The Generalized Method of Moments for Mixture and Mixed Models

by

Zhiyue Huang

A thesis
presented to the University of Waterloo
in fulfillment of the
requirement for the degree of
Doctor of Philosophy
in
Statistics

Waterloo, Ontario, Canada, 2015

I hereby declare that I am the sole author of this thesis. This is a true copy of the thesis, including any required final revisions, as accepted by my examiners.

I understand that my thesis may be made electronically available to the public.

Abstract

Mixture models can be found in a wide variety of statistical applications. However, undertaking statistical inference in mixture models, especially non-parametric mixture models, can be challenging. A general, or nonparametric, mixture model has effectively an infinite dimensional parameter space. In frequentist statistics, the maximum likelihood estimator with an infinite dimensional parameter may not be consistent or efficient in the sense that the Cramer-Rao bound is not attained even asymptotically. In Bayesian statistics, a prior on an infinite dimensional space is not well defined and can be highly informative even with large amounts of data.

In this thesis, we mainly consider mixture and mixed-effects models, when the mixing distribution is non-parametric. Following the dimensionality reduction idea in [Marriott, 2002], we propose a reparameterization-approximation framework with a complete orthonormal basis in a Hilbert space. The parameters in the reparameterized models are interpreted as the generalized moments of a mixing distribution. We consider different orthonormal bases, including the families of orthogonal polynomials and the eigenfunctions of positive self-adjoint integral operators. We also study the approximation errors of the truncation approximations of the reparameterized models in some special cases.

The generalized moments in the truncated approximations of the reparameterized models have a natural parameter space, called the generalized moment space. We study the geometric properties of the generalized moment space and obtain two important geometric properties: the positive representation and the gradient characterization. The positive representation reveals the identifiability of the mixing distribution by its generalized moments and provides an upper bound of the number of the support points of the mixing distribution. On the other hand, the gradient characterization provides the foundation of the class of gradient-based algorithms when the feasible set is the generalized moment space.

Next, we aim to fit a non-parametric mixture model by a set of generalized moment conditions, which are from the proposed reparameterization-approximation procedure. We propose a new estimation method, called the generalized method of moments for mixture models. The proposed estimation method involves minimizing a quadratic objective function over the generalized moment space. The proposed estimators can be easily computed through the gradient-based algorithms. We show the convergence rate of the mean squared error of the proposed estimators, as the sample size goes to infinity. Moreover, we design the quadratic objective function to ensure that the proposed estimators are robust to the outliers. Compared to the other existing estimation methods for mixture models, the GMM for mixture models is more computationally friendly and robust to outliers.

Lastly, we consider the hypothesis testing problem on the regression parameter in a mixed-effects model with univariate random effects. Through our new procedures, we obtain a series of estimating equations parameterized in the regression parameter and the generalized moments of the random-effects distribution. These parameters are estimated under the framework of the generalized method of moments. In the case that the number of the generalized moments diverges with the sample size and the dimension of the regression parameter is fixed, we compute the convergence rate of the generalized method of moments estimators for the mixed-effects models with univariate random effects. Since the regularity conditions in [Wilks, 1938] fail under our context, it is challenging to construct an asymptotically χ^2 test statistic. We propose using ensemble inference, in which an asymptotically χ^2 test statistic is constructed from a series of the estimators obtained from the generalized estimating equations with different working correlation matrices.

Acknowledgements

First of all, I would like to express my most gratitude to my supervisor Professor Paul Marriott. This thesis would not have been possible without his guidance and support. I greatly appreciate his stimulated discussion, inspirations and suggestions that help me to develop my research interests and shape my research skills.

I owe my thanks to my thesis committee members, Professor Pengfei Li, Professor Shoja'eddin Chenouri, Professor Dinghai Xu, and Professor Jun Zhang, for their valuable help and insightful comments.

Many thanks to the faculty and staff members of the Department of Statistics and Actuarial in University of Waterloo for creating a supportive environment for my study, particularly Mary Lou Dufton, Professor Mu Zhu and Professor Changbao Wu.

My thanks also go to my friends in Waterloo, especially to Jiheng Wang and Zhenhao Li for their everlasting friendship, Xichen She for being such an awesome office mate, and Jiaxi Liang and Lu Xin for the basketball games on many Thursday afternoons in PAC or CIF.

Finally, I want to thank my family members for their continuous and unconditional love, support and faith in me.

To All Who Believe in Me.

Contents

List of Figures

Guide to Notation

In this section, we provide brief explanation and representative examples of the notation used in this thesis. Matrices and column vectors are typically denoted using bold letters, e.g., $\boldsymbol{A}$, and transpose is denoted using $\mathbf{A}^{\mathrm{T}}$. For a parameter $\boldsymbol{\beta}$, we use $\tilde{\boldsymbol{\beta}}$ or $\hat{\boldsymbol{\beta}}$ to denote the esimators of the parameter, $\boldsymbol{\beta}^*$ to denote the true value of the parameter, and $\boldsymbol{\beta}_0$ to denote a value of the parameter. To establish the asymptotic results, we use the order notations, $O(\cdot)$, $o(\cdot)$, $O_p(\cdot)$ and $o_p(\cdot)$; see [Small, 2010, p.g. 4-16] for precise definitions. We list the following notations which are unified throughout this thesis.

symbol	**description**
$\boldsymbol{b}_n$	Random effects vector of the n^{th} individual in a mixed-effects model
$\mathcal{B}$	Support set of the random effects
C	Constant, which may vary between lines
$\mathcal{C}$	Generalized moment cone
$f_{\text{Mix}}(x;Q)$	Mixture models with mixing distribution Q
J_N+1	Dimension of the generalized moment vectors
$\boldsymbol{m}$	Generalized moment vectors for a mixture model
$\mathcal{M}$	Generalized moment space
K_N	Number of models in an ensemble inference
N	Sample size
p	Dimension of the regression parameter $\boldsymbol{\beta}$ in a mixed-effects model
Q	Mixing or random effects distribution
$\mathbb{R}$	Real space
$\mathcal{S}$	Sample space
pr	Probability functions

T_n	Number of visits of the n^{th} individual
$\boldsymbol{W}$	Weighting matrices
$\boldsymbol{x}_{nt}$, $\boldsymbol{X}_{nt}$	Covariate vector to the fixed effects on the n^{th} individual at t^{th} visit
y_{nt}, Y_{nt}	Response variable of the n^{th} individual at t^{th} visit
$\boldsymbol{z}_{nt}$, $\boldsymbol{Z}_{nt}$	Covariate vector to the random effects on the n^{th} individual at t^{th} visit
$\boldsymbol{\alpha}$	Generalized moments vector for a mixed-effects model
$\boldsymbol{\beta}$	Regression parameter vector in a mixed-effects model
λ_j	The j^{th} largest eigenvalue
θ	Mixing parameter
Θ	Support space of the mixing parameter θ

operators	**description**
det	Determinant operator of a matrix
diag	The operator to a vector that returns a square diagonal matrix with the elements of the vector on the main diagonal
$\mathbb{E}_X$	Expectation with respect to the random variable X
Var_X	Variance with respect random variable X
Cov	Covariance operator to two random variables
$\lVert\cdot\rVert_2$	2-norm of a matrix or L^2-norm of a vector

abbreviations	**description**
CMM	Conditional mixed method
CNM	Constrained Newton method with multiple exchange vertices
GLMM	Generalized linear mixed models
GEE	Generalized estimating equations
GMM	Generalized method of moments
LRTS	Likelihood ratio test statistics

MM	Method of Moments
MLE	Maximum likelihood estimators
MSE	Mean squared errors
NPMLE	Non-parametric maximum likelihood estimators
NEF-QVF	Natural exponential families with quadratic variance function
PQL	Penalized quasi-likelihood
QIF	Quadratic inference function
UMM	Unconditional mixed method

distributions	**description**
$\mathrm{Bin}(N, p)$	Binomial distribution with number of trials N and success probability p in each trial
$\mathrm{Exp}(\theta)$	Exponential distribution with mean θ
$\mathcal{N}(\theta, \sigma^2)$	Normal distribution with mean θ and variance σ^2
$\mathrm{Pois}(\theta)$	Poisson distribution with mean θ
χ^2_p	χ^2 distribution with degrees of freedom p

Chapter 1

Introduction to Mixture Models

1.1 Introductions

A mixture model is one which can be written as a convex combination of multiple distribution functions; a comprehensive review of these models can be found in [Lindsay, 1995]. Commonly used examples include finite mixture models with known or unknown components, with known or unknown order, and parametric or non-parametric mixture models. Here a parametric mixture is one where the mixing distribution is assumed to lie in a known parametric family, while non-parametric will mean the mixing distribution is unconstrained by any functional assumptions.

Much of the pioneering work on mixture models in statistics can be found in [Pearson, 1898], [Feller, 1943] and [Teicher, 1960]. Good modern references include [Titterington et al., 1985], [McLachlan and Basford, 1988], [Lindsay, 1995], [McLachlan and Peel, 2000], [Schlattmann, 2009] and [Mengersen et al., 2011].

Mixture models are useful in statistical modelling because of their flexibility [McLachlan and Basford, 1988] and the potential interpretation of the mixing process [Everitt et al., 2011]. However, mixture models create challenges for statistical inference. Firstly, mixtures may not be identifiable; see [Tallis and Chesson, 1982], [Lindsay and Roeder, 1993] and [Jasra et al., 2005]. Secondly, boundaries (and possible singularities) exist in the parameter space of a finite mixture; see [Leroux, 1992], [Chen and Kalbfleisch, 1996] and [Li et al., 2009] and this is also true for more general mixtures. Thirdly, a non-parametric mixture model can be thought of as having an infinite dimensional parameter space; see [Lindsay, 1980, 1983] and [Marriott, 2007]. Lastly, log-likelihood functions may not be convex; see [Gan and Jiang, 1999]. A more detailed discussion of these issues can be found in Section 1.2. Moreover, the challenges may also affect the convergence rates of associated computational algorithms, see Section 1.3 for details.

Using the geometry of mixture models is a useful approach to overcome these challenges. Lindsay [1983] studied the geometry of mixture models in an embedding space determined by the observed, and hence finite, sample (defined in Equation (1.1)) and gave the fundamental properties of the non-parametric maximum likelihood estimator for mixture models. These properties include identifiability and bounds on the number of support points in a non-parametric maximum likelihood estimator. It also leads to the class of gradient-based computational algorithms, which can be fast and stable; see [Böhning, 1995] and [Wang, 2007]. However, these fundamental properties of the non-parametric maximum likelihood are based on the observed sample. Developing asymptotic results on the non-parametric maximum likelihood estimator could be theoretical challenging since the size of the sample space is unbounded. On the other hand, Marriott [2002] considered the geometry of mixture models in an affine space and introduced the class of local mixture models, which can successfully reduce the number of parameters in a mixture model. However, a local mixture model may not provide a consistent estimator to the true mixture model; see Section 1.4 for details.

In this thesis, we develop estimation and inferential procedures for non-parametric mixture (or mixed-effects) models under the framework of the generalized method of moments. To deal with the dimensionality issue in a non-parametric mixture (or mixed-effects) model, we reduce the dimension of the parameter space through a reparameterization-approximation procedure with a complete orthonormal basis in a Hilbert space; see Chapter 2. The proposed reparameterization-approximation framework leads to the models with generalized moments as their parameters and a series of generalized moment conditions. Next, we study the geometric properties of the set of the generalized moments; see Chapter 3. They are helpful in studying the fundamental properties of the generalized method of moments and designing the computational algorithms which are associated with the proposed estimators in later chapters. The generalized moment conditions, which are obtained through the reparameterization-approximation procedure, can be used to fit a non-parametric mixture model; see Chapter 4. The proposed method is called the generalized method of moments for mixture models. It can be made robust to the outliers when weighting matrix is carefully designed. Then, we consider the class of mixture models with regression parameters, the mixed-effects models, under the framework of the generalized method of moments; see Chapter 5 and 6. The asymptotic theorems of the generalized method of moments estimators are established in the case that the dimension of the generalized moments diverges with the sample size; see Chapter 7. As will be pointed out later, the asymptotic results of the generalized method of moments can not be used for hypothesis testing problem on the regression parameters in a mixed-effects model with univariate random effects. Therefore, we propose to use the ensemble inference; see Chapter 8.

This chapter is organized as follows. In Section 1.2, we discuss important statistical inference issues for mixtures. In Section 1.3, we review the computational algorithms. In Section 1.4, we consider the underlying geometry. In Section 1.5, we look at the class of mixed-effects models. In Section 1.6, we list some real data examples. Lastly, we give an outline of this thesis.

1.2 Inference for Mixture Models

In this section, we discuss the difficulties of undertaking statistical inference in mixture models.

1.2.1 Identifiability

Titterington et al. [1985] described identifiability as “the existence of a unique characterization for any one of the class of models being considered”. Identifiability is defined in [Teicher, 1963] for finite, and [Tallis, 1969] for non-finite mixture. Some studies have been done on the identifiability of non-finite mixtures, such as [Tallis, 1969] for countable, and [Tallis and Chesson, 1982] for general mixture models. Because of their wide applicability, more research has been done on finite mixtures, and this work is summarized in [Titterington et al., 1985]. In this case, usually, we require the mixing proportions be strictly positive and the mixing parameters to be unequal; see [Chen et al., 2004]. With these constraints, many component distributions, including the Poisson [Teicher, 1961], exponential [Teicher, 1963], gamma [Teicher, 1963] and negative binomial [Yakowitz and Spragins, 1968], are shown to be identifiable as finite mixtures.

One special type of identifiability issue is the label switching problem when we are making inferences about the individual components; see [Stephens, 2000], [Jasra et al., 2005] and [Sperrin et al., 2010]. We could impose identifiability constraints on a particular set of parameters; see [Richardson and Green, 1997]. However, the successful of this often depends on the design and performance of the MCMC sampler, as argued by Celeux et al. [2000]. Other possible solutions are reviewed in [Jasra et al., 2005]. The following example shows another important form of lack of identifiability.

Example 1.1.
Consider a normal mixture of normal distributions such that $X \mid \theta \sim \mathcal{N}(\theta, \sigma_1^2)$, where $\theta \sim \mathcal{N}(0, \sigma_2^2)$ and $\mathcal{N}(\theta, \sigma^2)$ is the normal distribution with mean θ and variance σ^2. It can be shown that this model is equivalent to $\mathcal{N}(0, \sigma_1^2 + \sigma_2^2)$ for any σ_1^2 and σ_2^2. Therefore, this model is not identified.

1.2.2 The Parameter Space of Finite Mixture Models

The parameter space of a finite mixture model has boundaries, because the mixing proportions are non-negative and sum to one. Moreover, singularities exist, since we set the component parameters unequal for identifiability reasons. Undertaking statistical inference near the boundaries or singularities is challenging; see [Cheng and Traylor, 1995]. Firstly, the dimension of the parameter space is non-constant; see [Ryden, 1995], [James et al., 2001], [Chen and Khalili, 2008]. Secondly, log-likelihoods can not be approximated quadratically; see [Chen and Kalbfleisch, 1996], [Chen, 1998] and [Li et al., 2009]. Lastly, some estimators might be inconsistent; see [Kiefer and Wolfowitz, 1956], [Laird, 1978] and [Leroux, 1992]. To keep estimates away from boundaries or singularities, penalty functions have been used; see [Chen and Kalbfleisch, 1996], [Chen, 1998] and [Chen and Khalili, 2008].

The change of dimension of the parameter space causes problems for both frequentist and Bayesian theory. Whenever the order is underestimated, the model is mis-specified; see [Keribin, 2000]. On the other hand, if we overestimate the order, Chen [1995] showed that the convergence rate of an over-parameterized mixture could be lower than a finite mixture with a correctly specified order, when the sample size goes to infinity. In Bayesian statistics, this change of dimension makes the posterior distribution sensitive; see [Jasra et al., 2005].

To determine the order, many model selection techniques can be used, including the log-likelihood ratio test statistic (LRTS) (summarized in [Everitt et al., 2011]), the information criteria (summarized in [McLachlan and Peel, 2000]), the moment matrix [Lindsay, 1989b] and the non-smooth penalty functions [Chen and Khalili, 2008]. Overall, as noted by Everitt et al. [2011], "research on model selection criteria has not provided an unequivocal answer to the basic question of selecting the right number of components."

The failure of the quadratic approximation of log-likelihoods implies that the regularity conditions in [Wilks, 1938] are not satisfied. As a result, under the regularity conditions given in [Chen, 1995], the likelihood ratio statistic of a mixture model has a mixture of χ^2 distributions as its asymptotic distribution, compared to a single χ^2 distribution in the classical result. This also affects some information criteria depending on log-likelihoods, such as the AIC and the BIC; see [Ray and Lindsay, 2007]. Moreover, convergence rates can be very slow in some computational algorithms, especially the EM algorithm; see [Chen, 1998].

Example 1.2.

Consider the likelihood space of the family of binomial distribution $\mathrm{Bin}(2, \theta)$*, where* 2 *is the number of trials and* θ *denotes the probability of success in each trial. The LRTS for the test of one versus of two components has asymptotic distribution as*

$0.5\chi_0^2 + 0.5\chi_1^2$, where χ_p^2 is the χ^2 distribution with degrees of freedom p; see [Lindsay, 1995]. We simulate 1000 statistics under the true distribution $\text{Bin}(2, 0.5)$ with two levels of sample size, 30 and 100, and show the Q-Q plots of them versus $0.5\chi_0^2+0.5\chi_1^2$ in Figure 1.1.

Example 1.3.
Li et al. [2009] considered a mixture of two exponentials:

$$(1-\alpha)\text{Exp}(1) + \alpha\text{Exp}(\theta),$$

where $\text{Exp}(\theta)$ denotes the exponential distribution with mean θ. Consider testing the hypothesis $\alpha = 0$ versus $\alpha > 0$. Under the null hypothesis, the Fisher information is infinite when $\theta \geq 2$.

The boundaries and singularities can also lead to the unbounded likelihood and an inconsistent estimator; see [Kiefer and Wolfowitz, 1956] and [Cheng and Traylor, 1995]. However, many estimators for the mixing distributions, such as the penalized maximum likelihood estimator [Chen, 1998], the penalized minimum distance estimator [Chen and Kalbfleisch, 1996], and the maximum likelihood estimator (MLE) with an upper bound of the order [Leroux, 1992], are still consistent.

Example 1.4.
Kiefer and Wolfowitz [1956] considered the normal mixture with unknown parameters α and σ:

$$\alpha\mathcal{N}(0,1) + (1-\alpha)\mathcal{N}(0,\sigma^2),$$

where $\alpha \in (0,1)$ and $\sigma \geq 0$. Here the likelihood goes to infinity for any value of $\alpha \in (0,1)$ as the estimates of σ goes to 0.

1.2.3 Parameter Space of Non-parametric Mixture Models

A non-parametric mixture model can be considered to have an infinite dimensional parameter space, since an unknown mixing distribution is involved. When the underlying mixing process is not of direct interest, it can be viewed as an infinite

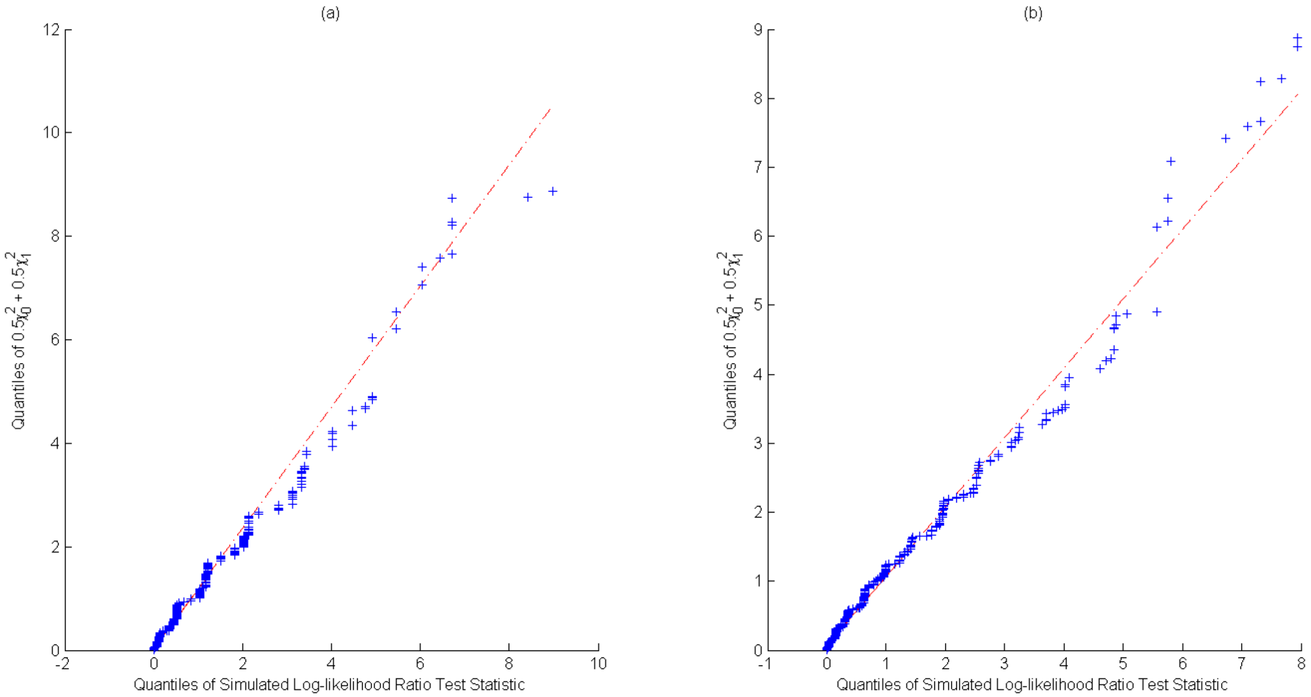

Figure 1.1: The Q-Q plots of the simulated LRTS versus $0.5\chi_0^2 + 0.5\chi_1^2$: (a) with sample size 30; (b) with sample size 100.

dimensional nuisance parameter. In frequentist statistics, the maximum likelihood estimate with an infinite dimensional nuisance parameter may not give a consistent estimator and may not be efficient in the sense that the Cramer-Rao bound is not attained even asymptotically; see [Neyman and Scott, 1948]. In Bayesian statistics, the prior on an infinite dimensional space may not be well defined; see [Marriott, 2007]. We can use the modified likelihood [Lindsay, 1980] or reduce the dimension of the parameter space, such as the semi-nonparametric approach in [Gallant and Nychka, 1987] and local mixture models in [Marriott, 2002]. Moreover, the non-parametrical maximum likelihood estimator (NPMLE) has also been proposed; see [Lindsay, 1995].

1.2.4 Non-convexity of Log-likelihood

Log-likelihoods of mixture models may have multi-modes. For example, Gan and Jiang [1999] gave the following example to illustrate this point.

Example 1.5.
Consider the normal mixture

$$0.4\mathcal{N}(\theta, 1) + 0.6\mathcal{N}(6, 4).$$

Let $\theta = -3$. We independently generate 5000 random variables and fit the model. The log-likelihood function, which is multi-modal, is plotted in Figure 1.2.

1.3 Computation in Mixture Models

The commonly used computational algorithms for mixture models include the EM algorithm, gradient based algorithms, the method of moments, and MCMC methods. In this section, we describe their strengths and weaknesses.

The EM-algorithm is popular for finding the MLE because, as pointed in [Redner and Walker, 1984], a finite mixture model with a known order is a special case of the model for incomplete data. It is reliable to find a local maximum but its converge rate is extremely slow; see [Titterington et al., 1985] and [Böhning et al., 1994].

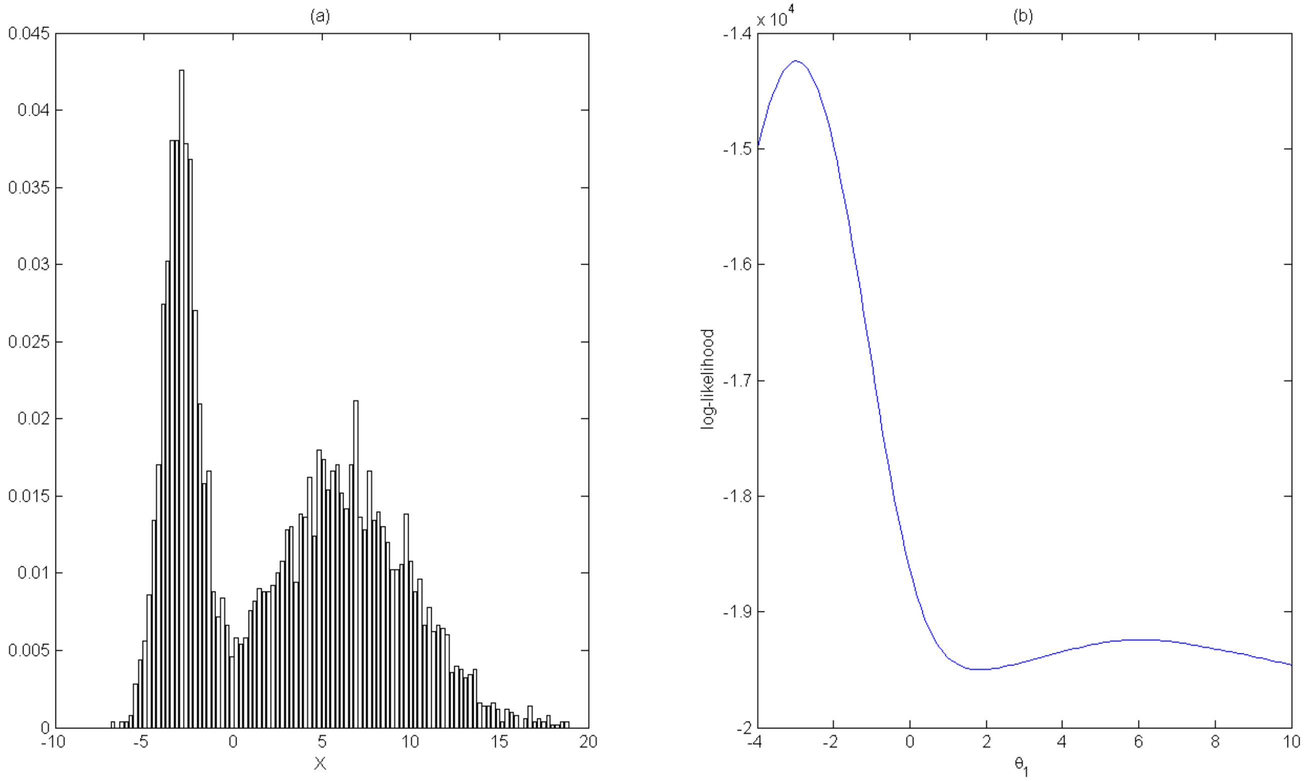

Figure 1.2: Plot of (a) the histogram; (b) the log-likelihood with multi-modes.

Gradient based algorithms are used for the NPMLE; see [Böhning, 1995]. They are based on the fundamental theorems of the NPMLE, which we discuss in the following section. Böhning [1995] reviewed some existing gradient based algorithms, including the vertex direction method, the vertex exchange method and the intra-simplex direction method. These algorithms converge faster than the EM algorithm; see [Böhning, 1995]. Wang [2007] proposed a faster algorithm called the constrained Newton method with multiple vertex exchange (CNM). At each iteration step, the CNM allows multiple points change in the support set, and thus increases the convergence rate of the algorithm.

The method of moments for mixture models can be used for finite mixture models with known orders; see [Lindsay, 1989b]. The mixing distribution is estimated from its moments, as summarized in [Titterington et al., 1985]. However, the moments are not easily obtained unless the mixture is with respect to the mean parameter in the family of the quadratic variance natural exponential distributions; see [Morris, 1982] and [Lindsay, 1989b]. Moreover, estimated component parameters are not necessarily in the parameter space, due to sampling variability, and thus adjustments are suggested by Lindsay [1989b].

MCMC methods for mixture models are popular, because it is hard to find a prior which makes the posterior belong to a tractable distributional family. A routine Bayesian analysis is proposed by [Diebolt and Robert, 1994] to finite mixture models with a known order. Later, Escobar and West [1995] and Richardson and Green [1997] studied the case in which orders are unknown. The computational effort of MCMC methods can be intense partly, because of the label switching problem; see Jasra et al. [2005]. Moreover, MCMC methods can suffer from convergence issues; see [Robert and Casella, 2004].

1.4 Geometry of Mixture Models

In this section, we see the role of geometry in both inferences and computation of mixture models. Firstly, we review Lindsay's geometry in the likelihood space based on a finite sample. Afterwards, we look at Marriott's mixture affine geometry.

1.4.1 Mixture Models in Likelihood Spaces

A framework for computing the NPMLE is given in [Lindsay, 1995]. Firstly, we construct the feasible region in likelihood space:

$$\begin{aligned}\mathcal{L} = \{\boldsymbol{L} =&(L_1(Q), \cdots, L_N(Q))^{\mathrm{T}} \in \mathbb{R}^N, \\ &\text{where } Q \text{ is a probability measure over } \Theta\}, \end{aligned} \tag{1.1}$$

and for each n,

$$L_n(Q) = \int_\Theta f(x_n; \theta) dQ(\theta)$$

and N is the number of observations. Secondly, we define and maximize the following log-likelihood over $\mathcal{L}$:

$$\ell(\boldsymbol{L}) = \mathbf{1}^{\mathrm{T}} \log(\boldsymbol{L}),$$

where $\mathbf{1} = (1, \ldots, 1)^{\mathrm{T}} \in \mathbb{R}^N$. Let $\ell(\boldsymbol{L})$ be maximized at $\hat{\boldsymbol{L}}$. Finally, we reconstruct the NPMLE $\hat{Q}_{\mathrm{NPMLE}}$ from the equation

$$\hat{\boldsymbol{L}} = \boldsymbol{L}(\hat{Q}_{\mathrm{NPMLE}}).$$

Lindsay's fundamental theorems are based on the facts that $\mathcal{L}$ is the convex hull of the likelihood curve in likelihood space

$$\left\{ \mathtt{f}(\theta) = (f(x_1; \theta), \cdots, f(x_N; \theta))^{\mathrm{T}} \in \mathbb{R}^N \mid \theta \in \Theta \right\}, \tag{1.2}$$

and $\ell(\boldsymbol{L})$ is strictly concave. With the condition that $\mathtt{f}(\theta)$ has full rank in $\mathbb{R}^N$, it follows that $\hat{\boldsymbol{L}}$ is on the boundary of $\mathcal{L}$ and $\hat{Q}_{\mathrm{NPMLE}}$ has at most N support points; see [Lindsay, 1995]. Moreover, there is a supporting hyperplane $\mathcal{H}$ of $\mathcal{L}$ at $\hat{\boldsymbol{L}}$ such that

$$\boldsymbol{p}^{\mathrm{T}} \frac{\partial}{\partial \boldsymbol{L}} \ell(\hat{\boldsymbol{L}}) \leq 0, \quad \text{if } \boldsymbol{p} \in \mathcal{L},$$

and

$$\boldsymbol{p}^{\mathrm{T}} \frac{\partial}{\partial \boldsymbol{L}} \ell(\hat{\boldsymbol{L}}) = 0, \quad \text{if } \boldsymbol{p} \in \mathcal{H}.$$

It follows that the gradient function, which is defined as

$$\mathcal{G}(\theta) = \left(\mathtt{f}(\theta)\right)^{\mathrm{T}} \frac{\partial}{\partial \boldsymbol{L}} \ell(\hat{\boldsymbol{L}}), \quad \text{for } \theta \in \Theta,$$

is non-positive and zeros are achieved at support points of $\hat{Q}_{\text{NPMLE}}$; see [Lindsay, 1995]. These results allow the development of the class of gradient based algorithms discussed in the previous section.

Example 1.2 (continued).
The probability function $\text{Bin}(2, \theta)$ *forms a curve in the likelihood space, as shown in Figure 1.3 (a). The convex hull of the curve is the feasible region* $\mathcal{L}$. *Moreover, we also plot its supporting hyperplane at* $\text{Bin}(2, 0.5)$. *In Figure 1.3 (b), we plot the cone structure, which leads to* $0.5\chi^2_0 + 0.5\chi^2_1$ *as the asymptotic distribution of the* LRTS.

Geometry is also involved in deriving the asymptotic distribution of the non-parametric mixture model likelihood ratio test statistic. If the model surface can be approximated by score tangent cones, Chernoff [1954] proved that the asymptotic distribution theory of the likelihood ratio test can be generated by projecting the empirical likelihood onto these cones; also see [Shapiro, 1985]. This result leads to the fact that the limiting distributions of some non-parametric tests are a mixture of χ^2-distributions, see ; see [Lindsay, 1995].

1.4.2 Mixture Models in Affine Spaces

Marriott [2002] constructed an affine space $(\mathcal{X}_{\text{Mix}}, \mathcal{V}_{\text{Mix}}, +)$ for mixture models, where $\mathcal{X}_{\text{Mix}}$ and $\mathcal{V}_{\text{Mix}}$ are subsets of certain functional space with the form

$$\mathcal{X}_{\text{Mix}} = \left\{ f(x) \mid \int f(x) dx = 1 \right\}, \quad \text{and} \quad \mathcal{V}_{\text{Mix}} = \left\{ v(x) \mid \int v(x) dx = 0 \right\}$$

and $+$ is the natural addition operation. The local mixture models are introduced in a finite dimensional space $(\mathcal{X}_{\text{Mix}}, \mathcal{V}'_{\text{Mix}}, +)$, where

$$\mathcal{V}'_{\text{Mix}} = \text{span}\left\{ v_j(x) \in \mathcal{V}_{\text{Mix}}, j = 1, 2, \cdots, J \right\};$$

see [Marriott, 2002]. Later, Marriott [2007] extended them to Hilbert spaces. When the mixing distribution is not of primary interest and the component distributions

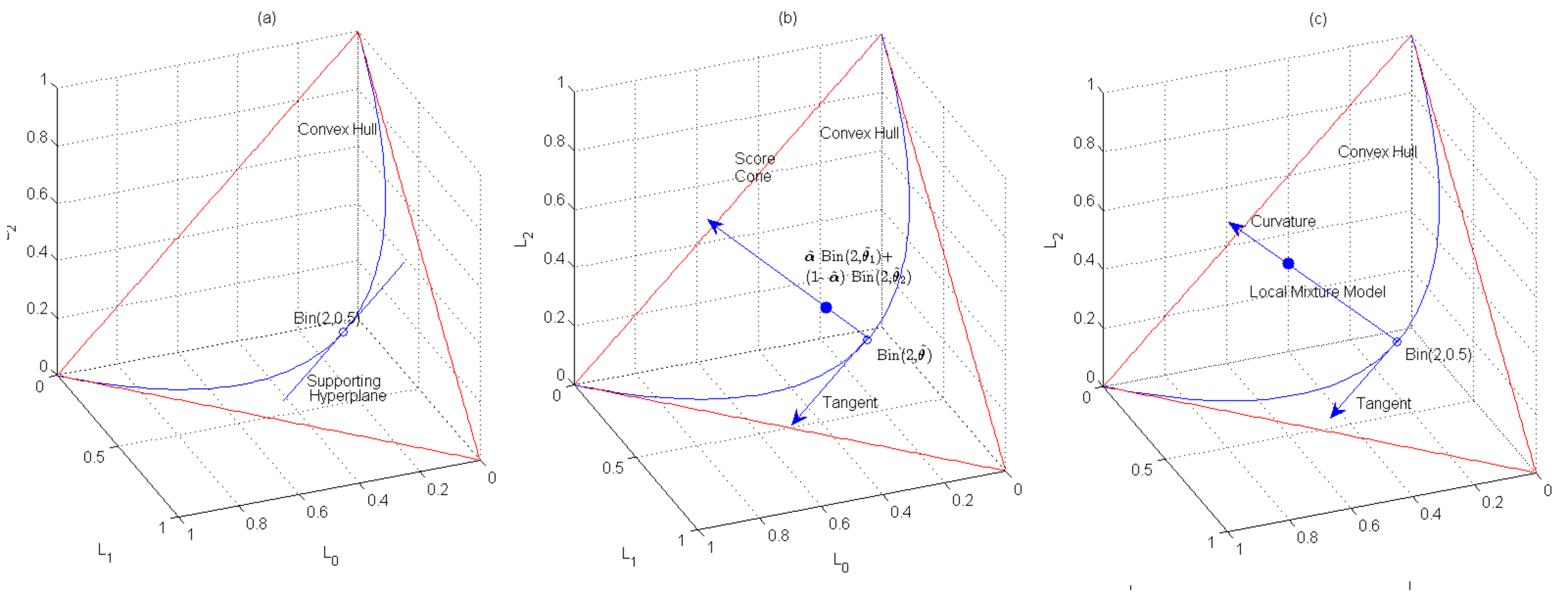

Figure 1.3: Plot of (a) the geometry in the likelihood space; (b) the score cone; (c) the geometry in the affine space.

are mixed locally, local mixture models are able to keep the flexibility with a small number of nuisance parameters; see [Marriott, 2002, 2007]. Other geometric properties of embedding a mixture model in an affine space can be seen in [Zhang, 2005], [Zhang and Hasto, 2006] and [Zhang, 2013].

Example 1.3. *(continued)*

The vector space $\mathcal{V}_{\text{Mix}}$ is spanned by its tangent and curvature vectors at $\theta = 0.5$ and approximated by the curvature only. And then, local mixture models locate on the curvature as shown in Figure 1.3 (c).

1.5 Mixed-Effects Models

Mixture models involving both fixed and random effects are called mixed-effects models; see [McCulloch and Neuhaus, 2005] for an introduction. Mixed-effects models are popular in longitudinal data analysis, because they are able to incorporate subject specific covariates and have a richer interpretation when the subject-specific effect is of interest; see [Diggle, 2002] and [Wu and Zhang, 2006].

Semi-parametric mixture models form a subclass of the mixed-effects models, where the random effects distribution is non-parametric. To fit semi-parametric mixture models, it is common to maximize the likelihood but this is computationally challenging and requires explicit full likelihood functions; see [Aitkin, 1999] and [Wang, 2010]. Methods based on subject-specific generalized estimating equations form another class of approaches for semi-parametric mixture models; see [Sutradhar and Godambe, 1997], [Vonesh et al., 2002] and [Wang et al., 2012]. This is more robust to the misspecification of the likelihood functions than the maximum likelihood methods, because no distribution assumption is made to the responses conditional on the random effects.

Inference on the regression parameter in a semi-parametric mixture model is more challenging than in a non-parametric mixture model. The challenges include the issued of identifiability, the boundary in the parameter space and the issue of dimension of the parameter space; see Section 1.2. When the random effects can be consistently predicated as the dimension of each response goes to infinity, inference on the regression parameter becomes possible in the methods based on subject-specific estimating equation methods; see [Vonesh et al., 2002] and [Wang et al., 2012].

A parametric distributional assumption, and typically a normal distribution is used, is made on the random effects distribution. Breslow and Clayton [1993] introduced the penalized quasi-likelihood (PQL) methods based on the Laplace approximation. This estimation method can be computed easily but is a biased estimator. Later, Lin and Breslow [1996] used a higher-order Laplace approximation and proposed a bias correction method for PQL estimators. The generalized estimating equation (GEE) method has been used for the mixed-effects model with normal random effects distribution in [Zeger et al., 1988].

Model misspecification on the random effects distribution has attracted much research interests. It was believed that inference on the regression parameters is quite robust; while inference on the random effects distribution itself is much less robust; see [Neuhaus et al., 1992]. However, the conditionally specified regression point estimators can result from using a simple random intercepts model when either the random effects distribution depends on measured covariates or there are autoregressive random effects; see in [Heagerty and Kurland, 2001].

1.6 Data Examples

To motivate both theoretical and methodological development in this thesis, a few real world dataset will be used for illustration throughout this thesis. This section begins with examples for mixture models and then describes examples for mixed-effects models.

1.6.1 Thailand Cohort Study Data

The description of the study is adapted from [Schlattmann, 2009]. To study the health status of 602 pre-school children, the number of times that a child who showed symptoms of fever, a cough, a runny nose, or these symptoms together, is recorded from June 1982 until September 1985.

The dataset has been studied in [Böhning et al., 1992] and [Schlattmann, 2009]. A Poisson distribution with $X \sim \text{Pois}(\theta)$ is often chosen as a parametric model for this kind of count data, where θ is the mean parameter. With the independent and identically distributed assumption, the maximum likelihood estimator of θ is 4.4485 in the single Poisson model. As we can see from Figure 1.4, this Poisson model does not fit the empirical distribution well.

Because a mixture of Poisson can take more variability into account than a single Poisson distribution, the following model is suggested. Consider a general mixture of Poisson distributions,

$$\int_{\Theta} \text{Pois}(\theta) \mathrm{d}Q(\theta),$$

where $Q(\theta)$ is an arbitrary probability measure over a compact set Θ.

1.6.2 Epileptic Seizures Data

Thall and Vail [1990] analyzed the data from a clinical trail of 59 epileptics, which aims to examine the effectiveness of the drug progabide in treating epileptic seizures. The outcomes are counts of epileptic seizures during four consecutive two-week periods. For each patient, the number of seizures in the eight weeks preceding entry into the trial and the age are recorded. Figure 1.5 displays a longitudinal plot of the data, where each trajectory represents a time series of a patient.

Let y_{nt} be the biweekly number of seizures for patient n at equally spaced time $t = 1, 2, 3, 4$, and let $\boldsymbol{x}_{nt}$ be the vector of covariates, including baseline seizure count, treatment, age and possibly the interaction between treatment and age. The following mixed model is used in [Breslow and Clayton, 1993]. For each n and t, it is assumed that y_{nt} follows Poisson distribution with mean $\mu_{nt}(\boldsymbol{b}_n)$, where

$$\log \mu_{nt}(\boldsymbol{b}_n) = \boldsymbol{x}_{nt}^{\mathrm{T}} \boldsymbol{\beta} + b_{n1} + v_t \times b_{n2} + b_{nt0},$$

where $\boldsymbol{\beta}$ is the regression parameter, v_t is the t^{th} visiting time, coded in $(-0.3, -0.1, 0.1, 0.3)$ from the first visiting time to the last, and $\boldsymbol{b}_n = (b_{n1}, b_{n2})^{\mathrm{T}} \in \mathbb{R}^2$ is a bi-variate normal

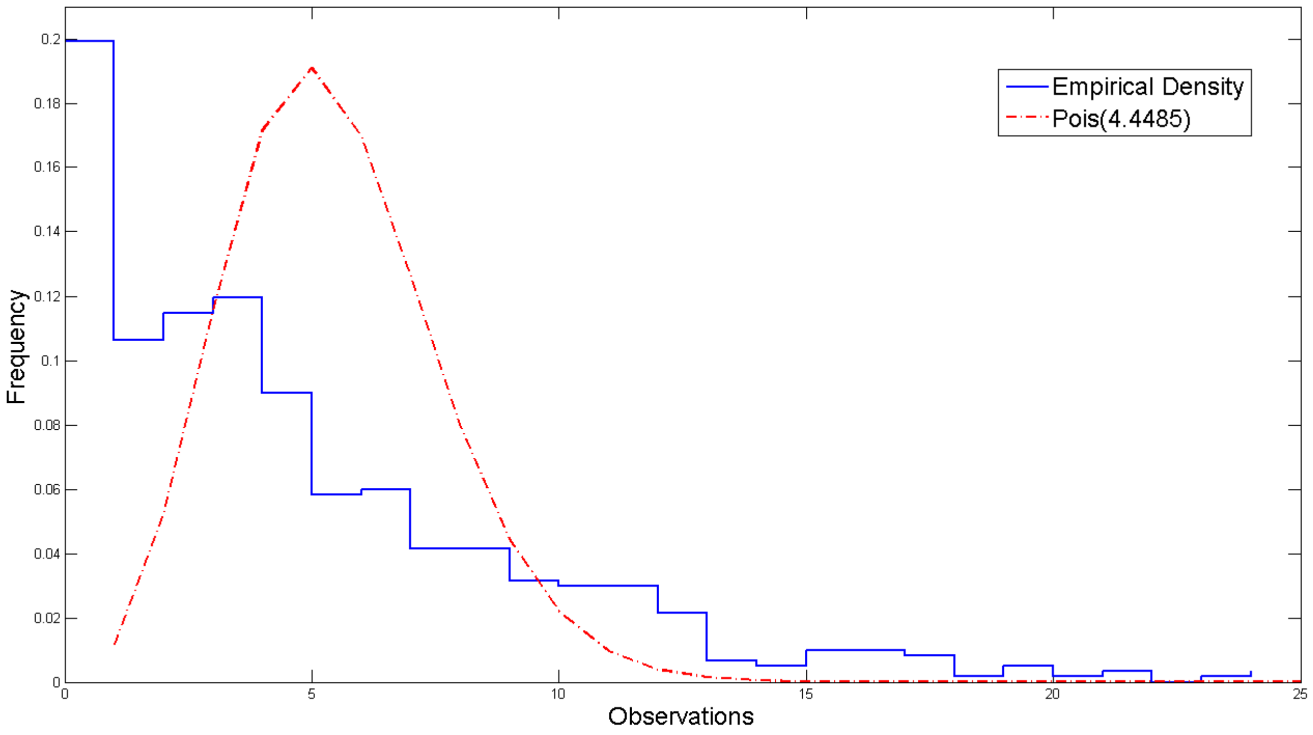

Figure 1.4: Plots of empirical density and Poisson density of the Thailand cohort study data.

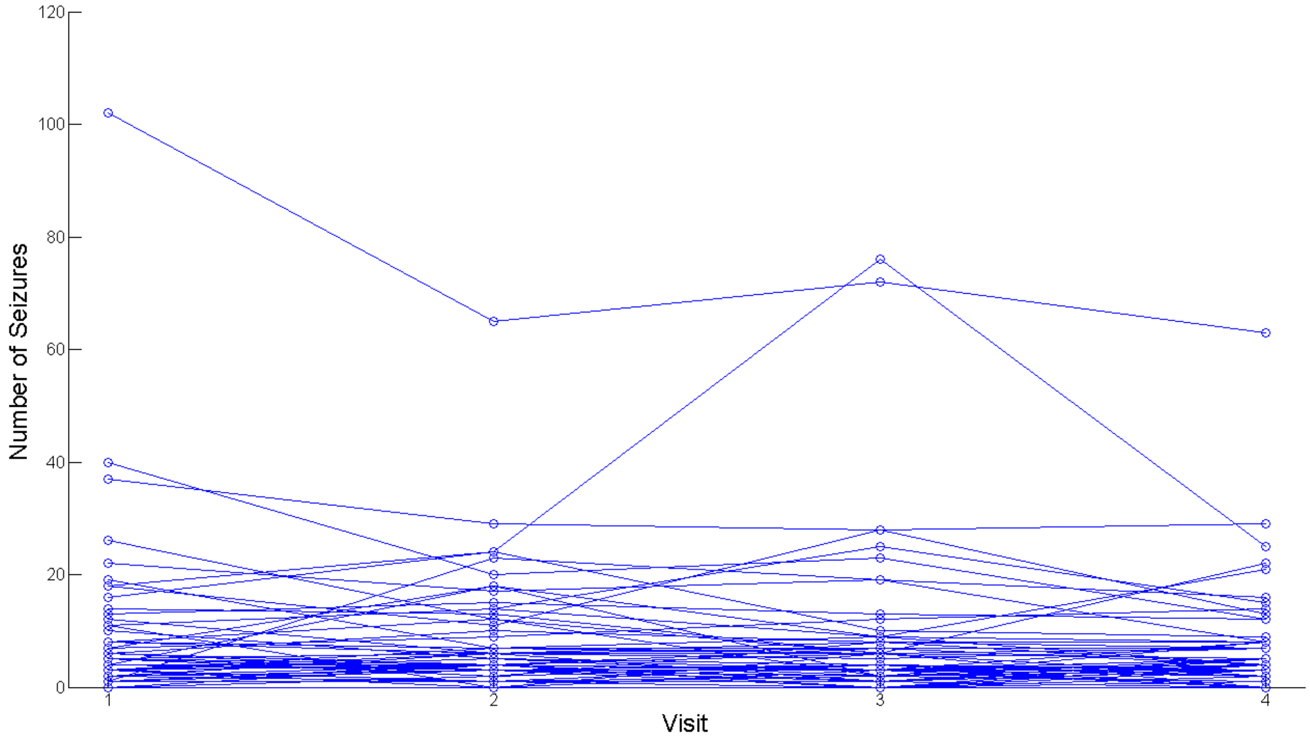

Figure 1.5: Plot of the epileptic seizures data.

random effect and b_{nt0} is additional random error terms that represent non-specific over-dispersion beyond that introduced by the subject-to-subject variation.

1.6.3 Retinal Surgery Data

Song and Tan [2000] considered data from a prospective study in ophthalmolgy where intraocular gas was used in complex retinal surgeries to provide internal tamponade of retinal breaks in the eye; also see [Song, 2007]. Three gas concentration levels were randomly administrated to 31 patients, who were then visited three to fifteen times over a three-month period after as injection. The outcome is the volume of the gas in the eyes of each patient at each follow-up visit, recorded as a percentage of the initial gas volume. The aim of this study is to estimate the decay rate of gas disappearance across three gas concentration levels. Figure 1.6 displays a longitudinal plot of the data, where each trajectory represents a time series of a patient.

Let y_{nt} be the percentage of gas volume for patient n at time t_n and let $\boldsymbol{x}_{nt}$ be the vector of covariates including the logarithm of time after surgery (in days) and its square, and the gas concentration level. Song [2007] suggested the following model. For each n and t, y_{nt} is assumed to follow a simplex distribution with mean $\mu_{nt}(b_n)$ and dispersion σ^2, where

$$\operatorname{logit}(\mu_{nt}(b_n)) = \boldsymbol{x}_{nt}^{\mathrm{T}}\boldsymbol{\beta} + b_{n0}, \tag{1.3}$$

and $\boldsymbol{\beta}$ is the regression parameter and $b_{n0} \in \mathbb{R}$ follows a normal distribution. Here the simplex distribution with mean $\theta \in (0,1)$ and dispersion parameter $\sigma^2 > 0$ has the density

$$f(x;\theta,\sigma^2) = \left(2\pi\sigma^2 \left(x(1-x)\right)^3\right)^{-1/2} \exp\left(-\frac{1}{2\sigma^2}\frac{(x-\theta)^2}{x(1-x)\mu^2(1-\theta)^2}\right);$$

see [Barndorff-Nielsen and Jørgensen, 1991].

1.7 Outline and Achievements of the Thesis

The structure and achievements of this thesis are as follows. In Chapter 2, a new reparameterization-approximation procedure for non-parametric mixture (or mixed-

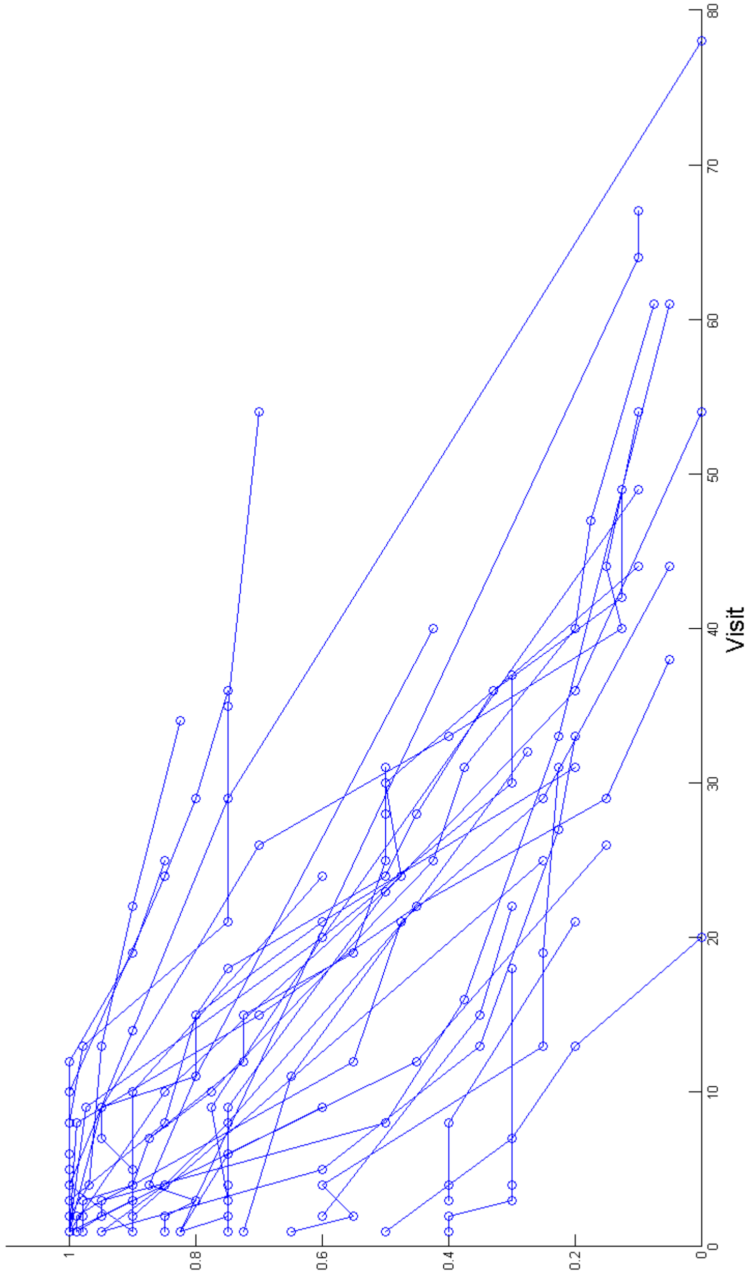

Figure 1.6: Plot of the retinal surgery data.

effects) models is proposed. By embedding the likelihood functions (or estimating functions) in a Hilbert space with countable dimension, the non-parametric mixing (or random-effects) distributions in these models are reparameterized by their generalized moments. Then, the reparameterized models are approximated by truncating the terms associated with higher order generalized moments. Though this new procedure, the dimension of the parameters in the considered models is successfully reduced from infinite to countable in the reparameterization step, and from countable to finite in the approximation step. The considered basis for Hilbert spaces include the eigenfunctions of an integral operator, and the families of orthogonal polynomials (the Chebyshev and Hermite polynomials). In this chapter, the orders of the residuals in the approximations are also computed as the number of generalized moments in a truncated model goes to infinity. Several examples are given, including the mixture of Poisson distributions, the mixture of normal distributions, and generalized linear mixture models with log-link functions, logit-link functions and tang-link functions. The materials of reparameterizing and approximating the mixture models in this chapter and the geometric properties of the generalized moment spaces in Chapter 3 are published in the journal paper "Parameterizing Mixture Models with Generalized Moments" accepted by the Annals of the Institute of Statistical Mathematics.

The major contribution in Chapter 3 is to derive two new important properties of the generalized moment spaces: the positive representation and the gradient characterization. The positive representation describes a sufficient and necessary condition that a probability measure can be uniquely determined by its generalized moments. It also gives an upper bound of the number of support points of a probability measure, when it can be uniquely reconstructed from its generalized moments. The gradient characterization is helpful in designing the class of the gradient-based computational algorithms to reconstruct probability measures from their generalized moments. Similar geometric properties are given in [Lindsay, 1995], when mixture models are in an embedding space determined by the observed sample. In this chapter, the generalized moment spaces induced by the power functions and the Chebyshev polynomials are studied in details as examples.

The generalized method of moments for mixture models is proposed as a new estimation method for mixture models in Chapter 4. The proposed method is based on the generalized moment conditions, which are obtained from the reparameterization-approximation procedure when the eigenfunctions of an integral operators are used as the basis for a Hilbert space. It involves reweighed projecting a sample generalized moments vector onto a generalized moment space. The weighting matrices in the proposed methods could be designed for different purposes. When weighting matrices are identity, the proposed method could be interpreted under the information geometry framework in [Zhang, 2013]. Another example is that the weighting matrices are designed to obtain the robustness of the proposed estimators when a data is contaminated with outliers. Because the geometry of the generalized moment spaces are well-studied in Chapter 3, computational algorithms, such as the CNM algorithm [Wang, 2007], can be easily adopted to compute the proposed estimators. The performance of the proposed estimators are investigated through simulation studies, and the proposed method is used to fit a model for the Thailand Cohort Study Data.

In Chapter 5, proposing the generalized method of moments for mixed-effects models with univariate random effects is the major contribution. This is a new estimation method for the considered models. Through the reparameterization-approximation procedure, the estimating functions, which are marginalized over the random effects, can be approximated by the functions with the regression parameters and the generalized moments of the random effects distribution. Next, the weighted L^2-norm of the vector of the approximated estimating functions is minimized over the parameter space (the generalized moment spaces for the generalized moments and the real space for the regression parameters). In this proposed method, distributional assumption on the random effects is not required. Therefore, the proposed estimators are robust to the misspecification of the likelihood functions. Simulation studies are conducted to investigative the performance of the proposed method and a random-effects model for the Retina Surgery Data is fitted by it.

The contribution in Chapter 6 is extending the method proposed in Chapter 5 to a Poisson regression model with random intercept and slope. Although the random effects distribution is bivariate, it is shown that the parameter space for the generalized moments is a generalized moment cone, which has same geometric properties as the generalized moment space. Through simulation studies, the robustness of the proposed estimators to the misspecification of the random effects distributions is studied. Moreover, a Poisson regression model with random intercept and slope for the Epileptic Seizures Data is fitted by the proposed method.

In Chapter 7, the asymptotic results of the generalized method of moments for mixed-effects models with univariate random effects are established. The major contribution include that the convergence rate of the proposed estimators is computed and the asymptotic normality in the proposed method is derived. These asymptotic results are obtained in the case where the dimension of the generalized moments diverges with the sample size and the dimension of the regression parameter is fixed. This is novel in the literature of mixed-effects models.

In Chapter 8, the idea of ensemble inference is firstly used to construct an asymptotically χ^2 test statistic for the hypothesis testing problems on the regression parameters in a mixed-effects model. Because the asymptotic results in Chapter 7 involve the generalized moments, which are unknown under the null hypothesis, there may not exist a pivotal statistic when only regression parameters are of interest. This is the motivation of using the ensemble idea. By using the generalized estimating functions with different weighting matrices, an asymptotically normal statistic is obtained in a space whose dimension is larger than the dimension of the generalized moment vectors. This asymptotically normal statistic is still not pivotal under the null hypothesis, because it involves the unknown generalized moments. However, by projecting this asymptotically normal statistic into a lower dimensional space, an asymptotically χ^2 test statistic, which is pivotal under the null hypothesis, can be constructed. Simulation studies show empirical evidence which supports this idea. And, the Epileptic Seizures Data is analyzed by the proposed inferential procedure.

Chapter 2

Reparameterization with Generalized Moments

2.1 Introduction

In this chapter, we consider the class of statistical models of the form

$$h_{\text{Mix}}(s;Q) = \mathbb{E}_\theta[h(s;\theta)] = \int_\Theta h(s;\theta)\mathrm{d}Q(\theta) < \infty, \tag{2.1}$$

where $h(s;\theta)$ is a known kernel function and is bounded over $(s,\theta) \in \mathcal{S} \times \Theta$, and Q is a probability measure over the set $\Theta \subseteq \mathbb{R}^q$. Both mixture and mixed-effects models can be written in the form of (2.1); see Section 2.3 and 2.4.

The model $h_{\text{Mix}}(s;Q)$ is nonparametric in the sense that no functional assumption is made on Q. Thus, the infinite dimensional parameter space could be a major challenge in statistical inference; see Section 1.2.3.

This chapter aims to solve this problem through a reparameterization-approximation procedure. In the reparameterization step, we reduce the infinite-dimensional parameter to a countable-dimensional parameter by rewriting the model $h_{\text{Mix}}(s;Q)$ in the form of a countable sum. Then, we further reduce the dimension of the parameter by truncating the higher-order terms.

We make the following contributions in this chapter. Firstly, we give a general framework for reparameterization of a mixture (or mixed-effects) model with a complete orthonormal basis in a Hilbert space. Further, the parameter in the reparameterized models can be interpreted as the generalized moments of the mixing distribution Q, which are induced by Chebyshev systems (see Definition 2.2.1).

Definition 2.1.1 (Generalized Moments)**.**
Let $\{u_j(\theta)\}_{j=0}^J$ with $u_0(\theta) \equiv 1$ form a Chebyshev system over Θ. For $j = 0, \ldots, J$, the j^{th} moment of a probability measure $Q(\theta)$ induced by $\{u_j(\theta)\}_{j=0}^J$ is defined as

$$m_j(Q) = \mathbb{E}_\theta[u_j(\theta)] = \int_\Theta u_j(\theta) \mathrm{d}Q(\theta) < \infty, \tag{2.2}$$

where θ has a probability measure Q on Θ.

Secondly, under the proposed reparameterization framework, we apply different orthonormal basises, including families of orthogonal polynomials and eigenfunctions from positive self-adjoint integral operators to mixture (or mixed-effects) models.

Lastly, we study the approximation error of the truncation approximation, when the basis in the expansion is the eigenfuctions of an integral operator and the Chebyshev or Hermite polynomials.

This chapter is organized as follows. In Section 2.2, we give an overview of the reparameterization-approximation procedure. The reparameterization technique is based on the orthonormal expansion of $h(s;\theta)$ in Hilbert spaces. Because $h(s;\theta)$ can be considered either a function of s given θ or a function of θ given s, we expand it in two different Hilbert spaces; see Subsection 2.2.1 and 2.2.2. To disclose the relationship between the new parameters and the mixing distribution Q, we interpreted them as the generalized moments of Q; see Subsection 2.2.3. In the last subsection, we describe the way to examine the quality of the truncation approximations.

In Section 2.3, we consider one-parameter mixture models (see Definition 2.3.1). Two types of orthonormal basis are considered. One is the class of orthogonal polynomials for natural exponential families with quadratic variance function (NEF-QVF); see Subsection 2.3.1. The other is the eigenfunctions from an integral operator for the exponential families; see Subsection 2.3.2. In the last subsection, we discuss the quality of the truncation approximations in detail.

In Section 2.4, we apply the reparameterization-approximation procedure to the inverse of the link functions in mixed-effects models. We use the Chebyshev polynomials (Subsection 2.4.1) and the Hermite polynomials (Subsection 2.4.2) for the case when the random effects are univariate. In Subsection 2.4.3, we consider the case when the random effects are multivariate and apply tensor product basises to a mixed-effects model with a bivariate random effect.

In the last part of this chapter, we give some foundational theory including the strictly totally positive kernel functions and the asymptotic coefficients of the expansions by the Chebyshev polynomials and the Hermite polynomials. Also, the proofs of theorems can be found in Appendix A.

2.2 Reparameterization Framework

2.2.1 Reparameterization with $L^2(\mathcal{S}, \nu_0)$

Consider a measure space $(\mathcal{S}, \Upsilon, \nu_0)$ and the $L^2(\mathcal{S}, \nu_0)$ space induced by it. We assume that ν_0 is a measure with support $\mathcal{S}$. We will also denote, where appropriate, $\nu_0(s) = h_0(s)\nu(s)$ with respect to a fixed measure ν, typically Lebesgue or a counting measure, and $h_0(s)$ is a strictly positive function of $s \in \mathcal{S}$.

Let $\{e_j(s)\}_{j=0}^{\infty}$ be a complete orthonormal system in $L^2(\mathcal{S}, \nu_0)$, i.e., for each i and j,

$$\langle e_i(s), e_j(s) \rangle_{L^2(\mathcal{S},\nu_0)} = \int_{\mathcal{S}} e_i(s) e_j(s) \mathrm{d}\nu_0 = \delta_{ij},$$

where δ_{ij} is the Kronecker delta. We make the following assumption.

Regularity Condition 2.A (Square integrable of $h(s;\theta)/h_0(s)$).
For each $\theta \in \Theta$, the function $h(s;\theta)/h_0(s) \in L^2(\mathcal{S}, \nu_0)$, i.e., for each $\theta \in \Theta$,

$$\int_{\mathcal{S}} \left(\frac{h(s;\theta)}{h_0(s)} \right)^2 \mathrm{d}\nu_0 < \infty.$$

According to the standard results in Hilbert spaces [Debnath and Mikusiński, 1999, p.g. 87-130], we have the expansion, for each $(s, \theta) \in \mathcal{S} \times \Theta$,

$$h(s; \theta) = \sum_{j=0}^{\infty} u_j(\theta) e_j(s) h_0(s),$$

where for each j,

$$u_j(\theta) = \langle e_j(s), h(s; \theta)/h_0(s) \rangle_{L^2(\mathcal{S}, \nu_0)} .$$

Additional to the boundedness of $h_{\text{Mix}}(s; Q)$ over $\mathcal{S}$, the order of the integral and the infinite sum in

$$h_{\text{Mix}}(s; Q) = \int_{\Theta} \sum_{j=0}^{\infty} u_j(\theta) e_j(s) h_0(s) \mathrm{d}Q$$

are exchangeable by the Fubini's theorem. Therefore, we have the reparameterization

$$h_{\nu_0}(s; \boldsymbol{m}_{\infty}) = \sum_{j=0}^{\infty} m_j(Q) e_j(s) h_0(s), \tag{2.3}$$

where $\boldsymbol{m}_{\infty} = (m_0, m_1, \ldots)^{\mathrm{T}} \in \mathbb{R}^{\infty}$ for each j,

$$m_j(Q) = \int_{\Theta} u_j(\theta) \mathrm{d}Q < \infty. \tag{2.4}$$

After reparameterization, we have a model with a countable-dimensional parameter $\boldsymbol{m}_{\infty}$.

We may approximate the countable sum in (2.3) with a finite sum by truncating the higher-order terms. It is

$$h_{\nu_0}(s; \boldsymbol{m}_J) = \sum_{j=0}^{J} m_j e_j(s) h_0(s), \tag{2.5}$$

where $\boldsymbol{m}_J = (m_0, \ldots, m_J)^{\mathrm{T}} \in \mathbb{R}^{J+1}$ and for each j, m_j is defined in (2.4). If $h_{\text{Mix}}(s; Q)$ can be approximated by $h_{\nu_0}(s; \boldsymbol{m}_J)$ appropriately, we successfully reduce the dimension of the parameter from infinity to finite with a loss which we will quantify.

2.2.2 Reparameterization with $L^2(\Theta, \mu_0)$

In the previous section, the reparameterization is based on the orthonormal expansion of $h(s;\theta)$ in the Hilbert space $L^2(\mathcal{S}, \nu_0)$. In this section, we consider the expansion in the $L^2(\Theta, \mu_0)$ space induced by the measure space $(\Theta, \Upsilon', \mu_0)$, where μ_0 is a measure with support Θ. Let $\mu_0(\theta) = w_0(\theta)\mu(\theta)$ with respect to a fixed measure μ, typically Lebesgue or a counting measure, and $w_0(\theta)$ is a strictly positive function of $\theta \in \Theta$.

Let $\{\upsilon_j(\theta)\}_{j=0}^{\infty}$ be a complete orthonormal system in $L^2(\Theta, \mu_0)$. We make the following assumption.

Regularity Condition 2.B (Square integrable of $h(s;\theta)/w_0(\theta)$)**.**
For each $s \in \mathcal{S}$, the function $h(s;\theta)/w_0(\theta) \in L^2(\Theta, \mu_0)$, i.e., for each $s \in \mathcal{S}$,

$$\int_{\Theta} \left(\frac{h(s;\theta)}{w_0(\theta)} \right)^2 \mathrm{d}\mu_0 < \infty.$$

Similar to the previous section, we have the expansion in $L^2(\Theta, \mu_0)$ that for each $(s, \theta) \in \mathcal{S} \times \Theta$,

$$h(s;\theta) = \sum_{j=0}^{\infty} \upsilon_j(\theta)\varphi_j(s)w_0(\theta),$$

where for each j,

$$\varphi_j(s) = \langle \upsilon_j(\theta), h(s;\theta)/w_0(\theta) \rangle_{L^2(\Theta,\mu_0)}.$$

Because $h_{\mathrm{Mix}}(s;Q)$ is bounded, we can change the order of the integrals by the Fubini's theorem and have the reparameterization

$$h_{\mu_0}(s; \boldsymbol{m}_\infty) = \sum_{j=0}^{\infty} m_j(Q)\varphi_j(s), \tag{2.6}$$

where $\boldsymbol{m}_\infty = (m_0, m_1, \ldots)^{\mathrm{T}} \in \mathbb{R}^\infty$ for each j,

$$m_j(Q) = \int_{\Theta} \upsilon_j(\theta) w_0(\theta) \mathrm{d}Q < \infty. \tag{2.7}$$

Again, we have a model with a countable-dimensional parameter $\boldsymbol{m}_\infty$ after reparameterization.

We further approximate (2.6) with a finite sum by truncating the higher order terms. It is

$$h_{\mu_0}(s;\boldsymbol{m}_J) = \sum_{j=0}^{J} m_j \varphi_j(s), \tag{2.8}$$

where $\boldsymbol{m}_J = (m_0, \ldots, m_J)^{\mathrm{T}} \in \mathbb{R}^{J+1}$ and for each j, m_j is defined in (2.7). When the truncation approximation $h_{\mu_0}(s;\boldsymbol{m}_J)$ is appropriate, the dimension of the model is reduced to finite.

Reparameterizating $h_{\mathrm{Mix}}(s;Q)$ with either $L^2(\mathcal{S},\nu_0)$ or $L^2(\Theta,\mu_0)$ depends on the kernel function $h(s;\theta)$. When $h(s;\theta)$ is a NEF-QVF, the reparameterization with $L^2(\mathcal{S},\nu_0)$ is natural; see Subsection 2.3.1. When $h(s;\theta)$ is the inverse of a link function in a mixed-effects model, reparameterization with $L^2(\Theta,\mu_0)$ would efficiently reduce the parameters in the truncation approximation models; see Subsection 2.4.1 and 2.4.2. In certain cases, the two reparameterizations are equivalent; see Subsection 2.3.2.

2.2.3 Interpretation of the Parameters

The generalized moments of a distribution have been defined in Equation (2.2). To interpret the parameters $\boldsymbol{m}_J$ in the truncation approximations (Equation (2.5) and (2.8)) as the generalized moments, we need $\{u_j(\theta)\}_{j=0}^J$ (or $\{\upsilon_j(\theta)w_0(\theta)\}_{j=0}^J$) to form a Chebyshev system (defined as follows) with $u_0(\theta) \equiv 1$ (or $\upsilon_0(\theta)w_0(\theta) \equiv 1$).

Definition 2.2.1 (Chebyshev Systems)**.**
The set of functions $\{u_j(\theta)\}_{j=0}^J$ is a Chebyshev system over $\Theta \subseteq \mathbb{R}$, if we have that $\det(u_i(\theta_j))_{i,j=0}^J > 0$ whenever $\theta_0 < \cdots < \theta_J$ and $\theta_j \in \Theta$, $j = 0, \ldots, J$.

The class of Chebyshev system is wide; see [Karlin and Studden, 1966, p.g. 9-20] for examples. An orthonormal basis is not necessarily a Chebyshev system. On the other hand, a Chebyshev system is not necessarily orthonormal. However, we can obtain an orthonormal basis from a Chebyshev system by the Gram-Schmidt process. We further prove that this orthonormal basis is still a Chebyshev system. The following result can not be found in the existing literature.

Theorem 2.2.1.
Let $\{u_j(\theta)\}_{j=0}^J$ with $u_0 \equiv 1$ be a set of functions in $L^2(\Theta, \mu_0)$ such that $\{u_j(\theta)\}_{j=0}^{J-1}$ and $\{u_j(\theta)\}_{j=0}^J$ form two Chebyshev systems over Θ. Also let $\{v_j(\theta)\}_{j=0}^J$ be the orthonormal basis obtained by applying a Gram-Schmidt process sequentially to $u_j(\theta)$, $j = 0, \dots, J$, in $L^2(\Theta, \mu_0)$. If each of $\{v_j(\theta)\}_{j=0}^J$ is multiply appropriately by ± 1, the set of orthonormal functions is converted into a Chebyshev system, defined in Definition 2.2.1.

The truncation approximations can be written as

$$h_{\nu_0}(s; \boldsymbol{m}_J) = e_0(s)h_0(s) + \sum_{j=1}^J m_j e_j(s) h_0(s),$$

when $u_0(\theta) \equiv 1$, or

$$h_{\mu_0}(s; \boldsymbol{m}_J) = \int_\Theta h(s;\theta)/w_0(\theta)\mathrm{d}\theta + \sum_{j=1}^J m_j \varphi_j(s),$$

when $v_0(\theta)w_0(\theta) \equiv 1$. According to the above expressions, the truncation approximations are locally defined by $\{e_j(s)h_0(s)\}_{j=1}^J$ at $e_0(s)h_0(s)$, or $\{\varphi_j(s)\}_{j=1}^J$ at $\int_\Theta h(s;\theta)/w_0(\theta)\mathrm{d}\theta$.

2.2.4 Examination of the Truncation Approximations

In this subsection, we describe the way to examine approximation qualities of the truncation approximations.

Let $\epsilon_{\nu_0,J}(s;Q)$ be the approximation error of $h_{\nu_0}(s;\boldsymbol{m}_J)$, i.e.,

$$\epsilon_{\nu_0,J}(s;Q) = \int_\Theta \epsilon_{\nu_0,J}(s;\theta)\mathrm{d}Q,$$

where

$$\epsilon_{\nu_0,J}(s;\theta) = \sum_{j=J+1}^\infty e_j(s)h_0(s)u_j(\theta).$$

Also let $\epsilon_{\mu_0,J}(s;Q)$ be the approximation error of $h_{\mu_0}(s;\boldsymbol{m}_J)$, i.e.,

$$\epsilon_{\mu_0,J}(s;Q) = \int_\Theta \epsilon_{\mu_0,J}(s;\theta)\mathrm{d}Q,$$

where

$$\epsilon_{\mu_0,J}(s;\theta) = \sum_{j=J+1}^{\infty} \varphi_j(s)\upsilon_j(\theta)w_0(\theta).$$

In the reparameterization with $L^2(\mathcal{S},\nu_0)$, it is natural to evaluate the approximation error $\epsilon_{\nu_0,J}(s;Q)$ using the norm defined in $L^2(\mathcal{S},\nu_0)$ that is

$$\begin{aligned}\left\|\frac{\epsilon_{\nu_0,J}(s;Q)}{h_0(s)}\right\|^2_{L^2(\mathcal{S},\nu_0)} &= \int_{\mathcal{S}} \left(\frac{\epsilon_{\nu_0,J}(s;Q)}{h_0(s)}\right)^2 \mathrm{d}\nu_0 \\ &= \sum_{j=J+1}^{\infty} \left(\int_\Theta u_j(\theta)\mathrm{d}Q\right)^2. \end{aligned} \tag{2.9}$$

However, this norm in $L^2(\mathcal{S},\nu_0)$ can not be applied to the evaluation of $\epsilon_{\mu_0,J}(s;Q)$, because $h(s;\theta)/w_0(\theta)$ may not belong to the space $L^2(\mathcal{S},\nu_0)$.

Note that $\epsilon_{\nu_0,J}(s;Q)$ (or $\epsilon_{\mu_0,J}(s;Q)$) can be written as a convex combination of $\epsilon_{\nu_0,J}(s;\theta)$ (or $\epsilon_{\mu_0,J}(s;\theta)$). We can study $\epsilon_{\nu_0,J}(s;\theta)$ (or $\epsilon_{\mu_0,J}(s;\theta)$) point-wisely over $(s,\theta)\in\mathcal{S}\times\Theta$. If $\epsilon_{\nu_0,J}(s;\theta)$ (or $\epsilon_{\mu_0,J}(s;\theta)$) is uniformly small over $\mathcal{S}\times\Theta$, we conclude that the truncation approximation $h_{\nu_0}(s;\boldsymbol{m}_J)$ (or $h_{\mu_0}(s;\boldsymbol{m}_J)$) is appropriate.

2.3 Reparameterization in Mixture Models

The importance and challenges in mixture models has been described in Chapter 1. This section considers the one-parameter mixture models defined as follows.

Definition 2.3.1 (One-parameter Mixture Models)**.**
Let $f(x;\theta)$ be a parametric density function which comes from a known family of distributions

$$\{f(x;\theta) \mid \theta\in\Theta\subseteq\mathbb{R}\}.$$

Let Q be a probability measure over Θ. Then, the distribution with the following density function is a one-parameter mixture model

$$f_{\mathrm{Mix}}(x;Q) = \int_{\Theta} f(x;\theta)\mathrm{d}Q(\theta),$$

where $f(x;\theta)$ is called the component distribution, $Q(\theta)$ the mixing distribution and θ the mixing parameter.

The one-parameter mixture models is a subclass of the statistical models in (2.1), because

$$f_{\mathrm{Mix}}(x;Q) = \mathbb{E}_{\theta}\left[f(x;\theta)\right],$$

by setting $s = x$ and $h(s;\theta) = f(s;\theta)$.

In Subsection 2.3.1 and 2.3.2, we give two examples of reparameterizing the one-parameter mixture models under the framework given in Section 2.2. In Subsection 2.3.3, we examine the approximation quality in each example.

2.3.1 Moments induced by Power Functions

We consider the one-parameter mixture models, in which the mixing parameter θ is the mean of the component distribution $f(x;\theta)$ and the mixing distribution $Q(\theta)$ has mean θ_0. Furthermore, the component distributions are natural exponential models with quadratic variance functions. This class includes the normal, Poisson, gamma, binomial and negative binomial families; see [Morris, 1982, 1983], and has the following formal definition.

Definition 2.3.2 (NEF-QVF).
If $f(x;\theta)$ is a natural exponential family in the mean parameterization, then $V_f(\theta)$, defined by $V_f(\theta) = \mathbb{E}_X[(X-\theta)^2]$, is called the variance function. If the variance function $V_f(\theta)$ is quadratic with the form $V_f(\theta) = c_0 + c_1\theta + c_2\theta^2$, then we say $f(x;\theta)$ is a natural exponential family with quadratic variance function.

When the mixing distribution is localized at θ_0, the mixture model $f_{\mathrm{Mix}}(x;Q)$ can be expanded by the Laplace expansion; see [Marriott, 2002]. Here we describe this

NEF-QVF	Polynomial
Normal	Hermite
Poisson	Poisson-Charlier
Gamma	Generalized Laguerre
Binomial	Krawtchouk
Negative Binomial	Meixner

Table 2.1: NEF-QVF and their associated orthonormal polynomials.

process in the view of expanding by orthonormal basis. Following [Morris, 1982], we define, for $j = 0, 1, \dots,$

$$P_j(x;\theta) = \frac{V_f^j(\theta)}{f(x;\theta)} \frac{\partial^j}{\partial \theta^j} f(x;\theta),$$

where $a_j = j! \prod_{i=0}^{j-1}(1 + ic_2) \equiv j! b_j$ and $V_f^j(\theta)$ is the j^{th} power of the variance function $V_f(\theta)$. The set of $\{P_j(x;\theta)\}_{j=0}^{\infty}$ forms an orthogonal polynomial system in the sense that

$$\langle P_i(x;\theta_0), P_j(x;\theta_0) \rangle_{L^2(\mathcal{S},\nu_0)} = \delta_{ij} a_j V_f^j(\theta_0),$$

where $\mathrm{d}\nu_0(x) = f(x;\theta_0)\mathrm{d}x$; see [Morris, 1982]. For each j, it can be shown through algebra that

$$\left\langle P_j(x;\theta_0), \frac{f(x;\theta)}{f(x;\theta_0)} \right\rangle_{L^2(\mathcal{S},\nu_0)} = \int_{\mathcal{S}} P_j(x;\theta_0) f(x;\theta)\mathrm{d}x = b_j(\theta - \theta_0)^j;$$

see [Morris, 1982]. Morris [1982] also pointed out that the orthogonal polynomial systems $\{P_j(x;\theta_0)\}$ are associated with different NEF-QVF; see Table 2.1.

For a given $\theta_0 \in \Theta$, a mixture of NEF-QVF can be reparameterized as

$$f_{\mathrm{Ty}}(x;\boldsymbol{m}_\infty) = f(x;\theta_0) + \sum_{j=2}^{\infty} m_j(Q) \frac{1}{j!} \frac{P_j(x;\theta_0)}{V_f^j(\theta_0)} f(x;\theta_0),$$

where for each $j = 1, 2, \dots,$

$$m_j(Q) = \int_{\Theta} (\theta - \theta_0)^j \mathrm{d}Q(\theta).$$

Its truncation approximation is

$$f_{\mathrm{Ty}}(x;\boldsymbol{m}_J)=f(x;\theta_0)+\sum_{j=2}^{J}m_j(Q)\frac{1}{j!}\frac{P_j(x;\theta_0)}{V_f^j(\theta_0)}f(x;\theta_0), \tag{2.10}$$

where $\boldsymbol{m}_J=(\theta_0,m_2,\dots,m_J)^{\mathrm{T}}\in\mathbb{R}^J$. Because $\boldsymbol{m}_J$ is induced by the set of the power functions $\{(\theta-\theta_0)^j\}_{j=0}^J$, which forms a Chebyshev system, $f_{\mathrm{Ty}}(x;\boldsymbol{m}_J)$ is parameterized in the moments induced by the power functions.

2.3.2 Moments induced by Eigenfunctions of an Integral Operator

This subsection considers the mixture models whose the component distributions $f(x;\theta)$ are in the exponential family and Θ is a compact set in $\mathbb{R}$. We further assume that, for each $\theta\in\Theta$, $f(x;\theta)/f_0(x)\in L^2(\mathcal{S},\nu_0)$ where $\mathrm{d}\nu_0=f_0(x)\mathrm{d}x$, and

$$f_0(x)=\frac{1}{|\Theta|}\int_\Theta f(x;\theta)\mathrm{d}\theta>0,\qquad\text{for } x\in\mathcal{S}$$

and $|\Theta|$ be the Lebesgue measure of Θ. Note that $f_0(x)$ is well defined because Θ is Lebesgue measurable under the compactness condition.

According to Equation (2.9), we see that the norm of $\epsilon_{\nu_0,J}(x;Q)/f_0(x)$ in $L^2(\mathcal{S},\nu_0)$ depends on the unknown mixing distribution Q. Therefore, it is hard to find an orthonormal basis in $L^2(\mathcal{S},\nu_0)$ which is optimal in the sense that the approximation error is minimized. However, we may minimize the following upper-bound instead. By the Cauchy-Schwarz inequality, we obtain the following upper-bound

$$\left\|\frac{\epsilon_{\nu_0,J}(s;Q)}{h_0(s)}\right\|^2_{L^2(\mathcal{S},\nu_0)}\le\int_\Theta\left(\frac{\mathrm{d}}{\mathrm{d}\theta}Q\right)^2\mathrm{d}\theta\times\sum_{j=J+1}^{\infty}\int_\Theta u_j^2(\theta)\mathrm{d}\theta.$$

When Q is either discrete or continuous on the compact set Θ, $(\mathrm{d}/\mathrm{d}\theta)\,Q$ is bounded. Therefore, minimizing this upper-bound is equivalent to minimizing

$$\sum_{j=J+1}^{\infty}\int_\Theta u_j^2(\theta)\mathrm{d}\theta. \tag{2.11}$$

The optimal orthonormal basis in $L^2(\mathcal{S}, \nu_0)$ can be found through a spectral decomposition of an integral operator. Let $\phi_j(x)$ be the eigenfunction associated with the j^{th} largest eigenvalue of the integral operator

$$(Ag)(s) = \int_{\mathcal{S}} g(x)K(x, x')\mathrm{d}x < \infty, \tag{2.12}$$

with the kernel function

$$K(x, x') = \int_{\Theta} \frac{f(x;\theta)}{f_0^{1/2}(x)} \frac{f(x';\theta)}{f_0^{1/2}(x')} \mathrm{d}\theta, \quad (x, x') \in \mathcal{S} \times \mathcal{S}.$$

This integral operator is positive and self-adjoint, and thus its eigenvalues are all positive; see [Debnath and Mikusiński, 1999, Section 4.4 and 4.6]. Because for any $\theta \in \Theta$

$$f(x;\theta) \leq \int_{\Theta} f(x;\theta)\mathrm{d}\theta = |\Theta| f_0(x),$$

we have, for each $\theta \in \Theta$,

$$\int_{\Theta}\int_{\mathcal{S}} f^2(x;\theta)/f_0(x)\mathrm{d}x\mathrm{d}\theta \leq \int_{\Theta}\int_{\mathcal{S}} |\Theta| f(x;\theta)\mathrm{d}x\mathrm{d}\theta = |\Theta|^2.$$

Therefore, the integral operator $A(\cdot)$ is Hilbert-Schmidt and thus compact. It follows that the set of $\{\phi_j(x)/f_0^{1/2}(x)\}_{j=0}^{\infty}$ forms the complete orthonormal basis in $L^2(\mathcal{S}, \nu_0)$ which minimizes (2.11), by the results of the functional principle decomposition in [Horváth and Kokoszka, 2012].

Expanding $f(x;\theta)$ with the basis $\{\phi_j(x)/f_0^{1/2}(x)\}_{j=0}^{\infty}$, we have

$$f(x;\theta) = f_0(x) + \sum_{j=1}^{\infty} \sqrt{\lambda_j}\gamma_j(\theta)\phi_j(x)f_0^{1/2}(x),$$

where for each j, λ_j is the j^{th} largest eigenvalue of $A(\cdot)$ and

$$\gamma_j(\theta) = \frac{1}{\sqrt{\lambda_j}} \int_{\mathcal{S}} f(x;\theta)\phi_j(x)f_0^{-1/2}(x)\mathrm{d}x. \tag{2.13}$$

The reparameterization of $f_{\text{Mix}}(x;Q)$ is

$$f_{\text{spec}}(x;\boldsymbol{m}_\infty) = f_0(x) + \sum_{j=1}^{\infty} \sqrt{\lambda_j} m_j \phi_j(x) f_0^{1/2}(x), \tag{2.14}$$

and the truncation approximation is

$$f_{\text{spec}}(x; \boldsymbol{m}_J) = f_0(x) + \sum_{j=1}^{J} \sqrt{\lambda_j} m_j \phi_j(x) f_0^{1/2}(x), \tag{2.15}$$

where for each j,

$$m_j = \int_{\Theta} \gamma_j(\theta) \mathrm{d}Q.$$

With the following theorem, we can show that the reparameterization with $L^2(\mathcal{S}, \nu_0)$ is equivalent to the reparameterization with $L^2(\Theta, \mu_0)$ in this case.

Theorem 2.3.1 (Representation of $\gamma_j(\theta)$)**.**
The eigenvalues λ_j, $j = 0, 1, \ldots,$ of $A(\cdot)$ are also the eigenvalues of the integral operator

$$(A'g)(s) = \int_{\Theta} g(\theta) K'(\theta, \theta') \mathrm{d}\theta < \infty,$$

with the kernel function

$$K'(\theta, \theta') = \int_{\mathcal{S}} \frac{f(x;\theta)}{f_0^{1/2}(x)} \frac{f(x;\theta')}{f_0^{1/2}(x)} \mathrm{d}x, \quad (\theta, \theta') \in \Theta \times \Theta. \tag{2.16}$$

Moreover, the function $\gamma_j(\theta)$ is the eigenfunction associated with the j^{th} largest eigenvalue of $A'(\cdot)$.

Proof. See the Appendix. □

According to Theorem 2.3.1, the set of functions $\{\gamma_j(\theta)\}_{j=0}^{\infty}$ forms an orthonormal basis in the space $L^2(\Theta, \mu)$, where μ is the Lebesgue measure, i.e., for each i and j,

$$\int_{\Theta} \gamma_i(\theta) \gamma_j(\theta) \mathrm{d}\mu = \delta_{ij}.$$

It is also true that, for each $j = 0, 1, \ldots,$

$$\begin{aligned} \langle f(x;\theta), \gamma_j(\theta) \rangle_{L^2(\Theta,\mu)} &= f_0^{1/2}(x) \frac{1}{\sqrt{\lambda_j}} (A'\phi_j)(x) \\ &= \sqrt{\lambda_j} \phi_j(x) f_0^{1/2}(x). \end{aligned} \tag{2.17}$$

Therefore, Equation (2.14) can also be viewed as a reparameterization with $L^2(\Theta, \mu)$.

Next, we show that $\{\gamma_j(\theta)\}_{j=0}^J$ forms a Chebyshev system with $\gamma_0(\theta) \equiv 1/\sqrt{|\Theta|}$. And thus, the parameter $\boldsymbol{m}_J$ in Equation (2.15) are the generalized moments of Q induced by the eigenfunctions of $A'(\cdot)$.

Theorem 2.3.2.
For each $J = 1, 2, \ldots$, the set $\{\gamma_j(\theta)\}_{j=0}^J$ forms a Chebyshev system over Θ. Moreover, $\gamma_0(\theta) \equiv 1/\sqrt{|\Theta|}$.

Proof. See the Appendix. □

According to Equation (2.15), we have

$$|f_{\text{spec}}(x; \boldsymbol{m}_J) - f_{\text{Mix}}(x; Q)| = \sum_{j=J+1}^{\infty} O(\sqrt{\lambda_j}), \tag{2.18}$$

when for each j, m_j is bounded. The decay of the eigenvalues is related to the smoothness of the kernel function $K'(\theta, \theta')$, $(\theta, \theta') \in \Theta \times \Theta$; see [Reade, 1983] and [Ha, 1986].

Proposition 2.3.1 ([Ha, 1986])**.**
If $K'(\theta, \theta')$ is positive definite and symmetric, and if the symmetric derivative

$$\frac{\partial^{2r}}{\partial\theta^r \partial\theta'^r} K(\theta, \theta')$$

exists and is continuous on $\Theta \times \Theta$, then for large j,

$$\lambda_j = O(j^{-2r-1}).$$

Applying Equation (A.3) to Equation (2.18), we have the following result.

Corollary 2.3.1.
Suppose that the symmetric derivative

$$\frac{\partial^{2r}}{\partial\theta^r \partial\theta'^r} K(\theta, \theta')$$

exists and is continuous on $\Theta \times \Theta$, where $K'(\theta, \theta')$ is defined in Equation (2.16). Then for large J,

$$|f_{\text{spec}}(x; \boldsymbol{m}_J) - f_{\text{Mix}}(x; Q)| = O(J^{-r}),$$

for each $x \in \mathcal{S}$.

