Finite Mixtures of Generalised Linear Mixed-Effect Models

Gillian L. Ainsworth

Doctor of Philosophy

Department of Epidemiology, Biostatistics and Occupational Health

McGill University

Montreal, Quebec

2020-12-11

A thesis submitted to the Faculty of Graduate Studies and Research in partial fulfilment of the requirements of the degree of Doctor of Philosophy

DEDICATION

I dedicate my thesis to:

My father, Bob Ainsworth, he dedicated his thesis to my sister and I; it has taken me a while to return the favour, thank-you for being patient.

My mother, Buffy Ainsworth (1952-2006), wish you were here to see this.

My grandfather John Halkirk Ainsworth (1909-1972), for his love of mathematics, which continues to inspire me.

My grandmother, Stuart Patterson aka Nana (1914-2009), who always assured me that girls could do math, and was always supportive.

PREFACE

This thesis includes original contributions to the methods for estimating and selecting parameters in finite mixtures of generalised linear mixed-effect models. For all chapters, I developed the ideas in collaboration with Erica Moodie and Abbas Khalili. All chapters were written by Gillian Ainsworth. All methodological work, derivations, programming, design of simulation studies, data analysis, and interpretation was carried out by Gillian Ainsworth. The data for chapter 5 were provided by the Scottish Early Rheumatoid Arthritis Study, for which an ethics approval was obtained from the West of Scotland Research Ethics Service by the original study. The thesis was reviewed and feedback provided by Erica Moodie, Abbas Khalili and Robert Platt.

ACKNOWLEDGEMENTS

First and foremost, I would like to thank my family, Bob Ainsworth, Heather Ainsworth, Stephen Ainsworth, Ian Stoski and Abigail Stoski, for celebrating with me during the achievements of graduate school and also commiserating through the difficult parts.

I must also thank my supervisor, Dr. Robert Platt, for his advice and guidance.

For their very useful and relevant comment on my protocol defence, I wish to thank Dr. Jim Hanley, Dr. Michal Abrahamowicz and Dr. Michael Kramer. For their comments, suggestions, and questions during my defence I thank Dr. Michal Abrahamowicz, Dr. Celia Greenwood, Dr. Alejandro Murua, Dr. Andrea Benedetti, and Dr. Paul McNicholas. I am grateful to Jagtar Singh Nigjjar and the SERA team for use of the SERA dataset. I must acknowledge Yeting Du for his work on finite mixtures of linear mixed-effects (FMLME) models which I have built on and referenced frequently.

The simulations and computational aspects of this thesis were labour intensive, and I wish to thank Tim Green and Dr. Dave Stephens for their help on coding. I am also grateful to Dr. Russell Steele, Dr. Johanna Neslehova, Dr. Erica Moodie and Dr. Abbas Khalili for their help. For answering all of my questions about biology from cells to diagnostic imaging, I am grateful to Robert Jiang. Thanks to Nick Verbree for motivating an example and extension to the model.

I am forever thankful to many others that helped me in many ways to survive graduate school: Pat Kirby, Leila Golparvar, Scarlett Kazimer, Luc Villandre, Nina

Sun, Golsa Dehghan, Libby Kirby, Geneviève Hébert-Jodoin, Mercedes Eustergerling, Clio Montrey, and Andrew Chiachia.

ABSTRACT

Complex datasets necessitate a thorough understanding of the research question and substantive area, and in some cases may require analyses that take into account the many features of the population of interest. Consider the Scottish Early Rheumatoid Arthritis or SERA study, which followed people in Scotland with rheumatoid arthritis (RA) over time, starting with a diagnosis of rheumatoid arthritis and then every six months thereafter. Outcomes of interest are often discrete and perhaps a binomial variable (such as if the patient has entered remission) or a count (such as the number of swollen joints). Clinicians are aware that treatments for RA vary in their effectiveness across the population, suggesting that there may be distinct subgroups within the overall population. Thus there are several key features that the analysis must take into account: (i) the discrete nature of the outcome, (ii) the longitudinal nature of the measurements, and (iii) the possibility of subpopulations across which treatment and covariate effects differ. In my thesis, I propose the use of a finite mixture of generalised linear mixed-effect models (FinMix GLMM) as an appropriate analytic approach. First, I develop the FinMix GLMM model and derive the maximum likelihood estimates. Next, I develop a penalised likelihood approach for both fixed and random effects selection in FinMix GLMM. Simulations illustrate the finite sample performance of the estimates. Lastly, a FinMix GLMM analysis (both with and without penalisation) to the SERA dataset demonstrates the real world application of this model.

ABRÉGÉ

Les ensembles de données complexes nécessitent une compréhension approfondie de la question de recherche et du domaine de fond et, dans certains cas, peuvent nécessiter des analyses tenant compte des nombreuses caractéristiques de la population d'intérêt. Prenons par exemple la population de personnes atteintes de polyarthrite rhumatoïde (RA) en Écosse, qui sont suivies au fil du temps en commençant par un diagnostic de polyarthrite rhumatoïde, puis tous les six mois par la suite (étude Scottish Early Rheumatoid Arthritis ou SERA). Les résultats d'intérêt sont souvent discrets et peuvent être une variable binomiale (comme si le patient est entré en rémission) ou un décompte (comme le nombre d'articulations enflées). Les cliniciens savent que les traitements de la PR varient dans leur efficacité dans la population, ce qui suggère qu'il peut y avoir des sous-groupes distincts au sein de la population globale. L'analyse doit donc prendre en compte plusieurs caractéristiques clés: (i) le caractère discret du résultat, (ii) la nature longitudinale des mesures, et (iii) la possibilité de sous-populations à travers lesquelles les effets du traitement et des covariables diffèrent. Dans ma thèse, je propose l'utilisation d'un mélange fini de modèles linéaires généralisés à effets mixtes (FinMix GLMM) comme approche analytique appropriée. Je développe d'abord le modèle FinMix GLMM et je dérive les estimations du maximum de vraisemblance. Ensuite, je développe une approche de vraisemblance pénalisée pour la sélection d'effets fixes et aléatoires dans FinMix GLMM. Des simulations sont fournies pour illustrer la performance d'échantillons

finis des estimations. Enfin, le FinMix GLMM est appliqué (avec et sans pénalisation) au jeu de données SERA.

TABLE OF CONTENTS

Page

LIST OF TABLES

Page

Page

Page

Page

Page

Page

LIST OF FIGURES

Figure | Page

Page

Page

Page

Page

LIST OF ALGORITHMS

LIST OF THEOREMS

LIST OF ACRONYMS AND ABBREVIATIONS

AIC Akaike Information Criterion. 18, 19, 174

ALASSO Adaptive Least Absolute Shrinkage and Selection Operator. 7, 18, 20–23, 54, 55, 57–59, 62, 66, 71, 76, 77, 91, 92, 95, 96, 100, 104, 111, 155, 161, 198, 201, 202, 205, 206, 312

BIC Bayesian Information Criterion. 19, 75, 77, 88, 89, 91, 92, 107, 131–133, 135, 173–175, 179–187, 314

BMI Body Mass Index. 4, 81, 82

CPP Cyclic Citrullinated Peptide. 83, 91, 93–98

CRP C-Reactive Protein. 82, 85

DAS28 Disease Activity Score on 28 Joints. 4, 81, 82, 85

DMARD Disease-Modifying Antirheumatic Drugs. 3, 79, 80

DNA Deoxyribonucleic Acid. 3

EM Expectation-Maximisation. 7, 14–16, 23, 32, 33, 40, 71, 88, 103, 315

ESR Erythrocyte Sedimentation Rate. 82, 85

FinMix GLMM Finite Mixture of Generalised Linear Mixed-Effect Model. 1, 5–7, 9, 12, 14, 15, 19, 23–28, 30, 34, 37, 40, 49–51, 54, 55, 57, 77, 81, 82, 88, 92, 99–107, 116, 130–133, 136–138, 144–146, 148, 155, 158, 159, 161, 173, 312

FMLME Finite Mixture of Linear Mixed-Effect. 10, 18, 27, 99

GEE Generalised Estimating Equations. 10, 11

GLM Generalised Linear Model. 4, 6, 9, 10, 15, 22, 23, 99, 106

GLMM Generalised Linear Mixed-Effect Model. 1, 4–7, 10, 11, 13, 15, 18, 19, 23, 24, 27, 28, 57, 88, 99, 101, 103, 106, 314, 317

HDL High-Density Lipoproteins. 82, 85, 91, 93–98

HIV Human Immunodeficiency Viruses. 3, 80

LARS Least-Angle Regression. 105

LASSO Least Absolute Shrinkage and Selection Operator. 7, 18–23, 54, 55, 57–59, 62, 66, 76, 77, 91–94, 100, 104, 107, 155, 161, 198–200, 203, 204, 312

MCEM Monte Carlo Expectation-Maximisation. 16–18, 23, 25, 31, 33, 34, 36, 51, 61, 62, 77, 99, 130, 133

MCMC Markov Chain Monte Carlo. 103

MICE Multivariate Imputation by Chained Equations. 83

MLE Maximum Likelihood Estimation. 14, 15, 17, 20, 23, 25, 30, 33, 34, 37, 41, 42, 49–51, 58, 65, 66, 71, 76, 88–91, 99, 100, 117, 130, 131, 134, 136, 137, 142, 144, 145, 173, 207, 312, 315

MPLE Maximum penalised Likelihood Estimation. 61, 67, 71–74, 93–98, 100, 130, 132, 133, 135, 136, 148, 155, 157–159, 312

MRI Magnetic Resonance Imaging. 22

MSE Mean Squared Error. 42–53, 75, 76, 100, 175–179, 188–197, 199–311

NSAID Nonsteroidal Anti-Inflammatory Drug. 79

RA Rheumatoid Arthritis. 1–3, 6, 13, 79–82, 90, 93, 180

SCAD Smoothly Clipped Absolute Deviation. 7, 18, 21–23, 55, 57, 59, 62, 66, 77, 91, 92, 97, 98, 100, 104, 105, 109, 155, 161, 198, 312

SERA Scottish Early Rheumatoid Arthritis Inception Cohort and Biobank. 2, 3, 5, 6, 80, 81, 86, 89, 180

T2T Treat to Target. 2, 79

TNFi Tumour Necrosis Factor Inhibitors. 3, 79

VAS Visual Analogue Scale. 82, 85

WCC White Cell Count. 82, 91, 93–98

CHAPTER 1
Introduction

1.1 Motivation

Complex datasets necessitate a thorough understanding of the research question, and substantive area, and in some cases may require analyses that take into account the many features of the population of interest. The increased use of electronic medical records has resulted in an increased gathering of data, requiring more complex models that take into account the nature of the data. Consider, for example, the population of people in Scotland with Rheumatoid Arthritis (RA) where a sample of these patients were followed over time, starting with a diagnosis of RA, and then every six months thereafter. Some of the outcomes of interest were discrete, and possibly a binomial variable (such as if the patient has entered remission) or a count (such as the number of swollen joints). Clinicians are aware that not all patients with RA present the same symptoms, and that treatments for RA vary in their effectiveness across the population, suggesting that there may be distinct subgroups within the overall population. Thus there are several key features that the analysis must take into account: (i) the discrete nature of the outcome, (ii) the longitudinal nature of the measurements, and (iii) the possibility of subpopulations across which treatment, and covariate effects differ. In my thesis, I propose the use of a Finite Mixture of Generalised Linear Mixed-Effect Model (FinMix GLMM) as an appropriate analytic approach. The combination of a finite number of Generalised Linear Mixed-Effect

Model (GLMM) allows for modelling both heterogeneity and correlation present in the data. A finite mixture captures the heterogeneity in the population and using a mixed-effect model handles the correlation induced by longitudinal data. This analysis would be a natural choice to study data collected in the Scottish Early Rheumatoid Arthritis Inception Cohort and Biobank (SERA), administered by the Scottish Collaborative Arthritis Research network. This longitudinal study contains many patients, and includes data both from questionnaires and blood samples, providing a large number of covariates to consider. A more in-depth description of the dataset and model follows.

1.1.1 Motivating Example

RA is an auto-immune disease that affects many people. This chronic disorder causes inflammation in the joints of patients, typically starting in the hands and feet. RA affects the lining of the joints, which leads to swelling, and eventually bone erosion, and even joint deformity. As the disease progresses, it affects other joints, usually the elbows, ankles, knees, shoulders, and hips. RA should not be confused with osteoarthritis, which is more common. The wearing away of cartilage causes osteoarthritis, whereas inflammation of the synovial membrane causes RA.

Unfortunately, there is currently no cure for RA. Treatment aims to reduce inflammation to manage pain, and slow or prevent joint damage (Guo et al., 2018). Experts currently favour a treatment program referred to as Treat to Target (T2T). The goal of T2T is to aggressively treat the patient to achieve either remission or a minimal level of disease activity. This minimises the symptoms that the patient encounters. Medication is a favoured form of treatment, and there are many types of drugs used to treat RA. These medications include, but are not limited too, non-steroidal anti-inflammatory drugs, steroids (often corticosteroids), disease-modifying anti-rheumatic drugs, immunosuppressants, and Tumour Necrosis Factor Inhibitors (TNFi). Because of the reliance on pharmacotherapy, drug toxicity has become an important adverse outcome of interest in addition to other detrimental side effects. In advanced stages, patients with RA may require surgery (such as joint replacement).

The SERA cohort contains patients from Scotland that have a diagnosis of RA. It is rich in many variables, and it has enrolled numerous patients. The goal of the study is to be able to accurately predict patient outcomes so that physicians can apply the best course of treatment. In addition to the demographic and questionnaire data being gathered bank of tissue, and blood samples were also collected to allow for analysis of Deoxyribonucleic Acid (DNA) or biomarkers in the future. Sixteen hospitals from around Scotland participated in this study.

During the first six months, the cohort enrolled 489 patients. Overall, recruitment resulted in a cohort of over 1100 patients in the study. In addition to patients with RA, recruitment also included several controls, which I did not consider in my analysis. In order to be included in the cohort, all patients must have had a new clinical diagnosis of RA or undifferentiated polyarthritis. Patients were excluded from the cohort if they had already been on Disease-Modifying Antirheumatic Drugs (DMARD) therapy for a time period greater than six months, had another rheumatological diagnosis, had Hepatitis B, had Hepatitis C, or were Human Immunodeficiency Viruses (HIV) positive. Patient data was collected at baseline, and every six months thereafter on demographic characteristics, employment status, clinical measurements, laboratory results, and radiographic findings. Possible outcomes of interest include clinical remission (defined as Disease Activity Score on 28 Joints (DAS28) less than 2.6), swollen joint count, and drug toxicity. Of particular interest as covariates are variables including erosion at presentation, Body Mass Index (BMI), age, and alcohol intake.

1.2 Generalised Linear Mixed-Effect Models

When observations are discrete or categorical, using a linear model (where observations are assumed to follow a Gaussian distribution) is not appropriate and a Generalised Linear Model (GLM) may be a more appropriate model. A few assumptions are required when using a GLM: (i) observations are independent, (ii) the mean of the observations is associated with a linear function of some covariates through a link function, and (iii) the variance of the observations is a function of the mean of the observations. Attention typically focuses on the binomial, and Poisson cases, although the theory allows for outcomes from any distribution from an exponential family. In practice, there are cases where the first assumption is violated, and the observations are not statistically independent. If many of the observations come from the same person, the same geographical area, or related individuals, they may be correlated. In this case, a GLMM is a natural extension.

GLMMs represent an important class of models for regression analysis of discrete longitudinal data. In longitudinal data, it is unrealistic to assume that repeated observations from an individual are independent as they come from the same person. Therefore, each of the n individuals in the dataset must be considered individually, and a scalar or vector random effect is incorporated to include the correlation induced by the longitudinal nature of the data.

1.3 Finite Mixture Models

One challenge that can arise in the analysis of data is that the sample may be drawn from a population with significant underlying heterogeneity that can be more accurately described as a combination of distinct subpopulations. Often this underlying heterogeneity is unobserved, which makes it difficult to capture using a single GLMM. Indeed, such a model may not adequately represent the effect of a covariate on the outcome if that covariate has different effects across the subpopulations. It is possible that different covariates will have various effects in different (possibly unobserved) subpopulations. Finite mixtures of GLMMs provide a natural way to model unobserved heterogeneity in such populations. When taking this approach, the population is separated into subpopulations, and a distinct GLMM is used for each subpopulation. In this research, I assumed that each individual comes from a particular subpopulation. Because the covariates for an individual can vary between visits, the expected mean for that individual can also differ between visits, however, the fixed and random effect parameters do not.

In the initial stages of a study, it is common to introduce or consider a large number of covariates. However, some of these covariates may not be associated with the response variable and those that are could have different associations depending on the subpopulation of the FinMix GLMM. That is, the impact of covariates may differ across the subpopulations. Identifying the important effects in the model, both in the random and fixed components, requires a reliable means of variable selection. It is on this topic that I focus on in this research. In the SERA cohort, different subpopulations may exist that are defined by genotypic or demographic variables including age, age at onset, and sex.

Because of the heterogeneity in response to treatment for RA, a mixture of models is appropriate for this application. As the SERA data are longitudinal, a mixed-effect model must be used. When the outcome is not continuous, a GLMM is the most reasonable choice for each subpopulation, which leads to a global model that is a FinMix GLMM.

Previous research considered this problem in the linear case, and it is natural to then extend to the GLM case. As of yet, this has not been done for any such discrete outcomes, and many of the results in the linear model are more complicated in the GLM case. This is often because of the link function, which is the identity function in linear regression.

I organised the remainder of this thesis as follows. Chapter 2 contains a literature review, Chapter 3 describes the maximum likelihood case, Chapter 4 extends to the penalised maximum likelihood case to perform model selection, Chapter 5 considers the SERA data in detail, and Chapter 6 concludes. Appendices A to J contain additional and supplementary information.

CHAPTER 2
Literature Review

This literature review contains three sections. The first section considers several models that relate to a Finite Mixture of Generalised Linear Mixed-Effect Model (FinMix GLMM), including Generalised Linear Mixed-Effect Model (GLMM)s as well as finite mixtures. Next, I considered a few similar models, and discussed how each of these differs from a FinMix GLMM. The second section concentrates on computational methods such as Expectation-Maximisation (EM), Newton-Raphson, and rejection sampling which I used for estimation of parameters in a FinMix GLMM in Chapters 3, and 4. The last section of the literature review focuses on variable selection techniques such as Least Absolute Shrinkage and Selection Operator (LASSO), Adaptive Least Absolute Shrinkage and Selection Operator (ALASSO), and Smoothly Clipped Absolute Deviation (SCAD).

Throughout this chapter, I used the following notation. Assume a set of n independent observations $(\mathbf{y}_i, \mathbb{X}_i, \mathbb{Z}_i)$, $i = 1, 2, \ldots, n$ where $\mathbf{y}_i = (y_{i1}, y_{i2}, \ldots, y_{in_i})^\top$, and $\mathbb{X}_i$ is an $n_i \times p$ matrix of fixed covariates. The variable n_i represents the number of replicates for the subject i. The matrix $\mathbb{Z}_i$ is comprised of $q \leq p$ columns of $\mathbb{X}_i$, that is $\mathbb{Z}_i$ is also a known matrix of covariates. In a FinMix GLMM, the covariates $\mathbb{X}_i$ are associated with fixed effects, and the covariates $\mathbb{Z}_i$ are associated with random effects. When the jth replicate for subject i is considered, the random variable Y_{ij} is used along with the corresponding vectors $\mathbf{x}_{ij}$, and $\mathbf{z}_{ij}$. Note that $\mathbb{X}_i$ always contains a column of ones as the first column in order to capture the intercept in the model, so $\mathbf{x}_{ij} = (x_{ij1}, x_{ij2}, x_{ij3}, \ldots, x_{ijp}) = (1, x_{ij2}, x_{ij3}, \ldots, x_{ijp})$. Consistent with the mixed-effect model literature, I assumed that all variables for which there is a random effect, there is also a fixed effect.

Next, I considered the unknown parameters. I assumed the population contains K distinct, and homogeneous subpopulations with $\boldsymbol{\pi} = (\pi_1, \pi_2, \ldots, \pi_{K-1})^\top$, the vector of so-called mixing proportions. Each π_k represents the proportion of the population contained in each of the subpopulations, $\pi_k > 0 \forall k$, and $\sum_{k=1}^{K} \pi_k = 1$. Separate regression parameters describe the relationship between the covariates, and the outcome in each of these subpopulations. The vector of regression coefficients for the fixed effects in subpopulation k is represented by $\boldsymbol{\beta}_k = (\beta_{k1}, \beta_{k2}, \beta_{k3}, \ldots, \beta_{kp})^\top$. Let $\mathbb{\Gamma}_k$ be a $q \times q$ lower triangle matrix and assume that the variable $\mathbf{b}_i$ follows a q dimensional multivariate standard Gaussian distribution. Then $\mathbb{\Gamma}_k \mathbf{b}_i$ is a vector of random effect which follows a q dimensional multivariate Gaussian distribution with mean zero, and variance-covariance matrix $\mathbb{D}_k = \mathbb{\Gamma}_k \mathbb{\Gamma}_k^\top$, and $\mathbb{\Gamma}_k$ is the Cholesky decomposition of $\mathbb{D}_k$. The vector containing just the lower triangle values of $\mathbb{\Gamma}_k$ is $\mathbb{\Gamma}_k^*$. Let $\boldsymbol{\theta}_k = (\boldsymbol{\beta}_k^\top, \mathbb{\Gamma}_k^{*\top})^\top$ be the vector of regression coefficients for both fixed, and random effects for subpopulation k. I grouped all of the relevant parameters into one vector $\boldsymbol{\theta} = (\pi_1, \pi_2, \ldots, \pi_{(K-1)}, \boldsymbol{\beta}_1^\top, \mathbb{\Gamma}_1^{*\top}, \boldsymbol{\beta}_2^\top, \mathbb{\Gamma}_2^{*\top}, \ldots, \boldsymbol{\beta}_K^\top, \mathbb{\Gamma}_K^{*\top})$. To facilitate variable selection in Chapter 4, I decomposed $\mathbb{\Gamma}_k$ into $\mathbb{\Gamma}_k = \mathbb{d}_k \mathbb{C}_k$ where $\mathbb{d}_k$ is a diagonal matrix, and $\mathbb{C}_k$ is a lower triangle matrix with 1s along the diagonal. Again, $\mathbb{d}_k^*$, and $\mathbb{C}_k^*$ are the vectorised version of their respective parameters, and $\underline{\boldsymbol{\theta}}_k = (\boldsymbol{\beta}_k^\top, \mathbb{d}_k^{*\top}, \mathbb{C}_k^{*\top})^\top$. As I have established the notation, the probability density function follows.

In a FinMix GLMM with K subpopulations, the conditional density function of the random variable $\mathbf{Y}_i|(\mathbb{X}_i, \mathbb{Z}_i, \mathbf{\Theta})$ at the realisation $\mathbf{y}_i$ is

$$\begin{aligned}
f_{\mathbf{y}_i}(\mathbf{y}_i|\mathbb{X}_i, \mathbb{Z}_i, \mathbf{\Theta}) &= \sum_{k=1}^{K} \pi_k f^{(k)}_{\mathbf{y}_i}(\mathbf{y}_i|\mathbb{X}_i, \mathbb{Z}_i, \boldsymbol{\beta}_k, \mathbb{\Gamma}_k) \\
f^{(k)}_{\mathbf{y}_i}(\mathbf{y}_i|\mathbb{X}_i, \mathbb{Z}_i, \boldsymbol{\beta}_k, \mathbb{\Gamma}_k) &= \int f^{(k)}_{\mathbf{y}_i|\mathbf{b}_i}(\mathbf{y}_i|\mathbf{b}_i, \mathbb{X}_i, \mathbb{Z}_i, \boldsymbol{\beta}_k, \mathbb{\Gamma}_k) \times f_{\mathbf{b}_i}(\mathbf{b}_i) d\mathbf{b}_i \\
&= \int \prod_{j=1}^{n_i} f^{(k)}_{y_{ij}|\mathbf{b}_i}(y_{ij}|\mathbf{b}_i, \mathbf{x}_{ij}, \mathbf{z}_{ij}, \boldsymbol{\beta}_k, \mathbb{\Gamma}_k) \times f_{\mathbf{b}_i}(\mathbf{b}_i) d\mathbf{b}_i. \quad (2.1)
\end{aligned}$$

I assumed that $Y_{ij}|(\mathbf{b}_i, \mathbf{x}_{ij}, \mathbf{z}_{ij}, \boldsymbol{\beta}_k, \mathbb{\Gamma}_k)$ follows an exponential family, and used an appropriate link function (g), specifically, the canonical link function. If just the kth subpopulation is considered, then $E[Y_{ij}|\mathbf{b}_i, \mathbf{x}_{ij}, \boldsymbol{\beta}_k, \mathbb{\Gamma}_k] = g(\mathbf{x}_{ij}\boldsymbol{\beta}_k + \mathbf{z}_{ij}\mathbb{\Gamma}_k\mathbf{b}_i)$. Combining all of the subpopulations, $E[Y_{ij}|\mathbf{b}_i, \mathbf{x}_{ij}, \boldsymbol{\theta}] = \sum_{k=1}^{K} \pi_k g(\mathbf{x}_{ij}\boldsymbol{\beta}_k + \mathbf{z}_{ij}\mathbb{\Gamma}_k\mathbf{b}_i)$.

2.1 Models That Relate to Finite Mixtures of Generalised Linear Mixed-Effect Models

A FinMix GLMM is an extension of several previously proposed models. A classic regression model is a linear regression (Montgomery et al., 2012). Using the previous notation, in a linear regression model, $K = 1$, $n_i = 1 \forall i$, g is the identity function, and $E[Y_i|\mathbf{x}_i, \boldsymbol{\beta}] = \mathbf{x}_i\boldsymbol{\beta}$. If the assumption that g is the identity function is relaxed, but $K = 1$, $n_i = 1 \forall i$, and $\mathbf{Y}_i|x_i, \boldsymbol{\beta}$ follows an exponential family, a Generalised Linear Model (GLM) (Nelder and Wedderburn, 1972; McCulloch, 2000) is produced, and $E[Y_i|\mathbf{x}_i, \boldsymbol{\beta}] = g(\mathbf{x}_i\boldsymbol{\beta})$. Consider next the possibility that $n_i > 1$, which allows for longitudinal data, and random effects by relaxing the assumption that all of the observations are independent. If the identity function is used for g, the outcome follows a Gaussian distribution, and group (or individual) specific effects are introduced then

$E[Y_{ij}|\mathbf{b}_i, \mathbf{x}_{ij}, \boldsymbol{\beta}, \mathbb{\Gamma}] = \mathbf{x}_{ij}\boldsymbol{\beta} + \mathbf{z}_{ij}\mathbb{\Gamma}\mathbf{b}_i$, which is a linear mixed-effects model (Jiang, 2007, Chapters 1–2). The inclusion of random effects in GLMs leads to a GLMM, and is equivalent to extending a linear mixed-effects model to link functions other than the identity function, and thus, outcomes follow an exponential family (Jiang, 2007, Chapters 3–4). Thus, for a GLMM, $E[Y_{ij}|\mathbf{b}_i, \mathbf{x}_{ij}, \boldsymbol{\beta}, \mathbb{\Gamma}] = g(\mathbf{x}_{ij}\boldsymbol{\beta} + \mathbf{z}_{ij}\mathbb{\Gamma}\mathbf{b}_i)$.

Many regression models have also been explored in the finite mixture literature, starting with Pearson (1894). McLachlan and Peel (2000), and McLachlan et al. (2019) provide an overview of finite mixtures of regression models. In a finite mixture model, a weighted average is used to combine multiple models. In a finite mixture of linear models, $E[Y_{ij}|\mathbf{x}_{ij}, \boldsymbol{\theta}]$ can be expressed as $\sum_{k=1}^{K} \pi_k(\mathbf{x}_{ij}\boldsymbol{\beta}_k)$. In a finite mixture of GLMs with K subpopulations, the conditional density function of $\mathbf{Y}_i|(\mathbb{X}_i, \mathbb{Z}_i, \boldsymbol{\theta})$ is $f_{\mathbf{y}_i}(\mathbf{y}_i|\mathbb{X}_i, \mathbb{Z}_i, \boldsymbol{\theta}) = \sum_{k=1}^{K} \pi_k f_{\mathbf{y}_i}^{(k)}(\mathbf{y}_i|\mathbb{X}_i, \mathbb{Z}_i, \boldsymbol{\beta}_k)$. I assumed that $Y_{ij}|(\mathbf{x}_{ij}, \mathbf{z}_{ij}, \boldsymbol{\beta}_k)$ follows an exponential family, and used an appropriate link function (g) such that $E[Y_{ij}|, \mathbf{x}_{ij}, \boldsymbol{\beta}_k,] = g(\mathbf{x}_{ij}\boldsymbol{\beta}_k)$. Combining each of these subpopulations yields $E[Y_{ij}|\mathbf{b}_i, \mathbf{x}_{ij}, \boldsymbol{\theta}] = \sum_{k=1}^{K} \pi_k g(\mathbf{x}_{ij}\boldsymbol{\beta}_k)$.

If several heterogeneous subpopulations are evident within a population where a linear mixed-effects model would be appropriate for each subpopulation, a Finite Mixture of Linear Mixed-Effect (FMLME) can be used (Du et al., 2013). In this case, $E[Y_{ij}|\mathbf{b}_i, \mathbf{x}_{ij}, \boldsymbol{\Theta}] = \sum_{k=1}^{K} \pi_k(\mathbf{x}_{ij}\boldsymbol{\beta}_k + \mathbf{z}_{ij}\mathbb{\Gamma}_k\mathbf{b}_i)$. In this thesis, I extend Du's model to allow for not only Gaussian data but data drawn from any exponential family, a significant generalisation.

Another popular approach to the analysis of longitudinal data is to use Generalised Estimating Equations (GEE) (Zeger et al., 1988). This approach differs from modelling using a linear mixed-effect model or GLMM as this alternative approach serves to estimate population average parameters; Jiang (2007, Section 4.2) contains more details. I did not pursue a GEE approach in this thesis, I specifically focused this research on patient-specific effects of treatments. However, because the two approaches are used in similar situations, comparisons between the two approaches have been discussed in the literature, such as by Evans et al. (2001), Localio et al. (2006), Gardiner et al. (2009), and Zhang et al. (2012). Additionally, Feng et al. (1996) explored other approaches to the analysis of longitudinal data.

Both linear mixed-effect, and GLMMs have been used in medical and epidemiological applications. Examples include estimation of the effects of changes in haemoglobin on the recovery from malaria (Sagara et al., 2014), modelling childhood ailments in Bolivia (Solis-Soto et al., 2013), and modelling of cardiovascular disease in Indigenous Americans (Chen et al., 2014). Other examples include studying outcomes following breast reconstruction modelled via a GLMM with patient-specific intercepts (Yuen et al., 2014), modelling fatty acid intake of adults in the Minneapolis metropolitan area using random effects for neighbourhoods (Honors et al., 2014), and studying malaria transmission using random effects at both the individual, and school level (Okebe et al., 2014).

Finite mixtures have also been used in medical research for several applications including modelling mixtures in borderline personality disorder using phenotypes (Hallquist and Pilkonis, 2012), detecting a binary trait locus (Deng et al., 2006), and predicting health care costs (Rein, 2005). Other examples include detecting tropical infectious diseases in Kenya (Fujii et al., 2014), modelling pollutant and exposure data (Li et al., 2013), and modelling the effects of five adolescent risk factors on the total number of sexual partners in adulthood (Lanza et al., 2011).

It should be noted that some of these examples describe their model as a mixed-effect model with a categorical latent variable. In my research, I have assumed that the latent variable in a mixed-effect model follows a Gaussian distribution, and is continuous. Tutz and Oelker (2017) provides a review of both types of models.

Before moving on to computational methods, consider a model similar to a FinMix GLMM.

2.1.1 Other Similar Models

In 2005, Hall, and Wang proposed a model that bears many similarities to a FinMix GLMM. The major difference between the research undertaken in the coming chapters, and the work of Hall and Wang (2005) is that the subpopulation membership is structured differently. In the model discussed in this thesis, I assumed that all of the repeated measures from a particular subject come from the same subpopulation. This is suitable for a situation where the repeated measures are from one person, and each person is a member of exactly 1 of the K subpopulations. In contrast, Hall and Wang (2005) consider settings where population membership is not constant, but rather an individual may belong to different populations at different measurement points. The example provided in Hall and Wang (2005) has cities as the unit of analysis, and the two subpopulations denote a disease outbreak or no disease outbreak. In this setting, allowing population membership to fluctuate with time is reasonable however in many conventional settings where, say, human health is concerned, study units are unlikely to switch subpopulations over time. For example, in Chapter 5, I considered modelling Rheumatoid Arthritis (RA) patient outcomes. While outcomes are expected to vary with time, the relationship between patient characteristics and outcomes are likely stable over time, so fixing population membership in time is the more reasonable model choice.

In addition, Hall, and Wang restricted their model to consider only $K = 2$. In contrast, the model presented over the following chapters is more general, with $K \geq 1$. Comparing the likelihood functions of the two models illustrates these differences.

In Hall and Wang (2005), with two subpopulations, the conditional density function of $\mathbf{Y}_i|(\mathbb{X}_i, \mathbb{Z}_i, \boldsymbol{\theta})$ is

$$\begin{aligned}
f_{\mathbf{y}_i}(\mathbf{y}_i|\mathbb{X}_i, \mathbb{Z}_i, \boldsymbol{\theta}) &= \int f_{\mathbf{y}_i|\mathbf{b}_i}(\mathbf{y}_i|\mathbf{b}_i, \mathbb{X}_i, \mathbb{Z}_i, \boldsymbol{\theta}) \times f_{\mathbf{b}_i}(\mathbf{b}_i) d\mathbf{b}_i \\
&= \int \prod_{j=1}^{n_i} f_{y_{ij}|\mathbf{b}_i}(y_{ij}|\mathbf{b}_i, \mathbf{x}_{ij}, \mathbf{z}_{ij}, \boldsymbol{\theta}) \times f_{\mathbf{b}_i}(\mathbf{b}_i) d\mathbf{b}_i \\
&= \int \prod_{j=1}^{n_i} f_{y_{ij}|\mathbf{b}_i}(y_{ij}|\mathbf{b}_i, \mathbf{x}_{ij}, \mathbf{z}_{ij}, \boldsymbol{\theta}) \times f_{\mathbf{b}_i}(\mathbf{b}_i) d\mathbf{b}_i \\
&= \int \prod_{j=1}^{n_i} [\pi_1 f_{y_{ij}|\mathbf{b}_i}(y_{ij}|\mathbf{b}_i, \mathbf{x}_{ij}, \mathbf{z}_{ij}, \boldsymbol{\beta}_1, \mathbb{\Gamma}_1) + \\
&\quad (1 - \pi_1) f_{y_{ij}|\mathbf{b}_i}(y_{ij}|\mathbf{b}_i, \mathbf{x}_{ij}, \mathbf{z}_{ij}, \boldsymbol{\beta}_2, \mathbb{\Gamma}_2)] \times f_{\mathbf{b}_i}(\mathbf{b}_i) d\mathbf{b}_i
\end{aligned}$$

Comparing this likelihood to (2.1), the variable π_1 is inside the integral allowing for a subject to belong to one subpopulation at one specific time point, and a different subpopulation at a later time point. While this is useful in certain situations, it is not a finite mixture model. The work in this thesis extends GLMMs by applying a GLMM to each subpopulation. Because I assume that all of the observations from a given subject are correlated, it is logical that all of those observations would come from the same subpopulation.

Other papers have shown similar special cases, but have not provided a general approach to FinMix GLMMs. Wang et al. (2002) shows a mixture of two or three Poisson regressions, but as with Hall and Wang (2005), the measures from a given subject can come from any of the subpopulations. Classification is the focus of Grun and Leisch (2008), they assume the outcome follows a multinomial logit distribution. In addition, they assumed that fixed effects are the same across all subpopulations, and only the random effects differ across subpopulations, with special attention given to the case with a random effect on the intercept. Dunson (2000) take a Bayesian approach to consider clustered mixed outcomes. While some of the models discussed in Dunson (2000) are similar to a FinMix GLMM, the majority of the models are different, and Dunson (2000) used a Bayesian approach rather than frequentist paradigm.

2.2 Computational Algorithm

Next, consider the computational tools that I used for estimation of the parameters in a FinMix GLMM. I implemented a Maximum Likelihood Estimation (MLE) approach, but there is no closed-form solution for the estimates in a FinMix GLMM.

The EM algorithm is ubiquitous in statistics. As noted by Dempster et al. (1977), Rao (1955) explored a special case of the EM algorithm for a multinomial outcome. While such special cases were published earlier, the general idea of the EM algorithm was first introduced by Dempster et al. (1977), and explored in McLachlan and Krishnan (2008). The motivation behind EM was to expand the possible settings in which one could compute an MLE using an iterative algorithm. Each iteration consists of two steps, an expectation step (or E-step) followed by a maximisation step (or M-step). In EM, one starts with incomplete data, and then estimates unobserved values, for example, nuisance parameters, or latent values, to form the complete data. Given the complete data, the MLE may be easier to compute. Because of the popularity of EM, it continues to be discussed in review articles such as Meng and van Dyk (1997), and Lange et al. (2014).

The EM algorithm is a popular computational choice in the literature for both GLMMs, and finite mixtures. Because one can view subpopulation membership as a missing variable, EM is a natural fit for finite mixture models, and Jansen (1993) explore EM for finite mixtures of GLMs. Similarly, the random effects in a GLMM can also be viewed as missing data and much has been written on the use of EM for mixed-effects models as well. For instance, Anderson and Hinde (1988) use EM for GLMMs, Laird et al. (1987) looks at using EM for situations with repeated measures, and Steele (1996) shows modifications to ease computation in the E-step when a GLMM is being used. Lindstrom and Bates (1988) provided a comparison of EM to Newton-Raphson in linear mixed-effect models, and Meng and van Dyk (1998) proposed computationally improvements.

Rather than considering the most general formulation of EM, the focus here is on EM as applied to FinMix GLMMs. In this thesis, EM is used to estimate the probability that a subject belongs to each of K distinct subpopulations leading to the estimation of π_k, and then again to estimate the random effects which informs the estimation of $\boldsymbol{\beta}_k$, and $\mathbb{\Gamma}_k$.

I started with an initial value for the parameter vector $\mathbf{\Theta}$, denoted $\mathbf{\Theta}^{(0)}$. To estimate the mixing proportions, $\pi_1, \pi_2, \ldots, \pi_K$, I first estimated the membership probability of each individual in each of the subpopulations, and denoted them τ_{ki}. This required calculating the likelihood for each individual, and each subpopulation using (2.1) under $\mathbf{\Theta}^{(0)}$. I then calculated τ_{ki} for each $k \in \{1, 2, \ldots, K\}$, and $i \in \{1, 2, \ldots, n\}$ according to:

$$\begin{aligned}
\tau_{ki}^{(t)} &= \frac{\pi_k^{(t)} f_{\mathbf{y}_i}^{(k)}(\mathbf{y}_i | \mathbb{X}_i, \mathbb{Z}_i, \boldsymbol{\beta}_k^{(t)}, \mathbb{\Gamma}_k^{(t)})}{\sum_{h=1}^{K} \pi_h^{(t)} f_{\mathbf{y}_i}^{(h)}(\mathbf{y}_i | \mathbb{X}_i, \mathbb{Z}_i, \boldsymbol{\beta}_h^{(t)}, \mathbb{\Gamma}_h^{(t)})} \\
&= \frac{\pi_k^{(t)} \int f_{\mathbf{y}_i|\mathbf{b}_i}^{(k)}(\mathbf{y}_i | \mathbf{b}_i, \mathbb{X}_i, \mathbb{Z}_i, \boldsymbol{\beta}_k, \mathbb{\Gamma}_k) \times f_{\mathbf{b}_i}(\mathbf{b}_i) d\mathbf{b}_i}{\sum_{h=1}^{K} \pi_h^{(t)} \int f_{\mathbf{y}_i|\mathbf{b}_i}^{(h)}(\mathbf{Y}_i | \mathbf{b}_i, \mathbb{X}_i, \mathbb{Z}_i, \boldsymbol{\beta}_h, \mathbb{\Gamma}_h) \times f_{\mathbf{b}_i}(\mathbf{b}_i) d\mathbf{b}_i} \\
&= \frac{\pi_k^{(t)} \int \prod_{j=1}^{n_i} f_{\mathbf{y}_i|\mathbf{b}_i}^{(k)}(y_{ij} | \mathbf{b}_i, \mathbf{x}_{ij}, \mathbf{z}_{ij}, \boldsymbol{\beta}_k, \mathbb{\Gamma}_k) \times f_{\mathbf{b}_i}(\mathbf{b}_i) d\mathbf{b}_i}{\sum_{h=1}^{K} \pi_h^{(t)} \int \prod_{j=1}^{n_i} f_{\mathbf{y}_i|\mathbf{b}_i}^{(h)}(y_{ij} | \mathbf{b}_i, \mathbf{x}_{ij}, \mathbf{z}_{ij}, \boldsymbol{\beta}_h, \mathbb{\Gamma}_h) \times f_{\mathbf{b}_i}(\mathbf{b}_i) d\mathbf{b}_i}.
\end{aligned}$$

As there is no closed-form solution to these integrals, I calculated an approximation using Monte Carlo integration (Robert and Casella, 2010, Chapter 3). That is, I generated L values of $\mathbf{b}_i$ from a standard Gaussian distribution, and replaced the integral by a summation over these L values. Because I used Monte Carlo in the E-step, this algorithm can be more accurately described as a Monte Carlo Expectation-Maximisation (MCEM) (Wei and Tanner, 1990).

Next, to compute $\mathbb{\Gamma}_k$, I used another EM loop, or more specifically, another MCEM loop. Dempster et al. (1977) showed this approach for the computation of the variance-covariance matrix in a mixed-effects model, and Laird (1982) explored it further. Again, I generated L estimates of $\mathbf{b}_i$ for each $i \in \{1, 2, \ldots, n\}$. Note that I generated $\mathbf{b}_i$ in this MCEM from $\mathbf{b}_i | \mathbf{y}_i, \mathbb{X}_i, \mathbb{Z}_i, \boldsymbol{\beta}_k, \mathbb{\Gamma}_k$, not from a standard Gaussian distribution or indeed from any standard distribution. As such, I used rejection sampling (McLachlan and Krishnan, 2008, Chapter 6) to generate values for the E-step in this situation.

Following the work of Gilks and Wild (1992), and Robert and Casella (2010, Section 2.3), I chose rejection sampling to generate L estimates of $\mathbf{b}_i$, as done in Ibrahim et al. (1999). I generated potential $\mathbf{b}_i$s by calculating the MLE for $\mathbf{b}_i|\mathbf{y}_i, \mathbb{X}_i, \mathbb{Z}_i, \boldsymbol{\beta}_k, \mathbb{F}_k$, and the Hessian matrix at this point. I then generated points from a Gaussian distribution with the mean set to the MLE for $\mathbf{b}_i$, and the variance set to the negative of the inverse of the calculated Hessian matrix. These points are then the proposed values for a rejection sampler, which I repeated multiple times. These Metropolis iterations improve the sample to make it more representative of the posterior distribution. Points that are accepted become the L potential $\mathbf{b}_i$s used in the next step. Note that I generated L potential $\mathbf{b}_i$ values for each n in each of the k calls to the inner MCEM used to estimate the random effects. Section C.4 shows this algorithm in detail.

The M-step of the inner MCEM algorithm maximises the log-likelihood of each of the K subpopulations, represented by the function $Q_k(\boldsymbol{\theta}_k)$, with respect to $\boldsymbol{\beta}_k$, and $\mathbb{F}_k$ using Newton-Raphson. The log-likelihood for each subpopulation is

$$Q_k(\boldsymbol{\theta}_k) = \sum_{i=1}^{n} \tau_{ki} \times \log[f_{\mathbf{y}_i}(\mathbf{y}_i|\mathbb{X}_i, \mathbb{Z}_i, \boldsymbol{\theta}_k)].$$

Newton-Raphson was first discussed as a root-finding method, see (Ypma, 1995; Hazewinkel, 1988, p. 1231), but when used to find the root of the first derivative of a function, it finds local optima. The problem of optimisation is longstanding, and there are many possible solutions. In this thesis, I applied Newton-Raphson to maximise the approximate likelihood or penalised likelihood. This likelihood was weighted by the values of τ_{ki} to incorporate the finite mixture structure of the population. Given the approximate likelihood, along with its first, and second derivatives, Newton-Raphson successively updates estimates according to $\boldsymbol{\theta}_k^{(s+1)} = \boldsymbol{\theta}_k^{(s)} - [Q_k''(\boldsymbol{\theta}_k^{(s)})]^{-1} Q_k'(\boldsymbol{\theta}_k^{(s)})$. In this thesis, I implemented Newton-Raphson in the M-step of the inner MCEM loop.

2.3 Variable Selection With Penalisation

The issue of variable selection in regression is a difficult and important problem. Given a data set with even a modest number of variables, there are a large number of possible models to explore, and considering all of them is computationally expensive (Buhlmann and van de Geer, 2011, Chapter 1). Variable selection is not a new subject in GLMMs or finite mixtures. Within the GLMM literature, variable selection has been considered in a variety of situations, including with finite support random effects distributions (Leung and Elashoff, 1996b); Bayesian variable selection (Chen et al., 2003); bootstrap tests for determining if a variance component is non-zero (Sinha, 2009); selection of both fixed and random effects using LASSO, ALASSO, and SCAD (Ibrahim et al., 2011); a GLMM specific Akaike Information Criterion (AIC) (Yu et al., 2013); using an L_1 penalty (Groll and Tutz, 2014); penalised quasi-likelihood (Pan and Huang, 2014); and an algorithm called GLMMLasso (Schelldorfer et al., 2014). The issue of variable selection has been discussed in the literature for finite mixtures of regression models by Khalili and Chen (2007), FMLMEs by Du et al. (2013), and the high-dimensional case of finite mixtures of Gaussian models in Devijver (2015). In this thesis, I focused on three penalties for variable selection, LASSO, ALASSO, and SCAD.

Prior to the advent of penalisation variable selection techniques, common algorithms for variable selection included stepwise regression, and all subsets selection. When all subsets regression was performed, then regression models were compared based on information criteria as discussed in Nishii (1984). A number of information criteria have been suggested including AIC (Akaike, 1998), Bayesian Information Criterion (BIC) (Schwarz, 1978), and generalised information criterion (Pu and Niu, 2006). However, these algorithms may not result in the same model being chosen (Huo and Ni, 2007), and there may be a lack of stability in model selection (Breiman, 1996). Lavergne et al. (2008) provides a specific formulation for information criteria in GLMMs. Another possibility is predictive cross-validation as proposed by Braun et al. (2014).

Certain penalty functions posses a number of desirable properties, including the oracle property, and consistency. These properties are not guaranteed by stepwise regression or all subsets selection. Buhlmann and van de Geer (2011, Chapter 6) considered this idea further. In addition, much of the literature on variable selection in mixed-effects models has considered variable selection on only the fixed effects, and not the random effects (Schelldorfer et al., 2014; Groll and Tutz, 2014).

The LASSO, while not the only penalty function that has been used, is the most commonly used penalty function (Tibshirani, 1996, 2011). LASSO uses an $L1$ penalty on the likelihood function, which together is often called the penalised likelihood. In general, the LASSO penalty function takes the form $p_{\lambda_{nk}}(\boldsymbol{\theta}) = \lambda_{nk}|\boldsymbol{\theta}|$. In the case of a FinMix GLMM, the penalty, when only the fixed effects are being examined, is $p_{\lambda_{nk}}(\boldsymbol{\theta}_k) = \lambda_{nk}\sum_{j=1}^{p}|\beta_{kj}|$ with $\frac{\partial p_{\lambda_{nk}}(\boldsymbol{\theta}_k)}{\partial \beta_{kj}} = \lambda_{nk}\text{sign}(\beta_{kj})$. Similarly, if a penalty is applied to both the fixed, and random effects, the penalty becomes $p_{\lambda_{nk}}(\underline{\boldsymbol{\theta}}_k) = \lambda_{nk} \sum_{j=1}^{p} |\beta_{kj}| + \lambda_{nk} \sum_{j=1}^{q} |d_{kj}| = \lambda_{nk} \sum_{j=1}^{p} |\beta_{kj}| + \lambda_{nk} \sum_{j=1}^{q} d_{kj}$, and the partial derivatives are $\frac{\partial p_{\lambda_{nk}}(\underline{\boldsymbol{\theta}}_k)}{\partial \beta_{kj}} = \lambda_{nk} \times \text{sign}(\beta_{kj})$, and $\frac{\partial p_{\lambda_{nk}}(\underline{\boldsymbol{\theta}}_k)}{\partial d_{kj}} = \lambda_{nk}$. The second derivatives of the penalty are equal to zero, both with respect to β_{kj}, and d_{kj}. Note that using LASSO results in a biased estimate (Fan and Li, 2001).

Two major shortcomings of the LASSO penalty are that the penalty is of the same magnitude regardless of the size of the parameter, and that the derivative of the penalty function is not continuous. Adaptive LASSO or ALASSO provides a solution to the first of these problems. This is done by adding weights to the penalty function as explained in Zou (2006). The choice of the weights must be specified by the analyst and is commonly chosen to be the inverse of the maximum likelihood estimate as proposed by Zou (2006). This, however, requires the calculation of the MLE, and as such could be considered a two-step estimation procedure. The ALASSO penalty function for a general model is $p_{\lambda_{nk}}(\boldsymbol{\theta}) = \lambda_{nk} w |\boldsymbol{\theta}|$.

If only fixed effects are considered, then $p_{\lambda_{nk}}(\boldsymbol{\theta}_k) = \lambda_{nk} \sum_{j=1}^{p} w_j |\beta_{kj}|$ with $\frac{\partial p_{\lambda_{nk}}(\boldsymbol{\theta}_k)}{\partial \beta_{kj}} = \lambda_{nk} w_j \times \text{sign}(\beta_{kj})$. Similarly, if a penalty is applied to both the fixed, and random effects, the penalty becomes $p_{\lambda_{nk}}(\underline{\boldsymbol{\theta}}_k) = \lambda_{nk} \sum_{j=1}^{p} w_j |\beta_{kj}| + \lambda_{nk} \sum_{j=1}^{q} w_{p+j} |d_{kj}| = \lambda_{nk} \sum_{j=1}^{p} w_j |\beta_{kj}| + \lambda_{nk} \sum_{j=1}^{q} w_{p+j} d_{kj}$, and the partial derivatives are $\frac{\partial p_{\lambda_{nk}}(\underline{\boldsymbol{\theta}}_k)}{\partial \beta_{kj}} = \lambda_{nk} w_j \times \text{sign}(\beta_{kj})$, and $\frac{\partial p_{\lambda_{nk}}(\underline{\boldsymbol{\theta}}_k)}{\partial d_{kj}} = \lambda_{nk} w_{p+j}$. Again, the second derivatives of the penalty with respect to both β_{kj}, and d_{kj} are equal to zero.

As previously noted, the inverse of the MLE is the most common choice for the weights in ALASSO, and the weights I used in this research. However, the theory in Zou (2006) is general to $w = |\hat{\boldsymbol{\theta}}|^{-\alpha}$ where $\alpha > 0$, and $\hat{\boldsymbol{\theta}}$ is a root-n-consistent estimator for the true value of $\boldsymbol{\theta}$.

While ALASSO reduces the bias on large values of parameter estimates, it does not change the shape of the derivative of the penalty function. As such, the first derivative of the penalty function for ALASSO is not continuous.

Fan and Li (2001) described the SCAD penalty as a solution to the problems with previous penalty functions and is

$$\frac{\partial p_{\lambda_{nk}}(\underline{\boldsymbol{\theta}})}{\partial \underline{\boldsymbol{\theta}}} = \lambda_{nk} \left\{ I(|\underline{\boldsymbol{\theta}}| \leq \lambda_{nk}) + \frac{(a\lambda_{nk} - |\underline{\boldsymbol{\theta}}|)_+}{(a-1)\lambda_{nk}} I(|\underline{\boldsymbol{\theta}}| > \lambda_{nk}) \right\}.$$

If only fixed effects are considered then

$$\frac{\partial p_{\lambda_{nk}}(\boldsymbol{\theta}_k)}{\partial \beta_{kj}} = \lambda_{nk} \left\{ I(|\beta_{kj}| \leq \lambda_{nk}) + \frac{(a\lambda_{nk} - |\beta_{kj}|)_+}{(a-1)\lambda_{nk}} I(|\beta_{kj}| > \lambda_{nk}) \right\}.$$

In the case where both fixed, and random effects are penalised,

$$\frac{\partial p_{\lambda_{nk}}(\theta_k)}{\partial \beta_{kj}} = \lambda_{nk} \left\{ I(|\beta_{kj}| \leq \lambda_{nk}) + \frac{(a\lambda_{nk} - |\beta_{kj}|)_+}{(a-1)\lambda_{nk}} I(|\beta_{kj}| > \lambda_{nk}) \right\},$$

and

$$\frac{\partial p_{\lambda_{nk}}(\underline{\boldsymbol{\theta}}_k)}{\partial d_{kj}} = \lambda_{nk} \left\{ I(d_{kj} \leq \lambda_{nk}) + \frac{(a\lambda_{nk} - d_{kj})_+}{(a-1)\lambda_{nk}} I(d_{kj} > \lambda_{nk}) \right\}.$$

Note that $(t)_+ = t \times I(t > 0)$, and $a > 2$.

These penalty functions are widely used in medical research. The LASSO has been used to perform variable selection in many medical applications, ranging from oncology (Olk-Batz et al. (2011), Wu et al. (2011)) to neurology (Zhou et al. (2012), Baradaran et al. (2013)), and has seen considerable use in genetics analyses, where dimensionality is often very high (Ghosh and Chinnaiyan (2005), Usai et al. (2009), Olk-Batz et al. (2011), Wimmer et al. (2013)). The ALASSO penalty function has been applied to various statistical models in the past. In addition to the high-dimensional linear regression case (Huang et al., 2008), ALASSO has been applied to general transformation models with right-censored data (Li and Gu, 2012), varying-coefficient partially linear measurement error models (Wang et al., 2013), zero-inflated count data (Zeng et al., 2014), joint models of longitudinal, and survival outcomes (He et al., 2015), Poisson regression (Ivanoff et al., 2016), and nonlinear mixed-effect pharmacokinetic models (Haem et al., 2017). While ALASSO has not been as widely used as LASSO, it has also been used for gene selection, including cancer classification (Algamal and Lee, 2015). The SCAD penalty has not been as popular as LASSO but has been used in a variety of settings including geostatistics (Chu et al., 2011), and denoising images (Chopra and Lian, 2010). In the medical field, SCAD has been used for variable selection in applications including genetics (Lu et al., 2011), liver fibrosis (Yan et al., 2011), and Magnetic Resonance Imaging (MRI) (Mehranian et al., 2013).

In conclusion, many regression models have been developed for analysing data, starting with classical linear regression, and building upon this framework to GLM, and mixed-effect models. These extensions allow for accurate modelling of more complex data, or situations that violate the underlying assumptions associated with linear regression. In some cases, however, there is underlying heterogeneity in a population, and using the same model for the entire population is not appropriate. In these cases, a finite mixture of regression models could be a suitable choice. Next, I introduce the computational methods that I used in this thesis, precisely EM, MCEM, and Newton-Raphson. Finally, I reviewed a variety of penalties that can be used for variable selection, specifically LASSO, ALASSO, and SCAD.

While many regression models have been explored in the literature, as of yet, a FinMix GLMM has not been explored. While many simpler models, including GLM, GLMM, and finite mixtures of regression models have been considered, combining them to produce a FinMix GLMM has not been done, making this a gap in the literature that this research fills. In addition, the use of penalised likelihood for performing variable selection has been studied extensively for certain circumstances (namely linear regression), but it has not been explored as fully for many other regression situations. Therefore, the inclusion of variable selection in this thesis is another important contribution to the field. Having considered the relevant literature, the next chapter considers the formulation, and computation of the MLE.

CHAPTER 3
Objective One: Finite Mixtures of Generalised Linear Mixed-Effect Models and Their Maximum Likelihood Estimates

3.1 Introduction

In this chapter, I developed the Finite Mixture of Generalised Linear Mixed-Effect Model (FinMix GLMM). This type of model can be used to analyse data that is both longitudinal in nature, and includes underlying heterogeneity in the population. Using a mixed-effect model accounts for the correlation between repeated measures in the longitudinal data, while a finite mixture of models allows for the modelling of the underlying heterogeneity in the population of interest. I assumed that the subpopulation membership of each subject (also called a patient or cluster) in the population is unknown. In addition, all of the observations from a given subject are assumed to be from the same subpopulation. I assumed that the distribution of the outcome within a given subpopulation follows a Generalised Linear Mixed-Effect Model (GLMM) and is therefore from an exponential distribution. In this work, I excluded from consideration the Gaussian distribution, as this was explored previously in Du et al. (2013). Combing these two elements, that is GLMMs with finite mixtures, allows for modelling of more complex datasets, and permits consideration of additional sources of heterogeneity.

In Section 3.2 I formally define the FinMix GLMM and discuss its identifiability conditions. In Section 3.3, I derive the maximum likelihood estimate, and described computational aspects of an estimation procedure based on the Monte Carlo Expectation-Maximisation (MCEM) algorithm. I then explored the performance of the Maximum Likelihood Estimation (MLE) via simulation in Section 3.4, and Section 3.5 concludes.

3.2 The Finite Mixture of Generalised Linear Mixed-Effect Model

I established the notation in the previous chapter but have reiterated here. Identifiability of the model is also considered.

3.2.1 Model Definition

Consider a set of n independent observations $(\mathbf{y}_i, \mathbb{X}_i)$ where subject i has n_i replicates. For subject i, the vector $\mathbf{y}_i$ contains the outcomes, and the matrices $\mathbb{X}_i$, and $\mathbb{Z}_i$ contain covariates. In the FinMix GLMM, the covariates $\mathbb{X}_i$ are associated with fixed effects, the covariates $\mathbb{Z}_i$ are associated with random effects, and all variables for which there is a random effect, there is also a fixed effect.

Consider next the unknown parameters. The mixing proportions for the K subpopulations are $\boldsymbol{\pi} = (\pi_1, \pi_2, \ldots, \pi_{K-1})$ where $\sum_{k=1}^{K} \pi_k = 1$. Similarly, separate regression parameters $\boldsymbol{\beta}_k$ and $\mathbb{F}_k$ describe the relationship between the covariates and the outcome for each subpopulation. Recall that $\mathbf{b}_i$ follows a q dimensional multivariate standard Gaussian distribution, and the vectorised version of $\mathbb{F}_k$ is $\mathbb{F}_k^*$. All of the estimated parameters are grouped into one vector $\boldsymbol{\Theta} = (\boldsymbol{\pi}, \boldsymbol{\beta}_1^\top, \mathbb{F}_1^{*\top}, \boldsymbol{\beta}_2^\top, \mathbb{F}_2^{*\top}, \ldots, \boldsymbol{\beta}_K^\top, \mathbb{F}_K^{*\top})^\top$. The vector $\boldsymbol{\theta}_k = (\boldsymbol{\beta}_k^\top, \mathbb{F}_k^{*\top})^\top$ contains the regression parameters for a single subpopulation.

Recall that the conditional density function of the random variable $\mathbf{Y}_i|(\mathbb{X}_i, \mathbb{Z}_i, \boldsymbol{\Theta})$ in a FinMix GLMM with K subpopulations, at the realisation $\mathbf{y}_i$ is

$$\begin{aligned} f_{\mathbf{y}_i}(\mathbf{y}_i|\mathbb{X}_i, \mathbb{Z}_i, \boldsymbol{\Theta}) &= \sum_{k=1}^{K} \pi_k f_{\mathbf{y}_i}^{(k)}(\mathbf{y}_i|\mathbb{X}_i, \mathbb{Z}_i, \boldsymbol{\beta}_k, \Gamma_k) \\ f_{\mathbf{y}_i}^{(k)}(\mathbf{y}_i|\mathbb{X}_i, \mathbb{Z}_i, \boldsymbol{\beta}_k, \Gamma_k) &= \int f_{\mathbf{y}_i|\mathbf{b}_i}^{(k)}(\mathbf{y}_i|\mathbf{b}_i, \mathbb{X}_i, \mathbb{Z}_i, \boldsymbol{\beta}_k, \Gamma_k) \times f_{\mathbf{b}_i}(\mathbf{b}_i) d\mathbf{b}_i \\ &= \int \prod_{j=1}^{n_i} f_{y_{ij}|\mathbf{b}_i}^{(k)}(y_{ij}|\mathbf{b}_i, \mathbf{x}_{ij}, \mathbf{z}_{ij}, \boldsymbol{\beta}_k, \Gamma_k) \times f_{\mathbf{b}_i}(\mathbf{b}_i) d\mathbf{b}_i. \end{aligned}$$

Recall that $Y_{ij}|(\mathbf{b}_i, \mathbf{x}_{ij}, \mathbf{z}_{ij}, \boldsymbol{\beta}_k, \Gamma_k)$ follows an exponential family, with an appropriate link function (g), I used the canonical link function in this thesis. Considering just the k^{th} subpopulation, $E[y_{ij}|\mathbf{b}_i, \mathbf{x}_{ij}, \boldsymbol{\beta}_k, \Gamma_k] = g(\mathbf{x}_{ij}\boldsymbol{\beta}_k + \mathbf{z}_{ij}\Gamma_k\mathbf{b}_i)$, and combining all subpopulations, $E[y_{ij}|\mathbf{b}_i, \mathbf{x}_{ij}, \boldsymbol{\Theta}] = \sum_{k=1}^{K} \pi_k g(\mathbf{x}_{ij}\boldsymbol{\beta}_k + \mathbf{z}_{ij}\Gamma_k\mathbf{b}_i)$. I discussed estimation of the parameters in Section 3.3.

Next, consider two well known exponential family distributions for discrete data. In the Poisson case, $Y_{ij}|\mathbf{b}_i, \mathbf{x}_{ij}, \mathbf{z}_{ij}, \boldsymbol{\beta}_k, \Gamma_k \sim Poisson(\xi_{ij})$ where $\log(\xi_{ij}) = \mathbf{x}_{ij}\boldsymbol{\beta}_k + \mathbf{z}_{ij}\Gamma_k\mathbf{b}_i$, and in the binomial case, $Y_{ij}|\mathbf{b}_i, \mathbf{x}_{ij}, \mathbf{z}_{ij}, \boldsymbol{\beta}_k, \Gamma_k, m_{ij} \sim binomial(m_{ij}, \varphi_{ij})$ where $\text{logit}(\varphi_{ij}) = \mathbf{x}_{ij}\boldsymbol{\beta}_k + \mathbf{z}_{ij}\Gamma_k\mathbf{b}_i$.

3.2.2 Identifiability

Identifiability is critical for valid statistical inference. In general, a statistical model is said to be non-identifiable if two distinct sets of parameters lead to the same probability distribution. In contrast, a parameter θ for a particular family of distributions $f(x|\theta), \theta \in \Theta$ with a sample space of $x \in S$ is identifiable if different values of θ correspond to different probability density functions. That is,

$$\theta \neq \eta \Rightarrow f(x|\theta) \neq f(x|\eta) \forall x \in S.$$

First, assume that a FinMix GLMM can be written in the form

$$f_{\mathbf{y}_i}(\mathbf{y}_i|\mathbb{X}_i, \mathbb{Z}_i, \mathbf{\Theta}) = \sum_{k=1}^{K} \pi_k f_{\mathbf{y}_i}^{(k)}(\mathbf{y}_i|\mathbb{X}_i, \mathbb{Z}_i, \boldsymbol{\beta}_k, \mathbb{\Gamma}_k).$$

Given a set of covariate matrices $(\mathbb{X}_1, \mathbb{X}_2, \ldots, \mathbb{X}_n)$, corresponding to the fixed effects, and $(\mathbb{Z}_1, \mathbb{Z}_2, \ldots, \mathbb{Z}_n)$, corresponding to the random effects, the FinMix GLMM is identifiable if for any two vectors $\mathbf{\Theta}$ and $\mathbf{\Theta}^\dagger$

$$\sum_{k=1}^{K} \boldsymbol{\pi}_k f_{\mathbf{y}_i}^{(k)}(\mathbf{y}_i|\mathbb{X}_i, \mathbb{Z}_i, \boldsymbol{\beta}_k, \mathbb{\Gamma}_k) = \sum_{k=1}^{K^\dagger} \boldsymbol{\pi}_k^\dagger f_{\mathbf{y}_i}^{(k)}(\mathbf{y}_i|\mathbb{X}_i, \mathbb{Z}_i, \boldsymbol{\beta}_k^\dagger, \mathbb{\Gamma}_k^\dagger)$$

for all possible values of $\mathbf{y}_i$, and each $i = 1, 2, \ldots, n$ then $K = K^\dagger$, and $\mathbf{\Theta} = \mathbf{\Theta}^\dagger$. For a finite mixture model to be identifiable, the models for each subpopulation must also be identifiable (Teicher, 1963; Atienza et al., 2006). First, I considered identifiability conditions for GLMMs, followed by identifiability conditions specific to finite mixtures.

As is the case in linear regression, $\det(X^\top X) \neq 0$. When categorical variables are present in the model, the model must be parameterised such that one category (often the most common category) is the baseline, or parameterised using a sum-to-zero constraint. Three additional factors are important for the identifiability of a FinMix GLMM, similar to a Finite Mixture of Linear Mixed-Effect (FMLME), and following from Hennig (2000): (i) density of the subpopulations, (ii) number of subpopulations K, (iii) design matrices of covariates.

Using the matrix

$$\Upsilon = \begin{bmatrix} \mathbb{X}_1, \mathbb{X}_2, \ldots \mathbb{X}_m \\ \mathbb{Z}_1, \mathbb{Z}_2, \ldots \mathbb{Z}_m \end{bmatrix}^\top.$$

Hennig's condition translates to the restriction that K must be smaller than the number of $(h_\Upsilon - 1)$-dimensional hyperplanes covered by the rows of Υ. Since it is typical in GLMMs that any covariate that has a random effect also has a fixed effect, this may be simplified by excluding the $\mathbb{Z}_i$s from the matrix Υ.

Following the work in Labouriau (2014), I provided two conditions for identifiability of a GLMM. A GLMM is identifiable if $\int g(\mathbf{x}_{ij}\boldsymbol{\beta}_k + \mathbf{z}_{ij}\mathbb{\Gamma}_k\mathbf{b}_i) \times f_{\mathbf{b}_i}(\mathbf{b}_i)d\mathbf{b}_i < \infty$ and $g(\mathbf{x}_{ij}\boldsymbol{\beta}_k + \mathbf{z}_{ij}\mathbb{\Gamma}_k\mathbf{b}_i)(\mathbb{\Gamma}_k)^{-1} = g(\mathbf{x}_{ij}\boldsymbol{\beta}_{k*} + \mathbf{z}_{ij}\mathbb{\Gamma}_{k*}\mathbf{b}_i)(\mathbb{\Gamma}_{k*})^{-1}\forall\mathbf{x}_{ij}, \forall\mathbf{z}_{ij} \implies \boldsymbol{\beta}_k = \boldsymbol{\beta}_{k*}, \mathbb{\Gamma}_k = \mathbb{\Gamma}_{k*}$.

There are three types of non-identifiability that are specific to finite mixture models. Label switching is a common type of non-identifiability in finite mixtures of models. This is also known as non-identifiability due to invariance in relabelling the components and occurs when the mixture components can be relabelled without changing the likelihood (Redner and Walker, 1984). A simple solution to this is to impose a strict ordering constraint on a single element, usually π_k.

Next, consider non-identifiability due to potential overfitting. In this case, two mixture models, one with K subpopulations, one with $K - 1$ subpopulations, are equivalent. Either one of the K subpopulations is empty, or two of the K subpopulations have equal regression parameters. Again, this can be solved by imposing a strict ordering constraint, and that $\pi_k > 0,\ \forall k \in \{1, 2, \ldots, K\}$, and checking that each subpopulation has distinct regression coefficients. One potential problem occurs when $\pi_k = \pi_{k^\dagger}$, in which case the user must impose an order.

Consider an example FinMix GLMM with $\boldsymbol{\Theta} = (\pi_1, \boldsymbol{\beta}_1^\top, \mathbb{\Gamma}_1^{*\top}, \boldsymbol{\beta}_2^\top, \mathbb{\Gamma}_2^{*\top})^\top$, and a set of covariate matrices $(\mathbb{X}_1, \mathbb{X}_2, \ldots, \mathbb{X}_n)$, corresponding to the fixed effects, and $(\mathbb{Z}_1,$

$\mathbb{Z}_2, \ldots, \mathbb{Z}_n$), corresponding to the random effects. Note that this model defines a two subpopulation distribution. Let $\boldsymbol{\Theta}^\dagger = (\pi_1^\dagger, \boldsymbol{\beta}_1^{\dagger\top}, \mathbb{\Gamma}_1^{\dagger*\top}, \boldsymbol{\beta}_2^{\dagger\top}, \mathbb{\Gamma}_2^{\dagger*\top})^\top = ((1-\pi_1), \boldsymbol{\beta}_2^\top, \mathbb{\Gamma}_2^{*\top}, \boldsymbol{\beta}_1^\top, \mathbb{\Gamma}_2^{*\top})^\top$. Then,

$$
\begin{aligned}
f_{\mathbf{y}_i}(\mathbf{y}_i|\mathbb{X}_i, \mathbb{Z}_i, \boldsymbol{\Theta}) &= \sum_{k=1}^{2} \pi_k f_{\mathbf{y}_i}^{(k)}(\mathbf{y}_i|\mathbb{X}_i, \mathbb{Z}_i, \boldsymbol{\beta}_k \mathbb{\Gamma}_k) \\
&= \pi_1 f_{\mathbf{y}_i}^{(1)}(\mathbf{y}_i|\mathbb{X}_i, \mathbb{Z}_i, \boldsymbol{\beta}_1, \mathbb{\Gamma}_1^*) + \pi_2 f_{\mathbf{y}_i}^{(2)}(\mathbf{y}_i|\mathbb{X}_i, \mathbb{Z}_i, \boldsymbol{\beta}_2, \mathbb{\Gamma}_2^*) \\
&= \pi_2 f_{\mathbf{y}_i}^{(2)}(\mathbf{y}_i|\mathbb{X}_i, \mathbb{Z}_i, \boldsymbol{\beta}_2, \mathbb{\Gamma}_2^*) + \pi_1 f_{\mathbf{y}_i}^{(1)}(\mathbf{y}_i|\mathbb{X}_i, \mathbb{Z}_i, \boldsymbol{\beta}_1, \mathbb{\Gamma}_1^*) \\
&= \pi_1^\dagger f_{\mathbf{y}_i}^{(1)}(\mathbf{y}_i|\mathbb{X}_i, \mathbb{Z}_i, \boldsymbol{\beta}_1^\dagger, \mathbb{\Gamma}_1^{\dagger*}) + \pi_2^\dagger f_{\mathbf{y}_i}^{(2)}(\mathbf{y}_i|\mathbb{X}_i, \mathbb{Z}_i, \boldsymbol{\beta}_2^\dagger, \Gamma_2^{\dagger*}) \\
&= \sum_{k=1}^{2} \pi_k' f_{\mathbf{y}_i}^{(k)}(\mathbf{y}_i|\mathbb{X}_i, \mathbb{Z}_i, \boldsymbol{\beta}_k^\dagger, \mathbb{\Gamma}_k^{\dagger*}) \\
&= f_{\mathbf{y}_i}(\mathbf{y}_i|\mathbb{X}_i, \mathbb{Z}_i, \boldsymbol{\Theta}^\dagger).
\end{aligned}
$$

Thus, this model is unidentifiable. This is an example of non-identifiability due to label switching, and can be avoided in many cases by imposing the requirement $\pi_1 > \pi_2$. This constraint does not solve this problem when $\pi_1 = \pi_2$, so in that case the additional requirement that $\beta_{11} > \beta_{21}$. If $\pi_1 = \pi_2$ and $\beta_{11} = \beta_{21}$, then impose the requirement $\beta_{12} > \beta_{22}$, and continuing on in this way if $\beta_{12} = \beta_{22}$. Note that if all of the parameters in both subpopulations are equal, the subpopulations are not distinct and there is only one subpopulation. As such, at least one parameter for the two subpopulations must be unequal, and the first of these is used to impose the constraint.

I assumed the model is identifiable if the above restrictions are met.

3.3 Maximum Likelihood Estimation

To calculate the MLE, I first considered the likelihood equation. I then examined the numerical computation of the MLE as well as specific details in the Poisson, and binomial cases.

3.3.1 Likelihood Equation

Let $(\mathbf{y}_i, \mathbb{X}_i), i \in \{1, 2, \ldots, n\}$ be a sample from a FinMix GLMM as described in (2.1). The likelihood function is given by

$$L_n(\mathbf{\Theta}) \quad = \quad \prod_{i=1}^{n} [f_{\mathbf{y}_i}(\mathbf{y}_i | \mathbb{X}_i, \mathbb{Z}_i, \mathbf{\Theta})]$$

and the log-likelihood as

$$\begin{aligned} \ell_n(\mathbf{\Theta}) &= \log[L_n(\mathbf{\Theta})] \\ &= \log\left\{\prod_{i=1}^{n}[f_{\mathbf{y}_i}(\mathbf{y}_i | \mathbb{X}_i, \mathbb{Z}_i, \mathbf{\Theta})]\right\} \\ &= \sum\nolimits_{i=1}^{n} \log[f_{\mathbf{y}_i}(\mathbf{y}_i | \mathbb{X}_i, \mathbb{Z}_i, \mathbf{\Theta})]. \end{aligned}$$

The MLE is defined as $\widehat{\mathbf{\Theta}} = \arg\max(L(\mathbf{\Theta}))$. Due to the complexity of the model, I approximated the likelihood by replacing the integral in (2.1) with a sum over a large number, L, of generated values of $\mathbf{b}_i$ for each of the n patients.

Many useful asymptotic properties of maximum likelihood estimators can be applied to these estimators. Specifically, consistency (Chen, 2017), and asymptotically following a Gaussian distribution are consequences of the maximum likelihood procedure. However, the usual regularity conditions are needed, and can be found in Appendix D.

3.3.2 Numerical Computation of the Maximum Likelihood Estimator

The MCEM algorithm is popular in many settings and has been widely used as detailed in Chapter 2. In this application, I applied the MCEM algorithm twice. The outer MCEM loop treats the subpopulation membership indicators as missing. That is, the information about which patient is in which subpopulation is unknown, and thus estimated. Estimating the probabilities of subgroup membership is in the outer MCEM. The inner MCEM loop treats the subject-specific random effect $\mathbf{b}_i$ as missing, and calculates the regression parameters $\boldsymbol{\beta}_k$, and $\mathbb{\Gamma}_k$. I summarised this algorithm in Section C.1.

Estimating the Subpopulation Membership; The Outer Monte Carlo Expectation-Maximisation Algorithm

To estimate the mixing proportions $\pi_1, \pi_2, \ldots, \pi_K$, I first estimated the membership probability of each individual in each of the subpopulations, which I denoted as τ_{ki}. I did this by calculating the likelihood for each individual, and each subpopulation, and then combining this information to find probabilities τ_{ki}. Specifically, I calculated τ_{ki} for each $k \in \{1, 2, \ldots, K\}$, and $i \in \{1, 2, \ldots, n\}$ according to:

$$\begin{aligned}
\tau_{ki} &= \frac{\hat{\pi}_k^{(t)} f_{\mathbf{y}_i}^{(k)}(\mathbf{y}_i|\mathbb{X}_i, \mathbb{Z}_i, \hat{\boldsymbol{\beta}}_k^{(t)}, \hat{\mathbb{\Gamma}}_k^{(t)})}{\sum_{h=1}^{K} \hat{\pi}_h^{(t)} f_{\mathbf{y}_i}^{(h)}(\mathbf{y}_i|\mathbb{X}_i, \mathbb{Z}_i, \hat{\boldsymbol{\beta}}_h^{(t)}, \hat{\mathbb{\Gamma}}_h^{(t)})} \\
&= \frac{\hat{\pi}_k^{(t)} \int f_{\mathbf{y}_i|\mathbf{b}_i}^{(k)}(\mathbf{y}_i|\mathbf{b}_i, \mathbb{X}_i, \mathbb{Z}_i, \hat{\boldsymbol{\beta}}_k, \hat{\mathbb{\Gamma}}_k) \times f_{\mathbf{b}_i}(\mathbf{b}_i) d\mathbf{b}_i}{\sum_{h=1}^{K} \hat{\pi}_h^{(t)} \int f_{\mathbf{y}_i|\mathbf{b}_i}^{(h)}(\mathbf{Y}_i|\mathbf{b}_i, \mathbb{X}_i, \mathbb{Z}_i, \hat{\boldsymbol{\beta}}_h, \hat{\mathbb{\Gamma}}_h) \times f_{\mathbf{b}_i}(\mathbf{b}_i) d\mathbf{b}_i} \\
&= \frac{\hat{\pi}_k^{(t)} \int \prod_{j=1}^{n_i} f_{\mathbf{y}_i|\mathbf{b}_i}^{(k)}(y_{ij}|\mathbf{b}_i, \mathbf{x}_{ij}, \mathbf{z}_{ij}, \hat{\boldsymbol{\beta}}_k, \hat{\mathbb{\Gamma}}_k) \times f_{\mathbf{b}_i}(\mathbf{b}_i) d\mathbf{b}_i}{\sum_{h=1}^{K} \hat{\pi}_h^{(t)} \int \prod_{j=1}^{n_i} f_{\mathbf{y}_i|\mathbf{b}_i}^{(h)}(y_{ij}|\mathbf{b}_i, \mathbf{x}_{ij}, \mathbf{z}_{ij}, \hat{\boldsymbol{\beta}}_h, \hat{\mathbb{\Gamma}}_h) \times f_{\mathbf{b}_i}(\mathbf{b}_i) d\mathbf{b}_i}. \qquad (3.1)
\end{aligned}$$

Because the $\mathbf{b}_i$s are unknown, the integrals in the numerator, and denominator of Equation (3.1) must be approximated. This forms the E-step of the outer Expectation-Maximisation (EM) algorithm. In order to approximate this integral, I generated L potential values of $\mathbf{b}_i$ for each $i \in \{1, 2, \ldots, n\}$, and $k \in \{1, 2, \ldots, K\}$ from a standard multivariate Gaussian distribution. A superscript on $\mathbf{b}_i$ denotes that the $\mathbf{b}_i^{(l)}$ was generated in the estimation process rather than the unobserved $\mathbf{b}_i$.

$$
\begin{aligned}
\tau_{ki} &= \frac{\hat{\pi}_k^{(t)} \int \prod_{j=1}^{n_i} f_{y_{ij}|\mathbf{b}_i}^{(k)}(y_{ij}|\mathbf{b}_i, \mathbf{x}_{ij}, \mathbf{z}_{ij}, \hat{\boldsymbol{\beta}}_k, \hat{\Gamma}_k) \times f_{\mathbf{b}_i}(\mathbf{b}_i) d\mathbf{b}_i}{\sum_{h=1}^{K} \hat{\pi}_h^{(t)} \int \prod_{j=1}^{n_i} f_{y_{ij}|\mathbf{b}_i}^{(h)}(y_{ij}|\mathbf{b}_i, \mathbf{x}_{ij}, \mathbf{z}_{ij}, \hat{\boldsymbol{\beta}}_h, \hat{\Gamma}_h) \times f_{\mathbf{b}_i}(\mathbf{b}_i) d\mathbf{b}_i} \\
&\approx \frac{\hat{\pi}_k^{(t)} \frac{1}{L} \sum_{l=1}^{L} \prod_{j=1}^{n_i} f_{y_{ij}|\mathbf{b}_i}^{(k)}(y_{ij}|\mathbf{b}_i^{(l)}, \mathbf{x}_{ij}, \mathbf{z}_{ij}, \hat{\boldsymbol{\beta}}_k, \hat{\Gamma}_k)}{\sum_{h=1}^{K} \hat{\pi}_h^{(t)} \frac{1}{L} \sum_{l=1}^{L} \prod_{j=1}^{n_i} f_{y_{ij}|\mathbf{b}_i}^{(h)}(y_{ij}|\mathbf{b}_i^{(l)}, \mathbf{x}_{ij}, \mathbf{z}_{ij}, \hat{\boldsymbol{\beta}}_h, \hat{\Gamma}_h)} \\
&= \frac{\hat{\pi}_k^{(t)} \sum_{l=1}^{L} \prod_{j=1}^{n_i} f_{y_{ij}|\mathbf{b}_i}^{(k)}(y_{ij}|\mathbf{b}_i^{(l)}, \mathbf{x}_{ij}, \mathbf{z}_{ij}, \hat{\boldsymbol{\beta}}_k, \hat{\Gamma}_k)}{\sum_{h=1}^{K} \hat{\pi}_h^{(t)} \sum_{l=1}^{L} \prod_{j=1}^{n_i} f_{y_{ij}|\mathbf{b}_i}^{(h)}(y_{ij}|\mathbf{b}_i^{(l)}, \mathbf{x}_{ij}, \mathbf{z}_{ij}, \hat{\boldsymbol{\beta}}_h, \hat{\Gamma}_h)}.
\end{aligned}
$$

Note that τ_{ki} is specific to patient i, and subgroup k but general to all visits $j = 1, 2, \ldots, n_i$ for patient i because I assumed each patient is a member of one, and only one subpopulation. In cases where both the numerator, and denominator were calculated to be zero or numerically very close to zero, I set $\tau_{ki} = \frac{1}{K}$ for all k. That is, when these calculations show that membership to each of the subpopulations is equally unlikely (given rounding error), I assigned an equal probability of belonging to any of the subpopulations, which can be thought of as a uniform prior distribution. This case occurs most frequently when the estimates in $\boldsymbol{\Theta}$ are poor or the true value for $\mathbf{b}_i$ is extreme.

I then estimated the mixing proportions by taking the empirical average of the individual subpopulation membership probabilities:

$$\hat{\pi}_k = \frac{1}{n} \sum_{i=1}^{n} \tau_{ki}.$$

Maximising the likelihood with respect to $\boldsymbol{\beta}_k$, and $\mathbb{\Gamma}_k$ for $k \in \{1, 2, \ldots, K\}$ forms the M-step of the outer MCEM algorithm. This involves calling the inner MCEM for each k (each subpopulation). I used the values of τ_{ki} as weights in the inner MCEM loops.

Estimating the Random Effects; The Inner Monte Carlo Expectation-Maximisation Algorithm

Consider the inner EM for a particular $k \in \{1, 2, \ldots, K\}$. Generate L estimates of $\mathbf{b}_i$s for each $i \in \{1, 2, \ldots, n\}$. Note that when I generated $\mathbf{b}_i$, I generated it from $\mathbf{b}_i | \mathbf{y}_i, \mathbb{X}_i, \mathbb{Z}_i, \boldsymbol{\beta}_k, \mathbb{\Gamma}_k$, not from a standard Gaussian distribution or indeed from any standard distribution.

Instead, I generated potential $\mathbf{b}_i$s by calculating the MLE for $\mathbf{b}_i | \mathbf{y}_i, \mathbb{X}_i, \mathbb{Z}_i, \boldsymbol{\beta}_k, \mathbb{\Gamma}_k$, and the Hessian matrix at this point. Next, I generated points from a multivariate Gaussian distribution with the mean set to $\hat{\mathbf{b}}_i$, the MLE for $\mathbf{b}_i$, and the variance-covariance matrix set to the negative of the inverse of the Hessian matrix. I then used these points as the proposed values for an accept/reject sampler which I repeated multiple times. These Metropolis iterations improve the sample to make it more representative of the posterior distribution. Points that were accepted became the L potential $\mathbf{b}_i$s used in the next step. Note that L potential $\mathbf{b}_i$ values were proposed for each n in each of the k calls to the inner MCEM loop. I denote these $\mathbf{b}_{ki}^{(l)}$. Note that

$\mathbf{b}_i|\mathbf{y}_i, \mathbb{Z}_i, \mathbb{X}_i, \boldsymbol{\beta}_1, \mathbb{\Gamma}_1 \overset{d}{\neq} \mathbf{b}_i|\mathbf{y}_i, \mathbb{Z}_i, \mathbb{X}_i, \boldsymbol{\beta}_2, \mathbb{\Gamma}_2$, so I generated different values of $\mathbf{b}_i$ for each of the K subpopulations. A more detailed algorithm can be found in Section C.4.

The M-step of the inner MCEM algorithm maximises the log-likelihood of each of the K subpopulations, represented by the function $Q_k(\boldsymbol{\theta}_k)$, with respect to $\theta_k = (\boldsymbol{\beta}_k^\top, \mathbb{\Gamma}_k^{*\top})^\top$. The log-likelihood for each subpopulation is

$$Q_k(\boldsymbol{\theta}_k) = \sum_{i=1}^{n} \tau_{ki} \times \log[f_{\mathbf{y}_i}(\mathbf{y}_i|\mathbb{X}_i, \mathbb{Z}_i, \boldsymbol{\theta}_k)].$$

Newton-Raphson was used to maximise the approximate likelihood. I weighted this likelihood by the values of τ_{ki}. Given the approximate likelihood, along with its first, and second derivatives, the next iteration of $\boldsymbol{\theta}_k$ was defined as $\boldsymbol{\theta}_k^{(s+1)} = \boldsymbol{\theta}_k^{(s)} - (\frac{1}{2})^\iota [Q_k''(\boldsymbol{\theta}_k^{(s)})]^{-1} Q_k'(\boldsymbol{\theta}_k^{(s)})$ where $\iota = 0$ in most cases, but can be changed to facilitate half step Newton-Raphson.

3.3.3 Examples

In this section, I derived the MLEs, and provided computational details for two specific FinMix GLMMs: a finite mixture of Poisson, and a finite mixture of binomial distributions.

Approximate Likelihood - Poisson Case

The approximate likelihood in the Poisson case uses the canonical link function, and a sum over generated values of $\mathbf{b}_{ki}^{(s,l)}$ to replace the integral. As previously, the superscript denotes that $\mathbf{b}_{ki}^{(s,l)}$ was generated as part of the estimation, and is not the latent $\mathbf{b}_i$. To simplify and condense the equations in the remainder of the chapter I shortened $\sum_{l=1}^{L}\sum_{i=1}^{n}\sum_{j=1}^{n_i}$ to $\sum_{l,i,j}$. The resulting likelihood equation is

$$\begin{aligned} Q_k(\boldsymbol{\theta}) &= Q_k(\boldsymbol{\beta}_k, \Gamma_k) = \frac{1}{L}\sum_{l,i,j}\tau_{ki}\left[y_{ij}\log(\xi_{kij}^{(s,l)}) - \xi_{kij}^{(s,l)}\right] \\ &= \frac{1}{L}\sum_{l,i,j}\tau_{ki}\left[y_{ij}x_{ij}\boldsymbol{\beta}_k + y_{ij}\mathbf{z}_{ij}\Gamma_k\mathbf{b}_{ki}^{(s,l)} - e^{\mathbf{x}_{ij}\boldsymbol{\beta}_k+\mathbf{z}_{ij}\Gamma_k\mathbf{b}_{ki}^{(s,l)}}\right]. \end{aligned}$$

To maximise this approximate likelihood, I used Newton-Raphson. This requires calculating both the first, and second derivatives with respect to θ_k. These derivatives are shown with respect to $\boldsymbol{\beta}_k$, and Γ_k separately, and are as follows:

$$\begin{aligned} \frac{\partial Q_k(\boldsymbol{\beta}_k, \Gamma_k)}{\partial \boldsymbol{\beta}_k} &= \frac{1}{L}\sum_{l,i,j}\tau_{ki}\left[y_{ij}\mathbf{x}_{ij} - \mathbf{x}_{ij}e^{\mathbf{x}_{ij}\boldsymbol{\beta}_k+\mathbf{z}_{ij}\Gamma_k\mathbf{b}_{ki}^{(s,l)}}\right] \\ &= \frac{1}{L}\sum_{l,i,j}\tau_{ki}\mathbf{x}_{ij}\left[y_{ij} - e^{\mathbf{x}_{ij}\boldsymbol{\beta}_k+\mathbf{z}_{ij}\Gamma_k\mathbf{b}_{ki}^{(s,l)}}\right]; \\ \frac{\partial Q_k(\boldsymbol{\beta}_k, \Gamma_k)}{\partial \Gamma_k^*} &= \frac{1}{L}\sum_{l,i,j}\tau_{ki}\left[y_{ij}\mathbf{z}_{ij}\mathbf{b}_{ki}^{(s,l)} - e^{\mathbf{x}_{ij}\boldsymbol{\beta}_k+\mathbf{z}_{ij}\Gamma_k\mathbf{b}_{ki}^{(s,l)}}\mathbf{z}_{ij}\mathbf{b}_{ki}^{(s,l)}\right] \\ &= \frac{1}{L}\sum_{l,i,j}\tau_{ki}\left[y_{ij} - e^{\mathbf{x}_{ij}\boldsymbol{\beta}_k+\mathbf{z}_{ij}\Gamma_k\mathbf{b}_{ki}^{(s,l)}}\right]\mathbf{z}_{ij}\mathbf{b}_{ki}^{(s,l)}; \\ \frac{\partial^2 Q_k(\boldsymbol{\beta}_k, \Gamma_k)}{\partial \boldsymbol{\beta}_k^\top \partial \boldsymbol{\beta}_k} &= \frac{1}{L}\sum_{l,i,j}\tau_{ki}\left[-\mathbf{x}_{ij}\mathbf{x}_{ij}^\top e^{\mathbf{x}_{ij}\boldsymbol{\beta}_k+\mathbf{z}_{ij}\Gamma_k\mathbf{b}_{ki}^{(s,l)}}\right]; \\ \frac{\partial^2 Q_k(\boldsymbol{\beta}_k, \Gamma_k)}{\partial \boldsymbol{\beta}_k^\top \partial \Gamma_k^*} &= \frac{1}{L}\sum_{l,i,j}\tau_{ki}\left[-\mathbf{x}_{ij}\mathbf{z}_{ij}\mathbf{b}_{ki}^{(s,l)}e^{\mathbf{x}_{ij}\boldsymbol{\beta}_k+\mathbf{z}_{ij}\Gamma_k\mathbf{b}_{ki}^{(s,l)}}\right]; \\ \frac{\partial^2 Q_k(\boldsymbol{\beta}_k, \Gamma_k)}{\partial \Gamma_k^{*\top} \partial \boldsymbol{\beta}_k} &= \frac{1}{L}\sum_{l,i,j}\tau_{ki}\left[-\mathbf{x}_{ij}\mathbf{z}_{ij}\mathbf{b}_{ki}^{(s,l)}e^{\mathbf{x}_{ij}\boldsymbol{\beta}_k+\mathbf{z}_{ij}\Gamma_k\mathbf{b}_{ki}^{(s,l)}}\right]; \\ \frac{\partial^2 Q_k(\boldsymbol{\beta}_k, \Gamma_k)}{\partial \Gamma_k^{*\top} \partial \Gamma_k^*} &= \frac{1}{L}\sum_{l,i,j}\tau_{ki}\left[-\mathbf{z}_{ij}\mathbf{b}_{ki}^{(s,l)}(\mathbf{z}_{ij}\mathbf{b}_{ki}^{(s,l)})^\top e^{\mathbf{x}_{ij}\boldsymbol{\beta}_k+\mathbf{z}_{ij}\Gamma_k\mathbf{b}_{ki}^{(s,l)}}\right]. \end{aligned}$$

Additional details on these derivations by way of element-wise calculations can be found in Section B.1.

Approximate Likelihood - Binomial Case

The computation for the binomial case is much the same. However, the link function is different, and the number of trials (m_{ij}) must be included in the likelihood as well.

$$
\begin{aligned}
Q_k(\boldsymbol{\theta}) &= Q_k(\boldsymbol{\beta}_k, \mathbb{\Gamma}_k) = \frac{1}{L}\sum_{l,i,j} \tau_{ki}\left[y_{ij}\log(\varphi_{kij}) + (m_{ij} - y_{ij})\log(1-\varphi_{kij})\right] \\
&= \frac{1}{L}\sum_{l,i,j} \tau_{ki}\left[y_{ij}(\mathbf{x}_{ij}\boldsymbol{\beta}_k + \mathbf{z}_{ij}\mathbb{\Gamma}_k\mathbf{b}_{ki}^{(s,l)}) - m_{ij}\log(e^{\mathbf{x}_{ij}\boldsymbol{\beta}_k + \mathbf{z}_{ij}\mathbb{\Gamma}_k\mathbf{b}_{ki}^{(s,l)}} + 1)\right]
\end{aligned}
$$

As with the Poisson case, I calculated the first, and second derivatives for their use in Newton-Raphson, to maximise the likelihood in the M-step of the inner MCEM.

$$
\begin{aligned}
\frac{\partial Q_k(\boldsymbol{\beta}_k, \mathbb{\Gamma}_k)}{\partial \boldsymbol{\beta}_k} &= \frac{1}{L}\sum_{l,i,j} \tau_{ki}\left[y_{ij}\mathbf{x}_{ij} - m_{ij}\mathbf{x}_{ij}\text{expit}(\mathbf{x}_{ij}\boldsymbol{\beta}_k + \mathbf{z}_{ij}\mathbb{\Gamma}_k\mathbf{b}_{ki}^{(s,l)})\right] \\
&= \frac{1}{L}\sum_{l,i,j} \tau_{ki}\left[y_{ij} - m_{ij}\text{expit}(\mathbf{x}_{ij}\boldsymbol{\beta}_k + \mathbf{z}_{ij}\mathbb{\Gamma}_k\mathbf{b}_{ki}^{(s,l)})\right]\mathbf{x}_{ij}; \\
\frac{\partial Q_k(\boldsymbol{\beta}_k, \mathbb{\Gamma}_k)}{\partial \mathbb{\Gamma}_k^*} &= \frac{1}{L}\sum_{l,i,j} \tau_{ki}\left[y_{ij}\mathbf{z}_{ij}\mathbf{b}_{ki}^{(s,l)} - m_{ij}\mathbf{z}_{ij}\mathbf{b}_{ki}^{(s,l)} \times \text{expit}(\mathbf{x}_{ij}\boldsymbol{\beta}_k + \mathbf{z}_{ij}\mathbb{\Gamma}_k\mathbf{b}_{ki}^{(s,l)})\right] \\
&= \frac{1}{L}\sum_{l,i,j} \tau_{ki}\left[y_{ij} - m_{ij}\text{expit}(\mathbf{x}_{ij}\boldsymbol{\beta}_k + \mathbf{z}_{ij}\mathbb{\Gamma}_k\mathbf{b}_{ki}^{(s,l)})\right] \times \mathbf{z}_{ij}\mathbf{b}_{ki}^{(s,l)}; \\
\frac{\partial^2 Q_k(\boldsymbol{\beta}_k, \mathbb{\Gamma}_k)}{\partial \boldsymbol{\beta}_k^\top \partial \boldsymbol{\beta}_k} &= \frac{1}{L}\sum_{l,i,j} \tau_{ki}\left[-\mathbf{x}_{ij}\mathbf{x}_{ij}^\top m_{ij}\frac{\exp(\mathbf{x}_{ij}\boldsymbol{\beta}_k + \mathbf{z}_{ij}\mathbb{\Gamma}_k\mathbf{b}_{ki}^{(s,l)})}{(\exp(\mathbf{x}_{ij}\boldsymbol{\beta}_k + \mathbf{z}_{ij}\mathbb{\Gamma}_k\mathbf{b}_{ki}^{(s,l)}) + 1)^2}\right]; \\
\frac{\partial^2 Q_k(\boldsymbol{\beta}_k, \mathbb{\Gamma}_k)}{\partial \boldsymbol{\beta}_k^\top \partial \mathbb{\Gamma}_k^*} &= \frac{1}{L}\sum_{l,i,j} \tau_{ki}\left[-\mathbf{x}_{ij}\mathbf{z}_{ij}\mathbf{b}_{ki}^{(s,l)} m_{ij}\frac{\exp(\mathbf{x}_{ij}\boldsymbol{\beta}_k + \mathbf{z}_{ij}\mathbb{\Gamma}_k\mathbf{b}_{ki}^{(s,l)})}{(\exp(\mathbf{x}_{ij}\boldsymbol{\beta}_k + \mathbf{z}_{ij}\mathbb{\Gamma}_k\mathbf{b}_{ki}^{(s,l)}) + 1)^2}\right]; \\
\frac{\partial^2 Q_k(\boldsymbol{\beta}_k, \mathbb{\Gamma}_k)}{\partial \mathbb{\Gamma}_k^{*\top} \partial \boldsymbol{\beta}_k} &= \frac{1}{L}\sum_{l,i,j} \tau_{ki}\left[-\mathbf{x}_{ij}\mathbf{z}_{ij}\mathbf{b}_{ki}^{(s,l)} m_{ij}\frac{\exp(\mathbf{x}_{ij}\boldsymbol{\beta}_k + \mathbf{z}_{ij}\mathbb{\Gamma}_k\mathbf{b}_{ki}^{(s,l)})}{(\exp(\mathbf{x}_{ij}\boldsymbol{\beta}_k + \mathbf{z}_{ij}\mathbb{\Gamma}_k\mathbf{b}_{ki}^{(s,l)}) + 1)^2}\right]; \\
\frac{\partial^2 Q_k(\boldsymbol{\beta}_k, \mathbb{\Gamma}_k)}{\partial \mathbb{\Gamma}_k^{*\top} \partial \mathbb{\Gamma}_k^*} &= \frac{1}{L}\sum_{l,i,j} \tau_{ki}\left[-\mathbf{z}_{ij}\mathbf{b}_{ki}^{(s,l)}(\mathbf{z}_{ij}\mathbf{b}_{ki}^{(s,l)})^\top m_{ij}\frac{\exp(\mathbf{x}_{ij}\boldsymbol{\beta}_k + \mathbf{z}_{ij}\mathbb{\Gamma}_k\mathbf{b}_{ki}^{(s,l)})}{(\exp(\mathbf{x}_{ij}\boldsymbol{\beta}_k + \mathbf{z}_{ij}\mathbb{\Gamma}_k\mathbf{b}_{ki}^{(s,l)}) + 1)^2}\right].
\end{aligned}
$$

Table 3–1: Simulation settings for MLE.

Outcome	K	n	Dim($\boldsymbol{\beta}$)	Dim($\mathbb{\Gamma}$)	Dim($\boldsymbol{\Theta}$)
Poisson	2	$100, 250, 500, 1000$	2	1	7
Poisson	2	$100, 250, 500, 1000$	5	1	13
Poisson	2	$100, 250, 500, 1000$	5	2	17
Poisson	3	$100, 250, 500, 1000$	2	1	11
Poisson	3	$100, 250, 500, 1000$	5	1	20
Poisson	3	$100, 250, 500, 1000$	5	2	26
Binomial	2	$100, 250, 500, 1000$	2	1	7
Binomial	2	$100, 250, 500, 1000$	5	1	13
Binomial	2	$100, 250, 500, 1000$	5	2	17
Binomial	3	$100, 250, 500, 1000$	2	1	11
Binomial	3	$100, 250, 500, 1000$	5	1	20
Binomial	3	$100, 250, 500, 1000$	5	2	26

Recall that expit is the inverse of the logit function, that is $\text{expit}(x) = \frac{\exp(x)}{\exp(x)+1}$. I have included additional details on these derivations in Section B.2.

3.4 Simulation Study

I designed and executed a simulation study to demonstrate the performance of the MLE of the FinMix GLMM parameters. I ran simulations for both Poisson, and binomial outcomes, with scenarios which included mixtures of two, and three subpopulations, and varied the number of fixed, and random effects. I considered four sample sizes $(100, 250, 500, 1000)$ for all simulations and ran each scenario 100 times. See Table 3–1 for a summary of the scenarios considered. In the Poisson case, $n_i \in \{8, 9, 10\} \forall i$, and in the binomial case, $n_i \in \{3, 4, 5, 6\} \forall i$.

In all cases, a multinomial distribution was used to generate subpopulation membership. I included an intercept in each model, so the first column of the matrix $\mathbb{X}$ is of 1s. I generated the covariates independently. In the following, $i \in \{1, 2, \ldots, n\}$, and $j \in \{1, 2, \ldots, n_i\}$. I generated the covariate data as follows:

$$\begin{aligned} x_{ij1} &= 1 \forall i \forall j; \\ x_{ij2} &\sim Gaussian(\mu = 1, \sigma = 0.5) \forall i \forall j \end{aligned}$$

for cases where $p = 2$. When $p = 5$, I used:

$$\begin{aligned} x_{ij3} &\sim Gaussian(\mu = 0, \sigma = 0.5) \forall i \forall j; \\ x_{ij4} &\sim Gaussian(\mu = -1, \sigma = 1) \forall i \forall j; \\ x_{ij5} &\sim Gaussian(\mu = 0, \sigma = 1) \forall i \forall j. \end{aligned}$$

For any variable for which there was a random effect, there was also a fixed effect, that is, all columns of $\mathbb{Z}$ are also columns of $\mathbb{X}$. Additionally, if $q = 1$ then $z_{ij1} = x_{ij1} \forall i \forall j$, and if $q = 2$ then $z_{ij1} = x_{ij1} \forall i \forall j$, and $z_{ij2} = x_{ij2} \forall i \forall j$. I generated the random effects $\mathbf{b}_i$ from a standard multivariate Gaussian distribution. Given the values of $\mathbb{X}_i$, $\mathbb{Z}_i$, the group membership, $\mathbf{b}_i$, $\boldsymbol{\beta}_k$, and $\mathbb{\Gamma}_k$ for all $k \in \{1, 2, \ldots, K\}$, I then generated the outcomes for the simulations. In the Poisson case, ξ_{ij} was calculated as $\log(\xi_{ij}) = \mathbf{X}_{ij}^\top \beta_k + \mathbf{Z}_{ij} \mathbb{\Gamma}_k \mathbf{b}_i$, and then $Y_{ij} \sim Poisson(\xi_{ij})$. In the binomial case, $\text{logit}(\varphi_{ij}) = \mathbf{x}_{ij}^\top \boldsymbol{\beta}_k + \mathbf{z}_{ij} \mathbb{\Gamma}_k \mathbf{b}_i$, and $Y_{ij} \sim binomial(m_{ij} = 10, \varphi_{ij})$.

Wherever possible, I used the same parameter settings in different simulation settings so that the results would be easier to compare, and I provided a summary of the parameter settings in Table 3–2. Consider first the simulations in which the outcome follows a Poisson distribution. Beginning with the cases where $K = 2$, that is when the population consisted of two heterogeneous subpopulations. In these cases, I set the mixing proportions to $\pi_1 = 0.6$, and $\pi_2 = 0.4$. When there were two fixed effects considered ($p = 2$), $\boldsymbol{\beta}_1^\top = (-0.75, 0.35)$, and $\boldsymbol{\beta}_2^\top = (0.60, -0.50)$. In cases where I estimated five fixed effects ($p = 5$), $\boldsymbol{\beta}_1^\top = (-0.75, 0.35, 0.10, -0.40, 0.00)$, and $\boldsymbol{\beta}_2^\top = (0.60, -0.50, -0.35, -0.15, 0.00)$. For the random effects, when I estimated one random effect ($q = 1$), $\mathbb{\Gamma}_1^{*\top} = (0.80)$, and $\mathbb{\Gamma}_2^{*\top} = (0.25)$. In the case where I estimated two random effects ($q = 2$), $\mathbb{\Gamma}_1^{*\top} = (0.80, -0.15, 0.20)$, and $\mathbb{\Gamma}_2^{*\top} = (0.25, 0.00, 0.30)$.

Turning now to the case where $K = 3$. In these simulations, I set the mixing proportions to $\pi_1 = 0.5$, $\pi_2 = 0.3$, and $\pi_3 = 0.2$. Given two fixed effects ($p = 2$), I used $\boldsymbol{\beta}_1^\top = (-0.75, 0.35)$, $\boldsymbol{\beta}_2^\top = (0.60, -0.50)$, and $\boldsymbol{\beta}_3^\top = (0.45, 0.75)$. For the case where five fixed effects were present in the model ($p = 5$), $\boldsymbol{\beta}_1^\top = (-0.75, 0.35, 0.10, -0.40, 0.00)$, $\boldsymbol{\beta}_2^\top = (0.60, -0.50, -0.35, -0.15, 0.00)$, and $\boldsymbol{\beta}_3^\top = (0.45, 0.75, -0.65, 0.20, 0.00)$. When one random effect was present in the model ($q = 1$), $\mathbb{\Gamma}_1^{*\top} = (0.80)$, $\mathbb{\Gamma}_2^{*\top} = (0.25)$, and $\mathbb{\Gamma}_3^{*\top} = (0.40)$. If I estimated two random effects ($q = 2$), the parameter settings were $\mathbb{\Gamma}_1^{*\top} = (0.80, -0.15, 0.20)$, $\mathbb{\Gamma}_2^{*\top} = (0.25, 0.00, 0.30)$, and $\mathbb{\Gamma}_3^{*\top} = (0.40, 0.25, 0.10)$.

In contrast, when the outcome followed a binomial distribution, I used the following parameter settings. When $K = 2$, I set the mixing proportions to $\pi_1 = 0.6$, and $\pi_2 = 0.4$. Given two fixed effects ($p = 2$), $\boldsymbol{\beta}_1^\top = (-0.55, 0.85)$, and $\boldsymbol{\beta}_2^\top = (0.25, -0.50)$. Alternatively, when I estimated five fixed effects ($p = 5$), $\boldsymbol{\beta}_1^\top = (-0.55, 0.85, 1.25, -0.70, 0.00)$, and $\boldsymbol{\beta}_2^\top = (0.25, -0.50, 1.35, -0.20, 0.00)$. For the random effects, when I estimated one random effect ($q = 1$), $\mathbb{\Gamma}_1^{*\top} = (1.60)$, and $\mathbb{\Gamma}_2^{*\top} = (1.05)$. However, when two random effects were in the model ($q = 2$), $\mathbb{\Gamma}_1^{*\top} = (1.60, -0.45, 1.00)$, and $\mathbb{\Gamma}_2^{*\top} = (1.05, 0.00, 1.40)$.

In the more complex case, where $K = 3$ I set the mixing proportions to $\pi_1 = 0.5$, $\pi_2 = 0.3$, and $\pi_3 = 0.2$. When two fixed effects were present in the model ($p = 2$),

$\boldsymbol{\beta}_1^\top = (-0.55, 0.85)$, $\boldsymbol{\beta}_2^\top = (0.25, -0.50)$, and $\boldsymbol{\beta}_3^\top = (-0.75, 0.35)$. If the model had five fixed effects ($p = 5$), $\boldsymbol{\beta}_1^\top = (-0.55, 0.85, 1.25, -0.70, 0.00)$, $\boldsymbol{\beta}_2^\top = (0.25, -0.50, 1.35, -0.20, 0.00)$, and $\boldsymbol{\beta}_3^\top = (-0.75, 0.35, -0.50, 0.55, 0.00)$. If the model had one random effect ($q = 1$), $\mathbb{\Gamma}_1^{*\top} = (1.60)$, $\mathbb{\Gamma}_2^{*\top} = (1.05)$, and $\mathbb{\Gamma}_3^{*\top} = (1.45)$. Alternatively, if I included two random effects ($q = 2$), the parameter settings were $\mathbb{\Gamma}_1^{*\top} = (1.60, -0.45, 1.00)$, $\mathbb{\Gamma}_2^{*\top} = (1.05, 0.00, 1.40)$, and $\mathbb{\Gamma}_3^{*\top} = (1.45, 0.40, 1.30)$.

The choice of starting values is important in the estimation of FinMix GLMM parameters because a poor choice can lead to slow convergence or non-convergence in both EM, and Newton-Raphson. In the following simulations, I generated an initial value for each generated dataset. I set the initial values of the mixing proportions to $\pi_k = \frac{1}{K}$, and the off-diagonal values of $\mathbb{\Gamma}$ to 0. For the other values in $\boldsymbol{\theta}$, namely $\boldsymbol{\beta}$, and the diagonal values of $\mathbb{\Gamma}$, I generated the initial value as the true value plus a random draw from a uniform distribution i.e. $\boldsymbol{\theta}^{(0)} = \boldsymbol{\theta} + U$ where $U \sim Uniform(-0.5, 0.5)$. Since the diagonal elements of $\mathbb{\Gamma}$ must be non-negative, in cases where the true value of a diagonal element of $\mathbb{\Gamma}$ was less than 0.5, I generated the starting value from $Uniform(0.1, 1.1)$. I included more detail on the sensitivity of the algorithm to the starting values in Section F.2.

To generate each dataset, first, I generated the dataset with $n = 1000$. I saved the data for the first 500 subjects as the $n = 500$ dataset. Similarly, the first 250 subjects' data became the $n = 250$ dataset, and the first 100 subjects' data is the $n = 100$ dataset. I did this to make the datasets for each of the possible sample sizes more comparable.

Table 3–2: Parameter settings for MLE. Parameter values for both Poisson, and binomial outcomes, and each of the six simulation settings.

Outcome	*Poisson*						*Binomial*					
Parameter	1	2	3	4	5	6	7	8	9	10	11	12
π_1	0.6	0.6	0.6	0.5	0.5	0.5	0.6	0.6	0.6	0.5	0.5	0.5
π_2	0.4	0.4	0.4	0.3	0.3	0.3	0.4	0.4	0.4	0.3	0.3	0.3
π_3	–	–	–	0.2	0.2	0.2	–	–	–	0.2	0.2	0.2
β_{10}	-0.75	-0.75	-0.75	-0.75	-0.75	-0.75	-0.55	-0.55	-0.55	-0.55	-0.55	-0.55
β_{11}	0.35	0.35	0.35	0.35	0.35	0.35	0.85	0.85	0.85	0.85	0.85	0.85
β_{12}	–	0.10	0.10	–	0.10	0.10	–	1.25	1.25	–	1.25	1.25
β_{13}	–	-0.40	-0.40	–	-0.40	-0.40	–	-0.70	-0.70	–	-0.70	-0.70
β_{14}	–	0.00	0.00	–	0.00	0.00	–	0.00	0.00	–	0.00	0.00
Γ_{11}	0.80	0.80	0.80	0.80	0.80	0.80	1.60	1.60	1.60	1.60	1.60	1.60
Γ_{112}	–	–	-0.15	–	–	-0.15	–	–	-0.45	–	–	-0.45
Γ_{12}	–	–	0.20	–	–	0.20	–	–	1.00	–	–	1.00
β_{20}	0.60	0.60	0.60	0.60	0.60	0.60	0.25	0.25	0.25	0.25	0.25	0.25
β_{21}	-0.50	-0.50	-0.50	-0.50	-0.50	-0.50	-0.50	-0.50	-0.50	-0.50	-0.50	-0.50
β_{12}	–	-0.35	-0.35	–	-0.35	-0.35	–	1.35	1.35	–	1.35	1.35
β_{13}	–	-0.15	-0.15	–	-0.15	-0.15	–	-0.20	-0.20	–	-0.20	-0.20
β_{14}	–	0.00	0.00	–	0.00	0.00	–	0.00	0.00	–	0.00	0.00
Γ_{21}	0.25	0.25	0.25	0.25	0.25	0.25	1.05	1.05	1.05	1.05	1.05	1.05
Γ_{212}	–	–	0.00	–	–	0.00	–	–	0.00	–	–	0.00
Γ_{22}	–	–	0.30	–	–	0.30	–	–	1.40	–	–	1.40
β_{30}	–	–	–	0.45	0.45	0.45	–	–	–	-0.75	-0.75	-0.75
β_{31}	–	–	–	0.75	0.75	0.75	–	–	–	0.35	0.35	0.35
β_{32}	–	–	–	–	-0.65	-0.65	–	–	–	–	-0.50	-0.50
β_{33}	–	–	–	–	0.20	0.20	–	–	–	–	0.55	0.55
β_{34}	–	–	–	–	0.00	0.00	–	–	–	–	0.00	0.00
Γ_{31}	–	–	–	0.40	0.40	0.40	–	–	–	1.45	1.45	1.45
Γ_{312}	–	–	–	–	–	0.25	–	–	–	–	–	0.40
Γ_{32}	–	–	–	–	–	0.10	–	–	–	–	–	1.30

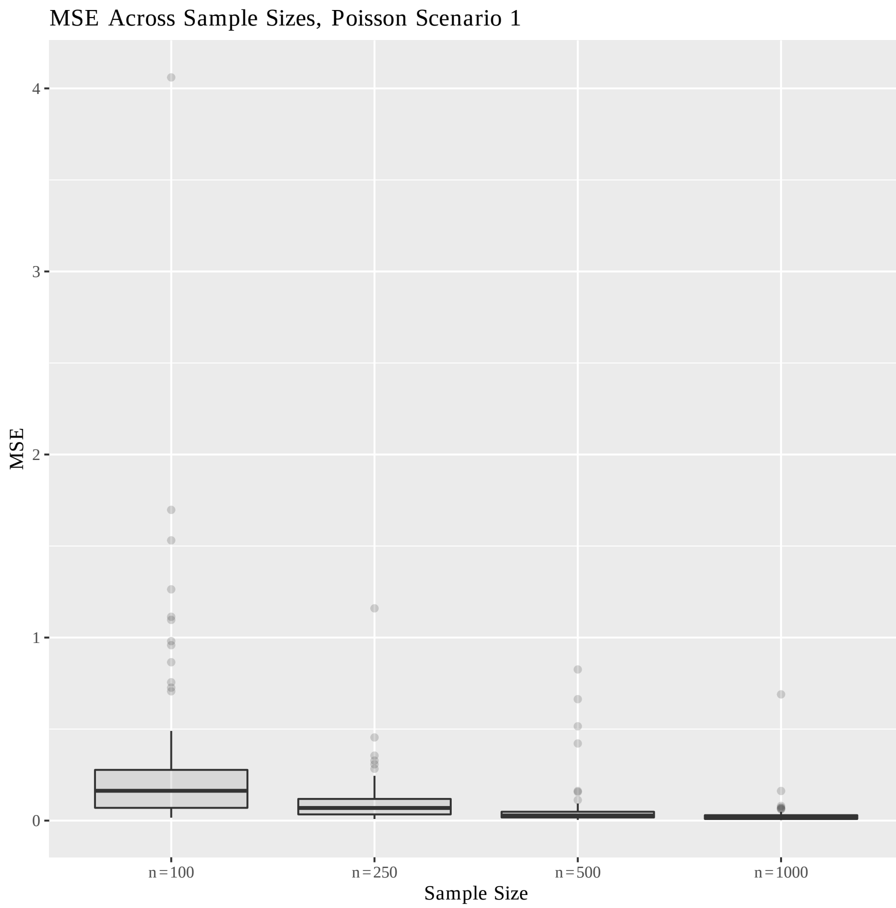

Figure 3–1: Mean Squared Error (MSE) across sample sizes, Poisson outcome, with $K = 2$, $p = 2$, and $q = 1$.

I summarised the results of these simulations in Tables E–1 to E–12, and reported the average bias, variance, and MSE for each parameter (multiplied by 100) in these tables. These simulations show that the MLE performs well in a variety of possible settings. The motivation for using the MLE is that it is a popular method of estimation. In addition, MLEs are consistent, efficient, and asymptotically follow a

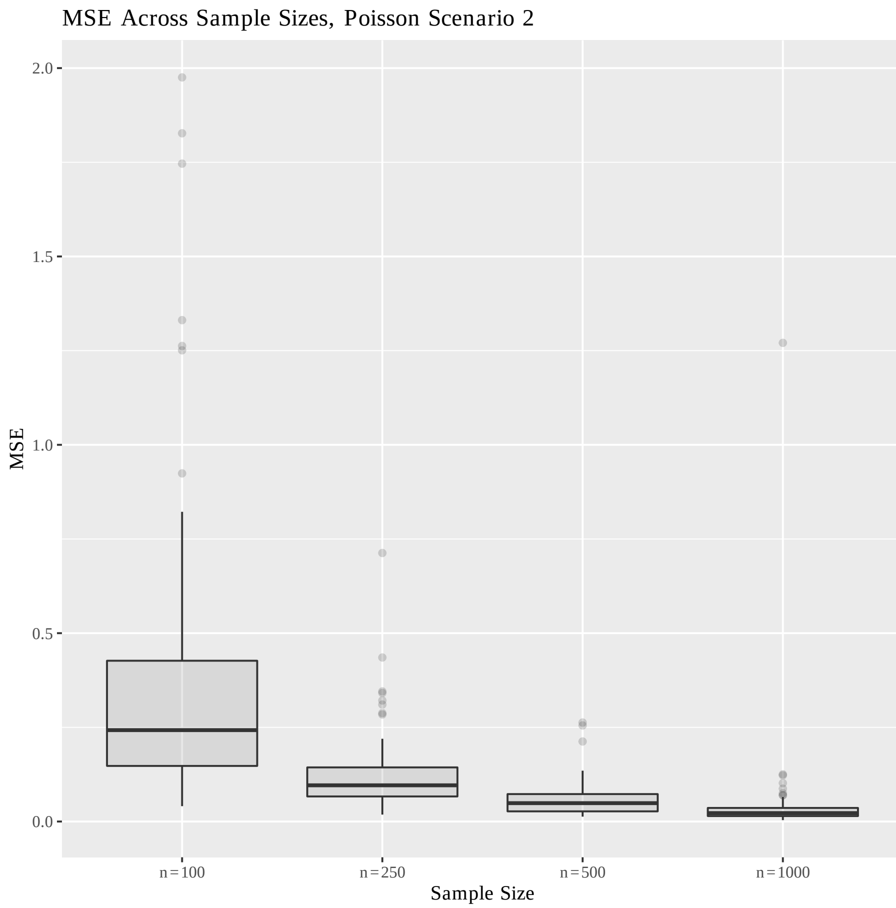

Figure 3–2: MSE across sample sizes, Poisson outcome, with $K = 2$, $p = 5$, and $q = 1$.

Gaussian distribution, and these simulations confirmed these theoretical properties. However, there were also unexpected findings from the simulation study.

The expected results that this simulation study confirmed were that bias, and variance of the estimates decreased with sample size, and the fewer parameters being estimated, the better the estimates. As the sample size increased, both the bias, and variance decreased, and thus the MSE also decreased. Similarly, the estimators for a

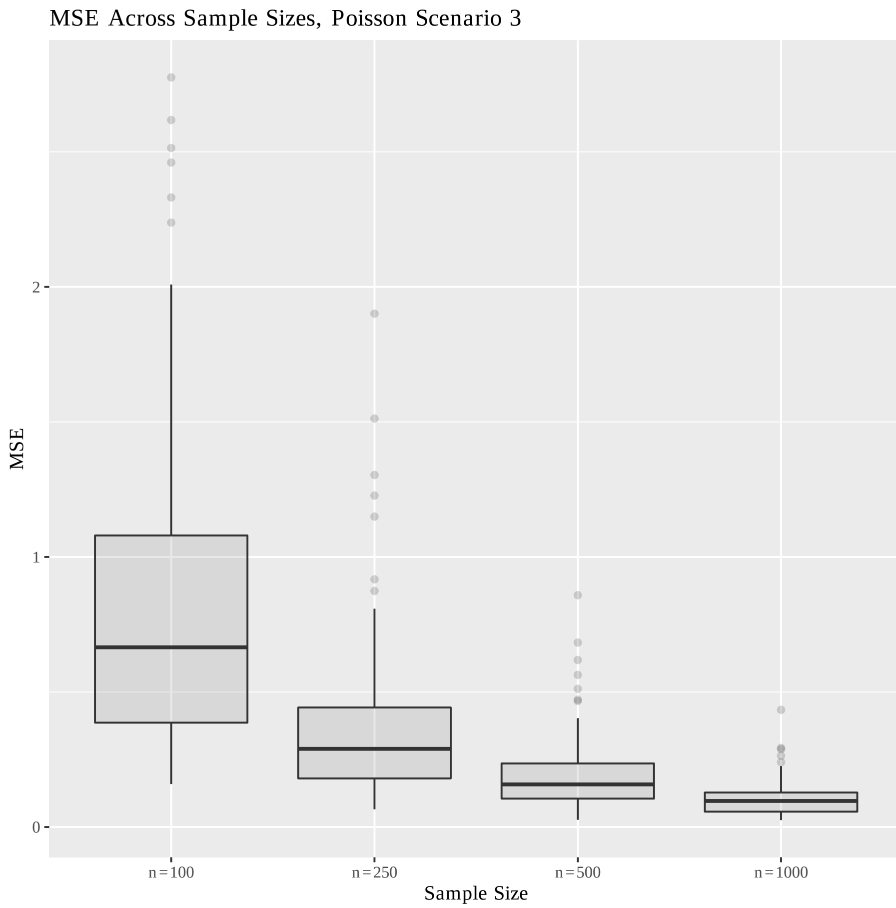

Figure 3–3: MSE across sample sizes, Poisson outcome, with $K = 2$, $p = 5$, and $q = 2$.

given subpopulation's parameters showed less variance, and bias, in general, the larger the subpopulation. That is, in the two subpopulation situation, the estimates in the larger subpopulation usually had smaller MSE than those in the smaller subpopulation regardless of the overall sample size n. While the largest sample size I considered in these simulations was $n = 1000$, I expect that this trend would continue in larger samples. Aside from the sample size comparison, the parameters that correspond to

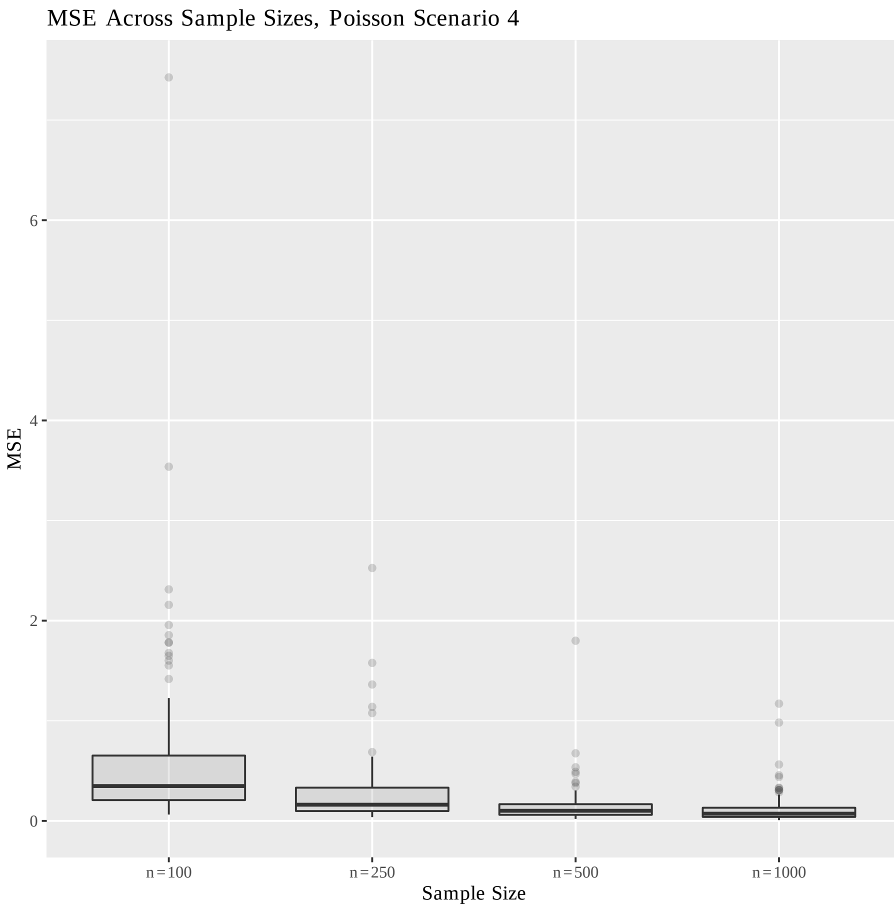

Figure 3–4: MSE across sample sizes, Poisson outcome, with $K = 3$, $p = 2$, and $q = 1$.

fixed effects for covariates that did not have a random effect exhibited the least bias, and variance, followed by fixed effects for covariates that had a random effect, then diagonal elements of the matrix $\mathbb{\Gamma}$, followed by lower-triangle elements of the matrix $\mathbb{\Gamma}$.

