parameter controlling overall penalty strength. Building on Equation 2.1 in the WGEE, we integrate this Exclusive Lasso penalty to achieve robust modeling of correlated data paired with sparse predictor selection. We implement the penalty function 2.8 to the main and missing data modules. We denote $d1, d2$ as the dimension of $\boldsymbol{\beta}(\boldsymbol{\xi})$ and let

$$\begin{aligned} &\mathbf{P}_\lambda = (\mathbf{P}_a^T, \mathbf{P}_b^T)^T, \\ &\mathbf{P}_a = (p_{a_1}(\| \xi_1 \|_1^2), p_{a_2}(\| \xi_2 \|_1^2), ..., p_{a_{d_1}}(\| \xi_{d_1} \|_1^2))^T, \\ &\mathbf{P}_b = (p_{b_1}(\| \beta_1 \|_1^2), p_{b_2}(\| \beta_2 \|_1^2), ..., p_{b_{d_2}}(\| \beta_{d_2} \|_1^2))^T, \end{aligned} \tag{2.8}$$

where $\mathbf{P}_\lambda$ is defined as the Exclusive Lasso penalty and $\lambda = (a_l, b_k)$ are tuning parameters for penalties for $\boldsymbol{\xi}(1 \leq l \leq d_1)$, $\boldsymbol{\beta}(1 \leq k \leq d_2)$. By adding penalty functions $\mathbf{P}_\lambda$ to the WGEE in Equation 2.4, we estimate $\boldsymbol{\theta} = (\boldsymbol{\beta}, \boldsymbol{\xi})^T$ and highlight a set of penalized weighted generalized estimating equations via Exclusive Lasso (PWGEE-eLasso):

$$\mathbf{Q}_n(\boldsymbol{\theta}) = \mathbf{U}_n(\boldsymbol{\theta}) - n\mathbf{P}_\lambda = \frac{1}{n} \sum_{i=1}^{n} D_i V_i^{-1} W_i (Y_i - \mathbf{g}_i) - \mathbf{P}_\lambda^T. \tag{2.9}$$

We set $p_\lambda(|\theta|) = 0$ for the intercept term, where θ represents either β_0 for the main and $\xi_{0j}(2 \leq j \leq m)$ for the missing data module, as these parameters θ_0 are usually not penalized. To optimize tuning parameters (λ), cross-validation is conducted.

The algorithm for fitting the model 2.9 follows a two-step procedure for parameter estimation (G. Lin et al., 2015). In the first step, weights are estimated using predicted probabilities derived from a logistic regression model, wherein the missing data indicator represents an indicator of missingness. In the logistic regression, the Exclusive Lasso is added to estimate $\boldsymbol{\xi}$ in equation 2.6:

$$\begin{aligned} \mathbf{Q}_n(p_{it}(\boldsymbol{\xi}_t)) &= L_n(\boldsymbol{\theta}) + p_a \\ &= \textstyle\prod_{i=1}^{n} P(\pi_{it}|\mathbf{X_{it}})^{\pi_{it}} \cdot P[1 - \pi_{it}|\mathbf{X_{it}}]^{1-\pi_{it}} + p_a(\boldsymbol{\xi}_t), \end{aligned} \tag{2.10}$$

where

$$P(\pi_{it}|\mathbf{X_{it}}) = \frac{1}{1 + exp(-(\sum_{i=1}^{n} X_{it}^T \boldsymbol{\beta}))}, \; 2 \leq t \leq m,$$

and L denotes the likelihood of a logistic regression.

Subsequently, an initial parameter estimate is calculated using an ordinary generalized linear model with $\mathbf{P_b}$ in equation 2.9, assuming independence of the responses. This assumption is applied explicitly in this step to derive an initial estimate without more comprehensive information. We solve equationl 2.9 using a coordinate descent algorithm for the Exclusive Lasso problem as well as an algorithm to compute the proximal operator (Campbell & Allen, 2017).

Then, following a multi-step approach, the weighted GEE fitting algorithm (G. Lin et al., 2015) is further implemented to provide robust parameter estimation:

1. The obtained parameter estimates and standardized residuals of model are used to construct a working correlation matrix with a predetermined structure.
2. The estimated covariance matrix is then calculated using the working correlation matrix.
3. The parameter estimates are updated using the information from the estimated covariance matrix and the inverse probability weights.
4. The first to third steps are repeated until convergence is achieved, ensuring a refined and accurate estimation of the model parameters.

2.3 Asymptotic Properties

We assume the following regularity conditions:

$$\mathbf{C}_n = \frac{1}{n}\sum_{i=1}^{n} \mathbf{X}_i \mathbf{X}_i^T \to \mathbf{C}, \tag{2.11}$$

where C is a nonsingular definite matrix and

$$\frac{1}{n}\max_{1\le i\le n} \mathbf{X}_i \mathbf{X}_i^T \to 0. \tag{2.12}$$

In practical applications, the covariates usually undergo scaling such that all diagonal elements of matrix C_n are identically set to one.

To consider the consistency of $\hat{\boldsymbol{\theta}}$, we define the following objective function:

$$\mathbf{Q}_n(\boldsymbol{\theta}) = \frac{1}{n}\sum_{i=1}^{n} D_i V_i^{-1} W_i (Y_i - \mathbf{g}_i) - \frac{\lambda_n}{2n}\frac{\partial}{\partial \boldsymbol{\theta}} \sum_{g\in G}\left(\sum_{j\in g} |\boldsymbol{\theta}_j|\right)^2 \tag{2.13}$$

where

$$\boldsymbol{\theta} = (\boldsymbol{\xi}, \boldsymbol{\beta})^T, \quad D_i = \tfrac{\partial \mathbf{g}_i}{\partial \boldsymbol{\beta}}, \quad \mathbf{g}_i = (g_{i1}, ..., g_{im})^T,$$
$$V_i = A_i^{1/2} R(\boldsymbol{\alpha}) A_i^{1/2}, \quad A_{it} = Var(Y_{it}|\mathbf{X_{it}}), \quad A_i = \text{diag}_t(A_{it}),$$

a working correlation matrix $R(\boldsymbol{\alpha})$ is parameterized by $\boldsymbol{\alpha}$ and A_i is defined by a $m \times m$ diagonal matrix with A_{it} as the tth diagonal element, $\text{diag}_t(A_{it})$.

Define

$$\mathbf{L}_n(\boldsymbol{\theta}) = \frac{1}{n}\sum_{i=1}^{n} \widehat{W}_i(Y_i - \mathbf{X}_i^T\boldsymbol{\theta})^2 + \frac{\lambda_n}{2n}\sum_{g\in G}\left(\sum_{j\in g}|\boldsymbol{\theta}_j|\right)^2 \tag{2.14}$$

Notice that $\mathbf{L}_n(\boldsymbol{\theta})$ is convex and $\mathbf{Q}_n(\theta) = \frac{\partial \mathbf{L}_n(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}$.

Solving $\mathbf{Q}_n(\boldsymbol{\theta}) = \mathbf{0}$ is equivalent to minimizing the $\mathbf{L}_n(\boldsymbol{\theta})$. Thus, it suffices to show that the minimizer of $\mathbf{L}_n(\boldsymbol{\theta})$ converges.

Thus, we will show the following proof that $\hat{\boldsymbol{\theta}}$ is consistent provided $\lambda_n = o(n)$. We define the random function $\mathbf{L}_n$:

$$\mathbf{L}_n(\hat{\boldsymbol{\theta}}) = \frac{1}{n}\sum_{i=1}^{n} \widehat{W}_i(Y_i - \mathbf{X}_i^T\hat{\boldsymbol{\theta}})^2 + \frac{\lambda_n}{2n}\sum_{g\in G}\left(\sum_{j\in g}|\boldsymbol{\theta}_j|\right)^2 \tag{2.15}$$

$$= \frac{1}{n}\sum_{i=1}^{n} (W_i + o(1/\sqrt{n}))(Y_i - \mathbf{X}_i^T\hat{\boldsymbol{\theta}})^2 + \frac{\lambda_n}{2n}\sum_{g\in G}\left(\sum_{j\in g}|\boldsymbol{\theta}_j|\right)^2 \tag{2.16}$$

$$= \frac{1}{n}\sum_{i=1}^{n} W_i(Y_i - \mathbf{X}_i^T\hat{\boldsymbol{\theta}})^2 + o(1/\sqrt{n}) + \frac{\lambda_n}{2n}\sum_{g\in G}\left(\sum_{j\in g}|\boldsymbol{\theta}_j|\right)^2 \tag{2.17}$$

$$= \left\{\frac{1}{n}\sum_{i=1}^{n} W_i(Y_i - \mathbf{X}_i^T\hat{\boldsymbol{\theta}})^2 + \frac{\lambda_n}{2n}\sum_{g\in G}\left(\sum_{j\in g}|\boldsymbol{\theta}_j|\right)^2\right\} + o(1/\sqrt{n}), \tag{2.18}$$

which is minimized at $\hat{\boldsymbol{\theta}}$. By the theorem on asymptotic normality of $\hat{\boldsymbol{\theta}}_{MLE}$, if $\hat{\boldsymbol{\theta}}$ satisfies

$$\sum_{i=1}^{n} \frac{\partial}{\partial \boldsymbol{\theta}^T} log\mathbf{L}_n(Y_i; \hat{\boldsymbol{\theta}}) = \mathbf{0} \text{ and } \hat{\boldsymbol{\theta}} \to_p \boldsymbol{\theta} \text{ as } n \to \infty, \text{ then}$$
$$\sqrt{n}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}) \to_d N(\mathbf{0}, I(\boldsymbol{\theta})^{-1}) \quad \text{as } n \to \infty,$$

thus, with $\hat{\boldsymbol{\theta}} - \boldsymbol{\theta} = o(1/\sqrt{n})$, we define $\widehat{W}_i$ as

$$\widehat{W}_i = exp(\mathbf{X}_i\hat{\boldsymbol{\theta}})/(1 + exp(\mathbf{X}_i\hat{\boldsymbol{\theta}})) = \frac{exp(\mathbf{X}_i(\boldsymbol{\theta} + o(1/\sqrt{n}))}{(1 + exp(\mathbf{X}_i(\boldsymbol{\theta} + o(1/\sqrt{n}))))},$$

then we approximate

$$\begin{aligned} exp(\mathbf{X}_i(\boldsymbol{\theta} + o(1/\sqrt{n})) &= exp(o(1/\sqrt{n})) * exp(\mathbf{X}_i\boldsymbol{\theta}) \\ &\approx (1 + o(1/\sqrt{n})) * exp(\mathbf{X}_i\boldsymbol{\theta}) \\ &= exp(\mathbf{X}_i\boldsymbol{\theta}) + o(1/\sqrt{n}), \end{aligned}$$

thus,

$$\widehat{W_i} = W_i + o(1/\sqrt{n}). \tag{2.19}$$

Theorem 2.3.1. *Let $\lambda_n/n \to \lambda_0 \geq 0$ $(1 \geq l \geq d)$. Under the regularity conditions 2.11, 2.12, $\hat{\boldsymbol{\theta}} \to_p argmin(\mathbf{L})$, where*

$$\mathbf{L}(\boldsymbol{\theta}) = W_i(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta})^T \mathbf{C}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}) + \frac{\lambda_0}{2} \sum_{g \in G} \left(\sum_{j \in g} |\boldsymbol{\theta}_j| \right)^2. \tag{2.20}$$

Thus, if $\lambda_n = 0(n)$, $argmin(\mathbf{L}) = \boldsymbol{\theta}$, so that, $\hat{\boldsymbol{\theta}}$ is consistent.

Proof. We will show that

$$\sup_{\boldsymbol{\theta} \in \mathbf{K}} |\mathbf{L}_n(\boldsymbol{\theta}) - \mathbf{L}(\boldsymbol{\theta}) - \sigma^2| \to_p 0 \tag{2.21}$$

for any compact set $\mathbf{K}$ and that

$$\hat{\boldsymbol{\theta}} = O_p(1). \tag{2.22}$$

Under Equation 2.21 and 2.22,

$$argmin(\mathbf{L}_n) \to_p argmin(\mathbf{L}).$$

According to Andersen and Gill, 1982; Pollard, 1991, since $\mathbf{L}_n$ is convex on $\boldsymbol{\theta}$, by applying standard results of the convexity lemma, Equation 2.21 and 2.22 follow from the pointwise convergence in probability of $\mathbf{L}_n(\boldsymbol{\theta})$ to $\mathbf{L}(\boldsymbol{\theta}) + \sigma^2$ ■

Theorem 2.3.2. *Let $\lambda_n/\sqrt{n} \to \lambda_0 \geq 0$ $(1 \geq l \geq d)$. Under the regularity conditions 2.11, 2.12, then*

$$\sqrt{n}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}) \to_d argmin(\mathbf{Z}) \tag{2.23}$$

where

$$\mathbf{Z}(\mathbf{u}) = -2W_i \mathbf{u}^T \boldsymbol{\Phi} + W_i \mathbf{u}^T \mathbf{C} \mathbf{u} + \lambda_0 \sum_{ginG} \Bigg[\sum_{j \in g \cap \mathcal{A}^-} (2\theta_j)(\mathbf{u}_j) + \tag{2.24}$$

$$\sum_{j \neq j' \in g \cap \mathcal{A}^-} |\theta_j \theta_{j'} \mathbf{u}_j \mathbf{u}_{j'}| + \sum_{j \in g \cap \mathcal{A}} \sum_{j' \in g \cap \mathcal{A}^-} |\mathbf{u}_j \theta_{j'}| \Bigg], \tag{2.25}$$

where $\boldsymbol{\Phi} \sim N(\mathbf{0}, \sigma^2 \mathbf{C})$, $\mathcal{A} = \{j; \theta_j = 0\}$, and $\mathcal{A}^- = \{j; \theta_j \neq 0\}$.

Proof. We define $\mathbf{Z}_n(\mathbf{u})$ by following function that

$$\begin{aligned} \mathbf{Z}_n(\mathbf{u}) \quad &= \sum_{i=1}^{n} W_i \left[(\epsilon_i - \mathbf{u}^T \mathbf{X}_i/\sqrt{n})^2 - \epsilon_i^2 \right] \\ &+ \lambda_n/2 \sum_{g \in G} \left[\left(\sum_{j \in g} |\theta_j + u_j/\sqrt{n}| \right)^2 - \left(\sum_{j \in g} |\theta_j| \right)^2 \right], \end{aligned} \tag{2.26}$$

where $\mathbf{u} = (u_1..., u_p)^T$, and $\mathbf{Z}_n$ is minimized at $\sqrt{n}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta})$. We first note that

$$\sum_{i=1}^{n} W_i \left[(\epsilon_i - \mathbf{u}^T \mathbf{X}_i/\sqrt{n})^2 - \epsilon_i^2 \right] \to_d -2W_i \mathbf{u}^T \boldsymbol{\Phi} + W_i \mathbf{u}^T \mathbf{C}\mathbf{u},$$

with finite-dimensional convergence holding trivially. Let $\mathcal{A} = \{j; \theta_j = 0\}$, $\mathcal{A}^- = \{j; \theta_j \neq 0\}$.

$$\begin{aligned}
&\frac{\lambda_n}{2} \sum_{g \in G} \left[\left(\sum_{j \in g} |\theta_j + \mathbf{u}_j/\sqrt{n}| \right)^2 - \left(\sum_{j \in g} |\theta_j| \right)^2 \right] \\
&= \frac{\lambda_n}{2} \sum_{g \in G} \left[\left(\sum_{j \in g \cap \mathcal{A}} + \sum_{j \in g \cap \mathcal{A}^-} |\theta_j + \mathbf{u}_j/\sqrt{n}| \right)^2 - \left(\sum_{j \in g \cap \mathcal{A}} + \sum_{j \in g \cap \mathcal{A}^-} |\theta_j| \right)^2 \right] \\
&= \frac{\lambda_n}{2} \sum_{g \in G} \left[\left(\sum_{j \in g \cap \mathcal{A}} |\mathbf{u}_j/\sqrt{n}| + \sum_{j \in g \cap \mathcal{A}^-} |\theta_j + \mathbf{u}_j/\sqrt{n}| \right)^2 - \left(\sum_{j \in g \cap \mathcal{A}^-} |\theta_j| \right)^2 \right] \\
&= \frac{\lambda_n}{2} \sum_{g \in G} \left[\left(\sum_{j \in g \cap \mathcal{A}} |\mathbf{u}_j/\sqrt{n}| \right)^2 + \left(\sum_{j \in g \cap \mathcal{A}^-} |\theta_j + \mathbf{u}_j/\sqrt{n}| \right)^2 \right. \\
&\left. + \sum_{j \in g \cap \mathcal{A}} \sum_{j' \in g \cap \mathcal{A}^-} |\mathbf{u}_j/\sqrt{n}||\theta_{j'} + \mathbf{u}_{j'}/\sqrt{n}| - \left(\sum_{j \in g \cap \mathcal{A}^-} |\theta_j| \right)^2 \right] \\
&= \frac{\lambda_n}{2} \sum_{g \in G} \left[\sum_{j \in g \cap \mathcal{A}} (\mathbf{u}_j/\sqrt{n})^2 + \sum_{j \neq j' \in g \cap \mathcal{A}} |\mathbf{u}_j/\sqrt{n}||\mathbf{u}_{j'}/\sqrt{n}| \right. \\
&+ \sum_{j \in g \cap \mathcal{A}^-} (\theta_j + \mathbf{u}_j/\sqrt{n})^2 + \sum_{j \neq j' \in g \cap \mathcal{A}^-} |\theta_j + \mathbf{u}_j/\sqrt{n}||\theta_{j'} + \mathbf{u}_{j'}/\sqrt{n}| \\
&\left. + \sum_{j \in g \cap \mathcal{A}} \sum_{j' \in g \cap \mathcal{A}^-} |\mathbf{u}_j/\sqrt{n}||\theta_{j'} + \mathbf{u}_{j'}/\sqrt{n}| - \sum_{j \in g \cap \mathcal{A}^-} (\theta_j)^2 - \sum_{j \neq j' \in g \cap \mathcal{A}^-} |\theta_j||\theta_{j'}| \right] \\
&:= \frac{1}{2} \sum_{g \in G} \Big[I_1 + I_2 + I_3 + I_4 + I_5 + I_6 + I_7 \Big],
\end{aligned}$$

where

$$\begin{aligned}
&I_1 = \sum_{j \in g \cap \mathcal{A}} (\mathbf{u}_j/\sqrt{n})^2, \quad I_2 = \sum_{j \neq j' \in g \cap \mathcal{A}} |\mathbf{u}_j/\sqrt{n}||\mathbf{u}_{j'}/\sqrt{n}|, \quad I_3 = \sum_{j \in g \cap \mathcal{A}^-} (\theta_j + \mathbf{u}_j/\sqrt{n})^2, \\
&I_4 = \sum_{j \neq j' \in g \cap \mathcal{A}^-} |\theta_j + \mathbf{u}_j/\sqrt{n}||\theta_{j'} + \mathbf{u}_{j'}/\sqrt{n}|, \quad I_5 = \sum_{j \in g \cap \mathcal{A}} \sum_{j' \in g \cap \mathcal{A}^-} |\mathbf{u}_j/\sqrt{n}||\theta_{j'} + \mathbf{u}_{j'}/\sqrt{n}|, \\
&I_6 = \sum_{j \in g \cap \mathcal{A}^-} -(\theta_j)^2, \quad \text{and} \quad I_7 = \sum_{j \neq j' \in g \cap \mathcal{A}^-} -|\theta_j||\theta_{j'}|.
\end{aligned}$$

Thus, $I_1 + I_2 \to_p 0$ and $I_3 + I_6$ follows that

$$
\begin{aligned}
&\lambda_n \Big[\sum_{j \in g \cap \mathcal{A}^-} (\theta_j + \mathbf{u}_j/\sqrt{n})^2 - \sum_{j \in g \cap \mathcal{A}^-} (\theta_j)^2 \Big] \\
&= \lambda_n \sum_{j \in g \cap \mathcal{A}^-} \Big[(\theta_j + \mathbf{u}_j/\sqrt{n})^2 - (\theta_j)^2 \Big] \\
&= \lambda_n \sum_{j \in g \cap \mathcal{A}^-} (2\theta_j + \mathbf{u}_j/\sqrt{n})(\mathbf{u}_j/\sqrt{n}) \\
&= \lambda_n \sum_{j \in g \cap \mathcal{A}^-} (2\theta_j)(\mathbf{u}_j/\sqrt{n}) + \lambda_n \sum_{j \in g \cap \mathcal{A}^-} (\mathbf{u}_j/\sqrt{n})^2 \\
&= \lambda_n/\sqrt{n} \sum_{j \in g \cap \mathcal{A}^-} (2\theta_j)(\mathbf{u}_j) + \lambda_n/\sqrt{n} \sum_{j \in g \cap \mathcal{A}^-} (\mathbf{u}_j^2)/\sqrt{n} \\
&\to \lambda_0 \sum_{j \in g \cap \mathcal{A}^-} (2\theta_j)(\mathbf{u}_j)
\end{aligned}
$$

Then $I_4 + I_7$ follows that

$$
\begin{aligned}
&\lambda_n \Big[\sum_{j \neq j' \in g \cap \mathcal{A}^-} |\theta_j + \mathbf{u}_j/\sqrt{n}||\theta_{j'} + \mathbf{u}_{j'}/\sqrt{n}| - \sum_{j \neq j' \in g \cap \mathcal{A}^-} |\theta_j||\theta_{j'}| \Big] \\
&= \frac{\lambda_n}{\sqrt{n}} \sum_{j \neq j' \in g \cap \mathcal{A}^-} |\sqrt{n}\theta_j + \mathbf{u}_j||\theta_{j'} + \mathbf{u}_{j'}/\sqrt{n}| - \lambda_n \sum_{j \neq j' \in g \cap \mathcal{A}^-} |\theta_j||\theta_{j'}| \\
&= \frac{\lambda_n}{\sqrt{n}} \sum_{j \neq j' \in g \cap \mathcal{A}^-} sign(\theta_j)(\sqrt{n}\theta_j + \mathbf{u}_j) sign(\theta_{j'})(\theta_{j'} + \mathbf{u}_{j'}/\sqrt{n}) - \lambda_n \sum_{j \neq j' \in g \cap \mathcal{A}^-} |\theta_j||\theta_{j'}| \\
&= \frac{\lambda_n}{\sqrt{n}} \sum_{j \neq j' \in g \cap \mathcal{A}^-} \Big[sign(\theta_j) sign(\theta_{j'})(\theta_j \mathbf{u}_{j'})(\theta_{j'} \mathbf{u}_j) + \mathbf{u}_j \mathbf{u}_{j'}/\sqrt{n} \Big] \\
&\to \lambda_0 \sum_{j \neq j' \in g \cap \mathcal{A}^-} |\theta_j \theta_{j'} \mathbf{u}_j \mathbf{u}_{j'}|
\end{aligned}
$$

Finally, I_5 converges

$$
\begin{aligned}
&\lambda_n \sum_{j \in g \cap \mathcal{A}} \sum_{j' \in g \cap \mathcal{A}^-} |\mathbf{u}_j/\sqrt{n}||\theta_{j'} + \mathbf{u}_{j'}/\sqrt{n}| \\
&= \lambda_n/\sqrt{n} \sum_{j \in g \cap \mathcal{A}} \sum_{j' \in g \cap \mathcal{A}^-} |\mathbf{u}_j| sign(\theta_j)(\theta_{j'} + \mathbf{u}_{j'}/\sqrt{n}) \\
&\to \lambda_0 \sum_{j \in g \cap \mathcal{A}} \sum_{j' \in g \cap \mathcal{A}^-} |\mathbf{u}_j \theta_{j'}|
\end{aligned}
$$

Thus, we have

$$\frac{\lambda_n}{2}\sum_{g\in G}\left[\left(\sum_{j\in g}|\theta_j+\mathbf{u}_j/\sqrt{n}|\right)^2-\left(\sum_{j\in g}|\theta_j|\right)^2\right]\rightarrow$$
$$\lambda_0\sum_{ginG}\left[\sum_{j\in g\cap\mathcal{A}^-}(2\theta_j)(\mathbf{u}_j)+\sum_{j\neq j'\in g\cap\mathcal{A}^-}|\theta_j\theta_{j'}\mathbf{u}_j\mathbf{u}_{j'}|+\sum_{j\in g\cap\mathcal{A}}\sum_{j'\in g\cap\mathcal{A}^-}|\mathbf{u}_j\theta_{j'}|\right]$$

Thus, $\mathbf{Z}_n(\mathbf{u})\rightarrow_d\mathbf{Z}(\mathbf{u})$ with the finite-dimensional convergence holding tirivially. Since $\mathbf{Z}_n$ is convex and $\mathbf{Z}$ has a unique minimum, it satisfies that (Geyer, 1996),

$$argmin(\mathbf{Z_n})=\sqrt{n}(\hat{\boldsymbol{\theta}}-\boldsymbol{\theta})\rightarrow_d argmin(\mathbf{Z}). \tag{2.27}$$

We note that if $\lambda_0=0$, then $argmin(\mathbf{Z})=\mathbf{C}^{-1}\mathbf{\Phi}\sim N(\mathbf{0},\sigma^2\mathbf{C}^{-1})$. ■

2.4 Simulation

We conduct a simulation study using continuous y_{it} with time in-variant independent covariates, $\mathbf{x_{it}}=\mathbf{x_i}$, which follows a multivariate standard normal distribution with a Toeplitz covariance matrix. We generate repeated continuous y_{it}'s given $\mathbf{x_i}$ from a marginal parametric generalized linear model (GLM) under an AR(1) correlation structure, where

$$y_{it}|\mathbf{x}_i\sim N(\mu_{it},1),\quad \mu_{it}=\mathbf{g}(\mathbf{x}_i^T\boldsymbol{\beta})=\mathbf{x}_i^T\boldsymbol{\beta}. \tag{2.28}$$

We consider three scenarios for the simulation studies: 1) the true parameter is non-zero at one index in each group and zero at the others. 2) the true parameter is non-zero at two indices in each group. 3) the true parameter may be non-zero at more than one index in each group. In each scenario, we examine the impact of varying sample sizes, n, on the performance of predictors measured in p dimensions. We consider sample sizes of either 100 or 300. The corresponding predictors are as follows: 1) For $n=100$, the predictors are $p=(50,100,200)$; 2) For $n=300$, the predictors are $p=(50,100,200)$; and 3) For $n=300$, the predictors are $p=(150,300,600)$. We set five groups, which are of equal or unequal size, denoted as $g=5$. Each dataset simulation is performed 100 times.

In each simulation study, the two main covariance design matrices are set to test the robustness of the Exclusive Lasso penalty regarding within- and between-group correlations while accounting for domain-based variables. All covariance design matrices are drawn from a multivariate normal distribution $N(0,\Sigma)$, where Σ is a Toeplitz covariance matrix with entries $\Sigma_{ij}=w^{|i-j|}$ for correlations among variables within groups, and $\Sigma_{lm}=b^{|l-m|}$ for correlations among variables between groups. Here, $(i,j)\in p$ represents the ith and jth components of $\mathbf{x}_i$, while $(l,m)\in g$ denotes the lth and mth groups. The correlation levels w and b are between 0 and 1. The covariance matrix in the first model sets the constant $w=0.6$ and

$b = 0.6$ to test moderate correlation within and between groups. The covariance matrix in the second model sets the high correlation within and between groups, $w = 0.9$ and $b = 0.9$.

For p variables, each variable is assigned a group index corresponding to one of five unique groups. When the groups are of equal size, the covariates are evenly split among all five groups. Conversely, for unequal-sized groups, half of the variables are shared among the five groups, while the other half is divided among three groups. The coefficient vectors used in the simulation analysis are structured to mimic the scenarios of high-dimensional grouped predictors. Thus, we examine three distinct scenarios based on the number of non-zero coefficients within each group, allowing us to evaluate how effectively the Exclusive Lasso addresses both homogeneous and heterogeneous sparsity structures in variable selection. These scenarios increase in complexity, ranging from a simple structure with a single non-zero coefficient per group to more complex situations with varying numbers of non-zero coefficients across groups. In the first scenario, exactly one non-zero coefficient is assigned to the first index in each group. Thus, the first variable in each group has a coefficient of 1, while the others in the same group are zero. An intercept of 0.2 is included. The parameter vectors are:

$$\boldsymbol{\beta} = \Big(0.2, \underbrace{1, 0, \ldots, 0}_{\text{g1}}, \underbrace{1, 0, \ldots, 0}_{\text{g2}}, \underbrace{1, 0, \ldots, 0}_{\text{g3}}, \underbrace{1, 0, \ldots, 0}_{\text{g4}}, \underbrace{1, 0, \ldots, 0}_{\text{g5}}\Big)^T$$

In the second scenario, two covariates per group are significant, giving ten non-zero coefficients, each set equal to 1. All remaining coefficients are zero:

$$\boldsymbol{\beta} = \Big(0.2, \underbrace{1, 1, 0, \ldots, 0}_{\text{g1}}, \underbrace{1, 1, 0, \ldots, 0}_{\text{g2}}, \underbrace{1, 1, 0, \ldots, 0}_{\text{g3}}, \underbrace{1, 1, 0, \ldots, 0}_{\text{g4}}, \underbrace{1, 1, 0, \ldots, 0}_{\text{g5}}\Big)^T$$

In this third scenario, we assign three significant coefficients to group 1, two non-zero coefficients to group 2, and one non-zero coefficient each to groups 3, 4, and 5. Each non-zero coefficient equals 1:

$$\boldsymbol{\beta} = \Big(0.2, \underbrace{1, 1, 1, 0, \ldots, 0}_{\text{g1}}, \underbrace{1, 1, 0, \ldots, 0}_{\text{g2}}, \underbrace{1, 0, \ldots, 0}_{\text{g3}}, \underbrace{1, 0, \ldots, 0}_{\text{g4}}, \underbrace{1, 0, \ldots, 0}_{\text{g5}}\Big)^T$$

In every scenario, the regularization parameter λ is selected as $\lambda = \max_i |\mathbf{x}_i^T y|$, which is large enough to ensure that the correct structure is estimated.

The missing-data module consists of one-step transition probabilities p_{i2} and p_{i3} modeled by logistic regression 2.6. We specify $\xi_{01} = -2$, $\xi_{x1} = (0.5, 0..., 0)$, $\xi_{y1} = 1$, $\xi_{02} = -2$, $\xi_{x2} = (-0.5, 0, ..., 0)$, $\xi_{y2} = (0.5, 1)$. This leads to a missing proportion of 25% to 30% at $t = 2(= 3)$ for all three regression models in 2.28.

To conduct the comparison over our method (PWGEE-eLasso), we implement Monte Carlo (MC) replicating the experiment 100 times and consider the PWGEE method (Kowalski & Tu, 2008) with different penalties. We compared our method with the Penalized WGEE with Lasso (PWGEE-Lasso), SCAD (PWGEE-SCAD), Minimax Concave Penalty (PWGEE-MCP), and composite MCP (PWGEE-

cMCP). Mean squared error, MSE=$\| \hat{\boldsymbol{\theta}}_p - \boldsymbol{\theta} \|^2$, is used to measure the performance of point estimates from PWGEE-eLasso model selection, where $\boldsymbol{\theta}_p$ denotes the penalized estimate from PWGEE-eLasso, and $\boldsymbol{\theta}$ denotes the empirical estimate. In addition, to measure the performance of selection methods, we compute a set of metrics denoted by (C, IC, CG), where C is the number of zero coefficients that are correctly estimated by zero, IC is the number of non-zero coefficients that are incorrectly estimated by zero, and CG is the number of groups that are correctly selected.

2.5 Results

We consider the performance of the variable selections in a high-dimensional longitudinal study with missing information. We set the correlation as $w = b = 0.6$, or $w = b = 0.9$. The sample size is set as $n = (100, 300)$, and the number of predictors is set as $p = (n/2, n, 2n)$ in each scenario, respectively. In addition, we study another setup in which $p = (50, 100, 200)$ when $n = 300$. Results provide a detailed overview of the experimental outcomes, showcasing the impact of various factors on the performance metrics, including C (the number of zero coefficients that are correctly estimated by zero), IC (the number of non-zero coefficients that are incorrectly estimated by zero), GC (the number of groups that are correctly selected), and MSE (Mean Squared Error).

In scenario 1, we assess the estimation accuracy performance and the performance of selecting the first five important predictors in each group. We compare the results of our proposed PWGEE-Exclusive Lasso (P-eLasso) method to those of PWGEE-Lasso(P-Lasso), PWGEE-SCAD (P-SCAD), PWGEE-MCP (P-MCP) and PWGEE-cMCP (P-cMCP). In Scenario 2, we explore variable selection using data that include two significant variables within each equally sized group. Similarly, in scenario 3, data include more than one significant variable within each equally sized group.

Table 2.1 shows that P-eLasso consistently balances accurately identifying true zero coefficients (C) and maintaining non-zero coefficients (low IC), resulting in the lowest MSE across all examined scenarios. For instance, in Scenario 1, with correlations of 0.60 and $p = 200$, P-eLasso achieves $C = 183.79$, $IC = 0.00$, and $MSE = 0.10$. This performance outperforms alternatives like P-SCAD and P-MCP, which have higher IC and larger MSE values. Even in more challenging high-correlation scenarios ($corr = 0.90$), P-eLasso continues to outperform other methods. In Scenario 2, where the correlations are 0.90 and $p = 200$, it shows $C = 180.39$, $IC = 2.20$, and $MSE = 0.86$; in contrast, other methods show significantly higher MSE values, with some reaching as high as 9.62. While P-Lasso frequently identifies many non-zero coefficients, it tends to produce denser solutions (resulting in a lower C) when encountering strong correlations. On the other hand, methods based on SCAD, MCP, and cMCP may yield sparser solutions but carry the risk of inaccurately shrinking true signals, which can result in inflated MSE. Overall, P-eLasso effectively balances the accurate exclusion of zero coefficients (C) with the precise retention of non-zero coefficients (IC), enhancing its superior selection properties.

Table 2.1: C (the number of zero coefficients that are correctly estimated by zero), IC (the number of non-zero coefficients that are incorrectly estimated by zero), GC (the number of groups that are correctly selected), and MSE (Mean Squared Error) for PWGEE-Lasso(P-Lasso), PWGEE-SCAD (P-SCAD), PWGEE-MCP (P-MCP), PWGEE-cMCP (P-cMCP), and PWGEE-Exclusive Lasso (P-eLasso) method with n = 100 and p =(50, 100, 200) in each scenario.

scenarios	corr	methods	C	IC	CG	MSE	C	IC	CG	MSE	C	IC	CG	MSE
				p=50				p=100				p=200		
1	0.60	P-Lasso	31.44	0.02	**5.00**	0.17	67.40	**0.00**	**5.00**	0.20	149.63	0.00	**5.00**	0.24
		P-SCAD	32.08	0.19	4.98	0.40	67.52	0.15	**5.00**	0.75	151.78	0.20	**5.00**	1.04
		P-MCP	35.30	0.25	4.94	0.42	74.31	0.21	4.99	0.77	159.22	0.34	**5.00**	1.56
		P-cMCP	32.98	0.21	4.94	0.31	69.27	0.14	4.99	0.43	148.32	0.15	**5.00**	0.81
		P-eLasso	**38.53**	**0.01**	**5.00**	**0.10**	**87.43**	**0.00**	**5.00**	**0.09**	**183.79**	**0.00**	**5.00**	**0.10**
	0.90	P-Lasso	37.01	0.65	4.87	0.64	81.47	0.77	4.91	0.60	176.04	0.76	4.99	0.55
		P-SCAD	35.56	2.29	4.72	2.92	72.77	2.11	4.92	5.93	176.40	2.72	4.91	2.73
		P-MCP	36.60	2.24	4.62	2.83	75.73	2.12	4.88	5.53	178.52	2.61	4.84	3.69
		P-cMCP	35.68	2.28	4.32	2.69	73.05	2.13	4.87	5.69	177.95	2.66	4.73	2.72
		P-eLasso	**39.17**	**0.28**	**5.00**	**0.34**	**87.81**	**0.17**	**5.00**	**0.28**	**187.61**	**0.09**	**5.00**	**0.19**
2	0.60	P-Lasso	27.73	0.26	**5.00**	0.31	63.25	0.32	**5.00**	0.39	139.54	0.33	**5.00**	0.77
		P-SCAD	28.11	2.09	4.99	0.80	62.17	2.26	**5.00**	1.52	145.14	2.92	**5.00**	2.43
		P-MCP	30.89	2.57	4.91	0.81	68.32	2.70	**5.00**	1.61	149.11	3.20	**5.00**	3.41
		P-cMCP	29.01	2.23	4.88	0.71	64.47	2.07	**5.00**	1.07	141.92	2.65	**5.00**	2.24
		P-eLasso	**32.01**	**0.17**	**5.00**	**0.24**	**80.27**	**0.28**	**5.00**	**0.25**	**168.47**	**0.23**	**5.00**	**0.52**
	0.90	P-Lasso	33.15	3.25	4.86	1.34	76.22	3.26	4.97	1.39	168.29	3.20	4.99	5.50
		P-SCAD	31.35	5.93	4.70	5.52	68.75	5.96	4.95	11.71	170.93	6.83	4.93	7.53
		P-MCP	31.93	5.95	4.64	5.54	71.68	6.00	4.85	10.58	173.25	6.82	4.89	9.62
		P-cMCP	31.56	6.11	4.09	5.53	69.45	6.00	4.78	11.09	173.01	6.83	4.75	6.51
		P-eLasso	**33.60**	**2.35**	**5.00**	**1.12**	**80.81**	**2.48**	**5.00**	**1.02**	**180.39**	**2.20**	**5.00**	**0.86**
3	0.60	P-Lasso	28.68	**0.11**	**5.00**	0.25	64.64	**0.19**	**5.00**	0.32	143.30	**0.15**	**5.00**	**0.48**
		P-SCAD	29.55	0.77	4.99	0.57	63.42	1.02	4.99	1.21	147.69	1.47	**5.00**	1.64
		P-MCP	**32.72**	1.01	4.98	0.58	70.48	1.31	**5.00**	1.20	152.85	1.76	**5.00**	2.62
		P-cMCP	30.39	0.73	4.92	0.47	65.90	0.92	4.99	0.75	145.00	1.30	**5.00**	1.52
		P-eLasso	31.80	0.16	**5.00**	**0.23**	**76.37**	0.20	**5.00**	**0.26**	**167.57**	0.21	**5.00**	0.50
	0.90	P-Lasso	34.88	2.20	4.82	0.96	77.51	1.99	4.94	1.12	169.93	2.20	**5.00**	3.77
		P-SCAD	32.92	4.76	4.63	4.37	71.14	4.35	4.94	8.08	173.22	5.04	4.90	6.06
		P-MCP	33.15	4.70	4.53	4.44	71.95	4.36	4.96	8.19	175.59	5.10	4.84	7.16
		P-cMCP	33.54	4.53	4.09	4.01	71.31	4.30	4.75	8.27	175.18	5.00	4.82	4.68
		P-eLasso	**35.22**	**1.67**	**5.00**	**0.77**	**82.98**	**1.55**	**5.00**	**0.73**	**180.76**	**1.62**	**5.00**	**0.72**

Note: Bold symbols in the C and CG indicate the highest values, while the lowest values for IC and MSE are also bolded. The optimal values of C vary by scenario and are determined by the (p - number of non-zeros). Thus, for p = 50, C = (45, 40, 42); for p = 100, C = (95, 90, 92); for p = 200, C = (195, 190 192). The optimal value for IC is zero.

Table 2.2: C (the number of zero coefficients that are correctly estimated by zero), IC (the number of non-zero coefficients that are incorrectly estimated by zero), GC (the number of groups that are correctly selected), and MSE (Mean Squared Error) for PWGEE-Lasso(P-Lasso), PWGEE-SCAD (P-SCAD), PWGEE-MCP (P-MCP), PWGEE-cMCP (P-cMCP), and PWGEE-Exclusive Lasso (P-eLasso) method with n = 300 and p =(50, 100, 200) in each scenario.

scenarios	corr	methods	C	IC	CG	MSE	C	IC	CG	MSE	C	IC	CG	MSE
				p=50				p=100				p=200		
1	0.60	P-Lasso	28.19	**0.00**	**5.00**	0.11	66.53	**0.00**	**5.00**	0.10	144.15	**0.00**	**5.00**	0.11
		P-SCAD	31.12	**0.00**	**5.00**	0.17	68.82	**0.00**	**5.00**	0.20	144.29	**0.00**	**5.00**	0.31
		P-MCP	**34.47**	0.01	**5.00**	0.18	76.21	**0.00**	**5.00**	0.22	159.51	**0.00**	**5.00**	0.36
		P-cMCP	31.11	0.02	**5.00**	0.14	69.91	**0.00**	**5.00**	0.13	149.25	**0.00**	**5.00**	0.15
		P-eLasso	32.88	**0.00**	**5.00**	**0.09**	**81.16**	**0.00**	**5.00**	**0.07**	**181.32**	**0.00**	**5.00**	**0.05**
	0.90	P-Lasso	30.64	0.30	4.98	0.74	74.46	0.40	4.99	0.51	164.83	0.19	**5.00**	0.38
		P-SCAD	33.16	1.39	4.91	1.90	75.18	1.50	4.98	2.06	161.50	1.44	**5.00**	2.52
		P-MCP	**33.90**	1.34	4.88	1.85	77.21	1.44	4.94	2.01	166.12	1.49	4.99	2.54
		P-cMCP	33.21	1.46	4.59	1.82	74.68	1.51	4.89	1.96	157.95	1.33	**5.00**	2.42
		P-eLasso	31.37	**0.13**	**5.00**	**0.65**	**79.16**	**0.20**	**5.00**	**0.39**	**178.44**	**0.04**	**5.00**	**0.21**
2	0.60	P-Lasso	24.70	0.11	**5.00**	0.22	62.05	**0.03**	**5.00**	0.18	140.40	**0.00**	**5.00**	0.20
		P-SCAD	27.41	0.46	4.99	0.35	63.46	0.20	**5.00**	0.37	140.16	0.19	**5.00**	0.52
		P-MCP	**30.29**	0.60	4.99	0.35	71.65	0.32	**5.00**	0.38	155.53	0.36	**5.00**	0.56
		P-cMCP	26.70	0.49	4.99	0.35	65.07	0.22	**5.00**	0.31	144.53	0.22	**5.00**	0.36
		P-eLasso	26.80	**0.08**	**5.00**	**0.20**	**73.16**	**0.03**	**5.00**	**0.14**	**171.41**	0.01	**5.00**	**0.14**
	0.90	P-Lasso	27.04	2.61	4.99	1.59	71.11	2.32	**5.00**	1.13	160.49	1.95	**5.00**	0.91
		P-SCAD	29.00	4.89	4.86	3.91	71.63	5.07	4.98	3.96	156.18	5.18	4.99	5.11
		P-MCP	**29.58**	4.86	4.84	3.77	**73.27**	5.19	4.93	3.97	162.14	5.33	**5.00**	4.76
		P-cMCP	29.01	4.99	4.56	3.79	72.07	5.16	4.80	3.91	155.20	5.12	4.99	5.34
		P-eLasso	26.24	**2.03**	**5.00**	**1.55**	73.18	**1.70**	**5.00**	**0.99**	**170.15**	**1.61**	**5.00**	**0.70**
3	0.60	P-Lasso	25.19	**0.00**	**5.00**	**0.16**	62.81	**0.00**	**5.00**	**0.14**	140.98	**0.01**	**5.00**	0.16
		P-SCAD	28.12	0.08	**5.00**	0.26	64.97	0.03	**5.00**	0.28	141.95	0.02	**5.00**	0.40
		P-MCP	**31.14**	0.13	**5.00**	0.27	**72.71**	0.06	**5.00**	0.30	156.99	0.05	**5.00**	0.45
		P-cMCP	27.63	0.14	**5.00**	0.23	66.36	0.04	4.99	0.21	145.26	**0.01**	**5.00**	0.24
		P-eLasso	25.83	**0.00**	**5.00**	0.17	72.42	0.02	**5.00**	**0.14**	**169.63**	0.02	**5.00**	**0.13**
	0.90	P-Lasso	28.67	1.63	4.97	**1.26**	70.95	1.41	**5.00**	1.01	161.46	1.16	4.99	0.69
		P-SCAD	30.86	3.54	4.87	2.99	72.73	3.58	4.98	3.41	158.21	3.61	4.99	4.03
		P-MCP	**31.14**	3.54	4.85	3.00	**74.39**	3.58	4.97	3.34	162.23	3.65	4.98	4.01
		P-cMCP	30.26	3.39	4.53	3.01	72.39	3.63	4.85	3.32	157.46	3.52	4.97	4.05
		P-eLasso	26.86	**1.30**	**5.00**	1.33	71.80	**0.99**	**5.00**	**0.93**	**172.18**	**1.07**	**5.00**	**0.56**

Note: Bold symbols in the C and CG indicate the highest values, while the lowest values for IC and MSE are also bolded. The optimal values of C vary by scenario and are determined by the (p - number of non-zeros). Thus, for p = 50, C = (45, 40, 42); for p = 100, C = (95, 90, 92); for p = 200, C = (195, 190 192). The optimal value for IC is zero.

Table 2.3: C (the number of zero coefficients that are correctly estimated by zero), IC (the number of non-zero coefficients that are incorrectly estimated by zero), GC (the number of groups that are correctly selected), and MSE (Mean Squared Error) for PWGEE-Lasso(P-Lasso), PWGEE-SCAD (P-SCAD), PWGEE-MCP (P-MCP), PWGEE-cMCP (P-cMCP), and PWGEE-Exclusive Lasso (P-eLasso) method with n = 300 and p =(150, 300, 600) in each scenario.

scenarios	corr	methods	C	IC	CG	MSE	C	IC	CG	MSE	C	IC	CG	MSE
			p=150				p=300				p=600			
1	0.60	P-Lasso	104.82	**0.00**	**5.00**	0.10	225.28	**0.00**	**5.00**	0.13	527.72	**0.00**	**5.00**	0.09
		P-SCAD	106.43	**0.00**	**5.00**	0.24	220.40	**0.00**	**5.00**	0.47	529.51	**0.00**	**5.00**	0.13
		P-MCP	117.49	**0.00**	**5.00**	0.29	243.35	**0.00**	**5.00**	0.52	546.84	**0.00**	**5.00**	0.24
		P-cMCP	108.95	**0.00**	**5.00**	0.14	230.66	**0.00**	**5.00**	0.18	533.16	**0.00**	**5.00**	0.14
		P-eLasso	**130.07**	**0.00**	**5.00**	**0.06**	**280.49**	**0.00**	**5.00**	**0.05**	**581.11**	**0.00**	**5.00**	**0.05**
	0.90	P-Lasso	119.53	0.21	4.99	0.40	256.14	0.23	**5.00**	0.37	566.13	0.26	**5.00**	0.30
		P-SCAD	117.55	1.40	4.97	2.42	244.71	1.47	**5.00**	3.46	567.15	2.64	4.99	1.20
		P-MCP	121.88	1.45	4.97	2.24	254.46	1.59	**5.00**	3.11	576.47	2.65	4.98	1.24
		P-cMCP	116.02	1.33	4.94	2.18	239.62	1.39	**5.00**	3.41	572.53	2.64	4.98	1.14
		P-eLasso	**130.93**	**0.04**	**5.00**	**0.21**	**279.32**	**0.05**	**5.00**	**0.16**	**577.36**	**0.05**	**5.00**	**0.13**
2	0.60	P-Lasso	101.72	**0.02**	**5.00**	0.19	219.28	**0.01**	**5.00**	0.21	511.11	**0.03**	**5.00**	0.19
		P-SCAD	103.05	0.26	**5.00**	0.41	217.50	0.23	**5.00**	0.75	515.25	0.72	**5.00**	0.42
		P-MCP	114.01	0.40	**5.00**	0.45	238.21	0.34	**5.00**	0.82	535.86	0.96	**5.00**	0.64
		P-cMCP	105.32	0.30	**5.00**	0.34	222.95	0.23	**5.00**	0.41	519.22	0.62	**5.00**	0.44
		P-eLasso	**122.66**	0.03	**5.00**	**0.14**	**272.12**	0.03	**5.00**	**0.12**	**570.57**	0.07	**5.00**	**0.13**
	0.90	P-Lasso	115.91	2.17	**5.00**	1.06	254.32	1.90	**5.00**	0.85	563.01	1.85	**5.00**	0.65
		P-SCAD	112.78	5.11	**5.00**	4.89	242.05	5.22	**5.00**	6.53	565.43	6.58	4.98	2.38
		P-MCP	116.81	5.34	4.95	4.61	252.21	5.40	**5.00**	5.85	**573.55**	6.70	4.92	2.32
		P-cMCP	113.63	5.24	4.86	4.78	238.37	5.05	**5.00**	7.22	571.55	6.63	4.92	2.18
		P-eLasso	**121.48**	**1.70**	**5.00**	**0.84**	**273.09**	**1.49**	**5.00**	**0.56**	573.31	**1.36**	**5.00**	**0.49**
3	0.60	P-Lasso	103.10	**0.0**	**5.00**	0.15	218.47	**0.0**	**5.00**	0.17	517.67	**0.00**	**5.00**	0.14
		P-SCAD	104.65	0.03	**5.00**	0.29	218.41	0.03	**5.00**	0.59	520.96	0.09	**5.00**	0.21
		P-MCP	115.65	0.05	**5.00**	0.33	238.82	0.05	**5.00**	0.66	540.71	0.15	**5.00**	0.36
		P-cMCP	107.07	0.03	**5.00**	0.21	224.72	0.01	**5.00**	0.28	524.74	0.08	**5.00**	0.25
		P-eLasso	**120.20**	**0.02**	**5.00**	**0.13**	**268.87**	0.01	**5.00**	**0.13**	**567.09**	0.02	**5.00**	**0.12**
	0.90	P-Lasso	117.50	1.39	**5.00**	0.78	253.25	1.08	**5.00**	0.66	563.48	1.15	**5.00**	0.55
		P-SCAD	115.97	3.52	4.99	3.47	240.67	3.43	**5.00**	5.68	566.05	4.90	4.96	1.91
		P-MCP	118.71	3.57	4.98	3.46	249.84	3.64	**5.00**	5.08	**574.52**	4.95	4.89	1.88
		P-cMCP	116.11	3.47	4.92	3.32	238.22	3.40	4.99	5.83	571.70	4.98	4.95	1.81
		P-eLasso	**123.18**	**1.17**	**5.00**	**0.66**	**272.71**	**1.05**	**5.00**	**0.48**	571.31	**0.87**	**5.00**	**0.42**

Note: Bold symbols in the C and CG indicate the highest values, while the lowest values for IC and MSE are also bolded. The optimal values of C vary by scenario and are determined by the (p - number of non-zeros). Thus, for p = 150, C = (145, 140, 142); for p = 300, C = (295, 290, 292); for p = 600, C = (595, 590 592). The optimal value for IC is zero.

As the sample size increases to $n = 300$ in Table 2.2, all methods generally show improved estimation accuracy (lower MSE) and better variable selection outcomes (higher C and lower IC). This improvement highlights the advantages of increased data for more stable estimates. Notably, P-eLasso continues to excel in balancing the correct exclusion of zero coefficients (C) while preserving non-zero signals (IC), especially under moderate correlation ($w = b = 0.60$). For example, in Scenario 1, when $p = 200$ and $corr = 0.60$, the MSE of P-eLasso decreases from around 0.10 in the $n = 100$ column to as low as 0.05 in the $n = 300$ column. Even under higher correlation ($corr = 0.90$) in Scenarios 2 and 3, P-eLasso still achieves significantly smaller MSE compared to the other methods. While P-Lasso also benefits from the larger sample size, showing zero misclassifications ($IC = 0$) in certain low-correlation conditions, its performance in high-correlation or more complex signal structures (Scenarios 2 and 3) is less consistent compared to P-eLasso. Methods using SCAD, MCP, and cMCP also show similar improvements with a larger sample size, but they are still more likely to incorrectly shrink true signals in highly correlated situations, as indicated by higher IC and MSE. As n increases, the performance gap narrows somewhat. However, P-eLasso consistently retains an advantage due to its ability to identify parameters for both individual coefficients and entire groups with minimal bias and variance.