To illustrate the construction of the moments induced by the eigenfuncations of $A'(\cdot)$, we give the following two examples.

Example 2.1 (Mixture of Poisson)**.**
Let $\Theta = [0, 25]$ and

$$f_0(x) = \frac{1}{25} \int_0^{25} \text{Pois}(x; \theta) \text{d}\theta,$$

where $\text{Pois}(x; \theta)$ is the probability function of the Poisson distribution with mean θ. Figure 2.1 shows the largest 10 eigenvalues of the integral operator $A(\cdot)$, the functions $\phi_j(x)$ and its associated $\gamma_j(\theta)$ corresponding to the largest 4 eigenvalues.

Example 2.2 (Mixture of Normal)**.**
For each fixed $\sigma^2 \geq 0$, let $\Theta = [0, 0.7]$ and

$$f_0(x) = \frac{1}{0.7} \int_0^{0.7} \mathcal{N}(x; \theta, \sigma^2) \text{d}\theta,$$

where $\mathcal{N}(x; \theta, \sigma^2)$ is the probability density function of the normal with mean θ and variance σ^2. For $\sigma^2 = 0.07^2$, Figure 2.2 shows the largest 10 eigenvalues of the integral operator $A(\cdot)$, the functions $\phi_j(x)$ and its associated $\gamma_j(\theta)$ corresponding to the largest 4 eigenvalues.

2.3.3 Quality of Truncation Approximation

We consider the quality of the two truncation approximations from two aspects: the non-negative sets and approximation error.

As an approximation to a probability function, we wish the truncation approximation $f_{\nu_0}(x; \boldsymbol{m}_J)$ to behave like a probability function. In other words, the truncation approximation $f_{\nu_0}(x; \boldsymbol{m}_J)$ should satisfy the following two conditions: for each $J = 1, 2, \ldots,$

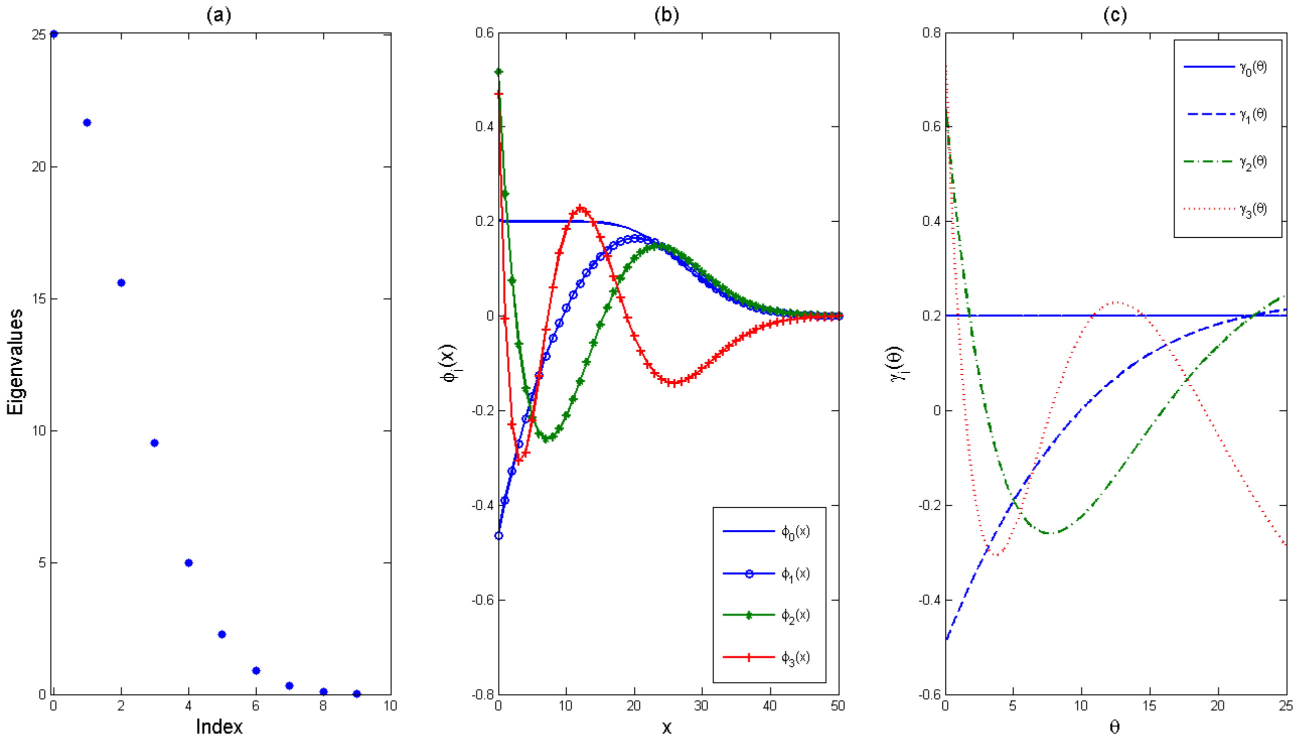

Figure 2.1: Plot of (a) the largest 10 eigenvalues; (b) the functions $\phi_j(x)$; (c) the associated $\gamma_j(\theta)$; in the mixture of Poisson for $j = 0, 1, 2, 3$.

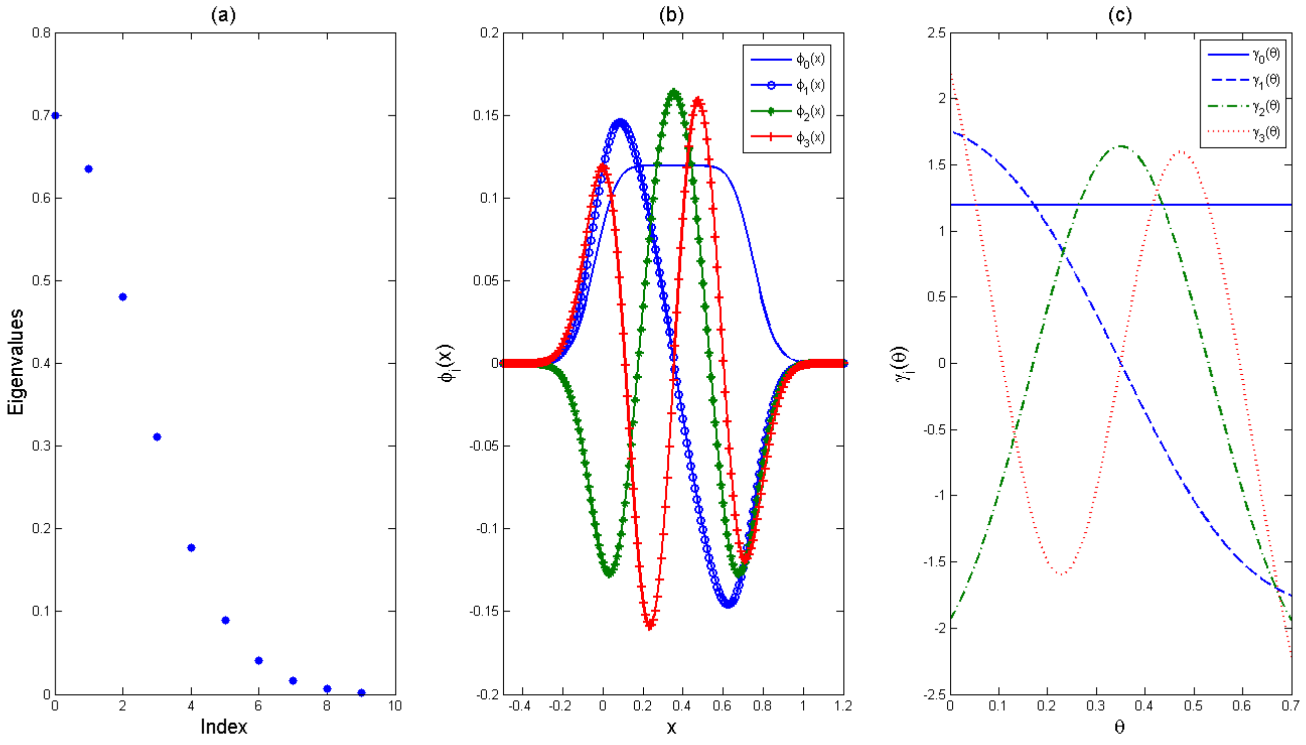

Figure 2.2: Plot of (a) the largest 10 eigenvalues; (b) the functions $\phi_j(x)$; (c) the associated $\gamma_j(\theta)$; in the mixture of normal for $j = 0, 1, 2, 3$.

1. the integral of $f_{\nu_0}(x; \boldsymbol{m}_J)$ with respect to x over $\mathcal{S}$ is one;
2. for each $x \in \mathcal{S}$, the truncation approximation $f_{\nu_0}(x; \boldsymbol{m}_J)$ must be strictly positive.

The first condition holds for either (2.10) or (2.15) because we have, for each $j = 1, 2, \ldots$,

$$\int_{\mathcal{S}} P_j(x; \theta_0) f(x; \theta_0) \mathrm{d}x = \langle P_j(x; \theta_0), P_0(x; \theta_0) \rangle_{L^2(\mathcal{S}, \nu_0)} = 0$$

and

$$\int_{\mathcal{S}} \phi_j(x) f_0^{1/2}(x) \mathrm{d}x = \frac{1}{\sqrt{\lambda_j}} \int_{\Theta} f(x; \theta) \gamma_j(\theta) \mathrm{d}\theta = 0.$$

However, the second condition is not always true. So, Marriott [2002, 2007] suggests to add the non-negative conditions as constraints on the parameter space of $\boldsymbol{m}_J$.

We return to Example 2.1 and 2.2 to examine the quality of the truncation approximations. Because each truncation approximation can be expressed as

$$f_{\nu_0}(x; \boldsymbol{m}_J) = \int_{\Theta} f_{\nu_0}(x; \boldsymbol{u}_J(\theta)) \mathrm{d}Q,$$

the non-negativeness of $f_{\nu_0}(x; \boldsymbol{u}_J(\theta))$ implies the non-negativeness of $f_{\nu_0}(x; \boldsymbol{m}_J)$, where $\boldsymbol{u}_J(\theta) = (u_1(\theta), \ldots, u_J(\theta))^{\mathrm{T}} \in \mathbb{R}^J$. And thus, we examine the negative region of $f_{\nu_0}(x; \boldsymbol{u}_J(\theta))$ over $\mathcal{S} \times \Theta$. We also consider the point-wise approximation error $\epsilon_{\nu_0, J}(s; \theta)$ over $\mathcal{S} \times \Theta$ defined in Subsection 2.2.4

Example 2.1 (continued).
Let $\boldsymbol{u}_4(\theta) = (u_0(\theta), \ldots, u_4(\theta))^{\mathrm{T}} \in \mathbb{R}^5$. We consider the cases where $\boldsymbol{u}_4(\theta)$ is induced by either the power functions with $\theta_0 = 12.5$ or the eigenfunctions of $A'(\cdot)$. The function $f_{\nu_0}(x; \boldsymbol{u}_J(\theta))$ is denoted by $f_{\mathrm{Ty}}(x; \boldsymbol{u}_J(\theta))$ and $f_{\mathrm{spec}}(x; \boldsymbol{u}_J(\theta))$ correspondingly. Figure 2.3 shows the negative region of $f_{\nu_0}(x; \boldsymbol{u}_4(\theta))$ over $\mathcal{S} \times [0, 25]$ in these two cases.

Figure 2.4 examines the approximation error of each component. Various issues of the reparameterization with power moments are seen from panel (a). Firstly, the quality of the approximation is non-uniform at each point in the sample space. Secondly, the approximation is poor when θ is away from $\theta = \theta_0$. This is due to the *nature of the underlying Laplace approximation where a polynomial approximation only behaves well in a small neighborhood of* $\theta = \theta_0$*. On the other hand, from the panel (b), we see that the quality of the approximation is almost uniform at each point* $(x, \theta) \in \mathcal{S} \times \Theta$*, when the moments are induced by the eigenfunctions of* $A'(\cdot)$*.*

The above discussions are also supported in Figure 2.5. The approximation $f_{\text{Ty}}(x; \boldsymbol{u}_4(\theta))$ *to the probability function* $\text{Pois}(x; \theta)$ *is not appropriate when* θ *is away from* $\theta_0 = 12.5$*; see the panel (a) and (c). The approximation* $f_{\text{spec}}(x; \boldsymbol{u}_4(\theta))$ *is not as good as* $f_{\text{Ty}}(x; \theta)$ *when* θ *is in a neighborhood of* $\theta_0 = 12.5$*; see the panel (b). However, it is able to characterize the shape of the probability function* $\text{Pois}(x; \theta)$ *when* θ *is away from* $\theta_0 = 12.5$*; see the panel (a) and (c).*

Example 2.2 (continued)**.**
Consider a fixed $\sigma^2 = 0.07^2$*. Again consider* $\boldsymbol{u}_4(\theta) = (u_1(\theta), \ldots, u_4(\theta))^{\mathrm{T}} \in \mathbb{R}^4$ *and* $\boldsymbol{u}_4(\theta)$ *is induced by either the power functions with* $\theta_0 = 0.35$ *or the eigenfunctions of* $A'(\cdot)$*. Figure 2.6 shows the negative regions of* $f_{\nu_0}(x; \boldsymbol{u}_4(\theta))$ *over* $\mathcal{S} \times [0, 0.7]$ *under these two types of reparameterizations. Also, Figure 2.7 gives the contour plots of* $\epsilon_{\nu_0,4}(x; \theta)$ *over* $\mathcal{S} \times [0, 0.7]$*. From the panel (a), we see the non-uniform and local approximation properties of the reparameterization with the moments induced by power functions. On the other hand, the quality of the approximation is more uniform, when the moments are induced by the eigenfunctions of* $A'(\cdot)$*. This is also supported by Figure 2.8.*

2.4 Reparameterization in Mixed-Effects Models

Generalized linear mixed models has been widely used in longitudinal studies; see [Diggle, 2002]. The following model is defined as the class of the generalized linear mixed models (GLMMs).

Definition 2.4.1 (Generalized Linear Mixed Models)**.**
Let $\boldsymbol{Y}_n = (Y_{n1}, \ldots, Y_{nT_n})^{\mathrm{T}} \in \mathbb{R}^{T_n}$ *be a response vector,* $\boldsymbol{X}_n = (\boldsymbol{X}_{n1}^{\mathrm{T}}, \ldots, \boldsymbol{X}_{nT_n}^{\mathrm{T}})^{\mathrm{T}} \in \mathbb{R}^{T_n \times p}$ *be the covariates matrix to the fixed effects,* $\boldsymbol{Z}_n = (\boldsymbol{Z}_{n1}^{\mathrm{T}}, \ldots, \boldsymbol{Z}_{nT_n}^{\mathrm{T}})^{\mathrm{T}} \in \mathbb{R}^{T_n \times q}$ *be the covariates matrix to the random effects, and* $\boldsymbol{b}_n = (b_{n1}, \ldots, b_{nq})^{\mathrm{T}} \in \mathbb{R}^q$ *be the*

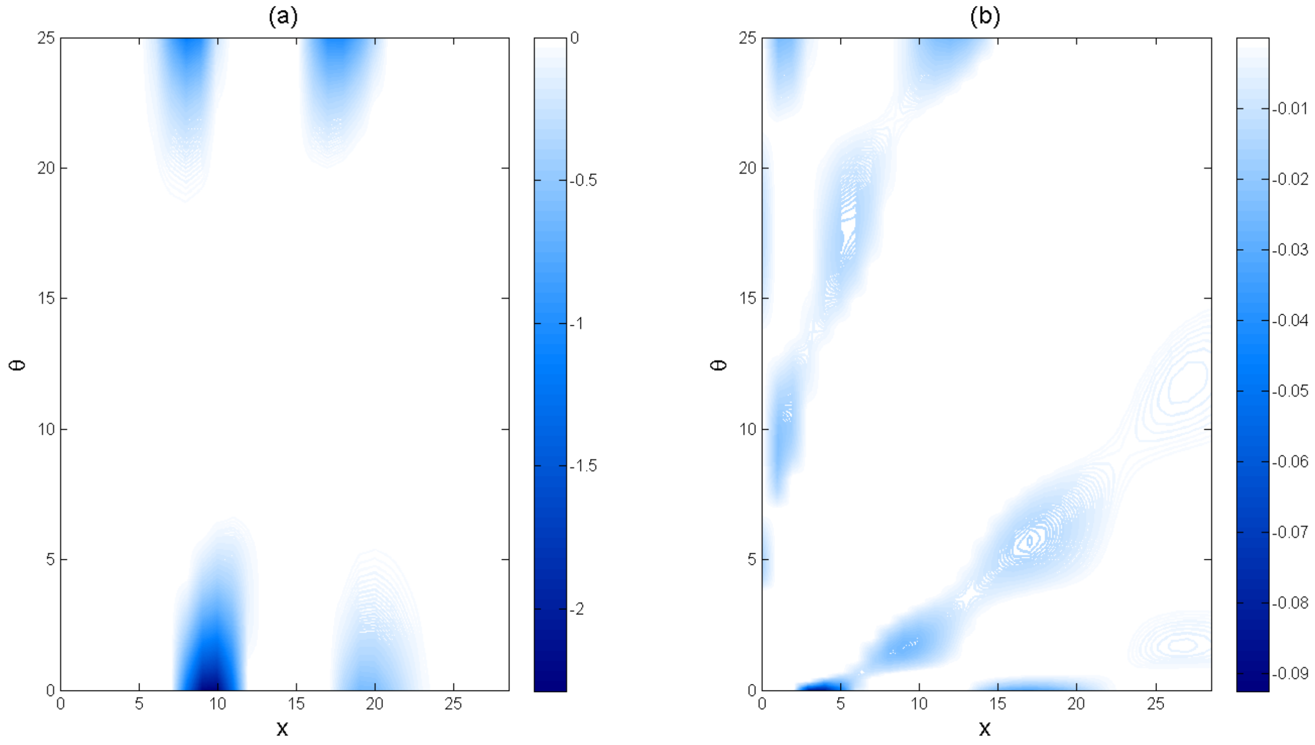

Figure 2.3: Plots of the negative regions of $f_{\nu_0}(x; \boldsymbol{u}_4(\theta))$ when the moments are induced by (a) power functions (b) the eigenfunctions of $A'(\cdot)$; for the mixture of $\text{Pois}(\theta)$.

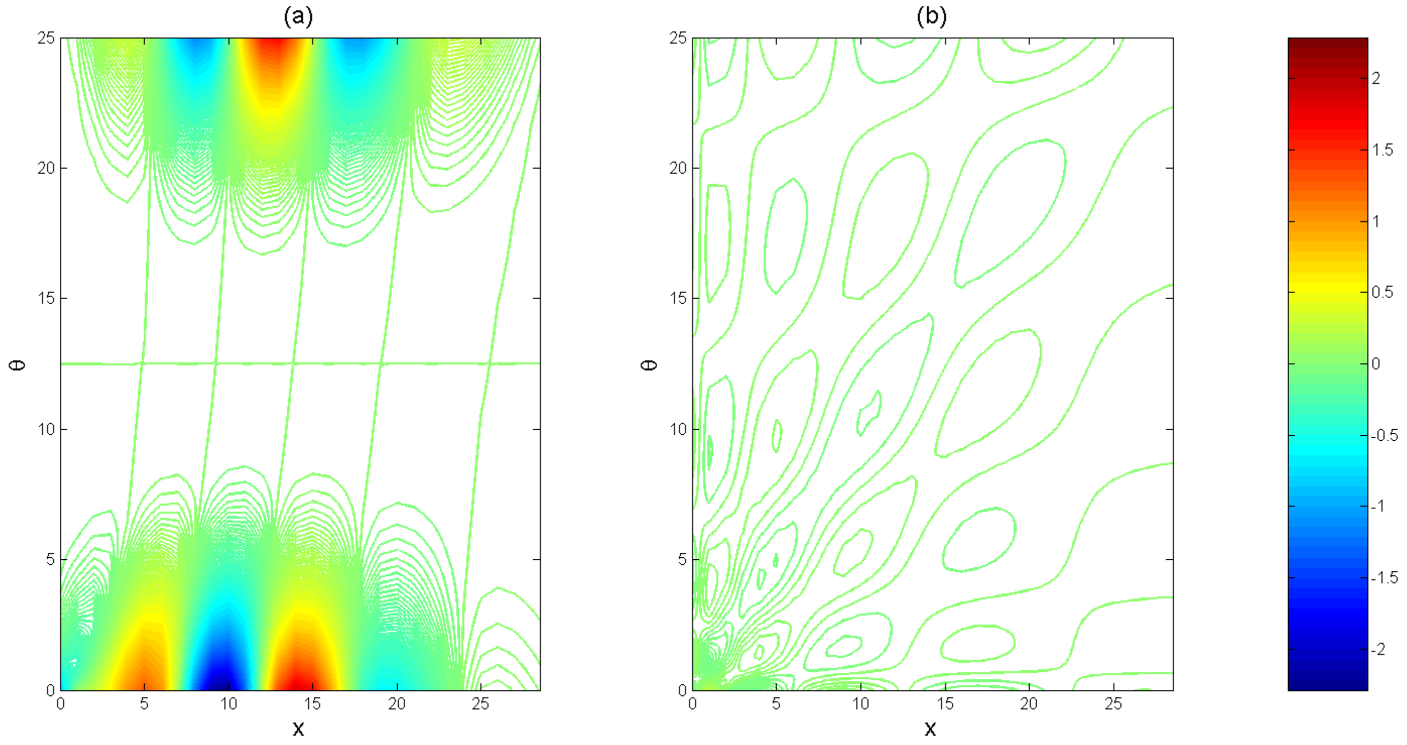

Figure 2.4: Contour plots of $\epsilon_{\nu_0,4}(x;\theta)$ when the moments are induced by (a) power functions (b) the eigenfunctions of $A'(\cdot)$; for the mixture of Pois(θ).

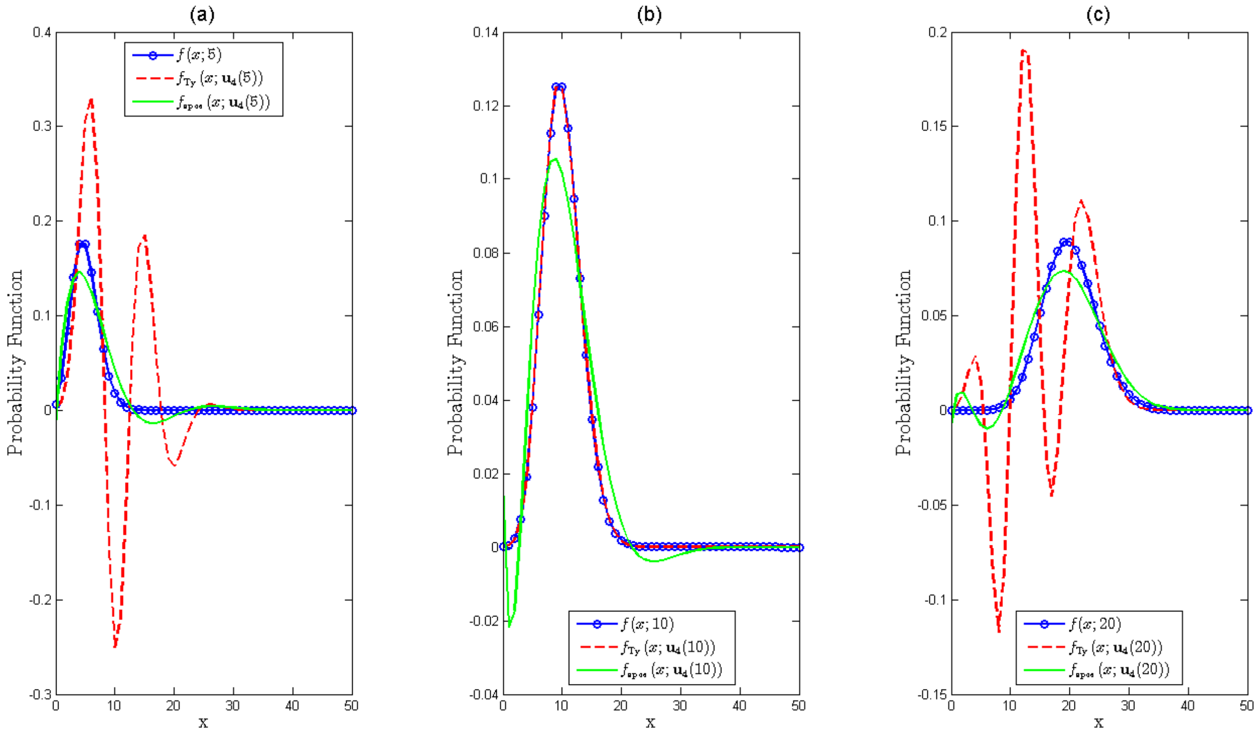

Figure 2.5: Plots of $f(x;\theta)$, $f_{\mathrm{Ty}}(x;\boldsymbol{u}_4(\theta))$ and $f_{\mathrm{spec}}(x;\boldsymbol{u}_4(\theta))$ when (a) $\theta = 5$, (b) $\theta = 10$ and (c) $\theta = 20$ in a mixture of Pois(θ).

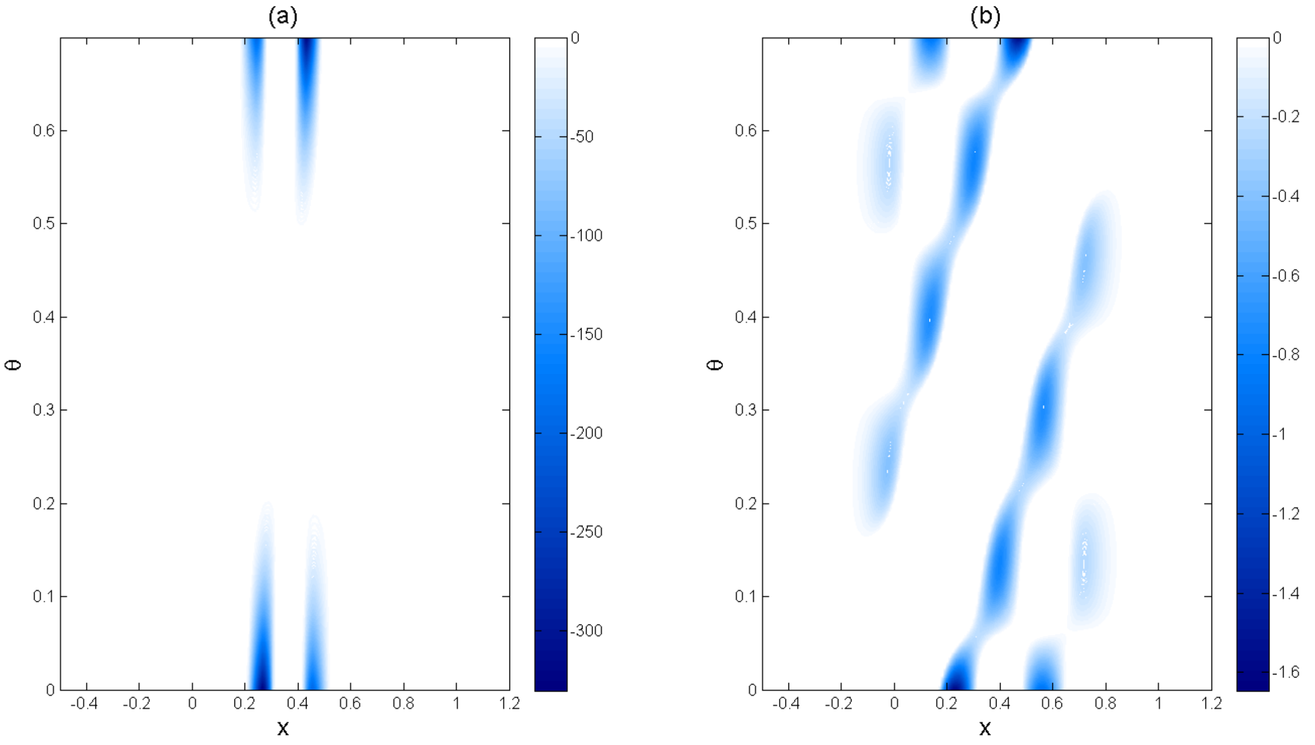

Figure 2.6: Plots of the negative regions of $f_{\nu_0}(x; \boldsymbol{u}_4(\theta))$ when the moments are induced by (a) power functions (b) the eigenfunctions of $A'(\cdot)$; for the mixture of $\mathcal{N}(\theta, \sigma^2)$ and $\sigma = 0.07$.

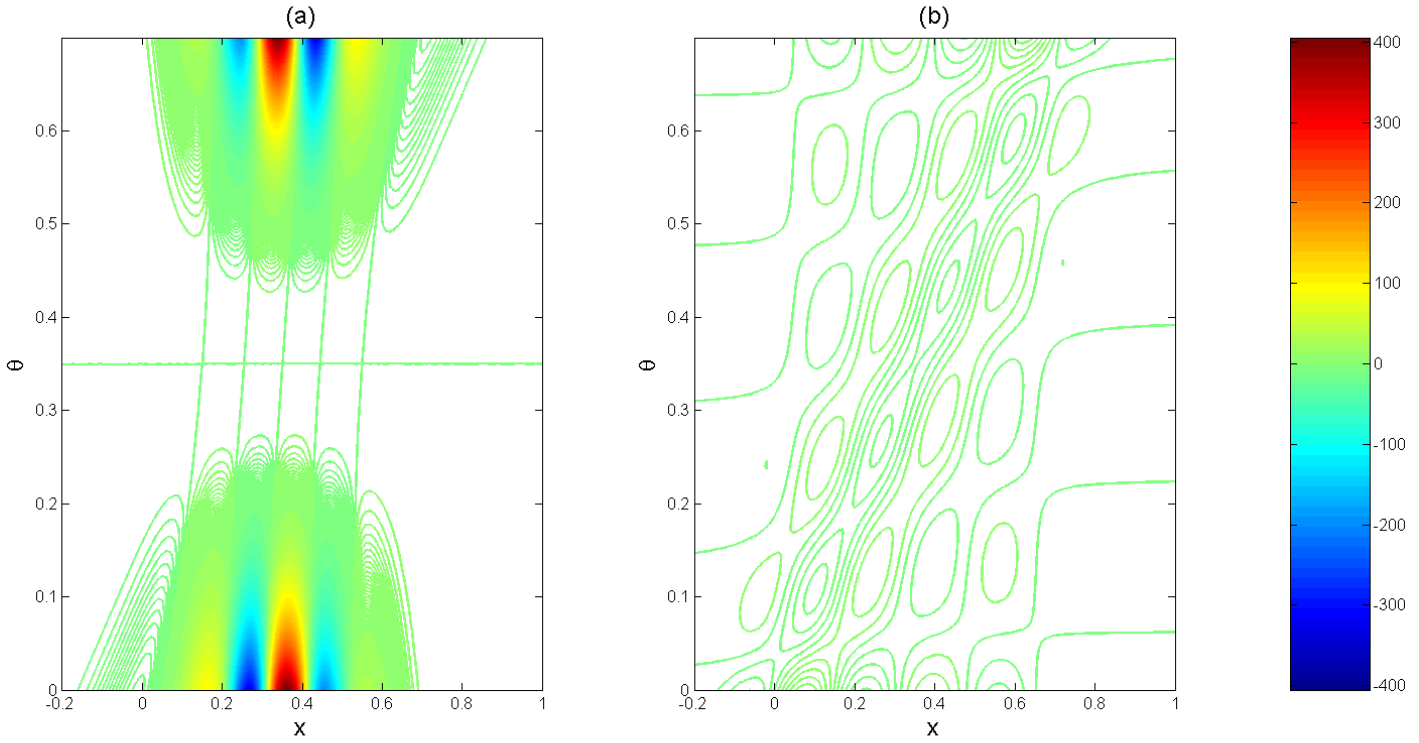

Figure 2.7: Contour plots of $\epsilon_{\nu_0,4}(x;\theta)$ when the moments are induced by (a) power functions (b) the eigenfunctions of $A'(\cdot)$; for the mixture of $\mathcal{N}(x;\theta,\sigma^2)$ and $\sigma = 0.07$.

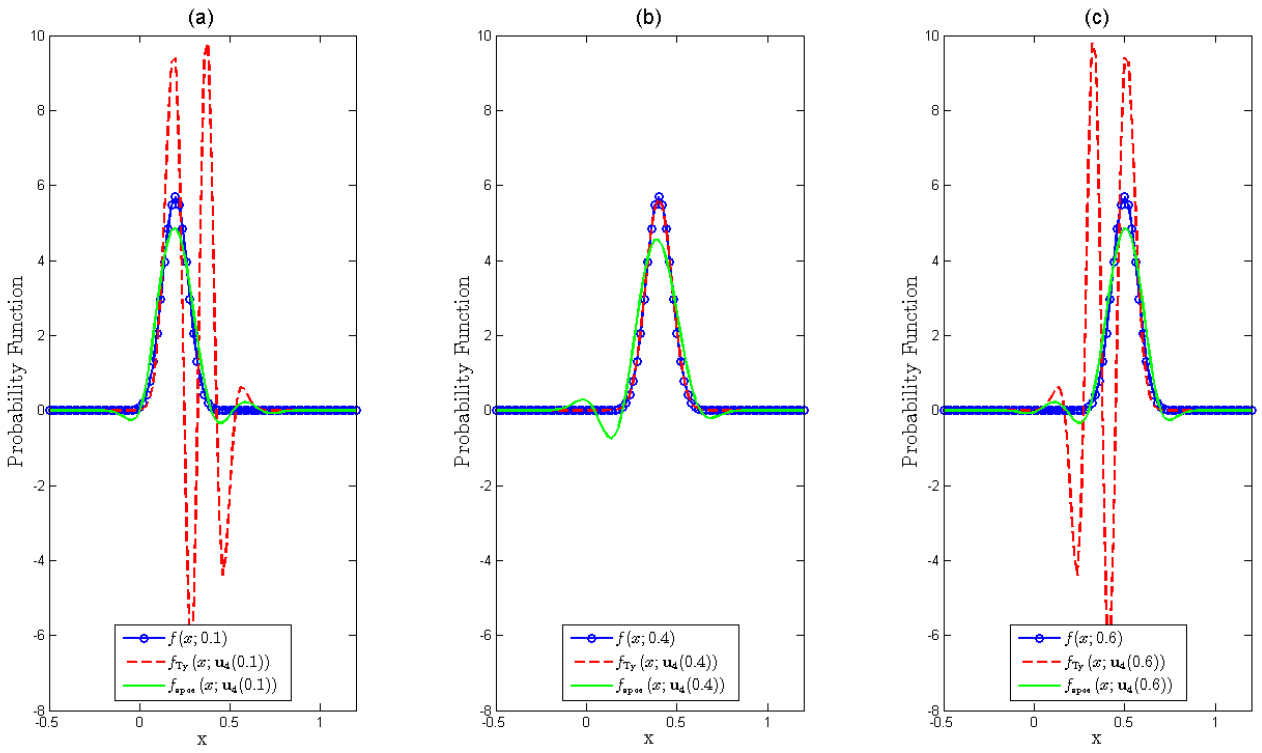

Figure 2.8: Plots of $f(x;\theta)$, $f_{\mathrm{Ty}}(x;\boldsymbol{u}_4(\theta))$ and $f_{\mathrm{spec}}(x;\boldsymbol{u}_4(\theta))$ when (a) $\theta = 0.2$, (b) $\theta = 0.4$ and (c) $\theta = 0.6$ in a mixture of $\mathcal{N}(\theta, \sigma^2)$ and $\sigma = 0.07$.

random effects vector. Conditional on $(\boldsymbol{X}_n, \boldsymbol{Z}_n, \boldsymbol{b}_n)$, we assume that $\boldsymbol{Y}_n$ follows a multivariate distribution with mean

$$\mathbb{E}\left[\boldsymbol{Y}_n \mid \boldsymbol{X}_n, \boldsymbol{Z}_n, \boldsymbol{b}_n\right] = g^{-1}\left(\boldsymbol{X}_n^{\mathrm{T}}\boldsymbol{\beta} + \boldsymbol{Z}_n^{\mathrm{T}}\boldsymbol{b}_n\right), \tag{2.19}$$

where $g^{-1}(\cdot)$ is the inverse of the link function $g(\cdot)$ and $\boldsymbol{\beta} \in \mathbb{R}^p$ is the regression parameter.

In Subsection 2.4.1 and 2.4.2, we consider the case that the random effect b_n is univariate. When the range of b_n is compact, the GLMMs (2.19) can be reparameterized in the generalized moments induced by the Chebyshev polynomials. On the other hand, the generalized moments induced by the Hermite polynomials are used, when the range of b_n is the real line. In Subsection 2.4.3, we consider the case that the random effect is multivariate.

2.4.1 Moments induced by the Chebyshev Polynomials

In this subsection, we consider the GLMMs which have univariate random effects, i.e.,

$$\mathbb{E}\left[\boldsymbol{Y}_{nt} \mid \boldsymbol{X}_{nt}, Z_{nt}, b_n\right] = g^{-1}\left(\boldsymbol{X}_{nt}^{\mathrm{T}}\boldsymbol{\beta} + Z_{nt}b_n\right), \tag{2.20}$$

where $Z_{nt} \in \mathbb{R}$ and $b_n \in \mathcal{B} = [-1, 1]$. Let $s = \boldsymbol{X}_n^{\mathrm{T}}\boldsymbol{\beta}$, $\theta = b_n$ and $h(s, \theta) = g^{-1}(s + Z_{nt}\theta)$. The mean of $Y_{nt} \mid (\boldsymbol{X}_{nt}, Z_{nt})$ is

$$U(s, Z_{nt}; Q) = \mathbb{E}[Y_{nt} \mid \boldsymbol{X}_{nt}, Z_{nt}] = \mathbb{E}_\theta\left[g^{-1}(s + Z_{nt}\theta)\right],$$

which is the form of Equation (2.1).

For any random effect b_n defined on $[a_l, a_u] \neq [-1, 1]$, we can have a new random effect

$$b_n' = \frac{2}{a_u - a_l}b_n - \frac{a_l + a_u}{a_u - a_l}, \tag{2.21}$$

which has the range $[-1, 1]$. Then, the model can be written as

$$\mathbb{E}\left[\boldsymbol{Y}_{nt} \mid \boldsymbol{X}_{nt}, Z_{nt}, b_n'\right] = g^{-1}\left(\boldsymbol{X}_{nt}^{\mathrm{T}}\boldsymbol{\beta} + Z_{nt}\frac{(a_u - a_l)b_n'}{2} + Z_{nt}\frac{a_l + a_u}{2}\right),$$

which is the form of Equation (2.20).

Consider the space $L^2([-1,1],\mu_0)$, where $\mathrm{d}\mu_0 = (1-b^2)^{-1/2}\mathrm{d}b$. The Chebyshev polynomials $\{T_j(x)\}_{j=0}^{\infty}$ (defined as follows) form a complete orthogonal system in the space $L^2([-1,1],\mu_0)$; see [Boyd, 2001, p.g. 64]. Figure 2.9 shows the functions $T_j(x)$ and $T_j(x)(1-x^2)^{-1/4}$, $j = 0,\ldots,3$, where for each i and j,

$$\int_{-1}^{1} T_i(x)(1-x^2)^{-1/4} \times T_j(x)(1-x^2)^{-1/4}\mathrm{d}x = \delta_{ij}. \tag{2.22}$$

Definition 2.4.2 (Chebyshev Polynomials)**.**
The Chebyshev polynomial $T_j(x)$ of the first kind is a polynomial in x of degree j, defined by the relation for $j = 2, 3, \ldots,$

$$T_{j+1}(x) = 2x\,T_j(x) - T_{j-1}(x)$$

with $T_0(x) = 1$ and $T_1(x) = x$.

Assume that $g^{-1}(s + Z_{nt}b_n)$ belongs to $L^2([-1,1],\mu_0)$ for each $s \in \mathcal{S}$. Then the Chebyshev expansion of $g^{-1}(s + Z_{nt}b_n)$ is

$$g^{-1}(s + Z_{nt}b_n) = \frac{1}{2}c_0(s, Z_{nt})\,T_0(b_n) + \sum_{j=1}^{\infty} c_j(s, Z_{nt})\,T_j(b_n),$$

where for each j,

$$c_j(s, Z_{nt}) = \frac{2}{\pi}\int_{-1}^{1}(1-b^2)^{-1/2}g^{-1}(s + Z_{nt}b)\,T_j(b)\mathrm{d}b.$$

The reparameterization of $U(s, Z_{nt}; Q)$ is

$$U_{\text{Chebyshev}}(s, Z_{nt}; \boldsymbol{m}_\infty) = c_0(s, Z_{nt}) + \sum_{j=1}^{\infty} c_j(s, Z_{nt})m_j$$

and the truncation approximation is

$$U_{\text{Chebyshev}}(s, Z_{nt}; \boldsymbol{m}_J) = c_0(s, Z_{nt}) + \sum_{j=1}^{J} c_j(s, Z_{nt})m_j$$

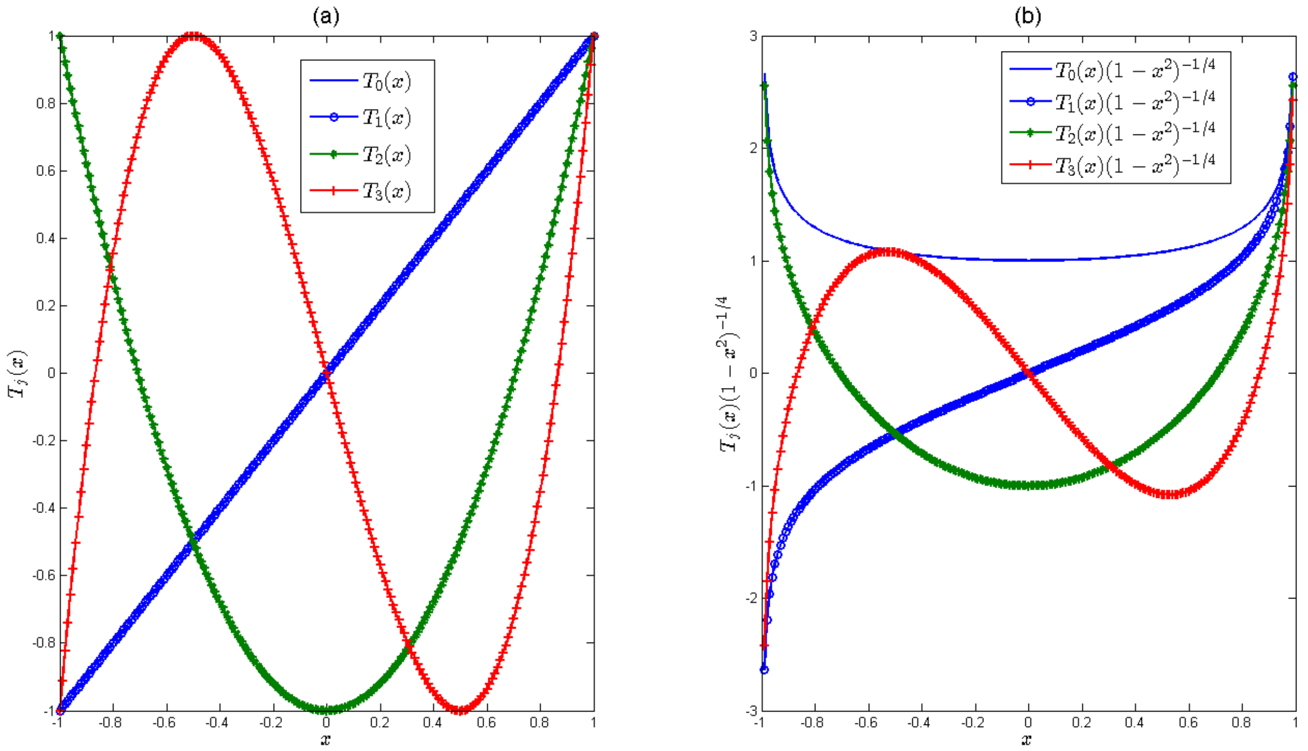

Figure 2.9: Plots of (a) the functions $T_j(x)$; (b) the functions $T_j(x)(1-x^2)^{-1/4}$, $j = 0, 1, 2, 3$.

where for each $j \in \{1, 2, \ldots\}$,

$$m_j = \int_{-1}^{1} T_j(b) \mathrm{d}Q.$$

For each $b_n \in \mathcal{B}$, the truncation approximation of $g^{-1}(s + Z_{nt} b_n)$ is

$$U_{\text{Chebyshev}}(s, Z_{nt};\, T_J(b_n)) = c_0(s, Z_{nt}) + \sum_{j=1}^{J} c_j(s, Z_{nt})\, T_j(b_n),$$

where $\boldsymbol{T}_J(b) = (\, T_0(b), \ldots,\, T_J(b))^{\mathrm{T}} \in \mathbb{R}^{J+1}$.

The Chebyshev polynomials can be described as the result of the Gram-Schmidt orthogonalization of the set of powers functions, $\{1, x, x^2, \ldots\}$ on $[-1, 1]$ with the measure $\mathrm{d}\mu_0 = (1 - x^2)^{-1/2}\mathrm{d}x$; see [Walter and Shen, 2001, p.g. 114]. Therefore, $\{\, T_j(x)\}_{j=0}^{J}$ forms a Chebyshev system by Theorem 2.2.1. Additional to the fact that $T_0(x) \equiv 1$, the parameters $\boldsymbol{m}_J$ can be interpreted by the generalized moments of Q induced by the Chebyshev polynomials.

If $g^{-1}(s + Z_{nt} b_n)$ is continuously differentiable, finitely or infinitely many times, the Chebyshev expansion converges fast; see Proposition 2.4.1.

Proposition 2.4.1 ([Mason and Handscomb, 2002, Theorem 5.14])**.**
Let $\{c_j\}_{j=0}^{J}$ be the coefficients in a Chebyshev expansion on $[-1, 1]$; defined in Equation (A.2). If a function $f(\theta)$ has $r + 1$ continuous derivatives on $[-1, 1]$, then

$$\left| f(\theta) - \sum_{j=0}^{J} c_j\, T_j(\theta) \right| = O(J^{-r}).$$

Because

$$\begin{aligned} & |U(s, Z_{nt}; Q) - U_{\text{Chebyshev}}(s, Z_{nt}; \boldsymbol{m}_J)| \\ = & \left| \int_{-1}^{1} g^{-1}(s + Z_{nt} b_n) - U_{\text{Chebyshev}}(s, Z_{nt};\, T_J(b_n)) \mathrm{d}Q \right| \\ \leq & \int_{-1}^{1} \left| g^{-1}(s + Z_{nt} b_n) - U_{\text{Chebyshev}}(s, Z_{nt};\, T_J(b_n)) \right| \mathrm{d}Q, \end{aligned}$$

the truncation approximation $U_{\text{Chebyshev}}(s, Z_{nt}; \boldsymbol{m}_J)$ could converge to the true model $U(s, Z_{nt}; Q)$ fast; see Corollary 2.4.1.

Corollary 2.4.1.
If $g^{-1}(s + Z_{nt}b_n)$ has $r+1$ continuous derivatives with respect to b_n on $[-1,1]$, then

$$|U(s, Z_{nt}; Q) - U_{\text{Chebyshev}}(s, Z_{nt}; \boldsymbol{m}_J(Q))| = O(J^{-r})$$

for each $s \in \mathcal{S}$ and $Z_{nt} \in \mathbb{R}$.

Some special functions can have explicit Chebyshev expansion (Example 2.3), while other functions can be expanded numerically; see Example (2.4) and (2.5).

Example 2.3 (The Log-link Function)**.**
The log-link function $g(\cdot) = \log(\cdot)$ is canonical in Poisson regression model; see [Diggle, 2002]. According to [Mason and Handscomb, 2002, Equation (5.18)], we have the expansion that

$$\begin{aligned} g^{-1}(s + Z_{nt}b_n) = &\ \exp\left(s + Z_{nt}b_n\right) \\ = &\ \exp(s)\left(\text{Bessel}_0(Z_{nt})\, T_0(b_n) + 2\sum_{j=1}^{\infty} \text{Bessel}_j(Z_{nt})\, T_j(b_n)\right), \end{aligned}$$

where $s = \boldsymbol{X}_{nt}^{\mathrm{T}}\boldsymbol{\beta}$ and $\text{Bessel}_j(x)$ is the j^{th} modified Bessel function of the first kind. Suppose that the random effect b_n has the range $[-1,1]$. The truncation approximation of $U(s, Z_{nt}; Q)$ is

$$U_{\text{Chebyshev}}(s, Z_{nt}; \boldsymbol{m}_J) = \exp(s)\left(2\text{Bessel}_0(Z_{nt}) + 2\sum_{j=1}^{J} \text{Bessel}_j(Z_{nt}) m_j\right),$$

where for each j,

$$m_j = \int_{-1}^{1} T_j(b)\mathrm{d}Q.$$

Let $Z_{nt} = 1$ and $J = 3$. In Figure 2.10 (a), we examine the point-wise approximation error $\epsilon_{\mu_0,3}(s; b_n)$ (see Subsection 2.2.4) for $(s, b_n) \in [-5,5] \times [-1,1]$ and see that the approximation is uniformly appropriate over the $[-5,5] \times [-1,1]$; also see Figure 2.11.

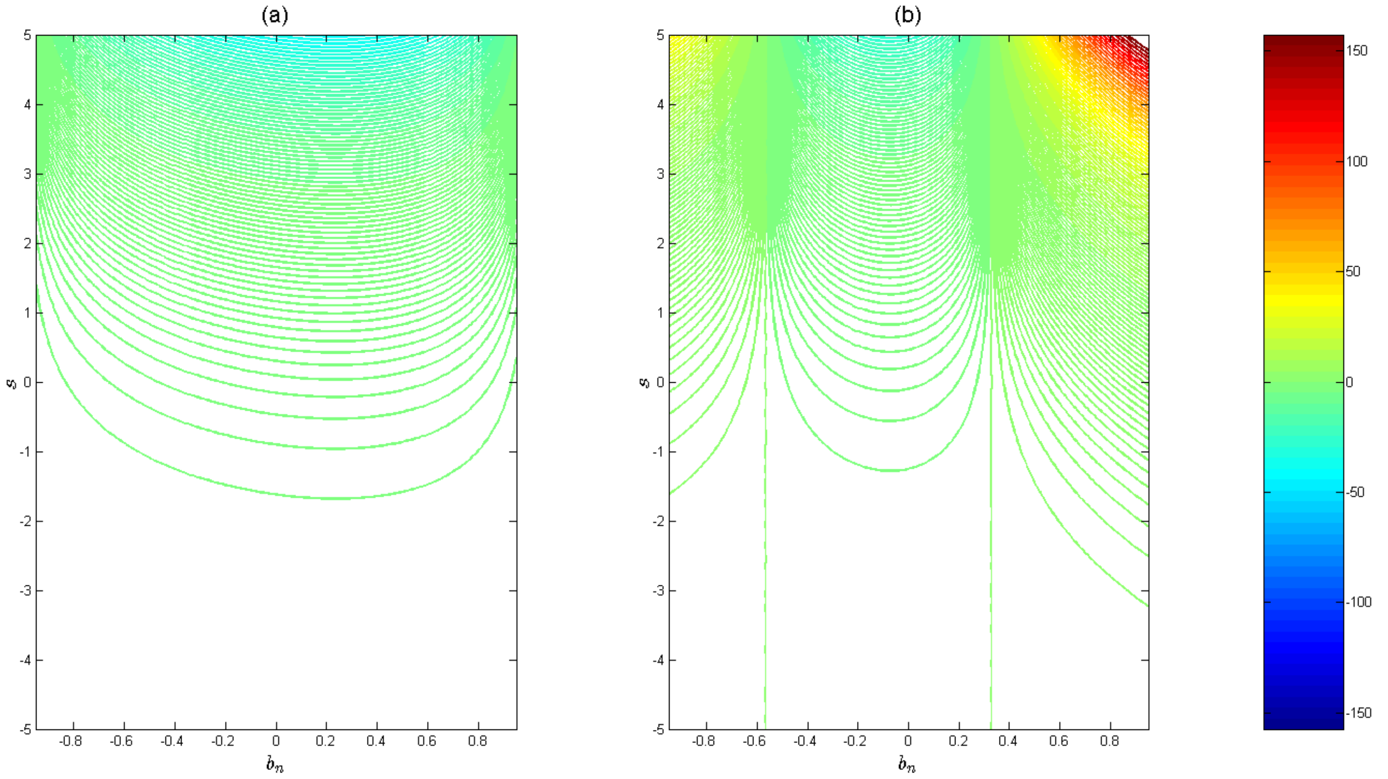

Figure 2.10: Contour plots of $\epsilon_{\mu_0,3}(s;b_n)$ when the moments are induced by (a) the Chebyshev polynomials (b) the Hermite polynomials; for the log-link function.

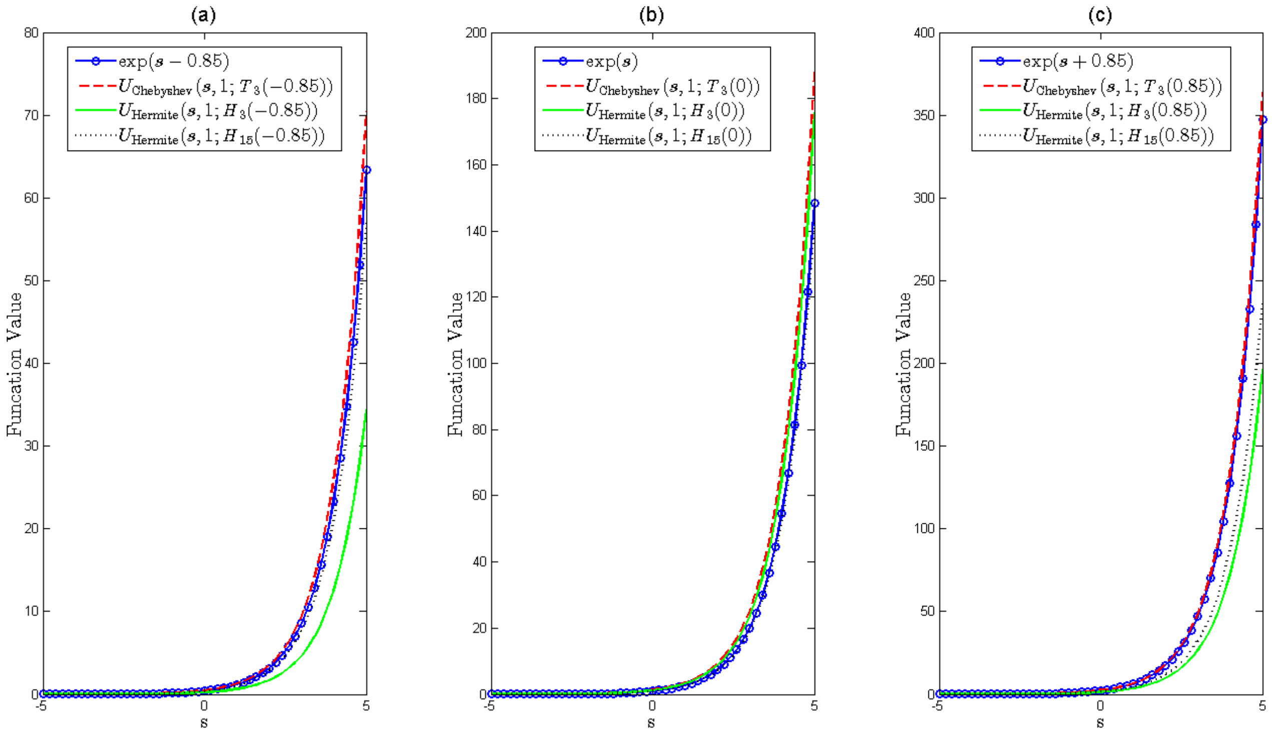

Figure 2.11: Plots of $g^{-1}(s + b_n)$, $U_{\text{Chebyshev}}(s, 1; T_3(b_n))$ $U_{\text{Hermite}}(s, 1; H_3(b_n))$ and $U_{\text{Hermite}}(s, 1; H_{15}(b_n))$ in a mixed-effects model with log-link function, when (a) $b_n = -0.85$, (b) $b_n = 0$ and (c) $b_n = 0.85$.

Example 2.4 (The Logit-link Function).

The logit-link function is common in the literature of regression models; see [Diggle, 2002]. The model is given by

$$\mathbb{E}[Y_{nt} \mid \boldsymbol{X}_{nt}, Z_{nt}, b_n] = \frac{1}{1 + \exp(-\boldsymbol{X}_{nt}^{\mathrm{T}}\boldsymbol{\beta} - Z_{nt}b_n)}$$

and

$$\mathbb{E}[Y_{nt} \mid \boldsymbol{X}_{nt}, Z_{nt}] = \int_{-1}^{1} \frac{1}{1 + \exp(-\boldsymbol{X}_{nt}^{\mathrm{T}}\boldsymbol{\beta} - Z_{nt}b_n)} \mathrm{d}Q(b_n).$$

Because $1/(1 + \exp(-s - Z_{nt}b_n)) \in [0, 1]$ *for any* Z_{nt}*,* b_n *and* s*,*

$$\int_{-1}^{1} (1 - b_n^2)^{-1/2}(1 + \exp(-s - Z_{nt}b_n))^{-2}\mathrm{d}b_n < \infty$$

for each $s \in \mathbb{R}$ *and* $Z_{nt} \in \mathbb{R}$*. Let* $s = \boldsymbol{X}_{nt}^{\mathrm{T}}\boldsymbol{\beta}$*,* $Z_{nt} = 1$ *and* $J = 3$*. In Figure 2.12 (a), we examine the point-wise approximation error* $\epsilon_{\mu_0,3}(s; b_n)$ *for* $(s, b_n) \in [-5, 5] \times [-1, 1]$ *and see that the approximation is appropriate uniformly over the* $[-5, 5] \times [-1, 1]$*. Figure 2.13 also supports that* $U_{\text{Chebyshev}}(s, 1; T_3(b_n))$ *approximates* $1/(1 + \exp(-s - b_n))$ *appropriately.*

Example 2.5 (The Tanh-link Function).

The hyperbolic link function is useful for modelling data that approaches an asymptote; see [Vos, 1991]. The model is given by the nonlinear regression

$$\mathbb{E}[Y_{nt} \mid \boldsymbol{X}_{nt}, Z_{nt}, b_n] = \tanh(\boldsymbol{X}_{nt}^{\mathrm{T}}\boldsymbol{\beta} + Z_{nt}b_n),$$

and

$$\mathbb{E}[Y_{nt} \mid \boldsymbol{X}_{nt}, Z_{nt}] = \int_{-1}^{1} \tanh(\boldsymbol{X}_{nt}^{\mathrm{T}}\boldsymbol{\beta} + Z_{nt}b_n)\mathrm{d}Q(b_n).$$

Because $\tanh(s + Z_{nt}b_n) \in [-1, 1]$ *for any* Z_{nt}*,* b_n *and* s*, it is true that*

$$\int_{-1}^{1} (1 - b_n^2)^{-1/2} \left(\tanh(s + Z_{nt}b_n)\right)^2 \mathrm{d}b_n < \infty$$

for each $s \in \mathbb{R}$ *and* $Z_{nt} \in \mathbb{R}$*. Let* $s = \boldsymbol{X}_{nt}^{\mathrm{T}}\boldsymbol{\beta}$*,* $Z_{nt} = 1$ *and* $J = 3$*. In Figure 2.14 (a), we examine the point-wise approximation error* $\epsilon_{\mu_0,3}(s; b_n)$ *for* $(s, b_n) \in [-5, 5] \times [-1, 1]$ *and see that the approximation is appropriate uniformly over the* $[-5, 5] \times [-1, 1]$*. In Figure 2.15, we also see that* $U_{\text{Chebyshev}}(s, 1; T_3(b_n))$ *approximates* $\tanh(s + b_n)$ *appropriately.*

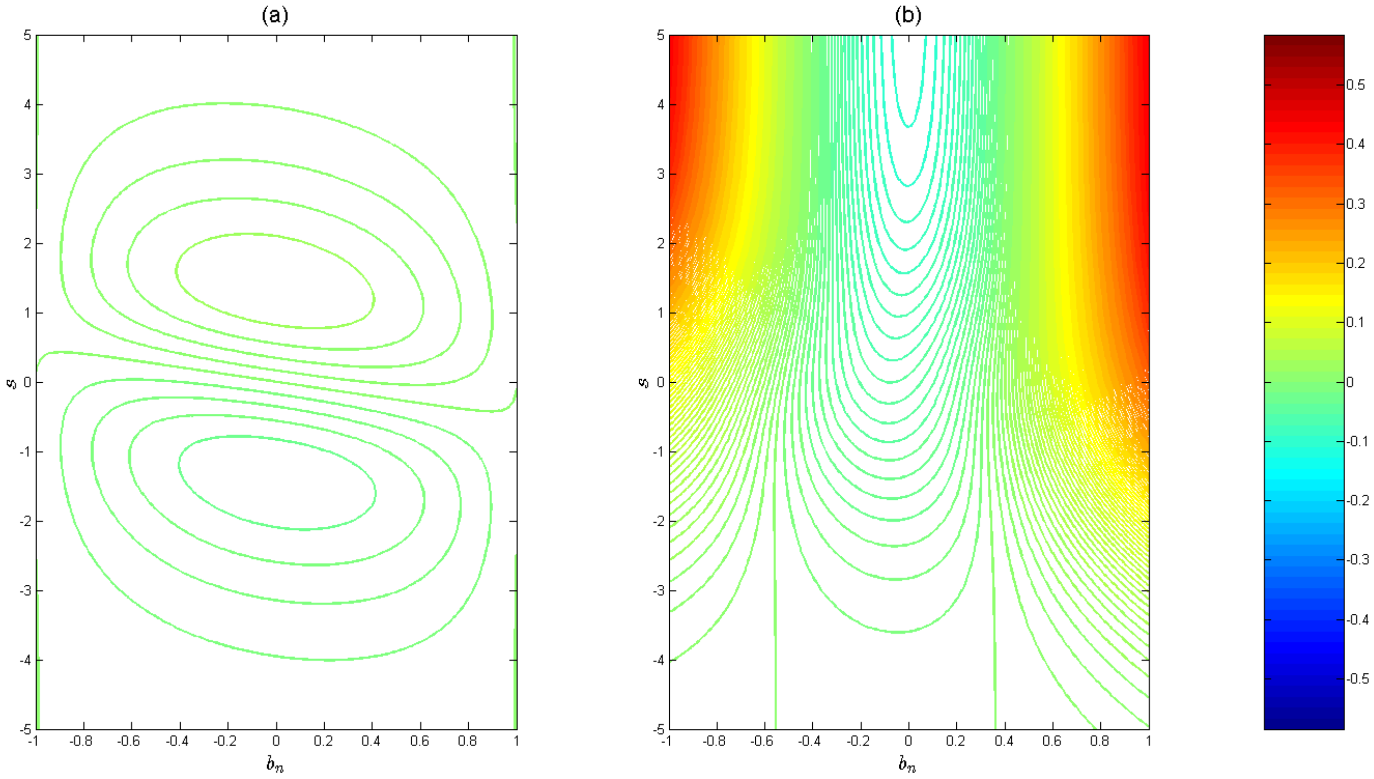

Figure 2.12: Contour plots of $\epsilon_{\mu_0,3}(s;b_n)$ when the moments are induced by (a) the Chebyshev polynomials (b) the Hermite polynomials; for the logit-link function.

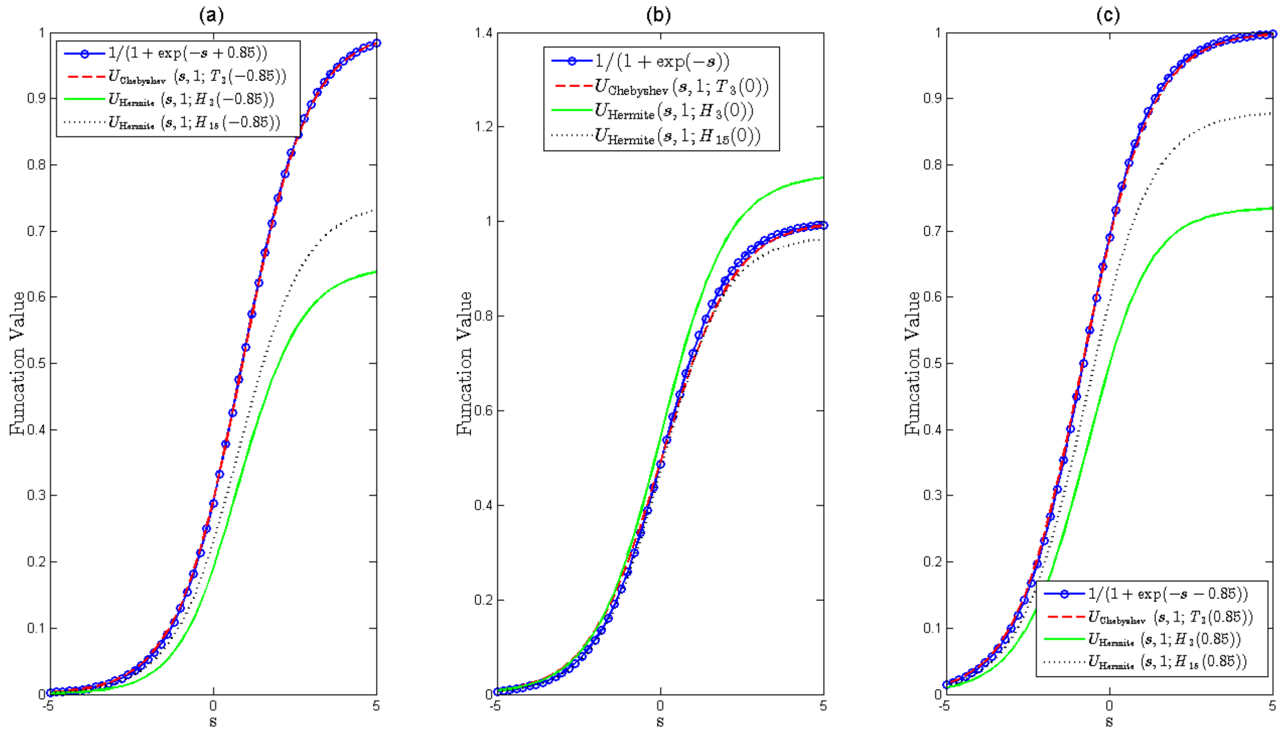

Figure 2.13: Plots of $g^{-1}(s+b_n)$, $U_{\text{Chebyshev}}(s,1;T_3(b_n))$ $U_{\text{Hermite}}(s,1;H_3(b_n))$ and $U_{\text{Hermite}}(s,1;H_{15}(b_n))$ in a mixed-effects model with logit-link function, when (a) $b_n=-0.85$, (b) $b_n=0$ and (c) $b_n=0.85$.

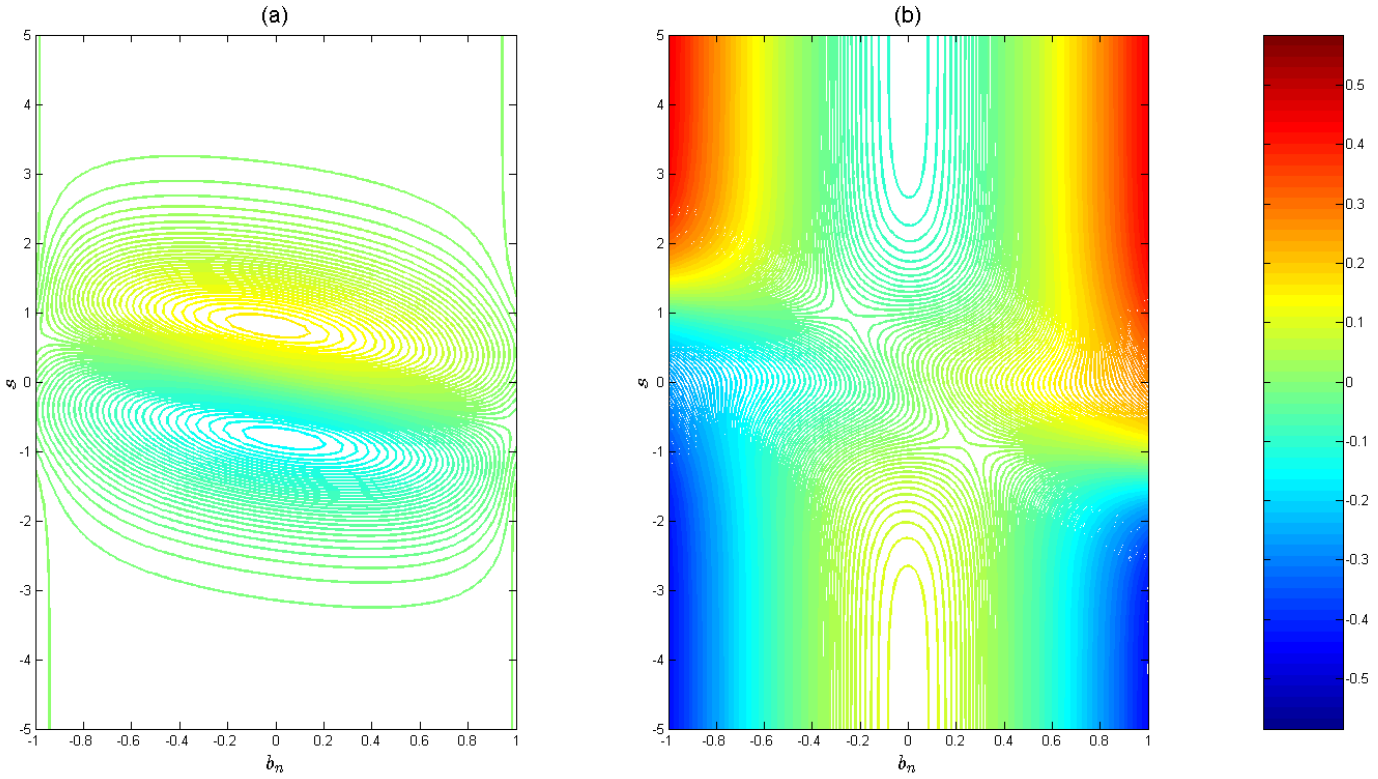

Figure 2.14: Contour plots of $\epsilon_{\mu_0,3}(s; b_n)$ when the moments are induced by (a) the Chebyshev polynomials (b) the Hermite polynomials; for the tanh-link function.

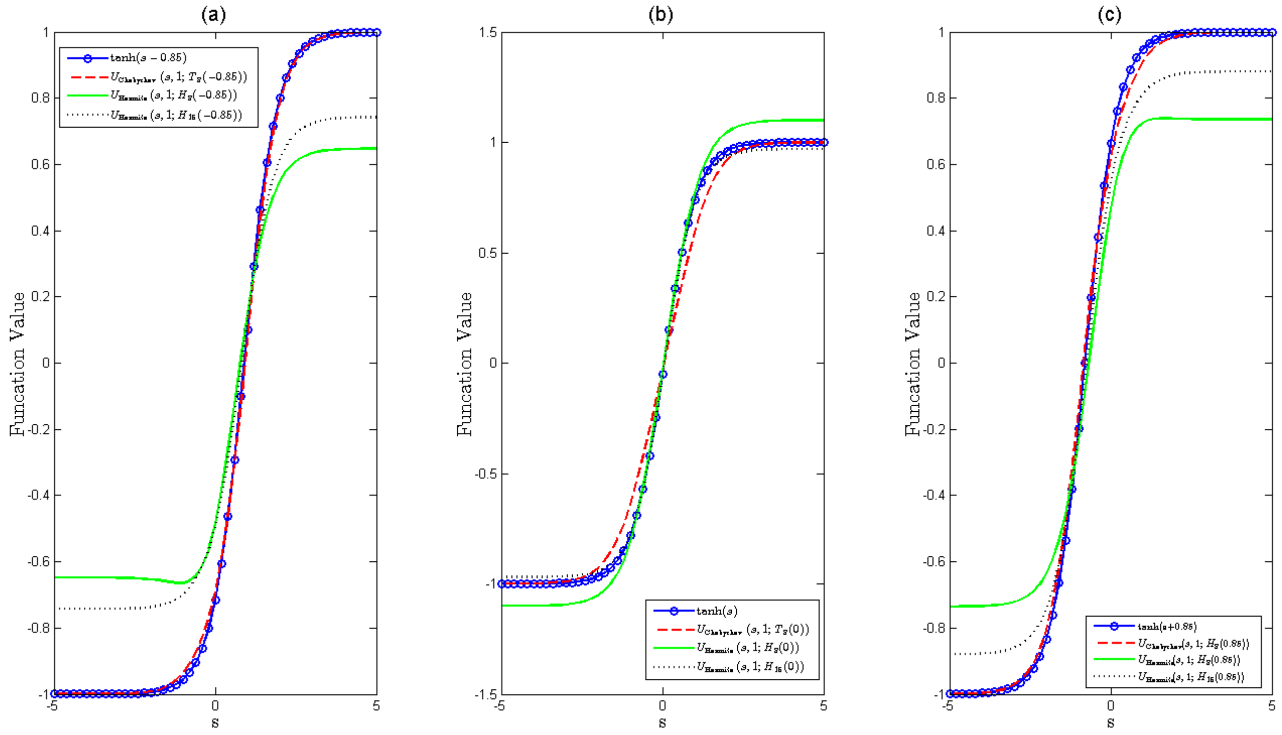

Figure 2.15: Plots of $g^{-1}(s+b_n)$, $U_{\mathrm{Chebyshev}}(s,1;T_3(b_n))$ $U_{\mathrm{Hermite}}(s,1;H_3(b_n))$ and $U_{\mathrm{Hermite}}(s,1;H_{15}(b_n))$ in a mixed-effects model with tanh-link function, when (a) $b_n=-0.85$, (b) $b_n=0$ and (c) $b_n=0.85$.

2.4.2 Moments induced by the Hermite Polynomials

In this subsection, we still consider the GLMMs with a univariate random effect $b_n \in \mathbb{R}$. Consider the space $L^2(\mathbb{R}, \mu_0)$, where $\mathrm{d}\mu_0 = \exp(-b^2)\mathrm{d}b$. The Hermite polynomials $\{H_j(x)\}_{j=0}^{\infty}$ (defined as follows) form a complete orthogonal system in the space $L^2(\mathbb{R}, \mu_0)$; see [Boyd, 2001, p.g. 64]. Figure 2.16 shows the functions $H_j(x)$ and $H_j(x)\exp(-x^2/2)$, $j = 0, \ldots, 3$, where for each i and j,

$$\int_{-\infty}^{\infty} H_i(x)\exp(-x^2/2) \times H_j(x)\exp(-x^2/2)\mathrm{d}x = \delta_{ij}.$$

Definition 2.4.3 (Hermite Polynomials)**.**
The Hermite polynomial $H_j(x)$ is a polynomial in x of degree j, defined by the relation for $j = 2, 3, \ldots,$

$$xH_j(x) = 1/2H_{j+1}(x) + jH_{j-1}(x)$$

with $H_0(x) = 1$ and $H_1(x) = 2x$.

Assume that $g^{-1}(s + Z_{nt}b_n)$ belongs to $L^2(\mathbb{R}, \mu_0)$ for each $s \in \mathcal{S}$. Then the Hermite expansion of $g^{-1}(s + Z_{nt}b_n)$ is

$$g^{-1}(s + Z_{nt}b_n) = \sum_{j=0}^{\infty} c_j(s, Z_{nt})H_j(b_n),$$

where for each j,

$$c_j(s, Z_{nt}) = \left(\pi^{1/2}2^j(j!)\right)^{-1/2} \int_{\mathbb{R}} \exp(-b^2)g^{-1}(s + Z_{nt}b)H_j(b)\mathrm{d}b.$$

Then, we have the reparameterization

$$U_{\text{Hermite}}(s, Z_{nt}; \boldsymbol{m}_{\infty}) = \sum_{j=0}^{\infty} c_j(s, Z_{nt})m_j$$

and the truncation approximation

$$U_{\text{Hermite}}(s, Z_{nt}; \boldsymbol{m}_J) = c_0(s, Z_{nt}) + \sum_{j=1}^{J} c_j(s, Z_{nt})m_j$$

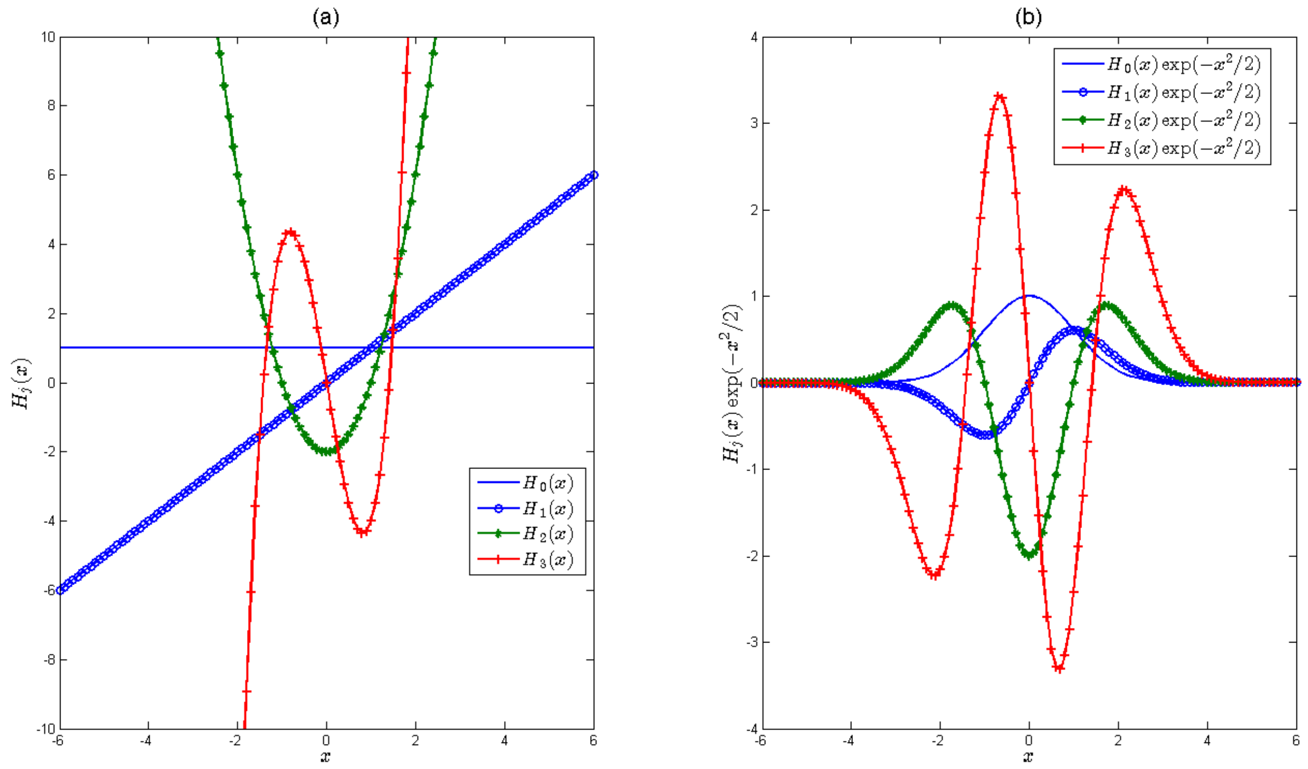

Figure 2.16: Plots of (a) the functions $H_j(x)$; (b) the functions $H_j(x)\exp(-x^2/2)$, $j = 0, 1, 2, 3$.

where for each $j = \{1, 2, \ldots\}$,

$$m_j = \int_{\mathbb{R}} H_j(b) \mathrm{d}Q.$$

For each $b_n \in \mathcal{B}$, the truncation approximation of $g^{-1}(s + Z_{nt}b_n)$ is

$$U_{\text{Hermite}}(s, Z_{nt}; \boldsymbol{H}_J(b_n)) = c_0(s, Z_{nt}) + \sum_{j=1}^{J} c_j(s, Z_{nt}) H_j(b_n),$$

where $\boldsymbol{H}_J(b) = (H_0(b), \ldots, H_J(b))^{\mathrm{T}} \in \mathbb{R}^{J+1}$.

The Hermite polynomials can be obtained from the Gram-Schmidt orthogonalization of the set of power functions $\{1, 2x, (2x)^2, \ldots\}$ on $\mathbb{R}$ with measure $\mathrm{d}\mu_0 = \exp(-x^2)\mathrm{d}x$; see [Walter and Shen, 2001, p.g. 121]. Therefore, $\{H_j(x)\}_{j=0}^{J}$ forms a Chebyshev system by Theorem 2.2.1. Additional to the fact that $H_0(x) \equiv 1$, the parameters $\boldsymbol{m}_J$ can be interpreted by the generalized moments of Q induced by the Hermite polynomials.

Similar to the Chebyshev expansion, the Hermite expansion converges fast, if $g^{-1}(s + Z_{nt}b_n)$ is continuously differentiable, finitely or infinitely many times; see Proposition 2.4.2.

Proposition 2.4.2.
Let $\{c_j\}_{j=0}^{J}$ be the coefficients in the Hermite expansion on $\mathbb{R}$. If $f(s)$ is such that $(\partial^r/\partial s^r) f(s)$ and $s^r f(s)$ are bounded and integrable on $\mathbb{R}$ for each $s \in \mathbb{R}$, then

$$\left| f(s) - \sum_{j=0}^{J} c_j H_j(s) \right| = O(J^{-r/2+1}).$$

Because

$$\begin{aligned} & |U(s, Z_{nt}; Q) - U_{\text{Hermite}}(s, Z_{nt}; \boldsymbol{m}_J)| \\ = & \left| \int_{\mathbb{R}} g^{-1}(s + Z_{nt}b_n) - U_{\text{Hermite}}(s, Z_{nt}; \boldsymbol{H}_J(b_n)) \mathrm{d}Q \right| \\ \leq & \int_{\mathbb{R}} \left| g^{-1}(s + Z_{nt}b_n) - U_{\text{Hermite}}(s, Z_{nt}; \boldsymbol{H}_J(b_n)) \right| \mathrm{d}Q, \end{aligned}$$

the truncation approximation $U_{\text{Hermite}}(s, Z_{nt}; \boldsymbol{m}_J)$ converges to $U(s, Z_{nt}; Q)$ fast; see Corollary 2.4.2.

Corollary 2.4.2.
If the function $g^{-1}(s+Z_{nt}b_n)$ is such that $(\partial^r/\partial b_n^r)\, g^{-1}(s+Z_{nt}b_n)$ and $b_n^r g^{-1}(s+Z_{nt}b_n)$ are bounded and integrable on $\mathbb{R}$ for each $b_n \in \mathbb{R}$ and $s \in \mathcal{S}$, then

$$|U(s, Z_{nt}; Q) - U_{\text{Hermite}}(s, Z_{nt}; \boldsymbol{m}_J(Q))| = O(J^{-r/2+1}),$$

for each $s \in \mathcal{S}$ and $Z_{nt} \in \mathbb{R}$.

We continue the following example to illustrate the Hermite expansion in GLMMs.

Example 2.3 (continued)**.**
According to [Lebedev, 1972, p.g. 74], we have the expansion that

$$\begin{aligned} g^{-1}(s + Z_{nt}b_n) &= \exp(s + Z_{nt}b_n) \\ &= \exp\left(s + Z_{nt}^2/4\right) \left(\sum_{j=0}^{\infty} \frac{1}{j!} \left(\frac{Z_{nt}}{2}\right)^j H_j(b_n) \right), \end{aligned}$$

where $s = \boldsymbol{X}_{nt}^{\mathrm{T}}\boldsymbol{\beta}$. The truncation approximation of $U_{nt}(s; Q)$ is

$$U_{\text{Hermite}}(s, Z_{nt}; \boldsymbol{m}_J) = \exp\left(s + Z_{nt}^2/4\right) \left(\sum_{j=0}^{\infty} \frac{1}{j!} \left(\frac{Z_{nt}}{2}\right)^j m_j \right),$$

where for each j,

$$m_j = \int_{\mathbb{R}} H_j(b)\mathrm{d}Q.$$

Let $Z_{nt} = 1$ and $J = 3$. Figure 2.10 (b) presents the point-wise approximation error $\epsilon_{\mu_0,3}(s; b_n)$ for $(s, b_n) \in [-5, 5]\times[-1, 1]$. And Figure 2.11 shows that $U_{\text{Hermite}}(s, 1; \boldsymbol{H}_3(b_n))$ can not appropriately approximate $\exp(s + b_n)$ *for some $b_n \in [-1, 1]$. We need to increase J to improve the quality of the approximation.*

Because for each s, Z_{nt} and b_n, it is true that

$$(1 + \exp(-s - Z_{nt}b_n))^{-2} \in [0, 1],$$

and

$$(\tanh(s + Z_{nt}b_n))^2 \in [0, 1],$$

we have

$$\int_{\mathbb{R}}(1+\exp(-s-Z_{nt}b_n))^{-2}\mathrm{d}\mu_0 < \infty$$

and

$$\int_{\mathbb{R}}\left(\tanh(s+Z_{nt}b_n)\right)^2\mathrm{d}\mu_0 < \infty$$

for each $s \in \mathbb{R}$ and $Z_{nt} \in \mathbb{R}$. Therefore, the Hermite expansions of the logit-link function and the tanh-link function are valid. The point-wise approximation errors $\epsilon_{\mu_0,3}(s;b_n)$ of the logit-link function and the tanh-link function, when the moments are induced by the Hermite polynomials, can be found in Figure 2.12 and 2.14 correspondingly. From these examples, it is observed that $U_{\text{Hermite}}(s,Z_{nt};\boldsymbol{m}_J)$ may not approximate as well as $U_{\text{Chebyshev}}(s,Z_{nt};\boldsymbol{m}_J)$; see also Figure 2.13 and 2.15. The reason is that the range of b_n is the real line in the Hermite expansion, while it is $[-1,1]$ in the Chebyshev expansion. This observation also supports the fact that the Hermite expansion has a slower convergence rate than the Chebyshev expansion; see Corollary 2.4.1 and 2.4.2.

2.4.3 Extension to Multivariate Random Effects

Many GLMMs have multivariate random effects; see [Diggle, 2002]. A univariate orthonormal basis for a function can be extended to a multivariate on by using tensor product (defined as below).

Definition 2.4.4 (Tensor Product Basis).
If $\{e_{kj}(x)\}_{j=0}^{\infty}$ is an orthonormal basis of a Hilbert space $\mathbb{H}_k$ for $k=1,\ldots,q$, the functions

$$\left\{\prod_{k=1}^{q} e_{kj_k}(x_k),\ \textit{for each } j_k = 0,1,\ldots\right\}$$

form an orthonormal basis for $\mathbb{H}_1 \times \cdots \times \mathbb{H}_q$, called the tensor product basis.

Here we reparameterize the inverse of the link function in a GLMM with a bivariate random effect by the tensor product basis induced by the Hermite polynomials.

Similar procedure can be extended to any orthonormal basis for a univariate functional space. Consider a GLMM with a bivariate random effect, i.e.,

$$\mathbb{E}\left[\boldsymbol{Y}_{nt} \mid \boldsymbol{X}_{nt}, \boldsymbol{Z}_{nt}, \boldsymbol{b}_n\right] = g^{-1}\left(\boldsymbol{X}_{nt}^{\mathrm{T}}\boldsymbol{\beta} + Z_{nt1}b_{n1} + Z_{nt2}b_{n2}\right),$$

where $\boldsymbol{Z}_{nt} = (Z_{nt1}, Z_{nt2})^{\mathrm{T}} \in \mathbb{R}^2$ and $\boldsymbol{b}_n = (b_{n1}, b_{n2})^{\mathrm{T}} \in \mathbb{R}^2$. Let $s = \boldsymbol{X}_{nt}^{\mathrm{T}}\boldsymbol{\beta} \in \mathcal{S}$. Assume that $g^{-1}\left(s + Z_{nt1}b_{n1} + Z_{nt2}b_{n2}\right)$ belongs to $L^2(\mathbb{R}, \mu_0) \times L^2(\mathbb{R}, \mu_0)$ for each $s \in \mathcal{S}$, where $\mathrm{d}\mu_0 = \exp(-b^2)\mathrm{d}b$.

Because the Hermite polynomials form an orthogonal basis of $L^2(\mathbb{R}, \mu_0)$, the tensor product basis induced by the Hermite polynomials for $L^2(\mathbb{R}, \mu_0) \times L^2(\mathbb{R}, \mu_0)$ is

$$\left\{H_{j_1 j_2}(\boldsymbol{b}_n) = H_{j_1}(b_{n1})H_{j_2}(b_{n2}), \text{ for each } j_1 = 0, 1, \dots \text{ and } j_2 = 0, 1, \dots\right\}.$$

Then, we have the bivariate Hermite expansion

$$g^{-1}(\boldsymbol{X}_{nt}^{\mathrm{T}}\boldsymbol{\beta} + \boldsymbol{Z}_{nt}^{\mathrm{T}}\boldsymbol{b}_n) = \sum_{j_1=0}^{\infty}\sum_{j_2=0}^{\infty} c_{j_1 j_2}(s, \boldsymbol{Z}_{nt})H_{j_1 j_2}(\boldsymbol{b}_n),$$

where for each j_1 and j_2,

$$c_{j_1 j_2}(s, \boldsymbol{Z}_{nt}) = \pi^{-1/2}\left(2^{(j_1+j_2)}j_1!j_2!\right)^{-1/2}\int_{\mathbb{R}^2} g^{-1}(\boldsymbol{X}_{nt}^{\mathrm{T}}\boldsymbol{\beta} + \boldsymbol{Z}_{nt}^{\mathrm{T}}\boldsymbol{b}_n)H_{j_1 j_2}(\boldsymbol{b}_n)\mathrm{d}\boldsymbol{b}_n.$$

The reparameterization is

$$U_{\text{Hermite}}(s, \boldsymbol{Z}_{nt}; \boldsymbol{m}_\infty) = \sum_{j_1=0}^{\infty}\sum_{j_2=0}^{\infty} c_{j_1 j_2}(s, \boldsymbol{Z}_{nt})m_{j_1 j_2},$$

and the truncation approximation is

$$U_{\text{Hermite}}(s, \boldsymbol{Z}_{nt}; \boldsymbol{m}_{J_1 J_2}) = \sum_{j_1=0}^{J_1}\sum_{j_2=0}^{J_2} c_{j_1 j_2}(s, \boldsymbol{Z}_{nt})m_{j_1 j_2},$$

where $\boldsymbol{m}_{J_1 J_2}$ is a vector in $\mathbb{R}^{J_1 J_2}$ whose elements

$$m_{j_1 j_2} = \int_{\mathbb{R}^2} H_{j_1 j_2}(\boldsymbol{b}_n)\mathrm{d}Q, \text{ for } j_1 = 0, \dots, J_1, j_2 = 0, \dots, J_2,$$

and Q is the probability measure of the vector $\boldsymbol{b}_n \in \mathbb{R}^2$. Note that the parameter $\boldsymbol{m}_{J_1 J_2}$ can not be interpreted as the generalized moments of Q in Definition 2.2, because the Chebyshev system is not defined on $\mathbb{R}^2$.