One surprising result of the simulations was the estimates in the binomial case. I expected that the estimates for a binomial outcome would be difficult to calculate,

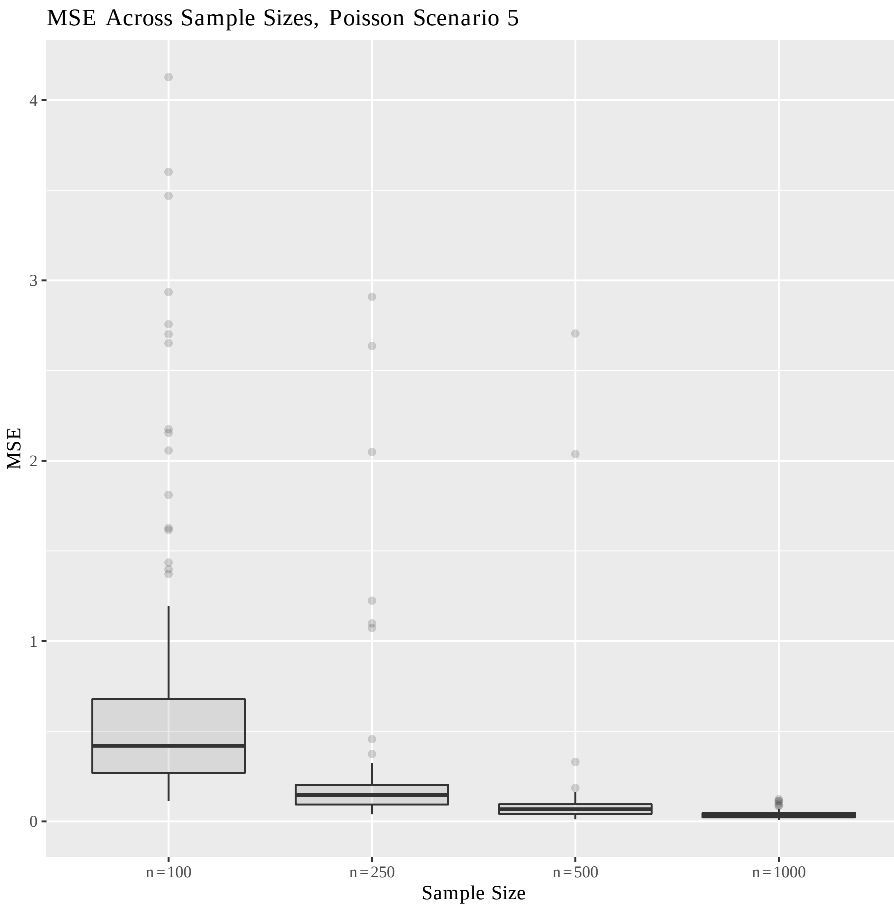

Figure 3–5: MSE across sample sizes, Poisson outcome, with $K = 5$, $p = 5$, and $q = 1$.

and that these estimates would show a large MSE. Many of the estimates in the binomial simulations showed smaller than expected bias, and variance. Another surprise from the simulation results was that the performance of the model under relatively small sample sizes was better than expected. In the $n = 100$, and $K = 3$ case, the expected number of subjects in the smallest subpopulation is just 20. It was

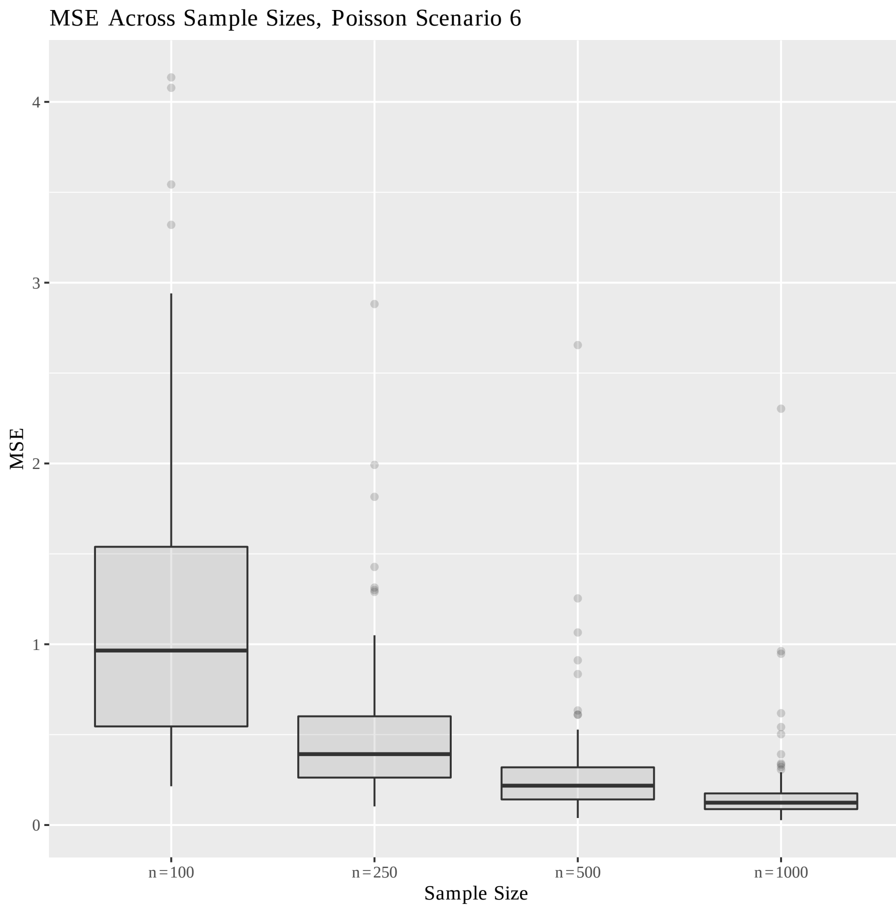

Figure 3–6: MSE across sample sizes, Poisson outcome, with $K = 3$, $p = 5$, and $q = 2$.

encouraging to see that even in this case, the estimation of parameters performed reasonably.

These simulation results suggest that this statistical model is applicable in a variety of situations. Because estimates of both a Poisson, and binomial outcome showed small MSE, outcomes that follow other exponential families could be explored. Given that the estimates in simulations where the outcome follows a binomial

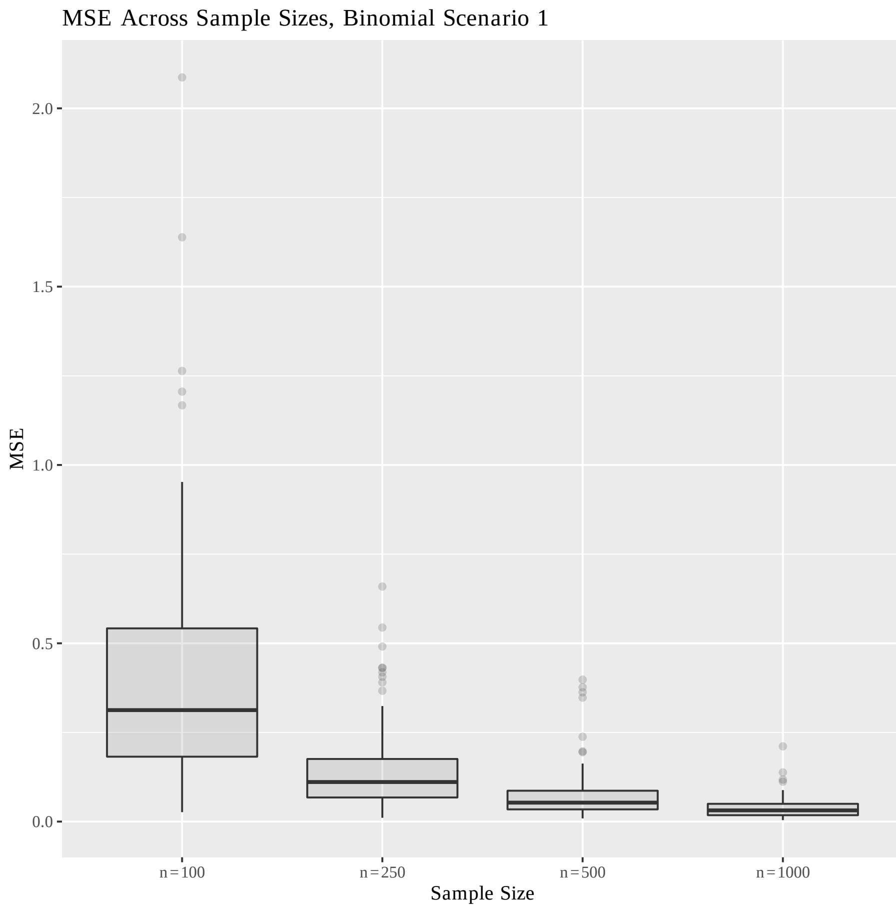

Figure 3–7: MSE across sample sizes, binomial outcome, with $K = 2$, $p = 2$, and $q = 1$.

distribution performed better than expected, one could apply this model with less hesitation to data with a binomial outcome. The performance of the model in small sample sizes is encouraging and makes this model more widely applicable. The fact that the estimation of the parameters corresponding to small subpopulations was possible indicates that such a model could be applied even if there are many distinct

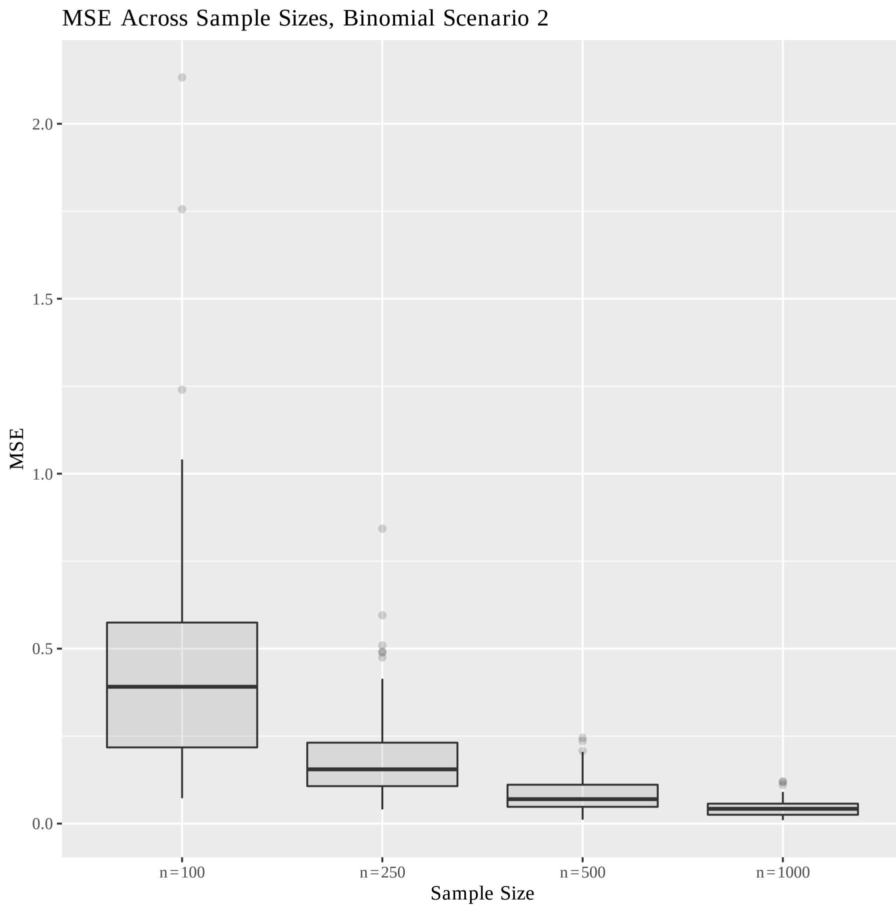

Figure 3–8: MSE across sample sizes, binomial outcome, with $K = 2$, $p = 5$, and $q = 1$.

subpopulations, or if a subpopulation has a small mixing proportion. This suggests that a FinMix GLMM is a widely applicable model for statistical analysis.

I included a few other possibilities of interest in Appendix F. I explored further to the case where different values of n_i were used in Section F.3. In the case where the outcome follows a binomial distribution, the value of m_{ij} need not be the same throughout. I explored this possibility in Section F.4. The MLE results for simulation

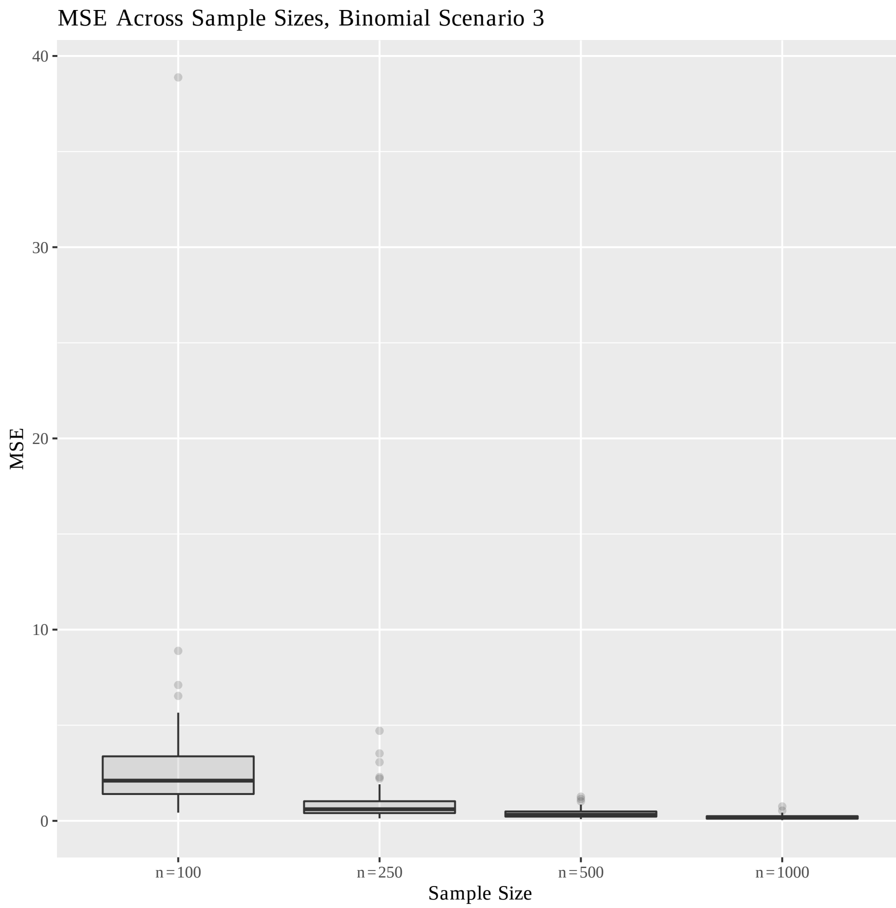

Figure 3–9: MSE across sample sizes, binomial outcome, with $K = 2$, $p = 5$, and $q = 2$.

scenarios from Chapter 4 are in Appendix H. These show the performance of the MLE in a variety of more complex settings.

3.5 Conclusion

In this chapter, I explored the form of the FinMix GLMM, including the likelihood equations, and details around identifiability. I also included an explanation of the derivations, and numerical calculations used to maximise the approximate likelihood. I

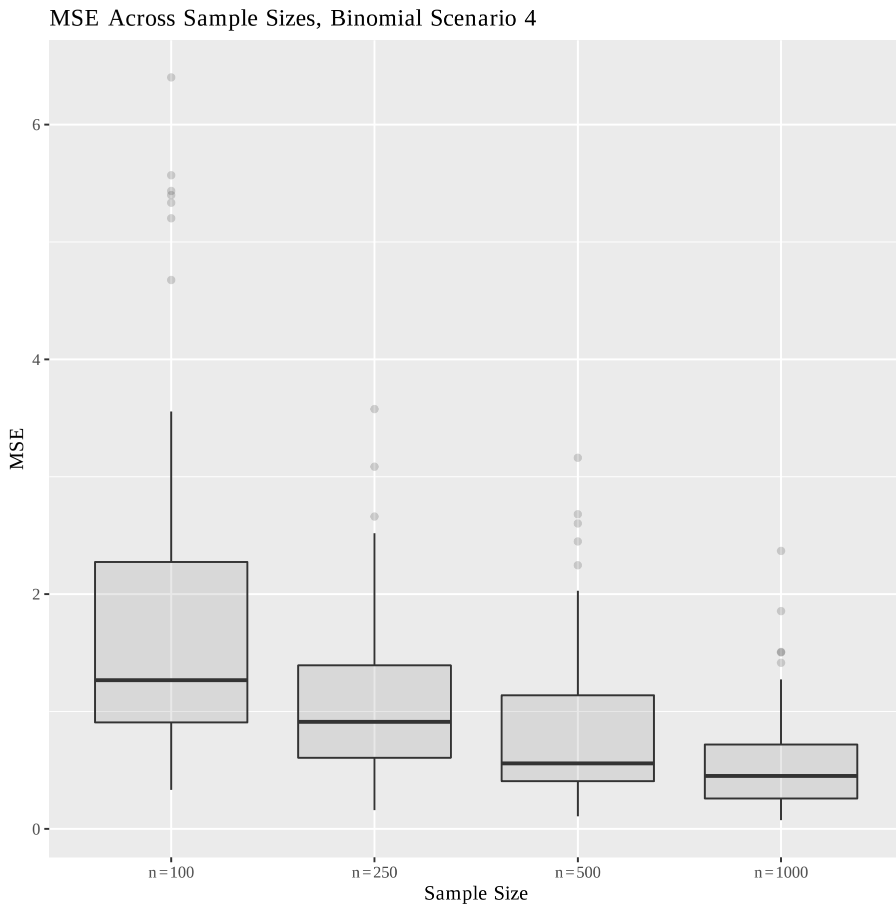

Figure 3–10: MSE across sample sizes, binomial outcome, with $K = 3$, $p = 2$, and $q = 1$.

then used these derivations in the implementation of an MCEM algorithm to estimate the MLE for both a Poisson or a binomial FinMix GLMM. Next, I performed a simulation study to verify, and illustrate that the proposed algorithm provides reasonable estimates in the MLE case. In the following chapter, Chapter 4, a penalised likelihood is used to select which variables are non-zero, and which variables are zero. In preparation for this, all simulations in this chapter with $p = 5$ contained one value

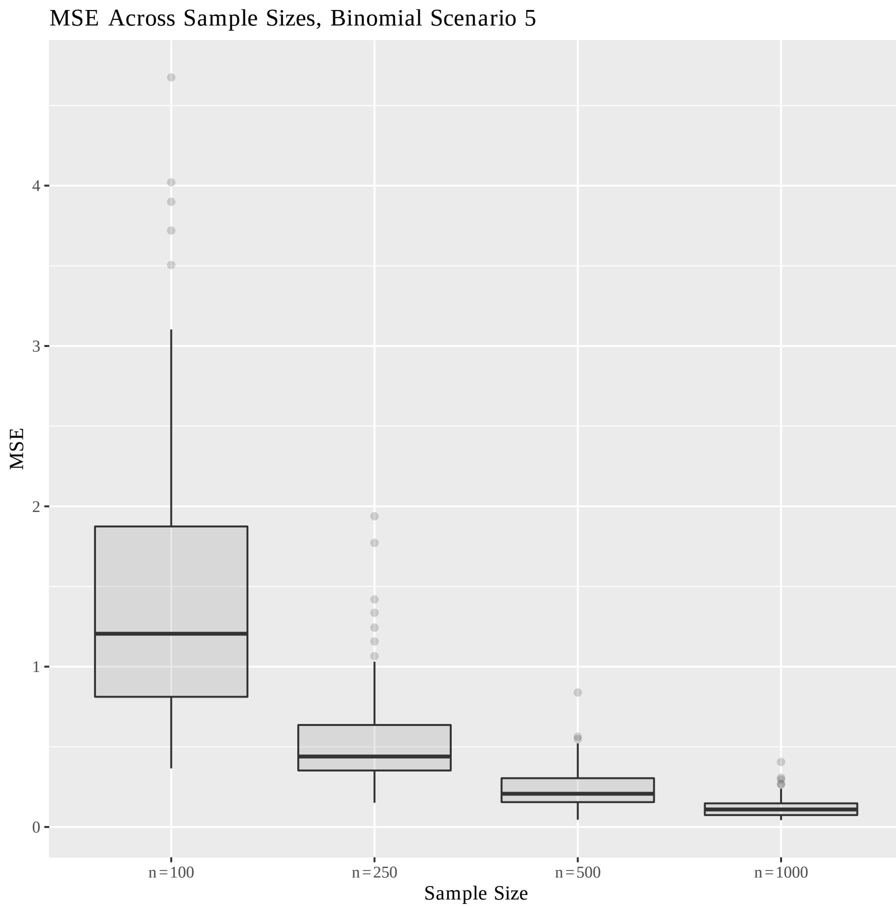

Figure 3–11: MSE across sample sizes, binomial outcome, with $K = 3$, $p = 5$, and $q = 1$.

of $\boldsymbol{\beta}$ for which the true value was 0 (the last value) in the vector $\boldsymbol{\beta}_k$. In addition, when $q = 2$ I set one of the off-diagonals in the matrix $\mathbb{\Gamma}$ to 0 as well. In this way, I considered the estimation of a parameter whose true value is 0, however, the case where the true value of a random effect is 0, and when the value of K is unknown are more complex, and I considered these in Section F.1.

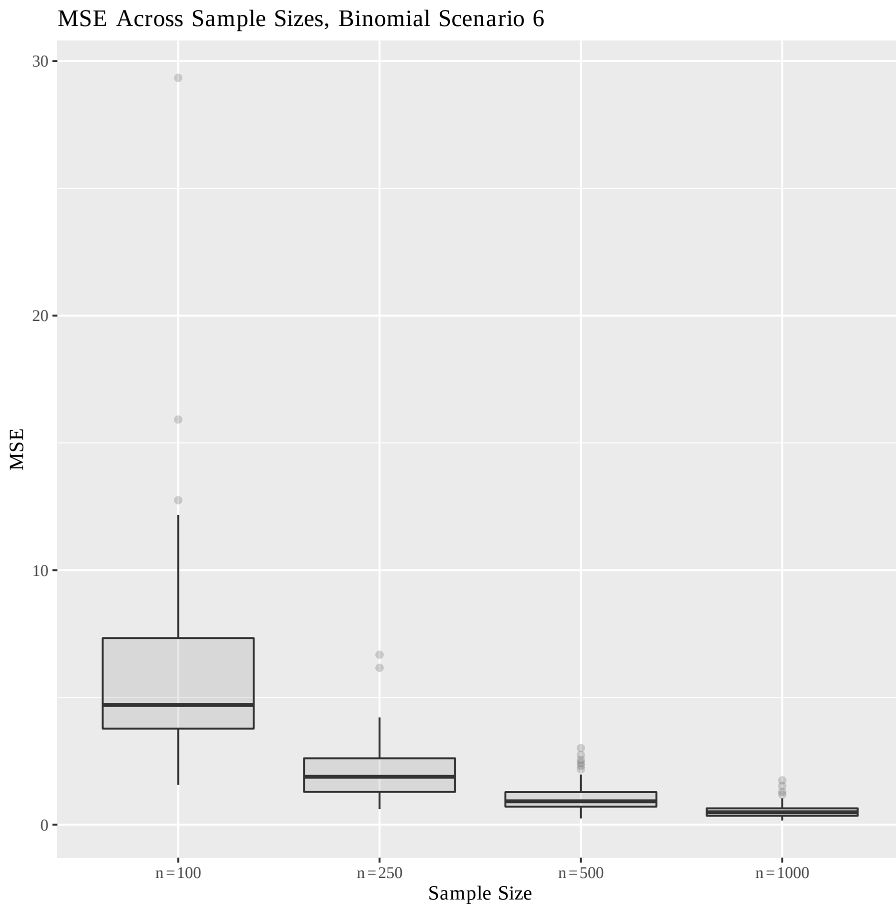

Figure 3–12: MSE across sample sizes, binomial outcome, with $K = 3$, $p = 5$, and $q = 2$.

CHAPTER 4
Objective Two: Optimisation of the Penalised Likelihood for Model Selection in Finite Mixtures of Generalised Linear Mixed-Effect Models

4.1 Introduction

In situations where the covariates in the model have been chosen a priori, the likelihood can be directly specified, and subsequently maximised, as seen in Chapter 3. However, there are many cases where a large number of covariates are under consideration, and a model with a small number of covariates is desired or the true model is assumed to have a small number of covariates. In this case, rather than the problem being solely one of estimating the values of the parameters, the problem also involves the selection of covariates or identifying which of the covariates to include in the model. The problem of variable selection is complex, and many solutions have been proposed. One popular approach to this problem is to add a penalty function to the likelihood equation and then maximise the resulting penalised likelihood. Using a penalised likelihood to perform variable selection provides a few advantageous properties, namely consistency, sparsity, and the oracle property, assuming that certain regularity conditions are satisfied. As such, it is this approach that I used in this thesis, and I explored three popular penalty functions for use in a Finite Mixture of Generalised Linear Mixed-Effect Model (FinMix GLMM). In this chapter, I combined the well known penalty functions of the Least Absolute Shrinkage and Selection Operator (LASSO) (Tibshirani, 1996), Adaptive Least Absolute Shrinkage and Selection Operator (ALASSO) (Zou, 2006), and the Smoothly Clipped Absolute Deviation (SCAD) (Fan and Li, 2001) penalties with the theory in the previous chapter to provide simultaneous model selection, and fitting of a FinMix GLMM.

The chapter proceeds with a discussion regarding changes to the parameterisation of the model in Section 4.2 including how to estimate the new parameters. Next, I considered penalty functions in Section 4.3 with special attention to LASSO, ALASSO, and SCAD. Section 4.4 discusses the asymptotic properties. I explored the numerical computation in Section 4.5 with appropriate changes to the algorithm to incorporate the penalty function. To verify the properties of the estimator, I undertook a simulation study, and I described the settings along with the results of this simulation study in Section 4.6 with detailed tables in Appendices G, and H. The last section, Section 4.7, provides a summary.

4.2 Reparameterisation of the Model

Reparameterisation of the model is only necessary when random effects are being penalised. I used the parameterisation shown in Chapter 3 in cases where I penalised only the fixed effects.

When more than one random effect is present in the model, both the variance of each of the random effects, and the covariances between them must be estimated. If random effects are to be penalised, I separated these parameters into two matrices such that $\mathbb{\Gamma} = \mathbb{d}\mathbb{C}$, where $\mathbb{C}$ is a lower triangle matrix with ones along the diagonal, and $\mathbb{d}$ is a diagonal matrix consisting of variances of the random effects. To simplify notation, I dropped the subscript k in this section. Recall that in the preceding chapter, these parameters were all contained in lower triangle matrices, denoted $\mathbb{\Gamma}_k$, and all estimated together. By selecting this parameterisation, if the variance of a random effect is set to be zero, the covariances of the random effects associated with that covariate are then automatically set to zero (by properties of diagonal matrices). This follows from the well known property of covariance that given any random variable W, and any constant $s \in \mathbb{R}$, $Cov(W, s) = 0$. To calculate the decomposition of $\mathbb{\Gamma}$, set $\mathbb{d}^* = diag(\Gamma_{11}, \Gamma_{22}, \ldots, \Gamma_{qq})$, and $C_{ih} = \Gamma_{ih}/\Gamma_{hh}$. Recall that $\mathbb{\Gamma}$ is a lower triangle matrix, so for $i < h, \Gamma_{ih} = 0$, thus $i < h \Rightarrow C_{ih} = 0$. Additionally, $C_{ii} = \Gamma_{ii}/\Gamma_{ii} = 1$ so the diagonal of $\mathbb{C}$ contains ones as desired. The vectorised versions of $\mathbb{d}$, and $\mathbb{C}$ are $\mathbb{d}^*$, and $\mathbb{C}^*$ respectively. All the regression coefficients for subpopulation k can then be represented by the vector $\underline{\boldsymbol{\theta}}_k = (\boldsymbol{\beta}_k^\top, \mathbb{d}_k^{*\top}, \mathbb{C}_k^{*\top})^\top$.

This reparameterisation of the model approach was also taken in Du et al. (2013), and Pan and Huang (2014), but this method is not unrivalled. Ibrahim et al. (2011) used group variable selection on each row of $\mathbb{\Gamma}$ instead and used the Euclidian norm of the row of $\mathbb{\Gamma}$ in the penalty function rather than considering the absolute value of a parameter. That is, for the e^{th} row of $\mathbb{\Gamma}$, $\mathbb{\Gamma}_e = [\Gamma_{e1}, \Gamma_{e2}, \ldots, \Gamma_{ee}, 0, 0, \ldots, 0] = \Gamma_{ee}[C_{e1}, C_{e2}, \ldots, C_{e(e-1)}, 1, 0, 0, \ldots, 0]$. Next consider the Euclidean norm of the row, $||\mathbb{\Gamma}_e|| = \sqrt{\Gamma_{ee}^2(C_{e1}^2 + C_{e2}^2 + \ldots + C_{e(e-1)}^2 + 1)}$. But $C_{e1}^2 + C_{e2}^2 + \ldots + C_{e(e-1)}^2 + 1 \geq 1 > 0$, thus $||\mathbb{\Gamma}_e|| = 0 \iff \Gamma_{ee}^2 = 0 \iff \Gamma_{ee} = 0$. Therefore, the group variable selection to select random effects as used by Ibrahim et al. (2011) is equivalent to the decomposition, and penalisation as shown in Du et al. (2013), and this thesis. Group variable selection is a considerably more complex generalisation of the penalised likelihood. As such, the reparameterisation of the model approach was chosen for this research.

4.3 Penalisation in the Generalised Linear Mixed-Effect Model Likelihood

In a FinMix GLMM, I applied the penalty to each of the Generalised Linear Mixed-Effect Model (GLMM)s separately. Building on the idea of maximum likelihood, the penalised maximum likelihood includes a penalty to decide which variables should be included in the model, and which should not be. The penalised likelihood function then takes the form

$$\ell^{\sharp}_{n\lambda_{nk}}(\boldsymbol{\theta}_k) = \ell_n(\boldsymbol{\theta}_k) - p_{\lambda_{nk}}(\boldsymbol{\theta}_k) \tag{4.1}$$

when just the fixed effects are penalised and

$$\ell^{\sharp}_{n\lambda_{nk}}(\underline{\boldsymbol{\theta}}_k) = \ell_n(\underline{\boldsymbol{\theta}}_k) - p_{\lambda_{nk}}(\underline{\boldsymbol{\theta}}_k)$$

when both fixed, and random effects are penalised, where $\ell_n(\boldsymbol{\theta}_k)$ or $\ell_n(\underline{\boldsymbol{\theta}}_k)$ is the likelihood function, and $p_{\lambda_{nk}}(\boldsymbol{\theta}_k)$ or $p_{\lambda_{nk}}(\underline{\boldsymbol{\theta}}_k)$ the penalty function.

I considered three penalisations below, the LASSO (Tibshirani, 1996), ALASSO (Zou, 2006), and SCAD (Fan and Li, 2001). While these are not the only possible penalty functions, these three are popular options.

As noted in Chapter 2, the penalty for the LASSO in a model with only fixed effects takes the form

$$p_{\lambda_{nk}}(\boldsymbol{\theta}_k) = \lambda_{nk} \sum_{h=1}^{p} |\beta_{kh}|$$

with

$$\frac{\partial p_{\lambda_{nk}}(\boldsymbol{\theta}_k)}{\partial \beta_{kh}} = \lambda_{nk} \text{sign}(\beta_{kh}).$$

Similarly, if I applied a penalty to both the fixed and random effects, following the work of Chen and Dunson (2003), and Bondell et al. (2010) the penalty becomes

$$p_{\lambda_{nk}}(\underline{\boldsymbol{\theta}}_k) = \lambda_{nk}\sum_{h=1}^{p}|\beta_{kh}| + \lambda_{nk}\sum_{h=1}^{q}|d_{kh}| = \lambda_{nk}\sum_{h=1}^{p}|\beta_{kh}| + \lambda_{nk}\sum_{h=1}^{q}d_{kh}$$

and the partial derivatives are

$$\frac{\partial p_{\lambda_{nk}}(\underline{\boldsymbol{\theta}}_k)}{\partial \beta_{kh}} = \lambda_{nk}\text{sign}(\beta_{kh})$$

and

$$\frac{\partial p_{\lambda_{nk}}(\underline{\boldsymbol{\theta}}_k)}{\partial d_{kh}} = \lambda_{nk}.$$

The second derivatives of the penalty are equal to zero, both with respect to β_{kh}, and d_{kh}.

The ALASSO penalty is similar to the LASSO penalty with the addition of weights w_h, more specifically, LASSO is ALASSO with all of the weights set to 1. However, this requires choosing both λ_{nk}, and w_h, which is sometimes considered to be a drawback of this method as this step can be computationally intensive (by generalised cross-validation for example). In this thesis, I chose the inverse of the Maximum Likelihood Estimation (MLE) estimate as the weight for a particular parameter. If only fixed effects are considered, this penalty is of the form

$$p_{\lambda_{nk}}(\boldsymbol{\theta}_k) = \lambda_{nk}\sum_{h=1}^{p} w_h|\beta_{kh}|$$

with

$$\frac{\partial p_{\lambda_{nk}}(\boldsymbol{\theta}_k)}{\partial \beta_{kh}} = \lambda_{nk}w_h\text{sign}(\beta_{kh}).$$

Similarly, if a penalty is applied to both the fixed and random effects, the penalty becomes

$$p_{\lambda_{nk}}(\underline{\boldsymbol{\theta}}_k) = \lambda_{nk}\sum_{h=1}^{p} w_h|\beta_{kh}| + \lambda_{nk}\sum_{h=1}^{q} w_{p+h}|d_{kh}| = \lambda_{nk}\sum_{h=1}^{p} w_h|\beta_{kh}| + \lambda_{nk}\sum_{h=1}^{q} w_{p+h}d_{kh}$$

and the partial derivatives are

$$\frac{\partial p_{\lambda_{nk}}(\underline{\boldsymbol{\theta}}_k)}{\partial \beta_{kh}} = \lambda_{nk} w_h \text{sign}(\beta_{kh})$$

and

$$\frac{\partial p_{\lambda_{nk}}(\underline{\boldsymbol{\theta}}_k)}{\partial d_{kh}} = \lambda_{nk} w_{p+h}.$$

Again, the second derivatives of the penalty are equal to zero, both with respect to β_{kh}, and d_{kh}.

One drawback of LASSO, and ALASSO is that all of the estimates are shrunk, regardless of their absolute value. In some circumstances, it is desirable to shrink larger estimates less, or not at all, and the SCAD penalty fulfils this requirement. The penalty for SCAD is not usually represented in a closed form, but rather the derivative is considered. If only fixed effects are considered then

$$\frac{\partial p_{\lambda_{nk}}(\boldsymbol{\theta}_k)}{\partial \beta_{kh}} = \lambda_{nk}\left\{ I(|\beta_{kh}| \leq \lambda_{nk}) + \frac{(a\lambda_{nk} - |\beta_{kh}|)_+}{(a-1)\lambda_{nk}} I(|\beta_{kh}| > \lambda_{nk}) \right\}.$$

In the case where both fixed and random effects are penalised,

$$\frac{\partial p_{\lambda_{nk}}(\underline{\boldsymbol{\theta}}_k)}{\partial \beta_{kh}} = \lambda_{nk}\left\{ I(|\beta_{kh}| \leq \lambda_{nk}) + \frac{(a\lambda_{nk} - |\beta_{kh}|)_+}{(a-1)\lambda_{nk}} I(|\beta_{kh}| > \lambda_{nk}) \right\}$$

and

$$\begin{aligned}\frac{\partial p_{\lambda_{nk}}(\underline{\boldsymbol{\theta}}_k)}{\partial d_{kh}} &= \lambda_{nk}\left\{I(|d_{kh}| \leq \lambda_{nk}) + \frac{(a\lambda_{nk} - |d_{kh}|)_+}{(a-1)\lambda_{nk}} I(|d_{kh}| > \lambda_{nk})\right\} \\ &= \lambda_{nk}\left\{I(d_{kh} \leq \lambda_{nk}) + \frac{(a\lambda_{nk} - d_{kh})_+}{(a-1)\lambda_{nk}} I(d_{kh} > \lambda_{nk})\right\}.\end{aligned}$$

Note that $(t)_+ = t \times I(t > 0)$, and $a > 2$. In this thesis, I used the value of $a = 3.7$ throughout, consistent with (Fan and Li, 2001). I did not consider other values for a, and choosing a using generalised cross-validation or other methods can be computationally intensive. This penalty also has second derivatives that equal zero.

In further derivations, I used the notation $\underline{\boldsymbol{\theta}}_k$ for the kth subpopulation, and equations as well as derivations of $\boldsymbol{\theta}_k$ were omitted. That is, the case where both fixed and random effects are penalised is presented. In addition, the hth element of $\underline{\boldsymbol{\theta}}_k$ is $\underline{\boldsymbol{\theta}}_{kh}$, and I denoted the penalty as $p_{\lambda_{nk}}(\underline{\boldsymbol{\theta}}_{kh})$. Furthermore, in all of the penalty functions considered, $p_{\lambda_{nk}}(\underline{\boldsymbol{\theta}}_{kh}) = p_{\lambda_{nk}}(|\underline{\boldsymbol{\theta}}_{kh}|)$, and thus $p_{\lambda_{nk}}(|\underline{\boldsymbol{\theta}}_{kh}|)$ is sometimes used in place of $p_{\lambda_{nk}}(\underline{\boldsymbol{\theta}}_{kh})$ in the literature.

In order to be of an appropriate size, it is pertinent to multiply the penalty function by n, the number of subjects in the sample (Tibshirani, 1996). Since, in my setting, I assumed the data comes from a finite mixture of models, the penalty function must take into account that the number of subjects in each of the distinct subpopulations, which is unknown. To scale this to an appropriate penalty for the size of each of the subpopulations, I multiplied the penalty function by $n_k = n \times \hat{\pi}_k$, the empirical estimate of $n \times \pi_k$, as shown in Khalili and Chen (2007). The penalised log-likelihood is thus,

$$\ell^{\sharp}_{n\lambda_{nk}}(\mathbf{\Theta}) \;=\; \ell_n(\mathbf{\Theta}) - n\sum_{k=1}^{K}\pi_k \times p_{\lambda_{nk}}(\underline{\boldsymbol{\theta}}_k)$$

for the entire dataset, and

$$\ell^{\sharp}_{n\lambda_{nk}}(\underline{\boldsymbol{\theta}}_k) \;=\; \ell_n(\underline{\boldsymbol{\theta}}_k, \tau_k) - n \times \pi_k \times p_{\lambda_{nk}}(\underline{\boldsymbol{\theta}}_k) \tag{4.2}$$

for each subpopulation. In some cases in the literature, the penalty parameter λ_{nk} is changed to λ^{α}_{nk}. While the code I wrote allows for this extension, I did not explore this possibility in detail in this thesis.

4.4 Asymptotic Properties

Given certain conditions on the model and penalty, the estimates calculated using Maximum penalised Likelihood Estimation (MPLE) possess many desirable asymptotic properties. The properties of interest are existence, consistency, sparsity, and that the distribution is asymptotically Gaussian. More information on the conditions, asymptotic properties, and proofs can be found in Appendix D.

4.5 Numerical Computation of the Penalised Maximum Likelihood Estimator

The estimation proceeds similarly to the unpenalised likelihood setting, using the Monte Carlo Expectation-Maximisation (MCEM) algorithm with the derivatives of Q updated with the relevant penalty. Specifically, the outer MCEM loop does not change, and only the M-step in the inner MCEM changes. The calculation of τ_{ki} does not change for any k or i, and the likelihood without a penalty function is used to calculate these values. The exact form of τ_{ki} is in Equation (3.1). Recall that I estimated both the subpopulation membership, and the mixing proportions in the outer MCEM, and therefore calculated them in the same manner regardless of the presence or form of the penalty function.

However, the calculation of τ_{ki} relies on $\hat{\boldsymbol{\beta}}_k$, and $\hat{\mathbb{\Gamma}}_k$, both of which I updated in the inner MCEM loop of the algorithm which takes into account the penalty function. Thus, while the penalty function does not explicitly appear in the calculations of the outer MCEM, I would not expect that the estimates for subpopulation membership and mixing proportions would be exactly identical regardless of the form of the penalty parameter.

4.5.1 Inner Monte Carlo Expectation-Maximisation

The inner loop of the MCEM is where I calculated most of the estimates in $\boldsymbol{\Theta}$. I calculated the parameter estimates for $\boldsymbol{\pi}$ in the outer MCEM, but the inner MCEM loops until the estimates for $\boldsymbol{\beta}_k$ and $\mathbb{\Gamma}_k$ reach convergence. Note that convergence of $\mathbb{\Gamma}_k$ is desired, not the convergence of $\mathbb{d}_k$, and $\mathbb{C}_k$. This is done because $\mathbb{\Gamma}_k$ contains the parameters of interest, and because of the reparameterisation, estimation of $\mathbb{d}_k$, and $\mathbb{C}_k$ could be less stable numerically. Recall that the penalty is only on the fixed and random effects, which are in $\boldsymbol{\beta}_k$, $\mathbb{d}_k$, and $\mathbb{C}_k$, and I estimated these in the inner MCEM. Therefore, I updated the inner MCEM to include the penalty. However, the penalty function may not be differentiable everywhere. Specifically, the LASSO, ALASSO, and SCAD penalties are not differentiable at 0. Therefore, I approximated the penalty function. There are many choices of approximating functions of varying functional forms. In this thesis, I chose a quadratic approximation as used by Fan and Li (2001), because it is a reasonably good approximation within a bounded set, while also being smooth, continuously differentiable, and providing ease of computation. Following the approach taken by Fan and Li (2001), I approximated the log-likelihood function with a Taylor polynomial, but I approximated the penalty function by a different quadratic function.

By using a quadratic approximation to the penalty function, the previously mentioned problems of singularity at the origin, and a lack of continuous second-order derivatives, or indeed first-order derivatives in some cases can be avoided. These issues exist among many possible penalty functions and this approach can be extended beyond the three penalty functions focused on here. I chose this approximation for three reasons. First, I approximated the likelihood with a quadratic function, so using a quadratic approximation for the penalty as well keeps the degree of the penalised likelihood consistent. Second, a quadratic polynomial function is continuous, continuously differentiable, and continuously integrable. Third, a quadratic approximation provides a reasonably close estimation of to the penalty function, given that one is near to the point around which the approximation is calculated.

Recall the weighted penalised likelihood Equation (4.2) where $\ell_n(\underline{\boldsymbol{\theta}}_k, \tau_k)$ is the likelihood equation for given values of $\underline{\boldsymbol{\theta}}_k$, and τ_k with $p_{\lambda_{nk}}(\underline{\boldsymbol{\theta}}_k)$ as the penalty function for the given values of $\underline{\boldsymbol{\theta}}_k$, and λ_{nk}. Recall from Section 3.3.3 that I provided the form of the approximate likelihood $Q(\underline{\boldsymbol{\theta}}_k, \tau_k)$ in the Poisson, and binomial cases, and since the specific values of $\mathbf{b}_{ki}$ are unknown, I used an approximation with L values for $\mathbf{b}_{ki}^{(l)}$. Given the form of the approximated likelihood, I considered the quadratic approximation to the penalty function next.

I assumed that the initial value of $\underline{\boldsymbol{\theta}}_k$, denoted as $\underline{\boldsymbol{\theta}}_k^{(0)}$, is close to the maximum of the weighted penalised likelihood equation $\ell^\sharp(\underline{\boldsymbol{\theta}}_k)$. I then considered the penalty function componentwise, that is, for each element of $\underline{\boldsymbol{\theta}}_k$ I approximated the penalty $p_{\lambda_{nk}}(\underline{\boldsymbol{\theta}}_{kh}) = p_{\lambda_{nk}}(|\underline{\boldsymbol{\theta}}_{kh}|)$. Using the chain rule,

$$\begin{aligned}
\frac{\partial p_{\lambda_{nk}}(\underline{\boldsymbol{\theta}}_{kh})}{\partial \underline{\boldsymbol{\theta}}_{kh}} &= \frac{\partial p_{\lambda_{nk}}(\underline{\boldsymbol{\theta}}_{kh})}{\partial \underline{\boldsymbol{\theta}}_{kh}} \text{sign}(\underline{\boldsymbol{\theta}}_{kh}) \\
&= \frac{\partial p_{\lambda_{nk}}(\underline{\boldsymbol{\theta}}_{kh})}{\partial \underline{\boldsymbol{\theta}}_{kh}} \times \frac{\underline{\boldsymbol{\theta}}_{kh}}{|\underline{\boldsymbol{\theta}}_{kh}|} \\
&\approx \left[\frac{\partial p_{\lambda_{nk}}(\underline{\boldsymbol{\theta}}_{kh}^{(0)})}{\partial \underline{\boldsymbol{\theta}}_{kh}} \times \frac{1}{|\underline{\boldsymbol{\theta}}_{kh}^{(0)}|}\right] \times \underline{\boldsymbol{\theta}}_{kh}.
\end{aligned}$$

However, this is an approximation to the derivative of the penalty function. Taking the anti-derivative of this approximation yields as an approximation for the penalty function,

$$p_{\lambda_{nk}}(|\underline{\boldsymbol{\theta}}_{kh}|) \approx p_{\lambda_{nk}}(|\underline{\boldsymbol{\theta}}_{kh}^{(0)}|) + \frac{1}{2}\left[\frac{\partial p_{\lambda_{nk}}(|\underline{\boldsymbol{\theta}}_{kh}^{(0)}|)}{\partial \underline{\boldsymbol{\theta}}_{kh}} \times \frac{1}{|\underline{\boldsymbol{\theta}}_{kh}^{(0)}|}\right] \{(\underline{\boldsymbol{\theta}}_{kh})^2 - (\underline{\boldsymbol{\theta}}_{kh}^{(0)})^2\}.$$

For ease of notation, let $U_{\lambda_{nk}}(\underline{\boldsymbol{\theta}}_k^{(s)}) = \Sigma_{\lambda_{nk}}(\underline{\boldsymbol{\theta}}_k^{(s)}) \times \underline{\boldsymbol{\theta}}_k^{(s)}$. Recall that $\varkappa$ is the number of parameters in each subpopulation, that is $\varkappa = length(\underline{\boldsymbol{\theta}}_k) = p + \frac{1}{2}q(q+1)$.

$$\Sigma_{\lambda_{nk}}(\underline{\boldsymbol{\theta}}_k^{(s)}) = diag\left(\frac{\partial p_{\lambda_{nk}}(|\underline{\boldsymbol{\theta}}_{k1}^{(s)}|)}{\partial \underline{\boldsymbol{\theta}}_{k1}} \frac{1}{|\underline{\boldsymbol{\theta}}_{k1}^{(s)}|}, \frac{\partial p_{\lambda_{nk}}(|\underline{\boldsymbol{\theta}}_{k2}^{(s)}|)}{\partial \underline{\boldsymbol{\theta}}_{k2}} \frac{1}{|\underline{\boldsymbol{\theta}}_{k2}^{(s)}|}, \ldots, \frac{\partial p_{\lambda_{nk}}(|\underline{\boldsymbol{\theta}}_{k\varkappa}^{(s)}|)}{\partial \underline{\boldsymbol{\theta}}_{k\varkappa}} \frac{1}{|\underline{\boldsymbol{\theta}}_{k\varkappa}^{(s)}|}\right).$$

In practice, to avoid dividing by 0 and for numerical stability, I added a small value (ϵ) to the denominator of all elements of $\Sigma_{\lambda_{nk}}(\underline{\boldsymbol{\theta}}_k^{(s)})$ as shown in Fan and Li (2001). I chose a small value for ϵ (specifically $\epsilon = 0.0001$) so it would have little impact on the bias of the parameter estimates. Thus,

$$\begin{aligned}\Sigma_{\lambda_{nk}}(\underline{\boldsymbol{\theta}}_k^{(s)}) &= diag\left(\frac{\partial p_{\lambda_{nk}}(|\underline{\boldsymbol{\theta}}_{k1}^{(s)}|)}{\partial \underline{\boldsymbol{\theta}}_{k1}}\frac{1}{|\underline{\boldsymbol{\theta}}_{k1}^{(s)}|+\epsilon}, \frac{\partial p_{\lambda_{nk}}(|\underline{\boldsymbol{\theta}}_{k2}^{(s)}|)}{\partial \underline{\boldsymbol{\theta}}_{k2}}\frac{1}{|\underline{\boldsymbol{\theta}}_{k2}^{(s)}|+\epsilon}, \ldots, \right.\\ &\left. \frac{\partial p_{\lambda_{nk}}(|\underline{\boldsymbol{\theta}}_{k\varkappa}^{(s)}|)}{\partial \underline{\boldsymbol{\theta}}_{k\varkappa}}\frac{1}{|\underline{\boldsymbol{\theta}}_{k\varkappa}^{(s)}|+\epsilon}\right)\end{aligned}$$

Given the penalised likelihood, and taking a second-order Taylor approximation of the penalised likelihood, the resulting equation to be maximised is:

$$\begin{aligned}\ell^{\sharp}_{n\lambda_{nk}}(\underline{\boldsymbol{\theta}}_k) &= \ell_n(\underline{\boldsymbol{\theta}}_k, \tau_k) - n\pi_k p_{\lambda_{nk}}(\underline{\boldsymbol{\theta}}_k)\\ &\approx Q_k(\underline{\boldsymbol{\theta}}_k^{(s)}) + Q'_k(\underline{\boldsymbol{\theta}}_k^{(s)})^\top(\underline{\boldsymbol{\theta}}_k^{(s+1)} - \underline{\boldsymbol{\theta}}_k^{(s)})\\ &+\frac{1}{2}(\underline{\boldsymbol{\theta}}_k^{(s+1)} - \underline{\boldsymbol{\theta}}_k^{(s)})^\top Q''_k(\underline{\boldsymbol{\theta}}_k^{(s)})(\underline{\boldsymbol{\theta}}_k^{(s+1)} - \underline{\boldsymbol{\theta}}_k^{(s)})\\ &-\frac{1}{2}n\underline{\boldsymbol{\theta}}_k^{(s+1)^\top}\Sigma_{\lambda_{nk}}(\underline{\boldsymbol{\theta}}_k^{(s)})\underline{\boldsymbol{\theta}}_k^{(s+1)}\end{aligned}$$

To maximise this approximation, I took the derivative, and set to 0, thus

$$Q'_k(\underline{\boldsymbol{\theta}}_k^{(s)}) + Q''_k(\underline{\boldsymbol{\theta}}_k^{(s)})(\underline{\boldsymbol{\theta}}_k^{(s+1)} - \underline{\boldsymbol{\theta}}_k^{(s)}) - n\Sigma_{\lambda_{nk}}(\underline{\boldsymbol{\theta}}_k^{(s)})\underline{\boldsymbol{\theta}}_k^{(s+1)} = 0$$

which I then rearranged to

$$\begin{aligned}\underline{\boldsymbol{\theta}}_k^{(s+1)} &= \underline{\boldsymbol{\theta}}_k^{(s)} - (Q''_k(\underline{\boldsymbol{\theta}}_k^{(s)}) - n\Sigma_{\lambda_{nk}})^{-1}(Q'_k(\underline{\boldsymbol{\theta}}_k^{(s)}) - n\Sigma_{\lambda_{nk}}(\underline{\boldsymbol{\theta}}_k^{(s)})\underline{\boldsymbol{\theta}}_k^{(s)})\\ &= \underline{\boldsymbol{\theta}}_k^{(s)} - (Q''_k(\underline{\boldsymbol{\theta}}_k^{(s)}) - n\Sigma_{\lambda_{nk}})^{-1}(Q'_k(\underline{\boldsymbol{\theta}}_k^{(s)}) - nU_{\lambda_{nk}}(\underline{\boldsymbol{\theta}}_k^{(s)})\end{aligned}$$

which I used in the Newton-Raphson algorithm. Following the work of Hunter and Li (2005), the root of this derivative is the maximiser of the likelihood. As in the MLE case, $(\frac{1}{2})^{\iota}$ can be added to the equation so that half step Newton-Raphson can be performed.

Once I defined the target function to be maximised, I applied this procedure to each of the three penalty functions (LASSO, ALASSO, and SCAD). Specifically, the matrix $\Sigma_{\lambda_{nk}}$, changes when the penalty changes.

In some cases, it is desirable to penalise only the fixed effects, and not the random effects. In these situations, I calculated the penalty for the fixed effects as above, but set the penalty for the random effects to 0. Additionally, as shown in Tibshirani (1996), I did not penalise the intercept (β_{k0}), and as in Du et al. (2013) I did not penalise C_{kih}. Thus, I set the corresponding elements of the diagonal of $\Sigma_{\lambda_{nk}}(\underline{\boldsymbol{\theta}}_k^{(s)})$ to zero. I summarised the non-zero diagonal elements of $\Sigma_{\lambda_{nk}}$ evaluated at $\underline{\boldsymbol{\theta}}_k^{(s)}$ in the following table.

Table 4–1: Non-zero diagonal element of $\Sigma_{\lambda_{nk}}$

Penalty	Value
LASSO	$\frac{\lambda_{nk}}{\lvert\underline{\boldsymbol{\theta}}_{kh}^{(s)}\rvert+\epsilon}$
ALASSO	$\frac{\lambda_{nk} w_h}{\lvert\underline{\boldsymbol{\theta}}_{kh}^{(s)}\rvert+\epsilon}$
SCAD	$\lambda_{nk}\left\{I(\lvert\underline{\boldsymbol{\theta}}_{kh}\rvert \leq \lambda_{nk}) + \frac{(a\lambda_{nk}-\lvert\underline{\boldsymbol{\theta}}_{kh}\rvert)_+}{(a-1)\lambda_{nk}} I(\lvert\underline{\boldsymbol{\theta}}_{kh}\rvert > \lambda_{nk})\right\} \frac{1}{\lvert\underline{\boldsymbol{\theta}}_{kh}^{(s)}\rvert+\epsilon}$

4.6 Simulation Study

The following table describes the outline for the simulation study. Keeping in line with the MLE simulations, I performed simulations for both Poisson and binomial outcomes, and for mixtures of 2, and 3 subpopulations. I varied the number of fixed and random effects (both zero and non-zero), and considered four sample sizes $(100, 250, 500, 1000)$ for the simulations in this chapter. I ran each simulation 100 times.

In all cases, I used a multinomial distribution to generate subpopulation mem-

Table 4–2: Simulation settings for MPLE

Outcome	K	n	Number of $\boldsymbol{\beta} \neq 0$	Number of $\boldsymbol{\beta} = 0$	Number of $\mathbb{d} \neq 0$	Number of $\mathbb{d} = 0$	$Dim(\Theta)$
Poisson	2	$100, 250, 500, 1000$	2	5	2	0	27
Poisson	2	$100, 250, 500, 1000$	2	15	2	0	47
Poisson	2	$100, 250, 500, 1000$	4	15	2	3	69
Poisson	3	$100, 250, 500, 1000$	2	5	2	0	41
Poisson	3	$100, 250, 500, 1000$	2	15	2	0	71
Poisson	3	$100, 250, 500, 1000$	4	15	2	3	104
Binomial	2	$100, 250, 500, 1000$	2	5	2	0	27
Binomial	2	$100, 250, 500, 1000$	2	15	2	0	47
Binomial	2	$100, 250, 500, 1000$	4	15	2	3	69
Binomial	3	$100, 250, 500, 1000$	2	5	2	0	41
Binomial	3	$100, 250, 500, 1000$	2	15	2	0	71
Binomial	3	$100, 250, 500, 1000$	4	15	2	3	104

bership. To include an intercept in the model, the first column of the matrix $\mathbb{X}$ is of 1s. I generated all the covariates independently, except for X_2, and X_3, where $Corr(X_2, X_3) = 0.5$. I introduced this collinearity to be more analogous to real data. In the following section, $i \in \{1, 2, \ldots, n\}$, and $j \in \{1, 2, \ldots, n_i\}$,. I generated the data as follows:

$$\begin{aligned}
x_{ij1} &= 1 \forall i \forall j; \\
x_{ij(2-p)} &\sim \mathit{Multivariate\ Gaussian}(\boldsymbol{\mu}, \boldsymbol{\Sigma}) \forall i \forall j; \\
\boldsymbol{\mu} &= \begin{bmatrix} 0 \\ 0 \\ \vdots \\ 0 \end{bmatrix}; \\
\boldsymbol{\Sigma} &= \begin{bmatrix} 1 & 0.5 & 0 & \dots & 0 \\ 0.5 & 1 & 0 & \dots & 0 \\ 0 & 0 & 1 & \dots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \dots & 1 \end{bmatrix}.
\end{aligned}$$

As before, for any variable for which there is a random effect, there is also a fixed effect, therefore columns of $\mathbb{Z}$ are also columns of $\mathbb{X}$. I again generated the random effects $\mathbf{b}_i$ from a standard Gaussian distribution. Given the values of $\mathbb{X}_i$, $\mathbb{Z}_i$, the group membership, $\mathbf{b}_i$, $\boldsymbol{\beta}_k$, and $\Gamma_k = \mathbb{d}_k \mathbb{C}_k$, I generated the outcome variables. In the Poisson case, ξ_{ij} was calculated as $\log(\xi_{ij}) = \mathbf{x}_{ij}\boldsymbol{\beta}_k + \mathbf{z}_{ij}\Gamma_k \mathbf{b}_i$, and then $Y_{ij} \sim \mathit{Poisson}(\xi_{ij})$. In the binomial case, $\text{logit}(\varphi_{ij}) = \mathbf{x}_{ij}\boldsymbol{\beta}_k + \mathbf{z}_{ij}\Gamma_k \mathbf{b}_i$ and $Y_{ij} \sim \mathit{binomial}(m_{ij} = 10, \varphi_{ij})$. I added additional parameters with a true value of zero to the vectors, and by including several irrelevant covariates I was able to explore variable selection through simulation.

As I presented Chapter 3, where possible, I used the same parameter settings across simulation settings, and a summary of the parameter settings can be found in Tables 4–3 to 4–5. Consider first the Poisson simulations where $K = 2$. I set the mixing proportions to $\pi_1 = 0.6$ and $\pi_2 = 0.4$. When I considered seven fixed effects ($p = 7$), $\boldsymbol{\beta}_1^\top = (0.65, 0.30, 0.00, 0.00, \ldots, 0.00)$ and $\boldsymbol{\beta}_2^\top = (0.20, -0.45, 0.00, 0.00, \ldots, 0.00)$, if I estimated seventeen fixed effects ($p = 17$), $\boldsymbol{\beta}_1^\top = (0.65, 0.30, 0.00, 0.00, \ldots, 0.00)$ and $\boldsymbol{\beta}_2^\top = (0.20, -0.45, 0.00, 0.00, \ldots, 0.00)$, and in the case where I included nineteen fixed effect parameters in the model ($p = 19$), $\boldsymbol{\beta}_1^\top = (0.65, 0.30, 0.15, 0.35, 0.00, 0.00, \ldots, 0.00)$ and $\boldsymbol{\beta}_2^\top = (0.20, -0.45, -0.10, 0.25, 0.00, 0.00, \ldots, 0.00)$. For the random effects, when I estimated two random effects ($q = 2$), $\mathbb{\Gamma}_1^{*\top} = (0.30, -0.25, 0.10)$ and $\mathbb{\Gamma}_2^{*\top} = 0.35, 0.20, 0.15)$. In the case where I estimated five random effects ($q = 5$), $\mathbb{\Gamma}_1^{*\top} = (0.30, -0.25, 0.10, 0.00, 0.00, \ldots, 0.00)$ and $\mathbb{\Gamma}_2^{*\top} = (0.35, 0.20, 0.15, 0.00, 0.00, \ldots, 0.00)$.

Turning now to the case where $K = 3$. In these simulations, I set the mixing proportions to $\pi_1 = 0.5$, $\pi_2 = 0.3$, and $\pi_3 = 0.2$. Given seven fixed effects ($p = 7$), I used $\boldsymbol{\beta}_1^\top = (0.65, 0.30, 0.00, 0.00, \ldots, 0.00)$, $\boldsymbol{\beta}_2^\top = (0.20, -0.45, 0.00, 0.00, \ldots, 0.00)$, and $\boldsymbol{\beta}_3^\top = (1.00, 0.15, 0.00, 0.00, \ldots, 0.00)$. For the case where seventeen fixed effects were present in the model ($p = 17$), $\boldsymbol{\beta}_1^\top = (0.65, 0.30, 0.00, 0.00, \ldots, 0.00)$, $\boldsymbol{\beta}_2^\top = (0.20, -0.45, 0.00, 0.00, \ldots, 0.00)$, $\boldsymbol{\beta}_3^\top = (1.00, 0.15, 0.00, 0.00, \ldots, 0.00)$. When I estimated nineteen fixed effects ($p = 19$), $\boldsymbol{\beta}_1^\top = (0.65, 0.30, 0.15, 0.35, 0.00, 0.00, \ldots, 0.00)$, $\boldsymbol{\beta}_2^\top = (0.20, -0.45, -0.10, 0.25, 0.00, 0.00, \ldots, 0.00)$, and $\boldsymbol{\beta}_3^\top = (1.00, 0.15, -0.65, -0.15, 0.00, 0.00, \ldots, 0.00)$. When two random effects were present in the model ($q = 2$), $\mathbb{\Gamma}_1^{*\top} = (0.30, -0.25, 0.10)$, $\mathbb{\Gamma}_2^{*\top} = (0.35, 0.20, 0.15)$, and $\mathbb{\Gamma}_3^{*\top} = (0.25, 0.00, 0.20)$. In the case where I estimated five random effects ($q = 5$), the parameter settings were $\mathbb{\Gamma}_1^{*\top} = (0.30, -0.25, 0.10, 0.00, 0.00, \ldots, 0.00)$, $\mathbb{\Gamma}_2^{*\top} = (0.35, 0.20, 0.15, 0.00, 0.00, \ldots, 0.00)$, and $\mathbb{\Gamma}_3^{*\top} = (0.25, 0.00, 0.20, 0.00, 0.00, \ldots, 0.00)$.

In contrast, when the outcome follows a binomial distribution, I used the following parameter settings. When $K = 2$, I again set the mixing proportions to $\pi_1 = 0.6$ and $\pi_2 = 0.4$. Given seven fixed effects ($p = 7$), $\boldsymbol{\beta}_1^\top = (0.95, 0.60, 0.00, 0.00, \ldots, 0.00)$ and $\boldsymbol{\beta}_2^\top = (-0.85, -0.15, 0.00, 0.00, \ldots, 0.00)$. Alternatively, when I estimated seventeen fixed effects ($p = 17$), $\boldsymbol{\beta}_1^\top = (0.95, 0.60, 0.00, 0.00, \ldots, 0.00)$ and $\boldsymbol{\beta}_2^\top = (-0.85, -0.15, 0.00, 0.00, \ldots, 0.00)$. When the number of fixed effects was increased to nineteen ($p = 19$), $\boldsymbol{\beta}_1^\top = (0.95, 0.60, -0.65, -0.25, 0.00, 0.00, \ldots, 0.00)$ and $\boldsymbol{\beta}_2^\top = (-0.85, -0.15, -0.75, 0.10, 0.00, 0.00, \ldots, 0.00)$. For the random effects, when I estimated two random effects ($q = 2$), $\mathbb{\Gamma}_1^{*\top} = (0.95, 0.85, 1.15)$ and $\mathbb{\Gamma}_2^{*\top} = (0.70, -0.70, 0.80)$. However, when I included five random effects in the model ($q = 5$), $\mathbb{\Gamma}_1^{*\top} = (0.95, 0.85, 1.15, 0.00, 0.00, \ldots, 0.00)$ and $\mathbb{\Gamma}_2^{*\top} = (0.70, -0.70, 0.80, 0.00, 0.00, \ldots, 0.00)$.

In the more complex case, where $K = 3$, I set the mixing proportions to $\pi_1 = 0.5$, $\pi_2 = 0.3$, and $\pi_3 = 0.2$. For a model with seven fixed effects ($p = 7$), $\boldsymbol{\beta}_1^\top = (0.95, 0.60, 0.00, 0.00, \ldots, 0.00)$, $\boldsymbol{\beta}_2^\top = (-0.85, -0.15, 0.00, 0.00, \ldots, 0.00)$, and $\boldsymbol{\beta}_3^\top = (-0.30, -0.90, 0.00, 0.00, \ldots, 0.00)$. If the model has seventeen fixed effects ($p = 17$), $\boldsymbol{\beta}_1^\top = (0.95, 0.60, 0.00, 0.00, \ldots, 0.00)$, $\boldsymbol{\beta}_2^\top = (-0.85, -0.15, 0.00, 0.00, \ldots, 0.00)$, and $\boldsymbol{\beta}_3^\top = (-0.30, -0.90, 0.00, 0.00, \ldots, 0.00)$. When I estimated nineteen fixed effects ($p = 19$), $\boldsymbol{\beta}_1^\top = (0.95, 0.60, -0.65, -0.25, 0.00, 0.00, \ldots, 0.00)$, $\boldsymbol{\beta}_2^\top = (-0.85, -0.15, -0.75, 0.10, 0.00, 0.00, \ldots, 0.00)$, and $\boldsymbol{\beta}_3^\top = (-0.30, -0.90, 0.80, -0.25, 0.00, 0.00, \ldots, 0.00)$. If the true model had two random effects ($q = 2$), $\mathbb{\Gamma}_1^{*\top} = (0.95, 0.85, 1.15)$, $\mathbb{\Gamma}_2^{*\top} = (0.70, -0.70, 0.80)$, and $\mathbb{\Gamma}_3^{*\top} = (1.75, 0.00, 0.85)$. Alternatively, when I included five random effects ($q = 5$), the parameter settings were $\mathbb{\Gamma}_1^{*\top} = (0.95, 0.85, 1.15, 0.00, 0.00, \ldots, 0.00)$, $\mathbb{\Gamma}_2^{*\top} = (0.70, -0.70, 0.80, 0.00, 0.00, \ldots, 0.00)$, and $\mathbb{\Gamma}_3^{*\top} = (1.75, 0.00, 0.85, 0.00, 0.00, \ldots, 0.00)$.

The choice of starting values was again important. I produced the initial values for the algorithm independently for each generated dataset. I set the initial values of the mixing proportions to $\pi_k = \frac{1}{K}$, and the off-diagonal values of $\mathbb{\Gamma}$ to 0. For the other values in $\boldsymbol{\theta}$, namely $\boldsymbol{\beta}$, and the diagonal values of $\mathbb{\Gamma}$, I generated the initial value as the true value plus a random draw from a uniform distribution ie. $\boldsymbol{\theta}^{(0)} = \boldsymbol{\theta} + U$ where $U \sim Uniform(-0.5, 0.5)$. If the true value of the diagonal of $\mathbb{\Gamma}$ was 0 or was under 0.5, I used a uniform distribution between 0.1, and 1.1 instead to guarantee that I did not generate a negative starting value. I generated initial values for $\mathbb{\Gamma}_k$ rather than initial values for $\mathbb{d}_k$, and $\mathbb{C}_k$ to be consistent with Chapter 3, and because I used $\mathbb{\Gamma}_k$ in the outer Expectation-Maximisation (EM). From these initial values, I calculated the MLE for each dataset. I then used the MLE as the starting value for the MPLE algorithm. The MLE is a popular choice as the starting value, especially for ALASSO (Pan and Shang, 2018).

To be consistent with the MLE case, to generate each dataset, I first generated the dataset with $n = 1000$. I saved the data for the first 500 subjects as the $n = 500$ dataset. Similarly, the first 250 subjects' data became the $n = 250$ dataset, and the first 100 subjects' data is the $n = 100$ dataset. Again, the goal was to make the datasets for each of the possible sample sizes easily comparable.

There are a variety of ways to choose the tuning parameter, and different values of the tuning parameter will result in different models. In this thesis, I generated

Table 4–3: Parameter settings for MPLE. Parameter values for both Poisson and binomial outcomes, and each of the six simulation settings, subpopulation 1.

Outcome	*Poisson*						*Binomial*					
Parameter	1	2	3	4	5	6	7	8	9	10	11	12
π_1	0.6	0.6	0.6	0.5	0.5	0.5	0.6	0.6	0.6	0.5	0.5	0.5
β_{10}	0.65	0.65	0.65	0.65	0.65	0.65	0.95	0.95	0.95	0.95	0.95	0.95
β_{11}	0.30	0.30	0.30	0.30	0.30	0.30	0.60	0.60	0.60	0.60	0.60	0.60
β_{12}	0.00	0.00	0.15	0.00	0.00	0.15	0.00	0.00	-0.65	0.00	0.00	-0.65
β_{13}	0.00	0.00	0.35	0.00	0.00	0.35	0.00	0.00	-0.25	0.00	0.00	-0.25
β_{14}	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
β_{15}	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
β_{16}	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
β_{17}	–	0.00	0.00	–	0.00	0.00	–	0.00	0.00	–	0.00	0.00
β_{18}	–	0.00	0.00	–	0.00	0.00	–	0.00	0.00	–	0.00	0.00
β_{19}	–	0.00	0.00	–	0.00	0.00	–	0.00	0.00	–	0.00	0.00
β_{110}	–	0.00	0.00	–	0.00	0.00	–	0.00	0.00	–	0.00	0.00
β_{111}	–	0.00	0.00	–	0.00	0.00	–	0.00	0.00	–	0.00	0.00
β_{112}	–	0.00	0.00	–	0.00	0.00	–	0.00	0.00	–	0.00	0.00
β_{113}	–	0.00	0.00	–	0.00	0.00	–	0.00	0.00	–	0.00	0.00
β_{114}	–	0.00	0.00	–	0.00	0.00	–	0.00	0.00	–	0.00	0.00
β_{115}	–	0.00	0.00	–	0.00	0.00	–	0.00	0.00	–	0.00	0.00
β_{116}	–	0.00	0.00	–	0.00	0.00	–	0.00	0.00	–	0.00	0.00
Γ_{11}	0.30	0.30	0.30	0.30	0.30	0.30	0.95	0.95	0.95	0.95	0.95	0.95
Γ_{112}	-0.25	-0.25	-0.25	-0.25	-0.25	-0.25	0.85	0.85	0.85	0.85	0.85	0.85
Γ_{12}	0.10	0.10	0.10	0.10	0.10	0.10	1.15	1.15	1.15	1.15	1.15	1.15
Γ_{113}	–	–	0.00	–	–	0.00	–	–	0.00	–	–	0.00
Γ_{123}	–	–	0.00	–	–	0.00	–	–	0.00	–	–	0.00
Γ_{13}	–	–	0.00	–	–	0.00	–	–	0.00	–	–	0.00
Γ_{114}	–	–	0.00	–	–	0.00	–	–	0.00	–	–	0.00
Γ_{124}	–	–	0.00	–	–	0.00	–	–	0.00	–	–	0.00
Γ_{134}	–	–	0.00	–	–	0.00	–	–	0.00	–	–	0.00
Γ_{14}	–	–	0.00	–	–	0.00	–	–	0.00	–	–	0.00
Γ_{115}	–	–	0.00	–	–	0.00	–	–	0.00	–	–	0.00
Γ_{125}	–	–	0.00	–	–	0.00	–	–	0.00	–	–	0.00
Γ_{135}	–	–	0.00	–	–	0.00	–	–	0.00	–	–	0.00
Γ_{145}	–	–	0.00	–	–	0.00	–	–	0.00	–	–	0.00
Γ_{15}	–	–	0.00	–	–	0.00	–	–	0.00	–	–	0.00

Table 4–4: Parameter settings for MPLE. Parameter values for both Poisson and binomial outcomes, and each of the six simulation settings, subpopulation 2.

Outcome	*Poisson*						*Binomial*					
Parameter	1	2	3	4	5	6	7	8	9	10	11	12
π_2	0.4	0.4	0.4	0.3	0.3	0.3	0.4	0.4	0.4	0.3	0.3	0.3
β_{20}	0.20	0.20	0.20	0.20	0.20	0.20	-0.85	-0.85	-0.85	-0.85	-0.85	-0.85
β_{21}	-0.45	-0.45	-0.45	-0.45	-0.45	-0.45	-0.15	-0.15	-0.15	-0.15	-0.15	-0.15
β_{22}	0.00	0.00	-0.10	0.00	0.00	-0.10	0.00	0.00	-0.75	0.00	0.00	-0.75
β_{23}	0.00	0.00	0.25	0.00	0.00	0.25	0.00	0.00	0.10	0.00	0.00	0.10
β_{24}	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
β_{25}	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
β_{26}	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
β_{27}	–	0.00	0.00	–	0.00	0.00	–	0.00	0.00	–	0.00	0.00
β_{28}	–	0.00	0.00	–	0.00	0.00	–	0.00	0.00	–	0.00	0.00
β_{29}	–	0.00	0.00	–	0.00	0.00	–	0.00	0.00	–	0.00	0.00
β_{210}	–	0.00	0.00	–	0.00	0.00	–	0.00	0.00	–	0.00	0.00
β_{211}	–	0.00	0.00	–	0.00	0.00	–	0.00	0.00	–	0.00	0.00
β_{212}	–	0.00	0.00	–	0.00	0.00	–	0.00	0.00	–	0.00	0.00
β_{213}	–	0.00	0.00	–	0.00	0.00	–	0.00	0.00	–	0.00	0.00
β_{214}	–	0.00	0.00	–	0.00	0.00	–	0.00	0.00	–	0.00	0.00
β_{215}	–	0.00	0.00	–	0.00	0.00	–	0.00	0.00	–	0.00	0.00
β_{216}	–	0.00	0.00	–	0.00	0.00	–	0.00	0.00	–	0.00	0.00
Γ_{21}	0.35	0.35	0.35	0.35	0.35	0.35	0.70	0.70	0.70	0.70	0.70	0.70
Γ_{212}	0.20	0.20	0.20	0.20	0.20	0.20	-0.70	-0.70	-0.70	-0.70	-0.70	-0.70
Γ_{22}	0.15	0.15	0.15	0.15	0.15	0.15	0.80	0.80	0.80	0.80	0.80	0.80
Γ_{213}	–	–	0.00	–	–	0.00	–	–	0.00	–	–	0.00
Γ_{223}	–	–	0.00	–	–	0.00	–	–	0.00	–	–	0.00
Γ_{23}	–	–	0.00	–	–	0.00	–	–	0.00	–	–	0.00
Γ_{214}	–	–	0.00	–	–	0.00	–	–	0.00	–	–	0.00
Γ_{224}	–	–	0.00	–	–	0.00	–	–	0.00	–	–	0.00
Γ_{234}	–	–	0.00	–	–	0.00	–	–	0.00	–	–	0.00
Γ_{24}	–	–	0.00	–	–	0.00	–	–	0.00	–	–	0.00
Γ_{215}	–	–	0.00	–	–	0.00	–	–	0.00	–	–	0.00
Γ_{225}	–	–	0.00	–	–	0.00	–	–	0.00	–	–	0.00
Γ_{235}	–	–	0.00	–	–	0.00	–	–	0.00	–	–	0.00
Γ_{245}	–	–	0.00	–	–	0.00	–	–	0.00	–	–	0.00
Γ_{25}	–	–	0.00	–	–	0.00	–	–	0.00	–	–	0.00

Table 4–5: Parameter settings for MPLE. Parameter values for both Poisson and binomial outcomes, and each of the six simulation settings, subpopulation 3.

Outcome	*Poisson*						*Binomial*					
Parameter	1	2	3	4	5	6	7	8	9	10	11	12
π_3	–	–	–	0.2	0.2	0.2	–	–	–	0.2	0.2	0.2
β_{30}	–	–	–	1.00	1.00	1.00	–	–	–	-0.30	-0.30	-0.30
β_{31}	–	–	–	0.15	0.15	0.15	–	–	–	-0.90	-0.90	-0.90
β_{32}	–	–	–	0.00	0.00	-0.65	–	–	–	0.00	0.00	0.80
β_{33}	–	–	–	0.00	0.00	-0.15	–	–	–	0.00	0.00	-0.25
β_{34}	–	–	–	0.00	0.00	0.00	–	–	–	0.00	0.00	0.00
β_{35}	–	–	–	0.00	0.00	0.00	–	–	–	0.00	0.00	0.00
β_{36}	–	–	–	0.00	0.00	0.00	–	–	–	0.00	0.00	0.00
β_{37}	–	–	–	–	0.00	0.00	–	–	–	–	0.00	0.00
β_{38}	–	–	–	–	0.00	0.00	–	–	–	–	0.00	0.00
β_{39}	–	–	–	–	0.00	0.00	–	–	–	–	0.00	0.00
β_{310}	–	–	–	–	0.00	0.00	–	–	–	–	0.00	0.00
β_{311}	–	–	–	–	0.00	0.00	–	–	–	–	0.00	0.00
β_{312}	–	–	–	–	0.00	0.00	–	–	–	–	0.00	0.00
β_{313}	–	–	–	–	0.00	0.00	–	–	–	–	0.00	0.00
β_{314}	–	–	–	–	0.00	0.00	–	–	–	–	0.00	0.00
β_{315}	–	–	–	–	0.00	0.00	–	–	–	–	0.00	0.00
β_{316}	–	–	–	–	0.00	0.00	–	–	–	–	0.00	0.00
Γ_{31}	–	–	–	0.25	0.25	0.25	–	–	–	1.75	1.75	1.75
Γ_{312}	–	–	–	0.00	0.00	0.00	–	–	–	0.00	0.00	0.00
Γ_{32}	–	–	–	0.20	0.20	0.20	–	–	–	0.85	0.85	0.85
Γ_{313}	–	–	–	–	–	0.00	–	–	–	–	–	0.00
Γ_{323}	–	–	–	–	–	0.00	–	–	–	–	–	0.00
Γ_{33}	–	–	–	–	–	0.00	–	–	–	–	–	0.00
Γ_{314}	–	–	–	–	–	0.00	–	–	–	–	–	0.00
Γ_{324}	–	–	–	–	–	0.00	–	–	–	–	–	0.00
Γ_{334}	–	–	–	–	–	0.00	–	–	–	–	–	0.00
Γ_{34}	–	–	–	–	–	0.00	–	–	–	–	–	0.00
Γ_{315}	–	–	–	–	–	0.00	–	–	–	–	–	0.00
Γ_{325}	–	–	–	–	–	0.00	–	–	–	–	–	0.00
Γ_{335}	–	–	–	–	–	0.00	–	–	–	–	–	0.00
Γ_{345}	–	–	–	–	–	0.00	–	–	–	–	–	0.00
Γ_{35}	–	–	–	–	–	0.00	–	–	–	–	–	0.00

a grid of possible values for the tuning parameter. Next, I estimated the model parameters for each value of the tuning parameter, and calculated the Bayesian Information Criterion (BIC) for that model. I then chose the model with the lowest BIC. More detail on this procedure is included in Section C.5.

The tables of these results can be found in Appendix G, and box plots showing an overview of the behaviour of the Mean Squared Error (MSE) are shown in Figures 4–1 to 4–2.

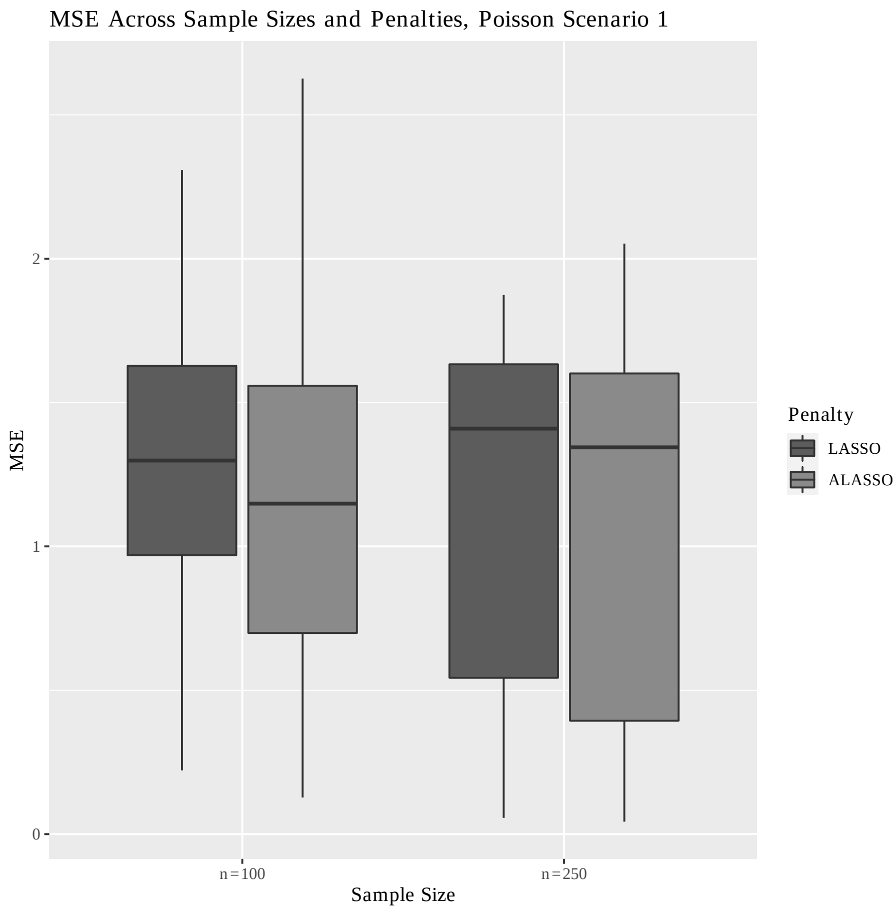

Figure 4–1: MSE across sample sizes, and penalties, Poisson outcome, with $K = 2$, $p = 2$, and $q = 1$.

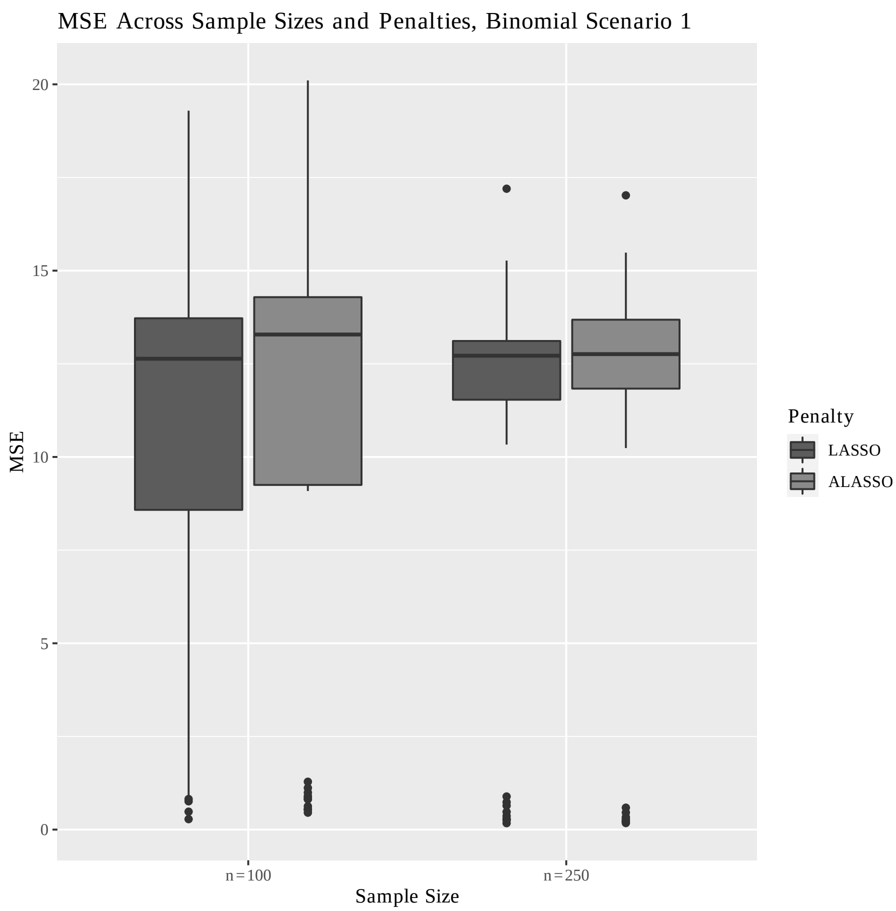

Figure 4–2: MSE across sample sizes, and penalties, binomial outcome, with $K = 2$, $p = 2$, and $q = 1$.

The Tables 4–6 and 4–7 show the proportion of correct selection for variables when I used two possible penalties, LASSO and ALASSO. I have also included in these tables the MLE when small values (those with a magnitude less than 0.01) were changed to 0, these have the row heading Small. I did not include the proportions of correct selection for the MLE, and oracle as they are 0%, and 100% respectively.

Table 4–6: Simulation 1 results proportion of parameters correctly classified, averaged over 50 runs. Outcome follows a Poisson distribution with $K = 2$, $p = 7$, $q = 2$, $\pi_1 = 0.6$, $\pi_2 = 0.4$, $\boldsymbol{\beta}_1^\top = (0.65, 0.30, 0.00, 0.00, 0.00, 0.00, 0.00)$, $\boldsymbol{\beta}_2^\top = (0.20, -0.45, 0.00, 0.00, 0.00, 0.00, 0.00)$, $\mathbb{\Gamma}_1^{*\top} = (0.30, -0.25, 0.10)$, and $\mathbb{\Gamma}_2^{*\top} = (0.35, 0.20, 0.15)$.

Penalty	Sample Size	Non-zero Fixed Effects	Zero Fixed Effects
LASSO	100	0.080	0.980
LASSO	250	0.270	0.956
ALASSO	100	0.150	0.980
ALASSO	250	0.380	0.964
Small	100	0.995	0.101
Small	250	1.000	0.166
Small	500	1.000	0.259
Small	1000	1.000	0.317

4.7 Conclusion

Using a penalised likelihood procedure for model selection offers many attractive statistical properties. In this chapter, I have extended the use of penalised likelihood to the FinMix GLMM setting. Because the penalisation can be applied to both the fixed and random effects, I changed the parameterisation of the variance components of the FinMix GLMM relative to the preceding chapter. I considered three penalty functions: LASSO, ALASSO, and SCAD. The penalty function does not affect the outer MCEM loop, and therefore, the only changes to the algorithm were in the inner MCEM loop. In order to use this approach with the MCEM algorithm as described in Chapter 3, I used a quadratic approximation to the penalty function. Given the equations for calculating the penalised likelihood, I considered the choice of the tuning parameter. In this case, I chose a grid of values for the tuning parameter, and chose the model with the best BIC. This model possesses a number of desirable asymptotic properties including consistency, sparsity, and the estimates follow a

Table 4–7: Simulation 7 results proportion of parameters correctly classified, averaged over 50 runs. Outcome follows a binomial distribution with $K = 2$, $p = 7$, $q = 2$, $\pi_1 = 0.6$, $\pi_2 = 0.4$, $\boldsymbol{\beta}_1^\top = (0.95, 0.60, 0.00, 0.00, 0.00, 0.00, 0.00)$, $\boldsymbol{\beta}_2^\top = (-0.85, -0.15, 0.00, 0.00, 0.00, 0.00, 0.00)$, $\mathbb{\Gamma}_1^{*\top} = (0.95, 0.85, 1.15)$, and $\mathbb{\Gamma}_2^{*\top} = (0.70, -0.70, 0.80)$.

Penalty	Sample Size	Non-zero Fixed Effects	Zero Fixed Effects
LASSO	100	0.250	0.918
LASSO	250	0.440	0.864
ALASSO	100	0.430	0.976
ALASSO	250	0.490	0.984
Small	100	0.985	0.101
Small	250	0.975	0.159
Small	500	0.985	0.217
Small	1000	0.990	0.339

Gaussian distribution asymptotically. Finally, the simulation study confirmed that this algorithm produces reasonable results in a variety of different settings. In addition to the mean squared error, I calculated the proportion of the simulations for which each parameter was correctly classified as being equal to zero or non-zero to verify the accuracy of the method.

CHAPTER 5
Real Data Analysis

Rheumatoid Arthritis (RA) is an auto-immune disease that affects many people, and dates back to the 1800s (Storey et al., 1994). This chronic disorder causes inflammation in the joints of patients, typically starting in the hands and feet. The lining of the joints is affected by RA, which leads to swelling and eventually, bone erosion and even joint deformity. As the disease progresses, other joints are affected, usually the elbows, ankles, knees, shoulders, and hips. More background information about RA can be found in Scott et al. (2010); Wasserman (2011); Meier et al. (2013); Smolen and Aletaha (2015); Smolen et al. (2016); Malmstrom et al. (2017). RA should not be confused with osteoarthritis, which is more common. The cause of osteoarthritis is the wearing away of cartilage, whereas the cause of RA is the inflammation of the synovial membrane.

Unfortunately, there is currently no cure for RA. Treatment aims to reduce inflammation, so that pain is reduced, and joint damage is slowed or prevented. The current treatment being favoured called Treat to Target (T2T). The goal of T2T is to aggressively treat the patient to either remission or a minimal level of disease activity. Medication is a favoured form of treatment, and there are many types of drugs used to treat RA. These medications include, but are not limited too, Nonsteroidal Anti-Inflammatory Drug (NSAID), steroids (often corticosteroids), Disease-Modifying Antirheumatic Drugs (DMARD), immunosuppressants, and Tumour Necrosis Factor

Inhibitors (TNFi). Because of the reliance on pharmacotherapy, drug toxicity has become an important adverse outcome of interest. In advanced stages, patients with RA may require surgery (such as joint replacement).

5.1 Scottish Early Rheumatoid Arthritis Inception Cohort and Biobank

The Scottish Early Rheumatoid Arthritis Inception Cohort and Biobank (SERA) contains patients from Scotland that have been diagnosed with RA and is administered by the Scottish Collaborative Arthritis Research network. It is rich in many variables, includes data both from questionnaires and blood samples, providing a large number of covariates to consider. The goal of the study is to be able to accurately predict patient outcomes so that the best course of treatment can be applied. A bank of tissue, and blood samples was also collected to allow for analysis of DNA or biomarkers in the future. Sixteen hospitals from around Scotland participated in this study.

During the first six months, the cohort enrolled 489 patients. The dataset I used for this analysis contained 1182 patients. While a number of controls were recruited for this study, but those subjects did not have a diagnosis of RA, therefore I did not consider them for this analysis. In order to be included in the cohort, all patients must have at least one swollen joint, as well as a new clinical diagnosis of RA or undifferentiated polyarthritis. Patients were excluded from the cohort if they had already been on DMARD therapy for a time period greater than six months, had another rheumatological diagnosis, had Hepatitis B, had Hepatitis C, or were Human Immunodeficiency Viruses (HIV) positive. Data was collected at baseline and every six months thereafter on demographic, employment, clinical measurements, laboratory measurements, and radiographic results. I considered several possible outcomes of interest, include clinical remission (defined as Disease Activity Score on 28 Joints (DAS28)< 2.6), swollen joint count, tender joint count, drug toxicity, and the number of steroids taken. Of particular interest as covariates were the presence of anti-Cyclic Citrullinated Peptide (CPP) antibodies, erosion at presentation, Body Mass Index (BMI), age, and alcohol intake. Further information about the SERA database can be found in Dale et al. (2016) and it has been used in published research, including Stalmach et al. (2014).

5.2 Outcome and Covariates

From the possible outcomes of interest, there were many options to consider. One outcome that is often of interest is DAS28, and clinical remission is often defined as DAS28< 2.6. However, DAS28 is not the only method used to define remission. Both the American College of Rheumatology, and the European League Against Rheumatism have suggested response criteria for RA (Ward et al., 2014), and the various composite measures do not necessarily agree (Smolen and Aletaha, 2015). I chose tender joint count out of 28 joints as the outcome for this analysis due to a variety of factors: It has been recommended in the literature (Felson et al., 1993), joint counts have been used for many years (Aletaha and Smolen, 2006), and tender joint count is an easy to understand outcome with potential values in $\{0, 1, 2, \ldots, 28\}$. The canonical link function (logit) was used to model this as a Finite Mixture of Generalised Linear Mixed-Effect Model (FinMix GLMM).

There are many factors that make a FinMix GLMM appropriate for this analysis. The outcome is a count with a fixed maximum, so modelling it as a binomial outcome is appropriate. The patients in the SERA cohort were followed longitudinally, so random effects are necessary as the outcomes from a single patient are likely correlated. In addition, RA is a heterogeneous disorder, especially with respect to treatment response, and drug-related toxicity (Dale et al., 2016), as well the presence or absence of antibodies to citrullinated protein antigens, rheumatoid factor, and other factors (Malmstrom et al., 2017). Genetic research supports the idea that RA is of a heterogeneous nature (Scott et al., 2010). Because of these factors, I used a FinMix GLMM.

There are a number of other variables in the dataset in addition to the outcome of interest. Every patient was assigned a unique patient number, and some information, namely marital status (Married, Single, Widowed, Living with partner, Divorced, Separated), race (White, Other, South East Asian, Indian Sub-Continent, Afro-Caribbean), weight, height, BMI, alcohol intake, gender (Female, Male), age, smoking status (Non-smoker, Ex-smoker, Current Smoker), and diagnosis (RA, Undifferentiated arthritis) was collected at baseline. Starting at baseline, and every visit thereafter, more information was collected, specifically, swollen joint count (out of 28), tender joint count (out of 28), DAS28 calculated using Erythrocyte Sedimentation Rate (ESR), DAS28 calculated using C-Reactive Protein (CRP), ESR, patient global health determined using a Visual Analogue Scale (VAS), CRP, assessor's global VAS, VAS pain score, hospital anxiety and depression scale, health assessment questionnaire score, EQ-5D score, employment status (Retired, Full-time employment, Part-time employment, Unemployed and not seeking work, Homemaker, Self-employed, Unemployed and seeking work, Student), total cholesterol, High-Density Lipoproteins (HDL) cholesterol, the ratio of total to HDL cholesterol, haemoglobin, total White Cell Count

(WCC), neutrophils, lymphocytes, monocytes, eosinophils, platelets, urea, creatinine, number of steroids taken, rheumatoid factor figure, and Cyclic Citrullinated Peptide (CPP). Medication information was also available. Patients in the dataset were on abatacept, azathioprine, gold, hydroxychloroquine, leflunomide, methotrexate, penicillamine, and sulfasalazine. For each patient, the following information was recorded for each medication they were on: the start date, stop date, dose, dose unit, frequency, route, reason stopped (if relevant), and the number of days in each interval that the patent was on that medication.

5.3 Other Statistical Considerations

Ideally, there would be no missing data, and thus no measures would need to be taken to account for the missing data. However, missing data is often encountered in applications. If all the data for a particular visit, and patient was missing (for example, if a patient missed an appointment, or was lost to follow up) then I removed that entire visit from the data set. I imputed missing variables using Multivariate Imputation by Chained Equations (MICE) with one imputation ($m = 1$), and five iterations. I did not perform a sensitivity analyses for missing data. In cases where the employment was missing, I used last observation carried forward rather than MICE. I made an effort in choosing methods that were statistically justified, and logical in data cleaning. Specifically, for factor variables, I chose the most common factor to be the baseline. I identified duplicates, and in cases where duplicates were identical, I removed one copy.

5.4 Introductory Tables

I included here two tables. The first table, Table 5–1, shows a summary of the patient information that does not change over time. The second table, Table 5–2, shows a summary of patient information separated by visit number.

Table 5–1: Summary information for all patients

	Overall
n	1168
Marital status (%)	
Married	680 (58.2)
Single	149 (12.8)
Widowed	115 (9.8)
Living with partner	99 (8.5)
Divorced	91 (7.8)
Separated	34 (2.9)
Race (%)	
White	1156 (99.0)
Other	6 (0.5)
South east asian	2 (0.2)
Indian sub-continent	4 (0.3)
Afro-caribbean	0 (0.0)
Weight In kg (mean (SD))	78.11 (17.43)
Height In m (mean (SD))	1.66 (0.10)
BMI (mean (SD))	28.21 (5.51)
Alcohol intake (mean (SD))	4.98 (8.93)
Gender = male (%)	412 (35.3)
Age (mean (SD))	57.90 (14.00)
Smoking status (%)	
Non-smoker	431 (36.9)
Ex-smoker	420 (36.0)
Current smoker	317 (27.1)
Diagnosis = undifferentiated arthritis (%)	168 (15.6)

Table 5–2: Information divided over six visits

	0	1	2	3	4	5	6
n	1114	1118	1118	1130	1125	1123	1120
Days in this interval (mean (SD))	0.00 (0.00)	152.12 (96.25)	138.98 (107.28)	110.43 (109.79)	85.70 (124.09)	43.26 (105.09)	24.68 (77.44)
Swollen joint count out of 28 (mean (SD))	7.04 (5.68)	2.49 (3.92)	2.04 (3.41)	2.06 (3.17)	1.87 (2.95)	1.72 (2.96)	1.34 (2.40)
Tender joint count out of 28 (mean (SD))	8.04 (7.13)	4.48 (6.07)	4.10 (6.04)	3.81 (5.18)	4.08 (5.66)	4.12 (5.56)	4.23 (5.60)
DAS28 using ESR (mean (SD))	4.91 (1.41)	3.51 (1.44)	3.34 (1.42)	3.34 (1.32)	3.32 (1.33)	3.29 (1.30)	3.25 (1.22)
DAS28 using CRP (mean (SD))	4.67 (1.34)	3.31 (1.34)	3.16 (1.32)	3.14 (1.21)	3.13 (1.23)	3.10 (1.21)	3.06 (1.17)
Erythrocyte sedimentation (mean (SD))	30.39 (25.56)	18.55 (18.71)	17.32 (17.69)	17.04 (18.30)	15.84 (16.64)	15.87 (17.64)	15.25 (15.70)
Patient global health VAS (mean (SD))	51.85 (27.67)	37.02 (27.96)	35.08 (27.58)	37.00 (27.94)	38.04 (28.75)	38.20 (27.97)	37.95 (28.00)
C reactive protein (mean (SD))	24.89 (34.43)	10.95 (18.28)	10.33 (17.90)	10.25 (17.72)	8.63 (12.30)	9.20 (16.59)	8.47 (16.38)
Assessor's global health VAS (mean (SD))	44.70 (24.30)	24.09 (21.19)	21.70 (20.13)	21.69 (20.33)	21.41 (19.86)	21.91 (19.62)	20.14 (18.83)
Pain VAS (mean (SD))	51.80 (28.05)	33.29 (26.60)	32.48 (26.99)	33.34 (26.99)	35.31 (27.55)	36.69 (27.74)	36.28 (27.06)
Hospital anxiety and depression scale (mean (SD))	5.52 (3.96)	4.63 (3.80)	4.48 (3.75)	4.77 (3.84)	4.89 (3.90)	5.29 (3.79)	5.19 (3.84)
Hospital anxiety and depression scale above 8 (%)	313 (28.1)	262 (23.4)	263 (23.5)	307 (27.2)	337 (30.0)	402 (35.8)	393 (35.1)
Hospital anxiety and depression scale above 11 (%)	140 (12.6)	95 (8.5)	94 (8.4)	114 (10.1)	122 (10.8)	144 (12.8)	155 (13.8)
Health assessment questionnaire score (mean (SD))	1.17 (0.79)	0.82 (0.76)	0.81 (0.75)	0.85 (0.76)	0.89 (0.75)	0.97 (0.75)	0.97 (0.74)
EQ 5D (mean (SD))	0.52 (0.32)	0.66 (0.28)	0.67 (0.28)	0.67 (0.28)	0.66 (0.27)	0.62 (0.29)	0.64 (0.27)
Employment status (%)							
Retired	417 (37.4)	433 (38.7)	447 (40.0)	455 (40.3)	468 (41.6)	469 (41.8)	467 (41.7)
Full time employment	349 (31.3)	338 (30.2)	323 (28.9)	322 (28.5)	305 (27.1)	303 (27.0)	305 (27.2)
Part time employment	168 (15.1)	175 (15.7)	168 (15.0)	170 (15.0)	167 (14.8)	169 (15.0)	170 (15.2)
Unemployed and not seeking work	42 (3.8)	55 (4.9)	58 (5.2)	55 (4.9)	58 (5.2)	61 (5.4)	55 (4.9)
Homemaker	48 (4.3)	43 (3.8)	54 (4.8)	57 (5.0)	52 (4.6)	50 (4.5)	52 (4.6)
Self-employed	51 (4.6)	36 (3.2)	33 (3.0)	36 (3.2)	36 (3.2)	31 (2.8)	34 (3.0)
Unemployed and seeking work	26 (2.3)	25 (2.2)	22 (2.0)	22 (1.9)	28 (2.5)	27 (2.4)	26 (2.3)
Student	13 (1.2)	13 (1.2)	13 (1.2)	13 (1.2)	11 (1.0)	13 (1.2)	11 (1.0)
Total cholesterol (mean (SD))	4.92 (1.12)	4.96 (1.06)	4.95 (1.03)	4.99 (1.05)	5.10 (1.12)	5.10 (1.13)	5.17 (1.10)
HDL cholesterol (mean (SD))	1.51 (3.09)	1.41 (0.57)	1.42 (0.56)	1.48 (0.81)	1.46 (0.77)	1.49 (0.75)	1.60 (3.07)
Ratio of total HDL cholesterol (mean (SD))	3.84 (1.28)	3.74 (1.24)	3.72 (1.24)	3.60 (1.14)	3.67 (1.30)	3.54 (1.14)	3.50 (1.19)
Haemaglobin (mean (SD))	13.21 (1.46)	13.26 (1.38)	13.34 (1.33)	13.26 (1.38)	13.41 (1.38)	13.35 (1.36)	13.32 (1.29)
Total white cell count (mean (SD))	8.36 (2.52)	7.42 (2.25)	7.07 (2.09)	7.23 (2.47)	7.30 (2.27)	7.20 (2.25)	7.20 (2.57)
Neutrophils (mean (SD))	5.62 (2.15)	4.82 (1.74)	4.49 (1.59)	4.52 (1.59)	4.47 (1.62)	4.24 (1.53)	4.15 (1.59)
Lymphocytes (mean (SD))	1.90 (0.72)	1.83 (0.72)	1.78 (0.70)	1.85 (0.73)	1.97 (0.86)	2.04 (0.82)	2.05 (0.85)
Monocytes (mean (SD))	0.59 (0.23)	0.55 (0.23)	0.53 (0.21)	0.55 (0.22)	0.55 (0.22)	0.57 (0.23)	0.56 (0.24)
Eosinophils (mean (SD))	0.20 (0.18)	0.19 (0.18)	0.19 (0.18)	0.19 (0.19)	0.19 (0.18)	0.19 (0.16)	0.18 (0.15)
Platelets (mean (SD))	325.94 (103.35)	289.52 (84.72)	279.56 (76.47)	277.19 (83.99)	267.42 (74.81)	260.93 (75.85)	259.99 (74.05)
Urea (mean (SD))	6.30 (6.35)	6.35 (7.23)	5.93 (5.64)	6.26 (7.54)	5.63 (5.36)	5.71 (6.43)	6.05 (7.99)
Creatinine (mean (SD))	68.51 (16.98)	69.11 (17.96)	68.32 (16.16)	68.17 (15.76)	68.78 (16.92)	67.87 (14.79)	67.09 (14.19)
Count steroids taken (mean (SD))	1.04 (1.40)	0.82 (1.06)	0.31 (0.69)	0.17 (0.53)	0.08 (0.35)	0.04 (0.28)	0.02 (0.26)
Rheumatoid factor figure (mean (SD))	169.65 (316.40)	757.05 (1193.06)	710.89 (1145.58)	632.64 (1076.43)	519.03 (993.05)	409.60 (955.04)	324.66 (731.39)
Cyclic citrullinated peptide (mean (SD))	171.04 (242.69)	232.07 (497.46)	230.27 (522.94)	210.64 (534.95)	192.69 (429.95)	183.53 (442.22)	152.48 (341.73)

I also included two figures to describe the dataset. The first figure, Figure 5–1, shows the proportion of patients with a given Swollen Joint Count for each visit number in the SERA dataset. The second figure, Figure 5–2, is analogous to the first figure but for Tender Joint Count. Note that some patients missed appointments or were lost to follow up, which explains why there were not the same number of patients for each visit number.

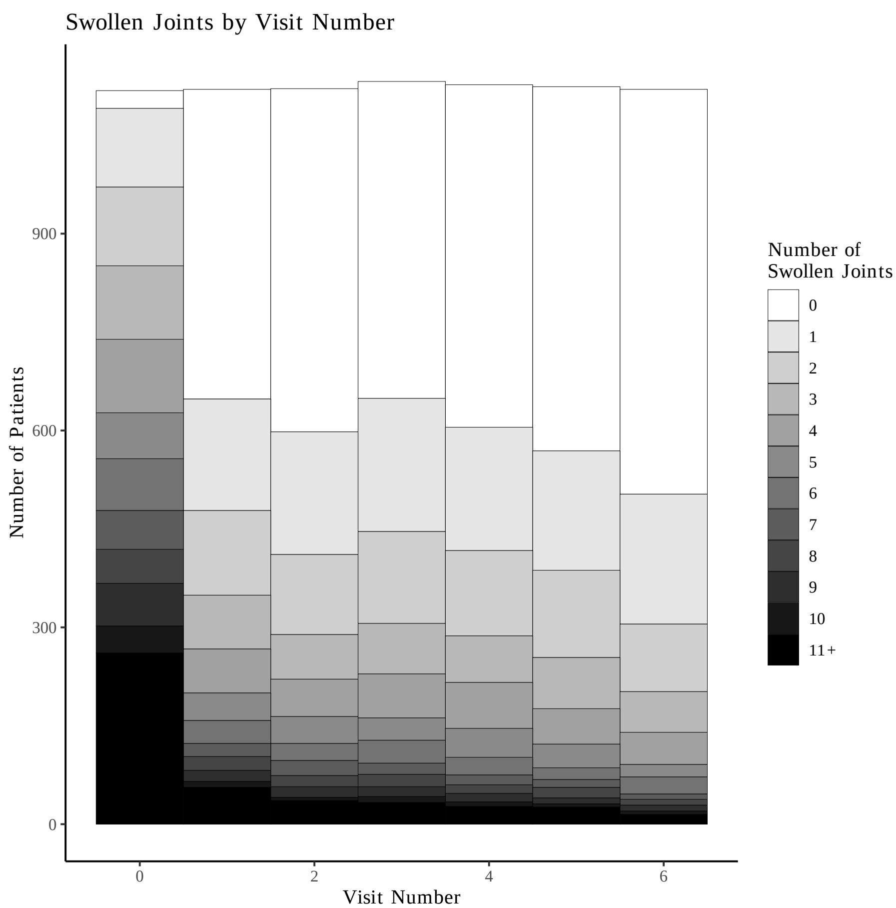

Figure 5–1: Number of swollen joints

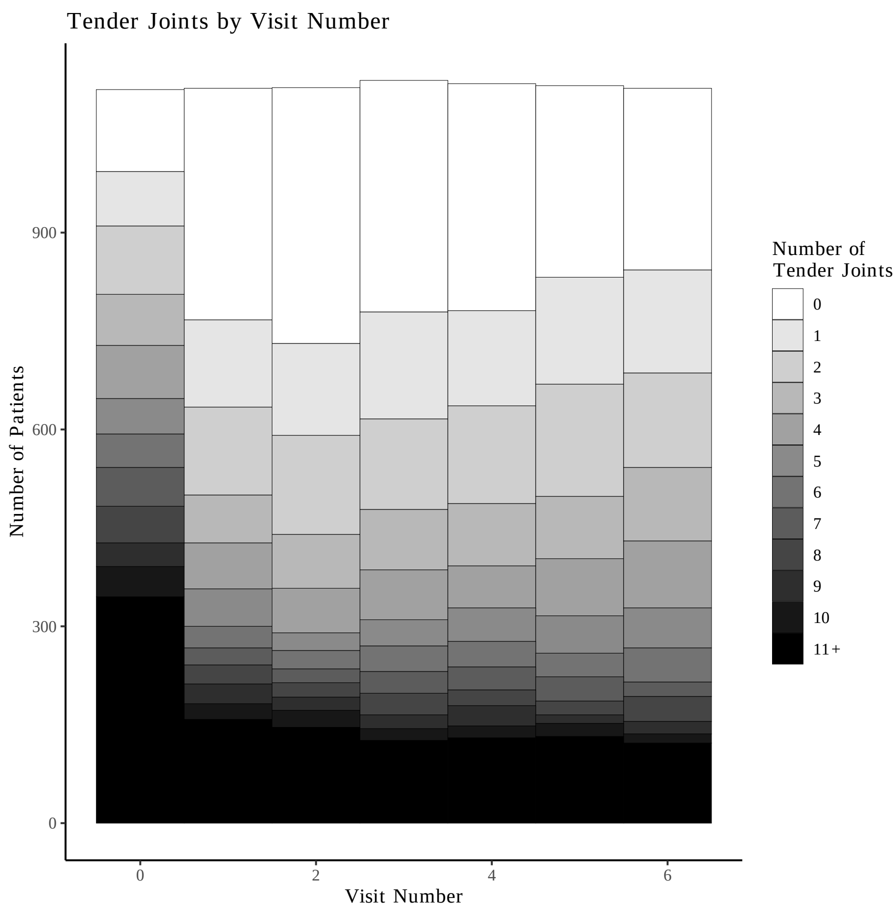

Figure 5–2: Number of tender joints

5.5 Maximum Likelihood Estimation

Using variables identified in the literature, a Maximum Likelihood Estimation (MLE) model was fit. I included the following variables (with relevant citations): (i) Visit Number, used as a proxy for disease duration, and time in treatment (Smolen and Aletaha, 2015), (ii) Smoking status, current smoker or not a current smoker (Smolen and Aletaha, 2015; Malmstrom et al., 2017), (iii) rheumatoid factor (Smolen and Aletaha, 2015). I included a fixed effect for each of the covariates, and a random effect on both the intercept and visit number. More specifically, π_k is the mixing proportion, β_{k0} is the fixed intercept, β_{k1} is the fixed effect for visit number, β_{k2} is the fixed effect for smoking status, and β_{k3} is the fixed effect for rheumatoid factor. To determine the value for K, a series of models were fit. First, I used the R Generalised Linear Mixed-Effect Model (GLMM) algorithm (glmer from the package lme4), and then took the result as the starting value for the inner Expectation-Maximisation (EM) of a FinMix GLMM. This corresponds to a FinMix GLMM with $K = 1$. Then, I took the results of that computation as the starting values for $\boldsymbol{\beta}_1$, and Γ_1, and the remaining starting values set to $\pi_1 = \frac{1}{2}$, $\boldsymbol{\beta}_2 = 0$, and $\Gamma_2 = I_q$. I then fit a FinMix GLMM, this time with $K = 2$, and both the inner, and outer EM loops being used. Following on in this way, I estimated parameters for a series of models, for each new K, I used the parameter estimates from the previously fit model starting values for $\boldsymbol{\beta}_k, k \in \{1, 2, \ldots, K-1\}$, and $\Gamma_k, k \in \{1, 2, \ldots, K-1\}$ but set $\pi_k = \frac{1}{K} \forall k$, $\boldsymbol{\beta}_K = 0$, and $\Gamma_K = I_q$. I continued to estimate the parameters until I found a minimum value for the Bayesian Information Criterion (BIC).

Once I had found the minimum value for the BIC, in this case, $K = 10$ this serves as the upper bound on K, and I chose the best model from $K \in \{1, 2, \ldots, 10\}$. I have included the values for BIC in Table 5–3. From these 10 models, I then calculated the standard errors for the parameters, as well as the corresponding p-value for inclusion in the model. I have included details of the calculations of standard errors can in Appendix J. Using these p-values, I chose the largest model for which all of the values of π_k were significant, in this case, $K = 5$. I have included the parameter values of this model, including their standard errors (where appropriate) in Table 5–4. In Table 5–5 I show the t-statistics and corresponding p-values for the parameter values of the MLE model.

Table 5–3: BIC over different values of K

K	BIC
1	46862.15
2	46097.65
3	45773.19
4	45461.07
5	45075.56
6	44717.19
7	44321.42
8	43945.43
9	43681.21
10	43452.87
11	43471.01

This analysis shows that time on treatment, smoking status, and rheumatoid factor are all important covariates that correlate to tender joint count. However, the relationship between these covariates, is not the same for all patients in the SERA cohort. This heterogeneity requires both random effects and a finite mixture of models

Table 5–4: Parameter values for MLE model with 95% confidence intervals

Variable	Subpopulation 1	Subpopulation 2	Subpopulation 3	Subpopulation 4	Subpopulation 5
Mixing Proportion (π_k)	0.60 (0.53, 0.65)	0.17 (0.12, 0.21)	0.10 (0.04, 0.15)	0.07 (0.01, 0.12)	0.05
Intercept (β_{k0})	-1.42 (-1.42, -1.41)	-0.42 (-0.43, -0.41)	-0.34 (-0.37, -0.32)	-0.17 (-0.22, -0.14)	-0.09 (-0.19, -0.02)
Visit Number (β_{k1})	-0.09 (-0.09, -0.09)	0.01 (0.01, 0.01)	-0.02 (-0.03, -0.02)	-0.01 (-0.02, 0.00)	-0.01 (-0.03, 0.01)
Smoking Status (β_{k2})	0.36 (0.36, 0.37)	-0.69 (-0.70, -0.69)	-0.06 (-0.08, -0.05)	0.04 (-0.02, 0.08)	0.04 (-0.08, 0.12)
Rheumatoid Factor (β_{k3})	-0.00 (0.00, 0.00)	0.00 (0.00, 0.00)	0.00 (0.00, 0.00)	0.00 (0.00, 0.00)	0.00 (0.00, 0.00)
Intercept (Γ_{k1})	1.47	1.36	1.16	1.04	1.02
Intercept/Visit Number (Γ_{k12})	-0.29	-0.07	-0.09	-0.05	-0.03
Visit Number (Γ_{k2})	0.19	0.86	0.71	0.79	0.90

Table 5–5: P-values for MLE model with t-statistics

Variable	Subpopulation 1	Subpopulation 2	Subpopulation 3	Subpopulation 4	Subpopulation 5
Mixing Proportion (π_k)	0.00 (16.90)	0.00 (6.93)	0.00 (3.15)	0.03 (2.19)	
Intercept (β_{k0})	0.00 (-423.49)	0.00 (-76.25)	0.00 (-23.35)	0.00 (-7.37)	0.07 (-1.79)
Visit Number (β_{k1})	0.00 (-117.47)	0.00 (7.31)	0.00 (-9.07)	0.03 (-2.16)	0.58 (-0.56)
Smoking Status (β_{k2})	0.00 (115.27)	0.00 (-130.73)	0.00 (-5.45)	0.17 (1.37)	0.55 (0.59)
Rheumatoid Factor (β_{k3})	0.00 (-887.68)	0.00 (359.99)	0.00 (17.13)	0.00 (-5.03)	0.25 (-1.15)

to represent. Across all subpopulations, the visit number was negatively associated with tender joint count in all but one subpopulation. This makes sense because the goal of treatment is to reduce the symptoms of RA including tender joint count. Smoking status was also associated with tender joint count, but this association was not the same across all subpopulations. In addition, the rheumatoid factor has a relatively small impact on tender joint count. From the random effects, it is clear that there is a significant heterogeneity between patients, even those within the same subpopulation. The mixing proportion suggests that over half of the population is in Subpopulation 1, and that all of the other subpopulations are significantly smaller.

In addition to this analysis, I performed another analysis in the same way with $p = 3$, excluding rheumatoid factor as a covariate. The fit of the $p = 3$ model was not as good as the model previously presented, so I have not included those results here. I also performed an analogous analysis with $q = 1$ and the only random effect corresponding to the intercept in the model. Again, this did not fit the data as well and these results have been omitted.

5.6 Maximum Penalised Likelihood Estimation

I confined this analysis to the $K = 2$ case. The previous analysis suggested that over half of the total population was in Subpopulation 1, and the other four subpopulations had estimated parameters that were more similar to each other than to the parameters for Subpopulation 1. I performed three analyses, one for each of the penalties. As in Chapter 4, the three penalties used were Least Absolute Shrinkage and Selection Operator (LASSO), Adaptive Least Absolute Shrinkage and Selection Operator (ALASSO), and Smoothly Clipped Absolute Deviation (SCAD). The following variables were included for consideration in the model: Visit number (β_{k1}), Ex-smoker (β_{k2}), Current smoker (β_{k3}), Rheumatoid factor figure (β_{k4}), Lymphocytes (β_{k5}), Erythrocyte sedimentation (β_{k6}), C reactive protein (β_{k7}), Ratio of total to HDL Cholesterol (β_{k8}), Haemaglobin (β_{k9}), Total WCC (β_{k10}), Neutrophils (β_{k11}), Monocytes (β_{k12}), Eosinophils (β_{k13}), Platelets (β_{k14}), Urea (β_{k15}), Creatinine (β_{k16}), CPP (β_{k17}). As in the MLE case, I included a random effect on both the intercept and visit number. As is common practice when a penalty is added to the likelihood equation, I standardised all of the variables to have mean 0 and standard deviation 1 before estimating the parameters. I used the algorithm in Section C.2 to find the lowest value of BIC for each penalty, then narrowed the range of potential values of $\boldsymbol{\lambda}$ around that lowest value. Using this smaller range and a finer grid, I again calculated the parameter values and BIC for each of the proposed values of $\boldsymbol{\lambda}$ and chose the setting with the lowest value of BIC. Since the original range for values of $\boldsymbol{\lambda}$ in the SCAD case was quite large, a much courser grid was used to cover it than in the LASSO or ALASSO case, and so I used a series of smaller and finer grids (three additional grids) rather than two as I did in the LASSO case. In this analysis, the lowest value for BIC in the LASSO case was 42803.23, in the ALASSO case was 43390.26, and in the SCAD case was 43233.47. Results of these three analysis are in Tables 5–6 to 5–11. I also considered the case where $K = 1$, and found that these models did not fit the data as well as the $K = 2$ models. The values for BIC in these cases were 45653.07 when I used the LASSO penalty, 45832.51 when I used the ALASSO penalty, and 45720.09 when I used the SCAD penalty.

In an ideal case, I would expect that the variables that are selected and the parameter estimates would be similar regardless of the penalty that was used. However, that was not the result of this analysis. There were some differences in the variables that were selected depending on the penalty that was used. Overall, the parameters were similar for all three cases. The results show that one subpopulation contains a majority of patients, but that there is another distinct subpopulation that differs significantly from the first. The values for the estimated fixed intercepts were similar for all three penalties. The following variables were selected in both subpopulations when I used both LASSO and SCAD: Erythrocyte sedimentation (β_{k6}), C reactive protein (β_{k7}), Haemaglobin (β_{k9}), Neutrophils (β_{k11}), Platelets (β_{k14}). The ALASSO penalty selected fewer covariates.

5.7 Conclusion

This analysis shows that a FinMix GLMM is a valuable regression tool and is useful for analysis of health related data. The MLE analysis suggested that there

Table 5–6: Parameter values for Maximum penalised Likelihood Estimation (MPLE) model with standard errors, LASSO penalty

Variable	Subpopulation 1	Subpopulation 2
Mixing Proportion (π_k)	0.70 (0.62, 0.75)	0.30
Intercept (β_{k0})	-2.06 (-2.06, -2.05)	-0.54 (-0.54, -0.54)
Visit Number (β_{k1})	-0.02 (-0.03, -0.02)	–
Ex-smoker (β_{k2})	–	–
Current Smoker (β_{k3})	–	–
Rheumatoid Factor (β_{k4})	-0.27 (-0.28, -0.27)	–
Lymphocytes (β_{k5})	0.07 (0.06, 0.07)	-0.11 (-0.12, -0.11)
Erythrocyte Sedimentation (β_{k6})	-0.09 (-0.10, -0.09)	0.04 (0.03, 0.04)
C Reactive Protein (β_{k7})	0.17 (0.17, 0.17)	-0.07 (-0.08, -0.07)
Ratio of Total to HDL Cholesterol (β_{k8})	0.15 (0.15, 0.15)	–
Haemaglobin (β_{k9})	-0.19 (-0.19, -0.19)	0.33 (0.33, 0.34)
Total WCC (β_{k10})	0.01 (0.01, 0.02)	–
Neutrophils (β_{k11})	0.51 (0.51, 0.52)	-0.31 (-0.32, -0.30)
Monocytes (β_{k12})	-0.28 (-0.28, -0.28)	0.03 (0.02, 0.03)
Eosinophils (β_{k13})	-0.04 (-0.04, -0.04)	-0.09 (-0.10, -0.09)
Platelets (β_{k14})	0.13 (0.12, 0.13)	0.01 (0.01, 0.02)
Urea (β_{k15})	0.02 (0.02, 0.03)	–
Creatinine (β_{k16})	-0.16 (-0.17, -0.16)	–
CPP (β_{k17})	0.07 (0.06, 0.07)	0.05 (0.04, 0.05)
Intercept (Γ_{k1})	0.96	1.05
Intercept/Visit Number (Γ_{k12})	0.17	0.17
Visit Number (Γ_{k2})	0.65	1.05

is significant heterogeneity in the population of patients with RA, not just between patients, but also multiple distinct subpopulations within the overall population of patients with RA.