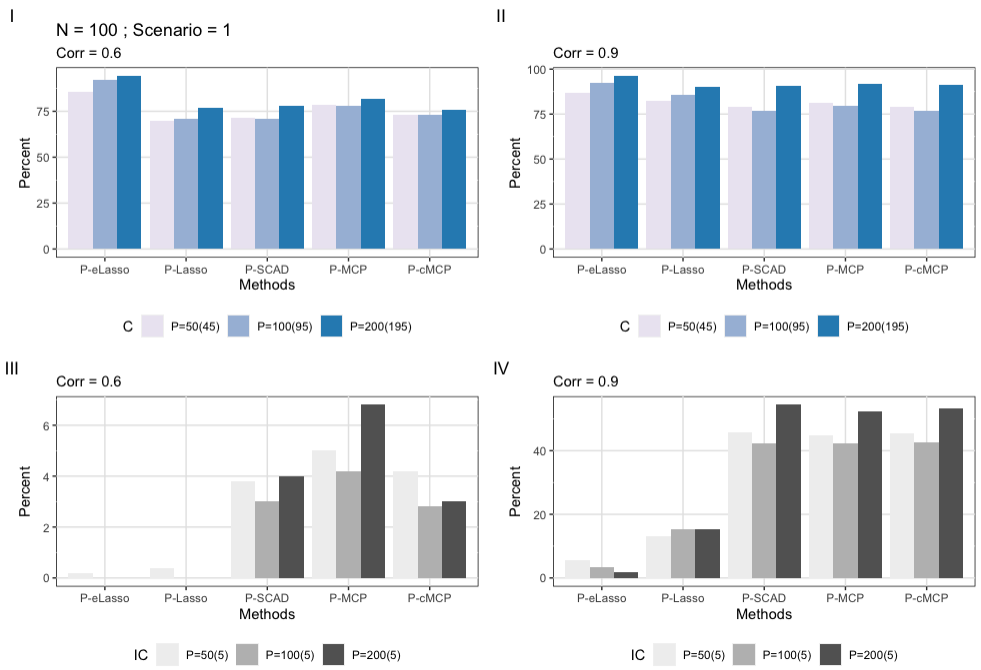
In Table 2.3, we observe that increasing the dimensionality while maintaining $n = 300$ results in a more challenging selection problem. Nevertheless, P-eLasso consistently demonstrates strong performance across all scenarios. In moderate-correlation settings ($corr = 0.60$), all methods generally retain or improve their ability to correctly exclude zero coefficients (C) compared to Table 2.2. P-eLasso notably achieves the lowest MSE by combining a high C with an almost zero IC. For example, in Scenario 1 with $p = 600$, it reaches a C of 581.11 and an IC of 0.00, resulting in an MSE of only 0.05. In contrast, other methods occasionally approach a zero IC under a moderate correlation (for example, $corr = 0.60$ for P-SCAD at $p = 150$), but they generally show a larger MSE than P-eLasso. When the correlation is high ($corr = 0.90$), it becomes more challenging to distinguish between signals. P-Lasso maintains decent performance in terms of C (often above 550 at $p = 600$), yet it still shows a slightly larger MSE compared to P-eLasso. P-SCAD, P-MCP, and P-cMCP tend to inaccurately set non-zero coefficients to zero, increasing IC and causing elevated MSE in cases of strong correlation and large p. P-eLasso consistently shows low IC values (0.04 to 1.70) and smaller MSE in Scenarios 2 and 3, even for $p = 600$. This pattern is consistent with the results for $p = 50, 100, 200$, but the gap in MSE becomes more pronounced at higher dimensionalities, especially under a strong correlation, emphasizing P-eLasso's advantage in preserving true signals and effectively excluding irrelevant features as p increases.

Overall, these results demonstrate that P-eLasso consistently achieves the best balance between a high C and a low IC, leading to the smallest MSE. While P-Lasso identifies many non-zero coefficients, its solutions are generally denser, which results in a slightly lower C. In contrast, P-SCAD, P-MCP, and P-cMCP provide sparser solutions than P-Lasso; however, they risk over-penalizing real signals, which can lead to higher IC and inflated MSE, especially in conditions of strong correlation or when the number of predictors is large. P-eLasso shows outstanding performance across various sample sizes and dimensions because it effectively reduces irrelevant predictors while preserving truly non-zero coefficients. This strength becomes particularly noticeable even with rising correlations and an increasing number of pre-

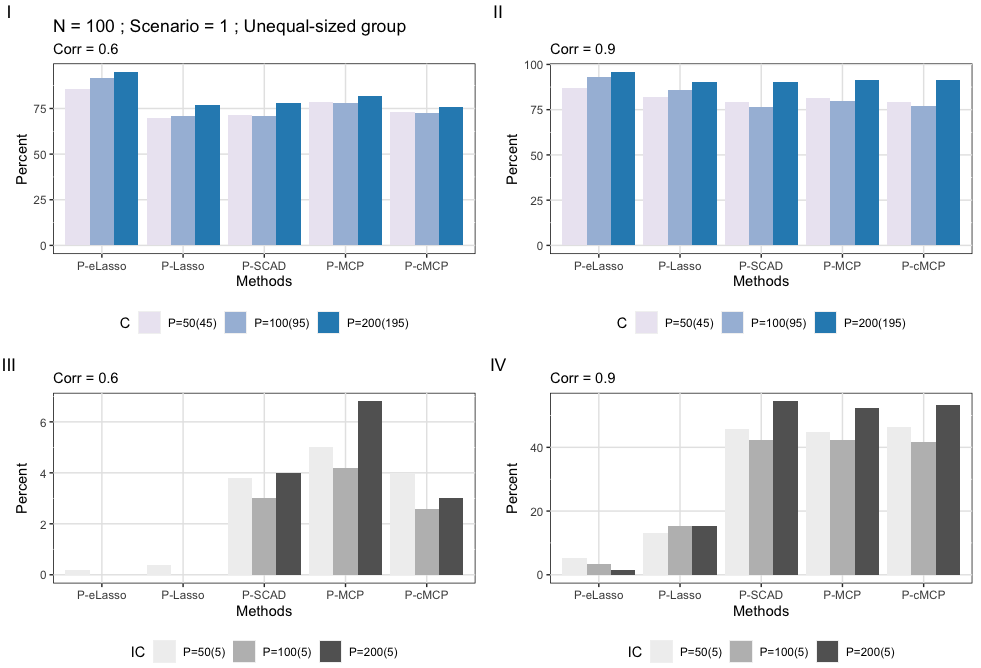
dictors. Moreover, the results in the Appendix tables vary from those shown in the main tables, as they focus on scenarios involving unequal group sizes instead of the equal-sized groups discussed in the main analysis. Despite these differences, performance patterns remain consistent (see Appendix Tables A.2, A.3, and A.3), reinforcing the robustness of P-eLasso regardless of group size configuration.

Figure 2.1, 2.2, and 2.3 summarizes the results from Tables 2.1 and A.1 (see the Appendix) for $n = 100$, illustrating the performance of P-Lasso, P-SCAD, P-MCP, P-cMCP, and P-eLasso under varying correlation levels ($Corr = 0.6$ and 0.9) and group structures (equal-sized vs. unequal-sized). Each figure corresponds to a different scenario (1-3), while the left and right columns represent varying correlation levels ($Corr = 0.6$ or 0.9). The upper row of the panels displays the percentage of correctly identified zero coefficients (C), while the lower row presents the percentage of incorrectly excluded true signals (IC) across various predictor dimensions, $p = (50, 100200)$. The numbers of true zero and non-zero coefficients are indicated in parentheses with p for C and IC, respectively. P-eLasso consistently achieves the highest C and the lowest IC, regardless of correlation or group size, demonstrating its robustness in identifying relevant predictors while excluding irrelevant ones. In contrast, Lasso and SCAD, along with MCP and cMCP, sometimes struggle with higher correlations or unequal-sized groups, as shown by their larger IC and lower C values.

Figure 2.4, 2.5, and 2.6 summarize results from Tables 2.2, A.2 (detailed in the Appendix). These figures also present both equal-sized and unequal-sized group structures in each scenario for $n = 300$ and ($p = 50, 100200$). Additionally, Figures 2.7, 2.8, and 2.9, based on Tables 2.3 and A.3, provide further information by increasing dimensionality for $p = (50, 100, 200)$ with $n = 300$. Across all scenarios and correlation levels, P-eLasso consistently exhibits higher C and lower IC than the compared methods, indicating lower selection errors and more accurate exclusion of irrelevant predictors. Conversely, P-Lasso usually detects numerous true signals but yields a lower C value, whereas P-SCAD, P-MCP, and P-cMCP might achieve sparser solutions at the risk of ignoring truly significant predictors, resulting in a higher IC. These trends hold true for both equal-sized and imbalanced groups, although the performance gap may vary slightly depending on the correlation strength and scenario. Overall, the bar plots in P-eLasso illustrate robustness in accurately differentiating between zero and non-zero coefficients, particularly in challenging high-correlation or high-dimensional settings.

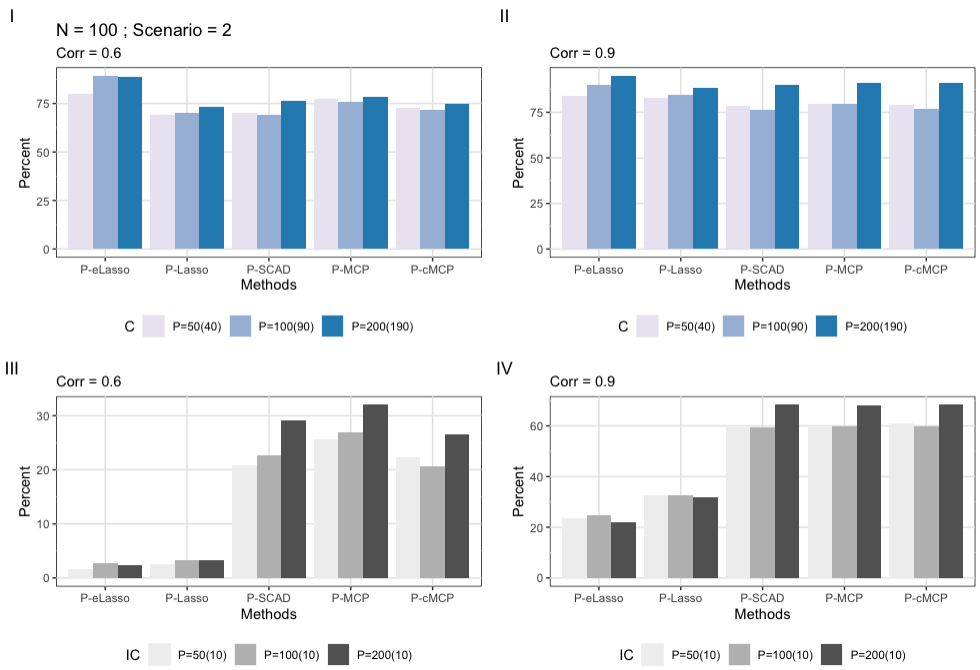


(a) Equal-Sized Group in Scenario 1, n=100, p = (50, 100, 200)

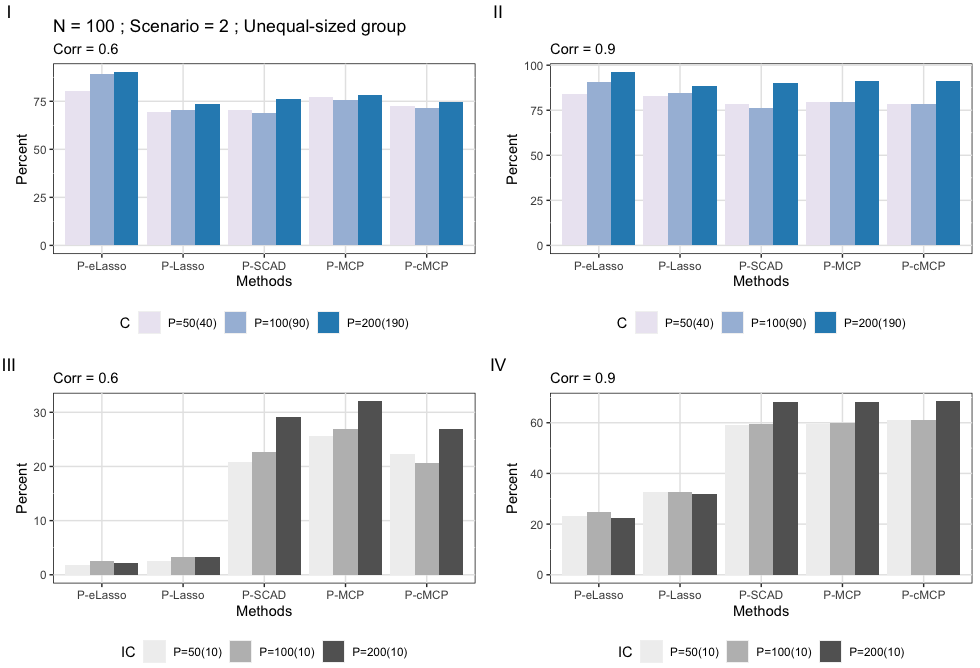


(b) Unequal-Sized Group in Scenario 1, n=100, p = (50, 100, 200)

Figure 2.1: Penalized Weighted Generalized Estimating Equations with various penalties: 1) Exclusive Lasso (P-eLasso), 2) Lasso (P-Lasso), 3) SCAD (P-SCAD), 4) MCP (P-MCP), and 5) Composite MCP (P-cMCP) tested in Scenario 1 with different data settings. Corr = 0.60 (Left), 0.90 (Right); n = 100; p = (50, 100, 200). The number of true zero and non-zero coefficients are indicated next to *p* for C and IC, respectively.

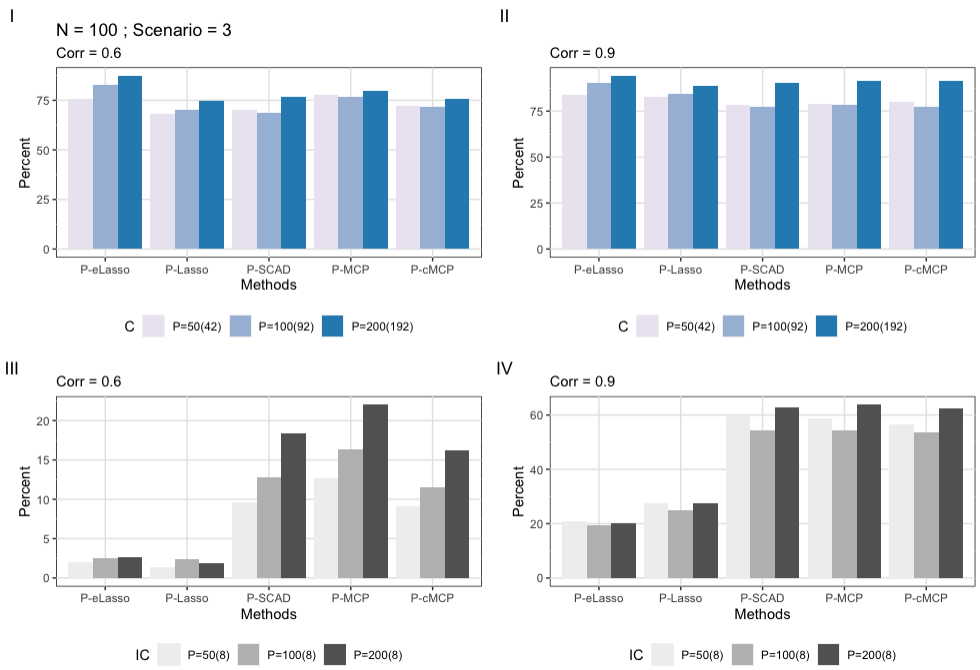


(a) Equal-Sized Group in Scenario 2, n=100, p = (50, 100, 200)

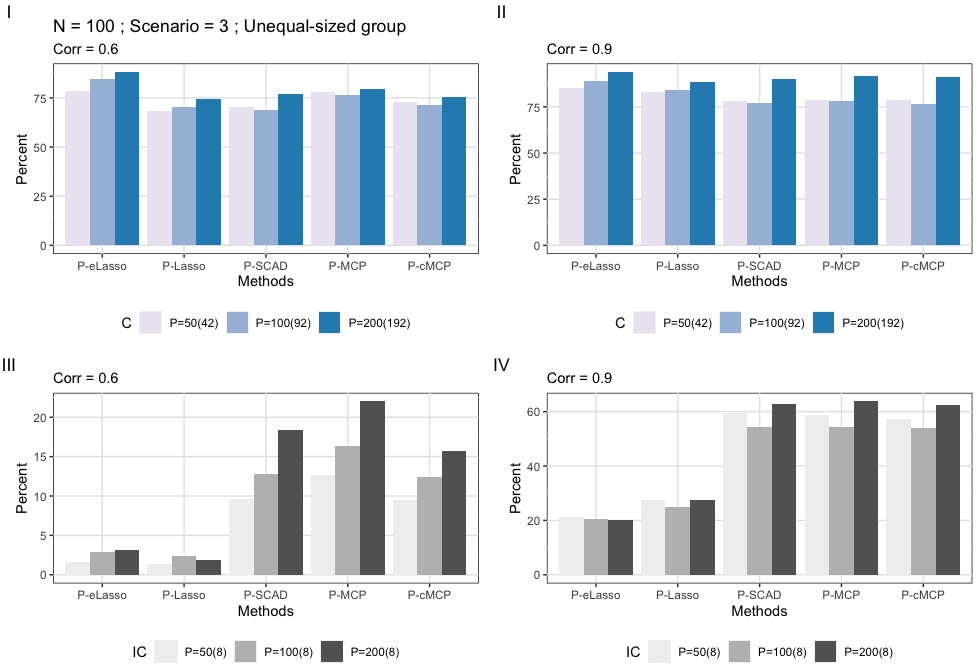


(b) Unequal-Sized Group in Scenario 2, n=100, p = (50, 100, 200)

Figure 2.2: Penalized Weighted Generalized Estimating Equations with various penalties: 1) Exclusive Lasso (P-eLasso), 2) Lasso (P-Lasso), 3) SCAD (P-SCAD), 4) MCP (P-MCP), and 5) Composite MCP (P-cMCP) tested in Scenario 2 with different data settings. Corr = 0.60 (Left), 0.90 (Right); n = 100; p = (50, 100, 200). The number of true zero and non-zero coefficients are indicated next to *p* for C and IC, respectively.

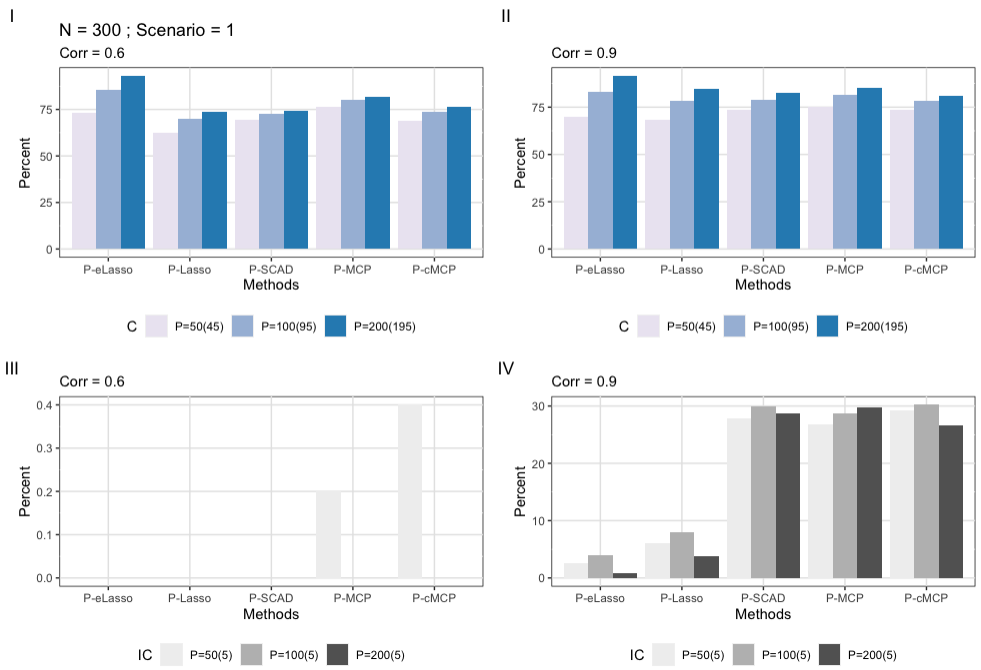


(a) Equal-Sized Group in Scenario 3, n=100, p = (50, 100, 200)

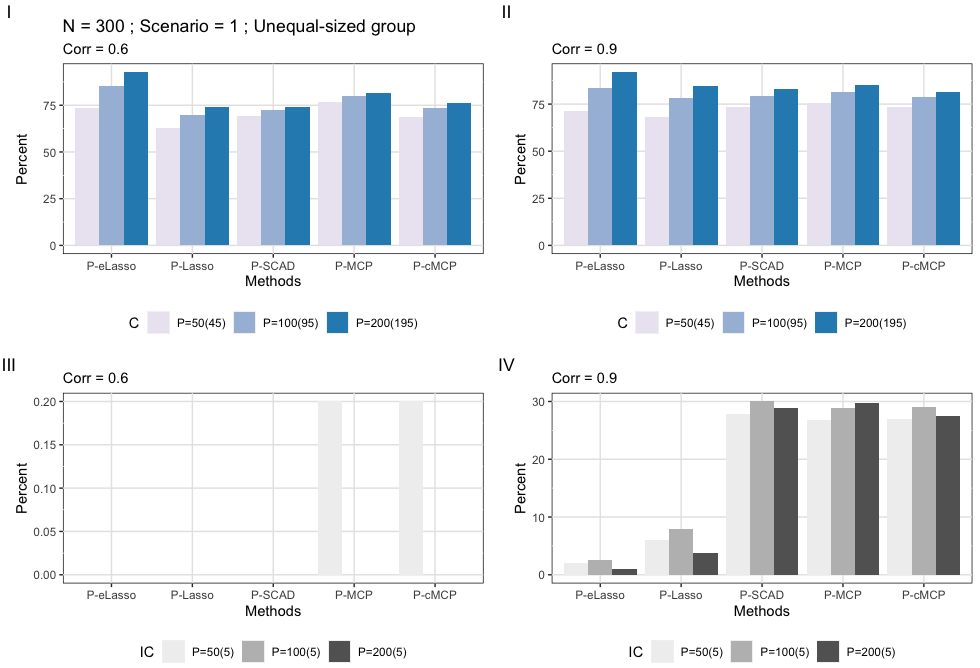


(b) Unequal-Sized Group in Scenario 3, n=100, p = (50, 100, 200)

Figure 2.3: Penalized Weighted Generalized Estimating Equations with various penalties: 1) Exclusive Lasso (P-eLasso), 2) Lasso (P-Lasso), 3) SCAD (P-SCAD), 4) MCP (P-MCP), and 5) Composite MCP (P-cMCP) tested in Scenario 3 with different data settings. Corr = (0.60 (Left), 0.90 (Right)); n = 100; p = (50, 100, 200). The number of true zero and non-zero coefficients are indicated next to *p* for C and IC, respectively.

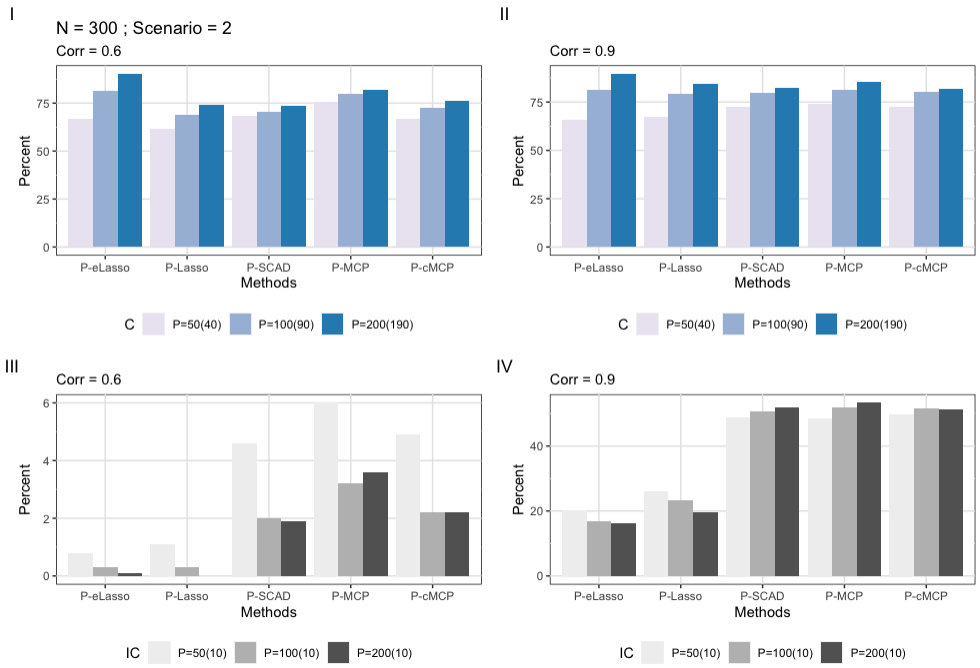


(a) Equal-Sized Group in Scenario 1, n=300, p = (50, 100, 200)

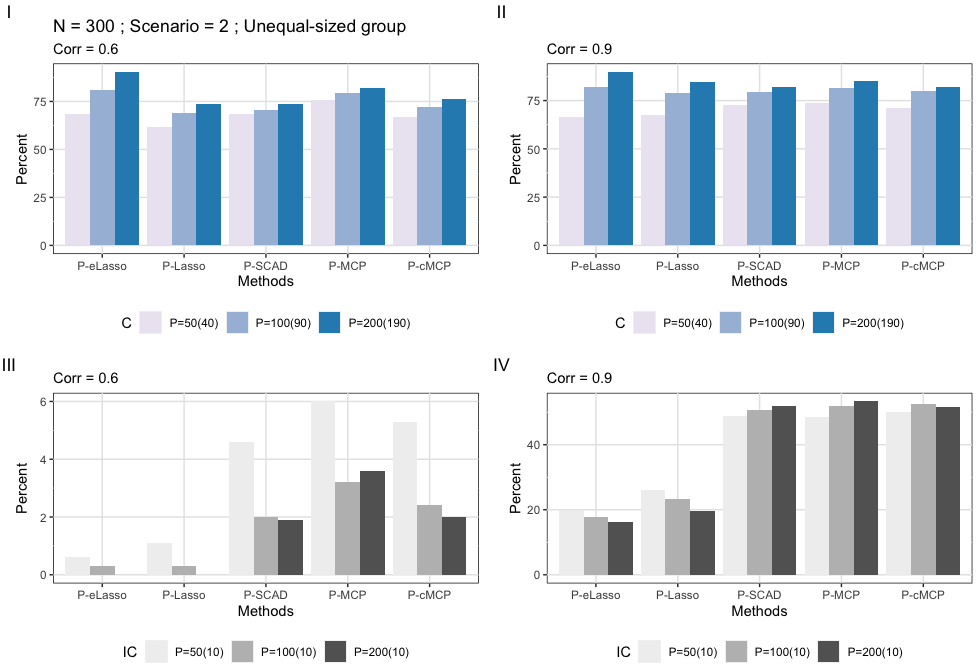


(b) Unequal-Sized Group in Scenario 1, n=300, p = (50, 100, 200)

Figure 2.4: Penalized Weighted Generalized Estimating Equations with various penalties: 1) Exclusive Lasso (P-eLasso), 2) Lasso (P-Lasso), 3) SCAD (P-SCAD), 4) MCP (P-MCP), and 5) Composite MCP (P-cMCP) tested in Scenario 1 with different data settings. Corr = 0.60 (Left), 0.90 (Right); n = 300; p = (50, 100, 200). The number of true zero and non-zero coefficients are indicated next to *p* for C and IC, respectively.

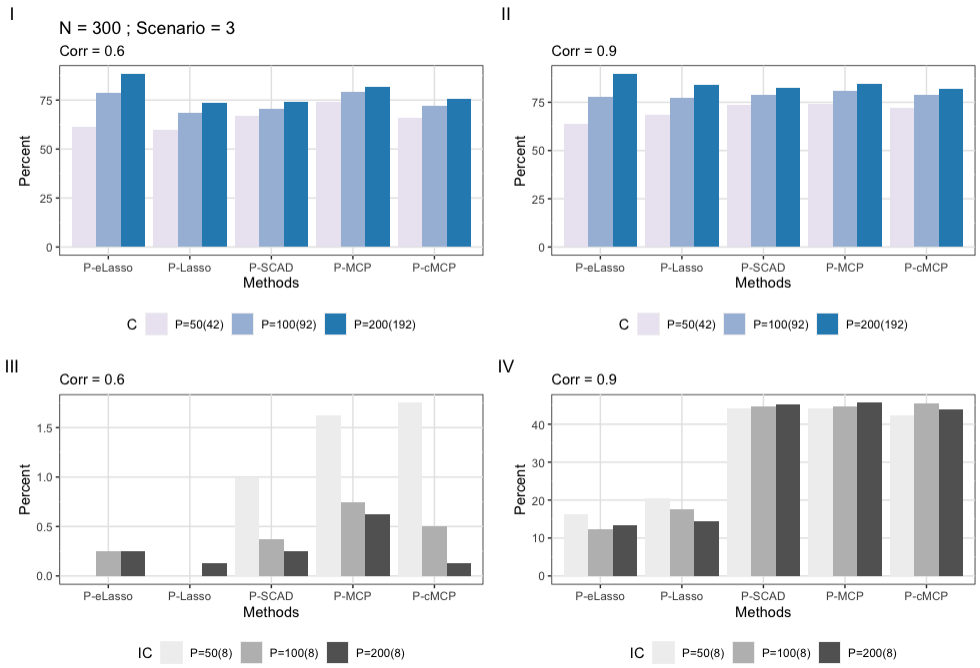


(a) Equal-Sized Group in Scenario 2, n=300, p = (50, 100, 200)

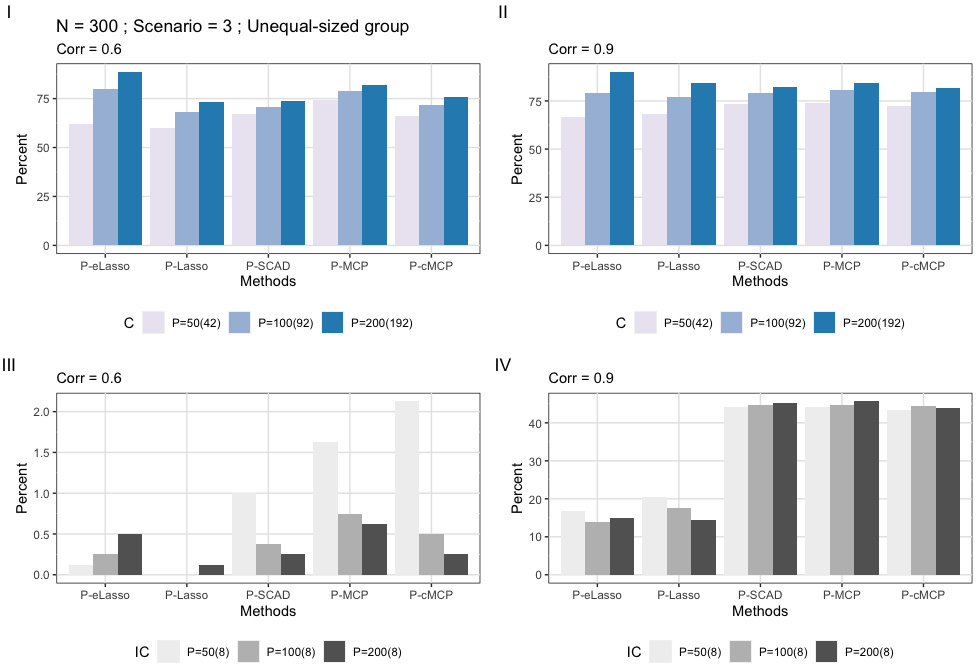


(b) Unequal-Sized Group in Scenario 2, n=300, p = (50, 100, 200)

Figure 2.5: Penalized Weighted Generalized Estimating Equations with various penalties: 1) Exclusive Lasso (P-eLasso), 2) Lasso (P-Lasso), 3) SCAD (P-SCAD), 4) MCP (P-MCP), and 5) Composite MCP (P-cMCP) tested in Scenario 2 with different data settings. Corr = (0.60 (Left), 0.90 (Right)); n = 300; p = (50, 100, 200). The number of true zero and non-zero coefficients are indicated next to p for C and IC, respectively.

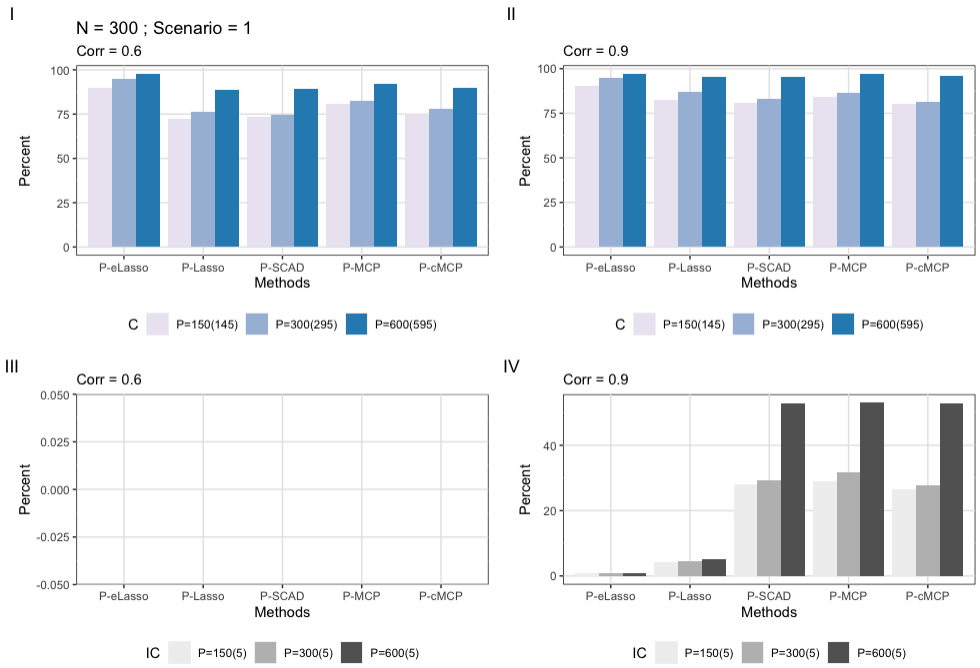


(a) Equal-Sized Group in Scenario 3, n=300, p = (50, 100, 200)

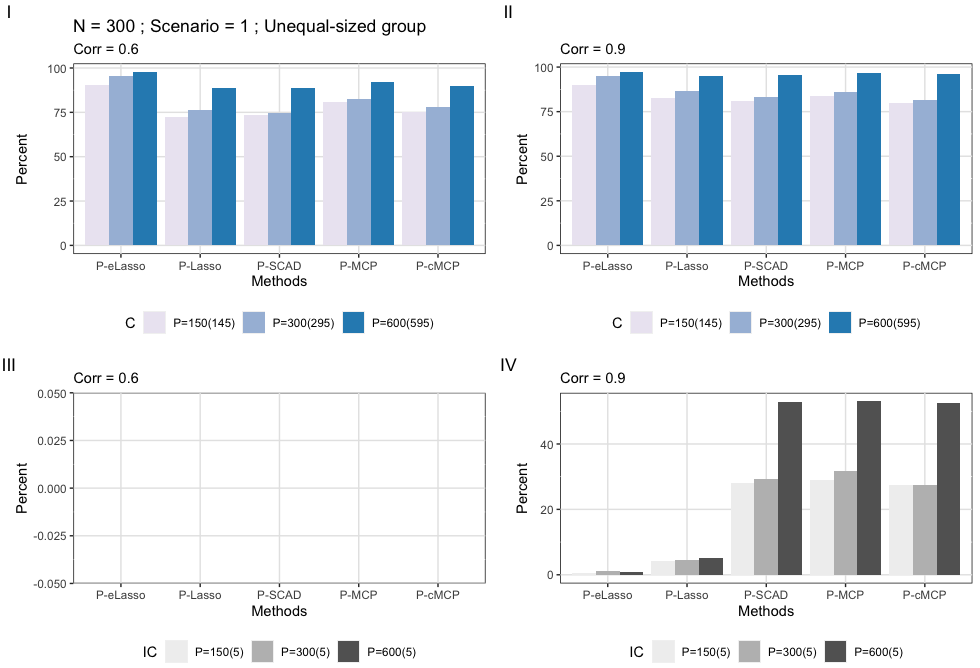


(b) Unequal-Sized Group in Scenario 3, n=300, p = (50, 100, 200)

Figure 2.6: Penalized Weighted Generalized Estimating Equations with various penalties: 1) Exclusive Lasso (P-eLasso), 2) Lasso (P-Lasso), 3) SCAD (P-SCAD), 4) MCP (P-MCP), and 5) Composite MCP (P-cMCP) tested in Scenario 3 with different data settings. Corr = (0.60 (Left), 0.90 (Right)); n = 300; p = (50, 100, 200). The number of true zero and non-zero coefficients are indicated next to *p* for C and IC, respectively.

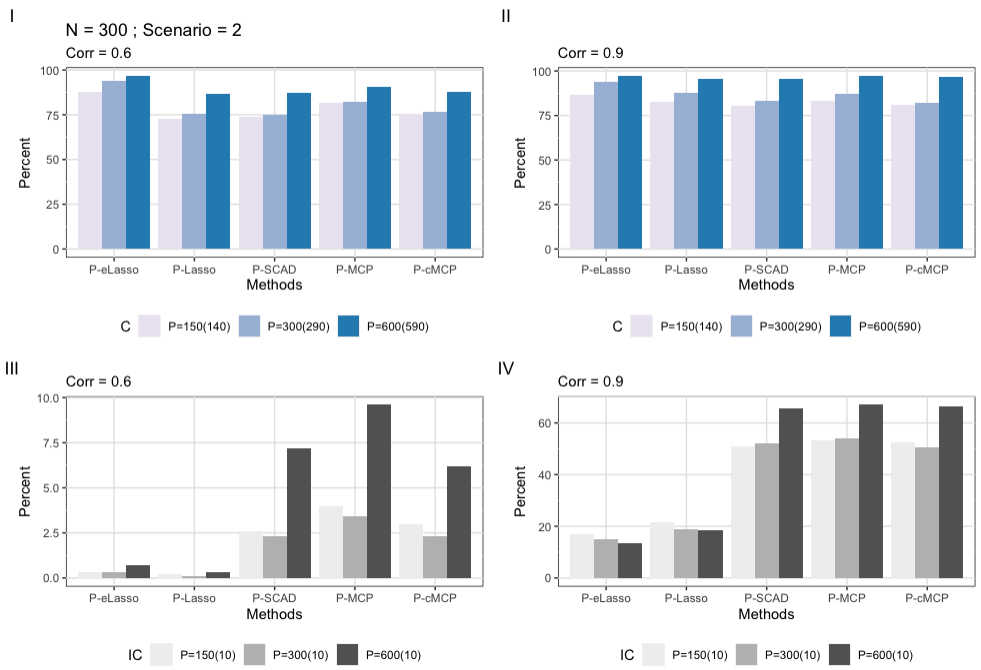


(a) Equal-Sized Group in Scenario 1, n=300, p = (150, 300, 600)

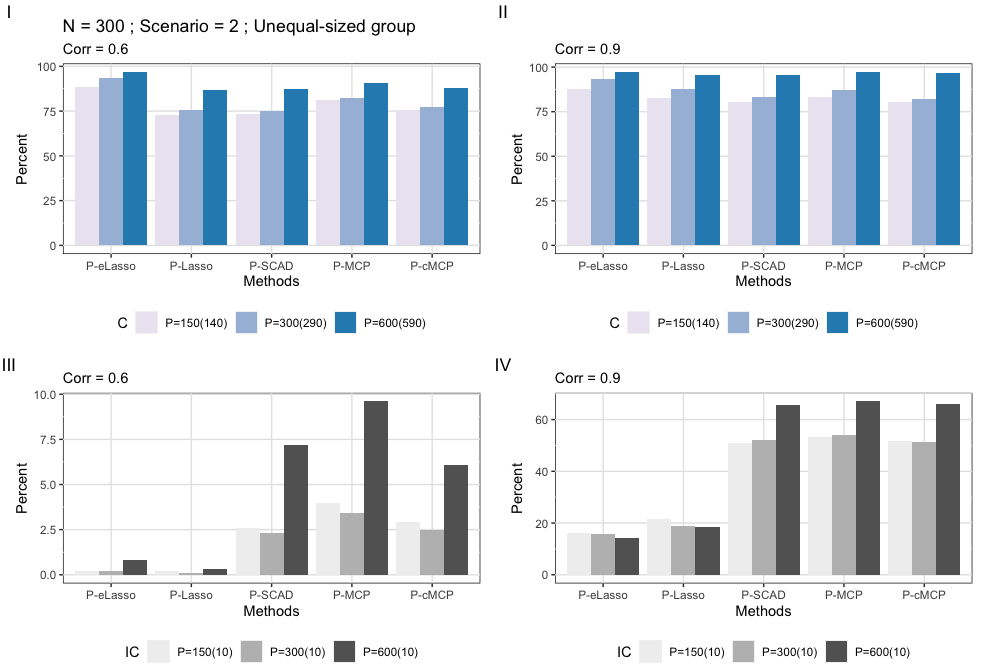


(b) Unequal-Sized Group in Scenario 1, n=300, p = (150, 300, 600)

Figure 2.7: Penalized Weighted Generalized Estimating Equations with various penalties: 1) Exclusive Lasso (P-eLasso), 2) Lasso (P-Lasso), 3) SCAD (P-SCAD), 4) MCP (P-MCP), and 5) Composite MCP (P-cMCP) tested in Scenario 1 with different data settings. Corr = (0.60 (Left), 0.90 (Right)); n = 300; Increased dimension, p = (150, 300, 600). The number of true zero and non-zero coefficients are indicated next to p for C and IC, respectively.

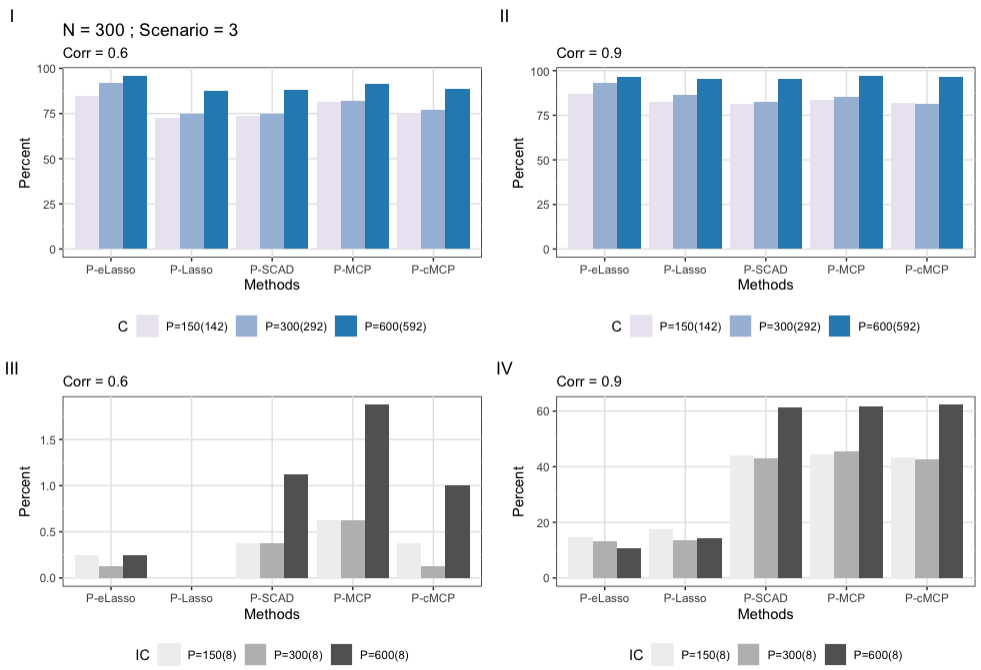


(a) Equal-Sized Group in Scenario 2, n=300, p = (150, 300, 600)

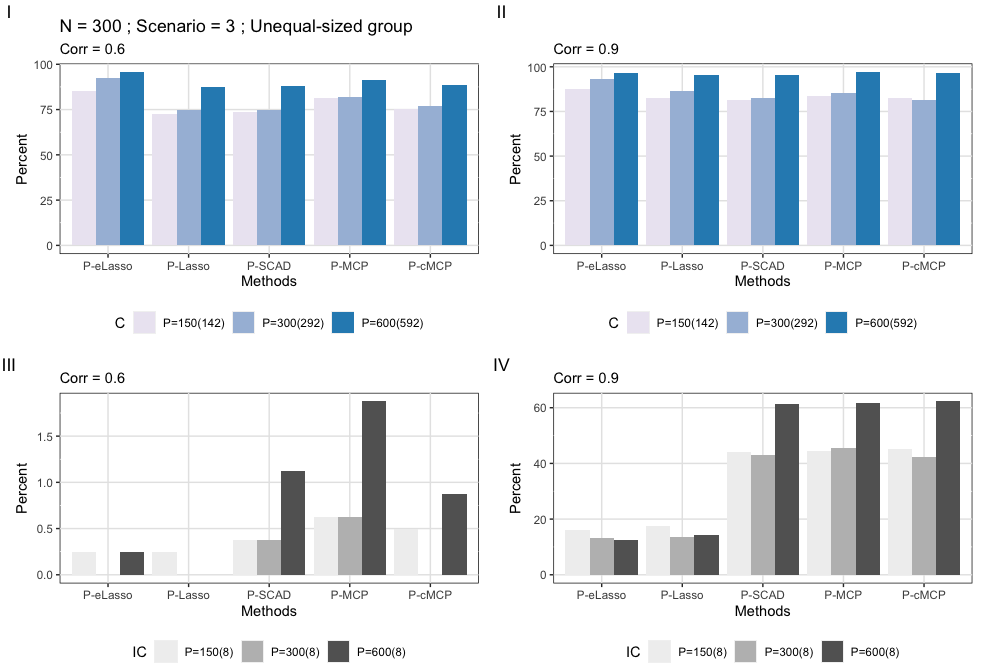


(b) Unequal-Sized Group in Scenario 2, n=300, p = (150, 300, 600)

Figure 2.8: Penalized Weighted Generalized Estimating Equations with various penalties: 1) Exclusive Lasso (P-eLasso), 2) Lasso (P-Lasso), 3) SCAD (P-SCAD), 4) MCP (P-MCP), and 5) Composite MCP (P-cMCP) tested in Scenario 2 with different data settings. Corr = (0.60 (Left), 0.90 (Right)); n = 300; Increased dimension, p = (150, 300, 600). The number of true zero and non-zero coefficients are indicated next to p for C and IC, respectively.



(a) Equal-Sized Group in Scenario 3, n=300, p = (150, 300, 600)



(b) Unequal-Sized Group in Scenario 3, n=300, p = (150, 300, 600)

Figure 2.9: Penalized Weighted Generalized Estimating Equations with various penalties: 1) Exclusive Lasso (P-eLasso), 2) Lasso (P-Lasso), 3) SCAD (P-SCAD), 4) MCP (P-MCP), and 5) Composite MCP (P-cMCP) tested in Scenario 3 with different data settings. Corr = (0.60 (Left), 0.90 (Right)); n = 300; Increased dimension, p = (150, 300, 600). The number of true zero and non-zero coefficients are indicated next to p for C and IC, respectively.

2.6 SDOH Data Application

We used county-level data from the 2009 to 2018 SDOH database from the Agency for Healthcare Research and Quality (AHRQ) (for Healthcare Research & Quality, 2020). AHRQ compiled and released the SDOH database in 2020 to better understand the relationship between community-level factors, healthcare quality and delivery, and individual health to address emerging health issues. The SDOH database is publicly available and integrates data from existing federal data sets and other public data sources spanning multiple years (for Healthcare Research & Quality, 2020). Variables in the data set are organized into 5 SDOH domains: (1) social context, such as age, race and ethnicity, and veteran status; (2) economic context, such as income and unemployment; (3) education, such as education levels; (4) physical infrastructure, such as housing, food insecurity, and transportation; and (5) health care context, such as health insurance coverage and health care access (for Healthcare Research & Quality, 2020). SDOH includes 345 variables across 30 specific topics. The social context covers six topics; the economic context consists of three topics; education encompasses five topics; physical infrastructure includes ten topics; healthcare consists of eight topics. Details are in Table 1.1 in Chapter 1.

Mortality data were obtained from the Interactive Atlas of Heart Disease and Stroke at the Centers for Disease Control and Prevention (CDC) (of Heart Disease & Stroke, n.d.), which were initially compiled from 2 data sources: (1) the National Vital Statistics System at the National Center for Health Statistics, and (2) the hospital discharge data from the Centers for Medicare & Medicaid Services' Medicare Provider Analysis and Review (MEDPAR) file (of Heart Disease & Stroke, n.d.). We linked the SDOH data with corresponding mortality data at the county level. All data and materials used in this analysis are publicly available at the AHRQ SDOH Database website and CDC website.

Our SDOH study data included 72 variables in 5 domains and 2,977 counties (detailed in Appendix Table A.4), not including US territories. Of these counties, 1,149 were classified as urban, while the remaining 1,828 were classified as rural. The rural-urban status of a county was determined according to the Urban-Rural Classification Scheme for Counties used by the National Center for Health Statistics (NCHS) in 2013 (Ingram & Franco, 2014). Counties in categories 1 through 4 were classified as urban, while counties in categories 5 and 6 were classified as rural (Ingram & Franco, 2014). We included this rural-urban indicator as an additional domain to address geographic disparities. In our previous study (Son et al., 2023), the analytic sample was based on the following inclusion and exclusion criteria. First, variables that were measured repeatedly for 10 years were included. Second, variables with more than 60% missing values were excluded. Third, we eliminated redundant variables with the same or similar definitions (e.g., percentage of native-born residents and percentage of foreign-born residents) to avoid duplication. After applying these criteria, we additionally dropped four variables due to significant baseline missing information in 2009. Our proposed method assumes a dropout missing pattern with no missing data observed at $t = 1$ (see Equation 2.3). Therefore, the following variables are not included:

1. Percentage of workers age 16 and over with at least 60-minute commute time by public transportation. 2. Beer, wine, and liquor stores per 1,000 people. 3. Community food services (targeting low-income or elderly) per 1,000 people. 4. Number of people living with diagnosed HIV / 1000.

In 2009, missing values were still included. Consequently, we excluded an additional 165 counties from the previous SDOH data (see details in Appendix Table A.6). Ultimately, our analysis of the SDOH study data reveals several variables with significant missing information regarding physical infrastructure and healthcare context. Of those, counties display a dropout pattern of missing data, starting with no missing data at baseline and increasing in the number of counties over the study period of a decade (see details in Appendix Table A.5). Each domain included different variables: the social context had 22 SDOH variables, the economic context included 25, the education featured 5, the physical infrastructure had 15, and the healthcare context contained 5.

Table 2.4: Selected Social Determinants of Healths Associated with Age-Adjusted CVD Mortality: PWGEE-eLasso, 2009-2018

Domain	**Variables**	**Estimates,** β
	Intercept	272.1746
1. Social context	% Population reporting Black race	0.3114
	% Population divorced or separated (ages 15 and over)	1.1313
	% Population that does not speak English at all (ages 5 and over)	-0.2687
	% Children living with a grandparent householder (ages 17 and under)	0.9175
	% Population reporting Hispanic ethnicity	-0.3351
	% Population who speak other languages (ages 5 and over)	-0.2242
2. Economic context	Gini index of income inequality	40.7584
	% Households that received food stamps/SNAP, past 12 months	0.5939
	% Employed working in manufacturing	0.0400
	% Population with income to poverty ratio: 2.00 or higher	-0.1903
	% Population with income to poverty ratio: <1.00	0.6141
	% Employed working in transportation and warehousing, and in utilities	0.3785
3. Education	% Population with some college or associate's degree (ages 25 and over)	-0.6896
	% Population with a bachelor's degree (ages 25 and over)	-2.0069
4. Physical infrastructure	Median home value of owner-occupied housing units	-0.0001
	% Workers taking a car, truck, or van to work (ages 16 and over)	0.0447
	% Housing units that are mobile homes	0.0191
	% Renter-occupied housing units with children	0.1157
	Full service restaurants per 1,000 people	-5.4385
5. Healthcare	Derived field that equals the ratio of enrollees over eligibles * 100	-0.3940
Geographic identifier	1: Rural, 0: Urban by NCHS 2013 Rural-Urban Classification Scheme	3.1804

We applied our proposed method with Exclusive Lasso (PWGEE-eLasso) in SDOH data. The results of the selection are summarized in Table 2.4. The final model includes variables from five domains: social context, economic context, education, physical infrastructure, and healthcare context. Several variables in the social context domain significantly contributed to the model. For instance, the percentage of the population reporting Black race (Estimates, β, = 0.31) and the percentage of children living with a grandparent householder (β = 0.92) were positively associated with the outcome. Conversely, the percentage of the population that does not speak English at all (β = -0.27) and the percentage of the population reporting Hispanic ethnicity (β = -0.34) were negatively associated with the outcome. In the economic

context, key variables include counties with a Gini index of income inequality (β = 40.76), which exhibited a strong positive association with the outcome. A higher proportion of the population with an income-to-poverty ratio of less than 1.00 (β = 0.61), households receiving food stamps/SNAP in the past 12 months (β = 0.59), and population employed in transportation and warehousing (β = 0.38) were more likely to experience a higher CVD mortality, respectively. However, the percentage of the population with an income-to-poverty ratio greater than 2.00 showed a negative association(β = -0.19). According to the previous analysis ((Son et al., 2023)), the collinearity problem may still persist among those poverty predictors. Regarding education, the percentage of the population with a bachelor's degree (β = -2.01) and those with some college or an associate's degree (β = -0.69) had a negative relationship with the outcome. It may not effectively address highly correlated variables when the number of variables in a specific domain is smaller compared to other domains. Among the physical infrastructure variables, full-service restaurants per 1,000 people (β = -5.44) had a substantial negative effect. Meanwhile, factors such as the percentage of workers taking a car, truck, or van to work (β = 0.04) and the percentage of housing units that are mobile homes (β = 0.02) were positively associated with the outcome. The derived field representing the ratio of enrollees to eligibles (β = -0.39) was negatively associated with the outcome, indicating that a higher ratio correlates with a decrease in the dependent variable in the healthcare context. Lastly, a geographic identifier, represented by a rural-urban classification, had a positive association with the outcome. Specifically, areas classified as rural (β = 3.18) were associated with higher dependent variable values.