If we further assume that b_{n1} and b_{n2} are independent and have distribution Q_1 and Q_2 on Θ correspondingly, we have

$$U_{\text{Hermite}}(s, \boldsymbol{Z}_{nt}; \boldsymbol{m}_{J_1}, \boldsymbol{m}'_{J_2}) = \sum_{j_1=0}^{J_1} \sum_{j_2=0}^{J_2} c_{j_1 j_2}(s, \boldsymbol{Z}_{nt}) m_{j_1} m'_{j_2},$$

where $\boldsymbol{m}_{J_1} = (m_1, \ldots, m_{J_1})^{\mathrm{T}} \in \mathbb{R}^{J_1}$ and $\boldsymbol{m}'_{J_2} = (m'_1, \ldots, m'_{J_2})^{\mathrm{T}} \in \mathbb{R}^{J_2}$, and

$$m_j = \int_\Theta H_j(b_{n1}) \mathrm{d}Q_1, \text{ for } j = 0, \ldots, J_1,$$

and

$$m'_j = \int_\Theta H_j(b_{n2}) \mathrm{d}Q_2 \text{ for } j = 0, \ldots, J_2.$$

Now, the parameters $\boldsymbol{m}_{J_1}$ and $\boldsymbol{m}_{J_2}$ can be interpreted as the generalized moments of Q_1 and Q_2 induced by the Hermite polynomials.

Appendix: A

A.1 Strictly Totally Positive Kernel Functions

The strictly totally positive kernel functions and Chebyshev systems are defined in the following ways in [Karlin and Studden, 1966]. In this subsection, the set Θ is assumed to be compact.

Definition A.1 (Strictly Totally Positive).
A real valued kernel function $K(s, \theta)$, $(s, \theta) \in \mathcal{S} \times \Theta \subseteq \mathbb{R}^2$, is called strictly totally positive of order r, if for each $J = 1, \ldots, r$, we have $\det(K(s_i, \theta_j))_{i,j=0}^{J} > 0$, whenever $s_0 < \cdots < s_r$, $\theta_0 < \cdots < \theta_r$ and $(s_i, \theta_j) \in \mathcal{S} \times \Theta$, $i, j = 0, \ldots, r$.(s,s')

Consider a kernel function for $(s, s') \in \mathcal{S} \times \mathcal{S}' \subseteq \mathbb{R}^2$,

$$K^*(s, s') = \int_\Theta L(s, \theta) M(s', \theta) \mathrm{d}\theta, \tag{A.1}$$

where $L(s, \theta), (s, \theta) \in \mathcal{S} \times \Theta \subset \mathbb{R}^2$ and $M(s', \theta), (s', \theta) \in \mathcal{S}' \times \Theta \subset \mathbb{R}^2$. The following proposition is proved in [Karlin and Studden, 1966].

Proposition A.1.
If the kernel function in (A.1) exists for each $(s, s') \in \mathcal{S} \times \mathcal{S}'$ *and* $L(s, \theta)$ *and* $M(s', \theta)$ *are strictly totally positive, then* $K(s, s')$ *is strictly totally positive.*

Pinkus [1996] further stated that the eigenfuctions from a strictly totally positive kernel function could also form a Chebyshev system.

Proposition A.2.
Let

$$(A'g)(\theta) = \int_{\Theta} g(\theta) K'(\theta, \theta') \mathrm{d}\theta,$$

be a compact, self-adjoint, positive integral operator in the form of (2.12). Moreover, the kernel function $K'(\theta, \theta')$ *is strictly totally positive over* $\Theta \times \Theta$*. Then, the integral operator* $A'(\cdot)$ *has the eigenvalues* $\lambda_0 > \lambda_1 > \cdots > 0$ *and associated eigenfunctions* $\phi_0(\theta), \phi_1(\theta), \ldots$*, which are continuous over* Θ*. For each* $J = 1, 2, \ldots$*, the set* $\{\phi_i(\theta)\}_{i=0}^{J}$ *forms a Chebyshev system over* Θ*. Moreover,* $\phi_0(\theta)$ *is strictly one sign on* Θ*.*

A.2 Asymptotic Coefficients of Orthogonal Polynomials Expansion

Consider the orthonormal polynomials $\{P_j(s)\}_{j=0}^{\infty}$ defined by the measure μ_0 on $\mathcal{S}$. That is for each i and j,

$$\int_{\mathcal{S}} P_i(s) P_j(s) \mathrm{d}\mu_0 = \delta_{ij}.$$

We have the expansion of a function $f(s)$ by $\{P_j(s)\}_{j=0}^{\infty}$ such that that

$$f(s) = \sum_{j=0}^{\infty} c_j P_j(s),$$

where for each j

$$c_j = \int_{\mathcal{S}} f(s) P_j(s) \mathrm{d}\mu_0 \tag{A.2}$$

is known as the j^{th} coefficient in the expansion.

When $\{P_j(s)\}_{j=0}^{\infty}$ is either the normalized Chebyshev or Hermite polynomials, the polynomial $P_j(s)$ is bounded for each s and j; see [Boyd, 2001, p.g. 47] and [Boyd, 1984]. Therefore, we have

$$\left| f(s) - \sum_{j=0}^{J} c_j P_j(s) \right| \leq M \sum_{j=J+1}^{\infty} |c_j|,$$

where M is a positive constant. When c_j decays fast, we may have

$$\left| f(s) - \sum_{j=0}^{J} c_j P_j(s) \right| = O(c_J);$$

see [Boyd, 1984]. For such a reason, it is important to study the asymptotic properties of c_J, as J goes to infinity; see [Boyd, 2001] and [Mason and Handscomb, 2002]. With the fact in [Orszag and Bender, 1999, p.g. 379] that, for large J and fixed r,

$$\sum_{j=J+1}^{\infty} \frac{1}{j^r} = O\left(\frac{1}{(r-1)J^{r-1}}\right), \tag{A.3}$$

Proposition 2.4.1 and 2.4.2 are obtained from the following two propositions.

Proposition A.3 ([Mason and Handscomb, 2002, Equation (5.100)])**.**
Let $\{c_j\}_{j=0}^{J}$ be the coefficients in a Chebyshev expansion on $[-1,1]$. If $f(s)$ has $r+1$ continuous derivatives on $[-1,1]$, then

$$|c_J| = O(J^{-r-1}).$$

Proposition A.4 ([Boyd, 1984])**.**
Let $\{c_j\}_{j=0}^{J}$ be the coefficients in a Hermite expansion on $\mathbb{R}$. If $f(s)$ is such that $(\partial^r/\partial s^r)\, f(s)$ and $s^r f(s)$ are bounded and integrable on $\mathbb{R}$ for all real s, then

$$|c_J| = O(J^{-r/2}).$$

A.3 Proof of Theorem 2.2.1

Proof. We use mathematical induction to prove this theorem. Note that $v_0(x)$ is a constant function and it forms a Chebyshev system over $\mathcal{S}$. All we need to show is that $\{v_j(s)\}_{j=0}^{J}$ forms a Chebyshev system if $\{v_j(s)\}_{j=0}^{J-1}$ forms a Chebyshev system.

Let $\boldsymbol{u}_{(J-1)}(\theta) = (u_0(\theta), \ldots, u_{J-1}(\theta))^{\mathrm{T}}$ and $\boldsymbol{v}_{(J-1)}(\theta) = (v_0(\theta), \ldots, v_{J-1}(\theta))^{\mathrm{T}}$ be two vectors in $\mathbb{R}^J$. Because $\{v_j(\theta)\}_{j=0}^{J-1}$ forms an orthonormal system for $\{u_j(\theta)\}_{j=0}^{J-1}$, there exists a unique $J \times J$ full-rank matrix $\boldsymbol{B}$ with respect to $L^2(\Theta, \mu_0)$ such that, for each $\theta \in \Theta$,

$$\boldsymbol{u}_{(J-1)}(\theta) = \boldsymbol{B}\boldsymbol{v}_{(J-1)}(\theta).$$

For any $\theta_0 < \cdots < \theta_{J-1}$, let $\boldsymbol{U}_{(J-1)} = \left[\boldsymbol{u}_{(J-1)}(\theta_0), \ldots, \boldsymbol{u}_{(J-1)}(\theta_{J-1})\right]$ be a $J \times J$ matrix and $\boldsymbol{V}_{(J-1)} = \left[\boldsymbol{v}_{(J-1)}(\theta_0), \ldots, \boldsymbol{v}_{(J-1)}(\theta_{J-1})\right]$ be a $J \times J$ matrix. We have

$$\boldsymbol{U}_{(J-1)} = \boldsymbol{B}\boldsymbol{V}_{(J-1)}.$$

It follows that

$$\det \boldsymbol{U}_{(J-1)} = \det \boldsymbol{B} \det \boldsymbol{V}_{(J-1)}.$$

Because both $\det \boldsymbol{U}_{(J-1)}$ and $\det \boldsymbol{V}_{(J-1)}$ are positive, we have $\det \boldsymbol{B} > 0$.

For any $\theta_0 < \cdots < \theta_J$, there exists a vector $\boldsymbol{C} = (\boldsymbol{c}_{(J-1)}^{\mathrm{T}}, c_J)^{\mathrm{T}} \in \mathbb{R}^{J+1}$ such that

$$\boldsymbol{u}_{(J)}^{\mathrm{T}} = \begin{bmatrix} \boldsymbol{c}_{(J-1)}^{\mathrm{T}} & c_J \end{bmatrix} \begin{bmatrix} \boldsymbol{V}_{(J-1,J)} \\ \boldsymbol{v}_{(J)}^{\mathrm{T}} \end{bmatrix},$$

where $\boldsymbol{u}_{(J)} = (u_J(\theta_0), \ldots, u_{(J)}(\theta_J))^{\mathrm{T}} \in \mathbb{R}^{J+1}$ and $\boldsymbol{v}_{(J)} = (v_J(\theta_0), \ldots, v_J(\theta_J))^{\mathrm{T}} \in \mathbb{R}^{J+1}$, and $\boldsymbol{c}_{(J-1)} \in \mathbb{R}^J$ and c_J is a scalar. Without losing generality, let $c_J > 0$. Let $\boldsymbol{U}_{(J-1,J)} = \left[\boldsymbol{u}_{(J-1)}(\theta_0), \ldots, \boldsymbol{u}_{(J-1)}(\theta_J)\right]$ be a $J \times (J+1)$ matrix and $\boldsymbol{V}_{(J-1,J)} = \left[\boldsymbol{v}_{(J-1)}(\theta_0), \ldots, \boldsymbol{v}_{(J-1)}(\theta_J)\right]$ be a $J \times (J+1)$ matrix. Then, we have

$$\begin{bmatrix} \boldsymbol{B} & \boldsymbol{0} \\ \boldsymbol{c}_{(J-1)}^{\mathrm{T}} & c_J \end{bmatrix} \begin{bmatrix} \boldsymbol{V}_{(J-1,J)} \\ \boldsymbol{v}_{(J)}^{\mathrm{T}} \end{bmatrix} = \begin{bmatrix} \boldsymbol{U}_{(J-1,J)} \\ \boldsymbol{u}_{(J)}^{\mathrm{T}} \end{bmatrix} = \boldsymbol{U}_{(J)}.$$

We can obtain that

$$\begin{aligned} \det \boldsymbol{U}_{(J)} &= \det \left(\begin{bmatrix} \boldsymbol{B} & \boldsymbol{0} \\ \boldsymbol{c}_{(J-1)}^{\mathrm{T}} & c_J \end{bmatrix} \begin{bmatrix} \boldsymbol{V}_{(J-1,J)} \\ \boldsymbol{v}_{(J)}^{\mathrm{T}} \end{bmatrix} \right) \\ &= c_J \det(\boldsymbol{B}) \det \boldsymbol{V}_{(J)}. \end{aligned}$$

Because $\det(\boldsymbol{B}) > 0$ and $c_J > 0$ and $\det \boldsymbol{U}_{(J)}$, we have $\det \boldsymbol{V}_{(J)} > 0$, for any $\theta_0 < \cdots < \theta_J$. In other words, $\{v_j(\theta)\}_{j=0}^{J}$ forms a Chebyshev system.

□

A.4 Proof of Theorem 2.3.1

Proof. We show that $\gamma_j(\theta)$ in (2.13) is an eigenfunction associated with the eigenvalue λ_j of $A'(\cdot)$. Because $\gamma_j(\theta)$ is bounded over Θ, by changing the order of integrals, we have, for each j,

$$\begin{aligned}
\lambda_j\gamma_j(\theta) &= \lambda_j \int_{\mathcal{S}} \frac{\phi_j(x)}{f_0^{1/2}(x)} f(x;\theta)\mathrm{d}x \\
&= \int_{\mathcal{S}} \frac{f(x;\theta)}{f_0^{1/2}(x)} \int_{\mathcal{S}} \phi_j(x')K(x,x')\mathrm{d}x'\mathrm{d}x \\
&= \int_{\Theta}\int_{\mathcal{S}} \frac{\phi_j(x)}{f_0^{1/2}(x)} \frac{f(x;\theta')}{f_0(x)} f_0(x)\mathrm{d}x \int_{\mathcal{S}} \frac{f(x';\theta)f(x';\theta')}{f_0(x)}\mathrm{d}x'\mathrm{d}\theta' \\
&= \int_{\Theta} \gamma_j(\theta')K'(\theta,\theta')\mathrm{d}\theta'.
\end{aligned}$$

□

A.5 Proof of Theorem 2.3.2

Proof. The concept of strictly totally positive is given in Definition A.1. The probability function $f(x;\theta)$ is an exponential family and thus is strictly totally positive; see [Lindsay and Roeder, 1993]. According to Proposition A.1, the kernel function $K'(\theta,\theta')$ is strictly totally positive. Therefore, the set of eigenfuctions $\{\gamma_j(\theta)\}_{j=0}^J$ forms a Chebyshev system over Θ and $\gamma_0(\theta)$ is strictly one-sign over Θ, by Proposition A.2.

Because $\gamma_0(\theta) \equiv 1/\sqrt{|\Theta|}$ is strictly positive over Θ, we need to show that it is an eigenfunction of the integral operator with the kernel $K'(\theta,\theta')$. We have

$$\int_{\Theta} \gamma_0(\theta)K'(\theta,\theta')\mathrm{d}\theta = \int_{\Theta} \gamma_0(\theta) \int_{\mathcal{S}} \frac{f(s;\theta)}{f_0^{1/2}(x)} \frac{f(x;\theta')}{f_0^{1/2}(x)}\mathrm{d}x\mathrm{d}\theta = |\Theta|^{1/2} = |\Theta|\gamma_0(\theta').$$

□

Chapter 3

Geometry of the Generalized Moment Space

3.1 Introduction

In the previous chapter, we introduce the truncation approximations of the reparameterized mixture (or mixed-effects) models, which are in the form of

$$h_{\nu_0}(s; \boldsymbol{m}_J) = e_0(s)h_0(s) + \sum_{j=1}^{J} m_j e_j(s)h_0(s),$$

where $\boldsymbol{m}_J = (m_0, \ldots, m_J)^{\mathrm{T}} \in \mathbb{R}^{J+1}$ and for each j,

$$m_j = \int_{\Theta} u_j(\theta)\mathrm{d}Q$$

and Q is a probability measure over Θ. Furthermore, $\{u_j(\theta)\}_{j=0}^{J}$ forms a Chebyshev system with $u_0(\theta) \equiv 1$. Note that the truncation approximation $h_{\mu_0}(s; \boldsymbol{m}_J)$ also have a similar expression. Because we do not use the orthonormal property of $\{e_j(s)\}_{j=0}^{J}$ in $L^2(\mathcal{S}, \nu_0)$, we only discuss $h_{\nu_0}(s; \boldsymbol{m}_J)$ in this chapter without losing generality.

The generalized moment space, defined as follows, is a natural parameter space of $\boldsymbol{m}_J$ in $h_{\nu_0}(s; \boldsymbol{m}_J)$.

Definition 3.1.1 (Generalized Moment Space)**.**
Let $\{u_j(\theta)\}_{j=0}^J$ *form a Chebyshev system over a compact set* $\Theta \subseteq \mathbb{R}$ *with* $u_0(\theta) \equiv 1$*. The generalized moment space in* $\mathbb{R}^{J+1}$ *induced by* $\{u_j(\theta)\}_{j=0}^J$ *is*

$$\mathcal{M}_J = \left\{ \boldsymbol{m}_J = (1, m_1, \ldots, m_J)^{\mathrm{T}} \in \mathbb{R}^{J+1} \mid \boldsymbol{m}_J = \int_\Theta \boldsymbol{u}_J(\theta) \mathrm{d}Q \right\}, \tag{3.1}$$

where $\boldsymbol{u}_J(\theta) = (u_0(\theta), \ldots, u_J(\theta))^{\mathrm{T}} \in \mathbb{R}^{J+1}$ *and* Q *is a probability measure over* Θ*.*

In this chapter, we study the geometry of the parameter space $\mathcal{M}_J$. As we will see, convex geometry provides a helpful tool to link the generalized moments $\boldsymbol{m}_J$ to the probability measure Q. We describe this link from two aspects: the positive reparameterization and the gradient characterization. The positive representation reveals the identifiability of Q by its generalized moments $\boldsymbol{m}_J$ and provides an upper bound of the number of the support points of Q; see Section 3.3. On the other hand, the gradient characterization provides the foundation of the class of gradient-based algorithms when the feasible set is the generalized moment space; see Section 3.4.

This chapter is organized as follows. In Section 3.2, we introduce the concept of the generalized moment cone and point out its connection to the generalized moment space. In Section 3.3 and 3.4, we describe the positive representation and the gradient characterization correspondingly. The proof of the theorems in this chapter can be found in the Appendix B.

3.2 Generalized Moment Cone

We introduce the generalized moment cone induced by a Chebyshev system. The generalized moment cone is of interest, because the boundary of the generalized moment space is a subset of the boundary of the generalized moment cone, whose geometry has been well studied; see [Karlin and Studden, 1966, Chapter 2].

Assume that each element of the Chebyshev system $\{u_j(\theta)\}_{j=0}^J$ is a continuous function of θ over $\Theta = [a, b]$ and $u_0(\theta) \equiv 1$. When θ moves from a to b, the trace of $\boldsymbol{u}_J(\theta) \in \mathbb{R}^{J+1}$ forms the moment curve Γ_J in $\mathbb{R}^{J+1}$.

Definition 3.2.1 (Generalized Moment Cone)**.**
Let $\{u_j(\theta)\}_{j=0}^J$ *form a Chebyshev system over a compact set* $\Theta \subseteq \mathbb{R}$ *with* $u_0(\theta) \equiv 1$*. The conical cone of the curve* Γ_J *is called the generalized moment cone induced by* $\{u_j(\theta)\}_{j=0}^J$*, that is*

$$\mathcal{C}_J = \left\{ \boldsymbol{c} = (c_0, \ldots, c_J)^{\mathrm{T}} \in \mathbb{R}^{J+1} \mid \boldsymbol{c} = \int_a^b \boldsymbol{u}_J(\theta) \mathrm{d}\sigma(\theta) \right\},$$

where $\sigma(\theta)$ *is a nondecreasing right continuous function of bounded variation and* $\theta \in [a, b]$*.*

The generalized moment cone contains the convex hull of Γ_{J+1}, which is the generalized moment space $\mathcal{M}_J$, because for each $\boldsymbol{m}_J \in \mathcal{M}_J$, the vector $\boldsymbol{m}_J = \int_a^b \boldsymbol{u}_J(\theta)\mathrm{d}Q(\theta)$, where $Q(\theta)$ is a probability measure over $[a, b]$. Moreover, we give the following result; also see Example 3.1.

Theorem 3.2.1.
If $u_0(\theta) \equiv 1$ *in a Chebyshev system* $\{u_j(\theta)\}_{j=0}^J$ *over* $[a, b]$*, then the boundary of* $\mathcal{M}_J$ *is a subset of the boundary of the generalized moment cone* $\mathcal{C}_J$ *induced by* $\{u_j(\theta)\}_{j=0}^J$*.*

Proof. See the Appendix. □

3.3 Positive Representation

As will be shown, a positive representation of a vector $\boldsymbol{m}_J \in \mathcal{M}_J$ corresponds to a probability measure Q. To illustrate the positive representation of a nonzero vector in $\mathcal{M}_J$, we need to first introduce the positive representation and its index.

Definition 3.3.1 (Positive Representation)**.**
A nonzero vector $\boldsymbol{c}$ *has a positive representation in a Chebyshev system* $\{u_j(\theta)\}_{j=0}^J$*, if it can be written in the form of*

$$\boldsymbol{c} = \sum_{i=1}^r a_i \boldsymbol{u}_J(\theta_i), \tag{3.2}$$

where $\boldsymbol{u}_J(\theta) \in \Gamma_J$*,* $a \leq \theta_1 < \cdots < \theta_r \leq b$ *and* $a_i > 0$*,* $i = 1, \ldots, J$*. If* $\sum_{i=1}^r a_i = 1$*, the positive representation (3.2) is called a convex representation.*

To evaluate the complexity of the positive representation of a nonzero vector $\boldsymbol{c} \in \mathbb{R}^{J+1}$, the index of a positive representation is introduced.

Definition 3.3.2 (Index of a Positive Representation).
Let

$$\mathcal{I}(\theta) = \begin{cases} 1, & \textit{if } \theta \in (a, b); \\ 1/2, & \textit{if } \theta = a \textit{ or } b. \end{cases}$$

If $\boldsymbol{c}$ has the positive representation (3.2), the index of $\boldsymbol{c}$, denoted by $\mathcal{I}(\boldsymbol{c})$, is $\sum_{i=1}^{r} \mathcal{I}(\theta_i)$.

According to Carathéodory's theorem, for each vector $\boldsymbol{m}_J \in \mathcal{M}_J$, there exists a convex representation of $\boldsymbol{m}_J$ by $\{u_j(\theta)\}_{j=0}^{J}$ with $r \leq J+1$. We have the following:

Theorem 3.3.1.
For each $\boldsymbol{m}_J \in \mathcal{M}_J$, the generalized moment space, there exists a probability measure $Q(\theta)$ such that $\boldsymbol{m}_J = \int_a^b \boldsymbol{u}_J(\theta) \mathrm{d}Q(\theta)$ and $Q(\theta)$ has at most $J+1$ support points over $[a, b]$.

If we further assume $\boldsymbol{m}_J$ is on the boundary of $\mathcal{M}_J$, the upper bound of the number of support points can be sharpened using the following proposition.

Proposition 3.3.1 ([Karlin and Studden, 1966, Theorem 2.1]).
A nonzero vector $\boldsymbol{c}$ is a boundary point of $\mathcal{C}_J$ the generalized moment cone, induced by $\{u_j(\theta)\}_{j=0}^{J}$ over $[a, b]$ if and only if $\mathcal{I}(\boldsymbol{c}) < (J+1)/2$. Moreover, its positive representation is unique with $r \leq (J+2)/2$.

With Proposition 3.3.1 and Theorem 3.2.1, we have the following.

Theorem 3.3.2.
If $\boldsymbol{m}_J$ is on the boundary of $\mathcal{M}_J$, there exists one unique probability measure $Q(\theta)$ such that $\boldsymbol{m}_J = \int_a^b \boldsymbol{u}_J(\theta) \mathrm{d}Q(\theta)$ and $Q(\theta)$ has at most $(J+2)/2$ support points.

Example 3.1.
Figure 3.1 shows the generalized moment cones $\mathcal{C}_2$ induced by the power functions $\{\theta^j\}_{j=0}^{2}$ and the Chebyshev polynomials $\{\mathfrak{T}_j(\theta)\}_{j=0}^{2}$, where $\theta \in [-1, 1]$. In each plot, the curve Γ_2 is induced by the corresponding Chebyshev system.

The boundary of $\mathcal{C}_2$ contains the boundary of $\mathcal{M}_2$; see Theorem 3.2.1. The boundary vectors of $\mathcal{M}_2$ are either $\boldsymbol{u}_2(\theta) \in \mathbb{R}^3$ or $(1-\alpha)\boldsymbol{u}_2(-1)+\alpha\boldsymbol{u}_2(1)$, where $0 < \alpha < 1$. Therefore, the index of a boundary vector is either 1 or $1/2$; see Theorem 3.3.2. On the other hand, if the index of a vector is less than $3/2$, it must locate on the boundary. Moreover, when $\boldsymbol{m}_2$ is on the boundary, it uniquely corresponds to a probability measure. For example, one point on Γ_2 is the image of $h(s;0)$ in $\mathbb{R}^3$, where $h(s;\theta)$ is the component of $h_{\text{Mix}}(s;Q)$ in Equation (2.1).

3.4 Gradient Characterization

The gradient characterization is useful for computational algorithms. In the literature of the NPMLE for mixture models, there exists a class of computational algorithms based on the same convex structure as considered here; see [Böhning et al., 1992] and [Wang, 2007]. This class has more stable computational speeds than the EM algorithm, which is also commonly used for mixture models.

In this subsection, we consider the following optimization problem

$$\min_{\boldsymbol{m}_J \in \mathcal{M}_J} \mathcal{L}(\boldsymbol{m}_J) \tag{3.3}$$

where $\mathcal{L}(\boldsymbol{m}_J)$ is an arbitrary loss function and strictly convex with respect to $\boldsymbol{m}_J$.

Since the optimization problem (3.3) is convex, its solution $\hat{\boldsymbol{m}}_J$ is unique in $\mathcal{M}_J$. There exists a supporting hyperplane of $\mathcal{M}_J$ at $\hat{\boldsymbol{m}}_J$ such that

$$\mathcal{H} = \left\{ \boldsymbol{h} = (1, h_1, \ldots, h_J)^{\mathrm{T}} \in \mathbb{R}^{J+1} \mid (\hat{\boldsymbol{m}}_J - \boldsymbol{h})^{\mathrm{T}} \bigtriangledown \mathcal{L}(\hat{\boldsymbol{m}}_J) = 0 \right\}.$$

The following theorem states the relationship between $\mathcal{H}$ and the support points of $\hat{Q}$ in Theorem 3.3.2.

Theorem 3.4.1.
Let $\hat{\Theta}$ be the set of support points of $\hat{Q}$. Then, if a point $\hat{\theta} \in [a,b]$ is an element of $\hat{\Theta}$, then $\boldsymbol{u}_J(\hat{\theta})$ is on the hyperplane $\mathcal{H}$. The converse also holds.

Proof. See the Appendix. □

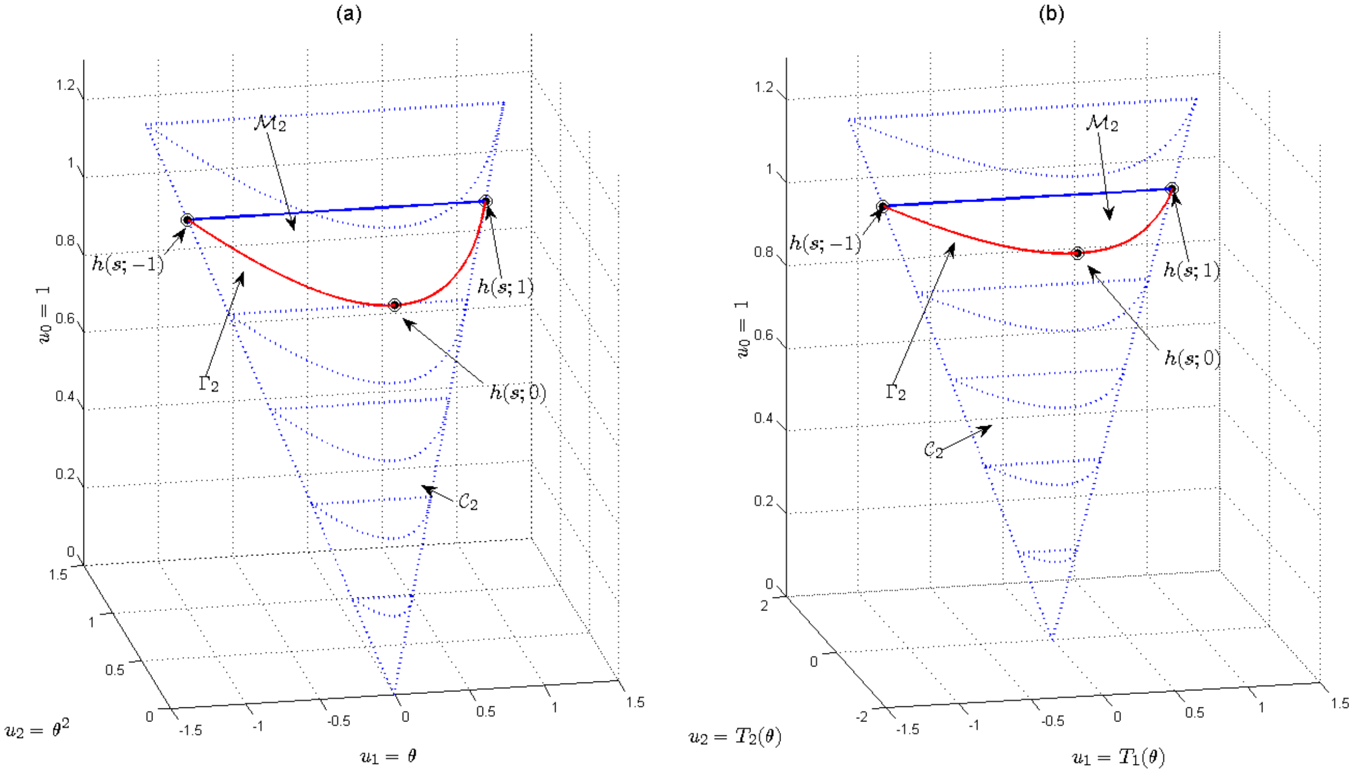

Figure 3.1: Plots of the moment cones induced by (a) $\{\theta^j\}_{J=0}^2$; and (b) $\{T_j(\theta)\}_{j=0}^2$.

The above theorem also implies that $\hat{\Theta}$ is the set of zeros of the gradient function of the objective function $\mathcal{L}(\boldsymbol{m}_J)$ which is defined as

$$\begin{aligned}\mathcal{D}(\hat{\boldsymbol{m}}_J, \boldsymbol{u}_J(\theta)) &= \frac{\partial}{\partial \epsilon}\mathcal{L}((1-\epsilon)\hat{\boldsymbol{m}}_J + \epsilon \boldsymbol{u}_J(\theta))\bigg|_{\epsilon=0} \\ &= (\boldsymbol{u}_J(\theta) - \hat{\boldsymbol{m}}_J)^{\mathrm{T}} \bigtriangledown \mathcal{L}(\hat{\boldsymbol{m}}_J).\end{aligned}$$

Moreover, we can use the gradient function to characterize $\hat{\boldsymbol{m}}_J$ as follows.

Theorem 3.4.2.
The following three statements are equivalent:

1. $\hat{\boldsymbol{m}}_J$ *minimizes* $\mathcal{L}(\boldsymbol{m}_J)$.

2. $\inf_\theta \mathcal{D}(\hat{\boldsymbol{m}}_J, \boldsymbol{u}_J(\theta)) = 0$.

3. $\hat{\boldsymbol{m}}_J$ *maximizes* $\inf_\theta \mathcal{D}(\boldsymbol{m}_J, \boldsymbol{u}_J(\theta))$.

Proof. See the Appendix. □

Now, we continue Example 3.1 to illustrate Theorem 3.4.1 and 3.4.2.

Example 3.1 (continued)**.**
In each panel of Figure 3.2, we see the images of the curve Γ_2 *induced by* $\{1, T_1(\theta), T_2(\theta)\}$ *and its convex hull* $\mathcal{M}_2$ *in the space of* $(m_1, m_2) \in \mathbb{R}^2$. *The contours show the identical values of the objective function*

$$\mathcal{L}(\boldsymbol{m}_2) = (\boldsymbol{t} - \boldsymbol{m}_2)^{\mathrm{T}}(\boldsymbol{t} - \boldsymbol{m}_2),$$

where $\boldsymbol{m}_2 = (1, m_1, m_2)^{\mathrm{T}} \in \mathbb{R}^3$ *and* $\boldsymbol{t} = (-0.7, -1, 0)^{\mathrm{T}} \in \mathbb{R}^3$. *Because* $\boldsymbol{t} \notin \mathcal{M}_2$, $\mathcal{L}(\boldsymbol{m}_2)$ *is strictly convex with respect to* $\boldsymbol{m}_2$. *The minimum value of* $\mathcal{L}(\boldsymbol{m}_2)$ *over* $\mathcal{M}_2$ *is* 0.1825. *As we see, the contour* $\mathcal{L}(\boldsymbol{m}_2) = 0.1825$ *has a unique intersection* $\hat{\boldsymbol{m}}_2$ *with* $\mathcal{M}_2$. *Moreover, the intersection* $\hat{\boldsymbol{m}}_2$ *is on the boundary of* $\mathcal{M}_2$. *In Figure 3.2(a), the solid line represents the supporting hyperplane* $\mathcal{H}$ *of* $\mathcal{M}_2$ *at* $\hat{\boldsymbol{m}}_2$. *Here we have* $\hat{\boldsymbol{m}}_2 = \boldsymbol{u}_2(\hat{\theta}) \in \mathcal{H}$; *see Theorem 3.4.1.*

Moreover, $\bigtriangledown\mathcal{L}(\hat{\boldsymbol{m}}_2)$ *is orthogonal to the supporting hyperplane. For any vector* $\boldsymbol{u}_2(\theta) \neq \hat{\boldsymbol{m}}_2$ *on* Γ_2, *we have the vector* $\boldsymbol{u}_2(\theta) - \hat{\boldsymbol{m}}_2$. *From Figure 3.2(a), it can be* *seen that the angle $\psi \in [0, \pi]$ between $\bigtriangledown\mathcal{L}(\hat{\boldsymbol{m}}_2)$ and $\boldsymbol{u}_2(\theta) - \hat{\boldsymbol{m}}_2$ is always acute. Therefore, we have*

$$\cos(\psi) = \frac{\mathcal{D}(\hat{\boldsymbol{m}}_2, \boldsymbol{u}_2(\theta))}{\sqrt{(\boldsymbol{u}_2(\theta) - \hat{\boldsymbol{m}}_2)^{\mathrm{T}}(\boldsymbol{u}_2(\theta) - \hat{\boldsymbol{m}}_2)}\sqrt{(\bigtriangledown\mathcal{L}(\hat{\boldsymbol{m}}_2))^{\mathrm{T}}(\bigtriangledown\mathcal{L}(\hat{\boldsymbol{m}}_J))}} > 0;$$

see Theorem 3.4.2 (2). It is also obvious that $\cos(\psi) = 0$ *if and only if* $\boldsymbol{u}_2(\theta) = \hat{\boldsymbol{m}}_2$. *In Figure 3.2(b), we see that for any $\boldsymbol{m}_2' \neq \hat{\boldsymbol{m}}_2$ in $\mathcal{M}_2$, there always exists a $\boldsymbol{u}_2(\theta) \in \Gamma_2$ such that the angle ψ' between $\bigtriangledown\mathcal{L}(\boldsymbol{m}_2')$ and $\boldsymbol{u}_2(\theta') - \boldsymbol{m}_2'$ is obtuse. It follows that* $\inf_\theta \mathcal{D}(\boldsymbol{m}_2', \boldsymbol{u}_2(\theta')) < 0$*; see Theorem 3.4.2 (3).*

Appendix: B

B.1 Proof of Theorem 3.2.1

Proof. We want to show that for each boundary vector $\boldsymbol{m}_J^*$ of $\mathcal{M}_J$, there exists a supporting hyperplane of $\mathcal{M}_J$ at $\boldsymbol{m}_J^*$ which is also a supporting hyperplane of $\mathcal{C}_J$ at $\boldsymbol{m}_J^*$.

Firstly, the convex hull $\mathcal{M}_J$ is the intersection of $\mathcal{C}_J$ and the hyperplane $\mathcal{H}_1 = \{\boldsymbol{h} = (1, h_1, \ldots, h_J)^{\mathrm{T}} \in \mathbb{R}^{J+1}\}$. Then, in $\mathcal{H}_1$, there exists a vector $\tilde{\boldsymbol{a}}_J = (\tilde{a}_1, \ldots, \tilde{a}_J)^{\mathrm{T}} \in \mathbb{R}^J$ such that for each $\boldsymbol{m}_J \in \mathcal{M}_J$, we have

$$\sum_{j=1}^{J} m_j \tilde{a}_j \geq \sum_{j=1}^{J} m_j^* \tilde{a}_j.$$

Let $\tilde{a}_0 = -\sum_{j=1}^{J} m_j^* \tilde{a}_j$. For each $\boldsymbol{m}_J \in \mathcal{M}_J$, we have

$$\sum_{j=0}^{J} m_j \tilde{a}_j \geq \sum_{j=0}^{J} m_j^* \tilde{a}_j.$$

Therefore, the vector $\tilde{\boldsymbol{a}} = (\tilde{a}_0, \tilde{a}_J^{\mathrm{T}})^{\mathrm{T}} \in \mathbb{R}^{J+1}$ determines the hyperplane

$$\{\boldsymbol{h} \in \mathbb{R}^{J+1} | (\boldsymbol{h} - \boldsymbol{m}_J^*)^{\mathrm{T}} \tilde{\boldsymbol{a}} = 0\} \tag{B.1}$$

as a supporting hyperplane of $\mathcal{M}_J$ at $\boldsymbol{m}_J^*$.

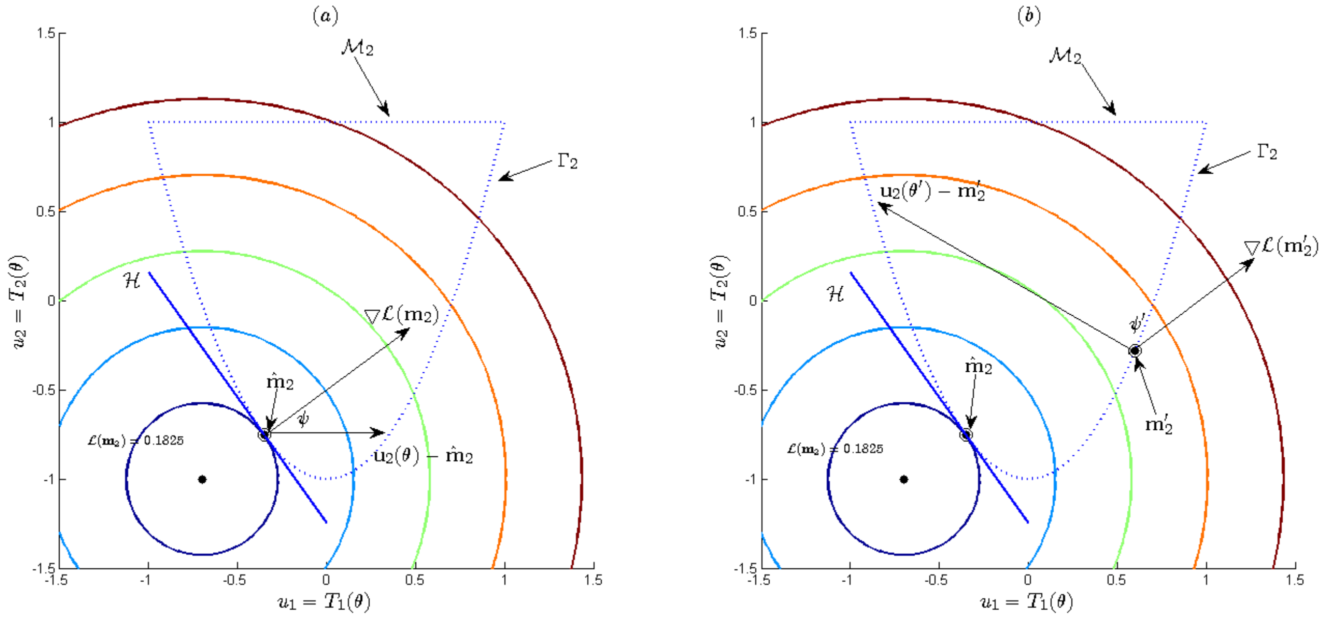

Figure 3.2: Plots of $\mathcal{M}_2$ induced by $\{T_j(\theta)\}_{j=0}^2$ and a visual interpretation of Theorem 3.4.1 and 3.4.2.

Note that any vector in $\mathcal{C}_J$ can be written as $\Delta\boldsymbol{m}_J$, where $\Delta \geq 0$ and $\boldsymbol{m}_J \in \mathcal{M}_J$. We have the inequality:

$$\begin{aligned}\sum_{j=0}^{J}(\Delta m_j - m_j^*)\tilde{a}_j &= (\Delta - 1)\tilde{a}_0 + \sum_{j=1}^{J}(\Delta m_j - m_j^*)\tilde{a}_j \\ &= (1-\Delta)\sum_{j=1}^{J} m_j^*\tilde{a}_j + \sum_{j=1}^{J}(\Delta m_j - m_j^*)\tilde{a}_j \\ &= \Delta\sum_{i=1}^{J}(m_j - m_j^*)\tilde{a}_j \geq 0,\end{aligned}$$

and thus the hyperplane (B.1) is also a supporting hyperplane of $\mathcal{C}_J$. □

B.2 Proof of Theorem 3.4.1

Proof. Because $\mathcal{H}$ is a supporting hyperplane of $\mathcal{M}_J$ at $\hat{\boldsymbol{m}}_J$, we have for any point $\boldsymbol{m}'_J \in \mathcal{M}_J$ but $\notin \mathcal{H}$,

$$(\hat{\boldsymbol{m}}_J - \boldsymbol{m}'_J)^{\mathrm{T}} \bigtriangledown \mathcal{L}(\hat{\boldsymbol{m}}_J) < 0.$$

Assume that there exists a $\hat{\theta} \in \hat{\Theta}$ such that $\boldsymbol{u}_J(\hat{\theta}) \notin \mathcal{H}$. Then,

$$\mathcal{D}(\hat{\boldsymbol{m}}_J, \boldsymbol{u}(\hat{\theta})) = (\hat{\boldsymbol{m}}_J - \boldsymbol{u}(\hat{\theta}))^{\mathrm{T}} \bigtriangledown \mathcal{L}(\hat{\boldsymbol{m}}_J) < 0.$$

In other words, the objective function can be decreased along the direction to $\boldsymbol{u}(\hat{\theta})$. Such statement conflicts to the fact that $\hat{\boldsymbol{m}}_J$ minimizes the objective function and is unique. Therefore, $\boldsymbol{u}_J(\hat{\theta})$ must locates on $\mathcal{H}$. □

B.3 Proof of Theorem 3.4.2

Proof. The first statement that $\hat{\boldsymbol{m}}_J$ minimizes the objective function $\mathcal{L}(\boldsymbol{m}_J)$ holds if and only if its path derivative from $\hat{\boldsymbol{m}}_J$ to any other $\boldsymbol{u}_J(\theta)$ is non-negative. In other words, we have $\inf_\theta \mathcal{D}(\hat{\boldsymbol{m}}_J, \boldsymbol{u}_J(\theta)) = 0$. Therefore, Statement 1 and Statement 2 are equivalent.

Because the objective function $\mathcal{L}(\boldsymbol{m}_J)$ is strictly convex along any path, for any $\theta \in \Theta$, we have

$$\mathcal{L}(\boldsymbol{m}_J) \geq \mathcal{L}(\boldsymbol{u}_J(\theta)) + \mathcal{D}(\boldsymbol{m}_J, \boldsymbol{u}_J(\theta)).$$

For some $\boldsymbol{m}'_J \neq \hat{\boldsymbol{m}}_J$, it is true

$$\mathcal{L}(\boldsymbol{m}'_J) \geq \mathcal{L}(\boldsymbol{u}_J(\theta)) + \mathcal{D}(\boldsymbol{m}'_J, \boldsymbol{u}_J(\theta)).$$

If $\inf_\theta \mathcal{D}(\boldsymbol{m}'_J, \boldsymbol{u}_J(\theta)) \geq 0$, we would have $\mathcal{L}(\boldsymbol{m}'_J) \geq \mathcal{L}(\hat{\boldsymbol{m}}_J)$. This is contradiction to the fact that $\hat{\boldsymbol{m}}_J$ minimizes $\mathcal{L}(\boldsymbol{m}_J)$ over $\mathcal{M}_J$. Therefore, Statement 1 and Statement 3 are equivalent. □

Chapter 4

The Generalized Method of Moments for Mixture Models

4.1 Introduction

Many existing methods can be used to fit one-parameter mixture model $f_{\mathrm{Mix}}(x; Q)$ in Definition 2.3.1. The commonly used ones include the method of moments (MM), minimum distance methods and the maximum likelihood method.

As early [Pearson, 1898], the MM has been used to fit a mixture of two normal distributions with different mean and variance; see [Lindsay, 1989a,b] for further developments. Because computing the MM estimators involves solving a polynomial equation system, it is computational friendly. And thus, it is often used as an initial value for iterative numerical algorithms for other estimation methods; see [Furman and Lindsay, 1994]. However, the MM can be used only when the component distributions are the NEF-QVF; see [Lindsay, 1989b]. Another issue of the MM is the potential loss of efficiency comparing to the other methods; [Titterington et al., 1985, p.g. 81].

A detailed review of the minimum distance estimators for mixture models can be found in [Titterington et al., 1985, p.g. 114-117]. A lot of distance measures, including the Kullbak-Leibler, Levy, chi-squared, modified chi-squared and averaged

L^2-norm measures, can be used to measure the difference between the mixture models and the empirical distributions; see [Titterington et al., 1985, p.g. 116]. Minimizing the Kullbak-Leibler distance between the empirical distribution and a mixture model is equivalent to maximizing the likelihood; see [Titterington et al., 1985, p.g. 115]. The minimum Hellinger distance method for mixture models has attracted many research interest, because the model complexity can be robustly estimated under error contamination; see [Cutler and Cordero-Braña, 1996] and [Woo and Sriram, 2006].

The MLE for mixture models is popular, partly because of the philosophy of likelihood-based inference; see [Titterington et al., 1985, p.g. 82]. However, due to non-regularity, there are many inference and computational challenges in the maximum likelihood methods for either finite mixture models or non-parametric mixture models; see Section 1.2.

This chapter aims to fit a non-parametric mixture model based on a set of generalized moment conditions (see Definition 4.2.1), which are from the reparameterization procedure introduced in Subsection 2.3.2. The proposed method is called the generalized method of moments (GMM) for mixture models. Computing the GMM estimator is a constrained quadratic minimization problem, which can be easily solved by the gradient-based algorithms; see Section 4.7. The mean squared error (MSE) of the GMM estimators converges to zero, as the sample size goes to infinity; see Section 4.5. Moreover, the GMM estimators are robust to the outliers when the quadratic objective functions are carefully designed; see Section 4.6.

The main contribution of this chapter is the introduction of the GMM estimator for mixture models. Asymptotic behaviour of the MSE and its robustness to outliers are also studied.

This chapter is organized as follows. In Section 4.2, the generalized moment conditions are introduced in our context. A set of countable generalized moment conditions is obtained from the reparameterization procedure of mixture models. In Section 4.3, we define the GMM for mixture models based on the generalized moment conditions. Also, we discuss the situation when the GMM estimator is not unique. In Section 4.4, we describe the GMM for mixture models in an information geometric view, when the weighting matrices are identity matrix. In Section 4.5, we show the convergence rate of the MSE of the GMM estimators with the sample size; see Theorem 4.5.1. In Section 4.6, we introduce a weighting matrix that leads to robust GMM estimators. Our work is supported by simulation studies in Section 4.7. Lastly, we apply the GMM to fit a mixture model for the Thailand cohort study data, which has been described in Subsection 1.6.1. The poofs of the theorems and lemmas and the MATLAB code for the proposed algorithms in this chapter can be seen in the Appendix C.

4.2 The Generalized Moment Conditions

The generalized method of moments was firstly proposed by Hansen [1982]. Later, it becomes popular in econometrics; see [Mátyás, 1999] and [Hall, 2005] for comprehensive introductions. This method is based on the a series of generalized moment conditions defined as follows in our context. As we will see in this section, the generalized moment conditions can be easily constructed for the considered mixture models.

Definition 4.2.1 (Generalized Moment Conditions).
Suppose that the random variable X is from a mixture model $f_{\text{Mix}}(x; Q^*)$. Let $(\phi(x), \gamma(\theta))$ be a pair of known functions such that

$$\mathbb{E}_X[\phi(X)] = \mathbb{E}_\theta[\gamma(\theta)], \tag{4.1}$$

where X follows the true mixture model $f_{\text{Mix}}(x; Q^*)$ and θ follows the true mixing distribution Q^*. Equation (4.1) is called a generalized moment condition.

Recall that we have the following reparameterization of $f_{\text{Mix}}(x; Q)$ under the assumptions given in Subsection 2.3.2

$$f_{\text{spec}}(x; \boldsymbol{m}) = f_0(x) + \sum_{j=1}^{\infty} m_j \phi_j(x) f_0^{1/2}(x),$$

where for each j,

$$m_j = \int_\Theta \sqrt{\lambda_j} \gamma_j(\theta) \mathrm{d}Q$$

and $\phi_j(x)$ is the eigenfunction associated with the j^{th} largest eigenvalue λ_j of the integral operator $A(\cdot)$ in Equation (2.12) and

$$\sqrt{\lambda_j}\gamma_j(\theta) = \int_{\mathcal{S}} f(x;\theta)\phi_j(x) f_0^{-1/2}(x)\mathrm{d}x.$$

By taking the expectation with respect to θ on both sides of the above equation and changing the order of integrals, we have, for each $j \in \{0, 1, \ldots\}$,

$$\int_{\Theta} \sqrt{\lambda_j}\gamma_j(\theta)\mathrm{d}Q = \int_{\mathcal{S}} \phi_j(x) f_0^{-1/2}(x) f_{\mathrm{Mix}}(x;Q)\mathrm{d}x.$$

It follows that, for each $j \in \{0, 1, \ldots\}$,

$$\mathbb{E}_X\left[\phi_j(X) f_0^{-1/2}(X) - m_j\right] = 0. \tag{4.2}$$

Therefore, there exist a countable number of generalized moment conditions for mixture models.

4.3 The Generalized Method of Moments

We use the first $J+1$ generalized moment conditions for the generalized method of moments, where J is a positive integer. Let $\boldsymbol{\gamma}_J(\theta) = (\gamma_1(\theta), \ldots, \gamma_J(\theta))^{\mathrm{T}}$, $\boldsymbol{\phi}_{f_0^{-1/2}}(x) = (\phi_1(x) f_0^{-1/2}(x), \ldots, \phi_J(x) f_0^{-1/2}(x))^{\mathrm{T}}$ and $\boldsymbol{m} = (m_1, \ldots, m_J)^{\mathrm{T}}$, which are the vectors in $\mathbb{R}^J$. Given a random sample $X_1, \ldots, X_N$, we estimate $\boldsymbol{m}$ by the sample average of $\boldsymbol{\phi}_{f_0^{-1/2}}(x)$ that

$$\bar{\boldsymbol{m}} = \frac{1}{N}\sum_{n=1}^{N} \boldsymbol{\phi}_{f_0^{-1/2}}(x_n) \in \mathbb{R}^J.$$

However, the simple estimator $\bar{\boldsymbol{m}}$ may not respect the constraints on $\boldsymbol{m}$. Therefore, we use the generalized method of moments, defined as following.

Definition 4.3.1 (The GMM Estimator for Mixture Models)**.**
Given a random sample $X_1, \ldots, X_N$ from a mixture model $f_{\mathrm{Mix}}(x;Q)$ and a fixed J, the generalized method of moments estimator for mixture models with order J, denoted

by $\hat{Q}_{\mathrm{GMM},J}$, is the solution to the following optimization problem

$$\begin{aligned} \min_{Q} \quad & (\bar{\boldsymbol{m}} - \boldsymbol{m})^{\mathrm{T}} \boldsymbol{W}_{(J)} (\bar{\boldsymbol{m}} - \boldsymbol{m}) \\ s.t. \quad & \boldsymbol{m}(Q) = \int_{\Theta} \boldsymbol{\Lambda}^{1/2} \boldsymbol{\gamma}(\theta) \mathrm{d}Q \in \mathbb{R}^{J}, \\ & Q \textit{ is a probability measure over } \Theta = [a, b], \end{aligned} \tag{4.3}$$

where $\boldsymbol{W}_{(J)}$ is a $J \times J$ positive definite matrix and $\boldsymbol{\Lambda}$ is a $J \times J$ diagonal matrix whose j^{th} diagonal element is λ_j.

The positive definite matrix $\boldsymbol{W}_{(J)}$ is called the weighting matrix. There are various choices of the weighting matrix $\boldsymbol{W}_{(J)}$. One simple choice is the identity matrix. Another popular choice is the inverse of the covariance matrices of $\boldsymbol{\phi}_{f_0^{-1/2}}(X)$, which provides the most efficient GMM estimator under the regularity conditions in [Mátyás, 1999, Section 1.3]. In our context, the inverse of the covariance matrix may not provide the most efficient GMM estimator due to the existence of the boundaries in the parameter space. It is important to choose a suitable weighting matrix; see Section 4.6 in which the weighting matrix is chosen for the robustness property.

Note that $\gamma_0(\theta) \equiv 1$. The feasible set of the optimization problem (4.3) is equivalent to $\boldsymbol{m}_J \in \mathcal{M}_J$, where $\boldsymbol{m}_J = (1, \boldsymbol{m}^{\mathrm{T}})^{\mathrm{T}} \in \mathbb{R}^{J+1}$ and $\mathcal{M}_J$ is the generalized moment space induced by $\{\sqrt{\lambda_j}\gamma_j(\theta)\}_{j=0}^{J}$ in Definition 3.1.1. Therefore, the vector $\hat{\boldsymbol{m}}_J = (1, \hat{\boldsymbol{m}}^{\mathrm{T}})^{\mathrm{T}} \in \mathbb{R}^{J+1}$ is the projection of $\bar{\boldsymbol{m}}_J = (1, \bar{\boldsymbol{m}}^{\mathrm{T}})^{\mathrm{T}} \in \mathbb{R}^{J+1}$ onto the generalized moment space $\mathcal{M}_J$, where $\hat{\boldsymbol{m}} = \boldsymbol{m}(\hat{Q}_{\mathrm{GMM},J})$. Furthermore, the GMM estimator $\hat{Q}_{\mathrm{GMM},J}$ is the positive representation of $\hat{\boldsymbol{m}}_J$.

Because $\boldsymbol{W}_{(J)}$ is positive definite, finding $\hat{\boldsymbol{m}}_J$ is a convex optimization problem and there exists a unique $\hat{\boldsymbol{m}}_J \in \mathcal{M}_J$. However, the uniqueness of $\hat{Q}_{\mathrm{GMM},J}$ depends on $\hat{\boldsymbol{m}}_J$. When $\hat{\boldsymbol{m}}_J$ is on the boundary of $\mathcal{M}_J$, the GMM estimator $\hat{Q}_{\mathrm{GMM},J}$ is unique and has at most $J/2$ support points over Θ; see Theorem 3.3.2. And, we use $f_{\mathrm{Mix}}(x; \hat{Q}_{\mathrm{GMM},J})$ to fit the mixture model. Otherwise, there is no unique $\hat{Q}_{\mathrm{GMM},J}$ and more generalized moment conditions are necessary to obtain a unique $\hat{Q}_{\mathrm{GMM},J}$.

Given a random sample $X_1, \ldots, X_N$, we have a series of vectors $\hat{\boldsymbol{m}}_J \in \mathbb{R}^{J+1}$, where $J = 2, 3, \ldots$. Let $\mathcal{J}$ be the smallest integer such that $\hat{\boldsymbol{m}}_{\mathcal{J}}$ is not an interior point of

$\mathcal{M}_{\mathcal{J}}$. Note that $\mathcal{J}$ is a random variable, because it depends on the random sample. Let J_N be a number which increases with the sample size N. Also let $\mathcal{A}_1$ be the event that $\mathcal{J} \geq J_N$. In the event $\mathcal{A}_1$, we fit the mixture model with

$$\begin{aligned} f_{\text{spec}}(x; \bar{\boldsymbol{m}}) &= f_0(x) + \bar{\boldsymbol{m}}^{\mathrm{T}} \boldsymbol{\phi}_{f_0^{1/2}}(x) \\ &= f_0(x) + \frac{1}{N} \sum_{n=1}^{N} \boldsymbol{\phi}^{\mathrm{T}}_{f_0^{-1/2}}(x_n) \boldsymbol{\phi}_{f_0^{1/2}}(x), \end{aligned}$$

where $\boldsymbol{\phi}_{f_0^{1/2}}(x) = (\phi_1(x) f_0^{1/2}(x), \ldots, \phi_{J_N}(x) f^{1/2}(x))^{\mathrm{T}} \in \mathbb{R}^{J_N}$ and $\boldsymbol{\phi}_{f_0^{-1/2}}(x) \in \mathbb{R}^{J_N}$. In summary, the fitted model is

$$\hat{f}_{\text{GMM}}(x) = \begin{cases} f_{\text{spec}}(x; \bar{\boldsymbol{m}}), & \text{if the event } \mathcal{A}_1 \text{ happens,} \\ f_{\text{Mix}}(x; \hat{Q}_{\text{GMM},\mathcal{J}}), & \text{otherwise.} \end{cases} \tag{4.4}$$

4.4 An Information Geometric View of The Generalized Method of Moments

Zhang [2013] considered the information geometry of an affine submanifold formed by a parametric model. It is known that a divergence function can uniquely determine the information geometry of a statistical manifold, including a Riemannian metric given by the Fisher information and a pair of dual connections that preserve the metric under parallel transport by their joint actions; see [Zhang, 2013]. In this section, we rewrite the objective function in the optimization problem (4.3) as a divergence function under the framework given in [Zhang, 2013]. The Riemannian metric and the pair of dual connection follow the divergence function; see [Zhang, 2005] and [Zhang, 2013].

Let $G : \mathbb{R} \to \mathbb{R}$ be a strictly convex function. Its convex conjugate G_{conj} is given by

$$G_{\text{conj}}(t) = t \times (\partial G)^{-1}(t) - G\left((\partial G)^{-1}(t)\right),$$

where ∂G is the first order derivative of $G(t)$ with respect to t, and $(\partial G)^{-1}(t)$ is the inverse function of ∂G; see [Zhang, 2013]. The conjugate representation of a probability function is defined as follows.

Definition 4.4.1 (Conjugate Representation [Zhang, 2004])**.**
For a strictly increasing function $\rho : \mathbb{R} \to \mathbb{R}$, the ρ-representation of a probability function pr *is the mapping* $\mathrm{pr} \mapsto \rho(\mathrm{pr})$*. For a strictly increasing function $\tau : \mathbb{R} \to \mathbb{R}$, the τ-representation of the probability function,* $\mathrm{pr} \mapsto \tau(\mathrm{pr})$ *is conjugate to the ρ-representation of the probability function* pr *with respect to a smooth and strictly convex function $G : \mathbb{R} \to \mathbb{R}$, if*

$$\tau(\mathrm{pr}) = \partial G\left(\rho(\mathrm{pr})\right) \Leftrightarrow \rho(\mathrm{pr}) = \partial G_{\mathrm{conj}}\left(\tau(\mathrm{pr})\right).$$

In our context, given the initial measure $f_0(x)$, let

$$\rho(\mathrm{pr}) = \mathrm{pr}/f_0(x)$$

and

$$G(\rho(\mathrm{pr})) = \frac{1}{2}\left(\rho(\mathrm{pr})\right)^2 f_0(x). \tag{4.5}$$

Then, $\tau(\mathrm{pr}) = \mathrm{pr}$. Consider the ρ-representation of $f_{\mathrm{spec}}(x; \boldsymbol{m})$, we have

$$\rho(f_{\mathrm{spec}}(x; \boldsymbol{m})) = f_{\mathrm{spec}}(x; \boldsymbol{m})/f_0(x) = 1 + \boldsymbol{m}^{\mathrm{T}}\boldsymbol{\phi}_{f_0^{-1/2}}(x)$$

where $\boldsymbol{m} = (m_1, \ldots, m_{J_N})^{\mathrm{T}} \in \mathbb{R}^{J_N}$. The model $f_{\mathrm{spec}}(x; \boldsymbol{m})$ is called ρ-affine, because its ρ-representation can be embedded into a countable-dimensional affine space; see [Zhang, 2013]. Here we generalize the dimension of the affine space in the definition of ρ-affine from finite to countable. The parameter $\boldsymbol{m} \in \mathbb{R}^{J_N}$ is called the natural parameter of $f_{\mathrm{spec}}(x; \boldsymbol{m})$. On the other hand, for any $f_{\mathrm{Mix}}(x; Q)$, the expectation parameter of $f_{\mathrm{spec}}(x; \boldsymbol{m})$ is defined as the projection of $\tau\left(f_{\mathrm{Mix}}(x; Q)\right)$ onto the functions $\boldsymbol{\phi}_{f_0^{-1/2}}(x)$, i.e.

$$\int_{\mathcal{S}} \boldsymbol{\phi}_{f_0^{-1/2}}(x)\tau\left(f_{\mathrm{Mix}}(x; Q)\right) \mathrm{d}x;$$

see [Zhang, 2013]. By the generalized moment conditions in Equation (4.2), we have that the natural parameter and expectation parameter are identical in $f_{\mathrm{spec}}(x; \boldsymbol{m})$.

Definition 4.4.2 ([Zhang, 2013])**.**
Let $G : \mathbb{R} \to \mathbb{R}$ be smooth and strictly convex, and $\rho : \mathbb{R} \to \mathbb{R}$ be strictly increasing.

The canonical divergence function between two probability functions pr *and* pr′ *of* $x \in \mathcal{S}$ *is*

$$\mathfrak{A}_G(\rho(\mathrm{pr}), \tau(\mathrm{pr}')) = \int \left(G(\rho(\mathrm{pr})) + G_{\mathrm{conj}}(\tau(\mathrm{pr}')) - \rho(\mathrm{pr})\tau(\mathrm{pr}')\right) \mathrm{d}\mu,$$

where μ *is a measure such that* $\mathrm{d}\mu = \mu(\mathrm{d}x)$*.*

In [Zhang, 2013], the formula of the canonical divergence function is given, when the parametric model is ρ-affine. By Corollary 11 in [Zhang, 2013], we find that minimizing the objective function in the GMM with the identity weighting matrix is equivalent to minimizing a canonical divergence function.

Corollary 4.4.1.
Let

$$\Phi(\boldsymbol{m}; \mu) = \int_{\mathcal{S}} G\left(\rho(f_{\mathrm{spec}}(x; \boldsymbol{m}))\right) \mathrm{d}\mu,$$

where μ *is a probability measure of* x *defined on* $\mathcal{S}$*,* G *is defined in Equation (4.5) and* $\rho(f_{\mathrm{spec}}(x; \boldsymbol{m})) = f_{\mathrm{spec}}(x; \boldsymbol{m})/f_0(x)$*. Then, the canonical divergence function between the empirical distribution and* $f_{\mathrm{Mix}}(x; Q)$ *is*

$$\mathfrak{A}_\Phi(\boldsymbol{m}, \bar{\boldsymbol{m}}) = \frac{1}{2}\boldsymbol{m}^{\mathrm{T}}\boldsymbol{m} + \frac{1}{2}\bar{\boldsymbol{m}}^{\mathrm{T}}\bar{\boldsymbol{m}} - \bar{\boldsymbol{m}}^{\mathrm{T}}\boldsymbol{m} = \frac{1}{2}\left(\boldsymbol{m} - \bar{\boldsymbol{m}}\right)^{\mathrm{T}}\left(\boldsymbol{m} - \bar{\boldsymbol{m}}\right).$$

4.5 The Quality of Point Estimators

Suppose that the parameter $\tau = \mathbb{E}[s(X)] < \infty$ is of interest, where X follows $f_{\mathrm{Mix}}(x; Q^*)$ and $s(x) \in L^2(\mathcal{S}, \nu_0)$ is a known function of x. Given $\hat{f}_{\mathrm{GMM}}(x)$, the GMM estimator of τ is

$$\hat{\tau}_{\mathrm{GMM}} = \int_{\mathcal{S}} s(x)\hat{f}_{\mathrm{GMM}}(x)\mathrm{d}x. \tag{4.6}$$

According to the following theorem, we know that the mean squared error of $\hat{\tau}_{\mathrm{GMM}}$ converges to zero as the sample size goes to infinity.

Theorem 4.5.1.
Let $X_1, \ldots, X_N$ *be a random sample from a mixture of exponential families* $f_{\mathrm{Mix}}(x; Q^*)$*,*

where the mixing distribution Q^* is defined on a compact set $\Theta = [a, b]$. Given a set of weighting matrices $\{\boldsymbol{W}_{(J)}, J = 2, 3, \ldots\}$, suppose that $\sup_J \|\boldsymbol{W}_{(J)}\|_2$ is bounded. Further suppose that the covariance matrix $\boldsymbol{\phi}_{f_0^{-1/2}}(X) \in \mathbb{R}^J$ is non-singular for each J. Then, for each positive integer r, the mean squared error of the GMM estimator $\hat{\tau}_{\mathrm{GMM}}$ has the optimal convergence rate $O(N^{-r/(r+1)})$, when $J_N^{(2r+2)} N^{-1} = O(1)$, i.e.,

$$\mathbb{E}\left[(\tau - \hat{\tau}_{\mathrm{GMM}})^2\right] = O(N^{-r/(r+1)}).$$

To prove the above theorem, we study the MSE of $\hat{\tau}_{\mathrm{GMM}}$ in the three possible events $\mathcal{A}_1$, $\mathcal{A}_2$ and $\mathcal{A}_3$. The events $\mathcal{A}_2$ and $\mathcal{A}_3$ are the two possible sub-events of the complement of $\mathcal{A}_1$. Let J^* be the smallest integer such that $\boldsymbol{m}^*_{J^*}$ is a boundary point of $\mathcal{M}_{J^*}$.

1. In the event $\mathcal{A}_1$, $J_N \leq \mathcal{J}$. This implies that $\hat{\boldsymbol{m}}_{J_N}$ is an interior point of $\mathcal{M}_{\mathcal{J}}$.
2. In the event $\mathcal{A}_2$, $\mathcal{J} < \min\{J^*, J_N\}$. This implies that $\hat{\boldsymbol{m}}_{\mathcal{J}}$ is not an interior point of $\mathcal{M}_{\mathcal{J}}$ but $\boldsymbol{m}^*_{\mathcal{J}}$ is.
3. In the event $\mathcal{A}_3$, $J^* \leq \mathcal{J} < J_N$. This implies that neither $\hat{\boldsymbol{m}}_{\mathcal{J}}$ nor $\boldsymbol{m}^*_{\mathcal{J}}$ is an interior point of $\mathcal{M}_{\mathcal{J}}$.

With the following lemmas and

$$\mathbb{E}\left[(\tau - \hat{\tau}_{\mathrm{GMM}})^2\right] = \sum_{i=1}^{3} \mathrm{pr}(\mathcal{A}_i)\mathbb{E}\left[(\tau - \hat{\tau}_{\mathrm{GMM}})^2 \mid \mathcal{A}_i\right],$$

we have Theorem 4.5.1, where for each i, $\mathrm{pr}(\mathcal{A}_i)$ is the probability that the event $\mathcal{A}_i$ happens.

Lemma 4.5.1.

Under the conditions of Theorem 4.5.1, for each positive integer r, it is true that

$$\mathrm{pr}(\mathcal{A}_1)\mathbb{E}\left[(\tau - \hat{\tau}_{\mathrm{GMM}})^2 \mid \mathcal{A}_1\right] = O(\max\{J_N^2 N^{-1}, J_N^{-2r}\}).$$

Proof. See the Appendix. □

Lemma 4.5.2.
Under the conditions of Theorem 4.5.1, it is true that

$$\mathrm{pr}(\mathcal{A}_2)\mathbb{E}\left[(\tau - \hat{\tau}_{\mathrm{GMM}})^2 \mid \mathcal{A}_2\right] = O(J_N^2 N^{-1}).$$

Proof. See the Appendix. □

Lemma 4.5.3.
Under the conditions of Theorem 4.5.1, it is true that

$$\mathrm{pr}(\mathcal{A}_3)\mathbb{E}\left[(\tau - \hat{\tau}_{\mathrm{GMM}})^2 \mid \mathcal{A}_3\right] = O(J_N^2 N^{-1}).$$

Proof. See the Appendix. □

From the proof of Lemma 4.5.1, we observe the trade-off between bias and variance as J_N varies. In the event $\mathcal{A}_1$, we have that the variance of $\hat{\tau}_{\mathrm{GMM}}$ is $O(J_N^2 N^{-1})$ and the squared bias is $O(J_N^{-2r})$; see the proof of Lemma 4.5.1 in the Appendix. With the increase of J_N, the variance of $\hat{\tau}_{\mathrm{GMM}}$ in $\mathcal{A}_1$ increases and the bias of $\hat{\tau}_{\mathrm{GMM}}$ decreases. Furthermore, in additional to Lemma 4.5.2 and 4.5.3, the convergence rate of the MSE of $\hat{\tau}_{\mathrm{GMM}}$ is minimized by $J_N = N^{1/(2r+2)}$.

4.6 Robustness Property

In this section, we study the robustness property of the GMM estimators to outliers. The influence function is a common tool to measure the robustness; see [Hampel, 1974]. However, the GMM estimators with constraints do not have explicit influence functions. Instead, we consider the robustness of the gradient functions in Section 3.4.

Given a random sample $X_1, \ldots, X_{N_1}$ from $f_{\mathrm{Mix}}(x; Q^*)$ and a fixed J, the GMM estimator is determined by the gradient function

$$\mathcal{D}(\boldsymbol{m}, \theta) = \left(\boldsymbol{\Lambda}^{1/2}\boldsymbol{\gamma}(\theta) - \boldsymbol{m}\right)^{\mathrm{T}} \boldsymbol{W}_{(J)} \left(\boldsymbol{m} - \bar{\boldsymbol{m}}\right),$$

by Theorem 3.4.2. If this random sample is further contaminated by N_2 random variables from Δ_z, where Δ_z is the probability measure with mass 1 at the single contaminated data point z, the gradient function for the new data is

$$\tilde{\mathcal{D}}(\boldsymbol{m}, \theta) = \left(\boldsymbol{\Lambda}^{1/2}\boldsymbol{\gamma}(\theta) - \boldsymbol{m}\right)^{\mathrm{T}} \boldsymbol{W}_{(J)} \left(\boldsymbol{m} - \tilde{\boldsymbol{m}}\right),$$

where $\tilde{\boldsymbol{m}} = (1-\alpha)\bar{\boldsymbol{m}} + \alpha\boldsymbol{\phi}_{f_0^{-1/2}}(z)$ and $\alpha = N_2/(N_1+N_2)$. Because the elements of $\boldsymbol{\phi}_{f_0^{-1/2}}(z)$ may not be bounded as $f_0(z)$ goes to zero, the gradient function is not robust to the outliers when $\boldsymbol{W}_{(J)}$ is the identity matrix. If we choose a weighting matrix $\boldsymbol{W}_{(J)}$ such that $\boldsymbol{W}_{(J)}\boldsymbol{\phi}_{f_0^{-1/2}}(z)$ converges to a constant vector as $f_0(z)$ goes to zero, we may achieve the robustness to the outliers.

For a fixed J, let $\boldsymbol{L}_{(J)}(\theta) = (L_1(\theta), \ldots, L_J(\theta))^{\mathrm{T}} \in \mathbb{R}^J$, where for each $j = 1, \ldots, J$,

$$L_j(\theta) = \int_{\mathcal{S}} f(y;\theta)\phi_j(y)\mathrm{d}y.$$

Note that $\{\phi_j(x)f_0^{-1/2}(x)\}_{j=0}^{\infty}$ is a complete orthonormal basis in $L^2(\mathcal{S}, \nu_0)$ and $\phi_0(x) = f_0^{1/2}(x)$; see Subsection 2.3.2. For each $\theta \in \Theta$, $f(x;\theta)f_0^{-1/2}(x) \in L^2(\mathcal{S}, \nu_0)$. Therefore, $L_j(\theta)$ is finite for each $j \in \{1, \ldots, J\}$ and $\theta \in [a, b]$. We also have the expansion in $L^2(\mathcal{S}, \nu_0)$

$$f(x;\theta)f_0^{-1/2}(x) = \int_{\mathcal{S}} f(y;\theta)f_0^{1/2}(y)\mathrm{d}y + \boldsymbol{L}_{(\infty)}^{\mathrm{T}}(\theta)\boldsymbol{\phi}_{f_0^{-1/2}}(x), \tag{4.7}$$

where $\boldsymbol{\phi}_{f_0^{-1/2}}(x) = (\phi_1(x)f_0^{-1/2}(x), \phi_2(x)f_0^{-1/2}(x), \ldots)^{\mathrm{T}} \in \mathbb{R}^{\infty}$. We call the matrix

$$\boldsymbol{W}_{(J)}^{\mathrm{Robust}} = \int_{\Theta} \boldsymbol{L}_{(J)}(\theta)\boldsymbol{L}_{(J)}^{\mathrm{T}}(\theta)\mathrm{d}\theta \tag{4.8}$$

the robust weighting matrix.

Theorem 4.6.1.
As $f_0(z)$ goes to 0, the vector

$$\boldsymbol{W}_{(\infty)}^{\mathrm{Robust}}\boldsymbol{\phi}_{f_0^{-1/2}}(z) = -\int_{\Theta} \boldsymbol{L}(\theta) \int_{\mathcal{S}} f(y;\theta)f_0^{1/2}(y)\mathrm{d}y\mathrm{d}\theta,$$

which is a constant vector in $\mathbb{R}^{\infty}$.

To illustrate the role of the robust weighting matrix plays, we consider the example of a mixture of Poisson distributions.

Example 4.1.
The basis functions $\{\phi_j(x)\}_{j=0}^{\infty}$ *for the mixture of Poisson have been illustrated in Example 2.1. Figure 4.1 shows the functions* $\{\phi_j(x)f_0^{-1/2}(x)\}_{j=1}^{4}$ *and the first four elements of* $\boldsymbol{W}_{(J)}^{\text{Robust}}\boldsymbol{\phi}_{f_0^{-1/2}}(x)$*, when* $J = 18$*. As we can see, for each* j*,* $\phi_j(x)f_0^{-1/2}(x)$ *goes to a large number as* $f_0(x)$ *goes to zero. And thus, a contaminated data* z *with small* $f_0(z)$ *is able to change* $\bar{\boldsymbol{m}}$ *significantly. On the other hand,* $\boldsymbol{W}_{(J)}^{\text{Robust}}\boldsymbol{\phi}_{f_0^{-1/2}}(x)$ *converges to a constant vector as* $f_0(x)$ *goes to zero. This controls the effects of a contaminated data* z *on* $\bar{\boldsymbol{m}}$*. And thus, the robustness to the outliers is expected.*

4.7 Computational Algorithms

Because the optimization framework of the GMM matches with the geometry discussed in Chapter 3, we have the gradient characterization of the $\hat{\boldsymbol{m}}_J$, where J is a fixed integer. This allows us to adopt the gradient-based algorithms to compute the GMM estimators. Existing gradient-based computational algorithms include the vertex directional method, vertex exchange method and intra simplex direction method; see [Böhning et al., 1992, Böhning, 1995] for a review. Wang [2007] proposed the constrained Newton method with multiple exchange vertices (CNM) algorithm. Empirically, Wang's algorithm is the fastest and most accurate comparing to the other algorithms. Therefore, we modify Wang's algorithm for the GMM estimator.

Algorithm 4.1 (The CNM for GMM)**.**
Set $s = 0$ *and fix* J*. From an initial estimate* $Q^{(0)}$ *with finite support* $\Theta^{(0)}$ *and* $\boldsymbol{m}^{(0)} = \boldsymbol{m}(Q^{(0)}) \neq \bar{\boldsymbol{m}}$*, repeat the following steps:*

1. *Compute all the local minimas* $\{\theta_j^{(s)}\}_{j=1}^{r^{(s)}}$ *of the function*

$$\mathcal{D}(\theta) = (\boldsymbol{\Lambda}^{1/2}\boldsymbol{\gamma}(\theta) - \boldsymbol{m}^{(s)})^{\mathrm{T}}\boldsymbol{W}_{(J)}(\boldsymbol{m}^{(s)} - \bar{\boldsymbol{m}})$$

   over $[a, b]$*. The iteration stops if the minimum of* $\mathcal{D}(\theta)$ *is zero.*

2. *Construct a set of candidate support points by*

$$\Theta^{(s),+} = \Theta^{(s)} \cup \{\theta_j^{(s)}\}_{j=1}^{r^{(s)}}.$$

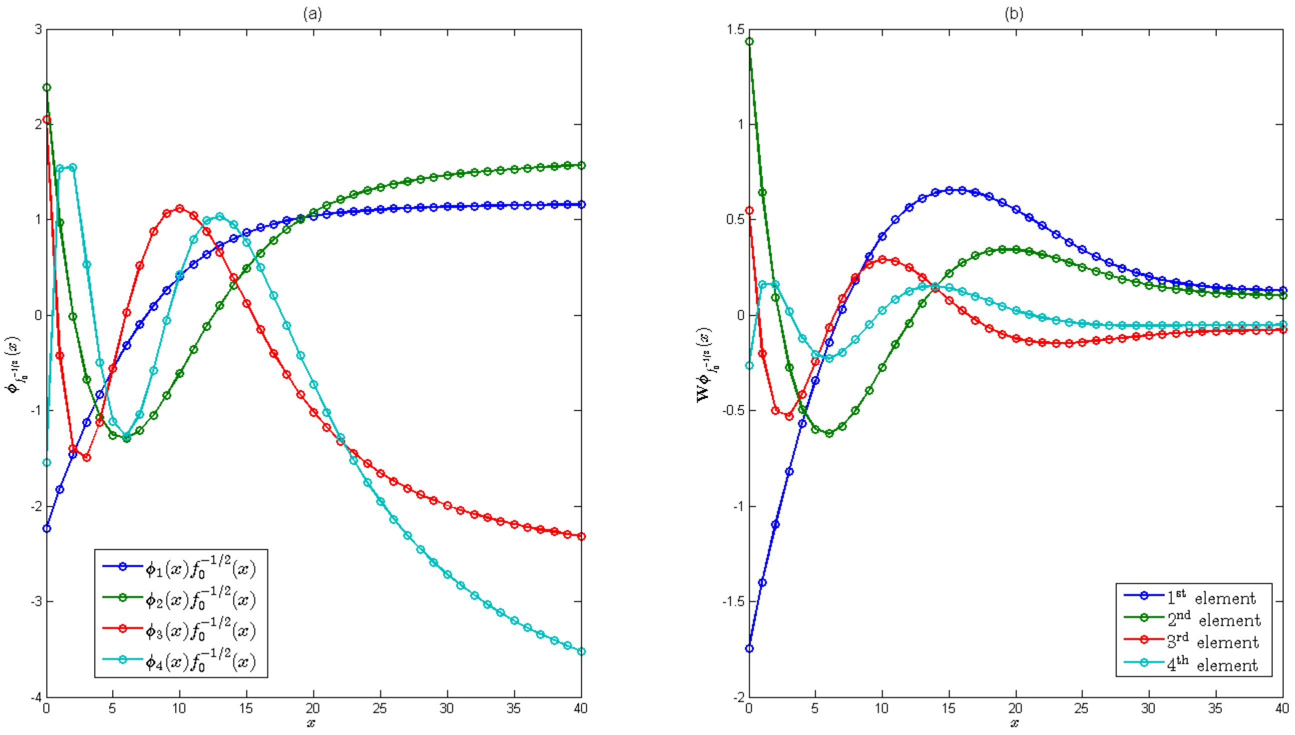

Figure 4.1: Plots of (a) the functions $\phi_j(x)f_0^{-1/2}(x)$, $j = 1, 2, 3, 4$ and (b) the first four elements of $\boldsymbol{W}_{(J)}^{\text{Robust}}\boldsymbol{\phi}_{f_0^{-1/2}}(x)$ with $J = 18$.

Let $r^{(s),+}$ be the number of elements in $\Theta^{(s),+}$.

3. *Solve the optimization problem*

$$\begin{aligned} \min \quad & \left(\sum_{i=1}^{r^{(s),+}} \alpha_i \boldsymbol{\Lambda}^{1/2} \boldsymbol{\gamma}(\theta_i) - \bar{\boldsymbol{m}} \right)^{\mathrm{T}} \boldsymbol{W}_{(J)} \left(\sum_{i=1}^{r^{(s),+}} \alpha_i \boldsymbol{\Lambda}^{1/2} \boldsymbol{\gamma}(\theta_i) - \bar{\boldsymbol{m}} \right) \\ s.t. \quad & \sum_{i=1}^{r^{(s),+}} \alpha_i = 1, \\ & \alpha_i \geq 0, \quad i = 1, \ldots, r^{(s),+}, \end{aligned}$$

where $\theta_i \in \Theta^{(s),+}$. We denote its solution by $\boldsymbol{\alpha}^{(s)} = (\alpha_1^{(s)}, \ldots, \alpha_{r^{(s),+}}^{(s)})^{\mathrm{T}}$. If the minimum is zero, stop the interation and return $\bar{\boldsymbol{m}} \in \mathbb{R}^J$.

4. *Discard all θ_is with zero $\alpha_i^{(s)}$, update $Q^{(s)}$, $\Theta^{(s)}$ and $\boldsymbol{m}^{(s)} = \boldsymbol{m}(Q^{(s)})$, and set $s = s + 1$.*

The convergence of the algorithm is shown in [Wang, 2007]. Further note that the optimization problem in Step 3 at each iteration is a constrained quadratic programming problem. Computational algorithms for the quadratic programming problem can be found in [Antoniou and Lu, 2007]. The MATLAB code for Algorithm 4.1 can be seen in the Appendix.

4.8 Simulation Studies

In this section, we study the performance of the GMM estimators through simulations. Four mixtures of Poisson distributions with different types of mixing distributions (listed as follows) are considered; see Figure 4.2 for the shapes of the considered models.

1. Let $Q_1^*(\theta)$ be the uniform distribution of θ defined on $[7, 13] \subset \mathbb{R}$. This is an example when the mixing distribution is continuous.

2. Let $Q_2^*(\theta) = 0.5I(\theta \leq 3) + 0.5I(\theta \leq 9)$. This is an example when a finite mixture model is regular in the sense that the elements of $\boldsymbol{\alpha}$ is away from 0 and the component distributions are different from each other.

3. Let $Q_3^*(\theta) = 0.5I(\theta \leq 4.9) + 0.5I(\theta \leq 5.1)$. Here Pois(4.9) and Pois(5.1) are closely linearly dependent. This is an example when mixing distribution is defined locally at 5; see [Marriott, 2002].

4. Let $Q_4^*(\theta) = 0.99I(\theta \leq 3) + 0.01I(\theta \leq 9)$. We consider 0.01 is a reasonable small positive number. This is an example of the contamination mixing; see [Tukey, 1960].

We compare the performance of the GMM estimators and the NPMLE. In the GMM, we set $\Theta = [0, 20]$ and the weighting matrix is $\boldsymbol{W}_{(J)}^{\text{Robust}}$ in Equation (4.8), where $J+1$ is the number of the used generalized moment conditions. The considered sample size levels are 20, 50, 100 and 200. The number of repetition is 1000 in each simulation.

We are interested in the point-wise MSE of the fitted mixture models: $\hat{f}_{\text{GMM}}(x)$ and $f_{\text{Mix}}(x; \hat{Q}_{\text{NPMLE}})$, where $\hat{Q}_{\text{NPMLE}}$ is the NPMLE for mixture models. In each repetition, there exists a finite $\mathcal{J}$ such that $\hat{\boldsymbol{m}}_{\mathcal{J}}$ is not an interior point of $\mathcal{M}_{\mathcal{J}}$. In other words, $\hat{f}_{\text{GMM}}(x) = f_{\text{Mix}}(x; \hat{Q}_{\text{GMM},\mathcal{J}})$. In Figure 4.3 to 4.6, we present the point-wise MSE of each fitted model. As we can see, $f_{\text{Mix}}(x; \hat{Q}_{\text{NPMLE}})$ has the smaller point-wise MSE over $x \in \mathbb{R}$ in general. However, the GMM estimator performs nearly as well as the NPMLE.

In Table 4.1 and 4.2, we give the empirical cumulative distribution function of $\mathcal{J}$ in each case and each sample size level. The J^* is ∞ in the case where Q_1^* is the true mixing distribution, and 4 in the other three cases. With the increase of the sample size, the empirical probability that $\mathcal{J} < J^*$ non-increases; see Table 4.1 and 4.2. According to Table 4.1, the expectation of $\mathcal{J}$ increases with the increase of the sample size, when Q_1^* is the true mixing distribution. This because that Q_1^* is a continuous function and $\boldsymbol{m}_J^*$ is always an interior point of $\mathcal{M}_J$ for any J. The observations from the tables imply that $\text{pr}(\mathcal{J} \leq J^*)$ decreases to zero as the sample size N goes to infinity; see Lemma 4.5.1. In Table 4.2, $\text{pr}(\mathcal{J} < J^*)$ remains large

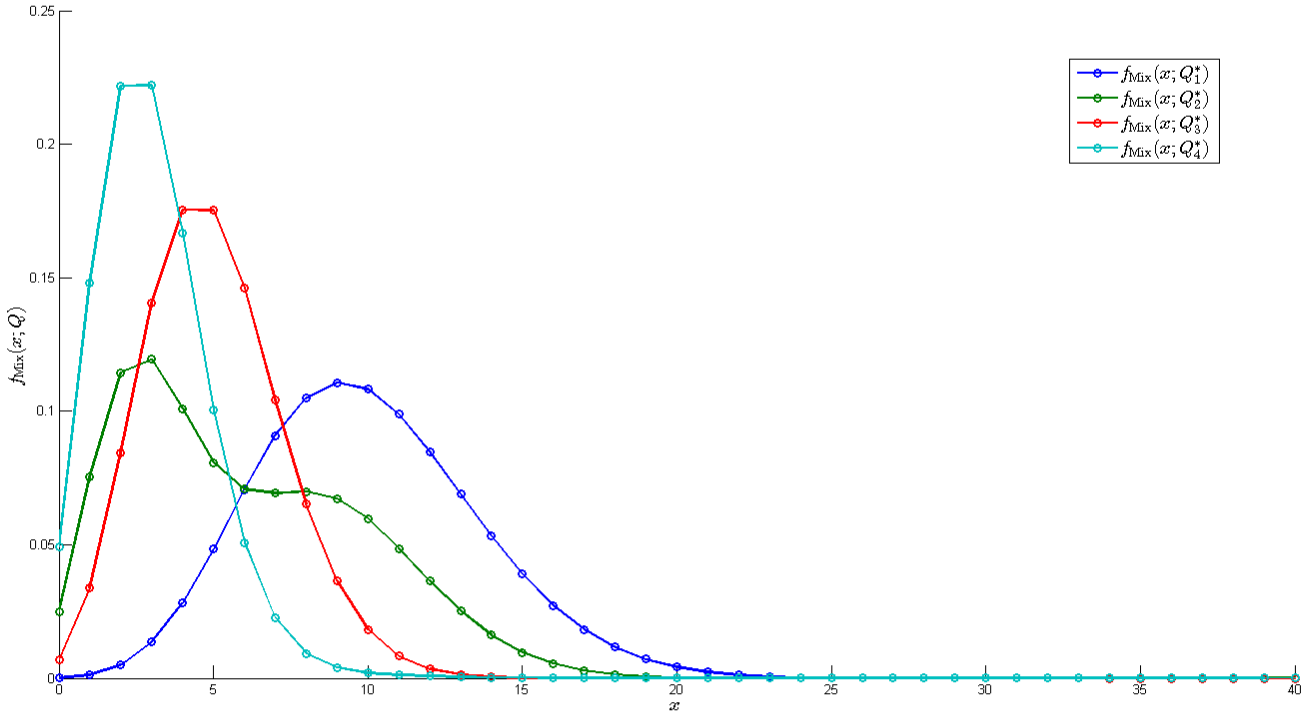

Figure 4.2: Plots of the mixtures of Poisson, $f_{\mathrm{Mix}}(x; Q_i^*)$, for $i = 1, 2, 3, 4$.