Table 5–7: P-values for MPLE model with t-statistics, LASSO penalty

Variable	Subpopulation 1	Subpopulation 2
Mixing Proportion (π_k)	0.00 (18.73)	
Intercept (β_{k0})	0.00 (-988.48)	0.00 (-335.58)
Visit Number (β_{k1})	0.00 (-15.10)	–
Ex-smoker (β_{k2})	–	–
Current Smoker (β_{k3})	–	–
Rheumatoid Factor (β_{k4})	0.00 (-162.38)	–
Lymphocytes (β_{k5})	0.00 (33.26)	0.00 (-37.11)
Erythrocyte Sedimentation (β_{k6})	0.00 (-48.91)	0.00 (17.14)
C Reactive Protein (β_{k7})	0.00 (119.37)	0.00 (-26.60)
Ratio of Total to HDL Cholesterol (β_{k8})	0.00 (117.89)	–
Haemaglobin (β_{k9})	0.00 (-120.13)	0.00 (176.62)
Total WCC (β_{k10})	0.00 (3.58)	–
Neutrophils (β_{k11})	0.00 (156.18)	0.00 (-46.21)
Monocytes (β_{k12})	0.00 (-163.02)	0.00 (15.09)
Eosinophils (β_{k13})	0.00 (-23.37)	0.00 (-49.44)
Platelets (β_{k14})	0.00 (73.32)	0.00 (8.44)
Urea (β_{k15})	0.00 (11.28)	–
Creatinine (β_{k16})	0.00 (-112.26)	–
CPP (β_{k17})	0.00 (58.39)	0.00 (38.39)

Table 5–8: Parameter values for MPLE model with standard errors, ALASSO penalty

Variable	Subpopulation 1	Subpopulation 2
Mixing Proportion (π_k)	0.70 (0.63, 0.75)	0.30
Intercept (β_{k0})	-2.03 (-2.03, -2.02)	-0.69 (-0.70, -0.69)
Visit Number (β_{k1})	–	–
Ex-smoker (β_{k2})	–	–
Current Smoker (β_{k3})	–	–
Rheumatoid Factor (β_{k4})	-0.01 (-0.02, -0.01)	–
Lymphocytes (β_{k5})	–	–
Erythrocyte Sedimentation (β_{k6})	–	–
C Reactive Protein (β_{k7})	–	–
Ratio of Total to HDL Cholesterol (β_{k8})	–	–
Haemaglobin (β_{k9})	–	–
Total WCC (β_{k10})	–	–
Neutrophils (β_{k11})	0.48 (0.47, 0.48)	-0.18 (-0.19, -0.17)
Monocytes (β_{k12})	–	–
Eosinophils (β_{k13})	–	–
Platelets (β_{k14})	–	–
Urea (β_{k15})	–	–
Creatinine (β_{k16})	–	–
CPP (β_{k17})	–	–
Intercept (Γ_{k1})	0.97	0.92
Intercept/Visit Number (Γ_{k12})	-0.16	0.16
Visit Number (Γ_{k2})	0.67	0.96

Table 5–9: P-values for MPLE model with t-statistics, ALASSO penalty

Variable	Subpopulation 1	Subpopulation 2
Mixing Proportion (π_k)	0.00 (18.91)	
Intercept (β_{k0})	0.00 (-952.11)	0.00 (-416.86)
Visit Number (β_{k1})	–	–
Ex-smoker (β_{k2})	–	–
Current Smoker (β_{k3})	–	–
Rheumatoid Factor (β_{k4})	0.00 (-8.73)	–
Lymphocytes (β_{k5})	–	–
Erythrocyte Sedimentation (β_{k6})	–	–
C Reactive Protein (β_{k7})	–	–
Ratio of Total to HDL Cholesterol (β_{k8})	–	–
Haemaglobin (β_{k9})	–	–
Total WCC (β_{k10})	–	–
Neutrophils (β_{k11})	0.00 (134.89)	0.00 (-28.49)
Monocytes (β_{k12})	–	–
Eosinophils (β_{k13})	–	–
Platelets (β_{k14})	–	–
Urea (β_{k15})	–	–
Creatinine (β_{k16})	–	–
CPP (β_{k17})	–	–

Table 5–10: Parameter values for MPLE model with standard errors, SCAD penalty

Variable	Subpopulation 1	Subpopulation 2
Mixing Proportion (π_k)	0.71 (0.64, 0.76)	0.29
Intercept (β_{k0})	-2.05 (-2.05, -2.05)	-1.17 (-1.17, -1.16)
Visit Number (β_{k1})	–	0.02 (0.02, 0.02)
Ex-smoker (β_{k2})	–	-0.14 (-0.14, -0.14)
Current Smoker (β_{k3})	–	-0.13 (-0.14, -0.13)
Rheumatoid Factor (β_{k4})	–	-0.23 (-0.24, -0.23)
Lymphocytes (β_{k5})	–	-0.16 (-0.16, -0.16)
Erythrocyte Sedimentation (β_{k6})	0.01 (0.01, 0.01)	-0.18 (-0.19, -0.18)
C Reactive Protein (β_{k7})	0.01 (0.01, 0.01)	0.02 (0.02, 0.03)
Ratio of Total to HDL Cholesterol (β_{k8})	–	0.49 (0.48, 0.49)
Haemaglobin (β_{k9})	–	0.15 (0.15, 0.15)
Total WCC (β_{k10})	0.01 (0.00, 0.02)	0.11 (0.10, 0.12)
Neutrophils (β_{k11})	0.01 (0.00, 0.02)	0.18 (0.17, 0.18)
Monocytes (β_{k12})	0.01 (0.01, 0.01)	-0.60 (-0.60, -0.60)
Eosinophils (β_{k13})	–	-0.34 (-0.35, -0.34)
Platelets (β_{k14})	–	0.45 (0.45, 0.45)
Urea (β_{k15})	–	-0.03 (-0.03, -0.02)
Creatinine (β_{k16})	–	-0.16 (-0.17, -0.16)
CPP (β_{k17})	–	-0.02 (-0.02, -0.02)
Intercept (Γ_{k1})	0.89	0.89
Intercept/Visit Number (Γ_{k12})	-0.09	0.09
Visit Number (Γ_{k2})	0.72	0.92

Table 5–11: P-values for MPLE model with t-statistics, SCAD penalty

Variable	Subpopulation 1	Subpopulation 2
Mixing Proportion (π_k)	0.00 (19.31)	
Intercept (β_{k0})	0.00 (-1087.73)	0.00 (-653.49)
Visit Number (β_{k1})	–	0.00 (9.77)
Ex-smoker (β_{k2})	–	0.00 (-70.44)
Current Smoker (β_{k3})	–	0.00 (-62.71)
Rheumatoid Factor (β_{k4})	–	0.00 (-102.16)
Lymphocytes (β_{k5})	–	0.00 (-76.14)
Erythrocyte Sedimentation (β_{k6})	0.00 (6.20)	0.00 (-78.31)
C Reactive Protein (β_{k7})	0.00 (6.59)	0.00 (12.94)
Ratio of Total to HDL Cholesterol (β_{k8})	–	0.00 (300.32)
Haemaglobin (β_{k9})	–	0.00 (79.00)
Total WCC (β_{k10})	0.07 (1.82)	0.00 (24.32)
Neutrophils (β_{k11})	0.01 (2.48)	0.00 (42.93)
Monocytes (β_{k12})	0.00 (6.57)	0.00 (-322.60)
Eosinophils (β_{k13})	–	0.00 (-128.98)
Platelets (β_{k14})	–	0.00 (229.82)
Urea (β_{k15})	–	0.00 (-11.75)
Creatinine (β_{k16})	–	0.00 (-89.39)
CPP (β_{k17})	–	0.00 (-14.88)

CHAPTER 6
Conclusion

The world is full of complex questions, and problems that can only be solved by appropriate statistical analysis. However, one must use the correct statistical techniques in order to answer these questions suitably. Therefore, the assumptions associated with a particular statistical model should be considered carefully when performing statistical analysis. As such, it is important to develop models that take into account the underlying properties of complex datasets.

While many regression models have been proposed, and studied in the past, a Finite Mixture of Generalised Linear Mixed-Effect Model (FinMix GLMM) is a novel addition. Previous literature has focused on linear regression, Generalised Linear Model (GLM), Generalised Linear Mixed-Effect Model (GLMM), and certain finite mixtures or regression models. While a few similar models have been considered, notable a Finite Mixture of Linear Mixed-Effect (FMLME), a FinMix GLMM has not been previously studied. As such, the FinMix GLMM is a useful extension to the current literature.

In Chapter 3 I showed the theory, and an algorithm for calculating the Maximum Likelihood Estimation (MLE) of a FinMix GLMM. I carefully defined, and described the model, including the likelihood equation that was then used to facilitate calculation of the MLE. I described in detail the numerical computation of the MLE which implemented two nested Monte Carlo Expectation-Maximisation (MCEM) loops. In order to verify the properties of the MLE, I undertook a simulation study. In the cases considered, the estimates calculated were well behaved with small variances, and Mean Squared Error (MSE).

Once the form of a model has been chosen, the choice of variables to include in that model is another complex issue. I explored the question of model selection in Chapter 4, and performed the variable selection through penalisation of the likelihood with Least Absolute Shrinkage and Selection Operator (LASSO), Adaptive Least Absolute Shrinkage and Selection Operator (ALASSO), and Smoothly Clipped Absolute Deviation (SCAD) penalties. The motivation for choosing these penalties was their asymptotic properties. The addition of a penalty to the likelihood equation of a FinMix GLMM required additional considerations, including reparameterisation, and approximation, I incorporated these into the algorithm for estimating the Maximum penalised Likelihood Estimation (MPLE). Again, I performed a simulation study to verify, and illustrate the performance of the algorithm. Due to time constraints, some of the planned simulations were not completed.

Finally, I conducted real data analysis. This shows the usefulness of a FinMix GLMM for analysis of data from medical settings, and that the work of this research is not purely theoretical.

6.1 Further Work

There are many possible extensions or different cases to consider for this model, and as such, there are a number of possibilities for future work.

One interesting possibility that is beyond the scope of this thesis is time-to-event outcomes. Survival analysis for a finite mixture of models is a complex topic, so this possibility would require an extensive amount of work. Time-to-event outcomes are often of interest, and have been explored in the context of GLMMs by Yau and McGilchrist (1996), and Yau and McGilchrist (1997), and multivariate mixed-effect models can be used as described in Fieuws et al. (2007), Gueorguieva (2001), and Sammel et al. (1999). A finite mixture of these types of models would be an interesting extension. Hunsberger et al. (2009) considered finite mixtures of survival models. An additional extension is to allow coefficients to change over time as shown in Tutz and Kauermann (2003), and Zhang (2004).

Zhu and Lee (2003) discussed influential observations for GLMM, and deletion diagnostics in Ganguli et al. (2016), further work is possible in extending these ideas to FinMix GLMMs. Jiang and Zhang (2001), and Sinha (2004) considered robust maximum likelihood which could be investigated for FinMix GLMMs.

In this thesis, I assumed that the random effects follow a Gaussian distribution, and while this is a popular choice, it is not the only option that has been explored in the literature. A multivariate Student's t-distribution is a natural extension as the Student's t-distribution has heavier tails than a Gaussian distribution and was used in Bai et al. (2016). A mixture of Gaussian distributions was used in Verbeke and Lesaffre (1996) and motivated by the problem of model misspecification in this context. A much more complex approach using first-order Markov chains can be found in Farcomeni (2015), and a Dirchlet process was used in Guha (2008). I leave these possibilities to future work.

This thesis focused on finite mixtures of distributions from an exponential family, specifically Poisson, and binomial GLMMs, but there are many possibilities for further exploration using different distributions. One popular use of finite mixtures of models is in cases of overdispersion or zero-inflated regression. Cao and Yao (2012) looked at a mixture of binomial outcomes, and a degenerate random variable. Similarly, Lim et al. (2014), and Morgan et al. (2014) explored a zero-inflated Poisson regression. Young (2014) considered finite mixtures of regressions that include change points, and Bao and Hanson (2016) showed a mean-constrained finite mixture. I have left these possibilities for further work.

If the values of the covariates are informative to the mixing proportions, a finite mixture of experts is applicable, as used in Huang and Yao (2012), Khalili (2010), Wu and Yu (2016), Jacobs et al. (1991). This can also be described as modelling predictors of latent classes (Kim et al., 2016). The number of subpopulations, K, is an important consideration in finite mixtures. Kasahara and Shimotsu (2015) tested the number of subpopulations when using likelihood-based tests when the underlying distributions each follow a Gaussian distribution. In comparison, Li et al. (2016) also focuses on the Gaussian case but uses trimmed information criteria for more robust estimation. These possibilities could inspire further work on FinMix GLMMs.

In addition, the aforementioned theoretical considerations, many computational ideas may also be relevant. Specifically, coordinate descent is often used when optimising penalised likelihoods (Wei and Zhu, 2012), the use of antithetic variables as shown in Rubinstein and Samorodnitsky (1985) may provide improved computational efficiency, and importance sampling (Kuk, 1999) or sampling from a Student's t-distribution (Booth and Hobert, 1999) rather than a standard Gaussian in the calculation of τ_{ik} could improve the approximation of the integral in Equation (3.1).

Ergodic averaging, as explored in Fort et al. (2003), could provide less variable parameter estimates, and is an interesting possibility.

The focus of the thesis was not on the computational aspects of the algorithm, and as such, there are many improvements, and possibilities for further work to compare different approaches. A Monte Carlo approach using a Gibbs sampler for GLMMs is possible (Zeger and Karim, 1991; Gamerman, 1997; Burton, 2003; Chan et al., 2005; Christensen et al., 2006; Fan et al., 2008) as well as a similar Gibbs sampling approach (Leung and Elashoff, 1996a). These extensions could be tried with a FinMix GLMM. Sequential reduction (Ogden, 2015), matching (Benedetti et al., 2014), and profile likelihood (Jeon and Rabe-Hesketh, 2012) are additional possibilities. Kuk (1999) showed that importance sampling could improve the approximation of an integrated marginal likelihood function. A simulation-based estimator for GLMMs was shown in Li and Wang (2012) where both consistency, and the asymptotic distribution of the estimator was included. penalised quasi-likelihood, and simulated maximum likelihood were compared, and contrasted in Ng et al. (2006). I have not applied these approaches to a FinMix GLMM. In addition, I used the same value for L in all simulations, and the same tolerance for assessing convergence. As such, possibilities of increasing L as the iterations increase or changing the assessment of convergence based on K or the number of parameters being estimated are possibilities for further work. In addition, an Markov Chain Monte Carlo (MCMC) approach could be applicable to problems where I used Expectation-Maximisation (EM), Ryden (2008) provided a comparison between these methods. There are also many different maximisation algorithms that could be explored. Coordinate descent is especially popular in the variable selection literature (Wu and Lange, 2008), and could be applied to FinMix GLMMs.

Group variable selection is an interesting problem, and has been explored in the literature with group LASSO in Yuan and Lin (2006), and Meier et al. (2008), Adaptive group LASSO in Wang and Leng (2008), and group SCAD in Wang et al. (2007). Extensions to group LASSO include weighted group LASSO (Hirose and Konishi, 2012), standardized group LASSO (Simon and Tibshirani, 2012), and sparse group LASSO (Xie and Xu, 2014). I have left expanding these options to FinMix GLMMs for future work.

Several computational adjustments have been proposed for LASSO in Foster et al. (2008), Wu and Lange (2008), Guo et al. (2015), Lee et al. (2015), Laurin et al. (2016), and Rajaratnam et al. (2016), and I could consider these in the FinMix GLMM case. The LASSO penalty function has also been extended in various ways including fused LASSO (Tibshirani et al., 2005), relaxed LASSO (Meinshausen, 2007), Bayesian LASSO (Hans, 2009), random LASSO (Wang et al., 2011), forward-LASSO adaptive shrinkage (Radchenko and James, 2011), smooth-lasso (Hebiri and van de Geer, 2011), iteratively reweighted LASSO (Liu et al., 2014), component LASSO (Hussami and Tibshirani, 2015), moderately clipped LASSO (Kwon et al., 2015), and multiple imputation random LASSO (Liu et al., 2016). I have left using these penalties with a FinMix GLMM for future work.

Similarly, several possible extensions relating to ALASSO have been proposed, and could be applied to a FinMix GLMM. These include Bayesian adaptive LASSO (Leng et al., 2014), the distribution of the estimates (Potscher and Schneider, 2009), the potential for model misspecification (van de Geer et al., 2011), robustness to model misspecification (Lu et al., 2012), rates of convergence (Chatterjee and Lahiri, 2013), false discovery rate (Sampson et al., 2013), tuning parameter selection (Hui et al., 2015), and post variable-selection inference (Chatterjee et al., 2015).

Extensions of the SCAD penalty function have been published as well, but I have not considered these in the FinMix GLMM case. Wang and Li (2009) proposed a weighted Wilcoxon extension which is more robust to outliers. Kwon et al. (2011) explored a quadratic approximation extension to SCAD called Q-SCAD. Alternatively, a quadratic approximation was used by Choi and Park (2012) to improve efficiency. Other expansions include SCAD for constrained variables (Ng and Yu, 2014), varying-coefficients models with autoregressive errors (Qiu et al., 2015), and generalised additive models with non-polynomial dimensionality (Li et al., 2012). I have not explored these options for a FinMix GLMM.

In addition to the previously discussed penalty functions, a few others have been proposed including ridge regression (Hoerl and Kennard, 1970), Least-Angle Regression (LARS) (Efron et al., 2004), elastic net (Zou and Hastie, 2005), MSCAD (Chen and Khalili, 2008), VISA (Mkhadri and Ouhourane, 2015), and minimum ϕ-divergence estimation (Sakate and Kashid, 2014). Hui et al. (2017), shrink the fixed effect to zero only if the corresponding random effect is or has already been shrunk to zero, which is a desirable property. Yu and Wang (2019) described another interesting penalty where both the mixing proportions as well as the regression coefficients are penalised. A robust variable selection method using minimum-distance techniques has been suggested in Tang and Karunamuni (2018). Further exploration of these penalty functions is a possibility.

Two large issues in statistics that have not been considered in this thesis are measurement error and missing data. Torabi (2013) explored measurement error in covariates of a GLMM, while Noh et al. (2012) examined both measurement error, and missing data. Yao and Song (2015) looked at measurement error in finite mixtures of models.

Model misspecification is a large topic, and one that could be further developed with respect to FinMix GLMMs. The issues, and consequences of model misspecification in GLMMs are considered in Abad et al. (2010), McCulloch and Neuhaus (2011), Heggeseth and Jewell (2013). The work of Heagerty and Kurland (2001), Litiere et al. (2007), Alonso et al. (2008), Huang (2009), Cox and Wong (2010), and Neuhaus et al. (2013) all focus on misspecification of the random effects.

While I took a frequentist approach in this thesis, a Bayesian approach to modelling of GLMM, and LMM has been discussed by Kizilkaya and Tempelman (2005), Natarajan and Kass (2000), and Li et al. (2014). Alternatively, Wolfinger and Oconnell (1993) explored a pseudo-likelihood approach. Exploration of FinMix GLMMs using these methods is a possibility for further work. Eskandari and Ormoz (2016), and Hunter and Young (2012) explored generalised semi-parameteric models. Ormoz and Eskandari (2016) look at the problem of variable selection in finite mixtures of semi-parameteric models. Huang et al. (2013) showed a nonparameteric finite mixture of regression models. Wang et al. (2014) showed finite mixtures of GLMs in both the semi-parameteric, and nonparameteric cases.

A number of excellent suggestions came up during my thesis defence. I assumed that an individual must remain in the same subpopulation through time, and relaxing this assumption would be an interesting extension. Using Metropolis-Hasting sampling assumes a Gaussian distribution, and exploring if the results are sensitive to this assumption is left to future work. A more in depth consideration of the impact of correlated covariates on convergence and identifiability is also an idea for future simulations and consideration. Similarly, further simulations considering the case where one subpopulation is quite small, such as a rare form of a disease would be interesting. Another possible idea is to adjust the Bayesian Information Criterion (BIC) when LASSO is used as in Bhattacharya and McNicholas (2014).

These additional possibilities show that more research is possible in the area of FinMix GLMMs.

Appendices

APPENDIX A
Notation

a - Value used in the SCAD penalty

$\mathbf{b}_{ki}$ - Possible value of $\mathbf{b}_i$ assuming subpopulation k

$\mathbf{b}_i = (b_{i1}, b_{i2}, b_{i3}, \ldots, b_{iq})^\top \sim Gaussian(0, \boldsymbol{I}_q)$ - Vector (length q) of standard Gaussian random effects for subject i

$$\mathbb{b} = \begin{bmatrix} b_{11} & b_{21} & \ldots & b_{n1} \\ b_{12} & b_{22} & \ldots & b_{n2} \\ \vdots & \vdots & \ddots & \vdots \\ b_{1q} & b_{2q} & \ldots & b_{nq} \end{bmatrix}$$ - Matrix (size $n \times q$) of all standard Gaussian random effects

$$\mathbb{C}_k = \begin{bmatrix} 1 & 0 & \ldots & 0 \\ C_{k21} & 1 & \ldots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ C_{kq1} & C_{kq2} & \ldots & 1 \end{bmatrix} = \begin{bmatrix} 1 & 0 & \ldots & 0 \\ \frac{\Gamma_{k21}}{\Gamma_{k22}} & 1 \ldots & 0 & \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\Gamma_{kq1}}{\Gamma_{kqq}} & \frac{\Gamma_{kq2}}{\Gamma_{kqq}} & \ldots & 1 \end{bmatrix}$$ - Matrix (size $q \times q$) of decomposed covariances of random effects

$\mathbb{C}_k^* = (C_{k21}, C_{k31}, C_{k32}, \ldots C_{kq(q-1)})^\top$ - Vector (length $\frac{q(q+1)}{2}$) version of $\mathbb{C}_k$

$$\mathbb{D}_k = \mathbb{\Gamma}_k \mathbb{\Gamma}_k^\top = \begin{bmatrix} D_{k11} & D_{k21} & \ldots & D_{kq1} \\ D_{k21} & D_{k22} & \ldots & D_{kq2} \\ \vdots & \vdots & \ddots & \vdots \\ D_{kq1} & D_{kq2} & \ldots & D_{kqq} \end{bmatrix}$$ - Variance-covariance matrix (size $q \times q$) for random effects

$$\mathbb{d}_k = \begin{bmatrix} d_{k1} & 0 & \dots & 0 \\ 0 & d_{k2} & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & d_{kq} \end{bmatrix} = \begin{bmatrix} \Gamma_{k11} & 0 & \dots & 0 \\ 0 & \Gamma_{k22} & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & \Gamma_{kqq} \end{bmatrix}$$ - Diagonal matrix (size $q \times q$) of standard deviations of random effects

$\mathbb{d}_k^* = (d_{k1}, d_{k2}, \dots, d_{kq})^\top = (\Gamma_{k11}, \Gamma_{k22}, \dots, \Gamma_{kqq})^\top$ - Vector (length q) version of $\mathbb{d}_k$

df - Degrees of freedom

$g(\mathbf{x}_{ij}\boldsymbol{\beta}_k + \mathbf{z}_{ij}\Gamma_k \mathbf{b}_i)$ - Canonical link function

$\mathbb{H}_{\mathbf{b}_{ki}}$ - Hessian matrix (size $n \times q$) at $\mathbf{b}_{ki}$

$I_\ell(\boldsymbol{\Theta}) = \sum_{i=1}^n s(\mathbf{y}_i|\mathbb{X}_i, \mathbb{Z}_i, \boldsymbol{\Theta})[s(\mathbf{y}_i|\mathbb{X}_i, \mathbb{Z}_i, \boldsymbol{\Theta})]^\top$ - Empirical observed information matrix (size $(K-1+K*p) \times (K-1+K*p)$)

K - Number of subpopulations

L - Number of potential values generated to approximate the integral in the likelihood

$L_n(\boldsymbol{\Theta})$ - Likelihood function at $\boldsymbol{\Theta}$ from a sample of size n

$\ell_n(\boldsymbol{\Theta})$ - Log-likelihood function at $\boldsymbol{\Theta}$ from a sample of size n

$\ell_{n\lambda}^\sharp(\underline{\boldsymbol{\theta}}_k)$ - penalised log-likelihood function at $\underline{\boldsymbol{\theta}}_k$ from a sample of size n

$M = \sum_{i=1}^n n_i$ - Total number of observations

m_{ij} - Number of trials for the i^{th} subject at the j^{th} observation when the outcome follows a binomial distribution

$\mathbf{m}_i = (m_{i1}, m_{i2}, m_{i3}, \dots, m_{in_i})^\top$ - Vector (length n_i) containing the number of trials for the i^{th} subject when the outcome follows a binomial distribution

$\mathbf{m} = (m_{11}, m_{12}, m_{13}, \dots, m_{1n_1}, m_{21}, m_{22}, m_{23}, \dots, m_{2n_2}, \dots, m_{n1}, m_{n2}, m_{n3}, \dots, m_{nn_n})^\top$ - Vector (length M) containing the number of trials when the outcome follows a binomial distribution

$\boldsymbol{N} = (n_1, n_2, n_3, \ldots, n_n)^\top$ - Vector (length n) of number of observations for all subjects

n - Number of subjects

n_i - Number of observations for the i^{th} subject

n_k - Estimated number of subjects in subpopulation k

p - Number of fixed effects

$p_\lambda(\underline{\boldsymbol{\theta}}_k)$ - Penalty function at $\underline{\boldsymbol{\theta}}_k$ with tuning parameter λ

$Q_k(\boldsymbol{\theta}_k)$ - Approximate likelihood

q - Number of random effects

$s(\mathbf{y}_i|\mathbb{X}_i, \mathbb{Z}_i, \boldsymbol{\Theta}) = \sum_{k=1}^{K} \frac{\partial}{\partial \boldsymbol{\Theta}} \tau_{ki} \log[\pi_k f_{\mathbf{y}_i}(\mathbf{y}_i|\mathbb{X}_i, \mathbb{Z}_i, \boldsymbol{\Theta})]$ - Vector (length $K - 1 + K * p$) containing gradients of the complete log-likelihood, the complete-data score statistic

$U_\lambda(\underline{\boldsymbol{\theta}}_k)$ - Vector (length $\varkappa$) used in penalised maximum likelihood

w_h - Weight for parameter h, used in ALASSO

x_{ijh} - Value of the h^{th} covariate, for which there is a fixed effect, for the i^{th} subject at the j^{th} observation

$\mathbf{x}_{ij}$ - Vector (length p) of covariates for which there is a fixed effect, for the i^{th} subject at the j^{th} observation

$\mathbb{X}_i = \begin{bmatrix} X_{i11} & X_{i12} & \ldots & X_{i1p} \\ X_{i21} & X_{i22} & \ldots & X_{i2p} \\ \vdots & \vdots & \ddots & \vdots \\ X_{in_i1} & X_{in_i2} & \ldots & X_{in_ip} \end{bmatrix}$ - Matrix (size $n_i \times p$) of covariates for which there is a fixed effect, for the i^{th} subject

$$\mathbb{X} = \begin{bmatrix} X_{111} & X_{112} & \dots & X_{11p} \\ X_{121} & X_{122} & \dots & X_{12p} \\ \vdots & \vdots & \ddots & \vdots \\ X_{1n_11} & X_{1n_12} & \dots & X_{1n_1p} \\ X_{211} & X_{212} & \dots & X_{21p} \\ \vdots & \vdots & \ddots & \vdots \\ X_{nn_n1} & X_{nn_n2} & \dots & X_{nn_np} \end{bmatrix}$$ - Matrix (size $M \times p$) of covariates for which there is a fixed effect, for all subjects

Y_{ij} - Outcome for the i^{th} subject at the j^{th} observation (random variable)

y_{ij} - Outcome for the i^{th} subject at the j^{th} observation (realisation of the random variable)

$\mathbf{Y}_i = (Y_{i1}, Y_{i2}, Y_{i3}, \dots, Y_{in_i})^\top$ - Vector (length n_i) of outcomes for the i^{th} subject (random variable)

$\mathbf{y}_i = (y_{i1}, y_{i2}, y_{i3}, \dots, y_{in_i})^\top$ - Vector (length n_i) of outcomes for the i^{th} subject (realisation of the random variable)

$\mathbf{Y} = (Y_{11}, Y_{12}, Y_{13}, \dots, Y_{1n_1}, Y_{21}, Y_{22}, Y_{23}, \dots, Y_{2n_2}, \dots, Y_{n1}, Y_{n2}, Y_{n3}, \dots, Y_{nn_n})^\top$ - Vector (length M) of all outcomes (random variable)

$\mathbf{y} = (y_{11}, y_{12}, y_{13}, \dots, y_{1n_1}, y_{21}, y_{22}, y_{23}, \dots, y_{2n_2}, \dots, y_{n1}, y_{n2}, y_{n3}, \dots, y_{nn_n})^\top$ - Vector (length M) of all outcomes (realisation of the random variable)

z_{ijh} - Value of the h^{th} covariate, for which there is a random effect, for the i^{th} subject at the j^{th} observation

$\mathbf{z}_{ij}$ - Vector (length q) of covariates for which there is a random effect, for the i^{th} subject at the j^{th} observation

$$\mathbb{Z}_i = \begin{bmatrix} Z_{i11} & Z_{i12} & \dots & Z_{i1q} \\ Z_{i21} & Z_{i22} & \dots & Z_{i2q} \\ \vdots & \vdots & \ddots & \vdots \\ Z_{in_i1} & Z_{in_i2} & \dots & Z_{in_iq} \end{bmatrix}$$ - Matrix (size $n_i \times q$) of covariates for which there is a random effect, for the i^{th} subject

$$\mathbb{Z} = \begin{bmatrix} Z_{111} & Z_{112} & \dots & Z_{11q} \\ Z_{121} & Z_{122} & \dots & Z_{12q} \\ \vdots & \vdots & \ddots & \vdots \\ Z_{1n_11} & Z_{1n_12} & \dots & Z_{1n_1q} \\ Z_{211} & Z_{212} & \dots & Z_{21q} \\ \vdots & \vdots & \ddots & \vdots \\ Z_{nn_n1} & Z_{nn_n2} & \dots & Z_{nn_nq} \end{bmatrix}$$ - Matrix (size $M \times q$) of covariates for which there is a random effect, for all subjects

α - Exponent on tuning parameter in the penalty function

$\boldsymbol{\beta}_k = (\beta_{k1}, \beta_{k2}, \beta_{k3}, \dots, \beta_{kp})^\top$ - Vector (length p) of coefficients for fixed effects, in the k^{th} subpopulation

$$\mathbb{\Gamma}_k = \begin{bmatrix} \Gamma_{k11} & 0 & \dots & 0 \\ \Gamma_{k21} & \Gamma_{k22} & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ \Gamma_{kq1} & \Gamma_{kq2} & \dots & \Gamma_{kqq} \end{bmatrix}$$ - Transformation matrix (size $q \times q$, lower triangle) for random effects, in the k^{th} subpopulation $\mathbb{\Gamma}_k$

$\mathbb{\Gamma}_k^* = (\Gamma_{k11}, \Gamma_{k21}, \Gamma_{k22}, \Gamma_{k31}, \Gamma_{k32}, \Gamma_{k33}, \dots \Gamma_{kqq})^\top$ - Vector (length $\frac{q(q+1)}{2}$) version of $\mathbb{\Gamma}_k$

δ - Difference between successive values of λ_k in the grid of possible values for λ_k

ϵ - Value added to the denominator in $\mathbb{\Sigma}_\lambda$

η_{kij} - Sum of fixed, and random effects for the i^{th} subject at the j^{th} observation, assuming membership in the k^{th} subpopulation

$\boldsymbol{\eta}_{ki} = (\eta_{ki1}, \eta_{ki2}, \eta_{ki3}, \ldots, \eta_{kin_i})^\top = \mathbb{X}_i\boldsymbol{\beta}_k + \mathbb{Z}_i\mathbb{\Gamma}_k\mathbf{b}_i$ - Vector (length n_i) of sums of fixed and random effects for the i^{th} subject, assuming membership in the k^{th} subpopulation

$\boldsymbol{\eta}_k = (\eta_{k11}, \eta_{k12}, \eta_{k13}, \ldots, \eta_{k1n_1}, \eta_{k21}, \eta_{k22}, \eta_{k23}, \ldots, \eta_{k2n_2}, \ldots, \eta_{kn1}, \eta_{kn2}, \eta_{kn3}, \ldots, \eta_{knn_n})^\top$ - Vector (length M) of all sums of fixed and random effects

$\boldsymbol{\Theta} = (\boldsymbol{\pi}^\top, \boldsymbol{\beta}_1^\top, \mathbb{\Gamma}_1^{*\top}, \boldsymbol{\beta}_2^\top, \mathbb{\Gamma}_2^{*\top}, \ldots, \boldsymbol{\beta}_K^\top, \mathbb{\Gamma}_K^{*\top})^\top$ - Vector (length $K-1+K\times\varkappa$) of $\boldsymbol{\pi}$, $\boldsymbol{\beta}_k$, and $\mathbb{\Gamma}_k^*$

$\tilde{\boldsymbol{\Theta}} = (\boldsymbol{\pi}^\top, \boldsymbol{\beta}_1^\top, \boldsymbol{\beta}_2^\top, \ldots, \boldsymbol{\beta}_K^\top)^\top$ - Vector (length $K-1+K\times p$) of $\boldsymbol{\pi}$ and $\boldsymbol{\beta}_k$, used in the calculation of standard errors

$\boldsymbol{\theta}_k = (\boldsymbol{\beta}_k^\top, \mathbb{\Gamma}_k^{*\top})^\top$ - Vector (length $\varkappa$) of $\boldsymbol{\beta}_k$ and $\mathbb{\Gamma}_k^*$

$\underline{\boldsymbol{\theta}}_k = (\boldsymbol{\beta}_k^\top, \mathbb{d}_k^{*\top}, \mathbb{C}_k^{*\top})^\top$ - Vector (length $\varkappa$) of $\boldsymbol{\beta}_k$, $\mathbb{d}_k^*$, and $\mathbb{C}_k^*$

ι - Exponent on $\frac{1}{2}$ to facilitate half step Newton-Raphson

$\varkappa = p + \frac{1}{2}q(q+1)$ - Length of $\boldsymbol{\theta}_k$, number of parameters in the k^{th} subpopulation

$\boldsymbol{\lambda} = (\lambda_1, \lambda_2, \lambda_3, \ldots, \lambda_K)$ - Vector (length K) of tuning parameters used in penalised maximum likelihood

$\nu_{ki} = (\nu_{ki1}, \nu_{ki2}, \ldots, \nu_{kiq})^\top = \mathbb{\Gamma}_k\mathbf{b}_i \sim Gaussian(0, \mathbb{D}_k)$ - Row vector (length q) of random effects for the i^{th} subject

$\nu_k = \begin{bmatrix} \nu_{k11} & \nu_{k21} & \ldots & \nu_{kn1} \\ \nu_{k12} & \nu_{k22} & \ldots & \nu_{kn2} \\ \vdots & \vdots & \ddots & \vdots \\ \nu_{k1q} & \nu_{k2q} & \ldots & \nu_{knq} \end{bmatrix}$ - Matrix (size $n \times q$) of all transformed random effects

ξ_{kij} - Expected outcome for the i^{th} subject at the j^{th} observation, assuming the k^{th} subpopulation (Poisson)

$\boldsymbol{\xi}_{ki} = (\xi_{ki1}, \xi_{ki2}, \xi_{ki3}, \ldots, \xi_{kin_i})^\top$ - Vector (length n_i) of expected outcome for the i^{th} subject, $\log(\boldsymbol{\xi}_{ki}) = \boldsymbol{\eta}_{ki} = \mathbb{X}_i\boldsymbol{\beta}_k + \mathbb{Z}_i\mathbb{\Gamma}_k\mathbf{b}_i$, $\boldsymbol{\xi}_{ki} = \exp(\boldsymbol{\eta}_{ki}) = \exp(\mathbb{X}_i\boldsymbol{\beta}_k + \mathbb{Z}_i\mathbb{\Gamma}_k\mathbf{b}_i)$ (Poisson)

$\boldsymbol{\xi}_k = (\xi_{k11}, \xi_{k12}, \xi_{k13}, \ldots, \xi_{knn_n})^\top$ - Vector (length M) of all expected observations (Poisson)

π_k - Mixing proportion, the proportion of the underlying population in subpopulation k with $\sum_{k=1}^{K} \pi_k = 1$

$\boldsymbol{\pi} = (\pi_1, \pi_2, \pi_3, \ldots, \pi_{(K-1)})^\top$ - Vector (length $K - 1$) of mixing proportions

$\mathbb{\Sigma}_\lambda(\underline{\boldsymbol{\theta}}_k)$ or $\mathbb{\Sigma}_\lambda(\boldsymbol{\theta}_k)$ - Matrix (size $\varkappa \times \varkappa$) used in the penalised maximum likelihood

τ_{ki} - Membership probability for subject i in subpopulation k

$\tau_k = (\tau_{k1}, \tau_{k2}, \tau_{k3}, \ldots, \tau_{kn})^\top$ - Vector (length n) of membership probability for all subjects in subpopulation k

$\tau_i = (\tau_{1i}, \tau_{2i}, \tau_{3i}, \ldots, \tau_{Ki})^\top$ - Vector (length K) of membership probability for one subjects in each of the K subpopulations

φ_{kij} - Expected outcome for the i^{th} subject at the j^{th} observation, assuming the k^{th} subpopulation (binomial)

$\boldsymbol{\varphi}_{ki} = (\varphi_{ki1}, \varphi_{ki2}, \varphi_{ki3}, \ldots, \varphi_{kin_i})^\top$ - Vector (length n_i) of expected outcome for the i^{th} subject, $\text{logit}(\varphi_{bi}) = \boldsymbol{\eta}_{ki} = \mathbb{X}_i\boldsymbol{\beta}_k + \mathbb{Z}_i\mathbb{\Gamma}_k\mathbb{b}$, $\boldsymbol{\varphi}_{ki} = expit(\boldsymbol{\eta}_{ki}) = \exp(\mathbb{X}_i\boldsymbol{\beta}_k + \mathbb{Z}_i\mathbb{\Gamma}_k\mathbf{b}_i)$ (binomial)

$\boldsymbol{\varphi}_{=}(\varphi_{k11}, \varphi_{k12}, \varphi_{k13}, \ldots, \varphi_{knn_n})^\top$ - Vector (length M) of all expected observations (binomial)

APPENDIX B
Details of the Derivatives

This appendix provides details on the maximisation of the approximate likelihood described in this thesis. Recall that the likelihood of a FinMix GLMM is

$$
\begin{aligned}
L(\boldsymbol{\Theta}) &= \prod_{i=1}^{n} f_{\mathbf{y}_i}(\mathbf{y}_i|\mathbb{X}_i, \mathbb{Z}_i, \boldsymbol{\Theta}) \\
&= \prod_{i=1}^{n}\sum_{k=1}^{K} \pi_k f_{\mathbf{y}_i}^{(k)}(\mathbf{y}_i|\mathbb{X}_i, \mathbb{Z}_i, \boldsymbol{\theta}_k) \\
&= \prod_{i=1}^{n}\sum_{k=1}^{K} \pi_k \int f_{\mathbf{y}_i|\mathbf{b}_i}^{(k)}(\mathbf{y}_i|\mathbb{X}_i, \mathbb{Z}_i, \mathbf{b}_i, \boldsymbol{\theta}_k) f(\mathbf{b}_i) d\mathbf{b}_i \\
&= \prod_{i=1}^{n}\sum_{k=1}^{K} \pi_k \int \left[\prod_{j=1}^{n_i} f_{y_{ij}|\mathbf{b}_i}^{(k)}(y_{ij}|\mathbf{x}_{ij}, \mathbf{z}_{ij}, \mathbf{b}_i, \boldsymbol{\theta}_k)\right] f(\mathbf{b}_i) d\mathbf{b}_i \\
&\approx \prod_{i=1}^{n}\sum_{k=1}^{K} \pi_k \frac{1}{L}\sum_{\ell=1}^{L}\prod_{j=1}^{n_i} f_{y_{ij}|\mathbf{b}_i}^{(k)}(y_{ij}|\mathbf{x}_{ij}, \mathbf{z}_{ij}, \mathbf{b}_i^{(\ell)}, \boldsymbol{\theta}_k) \\
&= \prod_{i=1}^{n}\frac{1}{L}\sum_{k=1}^{K} \pi_k \sum_{\ell=1}^{L}\prod_{j=1}^{n_i} f_{y_{ij}|\mathbf{b}_i}^{(k)}(y_{ij}|\mathbf{x}_{ij}, \mathbf{z}_{ij}, \mathbf{b}_i^{(\ell)}, \boldsymbol{\theta}_k) \\
&= \frac{1}{L^n}\prod_{i=1}^{n}\sum_{k=1}^{K} \pi_k \sum_{\ell=1}^{L}\prod_{j=1}^{n_i} f_{y_{ij}|\mathbf{b}_i}^{(k)}(y_{ij}|\mathbf{x}_{ij}, \mathbf{z}_{ij}, \mathbf{b}_i^{(\ell)}, \boldsymbol{\theta}_k) \\
&\propto \prod_{i=1}^{n}\sum_{k=1}^{K} \pi_k \sum_{\ell=1}^{L}\prod_{j=1}^{n_i} f_{y_{ij}|\mathbf{b}_i}^{(k)}(y_{ij}|\mathbf{x}_{ij}, \mathbf{z}_{ij}, \mathbf{b}_i^{(\ell)}, \boldsymbol{\theta}_k).
\end{aligned}
$$

Thus, the log-likelihood is

$$\begin{aligned}
\ell(\boldsymbol{\Theta}) &\approx \log\left(\prod_{i=1}^{n}\sum_{k=1}^{K}\pi_k\frac{1}{L}\sum_{\ell=1}^{L}\prod_{j=1}^{n_i} f^{(k)}_{y_{ij}|\mathbf{b}_i}(y_{ij}|\mathbf{x}_{ij},\mathbf{z}_{ij},\mathbf{b}_i^{(\ell)},\boldsymbol{\theta}_k)\right)\\
&= \sum_{i=1}^{n}\log\left(\sum_{k=1}^{K}\pi_k\frac{1}{L}\sum_{\ell=1}^{L}\prod_{j=1}^{n_i} f^{(k)}_{y_{ij}|\mathbf{b}_i}(y_{ij}|\mathbf{x}_{ij},\mathbf{z}_{ij},\mathbf{b}_i^{(\ell)},\boldsymbol{\theta}_k)\right)\\
&= \sum_{i=1}^{n}\left[\log\left(\sum_{k=1}^{K}\pi_k\sum_{\ell=1}^{L}\prod_{j=1}^{n_i} f^{(k)}_{y_{ij}|\mathbf{b}_i}(y_{ij}|\mathbf{x}_{ij},\mathbf{z}_{ij},\mathbf{b}_i^{(\ell)},\boldsymbol{\theta}_k)\right)-\log(L)\right]\\
&= \sum_{i=1}^{n}\log\left(\sum_{k=1}^{K}\pi_k\sum_{\ell=1}^{L}\prod_{j=1}^{n_i} f^{(k)}_{y_{ij}|\mathbf{b}_i}(y_{ij}|\mathbf{x}_{ij},\mathbf{z}_{ij},\mathbf{b}_i^{(\ell)},\boldsymbol{\theta}_k)\right)-\sum_{i=1}^{n}\log(L)\\
&= \sum_{i=1}^{n}\log\left(\sum_{k=1}^{K}\pi_k\sum_{\ell=1}^{L}\prod_{j=1}^{n_i} f^{(k)}_{y_{ij}|\mathbf{b}_i}(y_{ij}|\mathbf{x}_{ij},\mathbf{z}_{ij},\mathbf{b}_i^{(\ell)},\boldsymbol{\theta}_k)\right)-n\log(L)\\
&\propto \sum_{i=1}^{n}\log\left(\sum_{k=1}^{K}\pi_k\sum_{\ell=1}^{L}\prod_{j=1}^{n_i} f^{(k)}_{y_{ij}|\mathbf{b}_i}(y_{ij}|\mathbf{x}_{ij},\mathbf{z}_{ij},\mathbf{b}_i^{(\ell)},\boldsymbol{\theta}_k)\right).
\end{aligned}$$

The remainder of this Appendix shows the details of the maximisation of the MLE for a single subpopulation k in both the Poisson and binomial case. The first two sections show these derivatives with respect to $\boldsymbol{\beta}_k$ and $\mathbb{\Gamma}_k^*$ so these correspond to the MLE described in Chapter 3 and Chapter 4 when only the fixed effects are penalised. The last two sections show these derivatives with respect to $\boldsymbol{\beta}_k$, $\mathbb{d}^*$, and $\mathbb{C}_k^*$ as described in Chapter 4 to facilitate penalisation of random effects.

B.1 Details of Maximisation of the Maximum Likelihood Estimate for a Single Subpopulation K, Poisson Case

In order to maximise the approximate likelihood using Newton-Raphson, I calculated the first, and second derivatives of the approximate likelihood. As an example, the details of the derivatives for each element of $\boldsymbol{\theta}_k$ are shown here. In the following example, $p = 3$ and $q = 2$. Thus, $\mathbf{x}_{ij}\boldsymbol{\beta}_k + \mathbf{z}_{ij}\mathbb{\Gamma}_k\mathbf{b}_i = x_{ij1}\beta_{k1} + x_{ij2}\beta_{k2} +$

$x_{ij3}\beta_{k3} + z_{ij1}\Gamma_{k11}b_{i1} + z_{ij2}\Gamma_{k12}b_{i1} + z_{ij2}\Gamma_{k22}b_{i2}$ and

$$\begin{aligned} Q_k(\boldsymbol{\theta}_k) &= \frac{1}{L}\sum_{l=1}^{L}\sum_{i=1}^{n}\sum_{j=1}^{n_i}[y_{ij}x_{ij1}\beta_{k1} + y_{ij}x_{ij2}\beta_{k2} \\ &\quad + y_{ij}x_{ij3}\beta_{k3} + y_{ij}z_{ij1}\Gamma_{k11}b_{i1} + y_{ij}z_{ij2}\Gamma_{k12}b_{i1} + y_{ij}z_{ij2}\Gamma_{k22}b_{i2} \\ &\quad - e^{x_{ij1}\beta_{k1}+x_{ij2}\beta_{k2}+x_{ij3}\beta_{k3}+z_{ij1}\Gamma_{k11}b_{i1}+z_{ij2}\Gamma_{k12}b_{i1}+z_{ij2}\Gamma_{k22}b_{i2}}.] \end{aligned}$$

In an effort to simplify the notation slightly, I shortened $\mathbf{b}_i^{(s,l)}$ to $\mathbf{b}_i$, did not include weights τ_{ki}, and condensed $\sum_{l=1}^{L}\sum_{i=1}^{n}\sum_{j=1}^{n_i}$ to $\sum_{l,i,j}$.

First, consider the first derivative.

$$\begin{aligned} Q_k'(\boldsymbol{\theta}_k) &= \frac{\partial Q_k(\boldsymbol{\theta}_k)}{\partial \theta_k} = \begin{bmatrix} \frac{\partial Q_k(\boldsymbol{\theta}_k)}{\partial \boldsymbol{\beta}_k} \\ \frac{\partial Q_k(\boldsymbol{\theta}_k)}{\partial \Gamma_k^*} \end{bmatrix} \\ &= \begin{bmatrix} \frac{1}{L}\sum_{l,i,j} x_{ij1}\left[y_{ij} - e^{x_{ij1}\beta_{k1}+x_{ij2}\beta_{k2}+x_{ij3}\beta_{k3}+z_{ij1}\Gamma_{k11}b_{i1}+z_{ij2}\Gamma_{k12}b_{i1}+z_{ij2}\Gamma_{k22}b_{i2}}\right] \\ \frac{1}{L}\sum_{l,i,j} x_{ij2}\left[y_{ij} - e^{x_{ij1}\beta_{k1}+x_{ij2}\beta_{k2}+x_{ij3}\beta_{k3}+z_{ij1}\Gamma_{k11}b_{i1}+z_{ij2}\Gamma_{k12}b_{i1}+z_{ij2}\Gamma_{k22}b_{i2}}\right] \\ \frac{1}{L}\sum_{l,i,j} x_{ij3}\left[y_{ij} - e^{x_{ij1}\beta_{k1}+x_{ij2}\beta_{k2}+x_{ij3}\beta_{k3}+z_{ij1}\Gamma_{k11}b_{i1}+z_{ij2}\Gamma_{k12}b_{i1}+z_{ij2}\Gamma_{k22}b_{i2}}\right] \\ \frac{1}{L}\sum_{l,i,j} b_{i1}z_{ij1}\left[y_{ij} - e^{x_{ij1}\beta_{k1}+x_{ij2}\beta_{k2}+x_{ij3}\beta_{k3}+z_{ij1}\Gamma_{k11}b_{i1}+z_{ij2}\Gamma_{k12}b_{i1}+z_{ij2}\Gamma_{k22}b_{i2}}\right] \\ \frac{1}{L}\sum_{l,i,j} b_{i1}z_{ij2}\left[y_{ij} - e^{x_{ij1}\beta_{k1}+x_{ij2}\beta_{k2}+x_{ij3}\beta_{k3}+z_{ij1}\Gamma_{k11}b_{i1}+z_{ij2}\Gamma_{k12}b_{i1}+z_{ij2}\Gamma_{k22}b_{i2}}\right] \\ \frac{1}{L}\sum_{l,i,j} b_{i2}z_{ij2}\left[y_{ij} - e^{x_{ij1}\beta_{k1}+x_{ij2}\beta_{k2}+x_{ij3}\beta_{k3}+z_{ij1}\Gamma_{k11}b_{i1}+z_{ij2}\Gamma_{k12}b_{i1}+z_{ij2}\Gamma_{k22}b_{i2}}\right] \end{bmatrix} \\ &= \frac{1}{L}\sum_{l,i,j} \begin{bmatrix} x_{ij1} \\ x_{ij2} \\ x_{ij3} \\ z_{ij1}b_{i1} \\ z_{ij2}b_{i1} \\ z_{ij2}b_{i2} \end{bmatrix} \left(y_{ij} - e^{x_{ij1}\beta_{k1}+x_{ij2}\beta_{k2}+x_{ij3}\beta_{k3}+z_{ij1}\Gamma_{k11}b_{i1}+z_{ij2}\Gamma_{k12}b_{i1}+z_{ij2}\Gamma_{k22}b_{i2}}\right) \end{aligned}$$

I was then able to calculate the second derivative.

$$
\begin{aligned}
Q_k''(\boldsymbol{\theta}_k) &= \frac{\partial^2 Q_k(\boldsymbol{\theta}_k)}{\partial \boldsymbol{\theta}_k^\top \partial \boldsymbol{\theta}_k} = \begin{bmatrix} \frac{\partial^2 Q_k(\boldsymbol{\theta}_k)}{\partial \boldsymbol{\beta}^\top \partial \boldsymbol{\beta}} & \frac{\partial^2 Q_k(\boldsymbol{\theta}_k)}{\partial \boldsymbol{\beta}^\top \partial \Gamma^*} \\ \frac{\partial^2 Q_k(\boldsymbol{\theta}_k)}{\partial \Gamma^{*\top} \partial \boldsymbol{\beta}} & \frac{\partial^2 Q_k(\boldsymbol{\theta}_k)}{\partial \Gamma^{*\top} \partial \Gamma^*} \end{bmatrix} \\
&= \frac{1}{L} \sum_{l,i,j} -e^{x_{ij1}\beta_{k1}+x_{ij2}\beta_{k2}+x_{ij3}\beta_{k3}+z_{ij1}\Gamma_{k11}b_{i1}+z_{ij2}\Gamma_{k12}b_{i1}+z_{ij2}\Gamma_{k22}b_{i2}} \times \\
&\begin{bmatrix}
(x_{ij1})^2 & x_{ij1}x_{ij2} & x_{ij1}x_{ij3} & x_{ij1}z_{ij1}b_{i1} & x_{ij1}z_{ij2}b_{i1} & x_{ij1}z_{ij2}b_{i2} \\
x_{ij1}x_{ij2} & (x_{ij2})^2 & x_{ij2}x_{ij3} & x_{ij2}z_{ij1}b_{i1} & x_{ij2}z_{ij2}b_{i1} & x_{ij2}z_{ij2}b_{i2} \\
x_{ij1}x_{ij3} & x_{ij2}x_{ij3} & (x_{ij3})^2 & x_{ij3}z_{ij1}b_{i1} & x_{ij3}z_{ij2}b_{i1} & x_{ij3}z_{ij2}b_{i2} \\
x_{ij1}z_{ij1}b_{i1} & x_{ij2}z_{ij1}b_{i1} & x_{ij3}z_{ij1}b_{i1} & (z_{ij1}b_{i1})^2 & z_{ij1}z_{ij2}(b_{i1})^2 & z_{ij1}z_{ij2}b_{i1}b_{i2} \\
x_{ij1}z_{ij2}b_{i1} & x_{ij2}z_{ij2}b_{i1} & x_{ij3}z_{ij2}b_{i1} & z_{ij1}z_{ij2}(b_{i1})^2 & (z_{ij2}b_{i1})^2 & (z_{ij2})^2 b_{i1}b_{i2} \\
x_{ij1}z_{ij2}b_{i2} & x_{ij2}z_{ij2}b_{i2} & x_{ij3}z_{ij2}b_{i2} & z_{ij1}z_{ij2}b_{i1}b_{i2} & (z_{ij2})^2 b_{i1}b_{i2} & (z_{ij2}b_{i2})^2
\end{bmatrix} \\
&= \frac{1}{L} \sum_{l,i,j} \begin{bmatrix} x_{ij1} \\ x_{ij2} \\ x_{ij3} \\ z_{ij1}b_{i1} \\ z_{ij2}b_{i1} \\ z_{ij2}b_{i2} \end{bmatrix} \begin{bmatrix} x_{ij1} \\ x_{ij2} \\ x_{ij3} \\ z_{ij1}b_{i1} \\ z_{ij2}b_{i1} \\ z_{ij2}b_{i2} \end{bmatrix}^\top \\
&\times \quad -e^{x_{ij1}\beta_{k1}+x_{ij2}\beta_{k2}+x_{ij3}\beta_{k3}+z_{ij1}\Gamma_{k11}b_{i1}+z_{ij2}\Gamma_{k12}b_{i1}+z_{ij2}\Gamma_{k22}b_{i2}}
\end{aligned}
$$

Also, consider the derivative with respect to $\mathbf{b}_i$ for a particular value of i, and k. Here, I used $q = 3$.

$$
\begin{aligned}
&\frac{\partial Q_k(\boldsymbol{\theta}_k)}{\partial \mathbf{b}_i} \\
&= \begin{bmatrix}
\sum_{j=1}^{n_i} (z_{ij1}\Gamma_{k11} + z_{ij2}\Gamma_{k12} + z_{ij3}\Gamma_{k13}) \left[y_{ij} - e^{x_{ij1}\beta_{k1}+x_{ij2}\beta_{k2}+x_{ij3}\beta_{k3}+z_{ij1}\Gamma_{k11}b_{i1}+z_{ij2}\Gamma_{k12}b_{i1}+z_{ij2}\Gamma_{k22}b_{i2}}\right] \\
\sum_{j=1}^{n_i} (z_{ij2}\Gamma_{k22} + z_{ij3}\Gamma_{k23}) \left[y_{ij} - e^{x_{ij1}\beta_{k1}+x_{ij2}\beta_{k2}+x_{ij3}\beta_{k3}+z_{ij1}\Gamma_{k11}b_{i1}+z_{ij2}\Gamma_{k12}b_{i1}+z_{ij2}\Gamma_{k22}b_{i2}}\right] \\
\sum_{j=1}^{n_i} (z_{ij3}\Gamma_{k33}) \left[y_{ij} - e^{x_{ij1}\beta_{k1}+x_{ij2}\beta_{k2}+x_{ij3}\beta_{k3}+z_{ij1}\Gamma_{k11}b_{i1}+z_{ij2}\Gamma_{k12}b_{i1}+z_{ij2}\Gamma_{k22}b_{i2}}\right]
\end{bmatrix} \\
&= \sum_{j=1}^{n_i} \begin{bmatrix} z_{ij1}\Gamma_{k11} + z_{ij2}\Gamma_{k12} + z_{ij3}\Gamma_{k13} \\ z_{ij2}\Gamma_{k22} + z_{ij3}\Gamma_{k23} \\ z_{ij3}\Gamma_{k33} \end{bmatrix} (y_{ij} - e^{x_{ij1}\beta_{k1}+x_{ij2}\beta_{k2}+x_{ij3}\beta_{k3}+z_{ij1}\Gamma_{k11}b_{i1}+z_{ij2}\Gamma_{k12}b_{i1}+z_{ij2}\Gamma_{k22}b_{i2}})
\end{aligned}
$$

The second derivatives are as follows.

$$\frac{\partial^2 Q_k(\boldsymbol{\theta}_k)}{\partial \mathbf{b}_i^\top \partial \mathbf{b}_i} = \sum_{j=1}^{n_i} \begin{bmatrix} z_{ij1}\Gamma_{k11} + z_{ij2}\Gamma_{k12} + z_{ij3}\Gamma_{k13} \\ z_{ij2}\Gamma_{k22} + z_{ij3}\Gamma_{k23} \\ z_{ij3}\Gamma_{k33} \end{bmatrix} \begin{bmatrix} z_{ij1}\Gamma_{k11} + z_{ij2}\Gamma_{k12} + z_{ij3}\Gamma_{k13} \\ z_{ij2}\Gamma_{k22} + z_{ij3}\Gamma_{k23} \\ z_{ij3}\Gamma_{k33} \end{bmatrix}^\top \times(-e^{\beta_{k1}+x_{ij2}\beta_{k2}+x_{ij3}\beta_{k3}+z_{ij1}\Gamma_{k11}b_{i1}+z_{ij2}\Gamma_{k12}b_{i1}+z_{ij2}\Gamma_{k22}b_{i2}})$$

B.2 Details of Maximisation of the Maximum Likelihood Estimate for a Single Subpopulation K, binomial Case

Similar to the previous section, $p = 3$, and $q = 2$. Thus, $\mathbf{x}_{ij}\boldsymbol{\beta}_k + \mathbf{z}_{ij}\mathbb{\Gamma}_k\mathbf{b}_i = x_{ij1}\beta_{k1} + x_{ij2}\beta_{k2} + x_{ij3}\beta_{k3} + z_{ij1}\Gamma_{k11}b_{i1} + z_{ij2}\Gamma_{k12}b_{i1} + z_{ij2}\Gamma_{k22}b_{i2}$, and

$$\begin{aligned} Q_k(\boldsymbol{\theta}_k) &= \frac{1}{L}\sum_{l=1}^{L}\sum_{i=1}^{n}\sum_{j=1}^{n_i}[y_{ij}x_{ij1}\beta_{k1} + y_{ij}x_{ij2}\beta_{k2} + y_{ij}x_{ij3}\beta_{k3} + y_{ij}z_{ij1}\Gamma_{k11}b_{i1} \\ &\quad + y_{ij}z_{ij2}\Gamma_{k12}b_{i1} + y_{ij}z_{ij2}\Gamma_{k22}b_{i2} \\ &\quad - m_{ij}\log(\exp[x_{ij1}\beta_{k1} + x_{ij2}\beta_{k2} + x_{ij3}\beta_{k3} + z_{ij1}\Gamma_{k11}b_{i1} \\ &\qquad + z_{ij2}\Gamma_{k12}b_{i1} + z_{ij2}\Gamma_{k22}b_{i2}] + 1)]. \end{aligned}$$

Again, in an effort to simplify the notation slightly, I shortened $\mathbf{b}_i^{(s,l)}$ to $\mathbf{b}_i$, did not include weights τ_{ki}, and condensed $\sum_{l=1}^{L}\sum_{i=1}^{n}\sum_{j=1}^{n_i}$ to $\sum_{l,i,j}$.

I calculated the first derivative.

$$
\begin{aligned}
Q'_k(\boldsymbol{\theta}_k) &= \frac{\partial Q_k(\boldsymbol{\theta}_k)}{\partial \boldsymbol{\theta}_k} = \begin{bmatrix} \frac{\partial Q_k(\boldsymbol{\theta}_k)}{\partial \boldsymbol{\beta}_k} \\ \frac{\partial Q_k(\boldsymbol{\theta}_k)}{\partial \mathbb{F}^*_k} \end{bmatrix} \\
&= \begin{bmatrix}
\frac{1}{L}\sum_{l,i,j} x_{ij1}\left[y_{ij} - m_{ij}\frac{e^{x_{ij1}\beta_{k1}+x_{ij2}\beta_{k2}+x_{ij3}\beta_{k3}+z_{ij1}\Gamma_{k11}b_{i1}+z_{ij2}\Gamma_{k12}b_{i1}+z_{ij2}\Gamma_{k22}b_{i2}}}{e^{x_{ij1}\beta_{k1}+x_{ij2}\beta_{k2}+x_{ij3}\beta_{k3}+z_{ij1}\Gamma_{k11}b_{i1}+z_{ij2}\Gamma_{k12}b_{i1}+z_{ij2}\Gamma_{k22}b_{i2}}+1}\right] \\
\frac{1}{L}\sum_{l,i,j} x_{ij2}\left[y_{ij} - m_{ij}\frac{e^{x_{ij1}\beta_{k1}+x_{ij2}\beta_{k2}+x_{ij3}\beta_{k3}+z_{ij1}\Gamma_{k11}b_{i1}+z_{ij2}\Gamma_{k12}b_{i1}+z_{ij2}\Gamma_{k22}b_{i2}}}{e^{x_{ij1}\beta_{k1}+x_{ij2}\beta_{k2}+x_{ij3}\beta_{k3}+z_{ij1}\Gamma_{k11}b_{i1}+z_{ij2}\Gamma_{k12}b_{i1}+z_{ij2}\Gamma_{k22}b_{i2}}+1}\right] \\
\frac{1}{L}\sum_{l,i,j} x_{ij3}\left[y_{ij} - m_{ij}\frac{e^{x_{ij1}\beta_{k1}+x_{ij2}\beta_{k2}+x_{ij3}\beta_{k3}+z_{ij1}\Gamma_{k11}b_{i1}+z_{ij2}\Gamma_{k12}b_{i1}+z_{ij2}\Gamma_{k22}b_{i2}}}{e^{x_{ij1}\beta_{k1}+x_{ij2}\beta_{k2}+x_{ij3}\beta_{k3}+z_{ij1}\Gamma_{k11}b_{i1}+z_{ij2}\Gamma_{k12}b_{i1}+z_{ij2}\Gamma_{k22}b_{i2}}+1}\right] \\
\frac{1}{L}\sum_{l,i,j} b_{i1}z_{ij1}\left[y_{ij} - m_{ij}\frac{e^{x_{ij1}\beta_{k1}+x_{ij2}\beta_{k2}+x_{ij3}\beta_{k3}+z_{ij1}\Gamma_{k11}b_{i1}+z_{ij2}\Gamma_{k12}b_{i1}+z_{ij2}\Gamma_{k22}b_{i2}}}{e^{x_{ij1}\beta_{k1}+x_{ij2}\beta_{k2}+x_{ij3}\beta_{k3}+z_{ij1}\Gamma_{k11}b_{i1}+z_{ij2}\Gamma_{k12}b_{i1}+z_{ij2}\Gamma_{k22}b_{i2}}+1}\right] \\
\frac{1}{L}\sum_{l,i,j} b_{i1}z_{ij2}\left[y_{ij} - m_{ij}\frac{e^{x_{ij1}\beta_{k1}+x_{ij2}\beta_{k2}+x_{ij3}\beta_{k3}+z_{ij1}\Gamma_{k11}b_{i1}+z_{ij2}\Gamma_{k12}b_{i1}+z_{ij2}\Gamma_{k22}b_{i2}}}{e^{x_{ij1}\beta_{k1}+x_{ij2}\beta_{k2}+x_{ij3}\beta_{k3}+z_{ij1}\Gamma_{k11}b_{i1}+z_{ij2}\Gamma_{k12}b_{i1}+z_{ij2}\Gamma_{k22}b_{i2}}+1}\right] \\
\frac{1}{L}\sum_{l,i,j} b_{i2}z_{ij2}\left[y_{ij} - m_{ij}\frac{e^{x_{ij1}\beta_{k1}+x_{ij2}\beta_{k2}+x_{ij3}\beta_{k3}+z_{ij1}\Gamma_{k11}b_{i1}+z_{ij2}\Gamma_{k12}b_{i1}+z_{ij2}\Gamma_{k22}b_{i2}}}{e^{x_{ij1}\beta_{k1}+x_{ij2}\beta_{k2}+x_{ij3}\beta_{k3}+z_{ij1}\Gamma_{k11}b_{i1}+z_{ij2}\Gamma_{k12}b_{i1}+z_{ij2}\Gamma_{k22}b_{i2}}+1}\right]
\end{bmatrix} \\
&= \frac{1}{L}\sum_{l,i,j}\begin{bmatrix} x_{ij1} \\ x_{ij2} \\ x_{ij3} \\ z_{ij1}b_{i1} \\ z_{ij2}b_{i1} \\ z_{ij2}b_{i2} \end{bmatrix} \\
&\times \left(y_{ij} - m_{ij}\frac{e^{x_{ij1}\beta_{k1}+x_{ij2}\beta_{k2}+x_{ij3}\beta_{k3}+z_{ij1}\Gamma_{k11}b_{i1}+z_{ij2}\Gamma_{k12}b_{i1}+z_{ij2}\Gamma_{k22}b_{i2}}}{e^{x_{ij1}\beta_{k1}+x_{ij2}\beta_{k2}+x_{ij3}\beta_{k3}+z_{ij1}\Gamma_{k11}b_{i1}+z_{ij2}\Gamma_{k12}b_{i1}+z_{ij2}\Gamma_{k22}b_{i2}}+1}\right)
\end{aligned}
$$

I also used the second derivative, so I derived it.

$$
\begin{aligned}
Q''_k(\boldsymbol{\theta}_k) &= \frac{\partial^2 Q_k(\boldsymbol{\theta}_k)}{\partial \boldsymbol{\theta}_k^\top \partial \boldsymbol{\theta}_k} = \begin{bmatrix} \frac{\partial^2 Q_k(\boldsymbol{\theta}_k)}{\partial \boldsymbol{\beta}_k^\top \partial \boldsymbol{\beta}_k} & \frac{\partial^2 Q_k(\boldsymbol{\theta}_k)}{\partial \boldsymbol{\beta}_k^\top \partial \Gamma_k^*} \\ \frac{\partial^2 Q_k(\boldsymbol{\theta}_k)}{\partial \Gamma_k^{*\top} \partial \boldsymbol{\beta}_k} & \frac{\partial^2 Q_k(\boldsymbol{\theta}_k)}{\partial \Gamma_k^{*\top} \partial \Gamma_k^*} \end{bmatrix} \\
&= \frac{1}{L} \sum_{l,i,j} -m_{ij} \frac{e^{x_{ij1}\beta_{k1}+x_{ij2}\beta_{k2}+x_{ij3}\beta_{k3}+z_{ij1}\Gamma_{k11}b_{i1}+z_{ij2}\Gamma_{k12}b_{i1}+z_{ij2}\Gamma_{k22}b_{i2}}}{(e^{x_{ij1}\beta_{k1}+x_{ij2}\beta_{k2}+x_{ij3}\beta_{k3}+z_{ij1}\Gamma_{k11}b_{i1}+z_{ij2}\Gamma_{k12}b_{i1}+z_{ij2}\Gamma_{k22}b_{i2}}+1)^2} \times \\
&\quad \begin{bmatrix}
(x_{ij1})^2 & x_{ij1}x_{ij2} & x_{ij1}x_{ij3} & x_{ij1}z_{ij1}b_{i1} & x_{ij1}z_{ij2}b_{i1} & x_{ij1}z_{ij2}b_{i2} \\
x_{ij1}x_{ij2} & (x_{ij2})^2 & x_{ij2}x_{ij3} & x_{ij2}z_{ij1}b_{i1} & x_{ij2}z_{ij2}b_{i1} & x_{ij2}z_{ij2}b_{i2} \\
x_{ij1}x_{ij3} & x_{ij2}x_{ij3} & (x_{ij3})^2 & x_{ij3}z_{ij1}b_{i1} & x_{ij3}z_{ij2}b_{i1} & x_{ij3}z_{ij2}b_{i2} \\
x_{ij1}z_{ij1}b_{i1} & x_{ij2}z_{ij1}b_{i1} & x_{ij3}z_{ij1}b_{i1} & (z_{ij1}b_{i1})^2 & z_{ij1}z_{ij2}(b_{i1})^2 & z_{ij1}z_{ij2}b_{i1}b_{i2} \\
x_{ij1}z_{ij2}b_{i1} & x_{ij2}z_{ij2}b_{i1} & x_{ij3}z_{ij2}b_{i1} & z_{ij1}z_{ij2}(b_{i1})^2 & (z_{ij2}b_{i1})^2 & (z_{ij2})^2 b_{i1}b_{i2} \\
x_{ij1}z_{ij2}b_{i2} & x_{ij2}z_{ij2}b_{i2} & x_{ij3}z_{ij2}b_{i2} & z_{ij1}z_{ij2}b_{i1}b_{i2} & (z_{ij2})^2 b_{i1}b_{i2} & (z_{ij2}b_{i2})^2
\end{bmatrix} \\
&= \frac{1}{L} \sum_{l,i,j} \begin{bmatrix} x_{ij1} \\ x_{ij2} \\ x_{ij3} \\ z_{ij1}b_{i1} \\ z_{ij2}b_{i1} \\ z_{ij2}b_{i2} \end{bmatrix} \begin{bmatrix} x_{ij1} \\ x_{ij2} \\ x_{ij3} \\ z_{ij1}b_{i1} \\ z_{ij2}b_{i1} \\ z_{ij2}b_{i2} \end{bmatrix}^\top \\
&\quad \times -m_{ij} \frac{e^{x_{ij1}\beta_{k1}+x_{ij2}\beta_{k2}+x_{ij3}\beta_{k3}+z_{ij1}\Gamma_{k11}b_{i1}+z_{ij2}\Gamma_{k12}b_{i1}+z_{ij2}\Gamma_{k22}b_{i2}}}{(e^{x_{ij1}\beta_{k1}+x_{ij2}\beta_{k2}+x_{ij3}\beta_{k3}+z_{ij1}\Gamma_{k11}b_{i1}+z_{ij2}\Gamma_{k12}b_{i1}+z_{ij2}\Gamma_{k22}b_{i2}}+1)^2}
\end{aligned}
$$

Also, consider the derivative with respect to $\mathbf{b}_i$ for a particular value of i, and k. Again I used $q = 3$.

$$
\begin{aligned}
\frac{\partial Q_k(\boldsymbol{\theta}_k)}{\partial \mathbf{b}_i} &= \sum_{j=1}^{n_i} \begin{bmatrix} z_{ij1}\Gamma_{k11} + z_{ij2}\Gamma_{k12} + z_{ij3}\Gamma_{k13} \\ z_{ij2}\Gamma_{k22} + z_{ij3}\Gamma_{k23} \\ z_{ij3}\Gamma_{k33} \end{bmatrix} \\
&\quad \times \left(y_{ij} - m_{ij} \frac{e^{x_{ij1}\beta_{k1}+x_{ij2}\beta_{k2}+x_{ij3}\beta_{k3}+z_{ij1}\Gamma_{k11}b_{i1}+z_{ij2}\Gamma_{k12}b_{i1}+z_{ij2}\Gamma_{k22}b_{i2}}}{e^{x_{ij1}\beta_{k1}+x_{ij2}\beta_{k2}+x_{ij3}\beta_{k3}+z_{ij1}\Gamma_{k11}b_{i1}+z_{ij2}\Gamma_{k12}b_{i1}+z_{ij2}\Gamma_{k22}b_{i2}}+1} \right)
\end{aligned}
$$

I then calculated the second derivative.

$$
\begin{aligned}
\frac{\partial^2 Q_k(\boldsymbol{\theta}_k)}{\partial \mathbf{b}_i^\top \partial \mathbf{b}_i} &= \sum_{j=1}^{n_i} \begin{bmatrix} z_{ij1}\Gamma_{k11} + z_{ij2}\Gamma_{k12} + z_{ij3}\Gamma_{k13} \\ z_{ij2}\Gamma_{k22} + z_{ij3}\Gamma_{k23} \\ z_{ij3}\Gamma_{k33} \end{bmatrix} \begin{bmatrix} z_{ij1}\Gamma_{k11} + z_{ij2}\Gamma_{k12} + z_{ij3}\Gamma_{k13} \\ z_{ij2}\Gamma_{k22} + z_{ij3}\Gamma_{k23} \\ z_{ij3}\Gamma_{k33} \end{bmatrix}^\top \\
& \quad \times - m_{ij} \frac{e^{x_{ij1}\beta_{k1}+x_{ij2}\beta_{k2}+x_{ij3}\beta_{k3}+z_{ij1}\Gamma_{k11}b_{i1}+z_{ij2}\Gamma_{k12}b_{i1}+z_{ij2}\Gamma_{k22}b_{i2}}}{(e^{x_{ij1}\beta_{k1}+x_{ij2}\beta_{k2}+x_{ij3}\beta_{k3}+z_{ij1}\Gamma_{k11}b_{i1}+z_{ij2}\Gamma_{k12}b_{i1}+z_{ij2}\Gamma_{k22}b_{i2}}+1)^2}
\end{aligned}
$$

B.3 Details of Maximisation of the Maximum Likelihood Estimate for a Single Subpopulation K, Poisson Case, $\boldsymbol{\gamma}_k$ Reparameterised

Following the reparameterisation of the matrix $\mathbb{\Gamma}_k$ described in Section 4.2, I again calculated the first, and second derivatives of the approximate likelihood. To illustrate, I show an example with $p = 3$, and $q = 3$ with differentiation performed for each element. In this example, $\mathbf{x}_{ij}\boldsymbol{\beta}_k + \mathbf{z}_{ij}\mathbb{\Gamma}_k\mathbf{b}_i = \mathbf{x}_{ij}\boldsymbol{\beta}_k + \mathbf{z}_{ij}\mathbb{d}_k\mathbb{C}_k\mathbf{b}_i = x_{ij1}\beta_{k1} + x_{ij2}\beta_{k2} + x_{ij3}\beta_{k3} + z_{ij1}d_{k1}b_{i1} + z_{ij2}d_{k2}(C_{k12}b_{i1} + b_{i2}) + z_{ij3}d_{k3}(C_{k13}b_{i1} + C_{k23}b_{i2} + b_{i3})$ and

$$
\begin{aligned}
Q_k(\underline{\boldsymbol{\theta}}_k) &= \frac{1}{L}\sum_{l=1}^{L}\sum_{i=1}^{n}\sum_{j=1}^{n_i} [y_{ij}x_{ij1}\beta_{k1} + y_{ij}x_{ij2}\beta_{k2} + y_{ij}x_{ij3}\beta_{k3} \\
&+ y_{ij}z_{ij1}d_{k1}b_{i1} + y_{ij}z_{ij2}d_{k2}(C_{k12}b_{i1} + b_{i2}) + y_{ij}z_{ij3}d_{k3}(C_{k13}b_{i1} + C_{k23}b_{i2} + b_{i3}) \\
&- e^{x_{ij1}\beta_{k1}+x_{ij2}\beta_{k2}+x_{ij3}\beta_{k3}+z_{ij1}d_{k1}b_{i1}+z_{ij2}d_{k2}(C_{k12}b_{i1}+b_{i2})+z_{ij3}d_{k3}(C_{k13}b_{i1}+C_{k23}b_{i2}+b_{i3})}\big] .
\end{aligned}
$$

As before, in an effort to simplify the notation slightly, I shortened $\mathbf{b}_i^{(s,l)}$ to $\mathbf{b}_i$, did not include weights τ_{ki}, condensed $\sum_{l=1}^{L}\sum_{i=1}^{n}\sum_{j=1}^{n_i}$ to $\sum_{l,i,j}$, and used $e^{x_{ij1}\beta_{k1}+\cdots}$ in place of $e^{x_{ij1}\beta_{k1}+x_{ij2}\beta_{k2}+x_{ij3}\beta_{k3}+z_{ij1}d_{k1}b_{i1}+z_{ij2}d_{k2}(C_{k12}b_{i1}+b_{i2})+z_{ij3}d_{k3}(C_{k13}b_{i1}+C_{k23}b_{i2}+b_{i3})}$.

Again, I calculated the first derivative.

$$
\begin{aligned}
Q'_k(\underline{\boldsymbol{\theta}}_k) &= \frac{\partial Q_k(\underline{\boldsymbol{\theta}}_k)}{\partial \underline{\boldsymbol{\theta}}_k} = \begin{bmatrix} \frac{\partial Q_k(\underline{\boldsymbol{\theta}}_k)}{\partial \boldsymbol{\beta}_k} \\ \frac{\partial Q_k(\underline{\boldsymbol{\theta}}_k)}{\partial \mathbb{d}^*} \\ \frac{\partial Q_k(\underline{\boldsymbol{\theta}}_k)}{\partial \mathbb{C}_k^*} \end{bmatrix} \\
&= \begin{bmatrix} \frac{1}{L}\sum_{l,i,j} x_{ij1}[y_{ij} - e^{x_{ij1}\beta_{k1}+\cdots}] \\ \frac{1}{L}\sum_{l,i,j} x_{ij2}[y_{ij} - e^{x_{ij1}\beta_{k1}+\cdots}] \\ \frac{1}{L}\sum_{l,i,j} x_{ij3}[y_{ij} - e^{x_{ij1}\beta_{k1}+\cdots}] \\ \frac{1}{L}\sum_{l,i,j}(z_{ij1}b_{i1})[y_{ij} - e^{x_{ij1}\beta_{k1}+\cdots}] \\ \frac{1}{L}\sum_{l,i,j} z_{ij2}(C_{k12}b_{i1} + b_{i2})[y_{ij} - e^{x_{ij1}\beta_{k1}+\cdots}] \\ \frac{1}{L}\sum_{l,i,j} z_{ij3}(C_{k13}b_{i1} + C_{k23}b_{i2} + b_{i3})[y_{ij} - e^{x_{ij1}\beta_{k1}+\cdots}] \\ \frac{1}{L}\sum_{l,i,j}(z_{ij2}d_{k2}b_{i1})[y_{ij} - e^{x_{ij1}\beta_{k1}+\cdots}] \\ \frac{1}{L}\sum_{l,i,j}(z_{ij3}d_{k3}b_{i1})[y_{ij} - e^{x_{ij1}\beta_{k1}+\cdots}] \\ \frac{1}{L}\sum_{l,i,j}(z_{ij3}d_{k3}b_{i2})[y_{ij} - e^{x_{ij1}\beta_{k1}+\cdots}] \end{bmatrix}
\end{aligned}
$$

$$= \frac{1}{L}\sum_{l,i,j}\begin{bmatrix} x_{ij1} \\ x_{ij2} \\ x_{ij3} \\ z_{ij1}b_{i1} \\ z_{ij2}(C_{k12}b_{i1}+b_{i2}) \\ z_{ij3}(C_{k13}b_{i1}+C_{k23}b_{i2}+b_{i3}) \\ z_{ij2}(d_{k2}b_{i1}) \\ z_{ij3}(d_{k3}b_{i1}) \\ z_{ij3}(d_{k3}b_{i2}) \end{bmatrix}$$

$$\times \quad \{y_{ij} - e^{x_{ij1}\beta_{k1}+x_{ij2}\beta_{k2}+x_{ij3}\beta_{k3}+z_{ij1}d_{k1}b_{i1}+z_{ij2}d_{k2}(C_{k12}b_{i1}+b_{i2})+z_{ij3}d_{k3}(C_{k13}b_{i1}+C_{k23}b_{i2}+b_{i3})}\}$$

The second derivative is more complex in this case.

$$Q''_k(\underline{\boldsymbol{\theta}}_k) = \frac{\partial^2 Q_k(\underline{\boldsymbol{\theta}}_k)}{\partial \underline{\boldsymbol{\theta}}_k^\top \partial \underline{\boldsymbol{\theta}}_k} = \begin{bmatrix} \frac{\partial Q_k(\underline{\boldsymbol{\theta}}_k)}{\partial \boldsymbol{\beta}_k^\top \partial \boldsymbol{\beta}_k} & \frac{\partial Q_k(\underline{\boldsymbol{\theta}}_k)}{\partial \boldsymbol{\beta}_k^\top \partial \mathbb{d}^*} & \frac{\partial Q_k(\underline{\boldsymbol{\theta}}_k)}{\partial \boldsymbol{\beta}_k^\top \partial \mathbb{C}_k^*} \\ \frac{\partial Q_k(\underline{\boldsymbol{\theta}}_k)}{\partial \mathbb{d}^{*\top} \partial \boldsymbol{\beta}_k} & \frac{\partial Q_k(\underline{\boldsymbol{\theta}}_k)}{\partial \mathbb{d}^{*\top} \partial \mathbb{d}^*} & \frac{\partial Q_k(\underline{\boldsymbol{\theta}}_k)}{\partial \mathbb{d}^{*\top} \partial \mathbb{C}_k^*} \\ \frac{\partial Q_k(\underline{\boldsymbol{\theta}}_k)}{\partial \mathbb{C}_k^{*\top} \partial \boldsymbol{\beta}_k} & \frac{\partial Q_k(\underline{\boldsymbol{\theta}}_k)}{\partial \mathbb{C}_k^{*\top} \partial \mathbb{d}^*} & \frac{\partial Q_k(\underline{\boldsymbol{\theta}}_k)}{\partial \mathbb{C}_k^{*\top} \partial \mathbb{C}_k^*} \end{bmatrix}$$

$$
\begin{aligned}
= \quad & \frac{1}{L}\sum_{l,i,j} \begin{bmatrix} x_{ij1} \\ x_{ij2} \\ x_{ij3} \\ z_{ij1}b_{i1} \\ z_{ij2}(C_{k12}b_{i1}+b_{i2}) \\ z_{ij3}(C_{k13}b_{i1}+C_{k23}b_{i2}+b_{i3}) \\ z_{ij2}(d_{k2}b_{i1}) \\ z_{ij3}(d_{k3}b_{i1}) \\ z_{ij3}(d_{k3}b_{i2}) \end{bmatrix} \begin{bmatrix} x_{ij1} \\ x_{ij2} \\ x_{ij3} \\ z_{ij1}b_{i1} \\ z_{ij2}(C_{k12}b_{i1}+b_{i2}) \\ z_{ij3}(C_{k13}b_{i1}+C_{k23}b_{i2}+b_{i3}) \\ z_{ij2}(d_{k2}b_{i1}) \\ z_{ij3}(d_{k3}b_{i1}) \\ z_{ij3}(d_{k3}b_{i2}) \end{bmatrix}^{\top} \\
\times \quad & -e^{x_{ij1}\beta_{k1}+x_{ij2}\beta_{k2}+x_{ij3}\beta_{k3}+z_{ij1}d_{k1}b_{i1}+z_{ij2}d_{k2}(C_{k12}b_{i1}+b_{i2})+z_{ij3}d_{k3}(C_{k13}b_{i1}+C_{k23}b_{i2}+b_{i3})} \\
+ \quad & \begin{bmatrix} 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & z_{ij2}b_{i1} & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & z_{ij3}b_{i1} & z_{ij3}b_{i2} \\ 0 & 0 & 0 & 0 & z_{ij2}b_{i1} & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & z_{ij3}b_{i1} & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & z_{ij3}b_{i2} & 0 & 0 & 0 \end{bmatrix} \\
\times \quad & \{y_{ij} - e^{x_{ij1}\beta_{k1}+x_{ij2}\beta_{k2}+x_{ij3}\beta_{k3}+z_{ij1}d_{k1}b_{i1}+z_{ij2}d_{k2}(C_{k12}b_{i1}+b_{i2})+z_{ij3}d_{k3}(C_{k13}b_{i1}+C_{k23}b_{i2}+b_{i3})}\}
\end{aligned}
$$

B.4 Details of Maximisation of the Maximum Likelihood Estimate for a Single Subpopulation K, binomial Case, γ_k Reparameterised

Similar to the previous section, I provided a binomial example next with $p = 3$, $q = 3$, and $\mathbf{x}_{ij}\boldsymbol{\beta}_k + \mathbf{z}_{ij}\mathbb{\Gamma}_k\mathbf{b}_i = \mathbf{x}_{ij}\boldsymbol{\beta}_k + \mathbf{z}_{ij}\mathbb{d}_k\mathbb{C}_k\mathbf{b}_i = x_{ij1}\beta_{k1} + x_{ij2}\beta_{k2} + x_{ij3}\beta_{k3} + z_{ij1}d_{k1}b_{i1} + z_{ij2}d_{k2}(C_{k12}b_{i1} + b_{i2}) + z_{ij3}d_{k3}(C_{k13}b_{i1} + C_{k23}b_{i2} + b_{i3})$ to illustrate the differentiation. Thus,

$$\begin{aligned} Q_k(\underline{\boldsymbol{\theta}}_k) \quad = \quad & \frac{1}{L}\sum_{l=1}^{L}\sum_{i=1}^{n}\sum_{j=1}^{n_i}[y_{ij}x_{ij1}\beta_{k1} + y_{ij}x_{ij2}\beta_{k2} + y_{ij}x_{ij3}\beta_{k3} + y_{ij}z_{ij1}d_{k1}b_{i1} \\ & + y_{ij}z_{ij2}d_{k2}(C_{k12}b_{i1} + b_{i2}) + y_{ij}z_{ij3}d_{k3}(C_{k13}b_{i1} + C_{k23}b_{i2} + b_{i3}) \\ & - m_{ij}\log(\exp[x_{ij1}\beta_{k1} + x_{ij2}\beta_{k2} + x_{ij3}\beta_{k3} + z_{ij1}d_{k1}b_{i1} \\ & + z_{ij2}d_{k2}(C_{k12}b_{i1} + b_{i2}) + z_{ij3}d_{k3}(C_{k13}b_{i1} + C_{k23}b_{i2} + b_{i3})] + 1)]. \end{aligned}$$

As in the previous section, to simplify the notation slightly, I shortened $\mathbf{b}_i^{(s,l)}$ to $\mathbf{b}_i$, did not include the weights τ_{ki}, condensed $\sum_{l=1}^{L}\sum_{i=1}^{n}\sum_{j=1}^{n_i}$ to $\sum_{l,i,j}$, and used $e^{x_{ij1}\beta_{k1}+\ldots}$ rather than $e^{x_{ij1}\beta_{k1}+x_{ij2}\beta_{k2}+x_{ij3}\beta_{k3}+z_{ij1}d_{k1}b_{i1}+z_{ij2}d_{k2}(C_{k12}b_{i1}+b_{i2})+z_{ij3}d_{k3}(C_{k13}b_{i1}+C_{k23}b_{i2}+b_{i3})}$.

Consider the first derivative.

$$Q'_k(\underline{\boldsymbol{\theta}}_k) = \frac{\partial Q_k(\underline{\boldsymbol{\theta}}_k)}{\partial \underline{\boldsymbol{\theta}}_k} = \begin{bmatrix} \frac{\partial Q_k(\underline{\boldsymbol{\theta}}_k)}{\partial \boldsymbol{\beta}_k} \\ \frac{\partial Q_k(\underline{\boldsymbol{\theta}}_k)}{\partial \mathbb{d}^*_k} \\ \frac{\partial Q_k(\underline{\boldsymbol{\theta}}_k)}{\partial \mathbb{C}^*_k} \end{bmatrix}$$

$$= \begin{bmatrix} \frac{1}{L}\sum_{l,i,j} x_{ij1}[y_{ij} - mij\frac{e^{x_{ij1}\beta_{k1}+\cdots}}{e^{x_{ij1}\beta_{k1}+\cdots}+1}] \\ \frac{1}{L}\sum_{l,i,j} x_{ij2}[y_{ij} - mij\frac{e^{x_{ij1}\beta_{k1}+\cdots}}{e^{x_{ij1}\beta_{k1}+\cdots}+1}] \\ \frac{1}{L}\sum_{l,i,j} x_{ij3}[y_{ij} - mij\frac{e^{x_{ij1}\beta_{k1}+\cdots}}{e^{x_{ij1}\beta_{k1}+\cdots}+1}] \\ \frac{1}{L}\sum_{l,i,j} z_{ij1}b_{i1}[y_{ij} - mij\frac{e^{x_{ij1}\beta_{k1}+\cdots}}{e^{x_{ij1}\beta_{k1}+\cdots}+1}] \\ \frac{1}{L}\sum_{l,i,j} z_{ij2}(C_{k12}b_{i1}+b_{i2})[y_{ij} - mij\frac{e^{x_{ij1}\beta_{k1}+\cdots}}{e^{x_{ij1}\beta_{k1}+\cdots}+1}] \\ \frac{1}{L}\sum_{l,i,j} z_{ij3}(C_{k13}b_{i1}+C_{k23}b_{i2}+b_{i3})[y_{ij} - mij\frac{e^{x_{ij1}\beta_{k1}+\cdots}}{e^{x_{ij1}\beta_{k1}+\cdots}+1}] \\ \frac{1}{L}\sum_{l,i,j} z_{ij2}(d_{k2}b_{i1})[y_{ij} - mij\frac{e^{x_{ij1}\beta_{k1}+\cdots}}{e^{x_{ij1}\beta_{k1}+\cdots}+1}] \\ \frac{1}{L}\sum_{l,i,j} z_{ij3}(d_{k3}b_{i1})[y_{ij} - mij\frac{e^{x_{ij1}\beta_{k1}+\cdots}}{e^{x_{ij1}\beta_{k1}+\cdots}+1}] \\ \frac{1}{L}\sum_{l,i,j} z_{ij3}(d_{k3}b_{i2})[y_{ij} - mij\frac{e^{x_{ij1}\beta_{k1}+\cdots}}{e^{x_{ij1}\beta_{k1}+\cdots}+1}] \end{bmatrix}$$

$$= \frac{1}{L}\sum_{l,i,j} \begin{bmatrix} x_{ij1} \\ x_{ij2} \\ x_{ij3} \\ z_{ij1}b_{i1} \\ z_{ij2}(C_{k12}b_{i1}+b_{i2}) \\ z_{ij3}(C_{k13}b_{i1}+C_{k23}b_{i2}+b_{i3}) \\ z_{ij2}(d_{k2}b_{i1}) \\ z_{ij3}(d_{k3}b_{i1}) \\ z_{ij3}(d_{k3}b_{i2}) \end{bmatrix}$$
$$\times \left\{ y_{ij} - m_{ij}\frac{e^{x_{ij1}\beta_{k1}+x_{ij2}\beta_{k2}+x_{ij3}\beta_{k3}+z_{ij1}d_{k1}b_{i1}+z_{ij2}d_{k2}(C_{k12}b_{i1}+b_{i2})+z_{ij3}d_{k3}(C_{k13}b_{i1}+C_{k23}b_{i2}+b_{i3})}}{e^{x_{ij1}\beta_{k1}+x_{ij2}\beta_{k2}+x_{ij3}\beta_{k3}+z_{ij1}d_{k1}b_{i1}+z_{ij2}d_{k2}(C_{k12}b_{i1}+b_{i2})+z_{ij3}d_{k3}(C_{k13}b_{i1}+C_{k23}b_{i2}+b_{i3})}+1} \right\}$$

Again, the second derivative is more complex with the reparameterised notation.

$$Q''_k(\underline{\boldsymbol{\theta}}_k) = \frac{\partial^2 Q_k(\underline{\boldsymbol{\theta}}_k)}{\partial \underline{\boldsymbol{\theta}}_k^\top \partial \underline{\boldsymbol{\theta}}_k} = \begin{bmatrix} \frac{\partial Q_k(\underline{\boldsymbol{\theta}}_k)}{\partial \boldsymbol{\beta}_k^\top \partial \boldsymbol{\beta}_k} & \frac{\partial Q_k(\underline{\boldsymbol{\theta}}_k)}{\partial \boldsymbol{\beta}_k^\top \partial \mathbb{d}^*} & \frac{\partial Q_k(\underline{\boldsymbol{\theta}}_k)}{\partial \boldsymbol{\beta}_k^\top \partial \mathbb{C}^*_k} \\ \frac{\partial Q_k(\underline{\boldsymbol{\theta}}_k)}{\partial \mathbb{d}^{*\top} \partial \boldsymbol{\beta}_k} & \frac{\partial Q_k(\underline{\boldsymbol{\theta}}_k)}{\partial \mathbb{d}^{*\top} \partial \mathbb{d}^*} & \frac{\partial Q_k(\underline{\boldsymbol{\theta}}_k)}{\partial \mathbb{d}^{*\top} \partial \mathbb{C}^*_k} \\ \frac{\partial Q_k(\underline{\boldsymbol{\theta}}_k)}{\partial \mathbb{C}^{*\top}_k \partial \boldsymbol{\beta}_k} & \frac{\partial Q_k(\underline{\boldsymbol{\theta}}_k)}{\partial \mathbb{C}^{*\top}_k \partial \mathbb{d}^*} & \frac{\partial Q_k(\underline{\boldsymbol{\theta}}_k)}{\partial \mathbb{C}^{*\top}_k \partial \mathbb{C}^*_k} \end{bmatrix}$$

$$
\begin{aligned}
= \quad & \frac{1}{L}\sum_{l,i,j} \begin{bmatrix} x_{ij1} \\ x_{ij2} \\ x_{ij3} \\ z_{ij1}b_{i1} \\ z_{ij2}(C_{k12}b_{i1}+b_{i2}) \\ z_{ij3}(C_{k13}b_{i1}+C_{k23}b_{i2}+b_{i3}) \\ z_{ij2}(d_{k2}b_{i1}) \\ z_{ij3}(d_{k3}b_{i1}) \\ z_{ij3}(d_{k3}b_{i2}) \end{bmatrix} \begin{bmatrix} x_{ij1} \\ x_{ij2} \\ x_{ij3} \\ z_{ij1}b_{i1} \\ z_{ij2}(C_{k12}b_{i1}+b_{i2}) \\ z_{ij3}(C_{k13}b_{i1}+C_{k23}b_{i2}+b_{i3}) \\ z_{ij2}(d_{k2}b_{i1}) \\ z_{ij3}(d_{k3}b_{i1}) \\ z_{ij3}(d_{k3}b_{i2}) \end{bmatrix}^{\top} \\
\times \quad & -m_{ij}\frac{e^{x_{ij1}\beta_{k1}+x_{ij2}\beta_{k2}+x_{ij3}\beta_{k3}+z_{ij1}d_{k1}b_{i1}+z_{ij2}d_{k2}(C_{k12}b_{i1}+b_{i2})+z_{ij3}d_{k3}(C_{k13}b_{i1}+C_{k23}b_{i2}+b_{i3})}}{\{e^{x_{ij1}\beta_{k1}+x_{ij2}\beta_{k2}+x_{ij3}\beta_{k3}+z_{ij1}d_{k1}b_{i1}+z_{ij2}d_{k2}(C_{k12}b_{i1}+b_{i2})+z_{ij3}d_{k3}(C_{k13}b_{i1}+C_{k23}b_{i2}+b_{i3})}+1\}^2} \\
+ \quad & \begin{bmatrix} 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & z_{ij2}b_{i1} & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & z_{ij3}b_{i1} & z_{ij3}b_{i2} \\ 0 & 0 & 0 & 0 & z_{ij2}b_{i1} & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & z_{ij3}b_{i1} & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & z_{ij3}b_{i2} & 0 & 0 & 0 \end{bmatrix} \\
\times \quad & \left\{ y_{ij} - m_{ij}\frac{e^{x_{ij1}\beta_{k1}+x_{ij2}\beta_{k2}+x_{ij3}\beta_{k3}+z_{ij1}d_{k1}b_{i1}+z_{ij2}d_{k2}(C_{k12}b_{i1}+b_{i2})+z_{ij3}d_{k3}(C_{k13}b_{i1}+C_{k23}b_{i2}+b_{i3})}}{e^{x_{ij1}\beta_{k1}+x_{ij2}\beta_{k2}+x_{ij3}\beta_{k3}+z_{ij1}d_{k1}b_{i1}+z_{ij2}d_{k2}(C_{k12}b_{i1}+b_{i2})+z_{ij3}d_{k3}(C_{k13}b_{i1}+C_{k23}b_{i2}+b_{i3})}+1} \right\}
\end{aligned}
$$

APPENDIX C
Detailed Algorithms

The complexity of a FinMix GLMM necessitates a multi-step algorithm to perform parameter estimation. I described these calculations in Chapter 3 and Chapter 4 for the MLE and MPLE cases respectively but I reproduced them as formal algorithms here. Section C.1 details the steps to calculate the MLE of a FinMix GLMM, and Sections C.2 and C.3 show the process when a penalty is added to the likelihood. The generation of $\mathbf{b}_{ki}|\mathbf{y}_i, \mathbb{X}_i, \mathbb{Z}_i, \boldsymbol{\beta}_k, \mathbb{\Gamma}_k$ is an important part of the E-step in the inner MCEM which I explored in more detail in Section C.4. The choice of the tuning parameter when using penalised regression is a complex, and important question. I described the algorithm used to choose $\boldsymbol{\lambda}$ in this thesis in C.5. Due to the nested loops, I added additional subscripts have to the loop counters (s_{in}, s_{out}), the tolerance (t_{in}, t_{out}), and the maximum number of iterations (M_{in}, M_{out}).

C.1 Algorithm for Estimation of Parameters Using Maximum Likelihood Estimation

Data: $\mathbf{y}$, $\mathbb{X}$, $\mathbb{Z}$, $\mathbf{\Theta}^{(0)}$, K, p, q, L, n, ι, t_{in}, t_{out},M_{in}, M_{out}

Result: $\widehat{\mathbf{\Theta}}$

Set $s_{out} := 0$;

repeat

- Increment $s_{out} := s_{out} + 1$;
- Decompose $\mathbf{\Theta}^{(s_{out})}$ into $\pi_k^{(s_{out})}$, and $\boldsymbol{\theta}_k^{(s_{out})} \forall k$;
- Outer E-Step: Calculate $\tau_{ki} \forall k \forall i$ using $\mathbf{b}_{ki} \sim Gaussian(\mu := 0, \mathbf{\Sigma} := \boldsymbol{I}_q)$;
- Outer M-Step: Calculate $\boldsymbol{\pi}_k^{(s_{out}+1)} \forall k$;
- **for** $k \in \{1, 2, \ldots, K\}$ **do**
  - Reset $s_{in} := 0$;
  - **repeat**
    - Increment $s_{in} := s_{in} + 1$;
    - Inner E-Step: Generate L values of $\mathbf{b}_{ki}^{(s_{in},l)}$;
    - Inner M-Step, recall that Q''_k, and Q'_k are weighted by τ_{ki}:
      $\boldsymbol{\theta}_k^{(s_{in}+1)} := \boldsymbol{\theta}_k^{(s_{in})} - (\frac{1}{2})^{\iota} (Q''_k(\boldsymbol{\theta}_k^{(s_{in})}))^{-1} Q'_k(\boldsymbol{\theta}_k^{(s_{in})})$;
  - **until** $\sqrt{(\boldsymbol{\theta}_k^{(s_{in}+1)} - \boldsymbol{\theta}_k^{(s_{in})})^2} < t^{in}$ *or* $s_{in} > M_{in}$;
- **end**
- Reconstruct $\mathbf{\Theta}^{(s_{out}+1)}$ from $\pi_k^{(s_{out}+1)}$, and $\boldsymbol{\theta}_k^{(s_{out}+1)}$;

until $\sqrt{(\mathbf{\Theta}^{(s_{out}+1)} - \mathbf{\Theta}^{(s_{out})})^2} < t^{out}$ *or* $s_{out} > M_{out}$;

if $s_{out} \leq M_{out}$ **then**

- Increment $s_{out} := s_{out} + 1$;
- Calculate $\ell_n^{(s_{out})} := \ell_n(\mathbf{\Theta}^{(s_{out})})$;
- Calculate BIC:= $(-2) \times \ell_n^{(s_{out})} + df \times \log(n)$;
- Return $\widehat{\mathbf{\Theta}} := \mathbf{\Theta}^{(s_{out})}$, $\ell_n^{(s_{out})}$, and BIC;

end

Algorithm 1: Estimate FinMix GLMM parameters in MLE case.

C.2 Algorithm for Estimation of Parameters Using Maximum Penalised Likelihood Estimation, Penalise Only Fixed Effects

Data: $\mathbf{y}$, $\mathbb{X}$, $\mathbb{Z}$, $\mathbf{\Theta}^{(0)}$, K, p, q, L, n, ι, $\boldsymbol{\lambda}$, α, t_{in}, t_{out},M_{in}, M_{out}

Result: $\widehat{\mathbf{\Theta}}$

Set $s_{out} := 0$, and $\boldsymbol{\lambda} := \boldsymbol{\lambda}^{\alpha}$;

repeat

  Increment $s_{out} := s_{out} + 1$;

  Decompose $\mathbf{\Theta}^{(s_{out})}$ to find $\pi_k^{(s_{out})}$, and $\boldsymbol{\theta}_k^{(s_{out})} \forall k$;

  Outer E-Step: Calculate $\tau_{ki} \forall k \forall i$ using $\mathbf{b}_{ki} \sim Gaussian(\mu := 0, \mathbf{\Sigma} := \boldsymbol{I}_q)$;

  Outer M-Step: Calculate $\boldsymbol{\pi}_k^{(s_{out}+1)} \forall k$;

  for $k \in \{1, 2, \ldots, K\}$ **do**

    Reset $s_{in} := 0$;

    Calculate $n_k^{(s_{in})} = n\pi_k^{(s_{in})}$;

    repeat

      Inner E-Step: Generate L values of $\mathbf{b}_{ki}^{(s_{in},l)}$;

      Calculate $\mathbb{\Sigma}_{\lambda_k}(\boldsymbol{\theta}^{(s_{in})})$, and $U_{\lambda_k}(\boldsymbol{\theta}_k^{(s_{in})}) := \mathbb{\Sigma}_{\lambda_k}(\boldsymbol{\theta}_k^{(s_{in})}) \times \boldsymbol{\theta}_k^{(s_{in})}$;

      Inner M-Step, recall that Q_k'', and Q_k' are weighted by τ_{ki}: $\boldsymbol{\theta}_k^{(s_{in}+1)} :=$

      $\boldsymbol{\theta}_k^{(s_{in})} - (\frac{1}{2})^{\iota}(Q_k''(\boldsymbol{\theta}_k^{(s_{in})}) - n_k^{(s_{in})}\mathbb{\Sigma}_{\lambda_k}(\boldsymbol{\theta}^{(s_{in})}))^{-1}(Q_k'(\boldsymbol{\theta}_k^{(s_{in})}) - n_k^{(s_{in})}U_{\lambda_k}(\boldsymbol{\theta}_k^{(s_{in})}))$;

    until $\sqrt{(\boldsymbol{\theta}_k^{(s_{in}+1)} - \boldsymbol{\theta}_k^{(s_{in})})^2} < t^{in}$ *or* $s_{in} > M_{in}$;

  end

until $\sqrt{(\mathbf{\Theta}^{(s_{out}+1)} - \mathbf{\Theta}^{(s_{out})})^2} < t^{out}$ *or* $s_{out} > M_{out}$;

if $s_{out} \leq M_{out}$ **then**

  Increment $s_{out} := s_{out} + 1$;

  Calculate $\ell_n^{(s_{out})} := \ell_n(\mathbf{\Theta}^{(s_{out})})$;

  Calculate BIC:$= (-2) \times \ell_n^{(s_{out})} + df \times \log(n)$;

  Return $\widehat{\mathbf{\Theta}} := \mathbf{\Theta}^{(s_{out})}$, $\ell_n^{(s_{out})}$, and BIC;

end

Algorithm 2: Estimate FinMix GLMM parameters in MPLE case when only the fixed effects are penalised.

C.3 Algorithm for Estimation of Parameters Using Maximum Penalised Likelihood Estimation, Penalise Both Fixed and Random Effects

Data: $\mathbf{y}$, $\mathbb{X}$, $\mathbb{Z}$, $\mathbf{\Theta}^{(0)}$, K, p, q, L, n, ι, $\boldsymbol{\lambda}$, α, t_{in}, t_{out},M_{in}, M_{out}

Result: $\widehat{\mathbf{\Theta}}$

Set $s_{out} := 0$, and $\boldsymbol{\lambda} := \boldsymbol{\lambda}^{\alpha}$;

repeat

  Increment $s_{out} := s_{out} + 1$;

  Decompose $\mathbf{\Theta}^{(s_{out})}$ to find $\pi_k^{(s_{out})}$, and $\boldsymbol{\theta}_k^{(s_{out})} \forall k$;

  Outer E-Step: Calculate $\tau_{ki} \forall k \forall i$ using $\mathbf{b}_{ki} \sim Gaussian(\mu := 0, \mathbf{\Sigma} := \boldsymbol{I}_q)$;

  Outer M-Step: Calculate $\boldsymbol{\pi}_k^{(s_{out}+1)} \forall k$;

  for $k \in \{1, 2, \ldots, K\}$ **do**

    Reset $s_{in} := 0$;

    Calculate $n_k^{(s_{in})} = n\pi_k^{(s_{in})}$;

    repeat

      Decompose $\mathbb{\Gamma}_k$ into $\mathbb{d}_k \mathbb{C}_k$, form $\underline{\boldsymbol{\theta}}_k$;

      Inner E-Step: Generate L values of $\mathbf{b}_{ki}^{(s_{in},l)}$;

      Calculate $\mathbb{\Sigma}_{\lambda_k}(\underline{\boldsymbol{\theta}}^{(s_{in})})$, and $U_{\lambda_k}(\underline{\boldsymbol{\theta}}_k^{(s_{in})}) := \mathbb{\Sigma}_{\lambda_k}(\underline{\boldsymbol{\theta}}_k^{(s_{in})}) \times \underline{\boldsymbol{\theta}}_k^{(s_{in})}$;

      Inner M-Step, recall that Q_k'', and Q_k' are weighted by τ_{ki}: $\underline{\boldsymbol{\theta}}_k^{(s_{in}+1)} :=$
       $\underline{\boldsymbol{\theta}}_k^{(s_{in})} - (\frac{1}{2})^{\iota}(Q_k''(\underline{\boldsymbol{\theta}}_k^{(s_{in})}) - n_k^{(s_{in})}\mathbb{\Sigma}_{\lambda_k}(\underline{\boldsymbol{\theta}}^{(s_{in})}))^{-1}(Q_k'(\underline{\boldsymbol{\theta}}_k^{(s_{in})}) - n_k^{(s_{in})}U_{\lambda_k}(\underline{\boldsymbol{\theta}}_k^{(s_{in})})$;

    until $\sqrt{(\boldsymbol{\theta}_k^{(s_{in}+1)} - \boldsymbol{\theta}_k^{(s_{in})})^2} < t^{in}$ *or* $s_{in} > M_{in}$;

  end

until $\sqrt{(\mathbf{\Theta}^{(s_{out}+1)} - \mathbf{\Theta}^{(s_{out})})^2} < t^{out}$ *or* $s_{out} > M_{out}$;

if $s_{out} \leq M_{out}$ **then**

  Increment $s_{out} := s_{out} + 1$;

  Calculate $\ell_n^{(s_{out})} := \ell_n(\mathbf{\Theta}^{(s_{out})})$;

  Calculate BIC:= $(-2) \times \ell_n^{(s_{out})} + df \times \log(n)$;

  Return $\widehat{\mathbf{\Theta}} := \mathbf{\Theta}^{(s_{out})}$, $\ell_n^{(s_{out})}$, and BIC;

end

Algorithm 3: Estimate FinMix GLMM parameters in MPLE case when both fixed, and random effects are penalised.

C.4 Generate $\mathbf{b}_{ki}^{(s_{in},l)}$

I used this algorithm in the E-step of the inner MCEM. Recall that $\mathbf{b}_{ki}|\mathbf{y}_i, \mathbb{X}_i, \mathbb{Z}_i, \boldsymbol{\beta}_k, \mathbb{\Gamma}_k$ does not follow a standard distribution. Therefore, I used rejection sampling to generate these values. However, there are many possible options when performing rejection sampling, so I provided here some additional details.

Data: $\mathbf{y}_i$, $\mathbb{X}_i$, $\mathbb{Z}_i$, $\boldsymbol{\theta}_k$, p, q, L,A^{met}

Result: Sample from $\mathbf{b}_{ki}^{(l)}|\mathbf{y}_i, \mathbb{X}_i, \mathbb{Z}_i, \boldsymbol{\beta}_k, \mathbb{\Gamma}_k$ of size L

Calculate $\mathbf{b}_{ki}^{(MLE)}|\mathbf{y}_i, \mathbb{X}_i, \mathbb{Z}_i, \boldsymbol{\beta}_k, \mathbb{\Gamma}_k$, and the Hessian matrix $\boldsymbol{H}_{\mathbf{b}_{ki}}$ at $\mathbf{b}_{ki}^{(MLE)}$;

Generate L potential $\mathbf{b}_{ki}^{(l)}$ from $Gaussian(\mu := \mathbf{b}_{ki}^{(MLE)}, \boldsymbol{\Sigma} := (-\boldsymbol{H}_{\mathbf{b}_{ki}})^{-1})$, these are the starting values of the sample;

for $a^{counter} \in \{1, 2, \ldots, A^{met}\}$ **do**

  Generate L potential values $V^{(l)} \sim Gaussian(\mu := \mathbf{b}_{ki}^{(MLE)}, \boldsymbol{\Sigma} := (-\boldsymbol{H}_{\mathbf{b}_{ki}})^{-1})$;

  Calculate $f_{\mathbf{b}_{ki}|\mathbf{y}_i,\mathbb{X}_i,\mathbb{Z}_i,\boldsymbol{\beta}_k\mathbb{\Gamma}_k}(V^{(l)})$, the likelihood of each $V^{(l)}$;

  Generate L values of $U^l \sim Uniform(a := 0, b := 1)$;

  If $f_U(U^l) < f_{\mathbf{b}_{ki}|\mathbf{y}_i,\mathbb{X}_i,\mathbb{Z}_i,\boldsymbol{\beta}_k\mathbb{\Gamma}_k}(V^{(l)})$ then $\mathbf{b}_{ki}^{(l)} := V^{(l)}$;

end

Return $\mathbf{b}_{ki}^{(l)}$;

Algorithm 4: Generate $\mathbf{b}_{ki}^{(s_{in},l)}|\mathbf{y}_i, \mathbb{X}_i, \mathbb{Z}_i, \boldsymbol{\beta}_k, \mathbb{\Gamma}_k$.

C.5 How to Choose $\boldsymbol{\lambda}$

The choice of $\boldsymbol{\lambda}$ is an important problem in variable selection, and one that is a source of continued research. There is an infinite number of possible choices for each λ_k, any non-negative number is a possibility. However, all values of λ_k greater than a certain value will provide the same estimates, where all the penalised parameters are set to zero, and the non-penalised parameters are the only non-zero estimates, that is, the mixing proportions $\boldsymbol{\pi}$, and the intercepts $\beta_{k1} \forall k \in \{1, 2, \ldots, K\}$. In the case where only the fixed effects are penalised, there will also be non-zero estimates for the random effects $\mathbb{\Gamma}$. Once I found these upper limits on each λ_k, I constructed a grid of possible values for $\boldsymbol{\lambda}$, and calculated the MPLE for each of them. This provides a number of possible values of $\boldsymbol{\lambda}$, from which I chose the one that provided the lowest BIC as the optimal value of $\boldsymbol{\lambda}$.

Data: $\mathbf{y}$, $\mathbb{X}$, $\mathbb{Z}$, $\boldsymbol{\Theta}^{(0)}$, K, p, q, L, n, ι, α, t_{in}, t_{out},M_{in}, M_{out}, δ

Result: Optimal $\boldsymbol{\lambda}$ with corresponding values of $\widehat{\boldsymbol{\Theta}}$, and BIC

Calculate MPLE, $\widehat{\boldsymbol{\Theta}}$;

for $k \in \{1, 2, \ldots, K\}$ **do**

  Set $\lambda_k := 0$;

  repeat

    Calculate MPLE using the current value of λ_k, and $\lambda_h = 0 \forall h \neq k$;

    Save $\boldsymbol{\lambda}$, $\widehat{\boldsymbol{\Theta}}$, and BIC;

    Update $\lambda_k := \lambda_k + \delta$;

  until *All penalised estimates for subpopulation* k *equal zero*;

end

Using these ranges for λ_k, form a grid of possible values for $\boldsymbol{\lambda}$;

for *All possible values for* $\boldsymbol{\lambda}$ **do**

  Calculate MPLE using current value of λ_k, and $\lambda_h = 0 \forall h \neq k$;

  Save $\boldsymbol{\lambda}$, $\widehat{\boldsymbol{\Theta}}$, and BIC;

end

Find the lowest value of BIC as well as the corresponding values of $\boldsymbol{\lambda}$, and $\widehat{\boldsymbol{\Theta}}$;

Return $\boldsymbol{\lambda}, \widehat{\boldsymbol{\Theta}}$, BIC;

Algorithm 5: Finding the best choice for $\boldsymbol{\lambda}$.

APPENDIX D
Regularity Conditions, Asymptotic Properties, and Proofs

Asymptotic properties are very important for regression using a FinMix GLMM. However, regularity conditions must be met before I can show these asymptotic properties. This Appendix focuses on these regularity conditions, asymptotic properties, and their proofs. I have separated this Appendix into two sections, one of the MLE of a FinMix GLMM, and one for MPLE of a FinMix GLMM. In each case, I stated the regularity conditions, and where relevant, verified thm. Then I stated the asymptotic properties and proved them.

D.1 Asymptotic Properties of the Maximum Likelihood Estimates

I proved two asymptotic properties of the MLE in this section, consistency and that the estimates follow a Gaussian distribution asymptotically. Prior to the proofs of the asymptotic properties, it is logical to start with the regularity conditions. There are six regularity conditions that are required, and I verify two of them below.

D.1.1 Regularity Conditions and Their Verifications

These regularity conditions for a FinMix GLMM follow from Casella and Berger (2002, p. 516). While these regularity conditions could be applied to any MLE, I have provided verifications that are specific to a FinMix GLMM.

Maximum Likelihood Estimation Regularity Condition 1

Observe $(\mathbf{y}_i, \mathbb{X}_i, \mathbb{Z}_i)$, $i \in \{1, 2, \ldots, n\}$ where $\mathbb{X}_i$ and $\mathbb{Z}_i$ are fixed covariates and $\mathbf{Y}_i|(\mathbb{X}_i, \mathbb{Z}_i, \mathbf{\Theta})$ are independent identically distributed such that

$$f_{\mathbf{y}_i}(\mathbf{y}_i|\mathbb{X}_i, \mathbb{Z}_i, \mathbf{\Theta}) = \sum_{k=1}^{K} \pi_k f_{\mathbf{y}_i}^{(k)}(\mathbf{y}_i|\mathbb{X}_i, \mathbb{Z}_i, \boldsymbol{\beta}_k, \mathbb{F}_k).$$

Maximum Likelihood Estimation Regularity Condition 2

The parameter $\mathbf{\Theta}$ is identifiable, that is, if

$$\sum_{k=1}^{K} \pi_k f_{\mathbf{y}_i}^{(k)}(\mathbf{y}_i|\mathbb{X}_i, \mathbb{Z}_i, \boldsymbol{\beta}_k, \mathbb{F}_k) = \sum_{k=1}^{K'} \pi'_k f_{y_i}^{(k)}(y_i|X_i, Z_i, \boldsymbol{\beta}'_k, \mathbb{F}'_k)$$

for all possible values of $\mathbf{y}_i$, and each $i = 1, 2, \ldots, n$ then $K = K'$, and $\mathbf{\Theta} = \mathbf{\Theta}'$. Identifiability is explored further in Section 3.2.2.

Maximum Likelihood Estimation Regularity Condition 3

The densities $f_{\mathbf{y}_i}(\mathbf{y}_i|\mathbb{X}_i, \mathbb{Z}_i, \mathbf{\Theta})$ have common support, and $f_{\mathbf{y}_i}(\mathbf{y}_i|\mathbb{X}_i, \mathbb{Z}_i, \mathbf{\Theta})$ is differentiable in $\mathbf{\Theta}$.

Verification of MLE Regularity Condition 3

I broke this condition into two parts, common support and differentiability. Recall that the support of the density $f_{\mathbf{y}_i}(\mathbf{y}_i)$ is the set of all points $\{\mathbf{y}_i|f_{\mathbf{y}_i}(\mathbf{y}_i) > 0\}$. The densities of a FinMix GLMM do have common support, though the support is dependent on the distribution of $\mathbf{Y}_i|(\mathbb{X}_i, \mathbb{Z}_i, \mathbf{\Theta})$. For example, if $\mathbf{Y}_i|(\mathbb{X}_i, \mathbb{Z}_i, \mathbf{\Theta}) \sim$ *Poisson* then $y_{ij} \in \{0, 1, \ldots\}$, and if $\mathbf{Y}_i|(\mathbb{X}_i, \mathbb{Z}_i, \mathbf{\Theta}) \sim$ *Gamma* then $0 < y_{ij} < \infty$. However, two special cases are more complex, and I examined these in more detail.

Considering the binomial distribution, if $m_{ij} = m, \forall i \forall j$, that is, the number of binomial trials is the same for all values of i, and j, the support of y_{ij} is $y_{ij} \in \{1, 2, \ldots, m\}$. This satisfies the condition in this special case, but not more generally. However, I assumed that in all cases m_{ij} is fixed, and known, therefore y_{ij} is $y_{ij} \in \{1, 2, \ldots, m_{ij}\}$, and one can consider $\frac{y_{ij}}{m_{ij}} \in [0, 1]$, which is the common support.

Another more complex case is when subpopulations have outcomes that follow different distributions. Even if the different K's probability density functions have different supports, the overall probability density function of the mixture will have common support for all values of $\boldsymbol{\Theta}$. As an example, assume without loss of generality that $K = 2$ where $f_{\mathbf{y}_i}^{(1)}(\mathbf{y}_i|\mathbb{X}_i, \mathbb{Z}_i, \boldsymbol{\beta}_k, \mathbb{\Gamma}_k)$, and $f_{\mathbf{y}_i}^{(2)}(\mathbf{y}_i|\mathbb{X}_i, \mathbb{Z}_i, \boldsymbol{\beta}_k, \mathbb{\Gamma}_k)$ have different supports. Examples of this special case are a zero-inflated Poisson, or when $f_{\mathbf{y}_i}^{(1)}(\mathbf{y}_i|\mathbb{X}_i, \mathbb{Z}_i, \boldsymbol{\beta}_k, \mathbb{\Gamma}_k) \sim \textit{Gamma}$ and $f_{\mathbf{y}_i}^{(2)}(\mathbf{y}_i|\mathbb{X}_i, \mathbb{Z}_i, \boldsymbol{\beta}_k, \mathbb{\Gamma}_k) \sim \textit{Gaussian}$. While $f_{\mathbf{y}_i}^{(1)}(\mathbf{y}_i|\mathbb{X}_i, \mathbb{Z}_i, \boldsymbol{\beta}_k, \mathbb{\Gamma}_k)$, and $f_{\mathbf{y}_i}^{(2)}(\mathbf{y}_i|\mathbb{X}_i, \mathbb{Z}_i, \boldsymbol{\beta}_k, \mathbb{\Gamma}_k)$ have different supports,

$$f_{\mathbf{y}_i}(\mathbf{y}_i|\mathbb{X}_i, \mathbb{Z}_i, \boldsymbol{\Theta}) = \pi_1 f_{\mathbf{y}_i}^{(1)}(\mathbf{y}_i|\mathbb{X}_i, \mathbb{Z}_i, \boldsymbol{\beta}_k, \mathbb{\Gamma}_k) + \pi_2 f_{\mathbf{y}_i}^{(2)}(\mathbf{y}_i|\mathbb{X}_i, \mathbb{Z}_i, \boldsymbol{\beta}_k, \mathbb{\Gamma}_k)$$

has the same support for all $\mathbf{y}_i$. More generally, the support for $f_{\mathbf{y}_i}(\mathbf{y}_i|\mathbb{X}_i, \mathbb{Z}_i, \boldsymbol{\Theta})$ is the union of the supports for $f_{\mathbf{y}_i}^{(1)}, f_{\mathbf{y}_i}^{(2)} \ldots, f_{\mathbf{y}_i}^{(K)}$, so this condition is satisfied for larger values of K.

Next, consider the differentiability of a FinMix GLMM probability density function with respect to $\boldsymbol{\Theta}$. Recall that

$$f_{\mathbf{y}_i}(\mathbf{y}_i|\mathbb{X}_i, \mathbb{Z}_i, \boldsymbol{\Theta}) = \sum_{k=1}^{K} \pi_k f_{\mathbf{y}_i}^{(k)}(\mathbf{y}_i|\mathbb{X}_i, \mathbb{Z}_i, \boldsymbol{\beta}_k, \mathbb{\Gamma}_k).$$

The derivative with respect to π_k, that is, a mixing proportion are

$$\frac{\partial}{\partial \pi_k} f_{\mathbf{y}_i}(\mathbf{y}_i|\mathbb{X}_i, \mathbb{Z}_i, \mathbf{\Theta}) = 0 + f^{(k)}_{\mathbf{y}_i}(\mathbf{y}_i|\mathbb{X}_i, \mathbb{Z}_i, \boldsymbol{\beta}_k, \mathbb{\Gamma}_k) + 0 \tag{D.1}$$

Derivatives of the approximate likelihood's for the Poisson, and binomial cases are in Section 3.3.2, and Appendix B but the derivations shown here are general to any exponential family. Having looked at the partial derivative with respect to π_k, it is sufficient to show that $f^{(k)}_{\mathbf{y}_i}(\mathbf{y}_i|\mathbb{X}_i, \mathbb{Z}_i, \boldsymbol{\beta}_k, \mathbb{\Gamma}_k)$ is differentiable with respect to $\boldsymbol{\beta}_k$, and $\mathbb{\Gamma}_k$, for each $k \in \{1, 2, \ldots, K\}$.

In this thesis, I assumed that $f^{(k)}_{\mathbf{y}_i|\mathbf{b}_i}(\mathbf{y}_i|\mathbf{b}_i, \mathbb{X}_i, \mathbb{Z}_i, \boldsymbol{\beta}_k, \mathbb{\Gamma}_k)$ follows a distribution that is from an exponential family, and used the canonical link function. Rewriting the probability density function for one subpopulation k as an exponential family using the form in McCullagh and Nelder (1989, p. 28) produces

$$f^{(k)}_{y_i|\mathbf{b}_i}(y_i|\mathbf{b}_i, X_i, Z_i, \beta_k, \Gamma_k) = \exp\left(\frac{y_{ij}\psi - b(\psi)}{a(\phi) + c(y, \phi)}\right).$$

I used the canonical link function, $\psi = \eta = \mathbf{x}_{ij}\boldsymbol{\beta}_k + \mathbf{z}_{ij}\mathbb{\Gamma}_k\mathbf{b}_i$, which is linear, and therefore differentiable. Next, I applied the chain rule,

$$\frac{\partial}{\partial \boldsymbol{\theta}_k} f^{(k)}_{y_i|\mathbf{b}_i}(y_i|\mathbf{b}_i, X_i, Z_i, \beta_k, \Gamma_k) = exp\left(\frac{y_{ij}\psi - b(\psi)}{a(\phi) + c(y, \phi)}\right)\frac{y_{ij} - b'(\psi)}{a(\phi) + c(y, \phi)}, \tag{D.2}$$

and

$$\begin{aligned} &\frac{\partial^2}{\partial \boldsymbol{\theta}_k^\top \partial \boldsymbol{\theta}_k} f^{(k)}_{y_i|\mathbf{b}_i}(y_i|\mathbf{b}_i, X_i, Z_i, \beta_k, \Gamma_k) \\ &= \frac{-b''(\psi)}{a(\phi)} \exp\left(\frac{y_{ij}\psi - b(\psi)}{a(\phi) + c(y, \phi)}\right) \\ &\quad + \left[\frac{y_{ij} - b'(\psi)}{a(\phi) + c(y, \phi)}\right]^2 \exp\left(\frac{y_{ij}\psi - b(\psi)}{a(\phi) + c(y, \phi)}\right) \end{aligned} \tag{D.3}$$

where

$$b'(\psi) = \frac{\partial b(\psi)}{\partial \boldsymbol{\theta}_k}$$

and

$$b''(\psi) = \frac{\partial^2 b(\psi)}{\partial \boldsymbol{\theta}_k^\top \partial \boldsymbol{\theta}_k}.$$

This still requires that $b'(\psi)$ and $b''(\psi)$ exist. According to the product rule, the product of differentiable functions is differentiable, so $f^{(k)}_{y_{ij}|\mathbf{b}_i}(y_{ij}|\mathbf{b}_i, x_{ij}, z_{ij}, \boldsymbol{\beta}_k, \mathbb{\Gamma}_k) \times f_{\mathbf{b}_i}(\mathbf{b}_i)$ is differentiable with respect to $\boldsymbol{\beta}_k$, and $\mathbb{\Gamma}_k$. Next, I used Leibniz's rule for differentiation under the integral sign.

$$\frac{\partial}{\partial \boldsymbol{\theta}_k} f^{(k)}_{\mathbf{y}_i}(\mathbf{y}_i|\mathbb{X}_i, \mathbb{Z}_i, \boldsymbol{\beta}_k, \mathbb{\Gamma}_k) = \int \frac{\partial}{\partial \boldsymbol{\theta}_k} f^{(k)}_{\mathbf{y}_i|\mathbf{b}_i}(\mathbf{y}_i|\mathbf{b}_i, \mathbb{X}_i, \mathbb{Z}_i, \boldsymbol{\beta}_k, \mathbb{\Gamma}_k) f_{\mathbf{b}_i}(\mathbf{b}_i) d\mathbf{b}_i$$

Also, recall that $\int_{-\infty}^{\infty} f(x)dx = \underset{A\to-\infty}{lim} \underset{R\to\infty}{lim} \int_A^R f(x)dx$. However, this integral could be a double, triple or higher-order integral, and goes from $-\infty$ to ∞ because $\mathbf{b}_i \sim Gaussian(\boldsymbol{\mu} = 0, \boldsymbol{\Sigma} = \boldsymbol{I}_q)$. So this rule may need to be applied multiple times. Thus $f_{\mathbf{y}_i}(\mathbf{y}_i|\mathbb{X}_i, \mathbb{Z}_i, \boldsymbol{\Theta})$ is differentiable in $\boldsymbol{\Theta}$, and Condition 3 is satisfied.