Table A.7 in the Appendix presented a comparison of implementing the proposed PWGEE via other penalties. Overall methods generally exhibited consistent directional associations with the outcome from the previous SDOH study (Son et al., 2023), although subtle differences in coefficient magnitudes are evident. For example, while the positive effects of the percentage of the population reporting Black race and children living with a grandparent householder were observed across all penalized models in the social context, the estimates under P-eLasso tended to be slightly more moderate than those obtained via P-Lasso and other non-convex penalties. Similarly, negative associations for variables such as the derived field representing the ratio of enrollees to eligibles were maintained across models in the healthcare context, though their effect sizes varied. These differences highlighted that although the underlying relationships between SDOH and CVD mortality remained robust, the choice of penalization technique could influence the relative weight and sparsity of predictors. Additionally, when the number of variables in a specific domain was notably smaller than in other domains, all methods, especially P-Lasso and P-cMCP, might not have effectively addressed highly correlated variables, leading to less consistent handling of multicollinearity. PWGEE-eLasso could help address geographic disparities by creating a geographical indicator, such as rural-urban classification, as a distinct domain. However, other approaches struggled to address this distinction if they were not specifically tailored for geographic classification.

2.7 Summary

The penalized weighted generalized estimating equations framework with an Exclusive Lasso penalty (PWGEE-eLasso) provides a structured approach to managing high-dimensional longitudinal data with missing observations. By integrating inverse probability weighting for missing data under MAR assumption with a penalty that enforces sparsity within predefined covariate groups, this method simultaneously achieves two primary objectives: consistent parameter estimation when data are incompletely observed and enhanced interpretability by selecting significant variables within each group rather than treating all variables individually or discarding an entire group of covariates.

Simulation studies demonstrate that a notable advantage of the Exclusive Lasso penalty is its ability to effectively handle overlapping or correlated features. Classical Lasso-based approaches often select multiple correlated covariates, whereas group-Lasso methods typically either eliminate all features or retain all features within a given group. The Exclusive Lasso addresses this gap by maintaining within-group sparsity while allowing for the possibility that each group can contribute at least one non-zero variable if it is relevant. Compared to SCAD, MCP, and cMCP, which can also induce sparsity, the Exclusive Lasso penalty demonstrates superior retention of key signals in the presence of multicollinearity without excessively penalizing important predictors. This minimizes estimation bias and increases the likelihood of selecting the correct non-zero covariates.

In practical scenarios such as the SDOH analysis, missing information is almost inevitable. IPW under the MAR assumption ensures consistent estimation of the population-averaged effects, provided that the missing mechanism is accurately modeled. Standard errors and covariance estimates naturally accommodate missingness, as the weighting is directly incorporated into the GEE framework. In addition, many longitudinal data exhibit complex dependencies over time, and a flexible working correlation can approximate these relationships. Thus, PWGEE-eLasso can be advantageous in these contexts because it accommodates various correlation structures commonly used within the GEE framework.

The real data analysis demonstrates the practical value of PWGEE-eLasso in identifying meaningful patterns across multiple SDOH domains. Insights into socio-economic factors, healthcare contexts, and demographic variables reveal how structural inequalities shape CVD outcomes at the county level. By preserving at least one predictor from each relevant group, the final model remains interpretable, providing policymakers and stakeholders with greater clarity on which factors merit targeted interventions.

Even though MAR is commonly assumed in statistical analyses dealing with missing data, future research should investigate alternative patterns of missingness in real-world data settings. Conducting sensitivity analyses and using alternative inference methods for non-ignorable missingness would improve the generalizability of PWGEE-eLasso. Utilizing more flexible or data-driven working correlation matrices could enhance estimation in scenarios with complex dependency patterns. Adaptive weighting or multi-penalty schemes could assist in refining the selection process, especially when groups vary in size or when prior information about covariate importance is available. Moreover, as the number of explanatory variables in the missingness model increases, penalization can also be implemented in the missing

data module. This dual-step penalization strategy aims to address large-scale data structures in both the outcome and missingness models simultaneously.

In summary, PWGEE-eLasso integrates robust missing data handling with principled feature management in longitudinal studies. Its simulation stability and interpretability in large-scale SDOH data analysis highlight its practical value for researchers analyzing complex, incomplete datasets with inherent group structures.

Chapter 3

Model-based Clustering of High-Dimensional Longitudinal Data via Exclusive Lasso Penalty

3.1 Introduction

Cardiovascular disease (CVD) is the leading cause of death in the United States, with an age-adjusted mortality rate for heart disease of 167.2 deaths per 100,000 individuals reported in 2022. (for Disease Control & (CDC), 2023). Numerous studies have explored CVD associated with the social determinants of health (SDOH) using comparable data sources, such as geocoded health records, national survey data, and census information. These studies aim to develop interventions that improve healthcare access and quality, as well as to assess risks at both the neighborhood and individual levels (McNeill et al., 2023).

Health disparities in CVD mortality have been reported to be associated with SDOH (for Disease Control, Prevention, et al., 2019; Frieden et al., 2013). The Centers for Disease Control and Prevention (CDC) and the American Heart Association highlight the importance of addressing SDOH in public health initiatives and healthcare practices to reduce disparities among different racial groups and regions (Banerjee, 2017; Benjamin et al., 2019; Hacker et al., 2022; White-Williams et al., 2020). Specifically, the CDC identifies five key domains of SDOH that need attention to improve overall health outcomes: economic stability, education access and quality, healthcare access and quality, neighborhood and built environment, and social and community context (Hacker et al., 2022). Understanding these domains is essential for designing targeted strategies aimed at reducing CVD mortality by addressing key risk factors associated with each domain and tackling existing disparities.

Recent studies have highlighted specific regional disparities in CVD at the state or county level (Son et al., 2023; Zelko et al., 2023). To address these geographic disparities in CVD, it may be necessary to integrate SDOH with advanced methodological approaches. This integration should focus on geographic groups to link the effects of various SDOH domains to similar longitudinal associations between SDOH and CVD mortality outcomes. This need arises from the findings in our previous study (Son et al., 2023).

Furthermore, since the CVD mortality rate varies by geographic and racial groups over time, US counties may exhibit different characteristics when subjects are clustered based on SDOH in a high-dimensional longitudinal data setting. New clustering strategies that utilize data on SDOH may help drive heterogeneous policy development aimed at reducing the CVD burden and improving CVD outcomes.

The clustering method for high-dimensional longitudinal data has addressed the issue of missing values. Although there has been recent attention to clustering high-dimensional longitudinal data, advances in these techniques remain limited, particularly in applications in SDOH research concerning CVD mortality. For instance, existing observational longitudinal data clustering methods have been developed to group longitudinal trajectories of the outcome variable (Genolini & Falissard, 2010; Jacques & Preda, 2013; McNicholas & Murphy, 2010; Tang & Qu, 2016). However, these methods struggle with handling missing values and effectively exploring the relationship between dependent and independent variables. Thus, alternative approaches such as linear mixed-effects models (LMM), generalized linear mixed-effects models, or semi-parametric mixed-effects models have been introduced to account for the subpopulation heterogeneity (Arribas-Gil et al., 2015; Komárek & Komárková, 2013; Proust-Lima et al., 2015). Nevertheless, challenges persist, particularly with increasing dimensionality of covariates needed for accurate estimations.

While most of the existing studies have focused on variable selection in the high-dimensional linear regression model, only a few investigated high-dimensional linear mixed model settings. Schelldorfer et al., 2010 introduced a penalized likelihood-based approach for selecting fixed effects while assuming a fixed structure for random effects. Subsequently, Y. Fan and Li, 2012 developed a two-step method for selecting both fixed and random effects. More recently, Li et al., 2018 proposed a regularization method that simultaneously performs estimation and variable selection for both fixed and random effects, allowing the dimension of fixed and random effects to diverge to infinity as the sample size increases. Furthermore, several studies have investigated the use of mixture of LMMs with variable selection in high-dimensional longitudinal data. Du et al., 2013 considered the same penalty function for both fixed and random effects selection within a finite mixture of LMMs. Yang and Wu, 2022 developed a clustering method for high-dimensional longitudinal data, which enables the dimensions of fixed and random effects to grow at an exponential rate relative to the sample size under a general class of penalty functions that satisfy specific regularity conditions. Their method allows different penalty functions to be utilized for fixed effects and the diagonal components of random effects, thus allowing various sets of fixed and random effects to adjust for subgroup heterogeneity. However, these methods do not consider the predefined domain heterogeneity among variables. Clusters can be determined by different sets of fixed or random effects within each domain. Therefore, our method will investigate extending prior studies with regularizations in fixed effects to identify subgroup clusters by US counties by incorporating at least one SDOH from each predefined domain.

3.2 Method

3.2.1 Linear Mixed-Effects Model

First, we define some notations for our proposed framework. Suppose the ith subject is measured at m_i time points, $i = 1, ..., n$. Repeated measures m_i may be missed, which means the number of m_i can vary from subject to subject. At time t_{ij}, $j = 1, ..., m_i$, we have $(y_{ij}, \mathbf{X}_{ij}, \mathbf{Z}_{ij})$ observarions, where y_{ij} represents the responses, $\mathbf{X}_{ij} \in \Re^{p_n}$ contains the fixed effects covariates, and $\mathbf{Z}_{ij} \in \Re^{q_n}$ includes the random effects covariates. Here, p_n and q_n denote the dimension of covariates, which increase at a certain rate as n at a certain rate. We use the following notations for the ith subject: $\mathbf{y}_i = (y_{i1}, y_{i2}, ..., y_{im_i})^T$, $\mathbf{X}_i = (\mathbf{X}_{i1}^T, ..., \mathbf{X}_{im_i}^T)^T$, and $\mathbf{Z}_i = (\mathbf{Z}_{i1}^T, ..., \mathbf{Z}_{im_i}^T)^T$. This framework considers the linear mixed-effects model (LMM) for the ith subject, and it has the following form:

$$\mathbf{y}_i = \mathbf{X}_i\boldsymbol{\beta} + \mathbf{Z}_i\mathbf{b}_i + \mathbf{e}_i, \tag{3.1}$$

where $\boldsymbol{\beta}$ is a $(p_n \times 1)$ vector of parameters referred to the constant fixed effects, $\mathbf{b}_i \sim N(\mathbf{0},\ \sigma^2\mathbf{D})$ is a $(q_n \times 1)$ vector of subject-specific random effects coefficients, and $\mathbf{e}_i \sim N(\mathbf{0},\ \mathbf{R}_i)$ is a vector of the i.i.d. random error. $\mathbf{D}$ is a $(q_n \times q_n)$ covariance matrix that specifies the among-unit sources, assumed identical for all units. The covariance matrix of the random error $\mathbf{R}_i$ characterizes variance and correlation due to within-unit sources, and here, $\mathbf{R}_i$ is assumed to be $\sigma^2\mathbf{I}_{m_i}$.

In LMM 3.1, $\mathbf{y}_i$ given $\mathbf{X}_i,\ \mathbf{Z}_i$ follows a multivariate normal distribution with a particular form of the covariance matrix, that is, $\mathbf{y}_i|(\mathbf{X}_i,\ \mathbf{Z}_i) \sim N(\mathbf{X}_i\boldsymbol{\beta},\ \sigma^2\mathbf{V}_i)$, where $\mathbf{V}_i = \mathbf{Z}_i\mathbf{D}\mathbf{Z}_i^T + \mathbf{I}_{m_i}$. Let $\boldsymbol{\Theta}$ include $\boldsymbol{\beta}, \mathbf{D}$, where $\boldsymbol{\Theta} = (\boldsymbol{\beta}, \mathbf{D})$. Excluding the constant terms, the overall log-likelihood of the entire data set under model 3.1 is

$$\begin{aligned} l_n(\boldsymbol{\Theta}, \sigma^2) \quad &= \quad -\frac{1}{2}\sum_{i=1}^{n} log|\sigma^2\mathbf{V}_i| \\ &\quad -\frac{1}{2\sigma^2}\sum_{i=1}^{n}(\mathbf{y}_i - \mathbf{X}_i\boldsymbol{\beta})^T\mathbf{V}_i^{-1}(\mathbf{y}_i - \mathbf{X}_i\boldsymbol{\beta}). \end{aligned} \tag{3.2}$$

To maximize the log-likeligood in 3.2 we derive the maximum likelihood estimator of $\boldsymbol{\sigma^2}$. The maximum likelihood estimator of $\boldsymbol{\sigma^2}$ is a function of $\boldsymbol{\beta}$ and $\mathbf{D}$ and is given by

$$\hat{\sigma}^2_{MLE}(\boldsymbol{\Theta}) = \frac{1}{N}\sum_{i=1}^{n}(\mathbf{y}_i - \mathbf{X}_i\boldsymbol{\beta})^T\mathbf{V}_i^{-1}(\mathbf{y}_i - \mathbf{X}_i\boldsymbol{\beta}), \tag{3.3}$$

where $N = \sum_{i=1}^{n} m_i$. Next, to derive the profile log-likelihood for $\boldsymbol{\beta}$ and $\mathbf{D}$ we substitute $\hat{\sigma}_{MLE}(\boldsymbol{\beta}, \mathbf{D})$ into Equation 3.2,

$$
\begin{aligned}
l_n(\boldsymbol{\Theta}, \sigma^2_{MLE}) &= -\frac{1}{2}\sum_{i=1}^{n}\left[log|\mathbf{V}_i| + log\left\{\sum_{i=1}^{n}\frac{1}{N}(\mathbf{y}_i - \mathbf{X}_i\boldsymbol{\beta})^T\mathbf{V}_i^{-1}(\mathbf{y}_i - \mathbf{X}_i\boldsymbol{\beta})\right\}\right] - \frac{N}{2} \\
&= -\frac{1}{2}\left[\sum_{i=1}^{n}log|\mathbf{V}_i| + log\frac{1}{N} + log\left\{\sum_{i=1}^{n}(\mathbf{y}_i - \mathbf{X}_i\boldsymbol{\beta})^T\mathbf{V}_i^{-1}(\mathbf{y}_i - \mathbf{X}_i\boldsymbol{\beta})\right\}\right] - \frac{N}{2} \\
&= -\frac{1}{2}\sum_{i=1}^{n}log|\mathbf{V}_i| - \frac{N}{2}log\left\{\sum_{i=1}^{n}(\mathbf{y}_i - \mathbf{X}_i\boldsymbol{\beta})^T\mathbf{V}_i^{-1}(\mathbf{y}_i - \mathbf{X}_i\boldsymbol{\beta})\right\} + C,
\end{aligned}
$$

where C represents a constant. Thus, the profile log-likelihood for $\boldsymbol{\beta}$ and $\mathbf{D}$ is given by

$$
\begin{aligned}
P_F(\boldsymbol{\Theta}) &= -\frac{1}{2}\sum_{i=1}^{n}log|\mathbf{V}_i| \\
&\quad -\frac{N}{2}log\left\{\sum_{i=1}^{n}(\mathbf{y}_i - \mathbf{X}_i\boldsymbol{\beta})^T\mathbf{V}_i^{-1}(\mathbf{y}_i - \mathbf{X}_i\boldsymbol{\beta})\right\}.
\end{aligned} \tag{3.4}
$$

In high-dimensional settings, estimating the parameters from 3.4 may require a sparsity assumption, meaning that only a small set of variables are associated with the response, i.e., only a few elements in beta coefficients are nonzero. Under this assumption, we expect many coefficients to be zero, effectively reducing dimensionality by discarding unimportant parameters. Accordingly, we can penalize the profile likelihood 3.4 with penalty functions applied to both $\boldsymbol{\beta}$ and $\mathbf{D}$, to shrink negligible fixed and random effect coefficients to zero while preserving true signals for inference.

Penalized Likelihood Estimation with Fixed Effects Selection

Penalty functions on $\boldsymbol{\beta}$ and $\mathbf{D}$ are applied to both fixed and random effects for simultaneous selection. By regularizing $\boldsymbol{\beta}$, any fixed effect estimated to be zero will be removed from the model. Based on our previous study (Son et al., 2023), we assume that initial CVD mortality rates differ by county and that there are no additional random effects. Therefore, we do not impose penalties on $\mathbf{D}$ for the random effects to simplify our framework.

To select the fixed effects, we regularize $\boldsymbol{\beta}$ in the profile log-likelihood function in 3.4, which leads us to define the following objective function:

$$
Q_n(\boldsymbol{\Theta}) = P_F(\boldsymbol{\Theta}) - \frac{1}{2}P_{\lambda_n}\sum_{g\in G}\left(\sum_{j\in g}|\beta_j|\right)^2, \tag{3.5}
$$

where P_{λ_n} is non-decreasing penalty functions contingent on the non-negative tuning parameters λ_n, and g denotes a collection of non-overlapping predefined domains for all p_n. The penalty functions

$\frac{1}{2}P_{\lambda_n}\sum_{g\in G}\left(\sum_{j\in g}|\beta_j|\right)^2$ regulates the sparsity of $\boldsymbol{\beta}$ using an Exclusive lasso penalty. Other penalty functions that can be considered are Lasso and SCAD.

3.2.2 Mixture of LMM with Variable Selection

We have previously provided background on the LMM with a selection of fixed effects for p_n and q_n. However, a single LMM may be inadequate for the county-level SDOH data analysis due to geographic disparities. To address this issue, we assume the counties can be divided into k clusters, each following a different LMM model. Thus, for those counties within the kth cluster, we reformulate the LMM as follows:

$$\mathbf{y}_i = \mathbf{X}_i\boldsymbol{\beta}_k + \mathbf{Z}_i\mathbf{b}_{ik} + \mathbf{e}_{ik}, \tag{3.6}$$

where $\boldsymbol{\beta}_k$ is the fixed effects coefficients of size p_{n_k} in cluster k, $\mathbf{b}_{ik} \sim N(\mathbf{0},\ \sigma_k^2\mathbf{D}_k)$ is the subject-specific random effects coefficients of size q_{n_k} in cluster k, and $\mathbf{e}_{ik}$ is an $(n_i \times 1)$ error vector with $N(\mathbf{0},\ \boldsymbol{\sigma}_k^2\mathbf{I}_{m_i})$. In practice, the proposed clustering framework aims to classify a sample of subjects into one of the K clusters, where K may be unknown, based on a predefined rule of similarities in their observed patterns. A straightforward approach is to assume that the observed data $(\mathbf{y}_i, \mathbf{X}_i, \mathbf{Z}_i)$ follows a mixture of LMM across K groups, and consider each mixture component to be one cluster. It is also assumed that each mixture component $k \in \{1, ..., K\}$ has cluster-specific parameters, denoted as $\tilde{\boldsymbol{\Theta}}_k = (\boldsymbol{\beta}_k, \mathbf{D}_k, \sigma_k^2)$. The mixing probability for any subject i to belong to cluster k is represented by $P(\tau_i = k) \equiv \pi_k$, with the constraint that $\sum_{k=1}^{K}\pi_k = 1$. Let $w_{ik} = \mathbf{1}_{\{\pi_i=k\}}$ be the binary indicator based on whether subject i belongs to cluster k. Additionally, let $\mathbf{W}_k = (w_{1k}, ..., w_{nK})$ and $\mathbf{W} = (\mathbf{W}_1, ..., \mathbf{W}_K)$.

Assume the binary indicators $\mathbf{W}$ can not be observed in addition to **y, X, Z**, then the observed likelihood in model 3.6 is that

$$\begin{aligned} f(\mathbf{y}_i|\mathbf{X}_i, \mathbf{Z}_i, \tilde{\boldsymbol{\Theta}}) &= \pi_k f_k(\mathbf{y}_i|\mathbf{X}_i, \mathbf{Z}_i, \boldsymbol{\Theta}_k) \\ &= \textstyle\sum_{k=1}^{K}\pi_k\boldsymbol{\phi}(\mathbf{y}_i|\mathbf{X}_i\boldsymbol{\beta}_k, \sigma_k^2\mathbf{V}_{ik}), \end{aligned} \tag{3.7}$$

where $\tilde{\boldsymbol{\Theta}} = (\boldsymbol{\Theta}_1, ..., \boldsymbol{\Theta}_K)$ is the vector of all unkonw parameters, $\boldsymbol{\Theta}_k = (\boldsymbol{\beta}, \mathbf{D}_k, \mathbf{w}_k, \pi_k)$ denotes the vector of parameters of the kth components, and $\boldsymbol{\phi}$ is the multivariate Gaussian function with a mean of $\mathbf{X}_i\boldsymbol{\beta}_k$ and a covariance of $\sigma_k^2\mathbf{V}_{ik} = \sigma_k^2(\mathbf{Z}_i^T\mathbf{D}_k\mathbf{Z}_i + \mathbf{I}_{mi})$.

Since it is challenging to maximaize the likelihood in 3.7 due to unknown cluster assignments, we adopt an Expectation-Maximization (EM) approach (Dempster et al., 1977). To illustrate how the EM algorithm operates, suppose we can observe $\mathbf{W}$ in **y, X, Z**. In this context, the complete log-likelihood is given by

$$
\begin{aligned}
&log\left\{\prod_{i=1}^{n}\prod_{k=1}^{K}\pi_k^{w_{ik}}\boldsymbol{\phi}(\mathbf{y}_i|\mathbf{X}_i\boldsymbol{\beta}_k,\sigma_k^2\mathbf{V}_{ik})^{w_{ik}}\right\} \\
&=\sum_{i=1}^{n}\sum_{k=1}^{K}\{w_{ik}log(\pi_k)+w_{ik}log\boldsymbol{\phi}(\mathbf{y}_i|\mathbf{X}_i\boldsymbol{\beta}_k,\sigma_k^2\mathbf{V}_{ik})\}.
\end{aligned}
\tag{3.8}
$$

The variance, σ_k^2, can be expressed as a function of $(\boldsymbol{\beta}_k, \mathbf{D}_k)$, hence, the full log-likelihood for the mixture of LMM (mixLMM) can be rewritten as the following:

$$
\begin{aligned}
l_k(\boldsymbol{\Theta}) &= \sum_{k=1}^{K}\sum_{i=1}^{n} w_{ik}log\pi_k \\
&\quad -\sum_{k=1}^{K}\left\{\frac{1}{2}\sum_{i=1}^{n} w_{ik}log|\mathbf{V}_{ik}|\right. \\
&\quad \left.+\frac{N}{2}log\left\{\sum_{i=1}^{n} w_{ik}(\mathbf{y}_i-\mathbf{X}_i\boldsymbol{\beta})^T\mathbf{V}_i^{-1}(\mathbf{y}_i-\mathbf{X}_i\boldsymbol{\beta})\right\}\right\} \\
&= \sum_{k=1}^{K}\sum_{i=1}^{n} w_{ik}log\pi_k+\sum_{k=1}^{K}P_F(\boldsymbol{\Theta}_k).
\end{aligned}
\tag{3.9}
$$

The estimate of $\boldsymbol{\Theta}_k$ can be easily obtained if the weights w are known. This can be done by simply minimizing the profile log-likelihood $P_F(\boldsymbol{\Theta}_k)$ with respect to $\boldsymbol{\Theta}_k = (\boldsymbol{\beta}, \mathbf{D}_k, \mathbf{w}_k, \pi_k)$. In practice, however, the cluster assignment for each subject i is typically unknown, making w unavailable. Therefore, this problem can be framed within a missing data context.

The proposed EM algorithm iteratively maximizes the full log-likelihood by alternating between E and M steps.

(1) **E-step**: Estimating the $\mathbf{W}$ based on current parameter values.

(2) **M-step**: Updating the parameter estimates $\boldsymbol{\Theta}_k$ by maximizing the expectation of the full log-likelihood.

Penalized Model-based Clustering

We now introduce a penalized model-based clustering method that focuses on fixed effect selection through Exclusive Lasso (mixLMM-eLasso) using the following objective function:

$$
\begin{aligned}
Q_n(\boldsymbol{\Theta}) &= \sum_{k=1}^{K}\sum_{i=1}^{n} w_{ik} log\pi_k + \sum_{k=1}^{K} P_F(\boldsymbol{\Theta}_k) \\
&\quad - \sum_{k=1}^{K}\left\{\frac{1}{2} P_{\lambda_{kn}} \sum_{k\in K}\left(\sum_{j\in k}|\boldsymbol{\beta}_k|\right)^2\right\} \qquad (3.10) \\
&= \sum_{k=1}^{K}\sum_{i=1}^{n} w_{ik} log\pi_k + \sum_{k=1}^{K} Q_n(\boldsymbol{\Theta}_k),
\end{aligned}
$$

where $Q_n(\boldsymbol{\Theta}_k)$ is the objective function within each cluster and is defined in 3.5. The selection and estimation of the regression coefficients depend on the regularization parameters. The regularization parameters are denoted by λ_{kn} for $\boldsymbol{\beta_k}$.

To minimize the objective function in equation 3.10, a coordinate descent algorithm embedded within the EM algorithm is used. Furthermore, the number of clusters in the data needs to be estimated. The Bayesian Information Criterion (BIC) is used for determining K and the optimal values of $\boldsymbol{\lambda}_{kn}$, where $\boldsymbol{\lambda}_{kn} = (\lambda_{1,1n}, ..., \lambda_{K,n}) \in \Re^K$. The BIC is referred by

$$BIC = -2l_k(\boldsymbol{\Theta}) + dlogN,$$

where d is the total number of nonzero parameters in the model and $N = \sum_{i=1}^{n} m_i$.

3.2.3 Algorithm

To perform selection within specified groups of variables, the coordinate descent algorithm is used to solve the Exclusive Lasso problem (Campbell & Allen, 2017). Additionally, for clustering and selections within the clusters, a nested EM algorithm is adopted to minimize the objective function represented by Equation 3.10. This process involves minimizing the conditional expectation of penalized log-likelihoods using a coordinate descent algorithm.

M-step

In the Maximization step, the parameter estimate $\hat{\boldsymbol{\Theta}}_k$ is updated. Initially, the estimates $\hat{w}_{ik}^{(r)}$ and $\hat{\pi}_{ik}^{(r)}$ are considered known, and $\hat{\boldsymbol{\Theta}}_k$ is adjusted by minimizing the conditional expectation of the objective function 3.10 through:

$$
\begin{aligned}
Q_n(\boldsymbol{\Theta}_k|\boldsymbol{\Theta}^{(r)}) &= -\sum_{k=1}^{K}\Bigg\{\sum_{i=1}^{n}\hat{w}_{ik}^{(r)}log\hat{\pi}_k^{(r)} - \sum_{k=1}^{K}\sum_{i=1}^{n}\frac{1}{2}\hat{w}_{ik}^{(r)}log|\mathbf{V}_{ik}| \\
&\quad -\frac{N}{2}\Bigg\{\sum_{i=1}^{n}\hat{w}_{ik}^{(r)}(\mathbf{y}_i - \mathbf{X}_i\boldsymbol{\beta}_k)^T\mathbf{V}_{ik}^{-1}(\mathbf{y}_i - \mathbf{X}_i\boldsymbol{\beta}_k)\Bigg\}\Bigg\} \qquad (3.11) \\
&\quad + \sum_{k=1}^{K}\frac{1}{2}P_{\lambda_{kn}}\sum_{g\in G}\Bigg(\sum_{j\in g}|\boldsymbol{\beta}_k|\Bigg)^2.
\end{aligned}
$$

Given the computational burden of coordinate descent for Exclusive Lasso within the EM algorithm, we reformulate the objective function 3.11 and derive the adjusted objective function as

$$
\begin{aligned}
Q_n^*(\boldsymbol{\Theta}_k|\boldsymbol{\Theta}^{(r)}) &= -\sum_{k=1}^{K}\Bigg\{\sum_{i=1}^{n}\hat{w}_{ik}^{(r)}log\hat{\pi}_k^{(r)} - \sum_{k=1}^{K}\sum_{i=1}^{n}\frac{1}{2}\hat{w}_{ik}^{(r)}log|\mathbf{V}_{ik}| \\
&\quad -\frac{N}{2}\Bigg\{\sum_{i=1}^{n}\hat{w}_{ik}^{(r)}(\mathbf{y}_i\mathbf{V}_{ik}^{-\frac{1}{2}} - \mathbf{X}_i\mathbf{V}_{ik}^{-\frac{1}{2}}\boldsymbol{\beta}_k)^T(\mathbf{y}_i\mathbf{V}_{ik}^{-\frac{1}{2}} - \mathbf{X}_i\mathbf{V}_{ik}^{-\frac{1}{2}}\boldsymbol{\beta}_k)\Bigg\}\Bigg\} \\
&\quad + \sum_{k=1}^{K}\frac{1}{2}P_{\lambda_{kn}}\sum_{g\in G}\Bigg(\sum_{j\in g}|\boldsymbol{\beta}_k|\Bigg)^2 \qquad (3.12) \\
&= -\sum_{k=1}^{K}\Bigg\{\sum_{i=1}^{n}\hat{w}_{ik}^{(r)}log\hat{\pi}_k^{(r)} - \sum_{k=1}^{K}\sum_{i=1}^{n}\frac{1}{2}\hat{w}_{ik}^{(r)}log|\mathbf{V}_{ik}| \\
&\quad -\frac{N}{2}\Bigg\{\hat{w}_{ik}^{(r)}\|\mathbf{y}_i^* - \mathbf{X}_i^*\boldsymbol{\beta}_k\|_2^2\Bigg\} - \frac{1}{2}P_{\lambda_{kn}}\sum_{g\in G}\Bigg(\sum_{j\in g}|\boldsymbol{\beta}_k|\Bigg)^2\Bigg\},
\end{aligned}
$$

where $\mathbf{y}_i^* = \mathbf{y}_i\mathbf{V}_{ik}^{-1/2}$, $\mathbf{X}_i^* = \mathbf{X}_i\mathbf{V}_{ik}^{-1/2}$, and $\|\mathbf{y}_i^* - \mathbf{X}_i^*\boldsymbol{\beta}_k\|_2^2$ is driven by the Cholesky decomposition. In adjusted objective function 3.12, the term, $\frac{N}{2}\Bigg\{\|\mathbf{y}_i^* - \mathbf{X}_i^*\boldsymbol{\beta}_k\|_2^2\Bigg\} + \frac{1}{2}P_{\lambda_{kn}}\sum_{g\in G}\Bigg(\sum_{j\in g}|\boldsymbol{\beta}_k|\Bigg)^2$, is referred to as Exclusive Lasso function. Thus, the parameter estimate $\hat{\boldsymbol{\Theta}}$ can be updated by minimizing the Equation 3.12 through the coordinate descent algorithm(Campbell & Allen, 2017) such that

$$
\hat{\Theta}_{kj}^{(r+1)} = argmin_{\Theta_{kj}}Q_n^*(\boldsymbol{\Theta}_k|\boldsymbol{\Theta}^{(r)}) \qquad (3.13)
$$

for $j = 1, ..., p_n$.

E-step

In the Expectation step, the subject-specific and overall clustering probabilities $\hat{w}_{ik}$ and $\hat{\pi}_{ik}$ from the current estimates $\hat{\mathbf{\Theta}}_k$ are updated based on the updated estimate $\hat{\mathbf{\Theta}}^{(r+1)}$ from the M-step:

$$\begin{aligned} \hat{w}_{ik}^{(r+1)} &= \frac{\hat{\pi}_k^{(r)} \phi_k(\mathbf{y}_i | \mathbf{X}_i, \hat{\Theta}_k^{(r+1)})}{\sum_{l=1}^{K} \hat{\pi}_l^{(r)} \phi_l(\mathbf{y}_i | \mathbf{X}_i, \hat{\Theta}_l^{(r+1)})}, \\ \hat{\pi}_k^{(r+1)} &= \frac{\sum_{i=1}^{n} \hat{w}_k^{(r+1)}}{n}. \end{aligned} \tag{3.14}$$

3.3 Asymptotic properties

Numerous studies have investigated convergence rates of p_n in linear regression models (Lan, 2006; Schelldorfer et al., 2010). Recent studies have also focused on convergence rates of q_n (Bickel & Levina, 2008; Bondell et al., 2010; Lam & Fan, 2009; Li et al., 2018; Yang & Wu, 2022). Furthermore, a penalized likelihood approach and a nested EM algorithm have been introduced to facilitate efficient numerical computations for finite mixtures of linear mixed effects (FMLME) models (Du et al., 2013). Building on this FMLME study, We will demonstrate the properties of consistency and sparsity in our proposed estimators.

First, decompose the parameter vector $\mathbf{\Theta} = (\mathbf{\Theta}_1, \mathbf{\Theta}_2)$ such that $\mathbf{\Theta}_2$ contains all zero effects from all the mixture components, and split the vector of true parameter values accordingly as $\mathbf{\Theta}_0 = (\mathbf{\Theta}_{10}, \mathbf{\Theta}_{20})$ with $\mathbf{\Theta}_{20} = 0$. Then, denote the elements of $\mathbf{\Theta}_{10}$ with a superscript such as $\beta_{kj}^{10} = (\beta_{11}^{10}, ..., \beta_{1p}^{10}, ..., \beta_{K1}^{10}, ..., \beta_{Kp}^{10})$, for $k = 1, ..., K$ and $j = 1, ..., p$. Our asymptotic results involve assumptions on the following quantities:

$$\begin{aligned} a_n &= max_{k,j} \left\{ \sqrt{n} | p_{\lambda_{kn}}(\beta_{kj}^{10}) | \right\} \\ b_n &= max_{k,j} \left\{ \sqrt{n} | p'_{\lambda_{kn}}(\beta_{kj}^{10}) | \right\} \\ c_n &= max_{k,j} \left\{ \sqrt{n} | p''_{\lambda_{kn}}(\beta_{kj}^{10}) | \right\}, \end{aligned}$$

where $p_{\lambda_{kn}}$ denotes the penalties imposed on the parameters from the kth mixture components, n is a measure of effective sample size for the kth subpopulation, and, $p'_{\lambda_{kn}}(\cdot)$ and $p''_{\lambda_{kn}}(\cdot)$ are the first and second derivatives of the penalty function $p_{\lambda_{kn}}(\theta)$ with respect to θ, respectively. Further, we assume that the penalty functions $p_{\lambda_{kn}}(\theta)$ satisfy the following conditions:

(P1) For all n and k, $p_{\lambda_{kn}}(0) = 0$ and $p_{\lambda_{kn}}(\theta)$ is symmetric and non-negative. In addition, it is non-decreasing and twice differentiable for all θ in $(0, \infty)$ except at a finite number of points, which allows for a limited number of special points where the curve is not smooth.

(P2) As $n \to \infty$, $a_n = o(1 + b_n)$, and $c_n = o(\sqrt{n})$.

(P3) For $N_n = \{\theta : 0 < \theta \leq n^{-1/2} log(n)\}$, $\lim_{n\to\infty} \inf_{\theta \in N_n} \sqrt{n} p'_{\lambda_{kn}}(\theta) = \infty$.

Suppose that the data $(\mathbf{y}_i, \mathbf{X}_i, \mathbf{Z}_i)$ for $i = 1, ..., n$, is a random sample from the mixture of LMM. We assume that each mixture component $k \in \{1, ..., K\}$ has cluster-specific parameters $\mathbf{\Theta} = (\mathbf{\Theta}_1, ..., \mathbf{\Theta}_K)$. Let $f(\mathbf{y}_i, \mathbf{X}_i, \mathbf{Z}_i; \mathbf{\Theta})$ be the joint density function of $(\mathbf{y}_i, \mathbf{X}_i, \mathbf{Z}_i)$ and $\mathbf{\Omega}$ be an open parameter space.

We define the regularity conditions for the consistency and sparsistency below:

(A1) $f(\mathbf{y}_i, \mathbf{X}_i, \mathbf{Z}_i; \mathbf{\Theta})$ is identifiable in $\mathbf{\Theta}$ up to permutation of the components of the mixture.

(A2) For each $\mathbf{\Theta}_0 \in \mathbf{\Omega}$, there exist $M_{1i}(\mathbf{y}_i, \mathbf{X}_i, \mathbf{Z}_i)$, $M_{2i}(\mathbf{y}_i, \mathbf{X}_i, \mathbf{Z}_i)$, and $M_{3i}(\mathbf{y}_i, \mathbf{X}_i, \mathbf{Z}_i)$ (possibly depending on $\mathbf{\Theta}_0$) such that for $\mathbf{\Theta}$ in a neighborhood of $\mathbf{\Theta}_0$,

$$\left|\frac{\partial log f(\mathbf{y}_i, \mathbf{X}_i, \mathbf{Z}_i; \mathbf{\Theta})}{\partial \theta_j}\right| < M_{1i}(\mathbf{y}_i, \mathbf{X}_i, \mathbf{Z}_i),$$
$$\left|\frac{\partial^2 log f(\mathbf{y}_i, \mathbf{X}_i, \mathbf{Z}_i; \mathbf{\Theta})}{\partial \theta_j \partial \theta_l}\right| < M_{2i}(\mathbf{y}_i, \mathbf{X}_i, \mathbf{Z}_i),$$
$$\left|\frac{\partial^3 log f(\mathbf{y}_i, \mathbf{X}_i, \mathbf{Z}_i; \mathbf{\Theta})}{\partial \theta_j \partial \theta_l \partial \theta_r}\right| < M_{3i}(\mathbf{y}_i, \mathbf{X}_i, \mathbf{Z}_i),$$

such that $E_{\mathbf{\Theta}_0}(M_{1i}(\mathbf{y}_i, \mathbf{X}_i, \mathbf{Z}_i)) < \infty$, $E_{\mathbf{\Theta}_0}(M_{2i}(\mathbf{y}_i, \mathbf{X}_i, \mathbf{Z}_i)) < \infty$, and $E_{\mathbf{\Theta}_0}(M_3 1i(\mathbf{y}_i, \mathbf{X}_i, \mathbf{Z}_i)) < \infty$.

(A3) The Fisher information matrix $I(\mathbf{\Theta}_0)$ is finite and positive definite.

Let $(\mathbf{y}_i, \mathbf{X}_i, \mathbf{Z}_i)$, $i = 1, ..., n$, be a random sample from the observed likelihood of LMM 3.7 satisfying the regularity conditions $(A1) - (A3)$. Theorem 3.3.1 states that when b_n is bounded (i.e., $b_n = O(1)$), there exists a local maximizer $\hat{\mathbf{\Theta}}_n$ of the penalized likelihood function defined in equation 3.10. This maximizer converges to 0 at a rate of $\sqrt{n}$. This rate of convergence can be achieved using the Lasso, Exclusive Lasso, and SCAD penalties, provided that the tuning parameters are chosen appropriately.

Theorem 3.3.1. *Suppose the penalty function $p_{\lambda_{kn}}(\theta)$ satisfies the conditions (P_1) and (P_2). Then there exists a local maximizer $\hat{\mathbf{\Theta}}_n$ of the penalized log-likelihood function in 3.10 such that $\|\hat{\mathbf{\Theta}} - \mathbf{\Theta}_0\| = O_p\{n^{-1/2}(1 + b_n)\}$.*

Proof. Let $r_n = n^{-1/2}(1+b_n)$. It suffices to show that for any small enough $\epsilon > 0$, there exists a constant M_ϵ such that for sufficiently large n,

$$Pr\left\{ \sup_{\|u\|=M_\epsilon} \mathbf{Q}_n(\mathbf{\Theta}_0 + r_n\mathbf{u}) < \mathbf{Q}_n(\mathbf{\Theta}_0) \right\} \geq 1-\epsilon.$$

So with large probability, there exists a local maximum in $\{\mathbf{\Theta}_0 + r_n\mathbf{u} : \|\mathbf{u}\| \leq M_\epsilon\}$. This local maximizer $\hat{\mathbf{\Theta}}_n$ satisfies $\|\hat{\mathbf{\Theta}} - \mathbf{\Theta}_0\| = O_p\{n^{-1/2}(1+b_n)\}$. Let

$$\begin{aligned}\Delta_n(\mathbf{u}) &= \mathbf{Q}_n(\mathbf{\Theta}_0 + r_n\mathbf{u}) - \mathbf{Q}_n(\mathbf{\Theta}_0) \\ &= \{l_n(\mathbf{\Theta}_0 + r_n\mathbf{u}) - l_n(\mathbf{\Theta}_0)\} - \{p_n(\mathbf{\Theta}_0 + r_n\mathbf{u}) - p_n(\mathbf{\Theta}_0)\}.\end{aligned}$$

By condition $(P1)$, $p_{\lambda_k n}(0) = 0$ and hence $p_n(\mathbf{\Theta}_0) = p_n(\mathbf{\Theta}_{10})$. Given that $p_n(\mathbf{\Theta}_0 + r_n\mathbf{u})$ is a sum of positive terms, removing terms corresponding to zero components only decreases its value. Therefore,

$$\begin{aligned}\Delta_n(\mathbf{u}) &\leq \{l_n(\mathbf{\Theta}_0 + r_n\mathbf{u}) - l_n(\mathbf{\Theta}_0)\} - \{p_n(\mathbf{\Theta}_{10} + r_n\mathbf{u}_I) - p_n(\mathbf{\Theta}_{10})\} \\ &\leq \{l_n(\mathbf{\Theta}_0 + r_n\mathbf{u}) - l_n(\mathbf{\Theta}_0)\} + |p_n(\mathbf{\Theta}_{10} + r_n\mathbf{u}_I) - p_n(\mathbf{\Theta}_{10})|,\end{aligned}$$

where $\mathbf{u}_I$ is the sub-vector of $\mathbf{u}$ that corresponds to the nonzero effects. By Taylor's expansion,

$$\begin{aligned}l_n(\mathbf{\Theta}_0 + r_n\mathbf{u}) - l_n(\mathbf{\Theta}_0) &= r_n l_n'(\mathbf{\Theta}_0)^T\mathbf{u} + \frac{r_n^2}{2}\mathbf{u}^T(l_n''(\mathbf{\Theta}_0))\mathbf{u} \\ &= \frac{(1+b_n)}{\sqrt{n}} l_n'(\mathbf{\Theta}_0)^T\mathbf{u} + \frac{(1+b_n)^2}{2n}\mathbf{u}^T(l_n''(\mathbf{\Theta}_0))\mathbf{u},\end{aligned}$$

where we omitted the remainder terms since they become negligible as $n \to \infty$ by regularity condition $(A2)$. For the Hessian matrix $l_n''(\mathbf{\Theta}_0)$, we have

$$\frac{1}{n} l_n''(\mathbf{\Theta}_0) \to_p -I(\mathbf{\Theta}_0).$$

Therefore,

$$\begin{aligned}l_n(\mathbf{\Theta}_0 + r_n\mathbf{u}) - l_n(\mathbf{\Theta}_0) &= \frac{(1+b_n)}{\sqrt{n}} l_n'(\mathbf{\Theta}_0)^T\mathbf{u} - \frac{(1+b_n)^2}{2n}\mathbf{u}^T I(\mathbf{\Theta}_0)\mathbf{u}(1+o_p(1)) \\ &= (1+b_n)O_p(1)\|\mathbf{u}\| - \frac{(1+b_n)^2}{2n}\mathbf{u}^T I(\mathbf{\Theta}_0)\mathbf{u}(1+o_p(1)), \qquad (3.15)\end{aligned}$$

as $\frac{1}{\sqrt{n}} l_n'(\mathbf{\Theta}_0) = O_p(1)$ by regularity conditions. On the other hand, by Taylor's expansion and the triangular inequality,

$$
\begin{aligned}
&|p_n(\boldsymbol{\Theta}_{10} + r_n \mathbf{u}_I) - p_n(\boldsymbol{\Theta}_{10})| \\
&= p_n'(\boldsymbol{\Theta}_{10})^T r_n \mathbf{u}_I + \frac{r_n^2}{2} \mathbf{u}_I^T p_n''(\boldsymbol{\Theta}_{10}) \mathbf{u}_I (1 + o(1)) \\
&\leq r_n |p_n'(\boldsymbol{\Theta}_{10})^T \mathbf{u}_I| + \frac{r_n^2}{2} |\mathbf{u}_I^T p_n''(\boldsymbol{\Theta}_{10}) \mathbf{u}_I| (1 + o(1)) \\
&\leq r_n \|p_n'(\boldsymbol{\Theta}_{10})^T\| \cdot \|\mathbf{u}_I\| + \frac{r_n^2}{2} \|diag(p_n''(\boldsymbol{\Theta}_{10}))\| \cdot \|\mathbf{u}_I\|^2 (1 + o(1)). \qquad (3.16)
\end{aligned}
$$

Let t_k be the total number of true nonzero fixed and random effects in the k-th component, and let $t = max\{t_k, k = 1, ..., K\}$. Let $\boldsymbol{\beta}^{10}$ denote the vectors of $\boldsymbol{\beta}_{kj}^{10}$'s. We notice that for the first term of 3.16,

$$
\|p_n'(\boldsymbol{\Theta}_{10})\| = \|p_n'(\pi_1^0, ..., \pi_K^0)\| + \|p_n'(\boldsymbol{\beta}^{10})\|,
$$

where $p_n'(\pi_1^0, ..., \pi_K^0)$ and $p_n'(\boldsymbol{\beta}^{10})$ are the gradients of the penalty function $p_n(\cdot)$ with respect to the parameters $(\pi_1^0, ..., \pi_K^0)$ and (β_{kj}), respectively, evaluated at the true value $\boldsymbol{\Theta}_{10}$.

Recall that

$$
p_n(\boldsymbol{\Theta}) = \sum_{k=1}^{K} n\{p_{\lambda_{kn}}(\beta_{kj})\},
$$

where $p_{\lambda_{kn}}(\beta_{kj}) = \dfrac{\lambda_k}{2} \sum\limits_{k \in K} \left\{ \sum\limits_{j \in k} |\beta_{kj}| \right\}^2$. Therefore,

$$
p_n'(\pi_1^0, ..., \pi_K^0) = \begin{bmatrix} n\Big\{p_{\lambda_{kn}}(\beta_{1j}^{10})\Big\} \\ \vdots \\ n\Big\{p_{\lambda_{kn}}(\beta_{Kj}^{10})\Big\} \end{bmatrix}.
$$

Hence,

$$
\begin{aligned}
\|p_n'(\pi_1^0, ..., \pi_K^0)\| &= n \sqrt{\sum_{k=1}^{K} \left\{ p_{\lambda_{kn}}(\beta_{kj}^{10}) \right\}^2} \\
&\leq n \sqrt{\sum_{k=1}^{K} \left\{ t_k \cdot \frac{a_n}{\sqrt{n}} \right\}^2} = a_n \sqrt{n} \sqrt{\sum_{k=1}^{K} t_k^2} \\
&\leq a_n \sqrt{n} \sqrt{\sum_{k=1}^{K} t^2} = a_n \sqrt{n} \sqrt{K} t.
\end{aligned}
$$

Furthermore,

$$
\begin{aligned}
\|p_n'(\boldsymbol{\beta}^{10})\| &= \|\nabla p_n(\beta_{11}^{10}, ..., \beta_{1p}^{10}, ..., \beta_{K1}^{10}, ..., \beta_{Kp}^{10})\| \\
&= n\|\pi_1 p_{\lambda_{1n}}'(\beta_{11}^{10}), ..., \pi_1 p_{\lambda_{1n}}'(\beta_{1p}^{10}), ..., \pi_K p_{\lambda_{Kn}}'(\beta_{K1}^{10}), ..., \pi_K p_{\lambda_{Kn}}'(\beta_{Kp}^{10})\| \\
&\leq n\|p_{\lambda_{1n}}'(\beta_{11}^{10}), ..., p_{\lambda_{1n}}'(\beta_{1p}^{10}), ..., p_{\lambda_{Kn}}'(\beta_{K1}^{10}), ..., p_{\lambda_{Kn}}'(\beta_{Kp}^{10})\| \\
&= n\sqrt{\sum_{k=1}^{K} p_{\lambda_{kn}}'(\beta_{kj}^{10})^2} \\
&\leq n\sqrt{\sum_{k=1}^{K} t_k \cdot \left[\frac{b_n}{\sqrt{n}}\right]^2} = \sqrt{n} b_n \sqrt{\sum_{k=1}^{K} t_k} \\
&\leq b_n\sqrt{n}\sqrt{K \cdot t}.
\end{aligned}
$$

Combining the two terms, we have

$$
\|p_n'(\boldsymbol{\Theta}_0)\| \leq a_n\sqrt{n}\sqrt{K}t + b_n\sqrt{n}\sqrt{K \cdot t}.
$$

After plugged into 3.16, this leads to

$$
\begin{aligned}
&|p_n(\boldsymbol{\Theta}_{10} + r_n\mathbf{u}_I) - p_n(\boldsymbol{\Theta}_{10})| \\
&\leq r_n(a_n\sqrt{n}\sqrt{K}t + b_n\sqrt{n}\sqrt{K \cdot t})\|\mathbf{u}\| + \frac{r_n^2}{2}\|diag(p_n''(\boldsymbol{\Theta}_{10}))\| \cdot \|\mathbf{u}_I\|^2(1 + o(1)) \\
&= \frac{1 + b_n}{\sqrt{n}}(a_n\sqrt{n}\sqrt{K}t + b_n\sqrt{n}\sqrt{K \cdot t})\|\mathbf{u}\| + \frac{(1 + b_n)^2}{2n}\|diag(p_n''(\boldsymbol{\Theta}_{10}))\| \cdot \|\mathbf{u}_I\|^2(1 + o(1)) \\
&= a_n(1 + b_n)\sqrt{K}t\|\mathbf{u}\| + b_n(1 + b_n)\sqrt{K \cdot t}\|\mathbf{u}\| + \frac{(1 + b_n)^2}{2n}\|diag(p_n''(\boldsymbol{\Theta}_{10}))\| \cdot \|\mathbf{u}_I\|^2(1 + o(1)).
\end{aligned} \tag{3.17}
$$

Furthermore,

$$\begin{aligned}\|diag(p_n''(\boldsymbol{\Theta}_{10}))\| &= n\sqrt{\sum_{k=1}^{K} p''_{\lambda_{kn}}(\beta_{kj}^0)^2\pi_k^2} \\ &\leq n\sqrt{\sum_{k=1}^{K} p'_{\lambda_{kn}}(\beta_{kj}^{10})^2} \\ &\leq n\sqrt{\sum_{k=1}^{K} t_k \cdot \left[\frac{c_n}{\sqrt{n}}\right]^2} = \sqrt{n}c_n\sqrt{\sum_{k=1}^{K} t_k} \\ &\leq c_n\sqrt{n}\sqrt{K\cdot t}.\end{aligned}$$

Combining this result with 3.17 gives

$$\begin{aligned}&|p_n(\boldsymbol{\Theta}_{10} + r_n\mathbf{u}_I) - p_n(\boldsymbol{\Theta}_{10})| \\ &\leq a_n(1+b_n)\sqrt{K}t\|\mathbf{u}\| + b_n(1+b_n)\sqrt{K\cdot t}\|\mathbf{u}\| + \frac{(1+b_n)^2}{2\sqrt{n}}c_n\sqrt{K\cdot t}\|\mathbf{u}\|^2(1+o(1)). \qquad (3.18)\end{aligned}$$

By condition $(P2)$, $a_n = o(1+b_n)$ and $c_n = o(\sqrt{n})$. Hence, the order comparison of 3.15 and 3.18 implies that $-\frac{(1+b_n)^2}{2}\mathbf{u}^T I(\boldsymbol{\Theta}_0)\mathbf{u}(1+o_p(1))$ dominates every other term for sufficiently large $\|\mathbf{u}\| = M_\epsilon$. Therefore, for any given $\epsilon > 0$, there exists sufficiently large M_ϵ such that $\lim_{n\to\infty} Pr\big\{\sup_{\|\mathbf{u}\|=M_\epsilon}\Delta_n(\mathbf{u}) < 0\big\} = 1$. □

Next, we show that the following Theorem 3.3.2 guarantees that, under mild conditions, the penalized likelihood estimators have the sparsity property, which allows for consistent variable selection, and they are asymptotically normally distributed.

Theorem 3.3.2. *Assume that the penalty function $p_{\lambda_{kn}}(\theta)$ satisfies the conditions $(P_1)-(P_3)$, and that K is known a priori. Then the following statements hold.*

(a) For any $\boldsymbol{\Theta} = (\boldsymbol{\Theta}_1, \boldsymbol{\Theta}_2)$ such that $\|\hat{\boldsymbol{\Theta}} - \boldsymbol{\Theta}_0\| = O(n^{-1/2})$, with probability tending to 1,

$$\mathbf{Q}_n(\boldsymbol{\Theta}_1, \boldsymbol{\Theta}_2) < \mathbf{Q}_n(\boldsymbol{\Theta}_1, \mathbf{0}).$$

(b) For any root-n consistent maximum penalized likelihood estimator $\hat{\boldsymbol{\Theta}}_n = (\hat{\boldsymbol{\Theta}}_{1n}, \hat{\boldsymbol{\Theta}}_{2n})$ of $\boldsymbol{\Theta}$,

(i) Sparsity: $Pr\{\hat{\boldsymbol{\Theta}}_{2n} = \mathbf{0}\} \to 1$, as $n \to \infty$.