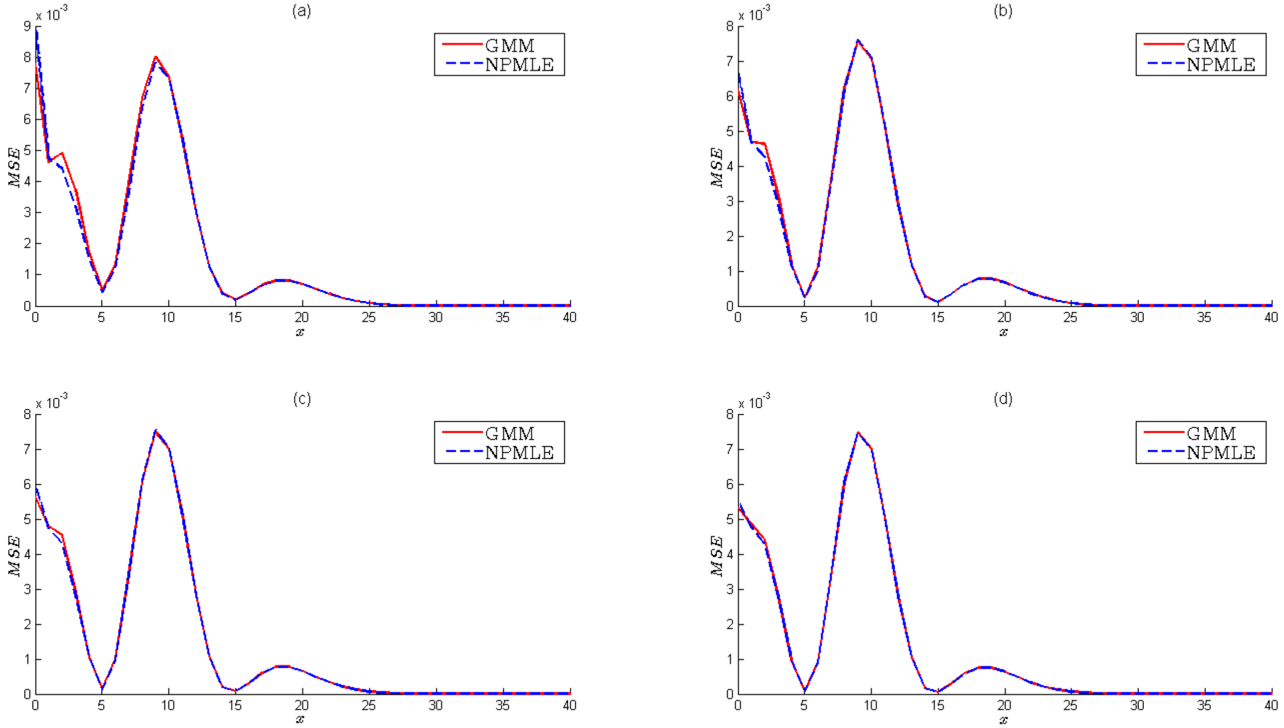

Figure 4.3: Plots of the point-wise MSE of $\hat{f}_{\mathrm{GMM}}(x)$ and $f_{\mathrm{Mix}}(x; \hat{Q}_{\mathrm{NPMLE}})$ when the random sample is from $f_{\mathrm{Mix}}(x; Q_1^*)$ and (a) $N = 20$; (b) $N = 50$; (c) $N = 100$ and (d) $N = 200$.

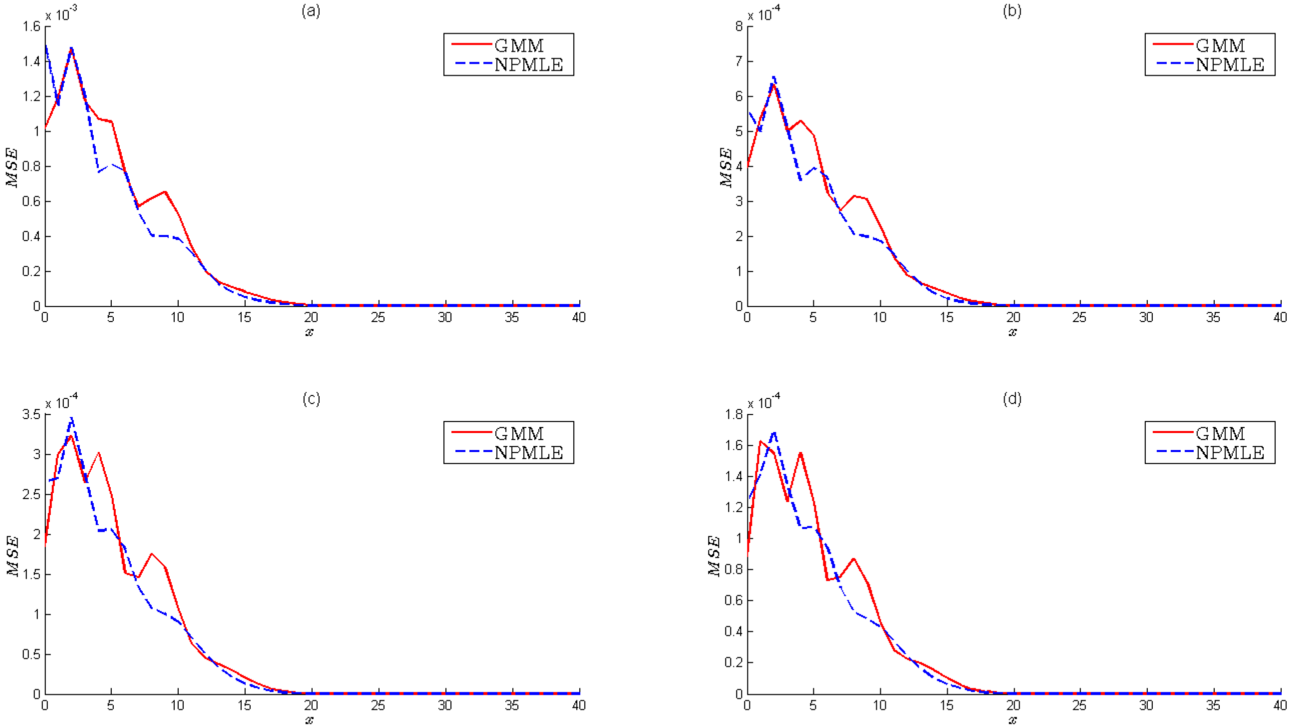

Figure 4.4: Plots of the point-wise MSE of $\hat{f}_{\text{GMM}}(x)$ and $f_{\text{Mix}}(x;\hat{Q}_{\text{NPMLE}})$ when the random sample is from $f_{\text{Mix}}(x;Q_2^*)$ and (a) $N=20$; (b) $N=50$; (c) $N=100$ and (d) $N=200$.

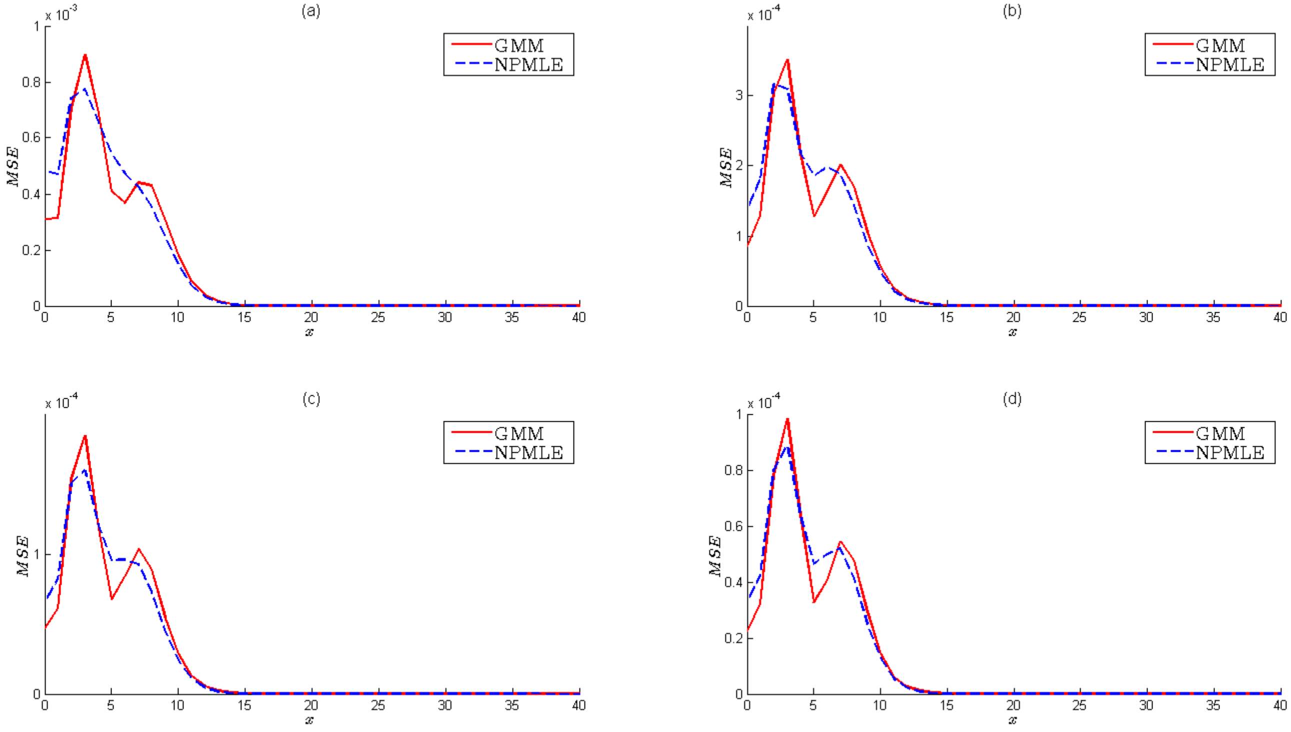

Figure 4.5: Plots of the point-wise MSE of $\hat{f}_{\text{GMM}}(x)$ and $f_{\text{Mix}}(x; \hat{Q}_{\text{NPMLE}})$ when the random sample is from $f_{\text{Mix}}(x; Q_3^*)$ and (a) $N = 20$; (b) $N = 50$; (c) $N = 100$ and (d) $N = 200$.

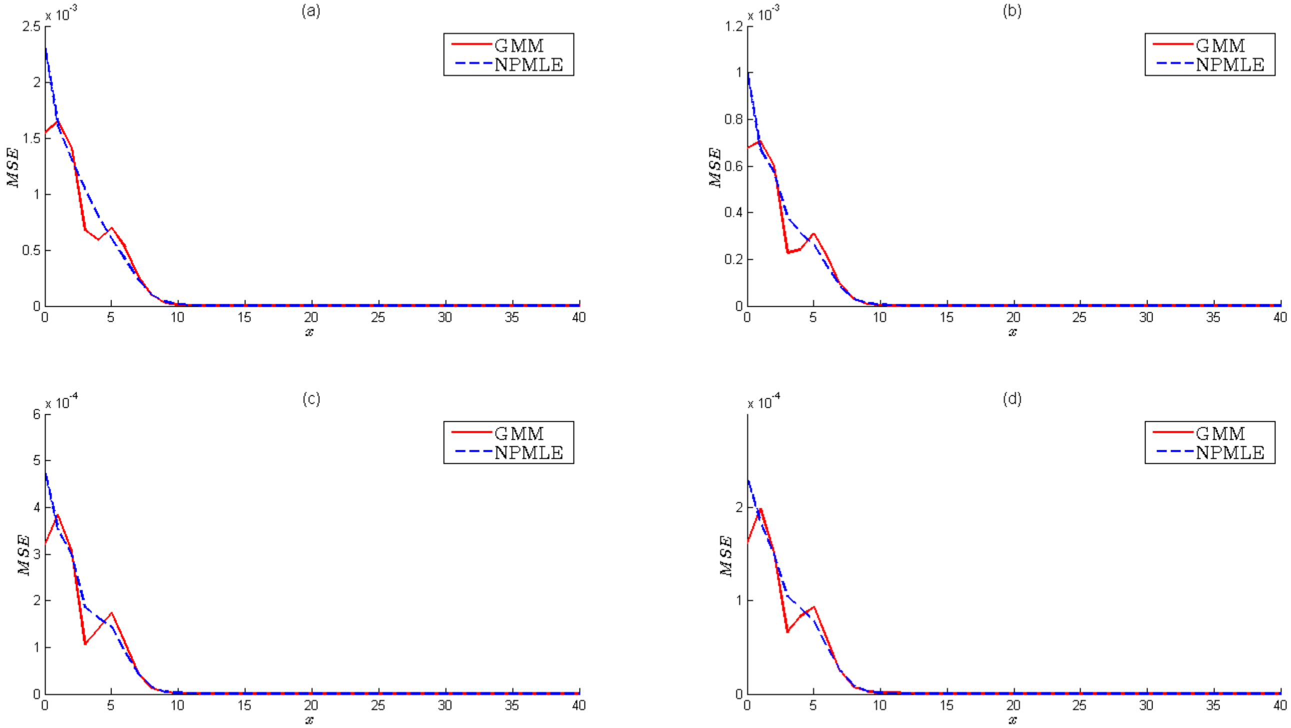

Figure 4.6: Plots of the point-wise MSE of $\hat{f}_{\mathrm{GMM}}(x)$ and $f_{\mathrm{Mix}}(x;\hat{Q}_{\mathrm{NPMLE}})$ when the random sample is from $f_{\mathrm{Mix}}(x;Q_4^*)$ and (a) $N=20$; (b) $N=50$; (c) $N=100$ and (d) $N=200$.

when $N = 200$. This is because that the local (Q_3^*) and contamination (Q_4^*) mixtures are close to a single-component Poisson distribution, and larger sample size is needed to reduce $\text{pr}(\mathcal{J} < J^*)$.

Next, we study the robustness of the GMM estimator. Let 5% of the data are from the degenerate distribution Δ_z, where $z = 40$. And the rests are from the true models $f_{\text{Mix}}(x; Q_i^*)$, for each $i \in \{1, \ldots, 4\}$. We fix the number of generalized moment conditions to 19, i.e., $J = 18$, and use the associated robust weighting matrix $\boldsymbol{W}_{(J)}^{\text{Robust}}$. Figure 4.7 to 4.10 show the point-wise MSE of the fitted models in each case. We see that $f(x; \hat{Q}_{\text{GMM},J})$ has the smaller point-wise MSE over $x \in \mathbb{R}$ in general.

4.9 Application to the Thailand Cohort Study Data

Consider the data on morbidity in northeast Thailand which has been described in Subsection 1.6.1. We fit a mixture of Poisson with $[a, b] = [0, 25]$ using the GMM. The number J are taken to be $1, \ldots, 18$ and the weighting matrix $\boldsymbol{W} = \boldsymbol{W}_{(J)}^{\text{Robust}}$ is associated with J. When $J \leq 7$, there is no unique GMM estimator for the mixing distribution. When $J \geq 8$, the results are summarized in Table 4.3. The fitted models with different Js are close, when $J \geq 8$. This implies that little information is contained in the higher order generalized moment conditions, when the robust weighting matrix is used. In Figure 4.11, we see that the fitted mixture model with ($J = 8$) successfully characterize the shape of the histogram.

4.10 Conclusion and Discussion

In this chapter, we have introduced the GMM estimators for mixture models and studied the asymptotic behavior of the MSE and the robustness property to the outliers. We can see that the GMM is a promising estimation method for non-parametric mixture models. In this section, we point out two possible future research directions in the GMM for mixture models.

The weighting matrix $\boldsymbol{W}$ plays an important role in determining the properties

	Q_1^*				Q_2^*			
$\mathcal{J}$	$N = 20$	$N = 50$	$N = 100$	$N = 200$	$N = 20$	$N = 50$	$N = 100$	$N = 200$
2	0.0000	0.0000	0.0000	0.0000	0.0030	0.0000	0.0000	0.0000
3	0.0060	0.0000	0.0000	0.0000	0.0180	0.0000	0.0000	0.0000
4	0.1520	0.0160	0.0000	0.0000	0.7180	0.6110	0.5780	0.5720
5	0.4720	0.1060	0.0180	0.0000	0.8940	0.8420	0.8240	0.8070
6	0.8010	0.4830	0.2210	0.0500	0.9700	0.9630	0.9570	0.9430
7	0.9750	0.8510	0.6460	0.4040	0.9980	0.9940	0.9940	0.9890
8	1.0000	0.9800	0.9250	0.8120	1.0000	1.0000	1.0000	1.0000
9	1.0000	1.0000	0.9960	0.9730	1.0000	1.0000	1.0000	1.0000
10	1.0000	1.0000	1.0000	0.9990	1.0000	1.0000	1.0000	1.0000
11	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000

Table 4.1: Empirical C.D.F of $\mathcal{J}$ in each simulation case when the true mixing distribution is Q_1^* or Q_2^* .

	Q_3^*				Q_4^*			
$\mathcal{J}$	$N = 20$	$N = 50$	$N = 100$	$N = 200$	$N = 20$	$N = 50$	$N = 100$	$N = 200$
2	0.5810	0.5710	0.5280	0.5320	0.5120	0.4350	0.3660	0.3250
3	0.7690	0.7400	0.7260	0.7160	0.7190	0.6750	0.5970	0.5150
4	0.9650	0.9510	0.9470	0.9040	0.9360	0.8730	0.8570	0.7940
5	0.9940	0.9920	0.9910	0.9820	0.9930	0.9660	0.9630	0.9400
6	0.9990	1.0000	0.9990	0.9980	1.0000	0.9980	0.9970	0.9950
7	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000

Table 4.2: Empirical C.D.F of $\mathcal{J}$ in each simulation case when the true mixing distribution is Q_3^* or Q_4^* .

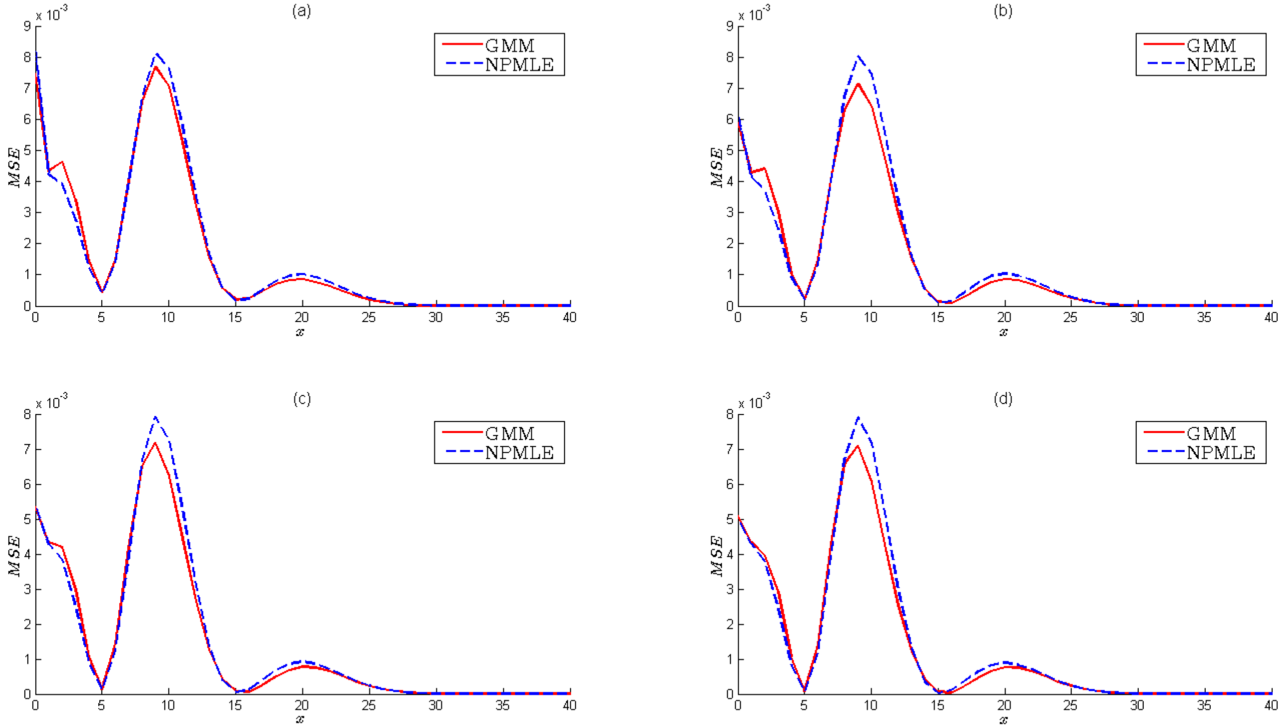

Figure 4.7: Plots of the point-wise MSE of $\hat{f}_{\text{GMM}}(x)$ and $f_{\text{Mix}}(x;\hat{Q}_{\text{NPMLE}})$ when the random sample is from $95\% f_{\text{Mix}}(x;Q_1^*) + 5\%\Delta_{40}$ and (a) $N = 20$; (b) $N = 50$; (c) $N = 100$ and (d) $N = 200$.

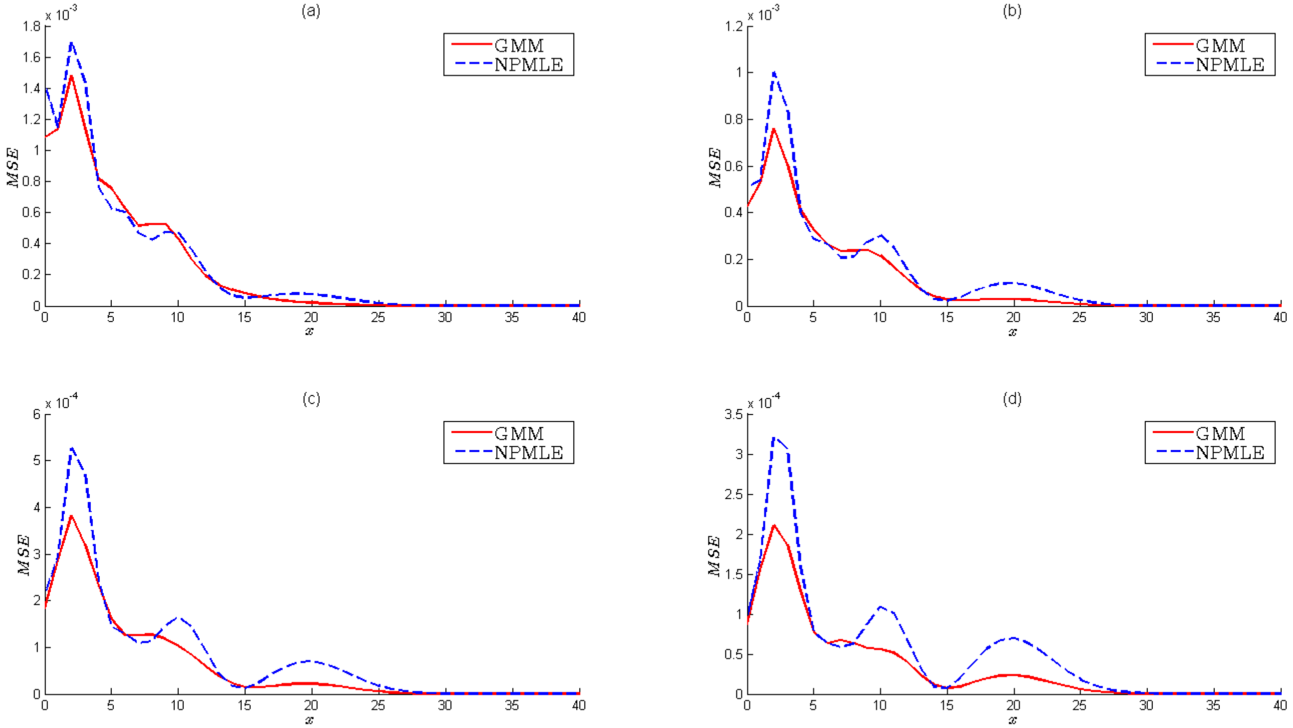

Figure 4.8: Plots of the point-wise MSE of $\hat{f}_{\text{GMM}}(x)$ and $f_{\text{Mix}}(x; \hat{Q}_{\text{NPMLE}})$ when the random sample is from $95\% f_{\text{Mix}}(x; Q_2^*) + 5\% \Delta_{40}$ and (a) $N = 20$; (b) $N = 50$; (c) $N = 100$ and (d) $N = 200$.

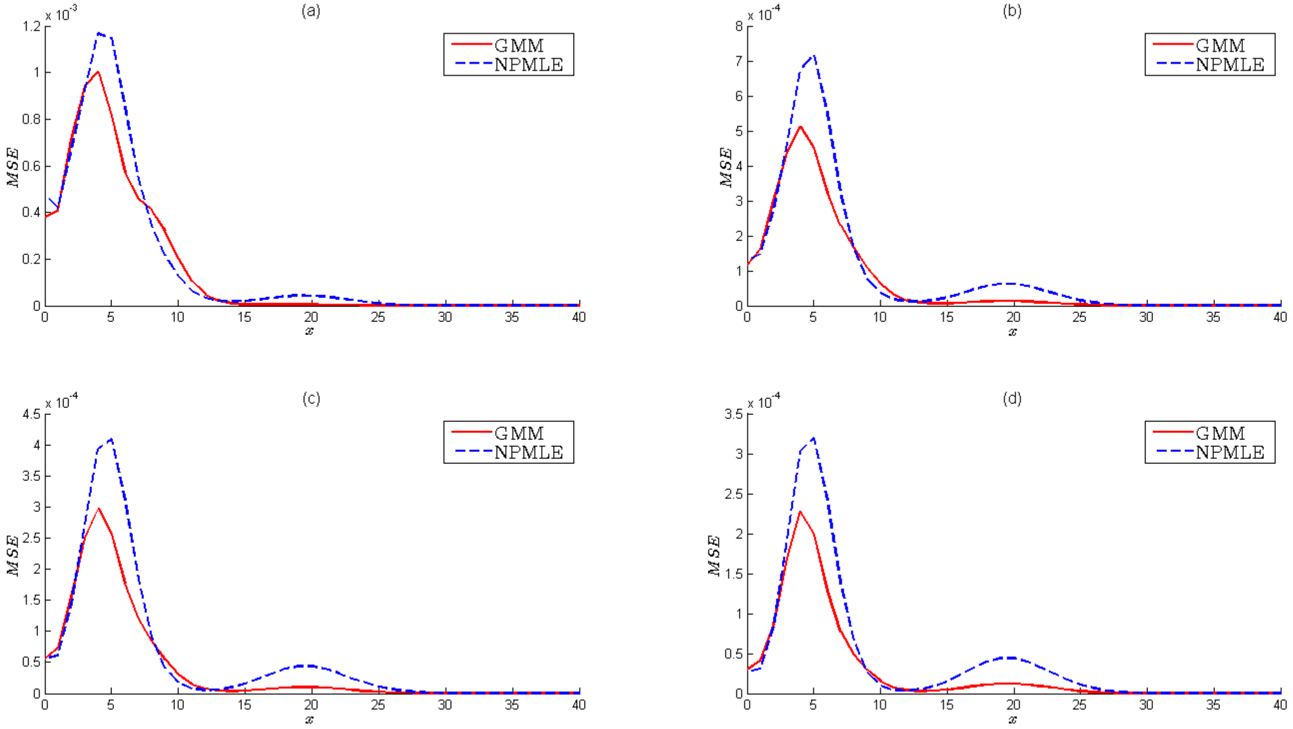

Figure 4.9: Plots of the point-wise MSE of $\hat{f}_{\text{GMM}}(x)$ and $f_{\text{Mix}}(x; \hat{Q}_{\text{NPMLE}})$ when the random sample is from $95\% f_{\text{Mix}}(x; Q_3^*) + 5\% \Delta_{40}$ and (a) $N = 20$; (b) $N = 50$; (c) $N = 100$ and (d) $N = 200$.

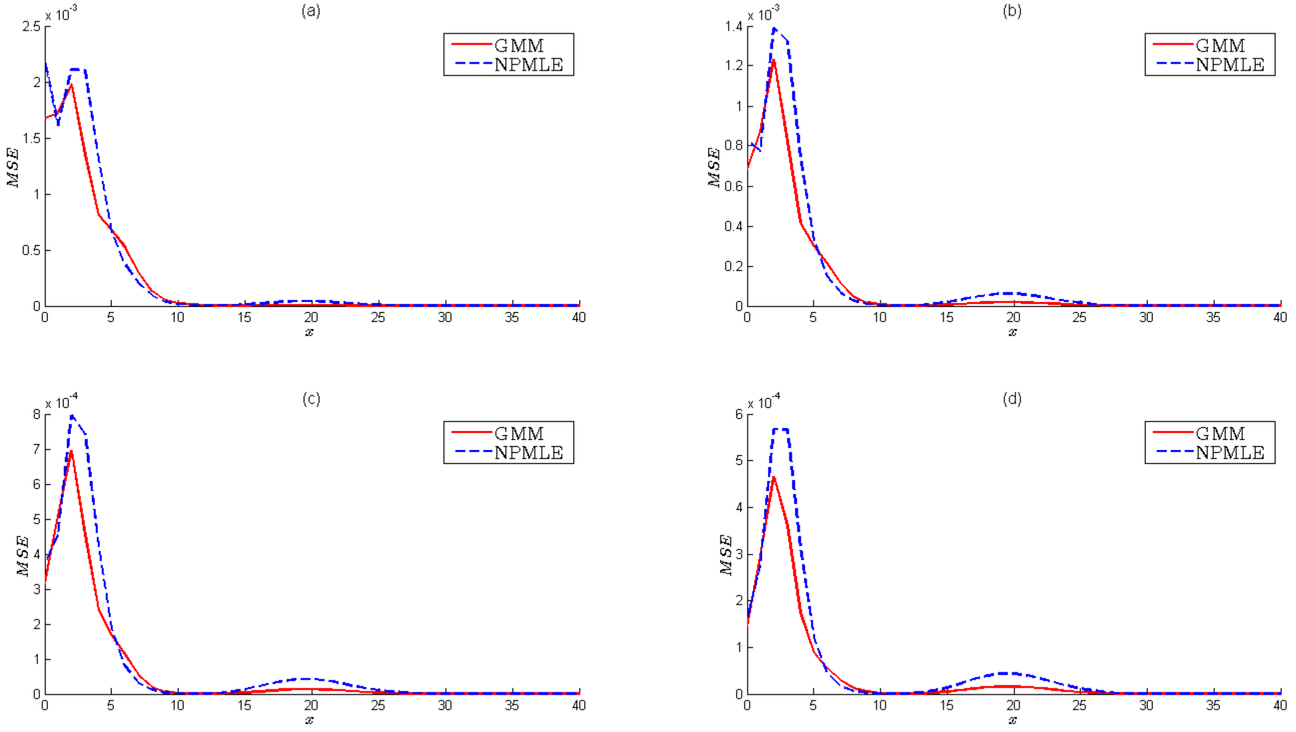

Figure 4.10: Plots of the point-wise MSE of $\hat{f}_{\text{GMM}}(x)$ and $f_{\text{Mix}}(x;\hat{Q}_{\text{NPMLE}})$ when the random sample is from $95\% f_{\text{Mix}}(x;Q_4^*) + 5\%\Delta_{40}$ and (a) $N = 20$; (b) $N = 50$; (c) $N = 100$ and (d) $N = 200$.

J	Mixing parameters	Mixing proportions	Log-likelihood (10^3)
8	(0.1375, 2.9050, 8.2425, 16.3075)	(0.1962, 0.4856, 0.2681, 0.0502)	-1.5538
9	(0.1425, 2.8375, 8.2775, 16.5100)	(0.1968, 0.4852, 0.2676, 0.0502)	-1.5538
10	(0.1425, 2.8550, 8.2775, 16.5025)	(0.1970, 0.4853, 0.2676, 0.0501)	-1.5538
11	(0.1425, 2.8450, 8.2825, 16.5200)	(0.1970, 0.4854, 0.2676, 0.0500)	-1.5538
12	(0.1375, 2.8400, 8.2750, 16.4925)	(0.1967, 0.4851, 0.2677, 0.0505)	-1.5538
13	(0.1425, 2.8300, 8.2775, 16.5075)	(0.1970, 0.4853, 0.2676, 0.0501)	-1.5538
14	(0.1425, 2.8300, 8.2775, 16.5075)	(0.1970, 0.4853, 0.2676, 0.0501)	-1.5538
15	(0.1425, 2.8300, 8.2775, 16.5075)	(0.1970, 0.4853, 0.2676, 0.0501)	-1.5538
16	(0.1425, 2.8550, 8.2775, 16.5075)	(0.1970, 0.4853, 0.2676, 0.0501)	-1.5538

Table 4.3: Some results of the fitted models to the Thailand cohort study data.

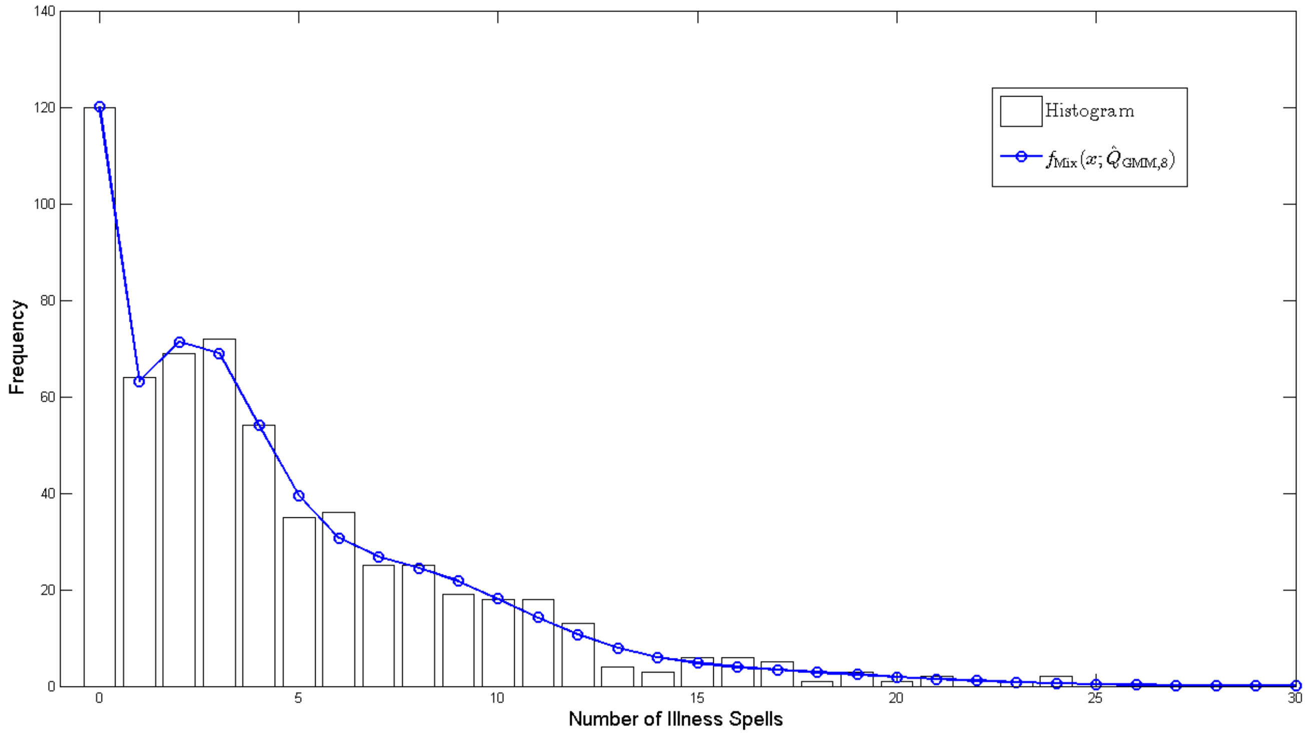

Figure 4.11: Plots of the observed frequency and the frequency of $f_{Mix}(x;\hat{Q}_{\text{GMM},8})$ and $\boldsymbol{W} = \boldsymbol{W}_{(8)}^{\text{Robust}}$ in the Thailand cohort study.

of the GMM estimators. One example is the robust weighting matrix given in Section 4.6. Naturally, we are also interested in some weighting matrices for the GMM estimators with less robustness but more efficiency. Suppose that we have one robust weighting matrix $\boldsymbol{W}_{\text{Robust}}$ and one efficient weighting matrix $\boldsymbol{W}_{\text{Efficient}}$. It is possible to balance the robustness and the efficiency of the GMM estimators by using a convex combination of the two weighted matrices, i.e.,

$$\boldsymbol{W} = (1-\alpha)\boldsymbol{W}_{\text{Robust}} + \alpha \boldsymbol{W}_{\text{Efficient}},$$

where $\alpha \in [0,1]$.

Due to the existence of the constraints on the feasible set, it is challenging to obtain the asymptotic distributions of $\hat{\tau}_{\text{GMM}}$ defined in Equation (4.6). And thus, it is challenging to construct interval estimators for the GMM estimator of τ. Some previous researches have shown the existence of the asymptotic normality in the NPMLE; see [Lambert and Tierney, 1984], [Van De Geer, 1997] and [Böhning and Patilea, 2005]. Because the similar geometric structures between the GMM estimators and the NPMLE, the previous results on the NPMLE direct possible paths to find possible asymptotic normality in the GMM estimators.

Appendix: C

C.1 Proof of Lemma 4.5.1

Proof. In the event $\mathcal{A}_1$, the mixture model is fitted by $f_{\text{spec}}(x; \bar{\boldsymbol{m}})$. For each $x \in \mathcal{S}$, the mean square error equals the sum of the variance and the squared bias of the estimator, i.e., for each $x \in \mathcal{S}$,

$$\begin{aligned}
&\operatorname{pr}(\mathcal{A}_1)\mathbb{E}\left[(\tau - \hat{\tau}_{\mathrm{GMM}})^2 \mid \mathcal{A}_1\right] \\
\leq &\sum_{i=1}^{3} \operatorname{pr}(\mathcal{A}_i)\mathbb{E}\left[\left(\int_{\mathcal{S}} s(x)\left(f_{\mathrm{Mix}}(x;Q^*) - f_0(x) - \boldsymbol{\phi}^{\mathrm{T}}_{f_0^{1/2}}(x)\bar{\boldsymbol{m}}\right)\mathrm{d}x\right)^2 \mid \mathcal{A}_i\right] \\
= &\mathbb{E}\left[\left(\int_{\mathcal{S}} s(x)\left(f_{\mathrm{Mix}}(x;Q^*) - f_0(x) - \boldsymbol{\phi}^{\mathrm{T}}_{f_0^{1/2}}(x)\bar{\boldsymbol{m}}\right)\mathrm{d}x\right)^2\right] \\
= &\mathbb{E}\left[\left(\int_{\mathcal{S}} s(x)\boldsymbol{\phi}^{\mathrm{T}}_{f_0^{1/2}}(x)(\boldsymbol{m}^* - \bar{\boldsymbol{m}})\mathrm{d}x\right)^2\right] \\
&+ \left(\int_{\mathcal{S}} s(x)\left(f(x;Q^*) - f_0(x) - \boldsymbol{\phi}^{\mathrm{T}}_{f_0^{1/2}}(x)\boldsymbol{m}^*\right)\mathrm{d}x\right)^2,
\end{aligned}$$

where $\boldsymbol{m}^*$ is the true values of $\boldsymbol{m}$. By the Cauchy-Schwarz inequality, the variance of the estimator is

$$\begin{aligned}
&\mathbb{E}\left[\left(\int_{\mathcal{S}} s(x)\boldsymbol{\phi}^{\mathrm{T}}_{f_0^{1/2}}(x)(\boldsymbol{m}^* - \bar{\boldsymbol{m}})\mathrm{d}x\right)^2\right] \\
= &\mathbb{E}\left[\left(\int_{\mathcal{S}} s(x)\boldsymbol{\phi}^{\mathrm{T}}_{f_0^{1/2}}(x)\mathrm{d}x\,(\bar{\boldsymbol{m}} - \boldsymbol{m}^*)\right)^2\right] \\
\leq &\mathbb{E}\left[\|\bar{\boldsymbol{m}} - \boldsymbol{m}^*\|_2^2 \left\|\int_{\mathcal{S}} s(x)\boldsymbol{\phi}_{f_0^{1/2}}(x)\mathrm{d}x\right\|_2^2\right] \\
= &\frac{1}{N}\sum_{j=1}^{J_N} \operatorname{Var}[\phi_j(X)f_0^{-1/2}(X)] \left\|\int_{\mathcal{S}} s(x)\boldsymbol{\phi}_{f_0^{1/2}}(x)\mathrm{d}x\right\|_2^2.
\end{aligned}$$

By the Cauchy-Schwarz inequality, we further have

$$\begin{aligned}
\left\|\int_{\mathcal{S}} s(x)\boldsymbol{\phi}_{f_0^{1/2}}(x)\mathrm{d}x\right\|_2^2 &= \sum_{j=1}^{J_N}\left(\int_{\mathcal{S}} s(x)\phi_j(x)f_0^{1/2}(x)\mathrm{d}x\right)^2 \\
&\leq \sum_{j=1}^{J_N}\int_{\mathcal{S}} s^2(x)f_0(x)\mathrm{d}x \int_{\mathcal{S}} \phi_j^2(x)\mathrm{d}x \\
&= \sum_{j=1}^{J_N}\int_{\mathcal{S}} s^2(x)f_0(x)\mathrm{d}x.
\end{aligned}$$

Under the assumption that $\int_{\mathcal{S}} s^2(x)f_0(x)\mathrm{d}x$ is bounded, say M, we have that

$$\left\|\int_{\mathcal{S}} s(x)\boldsymbol{\phi}_{f_0^{1/2}}(x)\mathrm{d}x\right\|_2^2 \leq J_N M.$$

Note that

$$f_{\text{Mix}}(x; Q^*) = \int_\Theta f(x;\theta)\mathrm{d}Q^* \leq |\Theta| f_0(x).$$

Therefore, for each j,

$$\text{Var}[\phi_j(X) f_0^{-1/2}(X)] \leq \mathbb{E}[\phi_j^2(X) f_0^{-1}(X)] \leq |\Theta| \int_\Theta \phi_j^2(x)\mathrm{d}x = |\Theta|.$$

Therefore, the variance of the estimator is $O(J_N^2/N) = O(N^{-r/(r+1)})$.

By the Cauchy-Schwarz inequality and Proposition 2.3.1, the squared bias

$$\begin{aligned}&\left(\int_{\mathcal{S}} s(x)\left(f(x;Q^*) - f_0(x) - \boldsymbol{\phi}_{f_0^{1/2}}^{\mathrm{T}}(x)\boldsymbol{m}^*\right)\mathrm{d}x\right)^2\\ &\leq \int_{\mathcal{S}} s^2(x) f_0(x)\mathrm{d}x \int_{\mathcal{S}} \left(f(x;Q^*) - f_0(x) - \boldsymbol{\phi}_{f_0^{1/2}}^{\mathrm{T}}(x)\boldsymbol{m}^*\right)^2 / f_0(x)\mathrm{d}x\\ &= O(J_N^{-2r}).\end{aligned}$$

In sum, the mean square error of $\hat{\tau}_{\text{GMM}}$ in $\mathcal{A}_1$ is $O(\max\{J_N^2 N^{-1}, J_N^{-2r}\})$. □

C.2 Proof of Lemma 4.5.2

Proof. In the event $\mathcal{A}_2$, the mixture model is fitted by $f_{\text{Mix}}(x; \hat{Q}_{\text{GMM},\mathcal{J}})$. Because Θ is compact and $f(x;\theta)$ is continuous with respect to $\theta \in \Theta$, there exists a finite number M such that, for any possible τ,

$$|\hat{\tau}_{\text{GMM}} - \tau| \leq M.$$

Therefore, the MSE of $\hat{\tau}_{\text{GMM}}$ conditional on the event $\mathcal{A}_2$ is bounded by M^2.

For a fixed J, let $\boldsymbol{\Sigma}$ be the covariance matrix of $\boldsymbol{\phi}_{f_0^{-1/2}}(X) \in \mathbb{R}^J$. Also, let

$$\boldsymbol{m}' = \arg \inf_{(1,\boldsymbol{m}^{\mathrm{T}})^{\mathrm{T}} \in \partial\mathcal{M}_J} (\boldsymbol{m} - \boldsymbol{m}^*)^{\mathrm{T}} \boldsymbol{\Sigma}^{-1} (\boldsymbol{m} - \boldsymbol{m}^*),$$

where $\partial\mathcal{M}_J$ is the boundary of $\mathcal{M}_J$. By Theorem 3.3.2, the true moments $\boldsymbol{m}_J^* = (1, (\boldsymbol{m}^*)^{\mathrm{T}})^{\mathrm{T}} \in \mathbb{R}^{J+1}$ is an interior point of $\mathcal{M}_J$. Therefore, $(\boldsymbol{m}' - \boldsymbol{m}^*)^{\mathrm{T}} \boldsymbol{\Sigma}^{-1} (\boldsymbol{m}' - \boldsymbol{m}^*) > 0$.

For any $\boldsymbol{m}_J = (1, \boldsymbol{m}^{\mathrm{T}})^{\mathrm{T}} \in \mathbb{R}^{J+1}$ which is not an interior point of $\mathcal{M}_J$, it is true that

$$(\boldsymbol{m} - \boldsymbol{m}^*)^{\mathrm{T}} \boldsymbol{\Sigma}^{-1} (\boldsymbol{m} - \boldsymbol{m}^*) \geq (\boldsymbol{m}' - \boldsymbol{m}^*)^{\mathrm{T}} \boldsymbol{\Sigma}^{-1} (\boldsymbol{m}' - \boldsymbol{m}^*).$$

Because $\bar{\boldsymbol{m}}_J$ is not an interior point of $\mathcal{M}_J$ in $\mathcal{A}_2$, we have

$$(\bar{\boldsymbol{m}} - \boldsymbol{m}^*)^{\mathrm{T}} \boldsymbol{\Sigma}^{-1} (\bar{\boldsymbol{m}} - \boldsymbol{m}^*) \geq (\boldsymbol{m}' - \boldsymbol{m}^*)^{\mathrm{T}} \boldsymbol{\Sigma}^{-1} (\boldsymbol{m}' - \boldsymbol{m}^*).$$

By the Chebyshev inequality for random vectors [Chen and Zhou, 1997, Theorem 2.1], we obtain that

$$\begin{aligned}
\mathrm{pr}\left(\hat{\boldsymbol{m}}_J \in \partial \mathcal{M}_J\right) &\leq \mathrm{pr}\left(N(\bar{\boldsymbol{m}} - \boldsymbol{m}^*)^{\mathrm{T}} \boldsymbol{\Sigma}^{-1} (\bar{\boldsymbol{m}} - \boldsymbol{m}^*) \geq N(\boldsymbol{m}' - \boldsymbol{m}^*)^{\mathrm{T}} \boldsymbol{\Sigma}^{-1} (\boldsymbol{m}' - \boldsymbol{m}^*)\right) \\
&\leq \frac{J}{N} \frac{1}{(\boldsymbol{m}' - \boldsymbol{m}^*)^{\mathrm{T}} \boldsymbol{\Sigma}^{-1} (\boldsymbol{m}' - \boldsymbol{m}^*)} \\
&= O\left(JN^{-1}\right).
\end{aligned}$$

Because, for each J, $\hat{\boldsymbol{m}}_J \in \partial \mathcal{M}_J$ implies $\hat{\boldsymbol{m}}_{J+1} \in \partial \mathcal{M}_{J+1}$, the two events $\hat{\boldsymbol{m}}_J \in \partial \mathcal{M}_J$ and $\mathcal{J} \leq J$ are equivalent. Therefore, we have

$$\begin{aligned}
\mathrm{pr}\left(\mathcal{A}_2\right) &= \sum_{J \leq J_N} \mathrm{pr}\left(\mathcal{J} = J\right) \\
&\leq \sum_{J \leq J_N} \mathrm{pr}\left(\mathcal{J} \leq J\right) \\
&= \sum_{J \leq J_N} \mathrm{pr}\left(\hat{\boldsymbol{m}}_J \in \partial \mathcal{M}_J\right) \\
&\leq \sum_{J \leq J_N} O\left(JN^{-1}\right) \\
&= O\left(J_N^2 N^{-1}\right).
\end{aligned}$$

□

C.3 Proof of Lemma 4.5.3

Proof. In the event $\mathcal{A}_3$, the fitted model is $f_{\mathrm{Mix}}(x; \hat{Q}_{\mathrm{GMM},\mathcal{J}})$. Let Θ^+ be the union of the support sets of Q^* and $\hat{Q}_{\mathrm{GMM}}$, and r^+ be the number of elements in Θ^+. Firstly, we show that $r^+ \leq \mathcal{J} + 1$.

When $\mathcal{J} = 2K$, by Theorem 3.3.2, we have two cases where the positive representation of a boundary vector of $\mathcal{M}_{\mathcal{J}}$ could have the largest possible number of support points:

1. The positive representation has K support points in (a, b).
2. The positive representation has $K-1$ support points in (a, b), and the two end points a and b are also its support points.

Therefore, the number of elements in Θ^+, denoted by r^+, is always less than $\mathcal{J}+1$, when $\mathcal{J}$ is even.

When $\mathcal{J} = 2K+1$, by Theorem 3.3.2, we have two cases where the positive representation of a boundary vector of $\mathcal{M}_{\mathcal{J}}$ could have the largest possible number of support points:

1. The positive representation has K support points in (a, b) and one of the end points a.
2. The positive representation has K support points in (a, b), and and one of the end points b.

Therefore, the number of elements in Θ^+ is always less than $\mathcal{J}+1$, when $\mathcal{J}$ is odd.

Let $\mathbf{\Gamma}_{\Theta^+}$ be a $(\mathcal{J}+1) \times r^+$ matrix whose i^{th} column is $\mathbf{\Lambda}_{\mathcal{J}} \boldsymbol{\gamma}_{\mathcal{J}}(\theta)$ where $\mathbf{\Lambda}_{\mathcal{J}}$ is the $(\mathcal{J}+1) \times (\mathcal{J}+1)$ diagonal matrix with the j^{th} diagonal element λ_{j-1}, and $\boldsymbol{\gamma}_{\mathcal{J}}(\theta) = (1, \gamma_1(\theta), \ldots, \gamma_{\mathcal{J}}(\theta))^{\mathrm{T}} \in \mathbb{R}^{\mathcal{J}+1}$. Also let $\boldsymbol{F}_{\Theta^+}(x)$ be a vector in $\mathbb{R}^{r^+}$ whose i^{th} element is $f(x; \theta_i)$, $\theta_i \in \Theta^+$. Because Q^* and $\hat{Q}_{\mathrm{GMM}}$ are two probability measures defined on Θ^+, each of them has an associated vector of weights, denoted by $\boldsymbol{\alpha}^* \in \mathbb{R}^{r^+}$ and $\hat{\boldsymbol{\alpha}}_{\mathrm{GMM}} \in \mathbb{R}^{r^+}$ correspondingly. Moreover, $\boldsymbol{\alpha}^*$ and $\hat{\boldsymbol{\alpha}}_{\mathrm{GMM}}$ are uniquely determined by $\boldsymbol{m}^*_{\mathcal{J}}$ and $\hat{\boldsymbol{m}}_{\mathcal{J}}$ by Theorem 3.3.2. Therefore, we have

$$\boldsymbol{\alpha}^* = \boldsymbol{C}\mathbf{\Gamma}^{\mathrm{T}}_{\Theta^+}\boldsymbol{m}^*_{\mathcal{J}}$$

and

$$\hat{\boldsymbol{\alpha}}_{\mathrm{GMM}} = \boldsymbol{C}\mathbf{\Gamma}^{\mathrm{T}}_{\Theta^+}\hat{\boldsymbol{m}}_{\mathcal{J}},$$

where $\boldsymbol{C} = \left(\boldsymbol{\Gamma}_{\Theta^+}^{\mathrm{T}} \boldsymbol{\Gamma}_{\Theta^+}\right)^{-1}$.

Because τ is finite and

$$\hat{f}_{\mathrm{GMM}}(x) = f_{\mathrm{Mix}}(x; \hat{Q}_{\mathrm{GMM}}) = \boldsymbol{F}_{\Theta^+}^{\mathrm{T}}(x) \hat{\boldsymbol{\alpha}}_{\mathrm{GMM}}$$

we have

$$\tau = \int_{\mathcal{S}} s(x) \hat{f}_{\mathrm{GMM}}(x) \mathrm{d}x = \int_{\mathcal{S}} s(x) \boldsymbol{F}_{\Theta^+}^{\mathrm{T}}(x) \hat{\boldsymbol{\alpha}}_{\mathrm{GMM}} \mathrm{d}x = \boldsymbol{T}_{\Theta^+}^{\mathrm{T}} \hat{\boldsymbol{\alpha}}_{\mathrm{GMM}},$$

where $\boldsymbol{T}_{\Theta^+} = \int_{\mathcal{S}} s(x) \boldsymbol{F}_{\Theta^+}(x) \mathrm{d}x \in \mathbb{R}^{r^+}$.

Let $\boldsymbol{\Gamma}_{\Theta^+} \boldsymbol{C} \boldsymbol{T}_{\Theta^+} = (t_0, \boldsymbol{t}_{\Theta^+}^{\mathrm{T}})^{\mathrm{T}} \in \mathbb{R}^{\mathcal{J}+1}$. By the Cauchy-Schwarz inequality and the fact that the first element of $\hat{\boldsymbol{m}}_{\mathcal{J}}$ and $\boldsymbol{m}_{\mathcal{J}}^*$ are 1s, we have

$$\begin{aligned}
\mathbb{E}\left[(\tau - \hat{\tau}_{\mathrm{GMM}})^2 \mid \mathcal{A}_3\right] &= \mathbb{E}\left[\left(\boldsymbol{T}_{\Theta^+}^{\mathrm{T}} (\hat{\boldsymbol{\alpha}}_{\mathrm{GMM}} - \boldsymbol{\alpha}^*)\right)^2 \mid \mathcal{A}_3\right] \\
&= \mathbb{E}\left[\left(\boldsymbol{T}_{\Theta^+}^{\mathrm{T}} \boldsymbol{C} \boldsymbol{\Gamma}_{\Theta^+}^{\mathrm{T}} (\hat{\boldsymbol{m}}_{\mathcal{J}} - \boldsymbol{m}_{\mathcal{J}}^*)\right)^2 \mid \mathcal{A}_3\right] \\
&= \mathbb{E}\left[\left(\boldsymbol{t}_{\Theta^+}^{\mathrm{T}} (\hat{\boldsymbol{m}} - \boldsymbol{m}^*)\right)^2 \mid \mathcal{A}_3\right] \\
&\leq \mathbb{E}\left[\left\| \boldsymbol{W}_{(\mathcal{J})}^{-1/2} \boldsymbol{t}_{\Theta^+} \right\|_2^2 \times \| \hat{\boldsymbol{m}} - \boldsymbol{m}^* \|_{\boldsymbol{W}_{(\mathcal{J})}}^2 \mid \mathcal{A}_3\right],
\end{aligned}$$

where for any vector $\boldsymbol{a} \in \mathbb{R}^{\mathcal{J}}$, $\|\boldsymbol{a}\|_{\boldsymbol{W}_{(\mathcal{J})}}^2 = \boldsymbol{a}^{\mathrm{T}} \boldsymbol{W}_{(\mathcal{J})} \boldsymbol{a}$. Here ${\boldsymbol{W}_{(\mathcal{J})}}^{-1/2}$ exists because $\boldsymbol{W}_{(\mathcal{J})}$ is non-singular and positive definite by the assumptions.

Because $f(x;\theta)$ is an exponential family distribution, both of $f(x;\theta)$, for each $x \in \mathcal{S}$, and $\gamma_j(\theta)$, for each $j \in \{1, \ldots, \mathcal{J}\}$, are continuous with respect to θ. Therefore, each element of $\boldsymbol{W}_{(\mathcal{J})}^{-1/2} \boldsymbol{t}_{\Theta^+}$ is also a continuous function of θ. Additional to the compactness of Θ, each element of $\boldsymbol{W}_{(\mathcal{J})}^{-1/2} \boldsymbol{t}_{\Theta^+}$ is bounded by a finite number, say M. Then $\|\boldsymbol{W}_{(\mathcal{J})}^{-1/2} \boldsymbol{t}_{\Theta^+}\|_2^2$ is bounded by $\mathcal{J} M^2$ and $J_N M^2$ by $\mathcal{J} \leq J_N$.

We use the non-expansive property in convex projection ([Deutsch, 2001, p.g. 72]) and have

$$
\begin{aligned}
\mathbb{E}\left[\|\hat{\boldsymbol{m}}-\boldsymbol{m}^*\|^2_{\boldsymbol{W}_{(\mathcal{J})}} \mid \mathcal{A}_3\right] &\leq \mathbb{E}\left[\|\bar{\boldsymbol{m}}-\boldsymbol{m}^*\|^2_{\boldsymbol{W}_{(\mathcal{J})}} \mid \mathcal{A}_3\right] \\
&\leq \mathbb{E}\left[\|\boldsymbol{W}_{(\mathcal{J})}\|_2^2 \, \|\bar{\boldsymbol{m}}-\boldsymbol{m}^*\|_2^2 \mid \mathcal{A}_3\right] \\
&\leq \sup_J \|\boldsymbol{W}_{(J)}\|_2^2 \times \mathbb{E}\left[\sum_{j=1}^{\mathcal{J}} (\bar{m}_j - m_j^*)^2 \mid \mathcal{A}_3\right] \\
&\leq \sup_J \|\boldsymbol{W}_{(J)}\|_2^2 \times \mathbb{E}\left[\sum_{j=1}^{J_N} (\bar{m}_j - m_j^*)^2 \mid \mathcal{A}_3\right].
\end{aligned}
$$

Note that

$$
f_{\mathrm{Mix}}(x; Q^*) = \int_\Theta f(x;\theta)\mathrm{d}Q^* \leq |\Theta| f_0(x).
$$

Therefore, for each j ,

$$
\mathrm{Var}[\phi_j(X) f_0^{-1/2}(X)] \leq \mathbb{E}[\phi_j^2(X) f_0^{-1}(X)] \leq |\Theta| \int_\Theta \phi_j^2(x)\mathrm{d}x = |\Theta|.
$$

In sum, we have

$$
\begin{aligned}
&\mathrm{pr}(\mathcal{A}_3)\mathbb{E}\left[(\tau - \hat{\tau}_{\mathrm{GMM}})^2 \mid \mathcal{A}_3\right] \\
\leq\ &\mathrm{pr}(\mathcal{A}_3)\mathbb{E}\left[\left\|\boldsymbol{W}_{(\mathcal{J})}^{-1/2}\boldsymbol{t}_{\Theta^+}\right\|_2^2 \times \|\hat{\boldsymbol{m}}-\boldsymbol{m}^*\|^2_{\boldsymbol{W}_{(\mathcal{J})}} \mid \mathcal{A}_3\right] \\
\leq\ &\mathrm{pr}(\mathcal{A}_3) J_N M^2 \times \sup_J \|\boldsymbol{W}_{(J)}\|_2^2 \times \mathbb{E}\left[\sum_{j=1}^{J_N} (\bar{m}_j - m_j^*)^2 \mid \mathcal{A}_3\right] \\
\leq\ &J_N M^2 \times \sup_J \|\boldsymbol{W}_{(J)}\|_2^2 \times \sum_{i=1}^{3} \mathrm{pr}(\mathcal{A}_i)\mathbb{E}\left[\sum_{j=1}^{J_N} (\bar{m}_j - m_j^*)^2 \mid \mathcal{A}_i\right] \\
=\ &J_N M^2 \times \sup_J \|\boldsymbol{W}_{(J)}\|_2^2 \times \mathbb{E}\left[\sum_{j=1}^{J_N} (\bar{m}_j - m_j^*)^2\right] \\
=\ &J_N M^2 \times \sup_J \|\boldsymbol{W}_{(J)}\|_2^2 \times \sum_{j=1}^{J_N} \mathrm{Var}\left[\phi_j(X) f_0^{-1/2}(X)\right] \\
\leq\ &\frac{1}{N} J_N^2 M^2 \times |\Theta| \times \sup_J \|\boldsymbol{W}_{(J)}\|_2^2 \\
=\ &O(J_N^2 N^{-1}).
\end{aligned}
$$

□

C.4 Proof of Theorem 4.6.1

Proof. Because of the compactness of Θ and Equation (4.7), we may change the order of the integrals and the infinite sum and obtain

$$\begin{aligned}
&\int_{\Theta} \boldsymbol{L}_{(\infty)}(\theta)\boldsymbol{L}^{\mathrm{T}}_{(\infty)}(\theta)\mathrm{d}\theta\boldsymbol{\phi}_{f_0^{-1/2}}(z) \\
=& \int_{\Theta} \boldsymbol{L}_{(\infty)}(\theta)\boldsymbol{L}^{\mathrm{T}}_{(\infty)}(\theta)\boldsymbol{\phi}_{f_0^{-1/2}}(z)\mathrm{d}\theta \\
=& \int_{\Theta} \boldsymbol{L}_{(\infty)}(\theta)\left(f(z;\theta)f_0^{-1/2}(z) - \int_{\mathcal{S}} f(y;\theta)f_0^{1/2}(y)\mathrm{d}y\right)\mathrm{d}\theta.
\end{aligned}$$

Because for each $\theta \in \Theta$

$$0 \leq f(z;\theta) \leq |\Theta| f_0(z),$$

we have that $f(z;\theta)f_0^{-1/2}(z)$ goes to zero as $f_0(z)$ goes to zero. Further note that, for each j, $L_j(\theta)\int_{\mathcal{S}} f(y;\theta)f_0^{1/2}(y)\mathrm{d}y$ is bounded for each $\theta \in \Theta$. We have that each element of $\boldsymbol{W}^{\mathrm{Robust}}_{(\infty)}\boldsymbol{\phi}_{f_0^{-1/2}}(z)$ converges to a constant as $f_0(z)$ goes to zero. □

C.5 MATLAB Code for Algorithm 4.1

```
function [mNew, weightshat, g, Ind] ...
    = GMMLCNM(Up, Mhatp, q, as, W)


e = 1e-10;
msn = Up*as';
pLs = find(as > e);
Ind = 0;
count = 1;


while Ind == 0
    ms = msn;
    cg = Mhatp - ms';
    Utheta =  ms(:,ones(1,length(q))) - Up;
    g  = cg* W*Utheta;
    dg = diff(g);
    signdg = sign(dg);
    dsigndg = diff(signdg);
    minL = find(dsigndg == 2)+1;
    L = minL;

    pLsnew = [1, pLs, L, length(q)];
    pLsnew = unique(pLsnew);
    Us = Up(:,pLsnew);

    warning off;
    options = optimset('display', 'off');
    aso = lsqlin(real(W^(1/2)*Us), ...
        real(W^(1/2)*Mhatp'), ...
        -eye(size(Us,2)), ...
        zeros(size(Us,2),1), ...
        ones(1,length(pLsnew)), ...
        1, [], [], [], options);
    msn = Us*aso;
    pLs = pLsnew(aso > e);
    count = count + 1;

    d = (ms-msn)'*W*(ms-msn);
    if d < e || count > 100
        Ind = 1;
    end

end

mNew = msn;
weightshat = zeros(length(q),1);
weightshat(pLsnew) = aso;
```

```
dn = (msn - Mhatp')'*W*(msn - Mhatp');
if max(abs(g)) < e || dn < e
    Ind = 0; % the solution is not unique
else
    Ind = 1; % the solution is unique
end
```

Chapter 5

The Generalized Method of Moments for Mixed-Effects Models with Univariate Random Effects

5.1 Introduction

Longitudinal data analysis has attracted considerable research interest in the past decades. A good review can be found in [Diggle, 2002] and [Fitzmaurice et al., 2012] and references therein. There are two classes of models for longitudinal data: the population-average models and the subject-specific models; see [Lee and Nelder, 2004] for a detailed discussion. The regression parameter has different interpretations in these models, except when the link function is linear. Usually, the subject-specific models are more useful when the main scientific objective is to make inferences about individuals rather than the populations; see [Fitzmaurice et al., 2012].

Semi-parametric mixture models are a subclass of the subject-specific models, where the distribution of the response conditional on the random effects is parametric and the random effects distribution is non-parametric. It avoids the possible sensitivity of the inference conclusions to the specification of random effects distributions; see [Neuhaus et al., 1992] and [Heagerty and Kurland, 2001].

To fit a semi-parametric mixture model, the maximum likelihoods method is commonly used. Under regularity conditions, the consistency of the MLE is established by Kiefer and Wolfowitz [1956]. However, finding the MLE is widely regarded as a computationally intensive problem; see [Aitkin, 1999] and [Wang, 2010] for some computational suggestions. Moreover, few results related to making inferences with the MLE for semi-parametric mixture models can be found in the literature.

Another class of approaches, including the (corrected) conditional mixed methods (CMM and CCMM) in [Sutradhar and Godambe, 1997], the penalized generalized weighted least squares method (PGWLS) [Jiang, 1999] and the mixed-effects quadratic inference function (QIF) methods [Wang et al., 2012], are based on the generalized estimating equations conditional on the random effects. This class of approaches involves the prediction of the random effects. Because the number of the random effects always increases with the sample size, it is questionable if there is sufficient information for all the random effects; see [Jiang, 1999]. Asymptotic results for the mixed-effects QIF estimators are established when the sample size and the cluster size go to infinity simultaneously; see [Wang et al., 2012]. However, the cluster size may not always be large enough in real applications; see the two real data examples in Section 1.6.2 and 1.6.3.

The unconditional mixed method (UMM) is based on the marginal generalized estimating equations; see [Sutradhar and Godambe, 1997]. In the UMM, the marginal estimating function is approximated by a function of the regression parameter and the variance of the random effects distribution. However, such approximation is valid only when the dispersion parameter of the random effects distribution is small. Similar idea has also been used to the likelihood functions, when the random effects distribution is normal; see [Breslow and Clayton, 1993].

The aim of this chapter is to fit a semi-parametric mixture model, when the random effects are univariate. We reparameterize the inverse link function into a function of the regression parameter and a countable set of the generalized moments of the random-effects distribution. Then, we use a truncation approximation of the reparameterized model and fit it using the GMM. The reparameterization-approximation procedure is described in Section 5.4.1.

The major contribution made in this chapter is the introduction of the GMM for mixed-effects models with univariate random effects; see Section 5.2 for details of the considered mixed-effects models. We apply the reparameterization-approximation procedure to the considered mixed-effects models and use the GMM to fit the model. As theoretically shown later in Chapter 6, the GMM estimator is consistent; see Section 5.7 for simulation evidences. Because the proposed method is based on the marginal estimating equations, the resulting estimator is robust to the misspecification of the likelihood functions; see the simulation results in Section 5.7.

This chapter is organized as follows. In Section 5.2, we give the response model, the mixed-effects models with univariate random effects, and its assumptions. The response model is based on the estimating equations conditional on the random effects. In Section 5.3, we review the UMM proposed by [Sutradhar and Godambe, 1997]. We discuss the limits of the UMM and the motivation for using the GMM. In Section 5.4, we introduce the GMM for mixed-effects models with univariate random effects. Firstly, we describe the reparameterization-approximation procedure to the considered mixed-effects models; see Section 5.4.1. Next, we give the definition of the GMM estimator for the considered mixed-effects model in Section 5.4.2. The GMM involves a minimization problem over a convex set and a computational algorithm is given in Section 5.5. In Section 5.6, we discuss the assessment of the fitted model using residual analysis. In the same section, we also give two possible estimates to the covariance matrix of the residuals. Our work is supported by simulations studies in Section 5.7. We use the proposed method of fit a model for the Retina Surgery Data in Section 5.8. Lastly, we ends this chapter with a discussion. The MATLAB code for the proposed algorithms in this chapter can be seen in Appendix D.

5.2 Response Model and its Assumptions

Our data setup is as follows. There are $n = 1, \ldots, N$ independent individuals, each with $t = 1, \ldots, T_n$ visits. At visit t, the complete data of the n^{th} individual is $(Y_{nt}, \boldsymbol{X}_{nt}, Z_{nt})$, where $Y_{nt} \in \mathbb{R}$ is the response, $\boldsymbol{X}_{nt} \in \mathbb{R}^p$ are the covariates to the fixed effects, and $Z_{nt} \in \mathbb{R}$ are the covariates to the random effects. We use the notation

$\boldsymbol{Y}_n = (Y_{n1}, \ldots, Y_{nT_n})^{\mathrm{T}} \in \mathbb{R}^{T_n}$, $\boldsymbol{X}_n = (\boldsymbol{X}_{n1}, \ldots, \boldsymbol{X}_{nT_n})$ and $\boldsymbol{Z}_n = (Z_{n1}, \ldots, Z_{nT_n})^{\mathrm{T}} \in \mathbb{R}^{T_n}$. Let $b_n \in \mathbb{R}$ be the univariate random effects. For each n, the sample space of b_n is $\mathcal{B}$.

Consider the epileptic seizures data, described in Section 1.6.2, as an example. There are 59 epileptics, which are considered as independent individuals. Therefore, $N = 59$. The number of epileptic seizures of each patient are observed 4 times. It means that, for each $n = 1, \ldots, 59$, $T_n = 4$. In the retina surgery data in Section 1.6.3. There are 31 patients, i.e., $N = 31$. However, the number of visits of each patient is different. For example, the first patient has 7 visits, while the second patient has 8 visits.

Suppose that the following model has been assumed:

1. Given the random effects b_n, the responses $Y_{nt} \mid (\boldsymbol{X}_{nt}, Z_{nt}, b_n)$, $t = 1, \ldots, T_n$, are independent of one another.

2. For each n and $t = 1, \ldots, T_n$, the mean of $Y_{nt} \mid (\boldsymbol{X}_{nt}, Z_{nt}, b_n)$ depends on the regression parameter $\boldsymbol{\beta}$ via the following linear predictor

$$g(\mathbb{E}[Y_{nt} \mid \boldsymbol{X}_{nt}, Z_{nt}, b_n]) = \boldsymbol{X}_{nt}^{\mathrm{T}}\boldsymbol{\beta} + Z_{nt}b_n, \tag{5.1}$$

   where $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)^{\mathrm{T}} \in \mathbb{R}^p$ is the regression parameter and $g(\cdot)$ is a known invertible link function.

3. For each n and t, the conditional variance of $Y_{nt} \mid (\boldsymbol{X}_{nt}, Z_{nt}, b_n)$ satisfies

$$\mathrm{Var}[Y_{nt} \mid \boldsymbol{X}_{nt}, Z_{nt}, b_n] = \sigma \times h \circ g^{-1}(\boldsymbol{X}_{nt}^{\mathrm{T}}\boldsymbol{\beta} + Z_{nt}b_n), \tag{5.2}$$

   where $h(\cdot)$ is a known variance function, σ is a constant and $g^{-1}(\cdot)$ is the inverse link function $g(\cdot)$. Here we assume that σ is known.

If a distribution assumption is further made on $Y_{nt} \mid (\boldsymbol{X}_{nt}, Z_{nt}, b_n)$, we can write down the likelihood for each n as

$$\mathrm{pr}_{\boldsymbol{\beta}}(\boldsymbol{Y}_n \mid \boldsymbol{X}_n, Z_n) = \int_{\mathcal{B}} \left\{ \prod_{t=1}^{T_n} \mathrm{pr}_{\boldsymbol{\beta}}(Y_{nt} \mid \boldsymbol{X}_{nt}, Z_{nt}, b_n) \right\} \mathrm{d}Q,$$

where $\mathrm{pr}_{\boldsymbol{\beta}}(Y_{nt} \mid \boldsymbol{X}_{nt}, Z_{nt}, b_n)$ has the mean $g^{-1}(\boldsymbol{X}_{nt}^{\mathrm{T}}\boldsymbol{\beta}+Z_{nt}b_n)$ and the variance $\sigma \times h \circ g^{-1}(\boldsymbol{X}_{nt}^{\mathrm{T}}\boldsymbol{\beta}+Z_{nt}b_n)$, and $Q(b)$ is a probability measure defined on $\mathcal{B}$. The above class of models is known as semi-parametric mixture models, which have wide applications in longitudinal data analysis; see [Diggle, 2002] and [Fitzmaurice et al., 2012] for details and examples. In this chapter, we focus on (5.1) and (5.2) without any distribution assumption on $Y_{nt} \mid (\boldsymbol{X}_{nt}, Z_{nt}, b_n)$ and b_n.

5.3 The Unconditional Mixed Method: A Review

Sutradhar and Godambe [1997] considered the class of random intercept models, where the intercepts are identically distributed with zero mean and variance υ^2, but no functional assumption is made on the random intercepts distribution. Here the variance υ^2 is unknown. The UMM is based on the condition that the unconditional mean and covariance matrix of the response vector $\boldsymbol{Y}_n$ can be expressed as (or be approximated by) functions of the regression parameter $\boldsymbol{\beta}$ and the variance parameter υ^2. Generally, the approximated unconditional means and variances can be obtained through the Laplace approximation on $\mathrm{pr}_{\boldsymbol{\beta}}(Y_{nt} \mid \boldsymbol{X}_{nt}, b_n)$, when υ^2 is small; see [Sutradhar and Rao, 1996] and [Sutradhar and Godambe, 1997]. A similar idea is also used to approximate probability functions in [Marriott, 2002].

In the UMM, the following steps are repeated iteratively until the convergence to the estimated values of $\boldsymbol{\beta}$ and υ^2:

1. Given the values of υ^2, the regression parameter $\boldsymbol{\beta}$ is estimated from the generalized estimating equations based on the approximated unconditional means and variances.

2. Given the values of υ^2 and $\boldsymbol{\beta}$, the random effects are predicated through the generalized estimating equations based on the approximated unconditional means and variances.

3. The variance υ^2 is estimated by using the predicated random effects.

The UMM empirically shows superior performance to the CMM and CCMM, which are proposed in the same paper by Sutradhar and Godambe. However, there are two major issues. Firstly, the failure of the Laplace approximation could lead to large bias to the estimators, when υ^2 is large. This is due to the natural of the Laplace approximation. Secondly, the predication of the random effects, whose number increases with the sample size, increases the computational load and causes the computational convergence issues, when the sample size is large.

We argue that the predication of all the random effects is necessary in the UMM. The UMM is based on the generalized estimating equations methods, in which no constraints is put on the non-negativeness of υ^2. To respect the non-negativeness of υ^2, it is reasonable to estimate υ^2 using the sample variance of the predicated random effects.

However, in many real applications, the regression parameter $\boldsymbol{\beta}$ is the one of interest. The predication of all the random effects would be unnecessary, if we can repeat the following steps iteratively until the convergence to the estimated value of $\boldsymbol{\beta}$ and υ^2:

1. Given the values of υ^2, the regression parameter $\boldsymbol{\beta}$ is estimated from the generalized estimating equations based on the approximated unconditional means and variance.

2. Given the values of $\boldsymbol{\beta}$, the variance υ^2 is estimated from the generalized estimating equations based on the approximated unconditional means and variance.

Making inference on the regression parameter $\boldsymbol{\beta}$ without predicating the random effects is the motivation of the GMM for mixed-effects models with univariate random effects.

5.4 The Generalized Method of Moments for Mixed-effects Models with Univariate Random Effects

With the spirit of working at the marginal level, we can derive a marginal mean of Y_{nt} from (5.1), i.e.

$$\begin{aligned}\mathbb{E}[Y_{nt}] &= \mathbb{E}_{\boldsymbol{X}_{nt},Z_{nt}}\left[\mathbb{E}_{b_n}\left[\mathbb{E}\left[Y_{nt} \mid \boldsymbol{X}_{nt}, Z_{nt}, b_n\right]\right]\right] \\ &= \int_{\mathcal{X}\times\mathcal{Z}}\int_{\mathcal{B}} g^{-1}(\boldsymbol{x}_{nt}^{\mathrm{T}}\boldsymbol{\beta} + z_{nt}b_n)\mathrm{d}Q \times \mathrm{pr}(\boldsymbol{x}_{nt}, z_{nt})\mathrm{d}(\boldsymbol{x}_{nt}, z_{nt}),\end{aligned}$$

where $\mathrm{pr}(\boldsymbol{x}_{nt}, z_{nt})$ is the joint probability function of $\boldsymbol{X}_{nt}$ and Z_{nt} with sample space $\mathcal{X} \times \mathcal{Z}$. For each n and t, let

$$U_{nt}(\boldsymbol{\beta}, Q) = Y_{nt} - \int_{\mathcal{B}} g^{-1}(\boldsymbol{X}_{nt}^{\mathrm{T}}\boldsymbol{\beta} + Z_{nt}b_n)\mathrm{d}Q,$$

where Q is a probability measure of the random effects b_n over $\mathcal{B}$. Let $\boldsymbol{U}_n(\boldsymbol{\beta}, Q) = (U_{n1}(\boldsymbol{\beta}, Q), \ldots, U_{nT_n}(\boldsymbol{\beta}, Q))^{\mathrm{T}} \in \mathbb{R}^{T_n}$. We then have the moment conditions, for each n,

$$\mathbb{E}_{\boldsymbol{Y}_n,\boldsymbol{X}_n,\boldsymbol{Z}_n}\left[\boldsymbol{U}_n(\boldsymbol{\beta}, Q)\right] = 0 \in \mathbb{R}^{T_n}.$$

Motivated by the discussion in Section 5.3, we propose our approach for mixed-effects models with univariate random effects. By the reparameterization-approximation procedure introduced in Section 2.4, we firstly approximate $U_{nt}(\boldsymbol{\beta}, Q)$ as a function of $\boldsymbol{\beta}$ and $\boldsymbol{\alpha}$, where $\boldsymbol{\alpha} \in \mathbb{R}^{J_N}$ depends on the random effects distribution Q and the dimension of $\boldsymbol{\alpha}$ grows with the sample size N. Here the parameter $\boldsymbol{\alpha}$ has a natural parameter space; see Section 5.4.1. Then, we use the GMM to estimate $\boldsymbol{\beta}$ and $\boldsymbol{\alpha}$; see Section 5.4.2.

5.4.1 The Reparameterization-Approximation Procedure with Orthogonal Polynomials

Recall that in Section 2.4, we have introduced the reparameterization-approximation procedure for the GLMM. Here we revisit the procedure in this current context.

Let $\{P_j(b)\}_{j=0}^{\infty}$ be an orthonormal polynomial system defined on $L^2(\mathcal{B}, \mu)$, where μ is a measure defined on $\mathcal{B}$. Assume that, for each n and t, $g^{-1}(\boldsymbol{X}_{nt}^{\mathrm{T}}\boldsymbol{\beta}+Z_{nt}b) \in L^2(\mathcal{B}, \mu)$, then we have the expansion

$$g^{-1}(\boldsymbol{X}_{nt}^{\mathrm{T}}\boldsymbol{\beta} + Z_{nt}b) = \sum_{j=0}^{\infty} \phi_{ntj}(\boldsymbol{\beta})P_j(b),$$

where for each j,

$$\phi_{ntj}(\boldsymbol{\beta}) = \int_{\mathcal{B}} g^{-1}(\boldsymbol{X}_{nt}^{\mathrm{T}}\boldsymbol{\beta} + Z_{nt}b)P_j(b)\mathrm{d}\mu. \tag{5.3}$$

By changing the order of the integrals, we have

$$U_{nt}(\boldsymbol{\beta}, Q) = Y_{nt} - \sum_{j=0}^{\infty} \phi_{ntj}(\boldsymbol{\beta})\alpha_j.$$

The truncation approximation of $U_{nt}(\boldsymbol{\beta}, Q)$ is defined by

$$U_{nt}(\boldsymbol{\beta}, \boldsymbol{\alpha}) = Y_{nt} - \sum_{j=0}^{J_N} \phi_{ntj}(\boldsymbol{\beta})\alpha_j, \tag{5.4}$$

where $\boldsymbol{\alpha} = (\alpha_0, \ldots, \alpha_{J_N})^{\mathrm{T}} \in \mathbb{R}^{J_N}$ and for each j,

$$\alpha_j = \int_{\mathcal{B}} P_j(b)\mathrm{d}Q$$

and J_N is an integer which can increase with the increase of the sample size N.

In matrix form, for each n, we have

$$\boldsymbol{U}_n(\boldsymbol{\beta}, \boldsymbol{\alpha}) = \boldsymbol{Y}_n - \boldsymbol{\Phi}_n^{\mathrm{T}}(\boldsymbol{\beta})\boldsymbol{\alpha}, \tag{5.5}$$

where $\boldsymbol{\Phi}_n(\boldsymbol{\beta})$ is a $J_N \times T_n$ matrix whose elements are $\phi_{ntj}(\boldsymbol{\beta})$. Furthermore, the vector $\boldsymbol{\alpha}$ is defined on the convex set

$$\mathcal{M} = \left\{ \boldsymbol{\alpha} = \int_{\mathcal{B}} \boldsymbol{P}(b)\mathrm{d}Q \in \mathbb{R}^{J_N} \right\} \tag{5.6}$$

where Q is any probability measure defined on $\mathcal{B}$ and

$$\boldsymbol{P}(b) = (P_0(b), \ldots, P_{J_N}(b))^{\mathrm{T}} \in \mathbb{R}^{J_N}$$

is a vector function of $b \in \mathcal{B}$. By Definition 3.1.1, $\mathcal{M}$ is the generalized moment space induced by $\{P_j(b)\}_{j=0}^{J_N}$.

On the other hand, assume that, for each n and t, $h \circ g^{-1}(\boldsymbol{X}_{nt}^{\mathrm{T}}\boldsymbol{\beta} + Z_{nt}b_n)$ and $g^{-1}(\boldsymbol{X}_{nt}^{\mathrm{T}}\boldsymbol{\beta} + Z_{nt}b)$ are in $L^2(\mathcal{B}, \mu)$. We also can use $\{P_j(b)\}_{j=0}^{J_N}$ in an approximation, i.e.,

$$\sigma \times h \circ g^{-1}(\boldsymbol{X}_{nt}^{\mathrm{T}}\boldsymbol{\beta} + Z_{nt}b_n) \approx \sigma \times \sum_{j=0}^{J_N} a_{ntj}(\boldsymbol{\beta})P_j(b),$$

and

$$\left(g^{-1}(\boldsymbol{X}_{n}^{\mathrm{T}}\boldsymbol{\beta} + Z_{nt}b)\right)^2 \approx \sum_{j=0}^{J_N} c_{nttj}(\boldsymbol{\beta})P_j(b),$$

where for each j,

$$a_{ntj}(\boldsymbol{\beta}) = \int_{\mathcal{B}} h \circ g^{-1}(\boldsymbol{X}_{nt}^{\mathrm{T}}\boldsymbol{\beta} + Z_{nt}b)P_j(b)\mathrm{d}\mu \tag{5.7}$$

and

$$c_{nttj}(\boldsymbol{\beta}) = \int_{\mathcal{B}} \left(g^{-1}(\boldsymbol{X}_{nt}^{\mathrm{T}}\boldsymbol{\beta} + Z_{nt}b)\right)^2 P_j(b)\mathrm{d}\mu. \tag{5.8}$$

By the law of total variance and the law of total expectation, we have

$$\begin{aligned}
&\mathrm{Var}[Y_{nt} \mid \boldsymbol{X}_{nt}, Z_{nt}] \\
&= \mathbb{E}_{b_n}\left[\mathrm{Var}[Y_{nt} \mid \boldsymbol{X}_{nt}, Z_{nt}, b_n]\right] + \mathrm{Var}_{b_n}[\mathbb{E}\left[Y_{nt} \mid \boldsymbol{X}_{nt}, Z_{nt}, b_n\right]] \\
&= \mathbb{E}_{b_n}\left[\mathrm{Var}[Y_{nt} \mid \boldsymbol{X}_{nt}, Z_{nt}, b_n]\right] + \mathbb{E}_{b_n}\left[\left(\mathbb{E}\left[Y_{nt} \mid \boldsymbol{X}_{nt}, Z_{nt}, b_n\right]\right)^2\right] \\
&\quad - \left(\mathbb{E}\left[Y_{nt} \mid \boldsymbol{X}_{nt}, Z_{nt}\right]\right)^2.
\end{aligned} \tag{5.9}$$

Changing the order of the integrals, we can approximate the terms as

$$\mathbb{E}_{b_n}\left[\mathrm{Var}[Y_{nt} \mid \boldsymbol{X}_{nt}, Z_{nt}, b_n]\right] \approx \sigma \times \sum_{j=0}^{J_N} a_{ntj}(\boldsymbol{\beta})\alpha_j,$$

$$\mathbb{E}_{b_n}\left[\left(\mathbb{E}\left[Y_{nt} \mid \boldsymbol{X}_{nt}, Z_{nt}, b_n\right]\right)^2\right] \approx \sum_{j=0}^{J_N} c_{nttj}(\boldsymbol{\beta})\alpha_j,$$

and

$$(\mathbb{E}\left[Y_{nt} \mid \boldsymbol{X}_{nt}, Z_{nt}\right])^2 \approx \left(\sum_{j=0}^{J_N} \phi_{ntj}(\boldsymbol{\beta})\alpha_j\right)^2 .$$

Therefore, the variance function of $Y_{nt} \mid (\boldsymbol{X}_{nt}, Z_{nt})$ is approximated by

$$V_{nt}(\boldsymbol{\beta}, \boldsymbol{\alpha}) = \sigma \times \boldsymbol{a}_{nt}^{\mathrm{T}}(\boldsymbol{\beta})\boldsymbol{\alpha} + \boldsymbol{c}_{ntt}^{\mathrm{T}}(\boldsymbol{\beta})\boldsymbol{\alpha} - \left(\boldsymbol{\phi}_{nt}^{\mathrm{T}}(\boldsymbol{\beta})\boldsymbol{\alpha}\right)^2 , \tag{5.10}$$

Note that the approximations

$$\mathbb{E}_{b_n}\left[\mathrm{Var}[Y_{nt} \mid \boldsymbol{X}_{nt}, Z_{nt}, b_n]\right] \approx \sigma \times \sum_{j=0}^{J_N} a_{ntj}(\boldsymbol{\beta})\alpha_j$$

and

$$\mathrm{Var}_{b_n}[\mathbb{E}\left[Y_{nt} \mid \boldsymbol{X}_{nt}, Z_{nt}, b_n\right]] \approx \sum_{j=0}^{J_N} c_{nttj}(\boldsymbol{\beta})\alpha_j - \left(\sum_{j=0}^{J_N} \phi_{ntj}(\boldsymbol{\beta})\alpha_j\right)^2$$

may not be valid due to the non-negative constraints. Instead, we use the approximation

$$\mathbb{E}_{b_n}\left[\mathrm{Var}[Y_{nt} \mid \boldsymbol{X}_{nt}, Z_{nt}, b_n]\right] \approx \max\left\{\epsilon, \sigma \times \sum_{j=0}^{J_N} a_{ntj}(\boldsymbol{\beta})\alpha_j\right\},$$
$$\mathrm{Var}_{b_n}[\mathbb{E}\left[Y_{nt} \mid \boldsymbol{X}_{nt}, Z_{nt}, b_n\right]] \approx \max\left\{\epsilon, \sum_{j=0}^{J_N} c_{nttj}(\boldsymbol{\beta})\alpha_j - \left(\sum_{j=0}^{J_N} \phi_{ntj}(\boldsymbol{\beta})\alpha_j\right)^2\right\},$$

where ϵ is a small positive number. Then, the variance function of $Y_{nt} \mid (\boldsymbol{X}_{nt}, Z_{nt})$ is approximated by

$$\begin{aligned} V_{\mathrm{adj},nt}(\boldsymbol{\beta}, \boldsymbol{\alpha}) = {} & \max\left\{\epsilon, \sigma \times \boldsymbol{a}_{nt}^{\mathrm{T}}(\boldsymbol{\beta})\boldsymbol{\alpha}\right\} \\ & + \max\left\{\epsilon, \boldsymbol{c}_{ntt}^{\mathrm{T}}(\boldsymbol{\beta})\boldsymbol{\alpha} - \left(\boldsymbol{\phi}_{nt}^{\mathrm{T}}(\boldsymbol{\beta})\boldsymbol{\alpha}\right)^2\right\}, \end{aligned}$$

where

$$\boldsymbol{a}_{nt}(\boldsymbol{\beta}) = (a_{nt0}(\boldsymbol{\beta}), \ldots, a_{ntJ_N}(\boldsymbol{\beta}))^{\mathrm{T}} \in \mathbb{R}^{J_N},$$
$$\boldsymbol{c}_{ntt}(\boldsymbol{\beta}) = (c_{ntt0}(\boldsymbol{\beta}), \ldots, c_{nttJ_N}(\boldsymbol{\beta}))^{\mathrm{T}} \in \mathbb{R}^{J_N}$$

and

$$\boldsymbol{\phi}_{nt}(\boldsymbol{\beta}) = (\phi_{nt0}(\boldsymbol{\beta}), \ldots, \phi_{ntJ_N}(\boldsymbol{\beta}))^{\mathrm{T}} \in \mathbb{R}^{J_N}.$$

As will be shown in Chapter 6, under some conditions, $V_{nt}(\boldsymbol{\beta}, \boldsymbol{\alpha})$ could be positive for each n and $t \in \{1, \ldots, T_n\}$, for large sample size N.

5.4.2 The Generalized Method of Moments

For each n, let

$$\boldsymbol{W}_n(\boldsymbol{\beta}, \boldsymbol{\alpha}) = \boldsymbol{V}_n^{-1/2}(\boldsymbol{\beta}, \boldsymbol{\alpha})\boldsymbol{R}_n^{-1}\boldsymbol{V}_n^{-1/2}(\boldsymbol{\beta}, \boldsymbol{\alpha})$$

where $\boldsymbol{V}_n(\boldsymbol{\beta}, \boldsymbol{\alpha})$ is a $T_n \times T_n$ diagonal matrix whose diagonal elements are $V_{nt}(\boldsymbol{\beta}, \boldsymbol{\alpha})$ defined in Equation (5.10) and $\boldsymbol{R}_n$ is a "working" correlation matrix. Common choices of the "working" correlation matrix include the independence, the exchangeable and the first order auto-regressive (AR(1)) correlation matrices; see [Liang and Zeger, 1986]. In the literature of the GEE methods and the GMM, the choice of the "working" correlation matrix will not change the consistency of the estimators but the efficiency.

Let $(\tilde{\boldsymbol{\beta}}, \tilde{\boldsymbol{\alpha}})$ be an initial estimator of $(\boldsymbol{\beta}, \boldsymbol{\alpha})$ and for each n, $\tilde{\boldsymbol{W}}_n = \boldsymbol{W}_n(\tilde{\boldsymbol{\beta}}, \tilde{\boldsymbol{\alpha}})$. The initial estimates will be discussed later in Section 6.3. We define the GMM for mixed-effects models with univariate random effects as follows.