Maximum Likelihood Estimation Regularity Condition 4

The parameter space Φ contains an open set ω of which the true parameter value $\boldsymbol{\Theta}_0$ is an interior point.

Maximum Likelihood Estimation Regularity Condition 5

For every $\mathbf{y}_i$, the density $f_{\mathbf{y}_i}(\mathbf{y}_i|\mathbb{X}_i, \mathbb{Z}_i, \boldsymbol{\Theta})$ is three times differentiable with respect to $\boldsymbol{\Theta}$, and $\int f_{\mathbf{y}_i}(\mathbf{y}_i|\mathbb{X}_i, \mathbb{Z}_i, \boldsymbol{\Theta}) d\mathbf{y}_i$ can be differentiated three times under the integral sign.

Verification of Maximum Likelihood Estimation Regularity Condition 5

This is trivial for $\pi_1, \pi_2, ..., \pi_{K-1}$, I have already shown the first derivatives in Equation (D.1), and all higher-order derivatives with respect to π_k are 0 for all $k \in \{1, 2, \ldots, K\}$. For $\boldsymbol{\beta}_k$, and $\mathbb{\Gamma}_k$, I have shown this up to a choice of $b(\psi)$ for the first, and second derivatives in Equations (D.2) and (D.3) respectively, and for a selection of common distributions in Table D–1. Using again the notation from McCullagh and Nelder (1989, p. 28),

$$
\begin{aligned}
&\frac{\partial^3}{\partial \boldsymbol{\theta}_k^3} f^{(k)}_{y_{ij}|\mathbf{b}_i}(y_{ij}|\mathbf{b}_i, x_{ij}, z_{ij}, \boldsymbol{\beta}_k, \mathbb{\Gamma}_k) \\
&= \frac{-b'''(\psi)}{a(\phi)+c(y_{ij},\phi)} \exp\left(\frac{y_{ij}\psi - b(\psi)}{a(\phi)+c(y_{ij},\phi)}\right) + \frac{b''(\psi)(y_{ij}-b'(\psi))}{(a(\phi)+c(y_{ij},\phi))^2} \exp\left(\frac{y_{ij}\psi - b(\psi)}{a(\phi)+c(y_{ij},\phi)}\right) \\
&\quad + \left(\frac{y_{ij}\psi - b(\psi)}{a(\phi)+c(y_{ij},\phi)}\right)^3 \exp\left(\frac{y_{ij}\psi - b(\psi)}{a(\phi)+c(y_{ij},\phi)}\right) - \frac{b''(\psi)}{a(\phi)+c(y_{ij},\phi)} \exp\left(\frac{y_{ij}\psi - b(\psi)}{a(\phi)+c(y_{ij},\phi)}\right) \\
&= \exp\left(\frac{y_{ij}\psi - b(\psi)}{a(\phi)+c(y_{ij},\phi)}\right)\left[\frac{-b'''(\psi)-b''(\psi)}{a(\phi)+c(y_{ij},\phi)} - \frac{b''(\psi)(y_{ij}-b'(\psi))}{(a(\phi)+c(y_{ij},\phi))^2} + \left(\frac{y_{ij}\psi - b(\psi)}{a(\phi)+c(y_{ij},\phi)}\right)^3\right]
\end{aligned}
$$

Recall that using the chain rule, and $\psi(\boldsymbol{\beta}_k, \mathbb{\Gamma}_k) = x_{ij}\boldsymbol{\beta}_k + z_{ij}\mathbb{\Gamma}_k\mathbf{b}_i$, the derivatives with respect to $\boldsymbol{\beta}_k$, and $\mathbb{\Gamma}_k$ exist, and are as follows:

$$
\begin{aligned}
\frac{\partial \psi(\boldsymbol{\beta}_k, \mathbb{\Gamma}_k)}{\partial \boldsymbol{\beta}_{kh}} &= x_{ijh}; \\
\frac{\partial \psi(\boldsymbol{\beta}_k, \mathbb{\Gamma}_k)}{\partial \mathbb{\Gamma}_{khg}} &= z_{ijg} b_{ih}, \ h \le g; \\
\frac{\partial \psi(\boldsymbol{\beta}_k, \mathbb{\Gamma}_k)}{\partial \mathbb{\Gamma}_{khg}} &= 0, \ h > g.
\end{aligned}
$$

Table D–1: Table of derivatives of $b(\psi)$ for different distributions

Distribution	$b(\psi)$	$b'(\psi)$	$b''(\psi)$	$b'''(\psi)$
Gaussian	$\frac{\psi^2}{2}$	ψ	1	0
Poisson	e^{ψ}	e^{ψ}	e^{ψ}	e^{ψ}
Binomial	$\log(1+e^{\psi})$	$\frac{e^{\psi}}{1+e^{\psi}}$	$\frac{e^{\psi}}{(1+e^{\psi})^2}$	$\frac{e^{\psi}(1-e^{\psi})}{(1+e^{\psi})^2}$
Gamma	$-\log(-\psi)$	$\frac{1}{\psi}$	$\frac{-1}{\psi^2}$	$\frac{2}{\psi^3}$
Inverse Gaussian	$-(-2\psi)^{\frac{1}{2}}$	$(-2\psi)^{\frac{-1}{2}}$	$(-2\psi)^{\frac{-3}{2}}$	$3(-2\psi)^{\frac{-5}{2}}$

Maximum Likelihood Estimation Regularity Condition 6

For any $\mathbf{\Theta}_0 \in \Phi$, there exists an open set $\Phi_0 \subseteq \Phi$ such that $\mathbf{\Theta}_0 \in \Phi_0$ and a function $M(y_i)$, both of which may depend on $\mathbf{\Theta}_0$ such that for all $\mathbf{\Theta} \in \Phi_0$ and $E_{\mathbf{\Theta}_0}[M(y_i)] < \infty$

$$\left| \frac{\partial^3}{\partial \mathbf{\Theta}^3} \log(f(y_i|X_i, Z_i, \mathbf{\Theta})) \right| \leq M(y_i).$$

D.1.2 Asymptotic Properties of the Maximum Likelihood Estimators and Their Proofs

There are two asymptotic properties of the MLE that are of interest here, consistency and that the asymptotic distribution of the estimators is Gaussian. Recall that the MLE is transform invariant. However, the following theorem, and proof do not include that generalisation.

Theorem 1 (Consistency of the MLE)**.** *The MLE for* $\mathbf{\Theta} \in \Phi$*, denoted* $\hat{\mathbf{\Theta}}$ *is consistent, that is* $\lim_{n\to\infty} P_{\mathbf{\Theta}}(||\hat{\mathbf{\Theta}} - \mathbf{\Theta}|| > \epsilon) = 0$.

The following proof follows from Kendall et al. (1994, Section 18.10).

Proof. Since $\hat{\mathbf{\Theta}}$ is the MLE, $\forall \mathbf{\Theta} \in \Phi$,

$$L(Y|X, \hat{\mathbf{\Theta}}) \geq L(Y|X, \mathbf{\Theta})$$

and also

$$\log L(Y|X, \hat{\mathbf{\Theta}}) \geq \log L(Y|X, \mathbf{\Theta}).$$

Assume that $\mathbf{\Theta}_0$ is the true value of the parameter vector and consider the random variable

$$\frac{L(Y|X, \mathbf{\Theta})}{L(Y|X, \mathbf{\Theta}_0)},$$

Using Jensen's inequality, for all $\mathbf{\Theta}^* \neq \mathbf{\Theta}_0$,

$$E_{\mathbf{\Theta}_0}\left\{\log\left[\frac{L(Y|X, \mathbf{\Theta}^*)}{L(Y|X, \mathbf{\Theta})}\right]\right\} < \log\left[E_{\mathbf{\Theta}_0}\left\{\frac{L(Y|X, \mathbf{\Theta}^*)}{L(Y|X, \mathbf{\Theta})}\right\}\right]$$

Expanding the expectation on the right side of the inequality,

$$\begin{aligned} &\int\int \ldots \int \frac{L(Y|X, \mathbf{\Theta}^*)}{L(Y|X, \mathbf{\Theta}_0)} L(Y|X, \mathbf{\Theta}_0) dy_1 dy_2 \ldots dy_n \\ &= \int\int \ldots \int L(Y|X, \mathbf{\Theta}^*) dy_1 dy_2 \ldots dy_n = 1, \end{aligned}$$

because the integral over a probability density function is always 1. Therefore, taking the log of that expectation, and $\log(1) = 0$,

$$\begin{aligned} E_{\mathbf{\Theta}_0} &\quad \left\{\log\left[\frac{L(Y|X, \mathbf{\Theta}^*)}{L(Y|X, \mathbf{\Theta})}\right]\right\} < 0 \\ \Longrightarrow &\quad E_{\mathbf{\Theta}_0}\{\log[L(Y|X, \mathbf{\Theta}^*)]\} - E_{\mathbf{\Theta}_0}\{\log[L(Y|X, \mathbf{\Theta})]\} < 0 \\ \Longrightarrow &\quad E_{\mathbf{\Theta}_0}\{\log[L(Y|X, \mathbf{\Theta}^*)]\} < E_{\mathbf{\Theta}_0}\{\log[L(Y|X, \mathbf{\Theta})]\} \\ \Longrightarrow &\quad E_{\mathbf{\Theta}_0}\left\{\frac{1}{n}\log[L(Y|X, \mathbf{\Theta}^*)]\right\} < E_{\mathbf{\Theta}_0}\left\{\frac{1}{n}\log[L(Y|X, \mathbf{\Theta})]\right\} \end{aligned}$$

For any value of $\boldsymbol{\Theta} \in \Phi$, $\frac{1}{n} \log L(Y|X, \boldsymbol{\Theta}) = \frac{1}{n} \sum_{i=1}^{n} \log(y_i|X_i, \boldsymbol{\Theta})$, which is the mean of n independent identically distributed y_i, and

$$E_{\boldsymbol{\Theta}_0}[\log f_{y_i}(y_i|X_i, \boldsymbol{\Theta})] = E_{\boldsymbol{\Theta}_0}\left[\frac{1}{n} \log L(Y|X, \boldsymbol{\Theta})\right].$$

By the strong law of large numbers, $\frac{1}{n} \log L(Y|X, \boldsymbol{\Theta})$ converges with probability 1 to its expectation. So $\frac{1}{n} \log L(Y|X, \boldsymbol{\Theta}^*) < \frac{1}{n} \log L(Y|X, \boldsymbol{\Theta}_0)$ and when $\boldsymbol{\Theta}^* \neq \boldsymbol{\Theta}_0$, $\lim\limits_{n\to\infty} P(\log L(Y|X, \boldsymbol{\Theta}^*) < \log) = 1$. However, when $\boldsymbol{\Theta} = \boldsymbol{\Theta}_0$, $\log L(Y|X, \hat{\boldsymbol{\Theta}}) \geq \log L(Y|X, \boldsymbol{\Theta}_0)$. Therefore, $P\left(\lim\limits_{n\to\infty} \hat{\boldsymbol{\Theta}} = \boldsymbol{\Theta}_0\right) = 1$. Therefore, the MLE is consistent. □

Theorem 2 (Asymptotic Distribution of the MLE)**.** *The MLE of a FinMix GLMM is asymptotically Gaussian, that is,* $\sqrt{n}(\hat{\boldsymbol{\Theta}} - \boldsymbol{\Theta}_0)$ *converges in distribution to a Gaussian distribution.*

The proof that the maximum likelihood estimator is asymptotically Gaussian uses a Taylor series around the true value of $\boldsymbol{\Theta}$, assumed to be $\boldsymbol{\Theta}_0$ as shown in Casella and Berger (2002, p. 472).

Proof. Let $\ell(\boldsymbol{\Theta}|Y, X) = \sum_{i=1}^{n} \log f_{y_i}(y_i|X_i, \boldsymbol{\Theta})$ be the log-likelihood function, and denote its derivatives with respect to $\boldsymbol{\Theta}$ as $\frac{\partial}{\partial \boldsymbol{\Theta}} \ell(\boldsymbol{\Theta}|Y, X)$, and $\frac{\partial^2}{\partial \boldsymbol{\Theta}^\top \partial \boldsymbol{\Theta}} \ell(\boldsymbol{\Theta}|Y, X)$. Using a first-order Taylor expansion of $\frac{\partial}{\partial \boldsymbol{\Theta}} \ell(\boldsymbol{\Theta}|Y, X)$ around the true value of $\boldsymbol{\Theta}$ denoted $\boldsymbol{\Theta}_0$,

$$\frac{\partial}{\partial \boldsymbol{\Theta}} \ell(\boldsymbol{\Theta}|Y, X) \approx \frac{\partial}{\partial \boldsymbol{\Theta}} \ell(\boldsymbol{\Theta}_0|Y, X) + (\boldsymbol{\Theta} - \boldsymbol{\Theta}_0) \frac{\partial^2}{\partial \boldsymbol{\Theta}^\top \partial \boldsymbol{\Theta}} \ell(\boldsymbol{\Theta}_0|Y, X).$$

Substituting in the MLE $\hat{\mathbf{\Theta}}$ provides

$$\begin{aligned} 0 &= \frac{\partial}{\partial \mathbf{\Theta}} \ell(\mathbf{\Theta}_0|Y,X) + (\hat{\mathbf{\Theta}} - \mathbf{\Theta}_0) \frac{\partial^2}{\partial \mathbf{\Theta}^\top \partial \mathbf{\Theta}} \ell(\mathbf{\Theta}_0|Y,X) \\ (\hat{\mathbf{\Theta}} - \mathbf{\Theta}_0) &= \frac{-\frac{\partial}{\partial \mathbf{\Theta}} \ell(\mathbf{\Theta}_0|Y,X)}{\frac{\partial^2}{\partial \mathbf{\Theta}^\top \partial \mathbf{\Theta}} \ell(\mathbf{\Theta}_0|Y,X)} \\ \sqrt{n}(\hat{\mathbf{\Theta}} - \mathbf{\Theta}_0) &= \frac{-\sqrt{n}\frac{\partial}{\partial \mathbf{\Theta}} \ell(\mathbf{\Theta}_0|Y,X)}{\frac{\partial^2}{\partial \mathbf{\Theta}^\top \partial \mathbf{\Theta}} \ell(\mathbf{\Theta}_0|Y,X)} \end{aligned}$$

Denote $\ell(\mathbf{\Theta}_0) = E\left[\frac{\partial}{\partial \mathbf{\Theta}} \ell(\mathbf{\Theta}_0|Y,X)\right]^2$. Using the Central Limit Theorem, and the Weak Law of Large Numbers,

$$-\frac{1}{\sqrt{n}} \frac{\partial}{\partial \mathbf{\Theta}} \ell(\mathbf{\Theta}_0|Y,X) \to Gaussian(\mu = 0, \mathbf{\Sigma} = I(\mathbf{\Theta}_0))$$

in distribution and

$$\frac{1}{n} \frac{\partial^2}{\partial \mathbf{\Theta}^\top \partial \mathbf{\Theta}} \ell(\mathbf{\Theta}_0|Y,X) \to \mathbf{\Theta}_0$$

in probability, so $\sqrt{n}(\hat{\mathbf{\Theta}} - \mathbf{\Theta}_0)$ converges in distribution to $Gaussian(\mu = 0, \mathbf{\Sigma} = \frac{1}{I(\mathbf{\Theta}_0)})$, which proves that the MLE is asymptotically Gaussian. $\square$

D.2 Asymptotic Properties of Maximum Penalised Likelihood Estimates

The regularity conditions, properties, and proofs follow in the same way as Du et al. (2013). Assume that the data follows a FinMix GLMM, and is of the form $(\mathbf{y}_i, \mathbb{X}_i, \mathbb{Z}_i)$, $\forall i \in \{1, 2, \ldots, n\}$. Let $\mathbf{\Theta} = (\pi_1, \pi_2, \ldots, \pi_{(K-1)}, \beta_{11}, \beta_{12}, \ldots, \beta_{1p}, \Gamma_{111}, \ldots, \Gamma_{1qq},$ $\ldots, \beta_{K1}, \beta_{K2}, \ldots, \beta_{Kp}, \Gamma_{K11}, \Gamma_{K21}, \ldots, \Gamma_{Kq(q-1)}, \Gamma_{Kqq})^\top$. Because this notation is cumbersome, $\mathbf{\Theta}$ is rewritten as $\mathbf{\Theta} = (\Theta_1, \Theta_2, \ldots, \Theta_S)^\top \in \Phi$. Note that $S = (K-1) + K(p + \frac{1}{2}q(q+1)) = (K-1) + K\varkappa$. Recall that the joint density of $(\mathbf{y}_i, \mathbb{X}_i, \mathbb{Z}_i)$ is $f(\mathbf{y}_i, \mathbb{X}_i, \mathbb{Z}_i|\mathbf{\Theta})$.

D.2.1 Maximum Penalised Likelihood Estimation Regularity Conditions

The following conditions are necessary for the ensuing theorems. I presented conditions on the model first, followed by conditions on the penalty function.

Maximum Penalised Likelihood Estimation Regularity Condition 1

The model $f(\mathbf{y}_i, \mathbb{X}_i, \mathbb{Z}_i|\boldsymbol{\Theta})$ is identifiable. More information on the identifiability of a FinMix GLMM can be found in Section 3.2.2.

Maximum Penalised Likelihood Estimation Regularity Condition 2

The joint density function is thrice differentiable such that for every possible $\boldsymbol{\Theta}^* \in \Phi$, there exist functions $G_{1i}(\mathbf{y}_i, \mathbb{X}_i, \mathbb{Z}_i)$, $G_{2i}(\mathbf{y}_i, \mathbb{X}_i, \mathbb{Z}_i)$, and $G_{3i}(\mathbf{y}_i, \mathbb{X}_i, \mathbb{Z}_i)$, such that $E_{\boldsymbol{\Theta}^*}[G_{1i}(\mathbf{y}_i, \mathbb{X}_i, \mathbb{Z}_i)] < \infty$, $E_{\boldsymbol{\Theta}^*}[G_{2i}(\mathbf{y}_i, \mathbb{X}_i, \mathbb{Z}_i)] < \infty$, and $E_{\boldsymbol{\Theta}^*}[G_{3i}(\mathbf{y}_i, \mathbb{X}_i, \mathbb{Z}_i)] < \infty$, and for all $\boldsymbol{\Theta}$ in a neighbourhood of $\boldsymbol{\Theta}^*$,

$$
\begin{aligned}
\left|\frac{\partial \log f(\mathbf{y}_i \mathbb{X}_i, \mathbb{Z}_i|\boldsymbol{\Theta})}{\partial \Theta_\iota}\right| &< G_{1i}(\mathbf{y}_i, \mathbb{X}_i, \mathbb{Z}_i), \\
\left|\frac{\partial^2 \log f(\mathbf{y}_i \mathbb{X}_i, \mathbb{Z}_i|\boldsymbol{\Theta})}{\partial \Theta_\iota \partial \Theta_\kappa}\right| &< G_{2i}(\mathbf{y}_i, \mathbb{X}_i, \mathbb{Z}_i), \\
\left|\frac{\partial^3 \log f(\mathbf{y}_i \mathbb{X}_i, \mathbb{Z}_i|\boldsymbol{\Theta})}{\partial \Theta_\iota \partial \Theta_\kappa \partial \Theta_\nu}\right| &< G_{3i}(\mathbf{y}_i, \mathbb{X}_i, \mathbb{Z}_i).
\end{aligned}
$$

Maximum Penalised Likelihood Estimation Regularity Condition 3

The Fisher information matrix, denoted as $I(\boldsymbol{\Theta})$, is both finite as well as positive definite for all possible $\boldsymbol{\Theta} \in \Phi$.

For ease of notation in this section, I decomposed the parameter $\boldsymbol{\Theta}$ into two components. Without loss of generality, assume that $\boldsymbol{\Theta}$ can be decomposed into $\boldsymbol{\Theta}^\top = (\boldsymbol{\Theta}^{\dagger\top}, \boldsymbol{\Theta}^{\ddagger\top})$ such that $\boldsymbol{\Theta}^\ddagger$ contains only zeros. Separate also the vector representing the true parameters in the same way, that is $\boldsymbol{\Theta}_0^\top = (\boldsymbol{\Theta}_0^{\dagger\top}, \boldsymbol{\Theta}_0^{\ddagger\top})$. All elements of $\boldsymbol{\Theta}_0^\dagger$

can be split into $\boldsymbol{\beta}_0^{\dagger}$, and $\mathbb{d}_0^{\dagger *}$ depending on if they are from $\boldsymbol{\beta}$, or $\mathbb{d}^*$ respectively. Similarly, all elements of $\boldsymbol{\Theta}_0^{\ddagger}$ can be split into $\boldsymbol{\beta}_0^{\ddagger}$, and $\mathbb{d}_0^{\ddagger *}$. Then, the following three values are defined for use in the subsequent asymptotic properties:

$$
\begin{aligned}
a_n &= max_{kh} \left\{ \sqrt{n} \left| p_{nk}(\boldsymbol{\beta}_0^{\dagger}) \right|, \sqrt{n} \left| p_{nk}(\mathbb{d}_0^{\dagger *}) \right| \right\}; \\
b_n &= max_{kh} \left\{ \sqrt{n} \left| \frac{\partial p_{nk}(\boldsymbol{\beta}_0^{\dagger})}{\partial \boldsymbol{\beta}} \right|, \sqrt{n} \left| \frac{\partial p_{nk}(\mathbb{d}_0^{\dagger *})}{\partial \mathbb{d}^*} \right| \right\}; \\
c_n &= max_{kh} \left\{ \sqrt{n} \left| \frac{\partial^2 p_{nk}(\boldsymbol{\beta}_0^{\dagger})}{\partial \boldsymbol{\beta}^{\top} \partial \boldsymbol{\beta}} \right|, \sqrt{n} \left| \frac{\partial^2 p_{nk}(\mathbb{d}_0^{\dagger *})}{\partial \mathbb{d}^{*\top} \partial \mathbb{d}^*} \right| \right\}.
\end{aligned}
$$

Penalty Condition 1

For all values of n, and k, the penalty function $p_{nk}(0) = 0$, the penalty function is symmetric, non-negative, non-decreasing, and twice differentiable on the open set $(0, \infty)$. That is, $p_{nk}(x) = p_{nk}(-x)$, $p_{nk}(x) \geq 0$, $x \leq y \Rightarrow p_{nk}(x) \leq p_{nk}(y)$, and $p''_{nk}(x)$ exists.

Penalty Condition 2

As $n \to \infty$, $a_n = o(1 + b_n)$ and $c_n = o(\sqrt{n})$. This means that the non-zero values of the penalty are asymptotically bounded. That is, for every $\epsilon > 0$, there exists an integer n_ϵ, potentially dependent on ϵ such that if $n > n_\epsilon$ then $|a_n| < \epsilon|1 + b_n|$. Similarly, for every $\epsilon > 0$, there exists an integer n_ϵ, potentially dependent on ϵ such that if $n > n_\epsilon$ then $|c_n| < \epsilon\sqrt{n}$. Formal definitions of asymptotic notation can be found in Bishop et al. (2007, Section 14.2).

Penalty Condition 3

For any $N_n = \{\varrho | 0 < \varrho \leq n^{-\frac{1}{2}} \log(n)\}$, $\lim\limits_{n \to \infty} \inf\limits_{\varrho \in N_n} \sqrt{n} \frac{dp_{nk}(\varrho)}{d\varrho} = \infty$.

D.2.2 Asymptotic Properties of Maximum Penalised Likelihood Estimators and Their Proofs

Consistency of the estimator in probability is included in the definition of existence.

Theorem 3 (Existence and consistency of the MPLE)**.** *Assume that* $\mathbf{\Theta}^\top = (\mathbf{\Theta}^\top, 0^\top)$ *are the true parameters of a FinMix GLMM. Assume that the MPLE regularity conditions 1, 2, and 3 are satisfied by this model and assume the penalty function* p_{nk} *satisfies penalty conditions 1, 2, and 3. Then, there exists* $\hat{\mathbf{\Theta}}_n$ *such that* $\hat{\mathbf{\Theta}}_n$ *is a local maximum of the penalised log-likelihood function, and also* $||\hat{\mathbf{\Theta}}_n - \mathbf{\Theta}_0|| = O_p\{n^{-\frac{1}{2}}(1 + b_n)\}$.

Recall that by the definition of O_p (Bishop et al., 2007, Section 14.4), $||\hat{\mathbf{\Theta}}_n - \mathbf{\Theta}_0|| = O_p\{n^{-\frac{1}{2}}(1 + b_n)\}$ means that for every $\eta > 0$, there exists a constant K_η, and an integer n_η, both of which could depend on η, such that for every $n > n_\eta$,

$$P\{||\hat{\mathbf{\Theta}}_n - \mathbf{\Theta}_0|| \leq K_\eta(n^{-\frac{1}{2}}(1 + b_n))\} \geq 1 - \eta.$$

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,

$$P\{sup||u|| < \ell^\sharp_{n\lambda}(\mathbf{\Theta}_0)\} \geq 1 - \epsilon$$

where $sup||u|| = M_\epsilon \ell^\sharp_{n\lambda}(\mathbf{\Theta}_0 + r_n u)$. Therefore, there is a large probability that there exists a local maximum of $\ell^\sharp_{n\lambda}(\mathbf{\Theta}_0)$ in $\{\mathbf{\Theta}_0 + r_n u : ||u|| \leq M_\epsilon\}$. This is an open set around the true value of the parameter vector. This local maximiser $\hat{\mathbf{\Theta}}_n$ satisfies

$||\hat{\mathbf{\Theta}}_n - \mathbf{\Theta}_0|| = O_p\{n^{-1/2}(1+b_n)\}$. Let

$$
\begin{aligned}
\Delta_n(u) &= \ell^{\sharp}_{n\lambda}(\mathbf{\Theta}_0 + r_n u) - \ell^{\sharp}_{n\lambda}(\mathbf{\Theta}_0) \\
&= \{\ell_n(\mathbf{\Theta}_0 + r_n u) - \ell_n(\mathbf{\Theta}_0)\} - \{p_{n\lambda}(\mathbf{\Theta}_0 + r_n u) - p_{n\lambda}(\mathbf{\Theta}_0)\}.
\end{aligned}
$$

Recall from penalty condition 1 that $p_{nk\lambda}(0) = 0$, so $p_{nk\lambda}(\mathbf{\Theta}_0^{\ddagger}) = 0$, and $p_{n\lambda}(\mathbf{\Theta}_0) = p_{n\lambda}(\mathbf{\Theta}_0^{\dagger})$. Thus, $p_{n\lambda}(\mathbf{\Theta}_0 + r_n u)$ is a sum of positive terms. Let u_I be the subvector of u that corresponds to the non-zero effect. Then,

$$
\begin{aligned}
\Delta_n(u) &\leq \{\ell_n(\mathbf{\Theta}_0 + r_n u) - \ell_n(\mathbf{\Theta}_0)\} - \{p_{n\lambda}(\mathbf{\Theta}_0^{\dagger} + r_n u_I) - p_{n\lambda}(\mathbf{\Theta}_0^{\dagger})\} \\
&\leq \{\ell_n(\mathbf{\Theta}_0 + r_n u) - \ell_n(\mathbf{\Theta}_0)\} - |p_{n\lambda}(\mathbf{\Theta}_0^{\dagger} + r_n u_I) - p_{n\lambda}(\mathbf{\Theta}_0^{\dagger})| \qquad \text{(D.4)}
\end{aligned}
$$

by the triangle inequality. A Taylor's expansion of $\ell_n(\mathbf{\Theta}_0)$ around $u = 0$ results in

$$
\begin{aligned}
\ell_n(\mathbf{\Theta}_0 + r_n u) - \ell_n(\mathbf{\Theta}_0) &\approx r_n \ell_n'(\mathbf{\Theta}_0)^{\top} u + \frac{1}{2} u^{\top}(\ell_n''(\mathbf{\Theta}_0)) u r_n^2 \\
&= \frac{(1+b_n)}{\sqrt{n}} \ell_n'(\mathbf{\Theta}_0)^{\top} u + \frac{(1+b_n)^2}{2m} u^{\top}(\ell_n''(\mathbf{\Theta}_0)) u.
\end{aligned}
$$

However, by model condition 2, the remainder term must converge to 0 as $n \to \infty$. Next, consider the Hessian matrix $\ell_n''(\mathbf{\Theta}_0)$. By properties of the Hessian matrix $\frac{1}{m}\ell_n''(\mathbf{\Theta}_0)$ converges in probability to $-I(\mathbf{\Theta}_0)$.

Therefore,

$$\begin{aligned}\ell_n(\mathbf{\Theta}_0 &+& r_n u) - \ell_n(\mathbf{\Theta}_0) \\ &=& \frac{(1+b_n)}{\sqrt{n}}\ell_n'(\mathbf{\Theta}_0)^\top u - \frac{(1+b_n)^2}{2}u^\top I(\mathbf{\Theta}_p)u\{1+o_p(1)\} \\ &=& (1+b_n)O_p(1)||u|| - \frac{(1+b_n)^2}{2}u^\top I(\mathbf{\Theta}_p)u\{1+o_p(1)\}, \qquad \text{(D.5)}\end{aligned}$$

since $\frac{1}{\sqrt{n}}\ell_n'(\mathbf{\Theta}_0) = O_p(1)$ by model condition 2. Given model condition 3, it is assumed that, $I(\mathbf{\Theta}_0)$ is positive definite. Thus, $I(\mathbf{\Theta}_0)$ has only positive eigenvalues. Let $\eta_{min} > 0$ be the smallest eigenvalue of $I(\mathbf{\Theta}_0)$. Since η_{min} is an eigenvalue, and $I(\mathbf{\Theta}_0)$ is positive definite, $u^\top I(\mathbf{\Theta}_0)u \geq \eta_{min}||u||^2$. Applying this result to (D.5), I have

$$\ell_n(\mathbf{\Theta}_0 + r_n u) - \ell_n(\mathbf{\Theta}_0) \leq (1+b_n)O_p(1)||u|| - \frac{(1+b_n)^2}{2}\eta_{min}||u||^2\{1+o_p(1)\} \quad \text{(D.6)}$$

In addition, by using a Taylor's expansion, and the triangle inequality,

$$\begin{aligned}&& |p_{n\lambda}(\mathbf{\Theta}_0^\dagger + r_n u_I) - p_{n\lambda}(\mathbf{\Theta}_0^\dagger)| \\ &=& p_{n\lambda}'(\mathbf{\Theta}_0^\dagger)^\top r_n u_I + \frac{r_n^2}{2}u_I^\top p_{n\lambda}''(\mathbf{\Theta}_0^\dagger)u_I\{1+o(1)\} \\ &\leq& r_n|p_{n\lambda}'(\mathbf{\Theta}_0^\dagger)^\top u_I| + \frac{r_n^2}{2}|u_I^\top p_{n\lambda}''(\mathbf{\Theta}_0^\dagger)u_I|\{1+o(1)\} \\ &\leq& r_n||p_{n\lambda}'(\mathbf{\Theta}_0^\dagger)^\top|| \times ||u_I|| + \frac{r_n^2}{2}||diag(p_{n\lambda}''(\mathbf{\Theta}_0^\dagger))|| \times ||u_I||^2\{1+o(1)\}. \quad \text{(D.7)}\end{aligned}$$

Let t_k be the total number of true non-zero fixed, and random effects in the k^{th} subpopulation, and let $t = max\{t_k, k = 1, \ldots, K\}$. Let $\boldsymbol{\beta}_0^\dagger$, and $\mathbb{d}_0^\dagger$ denote respectively the vectors of $\boldsymbol{\beta}_{kh}^\dagger$s, and $\mathbb{d}_0^\dagger$s from all K subpopulations. Consider the first term of

(D.7), $||p'_{n\lambda}(\boldsymbol{\Theta}_0^\dagger)||$. Separating $||p'_{n\lambda}(\boldsymbol{\Theta}_0^\dagger)||$ into parts,

$$||p'_{n\lambda}(\boldsymbol{\Theta}_0^\dagger)|| = ||p'_{n\lambda}(\pi_1, \ldots, \pi_K)|| + ||p'_{n\lambda}(\boldsymbol{\beta}_0^\dagger, \mathbb{d}_0^\dagger)||$$

because of the triangle inequality. Recall that the penalty function is always non-negative, so this is an equality, not an inequality. Recall that $p_{n\lambda}(\boldsymbol{\Theta}) = \sum_{k=1}^K \pi_k \times n(\sum_{h=1}^p p_{nk\lambda}(\beta_{kh}) + \sum_{h=1}^q p_{nk\lambda}(d_{kh}))$, therefore, the derivative of the penalty function with respect to the mixing proportions is

$$p'_{n\lambda}(\pi_1, \ldots, \pi_K) = \begin{pmatrix} n(\sum_{h=1}^p p_{nk\lambda}(\beta_{1h}) + \sum_{h=1}^q p_{nk\lambda}(d_{1h})) \\ n(\sum_{h=1}^p p_{nk\lambda}(\beta_{2h}) + \sum_{h=1}^q p_{nk\lambda}(d_{2h})) \\ \vdots \\ n(\sum_{h=1}^p p_{nk\lambda}(\beta_{Kh}) + \sum_{h=1}^q p_{nk\lambda}(d_{Kh})) \end{pmatrix}.$$

Recall that the penalty on the fixed, and random effects includes the value of the mixing proportions, so even though there is no penalty on the mixing proportions, this derivative is not zero. Thus,

$$\begin{aligned} ||p'_{n\lambda}(\pi_1, \ldots, \pi_K)|| &= n \times \sqrt{\sum_{k=1}^K \left[\sum_{h=1}^p p_{nk\lambda}(\beta_{0kh}^\dagger) + \sum_{h=1}^q p_{nk\lambda}(\mathbb{d}_{0kh}^\dagger)\right]^2} \\ &\leq n\sqrt{\sum_{k=1}^K \left[t_k \times \frac{a_n}{\sqrt{n}}\right]^2} \\ &= a_n\sqrt{n}\sqrt{\sum_{k=1}^K t_k^2} \\ &\leq a_n\sqrt{n}\sqrt{\sum_{k=1}^K t^2} = a_n\sqrt{n}\sqrt{K}t. \end{aligned}$$

Furthermore,

$$\begin{aligned}
||p'_{n\lambda}(\boldsymbol{\beta}_0^\dagger, \mathbb{d}_0^\dagger)|| &= ||\Delta p_{n\lambda}(\beta_{011}^\dagger, \ldots, \beta_{01p}^\dagger, \ldots, \beta_{0K1}^\dagger, \ldots, d_{011}^\dagger, \ldots, d_{01q}^\dagger, \ldots, d_{0K1}^\dagger, \ldots, d_{0Kq}^\dagger)|| \\
&= n||\pi_1 p'_{n1}(\beta_{011}^\dagger), \ldots, \pi_1 p'_{n1}(\beta_{01p}^\dagger), \ldots, \pi_K p'_{nK}(\beta_{0K1}^\dagger), \ldots, \pi_K p'_{nK}(\beta_{0Kp}^\dagger), \\
&\quad \pi_1 p'_{n1}(d_{011}^\dagger), \ldots, \pi_1 p'_{n1}(d_{01p}^\dagger), \ldots, \pi_K p'_{nK}(d_{0K1}^\dagger), \ldots, \pi_K p'_{nK}(d_{0Kp}^\dagger)|| \\
&\leq n||p'_{n1}(\beta_{011}^\dagger), \ldots, p'_{n1}(\beta_{01p}^\dagger), \ldots, p'_{nK}(\beta_{0K1}^\dagger), \ldots, p'_{nK}(\beta_{0Kp}^\dagger), \\
&\quad p'_{m1}(d_{011}^\dagger), \ldots, p'_{m1}(d_{01p}^\dagger), \ldots, p'_{mK}(d_{0K1}^\dagger), \ldots, p'_{mK}(d_{0Kp}^\dagger)|| \\
&= m\sqrt{\sum_{k=1}^{K}\sum_{h=1}^{p} p'_{nk\lambda}(\beta_{0kh}^\dagger)^2 + \sum_{k=1}^{K}\sum_{h=1}^{q} p'_{nk\lambda}(d_{0kh}^\dagger)^2} \\
&= m\sqrt{\sum_{k=1}^{K}\left(\sum_{h=1}^{p} p'_{nk\lambda}(\beta_{0kh}^\dagger)^2 + \sum_{h=1}^{q} p'_{nk\lambda}(d_{0kh}^\dagger)^2\right)} \\
&\leq m\sqrt{\sum_{k=1}^{K} t_k \times \left(\frac{b_n}{\sqrt{n}}\right)^2} = \sqrt{n} b_n \sqrt{\sum_{k=1}^{K} t_k} \\
&\leq b_n \sqrt{n}\sqrt{K \times t}.
\end{aligned}$$

Therefore,

$$||p'_{n\lambda}(\boldsymbol{\Theta}_0)|| \leq a_n \sqrt{n}\sqrt{K}t + b_n\sqrt{n}\sqrt{Kt}.$$

Using (D.7), this leads to

$$\begin{aligned}
|p_{n\lambda}(\boldsymbol{\Theta}_0^\dagger &+ r)mu_I) - p_{n\lambda}(\boldsymbol{\Theta}_0^\dagger)| \\
&\leq r_n(a_n\sqrt{n}\sqrt{K}t + b_n\sqrt{n}\sqrt{Kt})||u|| + \frac{r_n^2}{2}||diag(p''_{n\lambda}(\boldsymbol{\Theta}_0^\dagger))|| \times ||u_I||^2\{1+o(1)\} \\
&= \frac{1+b_n}{\sqrt{n}}(a_n\sqrt{n}\sqrt{K}t + b_n\sqrt{n}\sqrt{Kt})||u|| \\
&\quad + \frac{1}{2}\frac{(1+b_n)^2}{m}||diag(p''_{n\lambda}(\boldsymbol{\Theta}_0^\dagger))|| \times ||u_I||^2\{1+o(1)\} \\
&= a_n(1+b_n)\sqrt{K}t||u|| + b_n(1+b_n)\sqrt{Kt}||u|| \\
&\quad + \frac{1}{2}\frac{(1+b_n)^2}{m}||diag(p''_{n\lambda}(\boldsymbol{\Theta}_0^\dagger))|| \times ||u_I||^2\{1+o(1)\}. \qquad \text{(D.8)}
\end{aligned}$$

Furthermore,

$$\begin{aligned}
&||diag(p''_{n\lambda}(\boldsymbol{\Theta}_0^\dagger))|| \\
&= m\sqrt{\sum_{k=1}^{K}\sum_{h=1}^{p} p''_{nk\lambda}(\boldsymbol{\beta}_{kh}^\dagger)^2\pi_k^2 + \sum_{k=1}^{K}\sum_{h=1}^{q} p''_{nk\lambda}(\mathbb{d}_{kh}^\dagger)^2\pi_k^2} \\
&\leq m\sqrt{\sum_{k=1}^{K}\left(\sum_{h=1}^{p} p''_{nk\lambda}(\boldsymbol{\beta}_{kh}^\dagger)^2\pi_k^2 + \sum_{h=1}^{q} p''_{nk\lambda}(\mathbb{d}_{kh}^\dagger)^2\pi_k^2\right)} \\
&\leq m\sqrt{\sum_{k=1}^{K} t_k\left(\frac{c_n}{\sqrt{n}}\right)^2} = \sqrt{n}c_n\sqrt{\sum_{k=1}^{K} t_k} \\
&\leq c_n\sqrt{n}\sqrt{K}t.
\end{aligned}$$

Combining this result with (D.8) gives

$$\begin{aligned}
|p_{n\lambda}(\boldsymbol{\Theta}_0^\dagger + r_n u_I) - p_{n\lambda}(\boldsymbol{\Theta}_0^\dagger)| \leq\ & a_n(1+b_n)\sqrt{K}t||u|| \\
& + b_n(1+b_n)\sqrt{Kt}||u|| \\
& + \frac{1}{2}\frac{(1+b_n)^2}{\sqrt{n}}c_n\sqrt{Kt}||u||^2(1+o(1)). \qquad \text{(D.9)}
\end{aligned}$$

Combining (D.4), (D.6), and (D.9) results in

$$\begin{aligned}\Delta_n(u) \;\; \leq \;\; & (1+b_n)O_p(1)||u|| - \frac{(1+b_n)^2}{2}\eta_{min}||u||^2\{1+o_p(1)\} \\ & + a_n(1+b_n)\sqrt{K}t||u|| + b_n(1+b_n)\sqrt{K}t||u|| \\ & + \frac{1}{2}\frac{(1+b_n)^2}{\sqrt{n}}c_n\sqrt{Kt}||u||^2(1+o(1)).\end{aligned}$$

Dividing both sides of the above inequality by $(1+b_n)^2$ provides

$$\begin{aligned}\frac{\Delta_n(u)}{(1+b_n)^2} \;\; \leq \;\; & \frac{1}{1+b_n}O_p(1)||u|| - \frac{1}{2}\eta_{min}||u||^2\{1+o_p(1)\} \\ & + \frac{a_n}{1+b_n}\sqrt{K}t||u|| + \frac{b_n}{1+b_n}\sqrt{K}t||u|| \\ & + \frac{1}{2}\frac{c_n}{\sqrt{n}}\sqrt{Kt}||u||^2(1+o(1)).\end{aligned}$$

By penalty condition 2, $a_n = o(1+b_n)$, and $c_n = o(\sqrt{n})$. Applying these conditions to the second row of the above inequality, produces

$$\frac{\Delta_n(u)}{(1+b_n)^2} \leq \frac{1}{1+b_n}O_p(1)||u|| - \frac{1}{2}\eta_{min}||u||^2\{1+o_p(1)\} + \frac{b_n}{1+b_n}\sqrt{K}t||u|| + o(1) \quad \text{(D.10)}$$

Using (D.10),

$$
\begin{aligned}
&P\left\{\sup_{||u||=M_\epsilon} \Delta_n(u) < 0\right\} = Pr\left\{\sup_{||u||=M_\epsilon} \frac{\Delta_n(u)}{(1+b_n)^2} < 0\right\} \\
&\geq P\left\{\sup_{||u||=M_\epsilon} \left[\frac{1}{1+b_n}O_p(1)||u|| - \frac{1}{2}\eta_{min}||u||^2\{1+o_p(1)\} + \frac{b_n}{1+b_n}\sqrt{Kt}||u|| + o(1)\right] < 0\right\} \\
&= P\left\{\sup_{||u||=M_\epsilon} \left[\frac{1}{1+b_n}O_p(1)||u|| + \frac{b_n}{1+b_n}\sqrt{Kt}||u|| + o(1)\right] < \frac{1}{2}\eta_{min}||u||^2\{1+o_p(1)\}\right\} \\
&= P\left\{\sup_{M_\epsilon} \left[\frac{1}{1+b_n}O_p(1) + \frac{b_n}{1+b_n}\sqrt{Kt} + o(1)\right] < \frac{1}{2}\eta_{min}M_\epsilon\{1+o_p(1)\}\right\} \\
&= P\left\{\sup_{M_\epsilon} O_p(1) < \frac{1}{2}\eta_{min}M_\epsilon\{1+o_p(1)\}\right\} \\
&\geq 1-\epsilon
\end{aligned}
$$

for sufficiently large M_ϵ and n. Therefore, for any given $\epsilon > 0$, there exists a sufficiently large M_ϵ such that

$$
\lim_{m\to\infty} P\left\{\sup_{||u||=M_\epsilon} \ell^\sharp_{n\lambda}(\boldsymbol{\Theta}_0 + r_n u) - \ell^\sharp_{n\lambda}(\boldsymbol{\Theta}_0) < 0\right\} \geq 1-\epsilon
$$

which completes the proof. □

Theorem 3 states that when b_n is $O(1)$, there exists a local maximiser, $\hat{\boldsymbol{\Theta}}_n$, of the penalised likelihood function (4.1) which convergence to $\boldsymbol{\Theta}_0$ at a $\sqrt{n}$ rate. By choosing the value of the tuning parameter λ_{nk} carefully this can be achieved for the three penalty functions considered in this thesis. More specifically, if $\lambda_{nk} = O(n^{-1/2})$ for the LASSO, and ALASSO penalties, and $\lambda_{nk} \to 0$ for the SCAD penalty, then $b_n = O(1)$ for all three penalties, and $\sqrt{n}$ convergence can be achieved.

Theorem 4. *Assume that the observed data follow a FinMix GLMM satisfying the MPLE regularity conditions 1, 2, and 3. Assume also that the penalty function $p_{nk\lambda}$ satisfies penalty conditions 1, 2, and 3. Let the number of subpopulations K be known* *a priori. Then for any* $\boldsymbol{\Theta}$ *such that* $||\boldsymbol{\Theta} - \boldsymbol{\Theta}_0|| = O(n^{-1/2})$*, with probability tending to* 1*,*

$$\ell^{\sharp}_{n\lambda}\{(\boldsymbol{\Theta}^{\dagger}, \boldsymbol{\Theta}^{\ddagger})\} < p\ell^{\sharp}_{n\lambda}\{(\boldsymbol{\Theta}^{\dagger}, 0)\}.$$

Proof. Partition $\boldsymbol{\Theta} = (\boldsymbol{\Theta}^{\dagger}, \boldsymbol{\Theta}^{\ddagger})$ for any $\boldsymbol{\Theta}$ in the neighbourhood $||\boldsymbol{\Theta} - \boldsymbol{\Theta}_0|| = O(m^{-1/2})$. By the definition of $\ell^{\sharp}_{n\lambda}(\boldsymbol{\Theta})$

$$\begin{aligned}
&\ell^{\sharp}_{n\lambda}\{(\boldsymbol{\Theta}^{\dagger}, \boldsymbol{\Theta}^{\ddagger})\} - \ell^{\sharp}_{n\lambda}\{(\boldsymbol{\Theta}^{\dagger}, 0)\} \\
&\quad = [\ell_n\{(\boldsymbol{\Theta}^{\dagger}, \boldsymbol{\Theta}^{\ddagger})\} - \ell_n\{(\boldsymbol{\Theta}^{\dagger}, 0)\}] - [p_{n\lambda}\{(\boldsymbol{\Theta}^{\dagger}, \boldsymbol{\Theta}^{\ddagger})\} - p_{n\lambda}\{(\boldsymbol{\Theta}^{\dagger}, 0)\}].
\end{aligned}$$

Focusing on the first two terms, by the mean value theorem, there exists some ϵ such that $||\varepsilon|| < ||\boldsymbol{\Theta}^{\ddagger}|| = O(n^{-1/2})$, and

$$\ell_n\{(\boldsymbol{\Theta}^{\dagger}, \boldsymbol{\Theta}^{\ddagger})\} - \ell_n\{(\boldsymbol{\Theta}^{\dagger}, 0)\} = \left[\frac{\partial \ell_n\{(\boldsymbol{\Theta}^{\dagger}, \varepsilon)\}}{\partial \boldsymbol{\Theta}^{\ddagger}}\right]^{\top} \boldsymbol{\Theta}^{\ddagger}. \tag{D.11}$$

Then,

$$\begin{aligned}
||\frac{\partial \ell_n\{(\boldsymbol{\Theta}^{\dagger}, \varepsilon)\}}{\partial \boldsymbol{\Theta}^{\ddagger}} - \frac{\partial \ell_n\{(\boldsymbol{\Theta}^{\dagger}_0, 0)\}}{\partial \boldsymbol{\Theta}^{\ddagger}}|| \;\leq\; & ||\frac{\partial \ell_n\{(\boldsymbol{\Theta}^{\dagger}, \varepsilon)\}}{\partial \boldsymbol{\Theta}^{\ddagger}} - \frac{\partial \ell_n\{(\boldsymbol{\Theta}_1, 0)\}}{\partial \boldsymbol{\Theta}^{\ddagger}}|| \\
& + ||\frac{\partial \ell_n\{(\boldsymbol{\Theta}^{\dagger}, 0)\}}{\partial \boldsymbol{\Theta}^{\ddagger}} - \frac{\partial \ell_n\{(\boldsymbol{\Theta}^{\dagger}_0, 0)\}}{\partial \boldsymbol{\Theta}^{\ddagger}}||. \qquad \text{(D.12)}
\end{aligned}$$

But by the mean value theorem,

$$\frac{\partial \ell_n\{(\boldsymbol{\Theta}^{\dagger}, \varepsilon)\}}{\partial \boldsymbol{\Theta}^{\ddagger}} - \frac{\partial \ell_n\{(\boldsymbol{\Theta}_1, 0)\}}{\partial \boldsymbol{\Theta}^{\ddagger}} = \left[\frac{\partial^2 \ell_n\{(\boldsymbol{\Theta}^{\dagger}, \zeta_1)\}}{\partial \boldsymbol{\Theta}^{\ddagger\top} \partial \boldsymbol{\Theta}^{\ddagger}}\right] \times \varepsilon$$

for some $||\zeta_1|| \leq ||\varepsilon||$, and

$$\frac{\partial \ell_n\{(\boldsymbol{\Theta}^{\dagger}, 0)\}}{\partial \boldsymbol{\Theta}^{\ddagger}} - \frac{\partial \ell_n\{(\boldsymbol{\Theta}^{\dagger}_0, 0)\}}{\partial \boldsymbol{\Theta}^{\ddagger}} = \left[\frac{\partial^2 \ell_n\{(\zeta_2, 0)\}}{\partial \boldsymbol{\Theta}^{\dagger\top} \partial \boldsymbol{\Theta}^{\ddagger}}\right] \times (\boldsymbol{\Theta}^{\dagger} - \boldsymbol{\Theta}_0)$$

where $\zeta_2 = \boldsymbol{\Theta}_0^\dagger + t \times (\boldsymbol{\Theta}^\dagger - \boldsymbol{\Theta}_0^\dagger)$, for some $t \in [0,1]$. Applying these results to (D.12) and using MPLE model condition 2 results in

$$\begin{aligned}
& \left\| \frac{\partial \ell_n\{(\boldsymbol{\Theta}^\dagger, \varepsilon)\}}{\partial \boldsymbol{\Theta}^\ddagger} - \frac{\partial \ell_n\{(\boldsymbol{\Theta}_0^\dagger, 0)\}}{\partial \boldsymbol{\Theta}^\ddagger} \right\| \\
\leq & \left[\sum_{i=2}^{n} M_{2i}(y_i, X_i, Z_i) \right] \times ||\varepsilon|| + \left[\sum_{i=1}^{n} M_{2i}(y_i, X_i, Zi) \right] \times ||\boldsymbol{\Theta}^\dagger - \boldsymbol{\Theta}_0^\dagger|| \\
= & \; O_p(n) \times \left(||\varepsilon|| + ||\boldsymbol{\Theta}^\dagger - \boldsymbol{\Theta}_0^\dagger|| \right) \\
= & \; O_p(n) \times \{ O(n^{-\frac{1}{2}}) + O(n^{-\frac{1}{2}}) \} \\
= & \; O_p(n^{\frac{1}{2}}).
\end{aligned}$$

By the regularity conditions

$$\frac{\partial \ell_n\{(\boldsymbol{\Theta}_0^\dagger, 0)\}}{\partial \boldsymbol{\Theta}^\ddagger} = O_p(n^{\frac{1}{2}}),$$

therefore

$$\frac{\partial \ell_n\{(\boldsymbol{\Theta}_1, \varepsilon)\}}{\partial \boldsymbol{\Theta}^\ddagger} = O_p(n^{\frac{1}{2}}).$$

Using this result on (D.11) provides

$$\ell_n\{(\boldsymbol{\Theta}^\dagger, \boldsymbol{\Theta}^\ddagger)\} - \ell_n\{(\boldsymbol{\Theta}^\dagger, 0)\} = O_p(\sqrt{n}) \sum_{k=1}^{K} \left(\sum_{h=t_{\beta_k}+1}^{p} |\beta_{kh}| + \sum_{h=t_{d_k}+1}^{q} |d_{kh}| \right),$$

where t_{β_k} and t_{d_k} are the numbers of true non-zero fixed and random effects in component k respectively. On the other hand,

$$p_{n\lambda}\{(\boldsymbol{\Theta}^\dagger, \boldsymbol{\Theta}^\ddagger)\} - p_{n\lambda}\{(\boldsymbol{\Theta}^\dagger, 0)\} = \sum_{k=1}^{K} \left(\sum_{h=t_{\beta_k}+1}^{p} \pi_k n p_{nk\lambda}(\beta_{kh}) + \sum_{h=t_{d_k}+1}^{q} \pi_k n p_{nk\lambda}(d_{kh}) \right).$$

Therefore,

$$\begin{aligned}
\ell^{\sharp}_{n\lambda} & \{(\boldsymbol{\Theta}^{\dagger},\boldsymbol{\Theta}^{\ddagger})\} - \ell^{\sharp}_{n\lambda}\{(\boldsymbol{\Theta}^{\dagger},0)\} \\
= & \sum_{k=1}^{K}\left[\sum_{h=t_{\beta_k}+1}^{p}\{|\beta_{kh}|O_p(\sqrt{n}) - \pi_k m p_{nk\lambda}(\beta_{kh})\} + \sum_{h=t_{d_k}+1}^{q}\{|d_{kh}|O_p(\sqrt{n}) - \pi_k m p_{nk\lambda}(d_{kh})\}\right] \\
= & \sum_{k=1}^{K}\left[\sum_{h=t_{\beta_k}+1}^{p} A_{kh} + \sum_{h=t_{d_k}+1}^{q} B_{kh}\right].
\end{aligned}$$

By penalty condition 3, both A_{kh}, and B_{kh} are less than 0 in probability. Therefore,

$$Pr\left[\ell^{\sharp}_{n\lambda}\{(\boldsymbol{\Theta}^{\dagger},\boldsymbol{\Theta}^{\ddagger})\} - \ell^{\sharp}_{n\lambda}\{(\boldsymbol{\Theta}^{\dagger},0)\} < 0\right] \xrightarrow{p} 1.$$

This completes the proof. □

Also called consistency in selection or identifying zeros, sparsity is defined as $P(\hat{\boldsymbol{\Theta}}_n = 0) \to 1$ as $n \to \infty$.

Theorem 5 (Sparsity of the MPLE)**.** *Assume that the observed data follow a FinMix GLMM satisfying the MPLE regularity conditions 1, 2, and 3. Assume also that the penalty function $p_{nk\lambda}$ satisfies penalty conditions 1, 2, and 3. Let the number of subpopulations K be known a priori. Then for any $\sqrt{n}$-consistent maximum penalised likelihood estimator $\hat{\boldsymbol{\Theta}}_n$ of $\boldsymbol{\Theta}$, $Pr\{\hat{\boldsymbol{\Theta}}_2 = 0\} \to 1$ as $m \to \infty$.*

Proof. Let $(\hat{\boldsymbol{\Theta}}_1, 0)$ be a maximiser of the penalised log-likelihood function $\ell^{\sharp}_{n\lambda}\{(\boldsymbol{\Theta}^{\dagger},0)\}$ which is regarded as a function of $\boldsymbol{\Theta}^{\dagger}$. It suffices to show that in the neighbourhood $||\boldsymbol{\Theta} - \boldsymbol{\Theta}_0|| = O(m^{-1/2})$, the difference $\ell^{\sharp}_{n\lambda}\{(\boldsymbol{\Theta}^{\dagger},\boldsymbol{\Theta}^{\ddagger})\} - \ell^{\sharp}_{n\lambda}\{(\hat{\boldsymbol{\Theta}}_1,0)\} < 0$ with probability tending to 1 as $m \to \infty$. Recall that

$$\begin{aligned}
\ell^{\sharp}_{n\lambda}\{(\boldsymbol{\Theta}^{\dagger}, \boldsymbol{\Theta}^{\ddagger})\} &- \ell^{\sharp}_{n\lambda}\{(\hat{\boldsymbol{\Theta}}^{\dagger}, 0)\} \\
&= [\ell^{\sharp}_{n\lambda}\{(\boldsymbol{\Theta}^{\dagger}, \boldsymbol{\Theta}^{\ddagger})\} - \ell^{\sharp}_{n\lambda}\{(\boldsymbol{\Theta}^{\dagger}, 0)\}] + [\ell^{\sharp}_{n\lambda}\{(\boldsymbol{\Theta}^{\dagger}, 0)\} - \ell^{\sharp}_{n\lambda}\{(\hat{\boldsymbol{\Theta}}_1, 0)\}] \\
&\leq \ell^{\sharp}_{n\lambda}\{(\boldsymbol{\Theta}^{\dagger}, \boldsymbol{\Theta}^{\ddagger})\} - \ell^{\sharp}_{n\lambda}\{(\boldsymbol{\Theta}^{\dagger}, 0)\} \\
&< 0,
\end{aligned}$$

with probability tending to 1 by the previous Theorem. This completes the proof.

□

Theorem 6 (Asymptotic Distribution of the MPLE)**.** *Assume that the observed data follow a FinMix GLMM model satisfying the MPLE regularity conditions 1, 2, and 3. Assume also that the penalty function $p_{nk\lambda}$ satisfies penalty conditions 1, 2, and 3. Let the number of subpopulations K be known a priori. Then for any $\sqrt{n}$-consistent maximum penalised likelihood estimator $\hat{\boldsymbol{\Theta}}_n$ of $\boldsymbol{\Theta}$,*

$$\sqrt{n}\left[\left\{I_1(\boldsymbol{\Theta}^{\dagger}_0) + \frac{p''_{n\lambda}(\boldsymbol{\Theta}^{\dagger}_0)}{n}\right\}(\hat{\boldsymbol{\Theta}}_1 - \boldsymbol{\Theta}^{\dagger}_0) + \frac{p'_{n\lambda}(\boldsymbol{\Theta}^{\dagger}_0)}{n}\right] \xrightarrow{d} Gaussian(\mu = 0, \boldsymbol{\Sigma} = I_1(\boldsymbol{\Theta}^{\dagger}_0)),$$

where $I_1(\boldsymbol{\Theta}^{\dagger}_0)$ is the Fisher information knowing that $\boldsymbol{\Theta}^{\ddagger} = 0$.

Proof. Consider $\ell^{\sharp}_{n\lambda}\{(\boldsymbol{\Theta}^{\dagger}, 0)\}$ as a function of $\boldsymbol{\Theta}^{\dagger}$. Using the same argument as in Theorem 1, there exists a $\sqrt{n}$-consistent local maximiser of this function, say $\hat{\boldsymbol{\Theta}}_1$, which satisfies

$$\frac{\partial \ell^{\sharp}_{n\lambda}(\hat{\boldsymbol{\Theta}}_n)}{\partial \boldsymbol{\Theta}^{\dagger}} = \left\{\frac{\partial \ell_n(\boldsymbol{\Theta})}{\partial \boldsymbol{\Theta}^{\dagger}} - \frac{\partial p_{n\lambda}(\boldsymbol{\Theta})}{\partial \boldsymbol{\Theta}^{\dagger}}\right\}_{\hat{\boldsymbol{\Theta}}_n = (\hat{\boldsymbol{\Theta}}^{\dagger}, 0)} = 0. \qquad \text{(D.13)}$$

Since $\hat{\mathbf{\Theta}}_1$ is a $\sqrt{n}$-consistent estimator, using a Taylor's expansion around the true value

$$\frac{\partial \ell_n(\mathbf{\Theta})}{\partial \mathbf{\Theta}^\dagger}|_{\hat{\mathbf{\Theta}}_n=(\hat{\mathbf{\Theta}}^\dagger,0)} = \frac{\partial \ell_n(\mathbf{\Theta}_0^\dagger)}{\partial \mathbf{\Theta}^\dagger} + \left\{\frac{\partial^2 \ell_n(\mathbf{\Theta}_0^\dagger)}{\partial \mathbf{\Theta}^{\dagger\top}\partial \mathbf{\Theta}^\dagger} + o_p(m)\right\}(\hat{\mathbf{\Theta}}_1 - \mathbf{\Theta}_0^\dagger)$$

and

$$\frac{\partial p_{n\lambda}(\mathbf{\Theta})}{\partial \mathbf{\Theta}^\dagger}|_{\hat{\mathbf{\Theta}}_n=(\hat{\mathbf{\Theta}}^\dagger,0)} = p'_{n\lambda}(\mathbf{\Theta}_0^\dagger) + \left\{p''_{n\lambda}(\mathbf{\Theta}_0^\dagger) + o_p(m)\right\}(\hat{\mathbf{\Theta}}_1 - \mathbf{\Theta}_0^\dagger).$$

Substituting these into (D.13) results in

$$\left\{\frac{\partial \ell_n(\mathbf{\Theta}_0^\dagger)}{\partial \mathbf{\Theta}^\dagger} - p'_{n\lambda}(\mathbf{\Theta}_0^\dagger)\right\} + \left\{\frac{\partial^2 \ell_n(\mathbf{\Theta}_0^\dagger)}{\partial \mathbf{\Theta}^{\dagger\top}\partial \mathbf{\Theta}^\dagger} - p''_{n\lambda}(\mathbf{\Theta}_0^\dagger) + o_p(m)\right\}(\hat{\mathbf{\Theta}}_1 - \mathbf{\Theta}_0^\dagger) = 0.$$

Rearranging the terms and multiplying both sides by $\frac{1}{\sqrt{n}}$ provides

$$\frac{-\sqrt{n}}{n}\left\{\frac{\partial^2 \ell_n(\mathbf{\Theta}_0^\dagger)}{\partial \mathbf{\Theta}^{\dagger\top}\partial \mathbf{\Theta}^\dagger} - p''_{n}(\mathbf{\Theta}_0^\dagger) + o_p(n)\right\}\left(\hat{\mathbf{\Theta}}_1 - \mathbf{\Theta}_0^\dagger\right) = \frac{1}{\sqrt{n}}\left\{\frac{\partial \ell_n(\mathbf{\Theta}_0^\dagger)}{\partial \mathbf{\Theta}^\dagger} - p'_{n\lambda}(\mathbf{\Theta}_0^\dagger)\right\}.$$

Then, by the regularity conditions,

$$-\frac{1}{m}\frac{\partial^2 \ell_n(\mathbf{\Theta}_0^\dagger)}{\partial \mathbf{\Theta}^{\dagger\top}\partial \mathbf{\Theta}^\dagger} = I_1(\mathbf{\Theta}_0^\dagger) + o_p(1),$$

and

$$\frac{1}{\sqrt{n}}\frac{\partial \ell_n(\mathbf{\Theta}_0^\dagger)}{\partial \mathbf{\Theta}^\dagger} \to^d Gaussian(\mu = 0, \mathbf{\Sigma} = I_1(\mathbf{\Theta}_0^\dagger)).$$

Thus, by Slutsky's theorem,

$$\sqrt{n}\left[\left\{I_1(\mathbf{\Theta}_0^\dagger) + \frac{p''_{n\lambda}(\mathbf{\Theta}_0^\dagger)}{m}\right\}\left(\hat{\mathbf{\Theta}}_1 - \mathbf{\Theta}_0^\dagger\right) + \frac{p'_{n\lambda}(\mathbf{\Theta}_0^\dagger)}{m}\right] \xrightarrow{d} Gaussian(\mu = 0, \mathbf{\Sigma} = I_1(\mathbf{\Theta}_0^\dagger)).$$

This ends the proof. □

For the ALASSO, and SCAD penalties, sparsity can be achieved while maintaining root-n consistency, for a suitable choice of tuning parameters. For example, if I let $\lambda_{mk} = O(n^{-\frac{1}{2}})$ for the ALASSO penalty, and $\lambda_{nk} \to 0$, and $\sqrt{n}\lambda_{nk} \to \infty$ for the SCAD penalty, then root-n consistency, and sparsity can be achieved concurrently. This is, however, not true for the LASSO penalty. For LASSO, $b_n = \sqrt{n}\lambda_{nk}$. Therefore, root-$n$ consistency requires that $\sqrt{n}\lambda_{nk} = O(1)$. On the other hand, the sparsity property requires penalty assumption 2, which includes the condition that $\sqrt{n}\lambda_{nk} \to \infty$. These two requirements cannot be simultaneously satisfied.

Theorem 7 (Oracle Property of the MPLE)**.** *A FinMix GLMM has the oracle property if it asymptotically identifies the right subset model and has the optimal estimation rate.*

Proof. Given sparsity and that the distribution is asymptotically Gaussian, the oracle property follows (Fan and Li, 2012). □

APPENDIX E
Additional Tables From Chapter 3 (Bias, Variance, Mean Squared Error)

This appendix contains the tables of simulation results for Chapter 3.

Table E–1: Simulation 1 results multiplied by 100, averaged over 100 runs. Outcome follows a Poisson distribution with $K = 2$, $p = 2$, $q = 1$, $\pi_1 = 0.6$, $\pi_2 = 0.4$, $\boldsymbol{\beta}_1^\top = (-0.75, 0.35)$, $\boldsymbol{\beta}_2^\top = (0.60, -0.50)$, $\mathbb{\Gamma}_1^{*\top} = (0.80)$, and $\mathbb{\Gamma}_2^{*\top} = (0.25)$.

n	$n = 100$			$n = 250$			$n = 500$			$n = 1000$		
Measure	Bias	Var	MSE	Bias	Var	MSE	Bias	Var	MSE	Bias	Var	MSE
π_1	-4.3	1.3	1.5	-2.7	0.7	0.7	-2.4	0.4	0.5	-1.4	0.3	0.3
β_{10}	-3.6	10.4	10.4	-3.3	2.9	3.0	-1.8	1.4	1.4	-0.9	0.8	0.8
β_{11}	4.0	3.3	3.4	2.9	0.9	1.0	1.6	0.5	0.5	0.7	0.2	0.2
Γ_{11}	-2.9	3.2	3.3	-0.5	0.8	0.8	0.0	0.5	0.4	0.4	0.2	0.2
β_{20}	-6.1	5.5	5.8	-2.7	2.3	2.3	-3.7	1.4	1.5	-1.7	0.7	0.7
β_{21}	3.3	3.0	3.1	1.8	1.3	1.3	2.2	0.7	0.7	1.2	0.3	0.3
Γ_{21}	-0.3	2.1	2.1	0.8	1.1	1.1	1.5	0.8	0.8	0.5	0.4	0.4
Total			29.6			10.2			5.9			2.9

Table E–2: Simulation 2 results multiplied by 100, averaged over 100 runs. Outcome follows a Poisson distribution with $K = 2$, $p = 5$, $q = 1$, $\pi_1 = 0.6$, $\pi_2 = 0.4$, $\boldsymbol{\beta}_1^\top = (-0.75, 0.35, 0.10, -0.40, 0.00)$, $\boldsymbol{\beta}_2^\top = (0.60, -0.50, -0.35, -0.15, 0.00)$, $\mathbb{\Gamma}_1^{*\top} = (0.80)$, and $\mathbb{\Gamma}_2^{*\top} = (0.25)$.

n	$n = 100$			$n = 250$			$n = 500$			$n = 1000$		
Measure	Bias	Var	MSE	Bias	Var	MSE	Bias	Var	MSE	Bias	Var	MSE
π_1	-3.7	1.2	1.4	-2.1	0.6	0.6	-2.6	0.3	0.4	-2.2	0.2	0.2
β_{10}	-5.7	8.4	8.6	-4.5	3.0	3.2	-4.8	1.3	1.5	-3.5	1.1	1.2
β_{11}	3.4	3.8	3.9	2.4	0.9	0.9	2.3	0.4	0.5	1.6	0.5	0.6
β_{12}	-0.1	2.9	2.9	1.5	1.1	1.1	1.9	0.5	0.5	1.3	0.2	0.2
β_{13}	-1.0	0.9	0.9	-0.5	0.2	0.2	-0.9	0.1	0.1	-0.7	0.1	0.1
β_{14}	2.0	0.8	0.8	0.2	0.3	0.3	0.1	0.1	0.1	0.0	< 0.1	< 0.1
Γ_{11}	-3.2	3.1	3.2	0.5	0.8	0.8	1.6	0.5	0.5	0.8	0.2	0.2
β_{20}	-5.4	5.3	5.6	-3.5	1.8	1.9	-2.7	0.8	0.9	-2.3	0.4	0.4
β_{21}	2.0	2.7	2.7	1.3	0.8	0.8	1.8	0.4	0.5	2.4	0.5	0.5
β_{12}	2.4	2.1	2.1	1.3	0.8	0.8	1.4	0.3	0.3	1.8	0.2	0.2
β_{13}	-0.4	0.5	0.5	0.3	0.2	0.2	0.0	0.1	0.1	-0.5	0.1	0.1
β_{14}	-0.5	0.4	0.4	-0.4	0.2	0.2	-0.2	0.1	0.1	0.0	< 0.1	< 0.1
Γ_{21}	0.8	2.9	2.9	0.1	1.2	1.2	1.3	0.4	0.4	1.8	0.3	0.3
Total			35.8			12.0			5.8			4.2

Table E–3: Simulation 3 results multiplied by 100, averaged over 100 runs. Outcome follows a Poisson distribution with $K = 2$, $p = 5$, $q = 2$, $\pi_1 = 0.6$, $\pi_2 = 0.4$, $\boldsymbol{\beta}_1^\top = (-0.75, 0.35, 0.10, -0.40, 0.00)$, $\boldsymbol{\beta}_2^\top = (0.60, -0.50, -0.35, -0.15, 0.00)$, $\mathbb{\Gamma}_1^{*\top} = (0.80, -0.15, 0.20)$, and $\mathbb{\Gamma}_2^{*\top} = (0.25, 0.00, 0.30)$.

n	$n = 100$			$n = 250$			$n = 500$			$n = 1000$		
Measure	Bias	Var	MSE	Bias	Var	MSE	Bias	Var	MSE	Bias	Var	MSE
π_1	-5.2	1.4	1.7	-5.3	0.9	1.1	-4.0	0.5	0.7	-2.8	0.4	0.4
β_{10}	-4.3	13.8	13.9	-10.3	7.1	8.1	-7.5	3.1	3.6	-5.4	1.8	2.1
β_{11}	-2.3	6.0	6.0	4.8	2.8	3.0	4.3	1.4	1.6	2.8	0.8	0.9
β_{12}	-0.7	5.2	5.1	3.1	2.1	2.2	3.2	0.9	1.0	2.0	0.4	0.4
β_{13}	-1.1	1.4	1.4	-1.7	0.4	0.4	-1.7	0.2	0.3	-1.2	0.1	0.1
β_{14}	1.3	0.8	0.8	0.1	0.3	0.3	0.2	0.1	0.1	0.1	0.1	0.1
Γ_{11}	-2.3	8.9	8.9	1.8	4.7	4.7	1.7	2.1	2.2	1.7	0.8	0.9
Γ_{112}	-2.7	5.1	5.2	-2.3	2.4	2.4	-1.1	1.3	1.3	-0.3	0.7	0.7
Γ_{12}	-7.0	1.9	2.3	-5.4	1.2	1.5	-3.0	1.0	1.1	-2.5	0.5	0.6
β_{20}	-15.6	6.3	8.6	-6.5	2.3	2.7	-4.4	1.1	1.3	-3.3	0.6	0.7
β_{21}	9.9	5.6	6.5	5.6	1.9	2.2	2.7	0.9	0.9	2.2	0.5	0.5
β_{12}	4.2	3.0	3.2	2.0	0.8	0.8	1.4	0.4	0.4	0.6	0.3	0.3
β_{13}	-1.8	1.1	1.2	-1.7	0.3	0.4	-0.9	0.1	0.2	-0.6	0.1	0.1
β_{14}	-1.4	0.7	0.7	0.4	0.2	0.2	-0.2	0.1	0.1	0.3	< 0.1	< 0.1
Γ_{21}	2.6	6.5	6.5	1.1	2.1	2.1	1.1	1.3	1.3	0.7	0.7	0.7
Γ_{212}	-0.1	8.9	8.8	2.0	3.6	3.6	2.0	2.5	2.5	1.4	1.6	1.6
Γ_{22}	-11.8	1.8	3.2	-5.5	1.2	1.5	-4.4	0.7	0.9	-3.4	0.5	0.6
Total			83.9			37.2			19.5			10.6

Table E–4: Simulation 4 results multiplied by 100, averaged over 100 runs. Outcome follows a Poisson distribution with $K = 3$, $p = 2$, $q = 1$, $\pi_1 = 0.5$, $\pi_2 = 0.3$, $\pi_3 = 0.2$, $\boldsymbol{\beta}_1^\top = (-0.75, 0.35)$, $\boldsymbol{\beta}_2^\top = (0.60, -0.50)$, $\boldsymbol{\beta}_3^\top = (0.45, 0.75)$, $\mathbb{\Gamma}_1^{*\top} = (0.80)$, $\mathbb{\Gamma}_2^{*\top} = (0.25)$, and $\mathbb{\Gamma}_3^{*\top} = (0.40)$.

n	$n = 100$			$n = 250$			$n = 500$			$n = 1000$		
Measure	Bias	Var	MSE	Bias	Var	MSE	Bias	Var	MSE	Bias	Var	MSE
π_1	-7.0	1.5	2.0	-3.0	0.9	1.0	-1.6	0.6	0.6	-3.0	0.5	0.6
π_2	5.0	1.0	1.2	2.1	0.5	0.5	0.9	0.3	0.3	2.0	0.2	0.2
β_{10}	-10.9	18.7	19.7	-9.3	6.8	7.5	-6.4	4.7	5.0	-8.5	3.6	4.3
β_{11}	6.1	6.5	6.8	3.4	3.6	3.7	2.0	1.4	1.4	3.1	0.6	0.7
Γ_{11}	-10.7	7.7	8.8	-5.6	3.7	4.0	-6.2	2.4	2.7	-6.6	2.2	2.6
β_{20}	-9.8	7.0	7.9	-2.8	2.3	2.4	-1.9	1.1	1.1	-3.0	1.1	1.2
β_{21}	6.5	3.4	3.8	2.4	1.8	1.8	2.1	0.8	0.8	3.0	0.6	0.7
Γ_{21}	-1.9	2.8	2.8	-2.8	1.3	1.3	-2.4	0.6	0.6	-0.5	0.3	0.3
β_{30}	-0.3	6.7	6.6	0.2	2.3	2.3	0.5	1.6	1.6	1.5	0.6	0.6
β_{31}	-4.7	3.0	3.2	-4.0	2.5	2.7	-1.2	0.3	0.3	-2.6	1.0	1.0
Γ_{31}	-3.9	2.4	2.6	-1.4	1.1	1.1	-1.0	0.8	0.8	-0.4	0.3	0.3
Total			65.2			28.3			15.3			12.6

Table E–5: Simulation 5 results multiplied by 100, averaged over 100 runs. Outcome follows a Poisson distribution with $K = 3$, $p = 5$, $q = 1$, $\pi_1 = 0.5$, $\pi_2 = 0.3$, $\pi_3 = 0.2$, $\boldsymbol{\beta}_1^\top = (-0.75, 0.35, 0.10, -0.40, 0.00)$, $\boldsymbol{\beta}_2^\top = (0.60, -0.50, -0.35, -0.15, 0.00)$, $\boldsymbol{\beta}_3^\top = (0.45, 0.75, -0.65, 0.20, 0.00)$, $\mathbb{\Gamma}_1^{*\top} = (0.80)$, $\mathbb{\Gamma}_2^{*\top} = (0.25)$, and $\mathbb{\Gamma}_3^{*\top} = (0.40)$.

n	$n = 100$			$n = 250$			$n = 500$			$n = 1000$		
Measure	Bias	Var	MSE	Bias	Var	MSE	Bias	Var	MSE	Bias	Var	MSE
π_1	-4.9	1.5	1.7	-3.1	0.7	0.8	-1.3	0.5	0.5	-0.8	0.1	0.1
π_2	5.0	1.5	1.7	3.6	0.6	0.7	1.8	0.5	0.5	1.2	0.1	0.1
β_{10}	-11.8	12.1	13.4	-6.7	4.6	5.0	-2.5	1.8	1.9	-3.0	0.7	0.8
β_{11}	4.8	5.4	5.6	4.4	1.5	1.7	2.1	0.8	0.8	2.0	0.2	0.3
β_{12}	-2.0	7.8	7.8	0.7	1.7	1.7	-0.6	1.1	1.0	0.5	0.3	0.3
β_{13}	0.0	2.7	2.7	0.1	0.9	0.9	0.3	0.4	0.4	-0.2	0.1	0.1
β_{14}	-1.2	0.9	0.9	-0.6	0.3	0.3	-0.5	0.1	0.1	-0.5	< 0.1	< 0.1
Γ_{11}	-2.1	4.8	4.8	1.3	1.0	1.0	1.4	0.4	0.4	1.2	0.2	0.2
β_{20}	-9.3	7.6	8.4	-5.5	4.2	4.4	-2.5	1.8	1.8	-1.0	0.5	0.5
β_{21}	7.6	7.8	8.3	3.7	2.2	2.3	2.8	1.7	1.7	1.4	0.3	0.3
β_{12}	4.0	4.6	4.7	2.8	1.1	1.2	2.4	0.6	0.6	1.3	0.2	0.2
β_{13}	-0.8	1.4	1.4	-1.4	0.5	0.5	-0.1	0.3	0.3	-0.3	0.1	0.1
β_{14}	0.4	0.7	0.7	0.2	0.2	0.2	-0.1	0.1	0.1	0.0	0.1	0.1
Γ_{21}	2.5	4.3	4.3	3.8	2.7	2.8	1.6	0.8	0.8	0.4	0.2	0.2
β_{30}	0.9	2.7	2.7	1.2	0.8	0.8	0.8	0.4	0.4	0.6	0.2	0.2
β_{31}	0.8	0.8	0.8	-0.2	0.2	0.2	0.2	0.1	0.1	0.2	< 0.1	< 0.1
β_{32}	-1.6	0.8	0.8	-0.1	0.2	0.2	-0.3	0.1	0.1	-0.3	0.1	0.1
β_{33}	-0.2	0.2	0.2	0.0	< 0.1	< 0.1	0.0	< 0.1	< 0.1	0.1	< 0.1	< 0.1
β_{34}	-0.3	0.2	0.2	0.0	< 0.1	< 0.1	0.0	< 0.1	< 0.1	0.1	< 0.1	< 0.1
Γ_{31}	-3.1	0.8	0.9	-0.7	0.3	0.3	-1.1	0.1	0.2	-1.1	0.1	0.1
Total			71.8			25.0			12.0			3.8

Table E–6: Simulation 6 results multiplied by 100, averaged over 100 runs. Outcome follows a Poisson distribution with $K = 3$, $p = 5$, $q = 2$, $\pi_1 = 0.5$, $\pi_2 = 0.3$, $\pi_3 = 0.2$, $\boldsymbol{\beta}_1^\top = (-0.75, 0.35, 0.10, -0.40, 0.00)$, $\boldsymbol{\beta}_2^\top = (0.60, -0.50, -0.35, -0.15, 0.00)$, and $\boldsymbol{\beta}_3^\top = (0.45, 0.75, -0.65, 0.20, 0.00)$, $\mathbb{\Gamma}_1^{*\top} = (0.80, -0.15, 0.20)$, $\mathbb{\Gamma}_2^{*\top} = (0.25, 0.00, 0.30)$, and $\mathbb{\Gamma}_3^{*\top} = (0.40, 0.25, 0.10)$.

n	$n = 100$			$n = 250$			$n = 500$			$n = 1000$		
Measure	Bias	Var	MSE	Bias	Var	MSE	Bias	Var	MSE	Bias	Var	MSE
π_1	-4.3	1.7	1.9	-0.7	0.9	0.9	0.1	0.5	0.5	1.1	0.4	0.4
π_2	5.9	1.7	2.0	2.8	1.0	1.0	1.7	0.6	0.6	1.0	0.4	0.4
β_{10}	-4.9	13.4	13.5	-0.4	5.3	5.3	0.9	2.2	2.2	1.8	1.9	1.9
β_{11}	2.1	6.8	6.7	1.3	2.9	2.9	-0.4	1.2	1.2	-1.7	0.8	0.8
β_{12}	1.1	4.2	4.2	0.2	0.7	0.7	-0.1	0.3	0.3	-0.6	0.2	0.2
β_{13}	-0.3	0.7	0.7	0.1	0.3	0.2	0.1	0.1	0.1	0.2	0.1	0.1
β_{14}	-1.0	0.4	0.4	-0.3	0.1	0.1	0.0	0.1	0.1	0.0	< 0.1	< 0.1
Γ_{11}	-4.3	4.9	5.1	-2.0	1.9	1.9	-2.1	0.8	0.8	-1.2	0.4	0.4
Γ_{112}	-0.7	2.9	2.9	-0.7	0.9	0.9	0.1	0.5	0.5	-0.3	0.2	0.2
Γ_{12}	-9.3	1.6	2.4	-4.3	1.2	1.4	-2.8	0.7	0.8	-0.2	0.3	0.3
β_{20}	-11.6	9.0	10.2	-3.1	3.4	3.4	-2.4	2.6	2.7	0.1	1.4	1.4
β_{21}	17.1	7.9	10.7	9.2	3.6	4.4	7.6	2.1	2.7	6.1	1.9	2.2
β_{12}	4.2	4.6	4.8	-0.7	1.7	1.7	-0.1	1.1	1.1	-2.0	0.5	0.6
β_{13}	-0.3	1.9	1.9	0.7	0.9	0.9	1.1	0.5	0.5	2.4	0.3	0.4
β_{14}	0.0	1.1	1.1	0.3	0.3	0.3	0.3	0.2	0.2	-0.2	0.1	0.1
Γ_{21}	1.3	6.1	6.0	-5.5	4.3	4.6	-0.8	2.8	2.8	-1.5	2.1	2.1
Γ_{212}	0.1	7.6	7.5	5.4	4.8	5.0	-0.8	3.7	3.7	0.0	2.7	2.6
Γ_{22}	-5.4	4.7	4.9	-0.6	1.6	1.6	0.7	1.3	1.2	2.3	0.8	0.8
β_{30}	5.0	7.0	7.2	5.3	3.1	3.4	5.5	1.2	1.5	3.8	0.9	1.0
β_{31}	-0.2	7.6	7.5	2.6	2.7	2.7	3.6	1.3	1.4	3.8	0.5	0.6
β_{32}	-2.9	2.1	2.2	-1.4	0.5	0.5	-1.0	0.2	0.2	-0.7	0.1	0.1
β_{33}	-0.1	0.7	0.7	0.9	0.1	0.1	0.6	< 0.1	< 0.1	0.6	< 0.1	< 0.1
β_{34}	0.1	0.3	0.3	0.2	0.1	0.1	0.2	< 0.1	< 0.1	0.1	< 0.1	< 0.1
Γ_{31}	-5.8	4.6	4.9	0.5	2.1	2.1	-0.3	0.6	0.6	0.0	0.3	0.3
Γ_{312}	-9.2	2.9	3.7	-8.3	0.8	1.5	-7.7	0.6	1.1	-6.7	0.2	0.7
Γ_{32}	8.5	3.0	3.7	7.9	1.8	2.4	8.7	1.3	2.0	8.6	0.4	1.1
Total			117.3			50.1			28.8			18.8

Table E–7: Simulation 7 results multiplied by 100, averaged over 100 runs. Outcome follows a binomial distribution with $K = 2$, $p = 2$, $q = 1$, $\pi_1 = 0.6$, $\pi_2 = 0.4$, $\boldsymbol{\beta}_1^\top = (-0.55, 0.85)$, $\boldsymbol{\beta}_2^\top = (0.25, -0.50)$, $\mathbb{\Gamma}_1^{*\top} = (1.60)$, and $\mathbb{\Gamma}_2^{*\top} = (1.05)$.

n	$n = 100$			$n = 250$			$n = 500$			$n = 1000$		
Measure	Bias	Var	MSE	Bias	Var	MSE	Bias	Var	MSE	Bias	Var	MSE
π_1	-0.7	1.0	1.0	0.4	0.3	0.3	0.1	0.2	0.2	0.0	0.1	0.1
β_{10}	-0.4	8.9	8.8	1.6	3.1	3.1	2.7	1.6	1.7	2.0	0.8	0.8
β_{11}	2.7	3.1	3.1	0.3	1.1	1.1	-0.2	0.5	0.5	0.5	0.2	0.2
Γ_{11}	-6.5	5.4	5.8	-2.1	1.8	1.8	0.5	1.0	1.0	0.7	0.5	0.5
β_{20}	4.3	10.4	10.4	1.1	4.3	4.2	-2.1	2.0	2.1	-3.1	1.0	1.1
β_{21}	-2.4	3.6	3.6	-1.5	1.1	1.1	0.1	0.7	0.7	0.1	0.4	0.4
Γ_{21}	-3.0	7.1	7.2	-2.5	2.6	2.6	-2.4	1.4	1.4	-2.8	0.5	0.6
Total			39.8			14.4			7.5			3.8

Table E–8: Simulation 8 results multiplied by 100, averaged over 100 runs. Outcome follows a binomial distribution with $K = 2$, $p = 5$, $q = 1$, $\pi_1 = 0.6$, $\pi_2 = 0.4$, $\boldsymbol{\beta}_1^\top = (-0.55, 0.85, 1.25, -0.70, 0.00)$, $\boldsymbol{\beta}_2^\top = (0.25, -0.50, 1.35, -0.20, 0.00)$, $\mathbb{\Gamma}_1^{*\top} = (1.60)$, and $\mathbb{\Gamma}_2^{*\top} = (1.05)$.

n	$n = 100$			$n = 250$			$n = 500$			$n = 1000$		
Measure	Bias	Var	MSE	Bias	Var	MSE	Bias	Var	MSE	Bias	Var	MSE
π_1	-1.2	0.9	0.9	-0.7	0.4	0.4	-0.1	0.1	0.1	-0.1	0.1	0.1
β_{10}	-0.3	7.9	7.8	0.7	4.7	4.6	-1.8	1.9	1.9	-2.4	1.1	1.1
β_{11}	-1.4	3.0	3.0	0.8	1.2	1.2	0.4	0.5	0.5	0.7	0.2	0.2
β_{12}	0.3	2.2	2.2	0.5	0.7	0.7	-0.1	0.4	0.4	-0.2	0.2	0.2
β_{13}	0.9	0.8	0.8	0.4	0.3	0.3	0.5	0.1	0.1	0.5	0.1	0.1
β_{14}	0.2	0.6	0.6	0.8	0.2	0.2	0.6	0.1	0.1	0.0	< 0.1	< 0.1
Γ_{11}	-5.4	6.0	6.2	-2.9	2.2	2.3	-1.5	0.7	0.8	-1.2	0.4	0.4
β_{20}	1.8	9.7	9.6	-1.2	4.3	4.3	1.5	2.1	2.1	1.1	1.1	1.1
β_{21}	-0.7	3.2	3.2	-0.4	1.2	1.2	0.2	0.6	0.6	-0.1	0.4	0.3
β_{12}	2.2	2.7	2.7	0.8	1.3	1.3	0.4	0.6	0.6	-0.4	0.3	0.3
β_{13}	0.2	0.8	0.8	0.2	0.3	0.3	0.1	0.1	0.1	-0.1	0.1	0.1
β_{14}	-0.1	0.7	0.7	-0.3	0.3	0.3	-0.2	0.1	0.1	0.2	0.1	0.1
Γ_{21}	-0.5	6.9	6.9	0.8	2.1	2.1	-0.4	0.9	0.9	0.5	0.4	0.4
Total			45.5			19.1			8.4			4.4

Table E–9: Simulation 9 results multiplied by 100, averaged over 100 runs. Outcome follows a binomial distribution with $K = 2$, $p = 5$, $q = 2$, $\pi_1 = 0.6$, $\pi_2 = 0.4$, $\boldsymbol{\beta}_1^\top = (-0.55, 0.85, 1.25, -0.70, 0.00)$, $\boldsymbol{\beta}_2^\top = (0.25, -0.50, 1.35, -0.20, 0.00)$, $\mathbb{\Gamma}_1^{*\top} = (1.60, -0.45, 1.00)$, and $\mathbb{\Gamma}_2^{*\top} = (1.05, 0.00, 1.40)$.

n	$n = 100$			$n = 250$			$n = 500$			$n = 1000$		
Measure	Bias	Var	MSE	Bias	Var	MSE	Bias	Var	MSE	Bias	Var	MSE
π_1	-4.2	1.9	2.1	-4.1	0.9	1.0	-2.9	0.5	0.6	-2.4	0.3	0.4
β_{10}	-5.6	22.9	23.0	-6.6	6.8	7.1	-2.1	3.0	3.0	0.0	1.5	1.5
β_{11}	5.2	14.1	14.2	5.1	4.9	5.2	2.4	2.4	2.4	0.9	1.3	1.3
β_{12}	2.4	5.7	5.7	-0.7	1.6	1.6	-0.4	0.7	0.7	0.1	0.3	0.3
β_{13}	1.2	2.3	2.3	-1.3	0.6	0.6	-0.9	0.3	0.3	-1.3	0.1	0.1
β_{14}	-0.3	0.9	0.9	1.0	0.4	0.4	0.4	0.2	0.2	0.3	0.1	0.1
Γ_{11}	-1.7	24.2	24.0	-2.2	6.4	6.4	1.0	2.3	2.2	1.6	1.1	1.1
Γ_{112}	5.6	18.1	18.2	2.7	5.6	5.6	-1.3	2.7	2.7	-1.7	1.2	1.2
Γ_{12}	-13.8	9.7	11.5	-1.6	3.4	3.4	-1.2	1.3	1.3	-0.5	0.7	0.7
β_{20}	5.7	26.5	26.6	-1.3	7.4	7.3	-4.1	3.5	3.6	-3.9	2.0	2.1
β_{21}	-2.9	35.0	34.7	7.7	6.2	6.7	6.6	3.7	4.1	6.8	2.0	2.4
β_{12}	1.2	11.4	11.3	1.4	3.4	3.4	0.1	1.5	1.5	0.6	0.5	0.5
β_{13}	-5.6	3.2	3.5	-3.1	1.0	1.1	-2.1	0.5	0.5	-1.0	0.2	0.2
β_{14}	-0.1	1.7	1.6	-0.4	0.7	0.7	-0.2	0.3	0.3	0.0	0.1	0.1
Γ_{21}	-8.6	24.2	24.7	1.9	9.1	9.1	2.9	3.4	3.4	2.4	1.8	1.8
Γ_{212}	7.6	41.0	41.2	-2.0	17.3	17.1	-5.5	7.9	8.1	-4.7	4.3	4.4
Γ_{22}	-19.3	31.5	34.9	-9.3	5.8	6.6	-6.7	2.3	2.7	-3.3	1.2	1.3
Total			280.5			83.3			37.7			19.6

Table E–10: Simulation 10 results multiplied by 100, averaged over 100 runs. Outcome follows a binomial distribution with $K = 3$, $p = 2$, $q = 1$, $\pi_1 = 0.5$, $\pi_2 = 0.3$, $\pi_3 = 0.2$, $\boldsymbol{\beta}_1^\top = (-0.55, 0.85)$, $\boldsymbol{\beta}_2^\top = (0.25, -0.50)$, $\boldsymbol{\beta}_3^\top = (-0.75, 0.35)$, $\mathbb{\Gamma}_1^{*\top} = (1.60)$, $\mathbb{\Gamma}_2^{*\top} = (1.05)$, and $\mathbb{\Gamma}_3^{*\top} = (1.45)$.

n	$n = 100$			$n = 250$			$n = 500$			$n = 1000$		
Measure	Bias	Var	MSE	Bias	Var	MSE	Bias	Var	MSE	Bias	Var	MSE
π_1	-16.1	1.0	3.6	-17.0	1.0	3.9	-16.3	1.0	3.7	-17.0	0.9	3.8
π_2	-1.7	0.7	0.7	-1.1	0.7	0.7	-1.2	0.5	0.5	-1.0	0.3	0.3
β_{10}	17.5	29.4	32.2	16.2	21.5	23.9	14.7	14.8	16.8	14.3	9.1	11.1
β_{11}	9.4	7.3	8.1	9.2	4.4	5.2	5.8	2.3	2.6	6.1	1.3	1.7
Γ_{11}	-19.9	18.7	22.5	-16.6	11.2	13.9	-18.1	10.0	13.2	-14.3	6.4	8.4
β_{20}	16.9	28.6	31.2	11.0	11.5	12.6	9.2	5.1	5.9	9.5	2.9	3.8
β_{21}	-4.1	11.5	11.5	-2.2	4.8	4.8	-1.8	2.2	2.3	-1.1	0.6	0.6
Γ_{21}	-3.0	19.2	19.1	-1.4	8.1	8.0	-0.2	5.1	5.1	-0.7	1.7	1.7
β_{30}	-5.5	20.9	21.0	-1.0	15.9	15.7	-0.3	13.8	13.7	-2.9	11.5	11.5
β_{31}	11.0	12.7	13.8	14.9	7.2	9.4	14.9	6.0	8.2	15.0	4.3	6.5
Γ_{31}	-10.7	12.5	13.5	-4.0	10.0	10.0	-0.5	9.2	9.1	3.3	5.2	5.2
Total			177.1			108.2			80.8			54.6

Table E–11: Simulation 11 results multiplied by 100, averaged over 100 runs. Outcome follows a binomial distribution with $K = 3$, $p = 5$, $q = 1$, $\pi_1 = 0.5$, $\pi_2 = 0.3$, $\pi_3 = 0.2$, $\boldsymbol{\beta}_1^\top = (-0.55, 0.85, 1.25, -0.70, 0.00)$, $\boldsymbol{\beta}_2^\top = (0.25, -0.50, 1.35, -0.20, 0.00)$, $\boldsymbol{\beta}_3^\top = (-0.75, 0.35, -0.50, 0.55, 0.00)$, $\mathbb{\Gamma}_1^{*\top} = (1.60)$, $\mathbb{\Gamma}_2^{*\top} = (1.05)$, and $\mathbb{\Gamma}_3^{*\top} = (1.45)$.

n	$n = 100$			$n = 250$			$n = 500$			$n = 1000$		
Measure	Bias	Var	MSE	Bias	Var	MSE	Bias	Var	MSE	Bias	Var	MSE
π_1	-3.2	0.7	0.8	-1.2	0.3	0.4	-0.3	0.2	0.2	-0.2	0.1	0.1
π_2	2.4	0.8	0.8	0.7	0.3	0.3	0.2	0.1	0.1	0.3	0.1	0.1
β_{10}	3.7	14.9	14.9	4.7	4.9	5.0	1.7	2.1	2.1	-0.6	1.3	1.3
β_{11}	4.0	4.7	4.8	1.8	1.4	1.4	0.9	0.8	0.8	1.0	0.3	0.3
β_{12}	1.9	3.3	3.3	-0.3	0.9	0.9	-0.1	0.5	0.5	-0.5	0.2	0.2
β_{13}	-2.5	1.0	1.1	-1.9	0.3	0.3	-1.1	0.1	0.1	-0.5	0.1	0.1
β_{14}	-0.2	0.9	0.9	-0.6	0.2	0.2	-0.6	0.1	0.1	-0.3	< 0.1	< 0.1
Γ_{11}	-8.6	7.8	8.5	-4.2	2.6	2.8	-0.9	1.1	1.1	0.4	0.5	0.5
β_{20}	-2.5	19.9	19.7	1.1	7.8	7.7	3.9	3.0	3.1	4.8	1.6	1.9
β_{21}	-0.9	8.9	8.9	-1.2	2.6	2.6	-1.5	0.8	0.9	-1.1	0.4	0.4
β_{12}	0.5	4.3	4.3	0.3	2.1	2.1	0.6	0.8	0.8	0.5	0.4	0.4
β_{13}	-3.6	1.5	1.6	-1.1	0.4	0.4	-0.4	0.2	0.2	-0.2	0.1	0.1
β_{14}	2.4	1.7	1.7	0.2	0.4	0.4	0.3	0.1	0.1	0.0	0.1	0.1
Γ_{21}	-3.0	9.9	9.9	0.2	5.0	4.9	-1.0	1.9	1.9	-1.4	0.9	0.9
β_{30}	0.3	28.5	28.2	-2.2	9.9	9.9	-2.2	5.7	5.7	-1.1	2.8	2.8
β_{31}	-0.4	4.2	4.1	-1.3	2.0	2.0	0.1	0.9	0.9	0.3	0.6	0.5
β_{32}	1.2	5.3	5.2	2.7	2.5	2.5	1.8	1.2	1.2	0.8	0.5	0.5
β_{33}	0.4	1.3	1.3	0.7	0.7	0.7	0.7	0.2	0.3	0.4	0.1	0.1
β_{34}	-0.1	1.4	1.4	0.7	0.6	0.6	0.4	0.2	0.2	0.1	0.1	0.1
Γ_{31}	-17.2	17.8	20.6	-8.1	8.4	9.0	-5.8	3.4	3.8	-2.2	1.6	1.7
Total			142.1			54.1			24.1			12.2

Table E–12: Simulation 12 results multiplied by 100, averaged over 100 runs. Outcome follows a binomial distribution with $K = 3$, $p = 5$, $q = 2$, $\pi_1 = 0.5$, $\pi_2 = 0.3$, $\pi_3 = 0.2$, $\boldsymbol{\beta}_1^\top = (-0.55, 0.85, 1.25, -0.70, 0.00)$, $\boldsymbol{\beta}_2^\top = (0.25, -0.50, 1.35, -0.20, 0.00)$, $\boldsymbol{\beta}_3^\top = (-0.75, 0.35, -0.50, 0.55, 0.00)$, $\mathbb{\Gamma}_1^{*\top} = (1.60, -0.45, 1.00)$, $\mathbb{\Gamma}_2^{*\top} = (1.05, 0.00, 1.40)$, and $\mathbb{\Gamma}_3^{*\top} = (1.45, 0.40, 1.30)$.

n	$n = 100$			$n = 250$			$n = 500$			$n = 1000$		
Measure	Bias	Var	MSE	Bias	Var	MSE	Bias	Var	MSE	Bias	Var	MSE
π_1	-5.0	1.3	1.5	-2.1	0.8	0.8	-0.2	0.5	0.5	0.0	0.3	0.3
π_2	4.7	1.1	1.3	3.1	0.7	0.8	0.6	0.4	0.4	0.3	0.3	0.3
β_{10}	-13.1	15.4	16.9	0.3	6.3	6.2	0.7	2.9	2.9	0.6	1.7	1.6
β_{11}	12.6	16.6	18.0	4.7	5.5	5.7	3.8	2.8	3.0	3.3	1.6	1.7
β_{12}	0.0	7.2	7.1	2.7	1.5	1.6	0.9	0.7	0.7	1.2	0.4	0.4
β_{13}	1.1	4.0	4.0	-1.5	0.8	0.9	0.3	0.3	0.3	0.1	0.1	0.1
β_{14}	0.7	1.1	1.1	-0.5	0.5	0.5	-0.4	0.1	0.1	-0.6	0.1	0.1
Γ_{11}	-4.6	21.8	21.8	-1.9	6.9	6.9	-3.8	3.1	3.2	-2.2	1.4	1.4
Γ_{112}	6.6	23.1	23.3	0.8	8.7	8.7	1.1	3.8	3.8	-0.9	1.7	1.7
Γ_{12}	-15.0	12.7	14.8	-8.1	3.8	4.4	-4.4	1.3	1.5	-4.2	0.7	0.9
β_{20}	-3.0	25.9	25.7	-5.1	8.8	9.0	2.0	5.4	5.3	1.7	3.6	3.6
β_{21}	2.9	28.3	28.1	-1.6	11.4	11.3	-11.2	7.7	8.9	-9.4	4.2	5.0
β_{12}	1.3	10.3	10.2	-3.0	5.9	5.9	-0.2	2.3	2.3	0.0	1.1	1.1
β_{13}	-6.8	6.7	7.1	0.4	1.3	1.3	1.0	0.7	0.7	0.6	0.4	0.4
β_{14}	-0.3	2.2	2.2	1.1	1.1	1.1	1.0	0.5	0.5	0.6	0.2	0.2
Γ_{21}	-16.9	28.2	30.7	-2.2	15.0	14.9	-6.1	7.4	7.8	-4.8	4.0	4.2
Γ_{212}	30.5	39.0	48.0	20.3	24.0	27.9	15.9	14.1	16.5	16.4	6.1	8.8
Γ_{22}	-37.3	28.3	41.9	-19.0	8.6	12.2	-11.9	5.6	7.0	-9.1	2.7	3.5
β_{30}	1.8	39.0	38.7	4.8	11.7	11.8	1.4	5.6	5.6	0.1	3.2	3.2
β_{31}	0.9	54.4	53.9	1.7	13.6	13.5	2.7	5.3	5.3	1.9	2.7	2.7
β_{32}	-8.3	16.9	17.4	-3.1	4.0	4.0	-1.8	1.6	1.6	-2.0	0.7	0.8
β_{33}	0.5	3.0	3.0	0.9	1.0	1.0	0.6	0.4	0.4	0.5	0.2	0.2
β_{34}	-2.3	2.8	2.8	-0.3	0.8	0.8	-0.2	0.4	0.4	-0.4	0.2	0.2
Γ_{31}	-19.6	46.4	49.8	-12.1	14.3	15.6	2.0	8.8	8.7	1.8	3.9	3.9
Γ_{312}	9.9	56.6	57.0	1.3	20.7	20.5	-0.4	11.5	11.4	-2.3	5.5	5.5
Γ_{32}	-26.3	44.4	50.9	-6.1	22.0	22.1	-3.4	8.0	8.1	-0.2	3.6	3.6
Total			577.2			209.5			106.5			55.0

APPENDIX F
Supplement to Chapter 3

binomialThere are a number of special cases and further ideas regarding the behaviour of the MLE of a FinMix GLMM that are of interest. I explored a few of these possibilities in this Appendix. In particular, I considered fitting a model without random effects, exploring the behaviour of BIC for different values of K, different starting values, smaller values for n_i, and different values for m_{ij} (in the binomial case).

F.1 Fit a Model That Is More Complex Than the Truth

When the true value of a regression coefficient is 0, then the corresponding covariate does not contribute to the statistical model. As such, it is appropriate to remove such a variable from the model, making it simpler without sacrificing the performance of the model. This is the goal of the penalised maximum likelihood approach found in Chapter 4. However, in many cases, prior to optimising the penalised maximum likelihood equation, one must first optimise the unpenalised maximum likelihood equation. Therefore, it is important to confirm the performance of the algorithm when a model that is more complex than the truth is fit.

I explored the case where a particular fixed effect has a coefficient 0 but is still included in the model in Chapter 3. All models that contain 5 fixed effects contain at least one variable in each $\boldsymbol{\beta}_k$ that I set to 0. These results can be found in Tables E–2, E–3, E–5, E–6, E–8, E–9, E–11, and E–12. Similarly, all models that contain

2 random effects contain at least one covariance with a true value of 0. The results from these simulations are in Tables E–3, E–6, E–9, and E–12. Therefore, of interest here are the case where I estimated a random effect in the model unnecessarily (both when the coefficient for the corresponding fixed effect is zero and non-zero) and the case where I estimated more (or fewer) subpopulations than was in the model.

F.1.1 No Random Effect

In these simulations, I used a model with $K = 2$, $p = 5$, and $q = 2$. Tables E–3, and E–9 show similar cases. However, in this simulation, I set the coefficients for the fixed effects to $\boldsymbol{\beta}_1^\top = (-0.75, 0.35, 0.10, -0.40, 0.00)$, and $\boldsymbol{\beta}_2^\top = (0.60, 0.00, -0.35, -0.15, 0.00)$, and the random effects to $\mathbb{\Gamma}_1^{*\top} = (0.00, 0.00, 0.20)$, and $\mathbb{\Gamma}_2^{*\top} = (0.25, 0.00, 0.00)$.

These simulations show that the estimates for Gamma are reasonable even when the true value of Gamma is 0.

Turning to the binomial case, the parameters that have changed are $\boldsymbol{\beta}_1^\top = (-0.55, 0.85, 1.25, -0.70, 0.00)$, $\boldsymbol{\beta}_2^\top = (0.25, 0.00, 1.35, -0.20, 0.00)$, $\mathbb{\Gamma}_1^{*\top} = (0.00, 0.00, 1.00)$, and $\mathbb{\Gamma}_2^{*\top} = (1.05, 0.00, 0.00)$.

Similar to the case where the outcome is from a Poisson distribution, these simulations show that the estimates for Gamma are reasonable even when the true value of Gamma is 0.

F.1.2 Unknown Number of Subpopulations

Choosing the number of subpopulations can be done in various ways. One popular method is to fit models with a range of values for K and compare these models using information criteria such as AIC or BIC. As such, some simulations

Table F–1: Simulation 1 results multiplied by 100, averaged over 100 runs. Outcome follows a Poisson distribution with $K = 2$, $p = 5$, $q = 2$, $\pi_1 = 0.6$, $\pi_2 = 0.4$, $\boldsymbol{\beta}_1^\top = (-0.75, 0.35, 0.10, -0.40, 0.00)$, $\boldsymbol{\beta}_2^\top = (0.60, 0.00, -0.35, -0.15, 0.00)$, $\mathbb{\Gamma}_1^{*\top} = (0.00, 0.00, 0.20)$, and $\mathbb{\Gamma}_2^{*\top} = (0.25, 0.00, 0.00)$.

n	$n = 100$			$n = 250$			$n = 500$			$n = 1000$		
Measure	Bias	Var	MSE	Bias	Var	MSE	Bias	Var	MSE	Bias	Var	MSE
π_1	-2.2	1.2	1.2	-2.5	1.0	1.1	-2.6	1.0	1.1	-2.6	0.9	1.0
β_{10}	3.4	5.9	6.0	4.1	3.6	3.8	3.3	2.6	2.7	3.7	2.1	2.2
β_{11}	2.9	3.3	3.4	4.5	1.6	1.8	4.8	1.0	1.3	4.1	0.8	1.0
β_{12}	-6.5	4.0	4.4	-5.6	2.1	2.4	-4.6	1.9	2.1	-4.7	1.4	1.6
β_{13}	6.9	1.5	2.0	5.2	1.1	1.4	4.8	0.7	0.9	4.0	0.6	0.7
β_{14}	-2.7	1.6	1.6	-3.0	0.8	0.9	-2.5	0.5	0.5	-1.9	0.3	0.4
Γ_{11}	64.6	9.4	51.1	65.8	5.9	49.1	64.1	5.2	46.2	61.1	4.6	41.9
Γ_{112}	-19.2	3.6	7.2	-19.4	1.8	5.5	-18.6	1.4	4.9	-16.4	1.1	3.8
Γ_{12}	4.0	1.5	1.6	3.3	0.9	1.0	2.3	0.7	0.7	1.4	0.5	0.6
β_{20}	-20.8	5.3	9.6	-19.8	5.2	9.1	-20.3	5.0	9.1	-19.5	4.6	8.4
β_{21}	44.7	2.3	22.3	46.3	1.4	22.9	47.8	1.1	23.9	48.0	0.8	23.9
β_{12}	0.3	2.7	2.7	-0.1	1.2	1.2	0.1	0.9	0.9	-0.1	0.5	0.5
β_{13}	-0.3	0.8	0.8	-0.4	0.4	0.4	-0.3	0.3	0.3	-0.2	0.3	0.2
β_{14}	1.6	0.8	0.8	1.2	0.5	0.5	1.1	0.3	0.3	0.9	0.2	0.2
Γ_{21}	21.7	5.0	9.6	22.3	4.2	9.2	21.6	3.8	8.4	20.1	2.8	6.8
Γ_{212}	-12.3	1.9	3.4	-11.7	1.0	2.4	-10.6	0.8	2.0	-9.7	0.5	1.4
Γ_{22}	22.7	1.4	6.6	20.4	0.8	5.0	19.1	0.6	4.2	18.0	0.4	3.7
Total			134.3			117.4			109.3			98.1

were performed where the number of subpopulations in the model that was fit differed from the true number of subpopulations.

Again, these results can be compared to those in Tables E–3, and E–9. First, a model with $K = 1$ was fit, then one with $K = 3$. BIC was used to compare the three possibilities, and I showed the rank of the BIC for each $K \in \{1, 2, 3\}$ rather than tables of bias, variance, and MSE as shown for other simulation results. I used the same data values for outcomes, and covariates in these simulations, and in Chapter 4.

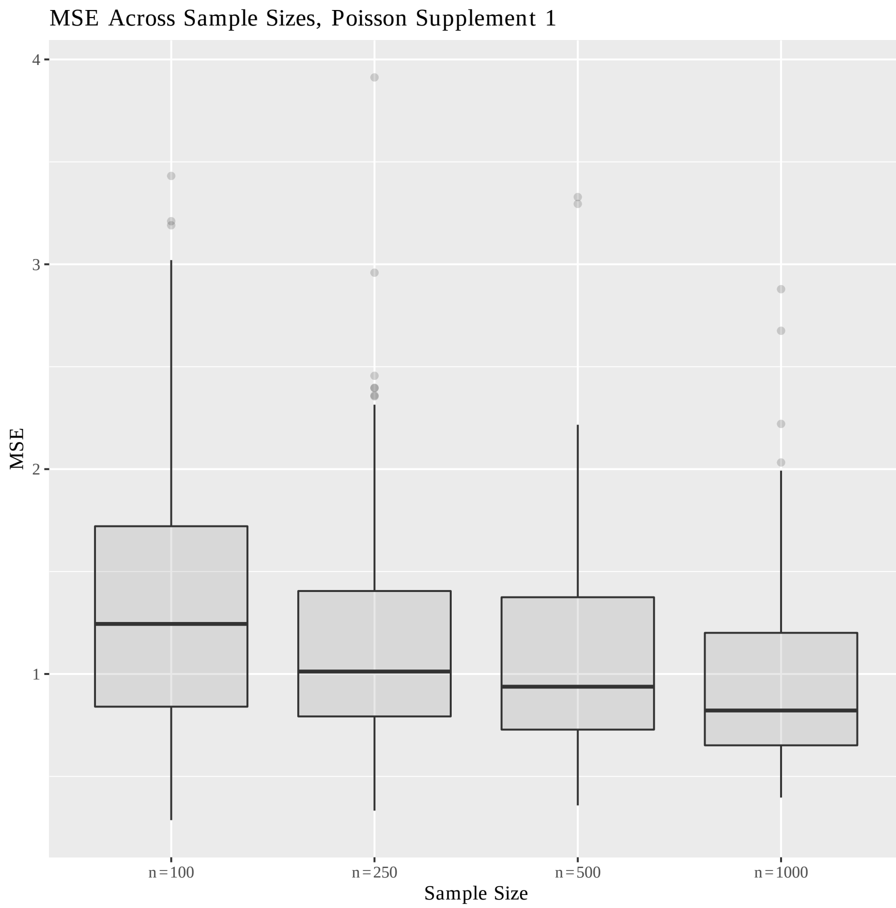

Figure F–1: MSE across sample sizes, Poisson outcome, one random effect set to zero.

These simulations show that the when n is larger, it is easier to identify the correct value of K. However, in these simulations $K = 2$, which is a relatively simple case.

F.2 Using the Same Starting Values for All Simulations

All previous simulations used randomised starting values and therefore I used a different starting value for each of the 100 simulations. I included this added randomisation in the previous results but is not inherent to the model. The following

Table F–2: Simulation 2 results multiplied by 100, averaged over 100 runs. Outcome follows a binomial distribution with $K = 2$, $p = 5$, $q = 2$, $\pi_1 = 0.6$, $\pi_2 = 0.4$, $\boldsymbol{\beta}_1^\top = (-0.55, 0.85, 1.25, -0.70, 0.00)$, and $\boldsymbol{\beta}_2^\top = (0.25, 0.00, 1.35, -0.20, 0.00)$, $\mathbb{\Gamma}_1^{*\top} = (0.00, 0.00, 1.00)$, and $\mathbb{\Gamma}_2^{*\top} = (1.05, 0.00, 0.00)$.

n	$n = 100$			$n = 250$			$n = 500$			$n = 1000$		
Measure	Bias	Var	MSE	Bias	Var	MSE	Bias	Var	MSE	Bias	Var	MSE
π_1	-5.4	0.9	1.1	-5.3	0.7	1.0	-4.7	0.6	0.8	-4.1	0.4	0.6
β_{10}	5.4	2.8	3.0	3.6	1.2	1.3	3.4	0.8	0.9	2.4	0.3	0.3
β_{11}	-5.8	4.1	4.4	-4.7	2.9	3.1	-2.1	2.1	2.2	-1.2	1.6	1.6
β_{12}	-1.7	2.4	2.4	-0.4	1.0	1.0	0.1	0.5	0.5	0.4	0.3	0.3
β_{13}	5.1	1.4	1.6	3.1	1.0	1.1	0.4	0.6	0.6	-1.4	0.3	0.3
β_{14}	-0.1	0.7	0.7	0.3	0.4	0.4	-0.1	0.3	0.3	0.4	0.1	0.1
Γ_{11}	43.4	10.2	28.9	30.1	6.2	15.3	19.4	5.6	9.3	8.7	2.0	2.7
Γ_{112}	21.7	8.8	13.5	32.3	7.5	17.8	40.1	6.8	22.8	49.2	4.2	28.3
Γ_{12}	-26.5	2.5	9.5	-23.4	1.8	7.2	-21.4	1.8	6.3	-20.5	1.5	5.7
β_{20}	-17.3	5.7	8.6	-16.4	4.4	7.0	-15.3	2.9	5.2	-11.2	2.0	3.3
β_{21}	18.9	4.1	7.6	16.3	2.7	5.3	12.6	1.5	3.1	10.1	1.0	2.0
β_{12}	-1.3	2.4	2.4	-1.9	1.3	1.3	-1.7	0.6	0.7	-0.7	0.3	0.3
β_{13}	-9.3	1.8	2.6	-7.5	1.3	1.8	-4.6	0.8	1.0	-3.0	0.3	0.4
β_{14}	-0.4	1.0	1.0	-0.5	0.6	0.6	0.3	0.3	0.3	-0.2	0.1	0.1
Γ_{21}	-5.8	4.8	5.1	-4.1	2.6	2.7	-1.2	1.2	1.2	0.1	0.6	0.5
Γ_{212}	-17.7	2.5	5.6	-18.6	1.3	4.7	-19.8	0.7	4.6	-18.1	0.6	3.8
Γ_{22}	65.4	2.2	44.9	63.1	1.7	41.5	57.9	1.3	34.8	52.7	1.0	28.8
Total			143.0			113.1			94.6			79.2

results use the same starting value for each simulation, to eliminate that as a source of variance. These results can be compared to Table E–3, and E–9. I used the same data values for outcomes, and covariates in these simulations as in Chapter 3.

These simulations showed that the starting value should be chosen carefully. In practice, one solution is to estimate the parameters of a model multiple times using multiple different starting values. However, this can be a computationally intensive process, but one that can be parallelised.

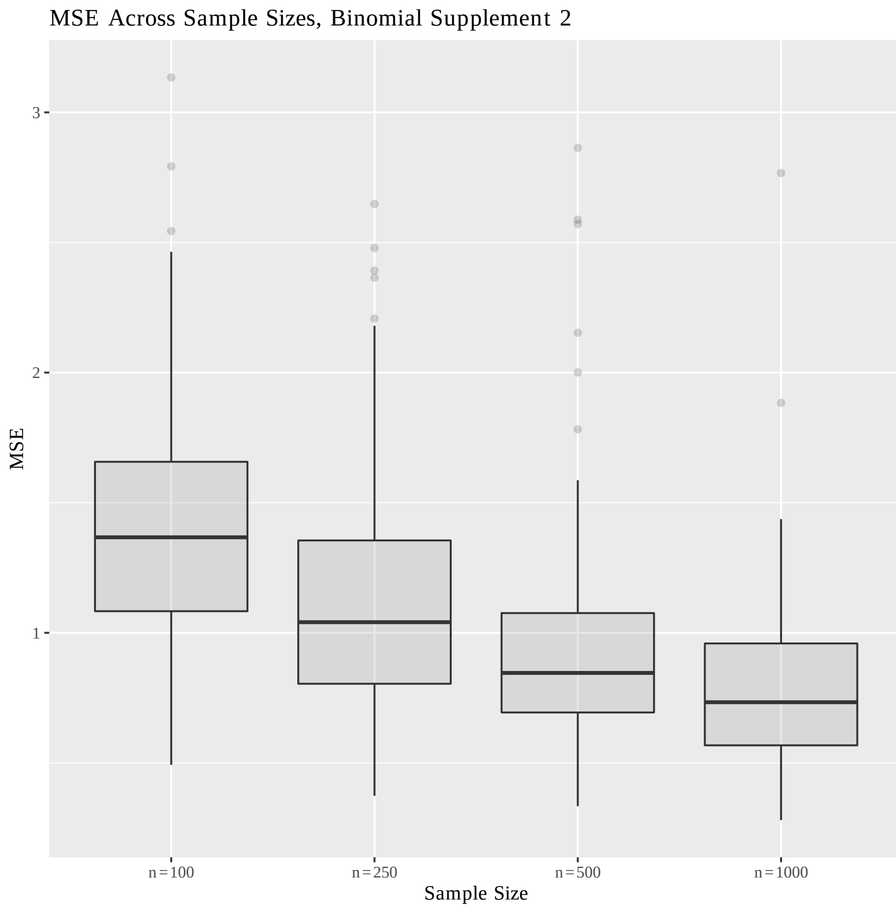

Figure F–2: MSE across sample sizes, Poisson outcome, one random effect set to zero.

F.3 Different Number of Visits per Patient

The number n represents the number of independent observations. However, for each subject n, there are n_i visits or observations. While these are not independent observations, changing the number of observations for each subject will change the overall total number of observation. Therefore, in these simulations, $n_i \in \{10, 11, 12\}$. These results can be compared to Table E–3, and E–9.

Table F–3: Simulation 3, and 4 results multiplied by 100, averaged over 100 runs. Outcome follows a Poisson distribution with $K = 2$, $p = 5$, $q = 2$, $\pi_1 = 0.6$, $\pi_2 = 0.4$, $\boldsymbol{\beta}_1^\top = (-0.75, 0.35, 0.10, -0.40, 0.00)$, $\boldsymbol{\beta}_2^\top = (0.60, -0.50, -0.35, -0.15, 0.00)$, $\mathbb{\Gamma}_1^{*\top} = (0.80, -0.15, 0.20)$, and $\mathbb{\Gamma}_2^{*\top} = (0.25, 0.00, 0.30)$, rank of BIC across different K. The true value of K used to generate the dataset was 2, but I considered different candidate values of $K \in \{1, 2, 3\}$ in the estimation of the parameters.

n	$n = 100$			$n = 250$			$n = 500$			$n = 1000$		
K	1	2	3	1	2	3	1	2	3	1	2	3
Lowest	68	32	0	9	91	0	0	100	0	0	100	0
Middle	32	68	0	83	9	8	37	0	63	0	0	100
Highest	0	0	100	8	0	92	63	0	37	100	0	0

Table F–4: Simulation 5, and 6 results multiplied by 100, averaged over 100 runs. Outcome follows a binomial distribution with $K = 2$, $p = 5$, $q = 2$, $\pi_1 = 0.6$, $\pi_2 = 0.4$, $\boldsymbol{\beta}_1^\top = (-0.55, 0.85, 1.25, -0.70, 0.00)$, and $\boldsymbol{\beta}_2^\top = (0.25, 0.00, 1.35, -0.20, 0.00)$, $\mathbb{\Gamma}_1^{*\top} = (0.00, 0.00, 1.00)$, and $\mathbb{\Gamma}_2^{*\top} = (1.05, 0.00, 0.00)$, rank of BIC across different K. The true value of K used to generate the dataset was 2, but I considered different candidate values of $K \in \{1, 2, 3\}$ in the estimation of the parameters.

n	$n = 100$			$n = 250$			$n = 500$			$n = 1000$		
K	1	2	3	1	2	3	1	2	3	1	2	3
Lowest	11	89	0	0	100	0	0	100	0	0	100	0
Middle	76	11	13	15	0	85	0	0	100	0	0	100
Highest	13	0	87	85	0	15	100	0	0	100	0	0

As expected, when more data was available, the parameter estimates exhibited smaller bias and MSE.

F.4 Variable m_{ij} in the binomial Case

In Chapter 3, all of the binomial simulations used $m_{ij} = 10 \forall i \in \{1, 2, \ldots, n\} \forall j \in \{1, 2, \ldots, n_i\}$. While this is reasonable in a number of situations, in order to explore the model further, I also considered the case where m_{ij} varies. I assumed that the values for m_{ij} are known in the model, as is typical in binomial regression.

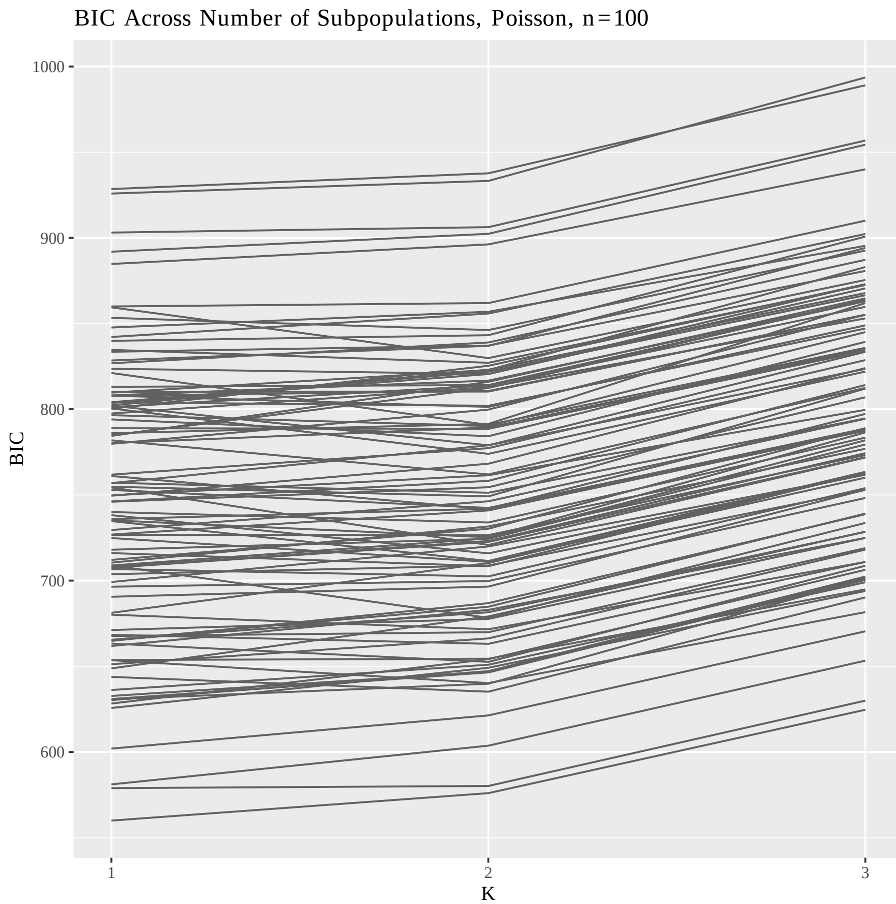

Figure F–3: BIC across different K, Poisson outcome, $n = 100$.