(ii) Asymptotic normality:

$$\sqrt{n}\left[\left\{\mathbf{I}_1(\boldsymbol{\Theta}_{10}) + \frac{p_n''(\boldsymbol{\Theta}_{10})}{n}\right\}(\hat{\boldsymbol{\Theta}}_{1n} - \boldsymbol{\Theta}_{10}) + \frac{p_n'(\boldsymbol{\Theta}_{10})}{n}\right] \to_d N(\mathbf{0}, \mathbf{I}_1(\boldsymbol{\Theta}_{10})),$$

where $I_1(\boldsymbol{\Theta}_{10})$ is the Fisher information matrix, and $p'(\boldsymbol{\Theta}_{10})$ and $p''(\boldsymbol{\Theta}_{10})$ are the first and second derivatives of the penalty function, respectively, knowing that $\boldsymbol{\Theta}_{20} = \mathbf{0}$.

Proof. (a) Consider the partitioning $\boldsymbol{\Theta} = (\boldsymbol{\Theta}_1, \boldsymbol{\Theta}_2)$ for any $\boldsymbol{\Theta}$ in the neighborhood $\|\boldsymbol{\Theta} - \boldsymbol{\Theta}_0\| = O(n^{-1/2})$. By the definition of $\mathbf{Q}_n(\boldsymbol{\Theta})$, we have

$$\begin{aligned}&\mathbf{Q}_n((\boldsymbol{\Theta}_1, \boldsymbol{\Theta}_2)) - \mathbf{Q}_n((\boldsymbol{\Theta}_1, 0)) \\ &= \big\{l_n((\boldsymbol{\Theta}_1, \boldsymbol{\Theta}_2)) - l_n((\boldsymbol{\Theta}_1, 0))\big\} - \big\{p_n((\boldsymbol{\Theta}_1, \boldsymbol{\Theta}_2)) - p_n((\boldsymbol{\Theta}_1, 0))\big\}.\end{aligned} \tag{3.19}$$

We now find the order of these two differences. By the mean value theorem,

$$l_n((\boldsymbol{\Theta}_1, \boldsymbol{\Theta}_2)) - l_n((\boldsymbol{\Theta}_1, 0)) = \left[\frac{\partial l_n((\boldsymbol{\Theta}_1, \xi))}{\partial \boldsymbol{\Theta}_2}\right]^T \boldsymbol{\Theta}_2, \tag{3.20}$$

for some $\|\xi\| \le \|\boldsymbol{\Theta}_2\| = O(n^{-1/2})$. Then,

$$\begin{aligned}&\left\|\tfrac{\partial l_n((\boldsymbol{\Theta}_1,\xi))}{\partial \boldsymbol{\Theta}_2} - \tfrac{\partial l_n((\boldsymbol{\Theta}_{10},0))}{\partial \boldsymbol{\Theta}_2}\right\| \\ &\le \left\|\tfrac{\partial l_n((\boldsymbol{\Theta}_1,\xi))}{\partial \boldsymbol{\Theta}_2} - \tfrac{\partial l_n((\boldsymbol{\Theta}_1,0))}{\partial \boldsymbol{\Theta}_2}\right\| \quad + \quad \left\|\tfrac{\partial l_n((\boldsymbol{\Theta}_1,0))}{\partial \boldsymbol{\Theta}_2} - \tfrac{\partial l_n((\boldsymbol{\Theta}_{10},0))}{\partial \boldsymbol{\Theta}_2}\right\|\end{aligned} \tag{3.21}$$

Applying the mean value theorem again,

$$\frac{\partial l_n((\boldsymbol{\Theta}_1, \xi))}{\partial \boldsymbol{\Theta}_2} - \frac{\partial l_n((\boldsymbol{\Theta}_1, 0))}{\partial \boldsymbol{\Theta}_2} = \left[\frac{\partial^2 l_n((\boldsymbol{\Theta}_1, \zeta_1))}{\partial^2 \boldsymbol{\Theta}_2}\right] \cdot \xi$$

for some $\|\zeta_1\| \le \|\xi\|$ and

$$\frac{\partial l_n((\boldsymbol{\Theta}_1, 0))}{\partial \boldsymbol{\Theta}_2} - \frac{\partial l_n((\boldsymbol{\Theta}_{10}, 0))}{\partial \boldsymbol{\Theta}_2} = \left[\frac{\partial^2 l_n((\zeta_2, 0))}{\partial \boldsymbol{\Theta}_1, \partial \boldsymbol{\Theta}_2}\right] \cdot (\boldsymbol{\Theta}_1 - \boldsymbol{\Theta}_0),$$

where $\zeta_2 = \boldsymbol{\Theta}_{10} + t \cdot (\boldsymbol{\Theta}_1 - \boldsymbol{\Theta}_{10})$, for some $t \in [0,1]$. Applying these results to 3.21 and using the regularity condition $(A2)$ we have

$$\begin{aligned}
&\left\| \frac{\partial l_n((\boldsymbol{\Theta}_1, \xi))}{\partial \boldsymbol{\Theta}_2} - \frac{\partial l_n((\boldsymbol{\Theta}_{10}, 0))}{\partial \boldsymbol{\Theta}_2} \right\| \\
&\leq \left[\sum_{i=1}^{n} M_{2i}(\mathbf{y}_i, \mathbf{X}_i, \mathbf{Z}_i)] \right] \cdot \|\epsilon\| + \left[\sum_{i=1}^{n} M_{2i}(\mathbf{y}_i, \mathbf{X}_i, \mathbf{Z}_i)] \right] \cdot \|\boldsymbol{\Theta}_1 - \boldsymbol{\Theta}_{10}\| \\
&= O_p(n) \cdot \Big\{ \|\epsilon\| + \|\boldsymbol{\Theta}_1 - \boldsymbol{\Theta}_{10}\| \Big\} \\
&= O_p(n) \cdot \Big\{ O(n^{-1/2}) + O(n^{-1/2}) \Big\} \\
&= O_p(n^{1/2}).
\end{aligned}$$

By the regularity condition, $\dfrac{\partial l_n((\boldsymbol{\Theta}_{10}, 0))}{\partial \boldsymbol{\Theta}_2} = O_p(n^{1/2})$ and hence $\dfrac{\partial l_n((\boldsymbol{\Theta}_1, \xi))}{\partial \boldsymbol{\Theta}_2} = O_p(n^{1/2})$. Applying this result to 3.20, we have

$$l_n((\boldsymbol{\Theta}_1, \boldsymbol{\Theta}_2)) - l_n((\boldsymbol{\Theta}_1, 0)) = O_p(\sqrt{n}) \sum_{k=1}^{K} \left[\sum_{k \in K} \Big\{ \sum_{j \in k} |\beta_{kj}| \Big\}^2 \right],$$

where $j = t_{\beta_k} + 1$, and t_{β_k} is the numbers of true nonzero fixed effects in component k. On the other hand,

$$p_n((\boldsymbol{\Theta}_1, \boldsymbol{\Theta}_2)) - p_n((\boldsymbol{\Theta}_1, 0)) = \sum_{k=1}^{K} \Big(\pi_k \cdot n \cdot p_{\lambda_{kn}}(\beta_{kj}) \Big).$$

Therefore,

$$\begin{aligned}
&\mathbf{Q}_n((\boldsymbol{\Theta}_1, \boldsymbol{\Theta}_2)) - \mathbf{Q}_n((\boldsymbol{\Theta}_1, 0)) \\
&= \sum_{k=1}^{K} \left[\sum_{k \in K} \Big\{ \sum_{j \in k} |\beta_{kj}| \Big\}^2 \cdot O_p(\sqrt{n}) - \pi_k \cdot n \cdot p_{\lambda_{kn}}(\beta_{kj}) \right].
\end{aligned}$$

By condition $(P3)$, it is less than 0 in probability. Therefore,

$$Pr\Big[\mathbf{Q}_n((\boldsymbol{\Theta}_1, \boldsymbol{\Theta}_2)) - \mathbf{Q}_n((\boldsymbol{\Theta}_1, 0)) < 0 \Big] \to_p 1.$$

This completes the proof of (a).

(b) Part (i). Let $(\hat{\boldsymbol{\Theta}}_{1n}, 0)$ be the maximizer of the penalized log-likelihood function $\mathbf{Q}((\boldsymbol{\Theta}_1, 0))$ which is regarded as a function of $\boldsymbol{\Theta}_1$. It suffices to show that in the neighborhood $\|\boldsymbol{\Theta} - \boldsymbol{\Theta}_1\| = O(n^{-1/2})$, $\mathbf{Q}_n((\boldsymbol{\Theta}_1, \boldsymbol{\Theta}_2)) - \mathbf{Q}_n((\hat{\boldsymbol{\Theta}}_{1n}, 0)) < 0$ with probability tending to 1 as $n \to \infty$. We

have that

$$\begin{aligned}
&\mathbf{Q}_n((\boldsymbol{\Theta}_1, \boldsymbol{\Theta}_2)) - \mathbf{Q}_n((\hat{\boldsymbol{\Theta}}_{1n}, 0)) \\
&= \Big\{\mathbf{Q}_n((\boldsymbol{\Theta}_1, \boldsymbol{\Theta}_2)) - \mathbf{Q}_n((\boldsymbol{\Theta}_1, \boldsymbol{\Theta}_0))\Big\} + \Big\{\mathbf{Q}_n((\boldsymbol{\Theta}_1, \boldsymbol{\Theta}_0)) - \mathbf{Q}_n((\hat{\boldsymbol{\Theta}}_{1n}, 0))\Big\} \\
&\leq \mathbf{Q}_n((\boldsymbol{\Theta}_1, \boldsymbol{\Theta}_2)) - \mathbf{Q}_n((\boldsymbol{\Theta}_1, \boldsymbol{\Theta}_0)) < 0
\end{aligned}$$

with probability tending to 1 by (a).

Part (ii). Regard $\mathbf{Q}_n((\boldsymbol{\Theta}_1, 0))$ as a function of $\boldsymbol{\Theta}_1$. Using the same argument as in Theorem 3.3.1, there exists a root-n consistent local maximizer of this function, say $\hat{\boldsymbol{\Theta}}_{1n}$, which satisfies the score-type equation

$$\mathbf{Q}_n'((\hat{\boldsymbol{\Theta}}_{1n}, 0)) = l_n'((\hat{\boldsymbol{\Theta}}_{1n}, 0)) - p_n'((\hat{\boldsymbol{\Theta}}_{1n}, 0)) = 0. \tag{3.22}$$

Since $\hat{\boldsymbol{\Theta}}_{1n}$ is a root-n consistent estimator, by Taylor's expansion around the true value, we have

$$\begin{aligned}
l_n'((\hat{\boldsymbol{\Theta}}_{1n}, 0)) &= l_n'((\boldsymbol{\Theta}_{10}, 0)) + \Big[l_n''((\boldsymbol{\Theta}_{10}, 0)) + o_p(n)\Big](\hat{\boldsymbol{\Theta}}_{1n}, -\boldsymbol{\Theta}_{10}) \\
p_n'((\hat{\boldsymbol{\Theta}}_{1n}, 0)) &= p_n'((\boldsymbol{\Theta}_{10}, 0)) + \Big[p_n''((\boldsymbol{\Theta}_{10}, 0)) + o_p(n)\Big](\hat{\boldsymbol{\Theta}}_{1n}, -\boldsymbol{\Theta}_{10}).
\end{aligned}$$

Substituting them into 3.22, we have

$$\begin{aligned}
&\Big[l_n'((\boldsymbol{\Theta}_{10}, 0)) - p_n'((\boldsymbol{\Theta}_{10}, 0))\Big] \\
&+ \Big[l_n''((\boldsymbol{\Theta}_{10}, 0)) - p_n''((\boldsymbol{\Theta}_{10}, 0)) + o_p(n)\Big](\hat{\boldsymbol{\Theta}}_{1n}, -\boldsymbol{\Theta}_{10}) = 0.
\end{aligned}$$

By rearranging the terms and multiplying both sides of the equation by $n^{-1/2}$, we get

$$\begin{aligned}
&- n^{-\frac{1}{2}}\Big[l_n''((\boldsymbol{\Theta}_{10}, 0)) - p_n''((\boldsymbol{\Theta}_{10}, 0)) + o_p(n)\Big](\hat{\boldsymbol{\Theta}}_{1n}, -\boldsymbol{\Theta}_{10}) \\
&= n^{-\frac{1}{2}}\Big[l_n'((\boldsymbol{\Theta}_{10}, 0)) - p_n'((\boldsymbol{\Theta}_{10}, 0))\Big].
\end{aligned}$$

Then, by the regularity conditions, $-\frac{1}{n}l_n''((\boldsymbol{\Theta}_{10}, 0)) = \mathbf{I}_1(\boldsymbol{\Theta}_{10}) + o_p(1)$ and $\frac{1}{\sqrt{n}}l_n'((\boldsymbol{\Theta}_{10}, 0)) \to_d N(0, \mathbf{I}_1(\boldsymbol{\Theta}_{10}))$, where $\mathbf{I}_1(\boldsymbol{\Theta}_{10})$ is the Fisher information matrix knowing that $\boldsymbol{\Theta}_{20} = 0$. Thus by Slutsky's theorem,

$$\sqrt{n}\Bigg[\Bigg\{\mathbf{I}_1(\boldsymbol{\Theta}_{10}) + \frac{p_n''(\boldsymbol{\Theta}_{10})}{n}\Bigg\}(\hat{\boldsymbol{\Theta}}_1 - \boldsymbol{\Theta}_{10}) + \frac{p_n'(\boldsymbol{\Theta}_{10})}{n}\Bigg] \to_d N(\mathbf{0}, \mathbf{I}_1(\boldsymbol{\Theta}_{10})),$$

where $p_n'(\boldsymbol{\Theta}_{10}) = p_n'((\boldsymbol{\Theta}_{10}, 0))$ and $p_n''(\boldsymbol{\Theta}_{10}) = p_n''((\boldsymbol{\Theta}_{10}, 0))$.

□

3.4 Simulation

We conduct empirical simulations to implement our proposed mixLMM with fixed effects selection across two clusters. The model-based clustering method with variable selection for high-dimensional longitudinal data was recently derived by Yang and Wu, 2022, and they were the first to address the clustering method with simultaneous effects selection. We adapt their method by substituting various penalties in our proposed approach. In this section, we evaluate our method's performance of variable selection using different penalties within the clusters.

We generate the data using the following LMM for cluster k, $k = 1, 2$:

$$\mathbf{y}_{ik} = \mathbf{X}_i\boldsymbol{\beta}_k + \mathbf{Z}_i\mathbf{b}_{ik} + \boldsymbol{\epsilon}_{ik}. \tag{3.23}$$

Group structure correlations in $\mathbf{X}_i$ are generated from a multivariate normal distribution $\mathbf{X}_i \sim N(0, \boldsymbol{\Sigma})$ to mimic the predefined domain-based SDOH data, where Σ is a Toeplitz covariance matrix with correlation entries $\Sigma_{ij} = w^{|i-j|}$ for variables within group correlation, and $\Sigma_{lm} = b^{|l-m|}$ for between group correlation. Here, $(i, j) \in p$ represents the ith and jth components of $\mathbf{X}_i$, while $(l, m) \in g$ denotes the lth and mth groups. The correlation levels w and b are between 0 and 1. The first covariance matrix model sets the constant $w = 0.6$ and $b = 0.6$ to test moderate correlation within and between groups. The second covariance matrix model sets the high correlation within and between groups, $w = 0.9$ and $b = 0.9$. The random errors were generated as $\boldsymbol{\epsilon}_{ik} \sim N(0, \boldsymbol{\sigma}_k^2)$, where

$$\boldsymbol{\sigma}_k^2 = \frac{\frac{1}{n}\sum_{i=1}^{n}\sum_{k=1}^{K}(\mathbf{X}_i\boldsymbol{\beta}_k + \mathbf{Z}_i\mathbf{b}_{ik})^2}{SNR}, \tag{3.24}$$

and the signal-to-noise ratio (SNR) is set to 1 in our simulations.

We set sample sizes of $n = (100, 200, 400)$ with repeated measurements of $m_i = 10$ for each dataset. Each simulated dataset was replicated 100 times. Additionally, we generated datasets with varying p-dimensional predictors. Thus, we considered the following combinations: 1) for $n = 100$, $p = (25, 50, 100)$; 2) for $n = 200$, $p = (25, 50, 100)$; and 3) for $n = 400$, $p = (100, 200, 400)$. The variables are divided into five groups, where $g = 5$, which can be equal-sized or unequal-sized.

For p variables, each variable is assigned a group index corresponding to one of five unique groups. When the groups are of equal size, the covariates are evenly split among all five groups. Conversely, for unequal-sized groups, half of the variables are shared among the five groups, while the other half is divided among three groups.

We examine three scenarios involving true fixed effect parameters: 1) one true non-zero fixed effect at the first index of each group, 2) two non-zero fixed effects at the first two indices for each group, and 3) one or more true non-zero fixed effects in each group. In Scenario 1, each simulated dataset contains five true non-zero parameters; Scenario 2 includes ten; and Scenario 3 has eight true non-zero parameters. These three distinct scenarios, based on the number of non-zero coefficients within each group, allow us

to evaluate how effectively the mixLMM-eLasso addresses both homogeneous and heterogeneous sparsity structures in grouped variable selection. These scenarios increase in complexity, ranging from a simple structure with a single non-zero coefficient per group to more complex situations with varying numbers of non-zero coefficients across structured groups. In the parameter vectors, the intercept β_0 for the fixed effects is set to 1. The true non-zero coefficient is assigned as 1 in cluster 1 and -1 in cluster 2. Therefore, in the first simulation scenario, the first indexed predictors from each group are significant for both equal- and unequal-sized groups:

In cluster 1,

$$\boldsymbol{\beta}_1 = \Big(1, \underbrace{1, 0, \ldots, 0}_{\text{g1}}, \underbrace{1, 0, \ldots, 0}_{\text{g2}}, \underbrace{1, 0, \ldots, 0}_{\text{g3}}, \underbrace{1, 0, \ldots, 0}_{\text{g4}}, \underbrace{1, 0, \ldots, 0}_{\text{g5}}\Big)^T;$$

In cluster 2,

$$\boldsymbol{\beta}_2 = \Big(1, \underbrace{-1, 0, \ldots, 0}_{\text{g1}}, \underbrace{-1, 0, \ldots, 0}_{\text{g2}}, \underbrace{-1, 0, \ldots, 0}_{\text{g3}}, \underbrace{-1, 0, \ldots, 0}_{\text{g4}}, \underbrace{-1, 0, \ldots, 0}_{\text{g5}}\Big)^T.$$

In the second scenario, the first two indices from each group are significant:

$$\boldsymbol{\beta}_1 = \Big(1, \underbrace{1, 1, 0, \ldots, 0}_{\text{g1}}, \underbrace{1, 1, 0, \ldots, 0}_{\text{g2}}, \underbrace{1, 1, 0, \ldots, 0}_{\text{g3}}, \underbrace{1, 1, 0, \ldots, 0}_{\text{g4}}, \underbrace{1, 1, 0, \ldots, 0}_{\text{g5}}\Big)^T,$$

$$\boldsymbol{\beta}_2 = \Big(1, \underbrace{-1, -1, 0, \ldots, 0}_{\text{g1}}, \underbrace{-1, -1, 0, \ldots, 0}_{\text{g2}}, \underbrace{-1, -1, 0, \ldots, 0}_{\text{g3}}, \underbrace{-1, -1, 0, \ldots, 0}_{\text{g4}}, \underbrace{-1, -1, 0, \ldots, 0}_{\text{g5}}\Big)^T.$$

In the third scenario, we set group 1 to have three non-zero parameters, group 2 to have two non-zero parameters, and the other three groups to each have exactly one non-zero parameter:

$$\boldsymbol{\beta}_1 = \Big(1, \underbrace{1, 1, 1, 0, \ldots, 0}_{\text{g1}}, \underbrace{1, 1, 0, \ldots, 0}_{\text{g2}}, \underbrace{1, 0, \ldots, 0}_{\text{g3}}, \underbrace{1, 0, \ldots, 0}_{\text{g4}}, \underbrace{1, 0, \ldots, 0}_{\text{g5}}\Big)^T,$$

$$\boldsymbol{\beta}_2 = \Big(1, \underbrace{-1, -1, -1, 0, \ldots, 0}_{\text{g1}}, \underbrace{-1, -1, 0, \ldots, 0}_{\text{g2}}, \underbrace{-1, 0, \ldots, 0}_{\text{g3}}, \underbrace{-1, 0, \ldots, 0}_{\text{g4}}, \underbrace{-1, 0, \ldots, 0}_{\text{g5}}\Big)^T.$$

The random effects for each group are represented by $\mathbf{b}_i = (b_{i0}, 0, ..., 0)$, where $b_{i0} \sim N(0, D)$. The intercepts b_{i0} for the random effects are sampled from $N(0, 1)$ for cluster 1 and $N(0, 2)$ for cluster 2, respectively. These intercepts are not subject to penalization during estimation.

The results are evaluated regarding the performance of clustering, variable selection, and parameter estimation. We compared the performance of our proposed method (mixLMM) after applying various

penalty functions, including Exclusive Lasso (mixLMM-eLasso), SCAD (mixLMM-SCAD), and Lasso (mixLMM-Lasso). Let λ_n and λ represent the tuning parameters in Lasso and SCAD, respectively. The following penalty functions are added to our proposed approach for comparison:

For Lasso,

$$P_{\lambda_n} \sum_{j=1}^{p} |\beta_j| \leq s,$$

where s is a predetermined free parameter that defines the level of regularization;

For SCAD,

$$P'_{\lambda}(|\beta_j|) = \lambda \Big\{ I(|\beta_j| \leq \lambda) + \frac{(a\lambda - |\beta|_j)^+}{(a-1)\lambda} I(|\beta_j| > \lambda) \Big\}, \quad a > 2,$$

where a is a known constant (J. Fan & Li, 2001), which can be set to 3.7 in common applications, and $I(\cdot)$ is a set indicator function.

We evaluate clustering accuracy by calculating the adjusted Rand Index (ARI), which measures the similarity between two cluster partitions while considering the probability of chance clustering (Hubert & Arabie, 1985). Given a set S of n elements, the ARI between two clusterings, $U = \{U_1, U_2, \ldots, U_r\}$ and $V = \{V_1, V_2, \ldots, V_s\}$, is defined as (Hubert & Arabie, 1985; Steinley, 2004)

$$\text{ARI} = \frac{\sum_{ij} \binom{n_{ij}}{2} - \left[\sum_i \binom{a_i}{2} \sum_j \binom{b_j}{2}\right] / \binom{n}{2}}{\frac{1}{2}\left[\sum_i \binom{a_i}{2} + \sum_j \binom{b_j}{2}\right] - \left[\sum_i \binom{a_i}{2} \sum_j \binom{b_j}{2}\right] / \binom{n}{2}},$$

where n_{ij} denotes the number of observations in common between U_i and V_j, a_i and b_j are the sizes of clusters U_i and V_j, respectively, and n is the total number of observations. The ARI ranges from -1 to 1, with higher values reflecting a more favorable clustering outcome. Additionally, the performance of selection methods is measured using (C, IC), where C represents the number of zero coefficients correctly estimated as zero, and IC is the number of non-zero coefficients incorrectly estimated as zero. At last, Mean squared error, MSE=$\| \hat{\boldsymbol{\beta}}_{pk} - \boldsymbol{\beta} \|^2$, is used to measure the performance of parameter estimates, where $\boldsymbol{\beta}_{pk}$ denotes the penalized estimates, and $\boldsymbol{\beta}$ denotes the true value of the estimates. Thus, we have

$$\text{MSE} = \frac{\sum_{k=1}^{K} \sum_{j=1}^{p} \| \hat{\boldsymbol{\beta}}_{jk} - \boldsymbol{\beta}_{jk} \|_2^2}{\sum_{k=1}^{K} \sum_{j=1}^{p} \| \boldsymbol{\beta}_{jk} \|_2^2}.$$

3.5 Results

We compared the performance of the proposed method after applying different penalties across various data settings characterized by sample size n, the number of variables p, and correlation structures. Additionally, we considered three scenarios (scn) involving true parameters in equal-sized and unequal-sized groups to assess the effectiveness of variable selection within the specified domains.

We first examined three scenarios that represent the data design with $n = 100$ and $p = (25, 50, 100)$ corresponding to $n/4$, $n/2$, and n. The moderate correlation within and between groups was set to 0.6, while the high correlation was set to 0.9. Each clustering method with Exclusive Lasso (mixLMM-eLasso), SCAD (mixLMM-SCAD), and Lasso (mixLMM-Lasso) was summarized using the average values of ARI, C, IC, and MSE (Bold values indicate the best ARI, C, IC, and MSE in the simulation).

In Table 3.1, the results showed that the mixLMM via Exclusive Lasso method consistently outperforms all other approaches across various scenarios and metrics, regardless of the correlation level or the dimensionality of the data. Its ability to balance clustering accuracy, sparsity, and stability made it a preferred choice for analyzing high-dimensional data. The mixLMM-eLasso achieved the highest ARI in all scenarios, demonstrating robust clustering performance across low and high correlation levels and varying dimensionalities. The mixLMM-SCAD and mixLMM-Lasso showed competitive ARI in some cases; however, their performance deteriorated as the dimensionality increased, especially in high-dimensional settings. Moreover, the IC and MSE were minimal across all scenarios in the mixLMM-eLasso, indicating its effectiveness in achieving accurate variable selection and parameter estimation while maintaining sparsity. The mixLMM-SCAD and, occasionally, mixLMM-Lasso performed better at accurately identifying true zero coefficients (C) and increased IC and MSE, particularly as p increases. They may tend to shrink coefficients to zero incorrectly, especially when the correlation between covariates is strong. This suggested that both methods are less effective at handling high-dimensional data compared to mixLMM-eLasso. The performance gap between mixLMM-eLasso and the other methods became more pronounced under a strong correlation (Corr = 0.9).

Table 3.2 displayed the results of variable selection and prediction error for a dataset with $n = 200$ and $p = (25, 50, 100)$. This table followed the same setup as Table 3.1, which highlighted the findings for $n = 100$. The mixLMM-eLasso consistently outperformed across all metrics, achieving the best ARI, lowest IC, and lowest MSE in every scenario and correlation level. In Table 3.2, mixLMM-eLasso demonstrated a slight improvement in ARI, reduced IC, and lower MSE compared to Table 3.1, highlighting the advantages of a larger sample size $(n = 200)$. While both mixLMM-SCAD and mixLMM-Lasso showed advantages from the expanded sample size in Table 3.2, these gains were less significant compared to mixLMM-eLasso's improvements. When two or more than one non-zero coefficients were allocated to each of the five equal-sized groups, mixLMM-eLasso remained the leading method in both Table 3.1 and 3.2, maintaining consistent clustering accuracy and sparsity across all values of p and $Corr$.

The results presented in Table 3.3 illustrated the performance of various methods on larger datasets with $n = 400$ and $p = (100, 200, 400)$. The table indicated that mixLMM-eLasso consistently outperformed mixLMM-SCAD and mixLMM-Lasso, achieving the highest ARI while minimizing IC and MSE. This demonstrated that mixLMM-eLasso maintained stable clustering accuracy and parameter estimation as the dimensionality increased. While mixLMM-SCAD and mixLMM-Lasso showed marginal improvements compared to smaller sample sizes $n = (100, 200)$, they still fell short of mixLMM-eLasso's performance in high-dimensional settings. In certain scenarios, mixLMM-SCAD provided better indications of true zero parameters (C) than mixLMM-Lasso; however, it struggled with higher values of IC and MSE as p and correlation increased. This may have been due to their penalty, which tends to

excessively shrink coefficients to zero and, therefore, incorrectly sets them to zero when the correlation between covariates is high and dimensionality increases (J. Fan & Lv, 2010; Zou & Hastie, 2005).

Thus, overall, an increase in n enhanced the statistical efficiency of all methods; however, the performance gap between mixLMM-eLasso and the other methods widened as p and correlation increased. These results highlighted the effectiveness of mixLMM-eLasso in high-dimensional data analysis by utilizing its simultaneous clustering and variable selection framework to attain superior performance across various scenarios. Additionally, we presented further simulations with unequal group sizes in the Appendix (see Tables B.1, B.2, and B.3), using the same data setup as for the equal group sizes. The results were shown in Table B.1 with $n = 100$, Table B.2 with $n = 200$, and Table B.3 with $n = 400$. These simulations yielded results similar to those of the equal-sized groups, where the mixLMM-eLasso outperformed all other methods.

Table 3.1: Comparison results averaged by ARI, C, IC, and MSE among the proposed methods(mixLMM) through Exclusive Lasso (mixLMM-eLasso), SCAD (mixLMM-SCAD), and Lasso (mixLMM-Lasso) using datasets for n=200, p=(25, 50 100) in equal-sized group.

Scn	method	Corr	ARI	C	IC	MSE	ARI	C	IC	MSE	ARI	C	IC	MSE
	mixLMM-		p=25				p=50				p=100			
1	eLasso	0.6	**0.939**	37.99	**0.00**	**0.011**	**0.940**	**88.07**	**0.00**	**0.006**	**0.941**	187.84	**0.00**	**0.003**
	SCAD		0.926	**38.03**	0.51	0.028	0.852	87.65	3.32	0.040	0.674	187.37	4.41	0.026
	Lasso		0.926	37.29	0.90	0.032	0.877	87.77	3.70	0.043	0.756	**188.59**	5.49	0.030
	eLasso	0.9	**0.947**	37.75	**0.00**	**0.016**	**0.949**	**87.76**	**0.06**	**0.008**	**0.949**	**187.66**	**0.00**	**0.004**
	SCAD		0.926	**38.29**	3.42	0.146	0.912	87.57	4.11	0.085	0.891	187.64	5.03	0.049
	Lasso		0.939	38.04	0.34	0.052	0.927	87.64	0.92	0.030	0.910	187.29	1.94	0.019
2	eLasso	0.6	**0.949**	26.29	**0.00**	**0.038**	**0.947**	76.52	**0.00**	**0.020**	**0.936**	175.42	**0.17**	**0.011**
	SCAD		0.916	**26.34**	3.55	0.136	0.862	74.46	6.53	0.105	0.758	173.59	10.12	0.074
	Lasso		0.922	26.23	3.02	0.104	0.889	**77.07**	6.15	0.082	0.747	**177.59**	11.97	0.067
	eLasso	0.9	**0.960**	27.43	**1.13**	**0.136**	**0.962**	77.69	**1.07**	**0.066**	**0.952**	**177.37**	**1.02**	**0.033**
	SCAD		0.939	**28.52**	11.22	0.584	0.929	76.60	11.25	0.310	0.896	174.61	12.45	0.172
	Lasso		0.958	27.81	2.75	0.214	0.955	**77.83**	3.63	0.113	0.931	176.79	4.74	0.059
3	eLasso	0.6	**0.945**	24.36	**1.05**	**0.036**	**0.947**	73.74	**1.07**	**0.021**	**0.937**	173.54	**1.19**	**0.012**
	SCAD		0.912	**28.97**	3.57	0.074	0.900	77.51	5.33	0.057	0.659	177.68	8.96	0.052
	Lasso		0.929	28.17	3.36	0.071	0.910	**79.33**	6.36	0.066	0.768	**180.46**	9.53	0.049
	eLasso	0.9	**0.957**	26.99	**1.49**	**0.111**	**0.957**	77.42	**1.30**	**0.054**	**0.957**	177.66	**1.31**	**0.027**
	SCAD		0.930	**30.96**	8.42	0.352	0.930	**79.75**	9.18	0.201	0.907	178.31	9.98	0.117
	Lasso		0.945	28.91	2.62	0.140	0.942	79.24	3.70	0.080	0.935	**178.41**	4.14	0.040

Note: Bold symbols in the ARI and C indicate the highest values, while the lowest values for IC and MSE are also bolded. The optimal values of C vary by scenario and are determined by the 2×(p - number of non-zeros). Thus, for p = 25, C = (45, 40, 42); for p = 50, C = (95, 90, 92); for p = 100, C = (195, 190 192). The optimal value for IC is zero; The method denotes mixLMM with tested penalties.

Table 3.2: Comparison results averaged by ARI, C, IC, and MSE among the proposed methods(mixLMM) through Exclusive Lasso (mixLMM-eLasso), SCAD (mixLMM-SCAD), and Lasso (mixLMM-Lasso) using datasets for n=200, p=(25, 50 100) in equal-sized group.

Scn	method	Corr	ARI	C	IC	MSE	ARI	C	IC	MSE	ARI	C	IC	MSE
	mixLMM-		p=25				p=50				p=100			
1	eLasso	0.6	**0.936**	38.69	**0.00**	**0.005**	**0.940**	**88.61**	**0.00**	**0.003**	**0.939**	188.38	**0.00**	**0.001**
	SCAD		0.923	**38.92**	0.30	0.011	0.913	87.87	1.00	0.013	0.837	**188.49**	3.03	0.016
	Lasso		0.924	37.46	0.50	0.016	0.915	86.08	1.40	0.017	0.877	185.10	3.20	0.017
	eLasso	0.9	**0.947**	38.22	**0.00**	**0.008**	**0.952**	87.92	**0.00**	**0.004**	**0.946**	**188.04**	**0.00**	**0.002**
	SCAD		0.923	**38.69**	1.69	0.069	0.931	**88.06**	2.34	0.047	0.913	187.22	3.01	0.029
	Lasso		0.945	38.14	0.02	0.024	0.949	87.37	0.02	0.013	0.933	186.76	0.51	0.008
2	eLasso	0.6	**0.955**	27.63	**0.00**	**0.018**	**0.947**	**77.27**	**0.00**	**0.010**	**0.955**	**176.94**	**0.00**	**0.005**
	SCAD		0.949	**28.21**	0.54	0.033	0.926	76.19	2.62	0.040	0.917	174.37	4.36	0.031
	Lasso		0.951	27.18	0.40	0.020	0.937	76.08	1.99	0.031	0.922	176.24	5.80	0.034
	eLasso	0.9	**0.961**	28.04	**0.18**	**0.074**	**0.956**	**77.57**	**0.22**	**0.037**	**0.964**	**177.97**	**0.16**	**0.018**
	SCAD		0.945	**28.95**	8.81	0.377	0.934	77.29	9.14	0.209	0.934	176.09	10.41	0.121
	Lasso		0.959	28.20	0.85	0.114	0.954	77.29	1.01	0.063	0.957	177.00	0.95	0.032
3	eLasso	0.6	**0.946**	24.41	**0.98**	**0.018**	**0.943**	73.64	**0.89**	**0.011**	**0.939**	172.78	**0.93**	**0.007**
	SCAD		0.943	**30.92**	2.39	0.021	0.933	**79.68**	2.57	0.013	0.858	**179.35**	5.18	0.022
	Lasso		0.944	28.57	1.91	0.020	0.935	78.45	2.80	0.021	0.863	178.64	7.15	0.034
	eLasso	0.9	**0.956**	26.03	**0.97**	**0.070**	**0.954**	75.91	**0.83**	**0.037**	**0.955**	176.47	**0.72**	**0.019**
	SCAD		0.945	**31.18**	6.90	0.220	0.939	**79.74**	7.27	0.129	0.937	**179.03**	8.65	0.080
	Lasso		0.955	28.67	1.60	0.073	0.950	78.10	1.72	0.040	0.948	177.77	1.71	0.021

Note: Bold symbols in the ARI and C indicate the highest values, while the lowest values for IC and MSE are also bolded. The optimal values of C vary by scenario and are determined by the 2×(p - number of non-zeros). Thus, for p = 25, C = (45, 40, 42); for p = 50, C = (95, 90, 92); for p = 100, C = (195, 190 192). The optimal value for IC is zero; The method denotes mixLMM with tested penalties.

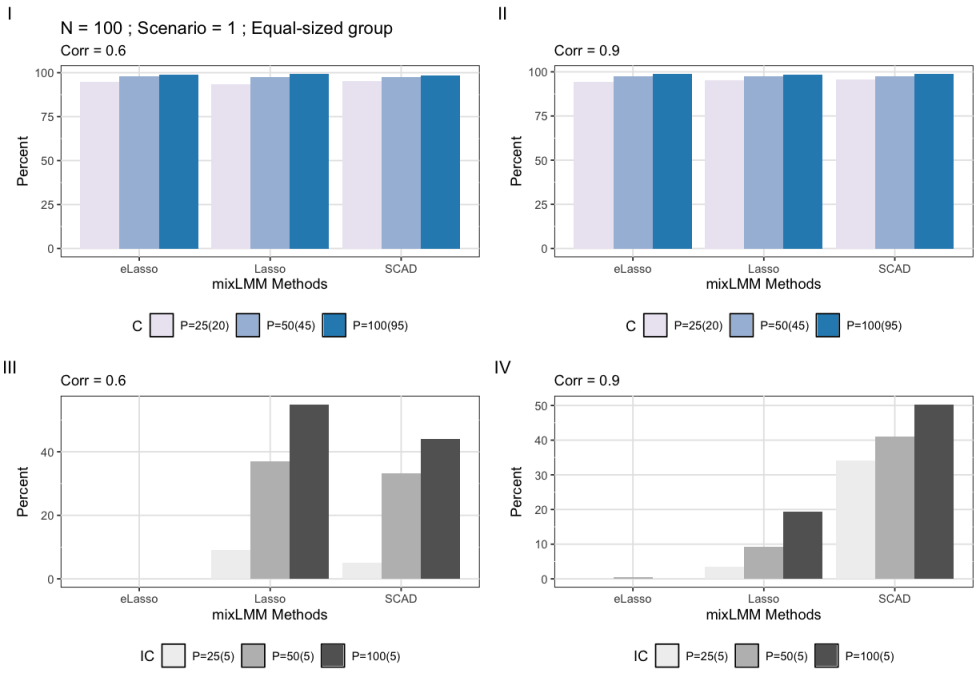
Table 3.3: Comparison results averaged by ARI, C, IC, and MSE among the proposed methods(mixLMM) through Exclusive Lasso (mixLMM-eLasso), SCAD (mixLMM-SCAD), and Lasso (mixLMM-Lasso) using datasets for n=400, p=(100, 200 400) in equal-sized group.

	method		p=100				p=200				p=400			
Scn	**mixLMM-**	Corr	ARI	C	IC	MSE	ARI	C	IC	MSE	ARI	C	IC	MSE
1	eLasso	0.6	**0.939**	**188.82**	**0.00**	**0.001**	**0.942**	**388.81**	**0.00**	**0.000**	**0.944**	**788.75**	**0.00**	**0.000**
	SCAD		0.899	186.86	1.30	0.007	0.879	387.38	2.80	0.007	0.827	788.15	3.50	0.005
	Lasso		0.909	182.46	1.50	0.008	0.886	382.56	3.20	0.008	0.824	781.45	4.59	0.006
	eLasso	0.9	**0.952**	**188.38**	**0.00**	**0.001**	**0.950**	**388.44**	**0.00**	**0.001**	**0.950**	**788.78**	**0.00**	**0.000**
	SCAD		0.934	187.49	0.84	0.009	0.934	386.58	1.67	0.008	0.902	784.90	2.04	0.004
	Lasso		0.945	186.45	**0.00**	0.003	0.931	385.65	0.70	0.003	0.904	784.63	1.89	0.003
2	eLasso	0.6	**0.957**	**177.59**	**0.00**	**0.003**	**0.955**	**377.43**	**0.00**	**0.001**	**0.956**	**777.42**	**0.20**	**0.001**
	SCAD		0.938	176.61	1.60	0.010	0.892	375.58	4.62	0.013	0.834	777.24	7.81	0.010
	Lasso		0.940	174.12	2.20	0.014	0.896	372.88	6.80	0.018	0.783	773.80	11.60	0.015
	eLasso	0.9	**0.964**	**177.89**	**0.01**	**0.009**	**0.963**	**378.48**	**0.00**	**0.005**	**0.966**	**778.38**	**0.03**	**0.002**
	SCAD		0.949	177.17	7.77	0.077	0.956	376.59	8.93	0.046	0.934	775.67	10.52	0.027
	Lasso		0.956	176.59	0.17	0.016	0.951	376.02	0.48	0.009	0.930	774.15	3.31	0.008
3	eLasso	0.6	**0.946**	172.35	**0.81**	**0.004**	**0.948**	372.93	**0.83**	**0.002**	**0.946**	771.99	**0.92**	**0.001**
	SCAD		0.937	**179.16**	2.56	0.005	0.904	**377.93**	5.36	0.010	0.803	**780.59**	8.29	0.009
	Lasso		0.920	173.81	3.58	0.013	0.881	372.68	6.04	0.013	0.824	777.23	9.97	0.012
	eLasso	0.9	**0.958**	175.59	**0.62**	0.013	**0.959**	376.07	**0.54**	0.007	**0.960**	776.00	**0.53**	**0.004**
	SCAD		0.946	**179.17**	5.76	0.040	0.942	**378.40**	7.09	0.026	0.934	**778.39**	8.50	0.017
	Lasso		0.955	177.82	1.67	**0.011**	0.946	375.03	1.38	**0.005**	0.930	773.53	3.40	0.005

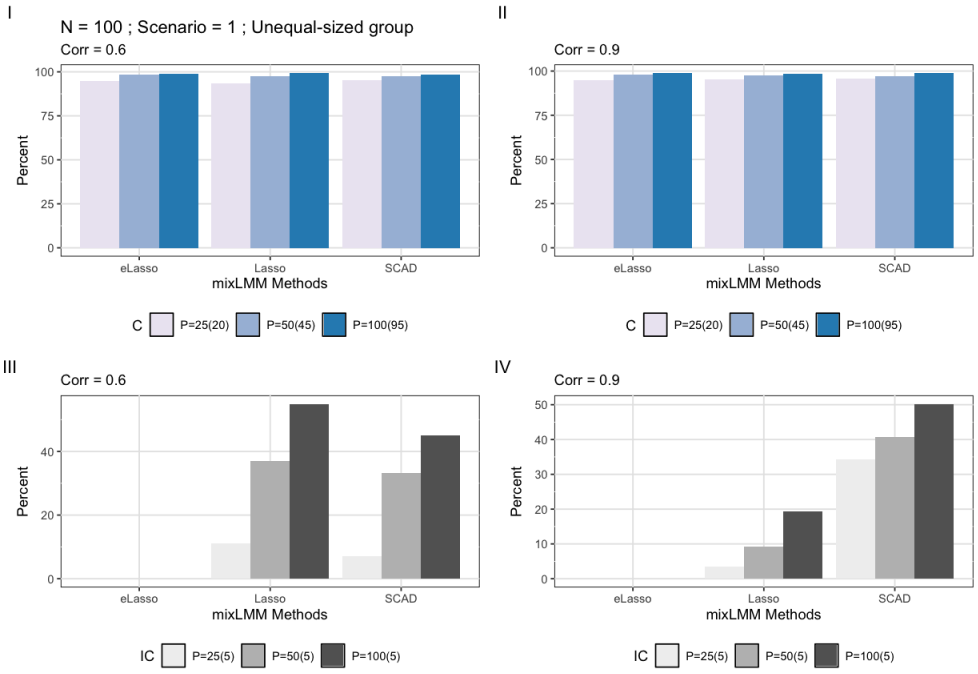
Note: Bold symbols in the ARI and C indicate the highest values, while the lowest values for IC and MSE are also bolded. The optimal values of C vary by scenario and are determined by the 2×(p - number of non-zeros). Thus, for p = 100, C = (190, 180, 184); for p = 200, C = (390, 380, 384); for p = 400, C = (790, 780 784). The optimal value for IC is zero; The method denotes mixLMM with tested penalties.

For $n = 100$, we further illustrated the testing performance of robust sparsity for accurate variable selection across various group structures using the C and IC criteria, as shown in Figures 3.1, 3.2, and 3.3, based on the results from Table 3.1 and Appendix Table B.1. The upper panels showed results for equal-sized groups (from Table 3.1), while the lower panels depicted results for unequal-sized groups (from Appendix Table B.1). Our observations indicated that the method by mixLMM-eLasso consistently demonstrated superior performance, achieving higher C values and lower IC values, particularly in high correlation (Corr = 0.9). This underscored the robustness of mixLMM-eLasso in accurately identifying relevant variables while minimizing classification errors. In contrast, mixLMM-SCAD and mixLMM-Lasso showed reduced performance compared to mixLMM-eLasso, exhibiting higher IC values and greater variability, especially under medium correlation (Corr $= 0.6$). The mixLMM-eLasso maintained its advantage for unequal-sized groups, although the performance gap narrowed, reflecting the challenges of variable selection in scenarios with group imbalance.

Similar to the findings for $n = 100$, the analysis for $n = 200$ and $n = 400$ (as shown in Figures 3.4, 3.5, 3.6, and Figures 3.7, 3.8, 3.9, respectively) confirmed the consistent superiority of mixLMM-eLasso in both equal- and unequal-sized groups. It attained a higher C criterion and a lower IC criterion, particularly under high correlation (Corr = 0.9), demonstrating robust classification accuracy with minimal variability. While both mixLMM-SCAD and mixLMM-Lasso also showed improvements with larger sample sizes, they continued to exhibit greater variability and higher IC values, especially under high correlation (Corr = 0.9). MixLMM-eLasso maintained its advantage in unequal-sized groups as well, highlighting its scalability and robustness across different group structures and sample sizes.

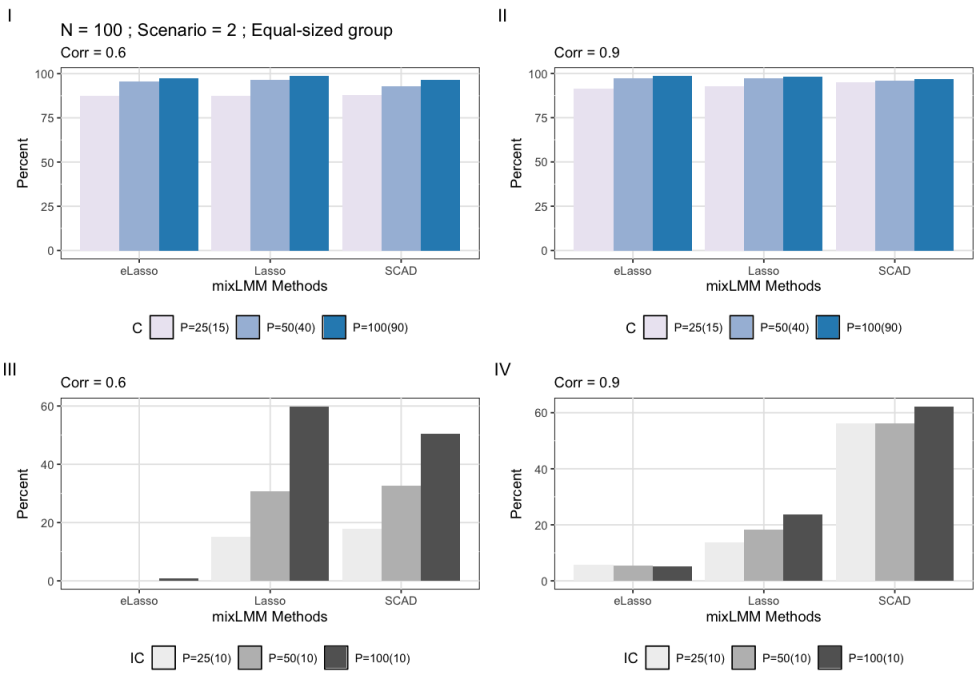


(a) Equal-Sized Group in Scenario 1, n=100, p = (25, 50, 100)

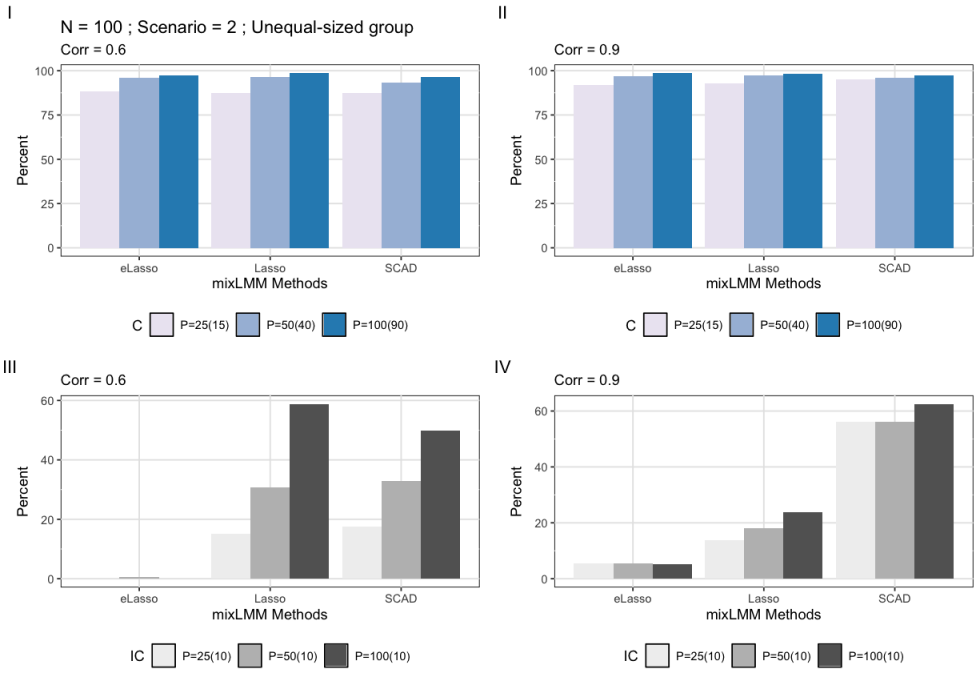


(b) Unequal-Sized Group in Scenario 1, n=100, p = (25, 50, 100)

Figure 3.1: Model-based Clustering with various penalties: 1) Exclusive Lasso (mixLMM-eLasso), 2) SCAD (mixLMM-SCAD), and 3) Lasso (mixLMM-Lasso) tested in Scenario 1 with different data settings. Corr = 0.60 (Left), 0.90 (Right); n = 100; p = (25, 50, 100). The number of true zero and non-zero coefficients in a cluster are indicated next to *p* for C and IC, respectively.
Note: On the x-axis, the penalties illustrate model-based clustering with the implemented penalties for comparisons.

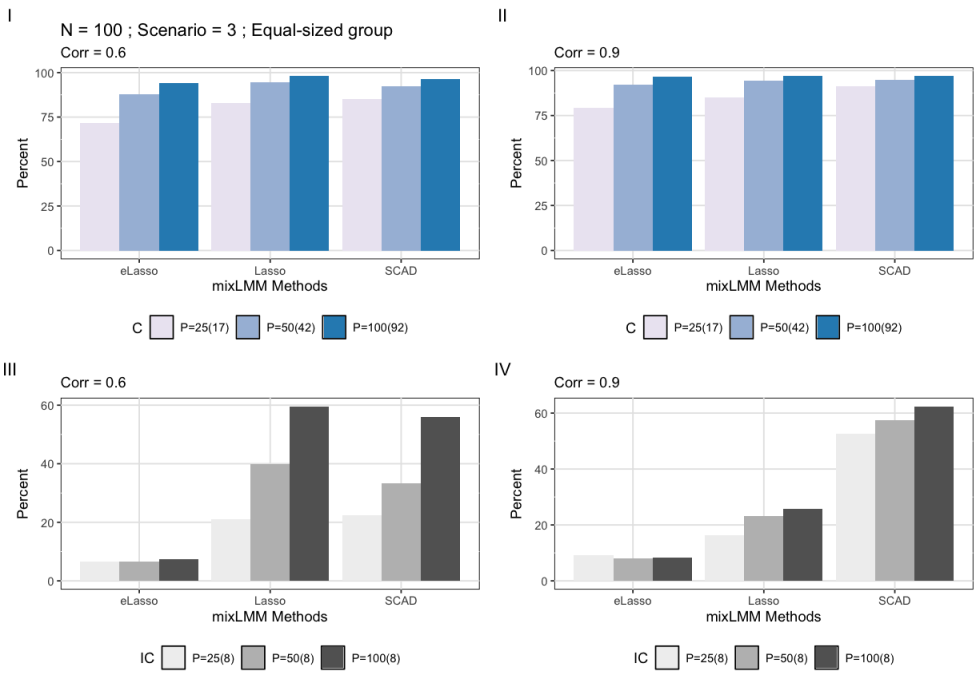


(a) Equal-Sized Group in Scenario 2, n=100, p = (25, 50, 100)

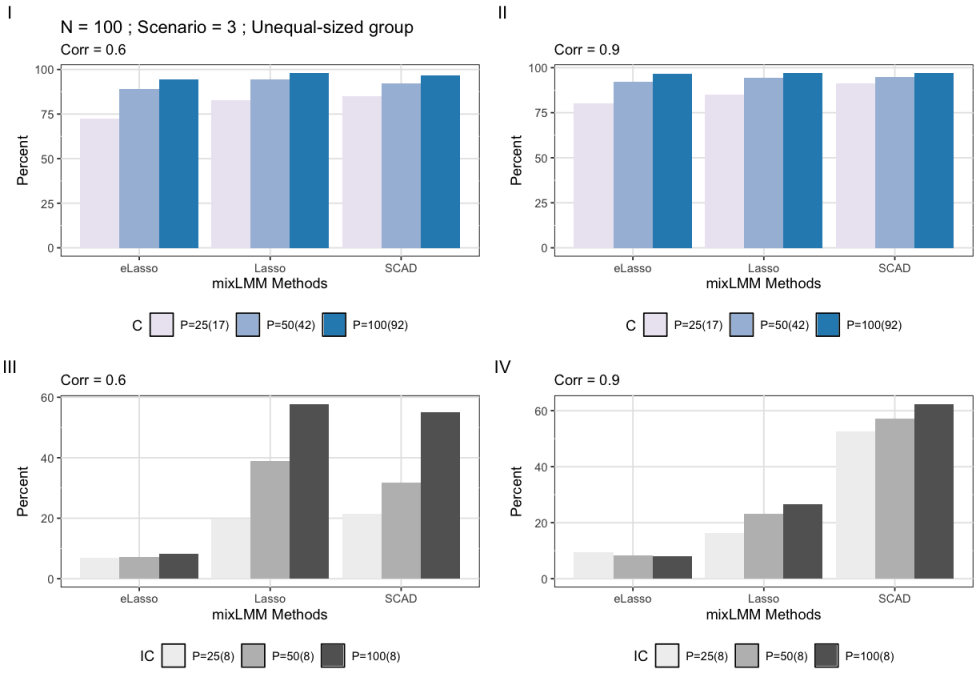


(b) Unequal-Sized Group in Scenario 2, n=100, p = (25, 50, 100)

Figure 3.2: Model-based Clustering with various penalties: 1) Exclusive Lasso (mixLMM-eLasso), 2) SCAD (mixLMM-SCAD), and 3) Lasso (mixLMM-Lasso) tested in Scenario 2 with different data settings. Corr = 0.60 (Left), 0.90 (Right); n = 100; p = (25, 50, 100). The number of true zero and non-zero coefficients in a cluster in a cluster is indicated next to *p* for C and IC, respectively.
Note: On the x-axis, the penalties illustrate model-based clustering with the implemented penalties for comparisons.