Definition 5.4.1 (The GMM for Mixed-Effects Models with Univariate Random Effects)**.**
Given a data set $(\boldsymbol{Y}_n, \boldsymbol{X}_n, \boldsymbol{Z}_n)$*,* $n = 1, \ldots, N$*, from the data setup in Section 5.2, the GMM estimator for mixed-effects models with univariate random effects, denoted by* $(\hat{\boldsymbol{\beta}}_{\mathrm{GMM}}, \hat{\boldsymbol{\alpha}}_{\mathrm{GMM}})$*, is the solution of the following optimization problem*

$$\begin{aligned} \min \quad & \frac{1}{N}\sum_{n=1}^{N} \boldsymbol{U}_n^{\mathrm{T}}(\boldsymbol{\beta}, \boldsymbol{\alpha})\tilde{\boldsymbol{W}}_n\boldsymbol{U}_n(\boldsymbol{\beta}, \boldsymbol{\alpha}) \\ s.t. \quad & \boldsymbol{\alpha} \in \mathcal{M}, \end{aligned} \tag{5.11}$$

where $\mathcal{M}$ *is defined in Equation (5.6).*

5.5 Computational Algorithms

To obtain the GMM estimators for mixed-effects models with univariate random effects, we propose the following computational algorithm for the optimization problem (5.11).

Algorithm 5.1 (The Alternating Parameter Algorithm)**.**
Set $s = 0$*. From an initial estimate* $\boldsymbol{\alpha}^{(0)} \in \mathbb{R}^J$*, repeat the following steps at the* $(s+1)^{\text{th}}$ *iteration:*

1. *Given* $\boldsymbol{\alpha}^{(s)}$*, solve the optimization problem*

$$\boldsymbol{\beta}^{(s+1)} = \arg\min_{\boldsymbol{\beta}} \frac{1}{N}\sum_{n=1}^{N} \boldsymbol{U}_n^{\mathrm{T}}(\boldsymbol{\beta}, \boldsymbol{\alpha}^{(s)})\tilde{\boldsymbol{W}}_n \boldsymbol{U}_n(\boldsymbol{\beta}, \boldsymbol{\alpha}^{(s)}). \tag{5.12}$$

2. *Update*

$$\boldsymbol{\alpha}^{(s+1)} = \arg\min_{\boldsymbol{\alpha}\in\mathcal{M}} \frac{1}{N}\sum_{n=1}^{N} \boldsymbol{U}_n^{\mathrm{T}}(\boldsymbol{\beta}^{(s+1)}, \boldsymbol{\alpha})\tilde{\boldsymbol{W}}_n \boldsymbol{U}_n(\boldsymbol{\beta}^{(s+1)}, \boldsymbol{\alpha}). \tag{5.13}$$

3. *Update* $s = s+1$*. The iteration stops, when*

$$\|\boldsymbol{\alpha}^{(s)} - \boldsymbol{\alpha}^{(s+1)}\|_2^2 + \|\boldsymbol{\beta}^{(s)} - \boldsymbol{\beta}^{(s+1)}\|_2^2 < \epsilon,$$

where ϵ *is a small positive number.*

The optimization problem (5.12) in Step 2 is a regular minimization problem, which is equivalent to solving the equation

$$\frac{1}{N}\sum_{n=1}^{N} \left(\frac{\partial}{\partial\boldsymbol{\beta}}\boldsymbol{U}_n(\boldsymbol{\beta}, \boldsymbol{\alpha}^{(s)})\right)^{\mathrm{T}} \tilde{\boldsymbol{W}}_n \boldsymbol{U}_n(\boldsymbol{\beta}, \boldsymbol{\alpha}^{(s)}) = 0,$$

by the Newton-Raphson method. On the other hand, given $\boldsymbol{\beta}^{(s+1)}$, the objective function of the optimization problem (5.13) is convex with respect to $\boldsymbol{\alpha}$ and can be solved by a modified version of the CNM for GMM in Algorithm 4.1.

Algorithm 5.2 (The CNM for GLMM)**.**
Set $s = 0$ *and given* $\boldsymbol{\beta}$*. From an initial estimate* $Q^{(0)}$ *with finite support* $\Theta^{(0)}$ *and* $\boldsymbol{\alpha}^{(0)} = \int_{\mathcal{B}_2} \boldsymbol{P}(\theta)\mathrm{d}Q^{(0)}$*, repeat the following steps:*

1. *Compute all the local minimas* $\{\theta_j^{(s)}\}_{j=1}^{r^{(s)}}$ *of the function*

$$\mathcal{D}(b) = \frac{1}{N}\sum_{n=1}^{N} \left(\boldsymbol{\alpha}^{(s)} - \boldsymbol{P}(b)\right)^{\mathrm{T}} \boldsymbol{\Phi}_n(\boldsymbol{\beta})\tilde{\boldsymbol{W}}_n \left(\boldsymbol{Y}_n - \boldsymbol{\Phi}_n^{\mathrm{T}}\boldsymbol{\alpha}^{(s)}\right)$$

over $\mathcal{B}_2$*. The iteration stops if the minimum of* $\mathcal{D}(b)$ *is zero.*

2. *Construct a set of candidate support points by*

$$\Theta^{(s),+} = \Theta^{(s)} \cup \{\theta_j^{(s)}\}_{j=1}^{r(s)}.$$

Let $r^{(s),+}$ *be the number of elements in* $\Theta^{(s),+}$.

3. *Solve the optimization problem*

$$\begin{aligned} \min \quad & \frac{1}{N}\sum_{n=1}^{N}\left(\boldsymbol{Y}_n - \sum_{i=1}^{r^{(s)}+1}\pi_i \boldsymbol{\Phi}_n^{\mathrm{T}}\boldsymbol{P}(b_i)\right)^{\mathrm{T}} \tilde{\boldsymbol{W}}_n \left(\boldsymbol{Y}_n - \sum_{i=1}^{r^{(s)}+1}\pi_i \boldsymbol{\Phi}_n^{\mathrm{T}}\boldsymbol{P}(b_i)\right) \\ s.t. \quad & \sum_{i=1}^{r^{(s),+}} \pi_i = 1, \\ & \pi_i \geq 0, \quad i = 1, \ldots, r^{(s),+}, \end{aligned}$$

where $b_i \in \Theta^{(s),+}$. *We denote its solution by* $\boldsymbol{\pi}^{(s)} = (\pi_1^{(s)}, \ldots, \pi_{r^{(s),+}}^{(s)})^{\mathrm{T}}$.

4. *Discard all* b_i*s with zero* $\pi_i^{(s)}$, *update* $Q^{(s)}$, $\Theta^{(s)}$ *and* $\boldsymbol{\alpha}^{(s)} = \boldsymbol{\alpha}(Q^{(s)})$, *and set* $s = s + 1$.

5.6 Residual Analysis and Correlation Structure Estimation

The adequacy of a fitted regression model can be assessed using residual analysis; see [Fitzmaurice et al., 2012, p.g. 267]. In this section, we discuss the analysis of transformed residuals and give two possible ways of estimating the correlation structure of the residuals.

Given the GMM estimates $(\hat{\boldsymbol{\beta}}_{\mathrm{GMM}}, \hat{\boldsymbol{\alpha}}_{\mathrm{GMM}})$, for each $n \in \{1, \ldots, N\}$, the fitted mean is $\boldsymbol{\Phi}_n^{\mathrm{T}}(\hat{\boldsymbol{\beta}}_{\mathrm{GMM}})\hat{\boldsymbol{\alpha}}_{\mathrm{GMM}} \in \mathbb{R}^{T_n}$ and the residual is

$$\hat{\boldsymbol{r}}_n = \boldsymbol{Y}_n - \boldsymbol{\Phi}_n^{\mathrm{T}}(\hat{\boldsymbol{\beta}}_{\mathrm{GMM}})\hat{\boldsymbol{\alpha}}_{\mathrm{GMM}} \in \mathbb{R}^{T_n}.$$

However, because the elements of $\hat{\boldsymbol{r}}_n$ are correlated and with different variances, we need to standardized them so that they have constant variance and zero correlation.

Let $\boldsymbol{\Sigma}_n(\boldsymbol{\beta}^*, Q^*)$ be the covariance matrix of $\boldsymbol{Y}_n \mid (\boldsymbol{X}_n, \boldsymbol{Z}_n)$, which is positive definite. The residual $\hat{\boldsymbol{r}}_n$ can be standardized by

$$\hat{\boldsymbol{e}}_n = \boldsymbol{\Sigma}_n^{-1/2}(\boldsymbol{\beta}^*, Q^*)\hat{\boldsymbol{r}}_n.$$

Then, the classical residual diagnostics for standard linear regression can be applied; see [Fitzmaurice et al., 2012, p.g. 267]. Note that it is not necessary to check the normality of the standardized residuals, because no distributional assumption is made for the residuals.

There are two possible ways to estimate the covariance matrix $\boldsymbol{\Sigma}_n(\boldsymbol{\beta}^*, Q^*)$: the sample average and the parametric version. The sample average version is suitable to a balanced design such that $T_n = T$. The covariance matrix can be estimated as

$$\hat{\boldsymbol{\Sigma}}_n = \hat{\boldsymbol{V}}_n^{1/2} \left(\frac{1}{N} \sum_{n=1}^{N} \hat{\boldsymbol{V}}_n^{-1/2} \hat{\boldsymbol{r}}_n \hat{\boldsymbol{r}}_n^{\mathrm{T}} \hat{\boldsymbol{V}}_n^{-1/2} \right) \hat{\boldsymbol{V}}_n^{1/2}, \tag{5.14}$$

where $\hat{\boldsymbol{V}}_n = \boldsymbol{V}_n(\hat{\boldsymbol{\beta}}_{\mathrm{GMM}}, \hat{\boldsymbol{\alpha}}_{\mathrm{GMM}})$ and $\boldsymbol{V}_n(\boldsymbol{\beta}, \boldsymbol{\alpha})$ is defined in Equation (5.10); see [Fitzmaurice et al., 2012, p.g. 357].

In the parametric version, we approximately estimate $\boldsymbol{\Sigma}_n(\boldsymbol{\beta}^*, Q^*)$ under an additional assumption that, for each n and $t, t' = 1, \ldots, T_n$,

$$g^{-1}(\boldsymbol{X}_{nt}^{\mathrm{T}}\boldsymbol{\beta} + Z_{nt}b) \times g^{-1}(\boldsymbol{X}_{nt'}^{\mathrm{T}}\boldsymbol{\beta} + Z_{nt'}b) \in L^2(\mathcal{B}, \mu).$$

By the law of total covariance, we have

$$\begin{aligned}
&\mathrm{Cov}[Y_{nt}, Y_{nt'} \mid \boldsymbol{X}_n, \boldsymbol{Z}_n] \\
&= \mathbb{E}_{b_n}\left[\mathrm{Cov}[Y_{nt}, Y_{nt'} \mid \boldsymbol{X}_n, \boldsymbol{Z}_n, b_n]\right] \\
&\quad + \mathrm{Cov}_{b_n}[\mathbb{E}\left[Y_{nt} \mid \boldsymbol{X}_n, \boldsymbol{Z}_n, b_n\right] \times \mathbb{E}\left[Y_{nt'} \mid \boldsymbol{X}_n, \boldsymbol{Z}_n, b_n\right]].
\end{aligned}$$

By the modelling assumption that Y_{nt} and $Y_{nt'}$ are independent conditional on b_n, $\mathrm{Cov}[Y_{nt}, Y_{nt'} \mid \boldsymbol{X}_n, \boldsymbol{Z}_n, b_n] = 0$. We further apply the law of expectation and have

$$\begin{aligned}
&\mathrm{Cov}[Y_{nt}, Y_{nt'} \mid \boldsymbol{X}_n, \boldsymbol{Z}_n] \\
&= \mathbb{E}_{b_n}\left[\mathbb{E}\left[Y_{nt} \mid \boldsymbol{X}_n, \boldsymbol{Z}_n, b_n\right] \times \mathbb{E}\left[Y_{nt'} \mid \boldsymbol{X}_n, \boldsymbol{Z}_n, b_n\right]\right] \\
&\quad - \mathbb{E}\left[Y_{nt} \mid \boldsymbol{X}_n, \boldsymbol{Z}_n\right] \times \mathbb{E}\left[Y_{nt'} \mid \boldsymbol{X}_n, \boldsymbol{Z}_n\right].
\end{aligned}$$

For each t and t', by changing the order of the integrals, we have the approximations

$$\mathbb{E}_{b_n}\left[\mathbb{E}\left[Y_{nt} \mid \boldsymbol{X}_n, \boldsymbol{Z}_n, b_n\right] \times \mathbb{E}\left[Y_{nt'} \mid \boldsymbol{X}_n, \boldsymbol{Z}_n, b_n\right]\right] \approx \sum_{j=0}^{J_N} c_{ntt'j}(\boldsymbol{\beta})\alpha_j,$$

and

$$\mathbb{E}\left[Y_{nt} \mid \boldsymbol{X}_n, \boldsymbol{Z}_n\right] \times \mathbb{E}\left[Y_{nt'} \mid \boldsymbol{X}_n, \boldsymbol{Z}_n\right]] \approx \left(\sum_{j=0}^{J_N} \phi_{ntj}(\boldsymbol{\beta})\alpha_j\right) \times \left(\sum_{j=0}^{J_N} \phi_{nt'j}(\boldsymbol{\beta})\alpha_j\right),$$

where, for each j,

$$c_{ntt'j}(\boldsymbol{\beta}) = \int_{\mathcal{B}} g^{-1}(\boldsymbol{X}_{nt}^{\mathrm{T}}\boldsymbol{\beta} + Z_{nt}b) \times g^{-1}(\boldsymbol{X}_{nt'}^{\mathrm{T}}\boldsymbol{\beta} + Z_{nt'}b)P_j(b)\mathrm{d}\mu. \tag{5.15}$$

In sum, the off-diagonal elements of $\boldsymbol{\Sigma}_n$ are estimated by

$$\begin{aligned} &\tilde{\Sigma}_{ntt'}(\hat{\boldsymbol{\beta}}_{\mathrm{GMM}}, \hat{\boldsymbol{\alpha}}_{\mathrm{GMM}}) \\ =& \sum_{j=0}^{J_N} c_{ntt'j}(\hat{\boldsymbol{\beta}}_{\mathrm{GMM}})\hat{\alpha}_{\mathrm{GMM},j} \\ &- \left(\sum_{j=0}^{J_N} \phi_{ntj}(\hat{\boldsymbol{\beta}}_{\mathrm{GMM}})\hat{\alpha}_{\mathrm{GMM},j}\right) \times \left(\sum_{j=0}^{J_N} \phi_{nt'j}(\hat{\boldsymbol{\beta}}_{\mathrm{GMM}})\hat{\alpha}_{\mathrm{GMM},j}\right), \end{aligned}$$

while the diagonal elements of $\boldsymbol{\Sigma}_n$ are estimated by $V_{nt}(\hat{\boldsymbol{\beta}}_{\mathrm{GMM}}, \hat{\boldsymbol{\alpha}}_{\mathrm{GMM}})$ in Equation (5.10). One possible issue of using the parametric version is that the resulting estimated covariance matrix may not be positive definite.

5.7 Simulation Studies

To evaluate the performance of the GMM estimator $(\hat{\boldsymbol{\beta}}_{\mathrm{GMM}}, \hat{\boldsymbol{\alpha}}_{\mathrm{GMM}})$, we consider the following models.

Model 5.A (A Poisson Regression Model with a Log-link Function)**.**
For each $n \in \{1, \ldots, N\}$ and $t \in \{1, \ldots, T_n\}$, the response $Y_{nt} \mid (\boldsymbol{X}_{nt}, Z_{nt}, b_n)$ follows a Poisson distribution with mean $\mu_{nt}(b_n)$, where $\mu_{nt}(b_n)$ depends on the regression parameter $\boldsymbol{\beta}$ via the log-link function

$$\mu_{nt}(b_n) = \mathbb{E}\left[Y_{nt} \mid \boldsymbol{X}_{nt}, Z_{nt}, b_n\right] = \exp(\boldsymbol{X}_{nt}^{\mathrm{T}}\boldsymbol{\beta} + Z_{nt}b_n).$$

Model 5.B (A Binomial Regression Model with a Logit-link Function)**.**
For each $n \in \{1, \ldots, N\}$ and $t \in \{1, \ldots, T_n\}$, the response $Y_{nt} \mid (\boldsymbol{X}_{nt}, Z_{nt}, b_n)$ follows a binomial distribution with the number of trials $B = 20$ and the mean $\mu_{nt}(b_n)$, where $\mu_{nt}(b_n)$ depends on the regression parameter $\boldsymbol{\beta}$ via the logit-link function

$$\mu_{nt}(b_n) = \mathbb{E}[Y_{nt} \mid \boldsymbol{X}_{nt}, Z_{nt}, b_n] = \frac{B}{1 + \exp(-\boldsymbol{X}_{nt}^{\mathrm{T}}\boldsymbol{\beta} - Z_{nt}b_n)}.$$

For each n, let T_n follow a discrete uniform distribution over $\{1, \ldots, 5\}$. For each n and t, let $\boldsymbol{X}_{nt} = (X_{nt1}, X_{nt2}, X_{nt3}, X_{nt4})^{\mathrm{T}} \in \mathbb{R}^4$ be the fixed effects, where X_{nt1} and X_{nt2} independently follow a continuous uniform distribution over $[-0.3, 0.3]$, X_{nt3} follows a Bernoulli distribution with success probability 0.5 and $X_{nt4} = 10 \times X_{nt1}X_{nt2}$ is considered as the interaction effects of X_{nt1} and X_{nt2}. For each n and t, $Z_{nt} = t/20$. The true value of the regression parameter $\boldsymbol{\beta}$ is $(-1, 2, 0.5, 0)^{\mathrm{T}} \in \mathbb{R}^4$. The distribution of the random effects $Q(b)$ is $0.4I(b \leq 0) + 0.1I(b \leq 1) + 0.5I(b \leq 2)$.

We use the Chebyshev polynomials (see Definition 2.4.2) defined on $\mathcal{B} = [-6, 6]$ as the orthonormal basis $\{P_j(b)\}_{j=0}^{J_N}$ in $L^2(\mathcal{B}, \mu)$, where $\mu = (1 - b^2)^{-1/2}\mathrm{d}b$. The approximation property has been studied in Section 2.4.1. For different sample sizes, the dimensions of the generalized moments $\boldsymbol{\alpha} \in \mathbb{R}^{J_N}$ are different, where $J_N = \lfloor 2N^{1/3} \rfloor$, with $\lfloor a \rfloor$ denoting the largest integer not greater than a. Three sample size levels are considered ($N = 50, 100$ and 200).

We consider two different working correlation matrices: the independence and the AR(1). The parameter in the AR(1) correlation matrix is 0.5. We also consider the case when the weighting matrix is the inverse of the true correlation matrix of $\boldsymbol{U}_n(\boldsymbol{\beta}, \boldsymbol{\alpha})$.

We compare the GMM estimator with the NPMLE in [Wang, 2010], as the NPMLE is considered as the most efficient estimator for the mixed-effects model with univariate random effects. To study the robustness of the GMM estimator to the misspecification of the likelihood function. We also use the following misspecified models to fit the simulated data. Model 5.C is used to fit the data from Model 5.A, and Model 5.D to fit the data from Model 5.B. Under our parameter setting, all of the considered models are well-defined.

Model 5.C (A Binomial Regression Model with a Log-link Function)**.**
For each $n \in \{1, \ldots, N\}$ and $t \in \{1, \ldots, T_n\}$, the response $Y_{nt} \mid (\boldsymbol{X}_{nt}, Z_{nt}, b_n)$ follows a binomial distribution with the number of trials $B = 20$ and the mean $\mu_{nt}(b_n)$, where $\mu_{nt}(b_n)$ depends on the regression parameter $\boldsymbol{\beta}$ via the log-link function

$$\mu_{nt}(b_n) = \mathbb{E}[Y_{nt} \mid \boldsymbol{X}_{nt}, Z_{nt}, b_n] = \exp(\boldsymbol{X}_{nt}^{\mathrm{T}}\boldsymbol{\beta} + Z_{nt}b_n).$$

Model 5.D (A Poisson Regression Model with a Logit-link Function)**.**
For each $n \in \{1, \ldots, N\}$ and $t \in \{1, \ldots, T_n\}$, the response $Y_{nt} \mid (\boldsymbol{X}_{nt}, Z_{nt}, b_n)$ follows a Poisson distribution with mean $\mu_{nt}(b_n)$, where $\mu_{nt}(b_n)$ depends on the regression parameter $\boldsymbol{\beta}$ via the logit-link function

$$\mu_{nt}(b_n) = \mathbb{E}\left[Y_{nt} \mid \boldsymbol{X}_{nt}, Z_{nt}, b_n\right] = \frac{B}{1 + \exp(-\boldsymbol{X}_{nt}^{\mathrm{T}}\boldsymbol{\beta} - Z_{nt}b_n)}$$

and $B = 20$.

The simulation results are summarized in Table 5.1 to 5.4. From these tables, we have the following observations.

1. Although the correlation matrices are misspecified, the GMM estimators could perform closely to the NPMLE; see Table 5.1 to 5.4. This implies that the lose of information is not significant in this simulation sutdy.

2. In general, the NPMLE has smaller MSE than the GMM estimators; see Table 5.1 to 5.4. This is because that the maximum likelihood estimator is efficient in general.

3. When the regression parameter is of interest, the GMM estimators could perform closely to the NPMLE; see Table 5.1 and 5.3.

4. The MSE of the GMM estimators for the generalized moments $\boldsymbol{\alpha}$ are much larger than the ones of the NPMLE; see Table 5.2 and 5.4. The reason is that the GMM estimators use the modelling information from the marginal mean, while the NPMLE use the modelling information from the conditional probability function.

				β_1		β_2		β_3		β_4	
	Model	N	J_N	BIAS	MSE	BIAS	MSE	BIAS	MSE	BIAS	MSE
GMM	Inv	50	7	-0.014	0.184	0.054	0.175	0.020	0.015	-0.002	0.062
		100	9	-0.006	0.088	0.024	0.084	0.017	0.008	-0.003	0.029
		200	11	0.004	0.042	0.010	0.041	0.013	0.004	-0.002	0.013
	Indep	50	7	0.003	0.179	0.004	0.165	0.011	0.014	-0.006	0.058
		100	9	0.004	0.089	0.001	0.086	0.012	0.007	-0.004	0.028
		200	11	0.010	0.043	-0.002	0.042	0.010	0.004	-0.001	0.013
	AR(1)	50	7	-0.017	0.218	0.013	0.221	0.008	0.019	-0.005	0.074
		100	9	-0.012	0.110	0.004	0.110	0.011	0.010	-0.003	0.035
		200	11	0.001	0.053	-0.002	0.050	0.010	0.005	-0.001	0.017
NPMLE	5.A	50	-	-0.001	0.180	0.025	0.170	-0.005	0.013	-0.007	0.060
		100	-	-0.002	0.086	0.009	0.081	-0.001	0.007	-0.006	0.028
		200	-	0.008	0.041	-0.000	0.040	0.001	0.003	-0.003	0.013
	5.C	50	-	-0.005	0.176	0.032	0.164	-0.004	0.013	-0.003	0.057
		100	-	-0.002	0.084	0.008	0.078	0.000	0.007	-0.004	0.027
		200	-	0.008	0.041	0.001	0.040	0.002	0.003	-0.003	0.013

Table 5.1: Simulation results of the regression parameter $\boldsymbol{\beta}$, when the random sample is from Model 5.A and $N = 50, 100$ and 200. The working correlation matrices used in the GMM include the inverse of the true correlation matrix of $\boldsymbol{U}_n(\boldsymbol{\beta}, \boldsymbol{\alpha})$ (Inv), the independence (Indep) and the first-order autoregressive (AR(1)) correlation matrix. The likelihoods in the NPMLE are constructed using Model 5.A and 5.C.

			α_1		α_2		α_3		α_4		α_5	
	Model	N	BIAS	MSE	BIAS	MSE	BIAS	MSE	BIAS	MSE	BIAS	MSE
GMM	Inv	50	-0.148	0.072	0.482	0.660	0.089	0.296	-0.056	0.177	-0.287	0.430
		100	-0.112	0.043	0.374	0.439	0.063	0.280	-0.065	0.145	-0.226	0.387
		200	-0.078	0.022	0.256	0.222	0.062	0.271	-0.030	0.126	-0.149	0.282
	Indep	50	-0.124	0.063	0.394	0.536	0.060	0.251	-0.035	0.166	-0.234	0.392
		100	-0.096	0.038	0.321	0.385	0.032	0.231	-0.035	0.127	-0.167	0.349
		200	-0.068	0.020	0.227	0.210	0.021	0.208	-0.017	0.126	-0.101	0.258
	AR(1)	50	-0.134	0.068	0.380	0.512	0.139	0.308	-0.037	0.183	-0.259	0.408
		100	-0.101	0.040	0.300	0.358	0.105	0.272	-0.019	0.151	-0.200	0.352
		200	-0.074	0.021	0.215	0.197	0.082	0.245	0.004	0.129	-0.135	0.269
NPMLE	5.A	50	-0.017	0.013	0.061	0.020	0.027	0.080	-0.052	0.103	-0.050	0.152
		100	-0.012	0.008	0.043	0.010	0.010	0.048	-0.037	0.065	-0.003	0.103
		200	-0.008	0.004	0.029	0.005	-0.003	0.026	-0.017	0.036	0.031	0.062
	5.C	50	-0.026	0.015	0.087	0.025	0.059	0.084	-0.086	0.109	-0.123	0.174
		100	-0.020	0.008	0.069	0.014	0.050	0.050	-0.078	0.070	-0.096	0.119
		200	-0.017	0.004	0.057	0.008	0.043	0.027	-0.064	0.040	-0.082	0.073

Table 5.2: Simulation results of the five generalized moments $\{\alpha_1, \ldots, \alpha_5\}$, when the random sample is from Model 5.A and $N = 50, 100$ and 200. The working correlation matrices used in the GMM include the inverse of the true correlation matrix of $\boldsymbol{U}_n(\boldsymbol{\beta}, \boldsymbol{\alpha})$ (Inv), the independence (Indep) and the first-order autoregressive (AR(1)) correlation matrix. The likelihoods in the NPMLE are constructed using Model 5.A and 5.C.

				β_1		β_2		β_3		β_4	
	Model	N	J_N	BIAS	MSE	BIAS	MSE	BIAS	MSE	BIAS	MSE
GMM	Inv	50	7	-0.038	0.061	0.070	0.068	0.018	0.007	0.003	0.020
		100	9	-0.032	0.029	0.054	0.034	0.014	0.003	0.006	0.009
		200	11	-0.020	0.014	0.034	0.017	0.008	0.002	0.004	0.005
	Indep	50	7	-0.037	0.061	0.069	0.068	0.017	0.007	0.005	0.020
		100	9	-0.032	0.029	0.054	0.035	0.013	0.003	0.006	0.009
		200	11	-0.019	0.014	0.033	0.017	0.008	0.002	0.004	0.005
	AR(1)	50	7	-0.046	0.071	0.088	0.089	0.014	0.007	0.004	0.024
		100	9	-0.038	0.033	0.067	0.045	0.011	0.004	0.005	0.011
		200	11	-0.026	0.017	0.040	0.022	0.007	0.002	0.005	0.006
NPMLE	5.B	50	-	-0.008	0.056	0.004	0.059	0.006	0.006	0.008	0.019
		100	-	-0.008	0.026	0.006	0.027	0.006	0.003	0.005	0.008
		200	-	-0.001	0.013	-0.003	0.013	0.002	0.001	0.004	0.005
	5.D	50	-	-0.022	0.068	0.034	0.057	-0.013	0.008	0.003	0.021
		100	-	-0.017	0.051	0.031	0.043	-0.008	0.004	0.011	0.015
		200	-	-0.004	0.042	0.017	0.041	-0.010	0.003	0.007	0.012

Table 5.3: Simulation results of the regression parameter $\boldsymbol{\beta}$, when the random sample is from Model 5.B and $N = 50, 100$ and 200. The working correlation matrices used in the GMM include the inverse of the true correlation matrix of $\boldsymbol{U}_n(\boldsymbol{\beta}, \boldsymbol{\alpha})$ (Inv), the independence (Indep) and the first-order autoregressive (AR(1)) correlation matrix. The likelihoods in the NPMLE are constructed using Model 5.B and 5.D.

			α_1		α_2		α_3		α_4		α_5	
	Model	N	BIAS	MSE	BIAS	MSE	BIAS	MSE	BIAS	MSE	BIAS	MSE
GMM	Inv	50	0.009	0.006	0.490	0.613	0.257	0.408	-0.058	0.206	-0.207	0.295
		100	0.007	0.003	0.375	0.383	0.209	0.351	-0.050	0.201	-0.172	0.257
		200	0.006	0.002	0.280	0.221	0.150	0.286	-0.021	0.170	-0.111	0.200
	Indep	50	0.009	0.006	0.492	0.616	0.256	0.409	-0.060	0.206	-0.209	0.301
		100	0.007	0.003	0.378	0.386	0.215	0.358	-0.054	0.202	-0.172	0.257
		200	0.007	0.002	0.282	0.222	0.150	0.289	-0.023	0.172	-0.114	0.202
	AR(1)	50	0.014	0.007	0.520	0.683	0.242	0.380	-0.073	0.219	-0.210	0.327
		100	0.011	0.004	0.402	0.425	0.204	0.358	-0.090	0.219	-0.177	0.278
		200	0.008	0.002	0.305	0.258	0.121	0.268	-0.069	0.206	-0.139	0.248
NPMLE	5.B	50	-0.002	0.005	0.022	0.006	0.030	0.034	-0.008	0.046	-0.031	0.072
		100	-0.003	0.003	0.010	0.003	0.022	0.017	0.012	0.026	-0.012	0.036
		200	-0.001	0.001	0.007	0.001	0.011	0.008	0.009	0.013	-0.001	0.020
	5.D	50	0.043	0.010	0.030	0.014	-0.100	0.044	-0.084	0.086	0.122	0.055
		100	0.033	0.005	0.009	0.004	-0.098	0.032	-0.037	0.039	0.158	0.058
		200	0.038	0.004	0.005	0.002	-0.120	0.032	-0.027	0.025	0.214	0.077

Table 5.4: Simulation results of the five generalized moments $\{\alpha_1, \dots, \alpha_5\}$, when the random sample is from Model 5.B and $N = 50, 100$ and 200. The working correlation matrices used in the GMM include the inverse of the true correlation matrix of $\boldsymbol{U}_n(\boldsymbol{\beta}, \boldsymbol{\alpha})$ (Inv), the independence (Indep) and the first-order autoregressive (AR(1)) correlation matrix. The likelihoods in the NPMLE are constructed using Model 5.B and 5.D.

We also can see the robustness of the GMM estimators to the misspecified likelihoods from the simulation results. When Model 5.C is used to fit a random sample from Model 5.A, the NPMLE performs as well as when Model 5.A is used; see Table 5.1 and 5.2. This is because that a non-parametric mixture of binomial distributions with large number of trials could appropriately approximate any discrete probability distributions; see [Wood, 1999]. On the other hand, when a random sample from Model 5.B is fitted by Model 5.D, the NPMLE performs worse than the GMM estimators, especially when the sample size is large; see Table 5.3 and 5.4. Note that a Poisson distribution can be used to approximate a binomial distribution with large number of trials when the success probability is either close to zero or one. In Figure 5.1, we show the simulated success probability when $N = 1000$, and see that few of the simulated success probabilities is close to zero or one. Therefore, it is inappropriate to use the mixture of Poisson distribution to approximate a binomial distribution in our simulation setting.

5.8 Application to the Retina Surgery Data

The retina surgery data has been analyzed in [Song and Tan, 2000] and [Qiu et al., 2008]. Let Y_{nt} be the percentage of gas volume for the n^{th} patient at time t_n and let $\boldsymbol{X}_{nt}$ be the vector of covariates including the logarithm of time after surgery (TIME) and its square, and the gas concentration level (LEVEL). The following model is used under the assumptions that $\boldsymbol{Y}_n \mid (\boldsymbol{X}_n, \boldsymbol{Z}_n)$, $n = 1, \ldots, N$, are independent to each other and conditional on the random effects $\boldsymbol{b}_n$, $Y_{nt_n} \mid (\boldsymbol{X}_n, \boldsymbol{Z}_n, \boldsymbol{b}_n)$ are independent to each other.

Model 5.E.

For each n and t_n, $Y_{nt_n} \mid (\boldsymbol{X}_{nt_n}, b_n)$ follows a distribution with mean $\mu_{nt}(b_n)$ such that

$$\operatorname{logit}(\mu_{nt_n}(b_n)) = \beta_1 \log(\text{TIME}) + \beta_2 \left(\log(\text{TIME})\right)^2 + \beta_3 \text{LEVEL} + b_n,$$

where $b_n \in \mathbb{R}$ has a probability measure Q defined on $\mathcal{B} = [-20, 50]$. The variance of $Y_{nt} \mid (\boldsymbol{X}_{nt}, b_n)$ is

$$\operatorname{Var}\left[Y_{nt} \mid \boldsymbol{X}_{nt}, b_n\right] = \left((1 - \mu_{nt}(b_n))\mu_{nt}(b_n) + 1\right)^{-1}.$$

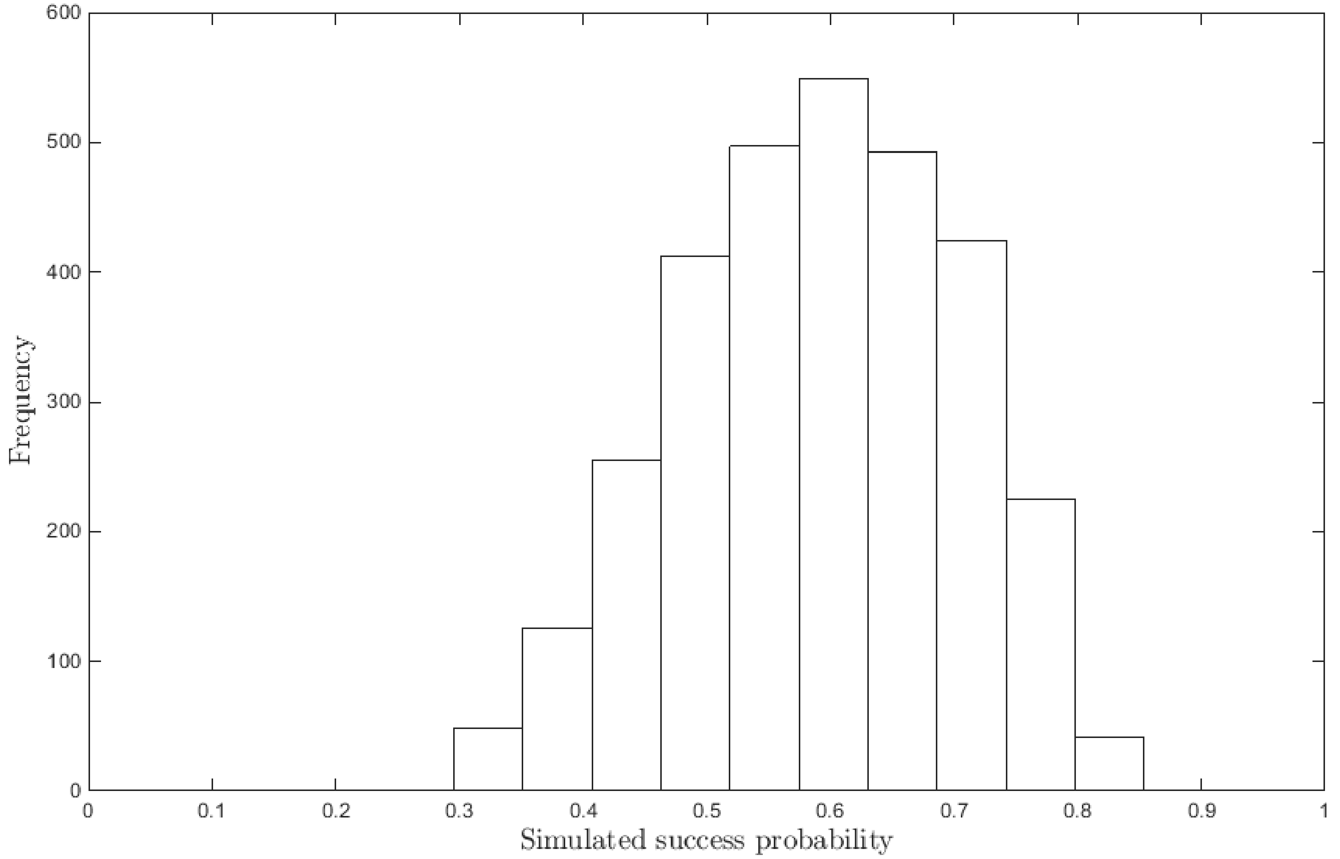

Figure 5.1: Histogram of the simulated random success probabilities, when $N = 1000$.

The Chebyshev polynomials defined on $\mathcal{B}$ is used to reparameterize and approximate Model 5.E. The numbers of generalized moments used in the truncation approximation models are $5, 7, 9$ and 11. We also consider two types of working correlation matrices: the independence and the AR(1). The parameter used in the AR(1) correlation matrix is 0.5. The estimated regression parameters are reported in Table 5.5. We also report the PQL estimates (with different approximation orders) from [Qiu et al., 2008], where a mixture of the simplex distribution is considered and the distribution of the random intercept b_n is normal. We notice that β_1 and β_2 could be estimated very differently by the two methods, because that two different models are used.

Panel (a) and (b) in Figure 5.2 display the standardized residuals against the responses and fitted means, when the model is fitted by the GMM with AR(1) correlation matrices and $J_N = 5$. The residuals are standardized by its working correlation matrix. The two red lines represent the 97.5% and 2.5% empirical quantiles of the standardized residuals. A linear trend is observed from Panel (a). It implies that some information in the residuals is not characterized by Model F. This is because that the working correlation matrices are misspecified. In Panel (b), we do not observe any pattern between the fitted mean and the standardized residuals. This implies that the correlation between them is small. Panel (c)-(e) in Figure 5.2 shows the fitted means and the proportions over time across three levels of gas concentration. We see that the fitted mean can successfully characterize the decaying trends of the proportions over time.

5.9 Discussion

In this chapter, we introduce the GMM for mixed-effects models with univariate random effects. By simulation, we see that the GMM estimator may not as efficient as the NPMLE but it is robust to the misspecified likelihood functions. As we will see in Chapter 6, it is challenging to evaluate the loss of efficiency. The major reason is that neither the GMM estimator nor the NPMLE has an explicit form of the covariance matrix due to the existence of the boundaries in the parameter space. In this section,

		J_N	log(TIME)	$(\log(\text{TIME}))^2$	LEVEL
GMM	AR(1)	5	0.46	-0.45	0.52
		7	0.38	-0.43	0.46
		9	0.32	-0.40	0.42
		11	0.41	-0.46	0.48
	Indep	5	0.31	-0.38	0.47
		7	0.45	-0.45	0.51
		9	0.45	-0.45	0.51
		11	0.44	-0.45	0.50
		order	log(TIME)	$(\log(\text{TIME}))^2$	LEVEL
PQL	-	1	0.06	-0.35	0.44
		2	0.05	-0.35	0.45
		4	0.14	-0.39	0.45
		6	0.14	-0.39	0.45

Table 5.5: Estimated regression parameters in Model 5.E to the retina surgery data

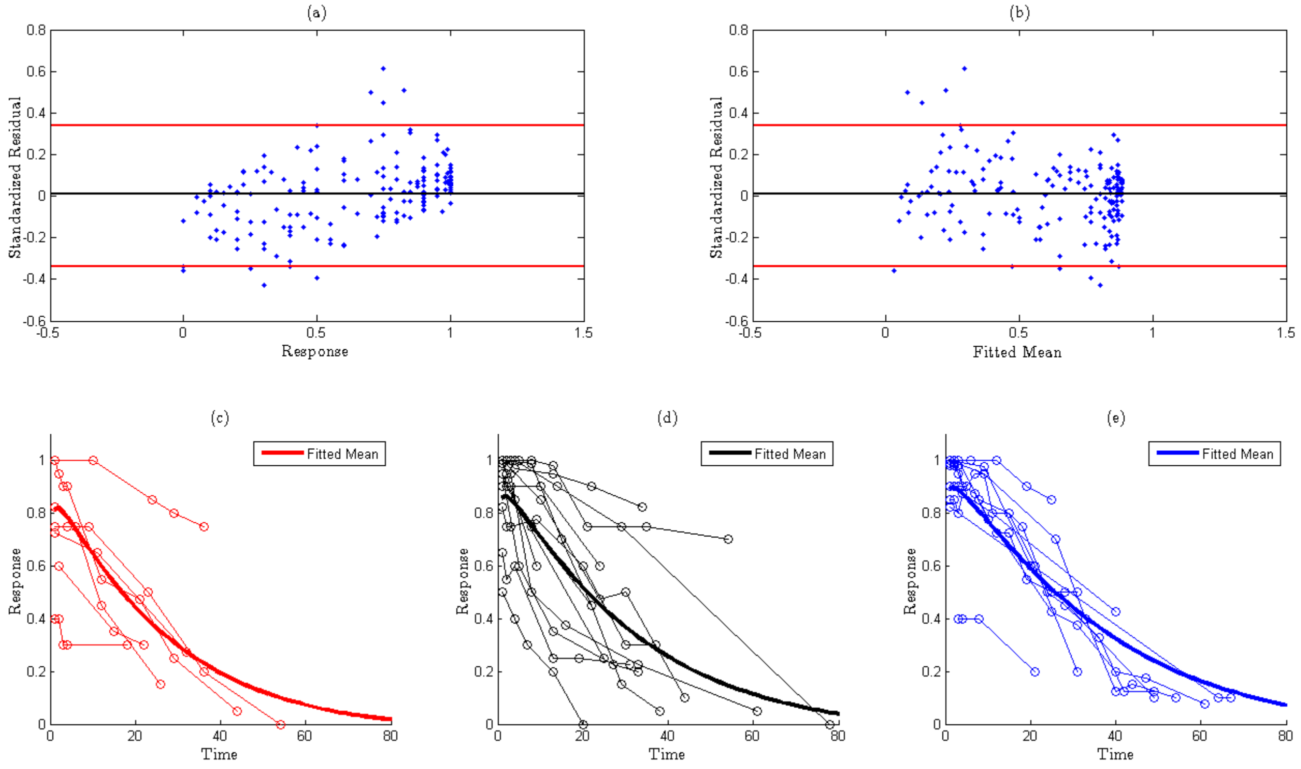

Figure 5.2: Plots of the residual analysis of the fitted Model 5.E to the epilepsy seizures data using the GMM with AR(1) working correlation matrix and $J_N = 5$.

we discuss the following possible future research direction.

Firstly, to study the subject-specific model, we need to predict random effects in some cases. Given the GMM estimator $(\hat{\boldsymbol{\beta}}_{\text{GMM}}, \hat{\boldsymbol{\alpha}}_{\text{GMM}})$, we may use the solution of the following optimization problem as the random effects predicator,

$$\begin{aligned} \min_{\{b_n\}_{n=1}^N} \quad & \frac{1}{N} \sum_{n=1}^{N} \boldsymbol{U}_n^{\mathrm{T}}(\hat{\boldsymbol{\beta}}_{\text{GMM}}, b_n) \boldsymbol{S}_n \boldsymbol{U}_n(\hat{\boldsymbol{\beta}}_{\text{GMM}}, b_n) \\ s.t. \quad & \frac{1}{N} \sum_{n=1}^{N} \boldsymbol{P}(b_n) = \hat{\boldsymbol{\alpha}}_{\text{GMM}}, \end{aligned} \tag{5.16}$$

where for each n, $\boldsymbol{S}_n$ is a $T_n \times T_n$ positive definite matrix, and

$$\boldsymbol{U}_n(\boldsymbol{\beta}, b_n) = \boldsymbol{Y}_n - g^{-1}(\boldsymbol{X}_n^{\mathrm{T}} \boldsymbol{\beta} + \boldsymbol{Z}_n b_n) \in \mathbb{R}^{T_n}.$$

The above optimization problem can be solved by the Lagrange multiplier method. However, the properties of the predicated random effects need further investigation.

In this thesis, we assume that the parameter σ is constant and known. This assumption is valid in the models considered in Section 5.7. An unknown or non-constant σ may lead to much more complex model. However, the GMM for σ requires future work.

Another important research direction is the extension to multivariate mixed-effects models. Recall that in Section 2.4, we have introduced the reparameterization-approximation procedure for the GLMM with multivariate random effects. Designing an efficient computational algorithm for multivariate mixed-effects models is challenge for two major reasons. Firstly, the Chebyshev system for multivariate functions is not well-defined. Secondly, few study has been done on the geometry of the generalized moment space for multivariate distributions. As a result, the positive representation and the gradient characterization, which are necessary for the gradient-based computational algorithms, are not established.

Lastly, we discuss the model selection problem. In the real examples, we considered difference combinations between the working correlation matrices and the number of the generalized moments. Different combination may lead to different point estimates in the GMM. It is natural to ask which fitted model to use. Also, the graphical analysis of the standardized residuals are subjective in this thesis. Numerical analysis of the residuals needs further investigation.

Appendix: D

D.1 MATLAB Code for Algorithm 5.1

```
function [betanew, as, out, objo] = GLM_GMM(DATA, q, ...
                    a0, beta0, V, W)


Ind1 = 0;
count1 = 1;
out = 1;
objo = 1e5;

while Ind1 == 1
    count1 = count1 + 1;

    [betanew, unused, H0, W] = GMM_NR(DATA, q,...
        a0, beta0, V);

    [asnew, obj] = GMM_CNM(DATA(:,end), q,...
        as,  H0, W);

    if count1 > 5e2
        Ind1 = 1;
        out = 0;
    end

    if  norm(betanew-beta0)<1e-5
        Ind1 = 1;
    else
        beta0 = betanew;
        as = asnew;
        objo = obj;
    end
end

function [betanew, obj, H0a, W, out] = GMML_NR(DATA, q,...
    aini, betaini, V, W)

% for logistic link functions

Ind = 0;
count = 1;
out = 1;
beta0 = betaini;
betap = length(beta0);
X = DATA(:,2:2+betap-1);
Z = DATA(:,2+betap);
S = DATA(:,end);
H0a = 1./(1+exp(-repmat(X*beta0,[1,length(q)])-Z*q));
H0 = H0a*(V*V');
U = H0*aini';
dU = (1-H0a).*H0a;
D = X'.*repmat(aini*(dU*(V*V'))',[size(X,2),1]);
C = D*W*(S-U);
G = D*W*D';
objini =(S-U)'*W*(S-U);

if( rcond(G) < 1e-5 )
    Ind = 1;
    out = 0;
    betanew = betaini;
    obj = objini;
end

while Ind == 0
    count = count + 1;
    betanew = beta0 + G\C;
    H0a = 1./(1+exp(-repmat(X*betanew,[1,length(q)])-Z*q));
    H0 = H0a*(V*V');
    U = H0*aini';
    dU = (1-H0a).*H0a;
    D = X'.*repmat(aini*(dU*(V*V'))',[size(X,2),1]);
    C = D*W*(S-U);
    G = D*W*D';
    obj =(S-U)'*W*(S-U);

    if obj < objini
        Ind = 1;
    else
        beta0 = betanew;
    end

    if count > 2e2
        betanew = betaini;
        obj = objini;
        Ind =1;
    end

end

H0a = 1./(1+exp(-repmat(X*betanew,[1,length(q)])-Z*q));
```

D.2 MATLAB Code for Algorithm 5.2

```
function [as, Dsmin] = GMM_CNM_adj(S, q, ainit, H0, W)

Ind = 0;
e = 1e-10;

pLs = find(ainit > 0);
as = ainit;

U = H0*as';

g = -(S-U)'*W*(H0-U(:,ones(1,length(q))));
Hc = H0'*W*H0;
Ac = H0'*W*S;
Dsmino = (S-U)'*W*(S-U);

while Ind == 0

    dg = diff(g);
    signdg = sign(dg);
    dsigndg = diff(signdg);
    minL = find(dsigndg == 2)+1;
    L = minL;
    pLsnew = [1 pLs L length(q)];
    pLsnew = unique(pLsnew);

    H = Hc(pLsnew, pLsnew);
    A = Ac(pLsnew,:);

    %warning off;
    options = optimset('Algorithm', ...
                             'interior-point-convex', ...
                             'display', 'off');

    Ast = quadprog((H+H')/2, -A, ...
                            -eye(size(H,2)), ...
                            zeros(1,size(H,2)), ...
                            ones(1,size(H,2)), 1, ...
                            [], [], [], options);

    as = zeros(1,length(as));
    as(pLsnew) = Ast;
    pLs = pLsnew(Ast > 0);

    U = H0*as';
    g = -(S-U)'*W*(H0-U(:,ones(1,length(q))));
    Dsmin = (S-U)'*W*(S-U);

    if abs(Dsmin-Dsmino) < 1e-5 || max(g) < e
        Ind = 1;
    else
        aso = as;
        Dsmino = Dsmin;
    end

end
```

Chapter 6

The Generalized Method of Moments for a Poisson Regression Model with Random Intercept and Slope

6.1 Introduction

In the previous chapter, we considered the case where the random effects in a generalized linear mixed model are univariate. Now, we consider the generalized method of moments for the following Poisson regression model.

Model 6.A.
For each $n \in \{1, \ldots, N\}$ and $t \in \{1, \ldots, T_n\}$, the response $Y_{nt} \mid (\boldsymbol{X}_{nt}, Z_{nt}, \boldsymbol{b}_n)$ follows a Poisson distribution with mean $\mu_{nt}(\boldsymbol{b}_n)$, where $\mu_{nt}(\boldsymbol{b}_n)$ depends on the regression parameter $\boldsymbol{\beta}$ via the log-link function

$$\log \mu_{nt}(\boldsymbol{b}_n) = \boldsymbol{X}_{nt}^{\mathrm{T}}\boldsymbol{\beta} + b_{n1} + Z_{nt} \times b_{n2},$$

and $\boldsymbol{X}_{nt} \in \mathbb{R}^p$ are the covariates for the fixed effects, $\boldsymbol{\beta} \in \mathbb{R}^p$ is the regression parameter, Z_{nt} are the covariates to the random effects and $\boldsymbol{b}_n = (b_{n1}, b_{n2})^{\mathrm{T}} \in \mathbb{R}^2$ are *random effects. Furthermore, b_{n1} and b_{n2} are assumed independent and their marginal distribution are Q_1 and Q_2 with the support sets $\mathcal{B}_1$ and $\mathcal{B}_2$ correspondingly.*

The main contribution in this chapter is to extend the GMM for mixed-effects models with univariate random effects to a Poisson regression model with random intercept and slope (Model 6.A). After the reparameterization-approximation procedure, we point out that the parameter space for Model 6.A is a generalized moment cone which share the same geometric properties as the generalized moment space; see Section 6.2. Therefore, the computational algorithms proposed in Chapter 5 can be easily modified to compute the GMM estimators for Model 6.A; see Section 6.3. The simulation studies in Section 6.4 provide empirical evidence that the GMM estimators in Model 6.A is consistency and is robust to the misspecification of the random-effects distribution. Also see Section 6.5 for a real data example.

We organize this chapter as follows. In Section 6.2, we describe the GMM for Model 6.A. In Section 6.3, we give the modified computational algorithms for Model 6.A. The finite sample performance of the GMM for Model 6.A is examined through simulations in Section 6.4. In Section 6.5, we fit the Epileptic Seizures Data by the proposed methods, which has been described in Section 1.6.2. Finally, we end this chapter with a discussion.

6.2 The Generalized Method of Moments

Let $\mathfrak{A}$ be the range of the function $\exp(b_{n1})$, for $b_{n1} \in \mathcal{B}_1$. By the assumption that b_{n1} and b_{n2} are independent, for each $n \in \{1, \dots, N\}$ and $t \in \{1, \dots, T_n\}$, we have

$$\begin{aligned}
\mathbb{E}\left[\boldsymbol{Y}_{nt} \mid \boldsymbol{X}_{nt}, Z_{nt}\right] &= \int_{\mathcal{B}_1}\int_{\mathcal{B}_2} \mu_{nt}(\boldsymbol{b}_n)\mathrm{d}Q_1\mathrm{d}Q_2 \\
&= \int_{\mathcal{B}_1}\int_{\mathcal{B}_2} \exp\left(\boldsymbol{X}_{nt}^{\mathrm{T}}\boldsymbol{\beta} + b_{n1} + Z_{nt}b_{n2}\right)\mathrm{d}Q_1\mathrm{d}Q_2 \\
&= \gamma_1 \times \int_{\mathcal{B}_2} \exp\left(Z_{nt}b_{n2}\right) \times \exp\left(\boldsymbol{X}_{nt}^{\mathrm{T}}\boldsymbol{\beta}\right)\mathrm{d}Q_2,
\end{aligned}$$

where

$$\gamma_1 = \int_{\mathcal{B}_1} \exp\left(b_{n1}\right)\mathrm{d}Q_1 \in \mathfrak{A}.$$

Assume that for each $n \in \{1, \ldots, N\}$ and $t \in \{1, \ldots, T_n\}$, the function of b_{n2}, $\exp(\boldsymbol{X}_{nt}^{\mathrm{T}}\boldsymbol{\beta} + Z_{nt}b_{n2})$, is an element of $L^2(\mathcal{B}, \mu)$. Given an orthonormal polynomial system $\{P_j(b)\}_{j=0}^{\infty}$ defined on $L^2(\mathcal{B}, \mu)$, we can reparameterize and approximate the expectation $\mathbb{E}\left[\boldsymbol{Y}_{nt} \mid \boldsymbol{X}_{nt}, Z_{nt}\right]$ as

$$\mathbb{E}\left[\boldsymbol{Y}_{nt} \mid \boldsymbol{X}_{nt}, Z_{nt}\right] \approx \gamma_1 \times \sum_{j=0}^{J_N} \phi_{ntj}(\boldsymbol{\beta})\alpha_j',$$

where for each $j \in \{0, \ldots, J_N\}$,

$$\phi_{ntj}(\boldsymbol{\beta}) = \int_{\mathcal{B}_2} \exp(Z_{nt}b_{n2}) \times \exp\left(\boldsymbol{X}_{nt}^{\mathrm{T}}\boldsymbol{\beta}\right) \times P_j(b_{n2})\mathrm{d}\mu$$

and

$$\alpha_j' = \int_{\mathcal{B}_2} P_j(b)\mathrm{d}Q_2.$$

For each n, let

$$\boldsymbol{U}_n(\boldsymbol{\beta}, \boldsymbol{\alpha}) = \boldsymbol{Y}_n - \boldsymbol{\Phi}_n^{\mathrm{T}}(\boldsymbol{\beta})\boldsymbol{\alpha},$$

where $\boldsymbol{\Phi}_n(\boldsymbol{\beta})$ is a $(J_N+1) \times T_n$ matrix whose elements are $\phi_{ntj}(\boldsymbol{\beta})$ and $\boldsymbol{\alpha} = (\alpha_0, \ldots, \alpha_{J_N})^{\mathrm{T}} \in \mathbb{R}^{J_N+1}$, for each $j \in \{1, \ldots, J_N\}$,

$$\alpha_j = \gamma_1 \times \alpha_j'.$$

The parameter space of $\boldsymbol{\alpha}$ is

$$\mathcal{C}(\mathfrak{A}) = \left\{\boldsymbol{\alpha} = \int_{\mathcal{B}_2} \boldsymbol{P}(b)\mathrm{d}Q' \in \mathbb{R}^{J_N}\right\},$$

where Q' is a nondecreasing right continuous function of bounded variation such that $\int_{\mathcal{B}_2} \mathrm{d}Q' \in \mathfrak{A}$ and $\boldsymbol{P}(b) = (P_0(b), \ldots, P_{J_N}(b))^{\mathrm{T}} \in \mathbb{R}^{J_N+1}$ is a vector function of $b \in \mathcal{B}_2$. The set $\mathcal{C}(\mathfrak{A})$ is known as a subset of the moment cone (see Definition 3.2.1). When $P_0(b) \equiv 1$, we have

$$\alpha_0 = \gamma_1 \times \int_{\mathcal{B}_2} \mathrm{d}Q_2 = \gamma_1.$$

On the other hand, by the law of total variance and the law of total expectation, we have

$$\begin{aligned}
&\operatorname{Var}[Y_{nt} \mid \boldsymbol{X}_{nt}, Z_{nt}] \\
=& \mathbb{E}_{\boldsymbol{b}_n}\left[\operatorname{Var}[Y_{nt} \mid \boldsymbol{X}_{nt}, Z_{nt}, \boldsymbol{b}_n]\right] + \operatorname{Var}_{\boldsymbol{b}_n}[\mathbb{E}\left[Y_{nt} \mid \boldsymbol{X}_{nt}, Z_{nt}, \boldsymbol{b}_n\right]] \\
=& \mathbb{E}_{\boldsymbol{b}_n}\left[\operatorname{Var}[Y_{nt} \mid \boldsymbol{X}_{nt}, Z_{nt}, \boldsymbol{b}_n]\right] + \mathbb{E}_{\boldsymbol{b}_n}\left[\left(\mathbb{E}\left[Y_{nt} \mid \boldsymbol{X}_{nt}, Z_{nt}, \boldsymbol{b}_n\right]\right)^2\right] \\
&- \left(\mathbb{E}\left[Y_{nt} \mid \boldsymbol{X}_{nt}, Z_{nt}\right]\right)^2.
\end{aligned}$$

Because $Y_{nt} \mid (\boldsymbol{X}_{nt}, Z_{nt}, \boldsymbol{b}_n)$ follows a Poisson distribution, we have

$$\operatorname{Var}[Y_{nt} \mid \boldsymbol{X}_{nt}, Z_{nt}, \boldsymbol{b}_n] = \mu_{nt}(\boldsymbol{b}_n).$$

By changing the order of the integrals, we can approximate the following terms as

$$\begin{aligned}
\mathbb{E}_{\boldsymbol{b}_n}\left[\operatorname{Var}[Y_{nt} \mid \boldsymbol{X}_{nt}, Z_{nt}, \boldsymbol{b}_n]\right] &= \int_{\mathcal{B}_1}\int_{\mathcal{B}_2} \mu_{nt}(\boldsymbol{b}_n) \mathrm{d}Q_1 \mathrm{d}Q_2 \\
&\approx \boldsymbol{\phi}_{nt}^{\mathrm{T}}(\boldsymbol{\beta})\boldsymbol{\alpha}, \\
\mathbb{E}_{\boldsymbol{b}_n}\left[\left(\mathbb{E}\left[Y_{nt} \mid \boldsymbol{X}_{nt}, Z_{nt}, \boldsymbol{b}_n\right]\right)^2\right] &= \gamma_2 \times \int_{\mathcal{B}_2} \left(\exp\left(\boldsymbol{X}_{nt}^{\mathrm{T}}\boldsymbol{\beta} + Z_{nt} b_{n2}\right)\right)^2 \mathrm{d}Q_2 \\
&\approx \gamma_2/\gamma_1 \times \boldsymbol{c}_{ntt}^{\mathrm{T}}(\boldsymbol{\beta})\boldsymbol{\alpha},
\end{aligned}$$

and

$$\left(\mathbb{E}\left[Y_{nt} \mid \boldsymbol{X}_{nt}, Z_{nt}\right]\right)^2 \approx \left(\boldsymbol{\phi}_{nt}^{\mathrm{T}}(\boldsymbol{\beta})\boldsymbol{\alpha}\right)^2,$$

where

$$\gamma_2 = \int_{\mathcal{B}_1} \exp(2b_{n1}) \mathrm{d}Q_1.$$

Here for each n and $t \in \{1, \ldots, T_n\}$,

$$\boldsymbol{\phi}_{nt}(\boldsymbol{\beta}) = (\phi_{nt0}(\boldsymbol{\beta}), \ldots, \phi_{ntJ_N}(\boldsymbol{\beta}))^{\mathrm{T}} \in \mathbb{R}^{J_N}$$

and

$$\boldsymbol{c}_{ntt}(\boldsymbol{\beta}) = (c_{ntt0}(\boldsymbol{\beta}), \ldots, c_{nttJ_N}(\boldsymbol{\beta}))^{\mathrm{T}} \in \mathbb{R}^{J_N},$$

where for each j,

$$c_{nttj}(\boldsymbol{\beta}) = \int_{\mathcal{B}_2} \left(\exp\left(\boldsymbol{X}_{nt}^{\mathrm{T}}\boldsymbol{\beta} + Z_{nt}b_{n2}\right)\right)^2 P_j(b_{n2})\mathrm{d}Q_2.$$

Therefore, the variance function of $Y_{nt} \mid (\boldsymbol{X}_{nt}, Z_{nt})$ is approximated by

$$V_{nt}(\boldsymbol{\beta}, \boldsymbol{\alpha}, \boldsymbol{\gamma}) = \boldsymbol{\phi}_{nt}^{\mathrm{T}}(\boldsymbol{\beta})\boldsymbol{\alpha} + \gamma_2/\gamma_1 \times \boldsymbol{c}_{ntt}^{\mathrm{T}}(\boldsymbol{\beta})\boldsymbol{\alpha} - \left(\boldsymbol{\phi}_{nt}^{\mathrm{T}}(\boldsymbol{\beta})\boldsymbol{\alpha}\right)^2, \tag{6.1}$$

and correspondingly, the adjusted approximation is

$$V_{\mathrm{adj},nt}(\boldsymbol{\beta}, \boldsymbol{\alpha}, \boldsymbol{\gamma}) = \max\left\{\epsilon, \boldsymbol{\phi}_{nt}^{\mathrm{T}}(\boldsymbol{\beta})\boldsymbol{\alpha}\right\} + \max\left\{\epsilon, \gamma_2/\gamma_1 \times \boldsymbol{c}_{ntt}^{\mathrm{T}}(\boldsymbol{\beta})\boldsymbol{\alpha} - \left(\boldsymbol{\phi}_{nt}^{\mathrm{T}}(\boldsymbol{\beta})\boldsymbol{\alpha}\right)^2\right\},$$

where $\boldsymbol{\gamma} = (\gamma_1, \gamma_2)^{\mathrm{T}} \in \mathbb{R}^2$ and ϵ is a small positive number; also see Section 5.4.1.

For each n, let $\boldsymbol{V}_n(\boldsymbol{\beta}, \boldsymbol{\alpha}, \boldsymbol{\gamma})$ be the $T_n \times T_n$ diagonal matrix whose diagonal elements are $V_{nt}(\boldsymbol{\beta}, \boldsymbol{\alpha}, \boldsymbol{\gamma})$, $t = 1, \ldots, T_n$. Given the initial estimators $(\tilde{\boldsymbol{\beta}}, \tilde{\boldsymbol{\alpha}}, \tilde{\boldsymbol{\gamma}})$, the GMM estimator for Model 6.A is

$$(\hat{\boldsymbol{\beta}}_{\mathrm{GMM}}, \hat{\boldsymbol{\alpha}}_{\mathrm{GMM}}) = \arg\min_{\boldsymbol{\beta}\in\mathbb{R}^p, \boldsymbol{\alpha}\in\mathcal{C}(\mathfrak{A})} \frac{1}{N}\sum_{n=1}^{N} \boldsymbol{U}_n^{\mathrm{T}}(\boldsymbol{\beta}, \boldsymbol{\alpha})\tilde{\boldsymbol{W}}_n\boldsymbol{U}_n(\boldsymbol{\beta}, \boldsymbol{\alpha}),$$

where for each n,

$$\tilde{\boldsymbol{W}}_n = \boldsymbol{V}_n^{-1/2}(\tilde{\boldsymbol{\beta}}, \tilde{\boldsymbol{\alpha}}, \tilde{\boldsymbol{\gamma}})\boldsymbol{R}_n^{-1}\boldsymbol{V}_n^{-1/2}(\tilde{\boldsymbol{\beta}}, \tilde{\boldsymbol{\alpha}}, \tilde{\boldsymbol{\gamma}}),$$

and $\boldsymbol{R}_n$ is the working correlation matrix.

The initial estimators of $(\boldsymbol{\alpha}, \boldsymbol{\beta}, \gamma_1)$ can be obtained by using fixed weighting matrices $\{\boldsymbol{W}_n\}_{n=1}^N$, when $P_0(b) \equiv 1$. To obtain the initial estimator of γ_2, we further assume that γ_2 is a function of γ_1. Because $\tilde{\gamma}_1 = \tilde{\alpha}_0$, we have $\tilde{\gamma}_2 = \gamma_2(\tilde{\gamma}_1)$. For example, when Q_1 is a normal distribution with mean zero and covariance σ^2, we have

$$\gamma_1 = \frac{1}{\sqrt{2\pi}\sigma}\int_{\mathbb{R}} \exp(b_{n1})\exp\left(-\frac{b_{n1}^2}{2\sigma^2}\right)\mathrm{d}b_{n1} = \exp\left(\sigma^2/2\right)$$

and

$$\gamma_2 = \frac{1}{\sqrt{2\pi}\sigma}\int_{\mathbb{R}} \exp(2b_{n1})\exp\left(-\frac{b_{n1}^2}{2\sigma^2}\right)\mathrm{d}b_{n1} = \exp\left(2\sigma^2\right).$$

Therefore, we have

$$\tilde{\gamma}_2 = \tilde{\gamma}_1^4.$$

In general, we do not need the distributional assumption of b_{n1} but a modelling assumption on γ_2 as a function of γ_1.

6.3 Computational Algorithms

To obtain the GMM estimators $(\hat{\boldsymbol{\beta}}_{\text{GMM}}, \hat{\boldsymbol{\alpha}}_{\text{GMM}})$ for Model 6.A, the following alternating parameter algorithm can be used.

Algorithm 6.1 (The Alternating Parameter Algorithm)**.**
Set $s = 0$. From an initial estimate $\boldsymbol{\alpha}^{(0)} \in \mathbb{R}^J$, repeat the following steps at the $(s+1)^{\text{th}}$ iteration:

1. *Given $\boldsymbol{\alpha}^{(s)}$, solve the optimization problem*

$$\boldsymbol{\beta}^{(s+1)} = \arg\min_{\boldsymbol{\beta}} \frac{1}{N} \sum_{n=1}^{N} \boldsymbol{U}_n^{\mathrm{T}}(\boldsymbol{\beta}, \boldsymbol{\alpha}^{(s)}) \tilde{\boldsymbol{W}}_n \boldsymbol{U}_n(\boldsymbol{\beta}, \boldsymbol{\alpha}^{(s)}). \tag{6.2}$$

2. *Update*

$$\boldsymbol{\alpha}^{(s+1)} = \arg\min_{\boldsymbol{\alpha} \in \mathcal{C}(\mathfrak{A})} \frac{1}{N} \sum_{n=1}^{N} \boldsymbol{U}_n^{\mathrm{T}}(\boldsymbol{\beta}^{(s+1)}, \boldsymbol{\alpha}) \tilde{\boldsymbol{W}}_n \boldsymbol{U}_n(\boldsymbol{\beta}^{(s+1)}, \boldsymbol{\alpha}). \tag{6.3}$$

3. *Update $s = s + 1$. The iteration stops, when*

$$\|\boldsymbol{\alpha}^{(s)} - \boldsymbol{\alpha}^{(s+1)}\|_2^2 + \|\boldsymbol{\beta}^{(s)} - \boldsymbol{\beta}^{(s+1)}\|_2^2 < \epsilon',$$

where ϵ' is a small positive number.

The optimization problem (6.2) in Step 1 can be solved by the Newton-Raphson method. On the other hand, given $\boldsymbol{\beta}^{(s+1)}$, the objective function of the optimization problem (6.3) is convex with respect to $\boldsymbol{\alpha}$. As described in Section 3.2, the parameter space $\mathcal{C}_{J_N}(\mathfrak{A})$ shares same boundary geometry with the generalized moment space $\mathcal{M}$ defined in Equation (5.6). And thus, the CNM algorithms in [Wang, 2007] also can be adopted for the optimization problem (6.3).

Algorithm 6.2 (The CNM for GLMM)**.**
Set $s = 0$ and given $\boldsymbol{\beta}$. From an initial estimate $Q^{(0)}$ with finite support $\Theta^{(0)}$ and $\boldsymbol{\alpha}^{(0)} = \int_{\mathcal{B}_2} \boldsymbol{P}(b) \mathrm{d}Q^{(0)}$, repeat the following steps:

1. *Compute all the local minimas* $\{\theta_j^{(s)}\}_{j=1}^{r^{(s)}}$ *of the function*

$$\mathcal{D}(b) = \frac{1}{N}\sum_{n=1}^{N}\left(\boldsymbol{\alpha}^{(s)} - \boldsymbol{P}(b)\right)^{\mathrm{T}} \boldsymbol{\Phi}_n(\boldsymbol{\beta})\tilde{\boldsymbol{W}}_n\left(\boldsymbol{Y}_n - \boldsymbol{\Phi}_n^{\mathrm{T}}\boldsymbol{\alpha}^{(s)}\right)$$

   over $\mathcal{B}_2$*. The iteration stops if the minimum of* $\mathcal{D}(b)$ *is zero.*

2. *Construct a set of candidate support points by*

$$\Theta^{(s),+} = \Theta^{(s)} \cup \{\theta_j^{(s)}\}_{j=1}^{r^{(s)}}.$$

   Let $r^{(s),+}$ *be the number of elements in* $\Theta^{(s),+}$*.*

3. *Solve the optimization problem*

$$\begin{aligned}\min \quad & \frac{1}{N}\sum_{n=1}^{N}\left(\boldsymbol{Y}_n - \sum_{i=1}^{r^{(s)}+1}\pi_i\boldsymbol{\Phi}_n^{\mathrm{T}}\boldsymbol{P}(b_i)\right)^{\mathrm{T}}\tilde{\boldsymbol{W}}_n\left(\boldsymbol{Y}_n - \sum_{i=1}^{r^{(s)}+1}\pi_i\boldsymbol{\Phi}_n^{\mathrm{T}}\boldsymbol{P}(b_i)\right)\\ s.t. \quad & \pi_i \geq 0, \quad i = 1, \ldots, r^{(s),+},\end{aligned}$$

   where $b_i \in \Theta^{(s),+}$*. We denote its solution by* $\boldsymbol{\pi}^{(s)} = (\pi_1^{(s)}, \ldots, \pi_{r^{(s),+}}^{(s)})^{\mathrm{T}}$*.*

4. *Discard all* b_i*s with zero* $\pi_i^{(s)}$*, update* $Q^{(s)}$*,* $\Theta^{(s)}$ *and* $\boldsymbol{\alpha}^{(s)} = \int_{\mathcal{B}_2}\boldsymbol{P}(b)\mathrm{d}Q^{(s)}$*, and set* $s = s + 1$*.*

6.4 Simulation Studies

To evaluate the performance of the GMM estimator $(\hat{\boldsymbol{\beta}}_{\mathrm{GMM}}, \hat{\boldsymbol{\alpha}}_{\mathrm{GMM}})$ for the Poisson regression model with random intercept and slope, we consider the following parameter setting.

For each $n \in \{1, \ldots, N\}$, let T_n follow a discrete uniform distribution over $\{1, \ldots, 5\}$. For each $n \in \{1, \ldots, N\}$ and $t \in \{1, \ldots, T_n\}$, let $\boldsymbol{X}_{nt} = (X_{nt1}, X_{nt2}, X_{nt3}, X_{nt4})^{\mathrm{T}} \in \mathbb{R}^4$ be the fixed effects, where X_{nt1} and X_{nt2} independently follow a continuous uniform distribution over $[-0.3, 0.3]$, X_{nt3} follows a Bernoulli distribution with success probability 0.5 and $X_{nt4} = 10 \times X_{nt1}X_{nt2}$ is considered as the interaction effects of X_{nt1} and X_{nt2}. For each n and t, $Z_{nt} = t/20$. The true value of the regression parameter $\boldsymbol{\beta}$ is $(-1, 2, 0.5, 0)^{\mathrm{T}} \in \mathbb{R}^4$. Three possible random effects distributions are considered.

1. The random effects b_{n1} and b_{n2} are independent to each other. Moreover, they have the same marginal distribution

$$Q(b) = 0.4I(b \leq 0) + 0.1I(b \leq 1) + 0.5I(b \leq 2).$$

2. The random effects $\boldsymbol{b} = (b_{n1}, b_{n2})^{\mathrm{T}} \in \mathbb{R}^2$ follow a bivariate normal distribution with mean zero and covariance matrix

$$\boldsymbol{\Sigma} = \begin{bmatrix} 1 & 0 \\ 0 & 0.1 \end{bmatrix}.$$

3. The random effects $\boldsymbol{b} = (b_{n1}, b_{n2})^{\mathrm{T}} \in \mathbb{R}^2$ follow a bivariate normal distribution with mean zero and covariance matrix

$$\boldsymbol{\Sigma} = \begin{bmatrix} 1 & 0.9 \times \sqrt{0.1} \\ 0.9 \times \sqrt{0.1} & 0.1 \end{bmatrix}.$$

We use the Chebyshev polynomials (see Definition 2.4.2) defined on $\mathcal{B} = [-6, 6]$ as the orthonormal basis $\{P_j(b)\}_{j=0}^{J_N}$ in $L^2(\mathcal{B}, \mu)$, where $\mu = (1 - b^2)^{-1/2}\mathrm{d}b$. The approximation property has been studied in Section 2.4.1. For different sample sizes, the dimensions of the generalized moments $\boldsymbol{\alpha} \in \mathbb{R}^{J_N}$ are different, where $J_N = \lfloor 2N^{1/3} \rfloor$. Three sample size levels are considered ($N = 50, 100$ and 200).

We consider two different working correlation matrices: the independence and the AR(1). The parameter in the AR(1) correlation matrices is 0.5. We also consider the case in which the working correlation matrix $\boldsymbol{R}_n$ is the true correlation matrix of $\boldsymbol{U}_n(\boldsymbol{\beta}^*, Q^*)$. The initial estimators $(\tilde{\boldsymbol{\alpha}}, \tilde{\boldsymbol{\beta}}, \tilde{\boldsymbol{\gamma}})$ are estimated by the GMM with $\boldsymbol{W}_n = \boldsymbol{I}_n$, where for each n, $\boldsymbol{I}_n$ is a $T_n \times T_n$ identity matrix. Moreover we let $\gamma_2 = \gamma_1^4$. The number of the repetitions is 1000. The NPMLE can not be easily computed in this case because the random-effects $\boldsymbol{b}_n$ is bivariate. Instead, we consider the Model 6.B, in which the random-effects distributions are bivariate normal, and fit it by the penalized quasi-likelihood (PQL) method [Breslow and Clayton, 1993]. The PQL estimators are calculated by the MATLAB code *fitgmle* in the Statistics and Machine Learning Toolbox.

Model 6.B.

For each $n \in \{1, \ldots, N\}$ and $t \in \{1, \ldots, T_n\}$, the response $Y_{nt} \mid (\boldsymbol{X}_{nt}, Z_{nt}, \boldsymbol{b}_n)$ follows a Poisson distribution with mean $\mu_{nt}(\boldsymbol{b}_n)$, where $\mu_{nt}(\boldsymbol{b}_n)$ depends on the regression parameter $\boldsymbol{\beta}$ via the log-link function

$$\log \mu_{nt}(\boldsymbol{b}_n) = \boldsymbol{X}_{nt}^{\mathrm{T}}\boldsymbol{\beta} + (\beta_0 + b_{n1}) + Z_{nt} \times (b_{n2} + \beta_b),$$

and $\boldsymbol{X}_{nt} \in \mathbb{R}^p$ are the covariates for the fixed effects, $\boldsymbol{\beta} \in \mathbb{R}^p$ is the regression parameter, β_0 and β_b are the mean of the random intercept and slope correspondingly, Z_{nt} are the covariates to the random effects and $\boldsymbol{b}_n = (b_{n1}, b_{n2})^{\mathrm{T}} \in \mathbb{R}^2$ are random effects. Here $\boldsymbol{b}_n \in \mathbb{R}^2$ is assumed to follow a bivariate normal distribution with mean zero and covariance matrix $\boldsymbol{\Sigma}$, where $\boldsymbol{\Sigma}$ is unknown.

The simulation results are summarized in Table 6.1-6.6. From these tables, we observe the followings.

1. When the true model is correctly specified by 6.B, the PQL estimators could have the smaller MSE than the GMM estimators; see Table 6.3-6.6. This is because that the PQL use the modelling information of the probability function $\mathrm{pr}(\boldsymbol{Y}_n \mid \boldsymbol{X}_n, \boldsymbol{Z}_n)$ while the GMM only use the modelling information of the mean condition $\mathbb{E}[\boldsymbol{Y}_n \mid \boldsymbol{X}_n, \boldsymbol{Z}_n]$; and the random effects distribution is parametric in Model 6.B, while it is non-parametric in Model 6.A.

2. When the regression parameter $\boldsymbol{\beta}$ is considered, the GMM estimators with the true correlation matrices has smaller bias and MSE than the ones with the working correlation matrices; see Table 6.1-6.6. This provide an empirical evidence that correctly modelling the with-in subject correlation could increase the efficiency of the GMM estimators.

3. When the random effects do not follow normal, the PQL estimators could be very biased; see Table 6.1 and 6.2. On the other hand, the GMM estimators are consistent and with smaller bias. The reason is that the random effects distribution is misspecified in Model 6.B.

4. When the random intercept b_{n1} and the random slope b_{n2} are correlated, we observe slightly larger bias and MSE in the GMM estimators; see Table 6.5-6.6. This is because that the independence assumption does not hold in Model 6.A.

6.5 Application to the Epileptic Seizures Data

In Section 1.6.2, we have described the epileptic seizures data, which has been analyzed by [Thall and Vail, 1990] and [Breslow and Clayton, 1993]. The response Y_{nt} is the biweekly number of seizures for the n^{th} patient at equally spaced times $t = 1, 2, 3, 4$. The covariates include baseline seizure count (BASE), treatment (TREAT), age (AGE) and possibly the interaction between treatment and age (INTER). Preliminary analysis indicated that the response were substantially lower during the fourth visit and thus an indicator (V4) is introduced to model such effect; see [Breslow and Clayton, 1993].

We consider the following two models under the independence assumptions that $\boldsymbol{Y}_n \mid (\boldsymbol{X}_n, \boldsymbol{Z}_n)$, $n = 1, \ldots, N$, are independent to each other and conditional on the random effects $\boldsymbol{b}_n$, and $Y_{nt} \mid (\boldsymbol{X}_n, \boldsymbol{Z}_n, \boldsymbol{b}_n)$, $t = 1, \ldots, T_n$, are independent to each other.

Model 6.C.
For each n and t, $Y_{nt} \mid (\boldsymbol{X}_{nt}, b_n)$ follows a Poisson distribution with mean $\mu_{nt}(b_n)$ such that

$$\begin{aligned}\log \mu_{nt}(b_n) = b_n + \beta_1 \text{BASE}_{nt} + \beta_2 \text{TREAT}_{nt} + \beta_3 \text{INTER}_{nt} \\ + \beta_4 \log(\text{AGE}_{nt}) + \beta_5 \text{V4},\end{aligned}$$

where $b_n \in \mathbb{R}$ has a probability measure Q defined on $\mathcal{B} = [-20, 20]$. The variance of $Y_{nt} \mid (\boldsymbol{X}_{nt}, b_n)$ is $\mu_{nt}(b_n)$.

Model 6.D.
For each n and t, $Y_{nt} \mid (\boldsymbol{X}_{nt}, \boldsymbol{b}_n)$ follows a Poisson distribution with mean $\mu_{nt}(\boldsymbol{b}_n)$

				β_1		β_2		β_3		β_4	
	Model	N	J_N	BIAS	MSE	BIAS	MSE	BIAS	MSE	BIAS	MSE
GMM	Inv	50	7	0.002	0.049	-0.003	0.051	0.006	0.053	-0.011	0.017
		100	9	-0.001	0.024	-0.005	0.023	0.004	0.024	-0.006	0.008
		200	11	0.003	0.012	-0.001	0.011	0.001	0.011	-0.005	0.004
	Indep	50	7	0.070	0.209	-0.108	0.217	-0.022	0.066	0.010	0.075
		100	9	0.033	0.085	-0.060	0.080	-0.011	0.029	0.009	0.030
		200	11	0.012	0.041	-0.032	0.041	-0.005	0.014	0.001	0.013
	AR(1)	50	7	0.033	0.097	-0.070	0.108	-0.025	0.064	-0.003	0.044
		100	9	0.016	0.043	-0.040	0.040	-0.013	0.027	0.000	0.015
		200	11	0.009	0.020	-0.021	0.019	-0.007	0.013	-0.001	0.006
PQL	6.B	50	-	-0.017	0.067	0.036	0.066	0.836	0.797	-0.005	0.023
		100	-	-0.018	0.032	0.026	0.030	0.852	0.774	-0.002	0.010
		200	-	-0.013	0.015	0.027	0.014	0.854	0.752	-0.002	0.005

Table 6.1: Simulation results of the regression parameter $\boldsymbol{\beta}$ in Case 1, and $N = 50, 100$ and 200. The working correlation matrices used in the GMM include the inverse of the true correlation matrix of $\boldsymbol{U}_n(\boldsymbol{\beta}, \boldsymbol{\alpha})$ (Inv), the independence (Indep) and the first-order autoregressive (AR(1)) correlation matrix.

			γ_1		α_1		α_2		α_3		α_4	
	Model	N	BIAS	MSE	BIAS	MSE	BIAS	MSE	BIAS	MSE	BIAS	MSE
GMM	Inv	50	0.204	0.735	-0.777	2.078	1.879	12.282	0.357	6.873	-0.077	3.627
		100	0.152	0.362	-0.523	1.018	1.254	6.086	0.348	6.592	-0.010	2.561
		200	0.109	0.179	-0.361	0.530	0.872	3.284	0.155	5.020	0.110	1.910
	Indep	50	0.329	0.933	-1.068	3.319	2.123	16.577	0.871	8.574	0.152	5.068
		100	0.246	0.470	-0.734	1.746	1.562	10.110	0.567	7.280	0.031	3.634
		200	0.167	0.221	-0.482	0.848	1.034	5.212	0.421	5.820	0.124	2.699
	AR(1)	50	0.275	0.910	-1.010	3.077	2.154	15.703	0.419	7.194	0.095	4.429
		100	0.197	0.439	-0.643	1.478	1.489	8.794	0.168	5.848	0.035	2.876
		200	0.135	0.213	-0.424	0.727	1.014	4.823	0.017	4.566	0.162	1.900
PQL	6.B	50	-0.023	0.113	2.820	8.241	0.899	1.089	-2.540	6.891	-1.812	3.764
		100	-0.026	0.077	2.833	8.227	0.824	0.919	-2.606	7.066	-1.737	3.439
		200	-0.051	0.053	2.889	8.478	0.851	0.902	-2.648	7.143	-1.781	3.484

Table 6.2: Simulation results of the five generalized moments $\{\gamma_1, \alpha_1, \ldots, \alpha_4\}$ in Case 1, and $N = 50, 100$ and 200. The working correlation matrices used in the GMM include the inverse of the true correlation matrix of $\boldsymbol{U}_n(\boldsymbol{\beta}, \boldsymbol{\alpha})$ (Inv), the independence (Indep) and the first-order autoregressive (AR(1)) correlation matrix.

				β_1		β_2		β_3		β_4	
	Model	N	J_N	BIAS	MSE	BIAS	MSE	BIAS	MSE	BIAS	MSE
GMM	Inv	50	7	-0.018	0.163	0.012	0.171	-0.014	0.199	0.008	0.056
		100	9	0.005	0.073	0.001	0.080	-0.002	0.081	0.011	0.025
		200	11	0.000	0.034	-0.002	0.039	0.001	0.036	0.004	0.011
	Indep	50	7	0.099	0.498	-0.169	0.484	-0.042	0.150	0.006	0.185
		100	9	0.079	0.236	-0.140	0.274	-0.032	0.079	0.017	0.092
		200	11	0.047	0.115	-0.085	0.140	-0.019	0.041	0.018	0.043
	AR(1)	50	7	0.063	0.325	-0.120	0.293	-0.051	0.146	0.002	0.118
		100	9	0.047	0.134	-0.092	0.146	-0.039	0.075	0.013	0.054
		200	11	0.029	0.060	-0.057	0.066	-0.023	0.038	0.014	0.024
PQL	6.B	50	-	-0.020	0.179	-0.002	0.176	-0.009	0.068	-0.000	0.060
		100	-	-0.007	0.082	-0.005	0.086	0.002	0.035	0.011	0.028
		200	-	-0.005	0.040	-0.003	0.043	0.003	0.017	0.005	0.013

Table 6.3: Simulation results of the regression parameter $\boldsymbol{\beta}$ in Case 2, and $N = 50, 100$ and 200. The working correlation matrices used in the GMM include the inverse of the true correlation matrix of $\boldsymbol{U}_n(\boldsymbol{\beta}, \boldsymbol{\alpha})$ (Inv), the independence (Indep) and the first-order autoregressive (AR(1)) correlation matrix.

			γ_1		α_1		α_2		α_3		α_4	
	Model	N	BIAS	MSE	BIAS	MSE	BIAS	MSE	BIAS	MSE	BIAS	MSE
GMM	Inv	50	0.195	0.376	-0.427	0.576	1.092	3.390	-0.227	1.351	-0.389	1.043
		100	0.167	0.180	-0.334	0.359	0.848	2.248	-0.184	1.269	-0.395	0.809
		200	0.153	0.099	-0.249	0.218	0.623	1.426	-0.164	1.106	-0.354	0.706
	Indep	50	0.185	0.369	-0.490	0.831	1.277	3.986	-0.273	1.660	-0.491	1.330
		100	0.184	0.205	-0.390	0.508	0.958	2.775	-0.120	1.343	-0.462	1.078
		200	0.179	0.126	-0.311	0.332	0.743	1.963	-0.088	1.186	-0.398	0.807
	AR(1)	50	0.154	0.335	-0.434	0.719	1.160	3.544	-0.306	1.469	-0.452	1.103
		100	0.163	0.187	-0.352	0.457	0.894	2.551	-0.218	1.261	-0.356	0.748
		200	0.160	0.114	-0.274	0.285	0.691	1.783	-0.204	1.137	-0.316	0.560
PQL	6.B	50	0.023	0.109	0.275	0.266	-0.118	0.421	-0.671	0.778	0.135	0.520
		100	0.017	0.055	0.162	0.130	-0.066	0.299	-0.455	0.439	0.096	0.468
		200	0.019	0.027	0.105	0.087	-0.060	0.176	-0.302	0.266	0.093	0.333

Table 6.4: Simulation results of the five generalized moments $\{\gamma_1, \alpha_1, \ldots, \alpha_4\}$ in Case 2, and $N = 50, 100$ and 200. The working correlation matrices used in the GMM include the inverse of the true correlation matrix of $\boldsymbol{U}_n(\boldsymbol{\beta}, \boldsymbol{\alpha})$ (Inv), the independence (Indep) and the first-order autoregressive (AR(1)) correlation matrix.

				β_1		β_2		β_3		β_4	
	Model	N	J_N	BIAS	MSE	BIAS	MSE	BIAS	MSE	BIAS	MSE
GMM	Inv	50	7	0.002	0.167	0.015	0.173	-0.005	0.206	-0.011	0.061
		100	9	0.006	0.076	0.012	0.072	0.001	0.084	0.010	0.026
		200	11	-0.000	0.032	0.006	0.034	0.003	0.039	0.004	0.012
	Indep	50	7	0.111	0.578	-0.197	0.611	-0.044	0.161	0.031	0.219
		100	9	0.093	0.260	-0.118	0.284	-0.034	0.083	0.029	0.101
		200	11	0.060	0.120	-0.066	0.150	-0.016	0.043	0.017	0.047
	AR(1)	50	7	0.078	0.342	-0.133	0.329	-0.050	0.154	0.017	0.124
		100	9	0.061	0.149	-0.086	0.149	-0.037	0.079	0.024	0.057
		200	11	0.040	0.063	-0.043	0.069	-0.018	0.040	0.011	0.024
PQL	6.B	50	-	0.000	0.181	0.011	0.183	0.002	0.076	-0.013	0.065
		100	-	0.005	0.084	0.007	0.079	0.008	0.038	0.006	0.029
		200	-	0.001	0.037	0.007	0.039	0.003	0.017	0.003	0.014

Table 6.5: Simulation results of the regression parameter $\boldsymbol{\beta}$ in Case 3, and $N = 50, 100$ and 200. The working correlation matrices used in the GMM include the inverse of the true correlation matrix of $\boldsymbol{U}_n(\boldsymbol{\beta}, \boldsymbol{\alpha})$ (Inv), the independence (Indep) and the first-order autoregressive (AR(1)) correlation matrix.