One possible motivation for exploring this case came from the SERA cohort, and the problem of modelling the number of tender or swollen joints in a patient with RA. While most patients have the same number of joints, a patient who has undergone certain joint replacement surgeries or amputations could have fewer joints. Therefore, in addition to the statistical exploration of the model, this special case is of interest.

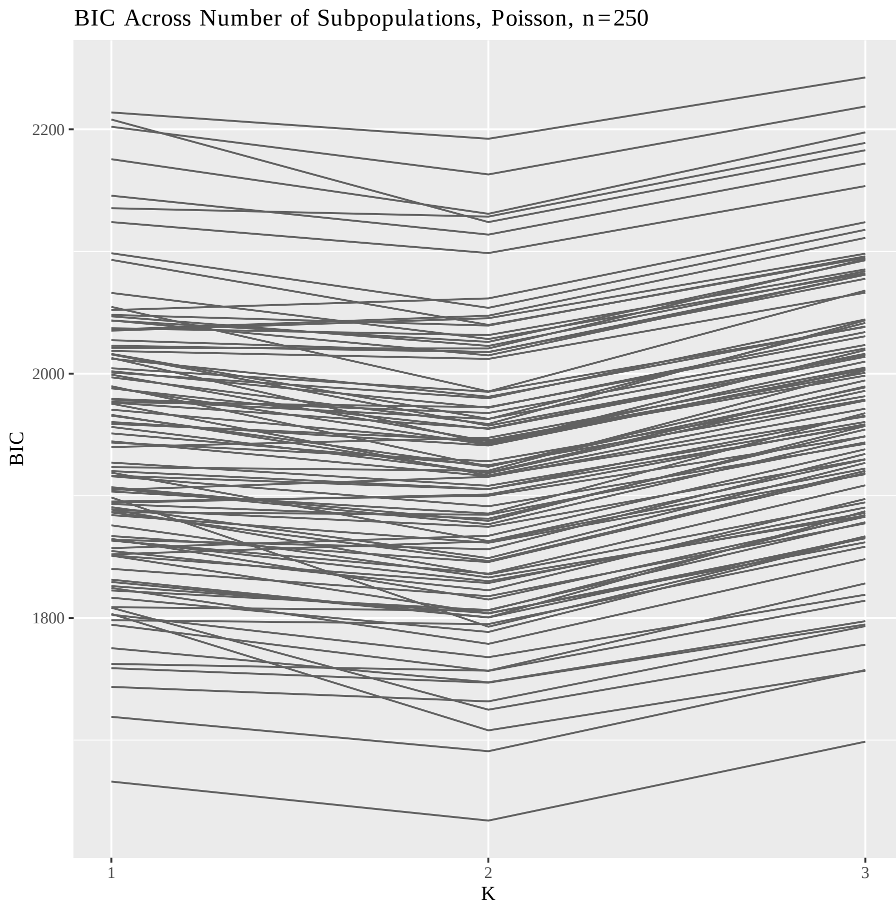

Figure F–4: BIC across different K, Poisson outcome, $n = 250$.

Again, I used $K = 2$, $p = 5$, and $q = 2$ and Table E–9 contains results for a similar case. I generated m_{ij} from a discrete uniform distribution with support $m_{ij} \in \{5, 6, \ldots, 15\}$.

It is reasonable to see an increase in variance induced by the additional inconsistency of m_{ij}. However, these simulations show that this algorithm provides reasonable estimates even in this situation.

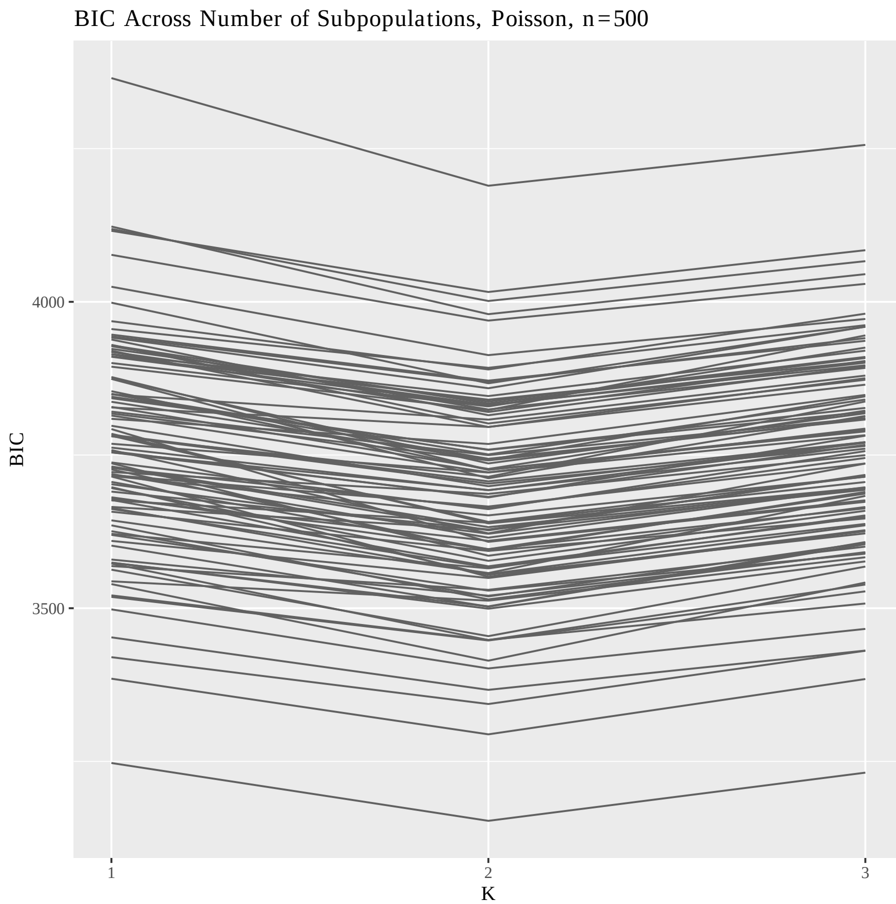

Figure F–5: BIC across different K, Poisson outcome, $n = 500$.

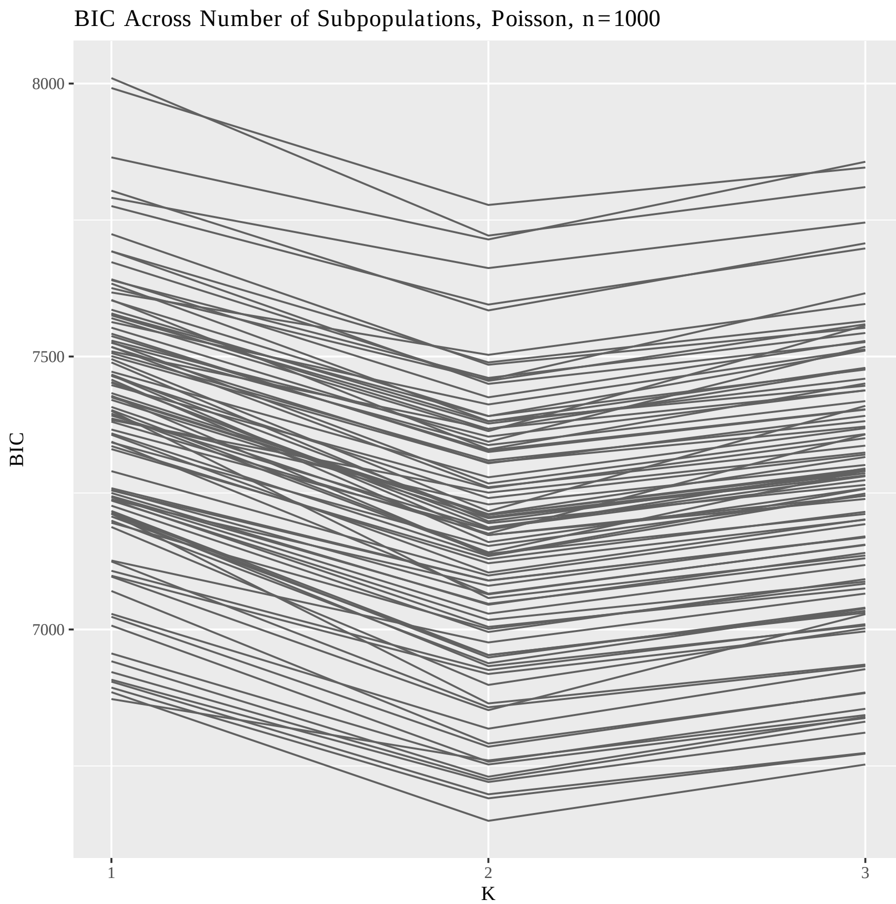

Figure F–6: BIC across different K, Poisson outcome, $n = 1000$.

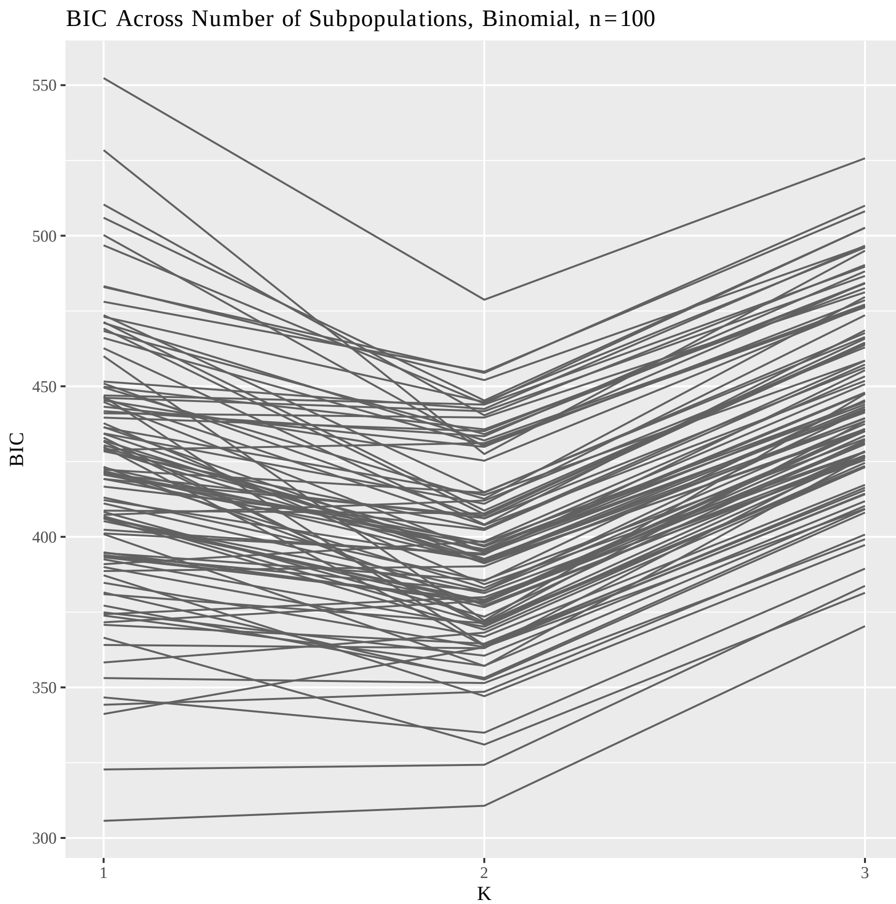

Figure F–7: BIC across different K, binomial outcome, $n = 100$.

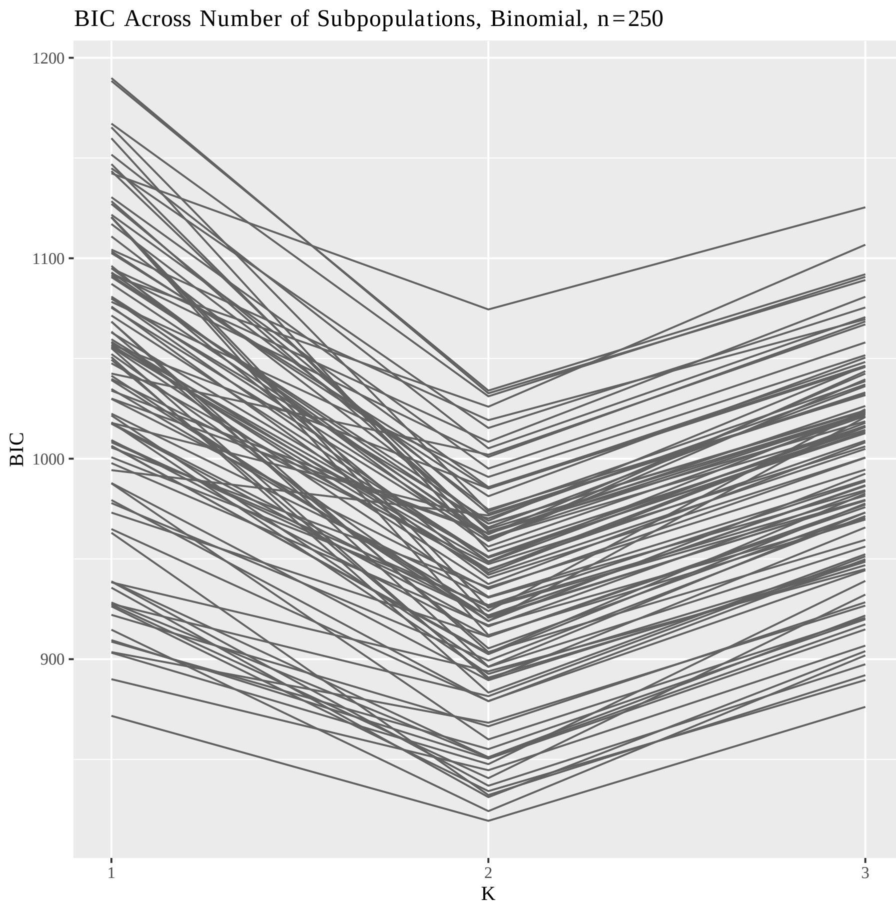

Figure F–8: BIC across different K, binomial outcome, $n = 250$.

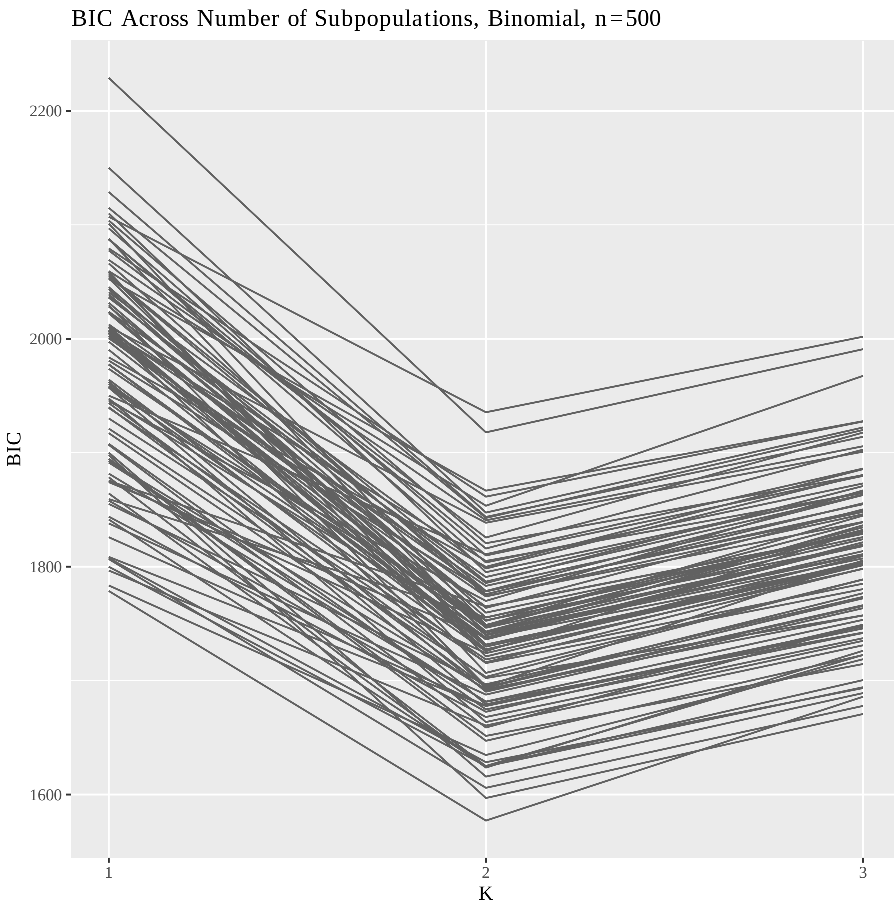

Figure F–9: BIC across different K, binomial outcome, $n = 500$.

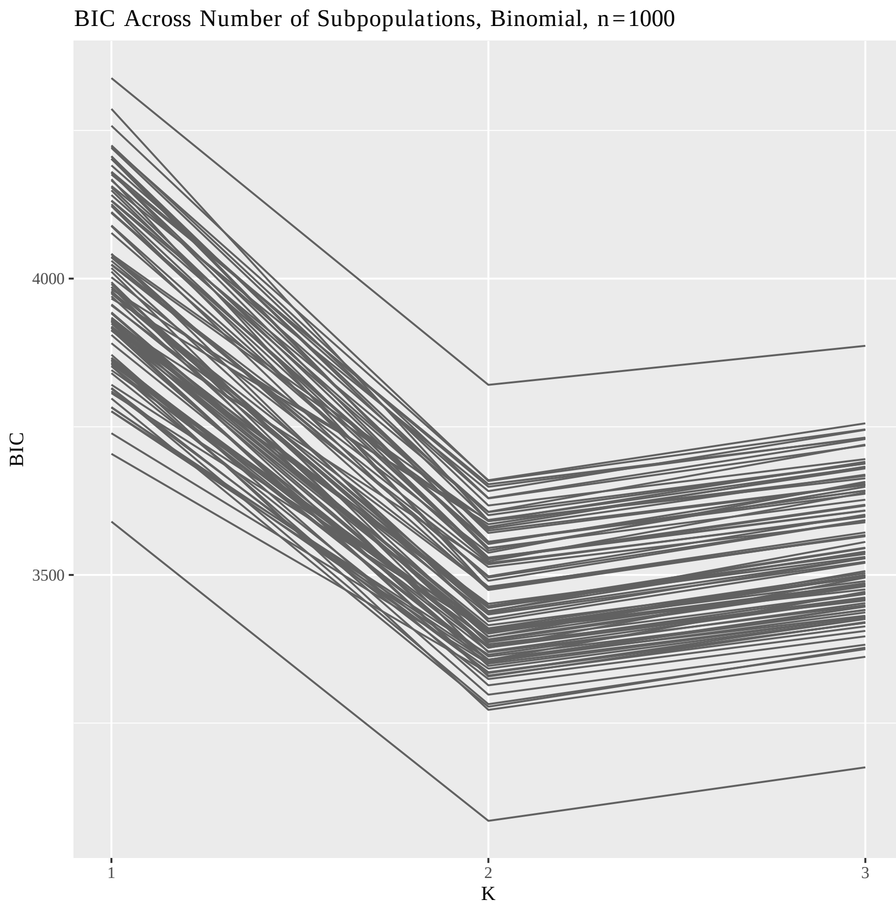

Figure F–10: BIC across different K, binomial outcome, $n = 1000$.

Table F–5: Simulation 7 results multiplied by 100, averaged over 100 runs. Outcome follows a Poisson distribution with $K = 2$, $p = 5$, $q = 2$, $\pi_1 = 0.6$, $\pi_2 = 0.4$, $\boldsymbol{\beta}_1^\top = (-0.75, 0.35, 0.10, -0.40, 0.00)$, $\boldsymbol{\beta}_2^\top = (0.60, -0.50, -0.35, -0.15, 0.00)$, $\mathbb{\Gamma}_1^{*\top} = (0.80, -0.15, 0.20)$, and $\mathbb{\Gamma}_2^{*\top} = (0.25, 0.00, 0.30)$. Starting value of $(0.50, 0.00, 0.00, 0.00, 0.00, 0.00, 1.00, 0.00, 1.00, -0.50, 0.00, 0.00, 0.00, 0.00, 1.00, 0.00, 1.00)$.

n	$n = 100$			$n = 250$			$n = 500$			$n = 1000$		
Measure	Bias	Var	MSE	Bias	Var	MSE	Bias	Var	MSE	Bias	Var	MSE
π_1	-6.9	< 0.1	0.5	-6.4	< 0.1	0.4	-6.4	< 0.1	0.4	-6.2	< 0.1	0.4
β_{10}	44.5	2.3	22.0	38.5	1.3	16.1	38.5	0.6	15.4	30.7	1.4	10.8
β_{11}	-29.5	1.2	9.9	-26.2	0.7	7.6	-26.0	0.4	7.2	-21.0	0.8	5.2
β_{12}	-21.8	0.9	5.6	-18.9	0.4	4.0	-18.2	0.3	3.6	-13.7	0.7	2.6
β_{13}	11.1	0.2	1.5	10.1	0.1	1.1	9.9	0.1	1.1	7.5	0.2	0.8
β_{14}	0.1	0.2	0.2	0.3	0.1	0.1	0.0	< 0.1	0.0	0.2	< 0.1	0.0
Γ_{11}	11.9	2.6	4.0	14.6	1.2	3.4	15.5	0.5	2.9	16.1	0.3	2.9
Γ_{112}	-19.7	1.6	5.5	-21.7	0.7	5.4	-23.1	0.4	5.7	-22.2	0.2	5.2
Γ_{12}	15.6	0.6	3.0	12.3	0.2	1.7	13.1	0.1	1.8	11.8	0.1	1.4
β_{20}	-46.1	1.5	22.7	-44.8	0.8	20.9	-45.8	0.5	21.5	-41.6	1.0	18.3
β_{21}	34.3	1.0	12.7	31.0	0.6	10.2	31.3	0.3	10.1	26.8	0.7	7.9
β_{12}	17.8	0.9	4.1	15.7	0.3	2.8	15.6	0.2	2.7	12.0	0.4	1.9
β_{13}	-10.9	0.2	1.4	-9.5	0.1	1.0	-9.0	0.1	0.9	-7.0	0.1	0.6
β_{14}	-0.4	0.2	0.2	0.2	0.1	0.1	-0.1	< 0.1	0.0	0.1	< 0.1	0.0
Γ_{21}	48.8	1.6	25.4	42.7	0.7	18.9	43.5	0.4	19.4	36.9	1.0	14.7
Γ_{212}	-23.3	1.5	6.9	-19.3	0.6	4.4	-21.2	0.4	4.9	-17.2	0.5	3.5
Γ_{22}	6.2	0.6	1.0	4.2	0.3	0.4	4.6	0.1	0.3	3.3	0.1	0.2
Total			126.6			98.5			97.8			76.4

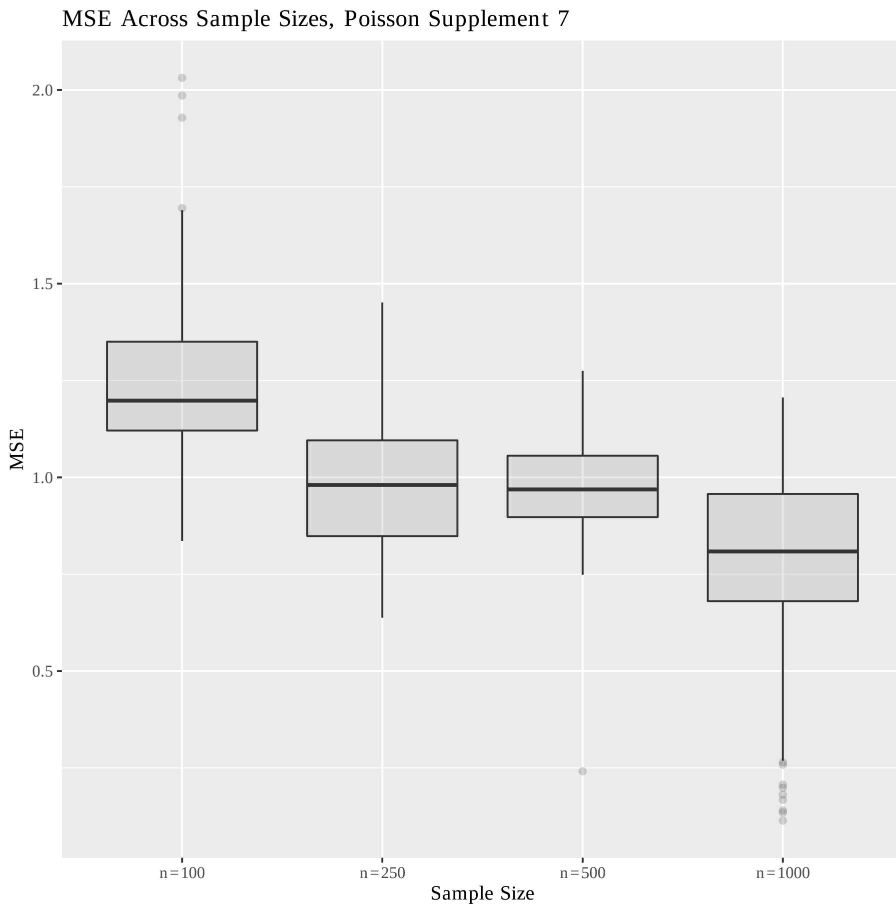

Figure F–11: MSE across sample sizes, Poisson outcome, with a constant starting value.

Table F–6: Simulation 8 results multiplied by 100, averaged over 100 runs. Outcome follows a binomial distribution with $K = 2$, $p = 5$, $q = 2$, $\pi_1 = 0.6$, $\pi_2 = 0.4$, $\boldsymbol{\beta}_1^\top = (-0.55, 0.85, 1.25, -0.70, 0.00)$, $\boldsymbol{\beta}_2^\top = (0.25, -0.50, 1.35, -0.20, 0.00)$, $\mathbb{\Gamma}_1^{*\top} = (1.60, -0.45, 1.00)$, and $\mathbb{\Gamma}_2^{*\top} = (1.05, 0.00, 1.40)$. Starting value of $(0.50, 0.00, 0.00, 0.00, 0.00, 0.00, 1.00, 0.00, 1.00, -0.50, 0.00, 0.00, 0.00, 0.00, 1.00, 0.00, 1.00)$.

n	$n = 100$			$n = 250$			$n = 500$			$n = 1000$		
Measure	Bias	Var	MSE	Bias	Var	MSE	Bias	Var	MSE	Bias	Var	MSE
π_1	-7.2	< 0.1	0.5	-7.6	< 0.1	0.6	-7.4	< 0.1	0.6	-7.5	< 0.1	0.6
β_{10}	5.7	3.4	3.7	2.3	1.0	1.0	3.3	0.5	0.6	4.5	0.3	0.5
β_{11}	-44.5	3.2	23.0	-43.4	0.8	19.7	-43.7	0.4	19.4	-44.3	0.2	19.9
β_{12}	4.0	1.5	1.6	2.9	0.5	0.6	2.5	0.2	0.3	2.9	0.1	0.2
β_{13}	17.7	0.4	3.5	17.5	0.1	3.2	17.8	0.1	3.2	17.7	< 0.1	3.2
β_{14}	-0.4	0.3	0.3	0.3	0.1	0.1	0.1	< 0.1	0.0	0.2	< 0.1	0.0
Γ_{11}	-5.3	5.7	5.9	-2.4	1.7	1.7	-2.4	0.8	0.8	-2.4	0.4	0.4
Γ_{112}	2.1	4.5	4.5	0.3	1.2	1.2	-0.6	0.5	0.5	-0.5	0.3	0.3
Γ_{12}	12.6	0.7	2.3	14.3	0.3	2.3	14.1	0.1	2.1	14.7	0.1	2.2
β_{20}	-6.0	2.7	3.0	-9.9	1.0	1.9	-9.2	0.5	1.3	-8.6	0.3	1.0
β_{21}	71.6	2.8	54.1	72.5	1.0	53.4	72.7	0.4	53.2	72.7	0.2	53.0
β_{12}	-5.4	1.5	1.8	-6.1	0.6	1.0	-6.1	0.3	0.7	-5.8	0.1	0.5
β_{13}	-27.0	0.4	7.6	-26.6	0.2	7.2	-26.6	0.1	7.2	-26.6	< 0.1	7.1
β_{14}	-0.4	0.2	0.2	0.1	0.1	0.1	0.0	0.1	0.1	0.1	< 0.1	0.0
Γ_{21}	39.0	4.5	19.7	40.9	1.5	18.3	41.2	0.8	17.8	40.6	0.4	16.9
Γ_{212}	-34.9	4.0	16.1	-35.3	1.3	13.7	-36.1	0.6	13.7	-35.3	0.3	12.8
Γ_{22}	-27.2	0.8	8.2	-25.5	0.4	6.9	-25.7	0.2	6.8	-24.4	0.1	6.1
Total			156.1			133.0			128.2			124.7

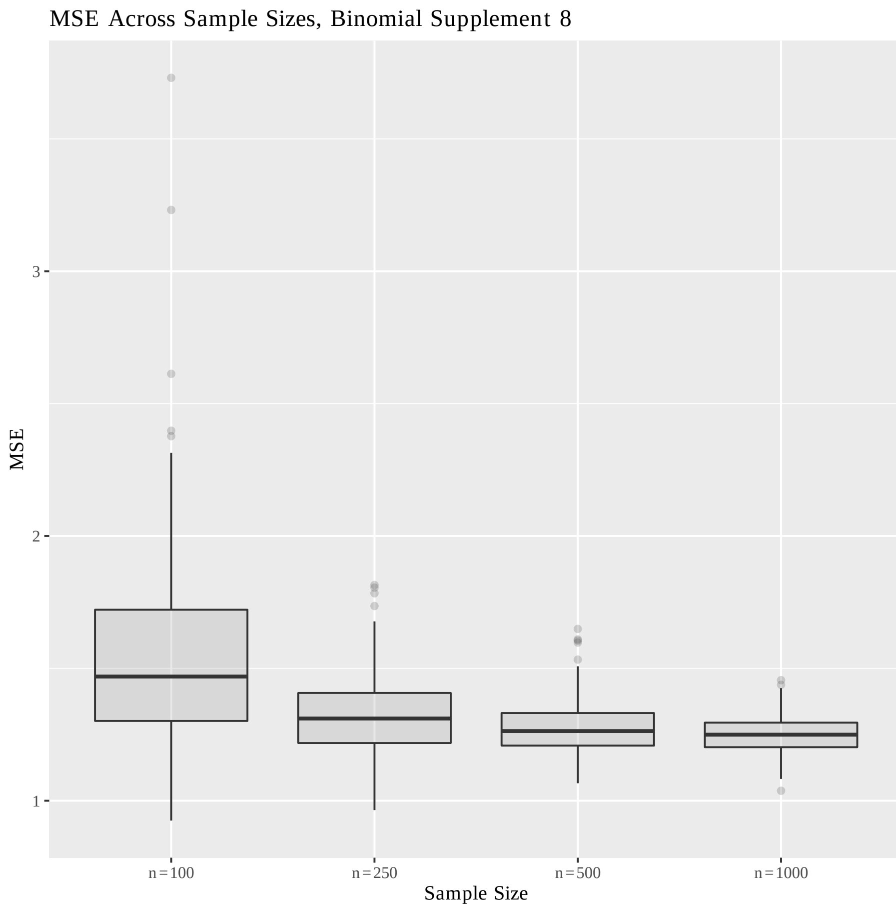

Figure F–12: MSE across sample sizes, binomial outcome, with constant starting value.

Table F–7: Simulation 9 results multiplied by 100, averaged over 100 runs. Outcome follows a Poisson distribution with $K = 2$, $p = 5$, $q = 2$, $\pi_1 = 0.6$, $\pi_2 = 0.4$, $\boldsymbol{\beta}_1^\top = (-0.75, 0.35, 0.10, -0.40, 0.00)$, $\boldsymbol{\beta}_2^\top = (0.60, -0.50, -0.35, -0.15, 0.00)$, $\mathbb{\Gamma}_1^{*\top} = (0.80, -0.15, 0.20)$, and $\mathbb{\Gamma}_2^{*\top} = (0.25, 0.00, 0.30)$, $n_i \in \{10, 11, 12\}$.

n	$n = 100$			$n = 250$			$n = 500$			$n = 1000$		
Measure	Bias	Var	MSE	Bias	Var	MSE	Bias	Var	MSE	Bias	Var	MSE
π_1	-5.1	1.4	1.6	-4.6	1.3	1.5	-4.6	1.1	1.3	-4.3	1.1	1.3
β_{10}	11.7	8.8	10.1	9.9	5.1	6.1	8.3	4.3	4.9	4.9	3.2	3.4
β_{11}	-6.2	2.7	3.1	-5.4	1.5	1.8	-4.5	1.4	1.6	-2.6	1.1	1.1
β_{12}	-4.6	2.1	2.3	-2.3	1.4	1.4	-1.1	1.1	1.1	-0.8	0.8	0.8
β_{13}	3.5	0.7	0.8	2.8	0.6	0.6	2.2	0.4	0.5	0.9	0.4	0.4
β_{14}	-0.2	0.6	0.6	0.0	0.4	0.4	-0.3	0.3	0.3	-0.1	0.2	0.2
Γ_{11}	7.1	4.9	5.4	7.3	1.8	2.4	8.9	0.9	1.7	9.4	0.9	1.7
Γ_{112}	-12.8	2.2	3.8	-12.9	0.7	2.4	-13.1	0.4	2.1	-12.4	0.4	1.9
Γ_{12}	11.6	1.2	2.6	10.8	0.6	1.8	10.2	0.4	1.4	8.8	0.3	1.0
β_{20}	-19.0	3.6	7.2	-18.2	2.9	6.1	-17.4	2.6	5.6	-15.4	2.1	4.5
β_{21}	11.6	2.1	3.5	10.0	1.4	2.4	9.4	1.4	2.3	8.1	1.1	1.7
β_{12}	3.2	1.3	1.3	2.8	0.8	0.9	2.3	0.6	0.6	1.8	0.4	0.5
β_{13}	-2.6	0.7	0.8	-1.9	0.4	0.5	-2.2	0.4	0.4	-1.6	0.3	0.3
β_{14}	-0.9	0.6	0.6	-0.2	0.3	0.3	-0.1	0.2	0.2	0.0	0.1	0.1
Γ_{21}	19.7	4.2	8.1	17.2	2.9	5.8	15.1	2.8	5.0	11.9	2.2	3.6
Γ_{212}	-6.9	2.4	2.9	-5.0	1.6	1.8	-4.2	1.3	1.4	-3.1	1.0	1.1
Γ_{22}	-2.9	1.4	1.4	-2.9	0.8	0.9	-2.6	0.6	0.6	-2.5	0.4	0.5
Total			56.1			37.1			31.1			24.1

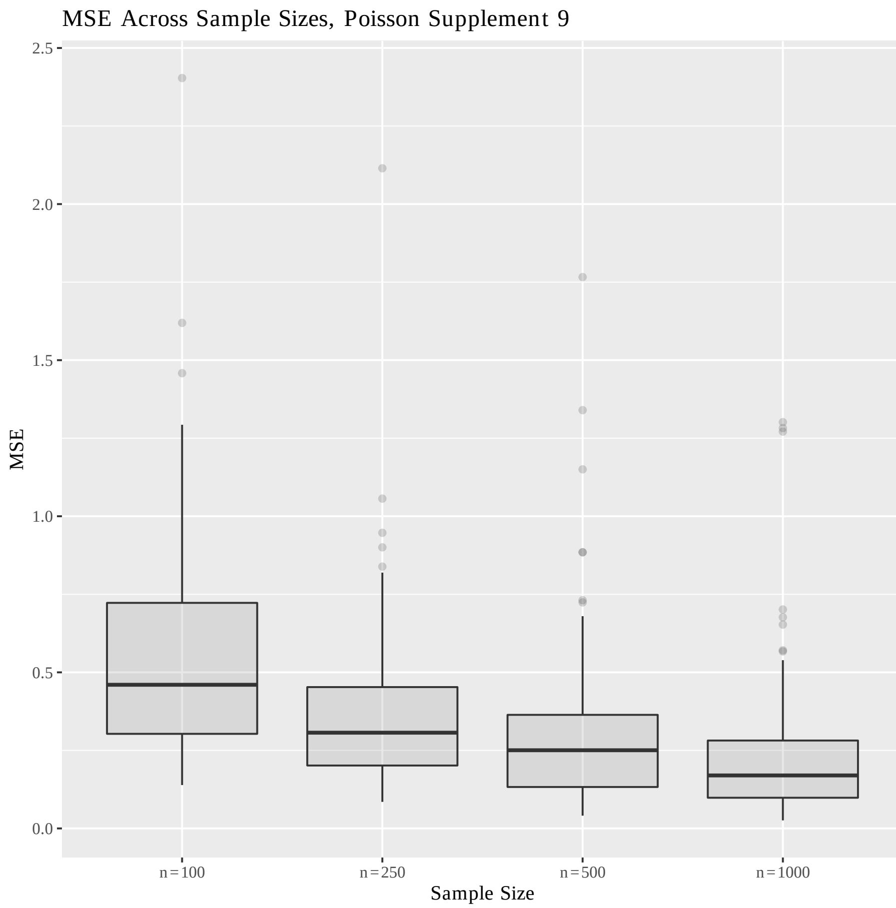

Figure F–13: MSE across sample sizes, Poisson outcome, with $n_i \in \{10, 11, 12\}$.

Table F–8: Simulation 10 results multiplied by 100, averaged over 100 runs. Outcome follows a binomial distribution with $K = 2$, $p = 5$, $q = 2$, $\pi_1 = 0.6$, $\pi_2 = 0.4$, $\boldsymbol{\beta}_1^\top = (-0.55, 0.85, 1.25, -0.70, 0.00)$, $\boldsymbol{\beta}_2^\top = (0.25, -0.50, 1.35, -0.20, 0.00)$, $\mathbb{\Gamma}_1^{*\top} = (1.60, -0.45, 1.00)$, and $\mathbb{\Gamma}_2^{*\top} = (1.05, 0.00, 1.40)$, $n_i \in \{10, 11, 12\}$.

n	$n = 100$			$n = 250$			$n = 500$			$n = 1000$		
Measure	Bias	Var	MSE	Bias	Var	MSE	Bias	Var	MSE	Bias	Var	MSE
π_1	-3.6	1.8	1.9	-2.7	1.3	1.3	-1.7	0.8	0.8	-0.8	0.6	0.6
β_{10}	-2.4	10.0	9.9	-2.1	7.6	7.6	-0.7	6.6	6.5	-0.5	4.9	4.9
β_{11}	-3.3	5.7	5.8	-1.7	4.0	4.0	-1.9	3.0	3.0	-2.6	2.2	2.2
β_{12}	0.1	1.0	1.0	-0.1	0.4	0.4	-0.1	0.2	0.2	-0.1	0.1	0.1
β_{13}	1.7	0.9	0.9	0.4	0.5	0.5	0.0	0.4	0.4	0.4	0.3	0.3
β_{14}	-1.3	0.4	0.4	-0.7	0.1	0.1	-0.6	0.1	0.1	-0.2	< 0.1	< 0.1
Γ_{11}	3.6	4.6	4.7	3.5	2.6	2.7	3.6	1.7	1.8	2.3	1.1	1.2
Γ_{112}	17.1	2.1	5.0	13.2	1.6	3.3	8.9	1.3	2.0	6.4	0.7	1.1
Γ_{12}	24.7	2.8	8.9	18.5	1.9	5.3	14.4	1.3	3.3	10.2	0.7	1.7
β_{20}	-6.8	7.1	7.5	-5.6	5.5	5.7	-4.1	5.0	5.1	-1.2	3.8	3.7
β_{21}	16.2	6.8	9.3	13.7	5.7	7.5	11.6	4.9	6.2	8.8	4.3	5.0
β_{12}	-1.9	1.2	1.2	-0.3	0.5	0.5	-0.1	0.3	0.3	-0.2	0.2	0.2
β_{13}	-6.5	1.7	2.1	-4.2	1.1	1.2	-2.6	0.7	0.8	-2.0	0.6	0.6
β_{14}	1.7	0.4	0.4	1.0	0.1	0.2	0.7	0.1	0.1	0.2	< 0.1	< 0.1
Γ_{21}	14.8	6.4	8.5	10.5	4.6	5.7	7.6	2.5	3.1	4.3	1.5	1.7
Γ_{212}	2.3	5.7	5.6	3.9	4.9	5.0	4.3	3.5	3.6	5.1	3.2	3.4
Γ_{22}	-15.9	3.3	5.8	-15.9	2.6	5.1	-14.9	1.8	4.0	-13.4	1.0	2.8
Total			79.0			56.1			41.3			29.4

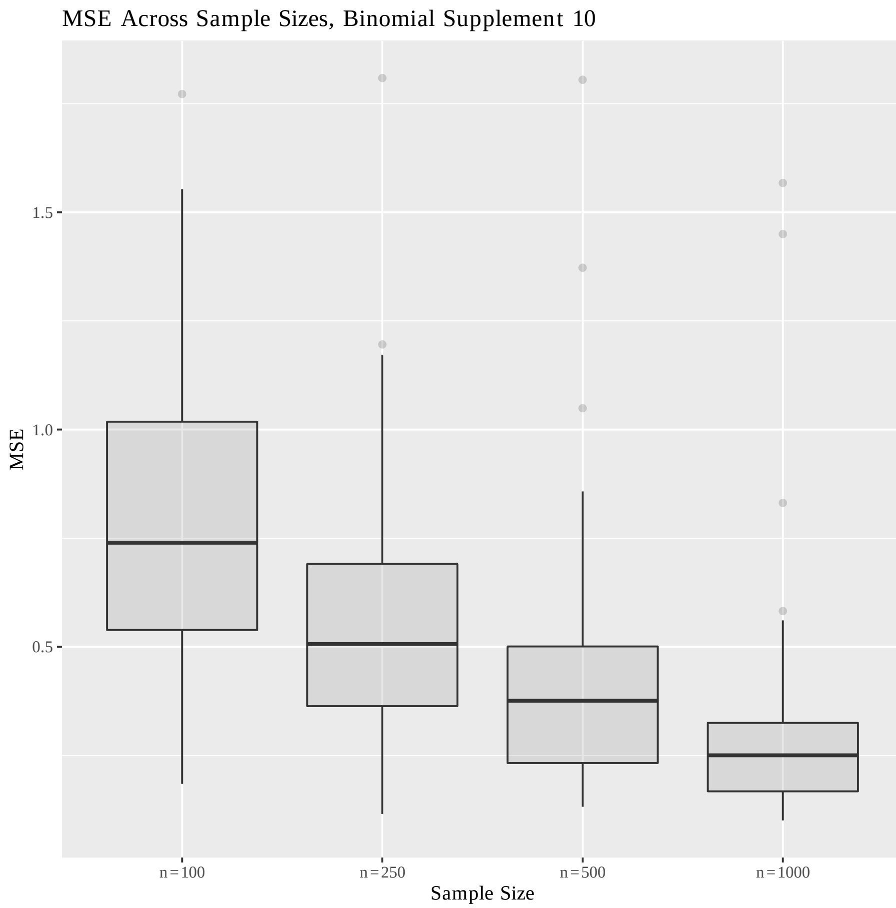

Figure F–14: MSE across sample sizes, binomial outcome, with $n_i \in \{10, 11, 12\}$.

Table F–9: Simulation 11 results multiplied by 100, averaged over 100 runs. Outcome follows a binomial distribution with $K = 2$, $p = 5$, $q = 2$, $\pi_1 = 0.6$, $\pi_2 = 0.4$, $\boldsymbol{\beta}_1^\top = (-0.55, 0.85, 1.25, -0.70, 0.00)$, $\boldsymbol{\beta}_2^\top = (0.25, -0.50, 1.35, -0.20, 0.00)$, $\mathbb{\Gamma}_1^{*\top} = (1.60, -0.45, 1.00)$, and $\mathbb{\Gamma}_2^{*\top} = (1.05, 0.00, 1.40)$, and $m_{ij} \in \{5, 6, \ldots, 15\}$.

n	$n = 100$			$n = 250$			$n = 500$			$n = 1000$		
Measure	Bias	Var	MSE	Bias	Var	MSE	Bias	Var	MSE	Bias	Var	MSE
π_1	-6.7	0.5	1.0	-6.9	0.5	1.0	-6.9	0.5	1.0	-6.8	0.5	0.9
β_{10}	1.6	11.0	10.9	-0.5	7.2	7.1	-1.8	5.0	5.0	-2.9	4.0	4.0
β_{11}	1.1	5.6	5.5	1.9	3.1	3.1	3.5	2.5	2.6	6.1	2.2	2.5
β_{12}	-0.7	2.9	2.9	-1.2	1.5	1.5	-0.6	1.1	1.1	-0.3	0.6	0.6
β_{13}	2.9	2.2	2.2	3.1	1.6	1.7	2.6	1.1	1.2	0.9	1.0	1.0
β_{14}	1.0	1.4	1.4	1.4	0.7	0.7	1.3	0.5	0.5	0.7	0.3	0.3
Γ_{11}	-4.2	9.1	9.2	-0.5	3.8	3.8	-2.1	2.5	2.5	-0.2	1.7	1.7
Γ_{112}	18.0	6.2	9.4	12.3	2.4	3.9	10.8	1.6	2.8	7.3	1.4	1.9
Γ_{12}	19.5	3.7	7.4	17.2	2.9	5.8	14.1	1.7	3.7	11.0	1.4	2.6
β_{20}	-11.7	8.6	9.9	-9.3	5.9	6.7	-10.1	4.8	5.8	-6.4	4.1	4.4
β_{21}	23.9	7.3	12.9	22.0	4.5	9.3	21.8	4.2	8.9	17.2	4.0	6.9
β_{12}	0.4	3.6	3.5	-1.6	2.0	2.0	-0.6	1.1	1.1	-0.7	0.6	0.6
β_{13}	-11.6	2.3	3.6	-11.1	1.6	2.8	-10.6	1.2	2.3	-8.8	1.1	1.8
β_{14}	-0.7	1.2	1.2	-1.3	0.9	0.9	-1.3	0.6	0.6	-0.5	0.3	0.3
Γ_{21}	17.7	8.5	11.5	21.6	3.5	8.1	19.2	2.5	6.1	14.6	2.1	4.2
Γ_{212}	-6.8	8.9	9.2	-13.0	4.3	6.0	-12.9	4.0	5.7	-10.5	3.8	4.8
Γ_{22}	-23.9	3.1	8.8	-25.0	2.1	8.4	-25.3	1.7	8.1	-23.3	1.4	6.8
Total			110.7			72.7			58.8			45.5

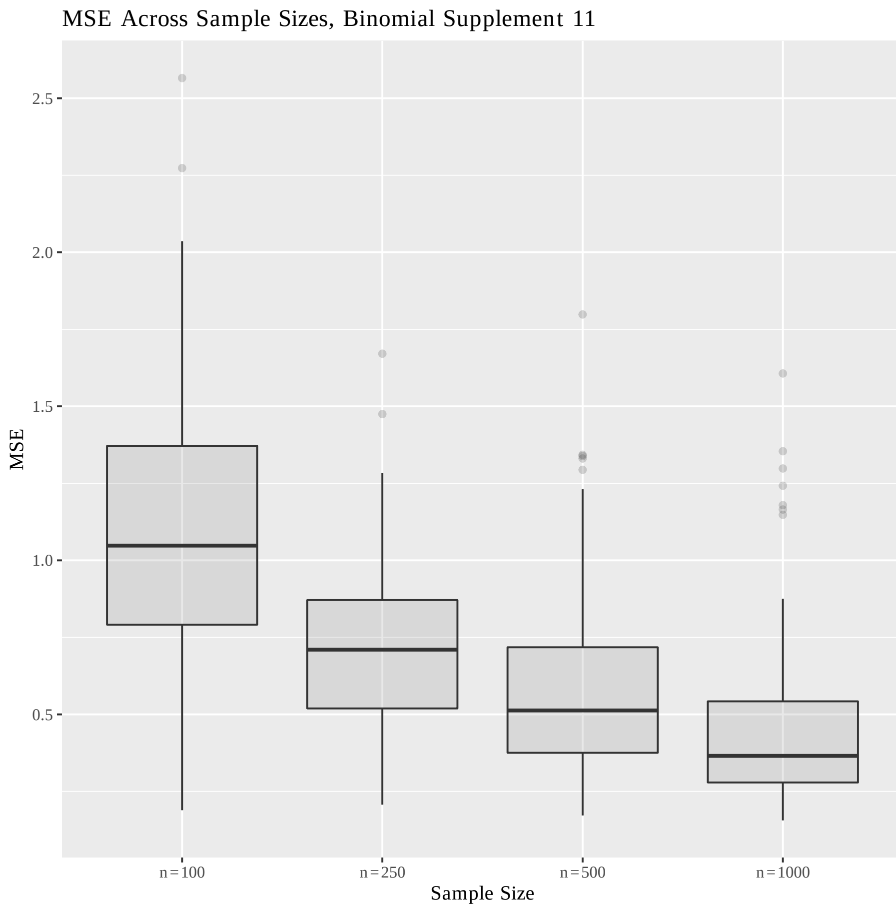

Figure F–15: MSE across sample sizes, binomial outcome, with $m_{ij} \in \{5, 6, \ldots, 15\}$.

APPENDIX G
Additional Tables From Chapter 4 (Bias, Variance, Mean Squared Error With Least Absolute Shrinkage and Selection Operator, Adaptive Least Absolute Shrinkage and Selection Operator, Smoothly Clipped Absolute Deviation)

This appendix contains the tables of simulation results for Chapter 4, in particular the results of simulations ran using the LASSO, ALASSO, and SCAD penalties.

Table G–1: Simulation 1 results multiplied by 100, averaged over 50 runs. Outcome follows a Poisson distribution with $K = 2$, $p = 7$, $q = 2$, $\pi_1 = 0.6$, $\pi_2 = 0.4$, $\boldsymbol{\beta}_1^\top = (0.65, 0.30, 0.00, 0.00, 0.00, 0.00, 0.00)$, $\boldsymbol{\beta}_2^\top = (0.20, -0.45, 0.00, 0.00, 0.00, 0.00, 0.00)$, $\mathbb{F}_1^{*\top} = (0.30, -0.25, 0.10)$, and $\mathbb{F}_2^{*\top} = (0.35, 0.20, 0.15)$. LASSO penalty.

n	$n = 100$			$n = 250$		
Measure	Bias	Var	MSE	Bias	Var	MSE
π_1	7.2	1.3	1.8	2.0	0.8	0.8
β_{10}	-19.5	5.5	9.2	-22.0	5.0	9.8
β_{11}	-28.0	0.4	8.2	-26.1	1.3	8.1
β_{12}	0.1	0.0	0.0	0.0	0.0	0.0
β_{13}	0.0	0.0	0.0	-0.1	0.0	0.0
β_{14}	-0.1	0.0	0.0	-0.1	0.0	0.0
β_{15}	0.0	0.0	0.0	-0.2	0.0	0.0
β_{16}	0.0	0.0	0.0	0.1	0.0	0.0
Γ_{11}	-19.2	3.2	6.9	-13.4	2.5	4.3
Γ_{112}	27.9	11.8	19.3	32.9	11.9	22.5
Γ_{12}	-1.5	3.9	3.8	1.3	2.2	2.1
β_{20}	29.5	21.2	29.4	38.3	12.3	26.7
β_{21}	43.5	1.4	20.2	43.3	2.3	21.0
β_{22}	-0.3	0.0	0.0	0.0	0.0	0.0
β_{23}	0.2	0.0	0.0	0.1	0.0	0.0
β_{24}	-0.4	0.1	0.1	-0.3	0.0	0.0
β_{25}	-0.1	0.0	0.0	-0.2	0.0	0.0
β_{26}	0.2	0.1	0.1	0.2	0.0	0.0
Γ_{21}	-7.4	5.6	6.1	-8.4	2.7	3.4
Γ_{212}	-20.4	12.9	16.8	-21.2	7.3	11.7
Γ_{22}	-10.3	2.6	3.6	-7.4	2.5	3.0
Total			125.7			113.6

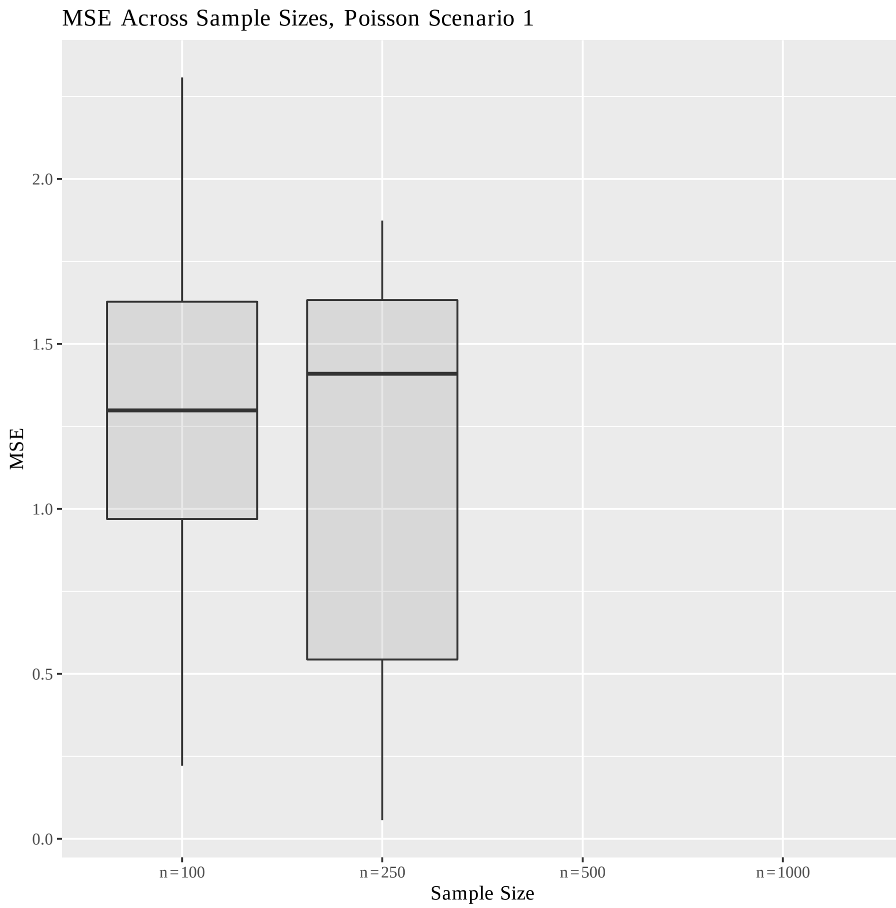

Figure G–1: MSE across sample sizes, Poisson outcome, LASSO penalty, with $K = 2$, $p = 7$, $q = 2$.

Table G–2: Simulation 1 results multiplied by 100, averaged over 50 runs. Outcome follows a Poisson distribution with $K = 2$, $p = 7$, $q = 2$, $\pi_1 = 0.6$, $\pi_2 = 0.4$, $\boldsymbol{\beta}_1^\top = (0.65, 0.30, 0.00, 0.00, 0.00, 0.00, 0.00)$, $\boldsymbol{\beta}_2^\top = (0.20, -0.45, 0.00, 0.00, 0.00, 0.00, 0.00)$, $\mathbb{\Gamma}_1^{*\top} = (0.30, -0.25, 0.10)$, and $\mathbb{\Gamma}_2^{*\top} = (0.35, 0.20, 0.15)$. ALASSO penalty.

n	$n = 100$			$n = 250$		
Measure	Bias	Var	MSE	Bias	Var	MSE
π_1	5.4	1.1	1.4	0.7	0.5	0.5
β_{10}	-17.8	5.8	8.8	-17.3	5.0	7.9
β_{11}	-26.6	0.7	7.7	-25.0	2.1	8.2
β_{12}	0.0	0.0	0.0	-0.1	0.0	0.0
β_{13}	0.0	0.0	0.0	-0.1	0.0	0.0
β_{14}	-0.1	0.0	0.0	-0.2	0.0	0.0
β_{15}	0.0	0.0	0.0	-0.1	0.0	0.0
β_{16}	0.0	0.0	0.0	0.1	0.0	0.0
Γ_{11}	-16.2	2.8	5.4	-9.2	2.1	2.9
Γ_{112}	27.2	11.8	18.9	28.6	9.6	17.6
Γ_{12}	-1.0	3.5	3.4	3.8	2.0	2.1
β_{20}	28.4	17.8	25.5	31.7	9.5	19.3
β_{21}	43.6	2.3	21.3	45.2	4.4	24.7
β_{22}	-0.3	0.0	0.0	-0.1	0.0	0.0
β_{23}	-0.1	0.1	0.1	0.2	0.0	0.0
β_{24}	-1.1	0.3	0.3	0.0	0.0	0.0
β_{25}	-0.1	0.0	0.0	-0.2	0.0	0.0
β_{26}	0.4	0.1	0.1	0.3	0.0	0.0
Γ_{21}	-7.1	4.6	5.1	-6.5	2.1	2.5
Γ_{212}	-21.3	9.9	14.2	-19.6	8.2	11.9
Γ_{22}	-7.7	2.2	2.8	-8.8	2.4	3.1
Total			115.1			100.8

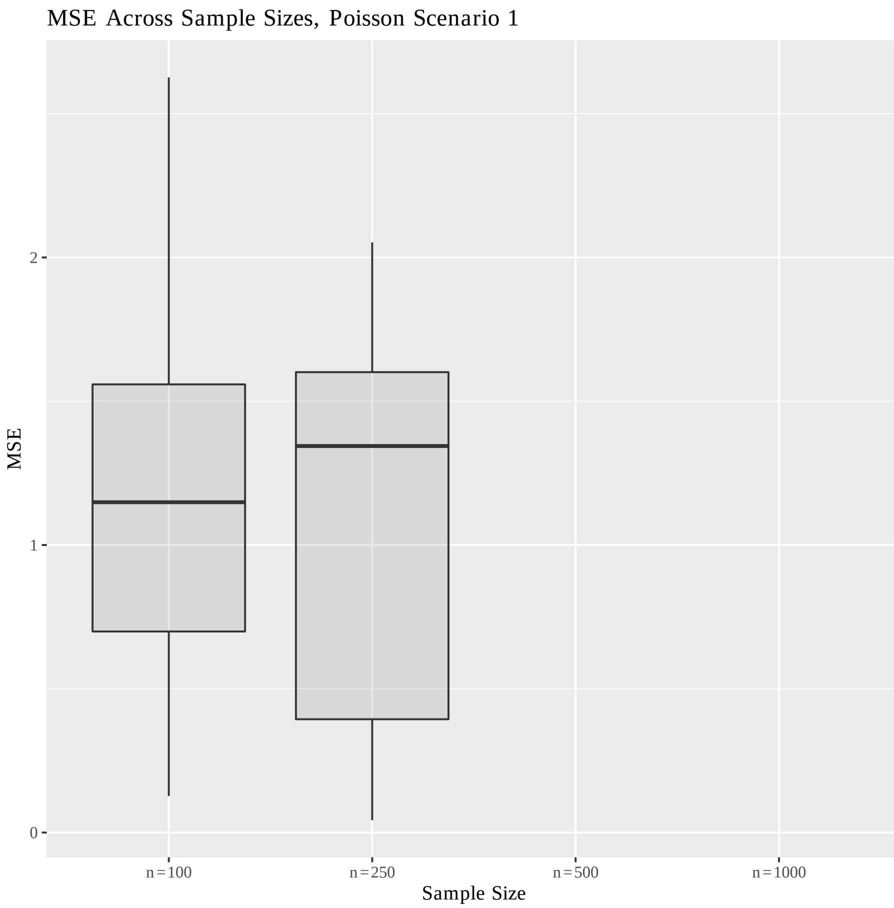

Figure G–2: MSE across sample sizes, Poisson outcome, ALASSO penalty, with $K = 2$, $p = 7$, $q = 2$.

Table G–3: Simulation 7 results multiplied by 100, averaged over 50 runs. Outcome follows a binomial distribution with $K = 2$, $p = 7$, $q = 2$, $\pi_1 = 0.6$, $\pi_2 = 0.4$, $\boldsymbol{\beta}_1^\top = (0.95, 0.60, 0.00, 0.00, 0.00, 0.00, 0.00)$, $\boldsymbol{\beta}_2^\top = (-0.85, -0.15, 0.00, 0.00, 0.00, 0.00, 0.00)$, $\mathbb{\Gamma}_1^{*\top} = (0.95, 0.85, 1.15)$, and $\mathbb{\Gamma}_2^{*\top} = (0.70, -0.70, 0.80)$. LASSO penalty.

n	$n = 100$			$n = 250$		
Measure	Bias	Var	MSE	Bias	Var	MSE
π_1	-0.8	0.4	0.4	-1.8	0.3	0.4
β_{10}	-113.8	51.9	180.5	-122.3	46.2	194.9
β_{11}	-56.6	3.0	35.0	-54.4	2.7	32.2
β_{12}	0.1	0.0	0.0	0.2	0.0	0.0
β_{13}	-0.5	0.0	0.0	-0.1	0.0	0.0
β_{14}	0.0	0.0	0.0	0.1	0.0	0.0
β_{15}	0.3	0.0	0.0	0.0	0.0	0.0
β_{16}	-0.5	0.1	0.1	0.0	0.0	0.0
Γ_{11}	-18.6	1.8	5.2	-18.2	0.8	4.1
Γ_{112}	-78.7	60.6	121.3	-98.1	30.9	126.5
Γ_{12}	-13.9	2.8	4.7	-13.6	2.4	4.2
β_{20}	165.0	80.9	351.6	178.7	70.0	387.9
β_{21}	40.3	16.4	32.3	67.7	12.4	58.0
β_{22}	0.9	0.4	0.4	0.6	0.2	0.2
β_{23}	0.5	0.1	0.1	0.4	0.1	0.1
β_{24}	0.4	0.2	0.2	-0.2	0.1	0.1
β_{25}	0.9	0.1	0.1	0.2	0.1	0.1
β_{26}	0.3	0.1	0.1	0.4	0.1	0.1
Γ_{21}	-2.7	4.3	4.3	1.5	2.8	2.7
Γ_{212}	143.0	66.7	269.9	140.1	39.9	235.5
Γ_{22}	29.1	12.4	20.6	36.8	3.1	16.6
Total			1026.8			1063.8

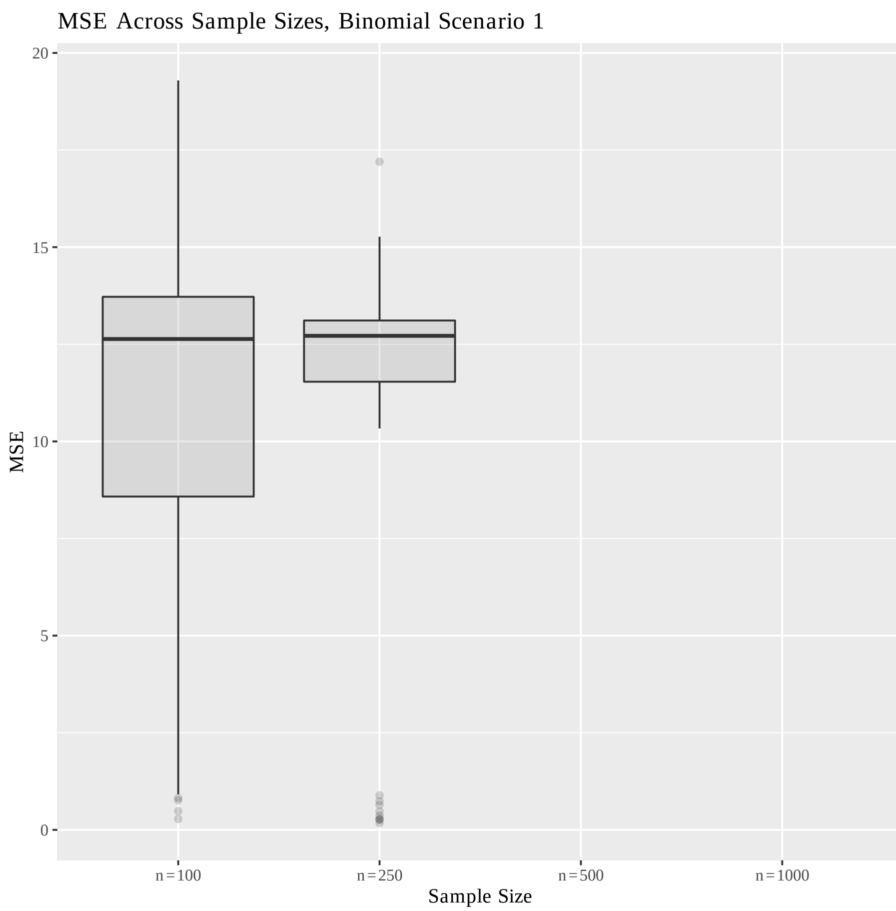

Figure G–3: MSE across sample sizes, binomial outcome, LASSO penalty, with $K = 2$, $p = 7$, $q = 2$.

Table G–4: Simulation 7 results multiplied by 100, averaged over 50 runs. Outcome follows a binomial distribution with $K = 2$, $p = 7$, $q = 2$, $\pi_1 = 0.6$, $\pi_2 = 0.4$, $\boldsymbol{\beta}_1^\top = (0.95, 0.60, 0.00, 0.00, 0.00, 0.00, 0.00)$, $\boldsymbol{\beta}_2^\top = (-0.85, -0.15, 0.00, 0.00, 0.00, 0.00, 0.00)$, $\mathbb{\Gamma}_1^{*\top} = (0.95, 0.85, 1.15)$, and $\mathbb{\Gamma}_2^{*\top} = (0.70, -0.70, 0.80)$. ALASSO penalty.

n	$n = 100$			$n = 250$		
Measure	Bias	Var	MSE	Bias	Var	MSE
π_1	-0.2	0.4	0.4	-1.7	0.4	0.4
β_{10}	-110.0	56.7	176.7	-125.1	43.6	199.2
β_{11}	-49.2	9.2	33.3	-51.1	5.5	31.5
β_{12}	0.1	0.0	0.0	0.3	0.1	0.1
β_{13}	-0.2	0.0	0.0	0.0	0.0	0.0
β_{14}	0.3	0.0	0.0	0.1	0.0	0.0
β_{15}	0.3	0.0	0.0	0.0	0.0	0.0
β_{16}	-0.5	0.1	0.1	0.0	0.0	0.0
Γ_{11}	-16.1	2.4	4.9	-16.5	1.0	3.7
Γ_{112}	-94.0	43.7	131.2	-106.1	26.0	138.1
Γ_{12}	-16.0	3.0	5.5	-14.2	2.2	4.2
β_{20}	170.0	82.3	369.6	181.7	66.3	395.2
β_{21}	74.1	28.5	82.9	84.8	12.6	84.2
β_{22}	1.2	0.4	0.4	0.0	0.0	0.0
β_{23}	0.0	0.0	0.0	-0.1	0.0	0.0
β_{24}	-0.6	0.1	0.1	-0.3	0.0	0.0
β_{25}	0.3	0.1	0.1	-0.1	0.0	0.0
β_{26}	0.4	0.1	0.1	0.4	0.1	0.1
Γ_{21}	-0.7	4.6	4.5	6.8	2.4	2.9
Γ_{212}	137.1	63.8	250.6	139.7	40.5	234.9
Γ_{22}	24.7	10.4	16.4	32.7	3.9	14.5
Total			1076.6			1108.9

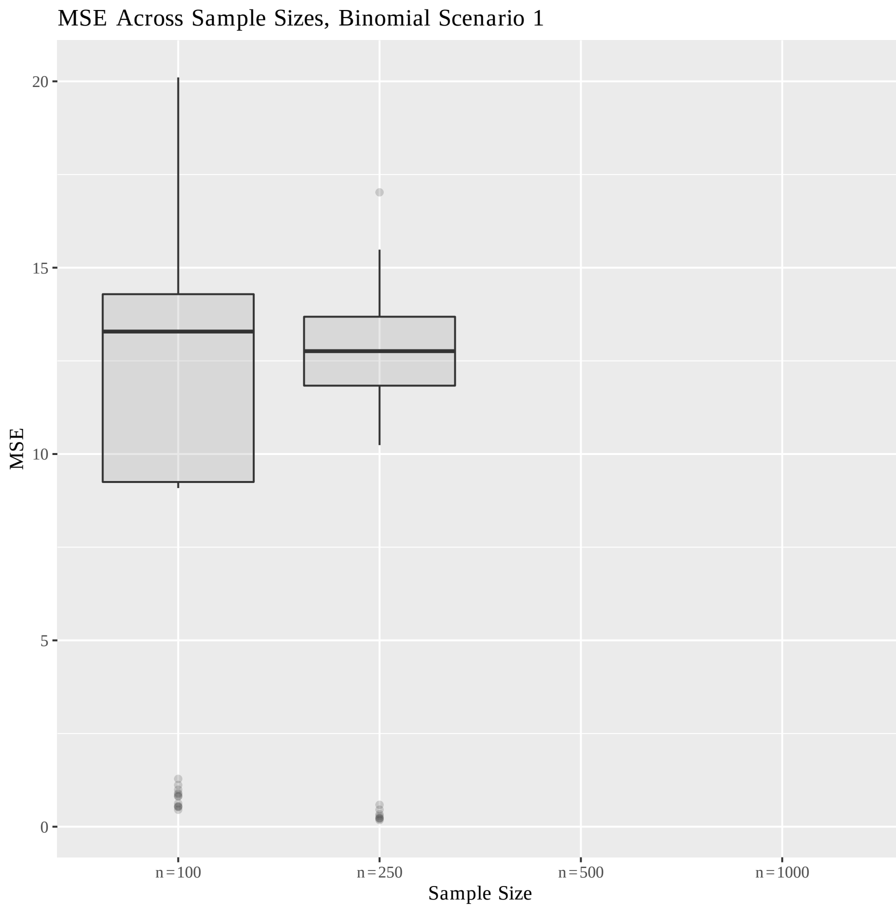

Figure G–4: MSE across sample sizes, binomial outcome, ALASSO penalty, with $K = 2$, $p = 7$, $q = 2$.

APPENDIX H
Additional Tables From Chapter 4 (Bias, Variance, Mean Squared Error With Maximum Likelihood Estimate, Small Values Changed to Zero, Oracle)

This appendix contains additional tables of simulation results from Chapter 4. These tables show the bias, variance, and MSE for the MLE (computed by setting $\lambda = 0$), the MLE with small values (those with a magnitude less than 0.01) changed to zero, and the oracle model.

Table H–1: Simulation 1 results multiplied by 100, averaged over 100 runs. Outcome follows a Poisson distribution with $K = 2$, $p = 7$, $q = 2$, $\pi_1 = 0.6$, $\pi_2 = 0.4$, $\boldsymbol{\beta}_1^\top = (0.65, 0.30, 0.00, 0.00, 0.00, 0.00, 0.00)$, $\boldsymbol{\beta}_2^\top = (0.20, -0.45, 0.00, 0.00, 0.00, 0.00, 0.00)$, $\mathbb{\Gamma}_1^{*\top} = (0.30, -0.25, 0.10)$, and $\mathbb{\Gamma}_2^{*\top} = (0.35, 0.20, 0.15)$. No penalty.

n	$n = 100$			$n = 250$			$n = 500$			$n = 1000$		
Measure	Bias	Var	MSE	Bias	Var	MSE	Bias	Var	MSE	Bias	Var	MSE
π_1	-3.1	1.3	1.4	-1.7	0.7	0.7	-2.3	0.3	0.3	-2.5	0.2	0.2
β_{10}	-2.9	1.5	1.6	-3.3	0.5	0.6	-3.3	0.2	0.4	-2.4	0.1	0.2
β_{11}	-0.8	1.2	1.2	1.2	0.7	0.7	1.2	0.3	0.3	1.4	0.1	0.1
β_{12}	0.1	0.5	0.5	-0.5	0.3	0.3	-0.1	0.1	0.1	-0.1	< 0.1	< 0.1
β_{13}	0.4	0.6	0.6	0.2	0.2	0.2	-0.2	0.1	0.1	-0.1	< 0.1	< 0.1
β_{14}	0.3	0.4	0.4	-0.2	0.2	0.2	-0.1	0.1	0.1	0.2	< 0.1	< 0.1
β_{15}	0.6	0.4	0.4	0.3	0.2	0.2	0.3	0.1	0.1	0.0	< 0.1	< 0.1
β_{16}	1.2	0.5	0.5	0.4	0.1	0.1	0.0	0.1	0.1	0.0	< 0.1	< 0.1
Γ_{11}	-5.0	1.4	1.7	-2.2	0.4	0.5	-0.3	0.2	0.2	-0.1	0.1	0.1
Γ_{112}	5.5	1.9	2.2	2.4	0.5	0.5	-0.2	0.2	0.2	-0.2	0.1	0.1
Γ_{12}	-4.6	0.7	0.9	-5.0	0.5	0.7	-4.0	0.3	0.5	-3.3	0.3	0.4
β_{20}	0.2	5.2	5.2	3.3	1.7	1.8	5.0	0.7	0.9	4.6	0.3	0.5
β_{21}	6.8	3.1	3.5	2.9	1.6	1.6	2.5	0.7	0.7	2.5	0.3	0.3
β_{22}	-1.3	1.3	1.4	0.3	0.6	0.6	0.2	0.2	0.2	0.4	0.1	0.1
β_{23}	-2.1	1.1	1.1	-1.8	0.4	0.4	-0.2	0.1	0.1	-0.2	0.1	0.1
β_{24}	-3.3	1.9	1.9	-0.5	0.6	0.6	-0.1	0.2	0.2	-0.2	0.1	0.1
β_{25}	-2.5	1.3	1.3	-1.9	0.5	0.5	-0.7	0.2	0.2	-0.3	0.1	0.1
β_{26}	-0.4	1.3	1.3	-0.3	0.3	0.3	0.7	0.2	0.2	0.4	0.1	0.1
Γ_{21}	0.3	2.6	2.6	3.3	1.5	1.6	4.0	0.6	0.8	4.0	0.4	0.5
Γ_{212}	-0.4	1.7	1.7	0.9	0.9	0.9	2.2	0.4	0.5	3.1	0.3	0.3
Γ_{22}	-6.6	2.0	2.5	-5.7	1.3	1.6	-3.7	0.6	0.7	-1.5	0.3	0.3
Total			34.0			14.5			6.6			3.8

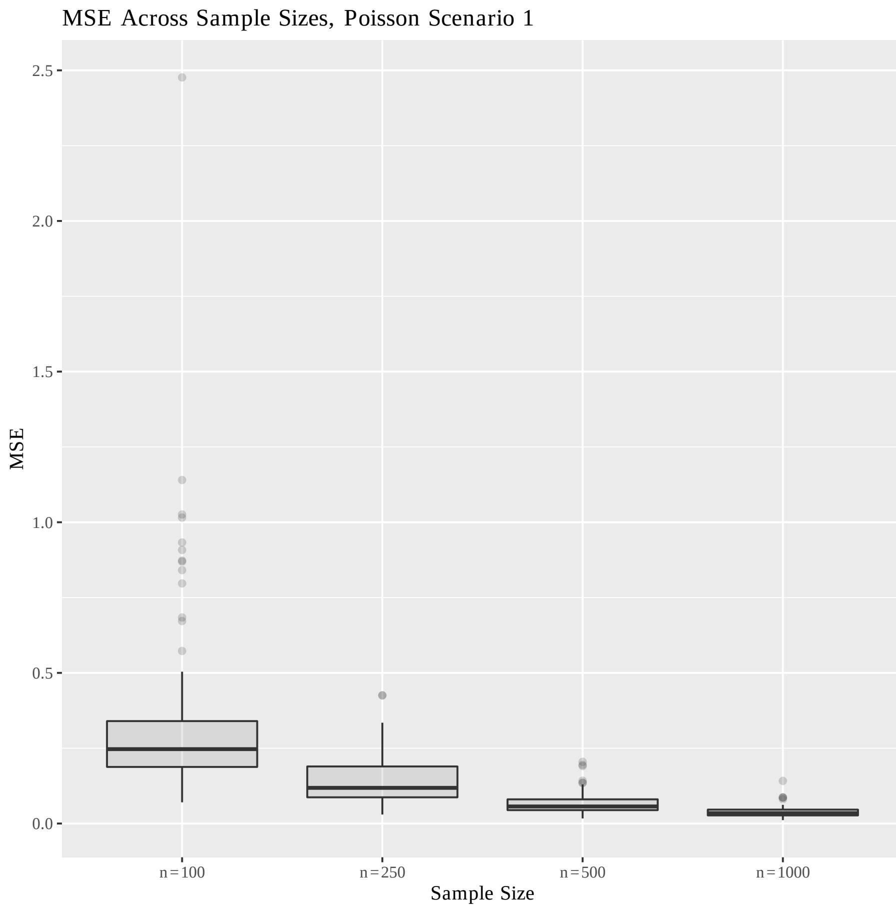

Figure H–1: MSE across sample sizes, Poisson outcome, no penalty, with $K = 2$, $p = 7$, $q = 2$.

Table H–2: Simulation 1 results multiplied by 100, averaged over 100 runs. Outcome follows a Poisson distribution with $K = 2$, $p = 7$, $q = 2$, $\pi_1 = 0.6$, $\pi_2 = 0.4$, $\boldsymbol{\beta}_1^\top = (0.65, 0.30, 0.00, 0.00, 0.00, 0.00, 0.00)$, $\boldsymbol{\beta}_2^\top = (0.20, -0.45, 0.00, 0.00, 0.00, 0.00, 0.00)$, $\mathbb{\Gamma}_1^{*\top} = (0.30, -0.25, 0.10)$, and $\mathbb{\Gamma}_2^{*\top} = (0.35, 0.20, 0.15)$. Small values changed to zero.

n	$n = 100$			$n = 250$			$n = 500$			$n = 1000$		
Measure	Bias	Var	MSE	Bias	Var	MSE	Bias	Var	MSE	Bias	Var	MSE
π_1	-3.1	1.3	1.4	-1.7	0.7	0.7	-2.3	0.3	0.3	-2.5	0.2	0.2
β_{10}	-2.9	1.5	1.6	-3.3	0.5	0.6	-3.3	0.2	0.4	-2.4	0.1	0.2
β_{11}	-0.8	1.2	1.2	1.2	0.7	0.7	1.2	0.3	0.3	1.4	0.1	0.1
β_{12}	0.1	0.5	0.5	-0.5	0.3	0.3	0.0	0.1	0.1	-0.2	< 0.1	< 0.1
β_{13}	0.4	0.6	0.6	0.2	0.2	0.2	-0.2	0.1	0.1	-0.1	< 0.1	< 0.1
β_{14}	0.3	0.4	0.4	-0.2	0.2	0.2	-0.1	0.1	0.1	0.2	< 0.1	< 0.1
β_{15}	0.7	0.4	0.4	0.3	0.2	0.2	0.2	0.1	0.1	0.0	< 0.1	< 0.1
β_{16}	1.2	0.5	0.5	0.4	0.1	0.1	0.0	0.1	0.1	-0.1	< 0.1	< 0.1
Γ_{11}	-5.0	1.4	1.7	-2.2	0.4	0.5	-0.3	0.2	0.2	-0.1	0.1	0.1
Γ_{112}	5.5	1.9	2.2	2.4	0.5	0.5	-0.2	0.2	0.2	-0.2	0.1	0.1
Γ_{12}	-4.6	0.7	0.9	-5.0	0.5	0.7	-4.0	0.3	0.5	-3.3	0.3	0.4
β_{20}	0.2	5.2	5.2	3.3	1.7	1.8	5.0	0.7	0.9	4.6	0.3	0.5
β_{21}	6.8	3.1	3.5	2.9	1.6	1.6	2.5	0.7	0.7	2.5	0.3	0.3
β_{22}	-1.3	1.3	1.4	0.2	0.6	0.6	0.2	0.2	0.2	0.4	0.1	0.1
β_{23}	-2.1	1.1	1.1	-1.8	0.4	0.4	-0.2	0.1	0.1	-0.3	0.1	0.1
β_{24}	-3.3	1.9	1.9	-0.5	0.6	0.6	-0.1	0.2	0.2	-0.3	0.1	0.1
β_{25}	-2.5	1.3	1.3	-1.9	0.5	0.5	-0.7	0.2	0.2	-0.3	0.1	0.1
β_{26}	-0.4	1.3	1.3	-0.3	0.3	0.3	0.7	0.2	0.2	0.5	0.1	0.1
Γ_{21}	0.3	2.6	2.6	3.3	1.5	1.6	4.0	0.6	0.8	4.0	0.4	0.5
Γ_{212}	-0.4	1.7	1.7	0.9	0.9	0.9	2.2	0.4	0.5	3.1	0.3	0.3
Γ_{22}	-6.6	2.0	2.5	-5.7	1.3	1.6	-3.7	0.6	0.7	-1.5	0.3	0.3
Total			34.0			14.5			6.6			3.8

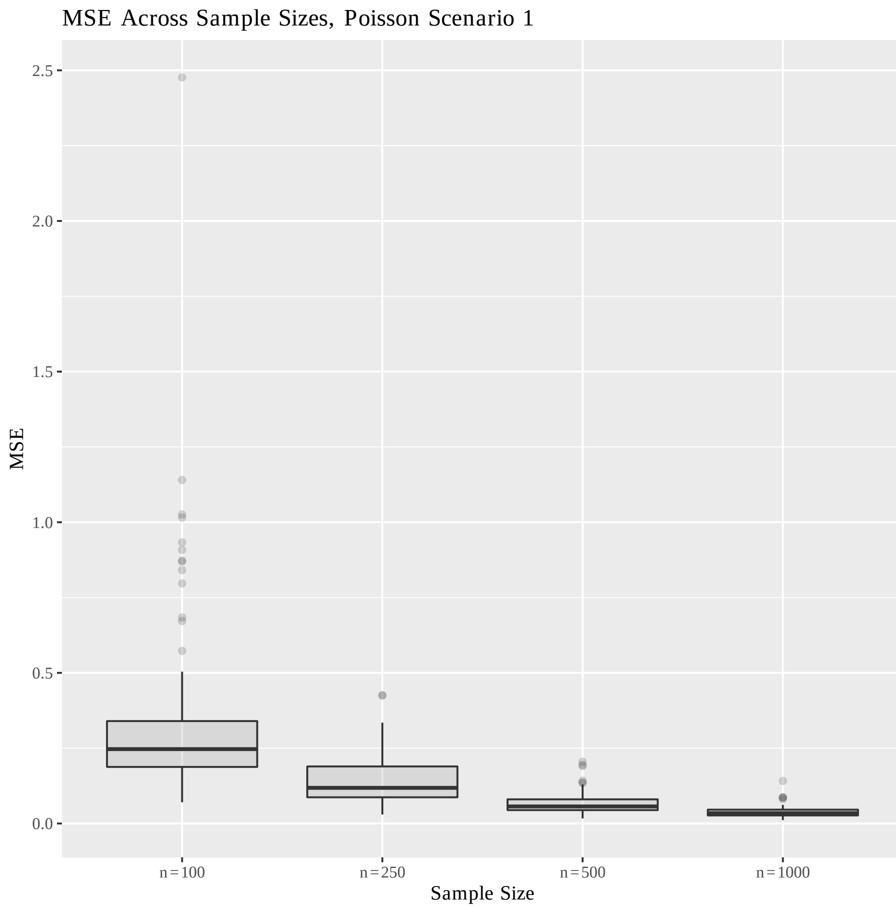

Figure H–2: MSE across sample sizes, Poisson outcome, small values changed to zero, with $K = 2$, $p = 7$, $q = 2$.

Table H–3: Simulation 1 results multiplied by 100, averaged over 100 runs. Outcome follows a Poisson distribution with $K = 2$, $p = 2$, $q = 2$, $\pi_1 = 0.6$, $\pi_2 = 0.4$, $\boldsymbol{\beta}_1^\top = (0.65, 0.30)$, $\boldsymbol{\beta}_2^\top = (0.20, -0.45)$, $\mathbb{\Gamma}_1^{*\top} = (0.30, -0.25, 0.10)$, and $\mathbb{\Gamma}_2^{*\top} = (0.35, 0.20, 0.15)$. Oracle model.

n	$n = 100$			$n = 250$			$n = 500$			$n = 1000$		
Measure	Bias	Var	MSE	Bias	Var	MSE	Bias	Var	MSE	Bias	Var	MSE
π_1	-2.0	1.0	1.0	-0.6	0.6	0.6	-1.7	0.2	0.3	-1.9	0.2	0.2
β_{10}	-2.0	1.0	1.0	-2.5	0.5	0.5	-2.7	0.3	0.3	-2.1	0.1	0.2
β_{11}	0.4	1.1	1.1	0.6	0.5	0.5	0.9	0.2	0.2	1.2	0.1	0.1
Γ_{11}	-2.8	1.1	1.1	-1.8	0.4	0.4	-0.4	0.2	0.2	0.1	0.1	0.1
Γ_{112}	5.2	1.2	1.4	2.2	0.5	0.5	-0.1	0.2	0.2	-0.2	0.1	0.1
Γ_{12}	-3.2	0.8	0.9	-3.6	0.5	0.7	-3.6	0.3	0.4	-2.7	0.3	0.3
β_{20}	1.7	4.0	4.0	2.7	1.4	1.5	4.2	0.7	0.8	3.9	0.3	0.5
β_{21}	2.9	2.8	2.9	1.9	1.3	1.3	2.2	0.6	0.7	2.0	0.2	0.3
Γ_{21}	1.5	2.5	2.5	3.4	1.4	1.5	3.6	0.6	0.8	3.3	0.4	0.5
Γ_{212}	-2.6	1.6	1.7	0.4	1.0	1.0	1.4	0.5	0.5	2.7	0.3	0.4
Γ_{22}	-4.9	1.9	2.2	-4.6	1.1	1.3	-2.9	0.8	0.8	-1.6	0.3	0.3
Total			19.7			9.8			5.2			2.9

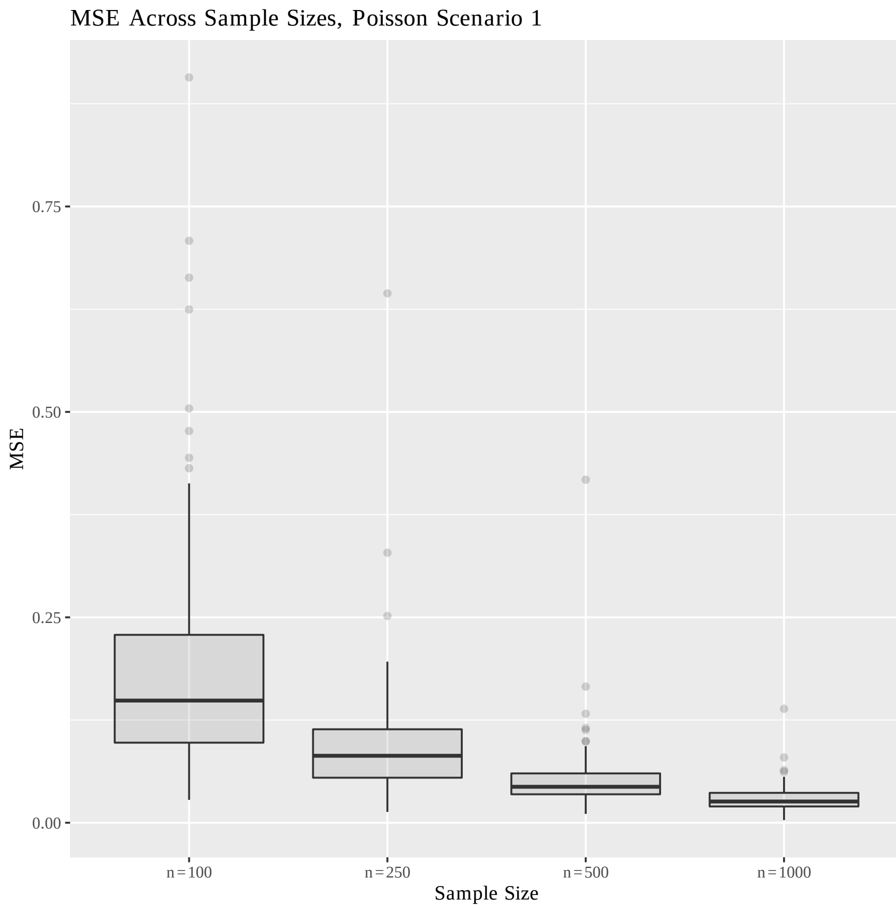

Figure H–3: MSE across sample sizes, Poisson outcome, oracle model, with $K = 2$, $p = 7$, $q = 2$.

Table H–4: Simulation 2 results multiplied by 100, averaged over 100 runs. Outcome follows a Poisson distribution with $K = 2$, $p = 17$, $q = 2$, $\pi_1 = 0.6$, $\pi_2 = 0.4$, $\boldsymbol{\beta}_1^\top = (0.65, 0.30, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00)$, $\boldsymbol{\beta}_2^\top = (0.20, -0.45, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00)$, $\Gamma_1^{*\top} = (0.30, -0.25, 0.10)$, and $\Gamma_2^{*\top} = (0.35, 0.20, 0.15)$. No penalty, part 1.

n	$n = 100$			$n = 250$			$n = 500$			$n = 1000$		
Measure	Bias	Var	MSE	Bias	Var	MSE	Bias	Var	MSE	Bias	Var	MSE
π_1	-6.1	1.7	2.0	-4.6	0.8	1.0	-2.9	0.4	0.5	-3.4	0.2	0.3
β_{10}	-3.3	2.1	2.2	-3.7	0.8	1.0	-2.1	0.2	0.3	-1.7	0.1	0.2
β_{11}	-5.1	2.9	3.1	-0.3	1.2	1.2	1.0	0.4	0.4	1.1	0.2	0.2
β_{12}	0.1	0.5	0.5	0.0	0.2	0.2	0.1	0.1	0.1	0.2	< 0.1	< 0.1
β_{13}	-0.2	0.6	0.6	-0.3	0.2	0.2	-0.1	0.1	0.1	-0.2	< 0.1	< 0.1
β_{14}	-0.1	0.7	0.6	0.1	0.2	0.2	-0.4	0.1	0.1	-0.3	< 0.1	< 0.1
β_{15}	0.2	0.7	0.7	0.6	0.2	0.2	0.1	0.1	0.1	-0.1	< 0.1	< 0.1
β_{16}	-1.5	0.5	0.5	-1.0	0.2	0.2	-0.3	0.1	0.1	-0.2	< 0.1	< 0.1
β_{17}	0.8	0.6	0.6	0.0	0.2	0.2	-0.1	0.1	0.1	-0.2	< 0.1	< 0.1
β_{18}	-0.1	0.5	0.5	-0.3	0.2	0.2	-0.2	0.1	0.1	0.0	< 0.1	< 0.1
β_{19}	-0.4	0.5	0.5	-0.3	0.2	0.2	0.4	0.1	0.1	0.2	< 0.1	< 0.1
β_{110}	0.6	0.7	0.7	-0.7	0.2	0.2	-0.3	0.1	0.1	0.0	< 0.1	< 0.1
β_{111}	0.2	0.5	0.5	-0.2	0.2	0.2	0.0	0.1	0.1	0.2	< 0.1	< 0.1
β_{112}	0.3	0.4	0.4	-0.2	0.2	0.2	0.1	0.1	0.1	0.0	< 0.1	< 0.1
β_{113}	-0.3	0.8	0.8	0.4	0.2	0.2	0.0	0.1	0.1	0.0	< 0.1	< 0.1
β_{114}	-0.3	0.6	0.6	0.0	0.1	0.1	-0.5	0.1	0.1	-0.4	< 0.1	< 0.1
β_{115}	-0.2	0.5	0.5	0.2	0.2	0.2	0.0	0.1	0.1	0.0	< 0.1	< 0.1
β_{116}	-0.2	0.6	0.6	-0.6	0.2	0.2	0.0	0.1	0.1	-0.1	< 0.1	< 0.1
Γ_{11}	-10.3	1.6	2.7	-6.3	1.0	1.4	-2.9	0.7	0.8	-1.0	0.1	0.1
Γ_{112}	9.0	1.9	2.7	5.0	1.4	1.6	1.5	0.6	0.6	-0.4	0.1	0.1
Γ_{12}	-5.7	0.7	1.0	-4.8	0.7	0.9	-4.7	0.3	0.5	-4.9	0.2	0.4

Table H–5: Simulation 2 results multiplied by 100, averaged over 100 runs. Outcome follows a Poisson distribution with $K = 2$, $p = 17$, $q = 2$, $\pi_1 = 0.6$, $\pi_2 = 0.4$, $\boldsymbol{\beta}_1^\top = (0.65, 0.30, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00)$, $\boldsymbol{\beta}_2^\top = (0.20, -0.45, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00)$, $\mathbb{\Gamma}_1^{*\top} = (0.30, -0.25, 0.10)$, and $\mathbb{\Gamma}_2^{*\top} = (0.35, 0.20, 0.15)$. No penalty, part 2.

n	$n = 100$			$n = 250$			$n = 500$			$n = 1000$		
Measure	Bias	Var	MSE	Bias	Var	MSE	Bias	Var	MSE	Bias	Var	MSE
β_{20}	-2.2	4.0	4.0	5.4	1.8	2.0	3.7	0.9	1.1	4.9	0.4	0.6
β_{21}	14.1	4.7	6.6	6.1	2.1	2.5	2.7	0.8	0.8	3.7	0.3	0.5
β_{22}	0.6	1.6	1.6	0.1	0.5	0.5	0.3	0.3	0.3	0.4	0.1	0.1
β_{23}	0.2	1.2	1.2	-0.5	0.3	0.3	-1.1	0.2	0.2	-0.2	0.1	0.1
β_{24}	0.6	1.3	1.3	0.0	0.3	0.3	-0.1	0.2	0.2	0.0	0.1	0.1
β_{25}	0.4	1.0	1.0	-0.8	0.3	0.3	-0.2	0.2	0.2	-0.1	0.1	0.1
β_{26}	2.9	1.2	1.2	1.2	0.4	0.4	0.3	0.2	0.2	0.0	0.1	0.1
β_{27}	-0.3	1.7	1.7	0.0	0.3	0.3	0.2	0.2	0.2	0.0	0.1	0.1
β_{28}	-0.4	1.3	1.3	-0.3	0.4	0.4	-0.3	0.2	0.1	-0.4	0.1	0.1
β_{29}	1.2	1.0	1.0	1.3	0.4	0.5	0.1	0.2	0.2	0.3	0.1	0.1
β_{210}	-0.5	1.4	1.4	0.0	0.4	0.4	-0.2	0.2	0.2	-0.2	0.1	0.1
β_{211}	0.1	1.4	1.4	-0.1	0.4	0.4	0.0	0.2	0.2	-0.2	0.1	0.1
β_{212}	0.2	0.8	0.8	0.0	0.3	0.3	0.0	0.1	0.1	0.0	0.1	0.1
β_{213}	1.8	1.3	1.3	0.7	0.5	0.5	0.3	0.2	0.2	0.0	0.1	0.1
β_{214}	0.3	1.0	1.0	0.8	0.3	0.3	0.6	0.2	0.2	0.1	0.1	0.1
β_{215}	1.1	1.3	1.3	0.0	0.3	0.3	0.2	0.1	0.1	0.1	0.1	0.1
β_{216}	0.3	1.6	1.6	-0.1	0.4	0.4	-0.2	0.2	0.2	0.1	0.1	0.1
Γ_{21}	-4.7	3.2	3.4	1.4	1.6	1.6	2.1	0.9	0.9	2.3	0.4	0.5
Γ_{212}	-0.6	2.5	2.5	1.5	1.1	1.1	0.3	0.8	0.8	1.1	0.3	0.3
Γ_{22}	-5.9	2.1	2.4	-6.7	1.0	1.4	-6.2	0.5	0.9	-1.6	0.3	0.3
Total			60.3			24.1			11.3			5.3

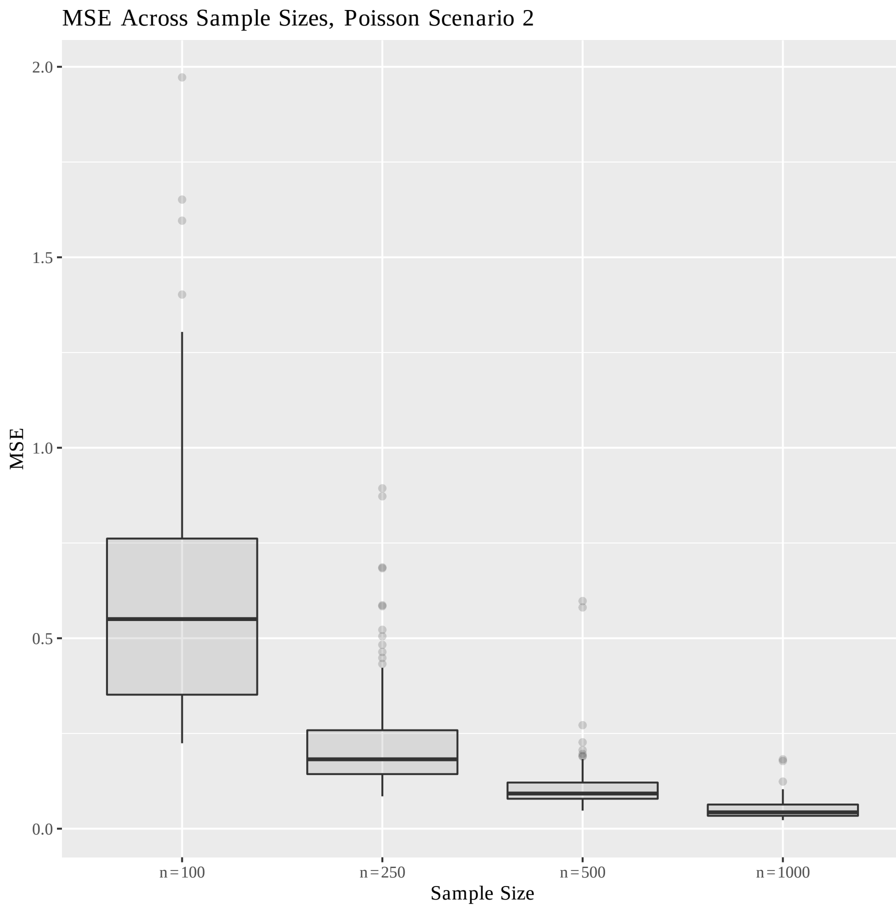

Figure H–4: MSE across sample sizes, Poisson outcome, no penalty, with $K = 2$, $p = 17$, $q = 2$.

Table H–6: Simulation 2 results multiplied by 100, averaged over 100 runs. Outcome follows a Poisson distribution with $K = 2$, $p = 17$, $q = 2$, $\pi_1 = 0.6$, $\pi_2 = 0.4$, $\boldsymbol{\beta}_1^\top = (0.65, 0.30, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00)$, $\boldsymbol{\beta}_2^\top = (0.20, -0.45, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00)$, $\mathbb{\Gamma}_1^{*\top} = (0.30, -0.25, 0.10)$, and $\mathbb{\Gamma}_2^{*\top} = (0.35, 0.20, 0.15)$. Small values changed to zero, part 1.

n	$n = 100$			$n = 250$			$n = 500$			$n = 1000$		
Measure	Bias	Var	MSE	Bias	Var	MSE	Bias	Var	MSE	Bias	Var	MSE
π_1	-6.1	1.7	2.0	-4.6	0.8	1.0	-2.9	0.4	0.5	-3.4	0.2	0.3
β_{10}	-3.3	2.1	2.2	-3.7	0.8	1.0	-2.1	0.2	0.3	-1.7	0.1	0.2
β_{11}	-5.1	2.9	3.1	-0.3	1.2	1.2	1.0	0.4	0.4	1.1	0.2	0.2
β_{12}	0.1	0.5	0.5	0.0	0.2	0.2	0.1	0.1	0.1	0.2	< 0.1	< 0.1
β_{13}	-0.2	0.6	0.6	-0.3	0.2	0.2	-0.1	0.1	0.1	-0.2	< 0.1	< 0.1
β_{14}	-0.2	0.7	0.6	0.1	0.2	0.2	-0.3	0.1	0.1	-0.3	< 0.1	< 0.1
β_{15}	0.2	0.7	0.7	0.6	0.2	0.2	0.2	0.1	0.1	-0.1	< 0.1	< 0.1
β_{16}	-1.5	0.5	0.5	-1.0	0.2	0.2	-0.3	0.1	0.1	-0.2	< 0.1	< 0.1
β_{17}	0.8	0.6	0.6	0.0	0.2	0.2	-0.1	0.1	0.1	-0.2	< 0.1	< 0.1
β_{18}	-0.1	0.5	0.5	-0.4	0.2	0.2	-0.2	0.1	0.1	-0.1	< 0.1	< 0.1
β_{19}	-0.4	0.5	0.5	-0.3	0.2	0.2	0.4	0.1	0.1	0.2	< 0.1	< 0.1
β_{110}	0.6	0.7	0.7	-0.6	0.2	0.2	-0.3	0.1	0.1	0.0	< 0.1	< 0.1
β_{111}	0.2	0.5	0.5	-0.2	0.2	0.2	0.0	0.1	0.1	0.2	< 0.1	< 0.1
β_{112}	0.4	0.4	0.4	-0.2	0.2	0.2	0.1	0.1	0.1	0.0	< 0.1	< 0.1
β_{113}	-0.3	0.8	0.8	0.3	0.2	0.2	0.1	0.1	0.1	0.1	< 0.1	< 0.1
β_{114}	-0.3	0.6	0.6	0.0	0.1	0.1	-0.5	0.1	0.1	-0.4	< 0.1	< 0.1
β_{115}	-0.2	0.5	0.5	0.2	0.2	0.2	0.0	0.1	0.1	0.0	< 0.1	< 0.1
β_{116}	-0.2	0.6	0.6	-0.6	0.2	0.2	0.0	0.1	0.1	-0.1	< 0.1	< 0.1
Γ_{11}	-10.3	1.6	2.7	-6.3	1.0	1.4	-2.9	0.7	0.8	-1.0	0.1	0.1
Γ_{112}	9.0	1.9	2.7	5.0	1.4	1.6	1.5	0.6	0.6	-0.4	0.1	0.1
Γ_{12}	-5.7	0.7	1.0	-4.8	0.7	0.9	-4.7	0.3	0.5	-4.9	0.2	0.4

Table H–7: Simulation 2 results multiplied by 100, averaged over 100 runs. Outcome follows a Poisson distribution with $K = 2$, $p = 17$, $q = 2$, $\pi_1 = 0.6$, $\pi_2 = 0.4$, $\boldsymbol{\beta}_1^\top = (0.65, 0.30, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00)$, $\boldsymbol{\beta}_2^\top = (0.20, -0.45, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00)$, $\mathbb{\Gamma}_1^{*\top} = (0.30, -0.25, 0.10)$, and $\mathbb{\Gamma}_2^{*\top} = (0.35, 0.20, 0.15)$. Small values changed to zero,part 2.

n	$n = 100$			$n = 250$			$n = 500$			$n = 1000$		
Measure	Bias	Var	MSE	Bias	Var	MSE	Bias	Var	MSE	Bias	Var	MSE
β_{20}	-2.2	4.0	4.0	5.4	1.8	2.0	3.7	0.9	1.1	4.9	0.4	0.6
β_{21}	14.2	4.7	6.6	6.1	2.1	2.5	2.7	0.8	0.8	3.7	0.3	0.5
β_{22}	0.6	1.6	1.6	0.0	0.5	0.5	0.3	0.3	0.3	0.4	0.1	0.1
β_{23}	0.1	1.2	1.2	-0.5	0.3	0.3	-1.1	0.2	0.2	-0.2	0.1	0.1
β_{24}	0.6	1.3	1.3	0.0	0.3	0.3	0.0	0.2	0.2	0.0	0.1	0.1
β_{25}	0.3	1.0	1.0	-0.8	0.3	0.3	-0.3	0.2	0.2	-0.1	0.1	0.1
β_{26}	2.9	1.2	1.2	1.2	0.4	0.4	0.3	0.2	0.2	0.0	0.1	0.1
β_{27}	-0.3	1.7	1.7	0.0	0.3	0.3	0.2	0.2	0.2	0.1	0.1	0.1
β_{28}	-0.4	1.3	1.3	-0.3	0.4	0.4	-0.3	0.1	0.1	-0.3	0.1	0.1
β_{29}	1.2	1.0	1.0	1.4	0.4	0.5	0.1	0.2	0.2	0.2	0.1	0.1
β_{210}	-0.5	1.4	1.4	0.0	0.4	0.4	-0.2	0.2	0.2	-0.2	0.1	0.1
β_{211}	0.1	1.4	1.4	-0.1	0.4	0.4	0.0	0.2	0.2	-0.2	0.1	0.1
β_{212}	0.2	0.8	0.8	0.0	0.3	0.3	-0.1	0.1	0.1	0.0	0.1	0.1
β_{213}	1.8	1.3	1.3	0.6	0.5	0.5	0.3	0.2	0.2	0.0	0.1	0.1
β_{214}	0.3	1.0	1.0	0.8	0.3	0.3	0.6	0.2	0.2	0.1	0.1	0.1
β_{215}	1.1	1.3	1.3	0.0	0.3	0.3	0.3	0.1	0.1	0.2	0.1	0.1
β_{216}	0.3	1.6	1.6	-0.1	0.4	0.3	-0.1	0.2	0.2	0.1	0.1	0.1
Γ_{21}	-4.7	3.2	3.4	1.4	1.6	1.6	2.1	0.9	0.9	2.3	0.4	0.5
Γ_{212}	-0.6	2.5	2.5	1.5	1.1	1.1	0.3	0.8	0.8	1.1	0.3	0.3
Γ_{22}	-5.9	2.1	2.4	-6.7	1.0	1.4	-6.2	0.5	0.9	-1.6	0.3	0.3
Total			60.3			24.1			11.3			5.2

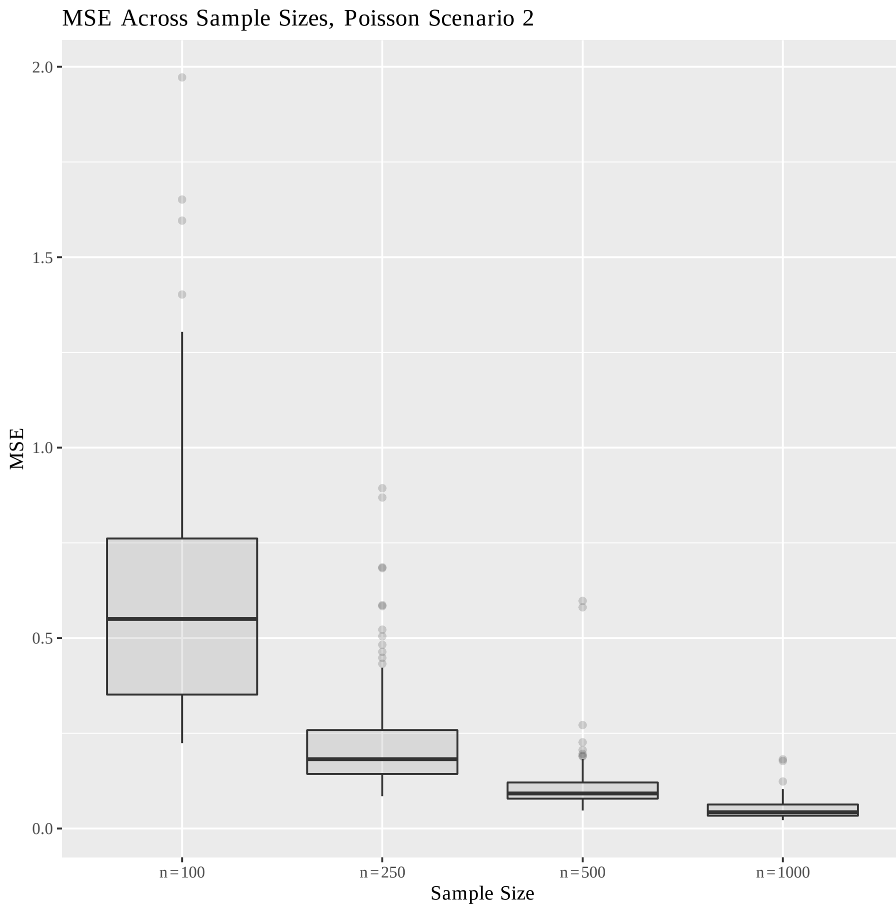

Figure H–5: MSE across sample sizes, Poisson outcome, small values changed to zero, with $K = 2$, $p = 17$, $q = 2$.

Table H–8: Simulation 2 results multiplied by 100, averaged over 100 runs. Outcome follows a Poisson distribution with $K = 2$, $p = 2$, $q = 2$, $\pi_1 = 0.6$, $\pi_2 = 0.4$, $\boldsymbol{\beta}_1^\top = (0.65, 0.30)$, $\boldsymbol{\beta}_2^\top = (0.20, -0.45)$, $\mathbb{\Gamma}_1^{*\top} = (0.30, -0.25, 0.10)$, and $\mathbb{\Gamma}_2^{*\top} = (0.35, 0.20, 0.15)$. Oracle model.

n	$n = 100$			$n = 250$			$n = 500$			$n = 1000$		
Measure	Bias	Var	MSE	Bias	Var	MSE	Bias	Var	MSE	Bias	Var	MSE
π_1	-3.8	0.9	1.1	-2.5	0.6	0.6	-2.4	0.3	0.4	-2.2	0.2	0.2
β_{10}	-0.5	1.3	1.2	-0.8	0.5	0.5	-0.9	0.2	0.2	-1.1	0.1	0.1
β_{11}	0.7	1.1	1.1	1.0	0.5	0.5	1.1	0.3	0.3	0.8	0.1	0.1
Γ_{11}	-5.7	1.3	1.6	-2.8	0.3	0.4	-1.7	0.4	0.4	-0.7	0.1	0.1
Γ_{112}	5.6	1.2	1.5	1.6	0.3	0.3	0.7	0.4	0.4	-0.4	0.1	0.1
Γ_{12}	-5.1	0.5	0.8	-4.5	0.4	0.6	-4.4	0.3	0.5	-4.1	0.2	0.4
β_{20}	0.9	2.7	2.6	2.5	1.2	1.2	3.2	0.6	0.7	3.5	0.3	0.5
β_{21}	4.2	2.7	2.8	1.5	1.3	1.3	2.1	0.5	0.6	2.5	0.2	0.3
Γ_{21}	-0.9	3.0	3.0	-0.3	1.5	1.5	1.8	0.6	0.6	1.3	0.3	0.4
Γ_{212}	-4.8	1.5	1.7	-1.4	0.9	0.9	-0.7	0.5	0.5	0.1	0.3	0.3
Γ_{22}	-5.3	2.0	2.3	-4.4	1.0	1.2	-3.9	0.5	0.7	-1.1	0.2	0.2
Total			19.7			9.1			5.2			2.7

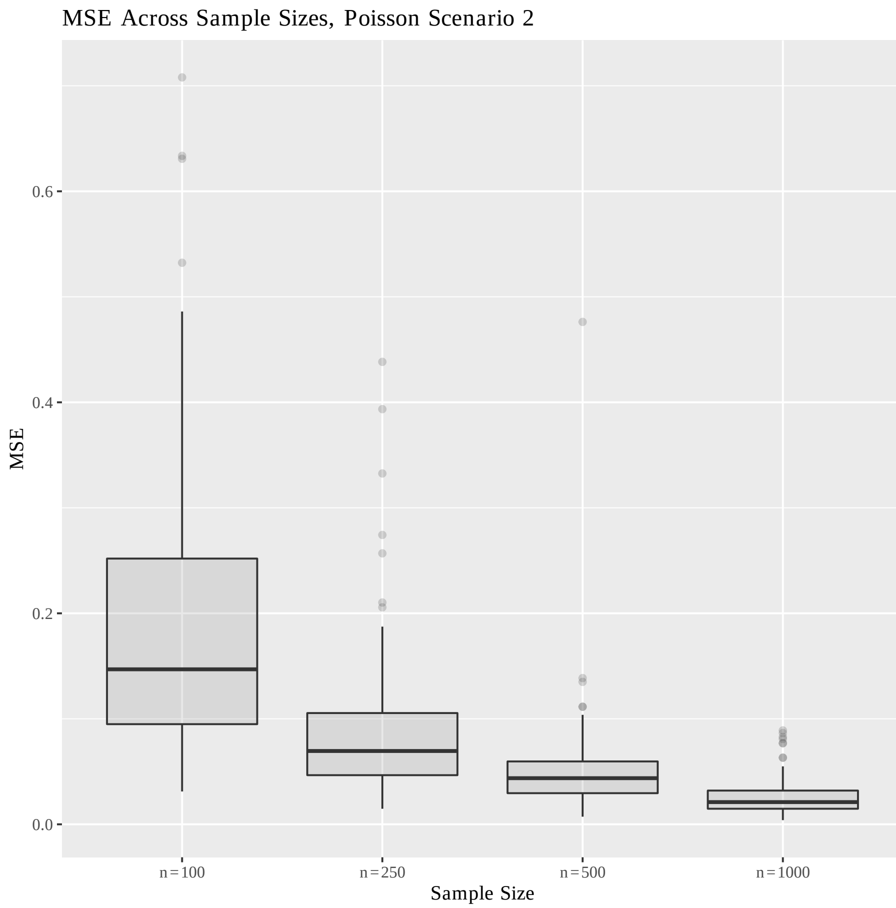

Figure H–6: MSE across sample sizes, Poisson outcome, oracle model, with $K = 2$, $p = 17$, $q = 2$.

Table H–9: Simulation 3 results multiplied by 100, averaged over 100 runs. Outcome follows a Poisson distribution with $K = 2$, $p = 19$, $q = 5$, $\pi_1 = 0.6$, $\pi_2 = 0.4$, $\boldsymbol{\beta}_1^\top = (0.65, 0.30, 0.15, 0.35, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00)$, $\boldsymbol{\beta}_2^\top = (0.20, -0.45, -0.10, 0.25, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00)$, $\mathbb{\Gamma}_1^{*\top} = (0.30, -0.25, 0.10, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00)$, and $\mathbb{\Gamma}_2^{*\top} = (0.35, 0.20, 0.15, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00)$. No penalty, part 1.

n	$n = 100$			$n = 250$			$n = 500$			$n = 1000$		
Measure	Bias	Var	MSE	Bias	Var	MSE	Bias	Var	MSE	Bias	Var	MSE
π_1	-3.2	1.0	1.1	-1.5	0.5	0.5	-1.2	0.2	0.2	-1.5	0.1	0.2
β_{10}	-6.0	1.7	2.1	-1.7	0.4	0.4	-1.3	0.2	0.2	-1.0	0.1	0.1
β_{11}	-2.5	1.5	1.6	0.5	0.5	0.5	1.3	0.2	0.2	1.2	0.1	0.1
β_{12}	-1.4	1.0	1.1	0.0	0.2	0.2	-0.3	0.1	0.1	0.0	< 0.1	< 0.1
β_{13}	0.7	0.6	0.6	0.6	0.2	0.2	-0.1	0.1	0.1	-0.8	< 0.1	< 0.1
β_{14}	-0.7	0.6	0.6	-0.1	0.1	0.1	-0.3	0.1	0.1	-0.2	< 0.1	< 0.1
β_{15}	-0.7	0.5	0.5	0.1	0.1	0.1	0.1	< 0.1	< 0.1	0.0	< 0.1	< 0.1
β_{16}	-0.3	0.4	0.4	-0.1	0.1	0.1	0.2	< 0.1	< 0.1	0.0	< 0.1	< 0.1
β_{17}	0.4	0.5	0.5	0.1	0.1	0.1	0.0	< 0.1	< 0.1	0.0	< 0.1	< 0.1
β_{18}	-0.6	0.4	0.4	-0.7	0.1	0.1	-0.1	0.1	0.1	-0.1	< 0.1	< 0.1
β_{19}	0.4	0.5	0.5	0.2	0.1	0.1	0.2	0.1	0.1	0.2	< 0.1	< 0.1
β_{110}	-0.4	0.5	0.5	0.0	0.1	0.1	-0.2	< 0.1	< 0.1	-0.2	< 0.1	< 0.1
β_{111}	-0.6	0.4	0.4	0.4	0.1	0.1	0.2	0.1	0.1	-0.1	< 0.1	< 0.1
β_{112}	-0.4	0.5	0.5	0.0	0.1	0.1	0.0	< 0.1	< 0.1	-0.1	< 0.1	< 0.1
β_{113}	0.1	0.4	0.4	0.4	0.1	0.1	0.0	< 0.1	< 0.1	0.1	< 0.1	< 0.1
β_{114}	-0.3	0.5	0.5	-0.4	0.1	0.1	-0.2	< 0.1	< 0.1	-0.2	< 0.1	< 0.1
β_{115}	-0.8	0.5	0.5	-0.5	0.1	0.1	-0.1	< 0.1	< 0.1	-0.1	< 0.1	< 0.1
β_{116}	0.7	0.4	0.4	0.4	0.1	0.1	0.0	0.1	0.1	0.3	< 0.1	< 0.1
β_{117}	-0.7	0.6	0.6	0.0	0.1	0.1	-0.2	0.1	0.1	-0.2	< 0.1	< 0.1
β_{118}	0.4	0.5	0.5	0.1	0.1	0.1	0.2	0.1	0.1	0.3	< 0.1	< 0.1
Γ_{11}	-10.3	2.2	3.2	-4.8	1.0	1.3	-2.9	0.5	0.6	-1.4	0.1	0.1
Γ_{112}	12.8	2.7	4.3	4.4	1.0	1.1	2.2	0.4	0.5	1.1	0.2	0.2
Γ_{12}	-1.7	1.1	1.1	-2.6	0.6	0.7	-2.9	0.3	0.4	-2.8	0.2	0.3
Γ_{113}	-0.7	0.6	0.6	-0.6	0.2	0.2	-0.1	0.1	0.1	-0.4	0.1	0.1
Γ_{123}	1.3	0.6	0.6	-0.3	0.4	0.4	0.1	0.3	0.3	-0.3	0.2	0.2
Γ_{13}	0.9	0.1	0.1	1.4	< 0.1	0.1	2.5	0.1	0.1	2.6	< 0.1	0.1
Γ_{114}	-1.3	0.4	0.5	-0.4	0.2	0.2	-0.1	0.1	0.1	-0.1	< 0.1	< 0.1
Γ_{124}	-0.1	0.3	0.3	-0.8	0.2	0.2	-1.0	0.1	0.2	-0.7	0.1	0.1
Γ_{134}	0.1	< 0.1	< 0.1	-0.3	0.1	0.1	-0.2	0.1	0.1	0.5	< 0.1	< 0.1
Γ_{14}	0.0	< 0.1	< 0.1	0.6	< 0.1	< 0.1	0.9	< 0.1	< 0.1	1.2	< 0.1	< 0.1
Γ_{115}	-0.9	0.5	0.5	-0.1	0.2	0.2	0.1	0.1	0.1	0.1	< 0.1	< 0.1
Γ_{125}	1.1	0.2	0.2	0.2	0.2	0.2	0.1	0.1	0.1	0.0	0.1	0.1
Γ_{135}	0.0	< 0.1	< 0.1	-0.3	0.1	0.1	-0.5	0.1	0.1	0.0	< 0.1	< 0.1
Γ_{145}	0.1	< 0.1	< 0.1	0.2	< 0.1	< 0.1	0.2	< 0.1	< 0.1	0.0	< 0.1	< 0.1
Γ_{15}	0.1	< 0.1	< 0.1	0.2	< 0.1	< 0.1	0.4	< 0.1	< 0.1	0.7	< 0.1	< 0.1

Table H–10: Simulation 3 results multiplied by 100, averaged over 100 runs. Outcome follows a Poisson distribution with $K = 2$, $p = 19$, $q = 5$, $\pi_1 = 0.6$, $\pi_2 = 0.4$, $\boldsymbol{\beta}_1^\top = (0.65, 0.30, 0.15, 0.35, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00)$, $\boldsymbol{\beta}_2^\top = (0.20, -0.45, -0.10, 0.25, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00)$, $\mathbb{\Gamma}_1^{*\top} = (0.30, -0.25, 0.10, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00)$, and $\mathbb{\Gamma}_2^{*\top} = (0.35, 0.20, 0.15, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00)$. No penalty, part 2.

n	$n = 100$			$n = 250$			$n = 500$			$n = 1000$		
Measure	Bias	Var	MSE	Bias	Var	MSE	Bias	Var	MSE	Bias	Var	MSE
β_{20}	-3.3	4.7	4.8	-2.9	1.5	1.5	1.3	0.6	0.6	0.6	0.3	0.3
β_{21}	6.2	5.4	5.7	1.4	1.3	1.3	-0.5	0.4	0.4	0.2	0.2	0.2
β_{22}	4.8	2.0	2.2	1.7	0.5	0.6	1.9	0.2	0.2	1.6	0.1	0.1
β_{23}	5.0	1.4	1.7	1.9	0.4	0.4	0.3	0.2	0.2	0.1	0.1	0.1
β_{24}	0.2	0.9	0.9	-0.4	0.3	0.3	-0.2	0.1	0.1	0.0	< 0.1	< 0.1
β_{25}	1.6	1.5	1.5	0.4	0.5	0.5	0.1	0.1	0.1	0.2	0.1	0.1
β_{26}	-1.0	1.2	1.2	0.5	0.3	0.3	0.1	0.1	0.1	0.2	0.1	0.1
β_{27}	1.2	1.3	1.3	-0.7	0.3	0.3	0.2	0.2	0.2	0.3	0.1	0.1
β_{28}	0.0	1.1	1.1	1.0	0.4	0.4	0.3	0.2	0.2	-0.1	0.1	0.1
β_{29}	0.0	1.8	1.8	0.8	0.4	0.4	0.3	0.2	0.2	0.0	0.1	0.1
β_{210}	0.1	1.4	1.4	0.3	0.4	0.4	0.2	0.2	0.2	0.1	0.1	0.1
β_{211}	-0.3	1.3	1.3	-2.1	0.3	0.3	-0.7	0.1	0.1	0.4	0.1	0.1
β_{212}	0.0	1.4	1.4	0.5	0.3	0.3	0.2	0.2	0.2	0.0	0.1	0.1
β_{213}	1.2	1.3	1.3	-0.2	0.4	0.4	0.3	0.1	0.1	0.3	0.1	0.1
β_{214}	0.9	1.1	1.1	0.0	0.3	0.3	-0.2	0.1	0.1	0.0	0.1	0.1
β_{215}	-1.3	1.2	1.2	0.0	0.4	0.4	0.6	0.2	0.2	0.1	0.1	0.1
β_{216}	-0.3	1.0	1.0	-0.8	0.3	0.3	-0.3	0.1	0.1	-0.3	0.1	0.1
β_{217}	1.4	1.2	1.2	1.3	0.4	0.4	0.3	0.2	0.2	0.2	0.1	0.1
β_{218}	-0.3	1.6	1.6	0.1	0.4	0.4	-0.4	0.2	0.2	-0.1	0.1	0.1
Γ_{21}	-2.7	2.7	2.7	1.1	1.5	1.5	3.0	0.7	0.8	3.2	0.3	0.4
Γ_{212}	-2.0	2.0	2.0	0.0	1.1	1.1	0.1	0.5	0.5	0.6	0.3	0.3
Γ_{22}	-5.9	2.2	2.6	-4.5	1.4	1.6	-4.6	0.7	0.9	-2.6	0.4	0.4
Γ_{213}	4.4	1.2	1.4	4.1	0.9	1.0	3.4	0.4	0.5	3.0	0.2	0.3
Γ_{223}	1.4	0.7	0.7	-0.6	0.8	0.8	-0.8	0.5	0.5	-0.9	0.3	0.3
Γ_{23}	0.8	0.1	0.1	2.4	0.1	0.2	3.6	0.2	0.4	3.8	0.2	0.3
Γ_{214}	0.4	0.8	0.8	0.4	0.4	0.4	0.4	0.3	0.3	1.1	0.2	0.2
Γ_{224}	2.3	0.4	0.4	1.3	0.5	0.5	1.6	0.3	0.3	1.3	0.2	0.2
Γ_{234}	0.3	< 0.1	< 0.1	1.3	0.1	0.1	1.3	0.2	0.2	1.0	0.2	0.2
Γ_{24}	0.2	< 0.1	< 0.1	0.7	0.1	0.1	1.4	0.1	0.1	1.7	0.1	0.1
Γ_{215}	0.2	0.6	0.6	-0.1	0.3	0.3	0.4	0.2	0.2	-0.1	0.1	0.1
Γ_{225}	0.9	0.4	0.4	0.7	0.4	0.4	-0.1	0.3	0.3	0.0	0.2	0.2
Γ_{235}	0.3	0.1	0.1	0.1	0.2	0.1	-0.7	0.2	0.2	0.7	0.1	0.1
Γ_{245}	-0.1	< 0.1	< 0.1	0.3	< 0.1	< 0.1	0.0	< 0.1	< 0.1	0.3	< 0.1	< 0.1
Γ_{25}	0.0	< 0.1	< 0.1	0.2	< 0.1	< 0.1	0.5	< 0.1	< 0.1	0.7	< 0.1	< 0.1
Total			70.5			25.6			13.1			7.4

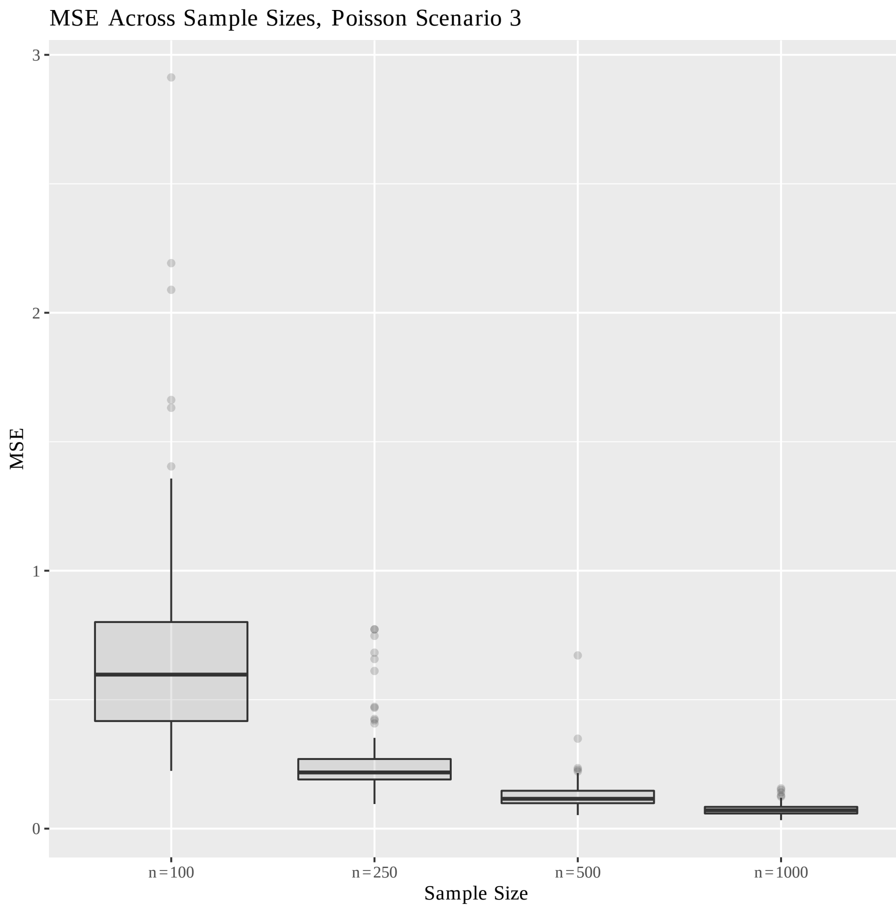

Figure H–7: MSE across sample sizes, Poisson outcome, no penalty, with $K = 2$, $p = 19$, $q = 5$.