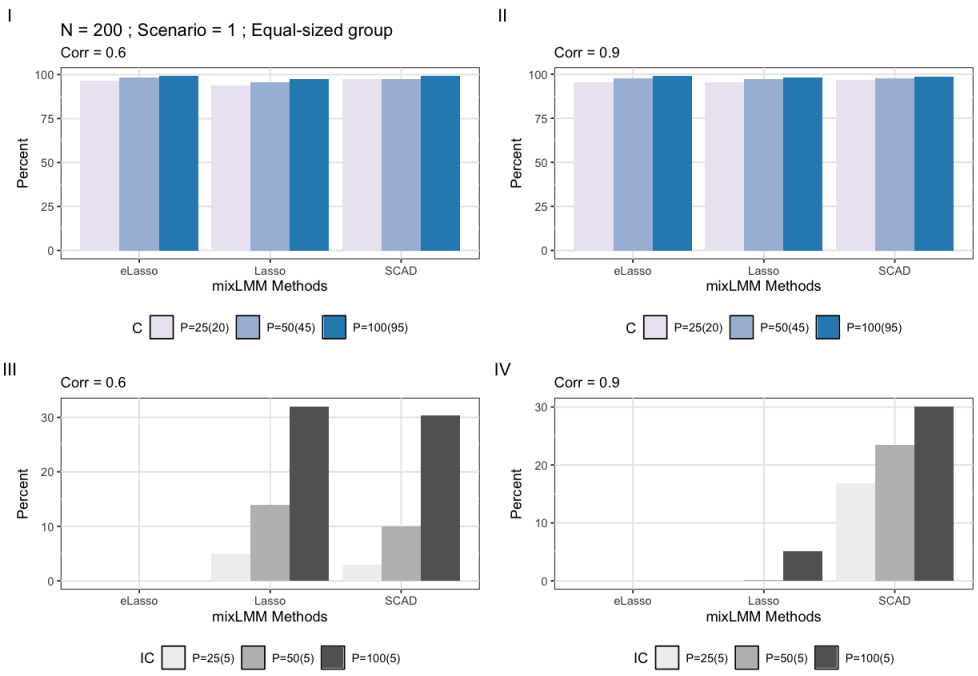


(a) Equal-Sized Group in Scenario 3, n=100, p = (25, 50, 100)

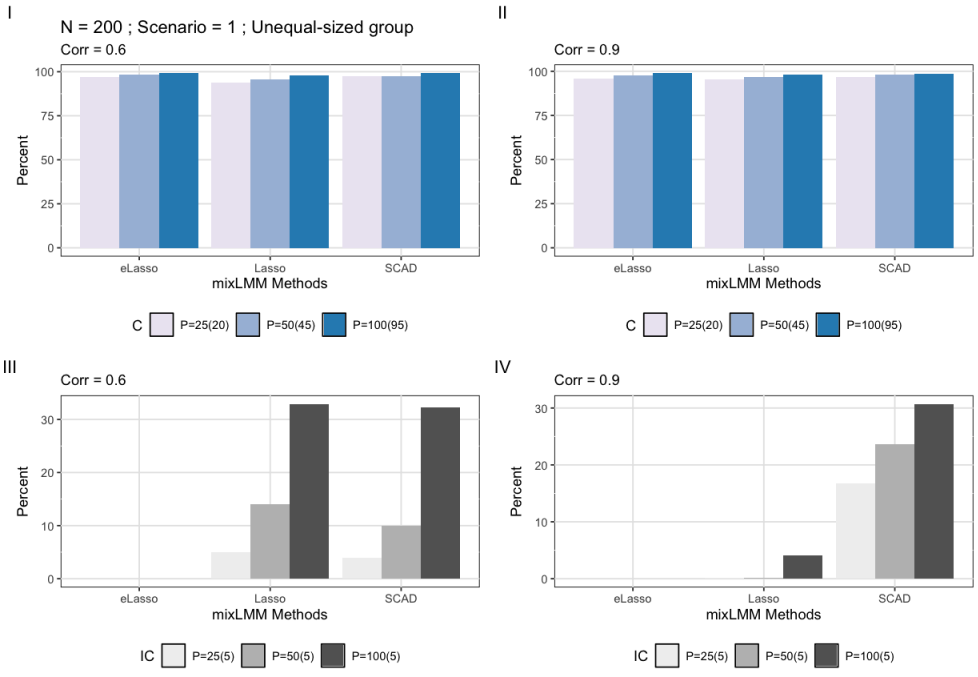


(b) Unequal-Sized Group in Scenario 3, n=100, p = (25, 50, 100)

Figure 3.3: Model-based Clustering with various penalties: 1) Exclusive Lasso (mixLMM-eLasso), 2) SCAD (mixLMM-SCAD), and 3) Lasso (mixLMM-Lasso) tested in Scenario 3 with different data settings. Corr = 0.60 (Left), 0.90 (Right); n = 100; p = (25, 50, 100). The number of true zero and non-zero coefficients in a cluster are indicated next to *p* for C and IC, respectively.
Note: On the x-axis, the penalties illustrate model-based clustering with the implemented penalties for comparisons.

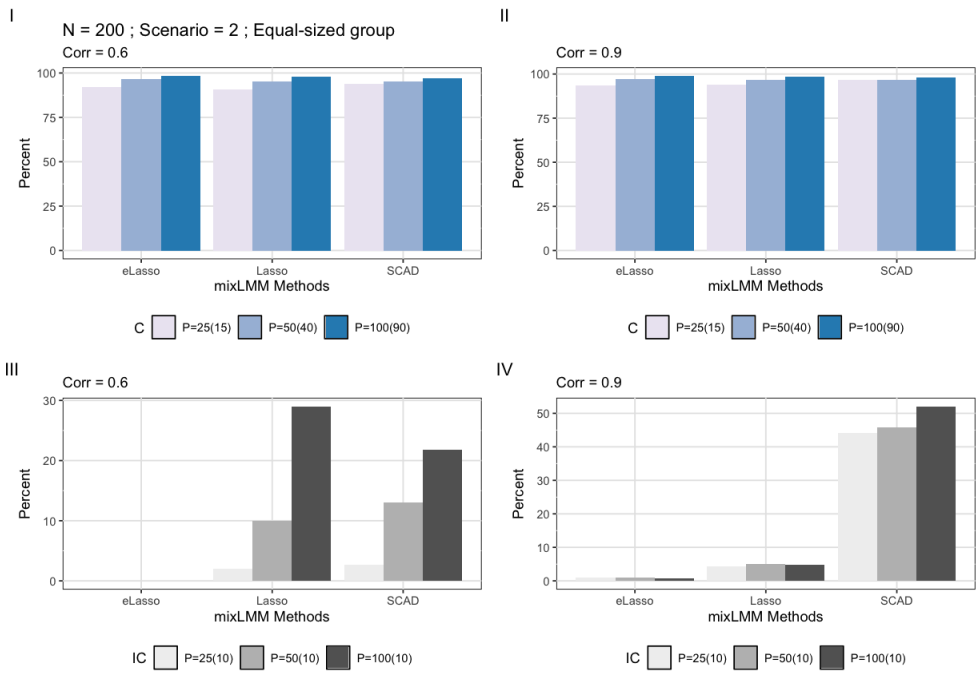


(a) Equal-Sized Group in Scenario 1, n=200, p = (25, 50, 100)

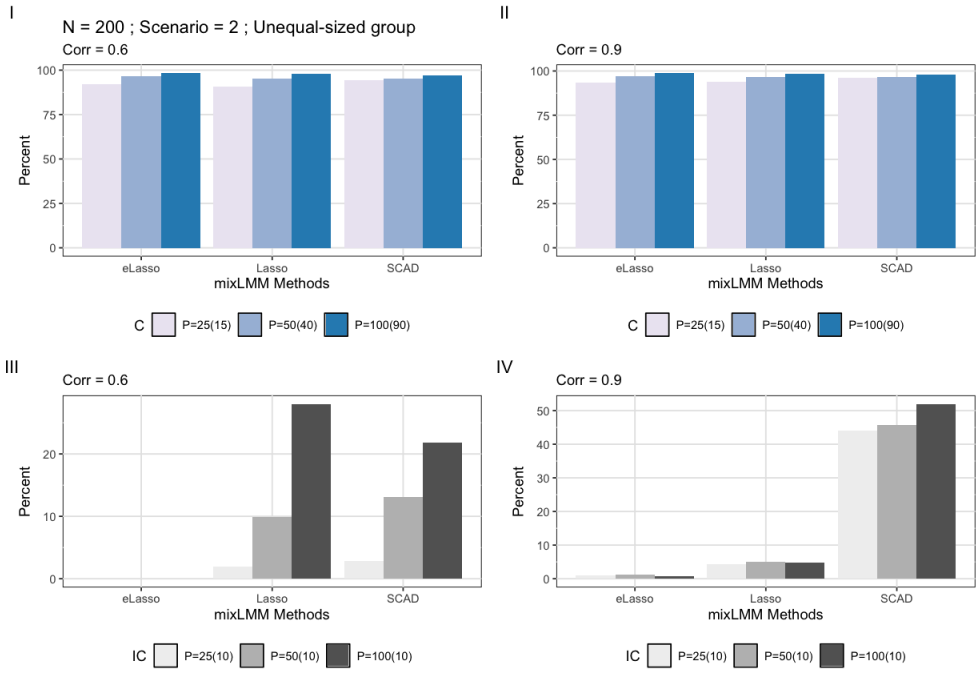


(b) Unequal-Sized Group in Scenario 1, n=200, p = (25, 50, 100)

Figure 3.4: Model-based Clustering with various penalties: 1) Exclusive Lasso (mixLMM-eLasso), 2) SCAD (mixLMM-SCAD), and 3) Lasso (mixLMM-Lasso) tested in Scenario 1 with different data settings. Corr = 0.60 (Left), 0.90 (Right); n = 200; p = (25, 50, 100). The number of true zero and non-zero coefficients in a cluster are indicated next to *p* for C and IC, respectively.
Note: On the x-axis, the penalties illustrate model-based clustering with the implemented penalties for comparisons.

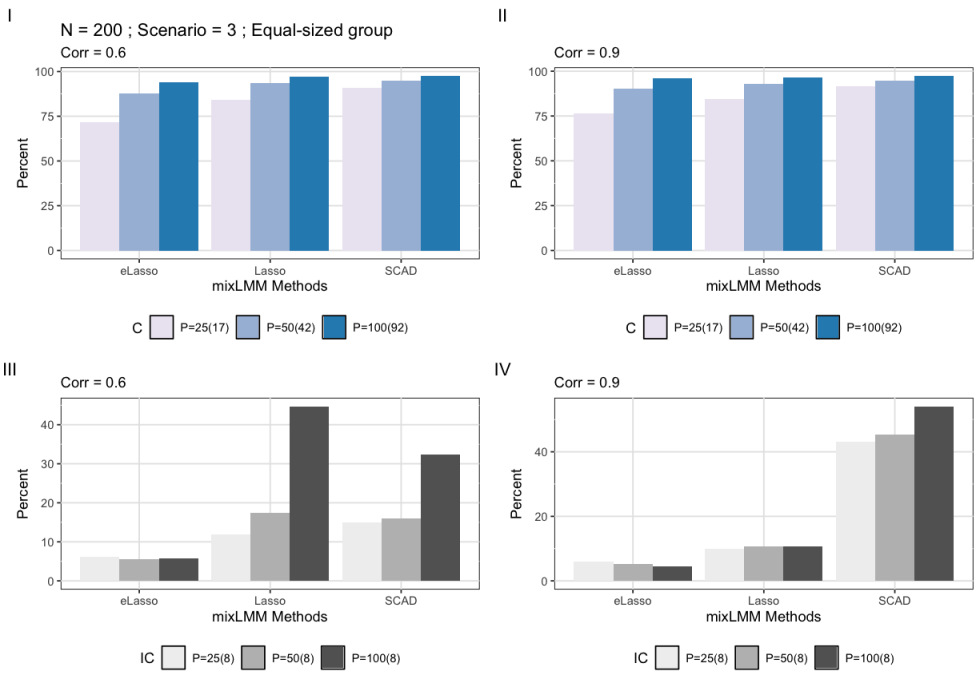


(a) Equal-Sized Group in Scenario 2, n=200, p = (25, 50, 100)

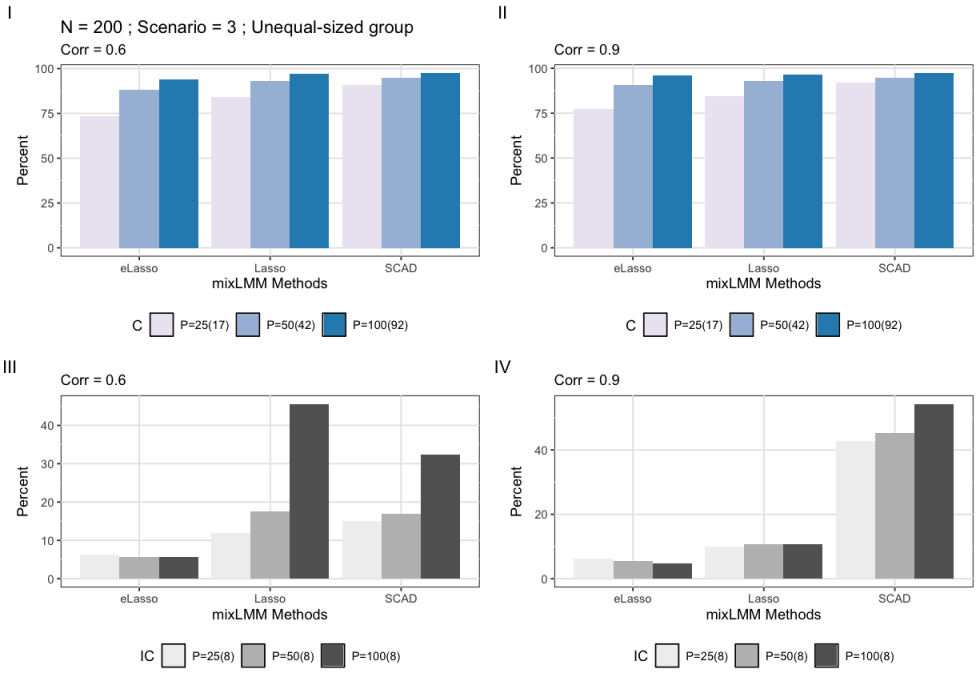


(b) Unequal-Sized Group in Scenario 2, n=200, p = (25, 50, 100)

Figure 3.5: Model-based Clustering with various penalties: 1) Exclusive Lasso (mixLMM-eLasso), 2) SCAD (mixLMM-SCAD), and 3) Lasso (mixLMM-Lasso) tested in Scenario 2 with different data settings. Corr = 0.60 (Left), 0.90 (Right); n = 200; p = (25, 50, 100). The number of true zero and non-zero coefficients in a cluster are indicated next to *p* for C and IC, respectively.
Note: On the x-axis, the penalties illustrate model-based clustering with the implemented penalties for comparisons.

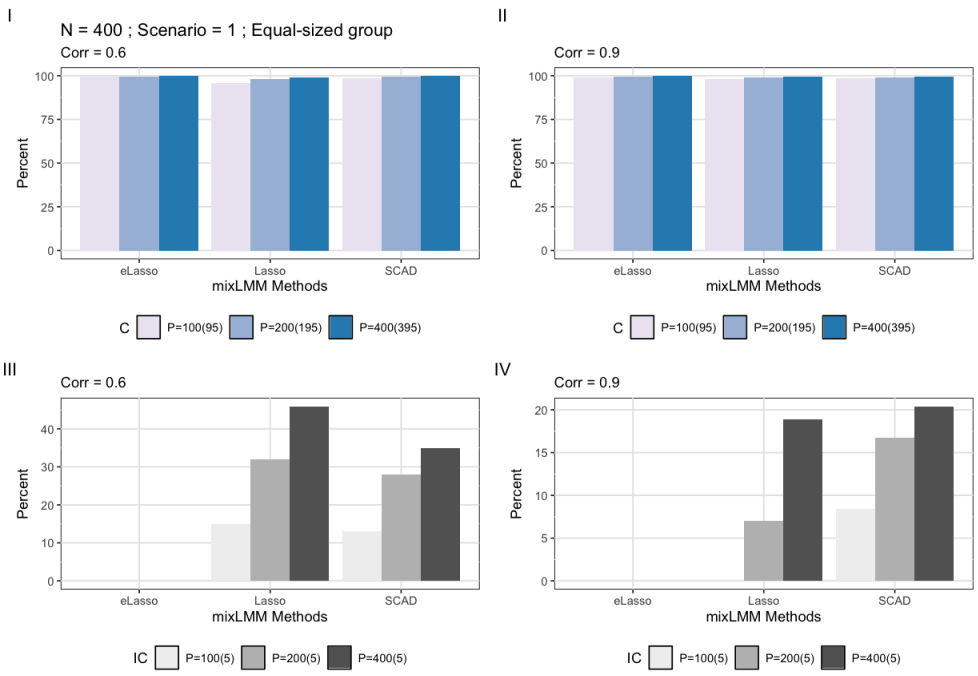


(a) Equal-Sized Group in Scenario 3, n=200, p = (25, 50, 100)

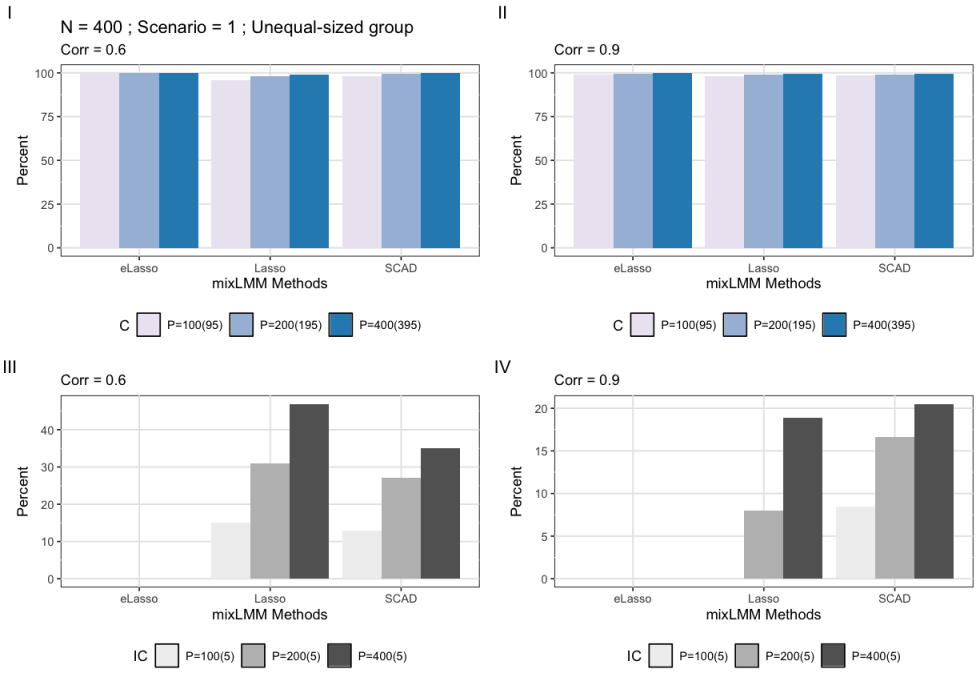


(b) Unequal-Sized Group in Scenario 3, n=200, p = (25, 50, 100)

Figure 3.6: Model-based Clustering with various penalties: 1) Exclusive Lasso (mixLMM-eLasso), 2) SCAD (mixLMM-SCAD), and 3) Lasso (mixLMM-Lasso) tested in Scenario 3 with different data settings. Corr = 0.60 (Left), 0.90 (Right); n = 200; p = (25, 50, 100). The number of true zero and non-zero coefficients in a cluster are indicated next to *p* for C and IC, respectively.
Note: On the x-axis, the penalties illustrate model-based clustering with the implemented penalties for comparisons.

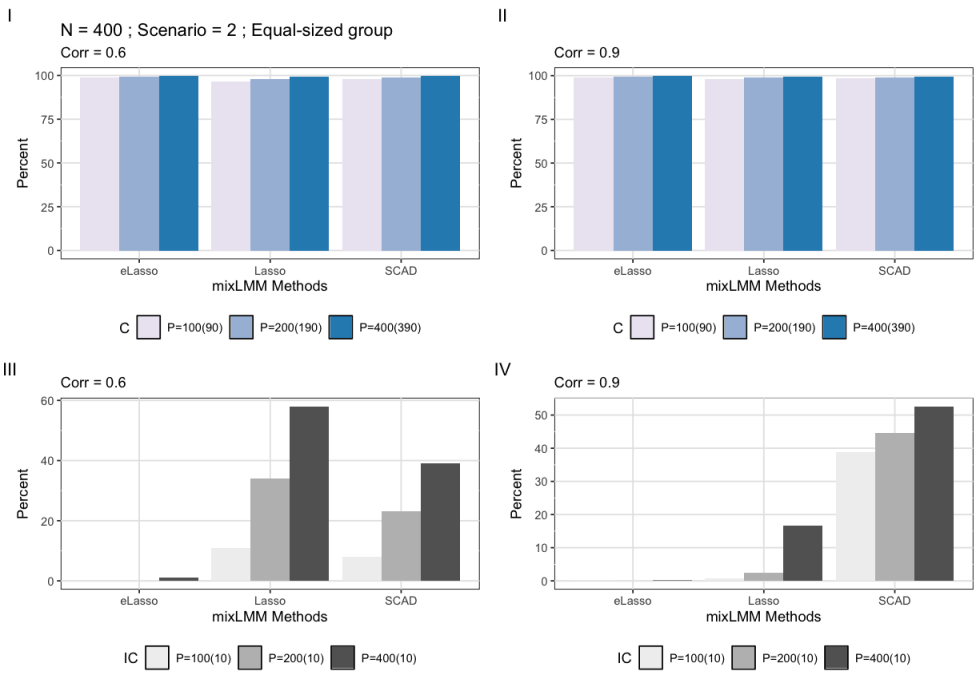


(a) Equal-Sized Group in Scenario 1, n=400, p = (100, 200, 400)

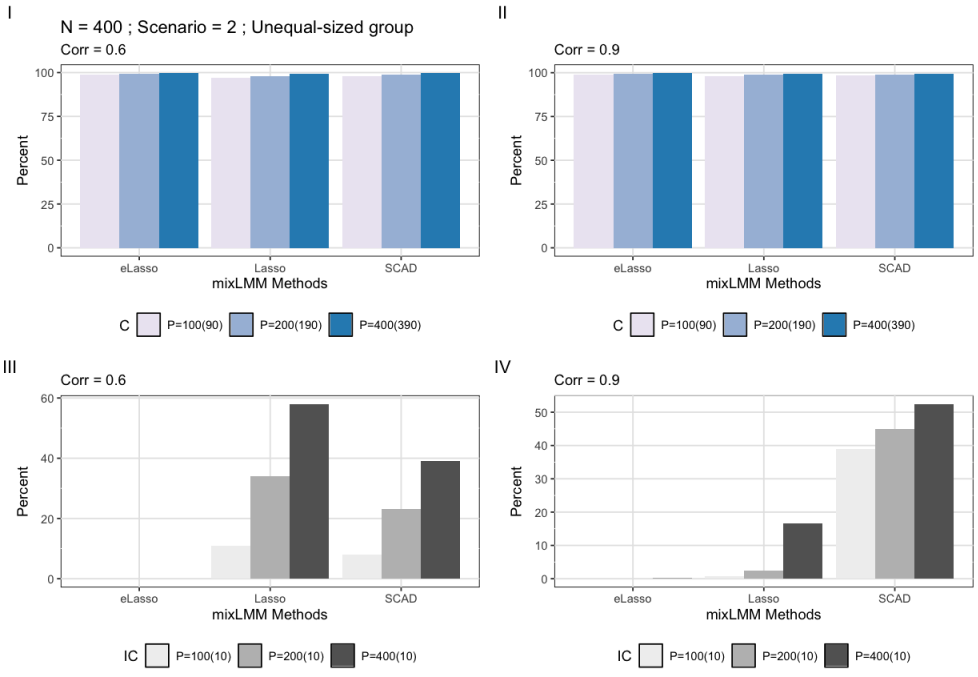


(b) Unequal-Sized Group in Scenario 1, n=400, p = (100, 200, 400)

Figure 3.7: Model-based Clustering with various penalties: 1) Exclusive Lasso (mixLMM-eLasso), 2) SCAD (mixLMM-SCAD), and 3) Lasso (mixLMM-Lasso) tested in Scenario 1 with different data settings. Corr = 0.60 (Left), 0.90 (Right); n = 400; p = (100, 200, 400). The number of true zero and non-zero coefficients in a cluster are indicated next to p for C and IC, respectively.
Note: On the x-axis, the penalties illustrate model-based clustering with the implemented penalties for comparisons.

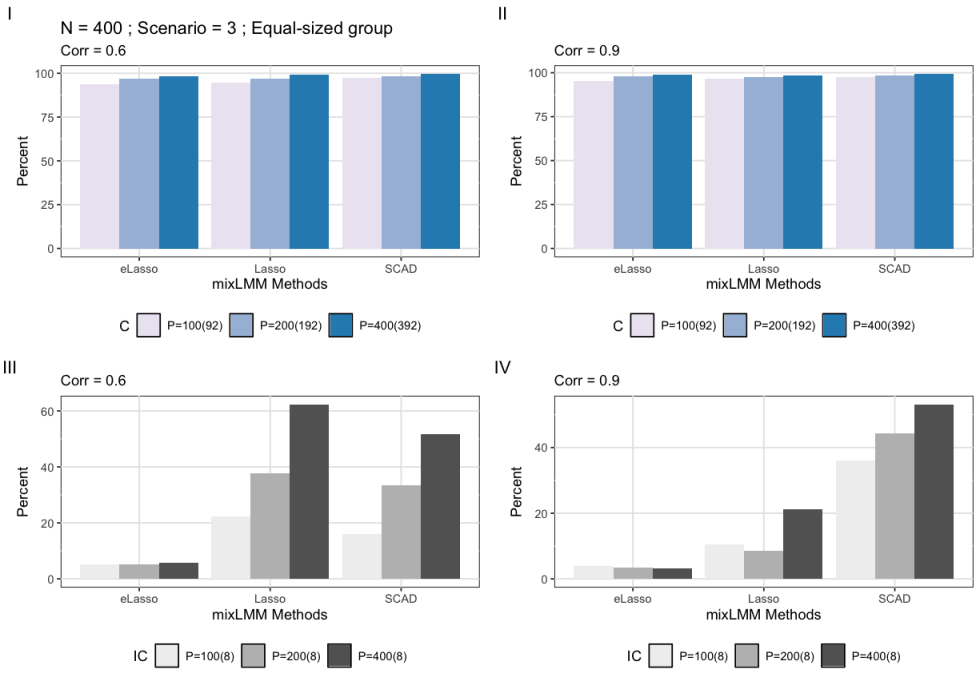


(a) Equal-Sized Group in Scenario 2, n=400, p = (100, 200, 400)

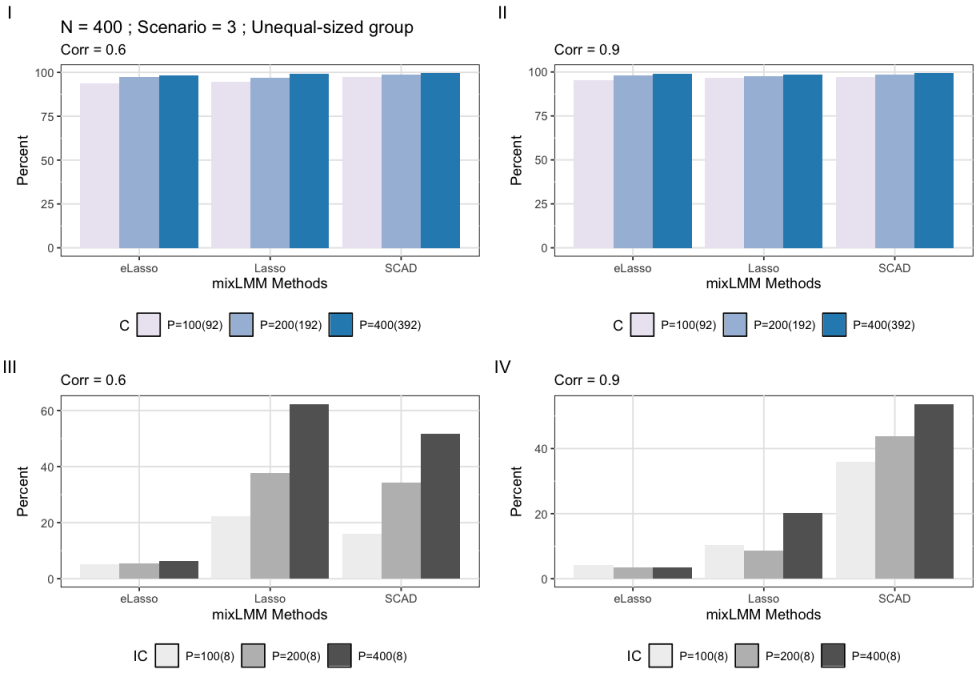


(b) Unequal-Sized Group in Scenario 2, n=400, p = (100, 200, 400)

Figure 3.8: Model-based Clustering with various penalties: 1) Exclusive Lasso (mixLMM-eLasso), 2) SCAD (mixLMM-SCAD), and 3) Lasso (mixLMM-Lasso) tested in Scenario 2 with different data settings. Corr = 0.60 (Left), 0.90 (Right); n = 400; p = (100, 200, 400). The number of true zero and non-zero coefficients in a cluster are indicated next to p for C and IC, respectively.
Note: On the x-axis, the penalties illustrate model-based clustering with the implemented penalties for comparisons.



(a) Equal-Sized Group in Scenario 3, n=400, p = (100, 200, 400)



(b) Unequal-Sized Group in Scenario 3, n=400, p = (100, 200, 400)

Figure 3.9: Model-based Clustering with various penalties: 1) Exclusive Lasso (mixLMM-eLasso), 2) SCAD (mixLMM-SCAD), and 3) Lasso (mixLMM-Lasso) tested in Scenario 3 with different data settings. Corr = 0.60 (Left), 0.90 (Right); n = 400; p = (100, 200, 400). The number of true zero and non-zero coefficients in a cluster are indicated next to *p* for C and IC, respectively.
Note: On the x-axis, the penalties illustrate model-based clustering with the implemented penalties for comparisons.

3.6 SDOH Data Application

We applied the proposed clustering method to the data, which considered the SDOH variables and CVD mortality in a longitudinal framework. The working SDOH (Son et al., 2023) dataset was described in Section 1.2.1. The dataset had less than 2% of missing information, and missing values were replaced using mean imputation, a commonly applied simple imputation technique in statistical analyses (Little & Rubin, 2019). After this imputation, the dataset included 78 variables across 3,224 counties. This consisted of 3,142 counties and 82 county-equivalents from US territories, including 78 municipalities in Puerto Rico, one district in Guam, and three main islands in the US Virgin Islands. Of these 3,142 counties, 1,166 were classified as urban, while the remaining 1,976 were classified as rural. The rural-urban status of a county was determined according to the Urban-Rural Classification Scheme for Counties used by the National Center for Health Statistics (NCHS) in 2013 (Ingram & Franco, 2014). Counties in categories 1 through 4 were classified as urban, while counties in categories 5 and 6 were classified as rural (Ingram & Franco, 2014); however, US territories were not included in this classification. We included this rural-urban indicator as an additional domain to address geographic disparities.

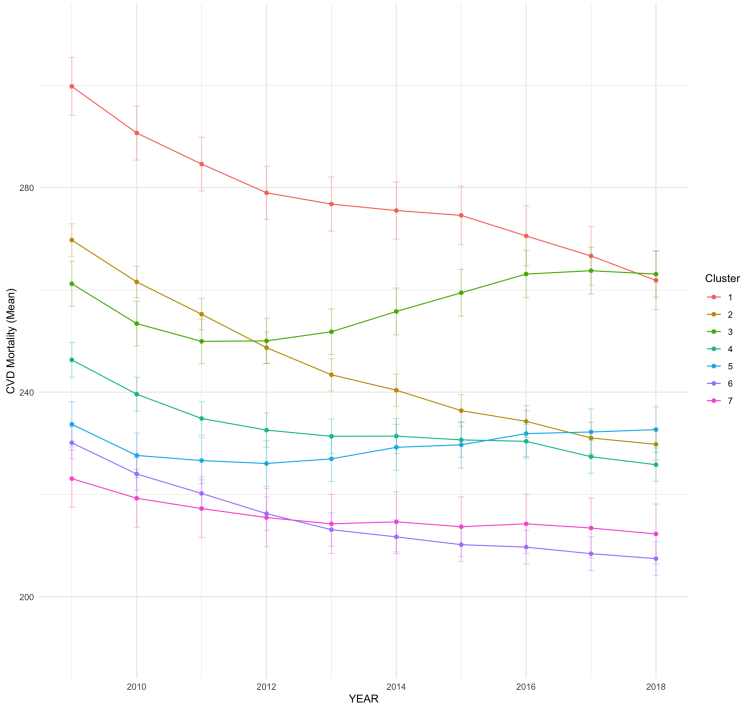


Figure 3.10: Mean age-adjusted cardiovascular disease mortality trajectories of the US counties from 2009 to 2018 in the seven clusters identified by the method.

The proposed clustering method with Exclusive Lasso (mixLMM-eLasso) was applied to our SDOH study data. To determine the optimal number of clusters, we calculated the BIC criterion, and the results can be found in the Appendix (see Table B.5). Based on the BIC, seven clusters of counties with similar CVD mortality trajectories associated with SDOH were identified, as shown in Figure 3.10. This figure illustrated trends in mean age-adjusted CVD mortalities per 100,000 people across seven clusters from 2009 to 2018, with corresponding statistics provided in the Appendix (see Table B.4). Cluster 1 consisted of 553 counties and consistently revealed the highest mortality rates. These rates steadily decreased from 299.79 per 100,000 people (95% CI: 294.12, 305.46) in 2009 to 261.85 per 100,000 people (95% CI: 256.13, 267.58) in 2018. Cluster 2, which included the largest group with 809 counties, exhibited a similar decreasing trend, with mortality declining from 269.74 per 100,000 people (95% CI: 266.56, 272.92) in 2009 to 229.80 per 100,000 people (95% CI: 226.70, 232.89) in 2018, highlighting substantial improvements over the decade. In contrast, Cluster 7, the smallest group with 170 counties, demonstrated the lowest mortality rates, with a gradual decline from 223.09 per 100,000 people (95% CI: 217.48, 228.69) in 2009 to 212.27 per 100,000 people (95% CI: 206.41, 218.13) in 2018. In addition, Cluster 3 showed a unique pattern, with a slight increase in mortality from 250.02 per 100,000 people (95% CI: 245.60, 254.44) in 2012 to 263.08 per 100,000 people (95% CI: 258.56, 267.60) in 2018, reflecting a deviation from the overall downward trend observed in other clusters. Similarly, Cluster 5 represented a somewhat unique pattern with minimal change over the years, showing a slight decrease followed by a slight increase from 226.05 per 100,000 people (95% CI: 221.58, 230.51) in 2012 to 232.67 per 100,000 people (95% CI: 228.19, 237.15) in 2018. These patterns suggested distinct trajectories in CVD mortality reduction across clusters, potentially associated with varying SDOH factors across all domains, as shown in Table 3.4.

Figure 3.11 showed the spatial distribution of US counties classified into seven clusters using the proposed method, with detailed results available in Table B.7 (see in the Appendix). However, the figure did not include the US territories. The clustering results highlighted the differences in county-level characteristics shown in Table B.6 (also in the Appendix). Additionally, it was noted that this rural-urban frequency table in B.6 did not cover the US territories as the Urban-Rural Classification Scheme for Counties (Ingram & Franco, 2014) did not provide the rural-urban status for those regions. Cluster 2, the largest group, accounted for 25.09% of counties and included 27.02% of urban counties and 23.63% of rural counties. This indicated a substantial presence in both urban and rural areas. Cluster 1, accounting for 17.15% of total counties, reflected a predominantly rural composition. The gap between rural (21.20%) and urban (7.98%) counties was notably larger in this cluster compared to other clusters, indicating the highest disparity in the proportion of counties between rural and urban areas. Similarly, Cluster 3, which made up 15.54% of the total, skewed toward rural areas, with 18.47% of total rural counties and 11.49% of total urban counties. In contrast, Cluster 6, which represented 11.17% of counties, was more urban-focused, featuring 19.47% of urban counties and only 6.53% of rural counties. Similarly, Cluster 5 (9.62%) exhibited a slightly urban-dominant distribution with 11.15% of total urban counties and 9.11% of total rural counties. Cluster 4 (16.16%) and Cluster 7 (5.27%) displayed mixed distributions, with Cluster 4 containing 14.07% of urban and 17.97% of rural counties, and Cluster 7 having the lowest proportions overall, with 8.83% of total urban and 3.09% of total rural. These results highlighted the geographic heterogeneity in

CVD mortality distribution, suggesting that SDOH associated with CVD mortality may also have varied significantly across clusters. Cluster-specific policies and interventions may have been needed to address the unique challenges faced by rural-dominant clusters like Clusters 1 and 3 and urban-dominant clusters like Cluster 6. Tailoring strategies to these patterns could have enhanced equity and effectiveness in CVD management across US counties.

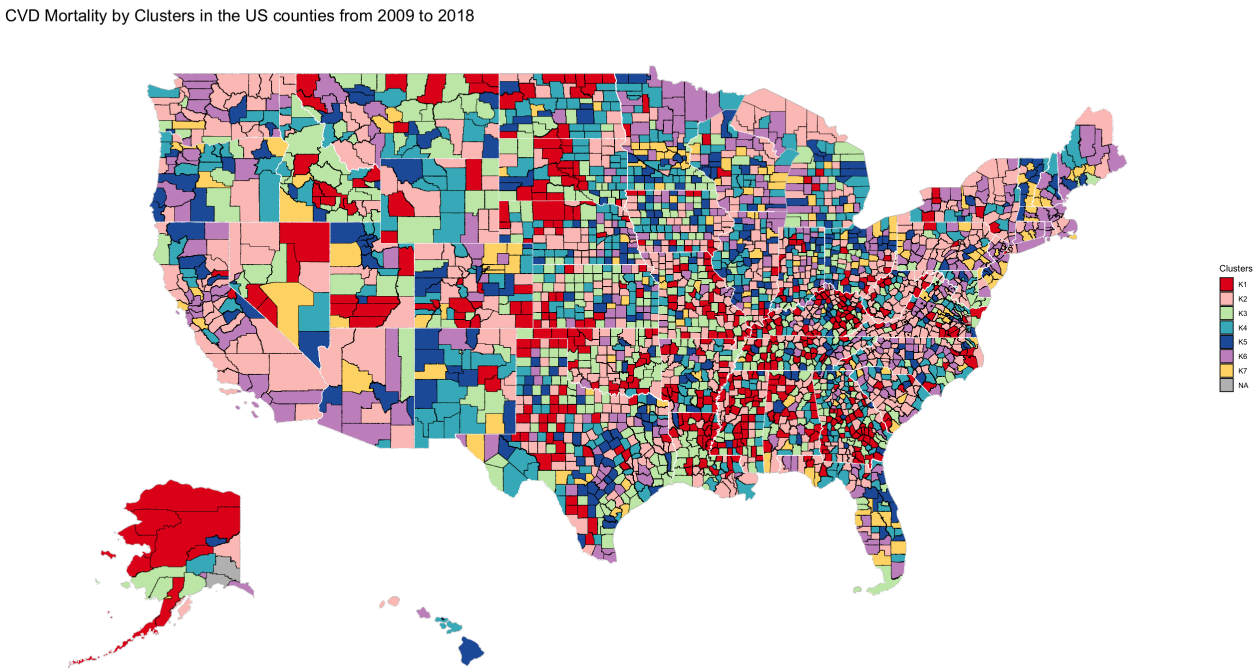


Figure 3.11: Geographic distribution of clusters based on age-adjusted cardiovascular disease mortality across counties in the United States.

Table 3.4 highlighted associations between SDOH variables and CVD mortality, providing insights into how these factors were related across seven clusters. The results also revealed distinct patterns of SDOH contributions within each domain. The intercepts for each cluster indicated the mean age-adjusted CVD mortality in each cluster, and they were assumed to vary by geographic distribution. To account for the heterogeneity, a random effect for the intercept was included in the proposed model. This approach captured the variability in baseline mortality rates across clusters while modeling the associations between SDOH variables and CVD mortality. Additionally, the geographic region for the rural and urban indicators was included as a measure of geographic disparity. While all clusters demonstrated varying rural-urban dynamics, most indicated that CVD mortality rates were higher in rural counties than urban ones.

Cluster 1, characterized by rural dominance, consistently showed the highest CVD mortality rates, although it had shown a steady decline over time. It had the highest baseline mean CVD mortality rate of 276.76 per 100,000 people, and most SDOH from the physical infrastructure and healthcare context were significant. For instance, the presence of community mental health centers (β=-2.6828) and the number of rural health clinics (β=-0.4263) showed a negative association with mortality rates, highlighting the benefits of accessible healthcare in rural areas. The number of people living with diagnosed HIV per 1,000 (β=24.7662) was significantly associated with increased CVD mortality. This could be due to the complex health challenges experienced by individuals with HIV, particularly in geographically isolated areas. Additionally, the physical infrastructure domain played a crucial role. Variables such as the number of beer, wine, and liquor stores (β=9.6685) and convenience stores (β=2.6932) per 1,000 people indicated potential lifestyle and environmental risks. The high density of liquor and convenience stores may have affected unhealthy behaviors in this rural-dominant cluster, like increased alcohol consumption and reliance on processed foods, both of which were associated with poor cardiovascular health. From an economic perspective, median household income (β=-0.3533) was linked to reduced mortality, suggesting that higher income levels could alleviate some of the disadvantages faced in these rural-dominant areas. However, limited education remained a critical issue; a higher percentage of individuals with less than a high school diploma (β=1.5853) was associated with elevated mortality rates.

Cluster 2, the largest cluster, demonstrated substantial improvements in reducing CVD mortality rates over the decade. This cluster, an urban-dominant cluster with significant rural representation, had a baseline mean CVD mortality of 244.49 per 100,000 people. The percentage of the civilian population from the social context, consisting of veterans (β=1.9300), was positively associated with mortality. Veterans faced unique health challenges, such as higher rates of chronic conditions like hypertension and diabetes, along with limited access to specialized healthcare services in both urban and rural areas. In terms of the physical infrastructure, the percentage of housing units with no vehicles available (β=0.2767) reflected unique challenges, as reliance on public transportation restricted access to healthcare in this cluster. Conversely, the high density of beer, wine, and liquor stores per 1,000 people (β=-2.4785) indicated a negative association with mortality in this cluster. This may have reflected counties where better regulations and infrastructure mitigated the harmful effects of excessive alcohol availability or where improved healthcare accessibility helped control heart disease. However, this effect may have differed across clusters,

where regulatory enforcement could have been weaker, indicating the need for further investigation into local dynamics.

In Cluster 3, the baseline mean CVD mortality was 256.42 per 100,000 people. This predominantly rural cluster uniquely showed an increasing trend in CVD mortality from 2012 onward, in contrast to the overall decline observed in the other clusters. Within the social context domain, the percentage of the population identifying as Native Hawaiian or Pacific Islander (β=5.5562) was strongly positively associated with CVD mortality. This association highlighted the unique health challenges faced by the population in this cluster. In the economic context domain, the percentage of the population that was unemployed (β=-1.3502) showed a negative association with CVD mortality. This finding may have reflected the specific dynamics of rural economies, where unemployment could occur alongside protective factors, such as informal caregiving networks, which may have been associated with health outcomes in differing ways. Furthermore, the density of full-service restaurants (β=-4.8642) per 1,000 people in the physical infrastructure was negatively associated with CVD mortality. This suggested that access to a variety of food options may have promoted healthier dietary habits and reduced cardiovascular risks. From the healthcare context domain, the number of Federally Qualified Health Centers (β=2.2065) was positively associated with CVD mortality. This finding may have indicated that these centers were located in areas without fully addressing healthcare disparities. Similarly, the number of rural health clinics (β=1.4253) also showed a positive association with mortality, reflecting potential limitations in the capacity or quality of the healthcare workforce provided by these facilities, particularly in preventive health services for complex chronic diseases like CVD.

Cluster 4, which indicated a mixed rural-urban composition and moderate CVD mortality rates, had a baseline mean CVD mortality rate of 233.65 per 100,000 people. In the economic context, the percentage of people employed in information services (β=2.9559) was possibly associated with the outcome. This link could be attributed to factors such as occupational stress, sedentary lifestyles, and long working hours that were often prevalent in information services jobs. In the healthcare context, the number of people living with diagnosed HIV per 1,000 individuals (β=1.2838) also showed a positive correlation with mortality. This correlation may have highlighted broader disparities in access to and quality of healthcare services within underserved populations, where individuals with chronic conditions faced challenges in accessing comprehensive care.

Cluster 5 maintained relatively stable mortality rates over time, characterized by minimal overall change but a slight increase in recent years. The baseline average CVD mortality rate was 229.51 per 100,000 people. Predominantly urban, the counties in this cluster displayed significant racial disparities, particularly regarding the percentage of Black residents (β=1.0650) within its social context compared to other clusters. In the physical infrastructure domain, the percentage of workers using public transportation, excluding taxicabs, (β=-2.2212) was negatively associated with mortality. This indicated the protective benefits of active transportation behaviors, such as walking or biking, to access public transit. Conversely, in Cluster 1, this association was slightly positive (β= 0.2784), suggesting that public transportation availability or usage patterns differed, potentially reflecting limited infrastructure quality, greater travel burden, or fewer overall benefits of transit systems in predominantly rural areas. Additionally, counties in Cluster 5 could

have benefited from well-developed public transportation networks that improved access to healthcare facilities and other essential services, thereby reducing mortality risks. In terms of healthcare context, the total number of community mental health centers (β=-3.7925) was also negatively associated with mortality. This suggested that increased access to these facilities may have helped address some of the health disparities faced by populations in this cluster, especially in rural areas where mental health resources were often scarce.

Clusters 6 and 7, both characterized by their urban dominance, displayed distinct patterns in the contributions of SDOH while also sharing similarities in some areas. Cluster 6, which showed moderate to low CVD mortality, had a baseline mean CVD mortality rate of 214.47 per 100,000 people. Cluster 7, the smallest group, consistently recorded the lowest mortality rates among all clusters, showing steady declines. It had a baseline CVD mortality rate of 216.77 per 100,000 people. In both clusters, a higher percentage of individuals with less than a high school diploma was positively associated with CVD mortality, with β coefficient values of 0.6571 for Cluster 6 and 0.3451 for Cluster 7. Meanwhile, the percentage of individuals holding a master's degree or higher was negatively associated with mortality in both clusters, with β coefficient values of -2.1020 in Cluster 6 and -0.9772 in Cluster 7. However, the economic context revealed different factors affecting CVD mortality. In both clusters, a higher percentage of workers employed in manufacturing was positively associated with CVD mortality. Nevertheless, economic vulnerabilities were highlighted by different variables in each cluster: Cluster 6 obsered a positive association between CVD mortality and the percentage of employees in construction (β=0.7175) and a negative association between CVD mortality and households receiving food stamps/SNAP (β=-0.4887). In contrast, CVD mortality in Cluster 7 was associated with the percentage of households with public assistance income or food stamps/SNAP (β=-1.7780) and the percentage of workers employed in wholesale trade (β=0.2732). Additional domains further distinguished the two clusters. In the physical infrastructure domain, the density of supermarkets and grocery stores (excluding convenience stores) per 1,000 people (β=-3.3162) was negatively associated with CVD mortality in Cluster 7, while this variable was not associated with the outcome in Cluster 6. In the social context domain, Cluster 6 had more ethnic and immigration characteristics associated with CVD mortality compared to Cluster 7.

Distinct patterns of SDOH emerged across different clusters, particularly within each specific domain. In rural-dominant clusters, such as Clusters 1 and 3, factors from the social context were prominently represented, highlighting the potential relationship between racial and social inequities. SDOH in the economic context, like median household income, consistently showed protective effects across clusters; however, these benefits were less pronounced in rural areas, possibly due to lower baseline income levels. The education domain underscored the far-reaching effects of limited educational attainment, especially in rural-dominant clusters. Additionally, factors related to physical infrastructure, such as housing and transportation, were significant in Cluster 1, where access to infrastructure was directly associated with mortality outcomes. In rural-dominant clusters, the healthcare context played a particularly crucial role, with both positive and negative associations reflecting disparities in healthcare access and quality. These findings illustrated how SDOH factors may have been associated with CVD mortality differently in rural and urban clusters, emphasizing the need for interventions tailored to the unique challenges of

each area. Policies aimed at improving educational access, increasing healthcare resources, and enhancing infrastructure were essential for reducing disparities and promoting health equity across various contexts.