			γ_1		α_1		α_2		α_3		α_4	
	Model	N	BIAS	MSE	BIAS	MSE	BIAS	MSE	BIAS	MSE	BIAS	MSE
GMM	Inv	50	0.190	0.346	-0.357	0.527	1.080	3.315	-0.356	1.403	-0.367	0.921
		100	0.180	0.182	-0.269	0.346	0.838	2.326	-0.315	1.181	-0.354	0.782
		200	0.149	0.097	-0.185	0.193	0.609	1.400	-0.193	1.061	-0.363	0.693
	Indep	50	0.208	0.385	-0.487	0.830	1.276	4.018	-0.243	1.692	-0.517	1.402
		100	0.220	0.230	-0.389	0.564	1.039	3.145	-0.196	1.461	-0.448	1.016
		200	0.194	0.137	-0.305	0.343	0.834	2.228	-0.123	1.273	-0.467	0.993
	AR(1)	50	0.196	0.365	-0.452	0.781	1.224	3.791	-0.342	1.624	-0.475	1.261
		100	0.193	0.213	-0.336	0.502	0.963	2.845	-0.341	1.388	-0.383	0.835
		200	0.165	0.116	-0.239	0.269	0.735	1.862	-0.287	1.165	-0.357	0.653
PQL	6.B	50	0.012	0.110	0.283	0.276	-0.116	0.444	-0.686	0.791	0.147	0.574
		100	0.018	0.053	0.161	0.145	-0.088	0.299	-0.451	0.447	0.147	0.492
		200	0.022	0.027	0.107	0.093	-0.079	0.185	-0.306	0.269	0.141	0.359

Table 6.6: Simulation results of the five generalized moments $\{\gamma_1, \alpha_1, \ldots, \alpha_4\}$ in Case 3, and $N = 50, 100$ and 200. The working correlation matrices used in the GMM include the inverse of the true correlation matrix of $\boldsymbol{U}_n(\boldsymbol{\beta}, \boldsymbol{\alpha})$ (Inv), the independence (Indep) and the first-order autoregressive (AR(1)) correlation matrix.

such that

$$\log \mu_{nt}(\boldsymbol{b}_n) = b_{n1} + \beta_1 \mathrm{BASE}_{nt} + \beta_2 \mathrm{TREAT}_{nt} + \beta_3 \mathrm{INTER}_{nt} \\ + \beta_4 \log(\mathrm{AGE}_{nt}) + b_{n2}\mathrm{TIME},$$

where $\boldsymbol{b}_n = (b_{n1}, b_{n2})^{\mathrm{T}} \in \mathbb{R}^2$*,* $b_{n1} \in \mathbb{R}$ *and* $b_{n2} \in \mathbb{R}$ *have probability measures* Q_1 *and* Q_2 *defined on* $\mathcal{B} = [-20, 20]$*, and the* TIME *effects is coded in* $(-0.3, -0.1, 0.1, 0.3)$*. The variance of* $Y_{nt} \mid (\boldsymbol{X}_{nt}, b_n)$ *is* $\mu_{nt}(b_n)$*. We further assume that* b_{n1} *and* b_{n2} *are independent.*

Model 6.C can be fitted by the GMM with one generalized moment $\alpha_0 = \int_{\mathcal{B}} \exp(b)\mathrm{d}Q$, while Model 6.D can be fitted by the GMM with the generalized moments $\boldsymbol{\alpha} = (\alpha_1, \alpha_2, \alpha_3, \alpha_4)^{\mathrm{T}} \in \mathbb{R}^4$, where for each $j \in \{1, \ldots, 4\}$,

$$\alpha_j = \int_{\mathcal{B}} \exp(\mathrm{TIME}_j \times b_{n2})\mathrm{d}Q_2 \times \int_{\mathcal{B}} \exp(b_{n1})\mathrm{d}Q_1$$

and TIME_js are associated with the coded TIME effects.

Table 6.7 presents the GMM estimates in Model 6.C and 6.D with two different working correlation matrices: the independence and the AR(1). We also give the PQL estimates from [Breslow and Clayton, 1993], where b_n in Model 6.D follows a normal distribution and $(b_{n1}, b_{n2})^{\mathrm{T}} \in \mathbb{R}^2$ follows a bivariate normal distribution.

In Figure 6.1 to 6.4, we use the standardized residuals to check the adequacy of the GMM. Here the estimated covariance matrices in Equation (5.14) are used. Panel (a), (b), (c) and (d) in each figure display to the plots of the standardized residuals against the coded visiting time, patients, responses and fitted means. The red lines represent the 97.5% and 2.5% empirical quantiles of the standardized residuals. The black straight line represents the mean of the standardized residuals. We see that the mean of the standardized residuals is close to zero and no trend is observed from any of the panels. This means that the considered models adequately fit the data through the GMM.

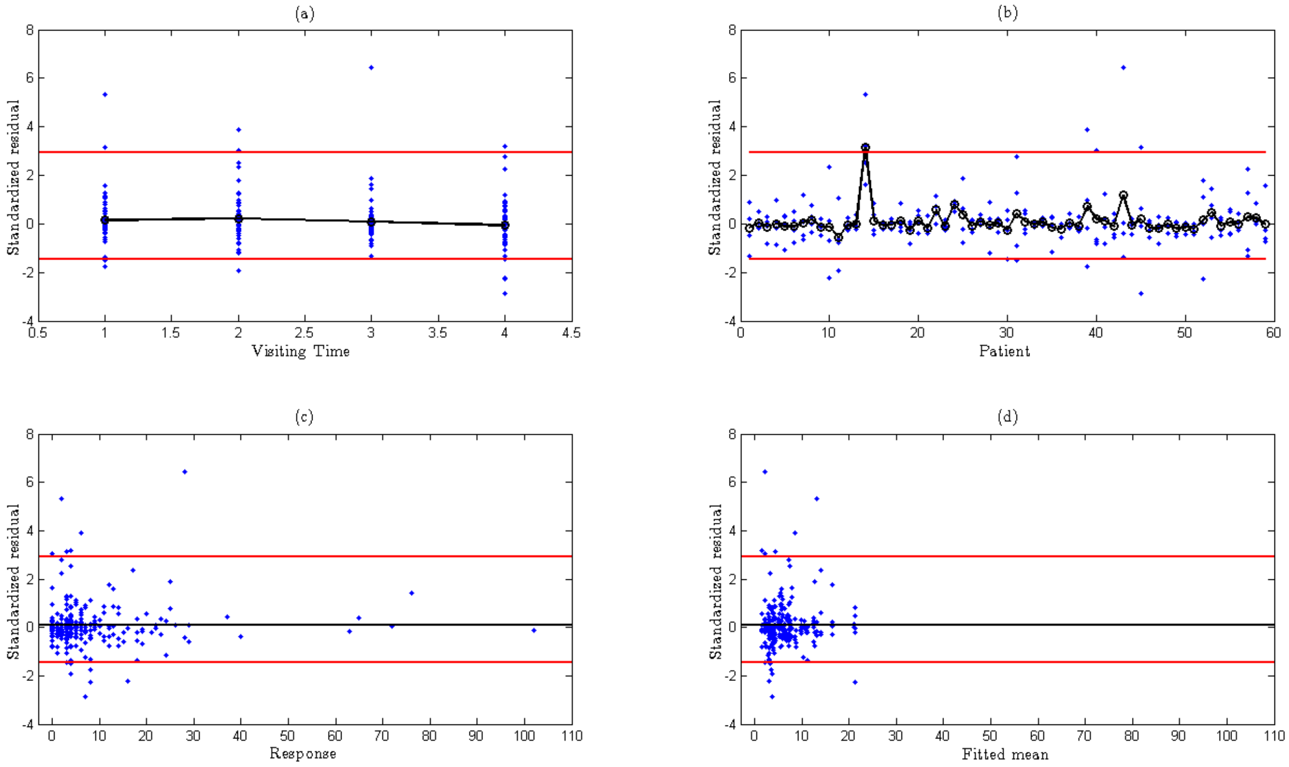

Figure 6.1: Plots of the residual analysis of the fitted Model 6.C to the epilepsy seizures data using the GMM with AR(1) working correlation matrix.

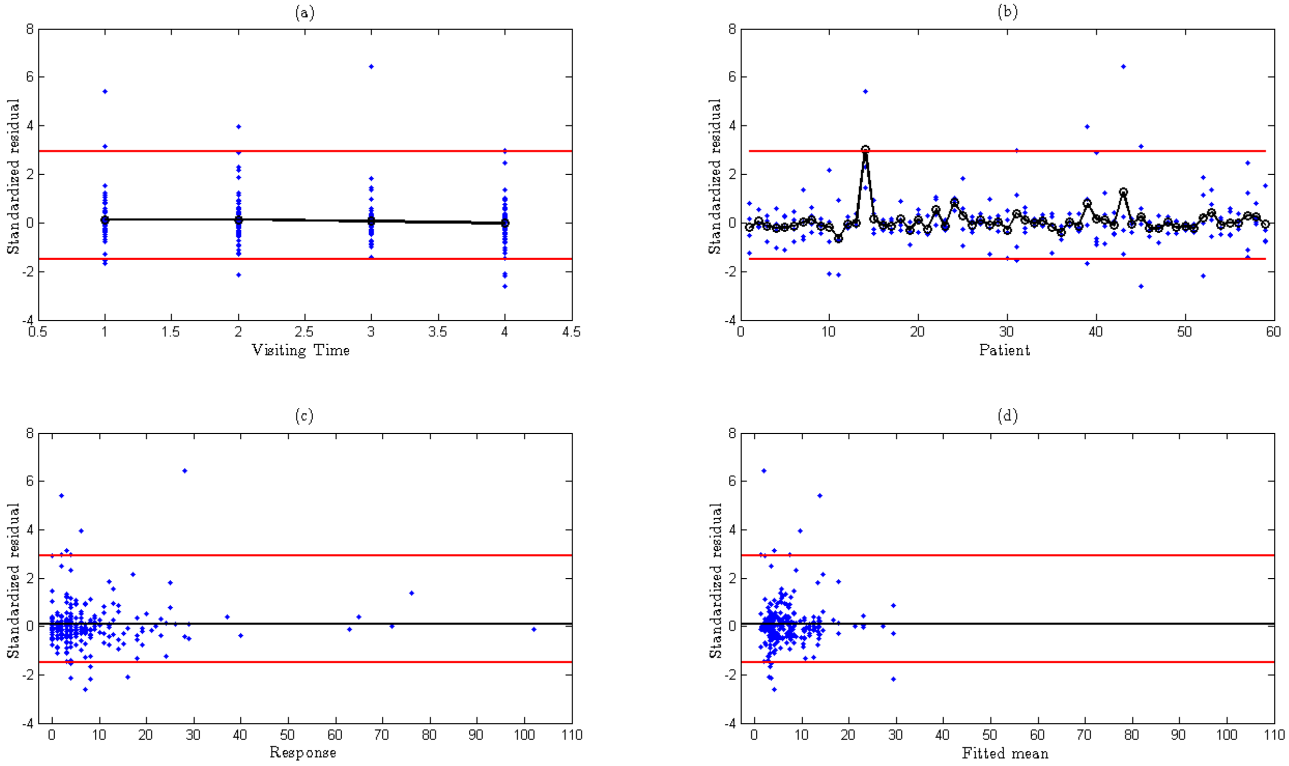

Figure 6.2: Plots of the residual analysis of the fitted Model 6.C to the epilepsy seizures data using the GMM with independent working correlation matrix.

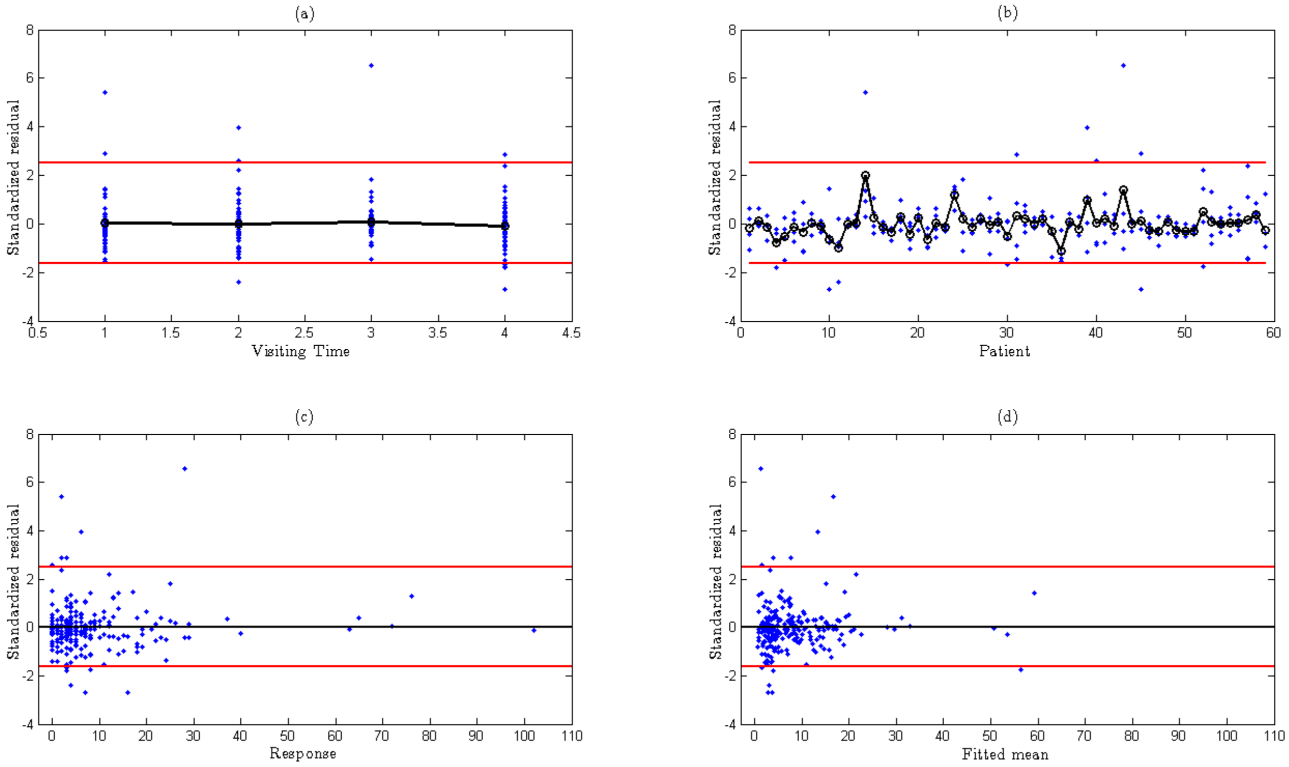

Figure 6.3: Plots of the residual analysis of the fitted Model 6.D to the epilepsy seizures data using the GMM with AR(1) working correlation matrix.

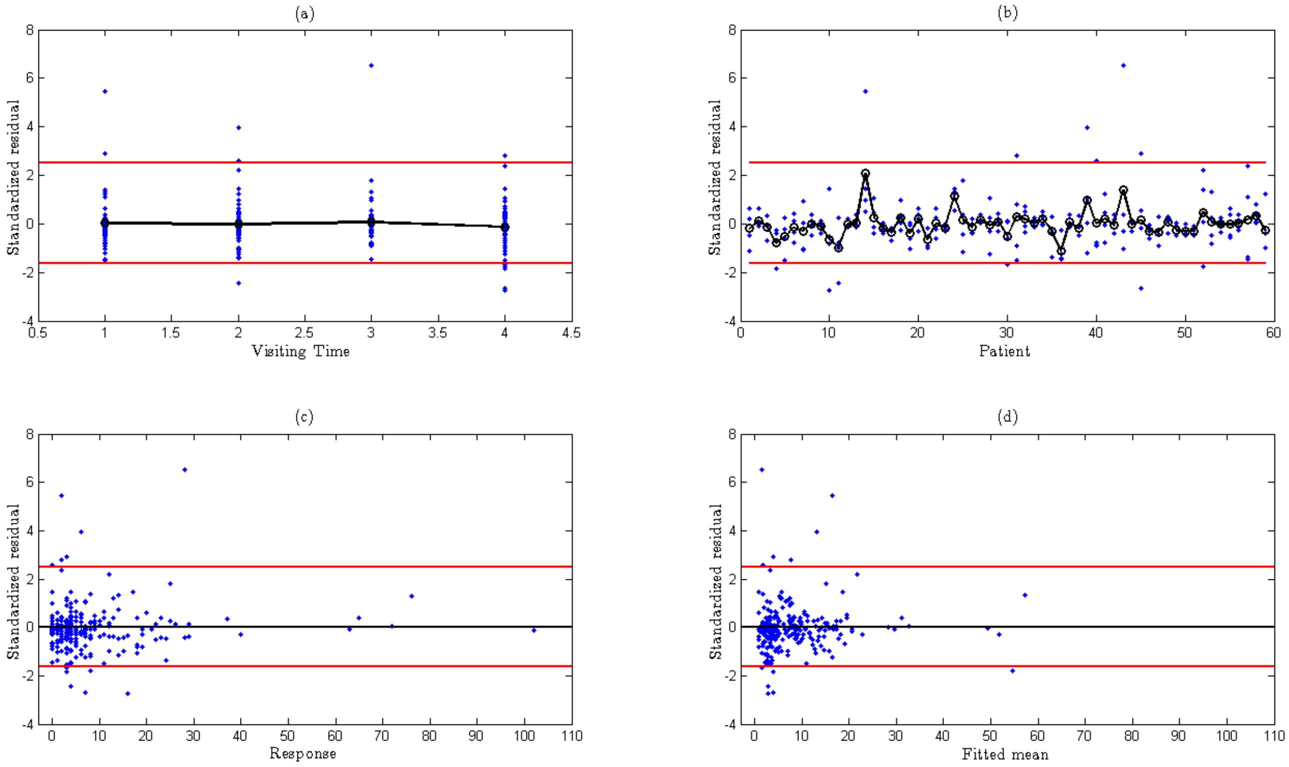

Figure 6.4: Plots of the residual analysis of the fitted Model 6.D to the epilepsy seizures data using the GMM with independent working correlation matrix.

	Model 6.C			Model 6.D		
	GMM		PQL	GMM		PQL
	AR(1)	Indep		AR(1)	Indep	
BASE	0.72	0.76	0.86	0.92	0.93	0.87
TREAT	-0.74	-0.78	-0.93	-1.20	-1.07	-0.91
INTER	0.18	0.25	0.34	0.49	0.44	0.33
Log(Age)	0.44	0.40	0.47	0.86	0.79	0.46
V4	-0.00	-0.07	-0.10	-	-	-

Table 6.7: Estimated regression parameters in the models to the epilepsy seizures data.

6.6 Conclusion and Discussion

In this chapter, we discussed the GMM for the Poisson regression models with random intercept and slope. Because the parameter space share same geometric properties, the computational algorithms proposed in Section 5.5 can be easily adopted for the GMM for Model 6.A. The simulation results indicate that the resulting estimators are consistent, when the models are correctly specific. Moreover, we compare the performance of the GMM with the QPL method in the simulation study. Because the GMM does not require the distribution assumption on the random effects, it could perform superior to the PQL, when the random effects distribution is not normal.

Model 6.A is more flexible than a Poisson regression model with univariate random-effects. However, there still exists a strong modelling assumption that the random intercept b_{n1} and random slope b_{n2} are independent to each other for each n. In the following paragraphs, we discuss the case where the independent assumption is relaxed.

Let

$$Q(\boldsymbol{b}_n) = Q_1(b_{n1}) \times Q_2(b_{n2} \mid b_{n1})$$

be the joint probability measure of (b_{n1}, b_{n2}) defined on $\mathcal{B}_1 \times \mathcal{B}_2$, where $Q_2(b_{n2} \mid b_{n1})$ is the distribution of $b_{n2} \mid b_{n1}$. Without the independent assumption, we have, for each

$n \in \{1, \ldots, N\}$ and $t \in \{1, \ldots, T_n\}$,

$$\begin{aligned}\mathbb{E}\left[Y_{nt} \mid \boldsymbol{X}_{nt}, Z_{nt}\right] &= \int_{\mathcal{B}_1 \times \mathcal{B}_2} \mu_{nt}(\boldsymbol{b}_n) \mathrm{d}Q(\boldsymbol{b}_n) \\ &= \int_{\mathcal{B}_1} \int_{\mathcal{B}_2} \exp\left(\boldsymbol{X}_{nt}^{\mathrm{T}} \boldsymbol{\beta} + b_{n1} + Z_{nt} b_{n2}\right) \mathrm{d}Q_2(b_{n2} \mid b_{n1}) \mathrm{d}Q_1(b_{n1}) \\ &= \int_{\mathcal{B}_1} \exp(b_{n1}) \int_{\mathcal{B}_2} \exp\left(\boldsymbol{X}_{nt}^{\mathrm{T}} \boldsymbol{\beta} + Z_{nt} b_{n2}\right) \mathrm{d}Q_2(b_{n2} \mid b_{n1}) \mathrm{d}Q_1(b_{n1}).\end{aligned}$$

By the reparameterization-approximation procedure, we have

$$\int_{\mathcal{B}_2} \exp\left(\boldsymbol{X}_{nt}^{\mathrm{T}} \boldsymbol{\beta} + Z_{nt} b_{n2}\right) \mathrm{d}Q_2(b_{n2} \mid b_{n1}) \approx \sum_{j=0}^{J_N} \boldsymbol{\phi}_{ntj}^{\mathrm{T}}(\boldsymbol{\beta}) \alpha_j'(b_{n1}),$$

where $\{P_j(b)\}_{j=0}^{J_N}$ is an orthonormal polynomial system defined on $(\mathcal{B}_2, \mu)$ and for each $j \in \{0, \ldots, J_N\}$,

$$\phi_{ntj}(\boldsymbol{\beta}) = \int_{\mathcal{B}_2} \exp\left(\boldsymbol{X}_{nt}^{\mathrm{T}} \boldsymbol{\beta} + Z_{nt} b_{n2}\right) P_j(b_{n2}) \mathrm{d}\mu$$

and

$$\alpha_j'(b_{n1}) = \int_{\mathcal{B}_2} P_j(b_{n2}) \mathrm{d}Q_2(b_{n2} \mid b_{n1}).$$

Then, the expectation of $Y_{nt} \mid (\boldsymbol{X}_{nt}, Z_{nt})$ is approximated by

$$\boldsymbol{U}_n(\boldsymbol{\beta}, \boldsymbol{\alpha}) = \boldsymbol{\Phi}_n^{\mathrm{T}}(\boldsymbol{\beta}) \boldsymbol{\alpha}, \tag{6.4}$$

where $\boldsymbol{\Phi}_n(\boldsymbol{\beta})$ is a $(J_N+1) \times T_n$ matrix whose elements are $\phi_{ntj}(\boldsymbol{\beta})$ and $\boldsymbol{\alpha} = (\alpha_0, \ldots, \alpha_{J_N})^{\mathrm{T}} \in \mathbb{R}^{J_N+1}$ and for each $j \in \{1, \ldots, J_N\}$,

$$\alpha_j = \int_{\mathcal{B}_{n1}} \alpha_j'(b_{n1}) \mathrm{dQ}_1(b_{n1}).$$

Modelling either the conditional distribution $Q_2(b_{n2} \mid b_{n1})$ or $\{\alpha_j'(b_{n1})\}_{j=0}^{J_N}$ will give us the approximation in Equation (6.4).

Appendix: E

E.1 MATLAB Code for Algorithm 6.1

```
function [beta, as, out, objo] = GLM_GMM(DATA, q, ...
                    a0, beta0, V, W)

Ind1 = 0;
count1 = 1;
out = 1;
objo = 1e5;

while Ind1 == 0
    count1 = count1 + 1;

    [betanew, unused, H0] = GMM_NR(DATA, q,...
        a0, beta0, V);

    [asnew, obj] = GMM_CNM_adj(DATA(:,end), q,...
        as, H0, W);

    if count1 > 5e2
        Ind1 = 1;
        out = 0;
    end

    if  norm(betanew-beta)<1e-5
        Ind1 = 1;
    else
        beta = betanew;
        as = asnew;
        objo = obj;
    end

end
```

```
function [betanew, obj, H0] = GMM_NR(DATA, q,...
    a0, beta0, V, W)

    % for log-link functions

Ind = 0;
count = 1;
betap = length(beta);
X = DATA(:,2:2+betap-1);
Z = DATA(:,2+betap);
S = DATA(:,end);
H0a = exp(repmat(X*beta,[1,length(q)])+Z*q);
H0 = H0a*(V*V');
U = H0*aini';
D = X'.*repmat(U',[size(X,2),1]);
C = D*W*(S-U);
G = D*W*D';
objini =(S-U)'*W*(S-U);

while Ind == 0
    count = count + 1;

    betanew = beta + G\C;

    H0a = exp(repmat(X*betanew,[1,length(q)])+Z*q);
    H0 = H0a*(V*V');
    U = H0*aini';

    D = X'.*repmat(U',[size(X,2),1]);
    C = D*W*(S-U);
    G = D*W*D';
    obj =(S-U)'*W*(S-U);
    if obj < objini
        Ind = 1;
    else
        beta = betanew;
    end

    if count > 1e2
        betanew = betaini;
        obj = objini;
        Ind =1;
    end

end

H0a = exp(repmat(X*betanew,[1,length(q)])+Z*q);
H0 = H0a*(V*V');
```

E.2 MATLAB Code for Algorithm 6.2

```
function [as, Dsmin] = GMM_CNM_adj(S, q,...
    ainit, H0, W)

Ind = 0;
e = 1e-10;

pLs = find(ainit > 0);
as = ainit;

U = H0*as';

g = -(S-U)'*W*(H0-U(:,ones(1,length(q))));
```

```
Hc = H0'*W*H0;
Ac = H0'*W*S;
Dsmino = (S-U)'*W*(S-U);

while Ind == 0

    dg = diff(g);
    signdg = sign(dg);
    dsigndg = diff(signdg);
    minL = find(dsigndg == 2)+1;
    L = minL;
    pLsnew = [1 pLs L length(q)];
    pLsnew = unique(pLsnew);

    H = Hc(pLsnew, pLsnew);
    A = Ac(pLsnew,:);

    %warning off;
    options = optimset('Algorithm', ....
             'interior-point-convex',...
             'display', 'off');

    Ast = quadprog((H+H')/2, -A, ...
                 -eye(size(H,2)),...
                 zeros(1,size(H,2)), ...
                 [], [], [], [], [], options);

    as = zeros(1,length(as));
    as(pLsnew) = Ast;
    pLs = pLsnew(Ast > 0);

    U = H0*as';
    g = -(S-U)'*W*(H0-U(:,ones(1,length(q))));
    Dsmin = (S-U)'*W*(S-U);

    if abs(Dsmin-Dsmino) < 1e-5 || max(g) < e
        Ind = 1;
    else
        aso = as;
        Dsmino = Dsmin;
    end

end
```

Chapter 7

Asymptotic Properties of the Generalized Method of Moments for Univariate Mixed-Effects Models

7.1 Introduction

In the previous chapter, we introduced the generalized method of moments estimator for univariate mixed-effects models; see Definition 5.4.1. In this chapter, we study the asymptotic properties of the GMM estimator.

The statistical theory of estimators with a diverging number of parameters has attracted interests from many researchers, especially with the advent of high-dimensional data in many scientific areas; see [Lam and Fan, 2008], [Chen et al., 2009] and [Wang, 2011]. Under the framework that the dimension of the regression parameter grows towards infinity with sample size, the asymptotic properties of many regular estimators have been studied; see the profile-kernel likelihood estimator [Lam and Fan, 2008], empirical likelihood estimators [Chen et al., 2009] and GEE estimators [Wang, 2011]. In this chapter, we consider the case where the dimension of the regression parameter

$\boldsymbol{\beta}$ is fixed but the dimension of the generalized moments vector $\boldsymbol{\alpha}$ diverges with the sample size N. To emphasize that the dimension of $\boldsymbol{\alpha}$ depends on the sample size, we add N as a subscript to $\boldsymbol{\alpha}$ and use the notation $\boldsymbol{\alpha}_N$ for the generalized moments vector. Although the estimation setting is different, similar techniques are used to establish the asymptotic results as in [Lam and Fan, 2008] and [Wang, 2011].

We make the following contributions in this chapter. Firstly, we show $N^{1/2}J_N^{-1/2}$ as the convergence rate of the GMM estimator; see Theorem 7.3.1. Here J_N is the dimension of the generalized moment vector $\boldsymbol{\alpha}_N$. The dimension J_N may diverge with the sample size N and $J_N^{1/2}N^{-1/2} = o(1)$. Secondly, we prove that the plug-in weighting matrices, obtained from the initial estimators, converges to non-random matrices asymptotically; see Theorem 7.4.1. Next, we derive the asymptotic normality for the GMM; see Theorem 7.5.1. Note that the regularity conditions in [Wilks, 1938] fail in the GMM in the sense that the dimension of the parameter diverges with the sample size and the true value of the parameter is a boundary point. However, according to Theorem 7.5.1, an asymptotically normal test statistics can be obtained in $\mathbb{R}^p$. Lastly, we show that the covariance matrix of $\boldsymbol{Y}_n \mid (\boldsymbol{X}_n, \boldsymbol{Z}_n)$ can be consistently estimated; see Theorem 7.6.1.

We organize this chapter as follows. In Section 7.2, we list the regularity conditions which are required to establish the asymptotic results in this chapter. We also give some lemmas which are straightforward to prove from the regularity conditions. In Section 7.3, we show the convergence rate of the GMM estimator. In Section 7.4, we show that the weighting matrices, which are obtained from the initial estimators, converges to non-random matrices as the sample size goes to infinity. In Section 7.5, we give the asymptotic normality theorem for the GMM. In Section 7.6, we show that the covariance matrix of $\boldsymbol{Y}_n \mid (\boldsymbol{X}_n, \boldsymbol{Z}_n)$ can be consistently estimated from the GMM estimators. Lastly, we discuss the challenges in using the asymptotic results for hypothesis testing problems on the regression parameters. The proofs of the lemmas can be found in Appendix E.

7.2 Regularity Conditions

In this section, the regularity conditions for the asymptotic results are listed as follows. Note that we use C_N as the notation of a finite number depending on the sample size N, but the value of C_N may vary between lines. Examples satisfying the following regularity conditions include the Poisson regression and the logistic regression models with the range of the random effects defined on a compact set.

Regularity Condition 7.A.
For every integer N, there exists a finite number C_N such that

$$\sup_{n\in\{1,\dots,N\}} \sup_{t\in\{1,\dots,T_n\}} \boldsymbol{X}_{nt}^{\mathrm{T}}\boldsymbol{X}_{nt} \leq C_N$$

with probability one.

Regularity Condition 7.B.
Let $J_N N^{-1}$ converge to zero, as N goes to infinity. For any function $h(b) \in L^2(\mathcal{B},\mu)$, there exists an expansion of $h(b)$ by an orthonormal system $\{P_j(b)\}_{j=0}^{J_N}$ in $L^2(\mathcal{B},\mu)$ such that

$$h(b) = \sum_{j=0}^{J_N} \int_{\mathcal{B}} h(b)P_j(b)\mathrm{d}\mu P_j(b) + o(J_N N^{-1}).$$

Regularity Condition 7.C.
The inverse link function $g^{-1}(s)$ is a smooth function of $s \in \mathbb{R}$. For each $\boldsymbol{\beta} \in \mathbb{R}^p$ and $(\boldsymbol{X}_{nt}, Z_{nt})$, $n = 1,\dots,N$ and $t = 1,\dots,T_n$, the following functions of $b \in \mathbb{R}$ are in the space $L^2(\mathcal{B},\mu)$:

1. $g^{-1}(\boldsymbol{X}_{nt}^{\mathrm{T}}\boldsymbol{\beta} + Z_{nt}b)$,
2. $\dot{g}^{-1}(\boldsymbol{X}_{nt}^{\mathrm{T}}\boldsymbol{\beta} + Z_{nt}b)$,
3. $\ddot{g}^{-1}(\boldsymbol{X}_{nt}^{\mathrm{T}}\boldsymbol{\beta} + Z_{nt}b)$,
4. $\dddot{g}^{-1}(\boldsymbol{X}_{nt}^{\mathrm{T}}\boldsymbol{\beta} + Z_{nt}b)$,

where $\dot{g}^{-1}(s)$, $\ddot{g}^{-1}(s)$ and $\dddot{g}^{-1}(s)$ are the first, second and third order derivatives of $g^{-1}(s)$ with respect to s.

Regularity Condition 7.D.

For every integer N and probability measure Q defined on $\mathcal{B}$, there exists a finite number $C_N(Q)$ by which the following functions of $\boldsymbol{\beta} \in \mathbb{R}^p$ are bounded:

1. $\sup_{n\in\{1,\ldots,N\}} \sup_{\{(\boldsymbol{X}_{nt},Z_{nt})\}_{t=1}^{T_n}} \int_{\mathcal{B}} \left(g^{-1}(\boldsymbol{X}_{nt}^{\mathrm{T}}\boldsymbol{\beta} + Z_{nt}b)\right)^2 \mathrm{d}Q$,
2. $\sup_{n\in\{1,\ldots,N\}} \sup_{\{(\boldsymbol{X}_{nt},Z_{nt})\}_{t=1}^{T_n}} \int_{\mathcal{B}} \left(\dot{g}^{-1}(\boldsymbol{X}_{nt}^{\mathrm{T}}\boldsymbol{\beta} + Z_{nt}b)\right)^2 \mathrm{d}Q$,
3. $\sup_{n\in\{1,\ldots,N\}} \sup_{\{(\boldsymbol{X}_{nt},Z_{nt})\}_{t=1}^{T_n}} \int_{\mathcal{B}} \left(\dddot{g}^{-1}(\boldsymbol{X}_{nt}^{\mathrm{T}}\boldsymbol{\beta} + Z_{nt}b)\right)^2 \mathrm{d}Q$.

Regularity Condition 7.E.

For each $(\boldsymbol{X}_{nt}, Z_{nt}, \boldsymbol{X}_{nt'}, Z_{nt})$, $n = 1, \ldots, N$ and $t, t' = 1, \ldots, T_n$, and $\boldsymbol{\beta} \in \mathbb{R}^p$, the following functions of $b \in \mathbb{R}$ are in the space $L^2(\mathcal{B}, \mu)$:

1. $g^{-1}(\boldsymbol{X}_{nt}^{\mathrm{T}}\boldsymbol{\beta} + Z_{nt}b) \times g^{-1}(\boldsymbol{X}_{nt'}^{\mathrm{T}}\boldsymbol{\beta} + Z_{nt'}b)$,
2. $g^{-1}(\boldsymbol{X}_{nt}^{\mathrm{T}}\boldsymbol{\beta} + Z_{nt}b) \times \dot{g}^{-1}(\boldsymbol{X}_{nt'}^{\mathrm{T}}\boldsymbol{\beta} + Z_{nt'}b)$,
3. $\dot{g}^{-1}(\boldsymbol{X}_{nt}^{\mathrm{T}}\boldsymbol{\beta} + Z_{nt}b) \times \dot{g}^{-1}(\boldsymbol{X}_{nt'}^{\mathrm{T}}\boldsymbol{\beta} + Z_{nt'}b)$,
4. $\ddot{g}^{-1}(\boldsymbol{X}_{nt}^{\mathrm{T}}\boldsymbol{\beta} + Z_{nt}b) \times g^{-1}(\boldsymbol{X}_{nt'}^{\mathrm{T}}\boldsymbol{\beta} + Z_{nt'}b)$.

Regularity Condition 7.F.

The function $h \circ g^{-1}(s)$ is a smooth function with respect to s. For each $\boldsymbol{\beta} \in \mathbb{R}^p$ and $(\boldsymbol{X}_{nt}, Z_{nt})$, $n = 1, \ldots, N$ and $t = 1, \ldots, T_n$, the following functions of $b \in \mathbb{R}$ are in the space $L^2(\mathcal{B}, \mu)$:

1. $h \circ g^{-1}(\boldsymbol{X}_{nt}^{\mathrm{T}}\boldsymbol{\beta} + Z_{nt}b)$,
2. $\dot{h} \circ g^{-1}(\boldsymbol{X}_{nt}^{\mathrm{T}}\boldsymbol{\beta} + Z_{nt}b)$,
3. $\ddot{h} \circ g^{-1}(\boldsymbol{X}_{nt}^{\mathrm{T}}\boldsymbol{\beta} + Z_{nt}b)$,

where $\dot{h} \circ g^{-1}(s)$ and $\ddot{h} \circ g^{-1}(s)$ are the first and second order derivatives of $h \circ g^{-1}(s)$ with respect to s.

Regularity Condition 7.G.

For every integer N and probability measure Q defined on $\mathcal{B}$, there exists a finite number $C_N(Q)$ by which the following functions of $\boldsymbol{\beta} \in \mathbb{R}^p$ are bounded:

1. $\sup_{n\in\{1,\dots,N\}} \sup_{\{(\boldsymbol{X}_{nt},Z_{nt})\}_{t=1}^{T_n}} \int_{\mathcal{B}} \left(h \circ g^{-1}(\boldsymbol{X}_{nt}^{\mathrm{T}}\boldsymbol{\beta} + Z_{nt}b)\right)^2 \mathrm{d}Q,$

2. $\sup_{n\in\{1,\dots,N\}} \sup_{\{(\boldsymbol{X}_{nt},Z_{nt})\}_{t=1}^{T_n}} \int_{\mathcal{B}} \left(\dot{h} \circ g^{-1}(\boldsymbol{X}_{nt}^{\mathrm{T}}\boldsymbol{\beta} + Z_{nt}b)\right)^2 \mathrm{d}Q,$

3. $\sup_{n\in\{1,\dots,N\}} \sup_{\{(\boldsymbol{X}_{nt},Z_{nt})\}_{t=1}^{T_n}} \int_{\mathcal{B}} \left(\ddot{h} \circ g^{-1}(\boldsymbol{X}_{nt}^{\mathrm{T}}\boldsymbol{\beta} + Z_{nt}b)\right)^2 \mathrm{d}Q.$

Regularity Condition 7.H.
For each $n \in \{1,\dots,N\}$ and $t \in \{1,\dots,T_n\}$, the function $U_{nt}(\boldsymbol{\beta},\boldsymbol{\alpha}_N)$ is Lipschitz continuous, i.e., for any two different parameter values $(\boldsymbol{\beta},\boldsymbol{\alpha}_N)$ and $(\boldsymbol{\beta}',\boldsymbol{\alpha}'_N)$, there exists a finite number $L_{nt} > 0$ such that

$$|U_{nt}(\boldsymbol{\beta},\boldsymbol{\alpha}_N) - U_{nt}(\boldsymbol{\beta}',\boldsymbol{\alpha}'_N)|^2 \leq L_{nt} \times \left(\|\boldsymbol{\alpha}_N - \boldsymbol{\alpha}'_N\|_2^2 + \|\boldsymbol{\beta} - \boldsymbol{\beta}'\|_2^2\right), \qquad (7.1)$$

where $U_{nt}(\boldsymbol{\beta},\boldsymbol{\alpha}_N)$ is defined in Equation (5.4).

Regularity Condition 7.I.
For each $n \in \{1,\dots,N\}$ and $t \in \{1,\dots,T_n\}$,

$$\mathbb{E}\left[U_{nt}(\boldsymbol{\beta}^*, Q^*)\right] = 0.$$

Moreover, for every integer N,

$$\sup_{n\in\{1,\dots,N\}} \sup_{t\in\{1,\dots,T_n\}} \mathbb{E}\left[U_{nt}^4(\boldsymbol{\beta}^*, Q^*)\right]$$

is bounded, where $\boldsymbol{\beta}^*$ is the true value of the regression parameter and Q^* is the true random-effects distribution defined on $\mathcal{B}$.

Regularity Condition 7.J.
For each $n \in \{1,\dots,N\}$ and $t,t' \in \{1,\dots,T_n\}$,

$$\tilde{w}_{ntt'} = w_{ntt'} + O_p(J_N^{1/2}N^{-1/2}). \qquad (7.2)$$

where $\tilde{w}_{ntt'}$ and $w_{ntt'}$ are the elements of $\tilde{\boldsymbol{W}}_n$ and $\boldsymbol{W}_n$ correspondingly.

The listed regularity conditions provide the following lemmas. The proofs of these lemmas can be found in Appendix E.1.

Lemma 7.2.1.
Assume that Regularity Condition 7.B and 7.C are satisfied. Then, for every integer N and $\boldsymbol{\beta} \in \mathbb{R}^p$, there exists a finite number C_N by which the following functions of $\boldsymbol{\beta} \in \mathbb{R}^p$ are bounded:

1. $\sup_{n\in\{1,\dots,N\}} \sup_{t\in\{1,\dots,T_n\}} \boldsymbol{\phi}_{nt}^{\mathrm{T}}(\boldsymbol{\beta})\boldsymbol{\phi}_{nt}(\boldsymbol{\beta})$,
2. $\sup_{n\in\{1,\dots,N\}} \sup_{t\in\{1,\dots,T_n\}} \dot{\boldsymbol{\phi}}_{nt}^{\mathrm{T}}(\boldsymbol{\beta})\dot{\boldsymbol{\phi}}_{nt}(\boldsymbol{\beta})$,
3. $\sup_{n\in\{1,\dots,N\}} \sup_{t\in\{1,\dots,T_n\}} \ddot{\boldsymbol{\phi}}_{nt}^{\mathrm{T}}(\boldsymbol{\beta})\ddot{\boldsymbol{\phi}}_{nt}(\boldsymbol{\beta})$,

where

$$\boldsymbol{\phi}_{nt}(\boldsymbol{\beta}) = \left(\phi_{nt0}(\boldsymbol{\beta}), \dots, \phi_{ntJ_N}(\boldsymbol{\beta})\right)^{\mathrm{T}} \in \mathbb{R}^{J_N+1}, \tag{7.3}$$

$$\dot{\boldsymbol{\phi}}_{nt}(\boldsymbol{\beta}) = \left(\dot{\phi}_{nt0}(\boldsymbol{\beta}), \dots, \dot{\phi}_{ntJ_N}(\boldsymbol{\beta})\right)^{\mathrm{T}} \in \mathbb{R}^{J_N+1}, \tag{7.4}$$

$$\ddot{\boldsymbol{\phi}}_{nt}(\boldsymbol{\beta}) = \left(\ddot{\phi}_{nt0}(\boldsymbol{\beta}), \dots, \ddot{\phi}_{ntJ_N}(\boldsymbol{\beta})\right)^{\mathrm{T}} \in \mathbb{R}^{J_N+1}, \tag{7.5}$$

and for each j, $\phi_{ntj}(\boldsymbol{\beta})$ is defined in Equation (5.3), and

$$\dot{\phi}_{ntj}(\boldsymbol{\beta}) = \int_{\mathcal{B}} \dot{g}^{-1}(\boldsymbol{X}_{nt}^{\mathrm{T}}\boldsymbol{\beta} + Z_{nt}b)P_j(b)\mathrm{d}\mu, \tag{7.6}$$

and

$$\ddot{\phi}_{ntj}(\boldsymbol{\beta}) = \int_{\mathcal{B}} \ddot{g}^{-1}(\boldsymbol{X}_{nt}^{\mathrm{T}}\boldsymbol{\beta} + Z_{nt}b)P_j(b)\mathrm{d}\mu. \tag{7.7}$$

Lemma 7.2.2.
Assume that Regularity Condition 7.B, 7.C and 7.D are satisfied. Then, for every integer N and $(\boldsymbol{\beta}, \boldsymbol{\alpha}_N) \in \mathbb{R}^p \times \mathcal{M}_{J_N+1}$, there exists a finite number C_N by which the following functions of $(\boldsymbol{\beta}, \boldsymbol{\alpha}_N) \in \mathbb{R}^p \times \mathcal{M}_{J_N+1}$ are bounded:

1. $\sup_{n\in\{1,\dots,N\}} \sup_{t\in\{1,\dots,T_n\}} \left|\boldsymbol{\phi}_{nt}^{\mathrm{T}}(\boldsymbol{\beta})\boldsymbol{\alpha}_N\right|$,
2. $\sup_{n\in\{1,\dots,N\}} \sup_{t\in\{1,\dots,T_n\}} \left|\dot{\boldsymbol{\phi}}_{nt}^{\mathrm{T}}(\boldsymbol{\beta})\boldsymbol{\alpha}_N\right|$,
3. $\sup_{n\in\{1,\dots,N\}} \sup_{t\in\{1,\dots,T_n\}} \left|\ddot{\boldsymbol{\phi}}_{nt}^{\mathrm{T}}(\boldsymbol{\beta})\boldsymbol{\alpha}_N\right|$,

4. $\sup_{n\in\{1,\ldots,N\}}\sup_{t\in\{1,\ldots,T_n\}}\left|\dddot{\boldsymbol{\phi}}_{nt}^{\mathrm{T}}(\boldsymbol{\beta})\boldsymbol{\alpha}_N\right|$,

where

$$\dddot{\boldsymbol{\phi}}_{nt}(\boldsymbol{\beta})=\left(\dddot{\phi}_{nt0}(\boldsymbol{\beta}),\ldots,\dddot{\phi}_{ntJ_N}(\boldsymbol{\beta})\right)^{\mathrm{T}}\in\mathbb{R}^{J_N+1} \tag{7.8}$$

and for each j,

$$\dddot{\phi}_{ntj}(\boldsymbol{\beta})=\int_{\mathcal{B}}\dddot{g}^{-1}(\boldsymbol{X}_{nt}^{\mathrm{T}}\boldsymbol{\beta}+Z_{nt}b)P_j(b)\mathrm{d}\mu. \tag{7.9}$$

Lemma 7.2.3.

Assume that Regularity Condition 7.B and 7.F are satisfied. Then, for every integer N and $\boldsymbol{\beta}\in\mathbb{R}^p$, there exists a finite number C_N by which the following functions of $\boldsymbol{\beta}\in\mathbb{R}^p$ are bounded:

1. $\sup_{n\in\{1,\ldots,N\}}\sup_{t\in\{1,\ldots,T_n\}}\boldsymbol{a}_{nt}^{\mathrm{T}}(\boldsymbol{\beta})\boldsymbol{a}_{nt}(\boldsymbol{\beta})$,

2. $\sup_{n\in\{1,\ldots,N\}}\sup_{t\in\{1,\ldots,T_n\}}\dot{\boldsymbol{a}}_{nt}^{\mathrm{T}}(\boldsymbol{\beta})\dot{\boldsymbol{a}}_{nt}(\boldsymbol{\beta})$,

where

$$\boldsymbol{a}_{nt}(\boldsymbol{\beta})=(a_{nt0}(\boldsymbol{\beta}),\ldots,a_{ntJ_N}(\boldsymbol{\beta}))^{\mathrm{T}}\in\mathbb{R}^{J_N+1} \tag{7.10}$$

and

$$\dot{\boldsymbol{a}}_{nt}(\boldsymbol{\beta})=(\dot{a}_{nt0}(\boldsymbol{\beta}),\ldots,\dot{a}_{ntJ_N}(\boldsymbol{\beta}))^{\mathrm{T}}\in\mathbb{R}^{J_N+1}, \tag{7.11}$$

and for each j, $a_{ntj}(\boldsymbol{\beta})$ is defined in Equation (5.7), and

$$\dot{a}_{ntj}(\boldsymbol{\beta})=\int_{\mathcal{B}}\dot{h}\circ g^{-1}(\boldsymbol{X}_{nt}^{\mathrm{T}}\boldsymbol{\beta}+Z_{nt}b)P_j(b)\mathrm{d}\mu.$$

Lemma 7.2.4.

Assume that Regularity Condition 7.B, 7.F and 7.G are satisfied. Then, for every integer N and $(\boldsymbol{\beta},\boldsymbol{\alpha}_N)\in\mathbb{R}^p\times\mathcal{M}_{J_N+1}$, there exists a finite number C_N by which the following functions of $(\boldsymbol{\beta},\boldsymbol{\alpha}_N)\in\mathbb{R}^p\times\mathcal{M}_{J_N+1}$ are bounded:

1. $\sup_{n\in\{1,\ldots,N\}}\sup_{t\in\{1,\ldots,T_n\}}\left|\boldsymbol{a}_{nt}^{\mathrm{T}}(\boldsymbol{\beta})\boldsymbol{\alpha}_N\right|$, *2.* $\sup_{n\in\{1,\ldots,N\}}\sup_{t\in\{1,\ldots,T_n\}}\left|\dot{\boldsymbol{a}}_{nt}^{\mathrm{T}}(\boldsymbol{\beta})\boldsymbol{\alpha}_N\right|$,

3. $\sup_{n\in\{1,\ldots,N\}}\sup_{t\in\{1,\ldots,T_n\}}\left|\ddot{\boldsymbol{a}}_{nt}^{\mathrm{T}}(\boldsymbol{\beta})\boldsymbol{\alpha}_N\right|$,

where

$$\ddot{\boldsymbol{a}}_{nt}(\boldsymbol{\beta}) = (\ddot{a}_{nt0}(\boldsymbol{\beta}),\ldots,\ddot{a}_{ntJ_N}(\boldsymbol{\beta}))^{\mathrm{T}} \in \mathbb{R}^{J_N+1}, \tag{7.12}$$

and for each j,

$$\ddot{a}_{ntj}(\boldsymbol{\beta}) = \int_{\mathcal{B}} \ddot{h}\circ g^{-1}(\boldsymbol{X}_{nt}^{\mathrm{T}}\boldsymbol{\beta}+Z_{nt}b)P_j(b)\mathrm{d}\mu.$$

Lemma 7.2.5.

Assume that Regularity Condition 7.B and 7.E are satisfied. Then, for every integer N and $\boldsymbol{\beta}\in\mathbb{R}^p$, there exists a finite number C_N by which the following functions of $\boldsymbol{\beta}\in\mathbb{R}^p$ are bounded:

1. $\sup_{n\in\{1,\ldots,N\}}\sup_{t\in\{1,\ldots,T_n\}}\boldsymbol{c}_{ntt'}^{\mathrm{T}}(\boldsymbol{\beta})\boldsymbol{c}_{ntt'}(\boldsymbol{\beta})$,

2. $\sup_{n\in\{1,\ldots,N\}}\sup_{t\in\{1,\ldots,T_n\}}\dot{\boldsymbol{c}}_{ntt'}^{\mathrm{T}}(\boldsymbol{\beta})\dot{\boldsymbol{c}}_{ntt'}(\boldsymbol{\beta})$,

where

$$\boldsymbol{c}_{ntt'}(\boldsymbol{\beta}) = (c_{ntt'0}(\boldsymbol{\beta}),\ldots,c_{ntt'J_N}(\boldsymbol{\beta}))^{\mathrm{T}} \in \mathbb{R}^{J_N+1} \tag{7.13}$$

and

$$\dot{\boldsymbol{c}}_{ntt'}(\boldsymbol{\beta}) = (\dot{c}_{ntt'0}(\boldsymbol{\beta}),\ldots,\dot{c}_{ntt'J_N}(\boldsymbol{\beta}))^{\mathrm{T}} \in \mathbb{R}^{J_N+1}, \tag{7.14}$$

and for each j, $c_{ntt'j}(\boldsymbol{\beta})$ is defined in Equation (5.8), and

$$\begin{aligned}\dot{c}_{ntt'j}(\boldsymbol{\beta}) = \int_{\mathcal{B}} \big(&\dot{g}^{-1}(\boldsymbol{X}_{nt}^{\mathrm{T}}\boldsymbol{\beta}+Z_{nt}b)\times g^{-1}(\boldsymbol{X}_{nt'}^{\mathrm{T}}\boldsymbol{\beta}+Z_{nt}b)\\ &+\dot{g}^{-1}(\boldsymbol{X}_{nt'}^{\mathrm{T}}\boldsymbol{\beta}+Z_{nt}b)\times g^{-1}(\boldsymbol{X}_{nt}^{\mathrm{T}}\boldsymbol{\beta}+Z_{nt}b)\big)\,P_j(b)\mathrm{d}\mu.\end{aligned}$$

Lemma 7.2.6.

Assume that Regularity Condition 7.B, 7.D and 7.E are satisfied. Then, for every integer N and $(\boldsymbol{\beta},\boldsymbol{\alpha}_N)\in\mathbb{R}^p\times\mathcal{M}_{J_N+1}$, there exists a finite number C_N by which the following functions of $(\boldsymbol{\beta},\boldsymbol{\alpha}_N)\in\mathbb{R}^p\times\mathcal{M}_{J_N+1}$ are bounded:

1. $\sup_{n\in\{1,\ldots,N\}}\sup_{t\in\{1,\ldots,T_n\}}\left|\boldsymbol{c}_{ntt'}^{\mathrm{T}}(\boldsymbol{\beta})\boldsymbol{\alpha}_N\right|$,

2. $\sup_{n\in\{1,\ldots,N\}}\sup_{t\in\{1,\ldots,T_n\}}\left|\dot{\boldsymbol{c}}_{ntt'}^{\mathrm{T}}(\boldsymbol{\beta})\boldsymbol{\alpha}_N\right|$,

3. $\sup_{n\in\{1,\ldots,N\}}\sup_{t\in\{1,\ldots,T_n\}}\left|\ddot{\boldsymbol{c}}_{ntt'}^{\mathrm{T}}(\boldsymbol{\beta})\boldsymbol{\alpha}_N\right|$,

where

$$\ddot{\boldsymbol{c}}_{ntt'}(\boldsymbol{\beta}) = \left(\ddot{c}_{ntt'0}(\boldsymbol{\beta}),\ldots,\ddot{c}_{ntt'J_N}(\boldsymbol{\beta})\right)^{\mathrm{T}} \in \mathbb{R}^{J_N+1} \tag{7.15}$$

and for each j,

$$\begin{aligned}\ddot{c}_{ntt'j}(\boldsymbol{\beta}) = \int_{\mathcal{B}} \big(&\dot{g}^{-1}(\boldsymbol{X}_{nt}^{\mathrm{T}}\boldsymbol{\beta}+Z_{nt}b)\times\dot{g}^{-1}(\boldsymbol{X}_{nt'}^{\mathrm{T}}\boldsymbol{\beta}+Z_{nt}b)\\ &+\ddot{g}^{-1}(\boldsymbol{X}_{nt}^{\mathrm{T}}\boldsymbol{\beta}+Z_{nt}b)\times g^{-1}(\boldsymbol{X}_{nt'}^{\mathrm{T}}\boldsymbol{\beta}+Z_{nt}b)\\ &+\dot{g}^{-1}(\boldsymbol{X}_{nt'}^{\mathrm{T}}\boldsymbol{\beta}+Z_{nt}b)\times\dot{g}^{-1}(\boldsymbol{X}_{nt}^{\mathrm{T}}\boldsymbol{\beta}+Z_{nt}b)\\ &+\ddot{g}^{-1}(\boldsymbol{X}_{nt'}^{\mathrm{T}}\boldsymbol{\beta}+Z_{nt}b)\times g^{-1}(\boldsymbol{X}_{nt}^{\mathrm{T}}\boldsymbol{\beta}+Z_{nt}b)\big)\,P_j(b)\mathrm{d}\mu.\end{aligned}$$

Regularity Condition 7.B determines the error rate of the truncation approximations obtained from $\{P_j(b)\}_{j=0}^{J_N}$; see Corollary 2.4.1 and 2.4.2 for more details.

By Regularity Condition 7.I, we have, for each $n\in\{1,\ldots,N\}$ and $t\in\{1,\ldots,T_n\}$,

$$|U_{nt}(\boldsymbol{\beta}^*,Q^*)| = O_p(1). \tag{7.16}$$

By the Markov inequality, for an arbitrary $a>0$,

$$\mathrm{pr}(|U_{nt}(\boldsymbol{\beta}^*,Q^*)|>a) \le \frac{\mathbb{E}[U_{nt}^4(\boldsymbol{\beta}^*,Q^*)]}{a^4} \le \frac{1}{a^4}\sup_{n\in\{1,\ldots,N\}}\sup_{t\in\{1,\ldots,T_n\}}\mathbb{E}\left[U_{nt}^4(\boldsymbol{\beta}^*,Q^*)\right].$$

Therefore, $|U_{nt}(\boldsymbol{\beta}^*,Q^*)|$ is bounded in probability. Furthermore, for every integer N, Regularity Condition 7.I also implies that

$$\sup_{n\in\{1,\ldots,N\}}\sup_{t\in\{1,\ldots,T_n\}}\mathbb{E}\left[U_{nt}^2(\boldsymbol{\beta}^*,Q^*)\right]$$

is bounded by the Cauchy-Schwarz inequality.

7.3 Existence and Consistency of the Generalized Method of Moments Estimator

In this section, we give the existence and consistency of the GMM estimators as the sample size goes to infinity. Simulation-based evidence of the consistency of the GMM estimator have been shown in Section 5.7. The proofs of the lemmas for Theorem 7.3.1 can be found in Appendix E.2.

Theorem 7.3.1 (Existence and Consistency of the GMM Estimator)**.**
Assume that Regularity Condition 7.A-7.J are satisfied and $J_N N^{-1} = o(1)$, as the sample size N goes to infinity. Then, there exists a local minima of the optimization problem (5.11), denoted by $(\hat{\boldsymbol{\beta}}_{\text{GMM}}, \hat{\boldsymbol{\alpha}}_{N,\text{GMM}})$, such that

$$\|\hat{\boldsymbol{\alpha}}_{N,\text{GMM}} - \boldsymbol{\alpha}_N^*\|_2^2 + \|\hat{\boldsymbol{\beta}}_{\text{GMM}} - \boldsymbol{\beta}^*\|_2^2 = O_p(J_N N^{-1}),$$

where $(\boldsymbol{\beta}^*, \boldsymbol{\alpha}_N^*)$ is the true value of the parameters, and

$$\boldsymbol{\alpha}_N^* = \int_{\mathcal{B}} \boldsymbol{P}(b) \mathrm{d}Q^*.$$

Proof. Let

$$\tilde{\mathcal{Q}}(\boldsymbol{\beta}, \boldsymbol{\alpha}_N) = \frac{1}{N} \sum_{n=1}^{N} \boldsymbol{U}_n^{\mathrm{T}}(\boldsymbol{\beta}, \boldsymbol{\alpha}_N) \tilde{\boldsymbol{W}}_n \boldsymbol{U}_n(\boldsymbol{\beta}, \boldsymbol{\alpha}_N).$$

and $\Delta_N = J_N^{1/2} N^{-1/2}$. We aim to show that, $\forall \epsilon > 0$, there exists a $C > 0$, depending on N_0, such that, for any $N \geq N_0$,

$$\mathrm{pr}\left(\inf_{\|\boldsymbol{v}\|_2 = C} \tilde{\mathcal{Q}}(\boldsymbol{\beta}^* + \Delta_N \boldsymbol{v}_{\boldsymbol{\beta}}, \boldsymbol{\alpha}_N^* + \Delta_N \boldsymbol{v}_{\boldsymbol{\alpha}}) > \tilde{\mathcal{Q}}(\boldsymbol{\beta}^*, \boldsymbol{\alpha}_N^*) \right) \geq 1 - \varepsilon,$$

where $\boldsymbol{v} = (\boldsymbol{v}_{\boldsymbol{\alpha}}^{\mathrm{T}}, \boldsymbol{v}_{\boldsymbol{\beta}}^{\mathrm{T}})^{\mathrm{T}} \in \mathbb{R}^{J_N + p + 1}$ and $\boldsymbol{\alpha}_N^* + \Delta_N \boldsymbol{v}_{\boldsymbol{\alpha}} \in \mathcal{M}_{J_N+1} \subset \mathbb{R}^{J_N+1}$. It implies that with probability 1, there is a local minima $(\hat{\boldsymbol{\beta}}_{\text{GMM}}, \hat{\boldsymbol{\alpha}}_{N,\text{GMM}})$ in the ball with radius $C\Delta_N$ at $(\boldsymbol{\beta}^*, \boldsymbol{\alpha}_N^*)$ such that $\|\hat{\boldsymbol{\alpha}}_{N,\text{GMM}} - \boldsymbol{\alpha}_N^*\|_2^2 + \|\hat{\boldsymbol{\beta}}_{\text{GMM}} - \boldsymbol{\beta}^*\|_2^2 = O_p(\Delta_N^2)$.

By Taylor's expansion at $\Delta_N = 0$, we have

$$\begin{aligned}
&\tilde{\mathcal{Q}}(\boldsymbol{\beta}^* + \Delta_N \boldsymbol{v}_{\boldsymbol{\beta}}, \boldsymbol{\alpha}_N^* + \Delta_N \boldsymbol{v}_{\boldsymbol{\alpha}}) - \tilde{\mathcal{Q}}(\boldsymbol{\beta}^*, \boldsymbol{\alpha}_N^*) \\
&= \Delta_N \partial \tilde{\mathcal{Q}}(\boldsymbol{\beta}^*, \boldsymbol{\alpha}_N^*) + \frac{\Delta_N^2}{2} \partial^2 \tilde{\mathcal{Q}}(\boldsymbol{\beta}^*, \boldsymbol{\alpha}_N^*) + \frac{\Delta_N^3}{3!} \partial^3 \tilde{\mathcal{Q}}(\breve{\boldsymbol{\beta}}, \breve{\boldsymbol{\alpha}}_N) \\
&\triangleq I_1 + I_2 + I_3,
\end{aligned}$$

where

$$\begin{aligned}
\partial\tilde{\mathcal{Q}}(\boldsymbol{\beta},\boldsymbol{\alpha}_N) &= \frac{\partial}{\partial\Delta_N}\tilde{\mathcal{Q}}(\boldsymbol{\beta}+\Delta_N\boldsymbol{v}_{\boldsymbol{\beta}},\boldsymbol{\alpha}_N+\Delta_N\boldsymbol{v}_{\boldsymbol{\alpha}})\bigg|_{\Delta_N=0},\\
\partial^2\tilde{\mathcal{Q}}(\boldsymbol{\beta},\boldsymbol{\alpha}_N) &= \frac{\partial^2}{\partial\Delta_N^2}\tilde{\mathcal{Q}}(\boldsymbol{\beta}+\Delta_N\boldsymbol{v}_{\boldsymbol{\beta}},\boldsymbol{\alpha}_N+\Delta_N\boldsymbol{v}_{\boldsymbol{\alpha}})\bigg|_{\Delta_N=0},\\
\partial^3\tilde{\mathcal{Q}}(\boldsymbol{\beta},\boldsymbol{\alpha}_N) &= \frac{\partial^3}{\partial\Delta_N^3}\tilde{\mathcal{Q}}(\boldsymbol{\beta}+\Delta_N\boldsymbol{v}_{\boldsymbol{\beta}},\boldsymbol{\alpha}_N+\Delta_N\boldsymbol{v}_{\boldsymbol{\alpha}})\bigg|_{\Delta_N=0},
\end{aligned}$$

and $(\breve{\boldsymbol{\beta}},\breve{\boldsymbol{\alpha}}_N)$ lies between $(\boldsymbol{\beta}^*,\boldsymbol{\alpha}_N^*)$ and $(\boldsymbol{\beta}^*+\Delta_N\boldsymbol{v}_{\boldsymbol{\beta}},\boldsymbol{\alpha}_N^*+\Delta_N\boldsymbol{v}_{\boldsymbol{\alpha}})$. In the following of the proof, we examine the asymptotic order of the three terms I_1, I_2 and I_3.

For each $n\in\{1,\ldots,N\}$ and $t,t'\in\{1,\ldots,T_n\}$, let $\tilde{w}_{ntt'}$ and $w_{ntt'}$ be the elements of $\tilde{\boldsymbol{W}}_n$ and $\boldsymbol{W}_n$ correspondingly. Also let

$$\partial U_{nt}(\boldsymbol{\beta},\boldsymbol{\alpha}_N) = \frac{\partial}{\partial\Delta_N}U_{nt}(\boldsymbol{\beta}+\Delta_N\boldsymbol{v}_{\boldsymbol{\beta}},\boldsymbol{\alpha}_N+\Delta_N\boldsymbol{v}_{\boldsymbol{\alpha}})\bigg|_{\Delta_N=0}, \tag{7.17}$$

for each $n\in\{1,\ldots,N\}$ and $t\in\{1,\ldots,T_n\}$. We have

$$\begin{aligned}
\Delta_N^{-1}I_1 &= \frac{2}{N}\sum_{n=1}^{N}\sum_{t,t'=1}^{T_n}\tilde{w}_{ntt'}\partial U_{nt}(\boldsymbol{\beta}^*,\boldsymbol{\alpha}_N^*)U_{nt'}(\boldsymbol{\beta}^*,\boldsymbol{\alpha}_N^*)\\
&= 2(I_{11}+I_{12}+I_{13}+I_{14}),
\end{aligned}$$

where

$$\begin{aligned}
I_{11} &= \frac{1}{N}\sum_{n=1}^{N}\sum_{t,t'=1}^{T_n}w_{ntt'}\partial U_{nt}(\boldsymbol{\beta}^*,\boldsymbol{\alpha}_N^*)U_{nt'}(\boldsymbol{\beta}^*,Q^*),\\
I_{12} &= \frac{1}{N}\sum_{n=1}^{N}\sum_{t,t'=1}^{T_n}(\tilde{w}_{ntt'}-w_{ntt'})\partial U_{nt}(\boldsymbol{\beta}^*,\boldsymbol{\alpha}_N^*)U_{nt'}(\boldsymbol{\beta}^*,Q^*),\\
I_{13} &= \frac{1}{N}\sum_{n=1}^{N}\sum_{t,t'=1}^{T_n}w_{ntt'}\partial U_{nt}(\boldsymbol{\beta}^*,\boldsymbol{\alpha}_N^*)\left(U_{nt'}(\boldsymbol{\beta}^*,\boldsymbol{\alpha}_N^*)-U_{nt'}(\boldsymbol{\beta}^*,Q^*)\right),
\end{aligned}$$

and

$$I_{14} = \frac{1}{N}\sum_{n=1}^{N}\sum_{t,t'=1}^{T_n}(\tilde{w}_{ntt'}-w_{ntt'})\partial U_{nt}(\boldsymbol{\beta}^*,\boldsymbol{\alpha}_N^*)\left(U_{nt}(\boldsymbol{\beta}^*,\boldsymbol{\alpha}_N^*)-U_{nt}(\boldsymbol{\beta}^*,Q^*)\right).$$

By Lemma F.5, we have

$$I_{11} = O_p(N^{-1/2} J_N^{1/2} \|\boldsymbol{v}\|_2),$$

By Lemma F.1 and Equation (7.2) and (7.16), we have

$$\begin{aligned}|I_{12}| &\leq \frac{1}{N} \sum_{n=1}^{N} \sum_{t,t'=1}^{T_n} |\tilde{w}_{ntt'} - w_{ntt'}| \, |\partial U_{nt}(\boldsymbol{\beta}^*, \boldsymbol{\alpha}_N^*)| \, |U_{nt'}(\boldsymbol{\beta}^*, Q^*)| \\ &= O_p(J_N^{1/2} N^{-1/2} \|\boldsymbol{v}\|_2)\end{aligned}$$

By Lemma F.1 and Regularity Condition 7.B-7.D, we have

$$\begin{aligned}|I_{13}| &\leq \frac{1}{N} \sum_{n=1}^{N} \sum_{t,t'=1}^{T_n} |w_{ntt'}| \, |\partial U_{nt}(\boldsymbol{\beta}^*, \boldsymbol{\alpha}_N^*)| \, |U_{nt}(\boldsymbol{\beta}^*, \boldsymbol{\alpha}_N^*) - U_{nt}(\boldsymbol{\beta}^*, Q^*)| \\ &= o(J_N^{1/2} N^{-1/2} \|\boldsymbol{v}\|_2).\end{aligned}$$

By Lemma F.1, Regularity Condition 7.B-7.D, and Equation (7.2), we have

$$\begin{aligned}|I_{14}| &\leq \frac{1}{N} \sum_{n=1}^{N} \sum_{t,t'=1}^{T_n} |\tilde{w}_{ntt'} - w_{ntt'}| \, |\partial U_{nt}(\boldsymbol{\beta}^*, \boldsymbol{\alpha}_N^*)| \, |(U_{nt}(\boldsymbol{\beta}^*, \boldsymbol{\alpha}_N^*) - U_{nt}(\boldsymbol{\beta}^*, Q^*))| \\ &= o_p(J_N N^{-1} \|\boldsymbol{v}\|_2).\end{aligned}$$

In sum,

$$\begin{aligned}I_1 &= O_p(\Delta_N J_N^{1/2} N^{-1/2} \|\boldsymbol{v}\|_2) + O_p(\Delta_N J_N^{1/2} N^{-1/2} \|\boldsymbol{v}\|_2) \\ &\quad + O_p(\Delta_N J_N^{1/2} N^{-1/2} \|\boldsymbol{v}\|_2) + O_p(J_N N^{-1} \|\boldsymbol{v}\|_2) \\ &= O_p(J_N N^{-1} \|\boldsymbol{v}\|_2). \qquad (7.18)\end{aligned}$$

Let

$$\partial^2 U_{nt}(\boldsymbol{\beta}, \boldsymbol{\alpha}_N) = \frac{\partial^2}{\partial \Delta_N^2} U_{nt}(\boldsymbol{\beta} + \Delta_N \boldsymbol{v}_{\boldsymbol{\beta}}, \boldsymbol{\alpha}_N + \Delta_N \boldsymbol{v}_{\boldsymbol{\alpha}}) \Big|_{\Delta_N = 0}, \qquad (7.19)$$

for each n and t. We have

$$\begin{aligned}\Delta_N^{-2} I_2 &= \frac{1}{N} \sum_{n=1}^{N} \sum_{t,t'=1}^{T_n} \tilde{w}_{ntt'} \left(\partial U_{nt}(\boldsymbol{\beta}^*, \boldsymbol{\alpha}_N^*) \partial U_{nt'}(\boldsymbol{\beta}^*, \boldsymbol{\alpha}_N^*) + \partial^2 U_{nt}(\boldsymbol{\beta}^*, \boldsymbol{\alpha}_N^*) U_{nt'}(\boldsymbol{\beta}^*, \boldsymbol{\alpha}_N^*) \right) \\ &= I_{21} + I_{22} + I_{23} + I_{24} + I_{25} + I_{26},\end{aligned}$$

where

$$I_{21} = \frac{1}{N}\sum_{n=1}^{N}\sum_{t,t'=1}^{T_n} w_{ntt'}\partial U_{nt}(\boldsymbol{\beta}^*, \boldsymbol{\alpha}_N^*)\partial U_{nt'}(\boldsymbol{\beta}^*, \boldsymbol{\alpha}_N^*)$$

$$I_{22} = \frac{1}{N}\sum_{n=1}^{N}\sum_{t,t'=1}^{T_n} (\tilde{w}_{ntt'} - w_{ntt'})\partial U_{nt}(\boldsymbol{\beta}^*, \boldsymbol{\alpha}_N^*)\partial U_{nt'}(\boldsymbol{\beta}^*, \boldsymbol{\alpha}_N^*)$$

$$I_{23} = \frac{1}{N}\sum_{n=1}^{N}\sum_{t,t'=1}^{T_n} w_{ntt'}\partial^2 U_{nt}(\boldsymbol{\beta}^*, \boldsymbol{\alpha}_N^*)U_{nt'}(\boldsymbol{\beta}^*, Q^*)$$

$$I_{24} = \frac{1}{N}\sum_{n=1}^{N}\sum_{t,t'=1}^{T_n} w_{ntt'}\partial^2 U_{nt}(\boldsymbol{\beta}^*, \boldsymbol{\alpha}_N^*)\left(U_{nt'}\left(\boldsymbol{\beta}^*, \boldsymbol{\alpha}_N^*\right) - U_{nt'}(\boldsymbol{\beta}^*, Q^*)\right)$$

$$I_{25} = \frac{1}{N}\sum_{n=1}^{N}\sum_{t,t'=1}^{T_n} (\tilde{w}_{ntt'} - w_{ntt'})\partial^2 U_{nt}(\boldsymbol{\beta}^*, \boldsymbol{\alpha}_N^*)U_{nt'}(\boldsymbol{\beta}^*, Q^*)$$

and

$$I_{26} = \frac{1}{N}\sum_{n=1}^{N}\sum_{t,t'=1}^{T_n} (\tilde{w}_{ntt'} - w_{ntt'})\partial^2 U_{nt}(\boldsymbol{\beta}^*, \boldsymbol{\alpha}_N^*)\left(U_{nt'}(\boldsymbol{\beta}^*, \boldsymbol{\alpha}_N^*) - U_{nt'}(\boldsymbol{\beta}^*, Q^*)\right).$$

By Lemma F.1, we have $I_{21} = O(\|\boldsymbol{v}\|_2^2)$. By Lemma F.1 and Equation (7.2), we have

$$\begin{aligned}|I_{22}| &\leq \frac{1}{N}\sum_{n=1}^{N}\sum_{t,t'=1}^{T_n} |\tilde{w}_{ntt'} - w_{ntt'}|\,|\partial U_{nt}(\boldsymbol{\beta}^*, \boldsymbol{\alpha}_N^*)|\,|\partial U_{nt'}(\boldsymbol{\beta}^*, \boldsymbol{\alpha}_N^*)| \\ &= O_p(J_N^{1/2}N^{-1/2}\|\boldsymbol{v}\|_2^2).\end{aligned}$$

By Lemma F.5, $I_{23} = O_p(J_N^{1/2}N^{-1/2}\|\boldsymbol{v}\|_2^2)$. By Lemma F.2 and Regularity Condition 7.B-7.D,

$$\begin{aligned}|I_{24}| &\leq \frac{1}{N}\sum_{n=1}^{N}\sum_{t,t'=1}^{T_n} |w_{ntt'}|\left|\partial^2 U_{nt}(\boldsymbol{\beta}^*, \boldsymbol{\alpha}_N^*)\right|\left|U_{nt'}\left(\boldsymbol{\beta}^*, \boldsymbol{\alpha}_N^*\right) - U_{nt'}(\boldsymbol{\beta}^*, Q^*)\right| \\ &= O(J_N N^{-1}\|\boldsymbol{v}\|_2^2),\end{aligned}$$

By Lemma F.2, Equation (7.2) and (7.16), we have

$$\begin{aligned}|I_{25}| &\leq \frac{1}{N}\sum_{n=1}^{N}\sum_{t,t'=1}^{T_n} |\tilde{w}_{ntt'} - w_{ntt'}|\left|\partial^2 U_{nt}(\boldsymbol{\beta}^*, \boldsymbol{\alpha}_N^*)\right|\left|U_{nt'}(\boldsymbol{\beta}^*, Q^*)\right| \\ &\leq O_p(J_N^{1/2}N^{-1/2}\|\boldsymbol{v}\|_2^2).\end{aligned}$$

By Lemma F.2, Regularity Condition 7.B-7.D, and Equation (7.2), we have

$$\begin{aligned}|I_{26}| &\leq \frac{1}{N}\sum_{n=1}^{N}\sum_{t,t'=1}^{T_n} |\tilde{w}_{ntt'} - w_{ntt'}| \left|\partial^2 U_{nt}(\boldsymbol{\beta}^*, \boldsymbol{\alpha}_N^*)\right| |U_{nt'}(\boldsymbol{\beta}^*, \boldsymbol{\alpha}_N^*) - U_{nt'}(\boldsymbol{\beta}^*, Q^*)| \\ &= O(J_N^{3/2} N^{-3/2}\|\boldsymbol{v}\|_2^2).\end{aligned}$$

By the condition that $J_N^{1/2}N^{-1/2} = o(1)$, we have

$$\begin{aligned}I_2 &= O(\Delta_N^2\|\boldsymbol{v}\|_2^2) + O_p(\Delta_N^2 J_N^{1/2} N^{-1/2}\|\boldsymbol{v}\|_2^2) \\ &\quad + O_p(\Delta_N^2 J_N^{1/2} N^{-1/2}\|\boldsymbol{v}\|_2^2) + O(\Delta_N^2 J_N N^{-1}\|\boldsymbol{v}\|_2^2) \\ &\quad + O_p(\Delta_N^2 J_N^{1/2} N^{-1/2}\|\boldsymbol{v}\|_2^2) + O_p(\Delta_N^2 J_N^{3/2} N^{-3/2}\|\boldsymbol{v}\|_2^2) \\ &= O_p(J_N N^{-1}\|\boldsymbol{v}\|_2^2). \end{aligned} \tag{7.20}$$

Let

$$\partial^3 U_{nt}(\boldsymbol{\beta}, \boldsymbol{\alpha}_N) = \frac{\partial^3}{\partial\Delta_N^3} U_{nt}(\boldsymbol{\beta} + \Delta_N \boldsymbol{v}_{\boldsymbol{\beta}}, \boldsymbol{\alpha}_N + \Delta_N \boldsymbol{v}_{\boldsymbol{\alpha}})\Big|_{\Delta_N=0}, \tag{7.21}$$

for each n and t. We have

$$\begin{aligned}6\Delta_N^{-3} I_3 &= \frac{3}{N}\sum_{n=1}^{N}\sum_{t,t'=1}^{T_n} \tilde{w}_{ntt'}\partial U_{nt}(\breve{\boldsymbol{\beta}}, \breve{\boldsymbol{\alpha}}_N)\partial^2 U_{nt'}(\breve{\boldsymbol{\beta}}, \breve{\boldsymbol{\alpha}}_N) \\ &\quad + \frac{1}{N}\sum_{n=1}^{N}\sum_{t,t'=1}^{T_n} \tilde{w}_{ntt'}\partial^3 U_{nt}(\breve{\boldsymbol{\beta}}, \breve{\boldsymbol{\alpha}}_N) U_{nt'}(\breve{\boldsymbol{\beta}}, \breve{\boldsymbol{\alpha}}_N) \\ &= I_{31} + I_{32} + I_{33} + I_{34},\end{aligned}$$

where

$$\begin{aligned}I_{31} &= \frac{3}{N}\sum_{n=1}^{N}\sum_{t,t'=1}^{T_n} w_{ntt'}\partial U_{nt}(\breve{\boldsymbol{\beta}}, \breve{\boldsymbol{\alpha}}_N)\partial^2 U_{nt'}(\breve{\boldsymbol{\beta}}, \breve{\boldsymbol{\alpha}}_N) \\ I_{32} &= \frac{3}{N}\sum_{n=1}^{N}\sum_{t,t'=1}^{T_n} (\tilde{w}_{ntt'} - w_{ntt'})\partial U_{nt}(\breve{\boldsymbol{\beta}}, \breve{\boldsymbol{\alpha}}_N)\partial^2 U_{nt'}(\breve{\boldsymbol{\beta}}, \breve{\boldsymbol{\alpha}}_N) \\ I_{33} &= \frac{1}{N}\sum_{n=1}^{N}\sum_{t,t'=1}^{T_n} w_{ntt'}\partial^3 U_{nt}(\breve{\boldsymbol{\beta}}, \breve{\boldsymbol{\alpha}}_N) U_{nt'}(\breve{\boldsymbol{\beta}}, \breve{\boldsymbol{\alpha}}_N)\end{aligned}$$

and

$$I_{34} = \frac{1}{N}\sum_{n=1}^{N}\sum_{t,t'=1}^{T_n}(\tilde{w}_{ntt'} - w_{ntt'})\partial^3 U_{nt}(\breve{\boldsymbol{\beta}}, \breve{\boldsymbol{\alpha}}_N)U_{nt'}(\breve{\boldsymbol{\beta}}, \breve{\boldsymbol{\alpha}}_N).$$

By Lemma F.1 and F.2, we have

$$\begin{aligned}|I_{31}| &\le \frac{3}{N}\sum_{n=1}^{N}\sum_{t,t'=1}^{T_n}|w_{ntt'}|\left|\partial U_{nt}(\breve{\boldsymbol{\beta}}, \breve{\boldsymbol{\alpha}}_N)\right|\left|\partial^2 U_{nt'}(\breve{\boldsymbol{\beta}}, \breve{\boldsymbol{\alpha}}_N)\right| \\ &= O_p(\|\boldsymbol{v}\|_2^3).\end{aligned}$$

By Equation (7.2), and Lemma F.1 and F.2, we have

$$\begin{aligned}|I_{32}| &\le \frac{3}{N}\sum_{n=1}^{N}\sum_{t,t'=1}^{T_n}|\tilde{w}_{ntt'} - w_{ntt'}|\left|\partial U_{nt}(\breve{\boldsymbol{\beta}}, \breve{\boldsymbol{\alpha}}_N)\right|\left|\partial^2 U_{nt'}(\breve{\boldsymbol{\beta}}, \breve{\boldsymbol{\alpha}}_N)\right| \\ &= O_p(J_N^{1/2}N^{-1/2}\|\boldsymbol{v}\|_2^3).\end{aligned}$$

By Regularity Condition 7.B-7.D and 7.H, Equation (7.16) and $J_N N^{-1} = o(1)$, we have

$$\begin{aligned}\left|U_{nt}(\breve{\boldsymbol{\beta}}, \breve{\boldsymbol{\alpha}}_N)\right| \le\ & \left|U_{nt}(\breve{\boldsymbol{\beta}}, \breve{\boldsymbol{\alpha}}_N) - U_{nt}(\boldsymbol{\beta}^*, \boldsymbol{\alpha}_N^*)\right| + |U_{nt}(\boldsymbol{\beta}^*, \boldsymbol{\alpha}_N^*) - U_{nt}(\boldsymbol{\beta}^*, Q^*)| \\ & + |U_{nt}(\boldsymbol{\beta}^*, Q^*)| \\ \le\ & L_{nt} \times (J_N^{1/2}N^{-1/2})\|\boldsymbol{v}\|_2 + o(J_N N^{-1}) + O_p(1) \\ =\ & O_p(1).\end{aligned}$$

By Lemma F.3, we have

$$\begin{aligned}|I_{33}| &\le \frac{1}{N}\sum_{n=1}^{N}\sum_{t,t'=1}^{T_n}|w_{ntt'}|\left|\partial^3 U_{nt}(\breve{\boldsymbol{\beta}}, \breve{\boldsymbol{\alpha}}_N)\right|\left|U_{nt'}(\breve{\boldsymbol{\beta}}, \breve{\boldsymbol{\alpha}}_N)\right| \\ &= O(\|\boldsymbol{v}\|_2^3).\end{aligned}$$

By Lemma F.3 and Equation (7.2), we have

$$\begin{aligned}|I_{34}| &\le \frac{1}{N}\sum_{n=1}^{N}\sum_{t,t'=1}^{T_n}|\tilde{w}_{ntt'} - w_{ntt'}|\left|\partial^3 U_{nt}(\breve{\boldsymbol{\beta}}, \breve{\boldsymbol{\alpha}}_N)\right|\left|U_{nt'}(\breve{\boldsymbol{\beta}}, \breve{\boldsymbol{\alpha}}_N)\right| \\ &= O_p(J_N N^{-1}\|\boldsymbol{v}\|_2^3)\end{aligned}$$

Therefore,

$$\begin{aligned}
I_3 &= O(\Delta_N^3 \|\boldsymbol{v}\|_2^3) + O_p(\Delta_N^3 J_N^{1/2} N^{-1/2} \|\boldsymbol{v}\|_2^3) \\
&\quad + O_p(\Delta_N^3 \|\boldsymbol{v}\|_2^3) + O(\Delta_N^3 J_N N^{-1} \|\boldsymbol{v}\|_2^3) \\
&= O_p(J_N^{3/2} N^{-3/2} \|\boldsymbol{v}\|_2^3). \qquad (7.22)
\end{aligned}$$

By Equation (7.18), (7.20) and (7.22), we find that

$$\tilde{\mathcal{Q}}(\boldsymbol{\beta}^* + \Delta_N \boldsymbol{v}_{\boldsymbol{\beta}}, \boldsymbol{\alpha}_N^* + \Delta_N \boldsymbol{v}_{\boldsymbol{\alpha}}) - \tilde{\mathcal{Q}}(\boldsymbol{\beta}^*, \boldsymbol{\alpha}_N^*)$$

is dominated by the term

$$\Delta_N^2 I_{21} = \Delta_N^2 \frac{1}{N} \sum_{n=1}^{N} \sum_{t,t'=1}^{T_n} w_{ntt'} \partial U_{nt}(\boldsymbol{\beta}^*, \boldsymbol{\alpha}_N^*) \partial U_{nt}(\boldsymbol{\beta}^*, \boldsymbol{\alpha}_N^*) > 0$$

by allowing $\|\boldsymbol{v}\|_2 = C$ to be large enough. It follows that

$$\tilde{\mathcal{Q}}(\boldsymbol{\beta}^* + \Delta_N \boldsymbol{v}_{\boldsymbol{\beta}}, \boldsymbol{\alpha}_N^* + \Delta_N \boldsymbol{v}_{\boldsymbol{\alpha}}) - \tilde{\mathcal{Q}}(\boldsymbol{\beta}^*, \boldsymbol{\alpha}_N^*)$$

converges to a positive number in probability. □

7.4 Convergence of the Plug-in Weighting Matrices

In Theorem 7.3.1, it is required that, for each n, $\tilde{\boldsymbol{W}}_n$ converges to $\boldsymbol{W}_n$ element-wise at rate $N^{1/2} J_N^{-1/2}$; see Regularity Condition 7.J. We may use arbitrary non-random weighting matrices to obtain the initial estimates $(\tilde{\boldsymbol{\beta}}, \tilde{\boldsymbol{\alpha}}_N)$ which converges to $(\boldsymbol{\beta}^*, \boldsymbol{\alpha}_N^*)$ at rate $N^{1/2} J_N^{-1/2}$. According to the following theorem, the plug-in weighting matrix $\boldsymbol{W}_n(\tilde{\boldsymbol{\beta}}, \tilde{\boldsymbol{\alpha}}_N)$, denoted by $\tilde{\boldsymbol{W}}_n$, satisfies Regularity Condition 7.J.

Let $(\boldsymbol{\beta}_0, \boldsymbol{\alpha}_{N,0}) \in \mathbb{R}^p \times \mathcal{M}_{J_N+1}$ such that

$$\|\tilde{\boldsymbol{\alpha}}_N - \boldsymbol{\alpha}_{N,0}\|_2^2 + \left\|\tilde{\boldsymbol{\beta}} - \boldsymbol{\beta}_0\right\|_2^2 = O_p(J_N N^{-1})$$

and Q_0 is a probability measure defined on $\mathcal{B}$ such that $\boldsymbol{\alpha}_{N,0} = \int_{\mathcal{B}} \boldsymbol{P}(b)\mathrm{d}Q_0$. Also for each n, let

$$\boldsymbol{W}_n(\boldsymbol{\beta}, Q) = \boldsymbol{V}_n^{-1/2}(\boldsymbol{\beta}, Q)\boldsymbol{R}_n^{-1}\boldsymbol{V}_n^{-1/2}(\boldsymbol{\beta}, Q),$$

and $\boldsymbol{V}_n(\boldsymbol{\beta}, Q)$ is a $T_n \times T_n$ diagonal matrix whose t^{th} diagonal element is

$$\begin{aligned} V_{nt}(\boldsymbol{\beta}, Q) = \sigma \int_{\mathcal{B}} h \circ g^{-1}(\boldsymbol{X}_{nt}^{\mathrm{T}}\boldsymbol{\beta} + Z_{nt}b)\mathrm{d}Q + \int_{\mathcal{B}} \left(g^{-1}(\boldsymbol{X}_{nt}^{\mathrm{T}}\boldsymbol{\beta} + Z_{nt}b)\right)^2 \mathrm{d}Q \\ - \left(\int_{\mathcal{B}} g^{-1}(\boldsymbol{X}_{nt}^{\mathrm{T}}\boldsymbol{\beta} + Z_{nt}b)\mathrm{d}Q\right)^2. \end{aligned} \tag{7.23}$$

Here neither $\boldsymbol{\beta}_0$ nor Q_0 is required to be the true parameter values.

Theorem 7.4.1 (Consistency of $\boldsymbol{W}_n(\boldsymbol{\beta}, \boldsymbol{\alpha}_N)$)**.**
Assume that Regularity Condition 7.A-7.G are satisfied. Further assume that the initial estimator $(\tilde{\boldsymbol{\beta}}, \tilde{\boldsymbol{\alpha}}_N)$ converges to $(\boldsymbol{\beta}_0, \boldsymbol{\alpha}_{N,0}) \in \mathbb{R}^p \times \mathcal{M}_{J_N+1}$ in the sense that

$$\|\tilde{\boldsymbol{\alpha}}_N - \boldsymbol{\alpha}_{N,0}\|_2^2 + \|\tilde{\boldsymbol{\beta}} - \boldsymbol{\beta}_0\|_2^2 = O_p(J_N N^{-1}).$$

Then, for each $n \in \{1, \ldots, N\}$, $\tilde{\boldsymbol{W}}_n$ converges in probability to $\boldsymbol{W}_n(\boldsymbol{\beta}_0, Q_0)$ element-wise at rate $J_N^{-1/2}N^{1/2}$, as the sample size N goes to infinity.

Proof. For each $n \in \{1, \ldots, N\}$, we have

$$\begin{aligned} &\boldsymbol{W}_n(\tilde{\boldsymbol{\beta}}, \tilde{\boldsymbol{\alpha}}_N) - \boldsymbol{W}_n(\boldsymbol{\beta}_0, Q_0) \\ &= \boldsymbol{V}_n^{-1/2}(\tilde{\boldsymbol{\beta}}, \tilde{\boldsymbol{\alpha}}_N)\boldsymbol{R}_n^{-1}\boldsymbol{V}_n^{-1/2}(\tilde{\boldsymbol{\beta}}, \tilde{\boldsymbol{\alpha}}_N) - \boldsymbol{V}_n^{-1/2}(\boldsymbol{\beta}_0, Q_0)\boldsymbol{R}_n^{-1}\boldsymbol{V}_n^{-1/2}(\boldsymbol{\beta}_0, Q_0) \\ &= \boldsymbol{R}_n^{-1/2}\left(\boldsymbol{V}_n^{-1}(\tilde{\boldsymbol{\beta}}, \tilde{\boldsymbol{\alpha}}_N) - \boldsymbol{V}_n^{-1}(\boldsymbol{\beta}_0, Q_0)\right)\boldsymbol{R}_n^{-1/2}, \end{aligned} \tag{7.24}$$

where the diagonal elements of $\boldsymbol{V}_n(\boldsymbol{\beta}, \boldsymbol{\alpha}_N)$ are

$$V_{nt}(\boldsymbol{\beta}, \boldsymbol{\alpha}_N) = \sigma \times \boldsymbol{a}_{nt}^{\mathrm{T}}(\boldsymbol{\beta})\boldsymbol{\alpha}_N + \boldsymbol{c}_{ntt}^{\mathrm{T}}(\boldsymbol{\beta})\boldsymbol{\alpha}_N - \left(\boldsymbol{\phi}_{nt}^{\mathrm{T}}(\boldsymbol{\beta})\boldsymbol{\alpha}_N\right)^2,$$

and $\boldsymbol{\phi}_{nt}(\boldsymbol{\beta})$, $\boldsymbol{a}_{nt}(\boldsymbol{\beta})$ and $\boldsymbol{c}_{ntt}(\boldsymbol{\beta})$ are defined in Equation (7.3), (7.10) and (7.13) correspondingly.