Table H–11: Simulation 3 results multiplied by 100, averaged over 100 runs. Outcome follows a Poisson distribution with $K = 2$, $p = 19$, $q = 5$, $\pi_1 = 0.6$, $\pi_2 = 0.4$, $\boldsymbol{\beta}_1^\top = (0.65, 0.30, 0.15, 0.35, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00)$, $\boldsymbol{\beta}_2^\top = (0.20, -0.45, -0.10, 0.25, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00)$, $\mathbb{\Gamma}_1^{*\top} = (0.30, -0.25, 0.10, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00)$, and $\mathbb{\Gamma}_2^{*\top} = (0.35, 0.20, 0.15, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00)$. Small values changed to zero, part 1.

n	$n = 100$			$n = 250$			$n = 500$			$n = 1000$		
Measure	Bias	Var	MSE	Bias	Var	MSE	Bias	Var	MSE	Bias	Var	MSE
π_1	-3.2	1.0	1.1	-1.5	0.5	0.5	-1.2	0.2	0.2	-1.5	0.1	0.2
β_{10}	-6.0	1.7	2.1	-1.7	0.4	0.4	-1.3	0.2	0.2	-1.0	0.1	0.1
β_{11}	-2.5	1.5	1.6	0.5	0.6	0.6	1.3	0.2	0.2	1.2	0.1	0.1
β_{12}	-1.4	1.0	1.1	0.0	0.2	0.2	-0.3	0.1	0.1	0.0	< 0.1	< 0.1
β_{13}	0.7	0.6	0.6	0.6	0.2	0.2	-0.1	0.1	0.1	-0.8	< 0.1	< 0.1
β_{14}	-0.7	0.6	0.6	-0.1	0.1	0.1	-0.3	0.1	0.1	-0.2	< 0.1	< 0.1
β_{15}	-0.7	0.5	0.5	0.1	0.1	0.1	0.2	< 0.1	< 0.1	0.1	< 0.1	< 0.1
β_{16}	-0.3	0.4	0.4	-0.1	0.1	0.1	0.2	< 0.1	< 0.1	0.0	< 0.1	< 0.1
β_{17}	0.4	0.5	0.5	0.1	0.1	0.1	0.0	< 0.1	< 0.1	0.0	< 0.1	< 0.1
β_{18}	-0.6	0.4	0.4	-0.7	0.1	0.1	-0.1	0.1	0.1	-0.1	< 0.1	< 0.1
β_{19}	0.4	0.5	0.5	0.2	0.1	0.1	0.2	0.1	0.1	0.3	< 0.1	< 0.1
β_{110}	-0.5	0.5	0.5	0.0	0.1	0.1	-0.2	< 0.1	< 0.1	-0.2	< 0.1	< 0.1
β_{111}	-0.6	0.4	0.4	0.4	0.1	0.1	0.2	0.1	0.1	-0.1	< 0.1	< 0.1
β_{112}	-0.4	0.5	0.5	0.0	0.1	0.1	0.0	< 0.1	< 0.1	-0.1	< 0.1	< 0.1
β_{113}	0.1	0.4	0.4	0.4	0.1	0.1	0.0	< 0.1	< 0.1	0.1	< 0.1	< 0.1
β_{114}	-0.4	0.5	0.5	-0.4	0.1	0.1	-0.2	< 0.1	< 0.1	-0.2	< 0.1	< 0.1
β_{115}	-0.9	0.5	0.5	-0.6	0.1	0.1	-0.1	< 0.1	< 0.1	-0.1	< 0.1	< 0.1
β_{116}	0.7	0.4	0.4	0.4	0.1	0.1	0.0	0.1	0.1	0.3	< 0.1	< 0.1
β_{117}	-0.7	0.6	0.6	0.1	0.1	0.1	-0.2	0.1	0.1	-0.2	< 0.1	< 0.1
β_{118}	0.4	0.5	0.5	0.2	0.1	0.1	0.1	0.1	0.1	0.3	< 0.1	< 0.1
Γ_{11}	-10.3	2.2	3.2	-4.8	1.0	1.3	-2.9	0.5	0.6	-1.4	0.1	0.1
Γ_{112}	12.8	2.7	4.3	4.4	1.0	1.1	2.2	0.4	0.5	1.1	0.2	0.2
Γ_{12}	-1.7	1.1	1.1	-2.6	0.6	0.7	-2.9	0.3	0.4	-2.8	0.2	0.3
Γ_{113}	-0.7	0.6	0.6	-0.6	0.2	0.2	-0.1	0.1	0.1	-0.4	0.1	0.1
Γ_{123}	1.3	0.6	0.6	-0.3	0.4	0.4	0.1	0.3	0.3	-0.3	0.2	0.2
Γ_{13}	0.9	0.1	0.1	1.4	< 0.1	0.1	2.5	0.1	0.1	2.6	< 0.1	0.1
Γ_{114}	-1.3	0.4	0.5	-0.4	0.2	0.2	-0.1	0.1	0.1	-0.1	< 0.1	< 0.1
Γ_{124}	-0.1	0.3	0.3	-0.8	0.2	0.2	-1.0	0.1	0.2	-0.7	0.1	0.1
Γ_{134}	0.1	< 0.1	< 0.1	-0.3	0.1	0.1	-0.2	0.1	0.1	0.5	< 0.1	< 0.1
Γ_{14}	0.0	< 0.1	< 0.1	0.6	< 0.1	< 0.1	0.9	< 0.1	< 0.1	1.2	< 0.1	< 0.1
Γ_{115}	-0.9	0.5	0.5	-0.1	0.2	0.2	0.1	0.1	0.1	0.1	< 0.1	< 0.1
Γ_{125}	1.1	0.2	0.2	0.2	0.2	0.2	0.1	0.1	0.1	0.0	0.1	0.1
Γ_{135}	0.0	< 0.1	< 0.1	-0.3	0.1	0.1	-0.5	0.1	0.1	0.0	< 0.1	< 0.1
Γ_{145}	0.1	< 0.1	< 0.1	0.2	< 0.1	< 0.1	0.2	< 0.1	< 0.1	0.0	< 0.1	< 0.1
Γ_{15}	0.1	< 0.1	< 0.1	0.2	< 0.1	< 0.1	0.4	< 0.1	< 0.1	0.7	< 0.1	< 0.1

Table H–12: Simulation 3 results multiplied by 100, averaged over 100 runs. Outcome follows a Poisson distribution with $K = 2$, $p = 19$, $q = 5$, $\pi_1 = 0.6$, $\pi_2 = 0.4$, $\boldsymbol{\beta}_1^\top = (0.65, 0.30, 0.15, 0.35, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00)$, $\boldsymbol{\beta}_2^\top = (0.20, -0.45, -0.10, 0.25, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00)$, $\mathbb{\Gamma}_1^{*\top} = (0.30, -0.25, 0.10, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00)$, and $\mathbb{\Gamma}_2^{*\top} = (0.35, 0.20, 0.15, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00)$. Small values changed to zero, part 2.

n	$n = 100$			$n = 250$			$n = 500$			$n = 1000$		
Measure	Bias	Var	MSE	Bias	Var	MSE	Bias	Var	MSE	Bias	Var	MSE
β_{20}	-3.3	4.7	4.8	-2.9	1.5	1.5	1.3	0.6	0.6	0.6	0.3	0.3
β_{21}	6.2	5.4	5.7	1.4	1.3	1.3	-0.5	0.4	0.4	0.2	0.2	0.2
β_{22}	4.8	2.0	2.2	1.7	0.5	0.6	1.9	0.2	0.2	1.6	0.1	0.1
β_{23}	5.0	1.4	1.7	1.9	0.4	0.4	0.3	0.2	0.2	0.1	0.1	0.1
β_{24}	0.2	0.9	0.9	-0.4	0.3	0.3	-0.2	0.1	0.1	0.1	< 0.1	< 0.1
β_{25}	1.6	1.5	1.5	0.4	0.5	0.5	0.1	0.1	0.1	0.2	0.1	0.1
β_{26}	-0.9	1.2	1.2	0.5	0.3	0.3	0.1	0.1	0.1	0.2	0.1	0.1
β_{27}	1.2	1.3	1.3	-0.6	0.3	0.3	0.2	0.2	0.2	0.3	0.1	0.1
β_{28}	0.0	1.1	1.1	0.9	0.4	0.4	0.3	0.2	0.2	-0.1	0.1	0.1
β_{29}	0.0	1.8	1.8	0.8	0.4	0.4	0.3	0.2	0.2	0.0	0.1	0.1
β_{210}	0.1	1.4	1.4	0.3	0.4	0.4	0.2	0.2	0.2	0.1	0.1	0.1
β_{211}	-0.3	1.3	1.3	-2.1	0.3	0.3	-0.7	0.1	0.1	0.4	0.1	0.1
β_{212}	0.0	1.4	1.4	0.5	0.3	0.3	0.2	0.2	0.2	0.1	0.1	0.1
β_{213}	1.3	1.3	1.3	-0.3	0.4	0.4	0.3	0.1	0.1	0.3	0.1	0.1
β_{214}	0.9	1.1	1.1	0.1	0.3	0.3	-0.2	0.1	0.1	0.0	0.1	0.1
β_{215}	-1.3	1.2	1.2	0.1	0.4	0.4	0.6	0.2	0.2	0.1	0.1	0.1
β_{216}	-0.4	1.0	1.0	-0.8	0.3	0.3	-0.3	0.1	0.1	-0.3	0.1	0.1
β_{217}	1.4	1.2	1.2	1.3	0.4	0.4	0.2	0.2	0.2	0.2	0.1	0.1
β_{218}	-0.3	1.6	1.6	0.1	0.4	0.4	-0.4	0.2	0.2	-0.1	0.1	0.1
Γ_{21}	-2.7	2.7	2.7	1.1	1.5	1.5	3.0	0.7	0.8	3.2	0.3	0.4
Γ_{212}	-2.0	2.0	2.0	0.0	1.1	1.1	0.1	0.5	0.5	0.6	0.3	0.3
Γ_{22}	-5.9	2.2	2.6	-4.5	1.4	1.6	-4.6	0.7	0.9	-2.6	0.4	0.4
Γ_{213}	4.4	1.2	1.4	4.1	0.9	1.0	3.4	0.4	0.5	3.0	0.2	0.3
Γ_{223}	1.4	0.7	0.7	-0.6	0.8	0.8	-0.8	0.5	0.5	-0.9	0.3	0.3
Γ_{23}	0.8	0.1	0.1	2.4	0.1	0.2	3.6	0.2	0.4	3.8	0.2	0.3
Γ_{214}	0.4	0.8	0.8	0.4	0.4	0.4	0.4	0.3	0.3	1.1	0.2	0.2
Γ_{224}	2.3	0.4	0.4	1.3	0.5	0.5	1.6	0.3	0.3	1.3	0.2	0.2
Γ_{234}	0.3	< 0.1	< 0.1	1.3	0.1	0.1	1.3	0.2	0.2	1.0	0.2	0.2
Γ_{24}	0.2	< 0.1	< 0.1	0.7	0.1	0.1	1.4	0.1	0.1	1.7	0.1	0.1
Γ_{215}	0.2	0.6	0.6	-0.1	0.3	0.3	0.4	0.2	0.2	-0.1	0.1	0.1
Γ_{225}	0.9	0.4	0.4	0.7	0.4	0.4	-0.1	0.3	0.3	0.0	0.2	0.2
Γ_{235}	0.3	0.1	0.1	0.1	0.2	0.1	-0.7	0.2	0.2	0.7	0.1	0.1
Γ_{245}	-0.1	< 0.1	< 0.1	0.3	< 0.1	< 0.1	0.0	< 0.1	< 0.1	0.3	< 0.1	< 0.1
Γ_{25}	0.0	< 0.1	< 0.1	0.2	< 0.1	< 0.1	0.5	< 0.1	< 0.1	0.7	< 0.1	< 0.1
Total			70.5			25.6			13.1			7.4

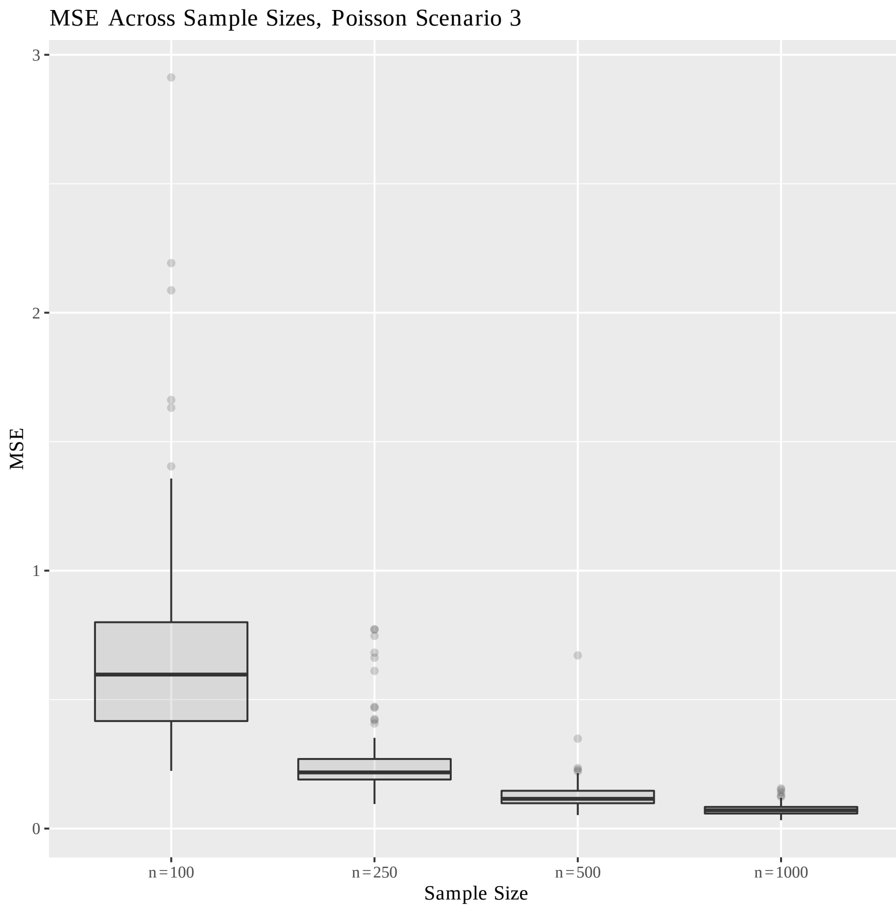

Figure H–8: MSE across sample sizes, Poisson outcome, small values changed to zero, with $K = 2$, $p = 19$, $q = 5$.

Table H–13: Simulation 3 results multiplied by 100, averaged over 100 runs. Outcome follows a Poisson distribution with $K = 2$, $p = 4$, $q = 2$, $\pi_1 = 0.6$, $\pi_2 = 0.4$, $\boldsymbol{\beta}_1^\top = (0.65, 0.30, 0.15, 0.35)$, $\boldsymbol{\beta}_2^\top = (0.20, -0.45, -0.10, 0.25)$, $\mathbb{\Gamma}_1^{*\top} = (0.30, -0.25, 0.10)$, and $\mathbb{\Gamma}_2^{*\top} = (0.35, 0.20, 0.15)$. Oracle model.

n	$n = 100$			$n = 250$			$n = 500$			$n = 1000$		
Measure	Bias	Var	MSE	Bias	Var	MSE	Bias	Var	MSE	Bias	Var	MSE
π_1	-1.8	0.8	0.8	-1.4	0.4	0.4	-1.7	0.2	0.2	-1.6	0.1	0.1
β_{10}	-0.5	0.8	0.8	-0.3	0.2	0.2	-0.2	0.1	0.1	-0.2	0.1	0.1
β_{11}	-0.1	0.9	0.9	0.5	0.3	0.3	0.4	0.2	0.2	0.1	0.1	0.1
β_{12}	1.1	0.3	0.4	1.2	0.1	0.1	1.1	0.1	0.1	0.9	< 0.1	< 0.1
β_{13}	1.3	0.3	0.3	1.0	0.1	0.2	0.2	0.1	0.1	-0.4	< 0.1	< 0.1
Γ_{11}	-4.4	1.1	1.2	-1.5	0.5	0.5	-0.9	0.2	0.2	-0.2	0.1	0.1
Γ_{112}	3.7	1.1	1.3	1.2	0.5	0.5	0.3	0.2	0.2	-0.1	0.1	0.1
Γ_{12}	-4.0	0.7	0.9	-2.9	0.5	0.6	-2.5	0.3	0.4	-2.1	0.2	0.3
β_{20}	-0.2	2.4	2.4	0.2	0.9	0.9	3.1	0.5	0.5	1.4	0.2	0.2
β_{21}	1.6	2.6	2.6	0.6	1.0	1.0	1.3	0.5	0.6	1.8	0.2	0.2
β_{22}	0.0	1.4	1.3	0.3	0.5	0.5	1.0	0.2	0.2	0.6	0.1	0.1
β_{23}	0.6	1.1	1.1	0.5	0.3	0.3	-0.3	0.1	0.1	-0.6	0.1	0.1
Γ_{21}	1.3	2.7	2.7	2.4	1.2	1.2	2.9	0.7	0.7	2.4	0.3	0.3
Γ_{212}	1.6	1.5	1.5	2.5	1.1	1.1	3.6	0.6	0.7	3.7	0.3	0.4
Γ_{22}	-8.3	1.5	2.2	-5.5	0.8	1.1	-2.9	0.5	0.6	-0.9	0.3	0.3
Total			20.4			9.0			5.0			2.4

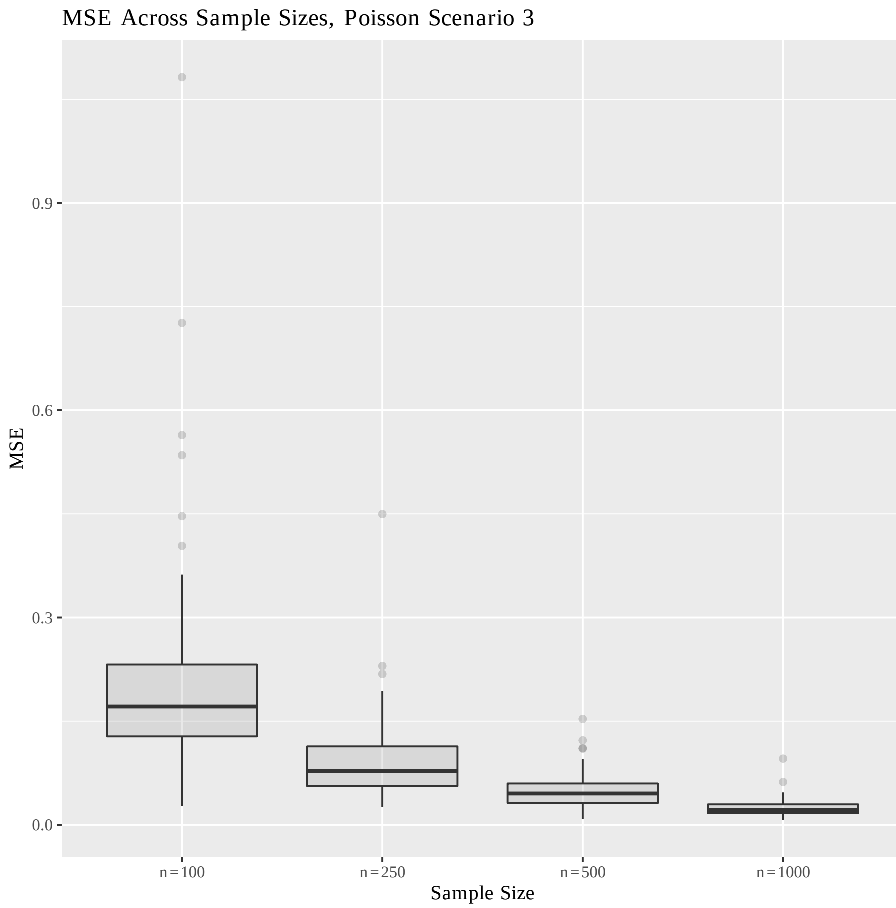

Figure H–9: MSE across sample sizes, Poisson outcome, oracle model, with $K = 2$, $p = 19$, $q = 5$.

Table H–14: Simulation 4 results multiplied by 100, averaged over 100 runs. Outcome follows a Poisson distribution with $K = 3$, $p = 7$, $q = 2$, $\pi_1 = 0.5$, $\pi_2 = 0.3$, $\pi_3 = 0.2$, $\boldsymbol{\beta}_1^\top = (0.65, 0.30, 0.00, 0.00, 0.00, 0.00, 0.00)$, $\boldsymbol{\beta}_2^\top = (0.20, -0.45, 0.00, 0.00, 0.00, 0.00, 0.00)$, $\boldsymbol{\beta}_3^\top = (1.00, 0.15, 0.00, 0.00, 0.00, 0.00, 0.00)$, $\mathbb{\Gamma}_1^{*\top} = (0.30, -0.25, 0.10)$, $\mathbb{\Gamma}_2^{*\top} = (0.35, 0.20, 0.15)$, and $\mathbb{\Gamma}_3^{*\top} = (0.25, 0.00, 0.20)$. No penalty, part 1.

n	$n = 100$			$n = 250$			$n = 500$			$n = 1000$		
Measure	Bias	Var	MSE	Bias	Var	MSE	Bias	Var	MSE	Bias	Var	MSE
π_1	-11.5	1.6	2.9	-13.6	1.2	3.1	-13.2	1.0	2.8	-13.4	0.8	2.6
π_2	3.0	1.6	1.7	7.3	1.1	1.6	7.5	0.7	1.3	9.3	0.5	1.4
β_{10}	-5.0	5.1	5.3	-1.8	1.7	1.7	0.3	1.3	1.3	1.2	0.5	0.5
β_{11}	2.2	5.3	5.3	0.5	1.5	1.5	1.9	0.7	0.8	2.8	0.5	0.5
β_{12}	-3.7	3.1	3.2	0.0	0.6	0.6	0.0	0.5	0.5	-1.4	0.3	0.3
β_{13}	1.8	2.4	2.4	-0.5	0.9	0.9	0.2	0.4	0.4	0.0	0.2	0.2
β_{14}	0.9	1.3	1.3	1.9	0.6	0.7	1.4	0.3	0.3	0.8	0.2	0.2
β_{15}	-0.5	1.2	1.2	0.1	0.9	0.9	0.0	0.3	0.3	-0.1	0.2	0.2
β_{16}	0.1	1.7	1.6	-0.3	0.8	0.8	0.3	0.4	0.4	-0.2	0.2	0.2
Γ_{11}	-5.9	2.9	3.2	0.3	2.4	2.4	3.0	1.7	1.7	3.3	0.4	0.5
Γ_{112}	11.0	2.5	3.7	6.4	2.2	2.6	4.3	1.1	1.3	-0.1	0.4	0.4
Γ_{12}	-8.7	0.2	0.9	-5.2	0.7	1.0	-2.7	0.6	0.7	-3.2	0.3	0.4

Table H–15: Simulation 4 results multiplied by 100, averaged over 100 runs. Outcome follows a Poisson distribution with $K = 3$, $p = 7$, $q = 2$, $\pi_1 = 0.5$, $\pi_2 = 0.3$, $\pi_3 = 0.2$, $\boldsymbol{\beta}_1^\top = (0.65, 0.30, 0.00, 0.00, 0.00, 0.00, 0.00)$, $\boldsymbol{\beta}_2^\top = (0.20, -0.45, 0.00, 0.00, 0.00, 0.00, 0.00)$, $\boldsymbol{\beta}_3^\top = (1.00, 0.15, 0.00, 0.00, 0.00, 0.00, 0.00)$, $\mathbb{\Gamma}_1^{*\top} = (0.30, -0.25, 0.10)$, $\mathbb{\Gamma}_2^{*\top} = (0.35, 0.20, 0.15)$, and $\mathbb{\Gamma}_3^{*\top} = (0.25, 0.00, 0.20)$. No penalty, part 2.

n	$n = 100$			$n = 250$			$n = 500$			$n = 1000$		
Measure	Bias	Var	MSE	Bias	Var	MSE	Bias	Var	MSE	Bias	Var	MSE
β_{20}	3.1	7.4	7.4	11.6	2.5	3.8	16.7	0.7	3.5	18.8	0.4	3.9
β_{21}	12.2	6.8	8.2	14.8	3.0	5.1	13.8	1.5	3.4	15.0	0.7	3.0
β_{22}	0.9	2.5	2.5	1.2	0.9	0.9	0.8	0.4	0.4	1.2	0.2	0.2
β_{23}	2.2	3.8	3.8	0.5	1.0	1.0	-0.5	0.3	0.3	0.2	0.1	0.1
β_{24}	-2.2	2.6	2.6	-0.5	0.9	0.9	-0.7	0.4	0.4	-0.3	0.2	0.2
β_{25}	-0.1	1.9	1.9	-0.2	0.8	0.8	-0.4	0.3	0.3	-0.1	0.1	0.1
β_{26}	0.2	1.9	1.9	-0.9	0.6	0.6	0.1	0.2	0.2	0.1	0.1	0.1
Γ_{21}	1.7	4.3	4.3	3.9	1.9	2.0	8.8	1.1	1.9	9.5	0.4	1.3
Γ_{212}	2.2	3.5	3.5	8.2	1.3	2.0	9.4	0.7	1.5	10.4	0.2	1.3
Γ_{22}	-8.7	2.0	2.8	-4.5	1.4	1.5	-4.1	0.8	1.0	-0.6	0.5	0.5

Table H–16: Simulation 4 results multiplied by 100, averaged over 100 runs. Outcome follows a Poisson distribution with $K = 3$, $p = 7$, $q = 2$, $\pi_1 = 0.5$, $\pi_2 = 0.3$, $\pi_3 = 0.2$, $\boldsymbol{\beta}_1^\top = (0.65, 0.30, 0.00, 0.00, 0.00, 0.00, 0.00)$, $\boldsymbol{\beta}_2^\top = (0.20, -0.45, 0.00, 0.00, 0.00, 0.00, 0.00)$, $\boldsymbol{\beta}_3^\top = (1.00, 0.15, 0.00, 0.00, 0.00, 0.00, 0.00)$, $\mathbb{\Gamma}_1^{*\top} = (0.30, -0.25, 0.10)$, $\mathbb{\Gamma}_2^{*\top} = (0.35, 0.20, 0.15)$, and $\mathbb{\Gamma}_3^{*\top} = (0.25, 0.00, 0.20)$. No penalty, part 3.

n	$n = 100$			$n = 250$			$n = 500$			$n = 1000$		
Measure	Bias	Var	MSE	Bias	Var	MSE	Bias	Var	MSE	Bias	Var	MSE
β_{30}	-15.4	3.7	6.0	-14.0	2.5	4.4	-19.6	2.6	6.4	-20.6	1.9	6.2
β_{31}	-6.2	3.8	4.1	-5.7	2.9	3.2	-2.7	3.5	3.5	0.7	1.9	1.9
β_{32}	0.0	3.8	3.8	-1.7	1.8	1.8	-0.5	0.9	0.8	0.2	0.5	0.5
β_{33}	-2.7	1.9	1.9	-0.6	1.1	1.1	-0.3	0.6	0.6	-0.4	0.4	0.4
β_{34}	2.5	1.7	1.7	-0.1	0.7	0.7	0.8	0.5	0.5	0.1	0.4	0.4
β_{35}	0.6	2.1	2.1	0.9	0.7	0.7	0.5	0.6	0.6	0.4	0.4	0.4
β_{36}	-0.5	1.9	1.9	-0.7	0.9	0.9	-1.0	0.7	0.7	-0.3	0.4	0.4
Γ_{31}	-10.7	2.0	3.1	-4.6	1.7	1.9	-0.1	1.5	1.5	3.4	0.8	0.9
Γ_{312}	1.0	2.3	2.3	-2.6	2.1	2.2	-6.6	2.0	2.5	-6.7	2.0	2.4
Γ_{32}	-15.6	1.0	3.4	-14.4	0.8	2.9	-10.8	1.0	2.2	-8.3	0.8	1.5
Total			102.0			56.1			44.1			33.1

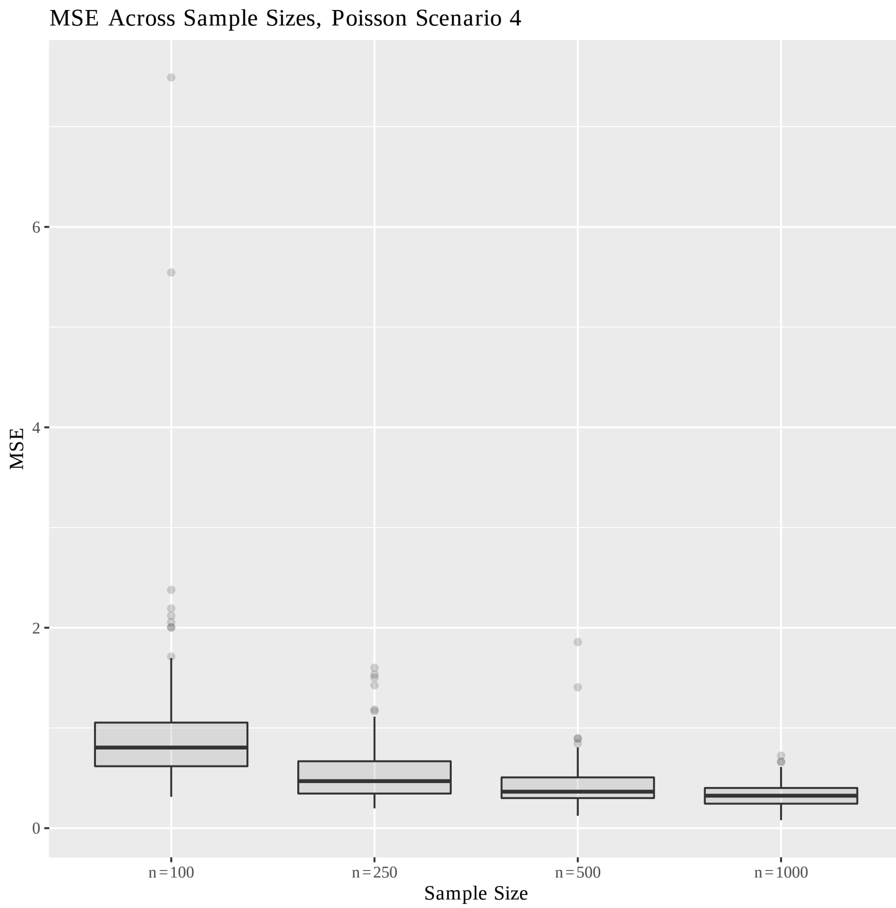

Figure H–10: MSE across sample sizes, Poisson outcome, no penalty, with $K = 3$, $p = 7$, $q = 2$.

Table H–17: Simulation 4 results multiplied by 100, averaged over 100 runs. Outcome follows a Poisson distribution with $K = 3$, $p = 7$, $q = 2$, $\pi_1 = 0.5$, $\pi_2 = 0.3$, $\pi_3 = 0.2$, $\boldsymbol{\beta}_1^\top = (0.65, 0.30, 0.00, 0.00, 0.00, 0.00, 0.00)$, $\boldsymbol{\beta}_2^\top = (0.20, -0.45, 0.00, 0.00, 0.00, 0.00, 0.00)$, $\boldsymbol{\beta}_3^\top = (1.00, 0.15, 0.00, 0.00, 0.00, 0.00, 0.00)$, $\mathbb{\Gamma}_1^{*\top} = (0.30, -0.25, 0.10)$, $\mathbb{\Gamma}_2^{*\top} = (0.35, 0.20, 0.15)$, and $\mathbb{\Gamma}_3^{*\top} = (0.25, 0.00, 0.20)$. Small values changed to zero, part 1.

n	$n = 100$			$n = 250$			$n = 500$			$n = 1000$		
Measure	Bias	Var	MSE	Bias	Var	MSE	Bias	Var	MSE	Bias	Var	MSE
π_1	-11.5	1.6	2.9	-13.6	1.2	3.1	-13.2	1.0	2.8	-13.4	0.8	2.6
π_2	3.0	1.6	1.7	7.3	1.1	1.6	7.5	0.7	1.3	9.3	0.5	1.4
β_{10}	-5.0	5.1	5.3	-1.8	1.7	1.7	0.3	1.3	1.3	1.2	0.5	0.5
β_{11}	2.2	5.3	5.3	0.5	1.5	1.5	1.9	0.7	0.8	2.8	0.5	0.5
β_{12}	-3.7	3.1	3.2	0.0	0.6	0.6	0.0	0.5	0.5	-1.4	0.3	0.3
β_{13}	1.8	2.4	2.4	-0.5	0.9	0.9	0.1	0.4	0.4	0.1	0.2	0.2
β_{14}	0.9	1.3	1.3	1.9	0.6	0.7	1.3	0.3	0.3	0.8	0.2	0.2
β_{15}	-0.5	1.2	1.2	0.1	0.9	0.9	0.0	0.3	0.3	-0.1	0.2	0.2
β_{16}	0.1	1.7	1.6	-0.4	0.8	0.8	0.3	0.4	0.4	-0.2	0.2	0.2
Γ_{11}	-5.9	2.9	3.2	0.3	2.4	2.4	3.0	1.7	1.7	3.3	0.4	0.5
Γ_{112}	11.0	2.5	3.7	6.4	2.2	2.6	4.3	1.1	1.3	-0.1	0.4	0.4
Γ_{12}	-8.7	0.2	0.9	-5.2	0.7	1.0	-2.7	0.6	0.7	-3.2	0.3	0.4

Table H–18: Simulation 4 results multiplied by 100, averaged over 100 runs. Outcome follows a Poisson distribution with $K = 3$, $p = 7$, $q = 2$, $\pi_1 = 0.5$, $\pi_2 = 0.3$, $\pi_3 = 0.2$, $\boldsymbol{\beta}_1^\top = (0.65, 0.30, 0.00, 0.00, 0.00, 0.00, 0.00)$, $\boldsymbol{\beta}_2^\top = (0.20, -0.45, 0.00, 0.00, 0.00, 0.00, 0.00)$, $\boldsymbol{\beta}_3^\top = (1.00, 0.15, 0.00, 0.00, 0.00, 0.00, 0.00)$, $\mathbb{\Gamma}_1^{*\top} = (0.30, -0.25, 0.10)$, $\mathbb{\Gamma}_2^{*\top} = (0.35, 0.20, 0.15)$, and $\mathbb{\Gamma}_3^{*\top} = (0.25, 0.00, 0.20)$. Small values changed to zero, part 2.

n	$n = 100$			$n = 250$			$n = 500$			$n = 1000$		
Measure	Bias	Var	MSE	Bias	Var	MSE	Bias	Var	MSE	Bias	Var	MSE
β_{20}	3.1	7.4	7.4	11.6	2.5	3.8	16.7	0.7	3.5	18.8	0.4	3.9
β_{21}	12.2	6.8	8.2	14.8	3.0	5.1	13.8	1.5	3.4	15.0	0.7	3.0
β_{22}	0.9	2.5	2.5	1.1	0.9	0.9	0.8	0.4	0.4	1.2	0.2	0.2
β_{23}	2.2	3.8	3.8	0.5	1.0	1.0	-0.5	0.3	0.3	0.2	0.1	0.1
β_{24}	-2.2	2.6	2.6	-0.5	0.9	0.9	-0.7	0.4	0.4	-0.3	0.2	0.2
β_{25}	-0.1	1.9	1.9	-0.2	0.8	0.8	-0.4	0.3	0.3	-0.1	0.1	0.1
β_{26}	0.2	1.9	1.9	-0.8	0.6	0.6	0.1	0.2	0.2	0.1	0.1	0.1
Γ_{21}	1.7	4.3	4.3	3.9	1.9	2.0	8.8	1.1	1.9	9.5	0.4	1.3
Γ_{212}	2.2	3.5	3.5	8.2	1.3	2.0	9.4	0.7	1.5	10.4	0.2	1.3
Γ_{22}	-8.7	2.0	2.8	-4.5	1.4	1.5	-4.1	0.8	1.0	-0.6	0.5	0.5

Table H–19: Simulation 4 results multiplied by 100, averaged over 100 runs. Outcome follows a Poisson distribution with $K = 3$, $p = 7$, $q = 2$, $\pi_1 = 0.5$, $\pi_2 = 0.3$, $\pi_3 = 0.2$, $\boldsymbol{\beta}_1^\top = (0.65, 0.30, 0.00, 0.00, 0.00, 0.00, 0.00)$, $\boldsymbol{\beta}_2^\top = (0.20, -0.45, 0.00, 0.00, 0.00, 0.00, 0.00)$, $\boldsymbol{\beta}_3^\top = (1.00, 0.15, 0.00, 0.00, 0.00, 0.00, 0.00)$, $\mathbb{\Gamma}_1^{*\top} = (0.30, -0.25, 0.10)$, $\mathbb{\Gamma}_2^{*\top} = (0.35, 0.20, 0.15)$, and $\mathbb{\Gamma}_3^{*\top} = (0.25, 0.00, 0.20)$. Small values changed to zero, part 3.

n	$n = 100$			$n = 250$			$n = 500$			$n = 1000$		
Measure	Bias	Var	MSE	Bias	Var	MSE	Bias	Var	MSE	Bias	Var	MSE
β_{30}	-15.4	3.7	6.0	-14.0	2.5	4.4	-19.6	2.6	6.4	-20.6	1.9	6.2
β_{31}	-6.2	3.8	4.1	-5.7	2.9	3.2	-2.7	3.5	3.5	0.7	1.9	1.9
β_{32}	0.0	3.8	3.8	-1.7	1.8	1.8	-0.5	0.9	0.8	0.2	0.5	0.5
β_{33}	-2.7	1.9	1.9	-0.6	1.1	1.1	-0.3	0.6	0.6	-0.4	0.4	0.4
β_{34}	2.5	1.7	1.7	-0.2	0.7	0.7	0.8	0.5	0.5	0.1	0.4	0.4
β_{35}	0.6	2.1	2.1	0.9	0.7	0.7	0.5	0.6	0.6	0.4	0.4	0.4
β_{36}	-0.5	1.9	1.9	-0.7	0.9	0.9	-1.1	0.7	0.7	-0.2	0.4	0.4
Γ_{31}	-10.7	2.0	3.1	-4.6	1.7	1.9	-0.1	1.5	1.5	3.4	0.8	0.9
Γ_{312}	1.0	2.3	2.3	-2.6	2.1	2.2	-6.6	2.0	2.5	-6.7	2.0	2.4
Γ_{32}	-15.6	1.0	3.4	-14.4	0.8	2.9	-10.8	1.0	2.2	-8.3	0.8	1.5
Total			102.0			56.1			44.0			33.1

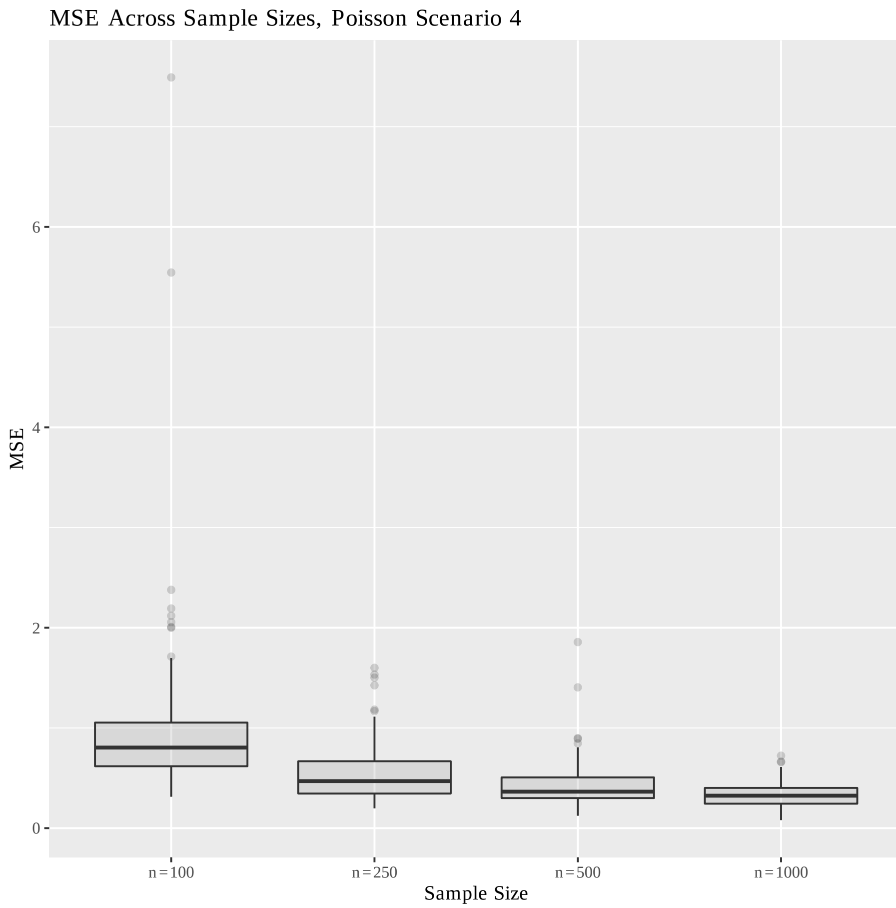

Figure H–11: MSE across sample sizes, Poisson outcome, small values changed to zero, with $K = 3$, $p = 7$, $q = 2$.

Table H–20: Simulation 4 results multiplied by 100, averaged over 100 runs. Outcome follows a Poisson distribution with $K = 3$, $p = 2$, $q = 2$, $\pi_1 = 0.5$, $\pi_2 = 0.3$, $\pi_3 = 0.2$, $\boldsymbol{\beta}_1^\top = (0.65, 0.30)$, $\boldsymbol{\beta}_2^\top = (0.20, -0.45)$, $\boldsymbol{\beta}_3^\top = (1.00, 0.15)$, $\mathbb{\Gamma}_1^{*\top} = (0.30, -0.25, 0.10)$, $\mathbb{\Gamma}_2^{*\top} = (0.35, 0.20, 0.15)$, and $\mathbb{\Gamma}_3^{*\top} = (0.25, 0.00, 0.20)$. Oracle model.

n	$n = 100$			$n = 250$			$n = 500$			$n = 1000$		
Measure	Bias	Var	MSE	Bias	Var	MSE	Bias	Var	MSE	Bias	Var	MSE
π_1	-12.5	1.4	2.9	-11.6	1.0	2.4	-11.6	0.7	2.1	-10.9	0.7	1.8
π_2	5.2	1.1	1.4	6.6	0.6	1.0	7.9	0.5	1.1	8.7	0.4	1.2
β_{10}	-1.5	3.6	3.6	1.0	1.0	1.0	1.2	0.8	0.8	1.8	0.3	0.3
β_{11}	1.5	2.3	2.3	2.1	0.8	0.9	2.2	0.6	0.6	2.6	0.4	0.5
Γ_{11}	1.4	3.7	3.6	1.7	1.1	1.2	2.6	1.1	1.2	3.6	0.4	0.5
Γ_{112}	6.5	3.3	3.7	4.6	1.2	1.4	3.0	0.7	0.8	1.4	0.3	0.3
Γ_{12}	-4.4	1.0	1.2	-3.7	0.5	0.7	-1.1	0.7	0.7	-1.1	0.4	0.4
β_{20}	8.8	3.3	4.1	13.8	1.3	3.1	15.2	0.9	3.2	17.2	0.6	3.5
β_{21}	13.0	3.5	5.2	13.2	1.7	3.4	14.4	1.2	3.2	15.3	0.6	2.9
Γ_{21}	4.7	3.0	3.2	6.1	1.5	1.9	7.9	0.8	1.4	8.3	0.4	1.1
Γ_{212}	7.6	2.5	3.0	7.7	1.2	1.8	9.4	0.6	1.5	10.4	0.2	1.3
Γ_{22}	-6.4	1.8	2.2	-2.6	1.4	1.4	-1.8	0.9	1.0	0.0	0.6	0.6
β_{30}	-8.5	3.3	4.0	-12.7	2.9	4.5	-14.8	2.5	4.7	-16.5	2.0	4.7
β_{31}	-3.5	3.3	3.4	-4.3	2.0	2.2	-2.2	1.9	1.9	-0.7	1.3	1.3
Γ_{31}	-4.4	2.7	2.8	-1.5	1.6	1.6	2.4	1.0	1.0	4.4	0.6	0.8
Γ_{312}	-4.3	1.9	2.1	-3.4	2.1	2.2	-6.3	1.8	2.2	-7.0	1.5	2.0
Γ_{32}	-14.9	0.7	2.9	-9.9	1.4	2.4	-6.8	1.2	1.6	-4.8	0.9	1.1
Total			51.6			33.0			28.8			24.3

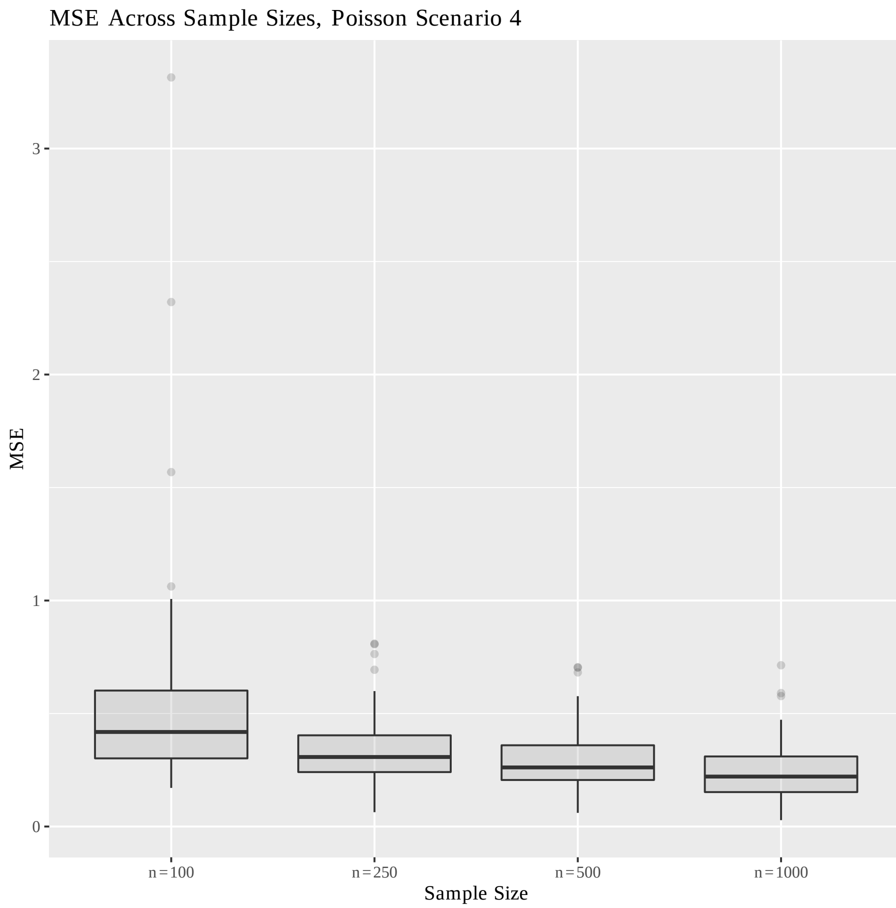

Figure H–12: MSE across sample sizes, Poisson outcome, oracle model, with $K = 3$, $p = 7$, $q = 2$.

Table H–21: Simulation 5 results multiplied by 100, averaged over 100 runs. Outcome follows a Poisson distribution with $K = 3$, $p = 17$, $q = 2$, $\pi_1 = 0.5$, $\pi_2 = 0.3$, $\pi_3 = 0.2$, $\boldsymbol{\beta}_1^\top = (0.65, 0.30, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00)$, $\boldsymbol{\beta}_2^\top = (0.20, -0.45, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00)$, $\boldsymbol{\beta}_3^\top = (1.00, 0.15, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00)$, $\mathbb{\Gamma}_1^{*\top} = (0.30, -0.25, 0.10)$, $\mathbb{\Gamma}_2^{*\top} = (0.35, 0.20, 0.15)$, and $\mathbb{\Gamma}_3^{*\top} = (0.25, 0.00, 0.20)$. No penalty, part 1.

n	$n = 100$			$n = 250$			$n = 500$			$n = 1000$		
Measure	Bias	Var	MSE	Bias	Var	MSE	Bias	Var	MSE	Bias	Var	MSE
π_1	-13.7	1.6	3.5	-13.8	1.3	3.2	-12.6	1.1	2.7	-11.2	0.9	2.1
π_2	0.8	1.4	1.4	4.7	1.1	1.3	6.0	0.6	0.9	7.8	0.4	1.0
β_{10}	-19.5	4.9	8.6	-7.4	3.1	3.6	-3.2	1.8	1.9	0.7	0.7	0.7
β_{11}	3.3	4.7	4.8	3.4	2.2	2.3	2.9	0.9	1.0	3.1	0.5	0.6
β_{12}	0.1	2.2	2.2	-1.1	0.7	0.7	-0.4	0.4	0.4	-0.4	0.2	0.2
β_{13}	0.1	2.1	2.0	-1.0	0.5	0.5	0.2	0.3	0.3	-0.5	0.1	0.1
β_{14}	0.1	2.6	2.5	0.8	0.7	0.7	-0.6	0.5	0.5	-0.1	0.2	0.2
β_{15}	-3.0	1.4	1.5	1.1	0.6	0.6	0.0	0.4	0.4	0.1	0.2	0.2
β_{16}	-0.1	1.9	1.9	1.3	0.6	0.6	-0.5	0.4	0.4	-0.8	0.1	0.1
β_{17}	-1.4	1.7	1.7	0.8	0.9	0.9	0.2	0.3	0.3	0.0	0.1	0.1
β_{18}	-2.1	2.1	2.1	-0.2	0.6	0.6	0.0	0.3	0.3	0.2	0.2	0.2
β_{19}	-0.6	1.6	1.5	1.1	0.6	0.6	0.8	0.6	0.6	0.7	0.2	0.2
β_{110}	-0.4	1.7	1.7	1.1	0.9	0.9	0.2	0.4	0.3	-0.1	0.2	0.2
β_{111}	-0.5	1.5	1.5	-0.8	0.7	0.7	0.6	0.3	0.3	-0.7	0.2	0.2
β_{112}	-0.8	1.8	1.8	0.3	0.7	0.7	-0.2	0.4	0.4	0.1	0.2	0.2
β_{113}	-2.2	2.6	2.6	1.5	0.6	0.6	-0.6	0.4	0.4	0.4	0.2	0.2
β_{114}	-1.2	2.1	2.1	-0.3	0.6	0.6	-0.9	0.4	0.4	0.1	0.2	0.2
β_{115}	3.6	1.7	1.8	0.5	0.8	0.8	-0.4	0.3	0.3	0.0	0.2	0.2
β_{116}	-1.2	1.5	1.5	2.0	0.6	0.7	1.4	0.3	0.3	0.3	0.2	0.2
Γ_{11}	-13.4	3.0	4.8	-8.2	1.6	2.3	-2.4	1.1	1.2	0.3	0.4	0.4
Γ_{112}	16.1	2.4	5.0	10.4	1.6	2.6	6.3	0.8	1.2	4.2	0.4	0.5
Γ_{12}	-8.3	0.4	1.0	-6.9	0.5	0.9	-4.9	0.4	0.6	-3.3	0.3	0.5

Table H–22: Simulation 5 results multiplied by 100, averaged over 100 runs. Outcome follows a Poisson distribution with $K = 3$, $p = 17$, $q = 2$, $\pi_1 = 0.5$, $\pi_2 = 0.3$, $\pi_3 = 0.2$, $\boldsymbol{\beta}_1^\top = (0.65, 0.30, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00)$, $\boldsymbol{\beta}_2^\top = (0.20, -0.45, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00)$, $\boldsymbol{\beta}_3^\top = (1.00, 0.15, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00)$, $\mathbb{\Gamma}_1^{*\top} = (0.30, -0.25, 0.10)$, $\mathbb{\Gamma}_2^{*\top} = (0.35, 0.20, 0.15)$, and $\mathbb{\Gamma}_3^{*\top} = (0.25, 0.00, 0.20)$. No penalty, part 2.

n	$n = 100$			$n = 250$			$n = 500$			$n = 1000$		
Measure	Bias	Var	MSE	Bias	Var	MSE	Bias	Var	MSE	Bias	Var	MSE
β_{20}	-5.7	18.1	18.2	3.5	3.6	3.7	8.9	1.6	2.4	12.1	0.9	2.3
β_{21}	11.1	10.4	11.5	12.7	4.1	5.7	11.6	1.6	3.0	11.3	0.6	1.9
β_{22}	1.1	5.8	5.7	-0.4	0.9	0.9	-0.3	0.5	0.4	0.2	0.1	0.1
β_{23}	-3.4	2.3	2.4	0.4	1.0	1.0	-0.3	0.3	0.3	0.5	0.1	0.1
β_{24}	2.2	3.9	3.9	-0.3	0.7	0.7	0.0	0.2	0.2	0.0	0.2	0.2
β_{25}	0.4	3.1	3.1	-1.8	0.6	0.6	-0.2	0.3	0.3	-0.7	0.1	0.1
β_{26}	-0.2	2.8	2.8	0.1	0.7	0.7	1.3	0.3	0.3	0.5	0.1	0.1
β_{27}	3.5	5.0	5.1	0.2	1.1	1.1	0.5	0.4	0.4	0.6	0.1	0.1
β_{28}	0.0	3.7	3.7	0.5	0.7	0.6	1.4	0.2	0.3	1.1	0.1	0.1
β_{29}	-2.9	3.6	3.6	0.5	0.8	0.8	0.0	0.3	0.3	0.0	0.1	0.1
β_{210}	-1.0	2.5	2.5	-0.6	0.7	0.7	-0.7	0.3	0.3	-0.4	0.1	0.1
β_{211}	-1.1	3.1	3.1	-0.7	0.8	0.8	-1.0	0.3	0.3	0.2	0.1	0.1
β_{212}	3.0	3.6	3.7	0.9	0.6	0.6	0.5	0.3	0.3	-0.4	0.1	0.1
β_{213}	2.1	2.6	2.6	0.6	0.9	0.9	0.5	0.3	0.3	0.3	0.1	0.1
β_{214}	-1.9	3.1	3.1	-1.2	0.7	0.7	-0.9	0.2	0.2	-0.3	0.1	0.1
β_{215}	-1.9	4.4	4.4	-1.6	0.6	0.7	0.4	0.3	0.3	-0.7	0.1	0.1
β_{216}	2.4	5.2	5.2	-1.1	0.7	0.7	0.4	0.3	0.3	-0.1	0.1	0.1
Γ_{21}	-10.4	3.7	4.7	-0.2	2.7	2.7	3.7	1.0	1.1	6.3	0.6	1.0
Γ_{212}	-7.5	2.5	3.0	3.8	1.5	1.6	5.0	0.9	1.1	7.6	0.4	1.0
Γ_{22}	-12.3	0.8	2.3	-7.1	1.7	2.2	-4.1	1.2	1.3	-1.8	0.6	0.6

Table H–23: Simulation 5 results multiplied by 100, averaged over 100 runs. Outcome follows a Poisson distribution with $K = 3$, $p = 17$, $q = 2$, $\pi_1 = 0.5$, $\pi_2 = 0.3$, $\pi_3 = 0.2$, $\boldsymbol{\beta}_1^\top = (0.65, 0.30, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00)$, $\boldsymbol{\beta}_2^\top = (0.20, -0.45, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00)$, $\boldsymbol{\beta}_3^\top = (1.00, 0.15, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00)$, $\mathbb{\Gamma}_1^{*\top} = (0.30, -0.25, 0.10)$, $\mathbb{\Gamma}_2^{*\top} = (0.35, 0.20, 0.15)$, and $\mathbb{\Gamma}_3^{*\top} = (0.25, 0.00, 0.20)$. No penalty, part 3.

n	$n = 100$			$n = 250$			$n = 500$			$n = 1000$		
Measure	Bias	Var	MSE	Bias	Var	MSE	Bias	Var	MSE	Bias	Var	MSE
β_{30}	-18.9	3.9	7.5	-12.7	3.3	4.8	-14.2	2.2	4.3	-14.1	2.7	4.6
β_{31}	-4.3	3.2	3.3	-4.6	2.9	3.0	-2.3	2.2	2.2	1.0	1.5	1.5
β_{32}	0.5	1.6	1.6	0.2	0.9	0.9	0.6	0.5	0.5	0.3	0.5	0.5
β_{33}	-0.7	1.2	1.2	-0.1	0.7	0.7	-0.6	0.4	0.4	0.0	0.3	0.3
β_{34}	-0.8	1.3	1.3	0.3	0.6	0.6	0.4	0.4	0.4	-0.1	0.3	0.3
β_{35}	0.6	1.5	1.5	0.1	0.6	0.6	-0.3	0.5	0.5	-0.6	0.3	0.3
β_{36}	-0.1	1.9	1.9	0.3	0.5	0.5	0.1	0.4	0.4	0.2	0.3	0.3
β_{37}	-0.2	1.3	1.3	0.2	0.8	0.8	0.5	0.4	0.4	-0.2	0.3	0.3
β_{38}	1.6	1.3	1.3	0.3	0.7	0.7	-0.9	0.4	0.4	-0.5	0.3	0.3
β_{39}	2.7	1.7	1.8	-0.9	0.6	0.6	0.4	0.4	0.4	0.2	0.3	0.3
β_{310}	-0.6	1.4	1.4	-0.2	0.8	0.8	-0.1	0.5	0.5	-0.7	0.3	0.3
β_{311}	2.3	1.2	1.2	-0.3	0.6	0.6	-0.1	0.3	0.3	0.4	0.3	0.3
β_{312}	1.6	2.1	2.1	1.1	0.6	0.6	-0.3	0.4	0.4	0.4	0.3	0.3
β_{313}	0.8	1.2	1.2	-1.1	0.8	0.8	-0.1	0.3	0.3	0.2	0.4	0.4
β_{314}	-2.4	1.0	1.1	0.2	0.6	0.6	0.7	0.5	0.5	0.9	0.4	0.4
β_{315}	-0.2	1.9	1.9	-0.6	0.6	0.6	-0.5	0.4	0.4	-0.1	0.3	0.3
β_{316}	1.5	1.3	1.3	-0.3	0.7	0.7	-1.2	0.4	0.4	0.1	0.3	0.3
Γ_{31}	-15.8	1.7	4.2	-10.4	1.6	2.7	-5.4	1.3	1.5	-1.9	0.9	1.0
Γ_{312}	-2.9	1.1	1.2	-3.3	1.8	1.9	-8.2	1.6	2.2	-11.9	1.1	2.5
Γ_{32}	-18.0	0.5	3.7	-16.5	0.5	3.2	-14.2	0.6	2.7	-11.1	0.6	1.8
Total			193.8			79.4			47.5			33.1

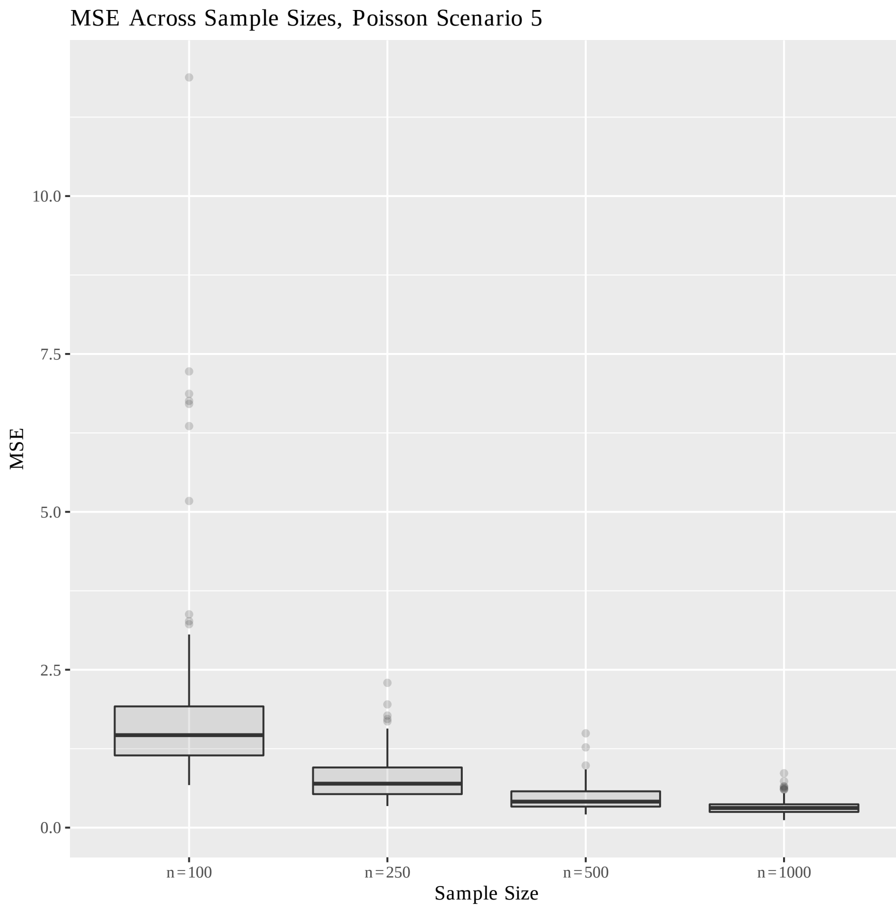

Figure H–13: MSE across sample sizes, Poisson outcome, no penalty, with $K = 3$, $p = 17$, $q = 2$.

Table H–24: Simulation 5 results multiplied by 100, averaged over 100 runs. Outcome follows a Poisson distribution with $K = 3$, $p = 17$, $q = 2$, $\pi_1 = 0.5$, $\pi_2 = 0.3$, $\pi_3 = 0.2$, $\boldsymbol{\beta}_1^\top = (0.65, 0.30, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00)$, $\boldsymbol{\beta}_2^\top = (0.20, -0.45, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00)$, $\boldsymbol{\beta}_3^\top = (1.00, 0.15, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00)$, $\mathbb{\Gamma}_1^{*\top} = (0.30, -0.25, 0.10)$, $\mathbb{\Gamma}_2^{*\top} = (0.35, 0.20, 0.15)$, and $\mathbb{\Gamma}_3^{*\top} = (0.25, 0.00, 0.20)$. Small values changed to zero, part 1.

n	$n = 100$			$n = 250$			$n = 500$			$n = 1000$		
Measure	Bias	Var	MSE	Bias	Var	MSE	Bias	Var	MSE	Bias	Var	MSE
π_1	-13.7	1.6	3.5	-13.8	1.3	3.2	-12.6	1.1	2.7	-11.2	0.9	2.1
π_2	0.8	1.4	1.4	4.7	1.1	1.3	6.0	0.6	0.9	7.8	0.4	1.0
β_{10}	-19.5	4.9	8.6	-7.4	3.1	3.6	-3.2	1.8	1.9	0.7	0.7	0.7
β_{11}	3.3	4.7	4.8	3.4	2.2	2.3	2.9	0.9	1.0	3.1	0.5	0.6
β_{12}	0.1	2.2	2.2	-1.1	0.7	0.7	-0.4	0.4	0.4	-0.4	0.2	0.2
β_{13}	0.1	2.1	2.0	-1.0	0.5	0.5	0.2	0.3	0.3	-0.5	0.1	0.1
β_{14}	0.1	2.6	2.5	0.8	0.7	0.7	-0.6	0.5	0.5	-0.1	0.2	0.2
β_{15}	-3.0	1.4	1.5	1.1	0.6	0.6	0.0	0.4	0.4	0.1	0.2	0.2
β_{16}	-0.1	1.9	1.9	1.3	0.6	0.6	-0.5	0.4	0.4	-0.8	0.1	0.1
β_{17}	-1.4	1.7	1.7	0.8	0.9	0.9	0.2	0.3	0.3	0.0	0.1	0.1
β_{18}	-2.1	2.1	2.1	-0.2	0.6	0.6	0.0	0.3	0.3	0.3	0.2	0.2
β_{19}	-0.6	1.6	1.5	1.1	0.6	0.6	0.8	0.6	0.6	0.7	0.2	0.2
β_{110}	-0.5	1.7	1.7	1.0	0.9	0.9	0.2	0.4	0.3	-0.2	0.2	0.1
β_{111}	-0.5	1.5	1.5	-0.8	0.7	0.7	0.6	0.3	0.3	-0.7	0.2	0.2
β_{112}	-0.8	1.8	1.8	0.4	0.7	0.7	-0.2	0.4	0.4	0.1	0.2	0.2
β_{113}	-2.1	2.6	2.6	1.5	0.6	0.6	-0.6	0.4	0.4	0.4	0.2	0.2
β_{114}	-1.2	2.1	2.1	-0.3	0.6	0.6	-0.9	0.4	0.4	0.1	0.2	0.2
β_{115}	3.6	1.7	1.8	0.5	0.8	0.8	-0.4	0.3	0.3	0.1	0.2	0.2
β_{116}	-1.2	1.5	1.5	2.0	0.6	0.7	1.4	0.3	0.3	0.3	0.2	0.2
Γ_{11}	-13.4	3.0	4.8	-8.2	1.6	2.3	-2.4	1.1	1.2	0.3	0.4	0.4
Γ_{112}	16.1	2.4	5.0	10.4	1.6	2.6	6.3	0.8	1.2	4.2	0.4	0.5
Γ_{12}	-8.3	0.4	1.0	-6.9	0.5	0.9	-4.9	0.4	0.6	-3.3	0.3	0.5

Table H–25: Simulation 5 results multiplied by 100, averaged over 100 runs. Outcome follows a Poisson distribution with $K = 3$, $p = 17$, $q = 2$, $\pi_1 = 0.5$, $\pi_2 = 0.3$, $\pi_3 = 0.2$, $\boldsymbol{\beta}_1^\top = (0.65, 0.30, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00)$, $\boldsymbol{\beta}_2^\top = (0.20, -0.45, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00)$, $\boldsymbol{\beta}_3^\top = (1.00, 0.15, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00)$, $\mathbb{\Gamma}_1^{*\top} = (0.30, -0.25, 0.10)$, $\mathbb{\Gamma}_2^{*\top} = (0.35, 0.20, 0.15)$, and $\mathbb{\Gamma}_3^{*\top} = (0.25, 0.00, 0.20)$. Small values changed to zero, part 2.

n	$n = 100$			$n = 250$			$n = 500$			$n = 1000$		
Measure	Bias	Var	MSE	Bias	Var	MSE	Bias	Var	MSE	Bias	Var	MSE
β_{20}	-5.7	18.1	18.2	3.5	3.6	3.7	8.9	1.6	2.4	12.1	0.9	2.3
β_{21}	11.1	10.4	11.5	12.7	4.1	5.7	11.6	1.6	3.0	11.3	0.6	1.9
β_{22}	1.1	5.8	5.7	-0.5	0.9	0.9	-0.3	0.5	0.4	0.3	0.1	0.1
β_{23}	-3.4	2.3	2.4	0.4	1.0	1.0	-0.3	0.3	0.3	0.5	0.1	0.1
β_{24}	2.2	3.9	3.9	-0.3	0.7	0.7	0.0	0.2	0.2	0.1	0.2	0.2
β_{25}	0.4	3.1	3.1	-1.8	0.6	0.6	-0.2	0.3	0.3	-0.7	0.1	0.1
β_{26}	-0.2	2.8	2.8	0.1	0.7	0.7	1.3	0.3	0.3	0.5	0.1	0.1
β_{27}	3.5	5.0	5.1	0.2	1.1	1.1	0.5	0.4	0.4	0.6	0.1	0.1
β_{28}	0.0	3.7	3.7	0.5	0.7	0.6	1.4	0.2	0.3	1.1	0.1	0.1
β_{29}	-3.0	3.6	3.6	0.6	0.8	0.8	0.0	0.3	0.3	0.1	0.1	0.1
β_{210}	-1.0	2.5	2.5	-0.6	0.7	0.7	-0.7	0.3	0.3	-0.4	0.1	0.1
β_{211}	-1.1	3.1	3.1	-0.7	0.8	0.8	-1.0	0.3	0.3	0.2	0.1	0.1
β_{212}	3.1	3.6	3.7	0.9	0.6	0.6	0.5	0.3	0.3	-0.4	0.1	0.1
β_{213}	2.1	2.6	2.6	0.6	0.9	0.9	0.5	0.3	0.3	0.3	0.1	0.1
β_{214}	-1.9	3.1	3.1	-1.2	0.7	0.7	-0.8	0.2	0.2	-0.4	0.1	0.1
β_{215}	-1.9	4.4	4.4	-1.6	0.6	0.7	0.4	0.3	0.3	-0.7	0.1	0.1
β_{216}	2.4	5.2	5.2	-1.1	0.7	0.7	0.4	0.3	0.3	-0.1	0.1	0.1
Γ_{21}	-10.4	3.7	4.7	-0.2	2.7	2.7	3.7	1.0	1.1	6.3	0.6	1.0
Γ_{212}	-7.5	2.5	3.0	3.8	1.5	1.6	5.0	0.9	1.1	7.6	0.4	1.0
Γ_{22}	-12.3	0.8	2.3	-7.1	1.7	2.2	-4.1	1.2	1.3	-1.8	0.6	0.6

Table H–26: Simulation 5 results multiplied by 100, averaged over 100 runs. Outcome follows a Poisson distribution with $K = 3$, $p = 17$, $q = 2$, $\pi_1 = 0.5$, $\pi_2 = 0.3$, $\pi_3 = 0.2$, $\boldsymbol{\beta}_1^\top = (0.65, 0.30, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00)$, $\boldsymbol{\beta}_2^\top = (0.20, -0.45, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00)$, $\boldsymbol{\beta}_3^\top = (1.00, 0.15, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00)$, $\mathbb{\Gamma}_1^{*\top} = (0.30, -0.25, 0.10)$, $\mathbb{\Gamma}_2^{*\top} = (0.35, 0.20, 0.15)$, and $\mathbb{\Gamma}_3^{*\top} = (0.25, 0.00, 0.20)$. Small values changed to zero, part 3.

n	$n = 100$			$n = 250$			$n = 500$			$n = 1000$		
Measure	Bias	Var	MSE	Bias	Var	MSE	Bias	Var	MSE	Bias	Var	MSE
β_{30}	-18.9	3.9	7.5	-12.7	3.3	4.8	-14.2	2.2	4.3	-14.1	2.7	4.6
β_{31}	-4.3	3.2	3.3	-4.6	2.8	3.0	-2.3	2.2	2.2	1.0	1.5	1.5
β_{32}	0.5	1.6	1.6	0.2	0.9	0.9	0.5	0.5	0.5	0.3	0.5	0.5
β_{33}	-0.7	1.2	1.2	0.0	0.7	0.7	-0.6	0.4	0.4	0.0	0.3	0.3
β_{34}	-0.8	1.3	1.3	0.3	0.6	0.6	0.4	0.4	0.4	-0.1	0.3	0.3
β_{35}	0.6	1.5	1.5	0.2	0.6	0.6	-0.3	0.5	0.5	-0.6	0.3	0.3
β_{36}	-0.1	1.9	1.9	0.3	0.5	0.5	0.1	0.4	0.4	0.2	0.3	0.3
β_{37}	-0.2	1.3	1.3	0.2	0.8	0.8	0.5	0.4	0.4	-0.2	0.3	0.3
β_{38}	1.6	1.3	1.3	0.3	0.7	0.7	-0.9	0.4	0.4	-0.4	0.3	0.3
β_{39}	2.7	1.7	1.8	-0.9	0.6	0.6	0.4	0.4	0.4	0.3	0.3	0.3
β_{310}	-0.7	1.4	1.4	-0.2	0.8	0.8	-0.1	0.5	0.5	-0.7	0.3	0.3
β_{311}	2.3	1.2	1.2	-0.3	0.6	0.6	-0.1	0.3	0.3	0.5	0.3	0.3
β_{312}	1.6	2.1	2.1	1.1	0.6	0.6	-0.3	0.4	0.4	0.4	0.3	0.3
β_{313}	0.8	1.2	1.2	-1.1	0.8	0.8	-0.1	0.3	0.3	0.2	0.4	0.4
β_{314}	-2.5	1.0	1.1	0.1	0.6	0.6	0.7	0.5	0.5	0.9	0.4	0.4
β_{315}	-0.2	1.9	1.9	-0.6	0.6	0.6	-0.5	0.4	0.4	-0.1	0.3	0.3
β_{316}	1.5	1.3	1.3	-0.3	0.7	0.7	-1.2	0.4	0.4	0.1	0.3	0.3
Γ_{31}	-15.8	1.7	4.2	-10.4	1.6	2.7	-5.4	1.3	1.5	-1.9	0.9	1.0
Γ_{312}	-2.9	1.1	1.2	-3.3	1.8	1.9	-8.2	1.6	2.2	-11.9	1.1	2.5
Γ_{32}	-18.0	0.5	3.7	-16.5	0.5	3.2	-14.2	0.6	2.7	-11.1	0.6	1.8
Total			193.8			79.4			47.5			33.1

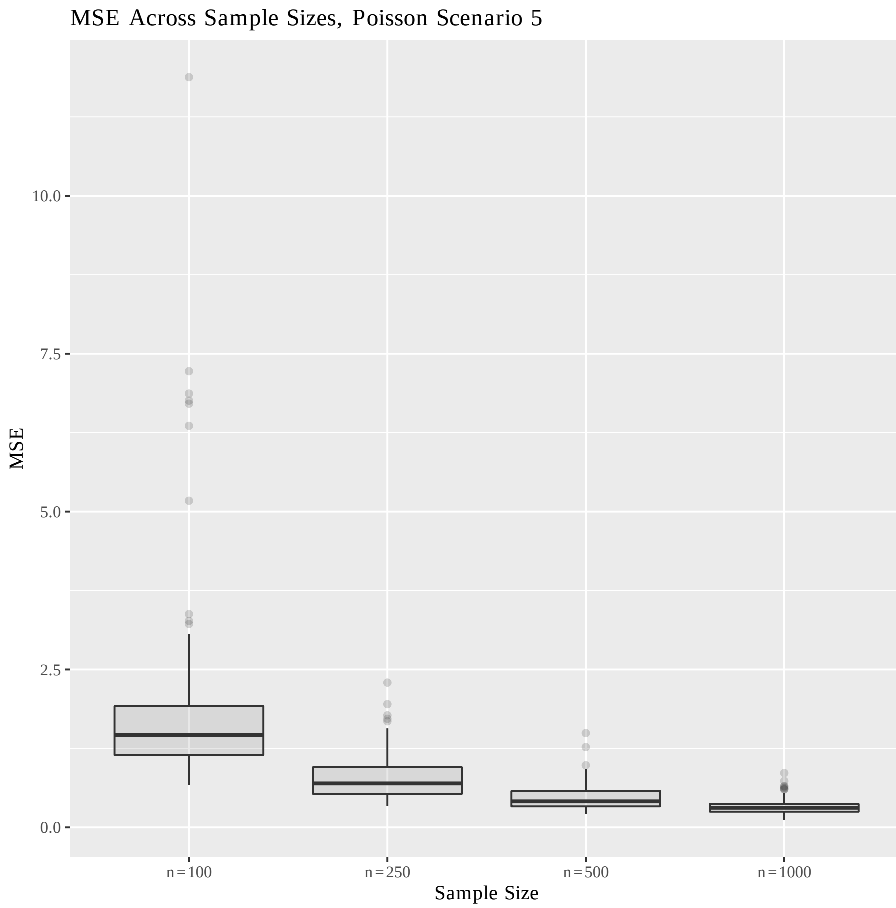

Figure H–14: MSE across sample sizes, Poisson outcome, small values changed to zero, with $K = 3$, $p = 17$, $q = 2$.

Table H–27: Simulation 5 results multiplied by 100, averaged over 100 runs. Outcome follows a Poisson distribution with $K = 3$, $p = 2$, $q = 2$, $\pi_1 = 0.5$, $\pi_2 = 0.3$, $\pi_3 = 0.2$, $\boldsymbol{\beta}_1^\top = (0.65, 0.30)$, $\boldsymbol{\beta}_2^\top = (0.20, -0.45)$, $\boldsymbol{\beta}_3^\top = (1.00, 0.15)$, $\mathbb{\Gamma}_1^{*\top} = (0.30, -0.25, 0.10)$, $\mathbb{\Gamma}_2^{*\top} = (0.35, 0.20, 0.15)$, and $\mathbb{\Gamma}_3^{*\top} = (0.25, 0.00, 0.20)$. Oracle model.

n	$n = 100$			$n = 250$			$n = 500$			$n = 1000$		
Measure	Bias	Var	MSE	Bias	Var	MSE	Bias	Var	MSE	Bias	Var	MSE
π_1	-9.9	1.0	2.0	-10.0	0.7	1.6	-9.0	0.5	1.3	-10.0	0.5	1.5
π_2	0.7	0.8	0.8	2.7	0.4	0.5	3.2	0.3	0.4	4.6	0.2	0.4
β_{10}	-11.1	2.4	3.6	-3.8	1.1	1.2	-5.0	1.0	1.2	-3.7	0.5	0.7
β_{11}	10.1	2.0	3.0	6.0	1.1	1.4	5.6	0.9	1.2	6.3	0.6	0.9
Γ_{11}	-9.0	2.5	3.3	-6.4	1.3	1.7	-5.3	0.9	1.2	-2.6	0.6	0.6
Γ_{112}	11.4	2.0	3.3	9.4	1.4	2.2	8.2	0.9	1.5	5.9	0.4	0.8
Γ_{12}	-5.5	0.5	0.8	-3.4	0.6	0.7	-1.9	0.6	0.7	-1.5	0.4	0.4
β_{20}	0.2	7.5	7.5	1.6	1.9	1.9	5.2	1.0	1.3	6.3	0.6	1.0
β_{21}	1.7	4.6	4.6	6.1	2.1	2.4	6.8	1.0	1.5	6.7	0.5	1.0
Γ_{21}	-2.2	3.9	3.9	-0.6	2.2	2.2	1.4	0.9	0.9	3.3	0.7	0.8
Γ_{212}	-1.0	2.6	2.5	2.6	1.3	1.4	2.3	0.9	1.0	4.6	0.6	0.8
Γ_{22}	-7.4	1.8	2.3	-4.1	1.3	1.5	-2.8	1.2	1.3	-0.8	0.4	0.4
β_{30}	-2.1	3.5	3.5	-1.9	2.3	2.3	-1.4	2.1	2.1	-1.7	1.7	1.7
β_{31}	-6.5	1.5	1.9	-7.0	1.3	1.8	-6.0	1.0	1.3	-4.1	1.0	1.1
Γ_{31}	-11.0	1.5	2.7	-6.6	1.0	1.4	-4.6	0.7	0.9	-3.0	0.5	0.6
Γ_{312}	0.1	1.6	1.6	-3.3	1.5	1.6	-4.6	1.0	1.2	-6.4	0.9	1.3
Γ_{32}	-14.3	0.6	2.6	-11.8	0.6	2.0	-7.6	0.8	1.3	-5.8	0.7	1.0
Total			50.1			27.8			20.4			15.1

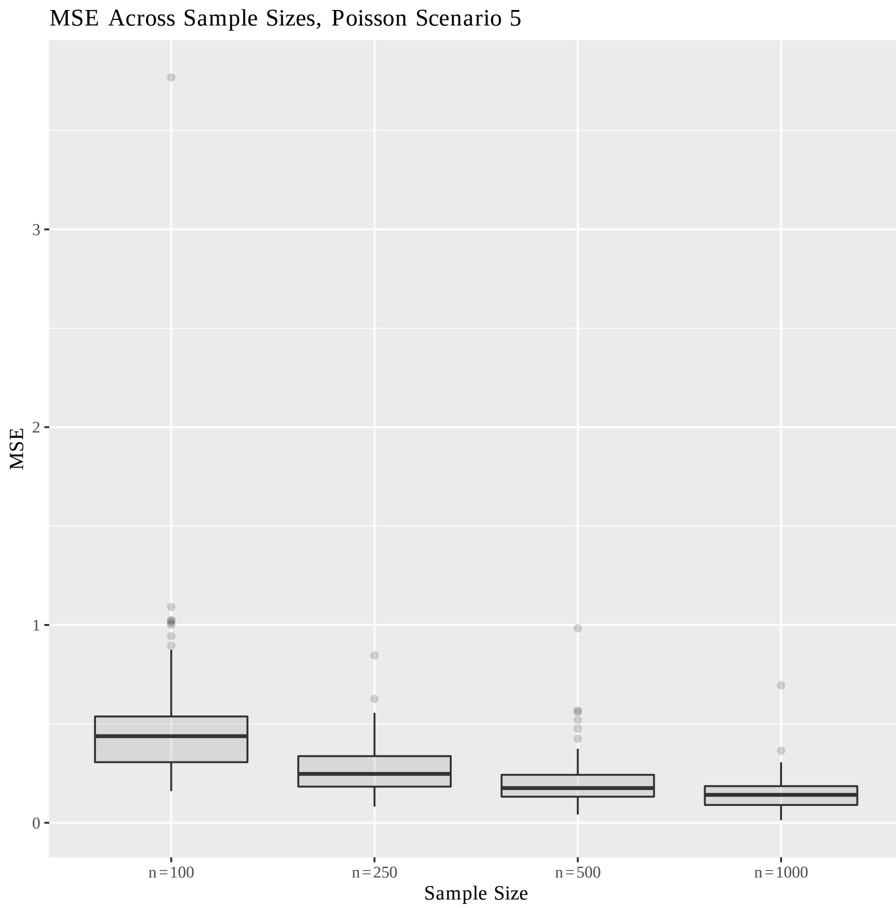

Figure H–15: MSE across sample sizes, Poisson outcome, oracle model, with $K = 3$, $p = 17$, $q = 2$.

Table H–28: Simulation 6 results multiplied by 100, averaged over 100 runs. Outcome follows a Poisson distribution with $K = 3$, $p = 19$, $q = 5$, $\pi_1 = 0.5$, $\pi_2 = 0.3$, $\pi_3 = 0.2$, $\boldsymbol{\beta}_1^\top = (0.65, 0.30, 0.15, 0.35, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00)$, $\boldsymbol{\beta}_2^\top = (0.20, -0.45, -0.10, 0.25, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00)$, $\boldsymbol{\beta}_3^\top = (1.00, 0.15, -0.65, -0.15, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00)$, $\Gamma_1^{*\top} = (0.30, -0.25, 0.10, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00)$, $\Gamma_2^{*\top} = (0.35, 0.20, 0.15, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00)$, and $\Gamma_3^{*\top} = (0.25, 0.00, 0.20, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00)$. No penalty, part 1.

n	$n = 100$			$n = 250$			$n = 500$			$n = 1000$		
Measure	Bias	Var	MSE	Bias	Var	MSE	Bias	Var	MSE	Bias	Var	MSE
π_1	-10.2	1.2	2.2	-7.6	0.8	1.4	-5.8	0.5	0.8	-4.3	0.2	0.4
π_2	3.8	0.8	1.0	4.0	0.9	1.0	5.4	0.3	0.6	4.3	0.2	0.4
β_{10}	-11.7	20.0	21.1	-5.2	1.2	1.4	-0.9	0.4	0.4	-1.0	0.1	0.1
β_{11}	-0.2	11.0	10.9	1.6	1.0	1.0	1.3	0.5	0.5	2.2	0.2	0.2
β_{12}	-2.8	2.6	2.7	-2.1	0.8	0.9	0.0	0.3	0.3	-0.6	0.2	0.2
β_{13}	-7.2	1.3	1.8	-3.5	0.5	0.6	-1.4	0.1	0.2	-1.2	0.1	0.1
β_{14}	1.6	2.1	2.1	-0.7	0.3	0.3	0.1	0.1	0.1	0.0	< 0.1	< 0.1
β_{15}	1.1	4.9	4.8	0.2	0.3	0.3	0.3	0.1	0.1	0.5	< 0.1	< 0.1
β_{16}	0.3	1.4	1.4	-0.3	0.3	0.3	0.1	0.1	0.1	-0.1	< 0.1	< 0.1
β_{17}	-0.1	1.8	1.8	0.4	0.3	0.3	0.4	0.1	0.1	0.3	< 0.1	< 0.1
β_{18}	1.0	1.3	1.3	0.8	0.4	0.4	0.1	0.1	0.1	0.1	< 0.1	< 0.1
β_{19}	1.5	1.0	1.0	0.4	0.3	0.3	-0.3	0.1	0.1	-0.1	< 0.1	< 0.1
β_{110}	-0.9	0.7	0.7	-0.9	0.3	0.3	0.1	0.1	0.1	-0.2	< 0.1	< 0.1
β_{111}	0.0	1.3	1.3	0.7	0.2	0.2	0.0	0.1	0.1	-0.1	< 0.1	< 0.1
β_{112}	-0.3	1.5	1.5	-0.1	0.3	0.3	-0.2	0.1	0.1	0.2	< 0.1	< 0.1
β_{113}	1.8	5.6	5.6	0.2	0.4	0.4	0.4	0.1	0.1	0.4	< 0.1	< 0.1
β_{114}	0.4	1.1	1.0	-0.9	0.4	0.4	-0.6	0.1	0.1	-0.3	< 0.1	< 0.1
β_{115}	0.3	3.1	3.1	1.0	0.3	0.3	0.2	0.1	0.1	0.2	< 0.1	< 0.1
β_{116}	-0.3	3.5	3.5	0.6	0.3	0.3	0.3	0.1	0.1	0.1	< 0.1	< 0.1
β_{117}	1.9	2.4	2.5	-0.4	0.3	0.3	0.1	0.1	0.1	0.3	< 0.1	< 0.1
β_{118}	-0.3	1.0	1.0	0.2	0.3	0.3	-0.3	0.1	0.1	-0.2	< 0.1	< 0.1
Γ_{11}	-13.1	2.3	4.0	-5.7	1.4	1.7	-2.2	0.3	0.4	-0.5	0.2	0.2
Γ_{112}	15.8	2.2	4.7	6.8	1.3	1.8	2.7	0.3	0.4	0.4	0.2	0.2
Γ_{12}	-3.8	1.0	1.1	-4.0	0.6	0.7	-4.8	0.3	0.5	-4.4	0.2	0.4
Γ_{113}	-7.2	1.5	2.0	-5.4	1.1	1.4	-2.5	0.4	0.5	-1.4	0.4	0.4
Γ_{123}	2.8	0.8	0.9	-0.4	0.3	0.3	-1.3	0.3	0.3	-1.4	0.2	0.3
Γ_{13}	2.0	0.3	0.3	1.0	0.1	0.1	1.1	< 0.1	0.1	1.7	0.1	0.1
Γ_{114}	-4.1	1.0	1.2	-4.3	0.6	0.8	-1.2	0.2	0.2	-1.3	0.1	0.2
Γ_{124}	0.8	0.5	0.5	-1.2	0.3	0.3	-1.4	0.2	0.2	-1.8	0.1	0.2
Γ_{134}	1.6	0.4	0.5	0.7	0.1	0.1	0.4	0.1	0.1	0.8	0.1	0.1
Γ_{14}	0.2	0.0	0.0	0.2	< 0.1	< 0.1	0.6	< 0.1	< 0.1	0.8	< 0.1	< 0.1
Γ_{115}	-0.8	0.5	0.5	-0.2	0.2	0.2	0.2	0.1	0.1	-0.1	0.1	0.1
Γ_{125}	0.4	0.2	0.2	0.5	0.3	0.3	0.2	0.2	0.2	0.3	0.1	0.1
Γ_{135}	0.6	0.1	0.1	0.3	< 0.1	< 0.1	0.7	0.1	0.1	0.1	< 0.1	< 0.1
Γ_{145}	-0.1	0.0	0.0	0.0	< 0.1	< 0.1	0.0	< 0.1	< 0.1	0.0	< 0.1	< 0.1
Γ_{15}	0.0	0.0	0.0	0.1	< 0.1	< 0.1	0.1	< 0.1	< 0.1	0.2	< 0.1	< 0.1

Table H–29: Simulation 6 results multiplied by 100, averaged over 100 runs. Outcome follows a Poisson distribution with $K = 3$, $p = 19$, $q = 5$, $\pi_1 = 0.5$, $\pi_2 = 0.3$, $\pi_3 = 0.2$, $\boldsymbol{\beta}_1^\top = (0.65, 0.30, 0.15, 0.35, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00)$, $\boldsymbol{\beta}_2^\top = (0.20, -0.45, -0.10, 0.25, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00)$, $\boldsymbol{\beta}_3^\top = (1.00, 0.15, -0.65, -0.15, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00)$, $\mathbb{\Gamma}_1^{*\top} = (0.30, -0.25, 0.10, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00)$, $\mathbb{\Gamma}_2^{*\top} = (0.35, 0.20, 0.15, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00)$, and $\mathbb{\Gamma}_3^{*\top} = (0.25, 0.00, 0.20, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00)$. No penalty, part 2.

n	$n = 100$			$n = 250$			$n = 500$			$n = 1000$		
Measure	Bias	Var	MSE	Bias	Var	MSE	Bias	Var	MSE	Bias	Var	MSE
β_{20}	0.3	13.1	13.0	3.0	2.6	2.6	4.3	1.4	1.5	5.0	0.5	0.8
β_{21}	14.5	6.8	8.8	9.5	3.1	4.0	8.3	1.4	2.0	6.4	0.6	1.0
β_{22}	-0.8	6.2	6.2	1.8	1.7	1.7	-0.6	0.8	0.8	1.6	0.4	0.4
β_{23}	-5.4	2.6	2.9	-1.5	0.9	0.9	-1.5	0.5	0.5	0.2	0.2	0.2
β_{24}	2.7	2.5	2.6	1.4	0.4	0.4	0.2	0.2	0.2	-0.1	0.1	0.1
β_{25}	-2.3	2.0	2.0	-0.4	0.7	0.7	-0.3	0.3	0.3	0.0	0.1	0.1
β_{26}	-1.7	3.2	3.2	0.3	0.6	0.6	0.2	0.2	0.2	0.3	0.1	0.1
β_{27}	0.7	2.0	2.0	-0.5	0.6	0.6	-0.5	0.2	0.2	-0.3	0.1	0.1
β_{28}	0.0	4.8	4.8	-2.0	0.7	0.7	-0.9	0.2	0.2	-0.6	0.1	0.1
β_{29}	0.3	1.8	1.8	-0.6	0.9	0.9	0.2	0.3	0.3	0.1	0.1	0.1
β_{210}	-1.1	3.0	3.0	1.0	0.6	0.6	0.2	0.2	0.2	0.4	0.1	0.1
β_{211}	0.6	1.8	1.8	0.7	0.8	0.8	-0.1	0.3	0.3	0.1	0.1	0.1
β_{212}	-1.8	2.4	2.4	0.7	1.0	1.0	-0.1	0.3	0.3	-0.3	0.1	0.1
β_{213}	-2.2	4.3	4.3	-0.2	1.2	1.1	-0.7	0.3	0.3	-0.6	0.1	0.2
β_{214}	-1.3	1.9	1.9	0.2	0.6	0.6	0.3	0.2	0.2	0.2	0.1	0.1
β_{215}	2.0	2.6	2.6	-0.1	0.8	0.8	0.5	0.2	0.2	0.0	0.1	0.1
β_{216}	2.2	1.3	1.3	1.0	0.9	0.9	0.0	0.2	0.2	0.1	0.1	0.1
β_{217}	3.2	2.6	2.6	0.1	0.6	0.5	-0.1	0.3	0.3	0.4	0.1	0.1
β_{218}	-0.5	2.4	2.3	-1.2	1.3	1.3	-0.5	0.3	0.3	-0.4	0.1	0.1
Γ_{21}	-4.1	4.9	5.0	2.1	2.2	2.3	2.3	1.3	1.3	2.1	0.5	0.6
Γ_{212}	-1.2	3.3	3.3	3.4	1.6	1.7	3.3	1.2	1.3	2.0	0.6	0.6
Γ_{22}	-4.0	2.8	2.9	-4.6	1.4	1.6	-1.7	0.8	0.8	-0.6	0.6	0.6
Γ_{213}	0.3	3.0	2.9	2.6	2.0	2.0	0.2	1.8	1.7	4.4	0.7	0.9
Γ_{223}	3.2	1.5	1.5	2.7	1.2	1.2	0.2	0.9	0.9	-0.4	0.7	0.7
Γ_{23}	2.0	0.3	0.3	4.0	0.6	0.7	4.7	0.4	0.6	6.0	0.4	0.7
Γ_{214}	-3.7	2.6	2.7	-1.7	1.0	1.0	-1.8	0.7	0.8	0.9	0.4	0.4
Γ_{224}	3.8	1.2	1.3	3.5	0.6	0.7	3.1	0.5	0.6	1.7	0.3	0.4
Γ_{234}	1.3	0.1	0.1	1.6	0.2	0.3	2.4	0.3	0.3	2.1	0.2	0.2
Γ_{24}	0.1	0.0	0.0	0.3	< 0.1	< 0.1	0.4	< 0.1	< 0.1	1.0	< 0.1	0.1
Γ_{215}	-0.7	1.0	1.0	0.4	0.7	0.6	-0.4	0.3	0.3	-0.1	0.2	0.2
Γ_{225}	0.0	0.6	0.6	0.5	0.5	0.5	0.6	0.3	0.3	0.7	0.3	0.3
Γ_{235}	-0.1	0.0	0.0	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.2	0.2
Γ_{245}	0.0	0.0	0.0	0.1	< 0.1	< 0.1	0.1	< 0.1	< 0.1	-0.1	< 0.1	< 0.1
Γ_{25}	0.0	0.0	0.0	0.1	< 0.1	< 0.1	0.0	< 0.1	< 0.1	0.2	< 0.1	< 0.1

Table H–30: Simulation 6 results multiplied by 100, averaged over 100 runs. Outcome follows a Poisson distribution with $K = 3$, $p = 19$, $q = 5$, $\pi_1 = 0.5$, $\pi_2 = 0.3$, $\pi_3 = 0.2$, $\boldsymbol{\beta}_1^\top = (0.65, 0.30, 0.15, 0.35, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00)$, $\boldsymbol{\beta}_2^\top = (0.20, -0.45, -0.10, 0.25, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00)$, $\boldsymbol{\beta}_3^\top = (1.00, 0.15, -0.65, -0.15, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00)$, $\Gamma_1^{*\top} = (0.30, -0.25, 0.10, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00)$, $\Gamma_2^{*\top} = (0.35, 0.20, 0.15, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00)$, and $\Gamma_3^{*\top} = (0.25, 0.00, 0.20, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00)$. No penalty, part 3.

n	$n = 100$			$n = 250$			$n = 500$			$n = 1000$		
Measure	Bias	Var	MSE	Bias	Var	MSE	Bias	Var	MSE	Bias	Var	MSE
β_{30}	-15.3	5.7	8.0	-8.5	3.1	3.8	-2.2	1.3	1.3	0.4	0.5	0.5
β_{31}	-0.1	4.5	4.4	-2.3	1.9	2.0	-0.9	0.8	0.8	-0.9	0.4	0.4
β_{32}	19.1	8.5	12.0	13.5	4.7	6.5	7.6	3.5	4.0	4.5	1.7	1.9
β_{33}	18.3	3.2	6.6	11.3	2.0	3.2	5.1	1.0	1.3	3.0	0.8	0.9
β_{34}	0.7	1.4	1.4	0.5	0.5	0.5	0.2	0.2	0.2	-0.1	0.1	0.1
β_{35}	1.2	2.1	2.1	-0.1	0.7	0.7	-0.4	0.2	0.2	-0.5	0.1	0.1
β_{36}	-4.4	1.9	2.1	-0.6	0.6	0.6	-0.3	0.3	0.3	0.1	0.1	0.1
β_{37}	0.1	1.5	1.5	1.0	0.5	0.5	-0.1	0.2	0.2	0.0	0.1	0.1
β_{38}	-1.9	1.3	1.3	-0.9	0.7	0.7	-0.9	0.3	0.3	0.1	0.1	0.1
β_{39}	0.5	1.3	1.3	-0.5	0.4	0.4	0.0	0.2	0.2	-0.2	0.1	0.1
β_{310}	-0.1	1.6	1.6	-0.1	0.5	0.5	0.0	0.2	0.2	-0.2	0.1	0.1
β_{311}	1.2	2.3	2.3	-1.5	0.6	0.6	0.0	0.2	0.2	0.0	0.1	0.1
β_{312}	2.2	1.8	1.9	0.1	0.5	0.5	0.1	0.3	0.3	0.5	0.1	0.1
β_{313}	1.5	1.6	1.6	-0.3	0.7	0.7	0.5	0.2	0.2	0.3	0.1	0.1
β_{314}	1.7	1.5	1.5	0.2	0.5	0.5	0.2	0.2	0.2	0.3	0.1	0.1
β_{315}	1.9	1.7	1.7	-0.3	0.3	0.3	-1.2	0.2	0.2	-0.7	0.1	0.1
β_{316}	1.3	2.4	2.4	-0.6	0.5	0.5	0.7	0.2	0.2	-0.6	0.1	0.1
β_{317}	0.4	2.0	2.0	0.4	0.5	0.5	0.7	0.2	0.2	0.1	0.1	0.1
β_{318}	2.6	2.5	2.5	-0.2	0.5	0.5	0.7	0.3	0.3	-0.6	0.1	0.1
Γ_{31}	-10.1	3.0	3.9	-5.4	1.5	1.8	-4.5	0.6	0.8	-3.4	0.4	0.5
Γ_{312}	-1.7	1.9	1.9	0.2	1.8	1.8	-2.3	0.8	0.9	-0.9	0.4	0.4
Γ_{32}	-16.2	0.7	3.3	-13.2	0.9	2.6	-8.6	0.8	1.5	-4.1	0.4	0.6
Γ_{313}	-4.0	2.5	2.6	-2.1	2.1	2.2	0.9	1.3	1.3	0.3	0.4	0.4
Γ_{323}	2.4	1.0	1.1	1.3	0.9	0.9	-0.9	0.5	0.5	-0.9	0.3	0.3
Γ_{33}	3.6	0.5	0.7	6.2	0.9	1.3	6.5	1.0	1.4	6.0	0.6	0.9
Γ_{314}	-5.7	1.5	1.8	-2.8	1.1	1.2	-0.7	0.5	0.5	-0.5	0.3	0.3
Γ_{324}	1.6	0.7	0.7	0.4	0.6	0.6	0.5	0.4	0.4	0.3	0.2	0.2
Γ_{334}	2.2	0.4	0.4	3.5	0.5	0.6	3.9	0.4	0.5	3.6	0.4	0.5
Γ_{34}	0.1	0.0	0.0	0.1	< 0.1	< 0.1	0.3	< 0.1	< 0.1	0.9	< 0.1	< 0.1
Γ_{315}	-1.0	1.0	1.0	-1.0	0.5	0.5	0.3	0.3	0.3	0.0	0.1	0.1
Γ_{325}	0.4	0.3	0.3	-0.4	0.2	0.2	0.3	0.2	0.2	0.1	0.1	0.1
Γ_{335}	-0.1	0.1	0.1	-0.1	0.1	0.1	0.3	0.1	0.1	0.6	0.2	0.2
Γ_{345}	0.0	0.0	0.0	0.0	< 0.1	< 0.1	0.0	< 0.1	< 0.1	-0.2	0.1	0.1
Γ_{35}	0.0	0.0	0.0	0.0	< 0.1	< 0.1	0.0	< 0.1	< 0.1	0.4	< 0.1	< 0.1
Total			255.8			89.7			44.6			23.3

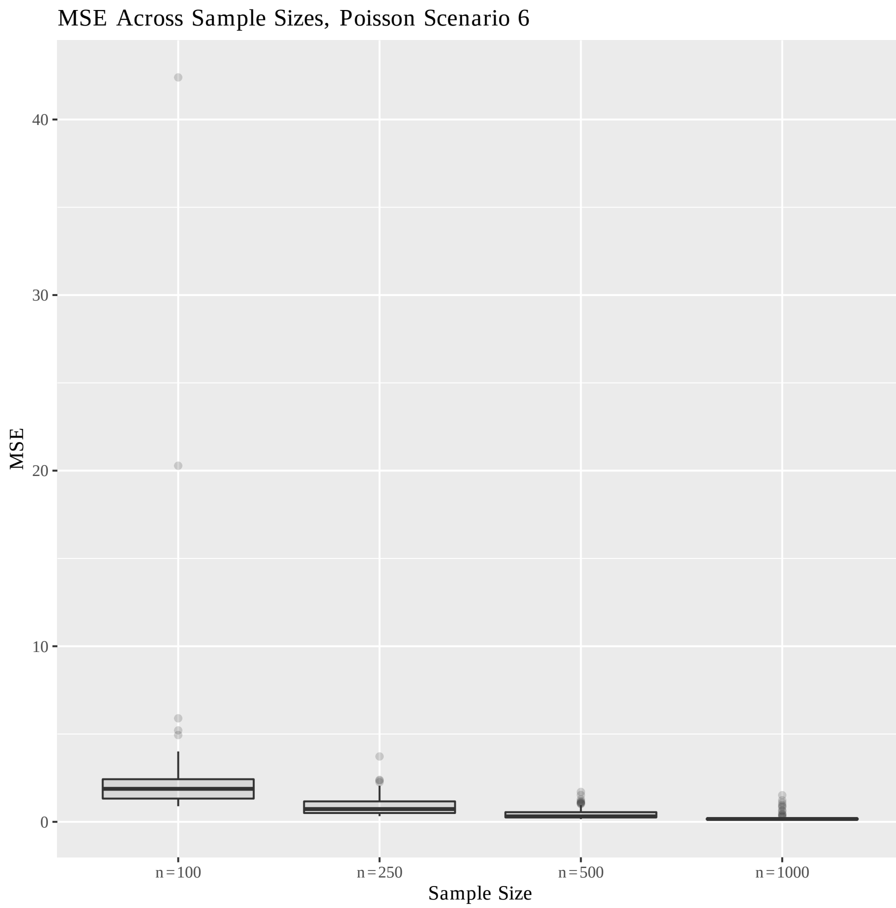

Figure H–16: MSE across sample sizes, Poisson outcome, no penalty, with $K = 3$, $p = 19$, $q = 5$.

Table H–31: Simulation 6 results multiplied by 100, averaged over 100 runs. Outcome follows a Poisson distribution with $K = 3$, $p = 19$, $q = 5$, $\pi_1 = 0.5$, $\pi_2 = 0.3$, $\pi_3 = 0.2$, $\boldsymbol{\beta}_1^\top = (0.65, 0.30, 0.15, 0.35, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00)$, $\boldsymbol{\beta}_2^\top = (0.20, -0.45, -0.10, 0.25, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00)$, $\boldsymbol{\beta}_3^\top = (1.00, 0.15, -0.65, -0.15, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00)$, $\mathbb{\Gamma}_1^{*\top} = (0.30, -0.25, 0.10, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00)$, $\mathbb{\Gamma}_2^{*\top} = (0.35, 0.20, 0.15, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00)$, and $\mathbb{\Gamma}_3^{*\top} = (0.25, 0.00, 0.20, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00)$. Small values changed to zero, part 1.

n	$n = 100$			$n = 250$			$n = 500$			$n = 1000$		
Measure	Bias	Var	MSE	Bias	Var	MSE	Bias	Var	MSE	Bias	Var	MSE
π_1	-10.2	1.2	2.2	-7.6	0.8	1.4	-5.8	0.5	0.8	-4.3	0.2	0.4
π_2	3.8	0.8	1.0	4.0	0.9	1.0	5.4	0.3	0.6	4.3	0.2	0.4
β_{10}	-11.7	20.0	21.1	-5.2	1.2	1.4	-0.9	0.4	0.4	-1.0	0.1	0.1
β_{11}	-0.2	11.0	10.9	1.6	1.0	1.0	1.3	0.5	0.5	2.2	0.2	0.2
β_{12}	-2.8	2.6	2.7	-2.1	0.8	0.9	0.0	0.3	0.3	-0.6	0.2	0.2
β_{13}	-7.2	1.3	1.8	-3.5	0.5	0.6	-1.4	0.1	0.2	-1.2	0.1	0.1
β_{14}	1.6	2.1	2.1	-0.8	0.3	0.3	0.2	0.1	0.1	-0.1	< 0.1	< 0.1
β_{15}	1.1	4.9	4.8	0.2	0.3	0.3	0.3	0.1	0.1	0.5	< 0.1	< 0.1
β_{16}	0.3	1.4	1.4	-0.2	0.3	0.3	0.1	0.1	0.1	-0.1	< 0.1	< 0.1
β_{17}	-0.2	1.8	1.8	0.4	0.3	0.3	0.4	0.1	0.1	0.3	< 0.1	< 0.1
β_{18}	1.0	1.3	1.3	0.8	0.4	0.4	0.0	0.1	0.1	0.1	< 0.1	< 0.1
β_{19}	1.5	1.0	1.0	0.4	0.3	0.3	-0.3	0.1	0.1	-0.1	< 0.1	< 0.1
β_{110}	-0.9	0.7	0.7	-0.9	0.3	0.3	0.1	0.1	0.1	-0.2	< 0.1	< 0.1
β_{111}	0.0	1.3	1.3	0.7	0.2	0.2	0.1	0.1	0.1	-0.1	< 0.1	< 0.1
β_{112}	-0.3	1.5	1.5	0.0	0.3	0.3	-0.1	0.1	0.1	0.2	< 0.1	< 0.1
β_{113}	1.8	5.6	5.6	0.2	0.4	0.4	0.4	0.1	0.1	0.3	< 0.1	< 0.1
β_{114}	0.4	1.1	1.0	-0.9	0.4	0.4	-0.5	0.1	0.1	-0.2	< 0.1	< 0.1
β_{115}	0.3	3.1	3.1	1.0	0.3	0.3	0.2	0.1	0.1	0.1	< 0.1	< 0.1
β_{116}	-0.3	3.5	3.5	0.6	0.3	0.3	0.3	0.1	0.1	0.1	< 0.1	< 0.1
β_{117}	1.9	2.4	2.5	-0.4	0.3	0.3	0.1	0.1	0.1	0.2	< 0.1	< 0.1
β_{118}	-0.3	1.0	1.0	0.2	0.3	0.3	-0.3	0.1	0.1	-0.2	< 0.1	< 0.1
Γ_{11}	-13.1	2.3	4.0	-5.7	1.4	1.7	-2.2	0.3	0.4	-0.5	0.2	0.2
Γ_{112}	15.8	2.2	4.7	6.8	1.3	1.8	2.7	0.3	0.4	0.4	0.2	0.2
Γ_{12}	-3.8	1.0	1.1	-4.0	0.6	0.7	-4.8	0.3	0.5	-4.4	0.2	0.4
Γ_{113}	-7.2	1.5	2.0	-5.4	1.1	1.4	-2.5	0.4	0.5	-1.4	0.4	0.4
Γ_{123}	2.8	0.8	0.9	-0.4	0.3	0.3	-1.3	0.3	0.3	-1.4	0.2	0.3
Γ_{13}	2.0	0.3	0.3	1.0	0.1	0.1	1.1	< 0.1	0.1	1.7	0.1	0.1
Γ_{114}	-4.1	1.0	1.2	-4.3	0.6	0.8	-1.2	0.2	0.2	-1.3	0.1	0.2
Γ_{124}	0.8	0.5	0.5	-1.2	0.3	0.3	-1.4	0.2	0.2	-1.8	0.1	0.2
Γ_{134}	1.6	0.4	0.5	0.7	0.1	0.1	0.4	0.1	0.1	0.8	0.1	0.1
Γ_{14}	0.2	0.0	0.0	0.2	< 0.1	< 0.1	0.6	< 0.1	< 0.1	0.8	< 0.1	< 0.1
Γ_{115}	-0.8	0.5	0.5	-0.2	0.2	0.2	0.2	0.1	0.1	-0.1	0.1	0.1
Γ_{125}	0.4	0.2	0.2	0.5	0.3	0.3	0.2	0.2	0.2	0.3	0.1	0.1
Γ_{135}	0.6	0.1	0.1	0.3	< 0.1	< 0.1	0.7	0.1	0.1	0.1	< 0.1	< 0.1
Γ_{145}	-0.1	0.0	0.0	0.0	< 0.1	< 0.1	0.0	< 0.1	< 0.1	0.0	< 0.1	< 0.1
Γ_{15}	0.0	0.0	0.0	0.1	< 0.1	< 0.1	0.1	< 0.1	< 0.1	0.2	< 0.1	< 0.1

Table H–32: Simulation 6 results multiplied by 100, averaged over 100 runs. Outcome follows a Poisson distribution with $K = 3$, $p = 19$, $q = 5$, $\pi_1 = 0.5$, $\pi_2 = 0.3$, $\pi_3 = 0.2$, $\boldsymbol{\beta}_1^\top = (0.65, 0.30, 0.15, 0.35, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00)$, $\boldsymbol{\beta}_2^\top = (0.20, -0.45, -0.10, 0.25, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00)$, $\boldsymbol{\beta}_3^\top = (1.00, 0.15, -0.65, -0.15, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00)$, $\mathbb{\Gamma}_1^{*\top} = (0.30, -0.25, 0.10, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00)$, $\mathbb{\Gamma}_2^{*\top} = (0.35, 0.20, 0.15, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00)$, and $\mathbb{\Gamma}_3^{*\top} = (0.25, 0.00, 0.20, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00)$. Small values changed to zero, part 2.

n	$n = 100$			$n = 250$			$n = 500$			$n = 1000$		
Measure	Bias	Var	MSE	Bias	Var	MSE	Bias	Var	MSE	Bias	Var	MSE
β_{20}	0.3	13.1	13.0	3.0	2.6	2.6	4.3	1.4	1.5	5.0	0.5	0.8
β_{21}	14.5	6.8	8.8	9.5	3.1	4.0	8.3	1.4	2.0	6.4	0.6	1.0
β_{22}	-0.8	6.2	6.2	1.8	1.7	1.7	-0.6	0.8	0.8	1.6	0.4	0.4
β_{23}	-5.4	2.6	2.9	-1.5	0.9	0.9	-1.5	0.5	0.5	0.2	0.2	0.2
β_{24}	2.7	2.5	2.6	1.4	0.4	0.4	0.1	0.2	0.2	-0.1	0.1	0.1
β_{25}	-2.3	2.0	2.0	-0.4	0.7	0.7	-0.3	0.3	0.3	0.0	0.1	0.1
β_{26}	-1.7	3.2	3.2	0.3	0.6	0.6	0.2	0.2	0.2	0.4	0.1	0.1
β_{27}	0.7	2.0	2.0	-0.5	0.6	0.6	-0.4	0.2	0.2	-0.3	0.1	0.1
β_{28}	-0.1	4.8	4.8	-2.0	0.7	0.7	-0.9	0.2	0.2	-0.6	0.1	0.1
β_{29}	0.3	1.8	1.8	-0.6	0.9	0.9	0.3	0.3	0.3	0.1	0.1	0.1
β_{210}	-1.1	3.0	3.0	0.9	0.6	0.6	0.2	0.2	0.2	0.4	0.1	0.1
β_{211}	0.6	1.8	1.8	0.7	0.8	0.8	-0.1	0.3	0.3	0.1	0.1	0.1
β_{212}	-1.8	2.4	2.4	0.7	1.0	1.0	-0.1	0.3	0.3	-0.3	0.1	0.1
β_{213}	-2.1	4.3	4.3	-0.1	1.2	1.1	-0.7	0.3	0.3	-0.6	0.1	0.2
β_{214}	-1.3	1.9	1.9	0.2	0.6	0.6	0.3	0.2	0.2	0.1	0.1	0.1
β_{215}	2.0	2.6	2.6	-0.1	0.8	0.8	0.5	0.2	0.2	0.0	0.1	0.1
β_{216}	2.2	1.3	1.3	1.0	0.9	0.9	0.0	0.2	0.2	0.0	0.1	0.1
β_{217}	3.2	2.6	2.6	0.1	0.6	0.5	-0.1	0.3	0.3	0.4	0.1	0.1
β_{218}	-0.5	2.4	2.3	-1.2	1.3	1.3	-0.6	0.3	0.3	-0.4	0.1	0.1
Γ_{21}	-4.1	4.9	5.0	2.1	2.2	2.3	2.3	1.3	1.3	2.1	0.5	0.6
Γ_{212}	-1.2	3.3	3.3	3.4	1.6	1.7	3.3	1.2	1.3	2.0	0.6	0.6
Γ_{22}	-4.0	2.8	2.9	-4.6	1.4	1.6	-1.7	0.8	0.8	-0.6	0.6	0.6
Γ_{213}	0.3	3.0	2.9	2.6	2.0	2.0	0.2	1.8	1.7	4.4	0.7	0.9
Γ_{223}	3.2	1.5	1.5	2.7	1.2	1.2	0.2	0.9	0.9	-0.4	0.7	0.7
Γ_{23}	2.0	0.3	0.3	4.0	0.6	0.7	4.7	0.4	0.6	6.0	0.4	0.7
Γ_{214}	-3.7	2.6	2.7	-1.7	1.0	1.0	-1.8	0.7	0.8	0.9	0.4	0.4
Γ_{224}	3.8	1.2	1.3	3.5	0.6	0.7	3.1	0.5	0.6	1.7	0.3	0.4
Γ_{234}	1.3	0.1	0.1	1.6	0.2	0.3	2.4	0.3	0.3	2.1	0.2	0.2
Γ_{24}	0.1	0.0	0.0	0.3	< 0.1	< 0.1	0.4	< 0.1	< 0.1	1.0	< 0.1	0.1
Γ_{215}	-0.7	1.0	1.0	0.4	0.7	0.6	-0.4	0.3	0.3	-0.1	0.2	0.2
Γ_{225}	0.0	0.6	0.6	0.5	0.5	0.5	0.6	0.3	0.3	0.7	0.3	0.3
Γ_{235}	-0.1	0.0	0.0	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.2	0.2
Γ_{245}	0.0	0.0	0.0	0.1	< 0.1	< 0.1	0.1	< 0.1	< 0.1	-0.1	< 0.1	< 0.1
Γ_{25}	0.0	0.0	0.0	0.1	< 0.1	< 0.1	0.0	< 0.1	< 0.1	0.2	< 0.1	< 0.1

Table H–33: Simulation 6 results multiplied by 100, averaged over 100 runs. Outcome follows a Poisson distribution with $K = 3$, $p = 19$, $q = 5$, $\pi_1 = 0.5$, $\pi_2 = 0.3$, $\pi_3 = 0.2$, $\boldsymbol{\beta}_1^\top = (0.65, 0.30, 0.15, 0.35, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00)$, $\boldsymbol{\beta}_2^\top = (0.20, -0.45, -0.10, 0.25, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00)$, $\boldsymbol{\beta}_3^\top = (1.00, 0.15, -0.65, -0.15, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00)$, $\Gamma_1^{*\top} = (0.30, -0.25, 0.10, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00)$, $\Gamma_2^{*\top} = (0.35, 0.20, 0.15, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00)$, and $\Gamma_3^{*\top} = (0.25, 0.00, 0.20, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00)$. Small values changed to zero, part 3.

n	$n = 100$			$n = 250$			$n = 500$			$n = 1000$		
Measure	Bias	Var	MSE	Bias	Var	MSE	Bias	Var	MSE	Bias	Var	MSE
β_{30}	-15.3	5.7	8.0	-8.5	3.1	3.8	-2.2	1.3	1.3	0.4	0.5	0.5
β_{31}	-0.1	4.5	4.5	-2.3	1.9	2.0	-0.9	0.8	0.8	-0.9	0.4	0.4
β_{32}	19.1	8.5	12.0	13.5	4.7	6.5	7.6	3.5	4.0	4.5	1.7	1.9
β_{33}	18.3	3.2	6.6	11.3	2.0	3.2	5.1	1.0	1.2	3.0	0.8	0.9
β_{34}	0.7	1.4	1.4	0.5	0.5	0.5	0.2	0.2	0.2	-0.1	0.1	0.1
β_{35}	1.2	2.1	2.1	-0.1	0.7	0.7	-0.4	0.2	0.2	-0.5	0.1	0.1
β_{36}	-4.4	1.9	2.1	-0.6	0.6	0.6	-0.3	0.3	0.3	0.0	0.1	0.1
β_{37}	0.1	1.5	1.5	0.9	0.5	0.5	-0.1	0.2	0.2	-0.1	0.1	0.1
β_{38}	-1.9	1.3	1.3	-0.9	0.7	0.7	-0.9	0.3	0.3	0.1	0.1	0.1
β_{39}	0.5	1.3	1.3	-0.5	0.4	0.4	0.0	0.2	0.2	-0.2	0.1	0.1
β_{310}	-0.1	1.6	1.6	-0.1	0.5	0.5	0.0	0.2	0.2	-0.2	0.1	0.1
β_{311}	1.2	2.3	2.3	-1.5	0.6	0.6	0.0	0.2	0.2	0.0	0.1	0.1
β_{312}	2.2	1.8	1.9	0.1	0.5	0.5	0.0	0.3	0.3	0.4	0.1	0.1
β_{313}	1.5	1.6	1.6	-0.3	0.7	0.7	0.4	0.2	0.2	0.2	0.1	0.1
β_{314}	1.7	1.5	1.5	0.2	0.5	0.5	0.2	0.2	0.2	0.3	0.1	0.1
β_{315}	2.0	1.7	1.7	-0.2	0.3	0.3	-1.2	0.2	0.2	-0.7	0.1	0.1
β_{316}	1.3	2.4	2.4	-0.6	0.5	0.5	0.7	0.2	0.2	-0.6	0.1	0.1
β_{317}	0.4	2.0	2.0	0.3	0.5	0.5	0.7	0.2	0.2	0.1	0.1	0.1
β_{318}	2.6	2.5	2.5	-0.2	0.5	0.5	0.7	0.3	0.3	-0.6	0.1	0.1
Γ_{31}	-10.1	3.0	3.9	-5.4	1.5	1.8	-4.5	0.6	0.8	-3.4	0.4	0.5
Γ_{312}	-1.7	1.9	1.9	0.2	1.8	1.8	-2.3	0.8	0.9	-0.9	0.4	0.4
Γ_{32}	-16.2	0.7	3.3	-13.2	0.9	2.6	-8.6	0.8	1.5	-4.1	0.4	0.6
Γ_{313}	-4.0	2.5	2.6	-2.1	2.1	2.2	0.9	1.3	1.3	0.3	0.4	0.4
Γ_{323}	2.4	1.0	1.1	1.3	0.9	0.9	-0.9	0.5	0.5	-0.9	0.3	0.3
Γ_{33}	3.6	0.5	0.7	6.2	0.9	1.3	6.5	1.0	1.4	6.0	0.6	0.9
Γ_{314}	-5.7	1.5	1.8	-2.8	1.1	1.2	-0.7	0.5	0.5	-0.5	0.3	0.3
Γ_{324}	1.6	0.7	0.7	0.4	0.6	0.6	0.5	0.4	0.4	0.3	0.2	0.2
Γ_{334}	2.2	0.4	0.4	3.5	0.5	0.6	3.9	0.4	0.5	3.6	0.4	0.5
Γ_{34}	0.1	0.0	0.0	0.1	< 0.1	< 0.1	0.3	< 0.1	< 0.1	0.9	< 0.1	< 0.1
Γ_{315}	-1.0	1.0	1.0	-1.0	0.5	0.5	0.3	0.3	0.3	0.0	0.1	0.1
Γ_{325}	0.4	0.3	0.3	-0.4	0.2	0.2	0.3	0.2	0.2	0.1	0.1	0.1
Γ_{335}	-0.1	0.1	0.1	-0.1	0.1	0.1	0.3	0.1	0.1	0.6	0.2	0.2
Γ_{345}	0.0	0.0	0.0	0.0	< 0.1	< 0.1	0.0	< 0.1	< 0.1	-0.2	0.1	0.1
Γ_{35}	0.0	0.0	0.0	0.0	< 0.1	< 0.1	0.0	< 0.1	< 0.1	0.4	< 0.1	< 0.1
Total			255.8			89.7			44.6			23.3

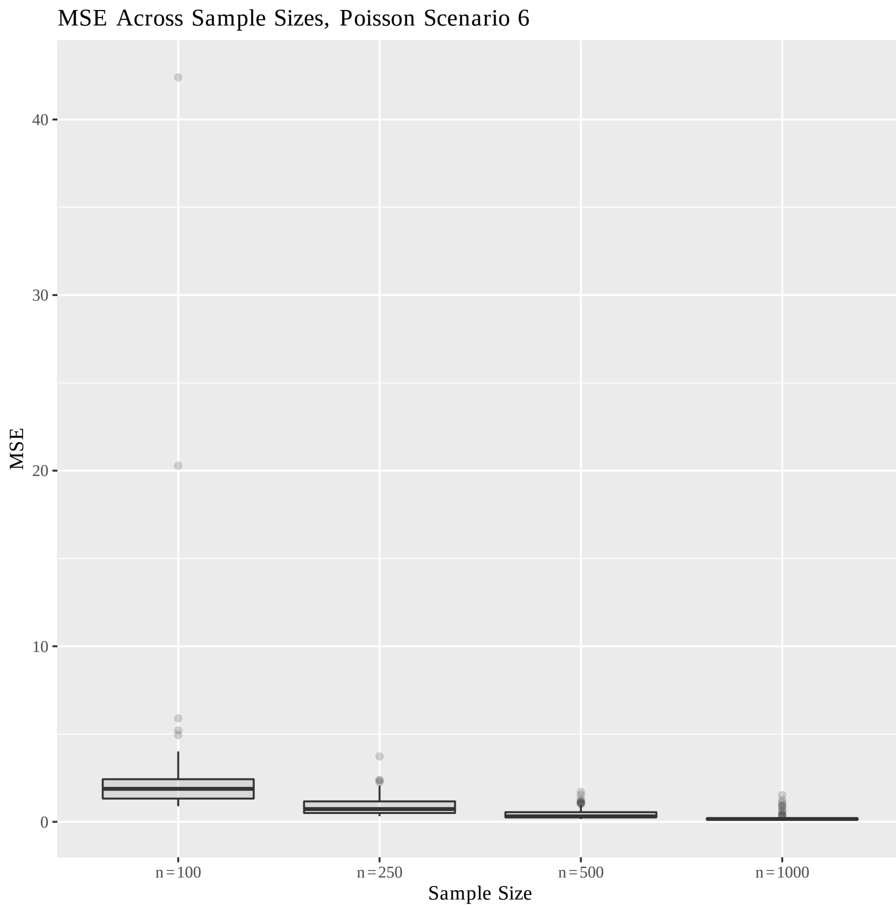

Figure H–17: MSE across sample sizes, Poisson outcome, small values changed to zero, with $K = 3$, $p = 19$, $q = 5$.