Table 3.4: Selected Social Determinants of Health Associated with Age-Adjusted CVD Mortality by County-level Clustering: mixLMM-EL, 2009–2018

Domains	Variables	K1	K2	K3	K4	K5	W6	K7
	Int	276.7620	244.4885	256.4248	233.6494	229.5073	214.4745	216.7652
Social	% Housing units with more than one occupant per room	-	-	-	-	-0.0014	-	-
context	% Population reporting Asian race	-	-	-	-	-0.0317	-0.0424	-0.6349
	% Population reporting Black race	-.4084	0.4714	0.6477	0.6587	1.0650	0.5829	0.7612
	% Families with Children that are single-parent Families	-	-	0.0698	-	-	-	-
	% Population that does not speak English at all (ages 5 and over)	-	-	-	-	-0.3375	-	-
	% Population that is foreign-born	-0.7631	-	-	-0.8383	-	-0.0677	-0.2732
	% Children living with a grandparent householder (ages 17 and under)	-	-	-	-	0.4601	-	-
	% Occupied housing units without fuel	-0.3900	-	-	-	-	-	-
	% Population reporting Hispanic ethnicity	-0.4569	-0.6203	-0.2484	-0.3423	-	-0.0957	-
	% Population reporting multiple races	-	-	-	-	-	-0.1468	-
	% Population reporting Native Hawaiian/Pacific Islander race	-	-	5.5562	-	-0.2529	-	-
	% Population who are not U.S. citizens and entered U.S. before 1990	-0.4616	-	-0.8767	-	-	-	-
	% Population who speak other languages (ages 5 and over)	-	-	-	-	-	-0.0545	-
	% Civilian Population consisting of veterans (ages 18 and over)	0.1744	1.9300	-0.6576	-	-	0.9636	-
Economic	Median household income (in dollars, inflation-adjusted to file data year) /1000	-0.3533	-0.1674	-	-0.2470	-	-	-
context	% Unmarried partner households that received food stamps/SNAP benefits	-	-0.1586	-0.0018	-	-	-	-
	% Employed working in public administration	-	-	-	-	-0.0435	-	-
	% Civilian Population in armed forces (ages 16 years and over)	0.8446	-	-	0.4130	-	-	-
	% Employed working in arts, entertainment, recreation, etc.	-	-	-	-	-	-	-
	% Employed working in construction	0.5123	0.5280	-	-	-	0.7175	-
	% Employed working in finance and insurance, real estate, and rental, etc.	-	-	-0.4683	-	-	-	-
	% Households that received food stamps/SNAP, past 12 months	-0.3562	-0.6349	-	-	-	-0.4887	-
	% Households with public assistance income or food stamps/SNAP	-	-	0.1780	-	-0.8084	-	-1.7780
	% Employed working in information services	0.4056	-	-2.1483	2.9559	-	-	-
	% Employed working in manufacturing	0.0031	0.1532	0.1194	0.4356	0.4380	0.7080	0.5038
	% Employed working in agriculture, forestry, fishing, etc. (ages 16 and over)	-0.7556	-0.1469	-0.1689	-	-	-	-
	% Employed working in other services, except public administration	-	-	0.4962	-	-	-	-
	% Population with income to poverty ratio: 1.25-1.99	-0.1478	-	-	-	-	-	-
	% Population with income to poverty ratio: <1.00	-	-	0.1889	-	-	-	-
	% Employed working in professional, scientific, management, etc.	-	-	-0.0307	-	-	-	-
	% Employed working in transportation and warehousing, and in utilities	0.0468	-	-	-	0.3085	-	-
	% Population that was unemployed (ages 16 years and over)	0.9503	0.1127	-1.3502	-	-1.2396	-	-0.0033
	% Employed working in wholesale trade	0.0237	1.2932	-	-	-	-	0.2732
Education	% Population with some college or associate's degree (ages 25 and over)	-	-	-0.1901	-0.0167	-	-	-
	% Population with a bachelor's degree (ages 25 and over)	-	-	-0.3119	-	-0.2891	-	-
	% Population with a master's or higher degree (ages 25 and over)	-0.4596	-2.4459	-0.5556	-0.5819	-0.2080	-2.1020	-0.9772
	% Population with only high school diploma (ages 25 and over)	0.51322	-	-	-	-	-	-
	% Population with less than high school education (ages 25 and over)	1.5853	2.1363		1.2619	0.2180	0.6571	0.3451
Physical	Median home value of owner-occupied housing units	-0.00010	-0.00006	-0.00002	-0.00005	-0.00003	-0.00001	-
Infrastructure	% Housing in structures with 10 or more units	-0.4236	-	-	-	-0.4568	-	-
	% Housing units that are mobile homes	0.1264	0.2387	0.4566	0.3299	-	0.7475	0.7655
	% Housing units with no vehicle available	-	0.2767	-	-	-	-	-
	% Workers (16 +) with a 60+ min public transit commute	0.0197	0.0074	0.0170	-0.0429	-	-	-0.0210
	% Workers taking public transportation, excluding taxicab (ages 16 and over)	0.2784	-	-0.4432	-	-2.2212	-	-0.0301
	% Housing units vacant	-0.4797	-0.3654	-	-0.0271	-	-0.0384	-
	Beer, wine and liquor stores per 1,000 people	9.6685	-2.4785	9.7171	-	-	-	-
	Convenience stores per 1,000 people	2.6932	-	-	-	-	-	-
	Full service restaurants per 1,000 people	-0.2455	-0.6055	-4.8642	-	-	-	-
	Supermarkets and other grocery (except convenience) stores per 1,000 people	-6.2250	-	-	-	-	-	-3.3162
Healthcare	Number of Federally Qualified Health Centers	-0.0610	-0.0761	2.2065	0.1479	0.3110	-0.0394	-0.0165
context	Total number of community mental health centers	-2.6828	2.4670	-0.0590	-	-3.7925	-0.0715	0.4188
	Number of rural health clinics	-0.4263	0.1236	1.4253	-0.2228	0.6498	-	0.4522
	Number of people living with diagnosed HIV / 1000	24.7662	0.1166	-	1.2838	-	0.3998	-
	Number of Medicare eligibles in the county	-0.0005	-	-	-	-	-	-
	Derived field that equals the ratio of enrollees over eligibles * 100	-0.4353	-0.4647	0.4623	-0.2404	0.0856	-0.2769	-0.1799
Region	1: Rural, 0: Urban by NCHS 2013 Rural-Urban Classification Scheme	5.0422	2.3003	6.9843	-1.2958	7.9543	0.3526	6.4239

3.7 Summary

This study introduces a novel clustering method specifically designed for high-dimensional longitudinal data with domain-based structures in covariates. It leverages a finite mixture of LMM combined with a composite penalty. This approach effectively addresses the challenges of selecting fixed effects within different domains and provides a new way to identify cluster-specific associations between the outcome and predictors in high-dimensional contexts. By utilizing an EM algorithm and a composite penalty that incorporates both l_1 and l_2 norms, this method provides an effective way of high-dimensional clustering across within- and between-grouped variables. The inclusion of l_1 and l_2 norms in Exclusive Lasso balances between variable sparsity and grouping effects, making the method robust to noise and collinearity in the data.

Building upon the framework established by Du et al., 2013, our method extends their work by focusing specifically on fixed-effect selection within domains. The proposed approach uses a composite penalty structure to facilitate the selection of domain-based variables, making it particularly suitable for applications where identifying domain-specific variables and interpreting their associations with the outcome variable within clusters is essential. This innovative method enhances clustering and variable selection in high-dimensional longitudinal data.

In the application of SDOH and CVD mortality data, the clustering results revealed distinct patterns of associations across different clusters. This highlights the variability in SDOH-CVD relationships. For instance, clusters dominated by rural areas showed higher baseline mortality rates, and they were linked to limited access to healthcare and lower educational attainment. Conversely, urban-dominant clusters were characterized by factors such as occupational stress and housing vulnerabilities. These findings point out the limitations of traditional geographic region-based analyses and emphasize the need for strategies informed by SDOH that are tailored to specific clusters. By identifying cluster-specific associations between SDOH and CVD mortality, the proposed method offers actionable insights for targeted interventions. Ultimately, these results suggest that strategies based on SDOH clusters have the potential to address health disparities more effectively than broad, one-size-fits-all approaches.

From a statistical perspective, this method can be adapted to handle various types of outcomes, including binary and count data. It also allows for the integration of different clustering techniques. The primary strength of this study is the development of a flexible and robust clustering method that incorporates domain-based covariate selection into longitudinal data analysis. Additionally, the method's ability to handle high-dimensional data with missing values enhances its practical utility. However, the dynamic changes in the associations between SDOH and CVD mortality over time have not been considered in our current approach. Future research could benefit from the inclusion of random effect selection, which could provide additional insights, especially in datasets where subject-specific variability is essential. Incorporating random effects in each domain would further extend the applicability of the method to a broader range of longitudinal models, allowing for more nuanced clustering and variable selection. Moreover, while the composite penalty offers a feasible solution for domain-based variable selection, its computational complexity rises with the number of domains and variables, potentially limiting scalability

for extremely large datasets. Future work could focus on developing scalable algorithms to overcome these challenges and enhance the method's applicability.

Chapter 4

Use of Model-base Clustering of High-Dimensional Longitudinal Data via Exclusive Lasso Penalty by Different Levels of SDOH Data

4.1 Introduction

The mortality rate of CVD in the US has substantially declined (Mensah et al., 2017). Despite this success, racial, ethnic, socioeconomic status, and regional disparities in CVD outcomes persist across the US (Graham, 2015; Post et al., 2022). Based on our data from 2009 to 2018, county-level data indicate an overall decline in CVD mortality rates; however, distinct geographic and racial differences in CVD mortality continue to exist (Dong et al., 2023; Son et al., 2023). Notably, a cluster of counties with high CVD mortality extends from southeastern Oklahoma through the Mississippi River valley to eastern Kentucky, often referred to as the heartland of the United States. Conversely, areas with the lowest CVD mortality rates include the San Francisco Bay area, central Colorado, northern Nebraska, central Minnesota, northeastern Virginia, and south Florida (Roth et al., 2017). Furthermore, our study highlights that rural counties consistently exhibit higher CVD mortality rates compared to urban counties (Son et al., 2023). Similarly, counties with a higher percentage of Black residents experience higher CVD mortality rates than those with a lower percentage of Black residents. (Dong et al., 2023; Son et al., 2023).

Health disparities in CVD mortality have been linked to SDOH (for Disease Control, Prevention, et al., 2019; Frieden et al., 2013). The Centers for Disease Control and Prevention (CDC) and the American Heart Association highlight the importance of addressing SDOH in public health initiatives and healthcare practices to reduce disparities among different racial groups and regions (Banerjee, 2017; Benjamin et al., 2019; Hacker et al., 2022; White-Williams et al., 2020). Specifically, the CDC identifies five key domains of SDOH that need attention to improve overall health outcomes: economic stability, education access and quality, healthcare access and quality, neighborhood and built environment, and social and

community context (Hacker et al., 2022). Understanding these domains is essential for designing targeted strategies aimed at reducing mortality by addressing key risk factors associated within each domain and tackling existing disparities (Y. Fu et al., 2023; Powell-Wiley et al., 2022).

Recent studies have highlighted specific regional disparities in CVD at the state or county level (Glynn et al., 2021; Patel et al., 2016; Son et al., 2023; Zelko et al., 2023). To understand and address these geographic disparities in CVD, it is necessary to analyze SDOH data with advanced methodological approaches. These approaches should focus on geographic groups to examine the effects of various SDOH domains and their associations with CVD mortality rates. This limitation has not been addressed in our previous study (Son et al., 2023). Furthermore, since the CVD mortality rates vary by geographic and racial groups over time, counties across the US may exhibit different characteristics when subjects are grouped based on SDOH in a high-dimensional longitudinal data setting. New clustering strategies incorporating data on SDOH may help drive region-specific interventions aimed at reducing CVD disparities and improving overall CVD outcomes.

Therefore, I proposed to analyze SDOH indicators and CVD mortality rates at the county level using the natural cluster of states. Understanding the specific factors influencing CVD mortality across regions, especially in the context of racial and geographic disparities, allows us to identify more effective, tailored interventions effectively. A targeted local-level analysis guides stakeholders and policymakers in developing precise strategies, ensuring that interventions are appropriately adjusted in intensity across sectors, leading to sustained positive health outcomes. (Roth et al., 2017). To enhance this approach, we will apply a model-based clustering method using Exclusive Lasso to refine our understanding of county-level variations within each state. By incorporating an additional algorithm that accounts for these variations, we can classify counties based on their unique characteristics. The selection of this algorithm will depend on predefined complexity thresholds within clusters, aiming to minimize intra-cluster variation and improve differentiation among counties at the state level.

4.2 Method

The Exclusive lasso penalized model-based clustering method is a technique that uses the regularization of mutually exclusive features within groups of fixed or random effects to cluster subpopulations by US counties. Each cluster includes at least one SDOH from each predefined domain. However, this method has limitations. As policies are typically made and implemented at the state level, the clusters identified by conventional clustering analysis may include too many states, which makes it less useful for real-world policy-making. In addition, while it classifies groups at the county level, it does not consider variations within counties in each state.

In this section, we propose a regularization method that incorporates a constraint into the model-based clustering framework to facilitate a more explainable clustering approach. By building on the Exclusive lasso penalized model-based clustering method, we introduce a constraint to the clustering method. This allows us to set cutoff values for the new cluster structure, which helps reduce the number of clusters in

each state and the state variation within each cluster. A modified EM algorithm will be implemented with this constraint.

The ith subject (county) is measured at m_i time points, $i = 1, ..., n$. Repeated measures m_i may be missed, which means the number of m_i can vary from subject to subject. At time t_{ij}, $j = 1, ..., m_i$, we have $(y_{ij}, \mathbf{X}_{ij}, \mathbf{Z}_{ij})$ observarions, where y_{ij} represents the responses as CVD mortality, $\mathbf{X}_{ij} \in \Re^{p_n}$ contains the fixed effects SDOH, and $\mathbf{Z}_{ij} \in \Re^{q_n}$ includes the random effects SDOH. Here, p_n and q_n denote the dimension of SDOH variables, which increase at a certain rate as n. We use the following notations for the ith subject: $\mathbf{y}_i = (y_{i1}, y_{i2}, ..., y_{im_i})^T$, $\mathbf{X}_i = (\mathbf{X}_{i1}^T, ..., \mathbf{X}_{im_i}^T)^T$, and $\mathbf{Z}_i = (\mathbf{Z}_{i1}^T, ..., \mathbf{Z}_{im_i}^T)^T$. This framework models the linear mixed-effects model (LMM) for the ith subject, and it has the form given by

$$\mathbf{y}_i = \mathbf{X}_i\boldsymbol{\beta}_i + \mathbf{Z}_i\mathbf{b}_i + \mathbf{e}_i, \tag{4.1}$$

where $\boldsymbol{\beta} \in \mathbb{R}$ is a vector of fixed-effects parameters, $\mathbf{b}_i \sim N(\mathbf{0},\ \sigma^2\mathbf{D})$ is a $(q_n \times 1)$ vector of subject-specific random effects, and $\mathbf{e}_i \sim N(\mathbf{0},\ \mathbf{R}_i)$ is a vector of the i.i.d. random error. $\mathbf{D}$ is a $(q_n \times q_n)$ covariance matrix that specifies the among-unit sources, and $\mathbf{R}_i$ captures within-subject variance and correlation. In many applications, $\mathbf{R}_i$ defines $\sigma^2\mathbf{I}_{m_i}$. In practical terms, $\boldsymbol{\beta}_i$ and $\mathbf{b}_i$ can vary across clusters based on different domains of the SDOH, reflecting the heterogeneous associations among different counties, i. We aim to capture both the average relationships of SDOH across the counties through $\boldsymbol{\beta}_i$ and the specific risk variations for each county with $\mathbf{b}_i$. Therefore, $\boldsymbol{\beta}_i$ indicates how changes in the SDOH variable, j, correlate with CVD mortality in relation to the overall average, while $\mathbf{b}_i$ addresses the unique characteristics of each county. With assumptions on $\mathbf{b}_i$ and $\mathbf{e}_i$ in model 4.1, $\mathbf{y}_i$ given $\mathbf{X}_i,\ \mathbf{Z}_i$ follows a multivariate normal distribution with a particular form of the covariance matrix, that is, $\mathbf{y}_i|\mathbf{X}_i,\ \mathbf{Z}_i \sim N(\mathbf{X}_i\boldsymbol{\beta},\ \sigma^2\mathbf{V}_i)$, where $\mathbf{V}_i = \mathbf{Z}_i\mathbf{D}\mathbf{Z}_i^T + \mathbf{I}_{m_i}$.

We provided background information on the LMM relating SDOH to CVD mortality at the county level. However, a single LMM may not sufficiently capture the underlying heterogeneity in these relationships due to significant geographic disparities across counties. To address this issue, we assume that each county belongs to one of K distinct clusters, where each cluster represents a unique subgroup characterized by specific associations between SDOH and CVD mortality. Typically, the cluster membership for each county remains uncertain. Statistically, this uncertainty can be modeled by assigning a mixing probability that reflects the likelihood of a county belonging to a specific cluster, where $k \in \{1, ..., K\}$. However, this mixing probability is also often unknown and can be estimated from the mixture of LMM parencitedu2013simultaneous, yang2022model.

For a mixture-LMM, each subject i belongs to one of k latent clusters, with the mixing probablilty $P(\tau_i = k) \equiv \pi_k$ subject to $\sum_{k=1}^{K} \pi_k = 1$. Let $w_{ik} = \mathbf{1}_{\{\pi_i = k\}}$ be the binary laten indicator for whether subject i belongs to cluster k, and let $\mathbf{W}_k = (w_{1k}, ..., w_{nK})$ and $\mathbf{W} = (\mathbf{W}_1, ..., \mathbf{W}_K)$. Clustering aims to classify a sample of subjects into one of the K groups based on a defined rule of similarity in their observed patterns. A simple approach is to assume that the observed data $(\mathbf{y}_i, \mathbf{X}_i, \mathbf{Z}_i)$ follows a mixture of LMM across K groups. Then, considering each mixture component to be kth cluster, the model becomes:

$$\mathbf{y}_i = \mathbf{X}_i\boldsymbol{\beta}_k + \mathbf{Z}_i\mathbf{b}_{ik} + \mathbf{e}_{ik}, \tag{4.2}$$

where $\boldsymbol{\beta}_k$ is the cluster-specific fixed effects vector of dimension p_{n_k}, $\mathbf{b}_{ik} \sim N(\mathbf{0},\ \sigma_k^2\mathbf{D}_k)$ is the random effects voectors of size q_{n_k}, and $\mathbf{e}_{ik} \sim N(\mathbf{0},\ \boldsymbol{\sigma}_k^2\mathbf{I}_{m_i})$. Each mixture cluster, k, thus has its own set of parameters $\boldsymbol{\Theta}_k = (\boldsymbol{\beta}_k, \mathbf{D}_k, \sigma_k^2)$. In addition, penalty functions on $\boldsymbol{\beta}$ and $\mathbf{D}$ are applied to both fixed and random effects for simultaneous selection in the mixture of LMM. By regularizing $\boldsymbol{\beta}$, any fixed effect estimated to be zero will be removed from the model. However, based on our previous study (Son et al., 2023), we assume that initial CVD mortality rates differ by county and that there are no additional random effects. Therefore, we do not impose penalties on $\mathbf{D}$ for the random effects.

Cluster assignments for each subject remain uncertain due to the value of w. The Expectation-Maximization (EM) algorithm (Dempster et al., 1977) is a widely used method to address this issue. Then, the number of clusters can be determined by comparing models fitted with various K-values using selection criteria like BIC. Our mixture of the LMM framework allows for separate estimation of fixed effects across multiple clusters, overcoming the limitations of a single LMM in data exhibiting strong geographic or contextual heterogeneity. A comprehensive explanation of this method is provided in the method section of **Chapter 3**.

Furthermore, to determine the optimal number of clusters within each state, we impose constraints on the EM algorithm. Specifically, the algorithm clusters counties at the state level by applying state-specific constraints, such as defined cut-off points. This method effectively reduces the total number of clusters per state, aligning with practical considerations since our work was originally structured at the county level. Consolidating county-level data into fewer, state-level clusters enhances the interpretation of results and allows for targeted policymaking interventions at the state level.

To refine the initially estimated clusters, it is important to ensure that the constraints for cutoff value denoted as f must not exceed 1/max(k), where $k = (1, ..., K)$ represents the number of clusters. This limitation effectively adjusts the clusters, computing the threshold for the number of counties in each state and classifying the clusters into two groups such that,

$$T_{S_l} = C_{S_l} \times f, \tag{4.3}$$

then

$$\begin{aligned} &G_1, \quad \text{when } C_{S_l,K} > T_{S_l}, \\ &G_2, \quad \text{otherwise}, \end{aligned} \tag{4.4}$$

where T represents the threshold defined by the constraint, S denotes an individual state, including US territories, with a state index ranging from $1 \leq l \leq 54$, and C_{S_l} indicates the total number of counties within each respective state S_l. Thus, clusters are categorized into G_1 and G_2. Group G_1 consists of counties from clusters in each state where the number of counties exceeds the threshold T_{S_l}, ensuring that these clusters are substantial enough for independent analysis. While G_2 includes the remaining clusters within the state that do not meet this threshold criterion. The estimation process for $\hat{\boldsymbol{\Theta}}_k$ is initiated

upon establishing specific thresholds for each state. This parameter undergoes an iterative updating process to refine estimates based on current groupings. Subsequently, both subject-specific and overall clustering probabilities, represented as $\hat{w}_{ik}$ and $\hat{\pi}_{ik}$ respectively, are recalculated using the updated estimate $\hat{\boldsymbol{\Theta}}^{(r+1)}$. This process is executed through the EM algorithm to enhance clustering accuracy. In the next step, counties within G_2 that meet or fall below their state's threshold are incorporated into the more robust clusters of G_1. In contrast, clusters comprising counties that surpass the threshold T_{S_l} remain unchanged, preserving their distinct groupings. The following steps elaborate on the methodological steps and implications of these clustering adjustments:

1. Number of counties by state is stratified by each cluster.
2. Given cutoff, T_{S_l} is calculated by each state.
3. At each state, clusters will be classified into either G_1 or G_2 according to 4.4.
4. Returning to the EM algorithm, counties in G_1 will be fixed and those in G_2 clustered to fixed clusters in G_1.
5. It will return until $\boldsymbol{\Theta}$ has converged or the $Q_n(\boldsymbol{\Theta})$ starts to decrease.

4.3 SDOH Data

In Chapter 3 (Aim 2), we employed model-based clustering using Exclusive Lasso with county-level SDOH data for empirical data analysis. We used BIC to determine the number of clustering models (as shown in the Appendix Table B.5). Each number of clustering models was run five times, and the model with seven clusters was selected based on the smallest BIC.

In this chapter, our objective was to reduce the number of clusters in each state for state-level clustering while also decreasing the variations among clusters. We aimed to achieve this by clustering counties, which include only a few counties ($\leq$ threshold), within clusters. Table 4.1 presents the frequency of county counts by clusters at the state level and specifies thresholds derived from tests under specific constraints. The cutoff value cannot exceed 1/7, where 7 is the number of clusters. The threshold of county numbers in each state is calculated by multiplying the total number of counties in each state by the cutoff. For example, the total number of counties in Georgia is 159, and its maximum threshold is calculated as 159 multiplied by 1/7, resulting in 22 (rounded down). If the cutoff exceeds 1/7, the threshold rises above 22, and the algorithm stops restricting which cluster a county can belong to. In other words, all counties in each cluster fall below the threshold, allowing any county to join any cluster. The results of clustering by constraints are shown in Tables 4.2 and 4.3. Following the threshold with a 1/7 cutoff in Georgia, clusters K_5, K_6, and K_7, which contain 22 or fewer counties, are grouped into clusters K_1-K_4 by the EM algorithm. In contrast, clusters K_5 to K_7 remain at zero. Otherwise, counties that exceed the threshold are fixed in clusters. Similarly, we can intuitively observe a reduction in the number of clusters as the cutoff values increase in other states. In addition, the tables show how the variation decreases across 7

clusters. For instance, in cluster K_1, Table 4.1 initially includes 39 states. However, as the constraints increase to 1/7, the number of states in K_1 decreases to 21 in Table 4.3. This trend is also evident in the other six clusters.

Table 4.1: The frequency of county by clusters at the state level and thresholds derived from tests under specific constraints

GEOID	State	K1	K2	K3	K4	K5	K6	K7	Total	Threshold			
										1/28	1/14	3/28	1/7
01	Alabama	20	18	12	7	4	3	3	67	2	4	7	9
02	Alaska	14	3	6	2	1	2	1	29	1	2	3	4
04	Arizona	0	4	1	1	3	5	1	15	0	1	1	2
05	Arkansas	18	18	16	13	6	3	1	75	2	5	8	10
06	California	2	24	1	5	6	17	3	58	2	4	6	8
08	Colorado	13	12	6	13	7	6	7	64	2	4	6	9
09	Connecticut	0	2	0	0	0	4	2	8	0	0	0	1
10	Delaware	0	1	0	0	0	1	1	3	0	0	0	0
11	District of Columbia	0	0	0	1	0	0	0	1	0	0	0	0
12	Florida	4	14	5	14	7	13	10	67	2	4	7	9
13	Georgia	47	33	27	34	9	6	3	159	5	11	17	22
15	Hawaii	0	1	0	1	2	1	0	5	0	0	0	0
16	Idaho	8	6	18	6	3	1	2	44	1	3	4	6
17	Illinois	6	30	9	19	19	12	7	102	3	7	10	14
18	Indiana	10	26	10	21	10	11	4	92	3	6	9	13
19	Iowa	9	28	19	18	13	10	2	99	3	7	10	14
20	Kansas	14	18	33	20	6	11	3	105	3	7	11	15
21	Kentucky	47	28	15	13	9	7	1	120	4	8	12	17
22	Louisiana	24	14	16	8	1	1	0	64	2	4	6	9
23	Maine	0	2	2	2	2	5	3	16	0	1	1	2
24	Maryland	0	7	7	5	1	4	0	24	0	1	2	3
25	Massachusetts	0	2	0	0	1	8	3	14	0	1	1	2
26	Michigan	3	21	10	20	10	14	5	83	2	5	8	11
27	Minnesota	2	5	11	25	14	17	13	87	3	6	9	12
28	Mississippi	34	20	16	10	2	0	0	82	2	5	8	11
29	Missouri	27	37	25	10	3	8	5	115	4	8	12	16
30	Montana	8	8	16	10	7	5	2	56	2	4	6	8
31	Nebraska	13	24	11	24	11	8	2	93	3	6	9	13
32	Nevada	4	4	3	2	2	1	1	17	0	1	1	2
33	New Hampshire	0	0	0	2	2	3	3	10	0	0	1	1
34	New Jersey	0	4	0	0	2	10	5	21	0	1	2	3
35	New Mexico	2	5	4	11	8	1	2	33	1	2	3	4
36	New York	6	27	0	8	3	17	1	62	2	4	6	8
37	North Carolina	7	37	8	15	8	21	4	100	3	7	10	14
38	North Dakota	12	8	8	16	6	1	2	53	1	3	5	7
39	Ohio	8	17	13	13	21	9	7	88	3	6	9	12
40	Oklahoma	20	23	25	2	3	3	1	77	2	5	8	11
41	Oregon	0	5	3	9	7	7	5	36	1	2	3	5
42	Pennsylvania	2	28	4	4	4	15	10	67	2	4	7	9
44	Rhode Island	0	1	0	0	0	4	0	5	0	0	0	0
45	South Carolina	5	18	2	9	2	6	4	46	1	3	4	6
46	South Dakota	16	10	15	18	4	2	1	66	2	4	7	9
47	Tennessee	26	19	24	13	6	5	2	95	3	6	10	13
48	Texas	34	53	55	36	39	27	10	254	9	18	27	36
49	Utah	6	4	3	6	7	1	2	29	1	2	3	4
50	Vermont	0	1	2	1	1	2	2	9	0	1	1	2
51	Virginia	3	17	11	14	10	17	3	75	4	9	14	19
53	Washington	0	6	8	3	6	10	6	39	1	2	4	5
54	West Virginia	3	10	3	11	2	5	1	35	1	3	5	7
55	Wisconsin	9	21	7	18	18	8	2	83	2	5	7	10
56	Wyoming	5	7	7	8	5	2	1	35	0	1	2	3
66	Guam	1	0	0	0	0	0	0	1	0	0	0	0
72	Puerto Rico	40	27	2	2	0	4	3	78	2	5	8	11
78	U.S. Virgin Islands	0	0	0	0	0	0	3	3	0	0	0	0
Total (County)		553	809	501	521	310	360	170					
Total (State)		39	50	42	47	47	49	45					

Table 4.2: County-level clustering with 1/28 and 1/14 cutoffs

GEOID	State	1/28 Cutoff							1/14 Cutoff						
		K1	K2	K3	K4	K5	W6	K7	K1	K2	K3	K4	K5	W6	K7
01	Alabama	20	18	12	7	4	3	3	20	19	14	14	0	0	0
02	Alaska	14	3	6	3	0	3	0	14	7	8	0	0	0	0
04	Arizona	0	4	1	1	3	5	1	0	4	0	0	5	6	0
05	Arkansas	18	18	16	14	6	3	0	18	19	16	15	7	0	0
06	California	0	25	0	6	7	17	3	0	25	0	6	7	20	0
08	Colorado	13	12	6	13	7	6	7	13	12	6	13	7	6	7
09	Connecticut	0	2	0	0	0	4	2	0	2	0	0	0	4	2
10	Delaware	0	1	0	0	0	1	1	0	1	0	0	0	1	1
11	D.C.	0	0	0	1	0	0	0	0	0	0	1	0	0	0
12	Florida	4	14	5	14	7	13	10	0	15	7	15	7	13	10
13	Georgia	47	33	27	34	11	7	0	47	35	33	44	0	0	0
15	Hawaii	0	1	0	1	2	1	0	0	1	0	1	2	1	0
16	Idaho	8	7	18	6	3	0	2	8	7	19	10	0	0	0
17	Illinois	6	30	9	19	19	12	7	0	34	11	19	20	18	0
18	Indiana	10	26	10	21	10	11	4	10	26	10	22	12	12	0
19	Iowa	9	28	19	18	14	11	0	9	28	19	18	13	12	0
20	Kansas	14	18	33	20	8	12	0	14	18	33	26	0	14	0
21	Kentucky	47	28	15	13	10	7	0	47	31	15	17	10	0	0
22	Louisiana	24	15	16	9	0	0	0	24	15	16	9	0	0	0
23	Maine	0	2	2	2	2	5	3	0	2	2	2	2	5	3
24	Maryland	0	7	7	5	1	4	0	0	7	8	5	0	4	0
25	Massachusetts	0	2	0	0	1	8	3	0	2	0	0	0	8	4
26	Michigan	3	21	10	20	10	14	5	0	22	12	21	11	17	0
27	Minnesota	0	6	12	25	14	17	13	0	0	14	29	14	17	13
28	Mississippi	34	20	18	10	0	0	0	34	20	18	10	0	0	0
29	Missouri	27	37	27	11	0	8	5	27	37	27	24	0	0	0
30	Montana	8	8	16	10	7	5	2	8	8	16	11	8	5	0
31	Nebraska	13	24	11	24	12	9	0	13	24	11	24	12	9	0
32	Nevada	4	4	3	2	2	1	1	4	5	3	3	2	0	0
33	New Hampshire	0	0	0	2	2	3	3	0	0	0	2	2	3	3
34	New Jersey	0	4	0	0	2	10	5	0	4	0	0	2	10	5
35	New Mexico	2	5	4	12	8	0	2	0	5	6	12	10	0	0
36	New York	6	27	0	8	4	17	0	6	27	0	11	0	18	0
37	North Carolina	7	37	8	15	8	21	4	0	41	11	16	9	23	0
38	North Dakota	12	8	8	17	6	0	0	12	8	8	17	7	0	0
39	Ohio	16	31	23	23	15	12	10	16	31	24	26	15	14	0
40	Oklahoma	17	19	22	12	8	3	2	17	19	24	14	8	4	0
41	Oregon	0	5	8	10	3	2	1	0	5	9	10	3	3	0
42	Pennsylvania	6	36	10	21	18	22	9	0	36	10	21	18	25	0
44	Rhode Island	0	0	0	0	0	1	1	0	0	0	0	0	1	1
45	South Carolina	18	18	16	12	2	2	0	18	18	16	13	2	3	0
46	South Dakota	3	3	13	6	5	0	0	3	3	13	7	5	0	0
47	Tennessee	41	24	22	11	0	6	3	41	24	23	13	0	6	0
48	Texas	56	47	30	37	16	22	15	56	50	37	47	0	0	0
49	Utah	6	4	3	6	7	0	3	6	4	3	6	10	0	0
50	Vermont	0	0	0	0	0	2	1	0	0	0	0	0	2	1
51	Virginia	5	16	6	16	4	10	4	0	18	8	17	5	11	0
53	Washington	4	12	13	14	10	15	7	4	12	15	15	12	18	0
54	West Virginia	8	3	10	5	2	0	0	8	3	10	6	2	0	0
55	Wisconsin	4	14	19	20	14	17	6	0	15	20	21	15	17	0
56	Wyoming	7	9	15	9	5	4	3	7	9	15	9	5	4	0
66	Guam	1	0	0	0	0	0	0	1	0	0	0	0	0	0
72	Puerto Rico	42	27	0	0	0	5	4	46	32	0	0	0	0	0
78	U.S. Virgin Islands	0	0	0	0	0	0	3	0	0	0	0	0	0	3
Total (County)		547	817	505	526	316	360	153	531	855	532	598	272	342	94
Total (State)		35	50	40	45	43	42	33	30	48	36	42	28	31	16

Table 4.3: County-level clustering with 3/28 and 1/7 cutoffs

GEOID	State	3/28 Cutoff							1/7 Cutoff						
		K1	K2	K3	K4	K5	W6	K7	K1	K2	K3	K4	K5	W6	K7
1	Alabama	21	29	17	0	0	0	0	20	30	17	0	0	0	0
2	Alaska	16	0	13	0	0	0	0	16	0	13	0	0	0	0
4	Arizona	0	4	0	0	5	6	0	0	4	0	0	5	6	0
5	Arkansas	18	19	21	17	0	0	0	18	19	19	19	0	0	0
6	California	0	37	0	0	0	21	0	0	37	0	0	0	21	0
8	Colorado	14	15	0	16	9	0	10	14	15	0	35	0	0	0
9	Connecticut	0	2	0	0	0	4	2	0	2	0	0	0	4	2
10	Delaware	0	1	0	0	0	1	1	0	1	0	0	0	1	1
11	D.C.	0	0	0	1	0	0	0	0	0	0	1	0	0	0
12	Florida	0	17	0	27	0	13	10	0	15	0	27	0	13	12
13	Georgia	47	35	33	44	0	0	0	47	36	33	43	0	0	0
15	Hawaii	0	1	0	1	2	1	0	0	1	0	1	2	1	0
16	Idaho	8	7	19	10	0	0	0	9	0	35	0	0	0	0
17	Illinois	0	36	0	27	21	18	0	0	41	0	32	29	0	0
18	Indiana	10	26	10	22	12	12	0	0	40	0	52	0	0	0
19	Iowa	0	35	22	25	17	0	0	0	34	23	42	0	0	0
20	Kansas	14	21	33	37	0	0	0	0	24	44	37	0	0	0
21	Kentucky	47	31	19	23	0	0	0	63	57	0	0	0	0	0
22	Louisiana	24	15	16	9	0	0	0	24	21	19	0	0	0	0
23	Maine	0	2	2	2	2	5	3	0	0	0	0	0	9	7
24	Maryland	0	7	8	5	0	4	0	0	7	7	6	0	4	0
25	Massachusetts	0	2	0	0	0	8	4	0	0	0	0	0	10	4
26	Michigan	0	22	12	23	11	15	0	0	22	0	44	0	17	0
27	Minnesota	0	0	14	29	14	17	13	0	0	0	38	19	17	13
28	Mississippi	34	20	18	10	0	0	0	35	25	22	0	0	0	0
29	Missouri	27	53	35	0	0	0	0	27	53	35	0	0	0	0
30	Montana	8	8	16	16	8	0	0	8	10	18	20	0	0	0
31	Nebraska	13	25	11	30	14	0	0	0	34	0	59	0	0	0
32	Nevada	4	5	3	3	2	0	0	4	6	7	0	0	0	0
33	New Hampshire	0	0	0	2	2	3	3	0	0	0	2	2	3	3
34	New Jersey	0	4	0	0	0	10	7	0	4	0	0	0	10	7
35	New Mexico	0	5	6	12	10	0	0	0	7	0	15	11	0	0
36	New York	0	33	0	11	0	18	0	0	42	0	0	0	20	0
37	North Carolina	0	43	0	34	0	23	0	0	43	0	33	0	24	0
38	North Dakota	12	8	8	18	7	0	0	12	8	9	24	0	0	0
39	Ohio	0	23	16	22	27	0	0	0	26	17	13	32	0	0
40	Oklahoma	20	28	29	0	0	0	0	20	28	29	0	0	0	0
41	Oregon	0	6	0	10	8	7	5	0	0	0	14	13	9	0
42	Pennsylvania	0	39	0	0	0	15	13	0	38	0	0	0	15	14
44	Rhode Island	0	1	0	0	0	4	0	0	1	0	0	0	4	0
45	South Carolina	7	18	0	14	0	7	0	0	25	0	21	0	0	0
46	South Dakota	16	11	16	23	0	0	0	16	11	15	24	0	0	0
47	Tennessee	26	21	31	17	0	0	0	26	34	35	0	0	0	0
48	Texas	34	56	55	60	49	0	0	0	87	93	0	74	0	0
49	Utah	7	6	0	6	10	0	0	7	0	0	12	10	0	0
50	Vermont	0	0	0	3	7	0	4	0	0	0	0	10	0	4
51	Virginia	26	49	23	35	0	0	0	26	67	40	0	0	0	0
53	Washington	0	28	0	0	0	11	0	0	28	0	0	0	11	0
54	West Virginia	10	28	6	11	0	0	0	11	29	0	15	0	0	0
55	Wisconsin	0	16	0	20	11	15	10	0	16	0	34	0	22	0
56	Wyoming	3	7	4	9	0	0	0	0	9	5	9	0	0	0
66	Guam	1	0	0	0	0	0	0	1	0	0	0	0	0	0
72	Puerto Rico	46	32	0	0	0	0	0	43	35	0	0	0	0	0
78	U.S. Virgin Islands	0	0	0	0	0	0	3	0	0	0	0	0	0	3
Total (County)		513	937	516	684	248	238	88	447	1072	535	672	207	221	70
Total (State)		27	47	29	38	21	23	14	21	42	21	27	11	20	11

Figure 4.1 presents a detailed analysis of data grouped into seven distinct clusters based on various constraints over the decade from 2009 to 2018. It provides insights into the variability within each cluster. The data reveals a trend of decreasing mean values of the CVD mortality rate in most clusters over the observed years, indicating a systematic change or underlying trend affecting the clusters. Standard deviations within each cluster remain relatively stable across the years, suggesting that the variability in the data within clusters does not significantly fluctuate over time. The trend of each cluster in each figure with various constraints shows similar trends to the initial analysis. We observe some changes in fluctuations and slopes, but generally, it retains the same characteristics of the downward trend in CVD mortality. For example, the mean of age-adjusted CVD mortality by cutoff 1/7, which contains the highest constraint, is depicted in Table C.1 in the Appendix. The clusters exhibit varying patterns of mortality rate changes, with most showing a gradual decline over time. In cluster K_1, the values decrease from the highest CVD mortality, 304.75 deaths per 100,000 people (95% CI, 298.15 - 311.36), to 268.41 deaths per 100,000 people (95% CI, 261.56 - 275.26), with a consistent decline over the sequence (Table C.1 in the Appendix). The overall decrease rate seems steady, although it may slow from 2012 to 2015. In cluster K_2, the values continue to decrease similar to K_1, and the rates in this cluster fluctuate less compared to K_1. While there are some small fluctuations in the rate of decrease, the overall trend is a consistent reduction downward. Cluster K_3 exhibits a more variable pattern. It shows a less pronounced decline compared to the first two clusters. Around 2016-2017, it plateaus before declining again toward 2018. This cluster has a medium-high CVD mortality rate, and the increase observed from 2012 to 2016 suggests that the reduction in CVD mortality is either slower or stalled during this period. Cluster K_5 follows a similar trend as K_3, but it has fewer fluctuations and remains consistently in the lower-middle range from 232.58 deaths per 100,000 people (95% CI, 227.55 - 237.61) in 2009 to 227.84 deaths per 100,000 people (95% CI, 222.81 - 232.87) in 2018. Cluster K_4 shows a moderate decline in mortality rates but with less variability and a smoother downward trajectory. Around 2013-2016, the trajectory is on a plateau, then declines again toward 2018. Cluster K_6 starts with a low mortality rate and experiences a slow but stable decrease, making it one of the clusters with the best CVD outcomes. Finally, K_7 consistently has the lowest CVD mortality rates, with only minor changes year by year, indicating a stable and positive CVD outcome.

The map in Figure 4.2 shows the geographic distribution of US counties grouped into seven distinct clusters based on CVD mortality from 2009 to 2018. Each county is color-coded according to the cluster it belongs to, reflecting different characteristics of CVD mortality patterns within each cluster. Compared to the map in Figure 3.11, both maps, represented by cutoffs 1/14 and 1/7, demonstrate a decrease in variation among counties within each state. These counties may represent specific regional trends in healthcare, lifestyle, or socioeconomic factors that define their CVD mortality patterns. The geographic distribution of the clustered CVD mortality with a 1/7 cutoff describes that counties in cluster K_1 are concentrated in the southeastern region, especially in states like Kentucky, West Virginia, and parts of Alabama and Mississippi. These areas may exhibit unique CVD mortality patterns driven by rural healthcare dynamics, local health infrastructure, or regional health behaviors. Counties in K_2 are scattered across the US, with notable regions in Texas, parts of the Midwest, and the West, while K_3 has a broad distribution, spanning the western and central regions. Cluster K_4 counties are primarily found in the West, Midwest, and Great

Plains. Counties in K_7 are clustered mainly in the northeastern states, including Maine, Vermont, and New Hampshire, and Florida in the south.

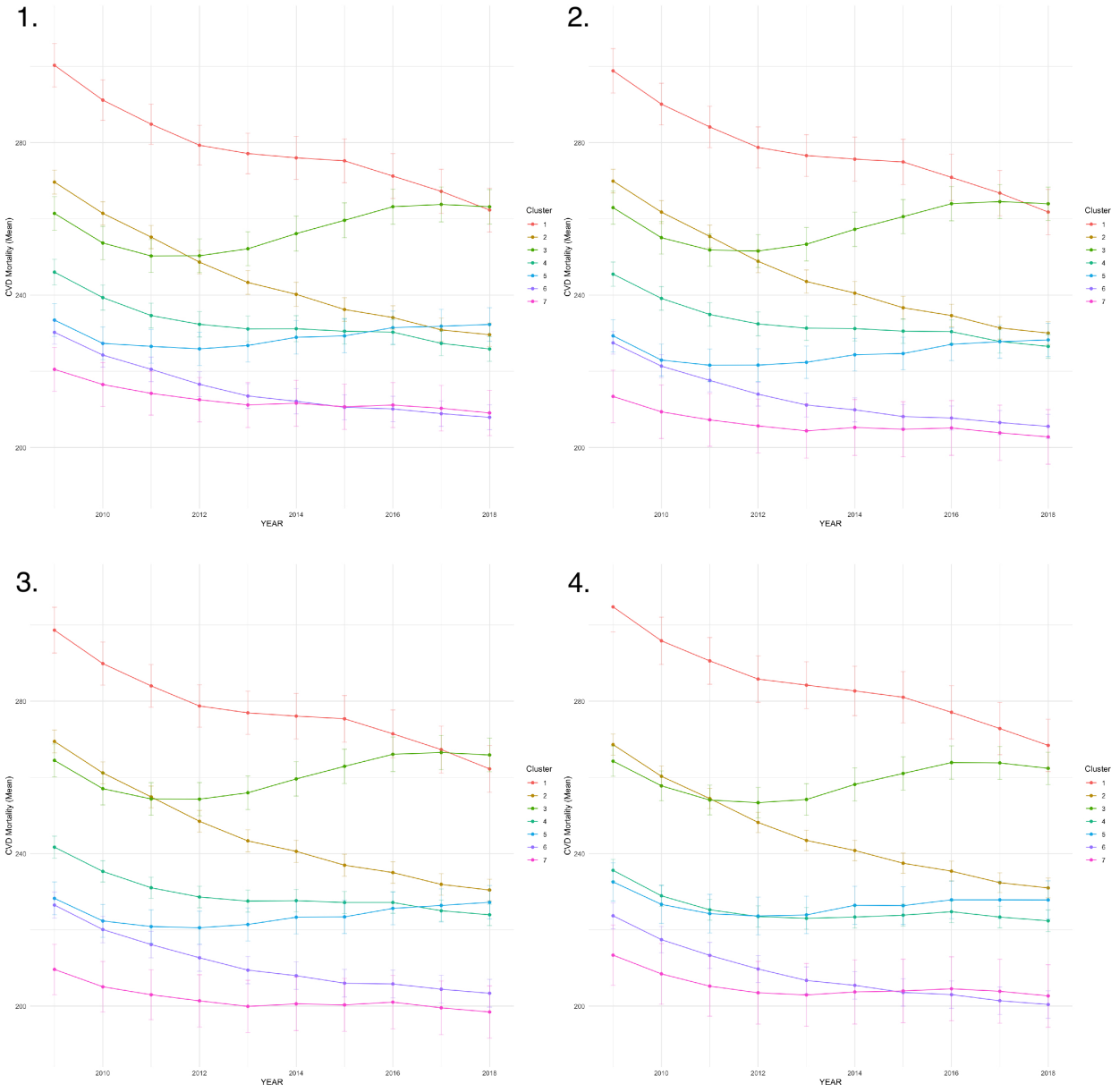


Figure 4.1: Mean age-adjusted cardiovascular disease mortality, per 100,000 people, trajectories of the US counties from 2009 to 2018 in the seven clusters identified by the method: 1. CVD mortality trajectories with cutoff 1/28 (Upper Left); 2. CVD mortality trajectories with cutoff 1/14 (Upper Right); 3. CVD mortality trajectories with cutoff 3/28 (Lower Left); 4. CVD mortality trajectories with cutoff 1/7 (Lower Right).

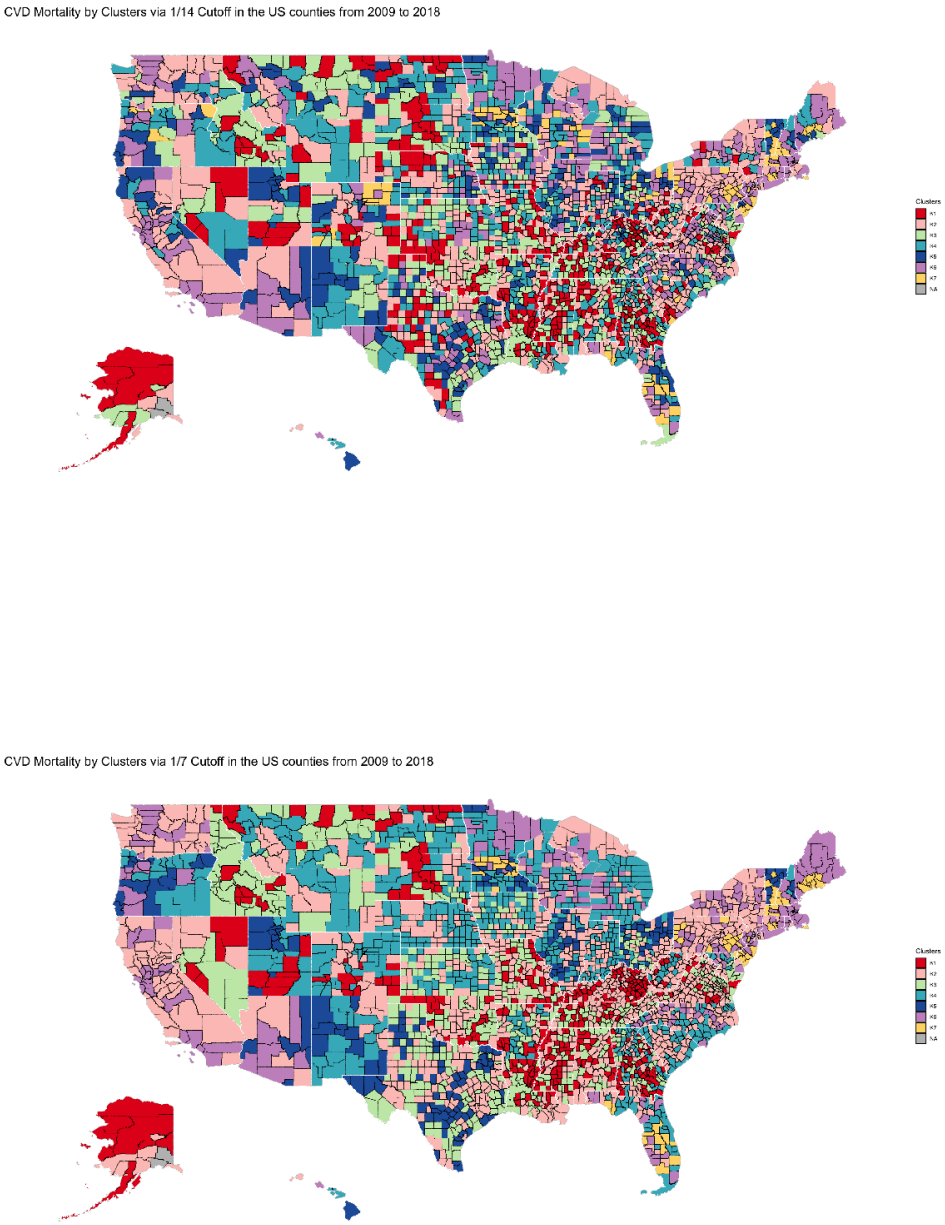


Figure 4.2: Geographic distribution of clusters based on age-adjusted cardiovascular disease mortality per 100,000 people across counties in the United States, clustering via 1/14 cutoff versus 1/7 cutoff.

The variables selected for each cluster with a 1/7 cutoff are listed in Table 4.4 (Additional Appendix Table C.2 provides the information for a 1/14 cutoff). The clusters in the table have a distinct intercept, which reflects the mean age-adjusted CVD mortality rate, allowing for varying intercepts across clusters. The geographic indicator for rural-urban status is selected for all clusters since it is the predominant variable in the geographic domain. After adjusting the model with the constraint, rural counties in all clusters have an increased overall CVD mortality rate compared to urban counties. Especially, the changes in CVD mortality between rural and urban counties are higher in cluster K_1 ($\beta = 6.53$), K_3 ($\beta = 5.99$), K_5 ($\beta = 6.67$), and K_7 ($\beta = 6.75$). Additionally, the percentage of Black residents, employment in manufacturing, population with graduate degrees, and those with less than a high school education, the total number of community mental health centers, the number of people living with diagnosed HIV per 1,000 people, and the ratio of Medicare coverage among the Medicare-eligible population are identified as significant variables associated with CVD mortality over time in overall clusters.

The selected variables vary across clusters, highlighting different SDOH factors associated with CVD mortality patterns in these groups. Cluster K_1 includes 447 counties, exhibiting the highest average CVD mortality intercept at 281.52 per 100,000 people. In this social context, a higher percentage of Black residents, single-parent families with children, children living with a grandparent householder aged 17 and under, and a civilian population consisting of veterans (age ge 18) are associated with a higher CVD death rate within this cluster. Conversely, a lower CVD mortality rate is linked to a higher percentage of housing units with more than one occupant per room, American Indian/Alaska Native and Hispanic ethnicity residents, a population not speaking English at all (age ge 5), foreign-born residents, occupied housing units without fuel, and a population that is not US citizens and entered the US before 1990. From the economic perspective, higher median household income and a greater percentage of jobs in arts, entertainment, recreation, accommodation, food services, finance, insurance, real estate, rental, and leasing, along with households that have received food stamps/Supplemental Nutrition Assistance Program (SNAP) in the past 12 months, are associated with a decrease in CVD mortality. Additionally, employment in agriculture, forestry, fishing, hunting, and mining (age $\geq$ 16) and a population with an income-to-poverty ratio of 1.25 to 1.99 also correlate with lower CVD mortality rates. Conversely, an increase in the percentage of the civilian population in the armed forces (age $\geq$ 16), as well as those employed in construction, manufacturing, other services (excluding public administration), transportation, warehousing, and utilities, along with unemployed individuals (age $\geq$ 16), is associated with higher CVD mortality rates. In the context of education, a higher percentage of people aged 25 and older who have only a high school diploma, along with those with lower education levels, is associated with increased CVD mortality. Conversely, an inverse association is observed with the percentage of the population holding a master's degree or higher in the same age group. From the domain of physical infrastructure, an increase in the median home value of owner-occupied housing units, as well as the presence of full-service restaurants and supermarkets per 1,000 people, is associated with a decrease in CVD mortality. Additionally, a higher percentage of housing in structures with 10 or more units and the percentage of vacant housing units also contribute to lower CVD mortality rates. Conversely, the percentage of housing units that are mobile homes is inversely related to mortality rates. Moreover, workers aged 16 and older who experience a public transit commute

longer than 60 minutes, along with a higher density of liquor stores, community food services aimed at low-income or elderly individuals, and convenience stores per 1,000 persons, are correlated with an increase in CVD mortality. Most variables in the healthcare context were selected. CVD mortalities increase with the number of federally qualified health centers and the number of people living with HIV diagnosed per 1,000 people. Conversely, the total number of community mental health centers, the number of Medicare-eligible individuals in the county, and the ratio of Medicare Advantage enrollees to original Medicare-eligible individuals are associated with a decrease in CVD mortality.

We can describe the selected variables similarly to those in cluster K_1. In cluster K_2, the largest study sample comprised 1,072 counties across 42 states, accounting for 33.25% of the total counties. In contrast, only 70 counties from 11 states were associated with cluster K_7. Most healthcare-related variables have been included across all clusters. Clusters K_3 and K_4 include all five education-related variables; however, the percentage of the population without a high school education shows an inverse relationship within cluster K_3. Likewise, the ratio of Medicare Advantage enrollees to original Medicare-eligible individuals demonstrates inverse relationships with CVD mortality across all other clusters.

Finally, we map the nonzero coefficients selected across all seven clusters, as shown in Figure 4.3. This mapping characterizes the spatial structures of these coefficients, highlighting how the selected estimates correspond to the various clusters within each county across the state. Figure 4.3 describes the spatial structures of estimates selected using a cutoff of 1/7, which were previously described. Additionally, Figure C.1 in the appendix presents the spatial distribution of the nonzero coefficients, chosen without any constraints, related to CVD mortality across US counties. They include factors such as the percentage of Black residents, employment in manufacturing, and individuals aged 25 or older with a master's degree or higher, the number of Federally Qualified Health Centers, the ratio of Medicare Advantage enrollees to individuals eligible for original Medicare, and the distinction between rural and urban regions.