For each $n \in \{1, \dots, N\}$ and $t \in \{1, \dots, T_n\}$, we consider the gradient function of $V_{nt}(\boldsymbol{\beta}, \boldsymbol{\alpha}_N)$ at $(\boldsymbol{\beta}_0, \boldsymbol{\alpha}_{N,0})$ along direction $(\boldsymbol{v}_{\boldsymbol{\beta}}, \boldsymbol{v}_{\boldsymbol{\alpha}})$ to $(\tilde{\boldsymbol{\beta}}, \tilde{\boldsymbol{\alpha}}_N)$, i.e.,

$$\begin{aligned} & V_{nt}(\boldsymbol{\beta}_0 + \Delta \boldsymbol{v}_{\boldsymbol{\beta}}, \boldsymbol{\alpha}_{N,0} + \Delta \boldsymbol{v}_{\boldsymbol{\alpha}}) \\ = & \, \sigma \times \boldsymbol{a}_{nt}^{\mathrm{T}}(\boldsymbol{\beta}_0 + \Delta \boldsymbol{v}_{\boldsymbol{\beta}}) \left(\boldsymbol{\alpha}_{N,0} + \Delta \boldsymbol{v}_{\boldsymbol{\alpha}}\right) + \boldsymbol{c}_{ntt}^{\mathrm{T}}(\boldsymbol{\beta}_0 + \Delta \boldsymbol{v}_{\boldsymbol{\beta}}) \left(\boldsymbol{\alpha}_{N,0} + \Delta \boldsymbol{v}_{\boldsymbol{\alpha}}\right) \\ & - \left(\boldsymbol{\phi}_{nt}^{\mathrm{T}}(\boldsymbol{\beta}_0 + \Delta \boldsymbol{v}_{\boldsymbol{\beta}}) \left(\boldsymbol{\alpha}_{N,0} + \Delta \boldsymbol{v}_{\boldsymbol{\alpha}}\right)\right)^2, \end{aligned}$$

where $\Delta = \|\tilde{\boldsymbol{\alpha}}_N - \boldsymbol{\alpha}_{N,0}\|_2^2 + \|\tilde{\boldsymbol{\beta}} - \boldsymbol{\beta}_0\|_2$.

By Taylor's expansion at $\Delta = 0$, we have

$$\begin{aligned} & V_{nt}(\boldsymbol{\beta}_0 + \Delta \boldsymbol{v}_{\boldsymbol{\beta}}, \boldsymbol{\alpha}_{N,0} + \Delta \boldsymbol{v}_{\boldsymbol{\alpha}}) - V_{nt}(\boldsymbol{\beta}_0, \boldsymbol{\alpha}_{N,0}) \\ = & \, \partial V_{nt}(\boldsymbol{\beta}_0, \boldsymbol{\alpha}_{N,0}) \Delta + \frac{1}{2} \partial^2 V_{nt}(\breve{\boldsymbol{\beta}}, \breve{\boldsymbol{\alpha}}_N) \Delta^2, \end{aligned}$$

where

$$\begin{aligned} & \partial V_{nt}(\boldsymbol{\beta}_0, \boldsymbol{\alpha}_{N,0}) \\ = & \, \frac{\partial}{\partial \Delta} V_{nt}(\boldsymbol{\beta}_0 + \Delta \boldsymbol{v}_{\boldsymbol{\beta}}, \boldsymbol{\alpha}_{N,0} + \Delta \boldsymbol{v}_{\boldsymbol{\alpha}}) \Big|_{\Delta = 0} \\ = & \, \left(\sigma \boldsymbol{a}_{nt}(\boldsymbol{\beta}_0) + \boldsymbol{c}_{ntt}(\boldsymbol{\beta}_0) - 2 \boldsymbol{\phi}_{nt}^{\mathrm{T}}(\boldsymbol{\beta}_0) \boldsymbol{\alpha}_{N,0} \boldsymbol{\phi}_{nt}(\boldsymbol{\beta}_0)\right)^{\mathrm{T}} \boldsymbol{v}_{\boldsymbol{\alpha}} \\ & + \left(\sigma \dot{\boldsymbol{a}}_{nt}^{\mathrm{T}}(\boldsymbol{\beta}) \boldsymbol{\alpha}_N + \dot{\boldsymbol{c}}_{ntt}^{\mathrm{T}}(\boldsymbol{\beta}) \boldsymbol{\alpha}_N - 2 \boldsymbol{\phi}_{nt}^{\mathrm{T}}(\boldsymbol{\beta}) \boldsymbol{\alpha}_N \dot{\boldsymbol{\phi}}_{nt}^{\mathrm{T}}(\boldsymbol{\beta}) \boldsymbol{\alpha}_N\right)^{\mathrm{T}} \boldsymbol{X}_{nt}^{\mathrm{T}} \boldsymbol{v}_{\boldsymbol{\beta}} \end{aligned} \tag{7.25}$$

and

$$\begin{aligned} & \partial^2 V_{nt}(\breve{\boldsymbol{\beta}}, \breve{\boldsymbol{\alpha}}_N) \\ = & \, \frac{\partial^2}{\partial \Delta^2} V_{nt}(\boldsymbol{\beta}_0 + \Delta \boldsymbol{v}_{\boldsymbol{\beta}}, \boldsymbol{\alpha}_{N,0} + \Delta \boldsymbol{v}_{\boldsymbol{\alpha}}) \Big|_{\Delta = \breve{\Delta}} \\ = & \, -2 \boldsymbol{v}_{\boldsymbol{\alpha}} \boldsymbol{\phi}_{nt}(\breve{\boldsymbol{\beta}}) \boldsymbol{\phi}_{nt}^{\mathrm{T}}(\breve{\boldsymbol{\beta}}) \boldsymbol{v}_{\boldsymbol{\alpha}} \\ & + 2 \boldsymbol{v}_{\boldsymbol{\beta}}^{\mathrm{T}} \boldsymbol{X}_{nt} \left(\sigma \dot{\boldsymbol{a}}_{nt}^{\mathrm{T}}(\breve{\boldsymbol{\beta}}) + \dot{\boldsymbol{c}}_{ntt}^{\mathrm{T}}(\breve{\boldsymbol{\beta}}) - 2 \dot{\boldsymbol{\phi}}_{nt}^{\mathrm{T}}(\breve{\boldsymbol{\beta}}) \breve{\boldsymbol{\alpha}}_N \dot{\boldsymbol{\phi}}_{nt}^{\mathrm{T}}(\breve{\boldsymbol{\beta}}) - 2 \boldsymbol{\phi}_{nt}^{\mathrm{T}}(\breve{\boldsymbol{\beta}}) \breve{\boldsymbol{\alpha}}_N \ddot{\boldsymbol{\phi}}_{nt}^{\mathrm{T}}(\breve{\boldsymbol{\beta}})\right) \boldsymbol{v}_{\boldsymbol{\alpha}} \\ & + \left(\sigma \ddot{\boldsymbol{a}}_{nt}^{\mathrm{T}}(\breve{\boldsymbol{\beta}}) \breve{\boldsymbol{\alpha}}_N + \ddot{\boldsymbol{c}}_{ntt}^{\mathrm{T}}(\breve{\boldsymbol{\beta}}) \breve{\boldsymbol{\alpha}}_N - 2 \dot{\boldsymbol{\phi}}_{nt}^{\mathrm{T}}(\breve{\boldsymbol{\beta}}) \breve{\boldsymbol{\alpha}}_N \dot{\boldsymbol{\phi}}_{nt}^{\mathrm{T}}(\breve{\boldsymbol{\beta}}) \breve{\boldsymbol{\alpha}}_N - 2 \boldsymbol{\phi}_{nt}^{\mathrm{T}}(\breve{\boldsymbol{\beta}}) \breve{\boldsymbol{\alpha}}_N \ddot{\boldsymbol{\phi}}_{nt}^{\mathrm{T}}(\breve{\boldsymbol{\beta}}) \breve{\boldsymbol{\alpha}}_N\right) \\ & \times \boldsymbol{v}_{\boldsymbol{\beta}}^{\mathrm{T}} \boldsymbol{X}_{nt} \boldsymbol{X}_{nt}^{\mathrm{T}} \boldsymbol{v}_{\boldsymbol{\beta}}. \end{aligned}$$

and $\breve{\boldsymbol{\beta}} = \boldsymbol{\beta}_0 + \breve{\Delta} \boldsymbol{v}_{\boldsymbol{\beta}}$ and $\breve{\boldsymbol{\alpha}} = \boldsymbol{\alpha}_0 + \breve{\Delta} \boldsymbol{v}_{\boldsymbol{\alpha}}$, and $(\breve{\boldsymbol{\beta}}, \breve{\boldsymbol{\alpha}})$ is on the line segment between $(\tilde{\boldsymbol{\beta}}, \tilde{\boldsymbol{\alpha}}_N)$ and $(\boldsymbol{\beta}_0, \boldsymbol{\alpha}_{N,0})$.

For each $n \in \{1, \ldots, N\}$ and $t \in \{1, \ldots, T_n\}$, we have

$$\begin{aligned}
& \left|\partial V_{nt}(\boldsymbol{\beta}_0, \boldsymbol{\alpha}_{N,0})\right| \\
\leq\ & \left|\left(\sigma \boldsymbol{a}_{nt}(\boldsymbol{\beta}_0) + \boldsymbol{c}_{ntt}(\boldsymbol{\beta}_0) - 2\boldsymbol{\phi}_{nt}^{\mathrm{T}}(\boldsymbol{\beta}_0)\boldsymbol{\alpha}_{N,0}\boldsymbol{\phi}_{nt}(\boldsymbol{\beta}_0)\right)^{\mathrm{T}} \boldsymbol{v}_{\boldsymbol{\alpha}}\right| \\
& + \left|\left(\sigma \dot{\boldsymbol{a}}_{nt}^{\mathrm{T}}(\boldsymbol{\beta})\boldsymbol{\alpha}_N + \dot{\boldsymbol{c}}_{ntt}^{\mathrm{T}}(\boldsymbol{\beta})\boldsymbol{\alpha}_N - 2\boldsymbol{\phi}_{nt}^{\mathrm{T}}(\boldsymbol{\beta})\boldsymbol{\alpha}_N\dot{\boldsymbol{\phi}}_{nt}^{\mathrm{T}}(\boldsymbol{\beta})\boldsymbol{\alpha}_N\right)^{\mathrm{T}} \boldsymbol{X}_{nt}^{\mathrm{T}}\boldsymbol{v}_{\boldsymbol{\beta}}\right| \\
\leq\ & \left\|\sigma \boldsymbol{a}_{nt}(\boldsymbol{\beta}_0) + \boldsymbol{c}_{ntt}(\boldsymbol{\beta}_0) - 2\boldsymbol{\phi}_{nt}^{\mathrm{T}}(\boldsymbol{\beta}_0)\boldsymbol{\alpha}_{N,0}\boldsymbol{\phi}_{nt}(\boldsymbol{\beta}_0)\right\|_2 \\
& + \left\|\sigma \dot{\boldsymbol{a}}_{nt}^{\mathrm{T}}(\boldsymbol{\beta})\boldsymbol{\alpha}_N + \dot{\boldsymbol{c}}_{ntt}^{\mathrm{T}}(\boldsymbol{\beta})\boldsymbol{\alpha}_N - 2\boldsymbol{\phi}_{nt}^{\mathrm{T}}(\boldsymbol{\beta})\boldsymbol{\alpha}_N\dot{\boldsymbol{\phi}}_{nt}^{\mathrm{T}}(\boldsymbol{\beta})\boldsymbol{\alpha}_N\right\|_2 \left\|\boldsymbol{X}_{nt}\right\|_2,
\end{aligned}$$

which is bounded, according to Lemma 7.2.1-7.2.6 and Regularity Condition 7.A.

On the other hand, for each $n \in \{1, \ldots, N\}$ and $t \in \{1, \ldots, T_n\}$, we have

$$\begin{aligned}
& \left|\partial^2 V_{nt}(\breve{\boldsymbol{\beta}}, \breve{\boldsymbol{\alpha}}_N)\right| \\
\leq\ & \left|2\boldsymbol{v}_{\boldsymbol{\alpha}}\boldsymbol{\phi}_{nt}(\breve{\boldsymbol{\beta}})\boldsymbol{\phi}_{nt}^{\mathrm{T}}(\breve{\boldsymbol{\beta}})\boldsymbol{v}_{\boldsymbol{\alpha}}\right| \\
& + \left|2\boldsymbol{v}_{\boldsymbol{\beta}}^{\mathrm{T}}\boldsymbol{X}_{nt}\left(\sigma\dot{\boldsymbol{a}}_{nt}^{\mathrm{T}}(\breve{\boldsymbol{\beta}}) + \dot{\boldsymbol{c}}_{ntt}^{\mathrm{T}}(\breve{\boldsymbol{\beta}}) - 2\dot{\boldsymbol{\phi}}_{nt}^{\mathrm{T}}(\breve{\boldsymbol{\beta}})\breve{\boldsymbol{\alpha}}_N\dot{\boldsymbol{\phi}}_{nt}^{\mathrm{T}}(\breve{\boldsymbol{\beta}}) - 2\boldsymbol{\phi}_{nt}^{\mathrm{T}}(\breve{\boldsymbol{\beta}})\breve{\boldsymbol{\alpha}}_N\ddot{\boldsymbol{\phi}}_{nt}^{\mathrm{T}}(\breve{\boldsymbol{\beta}})\right)\boldsymbol{v}_{\boldsymbol{\alpha}}\right| \\
& + \left|\sigma\ddot{\boldsymbol{a}}_{nt}^{\mathrm{T}}(\breve{\boldsymbol{\beta}})\breve{\boldsymbol{\alpha}}_N + \ddot{\boldsymbol{c}}_{ntt}^{\mathrm{T}}(\breve{\boldsymbol{\beta}})\breve{\boldsymbol{\alpha}}_N - 2\dot{\boldsymbol{\phi}}_{nt}^{\mathrm{T}}(\breve{\boldsymbol{\beta}})\breve{\boldsymbol{\alpha}}_N\dot{\boldsymbol{\phi}}_{nt}^{\mathrm{T}}(\breve{\boldsymbol{\beta}})\breve{\boldsymbol{\alpha}}_N - 2\boldsymbol{\phi}_{nt}^{\mathrm{T}}(\breve{\boldsymbol{\beta}})\breve{\boldsymbol{\alpha}}_N\ddot{\boldsymbol{\phi}}_{nt}^{\mathrm{T}}(\breve{\boldsymbol{\beta}})\breve{\boldsymbol{\alpha}}_N\right| \\
& \times \boldsymbol{v}_{\boldsymbol{\beta}}^{\mathrm{T}}\boldsymbol{X}_{nt}\boldsymbol{X}_{nt}^{\mathrm{T}}\boldsymbol{v}_{\boldsymbol{\beta}} \\
\leq\ & 2\left\|\boldsymbol{\phi}_{nt}(\breve{\boldsymbol{\beta}})\right\|_2 + 2\left\|\boldsymbol{X}_{nt}\right\|_2\left(\left\|\sigma\dot{\boldsymbol{a}}_{nt}(\breve{\boldsymbol{\beta}})\right\|_2 + \left\|\dot{\boldsymbol{c}}_{ntt}(\breve{\boldsymbol{\beta}})\right\| + 2\left|\dot{\boldsymbol{\phi}}_{nt}^{\mathrm{T}}(\breve{\boldsymbol{\beta}})\breve{\boldsymbol{\alpha}}_N\right|\left\|\dot{\boldsymbol{\phi}}_{nt}(\breve{\boldsymbol{\beta}})\right\|_2\right. \\
& \left. +2\left|\boldsymbol{\phi}_{nt}^{\mathrm{T}}(\breve{\boldsymbol{\beta}})\breve{\boldsymbol{\alpha}}_N\right|\left\|\ddot{\boldsymbol{\phi}}_{nt}(\breve{\boldsymbol{\beta}})\right\|_2\right) + \left(\left|\sigma\ddot{\boldsymbol{a}}_{nt}^{\mathrm{T}}(\breve{\boldsymbol{\beta}})\breve{\boldsymbol{\alpha}}_N\right| + \left|\ddot{\boldsymbol{c}}_{ntt}^{\mathrm{T}}(\breve{\boldsymbol{\beta}})\breve{\boldsymbol{\alpha}}_N\right|\right. \\
& \left. +2\left|\dot{\boldsymbol{\phi}}_{nt}^{\mathrm{T}}(\breve{\boldsymbol{\beta}})\breve{\boldsymbol{\alpha}}_N\dot{\boldsymbol{\phi}}_{nt}^{\mathrm{T}}(\breve{\boldsymbol{\beta}})\breve{\boldsymbol{\alpha}}_N\right| + 2\left|\boldsymbol{\phi}_{nt}^{\mathrm{T}}(\breve{\boldsymbol{\beta}})\breve{\boldsymbol{\alpha}}_N\ddot{\boldsymbol{\phi}}_{nt}^{\mathrm{T}}(\breve{\boldsymbol{\beta}})\breve{\boldsymbol{\alpha}}_N\right|\right) \times \left\|\boldsymbol{X}_{nt}\right\|_2^2,
\end{aligned}$$

which is also bounded, according to Lemma 7.2.1-7.2.6 and Regularity Condition 7.A. Therefore, we have

$$V_{nt}(\tilde{\boldsymbol{\beta}}, \tilde{\boldsymbol{\alpha}}_N) - V_{nt}(\boldsymbol{\beta}_0, \boldsymbol{\alpha}_{N,0}) = O_p(J_N^{1/2}N^{-1/2}).$$

By Regularity Condition 7.E and 7.F, we have

$$|V_{nt}(\boldsymbol{\beta}_0, \boldsymbol{\alpha}_{N,0}) - V_{nt}(\boldsymbol{\beta}_0, Q_0)| = o(J_N N^{-1}).$$

We have that

$$\begin{aligned}
&\left|V_{nt}(\tilde{\boldsymbol{\beta}}, \tilde{\boldsymbol{\alpha}}_N) - V_{nt}(\boldsymbol{\beta}_0, Q_0)\right| \\
\le\ & |V_{nt}(\boldsymbol{\beta}_0, \boldsymbol{\alpha}_{N,0}) - V_{nt}(\boldsymbol{\beta}_0, Q_0)| + \left|V_{nt}(\tilde{\boldsymbol{\beta}}, \tilde{\boldsymbol{\alpha}}_N) - V_{nt}(\boldsymbol{\beta}_0, \boldsymbol{\alpha}_{N,0})\right| \\
=\ & O_p(J_N^{1/2} N^{-1/2}).
\end{aligned}$$

Again, by Taylor's expansion, we have, for every large N, $n \in \{1, \ldots, N\}$ and $t \in \{1, \ldots, T_n\}$,

$$\begin{aligned}
\frac{1}{V_{nt}(\tilde{\boldsymbol{\beta}}, \tilde{\boldsymbol{\alpha}}_N)} &= \frac{1}{V_{nt}(\boldsymbol{\beta}_0, Q_0)} - \frac{1}{V_{nt}^2(\boldsymbol{\beta}_0, Q_0)} \Big(V_{nt}(\tilde{\boldsymbol{\beta}}, \tilde{\boldsymbol{\alpha}}_N) - V_{nt}(\boldsymbol{\beta}_0, Q_0)\Big) \\
&\quad + \frac{1}{\breve{V}_{nt}^3} \Big(V_{nt}(\tilde{\boldsymbol{\beta}}, \tilde{\boldsymbol{\alpha}}_N) - V_{nt}(\boldsymbol{\beta}_0, Q_0)\Big)^2 \\
&= \frac{1}{V_{nt}(\boldsymbol{\beta}_0, Q_0)} + O_p(J_N^{1/2} N^{-1/2}),
\end{aligned} \tag{7.26}$$

where $\breve{V}_{nt}$ is on the line segment between $V_{nt}(\boldsymbol{\beta}_0, Q_0)$ and $V_{nt}(\tilde{\boldsymbol{\beta}}, \tilde{\boldsymbol{\alpha}}_N)$.

By Equation (7.24) and (7.26), we have

$$\boldsymbol{W}_n(\tilde{\boldsymbol{\beta}}, \tilde{\boldsymbol{\alpha}}_N) - \boldsymbol{W}_n(\boldsymbol{\beta}_0, Q_0) = O_p(J_N^{1/2} N^{-1/2}).$$

This completes the proof. □

7.5 Asymptotic Normality in the Generalized Method of Moments

Let $\boldsymbol{D}_n(\boldsymbol{\beta}, \boldsymbol{\alpha}_N)$ be the diagonal matrix whose diagonal elements are $\dot{\boldsymbol{\phi}}_{nt}^{\mathrm{T}}(\boldsymbol{\beta})\boldsymbol{\alpha}_N$, $t = 1, \ldots, T_n$, where $\dot{\boldsymbol{\phi}}_{nt}(\boldsymbol{\beta})$ is defined in Equation (7.4). Then, we have

$$\frac{\partial}{\partial \boldsymbol{\beta}} \boldsymbol{U}_n(\boldsymbol{\beta}, \boldsymbol{\alpha}_N) = \boldsymbol{D}_n(\boldsymbol{\beta}, \boldsymbol{\alpha}_N) \boldsymbol{X}_n^{\mathrm{T}}.$$

From Section 5.4, it is known that the GMM estimator $(\hat{\boldsymbol{\beta}}_{\mathrm{GMM}}, \hat{\boldsymbol{\alpha}}_{N,\mathrm{GMM}})$ is a solution locally to $(\boldsymbol{\beta}^*, \boldsymbol{\alpha}_N^*)$ satisfying that

$$\frac{1}{N} \sum_{n=1}^{N} \boldsymbol{X}_n \boldsymbol{D}_n(\hat{\boldsymbol{\beta}}_{\mathrm{GMM}}, \hat{\boldsymbol{\alpha}}_{N,\mathrm{GMM}}) \tilde{\boldsymbol{W}}_n \boldsymbol{U}_n(\hat{\boldsymbol{\beta}}_{\mathrm{GMM}}, \hat{\boldsymbol{\alpha}}_{N,\mathrm{GMM}}) = 0 \tag{7.27}$$

and

$$\hat{\boldsymbol{\alpha}}_{N,\mathrm{GMM}} = \arg \min_{\boldsymbol{\alpha}_N \in \mathcal{M}_{J_N+1}} \frac{1}{N} \sum_{n=1}^{N} \boldsymbol{U}_n^{\mathrm{T}}(\hat{\boldsymbol{\beta}}_{\mathrm{GMM}}, \boldsymbol{\alpha}_N) \tilde{\boldsymbol{W}}_n \boldsymbol{U}_n(\hat{\boldsymbol{\beta}}_{\mathrm{GMM}}, \boldsymbol{\alpha}_N) \tag{7.28}$$

simultaneously.

For each n, let

$$\boldsymbol{G}_n(\boldsymbol{\beta}, \boldsymbol{\alpha}_N) = \begin{bmatrix} \boldsymbol{\Phi}_n^{\mathrm{T}}(\boldsymbol{\beta}) \\ \boldsymbol{D}_n(\boldsymbol{\beta}, \boldsymbol{\alpha}_N) \boldsymbol{X}_n^{\mathrm{T}} \end{bmatrix} \tag{7.29}$$

where $\boldsymbol{\Phi}_n(\boldsymbol{\beta})$ is a $(J_N + 1) \times T_n$ matrix whose t^{th} column is $\boldsymbol{\phi}_{nt}(\boldsymbol{\beta})$ defined in Equation (7.3). For each n, let

$$\boldsymbol{W}_n^* = \boldsymbol{W}_n(\boldsymbol{\beta}^*, Q^*),$$

and

$$\boldsymbol{D}_n^* = \boldsymbol{D}_n(\boldsymbol{\beta}^*, Q^*),$$

where $\boldsymbol{D}_n(\boldsymbol{\beta}, Q)$ is the diagonal matrix whose diagonal elements $\int_{\mathcal{B}} \dot{g}^{-1}(\boldsymbol{X}_{nt}^{\mathrm{T}}\boldsymbol{\beta} + Z_{nt} b_n) \mathrm{d}Q$, $t = 1, \ldots, T_n$.

Note that the true parameter value $\boldsymbol{\alpha}_N^*$ may on the boundary of $\mathcal{M}_{J_N+1}$, and thus the regularity conditions in [Wilks, 1938] fail and an asymptotic normality result may not be derived from the optimization problem (7.28). Under the regularity conditions listed in Section 7.2, we have the following theorem from Equation (7.27).

Theorem 7.5.1 (Asymptotic Normality in the GMM)**.**
Assume that Regularity Condition 7.A-7.J are satisfied. Further assume that $(\hat{\boldsymbol{\beta}}_{\mathrm{GMM}}, \hat{\boldsymbol{\alpha}}_{N,\mathrm{GMM}})$ *and the initial estimator* $(\tilde{\boldsymbol{\beta}}, \tilde{\boldsymbol{\alpha}}_N)$ *are in a neighbourhood of* $(\boldsymbol{\beta}^*, \boldsymbol{\alpha}_N^*)$ *such that*

$$\|\hat{\boldsymbol{\alpha}}_{N,\mathrm{GMM}} - \boldsymbol{\alpha}_N^*\|_2^2 + \left\|\hat{\boldsymbol{\beta}}_{\mathrm{GMM}} - \boldsymbol{\beta}^*\right\|_2^2 = O_p(J_N N^{-1})$$

and

$$\|\tilde{\boldsymbol{\alpha}}_N - \boldsymbol{\alpha}_N^*\|_2^2 + \left\|\tilde{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\right\|_2^2 = O_p(J_N N^{-1}),$$

as the sample size N goes to infinity. Given a series of working correlation matrix $\{\boldsymbol{R}_n\}_{n=1}^N$, if $J_N N^{-1/3} = o(1)$ as the sample size N goes to infinity, then

$$N^{-1/2}\sum_{n=1}^N \boldsymbol{X}_n\boldsymbol{D}_n(\hat{\boldsymbol{\beta}}_{\mathrm{GMM}}, \hat{\boldsymbol{\alpha}}_{N,\mathrm{GMM}})\tilde{\boldsymbol{W}}_n\boldsymbol{G}_n(\hat{\boldsymbol{\beta}}_{\mathrm{GMM}}, \hat{\boldsymbol{\alpha}}_{N,\mathrm{GMM}})\begin{bmatrix}\hat{\boldsymbol{\alpha}}_{N,\mathrm{GMM}} - \boldsymbol{\alpha}_N^* \\ \hat{\boldsymbol{\beta}}_{\mathrm{GMM}} - \boldsymbol{\beta}^*\end{bmatrix}$$

converges in distribution to a multivariate normal random vector in $\mathbb{R}^p$ with mean zero and covariance matrix $\boldsymbol{\Gamma}$, where the covariance matrix

$$\boldsymbol{\Gamma} = \lim_{N\to\infty}\frac{1}{N}\sum_{n=1}^N \boldsymbol{X}_n\boldsymbol{D}_n^*\boldsymbol{W}_n^*\boldsymbol{\Sigma}_n(\boldsymbol{\beta}^*, Q^*)\boldsymbol{W}_n^*\boldsymbol{D}_n^*\boldsymbol{X}_n^{\mathrm{T}},$$

and, for each n, $\boldsymbol{\Sigma}_n(\boldsymbol{\beta}^*, Q^*)$ is the covariance matrix of $\boldsymbol{Y}_n \mid (\boldsymbol{X}_n, \boldsymbol{Z}_n)$.

Proof. For each n, we have

$$\begin{aligned}&\boldsymbol{X}_n\boldsymbol{D}_n(\hat{\boldsymbol{\beta}}_{\mathrm{GMM}}, \hat{\boldsymbol{\alpha}}_{N,\mathrm{GMM}})\tilde{\boldsymbol{W}}_n\boldsymbol{U}_n(\hat{\boldsymbol{\beta}}_{\mathrm{GMM}}, \hat{\boldsymbol{\alpha}}_{N,\mathrm{GMM}})\\ =\ & I_{n1} + I_{n2} + I_{n3} + I_{n4} + I_{n5},\end{aligned}$$

where

$$\begin{aligned}I_{n1} &= \boldsymbol{X}_n\boldsymbol{D}_n^*\boldsymbol{W}_n^*\boldsymbol{U}_n(\boldsymbol{\beta}^*, Q^*),\\ I_{n2} &= \boldsymbol{X}_n\boldsymbol{D}_n(\hat{\boldsymbol{\beta}}_{\mathrm{GMM}}, \hat{\boldsymbol{\alpha}}_{N,\mathrm{GMM}})\tilde{\boldsymbol{W}}_n\left(\boldsymbol{U}_n(\hat{\boldsymbol{\beta}}_{\mathrm{GMM}}, \hat{\boldsymbol{\alpha}}_{N,\mathrm{GMM}}) - \boldsymbol{U}_n(\boldsymbol{\beta}^*, Q^*)\right),\\ I_{n3} &= \boldsymbol{X}_n\left(\boldsymbol{D}_n(\hat{\boldsymbol{\beta}}_{\mathrm{GMM}}, \hat{\boldsymbol{\alpha}}_{N,\mathrm{GMM}}) - \boldsymbol{D}_n^*\right)\boldsymbol{W}_n^*\boldsymbol{U}_n(\boldsymbol{\beta}^*, Q^*),\\ I_{n4} &= \boldsymbol{X}_n\boldsymbol{D}_n^*\left(\tilde{\boldsymbol{W}}_n - \boldsymbol{W}_n^*\right)\boldsymbol{U}_n(\boldsymbol{\beta}^*, Q^*)\end{aligned}$$

and

$$I_{n5} = \boldsymbol{X}_n\left(\boldsymbol{D}_n(\hat{\boldsymbol{\beta}}_{\mathrm{GMM}}, \hat{\boldsymbol{\alpha}}_{N,\mathrm{GMM}}) - \boldsymbol{D}_n^*\right)\left(\tilde{\boldsymbol{W}}_n - \boldsymbol{W}_n^*\right)\boldsymbol{U}_n(\boldsymbol{\beta}^*, Q^*).$$

By Lemma F.7 and F.11, we have

$$\begin{aligned}I_{n2} = &-\boldsymbol{X}_n\boldsymbol{D}_n(\hat{\boldsymbol{\beta}}_{\mathrm{GMM}}, \hat{\boldsymbol{\alpha}}_{N,\mathrm{GMM}})\tilde{\boldsymbol{W}}_n\boldsymbol{G}_n(\hat{\boldsymbol{\beta}}_{\mathrm{GMM}}, \hat{\boldsymbol{\alpha}}_{N,\mathrm{GMM}})\begin{bmatrix}\hat{\boldsymbol{\alpha}}_{N,\mathrm{GMM}} - \boldsymbol{\alpha}_N^* \\ \hat{\boldsymbol{\beta}}_{\mathrm{GMM}} - \boldsymbol{\beta}^*\end{bmatrix}\\ &+O_p(J_N^{3/2}N^{-1}).\end{aligned}$$

By Lemma F.8 and F.9, we have,

$$N^{-1/2}\sum_{n=1}^{N} I_{n3} = O_p(J_N N^{-1/2})$$

and

$$N^{-1/2}\sum_{n=1}^{N} I_{n4} = O_p(J_N N^{-1/2}).$$

Let I_{n5i} be the i^{th} element of I_{n5}, where $i = 1, \ldots, p$. By Lemma F.6 and Theorem 7.4.1 and Regularity Condition 7.A, we have, for each i,

$$\begin{aligned} I_{n5i}^2 \leq \|\boldsymbol{X}_{n\cdot i}\|_2^2 \|\boldsymbol{U}_n(\boldsymbol{\beta}^*, Q^*)\|_2^2 \times \lambda_{\max}\left(\left(\boldsymbol{D}_n(\hat{\boldsymbol{\beta}}_{\text{GMM}}, \hat{\boldsymbol{\alpha}}_{N,\text{GMM}}) - \boldsymbol{D}_n^*\right)^2\right) \\ \times \lambda_{\max}\left(\left(\tilde{\boldsymbol{W}}_n - \boldsymbol{W}_n^*\right)^2\right) \\ = O_p(J_N^2 N^{-2}), \end{aligned}$$

where, for each i, $\boldsymbol{X}_{n\cdot i}$ is the i^{th} row of $\boldsymbol{X}_n$, and, for any matrix $\boldsymbol{A}$, $\lambda_{\max}(\boldsymbol{A})$ is the largest eigenvalue of $\boldsymbol{A}$. Therefore, we have

$$N^{-1/2}\sum_{n=1}^{N} I_{n5} = O_p(J_N N^{-1/2}).$$

By Lemma F.11 and $J_N N^{-1/3} = o(1)$, we have

$$\begin{aligned} N^{-1/2}\sum_{n=1}^{N} \boldsymbol{X}_n \boldsymbol{D}_n(\hat{\boldsymbol{\beta}}_{\text{GMM}}, \hat{\boldsymbol{\alpha}}_{N,\text{GMM}}) \tilde{\boldsymbol{W}}_n \boldsymbol{G}_n(\hat{\boldsymbol{\beta}}_{\text{GMM}}, \hat{\boldsymbol{\alpha}}_{N,\text{GMM}}) \begin{bmatrix} \hat{\boldsymbol{\alpha}}_{N,\text{GMM}} - \boldsymbol{\alpha}_N^* \\ \hat{\boldsymbol{\beta}}_{\text{GMM}} - \boldsymbol{\beta}^* \end{bmatrix} \\ = N^{-1/2}\sum_{n=1}^{N} I_{n1}, \end{aligned}$$

which converges in distribution to a multivariate normal distribution by Lemma F.12. □

7.6 Consistency of the Estimation of the Covariance Structure of $\boldsymbol{Y}_n \mid (\boldsymbol{X}_n, \boldsymbol{Z}_n)$

In this section, we aim to show that the parametric version of the covariance matrix of $\boldsymbol{Y}_n \mid (\boldsymbol{X}_n, \boldsymbol{Z}_n)$ introduced in Section 5.6 can be consistently estimated by plug-in the GMM estimators.

Theorem 7.6.1.

Assume that Regularity Condition 7.A-7.I are satisfied and $J_N N^{-1} = o(1)$. Further assume that $(\hat{\boldsymbol{\beta}}_{\mathrm{GMM}}, \hat{\boldsymbol{\alpha}}_{N,\mathrm{GMM}})$ converges to $(\boldsymbol{\beta}^*, \boldsymbol{\alpha}_N^*)$ in the sense that

$$\|\hat{\boldsymbol{\alpha}}_{N,\mathrm{GMM}} - \boldsymbol{\alpha}_N^*\|_2^2 + \left\|\hat{\boldsymbol{\beta}}_{\mathrm{GMM}} - \boldsymbol{\beta}^*\right\|_2^2 = O_p(J_N N^{-1}).$$

Then, for each n, the parametric version $\tilde{\boldsymbol{\Sigma}}_n(\hat{\boldsymbol{\beta}}_{\mathrm{GMM}}, \hat{\boldsymbol{\alpha}}_{N,\mathrm{GMM}})$ converges in probability to $\boldsymbol{\Sigma}_n(\boldsymbol{\beta}^*, Q^*)$ element-wise at rate $N^{1/2} J_N^{-1/2}$, as the sample size N goes to infinity.

Proof. Firstly, the diagonal elements of $\tilde{\boldsymbol{\Sigma}}_n(\hat{\boldsymbol{\beta}}_{\mathrm{GMM}}, \hat{\boldsymbol{\alpha}}_{N,\mathrm{GMM}})$ converges to the variance of $Y_{nt} \mid (\boldsymbol{X}_{nt}, Z_{nt})$ at rate $N^{1/2} J_N^{-1/2}$, as has been shown in the proof of Theorem 7.4.1

Consider the off-diagonal element of $\boldsymbol{\Sigma}_n(\boldsymbol{\beta}^*, Q^*)$ that is

$$\boldsymbol{\Sigma}_{ntt'}(\boldsymbol{\beta}^*, Q^*) = \sum_{j=0}^{\infty} c_{ntt'j}(\boldsymbol{\beta}^*)\alpha_{N,j}^* - \left(\sum_{j=0}^{\infty} \phi_{ntj}(\boldsymbol{\beta}^*)\alpha_{N,j}^*\right) \times \left(\sum_{j=0}^{\infty} \phi_{nt'j}(\boldsymbol{\beta}^*)\alpha_{N,j}^*\right),$$

where for each $n \in \{1, \ldots, N\}$, $t, t' \in \{1, \ldots, T_n\}$ and $j \in \{1, \ldots, J_N\}$, $c_{ntt'j}(\boldsymbol{\beta})$ is defined in Equation (5.8) and $\phi_{ntj}(\boldsymbol{\beta})$ is defined in Equation (5.3).

For each $n \in \{1, \ldots, N\}$ and $t, t' \in \{1, \ldots, T_n\}$, by Taylor's expansion at $\boldsymbol{\beta}_0$, we have

$$\boldsymbol{c}_{ntt'}^{\mathrm{T}}(\hat{\boldsymbol{\beta}}_{\mathrm{GMM}})\hat{\boldsymbol{\alpha}}_{N,\mathrm{GMM}} - \sum_{j=0}^{\infty} c_{ntt'j}(\boldsymbol{\beta}^*)\alpha_{N,j}^* = I_{11} + I_{12} + I_{13},$$

where

$$I_{11} = \dot{\boldsymbol{c}}_{ntt'}^{\mathrm{T}}(\breve{\boldsymbol{\beta}})\boldsymbol{\alpha}_N^* \boldsymbol{X}_{nt}^{\mathrm{T}} \left(\hat{\boldsymbol{\beta}}_{\mathrm{GMM}} - \boldsymbol{\beta}^*\right),$$

$$I_{12} = \boldsymbol{c}_{ntt'}^{\mathrm{T}}(\hat{\boldsymbol{\beta}}_{\mathrm{GMM}})\left(\hat{\boldsymbol{\alpha}}_{N,\mathrm{GMM}} - \boldsymbol{\alpha}_N^*\right),$$

and

$$I_{13} = -\sum_{j=J_N+1}^{\infty} c_{ntt'j}(\boldsymbol{\beta}^*)\alpha^*_{N,j},$$

and $\breve{\boldsymbol{\beta}}$ is on the line segment between $\hat{\boldsymbol{\beta}}_{\text{GMM}}$ and $\boldsymbol{\beta}^*$.

By Lemma 7.2.6 and Regularity Condition 7.A, we have

$$\begin{aligned} I_{11}^2 &\leq \left| \dot{\boldsymbol{c}}^{\mathrm{T}}_{ntt'}(\breve{\boldsymbol{\beta}})\boldsymbol{\alpha}^*_N \boldsymbol{X}^{\mathrm{T}}_{nt} \left(\hat{\boldsymbol{\beta}}_{\text{GMM}} - \boldsymbol{\beta}^* \right) \right|^2 \\ &\leq \left\| \dot{\boldsymbol{c}}^{\mathrm{T}}_{ntt'}(\breve{\boldsymbol{\beta}})\boldsymbol{\alpha}^*_N \right\|_2^2 \left\| \boldsymbol{X}_{nt} \right\|_2^2 \left\| \hat{\boldsymbol{\beta}}_{\text{GMM}} - \boldsymbol{\beta}^* \right\|_2^2 \\ &= O_p(J_N N^{-1}). \end{aligned}$$

By Lemma 7.2.5, we have

$$\begin{aligned} I_{12}^2 &\leq \left\| \boldsymbol{c}_{ntt'}(\hat{\boldsymbol{\beta}}_{\text{GMM}}) \right\|_2^2 \left\| \hat{\boldsymbol{\alpha}}_{N,\text{GMM}} - \boldsymbol{\alpha}^*_N \right\|_2^2 \\ &= O_p(J_N N^{-1}). \end{aligned}$$

By Regularity Condition 7.B and 7.E, we have

$$I_{13} = o(J_N N^{-1}).$$

In sum, we have

$$\begin{aligned} I_1 &= O_p(J_N^{1/2} N^{-1/2}) + O_p(J_N^{1/2} N^{-1/2}) + o(J_N N^{-1}) \\ &= O_p(J_N^{1/2} N^{-1/2}). \end{aligned} \tag{7.30}$$

Next, we examine the asymptotic order of I_2. For each $n \in \{1, \ldots, N\}$ and $t \in \{1, \ldots, T_n\}$, by Taylor's expansion at $\boldsymbol{\beta}_0$, we have

$$\boldsymbol{\phi}^{\mathrm{T}}_{nt}(\hat{\boldsymbol{\beta}}_{\text{GMM}})\hat{\boldsymbol{\alpha}}_{\text{GMM}} - \sum_{j=0}^{\infty} \phi_{ntj}(\boldsymbol{\beta}^*)\alpha^*_{N,j} = I_{21} + I_{22} + I_{23},$$

where

$$\begin{aligned} I_{21} &= \dot{\boldsymbol{\phi}}^{\mathrm{T}}_{nt}(\breve{\boldsymbol{\beta}})\boldsymbol{\alpha}^*_N \boldsymbol{X}^{\mathrm{T}}_{nt} \left(\hat{\boldsymbol{\beta}}_{\text{GMM}} - \boldsymbol{\beta}^* \right), \\ I_{22} &= \boldsymbol{\phi}^{\mathrm{T}}_{nt}(\hat{\boldsymbol{\beta}}_{\text{GMM}}) \left(\hat{\boldsymbol{\alpha}}_{N,\text{GMM}} - \boldsymbol{\alpha}^*_N \right) \end{aligned}$$

and

$$I_{23} = -\sum_{j=J_N+1}^{\infty} \phi_{ntj}(\boldsymbol{\beta}^*)\alpha^*_{N,j}.$$

By Lemma 7.2.2 and Regularity Condition 7.A, we have

$$\begin{aligned} I_{21}^2 &\leq \left\| \dot{\boldsymbol{\phi}}_{nt}^{\mathrm{T}}(\breve{\boldsymbol{\beta}})\boldsymbol{\alpha}_N^* \right\|_2^2 \|\boldsymbol{X}_{nt}\|_2^2 \left\| \hat{\boldsymbol{\beta}}_{\mathrm{GMM}} - \boldsymbol{\beta}^* \right\|_2^2 \\ &= O_p(J_N N^{-1}). \end{aligned}$$

By Lemma 7.2.1, we have

$$\begin{aligned} I_{22}^2 &\leq \boldsymbol{\phi}_{nt}^{\mathrm{T}}(\hat{\boldsymbol{\beta}}_{\mathrm{GMM}})\boldsymbol{\phi}_{nt}(\hat{\boldsymbol{\beta}}_{\mathrm{GMM}}) \left\| \hat{\boldsymbol{\alpha}}_{N,\mathrm{GMM}} - \boldsymbol{\alpha}_N^* \right\|_2^2 \\ &= O_p(J_N N^{-1}). \end{aligned}$$

By Regularity Condition 7.B and 7.E, we have

$$I_{23} = o(J_N N^{-1}).$$

So, we have

$$\begin{aligned} I_2 &= O_p(J_N^{1/2} N^{-1/2}) + O_p(J_N^{1/2} N^{-1/2}) + o(J_N N^{-1}) \\ &= O_p(J_N^{1/2} N^{-1/2}). \end{aligned} \tag{7.31}$$

By Equation (7.30) and (7.31), we have the convergence of the off-diagonal elements, i.e., for each $t, t' \in \{1, \ldots, T_n\}$ and $t \neq t'$,

$$\tilde{\boldsymbol{\Sigma}}_{ntt'}(\hat{\boldsymbol{\beta}}_{\mathrm{GMM}}, \hat{\boldsymbol{\alpha}}_{N,\mathrm{GMM}}) = \boldsymbol{\Sigma}_{ntt'}(\boldsymbol{\beta}^*, Q^*) + O_p(J_N^{1/2} N^{-1/2}).$$

□

7.7 Discussion

In this chapter, we give the asymptotic properties of the GMM estimator for univariate mixed-effects models, including the convergence rates of the GMM estimators, the convergence rate of the plugging in weighting matrices, the asymptotic normality in the GMM and the consistency of the parametric version of the covariance structure of $\boldsymbol{Y}_n \mid (\boldsymbol{X}_n, \boldsymbol{Z}_n)$. We derive the asymptotic results in the case that the dimension J_N of the generalized moment vector diverges with the sample size. As we have seen, such divergence J_N slows the convergence rate of the GMM estimator. Moreover, an asymptotic normality result can be achieved when $J_N N^{-1/2} = o(1)$.

However, it is still challenging to use the results in this chapter to make inference for the regression parameter $\boldsymbol{\beta}$. One of the major reason is that the true value of the generalized moments $\boldsymbol{\alpha}_N^*$ is unknown; see [Silvapulle and Sen, 2005] for hypothesis testing problems in the presence of unknown nuisance parameters. Even if $\boldsymbol{\alpha}_N^*$ is known, it is still challenging to obtain the asymptotic distribution of $\hat{\boldsymbol{\alpha}}_{N,\mathrm{GMM}}$, because the boundary of parameter space of $(\boldsymbol{\beta}, \boldsymbol{\alpha}_N)$ is involved. In Chapter 8, we will propose a methodology to deal with such hypothesis testing problems.

Appendix: F

F.1 Proofs of the Lemmas in Section 7.2

Proof of Lemma 7.2.1

Proof. By Regularity Condition 7.B and 7.C, for every integer N, we have

$$\begin{aligned}
& \sup_{n\in\{1,\ldots,N\}} \sup_{t\in\{1,\ldots,T_n\}} \boldsymbol{\phi}_{nt}^{\mathrm{T}}(\boldsymbol{\beta})\boldsymbol{\phi}_{nt}(\boldsymbol{\beta}) \\
= & \sup_{n\in\{1,\ldots,N\}} \sup_{t\in\{1,\ldots,T_n\}} \sum_{j=0}^{J_N} \phi_{ntj}^2(\boldsymbol{\beta}) \\
\leq & \sup_{n\in\{1,\ldots,N\}} \sup_{t\in\{1,\ldots,T_n\}} \sum_{j=0}^{\infty} \phi_{ntj}^2(\boldsymbol{\beta}) \\
= & \sup_{n\in\{1,\ldots,N\}} \sup_{t\in\{1,\ldots,T_n\}} \int_{\mathcal{B}} \left(\sum_{j=0}^{\infty} \phi_{ntj}(\boldsymbol{\beta}) P_j(b) \right)^2 \mathrm{d}\mu \\
= & \sup_{n\in\{1,\ldots,N\}} \sup_{t\in\{1,\ldots,T_n\}} \int_{\mathcal{B}} \left(g^{-1}(\boldsymbol{X}_{nt}^{\mathrm{T}}\boldsymbol{\beta} + Z_{nt} b) \right)^2 \mathrm{d}\mu
\end{aligned}$$

is bounded. Similarly, we also can show that $\dot{\boldsymbol{\phi}}_{nt}^{\mathrm{T}}(\boldsymbol{\beta})\dot{\boldsymbol{\phi}}_{nt}(\boldsymbol{\beta})$ and $\ddot{\boldsymbol{\phi}}_{nt}^{\mathrm{T}}(\boldsymbol{\beta})\ddot{\boldsymbol{\phi}}_{nt}(\boldsymbol{\beta})$ is uniformly bounded over $n \in \{1,\ldots,N\}$ and $t \in \{1,\ldots,T_n\}$. □

Proof of Lemma 7.2.2

Proof. Because $\boldsymbol{\alpha}_N \in \mathcal{M}_{J_N+1}$, there exists a probability measure $Q_{\boldsymbol{\alpha}_N}$ defined on $\mathcal{B}$ such that $\boldsymbol{\alpha}_N = \int_{\mathcal{B}} \boldsymbol{P}(b)\mathrm{d}Q_{\boldsymbol{\alpha}_N}$ by Theorem 3.3.1. By Regularity Condition 7.B-7.D, for every integer N, we have that,

$$\begin{aligned}
& \sup_{n\in\{1,\ldots,N\}} \sup_{t\in\{1,\ldots,T_n\}} \left|\boldsymbol{\phi}_{nt}^{\mathrm{T}}(\boldsymbol{\beta})\boldsymbol{\alpha}_N\right| \\
= & \sup_{n\in\{1,\ldots,N\}} \sup_{t\in\{1,\ldots,T_n\}} \left|\sum_{j=0}^{J_N} \phi_{ntj}(\boldsymbol{\beta})\alpha_{N,j}\right| \\
= & \sup_{n\in\{1,\ldots,N\}} \sup_{t\in\{1,\ldots,T_n\}} \left|\int_{\mathcal{B}} g^{-1}(\boldsymbol{X}_{nt}^{\mathrm{T}}\boldsymbol{\beta} + Z_{nt}b)\mathrm{d}Q_{\boldsymbol{\alpha}_N} + o(J_N N^{-1})\right| \\
\leq & \sup_{n\in\{1,\ldots,N\}} \sup_{t\in\{1,\ldots,T_n\}} \left|\int_{\mathcal{B}} g^{-1}(\boldsymbol{X}_{nt}^{\mathrm{T}}\boldsymbol{\beta} + Z_{nt}b)\mathrm{d}Q_{\boldsymbol{\alpha}_N}\right| + \left|o(J_N N^{-1})\right|,
\end{aligned}$$

is bounded. Similarly, we also can show that $\left|\dot{\boldsymbol{\phi}}_{nt}^{\mathrm{T}}(\boldsymbol{\beta})\boldsymbol{\alpha}_N\right|$, $\left|\ddot{\boldsymbol{\phi}}_{nt}^{\mathrm{T}}(\boldsymbol{\beta})\boldsymbol{\alpha}_N\right|$ and $\left|\dddot{\boldsymbol{\phi}}_{nt}^{\mathrm{T}}(\boldsymbol{\beta})\boldsymbol{\alpha}_N\right|$ are uniformly bounded over $n \in \{1,\ldots,N\}$ and $t \in \{1,\ldots,T_n\}$. □

Proof of Lemma 7.2.3

Proof. By Regularity Condition 7.B and 7.F, for every integer N, we have

$$\begin{aligned}
& \sup_{n\in\{1,\ldots,N\}} \sup_{t\in\{1,\ldots,T_n\}} \boldsymbol{a}_{nt}^{\mathrm{T}}(\boldsymbol{\beta})\boldsymbol{a}_{nt}(\boldsymbol{\beta}) \\
= & \sup_{n\in\{1,\ldots,N\}} \sup_{t\in\{1,\ldots,T_n\}} \sum_{j=0}^{J_N} a_{ntj}^2(\boldsymbol{\beta}) \\
\leq & \sup_{n\in\{1,\ldots,N\}} \sup_{t\in\{1,\ldots,T_n\}} \sum_{j=0}^{\infty} a_{ntj}^2(\boldsymbol{\beta}) \\
= & \sup_{n\in\{1,\ldots,N\}} \sup_{t\in\{1,\ldots,T_n\}} \int_{\mathcal{B}} \left(\sum_{j=0}^{\infty} a_{ntj}(\boldsymbol{\beta})P_j(b)\right)^2 \mathrm{d}\mu \\
= & \sup_{n\in\{1,\ldots,N\}} \sup_{t\in\{1,\ldots,T_n\}} \int_{\mathcal{B}} \left(h \circ g^{-1}(\boldsymbol{X}_{nt}^{\mathrm{T}}\boldsymbol{\beta} + Z_{nt}b)\right)^2 \mathrm{d}\mu
\end{aligned}$$

is bounded. Similarly, we also can show that $\dot{\boldsymbol{a}}_{nt}^{\mathrm{T}}(\boldsymbol{\beta})\dot{\boldsymbol{a}}_{nt}(\boldsymbol{\beta})$ is uniformly bounded. □

Proof of Lemma 7.2.4

Proof. Because $\boldsymbol{\alpha}_N \in \mathcal{M}_{J_N+1}$, there exists a probability measure $Q_{\boldsymbol{\alpha}_N}$ defined on $\mathcal{B}$ such that $\boldsymbol{\alpha}_N = \int_{\mathcal{B}} \boldsymbol{P}(b)\mathrm{d}Q_{\boldsymbol{\alpha}_N}$ by Theorem 3.3.1. By Regularity Condition 7.B, 7.F and 7.G, for every integer N, we have

$$\begin{aligned}
& \sup_{n\in\{1,\ldots,N\}} \sup_{t\in\{1,\ldots,T_n\}} \left|\boldsymbol{a}_{nt}^{\mathrm{T}}(\boldsymbol{\beta})\boldsymbol{\alpha}_N\right| \\
= & \sup_{n\in\{1,\ldots,N\}} \sup_{t\in\{1,\ldots,T_n\}} \left|\sum_{j=0}^{J_N} a_{ntj}(\boldsymbol{\beta})\alpha_{N,j}\right| \\
= & \sup_{n\in\{1,\ldots,N\}} \sup_{t\in\{1,\ldots,T_n\}} \left|\int_{\mathcal{B}} h\circ g^{-1}(\boldsymbol{X}_{nt}^{\mathrm{T}}(\boldsymbol{\beta}) + Z_{nt}b)\mathrm{d}Q_{\boldsymbol{\alpha}_N} + o(J_N N^{-1})\right| \\
\leq & \sup_{n\in\{1,\ldots,N\}} \sup_{t\in\{1,\ldots,T_n\}} \left|\int_{\mathcal{B}} h\circ g^{-1}(\boldsymbol{X}_{nt}^{\mathrm{T}}(\boldsymbol{\beta}) + Z_{nt}b)\mathrm{d}Q_{\boldsymbol{\alpha}_N}\right| + \left|o(J_N N^{-1})\right|
\end{aligned}$$

is bounded. Similarly, we can also show that $\left|\dot{\boldsymbol{a}}_{nt}^{\mathrm{T}}(\boldsymbol{\beta})\boldsymbol{\alpha}_N\right|$ and $\left|\ddot{\boldsymbol{a}}_{nt}^{\mathrm{T}}(\boldsymbol{\beta})\boldsymbol{\alpha}_N\right|$ are uniformly bounded over $n \in \{1,\ldots,N\}$ and $t \in \{1,\ldots,T_n\}$. □

Proof of Lemma 7.2.5

Proof. Under Regularity Condition 7.B and 7.E, for every integer N, we have

$$\begin{aligned}
& \sup_{n\in\{1,\ldots,N\}} \sup_{t\in\{1,\ldots,T_n\}} \boldsymbol{c}_{nt}^{\mathrm{T}}(\boldsymbol{\beta})\boldsymbol{c}_{nt}(\boldsymbol{\beta}) \\
= & \sup_{n\in\{1,\ldots,N\}} \sup_{t\in\{1,\ldots,T_n\}} \sum_{j=0}^{J_N} c_{ntj}^2(\boldsymbol{\beta}) \\
\leq & \sup_{n\in\{1,\ldots,N\}} \sup_{t\in\{1,\ldots,T_n\}} \sum_{j=0}^{\infty} c_{ntj}^2(\boldsymbol{\beta}) \\
= & \sup_{n\in\{1,\ldots,N\}} \sup_{t\in\{1,\ldots,T_n\}} \int_{\mathcal{B}} \left(\sum_{j=0}^{\infty} c_{ntj}(\boldsymbol{\beta})P_j(b)\right)^2 \mathrm{d}\mu \\
= & \sup_{n\in\{1,\ldots,N\}} \sup_{t\in\{1,\ldots,T_n\}} \int_{\mathcal{B}} \left(g^{-1}(\boldsymbol{X}_{nt}^{\mathrm{T}}\boldsymbol{\beta} + Z_{nt}b) \times g^{-1}(\boldsymbol{X}_{nt'}^{\mathrm{T}}\boldsymbol{\beta} + Z_{nt}b)\right)^2 \mathrm{d}\mu
\end{aligned}$$

is bounded. Similarly, we also can show that $\dot{\boldsymbol{c}}_{ntt'}^{\mathrm{T}}(\boldsymbol{\beta})\dot{\boldsymbol{c}}_{ntt'}(\boldsymbol{\beta})$ is uniformly bounded over $n \in \{1,\dots,N\}$ and $t \in \{1,\dots,T_n\}$. □

Proof of Lemma 7.2.6

Proof. Because $\boldsymbol{\alpha}_N \in \mathcal{M}_{J_N+1}$, there exists a probability measure $Q_{\boldsymbol{\alpha}_N}$ defined on $\mathcal{B}$ such that $\boldsymbol{\alpha}_N = \int_{\mathcal{B}} \boldsymbol{P}(b)\mathrm{d}Q_{\boldsymbol{\alpha}_N}$ by Theorem 3.3.1. By Regularity Condition 7.B, 7.D and 7.E, for every integer N, we have that,

$$
\begin{aligned}
&\sup_{n\in\{1,\dots,N\}}\sup_{t\in\{1,\dots,T_n\}}\left|\boldsymbol{c}_{ntt'}^{\mathrm{T}}(\boldsymbol{\beta})\boldsymbol{\alpha}_N\right| \\
=&\sup_{n\in\{1,\dots,N\}}\sup_{t\in\{1,\dots,T_n\}}\left|\sum_{j=0}^{\infty} c_{ntt'j}(\boldsymbol{\beta})\alpha_j + o(J_N N^{-1})\right| \\
=&\sup_{n\in\{1,\dots,N\}}\sup_{t\in\{1,\dots,T_n\}}\left|\int_{\mathcal{B}} g^{-1}(\boldsymbol{X}_{nt}^{\mathrm{T}}\boldsymbol{\beta}+Z_{nt}b)\times g^{-1}(\boldsymbol{X}_{nt'}^{\mathrm{T}}\boldsymbol{\beta}+Z_{nt'}b)\mathrm{d}Q_{\boldsymbol{\alpha}_N} + o(J_N N^{-1})\right| \\
\leq&\sup_{n\in\{1,\dots,N\}}\sup_{t\in\{1,\dots,T_n\}}\left|\int_{\mathcal{B}} g^{-1}(\boldsymbol{X}_{nt}^{\mathrm{T}}\boldsymbol{\beta}+Z_{nt}b)\times g^{-1}(\boldsymbol{X}_{nt'}^{\mathrm{T}}\boldsymbol{\beta}+Z_{nt'}b)\mathrm{d}Q_{\boldsymbol{\alpha}_N}\right| + \left|o(J_N N^{-1})\right| \\
\leq&\sup_{n\in\{1,\dots,N\}}\sup_{t\in\{1,\dots,T_n\}}\left(\int_{\mathcal{B}}\left(g^{-1}(\boldsymbol{X}_{nt}^{\mathrm{T}}\boldsymbol{\beta}+Z_{nt}b)\right)^2\mathrm{d}Q_{\boldsymbol{\alpha}_N}\times\int_{\mathcal{B}}\left(g^{-1}(\boldsymbol{X}_{nt'}^{\mathrm{T}}\boldsymbol{\beta}+Z_{nt'}b)\right)^2\mathrm{d}Q_{\boldsymbol{\alpha}_N}\right)^{1/2} \\
&+\left|o(J_N N^{-1})\right|
\end{aligned}
$$

is bounded. Similarly, we also can show that $\left|\dot{\boldsymbol{c}}_{ntt'}^{\mathrm{T}}(\boldsymbol{\beta})\boldsymbol{\alpha}_N\right|$ and $\left|\ddot{\boldsymbol{c}}_{ntt'}^{\mathrm{T}}(\boldsymbol{\beta})\boldsymbol{\alpha}_N\right|$ are uniformly bounded over $n \in \{1,\dots,N\}$ and $t \in \{1,\dots,T_n\}$. □

F.2 Proofs of the Lemmas for Theorem 7.3.1

Proof of Lemma F.1

Lemma F.1.
Assume that Regularity Conditions 7.A-7.D are satisfied. Then, $\partial U_{nt}(\boldsymbol{\beta},\boldsymbol{\alpha}_N)$, defined in Equation (7.17), is $O(\|\boldsymbol{v}\|_2)$, for each $n \in \{1,\dots,N\}$, $t \in \{1,\dots,T_n\}$ and $(\boldsymbol{\beta},\boldsymbol{\alpha}_N) \in \mathbb{R}^p \times \mathcal{M}_{J_N+1}$.

Proof. For each $n \in \{1, \ldots, N\}$ and $t \in \{1, \ldots, T_n\}$, we have

$$U_{nt}(\boldsymbol{\beta} + \Delta \boldsymbol{v}_{\boldsymbol{\beta}}, \boldsymbol{\alpha}_N + \Delta \boldsymbol{v}_{\boldsymbol{\alpha}}) = Y_{nt} - \sum_{j=0}^{J_N} \phi_{ntj}(\boldsymbol{\beta} + \Delta \boldsymbol{v}_{\boldsymbol{\beta}})(\alpha_{N,j} + \Delta \boldsymbol{v}_{\boldsymbol{\alpha}_N,j}),$$

where $\boldsymbol{v}_{\boldsymbol{\alpha}_N,j}$ is the j^{th} element of $\boldsymbol{v}_{\boldsymbol{\alpha}}$.

For each $n \in \{1, \ldots, N\}$, $t \in \{1, \ldots, T_n\}$ and $j \in \{1, \ldots, J_N\}$, let

$$\bigtriangledown \phi_{ntj}(\boldsymbol{\beta}) = \dot{\phi}_{ntj}(\boldsymbol{\beta}) \times \boldsymbol{X}_{nt} \in \mathbb{R}^p \tag{F.1}$$

be the derivative of $\phi_{ntj}(\boldsymbol{\beta})$ with respect to $\boldsymbol{\beta}$, where $\dot{\phi}_{ntj}(\boldsymbol{\beta})$ is defined in Equation (7.6). We write

$$\bigtriangledown \boldsymbol{\phi}_{nt}(\boldsymbol{\beta}) = \boldsymbol{X}_{nt} \dot{\boldsymbol{\phi}}_{nt}^{\mathrm{T}}(\boldsymbol{\beta}),$$

where $\dot{\boldsymbol{\phi}}_{nt}(\boldsymbol{\beta})$ is defined in Equation (7.4).

The derivative of $U_{nt}(\boldsymbol{\beta} + \Delta \boldsymbol{v}_{\boldsymbol{\beta}}, \boldsymbol{\alpha}_N + \Delta \boldsymbol{v}_{\boldsymbol{\alpha}})$ with respect to Δ is

$$\begin{aligned} & \frac{\partial}{\partial \Delta} U_{nt}(\boldsymbol{\beta} + \Delta \boldsymbol{v}_{\boldsymbol{\beta}}, \boldsymbol{\alpha}_N + \Delta \boldsymbol{v}_{\boldsymbol{\alpha}}) \\ = & -\sum_{j=0}^{J_N} \left(\boldsymbol{v}_{\boldsymbol{\beta}}^{\mathrm{T}} \bigtriangledown \phi_{ntj}(\boldsymbol{\beta} + \Delta \boldsymbol{v}_{\boldsymbol{\beta}})(\alpha_{N,j} + \Delta \boldsymbol{v}_{\boldsymbol{\alpha}_N,j}) + \phi_{ntj}(\boldsymbol{\beta} + \Delta \boldsymbol{v}_{\boldsymbol{\beta}}) \boldsymbol{v}_{\boldsymbol{\alpha}_N,j} \right) \\ = & -\boldsymbol{v}^{\mathrm{T}} \left(\boldsymbol{\phi}_{nt}^{\mathrm{T}}(\boldsymbol{\beta} + \Delta \boldsymbol{v}_{\boldsymbol{\beta}}), (\bigtriangledown \boldsymbol{\phi}_{nt}(\boldsymbol{\beta} + \Delta \boldsymbol{v}_{\boldsymbol{\beta}})(\boldsymbol{\alpha}_N + \Delta \boldsymbol{v}_{\boldsymbol{\alpha}}))^{\mathrm{T}} \right)^{\mathrm{T}}, \end{aligned} \tag{F.2}$$

where $\boldsymbol{\phi}_{nt}(\boldsymbol{\beta}) \in \mathbb{R}^{J_N+1}$ is defined in Equation (7.3).

For each $n \in \{1, \ldots, N\}$ and $t \in \{1, \ldots, T_n\}$, let

$$\boldsymbol{h}_{nt}(\boldsymbol{\beta}, \boldsymbol{\alpha}_N) = \left(\boldsymbol{\phi}_{nt}^{\mathrm{T}}(\boldsymbol{\beta}), (\bigtriangledown \boldsymbol{\phi}_{nt}(\boldsymbol{\beta}) \boldsymbol{\alpha}_N)^{\mathrm{T}} \right)^{\mathrm{T}} \in \mathbb{R}^{p+J_N+1}, \tag{F.3}$$

where $(\boldsymbol{\beta}, \boldsymbol{\alpha}_N) \in \mathbb{R}^p \times \mathcal{M}_{J_N+1}$. When $\Delta = 0$, by the Cauchy-Schwarz inequality,

$$\begin{aligned} \left(\frac{\partial}{\partial \Delta} U_{nt}(\boldsymbol{\beta} + \Delta \boldsymbol{v}_{\boldsymbol{\beta}}, \boldsymbol{\alpha}_N + \Delta \boldsymbol{v}_{\boldsymbol{\alpha}}) \bigg|_{\Delta=0} \right)^2 &= \left(\boldsymbol{v}^{\mathrm{T}} \boldsymbol{h}_{nt}(\boldsymbol{\beta}, \boldsymbol{\alpha}_N) \right)^2 \\ &\leq \boldsymbol{h}_{nt}^{\mathrm{T}}(\boldsymbol{\beta}, \boldsymbol{\alpha}_N) \boldsymbol{h}_{nt}(\boldsymbol{\beta}, \boldsymbol{\alpha}_N) \|\boldsymbol{v}\|_2^2. \end{aligned}$$

By Lemma 7.2.1, we have that $\boldsymbol{\phi}_{nt}^{\mathrm{T}}(\boldsymbol{\beta})\boldsymbol{\phi}_{nt}(\boldsymbol{\beta})$ is bounded for each n and t. By Lemma 7.2.2 and Regularity Condition 7.A, we have

$$\begin{aligned}(\triangledown\boldsymbol{\phi}_{nt}(\boldsymbol{\beta})\boldsymbol{\alpha}_N)^{\mathrm{T}}\triangledown\boldsymbol{\phi}_{nt}(\boldsymbol{\beta})\boldsymbol{\alpha}_N &= \boldsymbol{\alpha}_N^{\mathrm{T}}\dot{\boldsymbol{\phi}}_{nt}(\boldsymbol{\beta})\boldsymbol{X}_{nt}^{\mathrm{T}}\boldsymbol{X}_{nt}\dot{\boldsymbol{\phi}}_{nt}^{\mathrm{T}}(\boldsymbol{\beta})\boldsymbol{\alpha}_N \\ &\le (\boldsymbol{X}_{nt}^{\mathrm{T}}\boldsymbol{X}_{nt})\left(\dot{\boldsymbol{\phi}}_{nt}^{\mathrm{T}}(\boldsymbol{\beta})\boldsymbol{\alpha}_N\right)^2\end{aligned}$$

is also bounded for each n and t. Therefore, for each $(\boldsymbol{\beta},\boldsymbol{\alpha}_N)\in\mathbb{R}^p\times\mathcal{M}_{J_N+1}$, $n\in\{1,\ldots,N\}$ and $t\in\{1,\ldots,T_n\}$, $\boldsymbol{h}_{nt}^{\mathrm{T}}(\boldsymbol{\beta},\boldsymbol{\alpha}_N)\boldsymbol{h}_{nt}(\boldsymbol{\beta},\boldsymbol{\alpha}_N)$ is bounded. It follows that, for each $n\in\{1,\ldots,N\}$, $t\in\{1,\ldots,T_n\}$ and $(\boldsymbol{\beta},\boldsymbol{\alpha}_N)\in\mathbb{R}^p\times\mathcal{M}_{J_N+1}$,

$$\partial U_{nt}(\boldsymbol{\beta},\boldsymbol{\alpha}_N)=O(\|\boldsymbol{v}\|_2).$$

□

Proof of Lemma F.2

Lemma F.2.
Assume that Regularity Condition 7.A-7.D are satisfied. Then, $\partial^2U_{nt}(\boldsymbol{\beta},\boldsymbol{\alpha}_N)$, defined in Equation (7.19), is $O(\|\boldsymbol{v}\|_2^2)$, for each $n\in\{1,\ldots,N\}$, $t\in\{1,\ldots,T_n\}$ and $(\boldsymbol{\beta},\boldsymbol{\alpha}_N)\in\mathbb{R}^p\times\mathcal{M}_{J_N+1}$.

Proof. For each $n\in\{1,\ldots,N\}$, $t\in\{1,\ldots,T_n\}$ and $j\in\{1,\ldots,J_N\}$, the Hessian matrix of $\phi_{ntj}(\boldsymbol{\beta})$ with respect to $\boldsymbol{\beta}$ is

$$\triangledown^2\phi_{ntj}(\boldsymbol{\beta})=\ddot{\phi}_{ntj}(\boldsymbol{\beta})\boldsymbol{X}_{nt}\boldsymbol{X}_{nt}^{\mathrm{T}}, \tag{F.4}$$

where $\ddot{\phi}_{ntj}(\boldsymbol{\beta})$ has been defined in Equation (7.7).

The second order derivative of $U_{nt}(\boldsymbol{\beta}+\Delta\boldsymbol{v}_{\boldsymbol{\beta}},\boldsymbol{\alpha}_N+\Delta\boldsymbol{v}_{\boldsymbol{\alpha}})$ with respect to Δ is

$$\begin{aligned}&\frac{\partial^2}{\partial\Delta^2}U_{nt}(\boldsymbol{\beta}+\Delta\boldsymbol{v}_{\boldsymbol{\beta}},\boldsymbol{\alpha}_N+\Delta\boldsymbol{v}_{\boldsymbol{\alpha}})\\ =&\ -\sum_{j=0}^{J_N}\left(\boldsymbol{v}_{\boldsymbol{\beta}}^{\mathrm{T}}\triangledown^2\phi_{ntj}(\boldsymbol{\beta}+\Delta\boldsymbol{v}_{\boldsymbol{\beta}})\boldsymbol{v}_{\boldsymbol{\beta}}(\alpha_{N,j}+\Delta\boldsymbol{v}_{\boldsymbol{\alpha}_N,j})+2\boldsymbol{v}_{\boldsymbol{\beta}}^{\mathrm{T}}\triangledown\phi_{ntj}(\boldsymbol{\beta}+\Delta\boldsymbol{v}_{\boldsymbol{\beta}})\boldsymbol{v}_{\boldsymbol{\alpha}_N,j}\right).\end{aligned}$$

When $\Delta = 0$, we have

$$\begin{aligned}
&\frac{\partial^2}{\partial \Delta^2} U_{nt}(\boldsymbol{\beta} + \Delta \boldsymbol{v}_{\boldsymbol{\beta}}, \boldsymbol{\alpha}_N + \Delta \boldsymbol{v}_{\boldsymbol{\alpha}})\bigg|_{\Delta=0} \\
= & -\sum_{j=0}^{J_N} \left(\boldsymbol{v}_{\boldsymbol{\beta}}^{\mathrm{T}} \bigtriangledown^2 \phi_{ntj}(\boldsymbol{\beta}) \boldsymbol{v}_{\boldsymbol{\beta}} \alpha_{N,j} + 2 \boldsymbol{v}_{\boldsymbol{\beta}}^{\mathrm{T}} \bigtriangledown \phi_{ntj}(\boldsymbol{\beta}) \boldsymbol{v}_{\boldsymbol{\alpha}_N,j} \right) \\
= & -\boldsymbol{v}_{\boldsymbol{\beta}}^{\mathrm{T}} \left(\sum_{j=0}^{J_N} \alpha_{N,j} \bigtriangledown^2 \phi_{ntj}(\boldsymbol{\beta}) \right) \boldsymbol{v}_{\boldsymbol{\beta}} - 2 \boldsymbol{v}_{\boldsymbol{\beta}}^{\mathrm{T}} \bigtriangledown \boldsymbol{\phi}_{nt}(\boldsymbol{\beta}) \boldsymbol{v}_{\boldsymbol{\alpha}}. \qquad \text{(F.5)}
\end{aligned}$$

By Lemma 7.2.2 and Regularity Condition 7.A, we have that

$$\begin{aligned}
\left| \boldsymbol{v}_{\boldsymbol{\beta}}^{\mathrm{T}} \left(\sum_{j=0}^{J_N} \alpha_{N,j} \bigtriangledown^2 \phi_{ntj}(\boldsymbol{\beta}) \right) \boldsymbol{v}_{\boldsymbol{\beta}} \right| &= \left| \ddot{\boldsymbol{\phi}}_{ntj}(\boldsymbol{\beta}) \boldsymbol{\alpha}_N \right| \boldsymbol{v}_{\boldsymbol{\beta}}^{\mathrm{T}} \boldsymbol{X}_{nt} \boldsymbol{X}_{nt}^{\mathrm{T}} \boldsymbol{v}_{\boldsymbol{\beta}} \\
&\leq \|\boldsymbol{v}_{\boldsymbol{\beta}}\|_2^2 \, \|\boldsymbol{X}_{nt}\|_2^2 \left| \ddot{\boldsymbol{\phi}}_{ntj}(\boldsymbol{\beta}) \boldsymbol{\alpha}_N \right| \\
&= O(\|\boldsymbol{v}\|_2^2). \qquad \text{(F.6)}
\end{aligned}$$

On the other hand, by Lemma 7.2.1, we have that

$$\begin{aligned}
\left(\boldsymbol{v}_{\boldsymbol{\beta}}^{\mathrm{T}} \bigtriangledown \boldsymbol{\phi}_{nt}(\boldsymbol{\beta}) \boldsymbol{v}_{\boldsymbol{\alpha}} \right)^2 &= \left(\boldsymbol{v}_{\boldsymbol{\beta}}^{\mathrm{T}} \boldsymbol{X}_{nt} \dot{\boldsymbol{\phi}}_{nt}^{\mathrm{T}}(\boldsymbol{\beta}) \boldsymbol{v}_{\boldsymbol{\alpha}} \right)^2 \\
&= \left(\boldsymbol{v}_{\boldsymbol{\beta}}^{\mathrm{T}} \boldsymbol{X}_{nt} \right)^2 \left(\dot{\boldsymbol{\phi}}_{nt}^{\mathrm{T}}(\boldsymbol{\beta}) \boldsymbol{v}_{\boldsymbol{\alpha}} \right)^2 \\
&\leq \|\boldsymbol{X}_{nt}\|_2^2 \left\| \dot{\boldsymbol{\phi}}_{nt}(\boldsymbol{\beta}) \right\|_2^2 \|\boldsymbol{v}_{\boldsymbol{\beta}}\|_2^2 \|\boldsymbol{v}_{\boldsymbol{\alpha}}\|_2^2 \\
&= O(\|\boldsymbol{v}\|_2^4). \qquad \text{(F.7)}
\end{aligned}$$

By Equation (F.5), (F.6) and (F.7), it follows that, for each $n \in \{1, \ldots, N\}$, $t \in \{1, \ldots, T_n\}$ and $(\boldsymbol{\beta}, \boldsymbol{\alpha}_N) \in \mathbb{R}^p \times \mathcal{M}_{J_N+1}$,

$$\partial^2 U_{nt}(\boldsymbol{\beta}, \boldsymbol{\alpha}_N) = O(\|\boldsymbol{v}\|_2^2).$$

□

Proof of Lemma F.3

Lemma F.3.

Assume that Regularity Condition 7.A-7.D are satisfied. Then, $\partial^3 U_{nt}(\boldsymbol{\beta}, \boldsymbol{\alpha}_N)$, defined in Equation (7.21), is $O(\|\boldsymbol{v}\|_2^3)$, for each $n \in \{1, \ldots, N\}$, $t \in \{1, \ldots, T_n\}$ and $(\boldsymbol{\beta}, \boldsymbol{\alpha}_N) \in \mathbb{R}^p \times \mathcal{M}_{J_N+1}$.

Proof. For each $n \in \{1, \ldots, N\}$, $t \in \{1, \ldots, T_n\}$ and $j \in \{1, \ldots, J_N\}$, let $\triangledown\left(\boldsymbol{v}_{\boldsymbol{\beta}}^{\mathrm{T}} \triangledown^2 \phi_{ntj}(\boldsymbol{\beta}) \boldsymbol{v}_{\boldsymbol{\beta}}\right)$ be the derivative of $\boldsymbol{v}_{\boldsymbol{\beta}}^{\mathrm{T}} \triangledown^2 \phi_{ntj}(\boldsymbol{\beta}) \boldsymbol{v}_{\boldsymbol{\beta}}$ with respect to $\boldsymbol{\beta}$, i.e.,

$$\triangledown\left(\boldsymbol{v}_{\boldsymbol{\beta}}^{\mathrm{T}} \triangledown^2 \phi_{ntj}(\boldsymbol{\beta}) \boldsymbol{v}_{\boldsymbol{\beta}}\right) = \dddot{\phi}_{ntj}(\boldsymbol{\beta}) \boldsymbol{v}_{\boldsymbol{\beta}}^{\mathrm{T}} \boldsymbol{X}_{nt} \boldsymbol{X}_{nt}^{\mathrm{T}} \boldsymbol{v}_{\boldsymbol{\beta}} \boldsymbol{X}_{nt},$$

where $\dddot{\phi}_{ntj}(\boldsymbol{\beta})$ is defined in Equation (7.9).

The third order derivative of $U_{nt}(\boldsymbol{\beta} + \Delta \boldsymbol{v}_{\boldsymbol{\beta}}, \boldsymbol{\alpha}_N + \Delta \boldsymbol{v}_{\boldsymbol{\alpha}})$ with respect to Δ is

$$\begin{aligned}
&\frac{\partial^3}{\partial \Delta^3} U_{nt}(\boldsymbol{\beta} + \Delta \boldsymbol{v}_{\boldsymbol{\beta}}, \boldsymbol{\alpha}_N + \Delta \boldsymbol{v}_{\boldsymbol{\alpha}}) \\
= &-\sum_{j=0}^{J_N} \boldsymbol{v}_{\boldsymbol{\beta}}^{\mathrm{T}} \left(\triangledown\left(\boldsymbol{v}_{\boldsymbol{\beta}}^{\mathrm{T}} \triangledown^2 \phi_{ntj}(\boldsymbol{\beta} + \Delta \boldsymbol{v}_{\boldsymbol{\beta}}) \boldsymbol{v}_{\boldsymbol{\beta}}\right) (\alpha_{N,j} + \Delta \boldsymbol{v}_{\boldsymbol{\alpha}_N, j})\right. \\
&\left. + 3 \boldsymbol{v}_{\boldsymbol{\beta}}^{\mathrm{T}} \triangledown^2 \phi_{ntj}(\boldsymbol{\beta} + \Delta \boldsymbol{v}_{\boldsymbol{\beta}}) \boldsymbol{v}_{\boldsymbol{\beta}} \boldsymbol{v}_{\boldsymbol{\alpha}_N, j}\right).
\end{aligned}$$

When $\Delta = 0$, we have

$$\begin{aligned}
&-\frac{\partial^3}{\partial \Delta^3} U_{nt}(\boldsymbol{\beta} + \Delta \boldsymbol{v}_{\boldsymbol{\beta}}, \boldsymbol{\alpha}_N + \Delta \boldsymbol{v}_{\boldsymbol{\alpha}})\Big|_{\Delta=0} \\
= &\sum_{j=0}^{J_N} \left(\boldsymbol{v}_{\boldsymbol{\beta}}^{\mathrm{T}} \triangledown \left(\boldsymbol{v}_{\boldsymbol{\beta}}^{\mathrm{T}} \triangledown^2 \phi_{ntj}(\boldsymbol{\beta}) \boldsymbol{v}_{\boldsymbol{\beta}}\right) \alpha_{N,j} + 3 \boldsymbol{v}_{\boldsymbol{\beta}}^{\mathrm{T}} \triangledown^2 \phi_{ntj}(\boldsymbol{\beta}) \boldsymbol{v}_{\boldsymbol{\beta}} \boldsymbol{v}_{\boldsymbol{\alpha}_N, j}\right).
\end{aligned} \tag{F.8}$$

By Lemma 7.2.2 and Regularity Condition 7.A, we have that

$$\begin{aligned}
\left|\sum_{j=0}^{J_N} \boldsymbol{v}_{\boldsymbol{\beta}}^{\mathrm{T}} \triangledown \left(\boldsymbol{v}_{\boldsymbol{\beta}}^{\mathrm{T}} \triangledown^2 \phi_{ntj}(\boldsymbol{\beta}) \boldsymbol{v}_{\boldsymbol{\beta}}\right) \alpha_{N,j}\right| &= \left|\dddot{\boldsymbol{\phi}}_{nt}^{\mathrm{T}}(\boldsymbol{\beta}) \boldsymbol{\alpha}_N \left(\boldsymbol{v}_{\boldsymbol{\beta}}^{\mathrm{T}} \boldsymbol{X}_{nt}\right)^3\right| \\
&\leq \left|\dddot{\boldsymbol{\phi}}_{nt}^{\mathrm{T}}(\boldsymbol{\beta}) \boldsymbol{\alpha}_N\right| \|\boldsymbol{X}_{nt}\|_2^3 \|\boldsymbol{v}\|_2^3 \\
&= O(\|\boldsymbol{v}\|_2^3).
\end{aligned} \tag{F.9}$$

By Lemma 7.2.1 and Regularity Condition 7.A, we have

$$\begin{aligned}
\left(\sum_{j=0}^{J_N} \boldsymbol{v}_{\boldsymbol{\beta}}^{\mathrm{T}} \triangledown^2 \phi_{ntj}(\boldsymbol{\beta}) \boldsymbol{v}_{\boldsymbol{\beta}} \boldsymbol{v}_{\boldsymbol{\alpha}_N, j}\right)^2 &= \left(\sum_{j=0}^{J_N} \boldsymbol{v}_{\boldsymbol{\alpha}_N, j} \ddot{\phi}_{ntj}(\boldsymbol{\beta})\right)^2 \left(\boldsymbol{v}_{\boldsymbol{\beta}}^{\mathrm{T}} \boldsymbol{X}_{nt} \boldsymbol{X}_{nt}^{\mathrm{T}} \boldsymbol{v}_{\boldsymbol{\beta}}\right)^2 \\
&\leq \|\boldsymbol{v}_{\boldsymbol{\alpha}}\|_2^2 \times \|\boldsymbol{v}_{\boldsymbol{\beta}}\|_2^4 \times \|\ddot{\boldsymbol{\phi}}_{nt}(\boldsymbol{\beta})\|_2^2 \times \left(\boldsymbol{X}_{nt}^{\mathrm{T}} \boldsymbol{X}_{nt}\right)^2 \\
&= O(\|\boldsymbol{v}\|_2^6).
\end{aligned} \tag{F.10}$$

By Equation (F.8), (F.9) and (F.10), it follows that, for each $n \in \{1, \ldots, N\}$, $t \in \{1, \ldots, T_n\}$ and $(\boldsymbol{\beta}, \boldsymbol{\alpha}_N) \in \mathbb{R}^p \times \mathcal{M}_{J_N+1}$,

$$\partial^3 U_{nt}(\boldsymbol{\beta}, \boldsymbol{\alpha}_N) = O(\|\boldsymbol{v}\|_2^3).$$

□

Proof of Lemma F.4

Lemma F.4.

For each $n \in \{1, \ldots, N\}$ and $t \in \{1, \ldots, T_n\}$, let $\mathcal{G}_{nt}$ be a function of $(\boldsymbol{X}_{nt}, Z_{nt})$. For every integer N, suppose that there exists a finite number C_N such that

$$\sup_{n \in \{1,\ldots,N\}} \sup_{t \in \{1,\ldots,T_n\}} \mathcal{G}_{nt} \leq C_N$$

with probability one. Assume that Regularity Condition 7.I is satisfied. Then, for each n,

$$\frac{1}{N} \sum_{n=1}^{N} \sum_{t,t'=1}^{T_n} w_{ntt'} \mathcal{G}_{nt} U_{nt'}(\boldsymbol{\beta}^*, Q^*) = O_p(N^{-1/2}).$$

Proof. Let $B_n = N^{-1/2} \sum_{t,t'=1}^{T_n} w_{ntt'} \mathcal{G}_{nt} U_{nt'}(\boldsymbol{\beta}^*, Q^*)$. Firstly, we show that the variance of B_n, denoted by σ_n^2, is bounded. We have that

$$\begin{aligned} N\mathrm{Var}\left[B_n\right] &\leq \mathbb{E}\left[\left(\sum_{t,t'=1}^{T_n} w_{ntt'} \mathcal{G}_{nt} U_{nt'}(\boldsymbol{\beta}^*, Q^*)\right)^2\right] \\ &\leq \mathbb{E}\left[\left(\sum_{t,t'=1}^{T_n} w_{ntt'}^2 \mathcal{G}_{nt} \mathcal{G}_{nt'}\right) \times \left(\sum_{t=1}^{T_n} U_{nt}^2(\boldsymbol{\beta}^*, Q^*)\right)\right]. \end{aligned}$$

Because $\boldsymbol{W}_n$ is a positive definite matrix and $\mathcal{G}_{nt}$ is bounded for each t, we further have

$$\begin{aligned} N\mathrm{Var}\left[B_n\right] &\leq C\mathbb{E}\left[\sum_{t=1}^{T_n} U_{nt}^2(\boldsymbol{\beta}^*, Q^*)\right] \\ &= C \sum_{t=1}^{T_n} \mathbb{E}\left[U_{nt}^2(\boldsymbol{\beta}^*, Q^*)\right], \end{aligned}$$

where $C > 0$ is a finite number. By Regularity Condition 7.I, $\sigma_n^2 = O(N^{-1})$.

Next, we use the Lindeberg-Feller Central Limit Theorem [Bauer, 1996, p.g. 234], to show that $\sum_{n=1}^N B_n$ is $O_p(1)$. It suffices to check the Lindeberg condition, that is, $\forall\epsilon > 0$,

$$\lim_{N\to\infty} \frac{1}{\sum_{n=1}^N \sigma_n^2} \sum_{n=1}^N \mathbb{E}[B_n^2 I(|B_n| > \epsilon)] = 0.$$

Given $\epsilon > 0$, by the Cauchy-Schwarz inequality, we have

$$\sum_{n=1}^N \mathbb{E}[B_n^2 I(|B_n| > \epsilon)] \leq \sum_{n=1}^N \left(\mathbb{E}\left[B_n^4\right] \times \mathrm{pr}\left(|B_n| > \epsilon\right)\right)^{1/2}.$$

Using Chebyshev's inequality, we have

$$\mathrm{pr}\left(|B_n| > \epsilon\right) \leq \frac{\sigma_n^2}{\epsilon^2} = O(N^{-1}).$$

On the other hand, by Regularity Condition 7.I,

$$\begin{aligned}
\mathbb{E}\left[B_n^4\right] &\leq N^{-2}\mathbb{E}\left[\left(\sum_{t,t'=1}^{T_n} w_{ntt'}\mathcal{G}_{nt}U_{nt'}(\boldsymbol{\beta}^*, Q^*)\right)^4\right] \\
&\leq N^{-2}\mathbb{E}\left[\left(\sum_{t,t'=1}^{T_n} w_{ntt'}^2\mathcal{G}_{nt}\mathcal{G}_{nt'}\right)^2 \times \left(\sum_{t=1}^{T_n} U_{nt}^2(\boldsymbol{\beta}^*, Q^*)\right)^2\right] \\
&= N^{-2}C^2\mathbb{E}\left[\left(\sum_{t=1}^{T_n} U_{nt}^2(\boldsymbol{\beta}^*, Q^*)\right)^2\right] \\
&\leq N^{-2}C^2 \sum_{t,t'=1}^{T_n} \mathbb{E}\left[U_{nt}^2(\boldsymbol{\beta}^*, Q^*)U_{nt'}^2(\boldsymbol{\beta}^*, Q^*)\right] \\
&= O(N^{-2}).
\end{aligned}$$

Therefore, we have $\frac{1}{\sum_{n=1}^N \sigma_n^2}\sum_{n=1}^N \mathbb{E}[B_n^2 I(|B_n| > \epsilon)] = O(N^{-1/2})$. This completes the proof. □

Proof of Lemma F.5

Lemma F.5.

Assume that Regularity Condition 7.A-7.D and 7.I are satisfied. Then,

$$\frac{1}{N}\sum_{n=1}^{N}\sum_{t,t'=1}^{T_n} w_{ntt'}\partial U_{nt}(\boldsymbol{\beta}^*,\boldsymbol{\alpha}_N^*)U_{nt'}(\boldsymbol{\beta}^*,Q^*) = O_p(N^{-1/2}J_N^{1/2}\|\boldsymbol{v}\|_2),$$

$$\frac{1}{N}\sum_{n=1}^{N}\sum_{t,t'=1}^{T_n} w_{ntt'}\partial^2 U_{nt}(\boldsymbol{\beta}^*,\boldsymbol{\alpha}_N^*)U_{nt'}(\boldsymbol{\beta}^*,Q^*) = O_p(J_N^{1/2}N^{-1/2}\|\boldsymbol{v}\|_2^2),$$

where $\partial U_{nt}(\boldsymbol{\beta}^*,\boldsymbol{\alpha}_N^*)$ is defined in Equation (7.17) and $\partial^2 U_{nt}(\boldsymbol{\beta}^*,\boldsymbol{\alpha}_N^*)$ is defined in Equation (7.19).

Proof. By Equation (F.2) and the Cauchy-Schwarz inequality, we have

$$\begin{aligned}
&\left(\frac{1}{N}\sum_{n=1}^{N}\sum_{t,t'=1}^{T_n} w_{ntt'}\partial U_{nt}(\boldsymbol{\beta}^*,\boldsymbol{\alpha}_N^*)U_{nt'}(\boldsymbol{\beta}^*,Q^*)\right)^2\\
=&\left(\boldsymbol{v}^{\mathrm{T}}\left(\frac{1}{N}\sum_{n=1}^{N}\sum_{t,t'=1}^{T_n} w_{ntt'}U_{nt'}(\boldsymbol{\beta}^*,Q^*)\boldsymbol{h}_{nt}(\boldsymbol{\beta}^*,\boldsymbol{\alpha}_N^*)\right)\right)^2\\
\le&\left\|\frac{1}{N}\sum_{n=1}^{N}\sum_{t,t'=1}^{T_n} w_{ntt'}U_{nt'}(\boldsymbol{\beta}^*,Q^*)\boldsymbol{h}_{nt}(\boldsymbol{\beta}^*,\boldsymbol{\alpha}_N^*)\right\|_2^2\|\boldsymbol{v}\|_2^2,
\end{aligned}$$

where $\boldsymbol{h}_{nt}(\boldsymbol{\beta},\boldsymbol{\alpha}_N)\in\mathbb{R}^{J_N+p+1}$ is defined in Equation (F.3).