Table H–34: Simulation 6 results multiplied by 100, averaged over 100 runs. Outcome follows a Poisson distribution with $K = 3$, $p = 4$, $q = 2$, $\pi_1 = 0.5$, $\pi_2 = 0.3$, $\pi_3 = 0.2$, $\boldsymbol{\beta}_1^\top = (0.65, 0.30, 0.15, 0.35)$, $\boldsymbol{\beta}_2^\top = (0.20, -0.45, -0.10, 0.25)$, $\boldsymbol{\beta}_3^\top = (1.00, 0.15, -0.65, -0.15)$, $\mathbb{\Gamma}_1^{*\top} = (0.30, -0.25, 0.10)$, $\mathbb{\Gamma}_2^{*\top} = (0.35, 0.20, 0.15)$, and $\mathbb{\Gamma}_3^{*\top} = (0.25, 0.00, 0.20)$. Oracle model.

n	$n = 100$			$n = 250$			$n = 500$			$n = 1000$		
Measure	Bias	Var	MSE	Bias	Var	MSE	Bias	Var	MSE	Bias	Var	MSE
π_1	-4.7	1.0	1.3	-3.5	0.4	0.6	-2.6	0.2	0.3	-2.8	0.1	0.2
π_2	2.7	1.1	1.1	3.1	0.4	0.4	2.7	0.2	0.3	3.1	0.1	0.2
β_{10}	-2.1	4.8	4.8	-0.8	0.6	0.6	0.7	0.2	0.2	0.4	0.1	0.1
β_{11}	4.0	2.5	2.7	2.0	0.6	0.6	0.4	0.2	0.2	0.3	0.1	0.1
β_{12}	1.8	0.9	0.9	1.7	0.2	0.2	1.6	0.1	0.1	1.1	0.1	0.1
β_{13}	-2.5	0.6	0.7	-1.4	0.2	0.2	-1.1	0.1	0.1	-0.2	< 0.1	< 0.1
Γ_{11}	-5.3	1.3	1.6	-2.5	0.4	0.4	-0.9	0.2	0.2	-0.4	0.1	0.1
Γ_{112}	6.5	1.7	2.1	1.6	0.4	0.4	-0.2	0.1	0.1	-1.1	0.1	0.1
Γ_{12}	-6.7	0.3	0.8	-5.8	0.3	0.6	-4.7	0.3	0.5	-5.1	0.2	0.4
β_{20}	-2.8	8.9	8.9	3.1	1.7	1.8	1.6	0.9	0.9	3.8	0.3	0.4
β_{21}	-2.7	10.5	10.5	4.2	1.6	1.7	3.7	1.0	1.2	5.3	0.5	0.7
β_{22}	4.4	4.0	4.1	0.6	0.9	0.9	0.4	0.3	0.3	0.3	0.2	0.2
β_{23}	-2.3	1.7	1.8	-0.3	0.7	0.7	-0.5	0.3	0.3	-0.6	0.1	0.1
Γ_{21}	-0.4	5.9	5.9	-0.4	2.1	2.0	-0.2	0.9	0.9	0.6	0.3	0.3
Γ_{212}	3.9	3.4	3.5	0.4	1.2	1.2	1.9	0.7	0.7	3.1	0.4	0.5
Γ_{22}	-4.4	2.2	2.4	-2.6	1.6	1.6	-1.8	0.9	0.9	0.5	0.4	0.4
β_{30}	-0.7	2.2	2.2	-1.1	1.0	1.0	0.2	0.4	0.4	1.2	0.2	0.2
β_{31}	1.8	2.4	2.4	1.5	0.6	0.6	1.2	0.3	0.3	1.4	0.1	0.1
β_{32}	-2.0	2.2	2.2	-2.6	0.4	0.5	-1.5	0.2	0.2	-0.1	0.1	0.1
β_{33}	2.0	0.8	0.8	0.0	0.3	0.3	-0.1	0.1	0.1	-0.1	0.1	0.1
Γ_{31}	-3.0	1.8	1.9	-2.5	0.7	0.8	-1.9	0.3	0.3	-1.5	0.2	0.2
Γ_{312}	-0.6	1.8	1.7	-1.3	1.0	1.0	-1.4	0.5	0.5	-1.1	0.2	0.2
Γ_{32}	-11.4	1.2	2.5	-7.7	0.9	1.5	-3.0	0.4	0.5	-1.1	0.1	0.2
Total			66.8			19.5			9.5			5.1

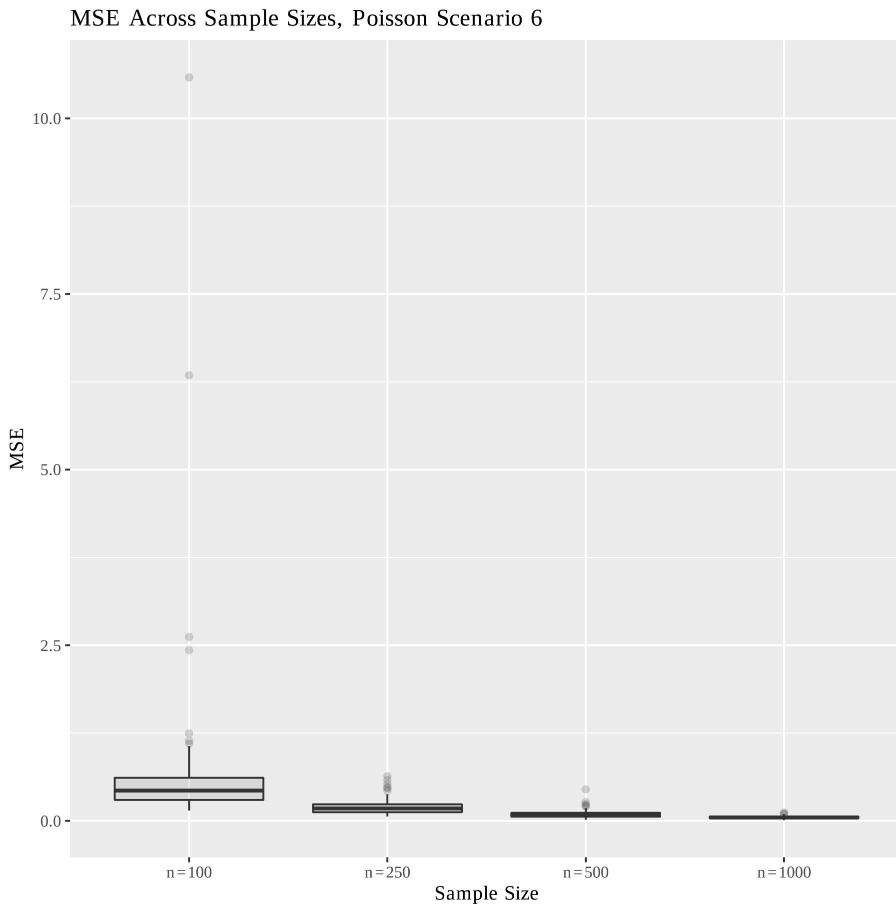

Figure H–18: MSE across sample sizes, Poisson outcome, oracle model, with $K = 3$, $p = 19$, $q = 5$.

Table H–35: Simulation 7 results multiplied by 100, averaged over 100 runs. Outcome follows a binomial distribution with $K = 2$, $p = 7$, $q = 2$, $\pi_1 = 0.6$, $\pi_2 = 0.4$, $\boldsymbol{\beta}_1^\top = (0.95, 0.60, 0.00, 0.00, 0.00, 0.00, 0.00)$, $\boldsymbol{\beta}_2^\top = (-0.85, -0.15, 0.00, 0.00, 0.00, 0.00, 0.00)$, $\mathbb{\Gamma}_1^{*\top} = (0.95, 0.85, 1.15)$, and $\mathbb{\Gamma}_2^{*\top} = (0.70, -0.70, 0.80)$. No penalty.

n	$n = 100$			$n = 250$			$n = 500$			$n = 1000$		
Measure	Bias	Var	MSE	Bias	Var	MSE	Bias	Var	MSE	Bias	Var	MSE
π_1	-11.3	1.0	2.3	-12.0	0.7	2.1	-12.1	0.4	1.8	-12.2	0.3	1.8
β_{10}	17.5	7.2	10.2	22.2	2.9	7.8	24.1	1.7	7.5	25.0	1.6	7.8
β_{11}	30.8	14.8	24.1	26.9	8.6	15.8	24.8	4.6	10.7	23.8	3.9	9.6
β_{12}	0.9	1.1	1.1	-0.3	0.4	0.4	0.5	0.2	0.2	0.1	0.1	0.1
β_{13}	0.2	1.1	1.1	-0.3	0.4	0.4	-0.2	0.1	0.1	-0.2	0.1	0.1
β_{14}	-0.3	1.0	1.0	0.4	0.4	0.4	0.1	0.1	0.1	0.0	< 0.1	< 0.1
β_{15}	1.0	0.7	0.7	1.0	0.3	0.3	0.8	0.1	0.1	0.4	0.1	0.1
β_{16}	0.3	0.8	0.8	0.1	0.3	0.3	0.0	0.1	0.1	-0.2	0.1	0.1
Γ_{11}	-5.8	6.8	7.1	-8.6	2.4	3.1	-7.5	1.2	1.8	-7.4	0.6	1.2
Γ_{112}	3.2	15.0	14.9	-0.2	5.2	5.1	-1.3	3.8	3.8	0.6	1.9	1.9
Γ_{12}	-11.4	7.2	8.4	-3.2	2.8	2.9	-1.4	1.2	1.2	-1.0	0.5	0.5
β_{20}	21.9	5.6	10.3	20.1	3.0	7.0	20.2	1.5	5.6	20.1	1.1	5.1
β_{21}	0.6	7.0	6.9	-0.5	2.3	2.3	-3.0	1.5	1.6	-3.0	0.8	0.9
β_{22}	-1.7	1.0	1.0	-0.9	0.3	0.3	-0.6	0.2	0.2	-0.5	0.1	0.1
β_{23}	-0.5	0.8	0.8	-0.7	0.3	0.3	-0.6	0.1	0.1	-0.3	0.1	0.1
β_{24}	0.1	0.7	0.7	-0.3	0.2	0.2	-0.1	0.1	0.1	0.0	< 0.1	< 0.1
β_{25}	-0.8	0.6	0.6	-0.2	0.3	0.3	-0.2	0.1	0.1	-0.5	< 0.1	< 0.1
β_{26}	-0.5	0.7	0.7	-0.5	0.2	0.2	-0.1	0.1	0.1	-0.2	< 0.1	< 0.1
Γ_{21}	6.2	3.8	4.1	7.7	1.5	2.1	8.5	0.8	1.5	8.4	0.5	1.2
Γ_{212}	19.4	6.5	10.2	20.0	3.4	7.4	19.5	1.7	5.5	21.6	1.0	5.7
Γ_{22}	1.5	5.7	5.7	7.6	2.2	2.8	11.4	1.2	2.5	13.2	0.6	2.3
Total			112.8			61.3			44.7			38.6

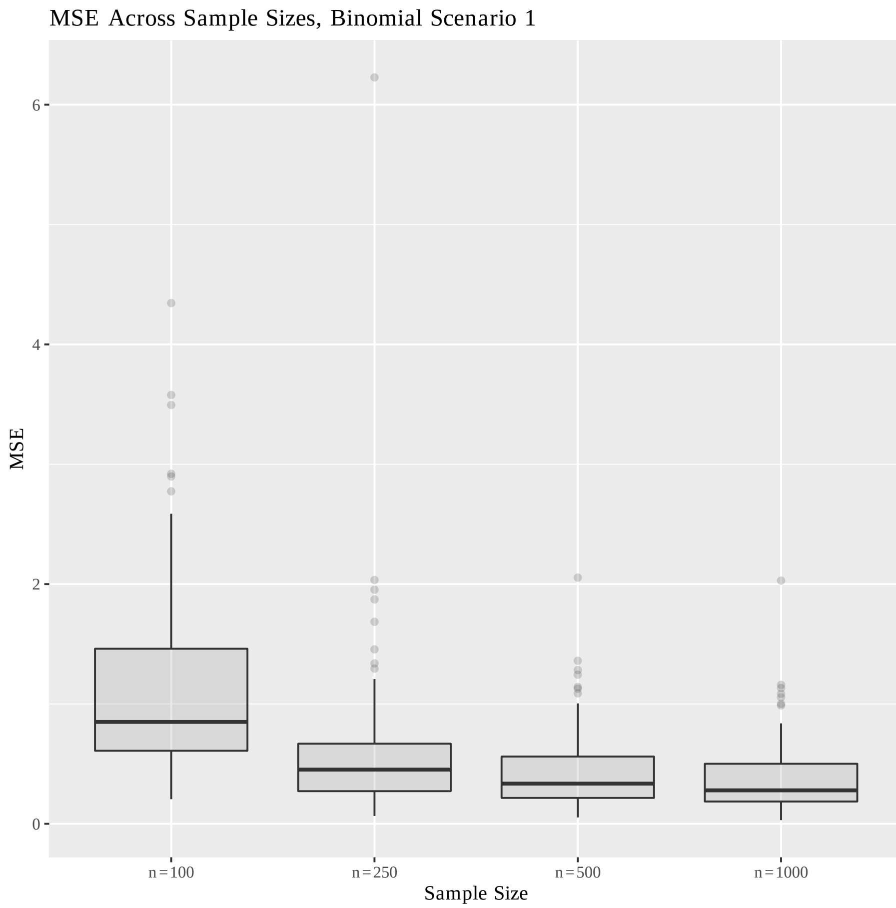

Figure H–19: MSE across sample sizes, binomial outcome, no penalty, with $K = 2$, $p = 7$, $q = 2$.

Table H–36: Simulation 7 results multiplied by 100, averaged over 100 runs. Outcome follows a binomial distribution with $K = 2$, $p = 7$, $q = 2$, $\pi_1 = 0.6$, $\pi_2 = 0.4$, $\boldsymbol{\beta}_1^\top = (0.95, 0.60, 0.00, 0.00, 0.00, 0.00, 0.00)$, $\boldsymbol{\beta}_2^\top = (-0.85, -0.15, 0.00, 0.00, 0.00, 0.00, 0.00)$, $\mathbb{\Gamma}_1^{*\top} = (0.95, 0.85, 1.15)$, and $\mathbb{\Gamma}_2^{*\top} = (0.70, -0.70, 0.80)$. Small values changed to zero.

n	$n = 100$			$n = 250$			$n = 500$			$n = 1000$		
Measure	Bias	Var	MSE	Bias	Var	MSE	Bias	Var	MSE	Bias	Var	MSE
π_1	-11.3	1.0	2.3	-12.0	0.7	2.1	-12.1	0.4	1.8	-12.2	0.3	1.8
β_{10}	17.5	7.2	10.2	22.2	2.9	7.8	24.1	1.7	7.5	25.0	1.6	7.8
β_{11}	30.8	14.8	24.1	26.9	8.6	15.8	24.8	4.6	10.7	23.8	3.9	9.6
β_{12}	0.9	1.1	1.1	-0.3	0.4	0.4	0.5	0.2	0.2	0.1	0.1	0.1
β_{13}	0.2	1.1	1.1	-0.4	0.4	0.4	-0.1	0.1	0.1	-0.3	0.1	0.1
β_{14}	-0.3	1.0	1.0	0.5	0.4	0.4	0.1	0.1	0.1	0.0	< 0.1	< 0.1
β_{15}	0.9	0.7	0.7	1.0	0.3	0.3	0.8	0.1	0.1	0.4	0.1	0.1
β_{16}	0.3	0.8	0.8	0.2	0.3	0.3	0.0	0.1	0.1	-0.2	0.1	0.1
Γ_{11}	-5.8	6.8	7.1	-8.6	2.4	3.1	-7.5	1.2	1.8	-7.4	0.6	1.2
Γ_{112}	3.2	15.0	14.9	-0.2	5.2	5.1	-1.3	3.8	3.8	0.6	1.9	1.9
Γ_{12}	-11.4	7.2	8.4	-3.2	2.8	2.9	-1.4	1.2	1.2	-1.0	0.5	0.5
β_{20}	21.9	5.6	10.3	20.1	3.0	7.0	20.2	1.5	5.6	20.1	1.1	5.1
β_{21}	0.5	7.0	6.9	-0.5	2.3	2.2	-3.0	1.5	1.6	-3.0	0.8	0.9
β_{22}	-1.7	1.0	1.0	-0.9	0.3	0.3	-0.6	0.2	0.2	-0.5	0.1	0.1
β_{23}	-0.5	0.8	0.8	-0.7	0.3	0.3	-0.6	0.1	0.1	-0.3	0.1	0.1
β_{24}	0.1	0.7	0.6	-0.3	0.2	0.2	0.0	0.1	0.1	0.0	< 0.1	< 0.1
β_{25}	-0.7	0.6	0.6	-0.2	0.3	0.3	-0.2	0.1	0.1	-0.5	< 0.1	< 0.1
β_{26}	-0.5	0.7	0.7	-0.4	0.2	0.2	-0.1	0.1	0.1	-0.1	< 0.1	< 0.1
Γ_{21}	6.2	3.8	4.1	7.7	1.5	2.1	8.5	0.8	1.5	8.4	0.5	1.2
Γ_{212}	19.4	6.5	10.2	20.0	3.4	7.4	19.5	1.7	5.5	21.6	1.0	5.7
Γ_{22}	1.5	5.7	5.7	7.6	2.2	2.8	11.4	1.2	2.5	13.2	0.6	2.3
Total			112.8			61.3			44.7			38.6

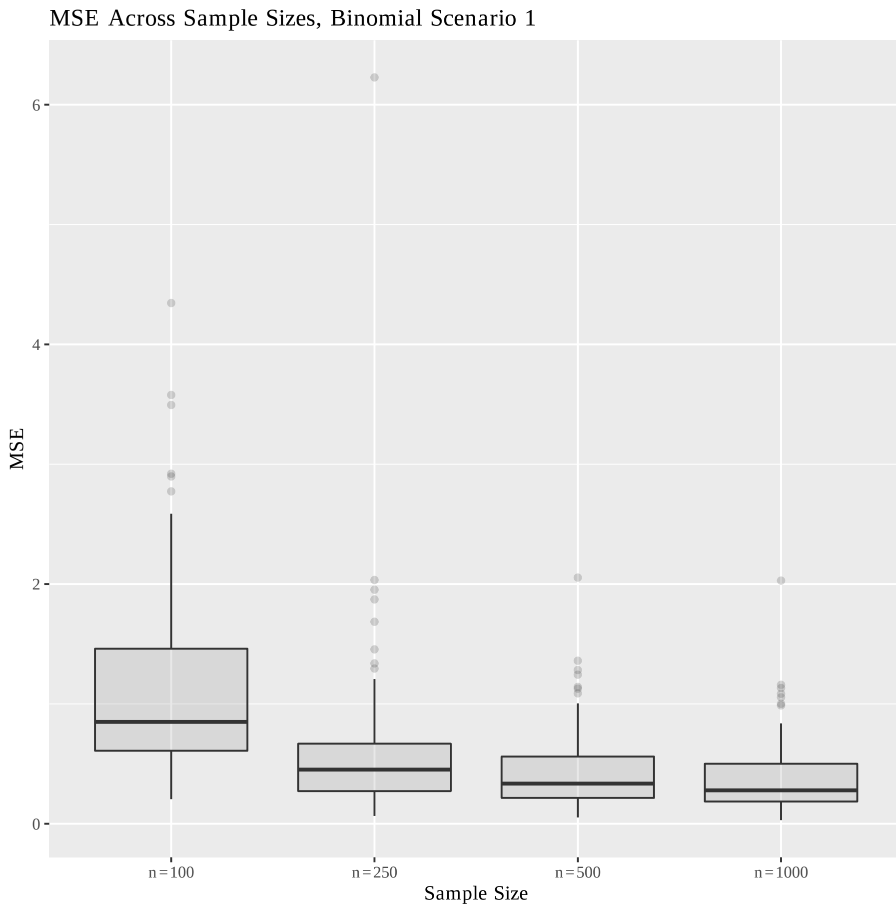

Figure H–20: MSE across sample sizes, binomial outcome, small values changed to zero, with $K = 2$, $p = 7$, $q = 2$.

Table H–37: Simulation 7 results multiplied by 100, averaged over 100 runs. Outcome follows a binomial distribution with $K = 2$, $p = 2$, $q = 2$, $\pi_1 = 0.6$, $\pi_2 = 0.4$, $\boldsymbol{\beta}_1^\top = (0.95, 0.60)$, $\boldsymbol{\beta}_2^\top = (-0.85, -0.15)$, $\mathbb{\Gamma}_1^{*\top} = (0.95, 0.85, 1.15)$, and $\mathbb{\Gamma}_2^{*\top} = (0.70, -0.70, 0.80)$. Oracle model.

n	$n = 100$			$n = 250$			$n = 500$			$n = 1000$		
Measure	Bias	Var	MSE	Bias	Var	MSE	Bias	Var	MSE	Bias	Var	MSE
π_1	-12.2	0.7	2.2	-12.0	0.5	1.9	-12.4	0.3	1.9	-12.6	0.3	1.9
β_{10}	22.3	5.0	10.0	22.5	1.6	6.7	25.5	1.6	8.0	26.7	1.5	8.6
β_{11}	32.6	11.1	21.7	26.6	5.8	12.9	26.0	4.7	11.4	25.5	4.1	10.6
Γ_{11}	-11.1	4.5	5.7	-9.0	1.8	2.6	-9.1	1.2	2.0	-9.3	0.6	1.5
Γ_{112}	-0.5	14.8	14.7	-3.3	5.1	5.1	-3.9	3.9	4.0	-2.4	2.1	2.1
Γ_{12}	-9.5	6.3	7.1	-1.8	2.0	2.0	-0.9	1.1	1.1	-0.8	0.5	0.5
β_{20}	20.2	4.3	8.3	19.4	2.4	6.2	19.8	1.4	5.3	19.8	1.1	5.0
β_{21}	-3.1	4.6	4.7	-2.1	1.7	1.8	-3.5	1.4	1.5	-3.8	0.8	0.9
Γ_{21}	5.1	2.7	2.9	7.0	1.3	1.8	7.9	0.7	1.4	7.9	0.5	1.1
Γ_{212}	19.0	5.6	9.1	19.2	2.6	6.3	20.2	1.4	5.4	22.0	0.9	5.7
Γ_{22}	4.7	5.1	5.3	8.9	2.0	2.8	12.2	1.1	2.6	13.8	0.6	2.5
Total			91.6			50.1			44.6			40.4

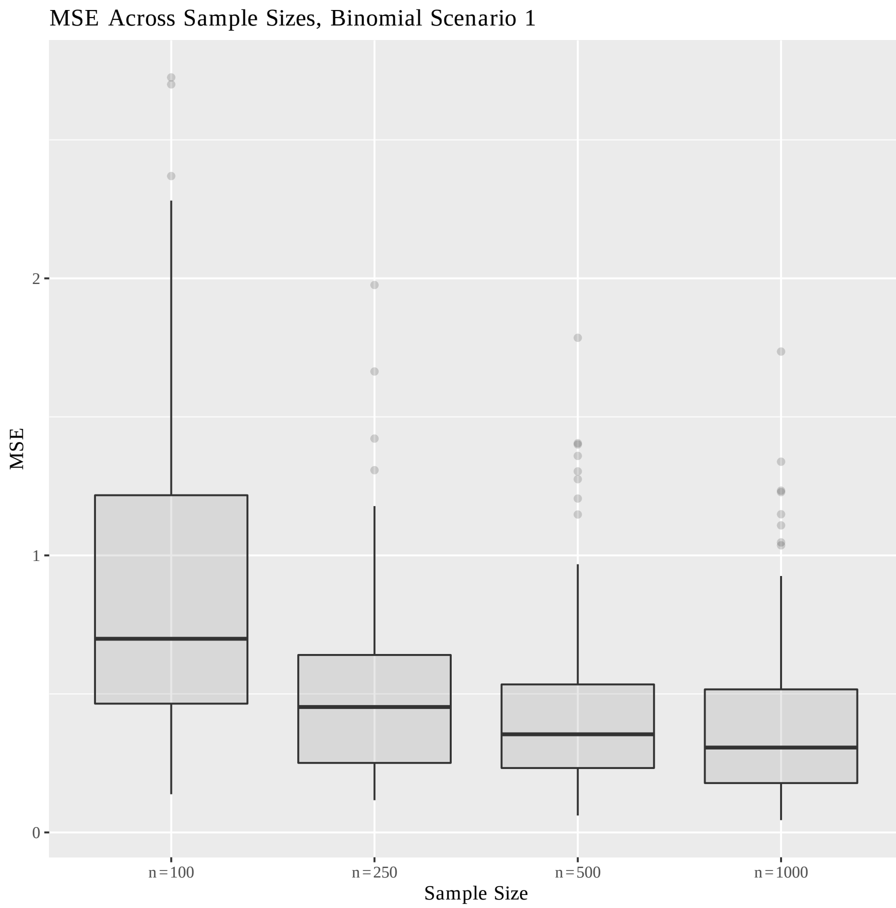

Figure H–21: MSE across sample sizes, binomial outcome, oracle model, with $K = 2$, $p = 7$, $q = 2$.

Table H–38: Simulation 8 results multiplied by 100, averaged over 100 runs. Outcome follows a binomial distribution with $K = 2$, $p = 17$, $q = 2$, $\pi_1 = 0.6$, $\pi_2 = 0.4$, $\boldsymbol{\beta}_1^\top = (0.95, 0.60, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00)$, $\boldsymbol{\beta}_2^\top = (-0.85, -0.15, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00)$, $\mathbb{\Gamma}_1^{*\top} = (0.95, 0.85, 1.15)$, and $\mathbb{\Gamma}_2^{*\top} = (0.70, -0.70, 0.80)$. No penalty, part 1.

n	$n = 100$			$n = 250$			$n = 500$			$n = 1000$		
Measure	Bias	Var	MSE	Bias	Var	MSE	Bias	Var	MSE	Bias	Var	MSE
π_1	-6.5	1.3	1.7	-7.9	0.6	1.3	-8.5	0.6	1.3	-8.1	0.5	1.1
β_{10}	12.3	8.3	9.8	14.7	3.4	5.5	16.9	3.3	6.1	15.8	2.5	5.0
β_{11}	23.5	14.9	20.2	27.6	6.7	14.2	27.0	5.0	12.2	27.8	3.9	11.6
β_{12}	1.9	1.0	1.0	0.4	0.4	0.4	0.0	0.2	0.2	-0.2	0.1	0.1
β_{13}	-0.3	0.8	0.8	-0.1	0.3	0.3	0.1	0.2	0.2	0.1	0.1	0.1
β_{14}	-0.8	0.8	0.8	0.1	0.3	0.3	-0.3	0.1	0.1	-0.1	0.1	0.1
β_{15}	0.8	0.8	0.7	-0.2	0.2	0.2	0.1	0.1	0.1	0.0	0.1	0.1
β_{16}	-0.9	0.5	0.5	-0.1	0.3	0.3	-0.7	0.1	0.1	-0.1	< 0.1	0.0
β_{17}	0.9	0.8	0.8	0.5	0.3	0.3	-0.2	0.1	0.1	0.1	< 0.1	0.0
β_{18}	-0.1	0.7	0.7	-0.3	0.3	0.3	-0.5	0.2	0.2	-0.1	0.1	0.1
β_{19}	0.4	0.8	0.8	-0.1	0.3	0.3	-0.3	0.1	0.1	-0.2	0.1	0.1
β_{110}	0.0	0.8	0.8	-0.4	0.3	0.3	-0.3	0.1	0.1	-0.1	0.1	0.1
β_{111}	-0.1	0.7	0.7	0.2	0.2	0.2	-0.2	0.1	0.1	0.0	< 0.1	0.0
Γ_{11}	-1.8	5.6	5.6	-2.6	3.0	3.0	-4.6	2.0	2.2	-5.1	1.3	1.6
Γ_{112}	-23.3	18.7	23.9	-16.6	6.6	9.3	-19.2	4.2	7.8	-19.7	2.8	6.6
Γ_{12}	-16.4	7.8	10.4	-12.1	1.7	3.1	-11.1	1.0	2.2	-9.6	0.5	1.4

Table H–39: Simulation 8 results multiplied by 100, averaged over 100 runs. Outcome follows a binomial distribution with $K = 2$, $p = 17$, $q = 2$, $\pi_1 = 0.6$, $\pi_2 = 0.4$, $\boldsymbol{\beta}_1^\top = (0.95, 0.60, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00)$, $\boldsymbol{\beta}_2^\top = (-0.85, -0.15, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00)$, $\mathbb{\Gamma}_1^{*\top} = (0.95, 0.85, 1.15)$, and $\mathbb{\Gamma}_2^{*\top} = (0.70, -0.70, 0.80)$. No penalty, part 2.

n	$n = 100$			$n = 250$			$n = 500$			$n = 1000$		
Measure	Bias	Var	MSE	Bias	Var	MSE	Bias	Var	MSE	Bias	Var	MSE
β_{20}	17.5	7.4	10.4	15.9	3.1	5.5	16.5	1.3	4.0	15.8	1.0	3.5
β_{21}	-11.8	8.1	9.4	-14.5	3.5	5.6	-15.6	1.1	3.6	-16.2	0.8	3.5
β_{22}	-1.1	1.6	1.6	-0.9	0.4	0.4	-0.4	0.2	0.2	-0.2	0.1	0.1
β_{23}	0.3	0.9	0.9	0.3	0.3	0.3	0.2	0.1	0.1	0.0	0.1	0.1
β_{24}	-0.7	0.9	0.8	0.0	0.3	0.3	-0.2	0.1	0.1	-0.2	0.1	0.1
β_{25}	-0.7	0.8	0.8	-0.3	0.2	0.2	-0.2	0.1	0.1	-0.3	< 0.1	0.0
β_{26}	-0.2	1.0	0.9	-0.3	0.3	0.3	0.0	0.2	0.2	-0.1	0.1	0.1
β_{27}	-1.4	0.8	0.8	0.1	0.3	0.3	0.1	0.1	0.1	0.0	0.1	0.1
β_{28}	-0.1	0.9	0.9	0.6	0.4	0.4	0.8	0.1	0.1	0.2	0.1	0.1
β_{29}	-0.5	0.9	0.9	0.1	0.3	0.3	0.4	0.1	0.1	0.2	0.1	0.1
β_{210}	-0.8	0.9	0.9	-0.2	0.2	0.2	0.0	0.1	0.1	-0.3	0.1	0.1
β_{211}	-0.4	0.7	0.7	-0.1	0.2	0.2	0.0	0.1	0.1	0.0	0.1	0.1
Γ_{21}	3.4	4.3	4.4	6.3	1.5	1.8	7.7	0.5	1.1	7.9	0.3	0.9
Γ_{212}	8.8	13.2	13.9	4.2	5.1	5.2	0.0	2.0	2.0	-0.4	1.3	1.3
Γ_{22}	-11.3	7.5	8.7	0.2	1.9	1.9	1.6	1.0	1.0	2.4	0.7	0.8
Total			135.2			62.2			46.1			38.6

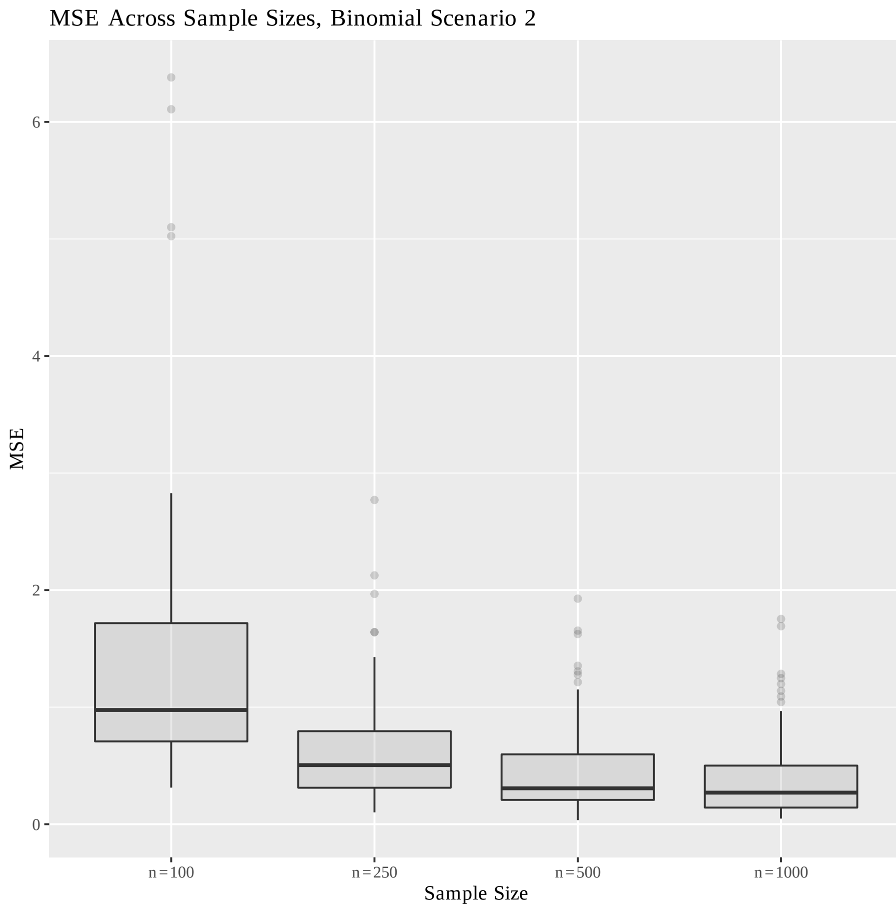

Figure H–22: MSE across sample sizes, binomial outcome, no penalty, with $K = 2$, $p = 17$, $q = 2$.

Table H–40: Simulation 8 results multiplied by 100, averaged over 100 runs. Outcome follows a binomial distribution with $K = 2$, $p = 17$, $q = 2$, $\pi_1 = 0.6$, $\pi_2 = 0.4$, $\boldsymbol{\beta}_1^\top = (0.95, 0.60, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00)$, $\boldsymbol{\beta}_2^\top = (-0.85, -0.15, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00)$, $\mathbb{\Gamma}_1^{*\top} = (0.95, 0.85, 1.15)$, and $\mathbb{\Gamma}_2^{*\top} = (0.70, -0.70, 0.80)$. Small values changed to zero, part 1.

n	$n = 100$			$n = 250$			$n = 500$			$n = 1000$		
Measure	Bias	Var	MSE	Bias	Var	MSE	Bias	Var	MSE	Bias	Var	MSE
π_1	-6.5	1.3	1.7	-7.9	0.6	1.3	-8.5	0.6	1.3	-8.1	0.5	1.1
β_{10}	12.3	8.3	9.8	14.7	3.4	5.5	16.9	3.3	6.1	15.8	2.5	5.0
β_{11}	23.5	14.9	20.2	27.6	6.7	14.2	27.0	5.0	12.2	27.8	3.9	11.6
β_{12}	1.9	1.0	1.0	0.5	0.4	0.4	0.0	0.2	0.2	-0.3	0.1	0.1
β_{13}	-0.3	0.8	0.8	-0.1	0.3	0.3	0.0	0.2	0.2	0.1	0.1	0.1
β_{14}	-0.8	0.8	0.8	0.1	0.3	0.3	-0.3	0.1	0.1	-0.2	0.1	0.1
β_{15}	0.8	0.7	0.7	-0.3	0.2	0.2	0.1	0.1	0.1	0.0	0.1	0.1
β_{16}	-0.9	0.5	0.5	-0.1	0.3	0.3	-0.7	0.1	0.1	-0.1	< 0.1	0.0
β_{17}	0.9	0.8	0.8	0.5	0.3	0.3	-0.2	0.1	0.1	0.1	< 0.1	0.0
β_{18}	-0.1	0.7	0.7	-0.3	0.3	0.3	-0.4	0.2	0.2	-0.1	0.1	0.1
β_{19}	0.4	0.8	0.8	-0.2	0.3	0.3	-0.3	0.1	0.1	-0.1	0.1	0.1
β_{110}	0.0	0.8	0.8	-0.5	0.3	0.3	-0.3	0.1	0.1	0.0	0.1	0.1
β_{111}	-0.1	0.7	0.7	0.2	0.2	0.2	-0.2	0.1	0.1	-0.1	< 0.1	0.0
Γ_{11}	-1.8	5.6	5.6	-2.6	3.0	3.0	-4.6	2.0	2.2	-5.1	1.3	1.6
Γ_{112}	-23.3	18.7	23.9	-16.6	6.6	9.3	-19.2	4.2	7.8	-19.7	2.8	6.6
Γ_{12}	-16.4	7.8	10.4	-12.1	1.7	3.1	-11.1	1.0	2.2	-9.6	0.5	1.4

Table H–41: Simulation 8 results multiplied by 100, averaged over 100 runs. Outcome follows a binomial distribution with $K = 2$, $p = 17$, $q = 2$, $\pi_1 = 0.6$, $\pi_2 = 0.4$, $\boldsymbol{\beta}_1^\top = (0.95, 0.60, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00)$, $\boldsymbol{\beta}_2^\top = (-0.85, -0.15, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00)$, $\mathbb{\Gamma}_1^{*\top} = (0.95, 0.85, 1.15)$, and $\mathbb{\Gamma}_2^{*\top} = (0.70, -0.70, 0.80)$. Small values changed to zero, part 2.

n	$n = 100$			$n = 250$			$n = 500$			$n = 1000$		
Measure	Bias	Var	MSE	Bias	Var	MSE	Bias	Var	MSE	Bias	Var	MSE
β_{20}	17.5	7.4	10.4	15.9	3.1	5.5	16.5	1.3	4.0	15.8	1.0	3.5
β_{21}	-11.8	8.1	9.4	-14.5	3.5	5.6	-15.6	1.1	3.6	-16.2	0.8	3.5
β_{22}	-1.1	1.6	1.6	-0.9	0.4	0.4	-0.4	0.2	0.2	-0.1	0.1	0.1
β_{23}	0.3	0.9	0.9	0.2	0.3	0.3	0.2	0.1	0.1	0.0	< 0.1	0.0
β_{24}	-0.7	0.9	0.8	0.0	0.3	0.3	-0.2	0.1	0.1	-0.2	0.1	0.1
β_{25}	-0.7	0.8	0.8	-0.3	0.2	0.2	-0.2	0.1	0.1	-0.4	< 0.1	0.0
β_{26}	-0.2	1.0	0.9	-0.2	0.3	0.3	0.0	0.2	0.2	-0.2	0.1	0.1
β_{27}	-1.4	0.8	0.8	0.2	0.3	0.3	0.1	0.1	0.1	0.0	0.1	0.1
β_{28}	0.0	0.9	0.9	0.6	0.4	0.4	0.8	0.1	0.1	0.2	0.1	0.1
β_{29}	-0.5	0.9	0.9	0.2	0.3	0.3	0.4	0.1	0.1	0.1	0.1	0.1
β_{210}	-0.7	0.9	0.9	-0.2	0.2	0.2	-0.1	0.1	0.1	-0.4	0.1	0.1
β_{211}	-0.4	0.7	0.7	-0.2	0.2	0.2	0.0	0.1	0.1	0.0	0.1	0.1
Γ_{21}	3.4	4.3	4.4	6.3	1.5	1.8	7.7	0.5	1.1	7.9	0.3	0.9
Γ_{212}	8.8	13.2	13.9	4.2	5.1	5.2	0.0	2.0	2.0	-0.4	1.3	1.3
Γ_{22}	-11.3	7.5	8.7	0.2	1.9	1.9	1.6	1.0	1.0	2.4	0.7	0.8
Total			135.2			62.2			46.1			38.5

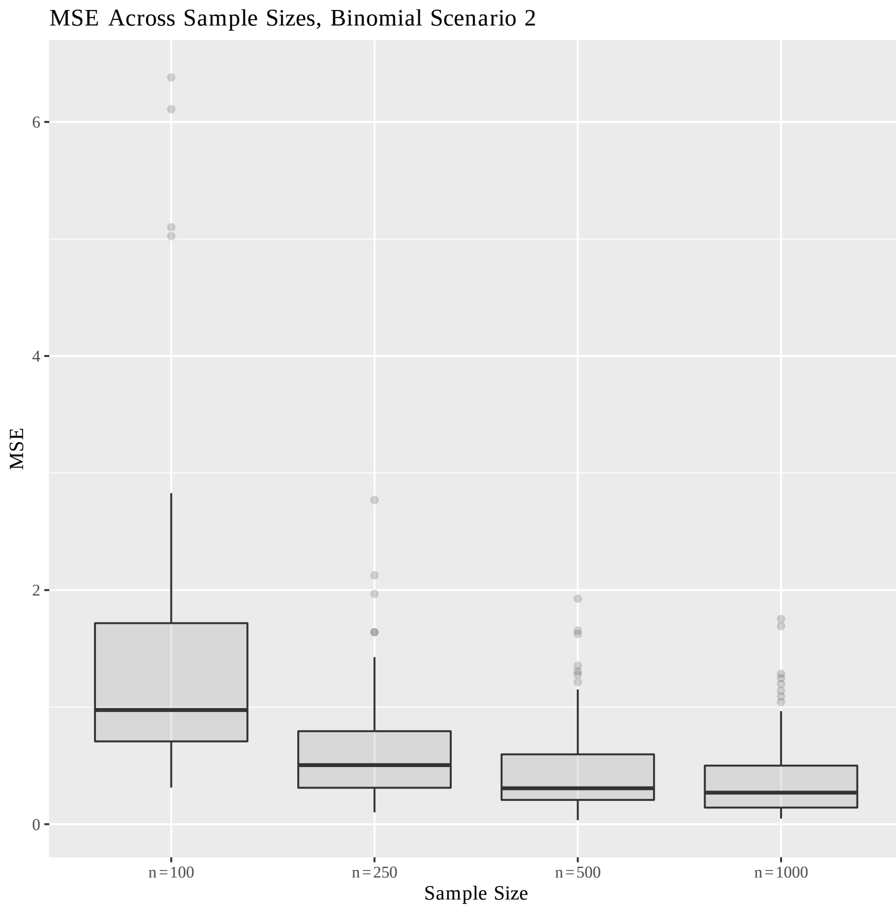

Figure H–23: MSE across sample sizes, binomial outcome, small values changed to zero, with $K = 2$, $p = 17$, $q = 2$.

Table H–42: Simulation 8 results multiplied by 100, averaged over 100 runs. Outcome follows a binomial distribution with $K = 2$, $p = 2$, $q = 2$, $\pi_1 = 0.6$, $\pi_2 = 0.4$, $\boldsymbol{\beta}_1^\top = (0.95, 0.60)$, $\boldsymbol{\beta}_2^\top = (-0.85, -0.15)$, $\mathbb{\Gamma}_1^{*\top} = (0.95, 0.85, 1.15)$, and $\mathbb{\Gamma}_2^{*\top} = (0.70, -0.70, 0.80)$. Oracle model.

n	$n = 100$			$n = 250$			$n = 500$			$n = 1000$		
Measure	Bias	Var	MSE	Bias	Var	MSE	Bias	Var	MSE	Bias	Var	MSE
π_1	-8.6	1.0	1.7	-9.1	0.6	1.4	-9.6	0.5	1.4	-9.4	0.4	1.3
β_{10}	19.2	5.8	9.4	18.6	3.4	6.8	20.4	2.9	7.0	19.5	2.4	6.2
β_{11}	27.1	11.7	19.0	30.7	5.9	15.3	29.5	4.5	13.1	30.9	3.6	13.1
Γ_{11}	-7.8	4.2	4.8	-6.1	2.4	2.7	-7.6	1.7	2.3	-8.1	1.3	1.9
Γ_{112}	-25.5	12.6	19.0	-21.7	5.8	10.4	-22.5	3.8	8.9	-23.7	2.9	8.6
Γ_{12}	-11.5	4.8	6.1	-10.8	1.8	2.9	-10.5	1.0	2.1	-9.4	0.5	1.4
β_{20}	15.9	4.4	6.9	15.7	2.3	4.7	16.5	1.3	4.0	16.5	1.0	3.7
β_{21}	-12.6	5.4	6.9	-16.3	2.8	5.4	-16.0	1.2	3.7	-16.7	0.7	3.4
Γ_{21}	3.6	2.4	2.5	6.2	1.1	1.4	7.1	0.5	1.0	7.6	0.2	0.8
Γ_{212}	8.9	11.2	11.8	4.3	3.5	3.7	1.9	2.0	2.1	1.7	1.3	1.3
Γ_{22}	-6.5	5.9	6.3	1.7	1.5	1.5	3.0	1.0	1.0	3.9	0.6	0.8
Total			94.3			56.3			46.7			42.5

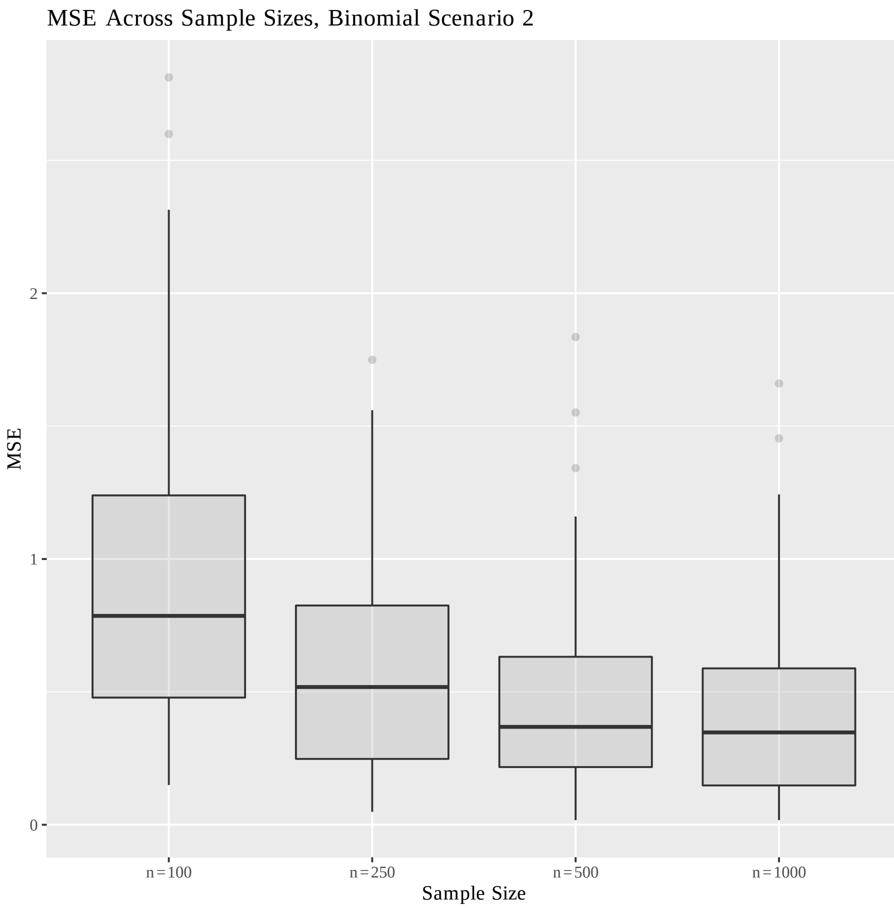

Figure H–24: MSE across sample sizes, binomial outcome, oracle model, with $K = 2$, $p = 17$, $q = 2$.

Table H–43: Simulation 9 results multiplied by 100, averaged over 100 runs. Outcome follows a binomial distribution with $K = 2$, $p = 19$, $q = 5$, $\pi_1 = 0.6$, $\pi_2 = 0.4$, $\boldsymbol{\beta}_1^\top = (0.95, 0.60, -0.65, -0.25, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00)$, $\boldsymbol{\beta}_2^\top = (-0.85, -0.15, -0.75, 0.10, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00)$, $\mathbb{\Gamma}_1^{*\top} = (0.95, 0.85, 1.15, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00)$, and $\mathbb{\Gamma}_2^{*\top} = (0.70, -0.70, 0.80, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00)$. No penalty, part 1.

n	$n = 100$			$n = 250$			$n = 500$			$n = 1000$		
Measure	Bias	Var	MSE	Bias	Var	MSE	Bias	Var	MSE	Bias	Var	MSE
π_1	-1.2	0.8	0.8	1.4	0.4	0.4	1.8	0.2	0.3	1.2	0.1	0.2
β_{10}	-3.9	5.4	5.5	-2.7	2.3	2.3	-1.5	1.3	1.3	-0.9	0.7	0.7
β_{11}	5.8	9.1	9.4	-2.2	3.4	3.5	-1.3	2.2	2.2	-0.6	1.2	1.2
β_{12}	-0.9	1.0	1.0	-0.4	0.4	0.4	-0.6	0.2	0.2	0.2	0.1	0.1
β_{13}	3.3	0.7	0.8	3.3	0.3	0.4	1.8	0.1	0.1	1.3	0.1	0.1
β_{14}	-0.6	0.4	0.4	0.3	0.2	0.2	-0.2	0.1	0.1	0.2	< 0.1	< 0.1
β_{15}	-0.3	0.6	0.6	-0.3	0.2	0.2	0.1	0.1	0.1	0.3	0.1	0.1
β_{16}	0.9	0.7	0.7	0.4	0.2	0.2	0.2	0.1	0.1	0.3	< 0.1	0.1
β_{17}	1.6	0.5	0.5	-0.1	0.3	0.3	0.3	0.1	0.1	0.1	0.1	0.1
β_{18}	-1.1	0.7	0.7	-0.8	0.2	0.2	-0.4	0.1	0.1	-0.3	0.1	0.1
β_{19}	0.3	0.7	0.7	-0.2	0.2	0.2	0.0	0.1	0.1	0.0	0.1	0.1
β_{110}	-0.5	0.6	0.6	-0.7	0.2	0.2	-0.7	0.1	0.1	-0.2	< 0.1	< 0.1
β_{111}	0.2	0.7	0.7	0.4	0.2	0.2	0.2	0.1	0.1	0.2	0.1	0.1
β_{112}	1.8	1.0	1.0	0.7	0.2	0.2	0.1	0.1	0.1	-0.1	0.1	0.1
β_{113}	1.1	0.7	0.7	0.8	0.2	0.2	0.8	0.1	0.1	0.5	< 0.1	< 0.1
Γ_{11}	1.1	5.3	5.3	2.5	1.1	1.1	2.1	0.6	0.6	1.0	0.3	0.3
Γ_{112}	-3.7	10.7	10.7	-0.9	3.9	3.9	-1.0	1.4	1.4	-0.3	0.7	0.7
Γ_{12}	-13.7	5.3	7.1	-3.0	1.3	1.4	-1.1	0.7	0.7	0.8	0.4	0.4
Γ_{113}	0.2	1.5	1.5	1.3	0.6	0.6	0.7	0.3	0.3	0.1	0.2	0.2
Γ_{123}	3.7	1.4	1.5	1.6	0.5	0.5	0.8	0.2	0.2	0.3	0.1	0.1
Γ_{13}	3.1	0.3	0.4	5.9	0.4	0.8	6.7	0.3	0.7	6.2	0.2	0.6
Γ_{114}	-3.6	1.6	1.7	-4.8	0.5	0.7	-4.2	0.2	0.4	-2.9	0.1	0.2
Γ_{124}	-2.2	1.1	1.1	-2.7	0.4	0.4	-1.6	0.2	0.2	-1.1	0.1	0.1
Γ_{134}	0.2	0.2	0.2	-0.4	0.3	0.3	-0.2	0.3	0.3	0.4	0.2	0.2
Γ_{14}	0.8	0.1	0.1	2.5	0.2	0.2	3.4	0.2	0.3	3.8	0.1	0.3
Γ_{115}	1.6	0.9	0.9	-0.1	0.4	0.4	0.1	0.2	0.2	-0.3	0.1	0.1
Γ_{125}	0.3	0.8	0.8	-0.8	0.3	0.3	-0.4	0.2	0.2	-0.2	0.1	0.1
Γ_{135}	-0.3	0.2	0.2	0.0	0.2	0.2	-0.2	0.2	0.2	-0.4	0.2	0.2
Γ_{145}	0.1	0.1	0.1	0.1	0.1	0.1	0.2	0.1	0.1	0.4	0.1	0.1
Γ_{15}	0.3	0.0	0.0	0.3	< 0.1	< 0.1	0.8	< 0.1	< 0.1	1.6	0.1	0.1

Table H–44: Simulation 9 results multiplied by 100, averaged over 100 runs. Outcome follows a binomial distribution with $K = 2$, $p = 19$, $q = 5$, $\pi_1 = 0.6$, $\pi_2 = 0.4$, $\boldsymbol{\beta}_1^\top = (0.95, 0.60, -0.65, -0.25, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00)$, $\boldsymbol{\beta}_2^\top = (-0.85, -0.15, -0.75, 0.10, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00)$, $\mathbb{\Gamma}_1^{*\top} = (0.95, 0.85, 1.15, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00)$, and $\mathbb{\Gamma}_2^{*\top} = (0.70, -0.70, 0.80, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00)$. No penalty, part 2.

n	$n = 100$			$n = 250$			$n = 500$			$n = 1000$		
Measure	Bias	Var	MSE	Bias	Var	MSE	Bias	Var	MSE	Bias	Var	MSE
β_{20}	9.7	8.1	8.9	3.4	2.6	2.7	2.6	1.1	1.1	1.7	0.5	0.6
β_{21}	-10.3	12.0	12.9	0.6	2.9	2.8	0.3	1.5	1.5	0.1	1.0	1.0
β_{22}	-0.5	1.8	1.8	2.8	0.5	0.6	1.6	0.3	0.3	1.6	0.1	0.2
β_{23}	-6.4	1.5	1.9	-3.4	0.6	0.8	-2.0	0.2	0.3	-1.4	0.1	0.1
β_{24}	1.2	0.9	0.9	0.6	0.5	0.5	0.6	0.3	0.3	-0.2	0.1	0.1
β_{25}	-0.6	1.3	1.3	0.7	0.5	0.5	0.6	0.2	0.2	0.4	0.1	0.1
β_{26}	-1.7	1.1	1.1	-1.3	0.4	0.4	-1.3	0.2	0.2	-0.5	0.1	0.1
β_{27}	0.7	1.2	1.2	0.3	0.4	0.4	0.0	0.2	0.2	-0.1	0.1	0.1
β_{28}	0.4	1.2	1.2	-0.1	0.4	0.4	-1.1	0.2	0.2	-0.4	0.1	0.1
β_{29}	0.4	1.0	1.0	0.6	0.5	0.5	0.2	0.2	0.2	0.0	0.1	0.1
β_{210}	-0.4	0.9	0.9	0.9	0.3	0.3	0.3	0.2	0.2	-0.1	0.1	0.1
β_{211}	-1.3	1.7	1.7	-0.2	0.4	0.4	-0.1	0.2	0.1	0.4	0.1	0.1
β_{212}	-0.9	1.1	1.1	-0.4	0.3	0.3	0.0	0.2	0.1	0.0	0.1	0.1
β_{213}	-0.7	1.2	1.2	-0.4	0.4	0.4	-0.1	0.2	0.2	-0.2	0.1	0.1
Γ_{21}	0.5	5.7	5.6	0.1	2.1	2.1	-0.6	0.9	0.9	-0.9	0.4	0.4
Γ_{212}	0.3	11.8	11.7	-2.7	3.9	4.0	-5.7	2.3	2.6	-3.4	0.8	0.9
Γ_{22}	-12.5	11.3	12.7	-9.6	4.8	5.7	-8.1	2.3	2.9	-4.6	0.8	1.0
Γ_{213}	0.8	2.9	2.9	2.4	0.9	0.9	2.6	0.4	0.4	1.7	0.2	0.2
Γ_{223}	2.8	1.9	1.9	0.4	0.9	0.9	0.9	0.4	0.4	0.5	0.2	0.2
Γ_{23}	2.9	0.3	0.4	5.7	0.5	0.8	6.8	0.4	0.9	6.9	0.3	0.8
Γ_{214}	-6.8	1.6	2.0	-4.4	0.8	1.0	-2.6	0.4	0.5	-2.1	0.2	0.2
Γ_{224}	-0.5	1.9	1.9	1.1	0.7	0.7	0.1	0.3	0.3	-0.5	0.2	0.2
Γ_{234}	-0.1	0.2	0.2	-2.4	0.4	0.5	-2.0	0.5	0.5	-2.0	0.3	0.3
Γ_{24}	1.4	0.2	0.2	2.1	0.2	0.2	3.0	0.2	0.3	4.5	0.3	0.5
Γ_{215}	0.1	1.4	1.4	-0.8	0.5	0.5	-0.2	0.3	0.3	-0.5	0.1	0.1
Γ_{225}	0.0	1.5	1.5	-0.4	0.6	0.6	0.2	0.3	0.3	-0.4	0.1	0.1
Γ_{235}	-0.2	0.3	0.3	-0.3	0.2	0.2	-0.1	0.2	0.2	0.3	0.2	0.2
Γ_{245}	-0.3	0.1	0.1	0.3	0.1	0.1	0.2	0.1	0.1	0.0	0.1	0.1
Γ_{25}	0.1	0.0	0.0	0.4	< 0.1	< 0.1	0.5	< 0.1	< 0.1	1.0	< 0.1	< 0.1
Total			135.7			49.6			26.5			14.2

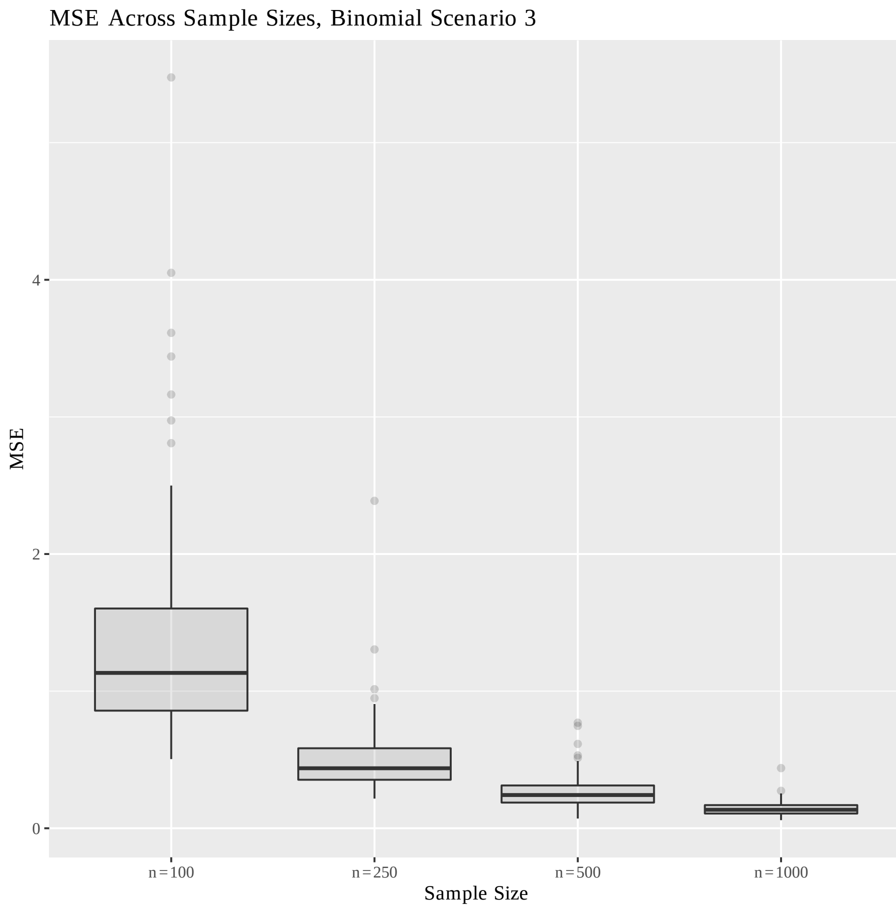

Figure H–25: MSE across sample sizes, binomial outcome, no penalty, with $K = 2$, $p = 19$, $q = 5$.

Table H–45: Simulation 9 results multiplied by 100, averaged over 100 runs. Outcome follows a binomial distribution with $K = 2$, $p = 19$, $q = 5$, $\pi_1 = 0.6$, $\pi_2 = 0.4$, $\boldsymbol{\beta}_1^\top = (0.95, 0.60, -0.65, -0.25, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00)$, $\boldsymbol{\beta}_2^\top = (-0.85, -0.15, -0.75, 0.10, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00)$, $\mathbb{\Gamma}_1^{*\top} = (0.95, 0.85, 1.15, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00)$, and $\mathbb{\Gamma}_2^{*\top} = (0.70, -0.70, 0.80, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00)$. Small values changed to zero, part 1.

n	$n = 100$			$n = 250$			$n = 500$			$n = 1000$		
Measure	Bias	Var	MSE	Bias	Var	MSE	Bias	Var	MSE	Bias	Var	MSE
π_1	-1.2	0.8	0.8	1.4	0.4	0.4	1.8	0.2	0.3	1.2	0.1	0.2
β_{10}	-3.9	5.4	5.5	-2.7	2.3	2.3	-1.5	1.3	1.3	-0.9	0.7	0.7
β_{11}	5.8	9.1	9.3	-2.2	3.4	3.5	-1.3	2.2	2.2	-0.6	1.2	1.2
β_{12}	-0.9	1.0	1.0	-0.4	0.4	0.4	-0.6	0.2	0.2	0.2	0.1	0.1
β_{13}	3.3	0.7	0.8	3.3	0.3	0.4	1.8	0.1	0.1	1.3	0.1	0.1
β_{14}	-0.6	0.4	0.4	0.3	0.2	0.2	-0.2	0.1	0.1	0.1	< 0.1	< 0.1
β_{15}	-0.2	0.6	0.6	-0.2	0.2	0.2	0.1	0.1	0.1	0.3	< 0.1	< 0.1
β_{16}	0.8	0.7	0.7	0.4	0.2	0.2	0.2	0.1	0.1	0.3	< 0.1	< 0.1
β_{17}	1.6	0.5	0.5	-0.1	0.3	0.2	0.3	0.1	0.1	0.2	0.1	0.1
β_{18}	-1.1	0.7	0.7	-0.8	0.2	0.2	-0.4	0.1	0.1	-0.3	0.1	0.1
β_{19}	0.4	0.7	0.7	-0.2	0.2	0.2	0.0	0.1	0.1	0.1	0.1	0.1
β_{110}	-0.5	0.6	0.6	-0.7	0.2	0.2	-0.7	0.1	0.1	-0.2	< 0.1	< 0.1
β_{111}	0.2	0.7	0.7	0.4	0.2	0.2	0.2	0.1	0.1	0.2	0.1	0.1
β_{112}	1.8	1.0	1.0	0.7	0.2	0.2	0.0	0.1	0.1	-0.1	0.1	0.1
β_{113}	1.1	0.7	0.7	0.8	0.2	0.2	0.8	0.1	0.1	0.5	< 0.1	< 0.1
Γ_{11}	1.1	5.3	5.3	2.5	1.1	1.1	2.1	0.6	0.6	1.0	0.3	0.3
Γ_{112}	-3.7	10.7	10.7	-0.9	3.9	3.9	-1.0	1.4	1.4	-0.3	0.7	0.7
Γ_{12}	-13.7	5.3	7.1	-3.0	1.3	1.4	-1.1	0.7	0.7	0.8	0.4	0.4
Γ_{113}	0.2	1.5	1.5	1.3	0.6	0.6	0.7	0.3	0.3	0.1	0.2	0.2
Γ_{123}	3.7	1.4	1.5	1.6	0.5	0.5	0.8	0.2	0.2	0.3	0.1	0.1
Γ_{13}	3.1	0.3	0.4	5.9	0.4	0.8	6.7	0.3	0.7	6.2	0.2	0.6
Γ_{114}	-3.6	1.6	1.7	-4.8	0.5	0.7	-4.2	0.2	0.4	-2.9	0.1	0.2
Γ_{124}	-2.2	1.1	1.1	-2.7	0.4	0.4	-1.6	0.2	0.2	-1.1	0.1	0.1
Γ_{134}	0.2	0.2	0.2	-0.4	0.3	0.3	-0.2	0.3	0.3	0.4	0.2	0.2
Γ_{14}	0.8	0.1	0.1	2.5	0.2	0.2	3.4	0.2	0.3	3.8	0.1	0.3
Γ_{115}	1.6	0.9	0.9	-0.1	0.4	0.4	0.1	0.2	0.2	-0.3	0.1	0.1
Γ_{125}	0.3	0.8	0.8	-0.8	0.3	0.3	-0.4	0.2	0.2	-0.2	0.1	0.1
Γ_{135}	-0.3	0.2	0.2	0.0	0.2	0.2	-0.2	0.2	0.2	-0.4	0.2	0.2
Γ_{145}	0.1	0.1	0.1	0.1	0.1	0.1	0.2	0.1	0.1	0.4	0.1	0.1
Γ_{15}	0.3	0.0	0.0	0.3	< 0.1	< 0.1	0.8	< 0.1	< 0.1	1.6	0.1	0.1

Table H–46: Simulation 9 results multiplied by 100, averaged over 100 runs. Outcome follows a binomial distribution with $K = 2$, $p = 19$, $q = 5$, $\pi_1 = 0.6$, $\pi_2 = 0.4$, $\boldsymbol{\beta}_1^\top = (0.95, 0.60, -0.65, -0.25, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00)$, $\boldsymbol{\beta}_2^\top = (-0.85, -0.15, -0.75, 0.10, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00)$, $\mathbb{\Gamma}_1^{*\top} = (0.95, 0.85, 1.15, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00)$, and $\mathbb{\Gamma}_2^{*\top} = (0.70, -0.70, 0.80, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00)$. Small values changed to zero, part 2.

n	$n = 100$			$n = 250$			$n = 500$			$n = 1000$		
Measure	Bias	Var	MSE	Bias	Var	MSE	Bias	Var	MSE	Bias	Var	MSE
β_{20}	9.7	8.1	8.9	3.4	2.6	2.7	2.6	1.1	1.1	1.7	0.5	0.6
β_{21}	-10.3	12.0	12.9	0.6	2.9	2.8	0.3	1.5	1.5	0.1	1.0	1.0
β_{22}	-0.5	1.8	1.8	2.8	0.5	0.6	1.6	0.3	0.3	1.6	0.1	0.2
β_{23}	-6.3	1.5	1.9	-3.4	0.6	0.8	-2.0	0.2	0.3	-1.4	0.1	0.1
β_{24}	1.2	0.9	0.9	0.6	0.5	0.5	0.6	0.3	0.3	-0.2	0.1	0.1
β_{25}	-0.6	1.3	1.3	0.7	0.5	0.5	0.6	0.2	0.2	0.4	0.1	0.1
β_{26}	-1.7	1.1	1.1	-1.2	0.4	0.4	-1.3	0.2	0.2	-0.5	0.1	0.1
β_{27}	0.8	1.2	1.2	0.3	0.4	0.4	0.0	0.2	0.2	-0.1	0.1	0.1
β_{28}	0.4	1.2	1.2	-0.1	0.4	0.4	-1.0	0.2	0.2	-0.3	0.1	0.1
β_{29}	0.4	1.0	1.0	0.6	0.5	0.5	0.2	0.2	0.2	0.0	0.1	0.1
β_{210}	-0.4	0.9	0.9	0.9	0.3	0.3	0.3	0.2	0.2	-0.2	0.1	0.1
β_{211}	-1.2	1.7	1.7	-0.2	0.4	0.4	-0.1	0.2	0.1	0.4	0.1	0.1
β_{212}	-0.9	1.1	1.1	-0.3	0.3	0.3	0.0	0.2	0.1	0.0	0.1	0.1
β_{213}	-0.7	1.2	1.1	-0.5	0.4	0.4	0.0	0.2	0.2	-0.2	0.1	0.1
Γ_{21}	0.5	5.7	5.6	0.1	2.1	2.1	-0.6	0.9	0.9	-0.9	0.4	0.4
Γ_{212}	0.3	11.8	11.7	-2.7	3.9	4.0	-5.7	2.3	2.6	-3.4	0.8	0.9
Γ_{22}	-12.5	11.3	12.7	-9.6	4.8	5.7	-8.1	2.3	2.9	-4.6	0.8	1.0
Γ_{213}	0.8	2.9	2.9	2.4	0.9	0.9	2.6	0.4	0.4	1.7	0.2	0.2
Γ_{223}	2.8	1.9	1.9	0.4	0.9	0.9	0.9	0.4	0.4	0.5	0.2	0.2
Γ_{23}	2.9	0.3	0.4	5.7	0.5	0.8	6.8	0.4	0.9	6.9	0.3	0.8
Γ_{214}	-6.8	1.6	2.0	-4.4	0.8	1.0	-2.6	0.4	0.5	-2.1	0.2	0.2
Γ_{224}	-0.5	1.9	1.9	1.1	0.7	0.7	0.1	0.3	0.3	-0.5	0.2	0.2
Γ_{234}	-0.1	0.2	0.2	-2.4	0.4	0.5	-2.0	0.5	0.5	-2.0	0.3	0.3
Γ_{24}	1.4	0.2	0.2	2.1	0.2	0.2	3.0	0.2	0.3	4.5	0.3	0.5
Γ_{215}	0.1	1.4	1.4	-0.8	0.5	0.5	-0.2	0.3	0.3	-0.5	0.1	0.1
Γ_{225}	0.0	1.5	1.5	-0.4	0.6	0.6	0.2	0.3	0.3	-0.4	0.1	0.1
Γ_{235}	-0.2	0.3	0.3	-0.3	0.2	0.2	-0.1	0.2	0.2	0.3	0.2	0.2
Γ_{245}	-0.3	0.1	0.1	0.3	0.1	0.1	0.2	0.1	0.1	0.0	0.1	0.1
Γ_{25}	0.1	0.0	0.0	0.4	< 0.1	< 0.1	0.5	< 0.1	< 0.1	1.0	< 0.1	< 0.1
Total			135.7			49.5			26.5			14.1

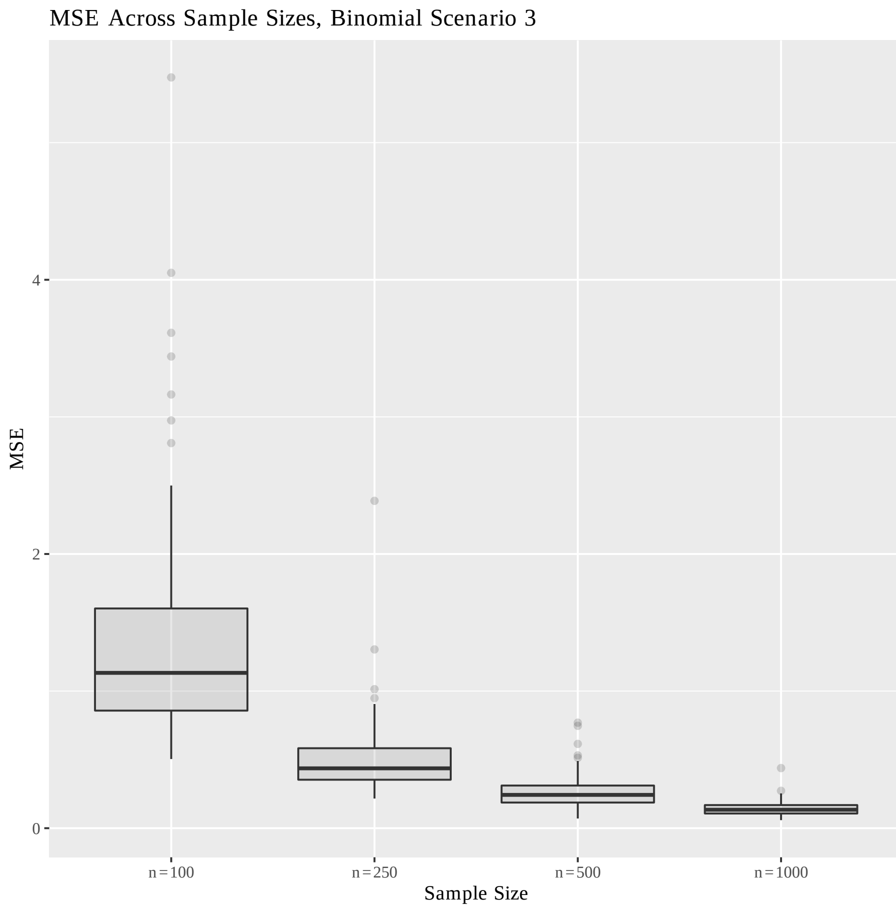

Figure H–26: MSE across sample sizes, binomial outcome, small values changed to zero, with $K = 2$, $p = 19$, $q = 5$.

Table H–47: Simulation 9 results multiplied by 100, averaged over 100 runs. Outcome follows a binomial distribution with $K = 2$, $p = 4$, $q = 2$, $\pi_1 = 0.6$, $\pi_2 = 0.4$, $\boldsymbol{\beta}_1^\top = (0.95, 0.60, -0.65, -0.25)$, $\boldsymbol{\beta}_2^\top = (-0.85, -0.15, -0.75, 0.10)$, $\mathbb{\Gamma}_1^{*\top} = (0.95, 0.85, 1.15)$, and $\mathbb{\Gamma}_2^{*\top} = (0.70, -0.70, 0.80)$. Oracle model.

n	$n = 100$			$n = 250$			$n = 500$			$n = 1000$		
Measure	Bias	Var	MSE	Bias	Var	MSE	Bias	Var	MSE	Bias	Var	MSE
π_1	-3.1	0.8	0.9	-3.1	0.4	0.5	-2.8	0.2	0.3	-3.2	0.1	0.2
β_{10}	1.7	5.0	5.0	6.5	1.9	2.3	6.9	1.0	1.5	7.2	0.6	1.1
β_{11}	8.6	9.2	9.9	8.7	3.9	4.6	9.3	2.5	3.4	10.2	1.3	2.3
β_{12}	0.8	0.8	0.8	1.6	0.4	0.4	0.7	0.1	0.2	1.1	0.1	0.1
β_{13}	-0.4	0.6	0.5	-0.2	0.2	0.2	-1.0	0.1	0.1	-0.9	0.1	0.1
Γ_{11}	-4.1	3.8	3.9	-4.3	1.0	1.2	-3.6	0.6	0.7	-4.3	0.3	0.4
Γ_{112}	-10.8	11.0	12.0	-10.6	4.7	5.8	-9.4	1.9	2.8	-8.5	0.9	1.6
Γ_{12}	-11.2	5.5	6.7	-5.9	1.1	1.5	-4.0	0.7	0.8	-2.9	0.4	0.5
β_{20}	8.6	4.9	5.6	8.2	1.4	2.0	8.8	0.8	1.6	7.6	0.4	0.9
β_{21}	-7.1	7.7	8.1	-6.5	2.9	3.3	-7.2	1.5	2.0	-7.8	0.8	1.4
β_{22}	0.8	1.3	1.3	2.4	0.5	0.5	1.6	0.2	0.2	1.9	0.1	0.1
β_{23}	-1.0	1.3	1.3	-1.2	0.4	0.4	-0.8	0.2	0.2	-1.0	0.1	0.1
Γ_{21}	-0.7	2.7	2.6	0.3	1.1	1.1	0.3	0.6	0.6	0.5	0.3	0.3
Γ_{212}	-1.3	8.9	8.8	-1.1	3.8	3.7	-2.5	1.7	1.8	-1.7	0.7	0.8
Γ_{22}	-7.3	9.9	10.3	0.5	3.8	3.7	1.8	1.5	1.5	4.4	0.6	0.8
Total			77.7			31.3			17.7			10.7

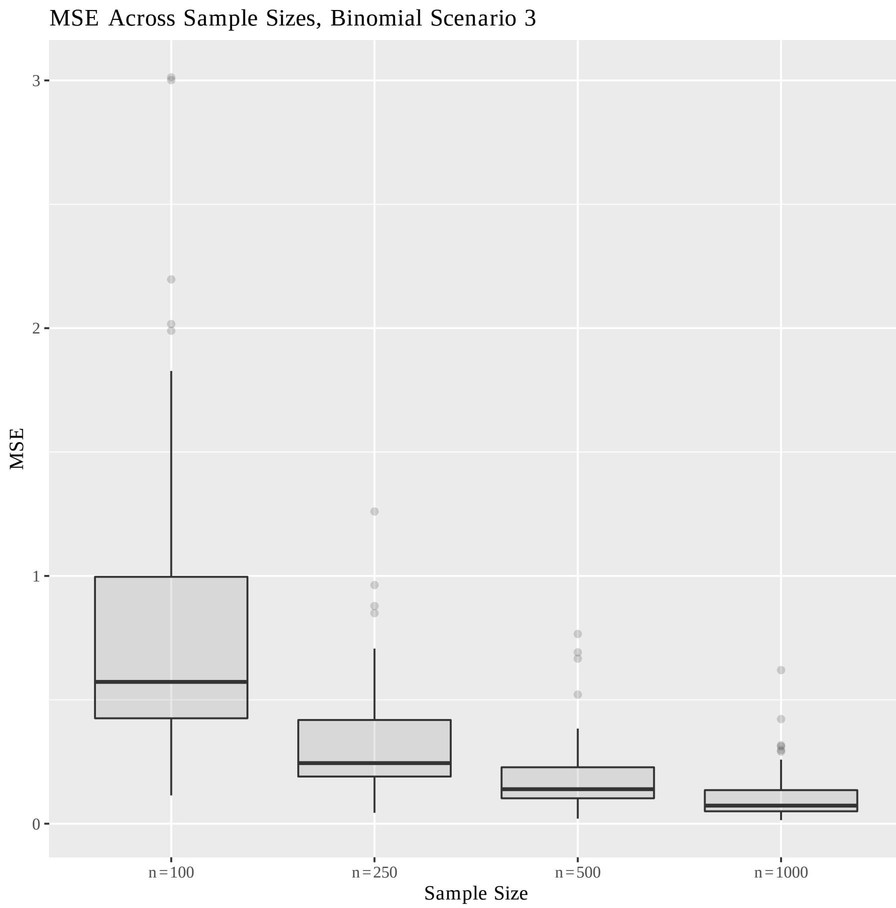

Figure H–27: MSE across sample sizes, binomial outcome, oracle model, with $K = 2$, $p = 19$, $q = 5$.

Table H–48: Simulation 10 results multiplied by 100, averaged over 100 runs. Outcome follows a binomial distribution with $K = 3$, $p = 7$, $q = 2$, $\pi_1 = 0.5$, $\pi_2 = 0.3$, $\pi_3 = 0.2$, $\boldsymbol{\beta}_1^\top = (0.95, 0.60, 0.00, 0.00, 0.00, 0.00, 0.00)$, $\boldsymbol{\beta}_2^\top = (-0.85, -0.15, 0.00, 0.00, 0.00, 0.00, 0.00)$, $\boldsymbol{\beta}_3^\top = (-0.30, -0.90, 0.00, 0.00, 0.00, 0.00, 0.00)$, $\mathbb{\Gamma}_1^{*\top} = (0.95, 0.85, 1.15)$, $\mathbb{\Gamma}_2^{*\top} = (0.70, -0.70, 0.80)$, and $\mathbb{\Gamma}_3^{*\top} = (1.75, 0.00, 0.85)$. No penalty, part 1.

n	$n = 100$			$n = 250$			$n = 500$			$n = 1000$		
Measure	Bias	Var	MSE	Bias	Var	MSE	Bias	Var	MSE	Bias	Var	MSE
π_1	-21.7	0.9	5.6	-24.2	0.4	6.2	-24.8	0.3	6.5	-23.7	0.2	5.8
π_2	5.2	1.7	1.9	7.6	1.2	1.8	10.3	0.9	1.9	8.1	0.7	1.3
β_{10}	37.5	15.5	29.4	38.0	4.6	19.0	43.6	3.8	22.7	41.1	2.5	19.4
β_{11}	49.1	32.8	56.6	59.4	12.4	47.7	64.3	10.2	51.4	60.6	6.2	42.9
β_{12}	1.8	3.9	3.9	1.9	1.1	1.2	-0.3	0.6	0.6	0.3	0.2	0.2
β_{13}	0.0	3.5	3.5	0.7	0.9	0.9	-0.6	0.5	0.5	-0.2	0.2	0.2
β_{14}	-2.7	2.9	2.9	-1.2	0.8	0.8	-0.1	0.4	0.4	0.8	0.2	0.2
β_{15}	2.3	2.4	2.4	1.0	1.0	1.0	1.0	0.4	0.4	0.4	0.2	0.2
β_{16}	1.3	2.3	2.3	0.4	0.8	0.8	-1.0	0.4	0.4	-0.4	0.2	0.2
Γ_{11}	-23.3	15.6	20.9	-20.3	6.8	10.9	-21.5	4.6	9.2	-20.7	2.7	6.9
Γ_{112}	-13.5	38.4	39.9	-11.0	33.7	34.5	-5.4	19.7	19.8	-7.1	7.3	7.7
Γ_{12}	-39.0	21.9	37.0	-25.9	11.4	18.0	-20.9	4.8	9.2	-15.9	1.3	3.8

Table H–49: Simulation 10 results multiplied by 100, averaged over 100 runs. Outcome follows a binomial distribution with $K = 3$, $p = 7$, $q = 2$, $\pi_1 = 0.5$, $\pi_2 = 0.3$, $\pi_3 = 0.2$, $\boldsymbol{\beta}_1^\top = (0.95, 0.60, 0.00, 0.00, 0.00, 0.00, 0.00)$, $\boldsymbol{\beta}_2^\top = (-0.85, -0.15, 0.00, 0.00, 0.00, 0.00, 0.00)$, $\boldsymbol{\beta}_3^\top = (-0.30, -0.90, 0.00, 0.00, 0.00, 0.00, 0.00)$, $\mathbb{\Gamma}_1^{*\top} = (0.95, 0.85, 1.15)$, $\mathbb{\Gamma}_2^{*\top} = (0.70, -0.70, 0.80)$, and $\mathbb{\Gamma}_3^{*\top} = (1.75, 0.00, 0.85)$. No penalty, part 2.

n	$n = 100$			$n = 250$			$n = 500$			$n = 1000$		
Measure	Bias	Var	MSE	Bias	Var	MSE	Bias	Var	MSE	Bias	Var	MSE
β_{20}	26.9	15.0	22.1	31.1	8.0	17.6	31.0	3.9	13.5	30.1	2.3	11.3
β_{21}	2.2	28.3	28.1	-2.0	9.1	9.1	-14.6	5.7	7.8	-15.8	4.6	7.0
β_{22}	0.0	3.0	2.9	0.0	1.0	1.0	-0.4	0.3	0.3	-0.2	0.2	0.2
β_{23}	1.2	2.8	2.8	-0.9	1.0	1.0	-1.8	0.3	0.3	-0.9	0.2	0.2
β_{24}	-2.4	2.8	2.8	0.6	0.8	0.8	1.2	0.4	0.4	0.2	0.2	0.2
β_{25}	2.7	2.6	2.6	0.4	1.0	1.0	-0.6	0.4	0.4	-0.1	0.2	0.2
β_{26}	-0.5	1.9	1.9	-0.6	1.0	1.0	0.4	0.3	0.3	0.7	0.2	0.2
Γ_{21}	19.0	13.3	16.8	24.9	8.2	14.4	22.6	5.4	10.5	21.3	4.3	8.8
Γ_{212}	32.9	28.6	39.2	26.8	15.4	22.5	30.6	8.0	17.3	28.0	6.0	13.8
Γ_{22}	-20.2	14.1	18.0	-3.7	6.4	6.5	8.6	2.2	2.9	11.3	1.4	2.7

Table H–50: Simulation 10 results multiplied by 100, averaged over 100 runs. Outcome follows a binomial distribution with $K = 3$, $p = 7$, $q = 2$, $\pi_1 = 0.5$, $\pi_2 = 0.3$, $\pi_3 = 0.2$, $\boldsymbol{\beta}_1^\top = (0.95, 0.60, 0.00, 0.00, 0.00, 0.00, 0.00)$, $\boldsymbol{\beta}_2^\top = (-0.85, -0.15, 0.00, 0.00, 0.00, 0.00, 0.00)$, $\boldsymbol{\beta}_3^\top = (-0.30, -0.90, 0.00, 0.00, 0.00, 0.00, 0.00)$, $\mathbb{\Gamma}_1^{*\top} = (0.95, 0.85, 1.15)$, $\mathbb{\Gamma}_2^{*\top} = (0.70, -0.70, 0.80)$, and $\mathbb{\Gamma}_3^{*\top} = (1.75, 0.00, 0.85)$. No penalty, part 3.

n	$n = 100$			$n = 250$			$n = 500$			$n = 1000$		
Measure	Bias	Var	MSE	Bias	Var	MSE	Bias	Var	MSE	Bias	Var	MSE
β_{30}	22.0	21.1	25.7	28.2	8.5	16.3	33.3	6.8	17.8	26.1	3.5	10.3
β_{31}	41.1	33.2	49.8	49.6	21.0	45.4	61.9	10.9	49.0	59.3	7.2	42.3
β_{32}	0.8	3.4	3.4	0.5	1.5	1.5	-0.1	0.6	0.6	-0.6	0.4	0.4
β_{33}	-2.0	3.3	3.3	0.5	1.6	1.6	2.4	0.9	0.9	0.7	0.4	0.4
β_{34}	4.0	4.1	4.3	1.8	1.9	2.0	-0.8	1.2	1.1	-0.2	0.5	0.5
β_{35}	-4.1	2.8	2.9	-1.9	1.0	1.0	-0.3	0.9	0.9	-0.7	0.5	0.5
β_{36}	0.7	2.2	2.2	-0.2	1.4	1.4	-0.7	0.8	0.8	-0.9	0.4	0.5
Γ_{31}	-34.0	24.4	35.7	-28.7	20.9	28.9	-25.6	10.0	16.4	-28.1	4.9	12.8
Γ_{312}	19.0	37.8	41.1	25.4	14.7	21.0	20.0	9.1	13.1	14.3	4.1	6.1
Γ_{32}	-0.2	23.0	22.8	24.0	18.0	23.6	30.3	6.9	16.0	29.4	3.0	11.6
Total			534.7			360.3			293.3			218.9

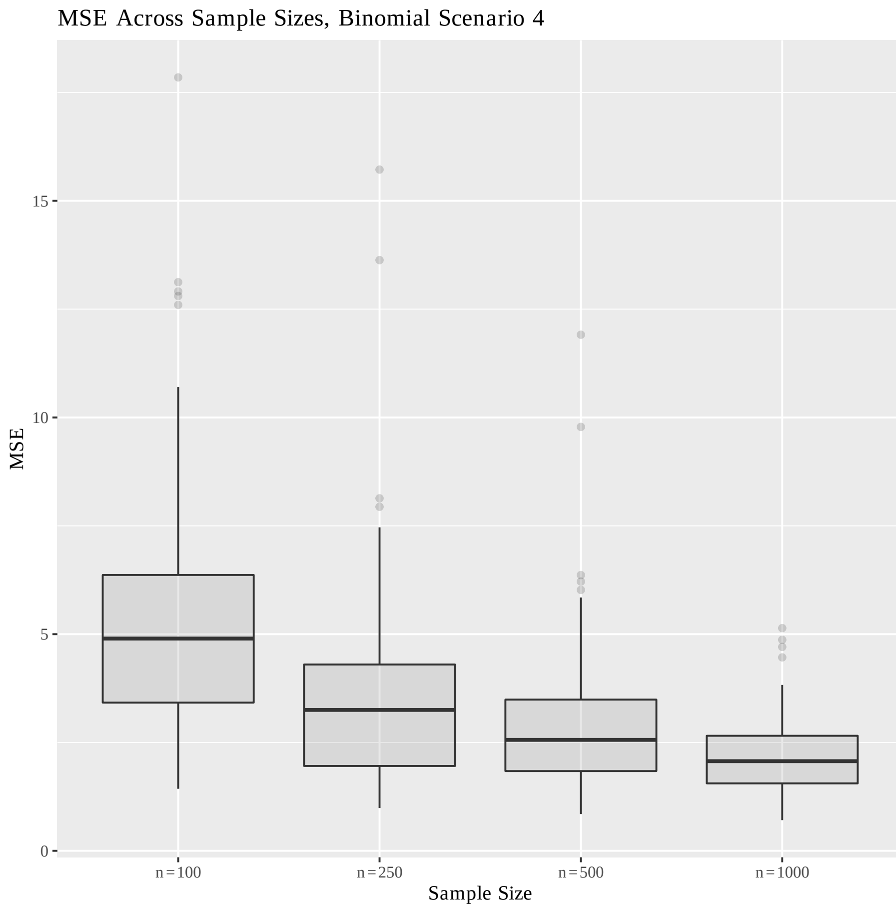

Figure H–28: MSE across sample sizes, binomial outcome, no penalty, with $K = 3$, $p = 7$, $q = 2$.

Table H–51: Simulation 10 results multiplied by 100, averaged over 100 runs. Outcome follows a binomial distribution with $K = 3$, $p = 7$, $q = 2$, $\pi_1 = 0.5$, $\pi_2 = 0.3$, $\pi_3 = 0.2$, $\boldsymbol{\beta}_1^\top = (0.95, 0.60, 0.00, 0.00, 0.00, 0.00, 0.00)$, $\boldsymbol{\beta}_2^\top = (-0.85, -0.15, 0.00, 0.00, 0.00, 0.00, 0.00)$, $\boldsymbol{\beta}_3^\top = (-0.30, -0.90, 0.00, 0.00, 0.00, 0.00, 0.00)$, $\mathbb{\Gamma}_1^{*\top} = (0.95, 0.85, 1.15)$, $\mathbb{\Gamma}_2^{*\top} = (0.70, -0.70, 0.80)$, and $\mathbb{\Gamma}_3^{*\top} = (1.75, 0.00, 0.85)$. Small values changed to zero, part 1.

n	$n = 100$			$n = 250$			$n = 500$			$n = 1000$		
Measure	Bias	Var	MSE	Bias	Var	MSE	Bias	Var	MSE	Bias	Var	MSE
π_1	-21.7	0.9	5.6	-24.2	0.4	6.2	-24.8	0.3	6.5	-23.7	0.2	5.8
π_2	5.2	1.7	1.9	7.6	1.2	1.8	10.3	0.9	1.9	8.1	0.7	1.3
β_{10}	37.5	15.5	29.4	38.0	4.6	19.0	43.6	3.8	22.7	41.1	2.5	19.4
β_{11}	49.1	32.8	56.6	59.4	12.4	47.7	64.3	10.2	51.4	60.6	6.2	42.9
β_{12}	1.8	3.9	3.9	1.9	1.1	1.2	-0.2	0.6	0.6	0.4	0.2	0.2
β_{13}	0.0	3.5	3.5	0.7	0.9	0.9	-0.6	0.5	0.5	-0.3	0.2	0.2
β_{14}	-2.7	2.9	2.9	-1.2	0.8	0.8	-0.1	0.4	0.4	0.8	0.2	0.2
β_{15}	2.3	2.4	2.4	1.0	1.0	1.0	1.0	0.4	0.4	0.4	0.2	0.2
β_{16}	1.3	2.3	2.3	0.5	0.8	0.8	-1.0	0.4	0.4	-0.5	0.2	0.2
Γ_{11}	-23.3	15.6	20.9	-20.3	6.8	10.9	-21.5	4.6	9.2	-20.7	2.7	6.9
Γ_{112}	-13.5	38.4	39.9	-11.0	33.7	34.5	-5.4	19.7	19.8	-7.1	7.3	7.7
Γ_{12}	-39.0	21.9	37.0	-25.9	11.4	18.0	-20.9	4.8	9.2	-15.9	1.3	3.8

Table H–52: Simulation 10 results multiplied by 100, averaged over 100 runs. Outcome follows a binomial distribution with $K = 3$, $p = 7$, $q = 2$, $\pi_1 = 0.5$, $\pi_2 = 0.3$, $\pi_3 = 0.2$, $\boldsymbol{\beta}_1^\top = (0.95, 0.60, 0.00, 0.00, 0.00, 0.00, 0.00)$, $\boldsymbol{\beta}_2^\top = (-0.85, -0.15, 0.00, 0.00, 0.00, 0.00, 0.00)$, $\boldsymbol{\beta}_3^\top = (-0.30, -0.90, 0.00, 0.00, 0.00, 0.00, 0.00)$, $\mathbb{F}_1^{*\top} = (0.95, 0.85, 1.15)$, $\mathbb{F}_2^{*\top} = (0.70, -0.70, 0.80)$, and $\mathbb{F}_3^{*\top} = (1.75, 0.00, 0.85)$. Small values changed to zero, part 2.

n	$n = 100$			$n = 250$			$n = 500$			$n = 1000$		
Measure	Bias	Var	MSE	Bias	Var	MSE	Bias	Var	MSE	Bias	Var	MSE
β_{20}	26.9	15.0	22.1	31.1	8.0	17.6	31.0	3.9	13.5	30.1	2.3	11.3
β_{21}	2.2	28.3	28.1	-2.0	9.1	9.1	-14.6	5.7	7.8	-15.8	4.6	7.0
β_{22}	0.0	3.0	2.9	-0.1	1.0	1.0	-0.4	0.3	0.3	-0.2	0.2	0.2
β_{23}	1.2	2.8	2.8	-0.9	1.0	1.0	-1.7	0.3	0.3	-0.9	0.2	0.2
β_{24}	-2.4	2.8	2.8	0.6	0.8	0.8	1.2	0.4	0.4	0.3	0.2	0.2
β_{25}	2.7	2.6	2.6	0.3	1.0	1.0	-0.6	0.4	0.4	-0.1	0.2	0.2
β_{26}	-0.6	1.9	1.9	-0.6	1.0	1.0	0.4	0.3	0.3	0.6	0.2	0.2
Γ_{21}	19.0	13.3	16.8	24.9	8.2	14.4	22.6	5.4	10.5	21.3	4.3	8.8
Γ_{212}	32.9	28.6	39.2	26.8	15.4	22.5	30.6	8.0	17.3	28.0	6.0	13.8
Γ_{22}	-20.2	14.1	18.0	-3.7	6.4	6.5	8.6	2.2	2.9	11.3	1.4	2.7

Table H–53: Simulation 10 results multiplied by 100, averaged over 100 runs. Outcome follows a binomial distribution with $K = 3$, $p = 7$, $q = 2$, $\pi_1 = 0.5$, $\pi_2 = 0.3$, $\pi_3 = 0.2$, $\boldsymbol{\beta}_1^\top = (0.95, 0.60, 0.00, 0.00, 0.00, 0.00, 0.00)$, $\boldsymbol{\beta}_2^\top = (-0.85, -0.15, 0.00, 0.00, 0.00, 0.00, 0.00)$, $\boldsymbol{\beta}_3^\top = (-0.30, -0.90, 0.00, 0.00, 0.00, 0.00, 0.00)$, $\mathbb{\Gamma}_1^{*\top} = (0.95, 0.85, 1.15)$, $\mathbb{\Gamma}_2^{*\top} = (0.70, -0.70, 0.80)$, and $\mathbb{\Gamma}_3^{*\top} = (1.75, 0.00, 0.85)$. Small values changed to zero, part 3.

n	$n = 100$			$n = 250$			$n = 500$			$n = 1000$		
Measure	Bias	Var	MSE	Bias	Var	MSE	Bias	Var	MSE	Bias	Var	MSE
β_{30}	22.0	21.1	25.7	28.2	8.5	16.3	33.3	6.8	17.8	26.1	3.5	10.3
β_{31}	41.1	33.2	49.8	49.6	21.0	45.4	61.9	10.9	49.0	59.3	7.2	42.3
β_{32}	0.8	3.4	3.4	0.5	1.5	1.5	-0.1	0.6	0.6	-0.6	0.4	0.4
β_{33}	-2.0	3.3	3.3	0.5	1.6	1.6	2.4	0.9	0.9	0.7	0.4	0.4
β_{34}	4.0	4.1	4.3	1.8	1.9	2.0	-0.7	1.2	1.1	-0.2	0.5	0.5
β_{35}	-4.0	2.8	2.9	-1.9	1.0	1.0	-0.3	0.9	0.9	-0.7	0.5	0.5
β_{36}	0.7	2.2	2.2	-0.1	1.4	1.4	-0.7	0.8	0.8	-0.9	0.4	0.4
Γ_{31}	-34.0	24.4	35.7	-28.7	20.9	28.9	-25.6	10.0	16.4	-28.1	4.9	12.8
Γ_{312}	19.0	37.8	41.1	25.4	14.7	21.0	20.0	9.1	13.1	14.3	4.1	6.1
Γ_{32}	-0.2	23.0	22.8	24.0	18.0	23.6	30.3	6.9	16.0	29.4	3.0	11.6
Total			534.7			360.3			293.3			218.9

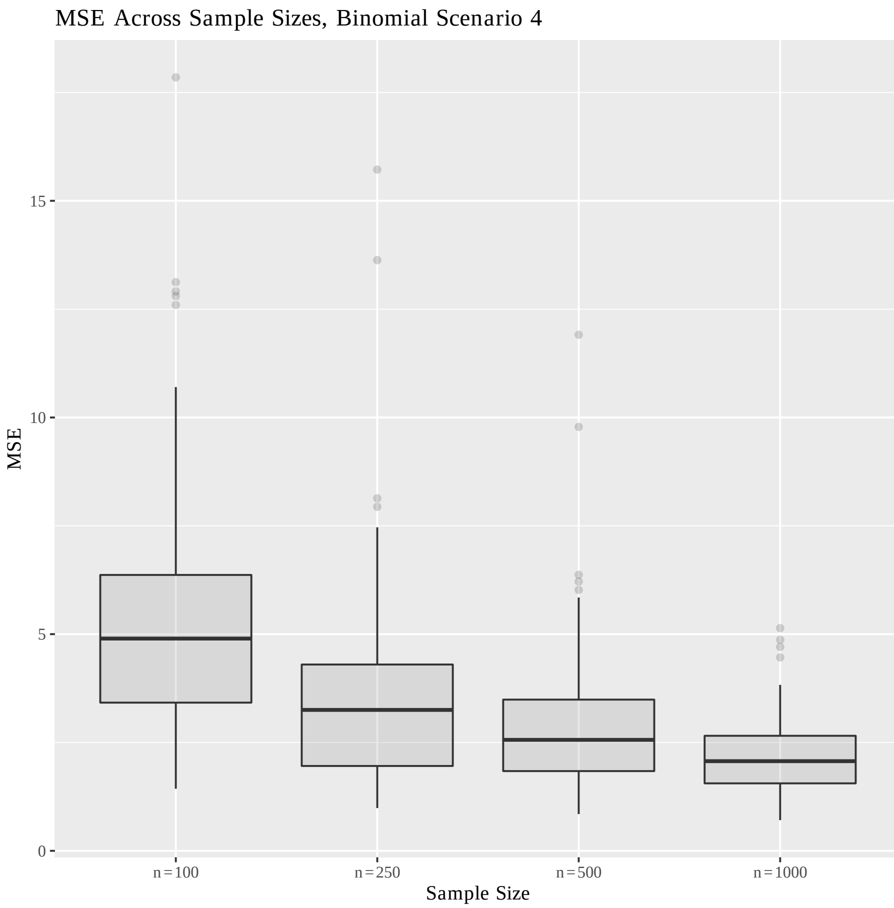

Figure H–29: MSE across sample sizes, binomial outcome, small values changed to zero, with $K = 3$, $p = 7$, $q = 2$.

Table H–54: Simulation 10 results multiplied by 100, averaged over 100 runs. Outcome follows a binomial distribution with $K = 3$, $p = 2$, $q = 2$, $\pi_1 = 0.5$, $\pi_2 = 0.3$, $\pi_3 = 0.2$, $\boldsymbol{\beta}_1^\top = (0.95, 0.60)$, $\boldsymbol{\beta}_2^\top = (-0.85, -0.15)$, $\boldsymbol{\beta}_3^\top = (-0.30, -0.90)$, $\mathbb{\Gamma}_1^{*\top} = (0.95, 0.85, 1.15)$, $\mathbb{\Gamma}_2^{*\top} = (0.70, -0.70, 0.80)$, and $\mathbb{\Gamma}_3^{*\top} = (1.75, 0.00, 0.85)$. Oracle model.

n	$n = 100$			$n = 250$			$n = 500$			$n = 1000$		
Measure	Bias	Var	MSE	Bias	Var	MSE	Bias	Var	MSE	Bias	Var	MSE
π_1	-21.9	0.6	5.4	-22.6	0.4	5.5	-22.7	0.3	5.5	-22.1	0.2	5.1
π_2	7.2	1.0	1.5	9.0	0.9	1.7	10.5	0.7	1.8	10.0	0.4	1.4
β_{10}	43.0	10.9	29.2	37.0	3.3	17.0	42.6	2.2	20.4	40.5	1.8	18.1
β_{11}	59.2	22.0	56.8	59.4	8.7	43.9	60.5	7.2	43.8	57.0	5.9	38.3
Γ_{11}	-16.7	22.8	25.4	-21.3	5.7	10.2	-21.9	2.7	7.5	-21.2	2.0	6.5
Γ_{112}	-12.7	31.9	33.2	-15.7	21.2	23.4	-9.9	9.2	10.1	-11.9	5.7	7.1
Γ_{12}	-24.8	22.4	28.3	-18.6	7.8	11.1	-15.3	2.5	4.8	-13.2	1.4	3.2
β_{20}	23.9	8.1	13.7	30.7	5.4	14.8	29.0	3.1	11.5	28.1	2.4	10.3
β_{21}	-9.7	13.1	14.0	-9.0	5.6	6.3	-14.8	3.9	6.0	-14.4	2.7	4.7
Γ_{21}	16.0	10.1	12.5	20.8	6.7	11.0	19.8	4.8	8.6	19.7	2.7	6.6
Γ_{212}	27.7	19.5	27.0	29.7	12.8	21.5	28.8	5.6	13.9	30.3	4.3	13.4
Γ_{22}	-9.2	15.4	16.1	4.3	4.0	4.1	10.5	1.8	2.8	12.4	1.2	2.7
β_{30}	24.8	14.0	20.0	24.8	7.4	13.5	26.5	5.5	12.5	24.9	5.4	11.5
β_{31}	43.9	20.7	39.8	46.6	7.8	29.5	54.1	8.7	37.8	50.6	5.4	31.0
Γ_{31}	-33.6	16.4	27.6	-25.1	12.4	18.6	-22.9	10.1	15.2	-28.2	5.8	13.6
Γ_{312}	11.7	25.0	26.2	17.0	13.3	16.0	15.1	8.1	10.3	15.3	4.7	7.0
Γ_{32}	7.0	17.2	17.5	24.9	11.3	17.4	29.6	6.8	15.5	31.3	3.7	13.5
Total			394.3			265.4			228.0			193.9

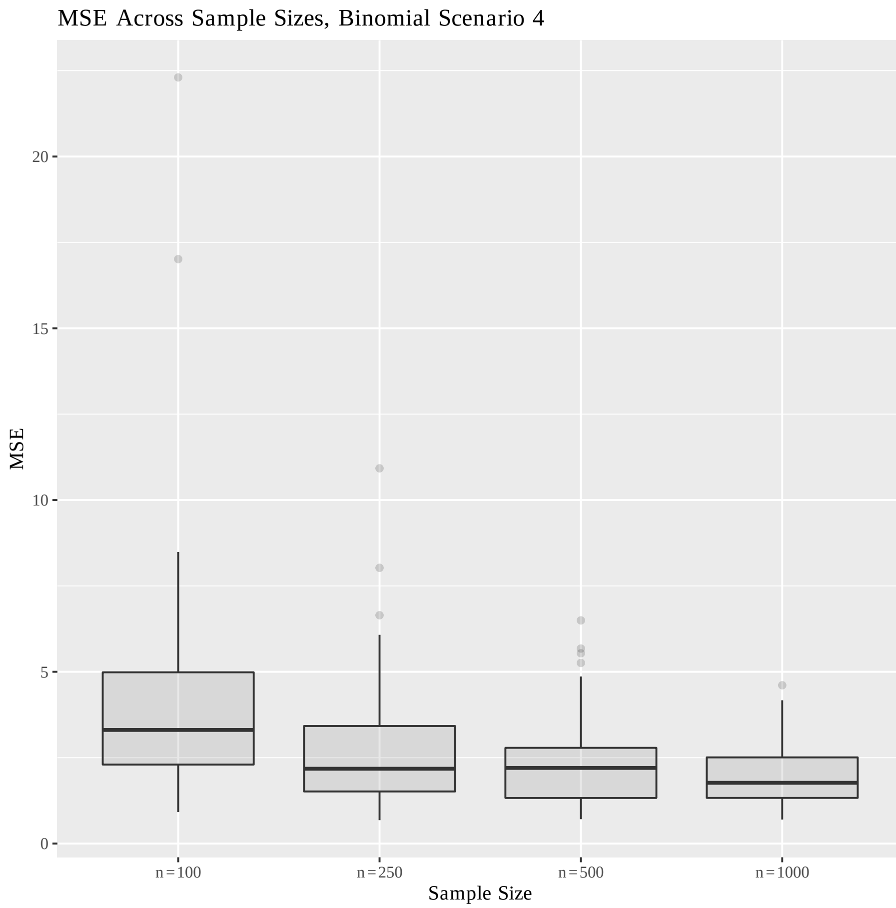

Figure H–30: MSE across sample sizes, binomial outcome, oracle model, with $K = 3$, $p = 7$, $q = 2$.

Table H–55: Simulation 11 results multiplied by 100, averaged over 100 runs. Outcome follows a binomial distribution with $K = 3$, $p = 17$, $q = 2$, $\pi_1 = 0.5$, $\pi_2 = 0.3$, $\pi_3 = 0.2$, $\boldsymbol{\beta}_1^\top = (0.95, 0.60, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00)$, $\boldsymbol{\beta}_2^\top = (-0.85, -0.15, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00)$, $\boldsymbol{\beta}_3^\top = (-0.30, -0.90, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00)$, $\mathbb{\Gamma}_1^{*\top} = (0.95, 0.85, 1.15)$, $\mathbb{\Gamma}_2^{*\top} = (0.70, -0.70, 0.80)$, and $\mathbb{\Gamma}_3^{*\top} = (1.75, 0.00, 0.85)$. No penalty, part 1.

n	$n = 100$			$n = 250$			$n = 500$			$n = 1000$		
Measure	Bias	Var	MSE	Bias	Var	MSE	Bias	Var	MSE	Bias	Var	MSE
π_1	-15.4	1.4	3.8	-14.3	0.8	2.9	-13.2	0.5	2.2	-13.5	0.3	2.1
π_2	-1.3	1.4	1.4	-2.4	0.9	0.9	-2.3	0.6	0.6	-3.3	0.5	0.6
β_{10}	19.3	16.1	19.6	13.0	4.8	6.5	14.3	3.2	5.2	16.7	2.7	5.4
β_{11}	34.5	38.7	50.2	18.8	13.6	17.0	19.2	5.8	9.4	24.3	5.7	11.6
β_{12}	1.0	3.3	3.2	0.4	0.8	0.8	-0.2	0.3	0.3	-0.2	0.2	0.2
β_{13}	0.9	2.7	2.7	1.1	0.8	0.8	-0.5	0.4	0.4	-0.1	0.2	0.2
β_{14}	1.9	2.8	2.8	-0.3	0.7	0.7	-0.2	0.3	0.3	0.0	0.1	0.1
β_{15}	-1.8	2.4	2.4	-0.3	0.7	0.7	-0.8	0.3	0.3	-0.7	0.2	0.2
β_{16}	-3.1	2.6	2.7	-0.5	0.6	0.6	0.0	0.3	0.3	0.4	0.1	0.1
β_{17}	0.5	2.4	2.4	0.3	0.7	0.7	-0.7	0.3	0.3	-0.3	0.1	0.1
β_{18}	0.1	2.3	2.2	1.8	0.7	0.7	0.8	0.3	0.3	0.4	0.1	0.1
β_{19}	0.3	1.7	1.6	1.6	0.5	0.6	0.7	0.3	0.3	0.0	0.1	0.1
β_{110}	-1.7	2.5	2.5	-1.4	0.8	0.8	-0.1	0.4	0.4	-0.3	0.1	0.1
β_{111}	0.2	2.7	2.7	-1.7	0.6	0.7	-0.1	0.3	0.3	-0.1	0.2	0.2
Γ_{11}	-5.5	17.8	17.9	0.4	6.9	6.9	-1.0	2.5	2.5	-4.2	1.3	1.5
Γ_{112}	0.0	50.2	49.7	15.8	15.0	17.3	18.7	6.0	9.4	8.0	6.7	7.2
Γ_{12}	-38.2	23.8	38.1	-8.8	8.2	8.9	-7.8	2.5	3.1	-5.6	1.2	1.5

Table H–56: Simulation 11 results multiplied by 100, averaged over 100 runs. Outcome follows a binomial distribution with $K = 3$, $p = 17$, $q = 2$, $\pi_1 = 0.5$, $\pi_2 = 0.3$, $\pi_3 = 0.2$, $\boldsymbol{\beta}_1^\top = (0.95, 0.60, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00)$, $\boldsymbol{\beta}_2^\top = (-0.85, -0.15, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00)$, $\boldsymbol{\beta}_3^\top = (-0.30, -0.90, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00)$, $\mathbb{\Gamma}_1^{*\top} = (0.95, 0.85, 1.15)$, $\mathbb{\Gamma}_2^{*\top} = (0.70, -0.70, 0.80)$, and $\mathbb{\Gamma}_3^{*\top} = (1.75, 0.00, 0.85)$. No penalty, part 2.

n	$n = 100$			$n = 250$			$n = 500$			$n = 1000$		
Measure	Bias	Var	MSE	Bias	Var	MSE	Bias	Var	MSE	Bias	Var	MSE
β_{20}	15.7	16.4	18.8	15.7	9.7	12.1	14.9	4.3	6.4	15.2	2.3	4.6
β_{21}	-10.6	23.9	24.8	-13.1	13.2	14.8	-16.3	5.7	8.3	-17.3	3.6	6.5
β_{22}	0.1	3.5	3.5	-0.4	1.7	1.7	-0.4	0.8	0.7	-0.8	0.4	0.4
β_{23}	-0.1	4.7	4.6	1.0	1.3	1.3	-0.7	0.4	0.4	-0.6	0.3	0.3
β_{24}	0.4	3.6	3.5	-0.7	1.4	1.4	-0.4	0.6	0.6	-0.3	0.3	0.3
β_{25}	-0.6	4.0	4.0	2.0	1.1	1.1	-0.3	0.5	0.5	-0.8	0.2	0.2
β_{26}	-0.1	4.8	4.7	-0.2	1.4	1.3	0.1	0.4	0.4	1.1	0.3	0.3
β_{27}	-1.5	2.0	2.0	0.9	1.5	1.5	0.3	0.4	0.4	-0.4	0.4	0.4
β_{28}	-0.7	4.3	4.2	2.1	1.1	1.1	1.2	0.5	0.5	0.8	0.3	0.3
β_{29}	-1.2	3.3	3.3	-0.3	1.2	1.2	-0.5	0.4	0.4	-0.1	0.3	0.3
β_{210}	3.4	3.1	3.2	1.0	1.6	1.6	1.0	0.4	0.4	0.3	0.2	0.2
β_{211}	1.3	2.8	2.8	0.2	1.1	1.1	1.1	0.5	0.5	0.0	0.3	0.3
Γ_{21}	2.6	18.4	18.2	9.1	8.3	9.0	4.3	3.7	3.9	2.5	2.0	2.1
Γ_{212}	19.3	36.6	40.0	12.5	17.7	19.1	3.5	9.2	9.2	2.2	5.2	5.2
Γ_{22}	-34.0	20.2	31.6	-27.6	11.6	19.1	-18.4	5.7	9.1	-9.1	2.8	3.6

Table H–57: Simulation 11 results multiplied by 100, averaged over 100 runs. Outcome follows a binomial distribution with $K = 3$, $p = 17$, $q = 2$, $\pi_1 = 0.5$, $\pi_2 = 0.3$, $\pi_3 = 0.2$, $\boldsymbol{\beta}_1^\top = (0.95, 0.60, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00)$, $\boldsymbol{\beta}_2^\top = (-0.85, -0.15, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00)$, $\boldsymbol{\beta}_3^\top = (-0.30, -0.90, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00)$, $\mathbb{\Gamma}_1^{*\top} = (0.95, 0.85, 1.15)$, $\mathbb{\Gamma}_2^{*\top} = (0.70, -0.70, 0.80)$, and $\mathbb{\Gamma}_3^{*\top} = (1.75, 0.00, 0.85)$. No penalty, part 3.

n	$n = 100$			$n = 250$			$n = 500$			$n = 1000$		
Measure	Bias	Var	MSE	Bias	Var	MSE	Bias	Var	MSE	Bias	Var	MSE
β_{30}	20.0	27.7	31.5	17.0	17.8	20.6	18.1	7.7	10.9	15.4	3.9	6.3
β_{31}	44.1	21.3	40.5	57.3	16.0	48.7	58.6	7.2	41.4	55.4	4.4	35.0
β_{32}	-0.2	3.3	3.3	-0.5	1.3	1.3	0.6	1.0	1.0	0.9	0.4	0.4
β_{33}	-0.2	2.8	2.7	-0.5	1.4	1.4	1.8	0.9	1.0	0.8	0.4	0.4
β_{34}	1.7	2.2	2.2	1.2	1.0	1.0	0.9	0.9	0.9	0.0	0.4	0.4
β_{35}	1.0	3.1	3.1	-2.2	1.2	1.2	-0.8	0.8	0.8	0.6	0.4	0.4
β_{36}	0.6	2.4	2.4	1.0	1.3	1.3	0.5	0.6	0.6	-1.3	0.4	0.4
β_{37}	4.0	3.1	3.2	0.0	1.1	1.1	0.4	0.7	0.7	-0.2	0.4	0.4
β_{38}	1.3	2.9	2.9	-1.0	1.3	1.3	-0.7	0.6	0.6	-1.2	0.3	0.3
β_{39}	0.9	2.9	2.9	-0.2	1.1	1.1	0.7	0.8	0.8	0.2	0.2	0.2
β_{310}	0.5	4.0	3.9	0.5	1.1	1.1	-0.8	0.6	0.6	-0.4	0.4	0.4
β_{311}	-1.6	3.0	3.0	-0.3	1.0	1.0	-1.0	0.7	0.7	-0.9	0.4	0.4
Γ_{31}	-39.0	39.6	54.4	-29.9	17.0	25.8	-22.3	7.6	12.5	-23.0	3.0	8.2
Γ_{312}	15.8	35.1	37.2	9.1	14.3	15.0	5.6	6.4	6.6	3.8	2.6	2.7
Γ_{32}	-10.6	28.2	29.1	9.3	14.6	15.4	23.3	8.4	13.7	24.6	3.0	9.1
Total			597.4			292.2			170.4			121.5

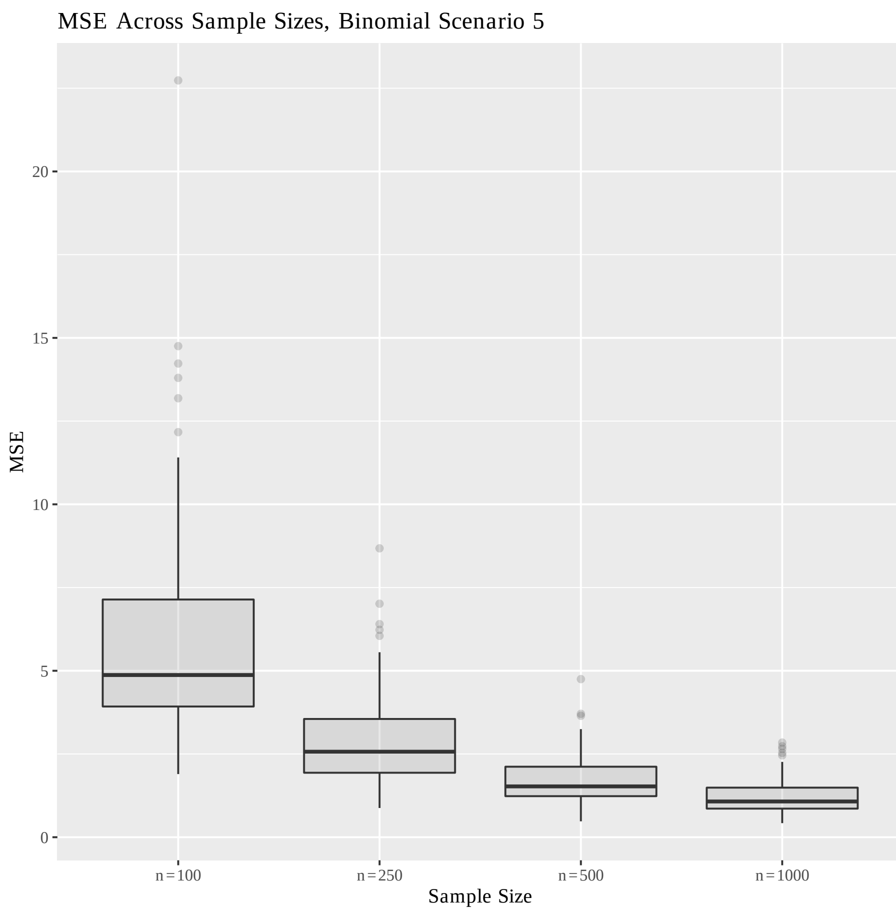

Figure H–31: MSE across sample sizes, binomial outcome, no penalty, with $K = 3$, $p = 17$, $q = 2$.