Table 4.4: Selected social determinants of health associated with age-adjusted CVD mortality in each cluster by 1/7 cutoff county-level clustering: Model-based clustering via Exclusive lasso with random effects in intercepts, 2009–2018

Domains	Variables	K1	K2	K3	K4	K5	W6	K7
	Int	281.5173	244.3994	257.6984	227.4570	228.0588	209.0151	209.7441
1	% Housing units with more than one occupant per room	-0.1895	0	0	0	0	0	0
1	% Population reporting American Indian/Alaska Native race	-0.0135	0	0	0	0	0	0
1	% Population reporting Asian race	0	0	0	0	-0.4806	-0.1099	-0.7479
1	% Population reporting Black race	0.2784	0.4295	0.6935	0.7863	0.9746	0.4836	0.6204
1	% Families with Children that are single-parent Families	0.0539	0	0.05396535	0	0	0	0
1	% Population that does not speak English at all (ages 5 and over)	-0.1886	0	0	0	-0.4286	0	0
1	% Population that is foreign-born	-0.4619	-0.0394	-0.5077	-0.6380	0	0	0
1	% Children living with a grandparent householder (ages 17 and under)	0.0103	0	0.1460	0	0.2014	0	0.3904
1	% Occupied housing units without fuel	-0.2480	0	0	0	0	0	-0.0319
1	% Population reporting Hispanic ethnicity	-0.5767	-0.5985	-0.0688	-0.2313	0	0	0
1	% Population reporting multiple races	0	0	0	0	0	-0.4620	0
1	% Population reporting Native Hawaiian/Pacific Islander race	0	-0.1545	4.4824	0	-0.2368	0	0
1	% Population who are not U.S. citizens and entered the U.S. before 1990	-0.2524	0	-0.1831	0	0	0.0379	0.4696
1	% Population who speak other languages (ages 5 and over)	0	0	0	-0.1318	-0.0273	0	0
1	% Civilian Population consisting of veterans (ages 18 and over)	0.1349	1.9593	-0.6942	0	0	1.2679	0
2	Median household income (in dollars) /1000	-0.3694	-0.2141	-0.0865	-0.0788	0	0	0
2	% Unmarried partner households that received food stamps/SNAP benefits	0	-0.1872	-0.0103	-0.0428	0	0	0
2	% Employed working in public administration	0	-0.0492	-0.0151	0	-0.0055	0	0
2	% Civilian Population in armed forces (ages 16 years and over)	0.8712	0	0.1504	0.0369	0	0	0
2	% Employed working in arts, entertainment, recreation, etc.	-0.0752	0	0	0	0.0248	0	0
2	% Employed working in construction	0.7018	0.1660	0.0473	0	0	0.4971	0.1406
2	% Employed working in finance and insurance, real estate, etc.	-0.1748	0	-0.0824	0	0	0	0
2	% Households that received food stamps/SNAP, past 12 months	-0.1768	-0.7203	0	0	0	-0.6180	-0.2044
2	% Households with public assistance income or food stamps/SNAP	0	0	0	0	-0.8760	0	-0.7822
2	% Employed working in information services	0	0	-0.5012	2.4148	0.3287	0	0.4650
2	% Employed working in manufacturing	0.0007	0.1454	0.1442	0.3733	0.3540	0.6239	0.3015
2	% Employed working in agriculture, forestry, fishing, etc. (ages 16 and over)	-0.7061	-0.0488	-0.2288	-0.1613	0	0	0
2	% Employed working in other services, except public administration	0.3180	0	0	-0.1973	0	0	0.2731
2	% Population with income to poverty ratio: 1.25-1.99	-0.1974	0	-0.1098	0	0	0	0
2	% Population with income to poverty ratio: <1.00	0	0	0.1436	0	0	0	0
2	% Employed working in professional, scientific, management, administrative, etc.	0	0	-0.0394	0	0	0	0
2	% Employed working in transportation and warehousing, and in utilities	0.0613	0.1371	0	0	0.3688	0	0
2	% Population that was unemployed (ages 16 years and over)	0.8257	0.1272	-0.9676	-0.5223	-1.0294	0	-0.2140
2	% Civilian veterans in labor force (ages 18–64)	0	0	-0.0029	0	0	0	0
2	% Employed working in wholesale trade	0	0.7144	0.6180	0	0	0.0825	0
3	% Population with some college or associate's degree (ages 25 and over)	0	0	-0.0489	-0.0265	0	0	0
3	% Population with a bachelor's degree (ages 25 and over)	0	0	-0.2129	-0.2163	0	0	0
3	% Population with a master's or higher degree (ages 25 and over)	-0.3743	-1.8717	-0.6684	-0.3960	-0.3789	-1.7894	-0.9483
3	% Population with only high school diploma (ages 25 and over)	0.3667	0	0.1780	0.1678	0.0375	0	0.0480
3	% Population with less than high school education (ages 25 and over)	1.6628	2.1361	-0.0131	0.8744	0.3266	0.6380	0.2165
4	Median home value of owner-occupied housing units	-0.00012	-0.00005	-0.00003	-0.00007	-0.00005	0	0
4	% Housing in structures with 10 or more units	-0.2798	-0.1710	0	-0.1538	0	0	0
4	% Workers with at least 60-minute commute time (ages 16 and over)	0	0	-0.0238	0	0	-0.1070	0
4	% Housing units that are mobile homes	0.0635	0.2240	0.2984	0.3081	0	0.7395	0.8406
4	% Housing units with no vehicle available	0	0.3094	0.0106	0.0370	0	0	0
4	Workers (16 +) with a 60+ min public transit commute	0.0224	0.0076	0	-0.0187	0.0067	-0.0099	-0.0452
4	% Workers taking public transportation, excluding taxicab (ages 16 and over)	0	0	0	0	-1.5002	-0.0406	0
4	% Rental units with rent equal to 30 percent or more of household income	0	0	-0.0698	0	0	0	0
4	% Housing units vacant	-0.4077	-0.3791	0	0	0	0	-0.0017
4	Beer, wine and liquor stores per 1,000 people	16.1077	-3.4531	9.1692	0	3.2268	0	0
4	Community food services (targeting low-income or elderly) per 1,000 people	10.6486	0	-4.1385	0	0	0	-0.2572
4	Convenience stores per 1,000 people	4.0803	0	0	1.2887	0	0	0
4	Full service restaurants per 1,000 people	-1.4146	0	-2.5332	0	0	-0.2352	0
4	Supermarkets and other grocery (except convenience) stores per 1,000 people	-16.2587	3.1321	-0.8939	0	-2.8575	0.4692	0
5	Number of Federally Qualified Health Centers	0.1070	-0.0045	1.5150	0.4217	0	-0.0554	0
5	Total number of community mental health centers	-1.8538	-0.3279	-0.3137	-0.0106	-1.4840	-0.1028	0.8425
5	Number of rural health clinics	0	0.1367	0.8509	-0.1625	0.7734	0	0.0377
5	Number of people living with diagnosed HIV / 1000	20.0988	0.4038	0.3568	0.5933	-0.9707	0.0930	0.0072
5	Number of Medicare eligibles in the county	-0.0006	0	0	0	0	0	-0.00001
5	Medicare, ratio of enrollees over Medicare- eligible, %	-0.4495	-0.3824	0.4043	-0.1036	-0.0128	-0.2488	-0.0667
6	1: Rural, 0: Urban by NCHS 2013 Rural-Urban Classification Scheme	6.5296	2.3053	5.9880	0.1211	6.6734	1.7759	6.7509

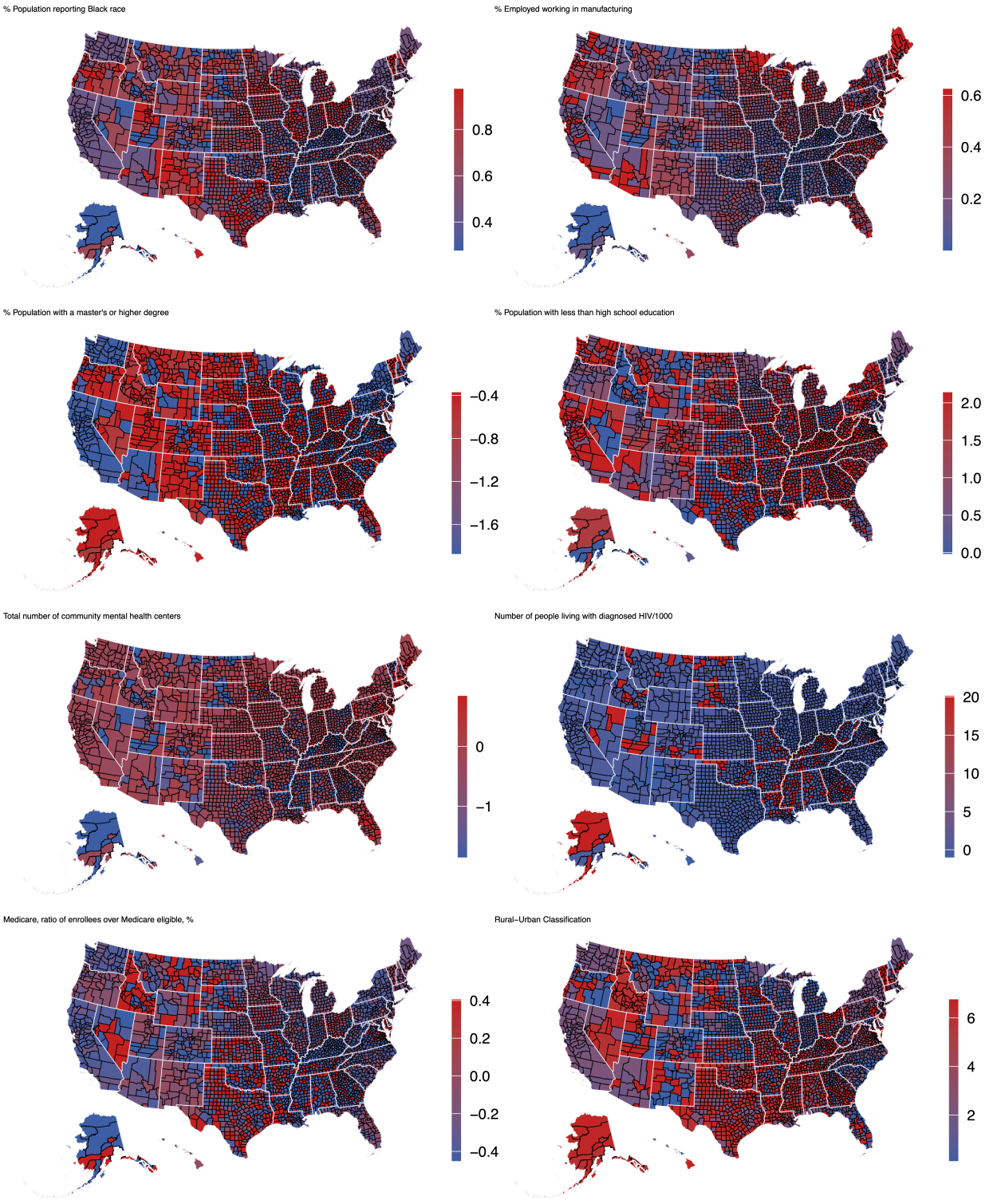


Figure 4.3: The spatial structures of nonzero coefficients with 1/7 cutoff threshold associated with age-adjusted cardiovascular disease mortality across counties in the United States, Left: % Population reporting Black race, % Population with a master's or higher degree, Total number of community mental health centers, Medicare, ratio of enrollees over Medicare eligible, %; Right: % Employed working in manufacturing, % Population with less than high school education, Number of people living with diagnosed HIV/1000, Rural-Urban Classification.

4.4 Summary

This chapter introduced a refined model-based clustering strategy using an Exclusive Lasso penalty to analyze high-dimensional, county-level SDOH data associated with CVD mortality. The approach builds upon a mixture of linear mixed-effects models to group counties into clusters exhibiting similar CVD mortality patterns associated with SDOH. Subsequently, additional state-level constraints are imposed to reduce the number of clusters and refine within-state variation.

The initial clustering employs a mixture of linear mixed-effects models for the county-level dataset. An Exclusive Lasso penalty enforces sparsity in selecting fixed-effect coefficients, identifying each cluster's most relevant SDOH variables linked to CVD mortality. A Bayesian Information Criterion determines the optimal number of clusters. The method introduces state-level constraints after identifying a global solution (e.g., seven clusters). A cutoff threshold limits the size of clusters within each state, ensuring practicality for policy-making. Counties in tiny clusters are reassigned to more extensive, fixed clusters via the EM algorithm. This systematic reduction in within-state cluster heterogeneity preserves the fundamental SDOH patterns elucidated by the model.

The final clustering results, validated with various cutoff values, demonstrate geographically coherent clusters exhibiting distinct CVD mortality trajectories from 2009 to 2018. Most clusters exhibit a general decline in mortality, although the magnitude and slope of reduction vary across groups. Rural counties consistently exhibit higher mortality rates compared to urban counties. Factors such as the percentage of Black residents, educational attainment, and the number of community mental health centers significantly influence the mortality trajectory. The relative importance and direction of these SDOH factors differ across the seven identified clusters.

By simultaneously balancing interpretability (fewer clusters in each state) and modeling accuracy (via Exclusive Lasso in a mixture framework), this method provides detailed insights into how various socioeconomic and health-related factors (i.e., SDOH variables) associate with CVD mortality at the county level within each state. The spatial mapping of nonzero coefficients and cluster assignments highlights localized disparities between rural and urban areas and identifies variables, such as employment in manufacturing, HIV prevalence, and educational attainment, that could inform more targeted health interventions.

This chapter illustrates how a penalized model-based clustering approach can efficiently handle large-scale longitudinal data, minimize unnecessary fragmentation within states, and provide more precise insights into the SDOH factors that correlate significantly with CVD mortality. The proposed framework serves as a valuable resource for health policymakers aiming to identify and prioritize strategies tailored to specific regions to address ongoing cardiovascular health disparities.

Chapter 5

Conclusion

This dissertation enhances methodological approaches for examining high-dimensional longitudinal data with missing information, focusing on CVD mortality disparities associated with SDOH. While overall CVD mortality rates in the US have significantly decreased, notable geographic, racial, ethnic, and socioeconomic disparities continue to persist.

The study addresses several key limitations observed in existing analytical methods. Initially, the proposed penalized weighted GEE with Exclusive Lasso regularization directly addresses missing data by employing IPW under MAR assumptions. This method integrates robust estimation procedures and effectively addresses the gaps left by traditional approaches that do not accommodate missingness in penalized longitudinal data analysis, ensuring more reliable and unbiased inferences. In addition, the Exclusive Lasso penalty is uniquely used to support domain-specific grouped variable selection. Unlike traditional penalization methods, such as Lasso or group Lasso, the Exclusive Lasso approach selectively identifies representative predictors from each defined domain, improving model interpretability while retaining critical information from complex covariate structures.

Given the inherent geographic and demographic diversity, a new model-based clustering technique for high-dimensional longitudinal data is proposed. This approach employs Exclusive Lasso regularization to identify variables across various county subpopulations, each influenced by different SDOH associated with CVD mortality. Additionally, the model-based clustering incorporates regularization for mutually exclusive features in domains to address variability within states, enhancing the precision of subgroup identification at the state level.

A limitation of this study is that the Exclusive Lasso mandates the selection of at least one variable from each domain, potentially leading to the inclusion of weak predictors when specific domains do not present any truly significant SDOH variables. This mandatory selection across all domains could introduce noise and reduce model precision. Furthermore, our analysis neglected to consider the spatial correlation among counties, which is a significant oversight, as geographic proximity often results in similar socioeconomic conditions and health outcomes. Future research can employ domain-flexible selection algorithms that allow for the exclusion of entire domains when supported by the data. This might involve developing hybrid approaches that combine Exclusive Lasso with other variable selection methods to establish domain-

specific significance thresholds. Furthermore, integrating spatial econometric modeling techniques such as spatial lag or spatial error models would address geographic interdependencies, potentially uncovering regional variations in SDOH impact that our current approach obscures.

In conclusion, our proposed methodological advancements significantly enhance analytical capabilities in addressing high-dimensional longitudinal data with missing values and multimodal covariate structures. By improving the interpretability and precision of statistical inferences, these methods provide valuable tools for stakeholders and policymakers, enabling targeted and evidence-based interventions to reduce disparities in CVD mortality across geographic and demographic groups within the US.

Appendix A

Penalized Weighted Generalized Estimating Equations via Exclusive Lasso Penalty

Table A.1: C (the number of zero coefficients that are correctly estimated by zero), IC (the number of non-zero coefficients that are incorrectly estimated by zero), GC (the number of groups that are correctly selected), and MSE (Mean Squared Error) for PWGEE-Lasso(P-Lasso), PWGEE-SCAD (P-SCAD), PWGEE-MCP (P-MCP), PWGEE-cMCP (P-cMCP), and PWGEE-Exclusive lasso (P-eLasso) method with n = 100 and p =(50, 100, 200), which has unequal group sizes, in each scenario.

scenarios	corr	methods	C	IC	CG	MSE	C	IC	CG	MSE	C	IC	CG	MSE
				p=50				p=100				p=200		
1	0.60	P-Lasso	31.44	0.02	**5.00**	0.17	67.40	**0.00**	**5.00**	0.20	149.63	**0.00**	**5.00**	0.24
		P-SCAD	32.08	0.19	4.98	0.40	67.52	0.15	**5.00**	0.75	151.78	0.20	**5.00**	1.04
		P-MCP	35.30	0.25	4.93	0.42	74.31	0.21	**5.00**	0.77	159.22	0.34	4.99	1.56
		P-cMCP	32.80	0.20	4.93	0.31	69.08	0.13	**5.00**	0.43	148.09	0.15	4.99	0.87
		P-eLasso	**38.46**	**0.01**	**5.00**	**0.10**	**87.03**	**0.00**	**5.00**	**0.10**	**185.13**	**0.00**	**5.00**	**0.10**
	0.90	P-Lasso	37.01	0.65	4.83	0.64	81.47	0.77	4.91	0.60	176.04	0.76	4.94	0.55
		P-SCAD	35.56	2.29	4.60	2.92	72.77	2.11	4.83	5.93	176.40	2.72	4.69	2.73
		P-MCP	36.60	2.24	4.54	2.83	75.73	2.12	4.79	5.53	178.52	2.61	4.59	3.69
		P-cMCP	35.56	2.32	4.19	2.73	72.93	2.09	4.70	5.84	178.02	2.66	4.52	2.52
		P-eLasso	**39.16**	**0.26**	**5.00**	**0.37**	**88.17**	**0.17**	**5.00**	**0.26**	**186.76**	**0.07**	**5.00**	**0.32**
2	0.60	P-Lasso	27.73	0.26	**5.00**	0.31	63.25	0.32	**5.00**	0.39	139.54	0.33	**5.00**	0.77
		P-SCAD	28.11	2.09	4.95	0.80	62.17	2.26	4.99	1.52	145.14	2.92	**5.00**	2.43
		P-MCP	30.89	2.57	4.88	0.81	68.32	2.70	**5.00**	1.61	149.11	3.20	**5.00**	3.41
		P-cMCP	29.01	2.23	4.85	0.71	64.53	2.06	4.97	1.06	141.66	2.70	**5.00**	2.29
		P-eLasso	**32.19**	**0.18**	**5.00**	**0.24**	**80.12**	**0.26**	**5.00**	**0.25**	**171.39**	**0.22**	**5.00**	**0.43**
	0.90	P-Lasso	33.15	3.25	4.79	1.34	76.22	3.26	4.97	1.39	168.29	3.20	4.96	5.50
		P-SCAD	31.35	5.93	4.63	5.52	68.75	5.96	4.84	11.71	170.93	6.83	4.70	7.53
		P-MCP	31.93	5.95	4.52	5.54	71.68	6.00	4.74	10.58	173.25	6.82	4.69	9.62
		P-cMCP	31.30	6.11	3.97	5.56	70.78	6.10	4.62	10.19	173.64	6.84	4.57	5.13
		P-eLasso	**33.70**	**2.33**	**5.00**	**1.07**	**81.60**	**2.47**	**5.00**	**0.95**	**182.73**	**2.23**	**5.00**	**0.63**
3	0.60	P-Lasso	28.68	**0.11**	4.99	0.25	64.64	**0.19**	**5.00**	0.32	143.30	**0.15**	**5.00**	0.48
		P-SCAD	29.55	0.77	4.95	0.57	63.42	1.02	4.98	1.21	147.69	1.47	**5.00**	1.64
		P-MCP	32.72	1.01	4.87	0.58	70.48	1.31	4.95	1.20	152.85	1.76	4.98	2.62
		P-cMCP	30.52	0.76	4.88	0.48	65.84	0.99	4.96	0.76	144.49	1.26	4.99	1.70
		P-eLasso	**32.88**	0.13	**5.00**	**0.22**	**77.69**	0.23	**5.00**	**0.26**	**169.22**	0.25	**5.00**	**0.46**
	0.90	P-Lasso	34.88	2.20	4.71	0.96	77.51	1.99	4.90	1.12	169.93	2.20	4.94	3.77
		P-SCAD	32.92	4.76	4.39	4.37	71.14	4.35	4.85	8.08	173.22	5.04	4.75	6.06
		P-MCP	33.15	4.70	4.39	4.44	71.95	4.36	4.83	8.19	175.59	5.10	4.65	7.16
		P-cMCP	33.02	4.57	3.89	4.26	70.17	4.33	4.65	8.82	174.93	5.01	4.54	4.30
		P-eLasso	**35.82**	**1.70**	**5.00**	**0.76**	**81.91**	**1.63**	**5.00**	**0.83**	**180.13**	**1.61**	**5.00**	**0.73**

Note: Bold symbols in the C and CG indicate the highest values, while the lowest values for IC and MSE are also bolded. The optimal values of C vary by scenario and are determined by the (p - number of non-zeros). Thus, for p = 50, C = (45, 40, 42); for p = 100, C = (95, 90, 92); for p = 200, C = (195, 190 192). The optimal value for IC is zero.

Table A.2: C (the number of zero coefficients that are correctly estimated by zero), IC (the number of non-zero coefficients that are incorrectly estimated by zero), GC (the number of groups that are correctly selected), and MSE (Mean Squared Error) for PWGEE-Lasso(P-Lasso), PWGEE-SCAD (P-SCAD), PWGEE-MCP (P-MCP), PWGEE-cMCP (P-cMCP), and PWGEE-Exclusive lasso (P-eLasso) method with n = 300 and p =(50, 100, 200), which has unequal group sizes, in each scenario.

scenarios	corr	methods	C	IC	CG	MSE	C	IC	CG	MSE	C	IC	CG	MSE
				p=50				p=100				p=200		
1	0.60	P-Lasso	28.19	**0.00**	**5.00**	0.11	66.53	**0.00**	**5.00**	0.10	144.15	**0.00**	**5.00**	0.11
		P-SCAD	31.12	**0.00**	**5.00**	0.17	68.82	**0.00**	**5.00**	0.20	144.29	**0.00**	**5.00**	0.31
		P-MCP	**34.47**	0.01	**5.00**	0.18	76.21	**0.00**	**5.00**	0.22	159.51	**0.00**	**5.00**	0.36
		P-cMCP	30.90	0.01	**5.00**	0.14	69.71	**0.00**	**5.00**	0.13	149.11	**0.00**	**5.00**	0.15
		P-eLasso	32.97	**0.00**	**5.00**	**0.09**	**81.16**	**0.00**	**5.00**	**0.07**	**180.88**	**0.00**	**5.00**	**0.05**
	0.90	P-Lasso	30.64	0.30	4.95	0.74	74.46	0.40	4.99	0.51	164.83	0.19	**5.00**	0.38
		P-SCAD	33.16	1.39	4.73	1.90	75.18	1.50	4.85	2.06	161.50	1.44	4.97	2.52
		P-MCP	**33.90**	1.34	4.72	1.85	77.21	1.44	4.85	2.01	166.12	1.49	4.94	2.54
		P-cMCP	32.96	1.35	4.49	1.80	74.68	1.45	4.77	1.92	159.03	1.37	4.95	2.37
		P-eLasso	32.03	**0.10**	**5.00**	**0.61**	**79.40**	**0.13**	**5.00**	**0.39**	**179.40**	**0.05**	**5.00**	**0.19**
2	0.60	P-Lasso	24.70	0.11	**5.00**	0.22	62.05	**0.03**	**5.00**	0.18	140.40	**0.00**	**5.00**	0.20
		P-SCAD	27.41	0.46	**5.00**	0.35	63.46	0.20	**5.00**	0.37	140.16	0.19	**5.00**	0.52
		P-MCP	**30.29**	0.60	4.99	0.35	71.65	0.32	**5.00**	0.38	155.53	0.36	**5.00**	0.56
		P-cMCP	26.77	0.53	4.99	0.34	65.10	0.24	**5.00**	0.31	144.95	0.20	**5.00**	0.36
		P-eLasso	27.31	**0.06**	**5.00**	**0.20**	**72.82**	**0.03**	**5.00**	**0.15**	**171.56**	**0.00**	**5.00**	**0.13**
	0.90	P-Lasso	27.04	2.61	4.97	**1.59**	71.11	2.32	4.97	1.13	160.49	1.95	4.99	0.91
		P-SCAD	29.00	4.89	4.81	3.91	71.63	5.07	4.88	3.96	156.18	5.18	4.99	5.11
		P-MCP	**29.58**	4.86	4.74	3.77	73.27	5.19	4.84	3.97	162.14	5.33	4.97	4.76
		P-cMCP	28.56	5.00	4.34	3.82	71.99	5.24	4.70	3.90	155.98	5.17	4.97	5.24
		P-eLasso	26.50	**1.99**	**5.00**	1.61	**73.73**	**1.77**	**5.00**	**0.95**	**170.52**	**1.61**	**5.00**	**0.71**
3	0.60	P-Lasso	25.19	**0.00**	**5.00**	**0.16**	62.81	**0.00**	**5.00**	**0.14**	140.98	**0.01**	**5.00**	0.16
		P-SCAD	28.12	0.08	4.99	0.26	64.97	0.03	**5.00**	0.28	141.95	0.02	**5.00**	0.40
		P-MCP	**31.14**	0.13	4.99	0.27	72.71	0.06	**5.00**	0.30	156.99	0.05	**5.00**	0.45
		P-cMCP	27.83	0.17	4.97	0.23	66.15	0.04	**5.00**	0.21	145.10	0.02	**5.00**	0.24
		P-eLasso	26.02	0.01	**5.00**	0.17	**73.44**	0.02	**5.00**	**0.14**	**170.02**	0.04	**5.00**	**0.13**
	0.90	P-Lasso	28.67	1.63	4.90	1.26	70.95	1.41	4.96	1.01	161.46	**1.16**	4.98	0.69
		P-SCAD	30.86	3.54	4.71	2.99	72.73	3.58	4.91	3.41	158.21	3.61	4.95	4.03
		P-MCP	**31.14**	3.54	4.68	3.00	**74.39**	3.58	4.84	3.34	162.23	3.65	4.92	4.01
		P-cMCP	30.39	3.46	4.36	3.04	73.24	3.55	4.64	3.22	156.77	3.51	4.95	4.13
		P-eLasso	28.02	**1.34**	**5.00**	**1.25**	72.77	**1.12**	**5.00**	**0.90**	**172.76**	1.19	**5.00**	**0.56**

Note: Bold symbols in the C and CG indicate the highest values, while the lowest values for IC and MSE are also bolded. The optimal values of C vary by scenario and are determined by the (p - number of non-zeros). Thus, for p = 50, C = (45, 40, 42); for p = 100, C = (95, 90, 92); for p = 200, C = (195, 190 192). The optimal value for IC is zero.

Table A.3: C (the number of zero coefficients that are correctly estimated by zero), IC (the number of non-zero coefficients that are incorrectly estimated by zero), GC (the number of groups that are correctly selected), and MSE (Mean Squared Error) for PWGEE-Lasso(P-Lasso), PWGEE-SCAD (P-SCAD), PWGEE-MCP (P-MCP), PWGEE-cMCP (P-cMCP), and PWGEE-Exclusive lasso (P-eLasso) method with n = 300 and p =(150, 300, 600), which has unequal group sizes, in each scenario.

scenarios	corr	methods	C	IC	CG	MSE	C	IC	CG	MSE	C	IC	CG	MSE
				p=150				p=300				p=600		
1	0.60	P-Lasso	104.82	**0.00**	**5.00**	0.10	225.28	**0.00**	**5.00**	0.13	527.72	**0.00**	**5.00**	0.09
		P-SCAD	106.43	**0.00**	**5.00**	0.24	220.40	**0.00**	**5.00**	0.47	529.51	**0.00**	**5.00**	0.13
		P-MCP	117.49	**0.00**	**5.00**	0.29	243.35	**0.00**	**5.00**	0.52	546.84	**0.00**	**5.00**	0.24
		P-cMCP	109.03	**0.00**	**5.00**	0.14	230.52	**0.00**	**5.00**	0.18	533.31	**0.00**	**5.00**	0.14
		P-eLasso	**131.01**	**0.00**	**5.00**	**0.06**	**281.95**	**0.00**	**5.00**	**0.05**	**581.31**	**0.00**	**5.00**	**0.05**
	0.90	P-Lasso	119.53	0.21	**5.00**	0.40	256.14	0.23	**5.00**	0.37	566.13	0.26	**5.00**	0.30
		P-SCAD	117.55	1.40	4.94	2.42	244.71	1.47	**5.00**	3.46	567.15	2.64	4.93	1.20
		P-MCP	121.88	1.45	4.92	2.24	254.46	1.59	**5.00**	3.11	576.47	2.65	4.85	1.24
		P-cMCP	115.52	1.37	4.90	2.23	239.99	1.38	**5.00**	3.40	572.48	2.63	4.86	1.14
		P-eLasso	**130.71**	**0.03**	**5.00**	**0.21**	**279.57**	**0.06**	**5.00**	**0.16**	**578.28**	**0.05**	**5.00**	**0.13**
2	0.60	P-Lasso	101.72	**0.02**	**5.00**	0.19	219.28	**0.01**	**5.00**	0.21	511.11	**0.03**	**5.00**	0.19
		P-SCAD	103.05	0.26	**5.00**	0.41	217.50	0.23	**5.00**	0.75	515.25	0.72	**5.00**	0.42
		P-MCP	114.01	0.40	**5.0**	0.45	238.21	0.34	**5.00**	0.82	535.86	0.96	**5.00**	0.64
		P-cMCP	105.50	0.29	**5.00**	0.33	223.74	0.25	**5.00**	0.41	519.17	0.61	**5.00**	0.44
		P-eLasso	**123.51**	**0.02**	**5.00**	**0.14**	**271.94**	0.02	**5.00**	**0.12**	**571.16**	0.08	**5.00**	**0.13**
	0.90	P-Lasso	115.91	2.17	**5.00**	1.06	254.32	1.90	**5.00**	0.85	563.01	1.85	**5.00**	0.65
		P-SCAD	112.78	5.11	4.96	4.89	242.05	5.22	4.99	6.53	565.43	6.58	4.90	2.38
		P-MCP	116.81	5.34	4.91	4.61	252.21	5.40	4.96	5.85	**573.55**	6.70	4.82	2.32
		P-cMCP	112.77	5.17	4.83	4.88	237.49	5.13	4.93	7.53	571.59	6.62	4.80	2.18
		P-eLasso	**122.61**	**1.63**	**5.00**	**0.82**	**271.01**	**1.56**	**5.00**	**0.58**	572.88	**1.41**	**5.00**	**0.49**
3	0.60	P-Lasso	103.10	**0.02**	**5.00**	0.15	218.47	**0.00**	**5.00**	0.17	517.67	**0.00**	**5.00**	0.14
		P-SCAD	104.65	0.03	**5.00**	0.29	218.41	0.03	**5.00**	0.59	520.96	0.09	**5.00**	0.21
		P-MCP	115.65	0.05	**5.00**	0.33	238.82	0.05	**5.00**	0.66	540.71	0.15	**5.00**	0.36
		P-cMCP	107.15	0.04	**5.00**	0.21	224.47	**0.00**	**5.00**	0.28	524.50	0.07	**5.00**	0.25
		P-eLasso	**120.93**	**0.02**	**5.00**	**0.13**	**269.43**	**0.00**	**5.00**	**0.13**	**566.51**	0.02	**5.00**	**0.12**
	0.90	P-Lasso	117.50	1.39	4.99	0.78	253.25	1.08	**5.00**	0.66	563.48	1.15	4.99	0.55
		P-SCAD	115.97	3.52	4.98	3.47	240.67	3.43	**5.00**	5.68	566.05	4.90	4.88	1.91
		P-MCP	118.71	3.57	4.94	3.46	249.84	3.64	4.98	5.08	**574.52**	4.95	4.83	1.88
		P-cMCP	116.94	3.60	4.84	3.25	237.30	3.39	4.96	6.06	571.69	4.98	4.78	1.81
		P-eLasso	**124.47**	**1.28**	**5.00**	**0.65**	**272.40**	**1.05**	**5.00**	**0.50**	572.04	**0.99**	**5.00**	**0.43**

Note: Bold symbols in the C and CG indicate the highest values, while the lowest values for IC and MSE are also bolded. The optimal values of C vary by scenario and are determined by the (p - number of non-zeros). Thus, for p = 150, C = (145, 140, 142); for p = 300, C = (295, 290, 292); for p = 600, C = (595, 590 592). The optimal value for IC is zero.

Table A.4: SDOH Domains and Topic Areas Represented in the SDOH Database

IND	Domain	Variable Name	Variable Label	Data Source
0	Identifier	YEAR	SDOH file year	
1	Social context	ACS_HH_SIZE	Average Household size	ACS
2	Social context	ACS_MEDIAN_AGE	Median age	ACS
3	Social context	ACS_PCT_1UP_PERS_1ROOM	% Housing units with more than one occupant per room	ACS
4	Social context	ACS_PCT_AIAN	% Population reporting American Indian/Alaska Native race	ACS
5	Social context	ACS_PCT_ASIAN	% Population reporting Asian race	ACS
6	Social context	ACS_PCT_BLACK	% Population reporting Black race	ACS
7	Social context	ACS_PCT_CHILD_1FAM	% Families with children that are single-parent families	ACS
8	Social context	ACS_PCT_CTZ_US_BORN	% Population consisting of U.S. citizens born in United States, Puerto Rico, or U.S. Islands	ACS
9	Social context	ACS_PCT_DIVORCE_SEPARAT	% Population divorced or separated (ages 15 and over)	ACS
10	Social context	ACS_PCT_ENGL_NOT_ALL	% Population that does not speak English at all (ages 5 and over)	ACS
11	Social context	ACS_PCT_FEMALE	% Population that is female	ACS
12	Social context	ACS_PCT_FOREIGN_BORN	% Population that is foreign-born	ACS
13	Social context	ACS_PCT_GRANDKID_TOT	% Children living with a grandparent Householder (ages 17 and under)	ACS
14	Social context	ACS_PCT_HH_1PERS	% Households with only one occupant	ACS
15	Social context	ACS_PCT_HH_NO_FUEL	% Occupied Housing units without fuel	ACS
16	Social context	ACS_PCT_HISPAN	% Population reporting Hispanic ethnicity	ACS
17	Social context	ACS_PCT_MULT_RACE	% Population reporting multiple races	ACS
18	Social context	ACS_PCT_NHPI	% Population reporting Native Hawaiian/Pacific Islander race	ACS
19	Social context	ACS_PCT_NONCTN_1990	% Population who are not U.S. citizens and entered U.S. before 1990	ACS
20	Social context	ACS_PCT_OTH_LANG	% Population who speak other languages (ages 5 and over)	ACS
21	Social context	ACS_PCT_VA	% Civilian Population consisting of veterans (ages 18 and over)	ACS
22	Social context	ACS_PCT_WHITE	% Population reporting White race	ACS
23	Economic context	ACS_GINI_INDEX	Gini index of income inequality	ACS
24	Economic context	ACS_MEDIAN_HH_INCOME_N	Median Household income (in dollars, inflation-adjusted to file data year) /1000	ACS
25	Economic context	ACS_PCT_1FAM_HH_FOOD_STMP	% Unmarried partner Households that received food stamps/ Supplemental Nutrition Assistance Program (SNAP) benefits	ACS
26	Economic context	ACS_PCT_ADMIN	% Employed working in public administration	ACS
27	Economic context	ACS_PCT_ARMED_FORCES	% Civilian Population in armed forces (ages 16 years and over)	ACS
28	Economic context	ACS_PCT_ART	% Employed working in arts, entertainment, recreation, and accommodation and food services	ACS
29	Economic context	ACS_PCT_CONSTRUCT	% Employed working in construction	ACS
30	Economic context	ACS_PCT_EDUC	% Employed working in educational services, and healthcare and social assistance	ACS
31	Economic context	ACS_PCT_FINANCE	% Employed working in finance and insurance, real estate, and rental and leasing	ACS
32	Economic context	ACS_PCT_FOOD_STAMP	% Households that received food stamps/SNAP, past 12 months	ACS
33	Economic context	ACS_PCT_HH_PUB_ASSIST	% Households with public assistance income or food stamps/SNAP	ACS
34	Economic context	ACS_PCT_INFORM	% Employed working in information services	ACS
35	Economic context	ACS_PCT_MANUFACT	% Employed working in manufacturing	ACS
36	Economic context	ACS_PCT_NATURE	% Employed working in agriculture, forestry, fishing and hunting, and mining (ages 16 and over)	ACS
37	Economic context	ACS_PCT_OTHER	% Employed working in other services, except public administration	ACS
38	Economic context	ACS_PCT_PERSON_INC124	% Population with income to poverty ratio: 1.00-1.24	ACS
39	Economic context	ACS_PCT_PERSON_INC199	% Population with income to poverty ratio: 1.25-1.99	ACS
40	Economic context	ACS_PCT_PERSON_INC200	% Population with income to poverty ratio: 2.00 or higher	ACS
41	Economic context	ACS_PCT_PERSON_INC99	% Population with income to poverty ratio: <1.00	ACS
42	Economic context	ACS_PCT_PROFESS	% Employed working in professional, scientific, management, administrative, and waste management services	ACS
43	Economic context	ACS_PCT_RETAIL	% Employed working in retail trade	ACS
44	Economic context	ACS_PCT_TRANSPORT	% Employed working in transportation and warehousing, and in utilities	ACS
45	Economic context	ACS_PCT_UNEMPLOY	% Population that was unemployed (ages 16 years and over)	ACS
46	Economic context	ACS_PCT_VA_LABOR_FORCE	% Civilian veterans in labor force (ages 18–64)	ACS
47	Economic context	ACS_PCT_WHOLESALE	% Employed working in wholesale trade	ACS
48	Education	ACS_PCT_ASSOCIATE_DGR	% Population with some college or associate's degree (ages 25 and over)	ACS
49	Education	ACS_PCT_BACHELOR_DGR	% Population with a bachelor's degree (ages 25 and over)	ACS
50	Education	ACS_PCT_GRADUATE_DGR	% Population with a master's or professional school degree or doctorate (ages 25 and over)	ACS
51	Education	ACS_PCT_HS_GRADUATE	% Population with only high school diploma (ages 25 and over)	ACS
52	Education	ACS_PCT_LT_HS	% Population with less than high school education (ages 25 and over)	ACS
53	Physical infrastructure	ACS_MEDIAN_HOME_VALUE	Median home value of owner-occupied Housing units	ACS
54	Physical infrastructure	ACS_MEDIAN_RENT	Median gross rent as a % Household income	ACS
55	Physical infrastructure	ACS_PCT_10UNITS	% Housing in structures with 10 or more units	ACS
56	Physical infrastructure	ACS_PCT_COMMT_60MINUP	% Workers with at least 60-minute commute time (ages 16 and over)	ACS
57	Physical infrastructure	ACS_PCT_DRIVE_2WORK	% Workers taking a car, truck, or van to Work (ages 16 and over)	ACS
58	Physical infrastructure	ACS_PCT_MOBILE_HOME	% Housing units that are mobile homes	ACS
59	Physical infrastructure	ACS_PCT_NO_VEH	% Housing units with no vehicle available	ACS
60	Physical infrastructure	ACS_PCT_PUBL_TRANSIT	% Workers taking public transportation, excluding taxicab (ages 16 and over)	ACS
61	Physical infrastructure	ACS_PCT_RENT_COST_30PCT	% Rental units with rent equal to 30 percent or more of Household income	ACS
62	Physical infrastructure	ACS_PCT_RENTED_HH	% Occupied Housing units: rented	ACS
63	Physical infrastructure	ACS_PCT_RENTER_HH_CHILD	% Renter-occupied Housing units with children	ACS
64	Physical infrastructure	ACS_PCT_VACANT_HH	% Housing units vacant	ACS
65	Physical infrastructure	CCBP_RATE_CS_PER_1000	Convenience stores per 1,000 people	CCBP
66	Physical infrastructure	CCBP_RATE_FSR_PER_1000	Full service restaurants per 1,000 people	CCBP
67	Physical infrastructure	CCBP_RATE_SOGS_PER_1000	Supermarkets and other grocery (except convenience) stores per 1,000 people	CCBP
68	Healthcare context	AHRF_FED_HLTH_CNT	Number of Federally Qualified Health Centers	AHRF
69	Healthcare context	AHRF_MENTL_HLTH_CNT	Total number of community mental health centers	AHRF
70	Healthcare context	AHRF_RURAL_H_CLINIC	Number of rural health clinics	AHRF
71	Healthcare context	MP_ELIGIBLES	Number of Medicare eligibles in the county	MP

Continued on next page

Table A.4 – continued from previous page

IND	Domain	Variable Name	Variable Label	Data Source
72	Healthcare context	MP_PERCPEN	Derived field that equals the ratio of enrollees over eligibles * 100	MP
73	Geography	UR2013	1: Rural, 0: Urban by NCHS 2013 Rural-Urban Classification Scheme	NCHS

Table A.5: Missing Variables by Years

Domain	Variables	2009	2010	2011	2012	2013	2014	2015	2016	2017	2018
2	Gini index of income inequality	.	.	.	.	.	.	.	.	.	1
2	Median household income (in dollars)/1000	.	.	.	.	.	.	.	.	.	1
2	% Unmarried partner households that received food stamps/SNAP	.	.	.	.	.	.	.	.	.	1
2	% Employed working in public administration	.	.	.	.	.	.	.	.	.	1
2	% Civilian population in armed forces (≥ 16yr)	.	.	.	.	.	.	.	.	.	1
2	% Employed working in arts, etc.	.	.	.	.	.	.	.	.	.	1
2	% Employed working in construction	.	.	.	.	.	.	.	.	.	1
2	% Employed working in educational services, etc.	.	.	.	.	.	.	.	.	.	1
2	% Employed working in finance and insurance, etc.	.	.	.	.	.	.	.	.	.	1
2	% Households that received food stamps/SNAP, past 12 months	.	.	.	.	.	.	.	.	.	1
2	% Households with public assistance income or food stamps/SNAP	.	.	.	.	.	.	.	.	.	1
2	% Employed working in information services	.	.	.	.	.	.	.	.	.	1
2	% Employed working in manufacturing	.	.	.	.	.	.	.	.	.	1
2	% Employed working in agriculture, forestry, etc.(≥ 16yr)	.	.	.	.	.	.	.	.	.	1
2	% Employed working in other services, except public administration	.	.	.	.	.	.	.	.	.	1
2	% Population with income to poverty ratio: 1.00-1.24	.	.	.	.	.	.	.	.	.	1
2	% Population with income to poverty ratio: 1.25-1.99	.	.	.	.	.	.	.	.	.	1
2	% Population with income to poverty ratio: ≥ 2.00	.	.	.	.	.	.	.	.	.	1
2	% Population with income to poverty ratio: <1.00	.	.	.	.	.	.	.	.	.	1
2	% Employed working in professional, scientific, etc.	.	.	.	.	.	.	.	.	.	1
2	% Employed working in retail trade	.	.	.	.	.	.	.	.	.	1
2	% Employed working in transportation, etc.	.	.	.	.	.	.	.	.	.	1
2	% Population that was unemployed (≥ 16yr)	.	.	.	.	.	.	.	.	.	1
2	% Civilian veterans in labor force (18–64yr)	.	.	.	.	.	.	.	.	.	1
2	% Employed working in wholesale trade	.	.	.	.	.	.	.	.	.	1
4	Median gross rent as a percentage of household income	.	.	.	.	.	.	.	.	.	1
4	% Workers with at least 60-minute commute time (≥ 16yr)	.	.	.	.	.	.	.	.	.	1
4	% Workers taking a car, truck, or van to work (≥ 16yr)	.	.	.	.	.	.	.	.	.	1
4	% Workers taking public transportation, excluding taxicab (≥ 16yr)	.	.	.	.	.	.	.	.	.	1
4	% Occupied housing units: rented	.	.	.	.	.	.	.	.	.	1
4	Convenience stores per 1,000 people	.	2	5	13	12	10	10	9	157	159
4	Full service restaurants per 1,000 people	.	5	13	12	16	14	17	19	174	173
4	Supermarkets and other grocery (except convenience) stores/1,000 people	.	5	8	16	21	19	24	29	684	714
5	Number of Medicare eligibles in the county	.	1	.	.	.	.	.	1	1	1
5	Derived field that equals the ratio of enrollees over eligibles * 100	.	7	8	6	17	46	49	46	47	41

Table A.6: Excluded Counties due to Missing at the Baseline (in 2009)

GEOID	Counties	States	GEOID	Counties	States	GEOID	Counties	States	GEOID	Counties	States	GEOID	Counties	States
1037	Coosa	Alabama	13007	Baker	Georgia	20187	Stanton	Kansas	31165	Sioux	Nebraska	48011	Armstrong	Texas
1085	Lowndes	Alabama	13061	Clay	Georgia	20195	Trego	Kansas	31171	Thomas	Nebraska	48033	Borden	Texas
2013	Maricopa	Arizona	13101	Echols	Georgia	20203	Wichita	Kansas	31183	Wheeler	Nebraska	48101	Cottle	Texas
2016	Navajo	Arizona	13239	Quitman	Georgia	21091	Hancock	Kentucky	32009	Esmeralda	Nevada	48111	Dallam	Texas
2050	Santa Cruz	Arizona	13259	Stevens	Georgia	21165	Menifee	Kentucky	32029	Storey	Nevada	48119	Delta	Texas
2060	Yuma	Arizona	13265	Taliaferro	Georgia	21201	Robertson	Kentucky	35021	Harding	New Mexico	48155	Foard	Texas
2068	Clark	Arkansas	13283	Washington	Georgia	22023	Cameron Parish	Louisiana	38007	Billings	North Dakota	48173	Glasscock	Texas
2070	Cleburne	Arkansas	13307	Webster	Georgia	22043	Grant Parish	Louisiana	38013	Burke	North Dakota	48205	Hartley	Texas
2100	Nevada	Arkansas	15005	Kalawao	Hawaii	22081	Red River Parish	Louisiana	38023	Divide	North Dakota	48261	Kenedy	Texas
2105	Ouachita	Arkansas	16025	Camas	Idaho	27167	Wilkin	Minnesota	38039	Griggs	North Dakota	48263	Kent	Texas
2130	Saline	Arkansas	16033	Clark	Idaho	28005	Amite	Mississippi	38043	Kidder	North Dakota	48269	King	Texas
2150	Van Buren	Arkansas	17059	Gallatin	Illinois	28055	Issaquena	Mississippi	38065	Oliver	North Dakota	48301	Loving	Texas
2158	White	Arkansas	20023	Cheyenne	Kansas	28063	Jefferson Davis	Mississippi	38085	Sioux	North Dakota	48311	McMullen	Texas
2164	Woodruff	Arkansas	20025	Clark	Kansas	28103	Noxubee	Mississippi	38087	Slope	North Dakota	48421	Sherman	Texas
2180	Yell	Arkansas	20033	Comanche	Kansas	28125	Sharkey	Mississippi	38095	Towner	North Dakota	48431	Sterling	Texas
2185	Calhoun	Arkansas	20039	Decatur	Kansas	28157	Wilkinson	Mississippi	39163	Vinton	Ohio	48433	Stonewall	Texas
2188	Carroll	Arkansas	20047	Edwards	Kansas	30011	Carter	Montana	40059	Harper	Oklahoma	48435	Sutton	Texas
2195	Clark	Arkansas	20049	Ellis	Kansas	30037	Golden Valley	Montana	40129	Roger Mills	Oklahoma	48443	Terrell	Texas
2198	Clay	Arkansas	20053	Ellsworth	Kansas	30045	Judith Basin	Montana	41049	Morrow	Oregon	49009	Daggett	Utah
2220	Conway	Arkansas	20063	Gove	Kansas	30051	Liberty	Montana	41069	Wheeler	Oregon	51021	Bland	Virginia
2230	Crawford	Arkansas	20067	Grant	Kansas	30069	Petroleum	Montana	45005	Allendale	South Carolina	51081	Greensville	Virginia
2240	Crittenden	Arkansas	20071	Greeley	Kansas	30101	Toole	Montana	46017	Buffalo	South Dakota	51097	King and Queen	Virginia
2261	Cross	Arkansas	20075	Hamilton	Kansas	30103	Treasure	Montana	46021	Campbell	South Dakota	51530	Buena Vista city	Virginia
2275	Dallas	Arkansas	20081	Haskell	Kansas	31005	Arthur	Nebraska	46063	Harding	South Dakota	53023	Garfield	Washington
2282	Desha	Arkansas	20083	Hodgeman	Kansas	31007	Banner	Nebraska	46069	Hyde	South Dakota	55078	Menominee	Wisconsin
2290	Drew	Arkansas	20093	Kearny	Kansas	31009	Blaine	Nebraska	46095	Mellette	South Dakota			
5025	Cleburne	Arkansas	20097	Kiowa	Kansas	31057	Dundy	Nebraska	46121	Potter	South Dakota			
6003	Alpine	California	20101	Lane	Kansas	31085	Hayes	Nebraska	46137	Ziebach	South Dakota			
8033	Dolores	Colorado	20119	Meade	Kansas	31091	Hooker	Nebraska	47033	Crockett	Tennessee			
8053	Hinsdale	Colorado	20129	Morton	Kansas	31103	Keya Paha	Nebraska	47067	Hancock	Tennessee			
8079	Mineral	Colorado	20135	Ness	Kansas	31113	Logan	Nebraska	47127	Moore	Tennessee			
8111	San Juan	Colorado	20163	Rooks	Kansas	31115	Loup	Nebraska	47135	Perry	Tennessee			
8115	Sedgwick	Colorado	20165	Rush	Kansas	31117	McPherson	Nebraska	47153	Sequatchie	Tennessee			
11001	District of Columbia		20171	Scott	Kansas	31133	Pawnee	Nebraska	47175	Van Buren	Tennessee			
12125	Union	Florida	20181	Sherman	Kansas	31149	Rock	Nebraska	48009	Archer	Texas			

Table A.7: Estimates of Selected Social Determinants of Healths Associated with Age-Adjusted CVD Mortality : PWGEE-Exclusive Lasso (P-eLasso), PWGEE-Lasso (P-Lasso), PWGEE-SCAD (P-SCAD), PWGEE-MCP (P-MCP), PWGEE-cMCP (P-cMCP), 2009-2018

Domain	SDOH	P-eLasso	P-Lasso	P-SCAD	P-MCP	P-cMCP
	Intercept	272.1746	251.7660	262.0719	216.0003	237.9212
1. Social context	Median age	.	.	-0.2923	-0.5307	-0.8611
	% population reporting Black race	0.3114	0.3464	.	0.3182	0.2741
	% population divorced or separated (ages $\geq$ 15)	1.1313	1.9075	2.9315	2.6944	2.6979
	% population that does not speak English at all (ages $\geq$ 5)	-0.2687	-1.6401	.	-3.6307	-4.0254
	% population that is female	.	.	0.4774	1.0845	1.3184
	% children living with a grandparent householder (ages $\leq$ 17)	0.9175	1.3664	1.9007	1.6355	1.8566
	% occupied housing units without fuel	.	.	-0.3063	-0.9674	-1.3354
	% population reporting Hispanic ethnicity	-0.3351	-0.4766	-1.0532	-0.7847	-0.7421
	% population reporting multiple races	.	.	1.6908	3.3803	3.6588
	% population who are not U.S. citizens and entered U.S. before 1990	.	-0.1970	.	.	.
	% population who speak other languages (ages $\geq$ 5)	-0.2242	-1.0254	-3.4467	-3.1026	-3.0637
	% the civilian population consisting of veterans (ages $\geq$18)	.	-1.0903	-3.6232	-3.5152	-3.5717
	% population reporting White race	.	.	-0.2410	.	.
2. Economic context	Gini index of income inequality	40.7584	35.6593	159.6749	191.8995	176.2410
	% civilian population in armed forces (ages $\geq$ 16)	.	.	2.3265	2.4682	2.7177
	% households that received food stamps/SNAP, past 12 months	0.5939	0.0200	.	.	.
	% employed working in manufacturing	0.0400	.	.	.	.
	% population with income to poverty ratio: 2.00 or higher	-0.1903	-0.0083	.	.	.
	% population with income to poverty ratio: <1.00	0.6141	0.7554	.	.	.
	% employed working in transportation and warehousing, and in utilities	0.3785	0.2748	0.7023	1.1556	0.4917
	% population that was unemployed (ages $\geq$ 16)	.	.	.	-0.3707	-0.0635
3. Education	% population with some college or associate's degree (ages $\geq$ 25)	-0.6896	-0.4342	-0.8798	-1.1510	-1.3114
	% population with a bachelor's degree (ages $\geq$ 25)	-2.0069	-2.3678	-3.1414	-3.3864	-2.8761
	% population with a master's or professional school degree or doctorate (ages $\geq$25)	.	.	.	.	-1.7095
	% population with only high school diploma (ages $\geq$ 25)	.	.	.	.	-0.0004
	% population with less than high school education (ages $\geq$ 25)	.	0.5375	.	.	.
4. Physical infrastructure	Median home value of owner-occupied housing units	-0.0001	-0.0001	-0.0001	-0.0001	-0.0001
	Median gross rent as a % household income	.	.	-0.0157	-0.3606	.
	% workers taking a car, truck, or van to work (ages $\geq$ 16)	0.0447	.	.	.	.
	% housing units that are mobile homes	0.0191	.	.	.	.
	% renter-occupied housing units with children	0.1157	0.1101	0.4305	0.5726	0.1495
	Full service restaurants per 1,000 people	-5.4385	-1.4014	.	.	.
5. Healthcare	Derived field that equals the ratio of enrollees over eligibles * 100	-0.3940	-0.2462	-0.5599	-0.5317	-0.5486
Geography	1: Rural, 0: Urban by NCHS 2013 Rural-Urban Classification Scheme	3.1804	.	.	.	.