Table H–58: Simulation 11 results multiplied by 100, averaged over 100 runs. Outcome follows a binomial distribution with $K = 3$, $p = 17$, $q = 2$, $\pi_1 = 0.5$, $\pi_2 = 0.3$, $\pi_3 = 0.2$, $\boldsymbol{\beta}_1^\top = (0.95, 0.60, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00)$, $\boldsymbol{\beta}_2^\top = (-0.85, -0.15, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00)$, $\boldsymbol{\beta}_3^\top = (-0.30, -0.90, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00)$, $\mathbb{\Gamma}_1^{*\top} = (0.95, 0.85, 1.15)$, $\mathbb{\Gamma}_2^{*\top} = (0.70, -0.70, 0.80)$, and $\mathbb{\Gamma}_3^{*\top} = (1.75, 0.00, 0.85)$. Small values changed to zero, part 1.

n	$n = 100$			$n = 250$			$n = 500$			$n = 1000$		
Measure	Bias	Var	MSE	Bias	Var	MSE	Bias	Var	MSE	Bias	Var	MSE
π_1	-15.4	1.4	3.8	-14.3	0.8	2.9	-13.2	0.5	2.2	-13.5	0.3	2.1
π_2	-1.3	1.4	1.4	-2.4	0.9	0.9	-2.3	0.6	0.6	-3.3	0.5	0.6
β_{10}	19.3	16.1	19.6	13.0	4.8	6.5	14.3	3.2	5.2	16.7	2.7	5.4
β_{11}	34.5	38.7	50.2	18.8	13.6	17.0	19.2	5.8	9.4	24.3	5.7	11.6
β_{12}	1.1	3.3	3.2	0.4	0.8	0.8	-0.2	0.3	0.3	-0.3	0.2	0.2
β_{13}	0.9	2.7	2.7	1.1	0.8	0.8	-0.5	0.4	0.4	-0.1	0.2	0.2
β_{14}	1.9	2.8	2.8	-0.3	0.7	0.7	-0.2	0.3	0.3	0.1	0.1	0.1
β_{15}	-1.8	2.4	2.4	-0.3	0.7	0.7	-0.8	0.3	0.3	-0.8	0.2	0.2
β_{16}	-3.1	2.6	2.7	-0.5	0.6	0.6	0.0	0.3	0.3	0.4	0.1	0.1
β_{17}	0.5	2.4	2.4	0.3	0.7	0.7	-0.7	0.3	0.3	-0.3	0.1	0.1
β_{18}	0.1	2.3	2.2	1.8	0.7	0.7	0.8	0.3	0.3	0.3	0.1	0.1
β_{19}	0.3	1.7	1.6	1.6	0.5	0.6	0.7	0.3	0.3	0.0	0.1	0.1
β_{110}	-1.7	2.5	2.5	-1.4	0.8	0.8	-0.1	0.4	0.4	-0.3	0.1	0.1
β_{111}	0.2	2.7	2.7	-1.7	0.6	0.7	-0.1	0.3	0.3	-0.1	0.2	0.2
Γ_{11}	-5.5	17.8	17.9	0.4	6.9	6.9	-1.0	2.5	2.5	-4.2	1.3	1.5
Γ_{112}	0.0	50.2	49.7	15.8	15.0	17.3	18.7	6.0	9.4	8.0	6.7	7.2
Γ_{12}	-38.2	23.8	38.1	-8.8	8.2	8.9	-7.8	2.5	3.1	-5.6	1.2	1.5

Table H–59: Simulation 11 results multiplied by 100, averaged over 100 runs. Outcome follows a binomial distribution with $K = 3$, $p = 17$, $q = 2$, $\pi_1 = 0.5$, $\pi_2 = 0.3$, $\pi_3 = 0.2$, $\boldsymbol{\beta}_1^\top = (0.95, 0.60, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00)$, $\boldsymbol{\beta}_2^\top = (-0.85, -0.15, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00)$, $\boldsymbol{\beta}_3^\top = (-0.30, -0.90, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00)$, $\mathbb{\Gamma}_1^{*\top} = (0.95, 0.85, 1.15)$, $\mathbb{\Gamma}_2^{*\top} = (0.70, -0.70, 0.80)$, and $\mathbb{\Gamma}_3^{*\top} = (1.75, 0.00, 0.85)$. Small values changed to zero, part 2.

n	$n = 100$			$n = 250$			$n = 500$			$n = 1000$		
Measure	Bias	Var	MSE	Bias	Var	MSE	Bias	Var	MSE	Bias	Var	MSE
β_{20}	15.7	16.4	18.8	15.7	9.7	12.1	14.9	4.3	6.4	15.2	2.3	4.6
β_{21}	-10.6	23.9	24.8	-13.1	13.2	14.8	-16.3	5.7	8.3	-17.3	3.6	6.5
β_{22}	0.1	3.5	3.5	-0.4	1.7	1.7	-0.3	0.7	0.7	-0.8	0.4	0.4
β_{23}	-0.1	4.7	4.6	1.0	1.3	1.3	-0.7	0.4	0.4	-0.6	0.3	0.3
β_{24}	0.4	3.6	3.5	-0.7	1.4	1.4	-0.4	0.6	0.6	-0.3	0.3	0.3
β_{25}	-0.6	4.0	4.0	2.0	1.1	1.1	-0.2	0.5	0.5	-0.8	0.2	0.2
β_{26}	-0.1	4.8	4.7	-0.3	1.4	1.3	0.0	0.4	0.4	1.1	0.3	0.3
β_{27}	-1.5	2.0	2.0	0.9	1.5	1.5	0.3	0.4	0.4	-0.4	0.4	0.4
β_{28}	-0.7	4.3	4.2	2.0	1.1	1.1	1.2	0.5	0.5	0.8	0.3	0.3
β_{29}	-1.2	3.3	3.3	-0.3	1.2	1.2	-0.5	0.4	0.4	-0.1	0.3	0.3
β_{210}	3.4	3.1	3.2	1.0	1.6	1.6	1.0	0.4	0.4	0.3	0.2	0.2
β_{211}	1.3	2.8	2.8	0.2	1.1	1.1	1.1	0.5	0.5	0.0	0.3	0.3
Γ_{21}	2.6	18.4	18.2	9.1	8.3	9.0	4.3	3.7	3.9	2.5	2.0	2.1
Γ_{212}	19.3	36.6	40.0	12.5	17.7	19.1	3.5	9.2	9.2	2.2	5.2	5.2
Γ_{22}	-34.0	20.2	31.6	-27.6	11.6	19.1	-18.4	5.7	9.1	-9.1	2.8	3.6

Table H–60: Simulation 11 results multiplied by 100, averaged over 100 runs. Outcome follows a binomial distribution with $K = 3$, $p = 17$, $q = 2$, $\pi_1 = 0.5$, $\pi_2 = 0.3$, $\pi_3 = 0.2$, $\boldsymbol{\beta}_1^\top = (0.95, 0.60, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00)$, $\boldsymbol{\beta}_2^\top = (-0.85, -0.15, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00)$, $\boldsymbol{\beta}_3^\top = (-0.30, -0.90, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00)$, $\mathbb{\Gamma}_1^{*\top} = (0.95, 0.85, 1.15)$, $\mathbb{\Gamma}_2^{*\top} = (0.70, -0.70, 0.80)$, and $\mathbb{\Gamma}_3^{*\top} = (1.75, 0.00, 0.85)$. Small values changed to zero, part 3.

n	$n = 100$			$n = 250$			$n = 500$			$n = 1000$		
Measure	Bias	Var	MSE	Bias	Var	MSE	Bias	Var	MSE	Bias	Var	MSE
β_{30}	20.0	27.7	31.5	17.0	17.8	20.6	18.1	7.7	10.9	15.4	3.9	6.3
β_{31}	44.1	21.2	40.5	57.3	16.0	48.7	58.6	7.2	41.4	55.4	4.4	35.0
β_{32}	-0.2	3.3	3.3	-0.5	1.3	1.3	0.6	1.0	1.0	0.9	0.4	0.4
β_{33}	-0.2	2.8	2.7	-0.6	1.4	1.4	1.7	0.9	1.0	0.9	0.4	0.4
β_{34}	1.7	2.2	2.2	1.1	1.0	1.0	0.8	0.9	0.9	0.0	0.4	0.4
β_{35}	1.0	3.1	3.1	-2.2	1.2	1.2	-0.8	0.8	0.8	0.6	0.4	0.4
β_{36}	0.6	2.4	2.4	1.0	1.3	1.3	0.4	0.6	0.6	-1.3	0.4	0.4
β_{37}	4.0	3.1	3.2	0.0	1.1	1.1	0.4	0.7	0.7	-0.2	0.4	0.4
β_{38}	1.3	2.9	2.9	-1.0	1.3	1.3	-0.7	0.6	0.6	-1.3	0.3	0.3
β_{39}	0.9	2.9	2.9	-0.2	1.1	1.1	0.7	0.8	0.8	0.2	0.2	0.2
β_{310}	0.5	4.0	3.9	0.5	1.1	1.1	-0.8	0.6	0.6	-0.4	0.4	0.4
β_{311}	-1.6	3.0	3.0	-0.3	1.0	1.0	-0.9	0.7	0.7	-0.9	0.4	0.4
Γ_{31}	-39.0	39.6	54.4	-29.9	17.0	25.8	-22.3	7.6	12.5	-23.0	3.0	8.2
Γ_{312}	15.8	35.1	37.2	9.1	14.3	15.0	5.6	6.4	6.6	3.8	2.6	2.7
Γ_{32}	-10.6	28.2	29.1	9.3	14.6	15.4	23.3	8.4	13.7	24.6	3.0	9.1
Total			597.4			292.2			170.4			121.4

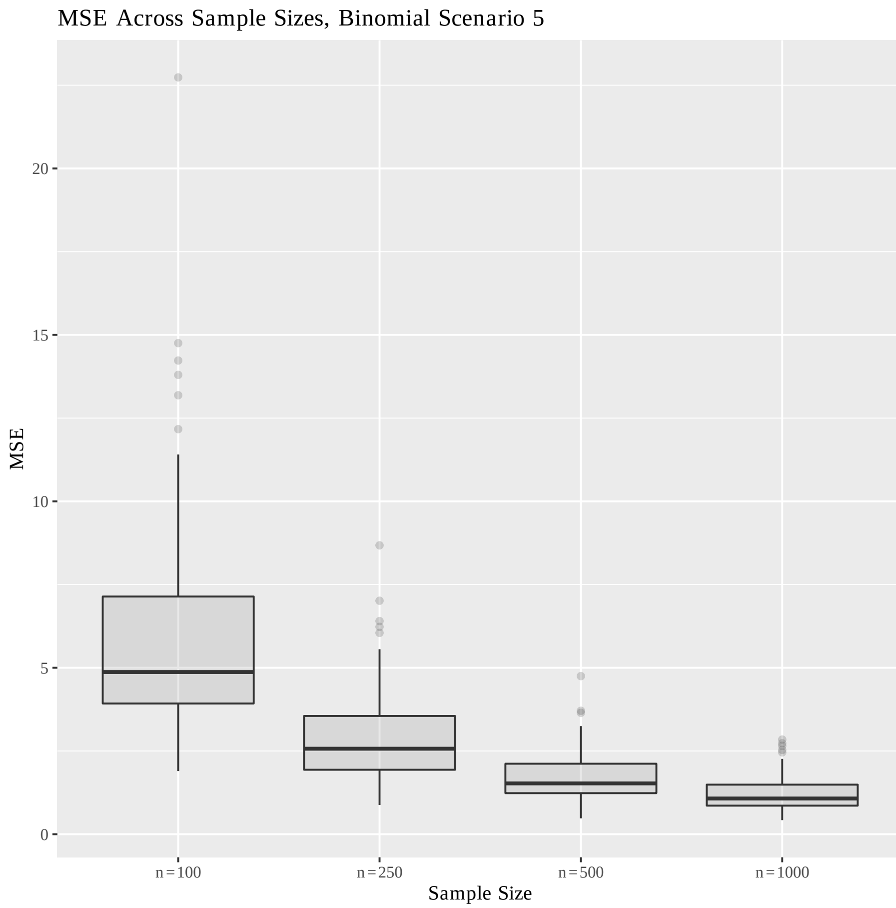

Figure H–32: MSE across sample sizes, binomial outcome, small values changed to zero, with $K = 3$, $p = 17$, $q = 2$.

Table H–61: Simulation 11 results multiplied by 100, averaged over 100 runs. Outcome follows a binomial distribution with $K = 3$, $p = 2$, $q = 2$, $\pi_1 = 0.5$, $\pi_2 = 0.3$, $\pi_3 = 0.2$, $\boldsymbol{\beta}_1^\top = (0.95, 0.60)$, $\boldsymbol{\beta}_2^\top = (-0.85, -0.15)$, $\boldsymbol{\beta}_3^\top = (-0.30, -0.90)$, $\mathbb{\Gamma}_1^{*\top} = (0.95, 0.85, 1.15)$, $\mathbb{\Gamma}_2^{*\top} = (0.70, -0.70, 0.80)$, and $\mathbb{\Gamma}_3^{*\top} = (1.75, 0.00, 0.85)$. Oracle model.

n	$n = 100$			$n = 250$			$n = 500$			$n = 1000$		
Measure	Bias	Var	MSE	Bias	Var	MSE	Bias	Var	MSE	Bias	Var	MSE
π_1	-12.7	0.8	2.4	-12.8	0.5	2.1	-11.9	0.3	1.7	-12.4	0.3	1.8
π_2	1.2	0.7	0.7	0.2	0.5	0.5	-1.3	0.4	0.4	0.7	0.4	0.4
β_{10}	14.2	7.8	9.7	13.9	3.6	5.5	15.1	2.3	4.6	18.4	2.2	5.5
β_{11}	26.1	15.2	21.8	23.0	8.1	13.3	20.6	5.1	9.3	24.2	3.9	9.7
Γ_{11}	-5.0	7.5	7.7	-4.9	3.5	3.7	-2.7	2.0	2.0	-5.9	1.5	1.9
Γ_{112}	5.5	26.2	26.2	8.3	9.6	10.2	11.4	5.6	6.9	3.3	4.3	4.3
Γ_{12}	-23.3	12.8	18.1	-7.3	4.5	5.0	-5.3	2.2	2.5	-3.4	0.9	1.0
β_{20}	15.1	11.0	13.1	15.1	5.1	7.3	13.9	2.6	4.4	17.4	1.7	4.7
β_{21}	-11.5	12.4	13.6	-18.0	7.8	10.9	-17.0	4.6	7.5	-20.1	2.1	6.1
Γ_{21}	7.5	12.6	13.1	4.8	3.9	4.0	4.3	2.1	2.2	4.2	0.9	1.1
Γ_{212}	6.5	23.1	23.3	5.7	10.9	11.1	0.0	8.1	8.0	5.1	4.8	5.1
Γ_{22}	-25.9	15.0	21.6	-15.3	6.7	9.0	-11.8	3.2	4.5	-4.1	1.5	1.6
β_{30}	22.1	25.9	30.5	18.2	14.0	17.1	14.7	5.7	7.8	13.2	3.2	4.9
β_{31}	32.6	20.0	30.5	48.9	8.1	31.9	54.5	4.7	34.4	52.0	3.2	30.2
Γ_{31}	-26.3	36.4	43.0	-20.1	24.0	27.9	-19.7	5.3	9.1	-16.9	3.0	5.8
Γ_{312}	3.1	39.5	39.2	9.7	13.9	14.7	4.5	4.2	4.4	3.7	2.4	2.5
Γ_{32}	6.8	30.9	31.1	17.9	11.1	14.2	24.8	6.5	12.6	26.0	4.4	11.1
Total			345.7			188.5			122.4			97.8

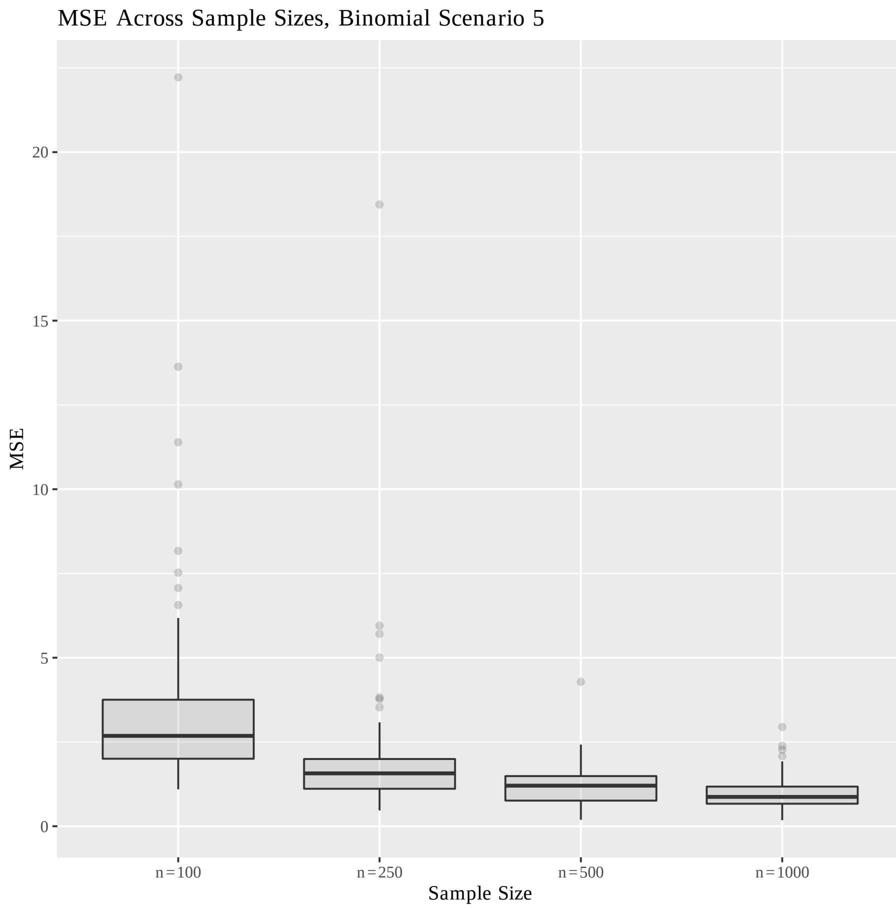

Figure H–33: MSE across sample sizes, binomial outcome, oracle model, with $K = 3$, $p = 17$, $q = 2$.

Table H–62: Simulation 12 results multiplied by 100, averaged over 100 runs. Outcome follows a binomial distribution with $K = 3$, $p = 19$, $q = 5$, $\pi_1 = 0.5$, $\pi_2 = 0.3$, $\pi_3 = 0.2$, $\boldsymbol{\beta}_1^\top = (0.95, 0.60, -0.65, -0.25, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00)$, $\boldsymbol{\beta}_2^\top = (-0.85, -0.15, -0.75, 0.10, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00)$, $\boldsymbol{\beta}_3^\top = (-0.30, -0.90, 0.80, -0.25, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00)$, $\mathbb{\Gamma}_1^{*\top} = (0.95, 0.85, 1.15, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00)$, $\mathbb{\Gamma}_2^{*\top} = (0.70, -0.70, 0.80, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00)$, and $\mathbb{\Gamma}_3^{*\top} = (1.75, 0.00, 0.85, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00)$. No penalty, part 1.

n	$n = 100$			$n = 250$			$n = 500$			$n = 1000$		
Measure	Bias	Var	MSE	Bias	Var	MSE	Bias	Var	MSE	Bias	Var	MSE
π_1	-8.1	0.9	1.5	-6.4	0.5	0.9	-5.7	0.3	0.6	-5.0	0.2	0.4
π_2	2.5	0.7	0.7	3.3	0.4	0.5	4.3	0.2	0.4	4.5	0.1	0.3
β_{10}	16.6	14.6	17.2	18.1	3.6	6.8	17.5	2.2	5.2	14.4	1.4	3.5
β_{11}	18.7	23.7	27.0	14.9	7.3	9.5	13.7	3.9	5.7	10.2	2.3	3.3
β_{12}	-3.5	3.6	3.6	-3.2	0.9	1.0	-1.4	0.4	0.4	-1.3	0.1	0.2
β_{13}	0.8	1.5	1.5	0.6	0.4	0.4	1.3	0.2	0.2	0.7	0.1	0.1
β_{14}	-0.7	1.3	1.3	-0.8	0.5	0.5	0.1	0.2	0.2	-0.1	0.1	0.1
β_{15}	-0.4	1.6	1.6	0.2	0.4	0.4	0.4	0.1	0.1	0.3	0.1	0.1
β_{16}	-0.3	1.6	1.6	0.0	0.5	0.4	-0.2	0.1	0.1	-0.3	0.1	0.1
β_{17}	-0.8	1.7	1.7	0.0	0.4	0.4	-0.1	0.2	0.2	0.0	0.1	0.1
β_{18}	-3.0	1.9	2.0	-0.4	0.3	0.3	-0.5	0.2	0.2	-0.4	0.1	0.1
β_{19}	0.8	1.8	1.8	0.1	0.3	0.3	0.3	0.1	0.1	-0.1	0.1	0.1
β_{110}	-0.5	1.4	1.4	-0.5	0.4	0.3	-0.2	0.2	0.2	0.0	0.1	0.1
β_{111}	1.1	1.3	1.3	0.0	0.3	0.3	0.3	0.1	0.1	-0.3	0.1	0.1
β_{112}	0.2	1.5	1.5	-0.8	0.3	0.3	-0.1	0.2	0.2	0.1	0.1	0.1
β_{113}	1.0	1.8	1.8	0.6	0.4	0.4	0.2	0.2	0.1	-0.1	0.1	0.1
Γ_{11}	-10.0	12.5	13.4	-15.0	2.8	5.1	-13.5	1.2	3.0	-12.3	0.6	2.1
Γ_{112}	-16.4	26.6	29.1	-9.5	7.3	8.2	-11.7	5.2	6.5	-7.4	2.2	2.8
Γ_{12}	-23.0	15.3	20.4	-9.2	2.9	3.7	-1.9	1.3	1.3	1.3	0.8	0.8
Γ_{113}	-1.4	4.0	4.0	0.4	1.0	1.0	0.9	0.5	0.5	0.1	0.2	0.2
Γ_{123}	3.0	2.9	3.0	-0.4	1.6	1.6	-0.9	0.4	0.4	-0.5	0.1	0.1
Γ_{13}	4.3	1.0	1.1	4.5	0.6	0.8	4.4	0.4	0.6	3.6	0.2	0.3
Γ_{114}	-7.3	2.1	2.6	-3.8	0.8	0.9	-2.4	0.4	0.5	-1.2	0.2	0.2
Γ_{124}	2.3	2.0	2.0	0.0	0.6	0.6	-0.8	0.3	0.3	-0.7	0.1	0.1
Γ_{134}	-0.5	0.3	0.3	0.3	0.3	0.3	-1.1	0.3	0.3	-0.7	0.2	0.2
Γ_{14}	0.3	0.0	0.0	1.6	0.2	0.2	2.5	0.2	0.2	3.7	0.2	0.4
Γ_{115}	0.0	1.6	1.6	-0.1	0.5	0.5	0.0	0.3	0.3	0.2	0.1	0.1
Γ_{125}	0.7	2.1	2.1	-0.9	0.5	0.5	-0.6	0.3	0.3	-0.1	0.1	0.1
Γ_{135}	0.0	0.4	0.4	-0.3	0.2	0.2	-0.5	0.3	0.3	0.5	0.2	0.2
Γ_{145}	-0.1	0.0	0.0	0.1	0.1	0.1	0.1	0.1	0.1	-0.5	0.1	0.1
Γ_{15}	0.0	0.0	0.0	0.5	0.1	0.1	0.7	0.1	0.1	1.3	0.1	0.1

Table H–63: Simulation 12 results multiplied by 100, averaged over 100 runs. Outcome follows a binomial distribution with $K = 3$, $p = 19$, $q = 5$, $\pi_1 = 0.5$, $\pi_2 = 0.3$, $\pi_3 = 0.2$, $\boldsymbol{\beta}_1^\top = (0.95, 0.60, -0.65, -0.25, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00)$, $\boldsymbol{\beta}_2^\top = (-0.85, -0.15, -0.75, 0.10, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00)$, $\boldsymbol{\beta}_3^\top = (-0.30, -0.90, 0.80, -0.25, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00)$, $\mathbb{F}_1^{*\top} = (0.95, 0.85, 1.15, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00)$, $\mathbb{F}_2^{*\top} = (0.70, -0.70, 0.80, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00)$, and $\mathbb{F}_3^{*\top} = (1.75, 0.00, 0.85, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00)$. No penalty, part 2.

n	$n = 100$			$n = 250$			$n = 500$			$n = 1000$		
Measure	Bias	Var	MSE	Bias	Var	MSE	Bias	Var	MSE	Bias	Var	MSE
β_{20}	23.5	9.1	14.5	7.9	4.2	4.8	6.2	2.0	2.4	3.3	1.0	1.1
β_{21}	-12.4	15.8	17.2	-6.3	6.2	6.5	-6.3	3.4	3.8	-5.2	1.8	2.1
β_{22}	1.8	4.0	3.9	-2.0	1.4	1.5	-0.2	0.3	0.3	0.3	0.2	0.2
β_{23}	-12.1	2.5	3.9	-6.1	0.8	1.1	-5.8	0.3	0.6	-4.8	0.1	0.4
β_{24}	-1.3	2.2	2.1	-0.2	0.5	0.5	-0.7	0.2	0.2	-0.3	0.1	0.1
β_{25}	-1.0	2.1	2.1	-0.2	0.4	0.4	-0.6	0.2	0.2	-0.4	0.1	0.1
β_{26}	2.9	2.2	2.2	1.2	0.5	0.5	0.8	0.2	0.2	0.8	0.1	0.1
β_{27}	0.5	2.2	2.1	-0.8	0.5	0.5	0.1	0.2	0.2	-0.3	0.1	0.1
β_{28}	-0.3	1.6	1.6	-0.9	0.5	0.5	-0.3	0.2	0.2	0.2	0.1	0.1
β_{29}	0.6	2.2	2.2	1.1	0.6	0.6	0.6	0.2	0.2	0.4	0.1	0.1
β_{210}	1.0	2.2	2.2	-0.7	0.5	0.5	0.2	0.2	0.2	0.3	0.1	0.1
β_{211}	1.9	2.1	2.1	-0.3	0.5	0.5	-0.2	0.2	0.2	0.0	0.1	0.1
β_{212}	-0.1	1.4	1.3	-0.1	0.7	0.7	0.3	0.2	0.2	0.0	0.1	0.1
β_{213}	0.2	1.8	1.8	-0.6	0.6	0.6	-0.8	0.2	0.2	-0.2	0.1	0.1
Γ_{21}	-5.0	9.1	9.3	-4.4	2.7	2.8	-3.6	1.3	1.4	-2.9	0.6	0.7
Γ_{212}	35.2	21.3	33.5	12.6	10.5	12.0	9.2	3.7	4.5	9.9	2.1	3.1
Γ_{22}	-28.0	11.6	19.3	-8.4	6.4	7.1	-0.6	1.7	1.7	2.4	0.9	1.0
Γ_{213}	10.4	7.3	8.3	3.2	1.3	1.4	1.8	0.4	0.4	1.9	0.2	0.2
Γ_{223}	2.9	3.3	3.4	1.4	1.6	1.6	1.4	0.5	0.5	1.2	0.3	0.3
Γ_{23}	4.9	1.2	1.4	5.7	0.8	1.1	5.9	0.5	0.8	5.6	0.4	0.7
Γ_{214}	-8.0	3.1	3.7	-5.0	0.9	1.2	-5.0	0.6	0.8	-3.5	0.2	0.3
Γ_{224}	2.3	1.9	1.9	-0.1	0.7	0.7	-0.5	0.5	0.5	-0.7	0.2	0.2
Γ_{234}	-0.9	0.6	0.6	-1.5	1.0	1.0	-0.8	0.6	0.6	-1.8	0.4	0.5
Γ_{24}	0.4	0.1	0.1	2.7	0.4	0.5	3.7	0.4	0.5	5.0	0.3	0.6
Γ_{215}	1.5	3.1	3.1	0.4	0.8	0.8	0.5	0.4	0.4	0.2	0.2	0.2
Γ_{225}	1.2	1.4	1.4	0.6	0.6	0.6	1.0	0.4	0.4	0.6	0.2	0.2
Γ_{235}	-1.1	0.2	0.3	0.0	0.3	0.3	-0.2	0.3	0.3	-0.7	0.2	0.2
Γ_{245}	0.2	0.0	0.0	-0.3	0.1	0.1	-0.1	0.1	0.1	-0.1	0.1	0.1
Γ_{25}	0.1	0.0	0.0	0.2	0.0	0.0	0.4	< 0.1	< 0.1	0.7	< 0.1	< 0.1

Table H–64: Simulation 12 results multiplied by 100, averaged over 100 runs. Outcome follows a binomial distribution with $K = 3$, $p = 19$, $q = 5$, $\pi_1 = 0.5$, $\pi_2 = 0.3$, $\pi_3 = 0.2$, $\boldsymbol{\beta}_1^\top = (0.95, 0.60, -0.65, -0.25, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00)$, $\boldsymbol{\beta}_2^\top = (-0.85, -0.15, -0.75, 0.10, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00)$, $\boldsymbol{\beta}_3^\top = (-0.30, -0.90, 0.80, -0.25, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00)$, $\mathbb{\Gamma}_1^{*\top} = (0.95, 0.85, 1.15, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00)$, $\mathbb{\Gamma}_2^{*\top} = (0.70, -0.70, 0.80, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00)$, and $\mathbb{\Gamma}_3^{*\top} = (1.75, 0.00, 0.85, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00)$. No penalty, part 3.

n	$n = 100$			$n = 250$			$n = 500$			$n = 1000$		
Measure	Bias	Var	MSE	Bias	Var	MSE	Bias	Var	MSE	Bias	Var	MSE
β_{30}	10.4	24.6	25.5	0.2	10.7	10.6	1.6	5.0	5.0	3.4	2.9	3.0
β_{31}	13.8	32.2	33.8	11.7	13.6	14.9	5.7	4.9	5.2	3.7	1.7	1.8
β_{32}	-41.3	15.7	32.6	-18.0	8.8	12.0	-8.1	3.9	4.5	-2.6	0.9	0.9
β_{33}	3.0	4.7	4.8	1.1	1.2	1.2	1.4	0.6	0.6	0.1	0.3	0.2
β_{34}	-5.2	5.8	6.1	-2.4	1.3	1.4	0.1	0.5	0.5	0.2	0.2	0.2
β_{35}	5.1	4.4	4.6	3.6	1.3	1.4	1.6	0.4	0.5	0.6	0.2	0.2
β_{36}	4.8	6.9	7.1	1.4	1.5	1.5	1.3	0.5	0.5	0.3	0.2	0.2
β_{37}	-1.7	4.8	4.8	-0.2	1.1	1.1	-0.4	0.4	0.4	-0.1	0.1	0.1
β_{38}	0.0	6.0	5.9	-0.1	1.3	1.3	-0.2	0.5	0.5	0.0	0.3	0.3
β_{39}	-2.6	6.8	6.8	-2.7	1.0	1.0	-0.7	0.4	0.4	-0.5	0.2	0.2
β_{310}	-1.8	4.6	4.6	0.6	1.4	1.4	0.0	0.5	0.5	0.3	0.2	0.2
β_{311}	-0.3	4.0	4.0	-1.0	1.0	1.0	-1.2	0.5	0.5	-0.1	0.2	0.2
β_{312}	4.7	4.9	5.1	0.8	1.3	1.3	0.5	0.5	0.5	0.3	0.2	0.2
β_{313}	-2.8	4.1	4.1	-1.4	1.1	1.1	-0.4	0.4	0.4	0.6	0.2	0.2
Γ_{31}	-16.5	28.1	30.5	-4.2	13.8	13.8	1.2	4.9	4.9	4.1	1.9	2.1
Γ_{312}	7.4	38.3	38.5	-3.0	12.7	12.7	0.5	5.7	5.7	-0.3	2.3	2.3
Γ_{32}	-9.5	32.2	32.7	1.6	13.1	13.0	-0.3	4.0	3.9	-0.2	1.6	1.6
Γ_{313}	-14.9	20.3	22.3	-0.8	4.8	4.8	-1.8	2.2	2.3	-0.9	0.9	0.9
Γ_{323}	-0.7	13.9	13.7	-4.8	7.6	7.7	-4.1	2.6	2.8	-3.8	0.9	1.1
Γ_{33}	8.3	2.6	3.2	18.4	4.2	7.6	15.8	2.9	5.4	13.1	1.2	2.9
Γ_{314}	-0.8	9.5	9.4	-1.8	1.8	1.8	0.3	1.0	1.0	0.3	0.4	0.4
Γ_{324}	-0.9	5.3	5.3	-0.2	2.0	2.0	0.2	0.9	0.9	1.0	0.4	0.4
Γ_{334}	-1.6	0.4	0.5	-0.5	0.7	0.7	-0.6	0.6	0.6	-0.7	0.5	0.5
Γ_{34}	0.3	0.0	0.0	1.6	0.3	0.3	2.5	0.3	0.4	3.2	0.3	0.4
Γ_{315}	-6.7	7.4	7.8	-3.9	2.3	2.4	-1.9	0.9	1.0	-0.8	0.4	0.4
Γ_{325}	-2.6	4.9	4.9	-0.2	1.2	1.2	0.9	0.6	0.6	-0.1	0.3	0.3
Γ_{335}	-0.7	0.8	0.8	-0.3	0.8	0.7	-0.7	0.7	0.7	0.7	0.5	0.5
Γ_{345}	-0.3	0.0	0.0	-0.8	0.2	0.3	-0.5	0.1	0.1	0.5	0.1	0.1
Γ_{35}	0.2	0.0	0.0	1.0	0.2	0.2	0.5	< 0.1	< 0.1	1.2	0.1	0.1
Total			612.6			217.2			100.9			51.1

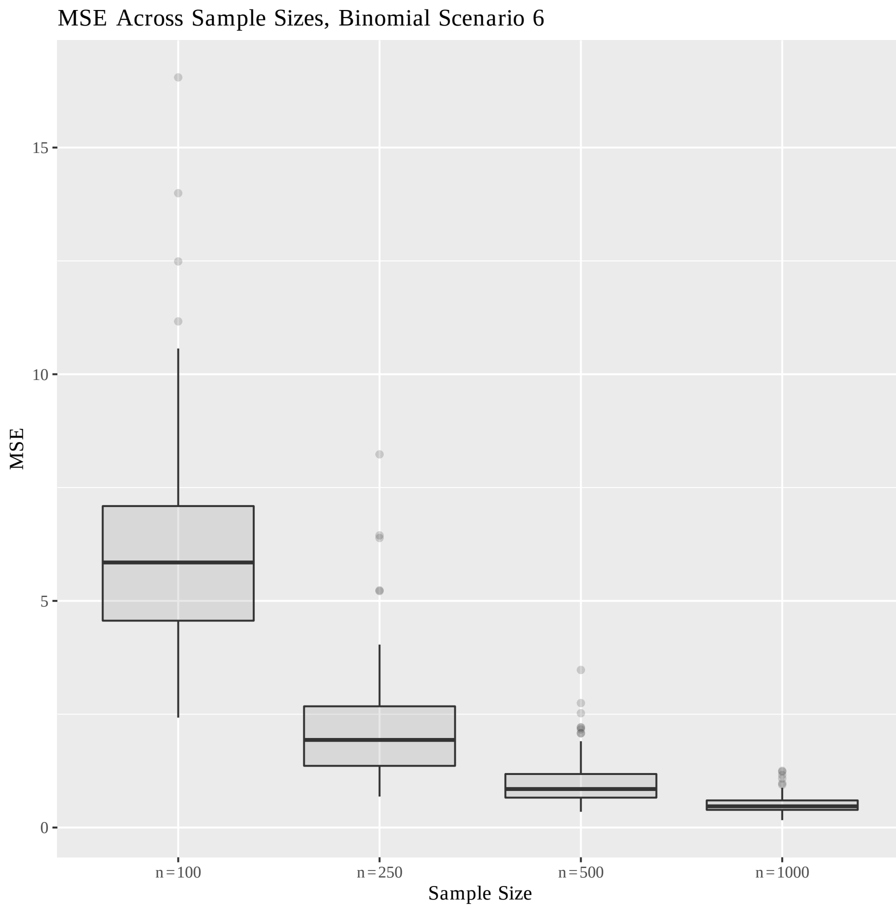

Figure H–34: MSE across sample sizes, binomial outcome, no penalty, with $K = 3$, $p = 19$, $q = 5$.

Table H–65: Simulation 12 results multiplied by 100, averaged over 100 runs. Outcome follows a binomial distribution with $K = 3$, $p = 19$, $q = 5$, $\pi_1 = 0.5$, $\pi_2 = 0.3$, $\pi_3 = 0.2$, $\boldsymbol{\beta}_1^\top = (0.95, 0.60, -0.65, -0.25, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00)$, $\boldsymbol{\beta}_2^\top = (-0.85, -0.15, -0.75, 0.10, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00)$, $\boldsymbol{\beta}_3^\top = (-0.30, -0.90, 0.80, -0.25, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00)$, $\mathbb{\Gamma}_1^{*\top} = (0.95, 0.85, 1.15, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00)$, $\mathbb{\Gamma}_2^{*\top} = (0.70, -0.70, 0.80, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00)$, and $\mathbb{\Gamma}_3^{*\top} = (1.75, 0.00, 0.85, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00)$. Small values changed to zero, part 1.

n	$n = 100$			$n = 250$			$n = 500$			$n = 1000$		
Measure	Bias	Var	MSE	Bias	Var	MSE	Bias	Var	MSE	Bias	Var	MSE
π_1	-8.1	0.9	1.5	-6.4	0.5	0.9	-5.7	0.3	0.6	-5.0	0.2	0.4
π_2	2.5	0.7	0.7	3.3	0.4	0.5	4.3	0.2	0.4	4.5	0.1	0.3
β_{10}	16.6	14.6	17.2	18.1	3.6	6.8	17.5	2.2	5.2	14.4	1.4	3.5
β_{11}	18.7	23.7	27.0	14.9	7.4	9.5	13.7	3.9	5.7	10.2	2.3	3.3
β_{12}	-3.5	3.6	3.6	-3.2	0.9	1.0	-1.4	0.4	0.4	-1.3	0.1	0.2
β_{13}	0.8	1.5	1.5	0.6	0.4	0.4	1.3	0.2	0.2	0.7	0.1	0.1
β_{14}	-0.7	1.3	1.3	-0.8	0.5	0.5	0.1	0.2	0.2	-0.1	0.1	0.1
β_{15}	-0.4	1.6	1.6	0.2	0.4	0.4	0.4	0.1	0.1	0.2	< 0.1	< 0.1
β_{16}	-0.3	1.6	1.6	0.0	0.5	0.4	-0.2	0.1	0.1	-0.3	0.1	0.1
β_{17}	-0.8	1.7	1.7	0.0	0.4	0.4	-0.1	0.2	0.2	-0.1	0.1	0.1
β_{18}	-2.9	1.9	2.0	-0.4	0.3	0.3	-0.5	0.2	0.2	-0.4	0.1	0.1
β_{19}	0.8	1.8	1.8	0.0	0.3	0.3	0.3	0.1	0.1	-0.1	0.1	0.1
β_{110}	-0.5	1.4	1.4	-0.5	0.4	0.3	-0.2	0.2	0.2	0.0	0.1	0.1
β_{111}	1.1	1.3	1.3	0.0	0.3	0.3	0.3	0.1	0.1	-0.3	0.1	0.1
β_{112}	0.2	1.5	1.5	-0.8	0.3	0.3	-0.1	0.2	0.2	0.1	0.1	0.1
β_{113}	1.0	1.8	1.8	0.6	0.4	0.4	0.2	0.1	0.1	-0.1	0.1	0.1
Γ_{11}	-10.0	12.5	13.4	-15.0	2.8	5.1	-13.5	1.2	3.0	-12.3	0.6	2.1
Γ_{112}	-16.4	26.6	29.1	-9.5	7.3	8.2	-11.7	5.2	6.5	-7.4	2.2	2.8
Γ_{12}	-23.0	15.3	20.4	-9.2	2.9	3.7	-1.9	1.3	1.3	1.3	0.8	0.8
Γ_{113}	-1.4	4.0	4.0	0.4	1.0	1.0	0.9	0.5	0.5	0.1	0.2	0.2
Γ_{123}	3.0	2.9	3.0	-0.4	1.6	1.6	-0.9	0.4	0.4	-0.5	0.1	0.1
Γ_{13}	4.3	1.0	1.1	4.5	0.6	0.8	4.4	0.4	0.6	3.6	0.2	0.3
Γ_{114}	-7.3	2.1	2.6	-3.8	0.8	0.9	-2.4	0.4	0.5	-1.2	0.2	0.2
Γ_{124}	2.3	2.0	2.0	0.0	0.6	0.6	-0.8	0.3	0.3	-0.7	0.1	0.1
Γ_{134}	-0.5	0.3	0.3	0.3	0.3	0.3	-1.1	0.3	0.3	-0.7	0.2	0.2
Γ_{14}	0.3	0.0	0.0	1.6	0.2	0.2	2.5	0.2	0.2	3.7	0.2	0.4
Γ_{115}	0.0	1.6	1.6	-0.1	0.5	0.5	0.0	0.3	0.3	0.2	0.1	0.1
Γ_{125}	0.7	2.1	2.1	-0.9	0.5	0.5	-0.6	0.3	0.3	-0.1	0.1	0.1
Γ_{135}	0.0	0.4	0.4	-0.3	0.2	0.2	-0.5	0.3	0.3	0.5	0.2	0.2
Γ_{145}	-0.1	0.0	0.0	0.1	0.1	0.1	0.1	0.1	0.1	-0.5	0.1	0.1
Γ_{15}	0.0	0.0	0.0	0.5	0.1	0.1	0.7	0.1	0.1	1.3	0.1	0.1

Table H–66: Simulation 12 results multiplied by 100, averaged over 100 runs. Outcome follows a binomial distribution with $K = 3$, $p = 19$, $q = 5$, $\pi_1 = 0.5$, $\pi_2 = 0.3$, $\pi_3 = 0.2$, $\boldsymbol{\beta}_1^\top = (0.95, 0.60, -0.65, -0.25, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00)$, $\boldsymbol{\beta}_2^\top = (-0.85, -0.15, -0.75, 0.10, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00)$, $\boldsymbol{\beta}_3^\top = (-0.30, -0.90, 0.80, -0.25, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00)$, $\mathbb{\Gamma}_1^{*\top} = (0.95, 0.85, 1.15, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00)$, $\mathbb{\Gamma}_2^{*\top} = (0.70, -0.70, 0.80, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00)$, and $\mathbb{\Gamma}_3^{*\top} = (1.75, 0.00, 0.85, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00)$. Small values changed to zero, part 2.

n	$n = 100$			$n = 250$			$n = 500$			$n = 1000$		
Measure	Bias	Var	MSE	Bias	Var	MSE	Bias	Var	MSE	Bias	Var	MSE
β_{20}	23.5	9.1	14.5	7.9	4.2	4.8	6.2	2.0	2.4	3.3	1.0	1.1
β_{21}	-12.5	15.8	17.2	-6.3	6.2	6.5	-6.3	3.4	3.8	-5.2	1.8	2.1
β_{22}	1.8	4.0	3.9	-2.0	1.4	1.5	-0.2	0.3	0.3	0.3	0.2	0.2
β_{23}	-12.1	2.5	3.9	-6.1	0.8	1.1	-5.9	0.3	0.6	-4.8	0.1	0.4
β_{24}	-1.3	2.2	2.1	-0.3	0.5	0.5	-0.7	0.2	0.2	-0.3	0.1	0.1
β_{25}	-0.9	2.1	2.1	-0.1	0.4	0.4	-0.6	0.2	0.2	-0.3	0.1	0.1
β_{26}	2.9	2.2	2.2	1.2	0.5	0.5	0.8	0.2	0.2	0.8	0.1	0.1
β_{27}	0.5	2.2	2.1	-0.8	0.5	0.5	0.0	0.2	0.2	-0.3	0.1	0.1
β_{28}	-0.3	1.6	1.6	-1.0	0.5	0.5	-0.3	0.2	0.2	0.2	0.1	0.1
β_{29}	0.6	2.2	2.2	1.1	0.6	0.6	0.6	0.2	0.2	0.4	0.1	0.1
β_{210}	1.1	2.2	2.2	-0.7	0.5	0.5	0.1	0.2	0.2	0.3	0.1	0.1
β_{211}	1.9	2.1	2.1	-0.3	0.5	0.5	-0.2	0.2	0.2	0.0	0.1	0.1
β_{212}	-0.1	1.4	1.3	-0.1	0.7	0.7	0.3	0.2	0.2	0.0	0.1	0.1
β_{213}	0.2	1.8	1.8	-0.6	0.6	0.6	-0.8	0.2	0.2	-0.2	0.1	0.1
Γ_{21}	-5.0	9.1	9.3	-4.4	2.7	2.8	-3.6	1.3	1.4	-2.9	0.6	0.7
Γ_{212}	35.2	21.3	33.5	12.6	10.5	12.0	9.2	3.7	4.5	9.9	2.1	3.1
Γ_{22}	-28.0	11.6	19.3	-8.4	6.4	7.1	-0.6	1.7	1.7	2.4	0.9	1.0
Γ_{213}	10.4	7.3	8.3	3.2	1.3	1.4	1.8	0.4	0.4	1.9	0.2	0.2
Γ_{223}	2.9	3.3	3.4	1.4	1.6	1.6	1.4	0.5	0.5	1.2	0.3	0.3
Γ_{23}	4.9	1.2	1.4	5.7	0.8	1.1	5.9	0.5	0.8	5.6	0.4	0.7
Γ_{214}	-8.0	3.1	3.7	-5.0	0.9	1.2	-5.0	0.6	0.8	-3.5	0.2	0.3
Γ_{224}	2.3	1.9	1.9	-0.1	0.7	0.7	-0.5	0.5	0.5	-0.7	0.2	0.2
Γ_{234}	-0.9	0.6	0.6	-1.5	1.0	1.0	-0.8	0.6	0.6	-1.8	0.4	0.5
Γ_{24}	0.4	0.1	0.1	2.7	0.4	0.5	3.7	0.4	0.5	5.0	0.3	0.6
Γ_{215}	1.5	3.1	3.1	0.4	0.8	0.8	0.5	0.4	0.4	0.2	0.2	0.2
Γ_{225}	1.2	1.4	1.4	0.6	0.6	0.6	1.0	0.4	0.4	0.6	0.2	0.2
Γ_{235}	-1.1	0.2	0.3	0.0	0.3	0.3	-0.2	0.3	0.3	-0.7	0.2	0.2
Γ_{245}	0.2	0.0	0.0	-0.3	0.1	0.1	-0.1	0.1	0.1	-0.1	0.1	0.1
Γ_{25}	0.1	0.0	0.0	0.2	0.0	0.0	0.4	< 0.1	< 0.1	0.7	< 0.1	< 0.1

Table H–67: Simulation 12 results multiplied by 100, averaged over 100 runs. Outcome follows a binomial distribution with $K = 3$, $p = 19$, $q = 5$, $\pi_1 = 0.5$, $\pi_2 = 0.3$, $\pi_3 = 0.2$, $\boldsymbol{\beta}_1^\top = (0.95, 0.60, -0.65, -0.25, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00)$, $\boldsymbol{\beta}_2^\top = (-0.85, -0.15, -0.75, 0.10, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00)$, $\boldsymbol{\beta}_3^\top = (-0.30, -0.90, 0.80, -0.25, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00)$, $\mathbb{\Gamma}_1^{*\top} = (0.95, 0.85, 1.15, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00)$, $\mathbb{\Gamma}_2^{*\top} = (0.70, -0.70, 0.80, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00)$, and $\mathbb{\Gamma}_3^{*\top} = (1.75, 0.00, 0.85, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00)$. Small values changed to zero, part 3.

n	$n = 100$			$n = 250$			$n = 500$			$n = 1000$		
Measure	Bias	Var	MSE	Bias	Var	MSE	Bias	Var	MSE	Bias	Var	MSE
β_{30}	10.4	24.6	25.5	0.2	10.7	10.6	1.6	5.0	5.0	3.4	2.9	3.0
β_{31}	13.8	32.2	33.8	11.7	13.7	14.9	5.7	4.9	5.2	3.7	1.7	1.8
β_{32}	-41.3	15.7	32.6	-18.0	8.8	12.0	-8.1	3.9	4.5	-2.6	0.9	0.9
β_{33}	3.0	4.7	4.8	1.1	1.2	1.2	1.4	0.6	0.6	0.1	0.3	0.2
β_{34}	-5.2	5.9	6.1	-2.4	1.3	1.4	0.2	0.5	0.5	0.2	0.2	0.2
β_{35}	5.1	4.4	4.6	3.6	1.3	1.4	1.6	0.4	0.5	0.6	0.2	0.2
β_{36}	4.8	6.9	7.1	1.4	1.5	1.5	1.3	0.5	0.5	0.3	0.2	0.2
β_{37}	-1.7	4.8	4.8	-0.2	1.1	1.1	-0.4	0.4	0.4	-0.1	0.1	0.1
β_{38}	0.1	6.0	5.9	-0.1	1.3	1.3	-0.2	0.5	0.5	0.1	0.3	0.3
β_{39}	-2.6	6.8	6.8	-2.7	1.0	1.0	-0.7	0.4	0.4	-0.5	0.2	0.2
β_{310}	-1.8	4.6	4.6	0.6	1.4	1.4	0.0	0.5	0.5	0.3	0.2	0.2
β_{311}	-0.3	4.0	4.0	-1.0	1.0	1.0	-1.2	0.5	0.5	0.0	0.2	0.2
β_{312}	4.7	4.9	5.1	0.8	1.3	1.3	0.5	0.5	0.5	0.4	0.2	0.2
β_{313}	-2.8	4.1	4.1	-1.4	1.1	1.1	-0.4	0.4	0.4	0.5	0.2	0.2
Γ_{31}	-16.5	28.1	30.5	-4.2	13.8	13.8	1.2	4.9	4.9	4.1	1.9	2.1
Γ_{312}	7.4	38.3	38.5	-3.0	12.7	12.7	0.5	5.7	5.7	-0.3	2.3	2.3
Γ_{32}	-9.5	32.2	32.7	1.6	13.1	13.0	-0.3	4.0	3.9	-0.2	1.6	1.6
Γ_{313}	-14.9	20.3	22.3	-0.8	4.8	4.8	-1.8	2.2	2.3	-0.9	0.9	0.9
Γ_{323}	-0.7	13.9	13.7	-4.8	7.6	7.7	-4.1	2.6	2.8	-3.8	0.9	1.1
Γ_{33}	8.3	2.6	3.2	18.4	4.2	7.6	15.8	2.9	5.4	13.1	1.2	2.9
Γ_{314}	-0.8	9.5	9.4	-1.8	1.8	1.8	0.3	1.0	1.0	0.3	0.4	0.4
Γ_{324}	-0.9	5.3	5.3	-0.2	2.0	2.0	0.2	0.9	0.9	1.0	0.4	0.4
Γ_{334}	-1.6	0.4	0.5	-0.5	0.7	0.7	-0.6	0.6	0.6	-0.7	0.5	0.5
Γ_{34}	0.3	0.0	0.0	1.6	0.3	0.3	2.5	0.3	0.4	3.2	0.3	0.4
Γ_{315}	-6.7	7.4	7.8	-3.9	2.3	2.4	-1.9	0.9	1.0	-0.8	0.4	0.4
Γ_{325}	-2.6	4.9	4.9	-0.2	1.2	1.2	0.9	0.6	0.6	-0.1	0.3	0.3
Γ_{335}	-0.7	0.8	0.8	-0.3	0.8	0.7	-0.7	0.7	0.7	0.7	0.5	0.5
Γ_{345}	-0.3	0.0	0.0	-0.8	0.2	0.3	-0.5	0.1	0.1	0.5	0.1	0.1
Γ_{35}	0.2	0.0	0.0	1.0	0.2	0.2	0.5	< 0.1	< 0.1	1.2	0.1	0.1
Total			612.6			217.2			100.9			51.1

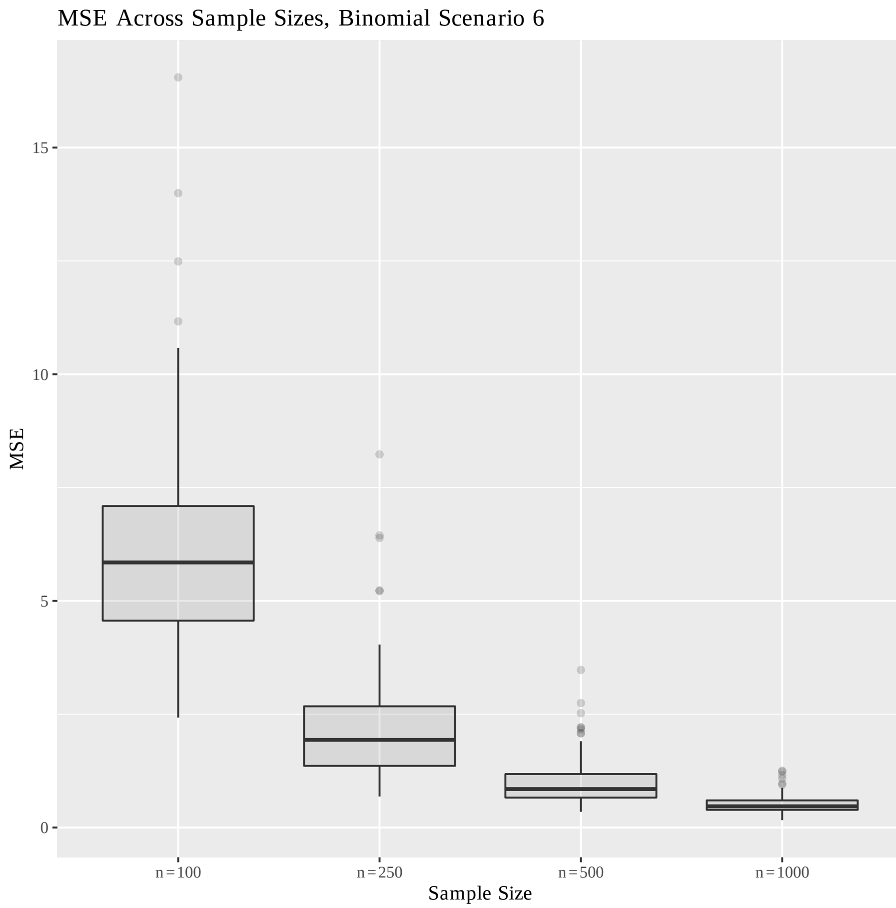

Figure H–35: MSE across sample sizes, binomial outcome, small values changed to zero, with $K = 3$, $p = 19$, $q = 5$.

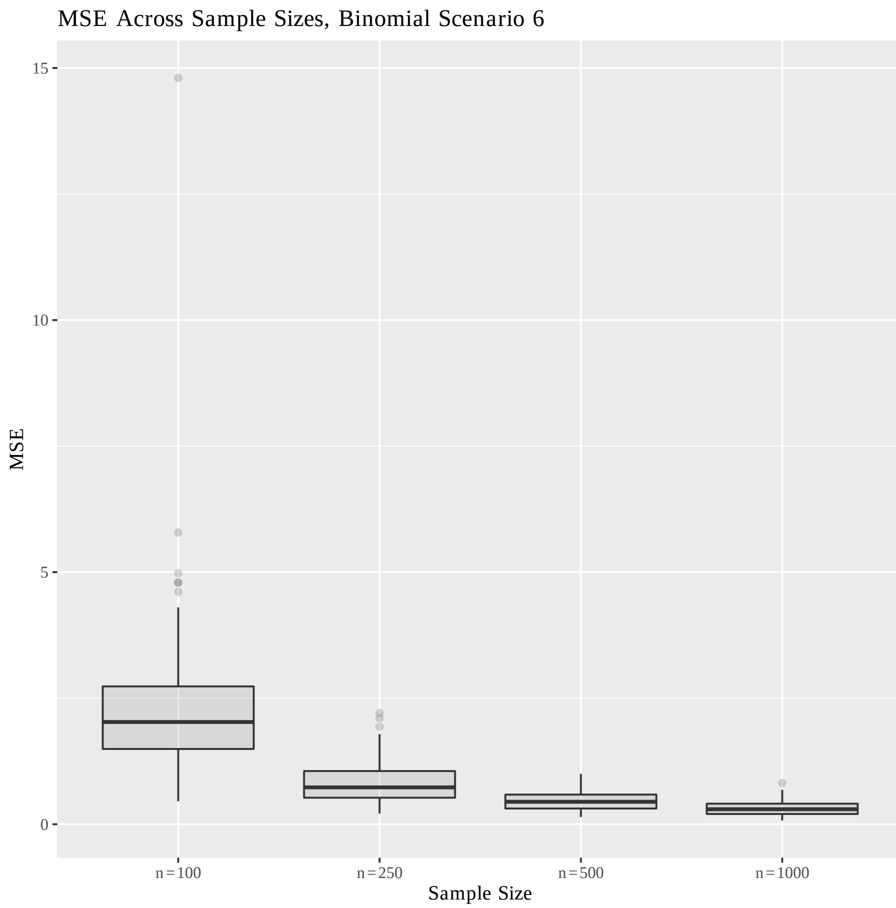

Figure H–36: MSE across sample sizes, binomial outcome, oracle model, with $K = 3$, $p = 19$, $q = 5$.

Table H–68: Simulation 12 results multiplied by 100, averaged over 100 runs. Outcome follows a binomial distribution with $K = 3$, $p = 4$, $q = 2$, $\pi_1 = 0.5$, $\pi_2 = 0.3$, $\pi_3 = 0.2$, $\boldsymbol{\beta}_1^\top = (0.95, 0.60, -0.65, -0.25)$, $\boldsymbol{\beta}_2^\top = (-0.85, -0.15, -0.75, 0.10)$, $\boldsymbol{\beta}_3^\top = (-0.30, -0.90, 0.80, -0.25)$, $\mathbb{\Gamma}_1^{*\top} = (0.95, 0.85, 1.15)$, $\mathbb{\Gamma}_2^{*\top} = (0.70, -0.70, 0.80)$, and $\mathbb{\Gamma}_3^{*\top} = (1.75, 0.00, 0.85)$. Oracle model.

n	$n = 100$			$n = 250$			$n = 500$			$n = 1000$		
Measure	Bias	Var	MSE	Bias	Var	MSE	Bias	Var	MSE	Bias	Var	MSE
π_1	-6.1	0.7	1.1	-5.8	0.3	0.6	-6.4	0.2	0.6	-6.5	0.1	0.5
π_2	6.2	0.8	1.1	6.2	0.3	0.6	6.7	0.1	0.6	6.8	0.1	0.5
β_{10}	19.6	5.3	9.1	19.7	2.0	5.9	21.1	1.1	5.6	19.6	0.9	4.8
β_{11}	18.5	11.3	14.6	16.3	3.9	6.5	18.5	1.8	5.2	18.1	1.4	4.6
β_{12}	-3.1	1.2	1.3	-2.0	0.5	0.5	-0.5	0.2	0.2	-0.3	0.1	0.1
β_{13}	-3.0	0.8	0.9	-2.5	0.3	0.3	-1.1	0.1	0.2	-0.9	0.1	0.1
Γ_{11}	-12.8	5.9	7.5	-12.3	2.2	3.7	-11.8	1.0	2.4	-11.8	0.4	1.8
Γ_{112}	-18.8	14.6	18.0	-16.6	6.1	8.8	-18.1	2.7	5.9	-17.1	1.7	4.6
Γ_{12}	-7.9	8.4	9.0	-2.0	1.7	1.7	0.7	1.0	1.0	2.0	0.5	0.6
β_{20}	16.2	5.8	8.4	9.5	2.1	3.0	10.2	1.0	2.0	8.3	0.5	1.2
β_{21}	-10.5	8.2	9.2	-8.5	3.4	4.1	-10.2	1.5	2.5	-10.3	0.9	2.0
β_{22}	-5.8	2.1	2.4	-2.2	0.5	0.6	0.1	0.2	0.2	0.4	0.1	0.1
β_{23}	-3.1	1.8	1.9	-3.2	0.5	0.6	-4.3	0.2	0.4	-4.6	0.1	0.3
Γ_{21}	-2.5	4.4	4.4	0.3	1.3	1.3	0.8	0.5	0.5	0.7	0.3	0.3
Γ_{212}	11.9	15.2	16.5	8.3	5.6	6.2	6.6	2.1	2.5	8.0	0.9	1.5
Γ_{22}	-7.8	10.0	10.5	0.0	3.9	3.9	3.4	1.3	1.4	5.0	0.6	0.9
β_{30}	2.7	19.3	19.2	-2.5	6.1	6.1	-1.8	4.3	4.3	0.0	2.1	2.1
β_{31}	-2.3	14.9	14.8	0.3	5.5	5.4	0.9	2.2	2.2	0.9	0.9	0.9
β_{32}	5.4	5.2	5.5	4.4	1.3	1.5	1.0	0.7	0.7	0.5	0.3	0.3
β_{33}	-1.3	4.9	4.9	0.7	1.0	1.0	0.8	0.5	0.5	0.0	0.2	0.2
Γ_{31}	-9.0	31.1	31.6	-5.3	8.8	9.0	-2.1	3.7	3.7	-1.1	1.7	1.7
Γ_{312}	6.2	22.5	22.7	-0.7	6.1	6.0	-1.7	2.4	2.4	-4.0	1.5	1.6
Γ_{32}	-24.3	15.0	20.8	-6.6	4.6	5.0	-2.8	2.3	2.3	-2.5	1.1	1.2
Total			235.2			82.3			47.2			31.9

APPENDIX I
Supplement to Chapter 4

I considered a number of additional possibilities for exploration of the MPLE that I have not included in Chapter 4. I explore some of these in this appendix, while others I leave to future work.

I.1 Choice of Starting Value

As previously stated, the choice of starting values when computing parameter estimates for a FinMix GLMM is important. When calculating the MPLE, the starting value that makes the most sense is the MLE. While it can be viewed as an additional step to calculate the MLE, this is a step that may already be planned. For example, the inverse of the MLE is a popular choice for the weights in ALASSO. To calculate the MLE for the simulations in 4 I generated starting values in the same way as for the simulations in Chapter 3. I also tried computing the parameter estimates for the MPLE using these randomly generated starting value rather than the MLE. I found that using these starting values resulted in a significantly longer run time to calculate the MPLE as the starting value was overall further away from the MPLE.

I.2 Hard Threshold

While LASSO, ALASSO, and SCAD are popular penalties, they are not the only possible penalties. There are many penalties that are more complex, but one simple penalty is the hard threshold (Antoniadis, 1997; Fan and Li, 2012). One desirable property of the hard threshold is that parameters are either shrunk to zero or are not shrunk at all.

The hard penalty in a model where only fixed effects are penalised takes the form

$$p_{\lambda_{nk}}(\boldsymbol{\theta}_k) = \lambda_{nk}^2 - (|\boldsymbol{\theta}_k| - \lambda_{nk})^2 I(|\boldsymbol{\theta}_k| \leq \lambda_{nk})$$

with

$$\frac{\partial p_{\lambda_{nk}}(\boldsymbol{\theta}_k)}{\partial \beta_{kh}} = -2(|\beta_{kh}| - \lambda_{nk}) I(|\beta_{kh}| \leq \lambda_{nk})$$

and

$$\frac{\partial p_{\lambda_{nk}}(\boldsymbol{\theta}_k)}{\partial \Gamma_{kh}} = 0.$$

Similarly, if a penalty is applied to both the fixed and random effects, the penalty becomes

$$\begin{aligned} p_{\lambda_{nk}}(\underline{\boldsymbol{\theta}}_k) &= \lambda_{nk}^2 - (|\underline{\boldsymbol{\theta}}_k| - \lambda_{nk})^2 I(|\underline{\boldsymbol{\theta}}_k| \leq \lambda_{nk}) \\ &= \sum_{h=1}^{p} \lambda_{nk}^2 - (|\beta_{kh}| - \lambda_{nk})^2 I(|\beta_{kh}| \leq \lambda_{nk}) \\ &\quad + \sum_{h=1}^{q} \lambda_{nk}^2 - (|d_{kh}| - \lambda_{nk})^2 I(|d_{kh}| \leq \lambda_{nk}) \end{aligned}$$

and the partial derivatives are

$$\frac{\partial p_{\lambda_{nk}}(\underline{\boldsymbol{\theta}}_k)}{\partial \beta_{kh}} = -2(|\beta_{kh}| - \lambda_{nk}) I(|\beta_{kh}| \leq \lambda_{nk})$$

and

$$\frac{\partial p_{\lambda_{nk}}(\underline{\boldsymbol{\theta}}_k)}{\partial d_{kh}} = -2(|d_{kh}| - \lambda_{nk}) I(|d_{kh}| \leq \lambda_{nk}).$$

The second derivatives of this penalty are equal to zero both with respect to β_{kh}, and d_{kh}.

Thus, the non-zero diagonal elements of $\Sigma_{\lambda_{nk}}(\underline{\boldsymbol{\theta}}_k^{(s)})$ are

$$\frac{-2 \times (|\underline{\boldsymbol{\theta}}_{kh}| - \lambda_{nk}) I(|\underline{\boldsymbol{\theta}}_{kh}| \leq \lambda_{nk})}{|\underline{\boldsymbol{\theta}}_{kh}^{(s)}| + \epsilon}$$

Due to time constraints, I did not perform simulations using this penalty.

I.3 All Values of λ_k Set to the Same Value

It would be convenient if the problem of finding an optimal value of $\boldsymbol{\lambda}$ could be simplified to finding one value, rather than a K dimensional vector of values. However, Du et al. (2013) found that this simplification did not perform well in the linear case, so there was not much hope that it would perform well in the GLMM case. Because of time constraints, I did not execute these simulations.

I.4 Exploration of the Grid of Possible Values of $\boldsymbol{\lambda}$

The choice of the tuning parameter is a difficult problem, made more difficult in the case of finite mixtures of regression models because one needs to find the optimal value of λ_k for each subpopulation. As such, an exploration of the behaviour of the BIC over different values of $\boldsymbol{\lambda}$ is of interest. Because of time constraints, I could not complete this exploration.

APPENDIX J
Calculation of Standard Errors

J.1 Calculating Standard Errors for a Finite Mixture of Generalised Linear Mixed-Effect Model

When estimating parameters in a model, it is useful to include a measure of the variability associated with a parameter estimate. Following from Louis (1982); McLachlan and Peel (2000); McLachlan and Krishnan (2008); Khalili and Vidyashankar (2018), I estimated the covariance matrix of the MLE, and then extracted the standard errors from that matrix. This estimation of this matrix is based on the second derivative of the complete data log-likelihood. This method is popular for calculating the standard errors when the EM algorithm is used. While this approach is not specific to finite mixtures of regression models, it is the standard approach to calculating standard errors for finite mixture models, and was used in Du et al. (2013). More specifically, as described in Khalili and Vidyashankar (2018), the standard errors are the square roots of the diagonal elements of the inverse of the empirical observed information matrix. The empirical observed information matrix is used to approximate the observed information matrix. Note that I calculate standard errors for the mixing parameters and the parameters that correspond to fixed effects but not on parameters that correspond to random effects. The first step was to calculate the gradient of the complete log-likelihood, I showed in Appendix D that these derivatives exist. I represented the gradient of the complete log-likelihood as

$s(\mathbf{y}_i|\mathbb{X}_i, \mathbb{Z}_i, \boldsymbol{\Theta})$ with

$$s(\mathbf{y}_i|\mathbb{X}_i, \mathbb{Z}_i, \boldsymbol{\Theta}) = \sum_{k=1}^{K} \frac{\partial}{\partial \tilde{\boldsymbol{\Theta}}} \tau_{ki} \log[\pi_k f_{\mathbf{y}_i}^{(k)}(\mathbf{y}_i|\mathbb{X}_i, \mathbb{Z}_i, \boldsymbol{\theta}_k)].$$

Next, I calculated the empirical observed information matrix

$$I_\ell(\boldsymbol{\Theta}) = \sum_{i=1}^{n} s(\mathbf{y}_i|\mathbb{X}_i, \mathbb{Z}_i, \boldsymbol{\Theta})[s(\mathbf{y}_i|\mathbb{X}_i, \mathbb{Z}_i, \boldsymbol{\Theta})]^\top.$$

I then calculated the inverse of the empirical observed information matrix, $I_\ell^{-1}(\boldsymbol{\Theta})$. The standard errors are the square roots of the diagonal elements of the inverse of the empirical observed information matrix.

As I did in Appendix B, I show here the details of this derivative for each element of $\boldsymbol{\Theta}$. As I described in the estimation of $\boldsymbol{\Theta}$, the values $\mathbf{b}_i$ are unknown. As such, I generated L potential values for $\mathbf{b}_i$ from a q-dimensional multivariate standard Gaussian distribution. These values, denoted $\mathbf{b}_i^{(l)}$, were used to approximate the integral over $\mathbf{b}_i$, and calculate $f_{\mathbf{y}_i}^{(k)}(\mathbf{y}_i|\mathbb{X}_i, \mathbb{Z}_i, \boldsymbol{\theta}_k)$. Recall that

$$\begin{aligned}
s(\mathbf{y}_i|\mathbb{X}_i, \mathbb{Z}_i, \boldsymbol{\Theta}) &= \sum_{k=1}^{K} \frac{\partial}{\partial \tilde{\boldsymbol{\Theta}}} \tau_{ki} \log[\pi_k f_{\mathbf{y}_i}^{(k)}(\mathbf{y}_i|\mathbb{X}_i, \mathbb{Z}_i, \boldsymbol{\theta}_k)] \\
&= \sum_{k=1}^{K} \tau_{ki} \frac{\partial}{\partial \tilde{\boldsymbol{\Theta}}} \left\{ \log[\pi_k] + \log[f_{\mathbf{y}_i}^{(k)}(\mathbf{y}_i|\mathbb{X}_i, \mathbb{Z}_i, \boldsymbol{\theta}_k)] \right\} \\
&= \sum_{k=1}^{K} \tau_{ki} \frac{\partial}{\partial \tilde{\boldsymbol{\Theta}}} \log(\pi_k) + \sum_{k=1}^{K} \tau_{ki} \frac{\partial}{\partial \tilde{\boldsymbol{\Theta}}} \log[f_{\mathbf{y}_i}^{(k)}(\mathbf{y}_i|\mathbb{X}_i, \mathbb{Z}_i, \boldsymbol{\theta}_k)].
\end{aligned}$$

Focusing on $\frac{\partial}{\partial \tilde{\boldsymbol{\Theta}}} \log(\pi_k)$, there are three cases to consider, $\frac{\partial}{\partial \pi_h} \log(\pi_k) = 0, \forall h \neq k$, $\frac{\partial}{\partial \boldsymbol{\beta}_h} \log(\pi_k) = 0, \forall h \forall k$, and $\frac{\partial}{\partial \pi_k} \log(\pi_k) = \frac{1}{\pi_k}$. Thus,

$$\sum_{k=1}^{K} \tau_{ki} \frac{\partial}{\partial \tilde{\boldsymbol{\Theta}}} \log(\pi_k) = \begin{bmatrix} \frac{\tau_{1i}}{\pi_1} \\ \frac{\tau_{2i}}{\pi_2} \\ \vdots \\ \frac{\tau_{(K-1)i}}{\pi_{K-1}} \\ 0 \\ \vdots \\ 0 \end{bmatrix}.$$

Considering next the second part of this equation, $\frac{\partial}{\partial \tilde{\boldsymbol{\Theta}}} f_{\mathbf{y}_i}^{(k)}(\mathbf{y}_i | \mathbb{X}_i, \mathbb{Z}_i, \boldsymbol{\theta}_k)$. Again, there are three cases to consider. First, $\frac{\partial}{\partial \pi_h} f_{\mathbf{y}_i}^{(k)}(\mathbf{y}_i | \mathbb{X}_i, \mathbb{Z}_i, \boldsymbol{\theta}_k) = 0, \forall h \forall k$, second, $\frac{\partial}{\partial \boldsymbol{\beta}_h} f_{\mathbf{y}_i}^{(k)}(\mathbf{y}_i | \mathbb{X}_i, \mathbb{Z}_i, \boldsymbol{\beta}_k, \mathbb{\Gamma}_k) = 0, \forall h \neq k$. The last case to consider is $\frac{\partial}{\partial \boldsymbol{\beta}_k} f_{\mathbf{y}_i}^{(k)}(\mathbf{y}_i | \mathbb{X}_i, \mathbb{Z}_i, \boldsymbol{\beta}_k, \mathbb{\Gamma}_k)$. Because a GLMM is being considered,

$$\begin{aligned} f_{\mathbf{y}_i}^{(k)}(\mathbf{y}_i | \mathbb{X}_i, \mathbb{Z}_i, \boldsymbol{\beta}_k, \mathbb{\Gamma}_k) &= \int f_{\mathbf{y}_i|\mathbf{b}_i}^{(k)}(\mathbf{y}_i | \mathbb{X}_i, \mathbb{Z}_i, \mathbf{b}_i, \boldsymbol{\beta}_k, \mathbb{\Gamma}_k) f(\mathbf{b}_i) d\mathbf{b}_i \\ &= \int \prod_{j=1}^{n_i} f_{y_{ij}|\mathbf{b}_i}^{(k)}(y_{ij} | \mathbf{x}_{ij}, \mathbf{z}_{ij}, \mathbf{b}_i, \boldsymbol{\beta}_k, \mathbb{\Gamma}_k) f(\mathbf{b}_i) d\mathbf{b}_i. \end{aligned}$$

I approximated this integral with

$$f_{\mathbf{y}_i}^{(k)}(\mathbf{y}_i | \mathbb{X}_i, \mathbb{Z}_i, \boldsymbol{\beta}_k, \mathbb{\Gamma}_k) \approx \frac{1}{L} \sum_{l=1}^{L} \prod_{j=1}^{n_i} f_{y_{ij}|\mathbf{b}_i}^{(k)}(y_{ij} | \mathbf{x}_{ij}, \mathbf{z}_{ij}, \mathbf{b}_i^{(l)}, \boldsymbol{\beta}_k, \mathbb{\Gamma}_k).$$

Therefore,

$$\begin{aligned}
\frac{\frac{\partial}{\partial \boldsymbol{\beta}_k} f_{\mathbf{y}_i}^{(k)}(\mathbf{y}_i|\mathbb{X}_i, \mathbb{Z}_i, \boldsymbol{\beta}_k, \mathbb{F}_k)}{f_{\mathbf{y}_i}^{(k)}(\mathbf{y}_i|\mathbb{X}_i, \mathbb{Z}_i, \boldsymbol{\beta}_k, \mathbb{F}_k)} &\approx \frac{\frac{\partial}{\partial \boldsymbol{\beta}_k} \frac{1}{L} \sum_{l=1}^{L} \prod_{j=1}^{n_i} f_{y_{ij}|\mathbf{b}_i}^{(k)}(y_{ij}|\mathbf{x}_{ij}, \mathbf{z}_{ij}, \mathbf{b}_i^{(l)}, \boldsymbol{\beta}_k, \mathbb{F}_k)}{\frac{1}{L} \sum_{l=1}^{L} \prod_{j=1}^{n_i} f_{y_{ij}|\mathbf{b}_i}^{(k)}(y_{ij}|\mathbf{x}_{ij}, \mathbf{z}_{ij}, \mathbf{b}_i^{(l)}, \boldsymbol{\beta}_k, \mathbb{F}_k)} \\
&= \frac{\frac{\partial}{\partial \boldsymbol{\beta}_k} \sum_{l=1}^{L} \prod_{j=1}^{n_i} f_{y_{ij}|\mathbf{b}_i}^{(k)}(y_{ij}|\mathbf{x}_{ij}, \mathbf{z}_{ij}, \mathbf{b}_i^{(l)}, \boldsymbol{\beta}_k, \mathbb{F}_k)}{\sum_{l=1}^{L} \prod_{j=1}^{n_i} f_{y_{ij}|\mathbf{b}_i}^{(k)}(y_{ij}|\mathbf{x}_{ij}, \mathbf{z}_{ij}, \mathbf{b}_i^{(l)}, \boldsymbol{\beta}_k, \mathbb{F}_k)} \\
&= \frac{\sum_{l=1}^{L} \frac{\partial}{\partial \boldsymbol{\beta}_k} \prod_{j=1}^{n_i} f_{y_{ij}|\mathbf{b}_i}^{(k)}(y_{ij}|\mathbf{x}_{ij}, \mathbf{z}_{ij}, \mathbf{b}_i^{(l)}, \boldsymbol{\beta}_k, \mathbb{F}_k)}{\sum_{l=1}^{L} \prod_{j=1}^{n_i} f_{y_{ij}|\mathbf{b}_i}^{(k)}(y_{ij}|\mathbf{x}_{ij}, \mathbf{z}_{ij}, \mathbf{b}_i^{(l)}, \boldsymbol{\beta}_k, \mathbb{F}_k)}.
\end{aligned}$$

Using the product rule for more than two factors,

$$\frac{\partial}{\partial \boldsymbol{\beta}_k} \prod_{j=1}^{n_i} f_{y_{ij}|\mathbf{b}_i}^{(k)}(y_{ij}|\mathbf{x}_{ij}, \mathbf{z}_{ij}, \mathbf{b}_i^{(l)}, \boldsymbol{\beta}_k, \mathbb{F}_k) =$$

$$\left(\prod_{j=1}^{n_i} f_{y_{ij}|\mathbf{b}_i}^{(k)}(y_{ij}|\mathbf{x}_{ij}, \mathbf{z}_{ij}, \mathbf{b}_i^{(l)}, \boldsymbol{\beta}_k, \mathbb{F}_k) \right) \times \left(\sum_{j=1}^{n_i} \frac{\frac{\partial}{\partial \boldsymbol{\beta}_k} f_{y_{ij}|\mathbf{b}_i}^{(k)}(y_{ij}|\mathbf{x}_{ij}, \mathbf{z}_{ij}, \mathbf{b}_i^{(l)}, \boldsymbol{\beta}_k, \mathbb{F}_k)}{f_{y_{ij}|\mathbf{b}_i}^{(k)}(y_{ij}|\mathbf{x}_{ij}, \mathbf{z}_{ij}, \mathbf{b}_i^{(l)}, \boldsymbol{\beta}_k, \mathbb{F}_k)} \right).$$

In the case of a Poisson outcome,

$$\frac{\frac{\partial}{\partial \boldsymbol{\beta}_k} f_{y_{ij}|\mathbf{b}_i}^{(k)}(y_{ij}|\mathbf{x}_{ij}, \mathbf{z}_{ij}, \mathbf{b}_i^{(l)}, \boldsymbol{\beta}_k, \mathbb{F}_k)}{f_{y_{ij}|\mathbf{b}_i}^{(k)}(y_{ij}|\mathbf{x}_{ij}, \mathbf{z}_{ij}, \mathbf{b}_i^{(l)}, \boldsymbol{\beta}_k, \mathbb{F}_k)} = \begin{bmatrix} x_{ij1} \\ x_{ij2} \\ \vdots \\ x_{ijp} \end{bmatrix} (y_{ij} - \exp(\mathbf{x}_{ij}\boldsymbol{\beta}_k + \mathbf{z}_{ij}\mathbb{F}_k \mathbf{b}_i^{(l)})).$$

When a binomial outcome is considered,

$$\frac{\frac{\partial}{\partial \boldsymbol{\beta}_k} f_{y_{ij}|\mathbf{b}_i}^{(k)}(y_{ij}|\mathbf{x}_{ij}, \mathbf{z}_{ij}, \mathbf{b}_i^{(l)}, \boldsymbol{\beta}_k, \mathbb{F}_k)}{f_{y_{ij}|\mathbf{b}_i}^{(k)}(y_{ij}|\mathbf{x}_{ij}, \mathbf{z}_{ij}, \mathbf{b}_i^{(l)}, \boldsymbol{\beta}_k, \mathbb{F}_k)} = \begin{bmatrix} x_{ij1} \\ x_{ij2} \\ \vdots \\ x_{ijp} \end{bmatrix} \frac{y_{ij} + (y_{ij} - m_{ij}) \exp(\mathbf{x}_{ij}\boldsymbol{\beta}_k + \mathbf{z}_{ij}\mathbb{F}_k \mathbf{b}_i^{(l)})}{1 + \exp(\mathbf{x}_{ij}\boldsymbol{\beta}_k + \mathbf{z}_{ij}\mathbb{F}_k \mathbf{b}_i^{(l)})}.$$

Putting all of this together, I calculated $s(\mathbf{y}_i|\mathbb{X}_i, \mathbb{Z}_i, \mathbf{\Theta})$, and therefore the standard errors. Once I calculated the standard errors, following from Khalili and Vidyashankar (2018), I computed the t-statistics with

$$t_{kh} = \frac{\beta_{kh}}{SE(\beta_{kh})}$$

or

$$t_k = \frac{\pi_k}{SE(\pi_k)}$$

and the p-value as $2 * P(t_n < -|t_{kh}|)$ or $2 * P(t_n < -|t_k|)$.

REFERENCES

Abad, A. A., Litiere, S., and Molenberghs, G. (2010). Testing for misspecification in generalized linear mixed models. *Biostatistics*, 11(4):771–786.

Akaike, H. (1998). Information theory and an extension of the maximum likelihood principle. In *Selected Papers of Hirotugu Akaike*, pages 199–213. Springer.

Aletaha, D. and Smolen, J. S. (2006). The definition and measurement of disease modification in inflammatory rheumatic diseases. *Rheumatic diseases clinics of North America*, 32(1):9–44.

Algamal, Z. Y. and Lee, M. H. (2015). Penalized logistic regression with the adaptive LASSO for gene selection in high-dimensional cancer classification. *Expert Systems with Applications*, 42(23):9326–9332.

Alonso, A., Litiere, S., and Molenberghs, G. (2008). A family of tests to detect misspecifications in the random-effects structure of generalized linear mixed models. *Computational Statistics and Data Analysis*, 52(9):4474–4486.

Anderson, D. A. and Hinde, J. P. (1988). Random Effects in Generalized Linear-Models and the EM Algorithm. *Communications in Statistics-Theory and Methods*, 17(11):3847–3856.

Antoniadis, A. (1997). Wavelets in statistics: a review. *Journal of the Italian Statistical Society*, 6(2):97–144.

Atienza, N., Garcia-Heras, J., and Munoz-Pichardo, J. (2006). A new condition for identifiability of finite mixture distributions. *Metrika*, 63(2):215–221.

Bai, X. Q., Chen, K., and Yao, W. X. (2016). Mixture of linear mixed models using multivariate t distribution. *Journal of Statistical Computation and Simulation*, 86(4):771–787.

Bao, J. S. and Hanson, T. E. (2016). A mean-constrained finite mixture of normals model. *Statistics and Probability Letters*, 117(C):93–99.

Baradaran, N., Tan, S. N., Liu, A., Ashoori, A., Palmer, S. J., Wang, Z. J., Oishi, M. M., and McKeown, M. J. (2013). Parkinson's disease rigidity: relation to brain connectivity and motor performance. *Frontiers in Neurology*, 4:67.

Benedetti, A., Platt, R., and Atherton, J. (2014). Generalized Linear Mixed Models for Binary Data: Are Matching Results from Penalized Quasi-Likelihood and Numerical Integration Less Biased? *PLoS One*, 9(1):e84601.

Bhattacharya, S. and McNicholas, P. D. (2014). A lasso-penalized bic for mixture model selection. *Advances in Data Analysis and Classification*, 8(1):45–61.

Bishop, Y. M., Fienberg, S. E., and Holland, P. W. (2007). *Discrete Multivariate Analysis: Theory and Practice*. Springer Science & Business Media.

Bondell, H. D., Krishna, A., and Ghosh, S. K. (2010). Joint Variable Selection for Fixed and Random Effects in Linear Mixed-Effects Models. *Biometrics*, 66(4):1069–1077.

Booth, J. G. and Hobert, J. P. (1999). Maximizing generalized linear mixed model likelihoods with an automated Monte Carlo EM algorithm. *Journal of the Royal Statistical Society, Series B*, 61(1):265–285.

Braun, J., Bove, D. S., and Held, L. (2014). Choice of generalized linear mixed models using predictive crossvalidation. *Computational Statistics and Data Analysis*, 75:190–202.

Breiman, L. (1996). Heuristics of instability and stabilization in model selection. *The annals of statistics*, 24(6):2350–2383.

Buhlmann, P. and van de Geer, S. A. (2011). *Statistics for High-Dimensional Data: Methods, Theory and Applications*. Springer series in statistics. Springer.

Burton, P. R. (2003). Correcting for nonrandom ascertainment in generalized linear mixed models (glmms), fitted using gibbs sampling. *Genetic Epidemiology*, 24(1):24–35.

Cao, J. and Yao, W. (2012). Semiparametric Mixture of Binomial Regression with a Degenerate Component. *Statistica Sinica*, 22(1):27–46.

Casella, G. and Berger, R. L. (2002). *Statistical Inference*, volume 2. Duxbury Pacific Grove, CA.

Chan, J. S. K., Kuk, A. Y. C., and Yam, C. H. K. (2005). Monte Carlo approximation through Gibbs output in generalized linear mixed models. *Journal of Multivariate Analysis*, 94(2):300–312.

Chatterjee, A., Gupta, S., and Lahiri, S. N. (2015). On the residual empirical process based on the ALASSO in high dimensions and its functional oracle property. *Journal of Econometrics*, 186(2):317–324.

Chatterjee, A. and Lahiri, S. N. (2013). Rates of Convergence of the Adaptive Lasso Estimators to the Oracle Distribution and Higher Order Refinements by the Bootstrap. *Annals of Statistics*, 41(3):1232–1259.

Chen, J. H. (2017). Consistency of the MLE under Mixture Models. *Statistical Science*, 32(1):47–63.

Chen, J. H. and Khalili, A. (2008). Order Selection in Finite Mixture Models With a Nonsmooth Penalty. *Journal of the American Statistical Association*, 103(484):1674–1683.

Chen, M. H., Ibrahim, J. G., Shao, Q. M., and Weiss, R. E. (2003). Prior elicitation for model selection and estimation in generalized linear mixed models. *Journal of Statistical Planning and Inference*, 111(1-2):57–76.

Chen, S. F., Yeh, F., Lin, J., Matsuguchi, T., Blackburn, E., Lee, E. T., Howard, B. V., and Zhao, J. Y. (2014). Short leukocyte telomere length is associated with obesity in American Indians: The strong heart family study. *Aging*, 6(5):380–389.

Chen, Z. and Dunson, D. B. (2003). Random effects selection in linear mixed models. *Biometrics*, 59(4):762–769.

Choi, H. and Park, C. (2012). Approximate penalization path for smoothly clipped absolute deviation. *Journal of Statistical Computation and Simulation*, 82(5):643–652.

Chopra, A. and Lian, H. (2010). Total variation, adaptive total variation and nonconvex smoothly clipped absolute deviation penalty for denoising blocky images. *Pattern Recognition*, 43(8):2609–2619.

Christensen, O. F., Roberts, G. O., and Skold, M. (2006). Robust Markov chain Monte Carlo methods for spatial generalized linear mixed models. *Journal of Computational and Graphical Statistics*, 15(1):1–17.

Chu, T. J., Zhu, J., and Wang, H. N. (2011). Penalized Maximum Likelihood Estimation and Variable Selection in Geostatistics. *Annals of Statistics*, 39(5):2607–2625.

Cox, D. R. and Wong, M. Y. (2010). A note on the sensitivity to assumptions of a generalized linear mixed model. *Biometrika*, 97(1):209–214.

Dale, J., Paterson, C., Tierney, A., Ralston, S. H., Reid, D. M., Basu, N., Harvie, J., McKay, N. D., Saunders, S., Wilson, H., et al. (2016). The Scottish Early Rheumatoid Arthritis (SERA) Study: an inception cohort and biobank. *BMC Musculoskeletal Disorders*, 17(1):1–8.

Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977). Maximum Likelihood from Incomplete Data Via EM Algorithm. *Journal of the Royal Statistical Society, Series B*, 39(1):1–38.

Deng, W. P., Chen, H. F., and Li, Z. H. (2006). A logistic regression mixture model for interval mapping of genetic trait loci affecting binary phenotypes. *Genetics*, 172(2):1349–1358.

Devijver, E. (2015). Finite mixture regression: A sparse variable selection by model selection for clustering. *Electronic Journal of Statistics*, 9(2):2642–2674.

Du, Y. T., Khalili, A., Neslehova, J. G., and 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.

Dunson, D. B. (2000). Bayesian latent variable models for clustered mixed outcomes. *Journal of the Royal Statistical Society, Series B*, 62(2):355–366.

Efron, B., Hastie, T., Johnstone, I., and Tibshirani, R. (2004). Least angle regression. *Annals of Statistics*, 32(2):407–451.

Eskandari, F. and Ormoz, E. (2016). Finite Mixture of Generalized Semiparametric Models: Variable Selection via Penalized Estimation. *Communications in Statistics-Simulation and Computation*, 45(10):3744–3759.

Evans, B. A., Feng, Z. D., and Peterson, A. V. (2001). A comparison of generalized linear mixed model procedures with estimating equations for variance and covariance parameter estimation in longitudinal studies and group randomized trials. *Statistics in Medicine*, 20(22):3353–3373.

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

Fan, Y., Leslie, D. S., and Wand, M. P. (2008). Generalised linear mixed model analysis via sequential Monte Carlo sampling. *Electronic Journal of Statistics*, 2:916–938.

Fan, Y. Y. and Li, R. Z. (2012). Variable Selection in Linear Mixed Effects Models. *Annals of Statistics*, 40(4):2043–2068.

Farcomeni, A. (2015). Generalized Linear Mixed Models Based on Latent Markov Heterogeneity Structures. *Scandinavian Journal of Statistics*, 42(4):1127–1135.

Felson, D. T., Anderson, J. J., Boers, M., Bombardier, C., Chernoff, M., Fried, B., Furst, D., Goldsmith, C., Kieszak, S., Lightfoot, R., et al. (1993). The American College of Rheumatology preliminary core set of disease activity measures for rheumatoid arthritis clinical trials. *Arthritis & Rheumatism: Official Journal of* *the American College of Rheumatology*, 36(6):729–740.

Feng, Z. D., McLerran, D., and Grizzle, J. (1996). A comparison of statistical methods for clustered data analysis with Gaussian error. *Statistics in Medicine*, 15(16):1793–1806.

Fieuws, S., Verbeke, G., and Mollenberghs, G. (2007). Random-effects models for multivariate repeated measures. *Statistical Methods in Medical Research*, 16(5):387–397.

Fort, G., Moulines, E., et al. (2003). Convergence of the Monte Carlo expectation maximization for curved exponential families. *The Annals of Statistics*, 31(4):1220–1259.

Foster, S. D., Verbyla, A. P., and Pitchford, W. S. (2008). A random model approach for the LASSO. *Computational Statistics*, 23(2):217–233.

Fujii, Y., Kaneko, S., Nzou, S. M., Mwau, M., Njenga, S. M., Tanigawa, C., Kimotho, J., Mwangi, A. W., Kiche, I., Matsumoto, S., Niki, M., Osada-Oka, M., Ichinose, Y., Inoue, M., Itoh, M., Tachibana, H., Ishii, K., Tsuboi, T., Yoshida, L. M., Mondal, D., Haque, R., Hamano, S., Changoma, M., Hoshi, T., Kamo, K., Karama, M., Miura, M., and Hirayama, K. (2014). Serological Surveillance Development for Tropical Infectious Diseases Using Simultaneous Microsphere-Based Multiplex Assays and Finite Mixture Models. *PLoS Neglected Tropical Diseases*, 8(7):e3040.

Gamerman, D. (1997). Sampling from the posterior distribution in generalized linear mixed models. *Statistics and Computing*, 7(1):57–68.

Ganguli, B., Sen Roy, S., Naskar, M., Malloy, E. J., and Eisen, E. A. (2016). Deletion diagnostics for the generalised linear mixed model with independent random effects.

Statistics in Medicine, 35(9):1488–1501.

Gardiner, J. C., Luo, Z. H., and Roman, L. A. (2009). Fixed effects, random effects and GEE: What are the differences? *Statistics in Medicine*, 28(2):221–239.

Ghosh, D. and Chinnaiyan, A. M. (2005). Classification and selection of biomarkers in genomic data using LASSO. *Journal of Biomedicine and Biotechnology*, 2005(2):147–154.

Gilks, W. R. and Wild, P. (1992). Adaptive Rejection Sampling for Gibbs sampling. *Journal of the Royal Statistical Society, Series C*, 41(2):337–348.

Groll, A. and Tutz, G. (2014). Variable selection for generalized linear mixed models by L 1-penalized estimation. *Statistics and Computing*, 24(2):137–154.

Grun, B. and Leisch, F. (2008). Identifiability of Finite Mixtures of Multinomial Logit Models with Varying and Fixed Effects. *Journal of Classification*, 25(2):225–247.

Gueorguieva, R. (2001). A multivariate generalized linear mixed model for joint modelling of clustered outcomes in the exponential family. *Statistical Modelling*, 1(3):177–193.

Guha, S. (2008). Posterior simulation in the generalized linear mixed model with semiparametric random effects. *Journal of Computational and Graphical Statistics*, 17(2):410–425.

Guo, P., Zeng, F., Hu, X., Zhang, D., Zhu, S., Deng, Y., and Hao, Y. (2015). Improved Variable Selection Algorithm Using a LASSO-Type Penalty, with an Application to Assessing Hepatitis B Infection Relevant Factors in Community Residents. *PLoS One*, 10(7):e0134151.

Guo, Q., Wang, Y., Xu, D., Nossent, J., Pavlos, N. J., and Xu, J. (2018). Rheumatoid arthritis: Pathological mechanisms and modern pharmacologic therapies. *Bone Research*, 6(1):1–14.

Haem, E., Harling, K., Ayatollahi, S. M., Zare, N., and Karlsson, M. O. (2017). Adjusted adaptive Lasso for covariate model-building in nonlinear mixed-effect pharmacokinetic models. *Journal of Pharmacokinetics and Pharmacodynamics*, 44(1):55–66.

Hall, D. B. and Wang, L. H. (2005). Two-component mixtures of generalized linear mixed effects models for cluster correlated data. *Statistical Modelling*, 5(1):21–37.

Hallquist, M. N. and Pilkonis, P. A. (2012). Refining the Phenotype of Borderline Personality Disorder: Diagnostic Criteria and Beyond. *Personality Disorders-Theory Research and Treatment*, 3(3):228–246.

Hans, C. (2009). Bayesian lasso regression. *Biometrika*, 96(4):835–845.

Hazewinkel, M. (1988). *Encyclopaedia of Mathematics: An Updated and Annotated Translation of the Soviet "Mathematical Encyclopaedia"*. Reidel.

He, Z. D., Tu, W. Z., Wang, S. J., Fu, H. D., and Yu, Z. S. (2015). Simultaneous Variable Selection for Joint Models of Longitudinal and Survival Outcomes. *Biometrics*, 71(1):178–187.

Heagerty, P. J. and Kurland, B. F. (2001). Misspecified maximum likelihood estimates and generalised linear mixed models. *Biometrika*, 88(4):973–985.

Hebiri, M. and van de Geer, S. (2011). The Smooth-Lasso and other $\ell(1) + \ell(2)$-penalized methods. *Electronic Journal of Statistics*, 5:1184–1226.

Heggeseth, B. C. and Jewell, N. P. (2013). The impact of covariance misspecification in multivariate Gaussian mixtures on estimation and inference: an application to longitudinal modeling. *Statistics in Medicine*, 32(16):2790–2803.

Hennig, C. (2000). Identifiability of models for clusterwise linear regression. *Journal of Classification*, 17(2):273–296.

Hirose, K. and Konishi, S. (2012). Variable selection via the weighted group lasso for factor analysis models. *Canadian Journal of Statistics*, 40(2):345–361.

Hoerl, A. E. and Kennard, R. W. (1970). Ridge Regression-Biased Estimation for Nonorthogonal Problems. *Technometrics*, 12(1):55–67.

Honors, M. A., Harnack, L. J., Zhou, X., and Steffen, L. M. (2014). Trends in fatty acid intake of adults in the Minneapolis-St Paul, MN Metropolitan Area, 1980-1982 through 2007-2009. *Journal of the American Hearth Association*, 3(5):e001023.

Huang, J., Ma, S. G., and Zhang, C. H. (2008). Adaptive Lasso for Sparse High-Dimensional Regression Models. *Statistica Sinica*, 18(4):1603–1618.

Huang, M. and Yao, W. X. (2012). Mixture of Regression Models With Varying Mixing Proportions: A Semiparametric Approach. *Journal of the American Statistical Association*, 107(498):711–724.

Huang, M. A., Li, R. Z., and Wang, S. L. (2013). Nonparametric Mixture of Regression Models. *Journal of the American Statistical Association*, 108(503):929–941.

Huang, X. Z. (2009). Diagnosis of Random-Effect Model Misspecification in Generalized Linear Mixed Models for Binary Response. *Biometrics*, 65(2):361–368.

Hui, F. K. C., Muller, S., and Welsh, A. H. (2017). Hierarchical Selection of Fixed and Random Effects in Generalized Linear Mixed Models. *Statistica Sinica*, 27(2):501–518.

Hui, F. K. C., Warton, D. I., and Foster, S. D. (2015). Tuning Parameter Selection for the Adaptive Lasso Using ERIC. *Journal of the American Statistical Association*, 110(509):262–269.

Hunsberger, S., Albert, P. S., and London, W. B. (2009). A finite mixture survival model to characterize risk groups of neuroblastoma. *Statistics in Medicine*, 28(8):1301–1314.

Hunter, D. R. and Li, R. (2005). Variable selection using MM algorithms. *Annals of Statistics*, 33(4):1617–1642.

Hunter, D. R. and Young, D. S. (2012). Semiparametric mixtures of regressions. *Journal of Nonparametric Statistics*, 24(1):19–38.

Huo, X. M. and Ni, X. L. (2007). When do stepwise algorithms meet subset selection criteria? *Annals of Statistics*, 35(2):870–887.

Hussami, N. and Tibshirani, R. J. (2015). A component lasso. *Canadian Journal of Statistics*, 43(4):624–646.

Ibrahim, J. G., Chen, M. H., and Lipsitz, S. R. (1999). Monte Carlo EM for missing covariates in parametric regression models. *Biometrics*, 55(2):591–596.

Ibrahim, J. G., Zhu, H. T., Garcia, R. I., and Guo, R. X. (2011). Fixed and Random Effects Selection in Mixed Effects Models. *Biometrics*, 67(2):495–503.

Ivanoff, S., Picard, F., and Rivoirard, V. (2016). Adaptive Lasso and group-Lasso for functional Poisson regression. *Journal of Machine Learning Research*, 17(1):1–46.

Jacobs, R. A., Jordan, M. I., Nowlan, S. J., and Hinton, G. E. (1991). Adaptive mixtures of local experts. *Neural computation*, 3(1):79–87.

Jansen, R. C. (1993). Maximum-Likelihood in a Generalized Linear Finite Mixture Model by Using the EM Algorithm. *Biometrics*, 49(1):227–231.

Jeon, M. and Rabe-Hesketh, S. (2012). Profile-Likelihood Approach for Estimating Generalized Linear Mixed Models With Factor Structures. *Journal of Educational and Behavioral Statistics*, 37(4):518–542.

Jiang, J. (2007). *Linear and Generalized Linear Mixed Models and Their Applications*. Springer series in statistics. Springer.

Jiang, J. M. and Zhang, W. H. (2001). Robust estimation in generalised linear mixed models. *Biometrika*, 88(3):753–765.

Kasahara, H. and Shimotsu, K. (2015). Testing the Number of Components in Normal Mixture Regression Models. *Journal of the American Statistical Association*, 110(512):1632–1645.

Kendall, M. G., Stuart, A., Ord, J. K., Arnold, S. F., and O'Hagan, A. (1994). *Kendall's Advanced Theory of Statistics.* Edward Arnold, London.

Khalili, A. (2010). New estimation and feature selection methods in mixture-of-experts models. *Canadian Journal of Statistics*, 38(4):519–539.

Khalili, A. and Chen, J. H. (2007). Variable selection in finite mixture of regression models. *Journal of the American Statistical Association*, 102(479):1025–1038.

Khalili, A. and Vidyashankar, A. N. (2018). Hypothesis testing in finite mixture of regressions: Sparsity and model selection uncertainty. *Canadian Journal of Statistics*, 46(3):429–457.

Kim, M., Vermunt, J., Bakk, Z., Jaki, T., and Van Horn, M. L. (2016). Modeling Predictors of Latent Classes in Regression Mixture Models. *Structural Equation Modeling A*, 23(4):601–614.

Kizilkaya, K. and Tempelman, R. J. (2005). A general approach to mixed effects modeling of residual variances in generalized linear mixed models. *Genetics Selection Evolution*, 37(1):31–56.

Kuk, A. Y. C. (1999). Laplace importance sampling for generalized linear mixed models. *Journal of Statistical Computation and Simulation*, 63(2):143–158.

Kwon, S., Choi, H., and Kim, Y. (2011). Quadratic approximation on SCAD penalized estimation. *Computational Statistics and Data Analysis*, 55(1):421–428.

Kwon, S., Lee, S., and Kim, Y. (2015). Moderately clipped LASSO. *Computational Statistics and Data Analysis*, 92:53–67.

Labouriau, R. (2014). A note on the identifiability of generalized linear mixed models. *arXiv preprint arXiv:1405.0673*.

Laird, N., Lange, N., and Stram, D. (1987). Maximum-Likelihood Computations with Repeated Measures - Application of the EM-Algorithm. *Journal of the American Statistical Association*, 82(397):97–105.

Laird, N. M. (1982). Computation of Variance-Components Using the EM Algorithm. *Journal of Statistical Computation and Simulation*, 14(3-4):295–303.

Lange, K., Chi, E. C., and Zhou, H. (2014). A Brief Survey of Modern Optimization for Statisticians. *International Statistical Review*, 82(1):46–70.

Lanza, S. T., Kugler, K. C., and Mathur, C. (2011). Differential Effects for Sexual Risk Behavior: An Application of Finite Mixture Regression. *The Open Family*

Studies Journal, 4(Suppl 1-M9):81–88.

Laurin, C., Boomsma, D., and Lubke, G. (2016). The use of vector bootstrapping to improve variable selection precision in Lasso models. *Statistical Applications in Genetics and Molecular Biology*, 15(4):305–320.

Lavergne, C., Martinez, M. J., and Trottier, C. (2008). Empirical model selection in generalized linear mixed effects models. *Computational Statistics*, 23(1):99–109.

Lee, S., Pawitan, Y., and Lee, Y. (2015). A random-effect model approach for group variable selection. *Computational Statistics and Data Analysis*, 89:147–157.

Leng, C. L., Tran, M. N., and Nott, D. (2014). Bayesian adaptive Lasso. *Annals of the Institute of Statistical Mathematics*, 66(2):221–244.

Leung, M. K. and Elashoff, R. M. (1996a). Estimation of a generalized linear mixed-effects model with a finite-support random-effects distribution via gibbs sampling. *Biometrical Journal*, 38(5):519–536.

Leung, M. K. and Elashoff, R. M. (1996b). Generalized linear mixed-effects models with a finite-support random-effects distribution: A Maximum-penalized-likelihood approach. *Biometrical Journal*, 38(2):135–151.

Li, B. Y., Bruyneel, L., and Lesaffre, E. (2014). A multivariate multilevel Gaussian model with a mixed effects structure in the mean and covariance part. *Statistics in Medicine*, 33(11):1877–1899.

Li, G. R., Xue, L. G., and Lian, H. (2012). SCAD-penalised generalised additive models with non-polynomial dimensionality. *Journal of Nonparametric Statistics*, 24(3):681–697.

Li, H. and Wang, L. (2012). A consistent simulation-based estimator in generalized linear mixed models. *Journal of Statistical Computation and Simulation*, 82(8):1085–1103.

Li, J. B. and Gu, M. G. (2012). Adaptive LASSO for general transformation models with right censored data. *Computational Statistics and Data Analysis*, 56(8):2583–2597.

Li, M., Xiang, S. J., and Yao, W. X. (2016). Robust estimation of the number of components for mixtures of linear regression models. *Computational Statistics*, 31(4):1539–1555.

Li, S., Batterman, S., Su, F. C., and Mukherjee, B. (2013). Addressing extrema and censoring in pollutant and exposure data using mixture of normal distributions. *Atmospheric Environment*, 77:464–473.

Lim, H. K., Li, W. K., and Yu, P. L. H. (2014). Zero-inflated Poisson regression mixture model. *Computational Statistics and Data Analysis*, 71:151–158.

Lindstrom, M. J. and Bates, D. M. (1988). Newton-Raphson and EM Algorithms for Linear Mixed-Effects Models for Repeated-Measures Data. *Journal of the American Statistical Association*, 83(404):1014–1022.

Litiere, S., Alonso, A., and Molenberghs, G. (2007). Type I and type II error under random-effects misspecification in generalized linear mixed models. *Biometrics*, 63(4):1038–1044.

Liu, Y., Wang, Y. J., Feng, Y., and Wall, M. M. (2016). Variable Selection and Prediction with Incomplete High-Dimensional Data. *Annals of Applied Statistics*, 10(1):418–450.

Liu, Y. X., Yang, T. F., Li, H. W., and Yang, R. Q. (2014). Iteratively reweighted LASSO for mapping multiple quantitative trait loci. *Briefings in Bioinformatics*, 15(1):20–29.

Localio, A. R., Berlin, J. A., and Ten Have, T. R. (2006). Longitudinal and repeated cross-sectional cluster-randomization designs using mixed effects regression for binary outcomes: Bias and coverage of frequentist and Bayesian methods. *Statistics in Medicine*, 25(16):2720–2736.

Louis, T. A. (1982). Finding the observed information matrix when using the EM algorithm. *Journal of the Royal Statistical Society, Series B*, 44(2):226–233.

Lu, T., Liang, H., Li, H. Z., and Wu, H. L. (2011). High-Dimensional ODEs Coupled With Mixed-Effects Modeling Techniques for Dynamic Gene Regulatory Network Identification. *Journal of the American Statistical Association*, 106(496):1242–1258.

Lu, W., Goldberg, Y., and Fine, J. P. (2012). On the robustness of the adaptive lasso to model misspecification. *Biometrika*, 99(3):717–731.

Malmstrom, V., Catrina, A. I., and Klareskog, L. (2017). The immunopathogenesis of seropositive rheumatoid arthritis: from triggering to targeting. *Nature Reviews Immunology*, 17(1):60.

McCullagh, P. and Nelder, J. (1989). *Generalized Linear Models, Second Edition*. Chapman and Hall/CRC Monographs on Statistics and Applied Probability Series. Chapman & Hall.

McCulloch, C. E. (2000). Generalized linear models. *Journal of the American Statistical Association*, 95(452):1320–1324.

McCulloch, C. E. and Neuhaus, J. M. (2011). Prediction of Random Effects in Linear and Generalized Linear Models under Model Misspecification. *Biometrics*, 67(1):270–279.

McLachlan, G. J. and Krishnan, T. (2008). *The EM Algorithm and Extensions*. Wiley series in probability and statistics. Wiley, 2nd edition.

McLachlan, G. J., Lee, S. X., and Rathnayake, S. I. (2019). *Finite Mixture Models*, volume 6 of *Annual Review of Statistics and Its Application*, pages 355–378. Annual Reviews.

McLachlan, G. J. and Peel, D. (2000). *Finite Mixture Models*. Wiley series in Probability and Statistics. Wiley.

Mehranian, A., Rad, H. S., Rahmim, A., Ay, M. R., and Zaidi, H. (2013). Smoothly Clipped Absolute Deviation (SCAD) regularization for compressed sensing MRI using an augmented Lagrangian scheme. *Magnetic Resonance Imaging*, 31(8):1399–1411.

Meier, F. M., Frerix, M., Hermann, W., and Müller-Ladner, U. (2013). Current immunotherapy in rheumatoid arthritis. *Immunotherapy*, 5(9):955–974.

Meier, L., van de Geer, S. A., and Buhlmann, P. (2008). The group lasso for logistic regression. *Journal of the Royal Statistical Society, Series B*, 70(1):53–71.

Meinshausen, N. (2007). Relaxed Lasso. *Computational Statistics and Data Analysis*, 52(1):374–393.

Meng, X. L. and van Dyk, D. (1997). The EM algorithm - An old folk-song sung to a fast new tune. *Journal of the Royal Statistical Society, Series B*, 59(3):511–540.

Meng, X. L. and van Dyk, D. (1998). Fast EM-type implementations for mixed effects models. *Journal of the Royal Statistical Society, Series B*, 60(3):559–578.

Mkhadri, A. and Ouhourane, M. (2015). A group VISA algorithm for variable selection. *Statistical Methods and Applications*, 24(1):41–60.

Montgomery, D. C., Peck, E. A., and Vining, G. G. (2012). *Introduction to Linear Regression Analysis*. John Wiley & Sons.

Morgan, C. J., Lenzenweger, M. F., Rubin, D. B., and Levy, D. L. (2014). A hierarchical finite mixture model that accommodates zero-inflated counts, non-independence, and heterogeneity. *Statistics in Medicine*, 33(13):2238–2250.

Natarajan, R. and Kass, R. E. (2000). Reference Bayesian methods for generalized linear mixed models. *Journal of the American Statistical Association*, 95(449):227–237.

Nelder, J. A. and Wedderburn, R. W. (1972). Generalized Linear Models. *Journal of the Royal Statistical Society, Series A*, 135(3):370–384.

Neuhaus, J. M., McCulloch, C. E., and Boylan, R. (2013). Estimation of covariate effects in generalized linear mixed models with a misspecified distribution of random intercepts and slopes. *Statistics in Medicine*, 32(14):2419–2429.

Ng, C. T. and Yu, C. W. (2014). Modified SCAD penalty for constrained variable selection problems. *Statistical Methodology*, 21:109–134.

Ng, E. S. W., Carpenter, J. R., Goldstein, H., and Rasbash, J. (2006). Estimation in generalised linear mixed models with binary outcomes by simulated maximum likelihood. *Statistical Modelling*, 6(1):23–42.

Nishii, R. (1984). Asymptotic Properties of Criteria for Selection of Variables in Multiple-Regression. *Annals of Statistics*, 12(2):758–765.

Noh, M., Wu, L., and Lee, Y. (2012). Hierarchical likelihood methods for nonlinear and generalized linear mixed models with missing data and measurement errors in covariates. *Journal of Multivariate Analysis*, 109:42–51.

Ogden, H. E. (2015). A sequential reduction method for inference in generalized linear mixed models. *Electronic Journal of Statistics*, 9(1):135–152.

Okebe, J., Affara, M., Correa, S., Muhammad, A. K., Nwakanma, D., Drakeley, C., and D'Alessandro, U. (2014). School-Based Countrywide Seroprevalence Survey Reveals Spatial Heterogeneity in Malaria Transmission in the Gambia. *PLoS One*, 9(10):e110926.

Olk-Batz, C., Poetsch, A. R., Nollke, P., Claus, R., Zucknick, M., Sandrock, I., Witte, T., Strahm, B., Hasle, H., Zecca, M., Stary, J., Bergstraesser, E., De Moerloose, B., Trebo, M., van den Heuvel-Eibrink, M. M., Wojcik, D., Locatelli, F., Plass, C., Niemeyer, C. M., Flotho, C., and European Working Group of Myelodysplastic Syndromes in, C. (2011). Aberrant DNA methylation characterizes juvenile myelomonocytic leukemia with poor outcome. *Blood*, 117(18):4871–80.

Ormoz, E. and Eskandari, F. (2016). Variable selection in finite mixture of semiparametric regression models. *Communications in Statistics-Theory and Methods*, 45(3):695–711.

Pan, J. M. and Shang, J. F. (2018). A simultaneous variable selection methodology for linear mixed models. *Journal of Statistical Computation and Simulation*, 88(17):3323–3337.

Pan, J. X. and Huang, C. (2014). Random effects selection in generalized linear mixed models via shrinkage penalty function. *Statistics and Computing*, 24(5):725–738.

Pearson, K. (1894). Contributions to the mathematical theory of evolution. *Philosophical Transactions of the Royal Society of London. A*, 185:71–110.

Potscher, B. M. and Schneider, U. (2009). On the distribution of the adaptive LASSO estimator. *Journal of Statistical Planning and Inference*, 139(8):2775–2790.

Pu, W. J. and Niu, X. F. (2006). Selecting mixed-effects models based on a generalized information criterion. *Journal of Multivariate Analysis*, 97(3):733–758.

Qiu, J., Li, D. G., and You, J. H. (2015). SCAD-penalized regression for varying-coefficient models with autoregressive errors. *Journal of Multivariate Analysis*, 137:100–118.

Radchenko, P. and James, G. M. (2011). Improved Variable Selection with Forward-Lasso Adaptive Shrinkage. *Annals of Applied Statistics*, 5(1):427–448.

Rajaratnam, B., Roberts, S., Sparks, D., and Dalal, O. (2016). Lasso regression: estimation and shrinkage via the limit of Gibbs sampling. *Journal of the Royal Statistical Society, Series B*, 78(1):153–174.

Rao, C. R. (1955). Estimation and tests of significance in factor analysis. *Psychometrika*, 20(2):93–111.

Redner, R. A. and Walker, H. F. (1984). Mixture Densities, Maximum-Likelihood and the EM Algorithm. *Siam Review*, 26(2):195–237.

Rein, D. B. (2005). A matter of classes: Stratifying health care populations to produce better estimates of inpatient costs. *Health Services Research*, 40(4):1217–1233.

Robert, C. P. and Casella, G. (2010). *Introducing Monte Carlo Methods with R*, volume 18. Springer.

Rubinstein, R. and Samorodnitsky, G. (1985). Variance reduction by the use of common and antithetic random variables. *Journal of Statistical Computation and Simulation*, 22(2):161–180.

Ryden, T. (2008). EM versus Markov chain Monte Carlo for Estimation of Hidden Markov Models: A Computational Perspective. *Bayesian Analysis*, 3(4):659–688.

Sagara, I., Piarroux, R., Djimde, A., Giorgi, R., Kayentao, K., Doumbo, O. K., and Gaudart, J. (2014). Delayed anemia assessment in patients treated with oral artemisinin derivatives for uncomplicated malaria: A pooled analysis of clinical trials data from Mali. *Malaria Journal*, 13(1):358.

Sakate, D. and Kashid, D. (2014). Variable selection via penalized minimum ϕ-divergence estimation in logistic regression. *Journal of Applied Statistics*, 41(6):1233–1246.

Sammel, M., Lin, X. H., and Ryan, L. (1999). Multivariate linear mixed models for multiple outcomes. *Statistics in Medicine*, 18(17-18):2479–2492.

Sampson, J. N., Chatterjee, N., Carroll, R. J., and Muller, S. (2013). Controlling the local false discovery rate in the adaptive Lasso. *Biostatistics*, 14(4):653–666.

Schelldorfer, J., Meier, L., and Buhlmann, P. (2014). GLMMLasso: An Algorithm for High-Dimensional Generalized Linear Mixed Models Using $\ell(1)$-Penalization. *Journal of Computational and Graphical Statistics*, 23(2):460–477.

Schwarz, G. (1978). Estimating Dimension of a Model. *Annals of Statistics*, 6(2):461–464.

Scott, D., Wolfe, F., and Huizinga, T. (2010). Rheumatoid arthritis.

Simon, N. and Tibshirani, R. (2012). Standardization and the Group Lasso Penalty. *Statistica Sinica*, 22(3):983–1001.

Sinha, S. K. (2004). Robust analysis of generalized linear mixed models. *Journal of the American Statistical Association*, 99(466):451–460.

Sinha, S. K. (2009). Bootstrap tests for variance components in generalized linear mixed models. *Canadian Journal of Statistics*, 37(2):219–234.

Smolen, J. S. and Aletaha, D. (2015). Rheumatoid arthritis therapy reappraisal: strategies, opportunities and challenges. *Nature Reviews Rheumatology*, 11(5):276.

Smolen, J. S., Breedveld, F. C., Burmester, G. R., Bykerk, V., Dougados, M., Emery, P., Kvien, T. K., Navarro-Compán, M. V., Oliver, S., Schoels, M., et al. (2016). Treating rheumatoid arthritis to target: 2014 update of the recommendations of an international task force. *Annals of the rheumatic diseases*, 75(1):3–15.

Solis-Soto, M. T., Patino, A., Nowak, D., and Radon, K. (2013). Association between environmental factors and current asthma, rhinoconjunctivitis and eczema symptoms in school-aged children from Oropeza Province - Bolivia: A cross-sectional study. *Environmental Health*, 12(1):95.

Stalmach, A., Johnsson, H., McInnes, I. B., Husi, H., Klein, J., Dakna, M., Mullen, W., Mischak, H., and Porter, D. (2014). Identification of urinary peptide biomarkers associated with rheumatoid arthritis. *PLoS One*, 9(8):e104625.

Steele, B. M. (1996). A modified EM algorithm for estimation in generalized mixed models. *Biometrics*, 52(4):1295–1310.

Storey, G. O., Comer, M., and Scott, D. L. (1994). Chronic arthritis before 1876: early British cases suggesting rheumatoid arthritis. *Annals of the Rheumatic Diseases*, 53(9):557–560.

Tang, Q. G. and Karunamuni, R. J. (2018). Robust variable selection for finite mixture regression models. *Annals of the Institute of Statistical Mathematics*, 70(3):489–521.

Teicher, H. (1963). Identifiability of finite mixtures. *The annals of Mathematical statistics*, 34(4):1265–1269.

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

Tibshirani, R. (2011). Regression shrinkage and selection via the lasso: a retrospective. *Journal of the Royal Statistical Society, Series B*, 73(3):273–282.

Tibshirani, R., Saunders, M., Rosset, S., Zhu, J., and Knight, K. (2005). Sparsity and smoothness via the fused lasso. *Journal of the Royal Statistical Society, Series B*, 67(1):91–108.

Torabi, M. (2013). Likelihood inference in generalized linear mixed measurement error models. *Computational Statistics and Data Analysis*, 57(1):549–557.

Tutz, G. and Kauermann, G. (2003). Generalized linear random effects models with varying coefficients. *Computational Statistics and Data Analysis*, 43(1):13–28.

Tutz, G. and Oelker, M. R. (2017). Modelling Clustered Heterogeneity: Fixed Effects, Random Effects and Mixtures. *International Statistical Review*, 85(2):204–227.

Usai, M. G., Goddard, M. E., and Hayes, B. J. (2009). Lasso with cross-validation for genomic selection. *Genetics Research*, 91(6):427–36.

van de Geer, S., Buhlmann, P., and Zhou, S. H. (2011). The adaptive and the thresholded Lasso for potentially misspecified models (and a lower bound for the Lasso). *Electronic Journal of Statistics*, 5:688–749.

Verbeke, G. and Lesaffre, E. (1996). A linear mixed-effects model with heterogeneity in the random-effects population. *Journal of the American Statistical Association*, 91(433):217–221.

Wang, H. S. and Leng, C. L. (2008). A note on adaptive group lasso. *Computational Statistics and Data Analysis*, 52(12):5277–5286.

Wang, H. Y., Zou, G. H., and Wan, A. T. K. (2013). Adaptive LASSO for varying-coefficient partially linear measurement error models. *Journal of Statistical Planning and Inference*, 143(1):40–54.

Wang, K., Yau, K. K. W., and Lee, A. H. (2002). A hierarchical Poisson mixture regression model to analyse maternity length of hospital stay. *Statistics in Medicine*, 21(23):3639–3654.

Wang, L. and Li, R. (2009). Weighted Wilcoxon-Type Smoothly Clipped Absolute Deviation Method. *Biometrics*, 65(2):564–571.

Wang, L. F., Chen, G., and Li, H. Z. (2007). Group SCAD regression analysis for microarray time course gene expression data. *Bioinformatics*, 23(12):1486–1494.

Wang, S. J., Nan, B., Rosset, S., and Zhu, J. (2011). Random Lasso. *Annals of Applied Statistics*, 5(1):468–485.

Wang, S. L., Yao, W. X., and Huang, M. (2014). A note on the identifiability of nonparametric and semiparametric mixtures of GLMs. *Statistics and Probability Letters*, 93:41–45.

Ward, M. M., Guthrie, L. C., and Alba, M. I. (2014). Brief Report: Rheumatoid Arthritis Response Criteria and Patient-Reported Improvement in Arthritis Activity: Is an American College of Rheumatology Twenty Percent Response Meaningful to Patients? *Arthritis & Rheumatology*, 66(9):2339–2343.

Wasserman, A. (2011). Diagnosis and management of rheumatoid arthritis. *American family physician*, 84(11):1245–1252.

Wei, F. R. and Zhu, H. X. (2012). Group coordinate descent algorithms for nonconvex penalized regression. *Computational Statistics and Data Analysis*, 56(2):316–326.

Wei, G. C. G. and Tanner, M. A. (1990). A Monte-Carlo Implementation of the EM Algorithm and the Poor Mans Data Augmentation Algorithms. *Journal of the American Statistical Association*, 85(411):699–704.

Wimmer, V., Lehermeier, C., Albrecht, T., Auinger, H. J., Wang, Y., and Schon, C. C. (2013). Genome-wide prediction of traits with different genetic architecture through efficient variable selection. *Genetics*, 195(2):573–87.

Wolfinger, R. and Oconnell, M. (1993). Generalized Linear Mixed Models - a Pseudo-Likelihood Approach. *Journal of Statistical Computation and Simulation*, 48(3-4):233–243.

Wu, T. T., Gong, H., and Clarke, E. M. (2011). A transcriptome analysis by lasso penalized Cox regression for pancreatic cancer survival. *Journal of Bioinformatics and Computational Biology*, 9(supp01):63–73.

Wu, T. T. and Lange, K. (2008). Coordinate Descent Algorithms for Lasso Penalized Regression. *Annals of Applied Statistics*, 2(1):224–244.

Wu, X. and Yu, C. L. (2016). Estimation of the mixtures of GLMs with covariate-dependent mixing proportions. *Communications in Statistics-Theory and Methods*, 45(24):7242–7257.

Xie, Z. X. and Xu, Y. (2014). Sparse group LASSO based uncertain feature selection. *International Journal of Machine Learning and Cybernetics*, 5(2):201–210.

Yan, F. R., Lin, J. G., and Liu, Y. (2011). Sparse Logistic Regression for Diagnosis of Liver Fibrosis in Rat by Using SCAD-Penalized Likelihood. *Journal of Biomedicine and Biotechnology*, pages 1–8.

Yao, W. X. and Song, W. X. (2015). Mixtures of Linear Regression with Measurement Errors. *Communications in Statistics-Theory and Methods*, 44(8):1602–1614.

Yau, K. K. W. and McGilchrist, C. A. (1996). Simulation study of the GLMM method applied to the analysis of clustered survival data. *Journal of Statistical Computation and Simulation*, 55(3):189–200.

Yau, K. K. W. and McGilchrist, C. A. (1997). Use of generalised linear mixed models for the analysis of clustered survival data. *Biometrical Journal*, 39(1):3–11.

Young, D. S. (2014). Mixtures of regressions with changepoints. *Statistics and Computing*, 24(2):265–281.

Ypma, T. J. (1995). Historical development of the Newton-Raphson method. *Siam Review*, 37(4):531–551.

Yu, C. L. and Wang, X. Y. (2019). A new model selection procedure for finite mixture regression models. *Communications in Statistics-Theory and Methods*, 48(24):1–20.

Yu, D. L., Zhang, X. Y., and Yau, K. K. W. (2013). Information based model selection criteria for generalized linear mixed models with unknown variance component parameters. *Journal of Multivariate Analysis*, 116:245–262.

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

Yuen, J. C., Yue, C. J., Erickson, S. W., Cooper, S., Boneti, C., Henry-Tillman, R., and Klimberg, S. (2014). Comparison between Freeze-dried and Ready-to-use AlloDerm in Alloplastic Breast Reconstruction. *Plastic and Reconstructive Surgery Global Open*, 2(3):e119.

Zeger, S. L. and Karim, M. R. (1991). Generalized Linear-Models with Random Effects - A Gibbs sampling approach. *Journal of the American Statistical Association*, 86(413):79–86.

Zeger, S. L., Liang, K. Y., and Albert, P. S. (1988). Models for Longitudinal Data - A Generalized Estimating Equation Approach. *Biometrics*, 44(4):1049–1060.

Zeng, P., Wei, Y. Y., Zhao, Y., Liu, J., Liu, L. Y., Zhang, R., Gou, J. W., Huang, S. P., and Chen, F. (2014). Variable selection approach for zero-inflated count data via adaptive lasso. *Journal of Applied Statistics*, 41(4):879–894.

Zhang, D. W. (2004). Generalized linear mixed models with varying coefficients for longitudinal data. *Biometrics*, 60(1):8–15.

Zhang, H., Yu, Q., Feng, C., Gunzler, D., Wu, P., and Tu, X. M. (2012). A new look at the difference between the GEE and the GLMM when modeling longitudinal count responses. *Journal of Applied Statistics*, 39(9):2067–2079.

Zhou, J., Liu, J., Narayan, V. A., and Ye, J. (2012). Modeling Disease Progression via Fused Sparse Group Lasso. *International Conference on Knowledge Discovery and Data Mining*, 2012:1095–1103.

Zhu, H. T. and Lee, S. Y. (2003). Local influence for generalized linear mixed models. *Canadian Journal of Statistics*, 31(3):293–309.

Zou, H. (2006). The Adaptive Lasso and Its Oracle Properties. *Journal of the American Statistical Association*, 101(476):1418–1429.

Zou, H. and Hastie, T. (2005). Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society, Series B*, 67(2):301–320.