Appendix B

Model-based Clustering of High-Dimensional Longitudinal Data via Exclusive Lasso Penalty

Table B.1: Comparison results averaged by ARI, C, IC, and MSE among the proposed methods(mixLMM) through Exclusive Lasso (mixLMM-eLasso), SCAD (mixLMM-SCAD), and Lasso (mixLMM-Lasso) using datasets for n=100, p=(25, 50 100) in unequal-sized group.

Scn	method	Corr	ARI	C	IC	MSE	ARI	C	IC	MSE	ARI	C	IC	MSE
			p=25				p=50				p=100			
1	eLasso	0.6	**0.937**	38.00	**0.00**	**0.011**	**0.939**	**88.40**	**0.00**	**0.006**	**0.942**	187.72	**0.00**	**0.003**
	SCAD		0.920	**38.03**	0.71	0.032	0.853	87.64	3.33	0.041	0.670	187.34	4.50	0.027
	Lasso		0.925	37.32	1.10	0.036	0.879	87.58	3.70	0.043	0.755	**188.60**	5.49	0.030
	eLasso	0.9	**0.946**	37.91	**0.00**	**0.016**	**0.951**	**88.02**	**0.00**	**0.007**	**0.949**	**187.76**	**0.00**	**0.004**
	SCAD		0.923	**38.29**	3.42	0.146	0.912	87.43	4.07	0.084	0.891	187.56	5.02	0.050
	Lasso		0.939	38.04	0.34	0.052	0.926	87.59	0.92	0.030	0.910	187.31	1.94	0.019
2	eLasso	0.6	**0.948**	**26.54**	**0.00**	**0.038**	**0.947**	76.70	**0.11**	**0.021**	**0.946**	175.50	**0.00**	**0.010**
	SCAD		0.917	26.29	3.52	0.135	0.865	74.55	6.56	0.106	0.752	173.46	9.95	0.073
	Lasso		0.922	26.23	3.02	0.104	0.889	**77.07**	6.15	0.082	0.757	**177.53**	11.76	0.066
	eLasso	0.9	**0.961**	27.53	**1.12**	**0.136**	**0.962**	77.41	**1.08**	**0.066**	**0.953**	**177.50**	**1.01**	**0.033**
	SCAD		0.939	**28.52**	11.22	0.584	0.931	76.63	11.24	0.310	0.904	174.81	12.50	0.171
	Lasso		0.958	27.82	2.75	0.214	0.955	**77.83**	3.63	0.113	0.931	176.79	4.74	0.059
3	eLasso	0.6	**0.946**	24.64	**1.11**	**0.035**	**0.946**	74.74	**1.16**	**0.021**	**0.925**	173.41	**1.32**	**0.012**
	SCAD		0.911	**28.88**	3.43	0.071	0.901	77.39	5.07	0.054	0.655	177.55	8.82	0.051
	Lasso		0.929	28.10	3.20	0.068	0.911	**79.26**	6.22	0.064	0.766	**180.21**	9.24	0.048
	eLasso	0.9	**0.959**	27.28	**1.51**	**0.110**	**0.957**	77.36	**1.33**	**0.055**	**0.957**	177.67	**1.30**	**0.027**
	SCAD		0.928	**31.06**	8.42	0.349	0.926	**79.70**	9.17	0.200	0.909	178.31	9.98	0.117
	Lasso		0.949	28.96	2.62	0.140	0.942	79.24	3.70	0.080	0.933	**178.44**	4.25	0.040

Note: Bold symbols in the ARI and C indicate the highest values, while the lowest values for IC and MSE are also bolded. The optimal values of C vary by scenario and are determined by the 2×(p - number of non-zeros). Thus, for p = 25, C = (45, 40, 42); for p = 50, C = (95, 90, 92); for p = 100, C = (195, 190 192). The optimal value for IC is zero; The method denotes mixLMM with tested penalties.

Table B.2: Comparison results averaged by ARI, C, IC, and MSE among the proposed methods(mixLMM) through Exclusive Lasso (mixLMM-eLasso), SCAD (mixLMM-SCAD), and Lasso (mixLMM-Lasso) using datasets for n=200, p=(25, 50 100) in unequal-sized group.

Scn	method	Corr	ARI	C	IC	MSE	ARI	C	IC	MSE	ARI	C	IC	MSE
			p=25				p=50				p=100			
1	eLasso	0.6	**0.937**	38.72	**0.00**	**0.005**	**0.940**	**88.46**	**0.00**	**0.003**	**0.938**	188.39	**0.00**	**0.001**
	SCAD		0.923	**38.99**	0.40	0.013	0.913	87.90	1.00	0.013	0.832	**188.47**	3.22	0.017
	Lasso		0.925	37.55	0.50	0.016	0.915	86.18	1.40	0.017	0.869	185.81	3.29	0.018
	eLasso	0.9	**0.947**	38.34	**0.00**	**0.008**	**0.952**	87.91	**0.00**	**0.004**	**0.945**	**187.99**	**0.00**	**0.002**
	SCAD		0.920	**38.68**	1.68	0.068	0.932	**88.11**	2.37	0.047	0.917	187.47	3.07	0.029
	Lasso		0.945	38.14	0.02	0.024	0.947	87.25	0.02	0.013	0.934	186.80	0.41	0.008
2	eLasso	0.6	**0.955**	27.57	**0.00**	**0.018**	**0.947**	**77.28**	**0.00**	**0.010**	**0.955**	**176.96**	**0.00**	**0.005**
	SCAD		0.949	**28.25**	0.56	0.033	0.926	76.15	2.61	0.040	0.916	174.34	4.36	0.031
	Lasso		0.951	27.20	0.40	0.029	0.937	76.15	1.99	0.031	0.923	176.30	5.60	0.033
	eLasso	0.9	**0.961**	28.05	**0.20**	**0.073**	**0.957**	**77.64**	**0.23**	**0.037**	**0.965**	**177.81**	**0.17**	**0.018**
	SCAD		0.942	**28.91**	8.83	0.377	0.935	77.28	9.14	0.209	0.938	176.10	10.39	0.122
	Lasso		0.959	28.20	0.85	0.114	0.954	77.26	1.01	0.063	0.959	177.14	0.95	0.032
3	eLasso	0.6	**0.947**	24.96	**1.02**	**0.017**	**0.943**	73.89	**0.92**	**0.011**	**0.940**	173.10	**0.92**	**0.006**
	SCAD		0.943	**30.94**	2.41	0.021	0.930	**79.71**	2.71	0.015	0.848	179.31	5.18	0.021
	Lasso		0.944	28.55	1.91	0.020	0.934	78.22	2.80	0.022	0.862	**178.71**	7.29	0.035
	eLasso	0.9	**0.957**	26.27	**0.98**	**0.070**	**0.953**	76.21	**0.87**	**0.037**	**0.954**	176.48	**0.76**	**0.019**
	SCAD		0.946	**31.22**	6.84	0.217	0.945	**79.72**	7.26	0.129	0.935	**179.02**	8.68	0.081
	Lasso		0.955	28.67	1.60	0.073	0.950	78.09	1.73	0.040	0.948	177.77	1.71	0.021

Note: Bold symbols in the ARI and C indicate the highest values, while the lowest values for IC and MSE are also bolded. The optimal values of C vary by scenario and are determined by the 2×(p - number of non-zeros). Thus, for p = 25, C = (45, 40, 42); for p = 50, C = (95, 90, 92); for p = 100, C = (195, 190 192). The optimal value for IC is zero; The method denotes mixLMM with tested penalties.

Table B.3: Comparison results averaged by ARI, C, IC, and MSE among the proposed methods(mixLMM) through Exclusive Lasso (mixLMM-eLasso), SCAD (mixLMM-SCAD), and Lasso (mixLMM-Lasso) using datasets for n=400, p=(100, 200, 400) in unequal-sized group.

Scn	method	Corr	ARI	C	IC	MSE	ARI	C	IC	MSE	ARI	C	IC	MSE
			p=100				p=200				p=400			
1	eLasso	0.6	**0.939**	**188.80**	**0.00**	**0.001**	**0.943**	**388.98**	**0.00**	**0.000**	**0.944**	**788.86**	**0.00**	**0.000**
	SCAD		0.897	186.58	1.30	0.007	0.880	387.37	2.70	0.007	0.827	788.04	3.50	0.005
	Lasso		0.909	182.42	1.50	0.008	0.886	382.87	3.10	0.008	0.832	781.18	4.69	0.006
	eLasso	0.9	**0.952**	**188.22**	**0.00**	**0.001**	**0.951**	**388.50**	**0.00**	**0.001**	**0.951**	**788.78**	**0.00**	**0.000**
	SCAD		0.933	187.48	0.85	0.009	0.932	386.64	1.66	0.008	0.904	784.99	2.05	0.004
	Lasso		0.945	186.46	**0.00**	0.003	0.930	385.73	0.80	0.004	0.904	784.65	1.89	0.003
2	eLasso	0.6	**0.957**	**177.89**	**0.00**	**0.003**	**0.955**	**377.56**	**0.00**	**0.001**	**0.956**	**777.54**	**0.00**	**0.001**
	SCAD		0.939	176.64	1.60	0.010	0.891	375.58	4.62	0.013	0.826	777.24	7.81	0.010
	Lasso		0.941	174.27	2.20	0.014	0.896	372.85	6.80	0.018	0.784	773.86	11.60	0.015
	eLasso	0.9	**0.963**	**177.77**	**0.01**	**0.009**	**0.963**	**378.39**	**0.00**	**0.005**	**0.965**	**778.22**	**0.02**	**0.002**
	SCAD		0.949	177.16	7.77	0.078	0.951	376.69	8.99	0.046	0.933	775.63	10.49	0.027
	Lasso		0.956	176.50	0.17	0.016	0.951	376.30	0.48	0.009	0.932	773.92	3.31	0.008
3	eLasso	0.6	**0.946**	172.85	**0.84**	**0.004**	**0.949**	373.33	**0.87**	**0.002**	**0.946**	772.05	**1.00**	**0.001**
	SCAD		0.937	**179.17**	2.56	0.005	0.902	**378.43**	5.50	0.011	0.813	**780.59**	8.29	0.009
	Lasso		0.921	173.78	3.58	0.013	0.882	372.68	6.04	0.013	0.824	777.22	9.97	0.012
	eLasso	0.9	**0.958**	175.19	**0.67**	0.013	**0.959**	376.03	**0.55**	0.007	**0.960**	775.54	**0.55**	**0.004**
	SCAD		0.946	**179.13**	5.75	0.040	0.944	**378.27**	7.02	0.026	0.937	**778.83**	8.59	0.017
	Lasso		0.955	177.73	1.67	**0.011**	0.946	375.14	1.38	**0.005**	0.933	773.50	3.24	0.005

Note: Bold symbols in the ARI and C indicate the highest values, while the lowest values for IC and MSE are also bolded. The optimal values of C vary by scenario and are determined by the 2×(p - number of non-zeros). Thus, for p = 100, C = (190, 180, 184); for p = 200, C = (390, 380, 384); for p = 400, C = (790, 780 784). The optimal value for IC is zero; The method denotes mixLMM with tested penalties.

Table B.4: Mean of Age-Adjusted CVD Mortality by County-level Clusters, 2009–2018

			Age-adjusted Cardiovascular Mortality Rate per 100,000 people									
cluster	**n**	**Estimates**	**2009**	**2010**	**2011**	**2012**	**2013**	**2014**	**2015**	**2016**	**2017**	**2018**
1	553	Mean	299.7869	290.6725	284.5985	278.9776	276.7732	275.5144	274.5635	270.5397	266.6263	261.8534
		SD	67.8654	63.3691	62.6746	62.1821	63.3762	66.8574	68.6705	70.2719	68.7916	68.5605
		95% CI	(294.1181, 305.4556)	(285.3793, 295.9657)	(279.3634, 289.8337)	(273.7835, 284.1716)	(271.4795, 282.0670)	(269.9299, 281.0990)	(268.8275, 280.2995)	(264.6699, 276.4095)	(260.8802, 272.3724)	(256.1266, 267.5802)
2	809	Mean	269.7396	261.5496	255.2260	248.6982	243.3983	240.3766	236.3857	234.2953	231.0360	229.7953
		SD	46.0338	44.5499	44.3158	44.8784	45.6186	45.9094	45.5532	44.9432	44.6287	44.8000
		95% CI	(266.5627, 272.9164)	(258.4751, 264.6241)	(252.1676, 258.2843)	(245.6010, 251.7953)	(240.2500, 246.5465)	(237.2083, 243.5449)	(233.2419, 239.5294)	(231.1937, 237.3969)	(227.9561, 234.1159)	(226.7036, 232.8870)
3	501	Mean	261.1980	253.4011	249.9383	250.0211	251.8101	255.7772	259.4319	263.0898	263.7471	263.0792
		SD	50.1689	49.5393	49.4442	50.3221	50.8786	52.2229	52.1871	52.4580	52.2378	51.5264
		95% CI	(256.7943, 265.6016)	(249.0527, 257.7496)	(245.5982, 254.2783)	(245.6040, 254.4382)	(247.3441, 256.2761)	(251.1932, 260.3611)	(254.8510, 264.0127)	(258.4852, 267.6944)	(259.1618, 268.3323)	(258.5563, 267.6020)
4	521	Mean	246.3103	239.5967	234.8454	232.5815	231.3681	231.4084	230.6646	230.3858	227.3727	225.8177
		SD	39.4513	38.3454	38.0876	38.8493	39.5066	40.1028	39.3972	38.3042	37.5646	37.8144
		95% CI	(242.9148, 249.7058)	(236.2964, 242.8970)	(231.5673, 238.1236)	(229.2378, 235.9252)	(227.9678, 234.7683)	(227.9568, 234.8600)	(227.2738, 234.0555)	(227.0890, 233.6826)	(224.1396, 230.6059)	(222.5631, 229.0723)
5	310	Mean	233.7210	227.6229	226.6207	226.0461	226.9555	229.2119	229.7052	231.8958	232.2097	232.6697
		SD	39.4372	39.0649	40.1989	39.9637	39.6417	40.1756	40.5101	39.9895	40.4259	40.0473
		95% CI	(229.3136, 238.1283)	(223.2572, 231.9886)	(222.1282, 231.1131)	(221.5799, 230.5123)	(222.5253, 231.3857)	(224.7221, 233.7018)	(225.1779, 234.2324)	(227.4267, 236.3649)	(227.6918, 236.7275)	(228.1942, 237.1452)
6	360	Mean	230.1055	223.9980	220.1913	216.2424	213.1052	211.6966	210.1780	209.6916	208.4124	207.4483
		SD	30.0589	30.5878	30.9114	31.4284	31.4386	31.7309	31.8979	31.9719	31.8356	31.5142
		95% CI	(226.9899, 233.2210)	(220.8276, 227.1684)	(216.9874, 223.3952)	(212.9849, 219.4999)	(209.8466, 216.3638)	(208.4077, 214.9855)	(206.8718, 213.4842)	(206.3778, 213.0054)	(205.1127, 211.7121)	(204.1819, 210.7147)
7	170	Mean	223.0892	219.2550	217.2339	215.4762	214.2539	214.6362	213.6909	214.2539	213.4303	212.2686
		SD	37.0229	37.4307	37.1696	37.6320	38.2259	38.7869	38.5821	38.2601	38.7848	38.6850
		95% CI	(217.4836, 228.6947)	(213.5878, 224.9223)	(211.6061, 222.8616)	(209.7785, 221.1740)	(208.4662, 220.0415)	(208.7636, 220.5088)	(207.8493, 219.5325)	(208.4610, 220.0467)	(207.5581, 219.3026)	(206.4114, 218.1257)

Note: n denotes the number of counties in each distint cluster. SD defines a standard deviation.

Table B.5: BIC Values for Different Numbers of Clusters

BIC	**Number of Clusters**							
No. of try	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**
1	1×10^7	275697.5	273242.8	271502.7	270913.5	270451.6	270179.1	270356.8
2	1×10^7	275708	273007.4	271626.6	270940	270640.6	270288.5	270313
3	1×10^7	275680.9	273069.3	271675.7	271440	270676.2	270417.8	270395.4
4	1×10^7	275665.8	273056.9	271544	270948.6	270637.1	**270173.5**	270227.7
5	1×10^7	275730.4	273135.9	271494.3	271011.4	270718.6	270498.1	270272.9

Table B.6: Number of Rural-Urban Counties within Clusters, excluding US territories

Cluster	**Urban (%)**	**Rural (%)**	**Total**
1	93 (7.98)	419 (21.20)	512
2	315 (27.02)	467 (23.63)	782
3	134 (11.49)	365 (18.47)	499
4	164 (14.07)	355 (17.97)	519
5	227 (11.15)	129 (6.53)	356
6	130 (19.47)	180 (9.11)	310
7	103 (8.83)	61 (3.09)	164
Total	1166 (100)	1976 (100)	3142

Table B.7: State-wise breakdown of county counts by cluster categories.

GEOID	State	K1	K2	K3	K4	K5	K6	K7	Total
01	Alabama	20	18	12	7	4	3	3	67
02	Alaska	14	3	6	2	1	2	1	29
04	Arizona	0	4	1	1	3	5	1	15
05	Arkansas	18	18	16	13	6	3	1	75
06	California	2	24	1	5	6	17	3	58
08	Colorado	13	12	6	13	7	6	7	64
09	Connecticut	0	2	0	0	0	4	2	8
10	Delaware	0	1	0	0	0	1	1	3
11	District of Columbia	0	0	0	1	0	0	0	1
12	Florida	4	14	5	14	7	13	10	67
13	Georgia	47	33	27	34	9	6	3	159
15	Hawaii	0	1	0	1	2	1	0	5
16	Idaho	8	6	18	6	3	1	2	44
17	Illinois	6	30	9	19	19	12	7	102
18	Indiana	10	26	10	21	10	11	4	92
19	Iowa	9	28	19	18	13	10	2	99
20	Kansas	14	18	33	20	6	11	3	105
21	Kentucky	47	28	15	13	9	7	1	120
22	Louisiana	24	14	16	8	1	1	0	64
23	Maine	0	2	2	2	2	5	3	16
24	Maryland	0	7	7	5	1	4	0	24
25	Massachusetts	0	2	0	0	1	8	3	14
26	Michigan	3	21	10	20	10	14	5	83
27	Minnesota	2	5	11	25	14	17	13	87
28	Mississippi	34	20	16	10	2	0	0	82
29	Missouri	27	37	25	10	3	8	5	115
30	Montana	8	8	16	10	7	5	2	56
31	Nebraska	13	24	11	24	11	8	2	93
32	Nevada	4	4	3	2	2	1	1	17
33	New Hampshire	0	0	0	2	2	3	3	10
34	New Jersey	0	4	0	0	2	10	5	21
35	New Mexico	2	5	4	11	8	1	2	33
36	New York	6	27	0	8	3	17	1	62
37	North Carolina	7	37	8	15	8	21	4	100
38	North Dakota	12	8	8	16	6	1	2	53
39	Ohio	8	17	13	13	21	9	7	88
40	Oklahoma	20	23	25	2	3	3	1	77
41	Oregon	0	5	3	9	7	7	5	36
42	Pennsylvania	2	28	4	4	4	15	10	67
44	Rhode Island	0	1	0	0	0	4	0	5
45	South Carolina	5	18	2	9	2	6	4	46
46	South Dakota	16	10	15	18	4	2	1	66
47	Tennessee	26	19	24	13	6	5	2	95
48	Texas	34	53	55	36	39	27	10	254
49	Utah	6	4	3	6	7	1	2	29
50	Vermont	0	1	2	1	1	2	2	9
51	Virginia	3	17	11	14	10	17	3	75
53	Washington	0	6	8	3	6	10	6	39
54	West Virginia	3	10	3	11	2	5	1	35
55	Wisconsin	9	21	7	18	18	8	2	83
56	Wyoming	5	7	7	8	5	2	1	35
66	Guam	1	0	0	0	0	0	0	1
72	Puerto Rico	40	27	2	2	0	4	3	78
78	U.S. Virgin Islands	0	0	0	0	0	0	3	3
Total (County)		553	809	501	521	310	360	170	3224
Total (State)		39	50	42	47	47	49	45	

Appendix C

Use of Model-base Clustering of High-Dimensional Longitudinal Data via Exclusive Lasso Penalty by Different Levels of SDOH Data

Table C.1: Mean of Age-Adjusted CVD Mortality by 1/7 Cut-off County-level Clusters, 2009–2018

			Age-adjusted Cardiovascular Mortality Rate per 100,000 people									
cluster	**n**	**Estimates**	**2009**	**2010**	**2011**	**2012**	**2013**	**2014**	**2015**	**2016**	**2017**	**2018**
1	447	Mean	304.7523	295.8413	290.5737	285.8079	284.2189	282.6958	281.0408	277.0917	272.8237	268.4093
		SD	71.0763	66.7627	66.0860	65.5670	66.1441	69.7681	72.4124	74.7987	73.6948	73.6528
		95% CI	(298.1454, 311.3592)	(289.6353, 302.0472)	(284.4307, 296.7168)	(279.7131, 291.9027)	(278.0705, 290.3673)	(276.2105, 289.1811)	(274.3096, 287.7719)	(270.1388, 284.0447)	(265.9734, 279.6740)	(261.5629, 275.2558)
2	1072	Mean	268.6067	260.2863	254.4522	248.1919	243.4732	240.8265	237.4837	235.4016	232.3813	230.9952
		SD	46.2955	44.6262	44.1383	44.2852	44.4991	44.6715	44.1898	43.5601	43.3562	43.4987
		95% CI	(265.8322, 271.3812)	(257.6119, 262.9607)	(251.8070, 257.0973)	(245.5379, 250.8459)	(240.8064, 246.1400)	(238.1493, 243.5036)	(234.8354, 240.1320)	(232.7910, 238.0121)	(229.7829, 234.9796)	(228.3883, 233.6020)
3	535	Mean	264.2712	257.7948	254.0622	253.3776	254.2062	258.1621	261.0447	263.8986	263.8008	262.3810
		SD	46.8487	46.0608	45.8430	46.9615	48.6153	51.0647	51.6164	51.2921	50.9182	49.9867
		95% CI	(260.2924, 268.2501)	(253.8829, 261.7067)	(250.1688, 257.9556)	(249.3892, 257.3660)	(250.0773, 258.3351)	(253.8252, 262.4990)	(256.6610, 265.4284)	(259.5424, 268.2548)	(259.4764, 268.1253)	(258.1357, 266.6263)
4	672	Mean	235.6188	228.9283	225.2618	223.5052	223.0188	223.3649	223.8531	224.7638	223.3579	222.3737
		SD	38.1039	36.4791	36.0183	36.7096	37.8610	38.3268	38.3507	37.8233	37.5587	37.4750
		95% CI	(232.7326, 238.5049)	(226.1652, 231.6913)	(222.5336, 227.9899)	(220.7247, 226.2857)	(220.1510, 225.8865)	(220.4619, 226.2679)	(220.9483, 226.7580)	(221.8990, 227.6287)	(220.5130, 226.2027)	(219.5352, 225.2122)
5	207	Mean	232.5821	226.6826	224.2541	223.6367	223.9256	226.4531	226.3536	227.8874	227.8734	227.8425
		SD	36.6961	36.5137	36.9289	36.1600	35.8816	36.4314	36.0948	36.0212	35.8324	36.7053
		95% CI	(227.5536, 237.6107)	(221.6791, 231.6862)	(219.1937, 229.3146)	(218.6816, 228.5918)	(219.0087, 228.8425)	(221.4609, 231.4454)	(221.4075, 231.2998)	(222.9514, 232.8235)	(222.9632, 232.7836)	(222.8127, 232.8723)
6	221	Mean	223.6851	217.4335	213.3045	209.7367	206.7244	205.4407	203.5828	203.0054	201.4063	200.4167
		SD	25.7137	26.1651	25.9044	26.5795	26.7175	27.0895	27.1780	27.0670	27.2326	27.0480
		95% CI	(220.2762, 227.0939)	(213.9648, 220.9022)	(209.8704, 216.7387)	(206.2130, 213.2603)	(203.1825, 210.2664)	(201.8495, 209.0320)	(199.9798, 207.1858)	(199.4171, 206.5937)	(197.7961, 205.0166)	(196.8310, 204.0025)
7	70	Mean	213.3631	208.4431	205.2302	203.4931	202.9274	203.7116	203.9774	204.5388	203.8874	202.6774
		SD	33.1974	33.3968	33.0356	34.5129	34.5273	35.3142	34.9275	35.2079	35.1865	34.3545
		95% CI	(205.4474, 221.2787)	(200.4799, 216.4063)	(197.3532, 213.1073)	(195.2638, 211.7224)	(194.6946, 211.1601)	(195.2913, 212.1320)	(195.6492, 212.3055)	(196.1437, 212.9338)	(195.4974, 212.2773)	(194.4858, 210.8689)

Note: n denotes the number of counties in each distint cluster. SD defines a standard deviation.

Table C.2: Selected social determinants of health associated with age-adjusted CVD mortality in each cluster by 1/14 cutoff county-level clustering: Model-based clustering via Exclusive lasso with random effects in intercepts, 2009–2018

Domains	Variables	K1	K2	K3	K4	K5	W6	K7
	Int	276.6916	244.7261	257.2023	233.3886	225.7982	212.7531	211.4275
1	% Housing units with more than one occupant per room	0	0	0	0	0	0	-0.0165
1	% Population reporting Asian race	0	0	0	0	-0.1058	-0.0950	-0.6253
1	% Population reporting Black race	0.3756	0.4497	0.6409	0.6672	1.0288	0.5762	0.7010
1	% Families with Children that are single-parent Families	0.0279	0	0.0861	0	0	0	0
1	% Population that does not speak English at all (ages 5 and over)	-0.1868	0	0	0	-0.3084	0	0
1	% Population that is foreign-born	-0.7811	0	0	-0.8865	0	-0.0552	-0.1739
1	% Children living with a grandparent householder (ages 17 and under)	0	0	0	0	0.3819	0	0
1	% Occupied housing units without fuel	-0.2844	0	0	0	0	0	0
1	% Population reporting Hispanic ethnicity	-0.4567	-0.5666	-0.2274	-0.3057	0	-0.0048	0
1	% Population reporting multiple races	0	0	0	0	0	-0.2423	0
1	% Population reporting Native Hawaiian/Pacific Islander race	0	0	5.5601	-0.0849	-0.2624	0	0
1	% Population who are not U.S. citizens and entered the U.S. before 1990	-0.0774	0	-0.8147	-0.0915	0	0	0
1	% Population who speak other languages (ages 5 and over)	0	0	0	0	0	0	-0.0052
1	% Civilian Population consisting of veterans (ages 18 and over)	0.0408	2.0556	-0.6133	0	0	1.0664	0.0387
2	Median household income (in dollars) /1000	-0.3358	-0.1748	0	-0.2230	0	0	0
2	% Unmarried partner households that received food stamps/SNAP benefits	0	-0.1366	-0.0120	0	0	0	-0.0004
2	% Employed working in public administration	0	0	0	-0.0229	0	0	0
2	% Civilian Population in armed forces (ages 16 years and over)	0.8621	0	0	0.3457	0	0	0
2	% Employed working in construction	0.5574	0.4894	0	0	0	0.5752	0.0809
2	% Employed working in finance and insurance, real estate, etc.	0	0	-0.4522	0	0	0	0
2	% Households that received food stamps/SNAP, past 12 months	-0.3261	-0.7405	0	0	0	-0.5241	0
2	% Households with public assistance income or food stamps/SNAP	-0.0726	0	0	0	-0.8127	0	-1.1433
2	% Employed working in information services	0.2305	0	-2.2572	2.9907	0	0	0
2	% Employed working in manufacturing	0.0177	0.1295	0.1262	0.3837	0.4092	0.7210	0.5077
2	% Employed working in agriculture, forestry, fishing, etc. (ages 16 and over)	-0.7732	-0.0883	-0.1458	0	0	0	0
2	% Employed working in other services, except public administration	0	0	0.7127	0	0	0	0
2	% Population with income to poverty ratio: 1.25-1.99	-0.1000	0	0	0	0	0	0
2	% Population with income to poverty ratio: <1.00	0	0	0.1859	0	0	0	0
2	% Employed working in professional, scientific, management, administrative, etc.	0	0	-0.0454	0	0	0	0
2	% Employed working in transportation and warehousing, and in utilities	0.0383	0	0	0	0.2355	0	0
2	% Population that was unemployed (ages 16 years and over)	0.9293	0.0592	-1.1274	-0.0288	-1.2616	0	-0.1005
2	% Employed working in wholesale trade	0.3539	1.1166	0	0	0	0	0.4549
3	% Population with some college or associate's degree (ages 25 and over)	0	0	-0.1832	-0.0119	0	0	0
3	% Population with a bachelor's degree (ages 25 and over)	0	0	-0.3475	0	-0.1323	0	0
3	% Population with a master's or highter degree (ages 25 and over)	-0.4561	-2.3631	-0.5953	-0.5655	-0.1505	-2.0155	-0.8791
3	% Population with only high school diploma (ages 25 and over)	0.4930	0	0	0	0.0358	0	0
3	% Population with less than high school education (ages 25 and over)	1.4860	2.0879	0	1.2639	0.2948	0.4909	0.3696
4	Median home value of owner-occupied housing units	-0.00011	-0.00006	-0.00001	-0.00006	-0.00003	-0.00001	0
4	% Housing in structures with 10 or more units	-0.3532	0	0	-0.0378	-0.4332	0	0
4	% Workers with at least 60-minute commute time (ages 16 and over)	0	0	0	0	0	-0.0028	0
4	% Housing units that are mobile homes	0.1622	0.2433	0.3848	0.2865	0	0.7600	0.8139
4	% Housing units with no vehicle available	0	0.30143	0	0	0	0	0
4	Workers (16 +) with a 60+ min public transit commute	0.0143	0.0116	0.0161	-0.0402	0	0	-0.0283
4	% Workers taking public transportation, excluding taxicab (ages 16 and over)	0.3540	0	-0.6969	0	-2.2075	0	0
4	% Rental units with rent equal to 30 percent or more of household income	0	0	-0.0228	0	0	0	0
4	% Housing units vacant	-0.4469	-0.3841	0	-0.0200	0	-0.0048	0
4	Beer, wine and liquor stores per 1,000 people	10.4898	-2.6260	7.1360	0	0	0	0
4	Convenience stores per 1,000 people	3.4677	0	0	0	0	0	0
4	Full service restaurants per 1,000 people	-0.1267	-0.4046	-4.6795	0	0	0	0
4	Supermarkets and other grocery (except convenience) stores per 1,000 people	-6.8065	0	0	0	0	0	-1.7719
5	Number of Federally Qualified Health Centers	-0.1200	-0.0746	2.2300	0.1338	0.2636	-0.0389	-0.0381
5	Total number of community mental health centers	-1.9253	1.7287	-0.0950	0	-4.0280	-0.0690	0.5185
5	Number of rural health clinics	-0.2207	0.0837	1.2492	-0.1563	0.5956	0	0.3225
5	Number of people living with diagnosed HIV / 1000	23.1298	0.0789	0	1.2689	0	0.2428	0
5	Number of Medicare eligibles in the county	-0.0005	0	0	0	0	0	0
5	Medicare, ratio of enrollees over Medicare- eligible, %	-0.4200	-0.4640	0.4354	-0.2307	0.0732	-0.2662	-0.1593
6	1: Rural, 0: Urban by NCHS 2013 Rural-Urban Classification Scheme	4.9288	2.0972	6.7817	-1.4872	7.7028	1.2883	6.3794

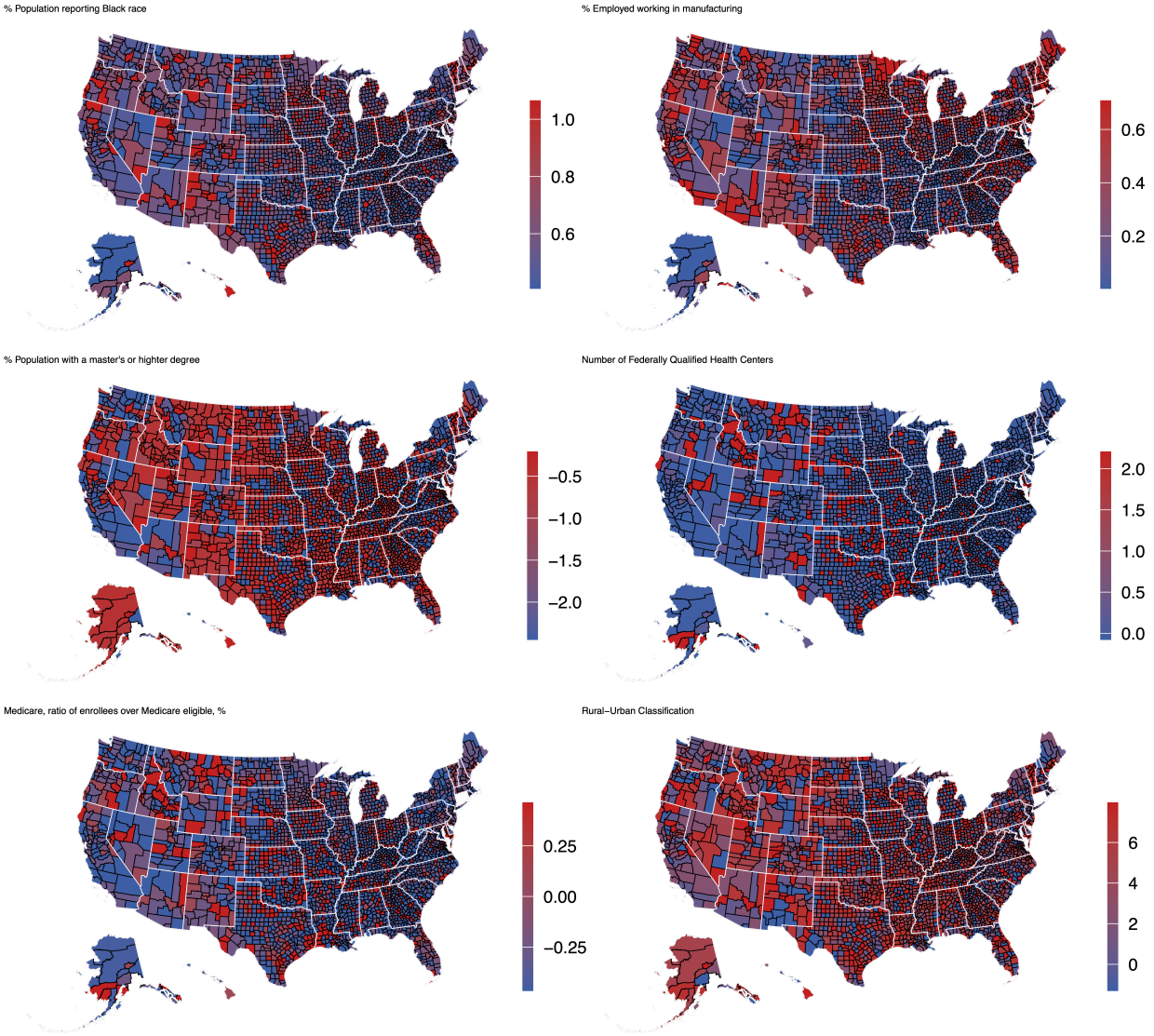


Figure C.1: The spatial structures of nonzero coefficients associated with age-adjusted cardiovascular disease mortality across counties in the United States, Left: % Population reporting Black race, % Population with a master's or higher degree, Medicare, ratio of enrollees over Medicare- eligible, %; Right: % Employed working in manufacturing, Number of Federally Qualified Health Centers, Rural-Urban Classification.

Bibliography

Andersen, P. K., & Gill, R. D. (1982). Cox's regression model for counting processes: A large sample study. *The annals of statistics*, 1100–1120.

Arribas-Gil, A., De la Cruz, R., Lebarbier, E., & Meza, C. (2015). Classification of longitudinal data through a semiparametric mixed-effects model based on lasso-type estimators. *Biometrics*, *71*(2), 333–343.

Banerjee, A. (2017). Bridging the global digital health divide for cardiovascular disease. *Circulation: Cardiovascular Quality and Outcomes*, *10*(11), e004297.

Benjamin, E. J., Muntner, P., Alonso, A., Bittencourt, M. S., Callaway, C. W., Carson, A. P., Chamberlain, A. M., Chang, A. R., Cheng, S., Das, S. R., et al. (2019). Heart disease and stroke statistics—2019 update: A report from the american heart association. *Circulation*, *139*(10), e56–e528.

Bickel, P. J., & Levina, E. (2008). Regularized estimation of large covariance matrices.

Blommaert, A., Hens, N., & Beutels, P. (2014). Data mining for longitudinal data under multicollinearity and time dependence using penalized generalized estimating equations. *Computational Statistics & Data Analysis*, *71*, 667–680.

Bondell, H. D., Krishna, A., & Ghosh, S. K. (2010). Joint variable selection for fixed and random effects in linear mixed-effects models. *Biometrics*, *66*(4), 1069–1077.

Breslow, N. E., & Clayton, D. G. (1993). Approximate inference in generalized linear mixed models. *Journal of the American Statistical Association*, *88*(421), 9–25. https://doi.org/10.1080/01621459.1993.10594284

Buckley, J., & James, I. (1979). Linear regression with censored data. *Biometrika*, *66*(3), 429–436.

Campbell, F., & Allen, G. I. (2017). Within group variable selection through the exclusive lasso.

Case, A., & Deaton, A. (2015). Rising morbidity and mortality in midlife among white non-hispanic americans in the 21st century. *Proceedings of the National Academy of Sciences*, *112*(49), 15078–15083.

Chen, Y., Yu, M.-H., & Wang, L. (2018). Identifying critical time points in longitudinal studies: Impact on disease prediction and intervention. *Statistical Methods in Medical Research*, *27*(4), 1123–1136. https://doi.org/10.1177/0962280217710835

Churchwell, K., Elkind, M. S., Benjamin, R. M., Carson, A. P., Chang, E. K., Lawrence, W., Mills, A., Odom, T. M., Rodriguez, C. J., Rodriguez, F., et al. (2020). Call to action: Structural racism as a fundamental driver of health disparities: A presidential advisory from the american heart association. *Circulation*, *142*(24), e454–e468.

Dempster, A. P., Laird, N. M., & Rubin, D. B. (1977). Maximum likelihood from incomplete data via the em algorithm. *Journal of the royal statistical society: series B (methodological)*, *39*(1), 1–22.

Diggle, P. J. (2002). *Analysis of longitudinal data*. Oxford University Press.

Dong, W., Motairek, I., Nasir, K., Chen, Z., Kim, U., Khalifa, Y., Freedman, D., Griggs, S., Rajagopalan, S., & Al-Kindi, S. G. (2023). Risk factors and geographic disparities in premature cardiovascular mortality in us counties: A machine learning approach. *Scientific Reports*, *13*(1), 2978.

Du, Y., Khalili, A., Nešlehová, J. G., & Steele, R. J. (2013). Simultaneous fixed and random effects selection in finite mixture of linear mixed-effects models. *Canadian Journal of Statistics*, *41*(4), 596–616.

Explorer, R. D. (n.d.). *Rural health information hub*. https://www.ruralhealthinfo.org/data-explorer?id=181%5C&year=2009 (accessed: 03.11.2022).

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

Fan, J., Li, R., Zhang, C.-H., & Zou, H. (2020). Statistical foundations of data science. *Annual Review of Statistics and Its Application*, *7*, 1–38. https://doi.org/10.1146/annurev-statistics-031219-041135

Fan, J., & Lv, J. (2010). A selective overview of variable selection in high dimensional feature space. *Statistica Sinica*, *20*(1), 101.

Fan, J., & Lv, J. (2011). Nonconcave penalized likelihood with np-dimensionality. *IEEE Transactions on Information Theory*, *57*(8), 5467–5484.

Fan, Y., & Li, R. (2012). Variable selection in linear mixed effects models. *Annals of statistics*, *40*(4), 2043.

Fitzmaurice, G. M., Laird, N. M., & Ware, J. H. (2012). *Applied longitudinal analysis*. John Wiley & Sons.

for Disease Control, C., & (CDC), P. (2023). Data brief no. 492: Selected tables [Accessed: 2024-12-04]. https://www.cdc.gov/nchs/products/databriefs/db492.htm%5C#section_4

for Disease Control, C., Prevention et al. (2019). Health, united states spotlight: Race and ethnic disparities in heart disease.

for Healthcare Research, A., & Quality, M., Rockville. (2020). *Social determinants of health database*. https://www.ahrq.gov/sdoh/data-analytics/sdoh-data.html (accessed: 08.19.2022).

Frieden, T. R., et al. (2013). Cdc health disparities and inequalities report-united states, 2013. foreword. *MMWR supplements*, *62*(3), 1–2.

Friedman, J., Hastie, T., & Tibshirani, R. (2010). Regularization paths for generalized linear models via coordinate descent. *Journal of statistical software*, *33*(1), 1.

Fu, W. J. (2003). Penalized estimating equations. *Biometrics*, *59*(1), 126–132.

Fu, Y., Yu, G., Maulana, N., & Thomson, K. (2023). Interventions to tackle health inequalities in cardiovascular risks for socioeconomically disadvantaged populations: A rapid review. *British Medical Bulletin*, *148*(1), 22–41.

Genolini, C., & Falissard, B. (2010). Kml: K-means for longitudinal data. *Computational Statistics*, *25*(2), 317–328.

Geyer, C. J. (1996). On the asymptotics of convex stochastic optimization. *Unpublished manuscript*, *37*.

Glynn, P. A., Molsberry, R., Harrington, K., Shah, N. S., Petito, L. C., Yancy, C. W., Carnethon, M. R., Lloyd-Jones, D. M., & Khan, S. S. (2021). Geographic variation in trends and disparities in heart failure mortality in the united states, 1999 to 2017. *Journal of the American Heart Association*, *10*(9), e020541.

Graham, G. (2015). Disparities in cardiovascular disease risk in the united states. *Current cardiology reviews*, *11*(3), 238–245.

Hacker, K., Auerbach, J., Ikeda, R., Philip, C., & Houry, D. (2022). Social determinants of health—an approach taken at cdc. *Journal of public health management and practice*, *28*(6), 589–594.

Hardin, J. W., & Hilbe, J. M. (2003). *Generalized estimating equations*. Chapman & Hall/CRC.

Hedeker, D., & Gibbons, R. D. (1997). Application of random-effects pattern-mixture models for missing data in longitudinal studies. *Psychological methods*, *2*(1), 64.

Hubbard, A. E., Ahern, J., Fleischer, N. L., Van der Laan, M., Lippman, S. A., Jewell, N., Bruckner, T., & Satariano, W. A. (2010). To gee or not to gee: Comparing population average and mixed models for estimating the associations between neighborhood risk factors and health. *Epidemiology*, *21*(4), 467–474.

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

Ingram, D. D., & Franco, S. J. (2014). *2013 nchs urban-rural classification scheme for counties*. US Department of Health; Human Services, Centers for Disease Control and…

Jacques, J., & Preda, C. (2013). Funclust: A curves clustering method using functional random variables density approximation. *Neurocomputing*, *112*, 164–171.

Javed, Z., Haisum Maqsood, M., Yahya, T., Amin, Z., Acquah, I., Valero-Elizondo, J., Andrieni, J., Dubey, P., Jackson, R. K., Daffin, M. A., et al. (2022). Race, racism, and cardiovascular health: Applying a social determinants of health framework to racial/ethnic disparities in cardiovascular disease. *Circulation: Cardiovascular Quality and Outcomes*, *15*(1), e007917.

Johnson, B. A., Lin, D., & Zeng, D. (2008). Penalized estimating functions and variable selection in semiparametric regression models. *Journal of the American Statistical Association*, *103*(482), 672–680.

Komárek, A., & Komárková, L. (2013). Clustering for multivariate continuous and discrete longitudinal data.

Kong, D., Xiong, H., & Ding, B. (2017). Exclusive feature learning on arbitrary structures via $\ell_{1,2}$-norm. *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, *31*(1), 1681–1687.

Kowalski, J., Hao, S., Chen, T., Liang, Y., Liu, J., Ge, L., Feng, C., & Tu, X. (2018). Modern variable selection for longitudinal semi-parametric models with missing data. *Journal of Applied Statistics*, *45*(14), 2548–2562.

Kowalski, J., & Tu, X. M. (2008). *Modern applied u-statistics*. John Wiley & Sons.

Laird, N. M., & Ware, J. H. (1982). Random-effects models for longitudinal data. *Biometrics*, *38*(4), 963–974.

Lam, C., & Fan, J. (2009). Sparsistency and rates of convergence in large covariance matrix estimation. *Annals of statistics*, *37*(6B), 4254.

Lan, L. (2006). Variable selection in linear mixed model for longitudinal data.

Li, Y., Wang, S., Song, P. X.-K., Wang, N., Zhou, L., & Zhu, J. (2018). Doubly regularized estimation and selection in linear mixed-effects models for high-dimensional longitudinal data. *Statistics and its Interface*, *11*(4), 721.

Liang, K.-Y., & Zeger, S. L. (1986). Longitudinal data analysis using generalized linear models. *Biometrika*, *73*(1), 13–22.

Lin, D. Y., & Ying, Z. (2001). Semiparametric and nonparametric regression analysis of longitudinal data. *Journal of the American Statistical Association*, *96*(453), 103–126.

Lin, G., Rodriguez, R. N., & SAS, I. (2015). Weighted methods for analyzing missing data with the gee procedure. *Paper SAS166-2015*, 1–8.

Little, R. J. A., & Rubin, D. B. (2019). *Statistical analysis with missing data* (3rd). John Wiley & Sons. https://doi.org/10.1002/9781119482260

Liu, X., & Deme, P. (2012). Modeling time-varying effects in longitudinal data analysis: A review. *Statistical Modelling*, *12*(3), 275–299. https://doi.org/10.1177/1471082X12437710

Mallinckrodt, C. (2013). *Preventing and treating missing data in longitudinal clinical trials: A practical guide*. Cambridge University Press.

McNeill, E., Lindenfeld, Z., Mostafa, L., Zein, D., Silver, D., Pagán, J., Weeks, W. B., Aerts, A., Des Rosiers, S., Boch, J., et al. (2023). Uses of social determinants of health data to address cardiovascular disease and health equity: A scoping review. *Journal of the American Heart Association*, *12*(21), e030571.

McNicholas, P. D., & Murphy, T. B. (2010). Model-based clustering of longitudinal data. *Canadian Journal of Statistics*, *38*(1), 153–168.

Mensah, G. A., Wei, G. S., Sorlie, P. D., Fine, L. J., Rosenberg, Y., Kaufmann, P. G., Mussolino, M. E., Hsu, L. L., Addou, E., Engelgau, M. M., et al. (2017). Decline in cardiovascular mortality: Possible causes and implications. *Circulation research*, *120*(2), 366.

Muller, S., Scealy, J. L., & Welsh, A. H. (2013). Model selection in linear mixed models. *Statistical Science*, *28*(2), 135–167. https://doi.org/10.1214/12-STS410

of Heart Disease, I. A., & Stroke. (n.d.). *Centers for disease control and prevention*. http://nccd.cdc.gov/DHDSPAtlas (accessed: 08.26.2022).

O'Kelly, M., & Ratitch, B. (2014). *Clinical trials with missing data: A guide for practitioners*. John Wiley & Sons.

Patel, S. A., Ali, M. K., Narayan, K. V., & Mehta, N. K. (2016). County-level variation in cardiovascular disease mortality in the united states in 2009–2013: Comparative assessment of contributing factors. *American journal of epidemiology*, *184*(12), 933–942.

Pollard, D. (1991). Asymptotics for least absolute deviation regression estimators. econometrics theory, 7, 186-199.

Post, W. S., Watson, K. E., Hansen, S., Folsom, A. R., Szklo, M., Shea, S., Barr, R. G., Burke, G., Bertoni, A. G., Allen, N., et al. (2022). Racial and ethnic differences in all-cause and cardiovascular disease mortality: The mesa study. *Circulation*, *146*(3), 229–239.

Powell-Wiley, T. M., Baumer, Y., Baah, F. O., Baez, A. S., Farmer, N., Mahlobo, C. T., Pita, M. A., Potharaju, K. A., Tamura, K., & Wallen, G. R. (2022). Social determinants of cardiovascular disease. *Circulation research*, *130*(5), 782–799.

Preisser, J. S., Lohman, K. K., & Rathouz, P. J. (2002). Performance of weighted estimating equations for longitudinal binary data with drop-outs missing at random. *Statistics in medicine*, *21*(20), 3035–3054.

Proust-Lima, C., Philipps, V., & Liquet, B. (2015). Estimation of extended mixed models using latent classes and latent processes: The r package lcmm. *arXiv preprint arXiv:1503.00890*.

Robins, J. M., Rotnitzky, A., & Zhao, L. P. (1995). Analysis of semiparametric regression models for repeated outcomes in the presence of missing data. *Journal of the american statistical association*, *90*(429), 106–121.

Robinson, W. S. (2009). Ecological correlations and the behavior of individuals. *International journal of epidemiology*, *38*(2), 337–341.

Roth, G. A., Dwyer-Lindgren, L., Bertozzi-Villa, A., Stubbs, R. W., Morozoff, C., Naghavi, M., Mokdad, A. H., & Murray, C. J. (2017). Trends and patterns of geographic variation in cardiovascular mortality among us counties, 1980-2014. *Jama*, *317*(19), 1976–1992.

Schelldorfer, J., Bühlmann, P., & DE GEER, S. V. (2010). Estimation for high-dimensional linear mixed-effects models using l1-penalization [Published online: February 19, 2010]. *Scandinavian Journal of Statistics*, *38*(2), 197–214. https://doi.org/10.1111/j.1467-9469.2011.00740.x

Son, H., Zhang, D., Shen, Y., Jaysing, A., Zhang, J., Chen, Z., Mu, L., Liu, J., Rajbhandari-Thapa, J., Li, Y., et al. (2023). Social determinants of cardiovascular health: A longitudinal analysis of cardiovascular disease mortality in us counties from 2009 to 2018. *Journal of the American Heart Association*, *12*(2), e026940.

Steinley, D. (2004). Properties of the hubert-arabie adjusted rand index. *Psychological Methods*, *9*(3), 386–396. https://doi.org/10.1037/1082-989X.9.3.386

Stokes, M. E., Davis, C. S., & Koch, G. G. (2012). *Categorical data analysis using sas* (3rd). SAS Institute Inc.

Tang, X., & Qu, A. (2016). Mixture modeling for longitudinal data. *Journal of Computational and Graphical Statistics*, *25*(4), 1117–1137.

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

Tsiatis, A. A. (2006). Semiparametric theory and missing data. https://api.semanticscholar.org/CorpusID:118005650

Verbeke, G., & Molenberghs, G. (2000). *Linear mixed models for longitudinal data*. Springer.

Virani, S. S., Alonso, A., Aparicio, H. J., Benjamin, E. J., Bittencourt, M. S., Callaway, C. W., Carson, A. P., Chamberlain, A. M., Cheng, S., Delling, F. N., et al. (2021). Heart disease and stroke statistics-2021 update: A report from the american heart association.

Wang, L., Zhou, J., & Qu, A. (2012). Penalized generalized estimating equations for high-dimensional longitudinal data analysis. *Biometrics*, *68*(2), 353–360.

White-Williams, C., Rossi, L. P., Bittner, V. A., Driscoll, A., Durant, R. W., Granger, B. B., Graven, L. J., Kitko, L., Newlin, K., Shirey, M., et al. (2020). Addressing social determinants of health in the care of patients with heart failure: A scientific statement from the american heart association. *Circulation*, *141*(22), e841–e863.

Yang, L., & Wu, T. T. (2022). Model-based clustering of high-dimensional longitudinal data via regularization. *Biometrics*.

Yuan, M., & Lin, Y. (2006). Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, *68*(1), 49–67.

Zelko, A., Salerno, P. R., Al-Kindi, S., Ho, F., Rocha, F. P., Nasir, K., Rajagopalan, S., Deo, S., & Sattar, N. (2023). Geographically weighted modeling to explore social and environmental factors affecting county-level cardiovascular mortality in people with diabetes in the united states: A cross-sectional analysis. *The American Journal of Cardiology*, *209*, 193–198.

Zhao, L., Zhang, L., & Wen, L. (2018). The exclusive lasso for group feature selection. *Pattern Recognition*, *79*, 494–507. https://doi.org/10.1016/j.patcog.2018.02.024

Zhou, Y., Jin, R., & Hoi, S. C.-.-H. (2010). Exclusive lasso for multi-task feature selection. *Proceedings of the thirteenth international conference on artificial intelligence and statistics*, 988–995.

Zhu, Y., & Xie, T. (2017). Big data and social determinants of health: A perspective from data science. *Journal of Public Health Management and Practice*, *23*, S75–S83. https://doi.org/10.1097/PHH.0000000000000581

Zou, H. (2006). The adaptive lasso and its oracle properties. *Journal of the American statistical association*, *101*(476), 1418–1429.

Zou, H., & Hastie, T. (2005). Regularization and variable selection via the elastic net. *Journal of the royal statistical society: series B (statistical methodology)*, *67*(2), 301–320.