By Lemma 7.2.1 and 7.2.2, we have that, for every integer N,

$$\sup_{n\in\{1,\dots,N\}}\sup_{t\in\{1,\dots,T_n\}}\boldsymbol{h}_{nt}^{\mathrm{T}}(\boldsymbol{\beta}^*,\boldsymbol{\alpha}_N^*)\boldsymbol{h}_{nt}(\boldsymbol{\beta}^*,\boldsymbol{\alpha}_N^*)$$

is bounded. So, for every N and $j\in\{1,\dots,J_N\}$,

$$\sup_{n\in\{1,\dots,N\}}\sup_{t\in\{1,\dots,T_n\}}|h_{ntj}(\boldsymbol{\beta}^*,\boldsymbol{\alpha}_N^*)|$$

is bounded, where $h_{ntj}(\boldsymbol{\beta},\boldsymbol{\alpha}_N)$ be the j^{th} element of $\boldsymbol{h}_{nt}(\boldsymbol{\beta},\boldsymbol{\alpha}_N)$. By Lemma F.4, we have

$$\begin{aligned}
&\left\| \frac{1}{N}\sum_{n=1}^{N}\sum_{t,t'=1}^{T_n} w_{ntt'}U_{nt'}(\boldsymbol{\beta}^*, Q^*)\boldsymbol{h}_{nt}(\boldsymbol{\beta}^*, \boldsymbol{\alpha}_N^*)\right\|_2^2 \\
&= \sum_{j=0}^{J_N+p}\left(\frac{1}{N}\sum_{n=1}^{N}\sum_{t,t'=1}^{T_n} w_{ntt'}U_{nt'}(\boldsymbol{\beta}^*, Q^*)h_{ntj}(\boldsymbol{\beta}^*, \boldsymbol{\alpha}_N^*)\right)^2 \\
&= O_p(N^{-1}J_N).
\end{aligned}$$

Therefore,

$$\frac{1}{N}\sum_{n=1}^{N}\sum_{t,t'=1}^{T_n} w_{ntt'}\partial U_{nt}(\boldsymbol{\beta}^*, \boldsymbol{\alpha}_N^*)U_{nt'}(\boldsymbol{\beta}^*, Q^*) = O_p(J_N^{1/2}N^{-1/2}\|\boldsymbol{v}\|_2).$$

By Equation (F.5), we have

$$\begin{aligned}
&\frac{1}{N}\sum_{n=1}^{N}\sum_{t,t'=1}^{T_n} w_{ntt'}\partial^2 U_{nt}(\boldsymbol{\beta}^*, \boldsymbol{\alpha}_N^*)U_{nt'}(\boldsymbol{\beta}^*, Q^*) \\
= &-\frac{1}{N}\sum_{n=1}^{N}\sum_{t,t'=1}^{T_n} w_{ntt'}\boldsymbol{v}_{\boldsymbol{\beta}}^{\mathrm{T}}\left(\sum_{j=0}^{J_N}\alpha_{N,j}^*\triangledown^2\phi_{ntj}(\boldsymbol{\beta}^*)\right)\boldsymbol{v}_{\boldsymbol{\beta}}U_{nt'}(\boldsymbol{\beta}^*, Q^*) \\
&-\frac{2}{N}\sum_{n=1}^{N}\sum_{t,t'=1}^{T_n} w_{ntt'}\boldsymbol{v}_{\boldsymbol{\beta}}^{\mathrm{T}}\triangledown\boldsymbol{\phi}_{nt}(\boldsymbol{\beta}^*)\boldsymbol{v}_{\boldsymbol{\alpha}}U_{nt'}(\boldsymbol{\beta}^*, Q^*),
\end{aligned}$$

where for each n, t and j, $\triangledown\boldsymbol{\phi}_{nt}(\boldsymbol{\beta}^*)$ is defined in Equation (F.4).

Consider the first term of on the right hand side of the above equation. By Regularity Condition 7.A and Lemma 7.2.2, we have, for every integer N, each element of $\sum_{j=0}^{J_N}\alpha_{N,j}^*\triangledown\boldsymbol{\phi}_{nt}(\boldsymbol{\beta}^*)$ is uniformly bounded over $n \in \{1,\dots,N\}$ and $t \in \{1,\dots,T_n\}$. So is the largest eigenvalue of $\sum_{j=0}^{J_N}\alpha_{N,j}^*\triangledown\boldsymbol{\phi}_{nt}(\boldsymbol{\beta}^*)$. By Lemma F.4, we have

$$\begin{aligned}
&\frac{1}{N}\sum_{n=1}^{N}\sum_{t,t'=1}^{T_n} w_{ntt'}\boldsymbol{v}_{\boldsymbol{\beta}}^{\mathrm{T}}\left(\sum_{j=0}^{J_N}\alpha_{N,j}^*\triangledown^2\phi_{ntj}(\boldsymbol{\beta}^*)\right)\boldsymbol{v}_{\boldsymbol{\beta}}U_{nt'}(\boldsymbol{\beta}^*, Q^*) \\
\le &\frac{1}{N}\sum_{n=1}^{N}\sum_{t,t'=1}^{T_n} w_{ntt'}U_{nt'}(\boldsymbol{\beta}^*, Q^*)\lambda_{\max}\left(\sum_{j=0}^{J_N}\alpha_{N,j}^*\triangledown^2\phi_{ntj}(\boldsymbol{\beta}^*)\right)\|\boldsymbol{v}_{\boldsymbol{\beta}}\|_2^2 \\
= &O_p(N^{-1/2}\|\boldsymbol{v}\|_2^2).
\end{aligned}$$

By Regularity Condition 7.A and Lemma 7.2.1, each element of $\boldsymbol{X}_{nt}$, denoted by X_{nti}, and $\dot{\phi}_{ntj}(\boldsymbol{\beta}^*)$ are uniformly bounded over $n \in \{1, \ldots, N\}$ and $t \in \{1, \ldots, T_n\}$. Then, by the Cauchy-Schwarz inequality and Lemma F.4, we also have

$$
\begin{aligned}
&\left(\frac{1}{N}\sum_{n=1}^{N}\sum_{t,t'=1}^{T_n} w_{ntt'}\boldsymbol{v}_{\boldsymbol{\beta}}^{\mathrm{T}}\triangledown\boldsymbol{\phi}_{nt}(\boldsymbol{\beta})\boldsymbol{v}_{\boldsymbol{\alpha}}U_{nt'}(\boldsymbol{\beta}^*, Q^*)\right)^2 \\
= &\left(\frac{1}{N}\sum_{n=1}^{N}\sum_{t,t'=1}^{T_n} w_{ntt'}\boldsymbol{v}_{\boldsymbol{\beta}}^{\mathrm{T}}\boldsymbol{X}_{nt}\dot{\boldsymbol{\phi}}_{nt}^{\mathrm{T}}(\boldsymbol{\beta}^*)\boldsymbol{v}_{\boldsymbol{\alpha}}U_{nt'}(\boldsymbol{\beta}^*, Q^*)\right)^2 \\
\leq &\left\|\frac{1}{N}\sum_{n=1}^{N}\sum_{t,t'=1}^{T_n} w_{ntt'}\boldsymbol{v}_{\boldsymbol{\beta}}^{\mathrm{T}}\boldsymbol{X}_{nt}U_{nt'}(\boldsymbol{\beta}^*, Q^*)\dot{\boldsymbol{\phi}}_{nt}(\boldsymbol{\beta}^*)\right\|_2^2 \|\boldsymbol{v}_{\boldsymbol{\alpha}}\|_2^2 \\
\leq &\sum_{j=0}^{J_N}\left(\boldsymbol{v}_{\boldsymbol{\beta}}^{\mathrm{T}}\frac{1}{N}\sum_{n=1}^{N}\sum_{t,t'=1}^{T_n} w_{ntt'}\boldsymbol{X}_{nt}U_{nt'}(\boldsymbol{\beta}^*, Q^*)\dot{\phi}_{ntj}(\boldsymbol{\beta}^*)\right)^2 \|\boldsymbol{v}_{\boldsymbol{\alpha}}\|_2^2 \\
\leq &\sum_{j=0}^{J_N}\|\boldsymbol{v}_{\boldsymbol{\beta}}\|_2^2\left\|\frac{1}{N}\sum_{n=1}^{N}\sum_{t,t'=1}^{T_n} w_{ntt'}\boldsymbol{X}_{nt}U_{nt'}(\boldsymbol{\beta}^*, Q^*)\dot{\phi}_{ntj}(\boldsymbol{\beta}^*)\right\|_2^2 \|\boldsymbol{v}_{\boldsymbol{\alpha}}\|_2^2 \\
\leq &\sum_{j=0}^{J_N}\|\boldsymbol{v}_{\boldsymbol{\beta}}\|_2^2\|\boldsymbol{v}_{\boldsymbol{\alpha}}\|_2^2\sum_{i=1}^{p}\left(\frac{1}{N}\sum_{n=1}^{N}\sum_{t,t'=1}^{T_n} w_{ntt'}X_{nti}U_{nt'}(\boldsymbol{\beta}^*, Q^*)\dot{\phi}_{ntj}(\boldsymbol{\beta}^*)\right)^2 \\
= &O_p(J_N N^{-1}\|\boldsymbol{v}\|_2^4).
\end{aligned}
$$

In sum,

$$
\frac{1}{N}\sum_{n=1}^{N}\sum_{t,t'=1}^{T_n} w_{ntt'}\partial^2 U_{nt}(\boldsymbol{\beta}^*, \boldsymbol{\alpha}_N^*)U_{nt}(\boldsymbol{\beta}^*, Q^*) = O_p(J_N^{1/2}N^{-1/2}\|\boldsymbol{v}\|_2^2).
$$

□

F.3 Proofs of the Lemmas for Theorem 7.5.1

Proof of Lemma F.6

Lemma F.6.

Assume that Regularity Condition 7.A-7.D are satisfied. Let $(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\alpha}}_N)$ *be an estimator which converges to* $(\boldsymbol{\beta}^*, \boldsymbol{\alpha}_N^*)$ *in the sense that*

$$
\|\hat{\boldsymbol{\alpha}}_N - \boldsymbol{\alpha}_N^*\|_2^2 + \|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\|_2^2 = O_p(J_N N^{-1}).
$$

Then, as the sample size N goes to infinity, for each $n \in \{1, \dots, N\}$ and $t \in \{1, \dots, T_n\}$,

$$\dot{\boldsymbol{\phi}}_{nt}^{\mathrm{T}}(\hat{\boldsymbol{\beta}})\hat{\boldsymbol{\alpha}}_N = \int_{\mathcal{B}} \dot{g}^{-1}(\boldsymbol{X}_{nt}^{\mathrm{T}}\boldsymbol{\beta}^* + Z_{nt}b_n)\mathrm{d}Q^* \\ + \begin{bmatrix} \hat{\boldsymbol{\alpha}}_N - \boldsymbol{\alpha}_N^* \\ \hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^* \end{bmatrix}^{\mathrm{T}} \begin{bmatrix} \dot{\boldsymbol{\phi}}_{nt}(\boldsymbol{\beta}^*) \\ \ddot{\boldsymbol{\phi}}_{nt}^{\mathrm{T}}(\boldsymbol{\beta}^*)\boldsymbol{\alpha}_N^*\boldsymbol{X}_{nt} \end{bmatrix} + O_p(J_N N^{-1}), \tag{F.11}$$

where $\dot{\boldsymbol{\phi}}_{nt}(\boldsymbol{\beta}) \in \mathbb{R}^{J_N+1}$ is defined in Equation (7.6) and $\ddot{\boldsymbol{\phi}}_{nt}(\boldsymbol{\beta})$ is defined in Equation (7.7).

Proof. By Taylor's expansion with respect to $\boldsymbol{\beta} \in \mathbb{R}^p$ locally at $\boldsymbol{\beta}^* \in \mathbb{R}^p$, we have

$$\begin{aligned} \dot{\boldsymbol{\phi}}_{nt}^{\mathrm{T}}(\hat{\boldsymbol{\beta}})\hat{\boldsymbol{\alpha}}_N = & \int_{\mathcal{B}} \dot{g}^{-1}(\boldsymbol{X}_{nt}^{\mathrm{T}}\boldsymbol{\beta}^* + Z_{nt}b_n)\mathrm{d}Q^* + \dot{\boldsymbol{\phi}}_{nt}^{\mathrm{T}}(\boldsymbol{\beta}^*)\left(\hat{\boldsymbol{\alpha}}_N - \boldsymbol{\alpha}_N^*\right) \\ & + \ddot{\boldsymbol{\phi}}_{nt}(\boldsymbol{\beta}^*)^{\mathrm{T}}\boldsymbol{\alpha}_N^*\boldsymbol{X}_{nt}^{\mathrm{T}}\left(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\right) \\ & + \ddot{\boldsymbol{\phi}}_{nt}(\boldsymbol{\beta}^*)^{\mathrm{T}}\left(\hat{\boldsymbol{\alpha}}_N - \boldsymbol{\alpha}_N^*\right)\boldsymbol{X}_{nt}^{\mathrm{T}}\left(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\right) \\ & + \left(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\right)^{\mathrm{T}}\boldsymbol{X}_{nt}\dddot{\boldsymbol{\phi}}_{nt}^{\mathrm{T}}(\breve{\boldsymbol{\beta}})\hat{\boldsymbol{\alpha}}_N\boldsymbol{X}_{nt}^{\mathrm{T}}\left(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\right) \\ & + \dot{\boldsymbol{\phi}}_{nt}^{\mathrm{T}}(\boldsymbol{\beta}^*)\boldsymbol{\alpha}_N^* - \int_{\mathcal{B}} \dot{g}^{-1}(\boldsymbol{X}_{nt}^{\mathrm{T}}\boldsymbol{\beta}^* + Z_{nt}b_n)\mathrm{d}Q^*, \end{aligned}$$

where $\breve{\boldsymbol{\beta}}$ is on the line segment between $\boldsymbol{\beta}^*$, and $\hat{\boldsymbol{\beta}}$ and $\dddot{\boldsymbol{\phi}}_{nt}(\boldsymbol{\beta}, \boldsymbol{\alpha}_N)$ is defined in Equation (7.8). By Lemma 7.2.1 and 7.2.2, and Regularity Condition 7.A, we have, for each $n \in \{1, \dots, N\}$ and $t \in \{1, \dots, T_n\}$,

$$\begin{aligned} & \left|\ddot{\boldsymbol{\phi}}_{nt}(\boldsymbol{\beta}^*)^{\mathrm{T}}\left(\hat{\boldsymbol{\alpha}}_N - \boldsymbol{\alpha}_N^*\right)\boldsymbol{X}_{nt}^{\mathrm{T}}\left(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\right)\right| \\ \leq & \left\|\ddot{\boldsymbol{\phi}}_{nt}(\boldsymbol{\beta}^*)\right\|_2 \|\boldsymbol{X}_{nt}\|_2 \|\hat{\boldsymbol{\alpha}}_N - \boldsymbol{\alpha}_N^*\|_2 \left\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\right\|_2 \\ \leq & O_p(J_N N^{-1}), \end{aligned}$$

and

$$\begin{aligned} & \left|\left(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\right)^{\mathrm{T}}\boldsymbol{X}_{nt}\dddot{\boldsymbol{\phi}}_{nt}^{\mathrm{T}}(\breve{\boldsymbol{\beta}})\hat{\boldsymbol{\alpha}}_N\boldsymbol{X}_{nt}^{\mathrm{T}}\left(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\right)\right| \\ \leq & \left|\dddot{\boldsymbol{\phi}}_{nt}^{\mathrm{T}}(\hat{\boldsymbol{\beta}})\breve{\boldsymbol{\alpha}}_N\right| \|\boldsymbol{X}_{nt}\|_2^2 \left\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\right\|_2^2 \\ = & O_p(J_N N^{-1}) \end{aligned}$$

and

$$\left|\dot{\boldsymbol{\phi}}_{nt}^{\mathrm{T}}(\boldsymbol{\beta}^*)\boldsymbol{\alpha}_N^* - \int_{\mathcal{B}} \dot{g}^{-1}(\boldsymbol{X}_{nt}^{\mathrm{T}}\boldsymbol{\beta}^* + Z_{nt}b_n)\mathrm{d}Q^*\right| = o(J_N N^{-1}).$$

Therefore, we have Equation (F.11). □

Proof of Lemma F.7

Lemma F.7.
Assume that Regularity Condition 7.A-7.D are satisfied. Let $(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\alpha}}_N)$ be an estimator which converges to $(\boldsymbol{\beta}^*, \boldsymbol{\alpha}_N^*)$ in the sense that

$$\|\hat{\boldsymbol{\alpha}}_N - \boldsymbol{\alpha}_N^*\|_2^2 + \|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\|_2^2 = O_p(J_N N^{-1}).$$

Then, as the sample size N goes to infinity, for each $n \in \{1, \ldots, N\}$ and $t \in \{1, \ldots, T_n\}$,

$$\begin{aligned} U_{nt}(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\alpha}}_N) = U_{nt}(\boldsymbol{\beta}^*, Q^*) &- \boldsymbol{\phi}_{nt}^{\mathrm{T}}(\boldsymbol{\beta}^*)\left(\hat{\boldsymbol{\alpha}}_N - \boldsymbol{\alpha}_N^*\right) \\ &- \dot{\boldsymbol{\phi}}_{nt}^{\mathrm{T}}(\boldsymbol{\beta}^*)\boldsymbol{\alpha}_N^* \boldsymbol{X}_{nt}^{\mathrm{T}}\left(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\right) + o_p(J_N N^{-1}), \end{aligned}$$

where $\boldsymbol{\phi}_{nt}(\boldsymbol{\beta})$ is defined in Equation (7.3) and $\dot{\boldsymbol{\phi}}_{nt}(\boldsymbol{\beta})$ is defined in Equation (7.4).

Proof. By Taylor's expansion with respect to $\boldsymbol{\beta} \in \mathbb{R}^p$ locally at $\boldsymbol{\beta}^* \in \mathbb{R}^p$, we have

$$\begin{aligned} U_{nt}(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\alpha}}_N) = U_{nt}(\boldsymbol{\beta}^*, Q^*) &- \boldsymbol{\phi}_{nt}^{\mathrm{T}}(\boldsymbol{\beta}^*)\left(\hat{\boldsymbol{\alpha}}_N - \boldsymbol{\alpha}_N^*\right) \\ &- \dot{\boldsymbol{\phi}}_{nt}^{\mathrm{T}}(\boldsymbol{\beta}^*)\boldsymbol{\alpha}_N^* \boldsymbol{X}_{nt}^{\mathrm{T}}\left(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\right) \\ &- \left(\hat{\boldsymbol{\alpha}}_N - \boldsymbol{\alpha}_N^*\right)^{\mathrm{T}} \dot{\boldsymbol{\phi}}_{nt}(\boldsymbol{\beta}^*)\boldsymbol{X}_{nt}^{\mathrm{T}}\left(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\right) \\ &- \left(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\right)^{\mathrm{T}} \boldsymbol{X}_{nt}\ddot{\boldsymbol{\phi}}_{nt}^{\mathrm{T}}(\breve{\boldsymbol{\beta}})\hat{\boldsymbol{\alpha}}_N \boldsymbol{X}_{nt}^{\mathrm{T}}\left(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\right) \\ &+ U_{nt}(\boldsymbol{\beta}^*, \boldsymbol{\alpha}_N^*) - U_{nt}(\boldsymbol{\beta}^*, Q^*), \end{aligned}$$

where $\breve{\boldsymbol{\beta}}$ is on the line segment between $\boldsymbol{\beta}^*$, and $\ddot{\boldsymbol{\phi}}_{nt}(\boldsymbol{\beta}, \boldsymbol{\alpha}_N)$ is defined in Equation (7.5). By Lemma 7.2.1 and 7.2.2, and Regularity Condition 7.A, we have, for each

$n \in \{1, \ldots, N\}$ and $t \in \{1, \ldots, T_n\}$,

$$\begin{aligned}
&\left|\left(\hat{\boldsymbol{\alpha}}_N - \boldsymbol{\alpha}_N^*\right)^{\mathrm{T}} \dot{\boldsymbol{\phi}}_{nt}(\boldsymbol{\beta}^*)\boldsymbol{X}_{nt}^{\mathrm{T}}\left(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\right)\right| \\
\leq\ &\left\|\dot{\boldsymbol{\phi}}_{nt}(\boldsymbol{\beta}^*)\right\|_2 \left\|\boldsymbol{X}_{nt}\right\|_2 \left\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\right\|_2 \left\|\hat{\boldsymbol{\alpha}}_N - \boldsymbol{\alpha}_N^*\right\|_2 \\
=\ &O_p(J_N N^{-1})
\end{aligned}$$

and

$$\begin{aligned}
&\left|\left(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\right)^{\mathrm{T}} \boldsymbol{X}_{nt}\ddot{\boldsymbol{\phi}}_{nt}^{\mathrm{T}}(\breve{\boldsymbol{\beta}})\boldsymbol{\alpha}_N^*\boldsymbol{X}_{nt}^{\mathrm{T}}\left(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\right)\right| \\
\leq\ &\boldsymbol{X}_{nt}^{\mathrm{T}}\boldsymbol{X}_{nt}\left|\ddot{\boldsymbol{\phi}}_{nt}^{\mathrm{T}}(\breve{\boldsymbol{\beta}})\boldsymbol{\alpha}_N^*\right| \left\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\right\|_2^2 \\
=\ &O_p(J_N N^{-1})
\end{aligned}$$

and

$$\left|U_{nt}(\boldsymbol{\beta}^*, \boldsymbol{\alpha}_N^*) - U_{nt}(\boldsymbol{\beta}^*, Q^*)\right| = o(J_N N^{-1}).$$

Therefore, we have

$$\begin{aligned}
&\left(\hat{\boldsymbol{\alpha}}_N - \boldsymbol{\alpha}_N^*\right)^{\mathrm{T}} \dot{\boldsymbol{\phi}}_{nt}(\boldsymbol{\beta}^*)\boldsymbol{X}_{nt}^{\mathrm{T}}\left(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\right) - U_{nt}(\boldsymbol{\beta}^*, \boldsymbol{\alpha}_N^*) + U_{nt}(\boldsymbol{\beta}^*, Q^*) \\
&+ \left(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\right)^{\mathrm{T}} \boldsymbol{X}_{nt}\ddot{\boldsymbol{\phi}}_{nt}^{\mathrm{T}}(\breve{\boldsymbol{\beta}})\hat{\boldsymbol{\alpha}}_N\boldsymbol{X}_{nt}^{\mathrm{T}}\left(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\right) \\
=\ &O_p(J_N N^{-1}).
\end{aligned}$$

□

Proof of Lemma F.8

Lemma F.8.

Assume that Regularity Condition 7.A-7.D are satisfied. Let $(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\alpha}}_N)$ be an estimator which converges to $(\boldsymbol{\beta}^*, \boldsymbol{\alpha}_N^*)$ in the sense that

$$\|\hat{\boldsymbol{\alpha}}_N - \boldsymbol{\alpha}_N^*\|_2^2 + \|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\|_2^2 = O_p(J_N N^{-1}).$$

Then, as the sample size N goes to infinity, for each $n \in \{1, \ldots, N\}$ and $i \in \{1, \ldots, p\}$,

$$\frac{1}{N}\sum_{n=1}^{N} \boldsymbol{X}_{n\cdot i}\left(\boldsymbol{D}_n(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\alpha}}_N) - \boldsymbol{D}_n^*\right)\boldsymbol{W}_n\boldsymbol{U}_n(\boldsymbol{\beta}^*, Q^*) = O_p(J_N N^{-1}),$$

where $\boldsymbol{X}_{n\cdot i}$ is the i^{th} row of $\boldsymbol{X}_n$ and $\boldsymbol{D}_n(\boldsymbol{\beta}, \boldsymbol{\alpha}_N)$ is a diagonal matrix whose diagonal elements are $\dot{\boldsymbol{\phi}}_{nt}^{\mathrm{T}}(\boldsymbol{\beta})\boldsymbol{\alpha}_N$.

Proof. By Lemma F.6, we have

$$\begin{aligned}
&\boldsymbol{X}_{n\cdot i}\left(\boldsymbol{D}_n(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\alpha}}_N) - \boldsymbol{D}_n^*\right)\boldsymbol{W}_n\boldsymbol{U}_n(\boldsymbol{\beta}^*, Q^*)\\
=&\left(\begin{bmatrix}\hat{\boldsymbol{\alpha}}_N - \boldsymbol{\alpha}_N^*\\ \hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\end{bmatrix}^{\mathrm{T}}\begin{bmatrix}\dot{\boldsymbol{\Phi}}_n(\boldsymbol{\beta}^*)\\ \boldsymbol{X}_n\mathrm{diag}\left(\ddot{\boldsymbol{\Phi}}_n^{\mathrm{T}}(\boldsymbol{\beta}^*)\boldsymbol{\alpha}_N^*\right)\end{bmatrix} + O_p(J_N N^{-1})\right)\mathrm{diag}(\boldsymbol{X}_{n\cdot i})\\
&\times\boldsymbol{W}_n\boldsymbol{U}_n(\boldsymbol{\beta}^*, Q^*),
\end{aligned}$$

where $\dot{\boldsymbol{\Phi}}_n(\boldsymbol{\beta})$ is the $(J_N+1)\times T_n$ matrix whose columns are $\{\dot{\boldsymbol{\phi}}_{nt}(\boldsymbol{\beta})\}_{t=1}^{T_n}$, $\ddot{\boldsymbol{\Phi}}_n(\boldsymbol{\beta})$ is the $(J_N+1)\times T_n$ matrix whose columns are $\{\ddot{\boldsymbol{\phi}}_{nt}(\boldsymbol{\beta})\}_{t=1}^{T_n}$, and for any vector $\boldsymbol{A}$, $\mathrm{diag}(\boldsymbol{A})$ is the diagonal matrix whose diagonal elements are $\boldsymbol{A}$.

For each n and i, let

$$\boldsymbol{d}_{ni} = \begin{bmatrix}\dot{\boldsymbol{\Phi}}_n(\boldsymbol{\beta}^*)\\ \boldsymbol{X}_n\mathrm{diag}\left(\ddot{\boldsymbol{\Phi}}_n^{\mathrm{T}}(\boldsymbol{\beta}^*)\boldsymbol{\alpha}_N^*\right)\end{bmatrix}\mathrm{diag}(\boldsymbol{X}_{n\cdot i})\boldsymbol{W}_n\boldsymbol{U}_n(\boldsymbol{\beta}^*, Q^*) \in \mathbb{R}^{J_N+1+p}.$$

By Lemma 7.2.1 and 7.2.2, and Regularity Condition 7.A, we have that, for every integer N, each element of

$$\begin{bmatrix}\dot{\boldsymbol{\Phi}}_n(\boldsymbol{\beta}^*)\\ \boldsymbol{X}_n\mathrm{diag}\left(\ddot{\boldsymbol{\Phi}}_n^{\mathrm{T}}(\boldsymbol{\beta}^*)\boldsymbol{\alpha}_N^*\right)\end{bmatrix}\mathrm{diag}(\boldsymbol{X}_{n\cdot i})$$

is uniformly bounded over $n \in \{1, \ldots, N\}$ and $t \in \{1, \ldots, T_n\}$. Then, by Lemma F.4, we have that each element of $\frac{1}{N}\sum_{n=1}^{N}\boldsymbol{d}_{ni}$ is $O_p(N^{-1/2})$. Therefore, we have

$$\left\|\frac{1}{N}\sum_{n=1}^{N}\boldsymbol{d}_{ni}\right\|_2 = O_p(J_N^{1/2}N^{-1/2}).$$

By the Cauchy-Schwarz inequality, we have

$$
\begin{aligned}
&\left| \frac{1}{N} \sum_{n=1}^{N} \begin{bmatrix} \hat{\boldsymbol{\alpha}}_N - \boldsymbol{\alpha}_N^* \\ \hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^* \end{bmatrix}^{\mathrm{T}} \boldsymbol{d}_{ni} \right| \\
= &\left| \begin{bmatrix} \hat{\boldsymbol{\alpha}}_N - \boldsymbol{\alpha}_N^* \\ \hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^* \end{bmatrix}^{\mathrm{T}} \frac{1}{N} \sum_{n=1}^{N} \boldsymbol{d}_{ni} \right| \\
\leq &\left(\|\hat{\boldsymbol{\alpha}}_N - \boldsymbol{\alpha}_N^*\|_2^2 + \left\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\right\|_2^2 \right)^{1/2} \left\| \frac{1}{N} \sum_{n=1}^{N} \boldsymbol{d}_{ni} \right\|_2 \\
= &O_p(J_N^{1/2} N^{-1/2}) \times O_p(J_N^{1/2} N^{-1/2}) \\
= &O_p(J_N N^{-1}).
\end{aligned}
$$

This completes the proof. □

Proof of Lemma F.9

Lemma F.9.
Assume that Regularity Condition 7.A-7.G are satisfied. Further assume that the initial estimator $(\tilde{\boldsymbol{\beta}}, \tilde{\boldsymbol{\alpha}}_N)$ converges to $(\boldsymbol{\beta}_0, \boldsymbol{\alpha}_{N,0})$ in the sense that

$$\|\tilde{\boldsymbol{\alpha}}_N - \boldsymbol{\alpha}_{N,0}\|_2^2 + \|\tilde{\boldsymbol{\beta}} - \boldsymbol{\beta}_0\|_2^2 = O_p(J_N N^{-1}),$$

where there exists a probability measure Q_0 defined on $\mathcal{B}$ such that $\boldsymbol{\alpha}_{N,0} = \int_{\mathcal{B}} \boldsymbol{P}(b) \mathrm{d}Q_0$. Then, as the sample size N goes to infinity, for each $n \in \{1, \ldots, N\}$ and $i \in \{1, \ldots, p\}$,

$$\frac{1}{N} \sum_{n=1}^{N} \boldsymbol{X}_{n \cdot i} \boldsymbol{D}_n^* \left(\tilde{\boldsymbol{W}}_n - \boldsymbol{W}_n(\boldsymbol{\beta}_0, Q_0) \right) \boldsymbol{U}_n(\boldsymbol{\beta}^*, Q^*) = O_p(J_N N^{-1}),$$

where $\boldsymbol{X}_{n \cdot i}$ is the i^{th} row of $\boldsymbol{X}_n$ and $\boldsymbol{D}_n(\boldsymbol{\beta}, \boldsymbol{\alpha}_N)$ be a diagonal matrix whose diagonal elements are $\dot{\boldsymbol{\phi}}_{nt}^{\mathrm{T}}(\boldsymbol{\beta}) \boldsymbol{\alpha}_N$.

Proof. By Equation (7.26), we have, for each $n \in \{1, \ldots, N\}$ and $t \in \{1, \ldots, T_n\}$,

$$
\begin{aligned}
&V_{nt}^{-1}(\tilde{\boldsymbol{\beta}}, \tilde{\boldsymbol{\alpha}}_N) - V_{nt}^{-1}(\boldsymbol{\beta}_0, Q_0) \\
= &- V_{nt}^{-2}(\boldsymbol{\beta}_0, Q_0) \left(\partial V_{nt}(\boldsymbol{\beta}_0, \boldsymbol{\alpha}_{N,0}) \Delta \right) + O_p(J_N N^{-1}),
\end{aligned}
$$

where $\Delta = O_p(J_N^{1/2}N^{-1/2})$ and $\partial V_{nt}(\boldsymbol{\beta}, \boldsymbol{\alpha})$ is defined in Equation (7.25).

Further, by Equation (7.24),

$$\begin{aligned}
&\boldsymbol{X}_{n\cdot i}\boldsymbol{D}_n^* \left(\tilde{\boldsymbol{W}}_n - \boldsymbol{W}_n(\boldsymbol{\beta}_0, Q_0)\right)\boldsymbol{U}_n(\boldsymbol{\beta}^*, Q^*) \\
&= \boldsymbol{X}_{n\cdot i}\boldsymbol{D}_n^*\boldsymbol{R}_n^{-1/2}\left(\boldsymbol{V}_n^{-1}(\tilde{\boldsymbol{\beta}}, \tilde{\boldsymbol{\alpha}}_N) - \boldsymbol{V}_n^{-1}(\boldsymbol{\beta}_0, Q_0)\right)\boldsymbol{R}_n^{-1/2}\boldsymbol{U}_n(\boldsymbol{\beta}^*, Q^*) \\
&= \Delta\left[V_{n1}^{-2}(\boldsymbol{\beta}_0, Q_0)\partial V_{n1}(\boldsymbol{\beta}_0, \boldsymbol{\alpha}_{N,0}), \ldots, V_{nT_n}^{-2}(\boldsymbol{\beta}_0, Q_0)\partial V_{nT_n}(\boldsymbol{\beta}_0, \boldsymbol{\alpha}_{N,0})\right] \\
&\quad \times \operatorname{diag}\left(\boldsymbol{X}_{n\cdot i}\boldsymbol{D}_n^*\boldsymbol{R}_n^{-1/2}\right) \times \boldsymbol{R}_n^{-1/2}\boldsymbol{U}_n(\boldsymbol{\beta}^*, Q^*) + O_p(J_N N^{-1}) \\
&= \Delta\begin{bmatrix}\boldsymbol{v}_{\boldsymbol{\alpha}} \\ \boldsymbol{v}_{\boldsymbol{\beta}}\end{bmatrix}^{\mathrm{T}}\boldsymbol{C}_n\operatorname{diag}\left(\boldsymbol{X}_{n\cdot i}\boldsymbol{D}_n^*\boldsymbol{R}_n^{-1/2}\right) \times \boldsymbol{R}_n^{-1/2}\boldsymbol{U}_n(\boldsymbol{\beta}^*, Q^*) + O_p(J_N N^{-1}),
\end{aligned}$$

where C_n is a $J_N + p + 1 \times T_n$ matrix whose t^{th} column is

$$\begin{bmatrix}\sigma\boldsymbol{a}_{nt}(\boldsymbol{\beta}_0) + \boldsymbol{c}_{ntt}(\boldsymbol{\beta}_0) - 2\boldsymbol{\phi}_{nt}^{\mathrm{T}}(\boldsymbol{\beta}_0)\boldsymbol{\alpha}_{N,0}\boldsymbol{\phi}_{nt}(\boldsymbol{\beta}_0) \\ \boldsymbol{X}_{nt}\left(\sigma\dot{\boldsymbol{a}}_{nt}^{\mathrm{T}}(\boldsymbol{\beta})\boldsymbol{\alpha}_N + \dot{\boldsymbol{c}}_{ntt}^{\mathrm{T}}(\boldsymbol{\beta})\boldsymbol{\alpha}_N - 2\boldsymbol{\phi}_{nt}^{\mathrm{T}}(\boldsymbol{\beta})\boldsymbol{\alpha}_N\dot{\boldsymbol{\phi}}_{nt}^{\mathrm{T}}(\boldsymbol{\beta})\boldsymbol{\alpha}_N\right)\end{bmatrix}.$$

For each n and i, let

$$\boldsymbol{d}'_{ni} = \boldsymbol{C}_n\operatorname{diag}\left(\boldsymbol{X}_{n\cdot i}\boldsymbol{D}_n^*\boldsymbol{R}_n^{-1/2}\right) \times \boldsymbol{R}_n^{-1/2}\boldsymbol{U}_n(\boldsymbol{\beta}^*, Q^*).$$

By Lemma 7.2.1-7.2.6, and Regularity Condition 7.A, we have that, for every integer N, each element of

$$\boldsymbol{C}_n\operatorname{diag}\left(\boldsymbol{X}_{n\cdot i}\boldsymbol{D}_n^*\boldsymbol{R}_n^{-1/2}\right)\boldsymbol{R}_n^{-1/2}$$

is uniformly bounded over $n \in \{1, \ldots, N\}$ and $t \in \{1, \ldots, T_n\}$; see the proof of Theorem 7.4.1 for more details. Then, by Lemma F.4, we have that each element of $\frac{1}{N}\sum_{n=1}^N \boldsymbol{d}'_{ni}$ is $O_p(N^{-1/2})$. Therefore, we have

$$\left\|\frac{1}{N}\sum_{n=1}^N \boldsymbol{d}'_{ni}\right\|_2 = O_p(J_N^{1/2}N^{-1/2}).$$

By the Cauchy-Schwarz inequality, we have

$$\begin{aligned}
\Delta\left|\frac{1}{N}\sum_{n=1}^N\begin{bmatrix}\boldsymbol{v}_{\boldsymbol{\alpha}} \\ \boldsymbol{v}_{\boldsymbol{\beta}}\end{bmatrix}^{\mathrm{T}}\boldsymbol{d}'_{ni}\right| &= \Delta\left|\begin{bmatrix}\boldsymbol{v}_{\boldsymbol{\alpha}} \\ \boldsymbol{v}_{\boldsymbol{\beta}}\end{bmatrix}^{\mathrm{T}}\frac{1}{N}\sum_{n=1}^N\boldsymbol{d}'_{ni}\right| \\
&\leq \Delta\left\|\frac{1}{N}\sum_{n=1}^N\boldsymbol{d}'_{ni}\right\|_2 \\
&= O_p(J_N N^{-1}).
\end{aligned}$$

This completes the proof. □

Proof of Lemma F.10

Lemma F.10.

Assume that Regularity Condition 7.A-7.D are satisfied. Let $(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\alpha}}_N)$ be an estimator which converges to $(\boldsymbol{\beta}^*, \boldsymbol{\alpha}_N^*)$ in the sense that

$$\|\hat{\boldsymbol{\alpha}}_N - \boldsymbol{\alpha}_N^*\|_2^2 + \|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\|_2^2 = O_p(J_N N^{-1}).$$

Then, as the sample size N goes to infinity, $\boldsymbol{G}_n(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\alpha}}_N)$ element-wise converges to $\boldsymbol{G}_n(\boldsymbol{\beta}^*, \boldsymbol{\alpha}_N^*)$ in probability with rate $N^{1/2} J_N^{-1/2}$.

Proof. By Lemma F.6, it is shown that $\boldsymbol{D}_n(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\alpha}}_N)$ element-wise converges to $\boldsymbol{D}_n(\boldsymbol{\beta}^*, \boldsymbol{\alpha}_N^*)$ in probability with rate $N^{1/2} J_N^{-1/2}$. The rest to show is that $\boldsymbol{\Phi}_n(\hat{\boldsymbol{\beta}})$ element-wise converges to $\boldsymbol{\Phi}_n(\boldsymbol{\beta}^*)$ in probability at rate $N^{1/2} J_N^{-1/2}$.

By Taylor's expansion, we have, for each $n \in \{1, \ldots, N\}$ and $t \in \{1, \ldots, T_n\}$ and $j \in \{0, \ldots, J_N\}$,

$$\begin{aligned} \phi_{ntj}(\hat{\boldsymbol{\beta}}) - \phi_{ntj}(\boldsymbol{\beta}^*) &= \left(\triangledown \phi_{ntj}(\breve{\boldsymbol{\beta}})\right)^{\mathrm{T}} (\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*) \\ &= \dot{\phi}_{ntj}(\boldsymbol{\beta}) \boldsymbol{X}_{nt}^{\mathrm{T}} (\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*), \end{aligned}$$

where $\triangledown \phi_{ntj}(\boldsymbol{\beta})$ is defined in Equation (F.1) and $\breve{\boldsymbol{\beta}}$ locates on the line segment between $\hat{\boldsymbol{\beta}}$ and $\boldsymbol{\beta}^*$. By Lemma 7.2.1 and Regularity Condition 7.A, we have, for each $n \in \{1, \ldots, N\}$ and $t \in \{1, \ldots, T_n\}$ and $j \in \{0, \ldots, J_N\}$,

$$\begin{aligned} \left|\phi_{ntj}(\hat{\boldsymbol{\beta}}) - \phi_{ntj}(\boldsymbol{\beta}^*)\right| &\leq \left|\dot{\phi}_{ntj}(\boldsymbol{\beta})\right| \|\boldsymbol{X}_{nt}\|_2 \left\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\right\|_2 \\ &= O_p(J_N^{1/2} N^{-1/2}). \end{aligned}$$

This completes the proof. □

Proof of Lemma F.11

Lemma F.11.

Assume that Regularity Condition 7.A-7.I are satisfied. Let $(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\alpha}}_N)$ be an estimator *which converges to* $(\boldsymbol{\beta}^*, \boldsymbol{\alpha}_N^*)$ *in the sense that*

$$\|\hat{\boldsymbol{\alpha}}_N - \boldsymbol{\alpha}_N^*\|_2^2 + \left\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\right\|_2^2 = O_p(J_N N^{-1}),$$

and $J_N^3 N^{-2} = o(1)$*. Then, as the sample size* N *goes to infinity, for each* $n \in \{1, \ldots, N\}$ *and* $i \in \{1, \ldots, p\}$*,*

$$\begin{aligned}&\boldsymbol{X}_{n\cdot i}\boldsymbol{D}_n(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\alpha}}_N)\tilde{\boldsymbol{W}}_n\boldsymbol{G}_n(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\alpha}}_N)\begin{bmatrix}\hat{\boldsymbol{\alpha}}_N - \boldsymbol{\alpha}_N^* \\ \hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\end{bmatrix} \\ &= \boldsymbol{X}_{n\cdot i}\boldsymbol{D}_n(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\alpha}}_N)\tilde{\boldsymbol{W}}_n\boldsymbol{G}_n(\boldsymbol{\beta}^*, \boldsymbol{\alpha}_N^*)\begin{bmatrix}\hat{\boldsymbol{\alpha}}_N - \boldsymbol{\alpha}_N^* \\ \hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\end{bmatrix} + O_p(J_N^{3/2}N^{-1})\end{aligned}$$

Proof. For each n, we have

$$\begin{aligned}&\boldsymbol{X}_{n\cdot i}\boldsymbol{D}_n(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\alpha}}_N)\tilde{\boldsymbol{W}}_n\boldsymbol{G}_n(\boldsymbol{\beta}^*, \boldsymbol{\alpha}_N^*)\begin{bmatrix}\hat{\boldsymbol{\alpha}}_N - \boldsymbol{\alpha}_N^* \\ \hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\end{bmatrix} \\ &= I_{ni1} + I_{ni2} + I_{ni3} + I_{ni4} + I_{ni5},\end{aligned}$$

where

$$\begin{aligned}I_{ni1} &= \boldsymbol{X}_{n\cdot i}\boldsymbol{D}_n(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\alpha}}_N)\tilde{\boldsymbol{W}}_n\boldsymbol{G}_n(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\alpha}}_N)\begin{bmatrix}\hat{\boldsymbol{\alpha}}_N - \boldsymbol{\alpha}_N^* \\ \hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\end{bmatrix}, \\ I_{ni2} &= \boldsymbol{X}_{n\cdot i}\boldsymbol{D}_n^*\boldsymbol{W}_n^*\left(\boldsymbol{G}_n(\boldsymbol{\beta}^*, \boldsymbol{\alpha}_N^*) - \boldsymbol{G}_n(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\alpha}}_N)\right) \times \begin{bmatrix}\hat{\boldsymbol{\alpha}}_N - \boldsymbol{\alpha}_N^* \\ \hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\end{bmatrix}, \\ I_{ni3} &= \boldsymbol{X}_{n\cdot i}\left(\boldsymbol{D}_n(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\alpha}}_N) - \boldsymbol{D}_n^*\right)\boldsymbol{W}_n^*\left(\boldsymbol{G}_n(\boldsymbol{\beta}^*, \boldsymbol{\alpha}_N^*) - \boldsymbol{G}_n(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\alpha}}_N)\right) \\ &\quad \times \begin{bmatrix}\hat{\boldsymbol{\alpha}}_N - \boldsymbol{\alpha}_N^* \\ \hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\end{bmatrix}, \\ I_{ni4} &= \boldsymbol{X}_{n\cdot i}\boldsymbol{D}_n^*\left(\tilde{\boldsymbol{W}}_n - \boldsymbol{W}_n^*\right)\left(\boldsymbol{G}_n(\boldsymbol{\beta}^*, \boldsymbol{\alpha}_N^*) - \boldsymbol{G}_n(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\alpha}}_N)\right) \\ &\quad \times \begin{bmatrix}\hat{\boldsymbol{\alpha}}_N - \boldsymbol{\alpha}_N^* \\ \hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\end{bmatrix}\end{aligned}$$

and

$$\begin{aligned}I_{ni5} = \boldsymbol{X}_{n\cdot i}&\left(\boldsymbol{D}_n(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\alpha}}_N) - \boldsymbol{D}_n^*\right)\left(\tilde{\boldsymbol{W}}_n - \boldsymbol{W}_n^*\right) \\ &\times \left(\boldsymbol{G}_n(\boldsymbol{\beta}^*, \boldsymbol{\alpha}_N^*) - \boldsymbol{G}_n(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\alpha}}_N)\right)\begin{bmatrix}\hat{\boldsymbol{\alpha}}_N - \boldsymbol{\alpha}_N^* \\ \hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\end{bmatrix}.\end{aligned}$$

By Lemma F.10, we have that, for each $n \in \{1, \ldots, N\}$,

$$\lambda_{\max}\left(\left(\boldsymbol{G}_n(\boldsymbol{\beta}^*, \boldsymbol{\alpha}_N^*) - \boldsymbol{G}_n(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\alpha}}_N)\right)\left(\boldsymbol{G}_n(\boldsymbol{\beta}^*, \boldsymbol{\alpha}_N^*) - \boldsymbol{G}_n(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\alpha}}_N)\right)^{\mathrm{T}}\right) = O_p(J_N^2 N^{-1}).$$

By Lemma 7.2.2 and Regularity Condition 7.A, and the Cauchy-Schwarz inequality, we have

$$\begin{aligned} I_{ni2}^2 \leq & \lambda_{\max}\left(\left(\boldsymbol{G}_n(\boldsymbol{\beta}^*, \boldsymbol{\alpha}_N^*) - \boldsymbol{G}_n(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\alpha}}_N)\right)\left(\boldsymbol{G}_n(\boldsymbol{\beta}^*, \boldsymbol{\alpha}_N^*) - \boldsymbol{G}_n(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\alpha}}_N)\right)^{\mathrm{T}}\right) \\ & \times \|\boldsymbol{X}_{n\cdot i}\boldsymbol{D}_n^*\boldsymbol{W}_n^*\|_2^2 \begin{bmatrix} \hat{\boldsymbol{\alpha}}_N - \boldsymbol{\alpha}_N^* \\ \hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^* \end{bmatrix}^{\mathrm{T}} \begin{bmatrix} \hat{\boldsymbol{\alpha}}_N - \boldsymbol{\alpha}_N^* \\ \hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^* \end{bmatrix} \\ = & O_p(J_N^3 N^{-2}). \end{aligned}$$

Similarly, we can show that

$$\begin{aligned} I_{ni3}^2 &= O_p(J_N^4 N^{-3}), \\ I_{ni4}^2 &= O_p(J_N^4 N^{-3}) \end{aligned}$$

and

$$I_{ni5}^2 = O_p(J_N^5 N^{-4})$$

by Lemma F.6 and F.10, Theorem 7.4.1 and Regularity Condition 7.A. This completes the proof. □

Proof of Lemma F.12

Lemma F.12.
Assume that Regularity Condition 7.A-7.C and 7.I are satisfied. Then, as the sample size N goes to infinity,

$$N^{-1/2}\sum_{n=1}^{N} \boldsymbol{X}_n \boldsymbol{D}_n^* \boldsymbol{W}_n^* \boldsymbol{U}_n(\boldsymbol{\beta}^*, Q^*) \tag{F.12}$$

converges in distribution to a multivariate normal random variable with mean zero and covariance matrix $\boldsymbol{\Gamma}$.

Proof. Let

$$\boldsymbol{B}_n = N^{-1/2}\boldsymbol{X}_n\boldsymbol{D}_n^*\boldsymbol{W}_n^*\boldsymbol{U}_n(\boldsymbol{\beta}^*, Q^*) \in \mathbb{R}^p$$

and, for each $i \in \{1, \dots, p\}$,

$$B_{ni} = N^{-1/2}\boldsymbol{X}_{n\cdot i}\boldsymbol{D}_n^*\boldsymbol{W}_n^*\boldsymbol{U}_n(\boldsymbol{\beta}^*, Q^*)$$

is the i^{th} element of $\boldsymbol{B}_n$, where $\boldsymbol{X}_{n\cdot i}$ is the i^{th} row of $\boldsymbol{X}_n$. We use the Lindeberg-Feller Central Limit Theorem [Bauer, 1996, p.g. 234] to prove this lemma by checking the Lindeberg condition.

Firstly, we show that for each $i \in \{1, \dots, p\}$, the variance of B_{ni}, denoted by σ_{ni}^2, is bounded. By the Cauchy-Schwaz inequality, we have

$$\begin{aligned}
& N\text{Var}[B_{ni}] \\
&= \mathbb{E}\left[(\boldsymbol{X}_{n\cdot i}\boldsymbol{D}_n^*\boldsymbol{W}_n^*\boldsymbol{U}_n(\boldsymbol{\beta}^*, Q^*))^2\right] \\
&= \mathbb{E}\left[\left(\boldsymbol{X}_{n\cdot i}\boldsymbol{D}_n^*\boldsymbol{W}_n^*\boldsymbol{D}_n^*\boldsymbol{X}_{n\cdot i}^{\text{T}}\right) \times \boldsymbol{U}_n^{\text{T}}(\boldsymbol{\beta}^*, Q^*)\boldsymbol{U}_n(\boldsymbol{\beta}^*, Q^*)\right].
\end{aligned}$$

Because $\boldsymbol{W}_n^*$ is positive definite matrix and the elements in $\boldsymbol{X}_{n\cdot i}$, $\boldsymbol{D}_n^*$ are all bounded by Regularity Condition 7.A and Lemma 7.2.1, we have

$$N\text{Var}[B_{ni}] \le C_i\mathbb{E}\left[\boldsymbol{U}_n^{\text{T}}(\boldsymbol{\beta}^*, Q^*)\boldsymbol{U}_n(\boldsymbol{\beta}^*, Q^*)\right],$$

where

$$C_i = \sup_{n\in\{1,\dots,N\}} \boldsymbol{X}_{n\cdot i}\boldsymbol{D}_n^*\boldsymbol{W}_n^*\boldsymbol{D}_n^*\boldsymbol{X}_{n\cdot i}^{\text{T}}.$$

Moreover, because $\mathbb{E}\left[\boldsymbol{U}_n^{\text{T}}(\boldsymbol{\beta}^*, Q^*)\boldsymbol{U}_n(\boldsymbol{\beta}^*, Q^*)\right]$ is bounded for each n, by Regularity Condition 7.I, we have $\sigma_{ni}^2 = O(N^{-1})$.

For any $\epsilon > 0$, by the Cauchy-Schwaz inequality, we have

$$\sum_{n=1}^N \mathbb{E}\left[\|\boldsymbol{B}_n\|_2^2 I(\|\boldsymbol{B}_n\| > \epsilon)\right] \le \sum_{n=1}^N \left(\mathbb{E}[\|\boldsymbol{B}_n\|_2^4] \times \text{pr}(\|\boldsymbol{B}_n\|_2 > \epsilon)\right)^{1/2}.$$

Using the Chebyshev's inequality, we have

$$\text{pr}\left(\|\boldsymbol{B}_n\| > \epsilon\right) \le N^{-1}\epsilon^{-2}\sum_{i=1}^p \sigma_{ni}^2 = O(N^{-1}).$$

On the other hand, by the Cauchy-Schwarz inequality and Regularity Condition 7.A and 7.I and Lemma 7.2.1, we have

$$\begin{aligned}
&\mathbb{E}[\|\boldsymbol{B}_n\|_2^4] \\
&= \mathbb{E}\left[\left(\sum_{i=1}^{p} B_{ni}^2\right)^2\right] \\
&\le p\sum_{i=1}^{p}\mathbb{E}\left[B_{ni}^4\right] \\
&\le pN^{-2}\sum_{i=1}^{p}\mathbb{E}\left[\left(\boldsymbol{X}_{n\cdot i}\boldsymbol{D}_n^*\boldsymbol{W}_n^*\boldsymbol{D}_n^*\boldsymbol{X}_{n\cdot i}^{\mathrm{T}}\right)^2\left(\boldsymbol{U}_n^{\mathrm{T}}(\boldsymbol{\beta}^*, Q^*)\boldsymbol{U}_n(\boldsymbol{\beta}^*, Q^*)\right)^2\right] \\
&\le p\sum_{i=1}^{p} C_i^2\mathbb{E}\left[\left(\boldsymbol{U}_n^{\mathrm{T}}(\boldsymbol{\beta}^*, Q^*)\boldsymbol{U}_n(\boldsymbol{\beta}^*, Q^*)\right)^2\right] \\
&= O(N^{-2}).
\end{aligned}$$

Therefore, $\sum_{n=1}^{N}\mathbb{E}\left[\|\boldsymbol{B}_n\|_2^2 I(\|\boldsymbol{B}_n\| > \epsilon)\right]$ is $O(N^{-1/2})$.

For each n, the covariance matrix of $\boldsymbol{B}_n$ is

$$\begin{aligned}
\mathrm{Cov}\left[\boldsymbol{B}_n\right] &= \mathbb{E}\left[\boldsymbol{B}_n\boldsymbol{B}_n^{\mathrm{T}}\right] \\
&= N^{-1}\mathbb{E}\left[\boldsymbol{X}_n\boldsymbol{D}_n^*\boldsymbol{W}_n^*\boldsymbol{\Sigma}_n(\boldsymbol{\beta}^*, \boldsymbol{Q}^*)\boldsymbol{W}_n^*\boldsymbol{D}_n^*\boldsymbol{X}_n^{\mathrm{T}}\right],
\end{aligned}$$

where $\boldsymbol{\Sigma}_n(\boldsymbol{\beta}^*, Q^*)$ is the covariance matrix of $\boldsymbol{Y}_n \mid (\boldsymbol{X}_n, \boldsymbol{Z}_n)$. Under the assumption that the elements of $\sum_{n=1}^{N}\mathrm{Cov}\left[\boldsymbol{B}_n\right]$ converges to a covariance matrix $\boldsymbol{\Gamma}$ as N goes to infinity, we complete the proof. □

Chapter 8

Ensemble Inference with the Generalized Method of Moments Estimators

8.1 Introduction

Given a data set $(\boldsymbol{Y}_n, \boldsymbol{X}_n, \boldsymbol{Z}_n)$, $n = 1, \ldots, N$, from the data setup in Section 5.2, we are interested in the inference about the regression parameter $\boldsymbol{\beta} \in \mathbb{R}^p$ under the framework of the generalized method of moments for mixed-effects models with univariate random effects. In other words, the regression parameter $\boldsymbol{\beta} \in \mathbb{R}^p$ is the parameter of interest, while the generalized moments $\boldsymbol{\alpha}_N \in \mathbb{R}^{J_N+1}$ are nuisance parameters. In this chapter, we use the generalized method of moments to construct a χ^2 test statistic for the following hypothesis testing problem

$$\mathrm{H}_0 : \boldsymbol{\beta} = \boldsymbol{\beta}_0, \tag{8.1}$$

where $\boldsymbol{\beta}_0 \in \mathbb{R}^p$ is a real vector.

From Theorem 7.5.1 and its proof, we know that, given an estimator $(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\alpha}}_N)$

satisfying Equation (7.27), the asymptotic distribution of

$$\boldsymbol{\eta} = N^{1/2} \left[\hat{\boldsymbol{A}}, \hat{\boldsymbol{B}}\right] \begin{bmatrix} \hat{\boldsymbol{\alpha}}_N - \boldsymbol{\alpha}_N^* \\ \hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^* \end{bmatrix} \in \mathbb{R}^p \tag{8.2}$$

is multivariate normal under the regularity conditions in Section 7.2, where

$$\hat{\boldsymbol{A}} = \frac{1}{N} \sum_{n=1}^{N} \boldsymbol{X}_n \boldsymbol{D}_n(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\alpha}}_N) \tilde{\boldsymbol{W}}_n \boldsymbol{\Phi}_n^{\mathrm{T}}(\hat{\boldsymbol{\beta}})$$

and

$$\hat{\boldsymbol{B}} = \frac{1}{N} \sum_{n=1}^{N} \boldsymbol{X}_n \boldsymbol{D}_n(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\alpha}}_N) \tilde{\boldsymbol{W}}_n \boldsymbol{D}_n(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\alpha}}_N) \boldsymbol{X}_n^{\mathrm{T}}.$$

Here for each $n \in \{1, \ldots, N\}$, $\boldsymbol{\Phi}_n(\boldsymbol{\beta})$ is $(J_N + 1) \times T_n$ matrix whose elements are $\phi_{ntj}(\boldsymbol{\beta})$ defined in Equation (5.3); $\boldsymbol{D}_n(\boldsymbol{\beta}, \boldsymbol{\alpha}_N)$ is the diagonal matrix whose diagonal elements are $\dot{\boldsymbol{\phi}}_{nt}^{\mathrm{T}}(\boldsymbol{\beta})\boldsymbol{\alpha}_N$, $t = 1, \ldots, T_n$, where for each t, $\dot{\boldsymbol{\phi}}_{nt}(\boldsymbol{\beta})$ is defined in Equation (7.4); the weighting matrix

$$\tilde{\boldsymbol{W}}_n = \tilde{\boldsymbol{V}}_n^{-1/2} \boldsymbol{R}_n^{-1} \tilde{\boldsymbol{V}}_n^{-1/2},$$

where $\tilde{\boldsymbol{V}}_n = \boldsymbol{V}_n(\tilde{\boldsymbol{\beta}}, \tilde{\boldsymbol{\alpha}}_N)$, $\boldsymbol{V}_n(\boldsymbol{\beta}, \boldsymbol{\alpha}_N)$ is a $T_n \times T_n$ diagonal matrix whose diagonal elements are $V_{nt}(\boldsymbol{\beta}, \boldsymbol{\alpha}_N)$ defined in Equation (5.10), and $\boldsymbol{R}_n$ is the working correlation matrix.

However, as we pointed out in Section 7.7, $\boldsymbol{\eta} \in \mathbb{R}^p$ may not be directly used for testing the hypothesis H_0 in (8.1), because $\boldsymbol{\alpha}_N^* \in \mathbb{R}^{J_N}$ is unknown. However, if we can find a unit vector $\boldsymbol{e} \in \mathbb{R}^p$ such that

$$\boldsymbol{e}^{\mathrm{T}} \hat{\boldsymbol{A}} = \boldsymbol{0} \in \mathbb{R},$$

and

$$\boldsymbol{e}^{\mathrm{T}} \hat{\boldsymbol{B}} \neq \boldsymbol{0} \in \mathbb{R},$$

then we can construct an asymptotically normal test statistic for H_0,

$$\boldsymbol{e}^{\mathrm{T}} \boldsymbol{\eta} = N^{1/2} \boldsymbol{e}^{\mathrm{T}} \hat{\boldsymbol{B}} \left(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\right) \in \mathbb{R}.$$

However, such unit vector $\boldsymbol{e} \in \mathbb{R}^p$ does not exist, when $\hat{\boldsymbol{A}}$ is a full rank $p \times (J_N + 1)$ matrix and $\hat{\boldsymbol{B}}$ is full rank $p \times p$ matrix. The reason is that the space spanned by the columns of $\hat{\boldsymbol{B}}$ is a subspace spanned by the columns of $\hat{\boldsymbol{A}}$.

We propose using the ensemble inference [Zhu, 2008] for testing the hypothesis H_0 under the framework of the generalized method of moments. We prove that the proposed test statistic asymptotically follows a χ^2 distribution. In the literature of mixture models, it is uncommon to see asymptotically χ^2 test statistics. Most of them have a mixture of χ^2 distributions as their asymptotic distributions; see [Lindsay, 1995] and [Li and Chen, 2010].

We organize this chapter as follows. In Section 8.2, we describe the procedure for using the ensemble inference. In Section 8.3, we establish the appropriate asymptotic theory. In Section 8.4, we conduct simulation studies to investigate the performance of the ensemble inference. In Section 8.5, we use the ensemble inference to analyze the Epileptic Seizures Data. Lastly, we end this chapter with a discussion.

8.2 Ensemble Inference

The ensemble idea that making inference with a collection of models rather than a single model is not new to the literature of statistics. The famous ensemble methods include the AdaBoost [Freund and Schapire, 1997] and the random forests [Breiman, 2001]; see [Zhu, 2008] for an insightful discussion. Our ensemble idea is to construct a χ^2 test statistic for H_0 with an initial estimator $(\tilde{\boldsymbol{\beta}}, \tilde{\boldsymbol{\alpha}}_N)$ and a collection of the generalized estimating equations with different working correlation matrices. Here the initial estimator is obtained from the GMM. This is a new application of the ensemble idea, by our knowledge, while the ensemble methods are often used to reduce the prediction errors in the existing statistical literature; see [Zhu, 2008].

Given a data set $(\boldsymbol{Y}_n, \boldsymbol{X}_n, \boldsymbol{Z}_n)$, $n = 1, \ldots, N$, from the data setup in Section 5.2, the ensemble inference for the null hypothesis testing H_0 in Equation (8.1) include the following steps:

1. Compute the initial estimator $(\tilde{\boldsymbol{\beta}}, \tilde{\boldsymbol{\alpha}}_N)$ by the GMM, where for each n, the

weighting matrix is the identity matrix.

2. Let K_N be the smallest integer such that $pK_N > J_N + 1$. Choose K_N random matrix processes $\mathfrak{R}_k$, $k = 1, \ldots, K_N$. For each $k \in \{1, \ldots, K_N\}$, generate a set of random correlation matrices $\{\boldsymbol{R}_n^{(k)}\}_{n=1}^N$ from the random matrix process $\mathfrak{R}_k$, where $\boldsymbol{R}_n^{(k)}$, $n = 1, \ldots, N$, are independent to each other. Moreover, for each $n \in \{1, \ldots, N\}$, $\boldsymbol{R}_n^{(k)}$ depends on the data $(\boldsymbol{Y}_n, \boldsymbol{X}_n, \boldsymbol{Z}_n)$ in the sense that $\boldsymbol{R}_n^{(k)}$ is a $T_n \times T_n$ matrix, where T_n is the dimension of $\boldsymbol{Y}_n$.

3. For each $k \in \{1, \ldots, K_N\}$, solve the estimating equations

$$\frac{1}{N}\sum_{n=1}^{N} \boldsymbol{X}_n \tilde{\boldsymbol{D}}_n \tilde{\boldsymbol{W}}_n^{(k)} \boldsymbol{U}_n(\boldsymbol{\beta}, \tilde{\boldsymbol{\alpha}}_N) = 0, \tag{8.3}$$

where for each $n \in \{1, \ldots, N\}$, $\tilde{\boldsymbol{D}}_n = \boldsymbol{D}_n(\tilde{\boldsymbol{\beta}}, \tilde{\boldsymbol{\alpha}}_N)$, $\boldsymbol{U}_n(\boldsymbol{\beta}, \boldsymbol{\alpha}_N)$ is defined in Equation (5.5), and

$$\tilde{\boldsymbol{W}}_n^{(k)} = \tilde{\boldsymbol{V}}_n^{-1/2} \left(\boldsymbol{R}_n^{(k)}\right)^{-1} \tilde{\boldsymbol{V}}_n^{-1/2}.$$

For each $k \in \{1, \ldots, K_N\}$, let $\hat{\boldsymbol{\beta}}^{(k)} \in \mathbb{R}^p$ be the solution of Equation (8.3).

4. Compute the matrix

$$\tilde{\boldsymbol{\Gamma}} = \begin{bmatrix} \tilde{\boldsymbol{\Gamma}}_{(1,1)} & \tilde{\boldsymbol{\Gamma}}_{(1,2)} & \ldots & \tilde{\boldsymbol{\Gamma}}_{(1,K_N)} \\ \tilde{\boldsymbol{\Gamma}}_{(2,1)} & \tilde{\boldsymbol{\Gamma}}_{(2,2)} & \ldots & \tilde{\boldsymbol{\Gamma}}_{(2,K_N)} \\ \vdots & \vdots & \ddots & \vdots \\ \tilde{\boldsymbol{\Gamma}}_{(K_N,1)} & \tilde{\boldsymbol{\Gamma}}_{(K_N,2)} & \ldots & \tilde{\boldsymbol{\Gamma}}_{(K_N,K_N)} \end{bmatrix}, \tag{8.4}$$

where for each $k, k' \in \{1, \ldots, K_N\}$,

$$\tilde{\boldsymbol{\Gamma}}_{(k,k')} = \frac{1}{N}\sum_{n=1}^{N} \boldsymbol{X}_n \tilde{\boldsymbol{D}}_n \tilde{\boldsymbol{W}}_n^{(k)} \tilde{\boldsymbol{\Sigma}}_n \tilde{\boldsymbol{W}}_n^{(k)} \tilde{\boldsymbol{D}}_n \boldsymbol{X}_n^{\mathrm{T}}. \tag{8.5}$$

Here for each $n \in \{1, \ldots, N\}$, $\tilde{\boldsymbol{\Sigma}}_n$ is an estimator of the correlation matrix of $\boldsymbol{U}_n(\boldsymbol{\beta}, \boldsymbol{\alpha}_N)$. In Section 5.6, we have given two possible ways to calculate $\tilde{\boldsymbol{\Sigma}}$.

5. Let $\tilde{\boldsymbol{A}}$ be a $pK_N \times (J_N + 1)$ matrix such that

$$\tilde{\boldsymbol{A}} = \left[\left(\tilde{\boldsymbol{A}}_{(1)}\right)^{\mathrm{T}}, \cdots, \left(\tilde{\boldsymbol{A}}_{(K_N)}\right)^{\mathrm{T}}\right]^{\mathrm{T}}, \tag{8.6}$$

where for each $k \in \{1, \ldots, K_N\}$,

$$\tilde{\boldsymbol{A}}_{(k)} = \frac{1}{N} \sum_{n=1}^{N} \boldsymbol{X}_n \tilde{\boldsymbol{D}}_n \tilde{\boldsymbol{W}}_n^{(k)} \boldsymbol{\Phi}_n^{\mathrm{T}}(\tilde{\boldsymbol{\beta}}). \tag{8.7}$$

Compute the projection matrix

$$\tilde{\boldsymbol{P}} = \boldsymbol{I}_{pK_N} - \tilde{\boldsymbol{\Gamma}}^{-1/2} \tilde{\boldsymbol{A}} \left(\tilde{\boldsymbol{A}}^{\mathrm{T}} \boldsymbol{\Gamma}^{-1} \tilde{\boldsymbol{A}} \right)^{-1} \tilde{\boldsymbol{A}}^{\mathrm{T}} \tilde{\boldsymbol{\Gamma}}^{-1/2}, \tag{8.8}$$

where $\boldsymbol{I}_{pK_N}$ is a $pK_N \times pK_N$ identity matrix.

6. Let $\tilde{\boldsymbol{B}}$ be a $pK_N \times pK_N$ matrix such that

$$\tilde{\boldsymbol{B}} = \begin{bmatrix} \tilde{\boldsymbol{B}}_{(1)} & & \boldsymbol{O} \\ & \ddots & \\ \boldsymbol{O} & & \tilde{\boldsymbol{B}}_{(K_N)} \end{bmatrix}, \tag{8.9}$$

where $\boldsymbol{O}$ is the zero matrix and for each $k \in \{1, \ldots, K_N\}$,

$$\tilde{\boldsymbol{B}}_{(k)} = \frac{1}{N} \sum_{n=1}^{N} \boldsymbol{X}_n \tilde{\boldsymbol{D}}_n \tilde{\boldsymbol{W}}_n^{(k)} \tilde{\boldsymbol{D}}_n \boldsymbol{X}_n^{\mathrm{T}}.$$

Compute the test statistic for H_0 such that

$$\zeta = N \boldsymbol{v}_{\boldsymbol{\beta}}^{\mathrm{T}} \tilde{\boldsymbol{B}}^{\mathrm{T}} \tilde{\boldsymbol{\Gamma}}^{-1/2} \tilde{\boldsymbol{P}} \tilde{\boldsymbol{\Gamma}}^{-1/2} \tilde{\boldsymbol{B}} \boldsymbol{v}_{\boldsymbol{\beta}}. \tag{8.10}$$

where

$$\boldsymbol{v}_{\boldsymbol{\beta}} = \left(\left(\hat{\boldsymbol{\beta}}^{(1)} - \boldsymbol{\beta}^* \right)^{\mathrm{T}}, \ldots, \left(\hat{\boldsymbol{\beta}}^{(K_N)} - \boldsymbol{\beta}^* \right)^{\mathrm{T}} \right)^{\mathrm{T}} \in \mathbb{R}^{pK_N}. \tag{8.11}$$

7. Reject the null hypothesis H_0, when ζ exceeds some critical value to be determined by its limiting distribution.

Definition 8.2.1 (Ensemble Statistics)**.**
The test statistic ζ in Equation (8.10) is called the ensemble test statistics.

There are two basic rules for the models in a good ensemble method [Zhu, 2008]:

1. The estimation or predication using each model is accurate or appropriate;

2. There are small correlation between individual models within the ensemble.

Our ensemble inference follows the exactly the same rule. Using the estimating equations, for each $k \in \{1, \dots, K_N\}$, we have $\hat{\boldsymbol{\beta}}^{(k)}$ converges to $\boldsymbol{\beta}^*$ in probability; see Theorem 8.3.1. This ensures the accuracy of the model, at least asymptotically.

Meanwhile, we generate the random working correlation matrices to reduce the correlation between $\hat{\boldsymbol{\beta}}^{(k)}$. To understand the role of random working correlation matrices, we consider the following projection problem, which is equivalent to the estimating equations (8.3),

$$\min_{\boldsymbol{\beta} \in \mathbb{R}^p} \quad \frac{1}{N} \sum_{n=1}^{N} \boldsymbol{U}_n^{\mathrm{T}}(\boldsymbol{\beta}, \tilde{\boldsymbol{\alpha}}_N) \tilde{\boldsymbol{W}}_n^{(k)} \boldsymbol{U}_n(\boldsymbol{\beta}, \tilde{\boldsymbol{\alpha}}_N).$$

The optimization problem can be interpreted by projection the response vector $\boldsymbol{Y}_n$ into the model space $\boldsymbol{\Phi}_n^{\mathrm{T}}(\boldsymbol{\beta}) \tilde{\boldsymbol{\alpha}}_N$, because

$$\boldsymbol{U}_n = \boldsymbol{Y}_n - \boldsymbol{\Phi}_n^{\mathrm{T}}(\boldsymbol{\beta}) \boldsymbol{\alpha}_N.$$

Geometrically, different working correlation matrices determine different projecting directions and different efficiencies of the corresponding estimates $\hat{\boldsymbol{\beta}}^{(k)}$ follows. Such a variety in efficiency allows us to construct an asymptotically normal statistic in a higher dimensional space ($\mathbb{R}^{pK_N}$). Then, we can obtain an asymptotically normal statistic in a lower dimensional subspace which is orthogonal to $\tilde{\boldsymbol{\alpha}}_N$.

We comment on each step of the ensemble inference. Firstly, the properties of the GMM estimator $(\tilde{\boldsymbol{\beta}}, \tilde{\boldsymbol{\alpha}}_N)$ in Step 1 have been studied in Chapter 5 and 7. It is known that $(\tilde{\boldsymbol{\beta}}, \tilde{\boldsymbol{\alpha}}_N)$ is consistent under the regularity conditions listed in Section 7.2. Moreover, asymptotic normality only exists in $\mathbb{R}^p$; see Theorem 7.5.1.

The following algorithm is used to generate the random correlation matrices in Step 2. In it, the Wishart distribution is used, because it is defined over symmetric, non-negative definite random matrices; see [Gelman et al., 2014, p.g. 582]. In Bayesian statistics, it is a commonly used prior for the covariance matrices; see [Gelman et al., 2014, p.g. 582]. The generating process for the ensemble inference should satisfy Regularity Condition 8.A-8.D. However, sufficient conditions for the generating processes require further investigation.

Algorithm 8.1.
Let $T_{\max} = \max_{n \in \{1,\ldots,N\}} T_n$ and $\boldsymbol{I}_t$ be the $t \times t$ identity matrix. For $k = 1, \ldots, K_N$, repeat the following steps:

1. *For each $t \in \{1, \ldots, T_{\max}\}$, generate a $t \times t$ random matrix $\boldsymbol{S}_t^{(k)}$ from the Wishart distribution with degrees of freedom $2t$ and scale matrix $\boldsymbol{I}_t$. For each $t = 1, \ldots, T_{\max}$, let $\boldsymbol{\rho}_t^{(k)}$ be the correlation matrix of $\boldsymbol{S}_t^{(k)}$.*

2. *For $n = 1, \ldots, N$, let*
$$\boldsymbol{R}_n^{(k)} = \boldsymbol{\rho}_{T_n}^{(k)},$$
where T_n is the dimension of $\boldsymbol{Y}_n$.

To solve Equation (8.3) in Step 3, we can use the Newton-Raphson method. The consistency of $\hat{\boldsymbol{\beta}}^{(k)}$, $k = 1, \ldots, K_N$, will be described in Theorem 8.3.1; see Section 8.3. It will be should later in Theorem 8.3.3 that

$$N^{1/2}\tilde{\boldsymbol{\Gamma}}^{-1/2}\left(\tilde{\boldsymbol{A}}\boldsymbol{v}_{\boldsymbol{\alpha}} + \tilde{\boldsymbol{B}}\boldsymbol{v}_{\boldsymbol{\beta}}\right) \in \mathbb{R}^{pK_N}$$

converges to a standard multivariate normal distribution in $\mathbb{R}^{pK_N}$, where

$$\boldsymbol{v}_{\boldsymbol{\alpha}} = \tilde{\boldsymbol{\alpha}}_N - \boldsymbol{\alpha}_N^* \in \mathbb{R}^{J_N+1}. \tag{8.12}$$

In Step 5 and 6, we aim to find the vector space which is complement the space spanned by the columns of $\tilde{\boldsymbol{A}}$. Because $pK_N > J_N + 1$, the complement space is determined by the projection matrix $\tilde{\boldsymbol{P}}$, whose rank is $pK_N - J_N - 1$. Let $\tilde{\boldsymbol{V}}$ be a $pK_N \times (pK_N - J_N - 1)$ matrix whose columns are the eigenvectors of $\tilde{\boldsymbol{P}}$. This leads to that

$$N^{1/2}\tilde{\boldsymbol{V}}^{\mathrm{T}}\tilde{\boldsymbol{\Gamma}}^{-1/2}\left(\tilde{\boldsymbol{A}}\boldsymbol{v}_{\boldsymbol{\alpha}} + \tilde{\boldsymbol{B}}\boldsymbol{v}_{\boldsymbol{\beta}}\right) = N^{1/2}\tilde{\boldsymbol{V}}^{\mathrm{T}}\tilde{\boldsymbol{\Gamma}}^{-1/2}\tilde{\boldsymbol{B}}\boldsymbol{v}_{\boldsymbol{\beta}}$$

asymptotically follows a multivariate normal distribution in $\mathbb{R}^{pK_N-J_N-1}$. And thus,

$$\begin{aligned}\zeta &= N\boldsymbol{v}_{\boldsymbol{\beta}}^{\mathrm{T}}\tilde{\boldsymbol{B}}^{\mathrm{T}}\tilde{\boldsymbol{\Gamma}}^{-1/2}\tilde{\boldsymbol{P}}\tilde{\boldsymbol{\Gamma}}^{-1/2}\tilde{\boldsymbol{B}}\boldsymbol{v}_{\boldsymbol{\beta}} \\ &= N\boldsymbol{v}_{\boldsymbol{\beta}}^{\mathrm{T}}\tilde{\boldsymbol{B}}^{\mathrm{T}}\tilde{\boldsymbol{\Gamma}}^{-1/2}\tilde{\boldsymbol{V}}\tilde{\boldsymbol{V}}^{\mathrm{T}}\tilde{\boldsymbol{\Gamma}}^{-1/2}\tilde{\boldsymbol{B}}\boldsymbol{v}_{\boldsymbol{\beta}}\end{aligned}$$

asymptotically follows a χ^2 distribution with degrees of freedom $pK_N - J_N - 1$. We will show that the asymptotic distribution of ζ does not depend on the specific choices of the initial estimators $(\tilde{\boldsymbol{\beta}}, \tilde{\boldsymbol{\alpha}}_N)$ and the set of random correlation matrices $\{\boldsymbol{R}_n^{(k)}\}_{n=1}^N$, $k = 1, \ldots, K_N$.

8.3 Asymptotic Theory

8.3.1 Existence and Consistency of $\hat{\boldsymbol{\beta}}^{(k)}$

Firstly, we show the existence of the roots $\hat{\boldsymbol{\beta}}^{(k)}$, $k = 1, \ldots, K_N$, and compute their convergence rates. As same as the GEE method [Liang and Zeger, 1986], with different working correlation matrices, the consistency of $\hat{\boldsymbol{\beta}}^{(k)}$ are consistent for each $k \in \{1, \ldots, K_N\}$. The regularity conditions are listed in Section 7.2. The proof can be found in Appendix G.2.

Theorem 8.3.1 (Existence and Consistency of $\hat{\boldsymbol{\beta}}^{(k)}$)**.**
Assume that Regularity Condition 7.A-7.J are satisfied and $J_N N^{-1} = o(1)$. Further assume that the initial estimator $(\tilde{\boldsymbol{\beta}}, \tilde{\boldsymbol{\alpha}}_N)$ converges to $(\boldsymbol{\beta}^*, \boldsymbol{\alpha}_N^*) \in \mathbb{R}^p \times \mathcal{M}$ in the sense that

$$\|\tilde{\boldsymbol{\alpha}}_N - \boldsymbol{\alpha}_N^*\|_2^2 + \left\|\tilde{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\right\|_2^2 = O_p(J_N N^{-1}),$$

as the sample size N goes to infinity. Then, for each $k \in \{1, \ldots, K_N\}$, Equation (8.3) has a root $\hat{\boldsymbol{\beta}}^{(k)}$ such that

$$\left\|\hat{\boldsymbol{\beta}}^{(k)} - \boldsymbol{\beta}^*\right\|_2^2 = O_p(J_N N^{-1}),$$

as the sample size N goes to infinity.

8.3.2 Asymptotic Distribution

Let

$$\boldsymbol{\Gamma}^* = \begin{bmatrix} \boldsymbol{\Gamma}^*_{(1,1)} & \boldsymbol{\Gamma}^*_{(1,2)} & \cdots & \boldsymbol{\Gamma}^*_{(1,K_N)} \\ \boldsymbol{\Gamma}^*_{(2,1)} & \boldsymbol{\Gamma}^*_{(2,2)} & \cdots & \boldsymbol{\Gamma}^*_{(2,K_N)} \\ \vdots & \vdots & \ddots & \vdots \\ \boldsymbol{\Gamma}^*_{(K_N,1)} & \boldsymbol{\Gamma}^*_{(K_N,2)} & \ldots, & \boldsymbol{\Gamma}^*_{(K_N,K_N)} \end{bmatrix}, \tag{8.13}$$

where for each $k, k' \in \{1, \ldots, K_N\}$,

$$\boldsymbol{\Gamma}^*_{(k,k')} = \lim_{N\to\infty} \frac{1}{N} \sum_{n=1}^{N} \boldsymbol{X}_n \boldsymbol{D}^*_n \boldsymbol{W}^{(k)}_n \boldsymbol{\Sigma}^*_n \boldsymbol{W}^{(k')}_n \boldsymbol{D}^*_n \boldsymbol{X}^{\mathrm{T}}_n, \tag{8.14}$$

and, for each n, $\boldsymbol{\Sigma}^*_n$ is the covariance matrix of $\boldsymbol{Y}_n \mid (\boldsymbol{X}_n, \boldsymbol{Z}_n)$,

$$\boldsymbol{W}^{(k)}_n = \boldsymbol{V}^{-1/2}_n(\boldsymbol{\beta}^*, Q^*) \boldsymbol{R}^{(k)}_n \boldsymbol{V}^{-1/2}_n(\boldsymbol{\beta}^*, Q^*),$$

and $\boldsymbol{V}_n(\boldsymbol{\beta}, Q)$ is a $T_n \times T_n$ diagonal matrix whose t^{th} diagonal element is $V_{nt}(\boldsymbol{\beta}, Q)$ defined in Equation (7.23).

Also let $\boldsymbol{A}^*$ be a $pK_N \times (J_N + 1)$ matrix such that

$$\boldsymbol{A}^* = \left[\left(\boldsymbol{A}^*_{(1)}\right)^{\mathrm{T}}, \cdots, \left(\boldsymbol{A}^*_{(K_N)}\right)^{\mathrm{T}}\right]^{\mathrm{T}}, \tag{8.15}$$

and

$$\boldsymbol{B}^* = \begin{bmatrix} \boldsymbol{B}^*_{(1)} & & \boldsymbol{O} \\ & \ddots & \\ \boldsymbol{O} & & \boldsymbol{B}^*_{(K_N)} \end{bmatrix}, \tag{8.16}$$

where for each $k \in \{1, \ldots, K_N\}$,

$$\boldsymbol{A}^*_{(k)} = \lim_{N\to\infty} \frac{1}{N} \sum_{n=1}^{N} \boldsymbol{X}_n \boldsymbol{D}^*_n \boldsymbol{W}^{(k)}_n \boldsymbol{\Phi}^{\mathrm{T}}_n(\boldsymbol{\beta}^*), \tag{8.17}$$

and

$$\boldsymbol{B}^*_{(k)} = \lim_{N\to\infty} \frac{1}{N} \sum_{n=1}^{N} \boldsymbol{X}_n \boldsymbol{D}^*_n \boldsymbol{W}^{(k)}_n \boldsymbol{D}^*_n \boldsymbol{X}^{\mathrm{T}}_n. \tag{8.18}$$

Theorem 8.3.2.
Assume that Regularity Condition 7.A-7.I are satisfied. Further assume that, for each $k \in \{1, \ldots, K_N\}$, the initial estimator $(\tilde{\boldsymbol{\beta}}, \tilde{\boldsymbol{\alpha}}_N)$ and $(\hat{\boldsymbol{\beta}}^{(k)}, \tilde{\boldsymbol{\alpha}}_N)$ converges to $(\boldsymbol{\beta}^*, \boldsymbol{\alpha}_N^*)$ in the sense that

$$\|\tilde{\boldsymbol{\alpha}}_N - \boldsymbol{\alpha}_N^*\|_2^2 + \left\|\hat{\boldsymbol{\beta}}^{(k)} - \boldsymbol{\beta}^*\right\|_2^2 = O_p(J_N N^{-1})$$

and

$$\|\tilde{\boldsymbol{\alpha}}_N - \boldsymbol{\alpha}_N^*\|_2^2 + \left\|\tilde{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\right\|_2^2 = O_p(J_N N^{-1}),$$

as the sample size N goes to infinity. If $J_N N^{-1/2} = o(1)$ as the sample size N goes to infinity, then

$$N^{1/2}\left(\boldsymbol{A}^* \boldsymbol{v}_{\boldsymbol{\alpha}} + \boldsymbol{B}^* \boldsymbol{v}_{\boldsymbol{\beta}}\right) \in \mathbb{R}^{pK_N}$$

converges in distribution to a multivariate normal random vector in $\mathbb{R}^{pK_N}$ with mean zero and covariance matrix $\boldsymbol{\Gamma}^*$, where $\boldsymbol{v}_{\boldsymbol{\beta}}$ and $\boldsymbol{v}_{\boldsymbol{\alpha}}$ are defined in Equation (8.11) and (8.12) correspondingly.

Because $\boldsymbol{\Gamma}^*$, $\boldsymbol{A}^*$ are $\boldsymbol{B}^*$ are unknown, in the rest of this subsection, we aim to show that

$$N^{1/2}\left(\tilde{\boldsymbol{\Gamma}}^{-1/2} \tilde{\boldsymbol{A}} \boldsymbol{v}_{\boldsymbol{\alpha}} + \tilde{\boldsymbol{\Gamma}}^{-1/2} \tilde{\boldsymbol{B}} \boldsymbol{v}_{\boldsymbol{\beta}}\right) \in \mathbb{R}^{pK_N}$$

asymptotically follows a standard multivariate normal distribution in $\mathbb{R}^{pK_N}$, where $\tilde{\boldsymbol{\Gamma}}$, $\tilde{\boldsymbol{A}}$ and $\tilde{\boldsymbol{B}}$ are defined in Equation (8.4), (8.6) and (8.9) correspondingly. The proof of the following theorem is given in Appendix G.6. Some of the required regularity conditions are given in Section 7.2 and G.1.

Theorem 8.3.3.
Assume that Regularity Condition 7.A-7.I and 8.A-8.D are satisfied and $J_N N^{-1/4} = o(1)$. Further assume that the initial estimator $(\tilde{\boldsymbol{\beta}}, \tilde{\boldsymbol{\alpha}}_N)$ converges to $(\boldsymbol{\beta}^*, \boldsymbol{\alpha}_N^*) \in \mathbb{R}^p \times \mathcal{M}$ in the sense that

$$\|\tilde{\boldsymbol{\alpha}}_N - \boldsymbol{\alpha}_N^*\|_2^2 + \left\|\tilde{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\right\|_2^2 = O_p(J_N N^{-1}),$$

as the sample size N goes to infinity. As the sample size N goes to infinity, the following statistic

$$N^{1/2}\tilde{\mathbf{\Gamma}}^{-1/2}\left(\tilde{\boldsymbol{A}}\boldsymbol{v}_{\boldsymbol{\alpha}}+\tilde{\boldsymbol{B}}\boldsymbol{v}_{\boldsymbol{\beta}}\right)\in\mathbb{R}^{pK_N}$$

converges in distribution to a standard multivariate normal distribution in $\mathbb{R}^{pK_N}$.

8.4 Simulation Studies

In this section, we conduct simulation studies to investigate the performance of the ensemble inference. The considered model is the Poisson regression model with a log-link function; see Model 5.A in Section 5.7. In Section 8.4.1, we describe the setups of the simulation studies. Firstly, we argue that the finite sample distribution of $\hat{\boldsymbol{\beta}}^{(k)}$ can not be appropriately approximated by normal distributions; see Section 8.4.2. Secondly, we show some empirical evidence that $\tilde{\mathbf{\Gamma}}^{-1/2}$ converges to $(\mathbf{\Gamma}^{*})^{-1/2}$ in the 2-norm; see Section 8.4.3. Next, we investigate the Type I errors in the ensemble inference; see Section 8.4.4. Lastly, we study the powers of the ensemble test statistics; see Section 8.4.5.

8.4.1 Simulation Setups

We set the parameter values in the Poisson regression model with a log-link function as follows. For each $n \in \{1,\ldots,N\}$, let T_n follow a discrete uniform distribution over $\{1,\ldots,4\}$. For each $n \in \{1,\ldots,N\}$ and $t \in \{1,\ldots,T_n\}$, let $\boldsymbol{X}_{nt} = (X_{nt1}, X_{nt2}, X_{nt3}, X_{nt4})^{\mathrm{T}} \in \mathbb{R}^4$ be the fixed effects, where X_{nt1} and X_{nt2} independently follow a continuous uniform distribution over $[-0.3, 0.3]$, X_{nt3} follows a Bernoulli distribution with success probability 0.5 and $X_{nt4} = 10 \times X_{nt1}X_{nt2}$ is considered as the interaction effects of X_{nt1} and X_{nt2}. For each n and t, Z_{nt} follows a continuous distribution over $[-1, 1]$. The true value of the regression parameter $\boldsymbol{\beta}$ is $(-1, 2, 0.5, 0)^{\mathrm{T}} \in \mathbb{R}^4$. The distribution of the random effects $Q(b)$ is $0.4I(b \leq -2) + 0.1I(b \leq 0) + 0.5I(b \leq 1)$.

We use the Chebyshev polynomials (see Definition 2.4.2) defined on $\mathcal{B} = [-3, 3]$ as the orthonormal basis $\{P_j(b)\}_{j=0}^{J_N}$ in $L^2(\mathcal{B}, \mu)$, where $\mu = (1 - b^2)^{-1/2}\mathrm{d}b$. The approximation property has been studied in Section 2.4.2. The dimensions of the generalized moments $\boldsymbol{\alpha} \in \mathbb{R}^{J_N}$ depends on the sample size N, where $J_N = \lfloor 2N^{1/5} \rfloor$, with $\lfloor a \rfloor$ denoting the largest integer not greater than a. The random working correlation matrices are generated from Algorithm 8.1, where $K_N = 2$. We consider six sample size levels ($N = 200, 300, 400, 500, 600$ and 700).

8.4.2 Finite Sample Distribution of $\hat{\boldsymbol{\beta}}^{(k)}$

Firstly, we argue that the elements of the estimators $\hat{\boldsymbol{\beta}}^{(k)} = (\hat{\beta}_1^{(k)}, \ldots, \hat{\beta}_4^{(k)})^{\mathrm{T}} \in \mathbb{R}^4$, $k = 1, 2$, do NOT follow normal distributions; see Figure 8.1 as an example. From this figure, we see that a normal distribution does not appropriately approximate the finite sample distribution of $\hat{\beta}_i^{(k)}$, $i = 1, \ldots, 4$ and $k = 1, 2$.

To evaluate the approximation quantitatively, for each $k \in \{1, 2\}$ and $i \in \{1, \ldots, 4\}$, we use the one-sample Kolmogorov-Smirnov (K-S) test for the null hypothesis that the finite sample distribution of the standardized $\hat{\beta}_i^{(k)}$ can be fitted by $\mathcal{N}(0, 1)$, where $\hat{\beta}_i^{(k)}$ is standardized by

$$\frac{1}{\sigma_{\mathrm{E},i,k}} \left(\hat{\beta}_i^{(k)} - \beta_i^* \right) \tag{8.19}$$

and $\sigma_{\mathrm{E},i,k}$ is the simulated variance of $\hat{\beta}_i^{(k)}$. Some of the results are reported in Table 8.1. We see that the null hypothesis is rejected in some of the cases when the significant level is 0.05. For examples, the p-values of the K-S test statistics for the finite sample distributions of $\hat{\beta}_3^{(1)}$ are smaller than the significant level 0.05 at each sample size level. The boundary effects in $\tilde{\boldsymbol{\alpha}}_N \in \mathbb{R}^{J_N}$ is the major reason that $\hat{\beta}_i^{(k)}$s are not normally distributed. In Figure 8.2, we plot the finite sample distributions of the initial estimators of $\tilde{\alpha}_{N,j}$, $j = 1, \ldots, 4$ and $N = 600$. The boundary effects of the generalized moment spaces can be observed easily.

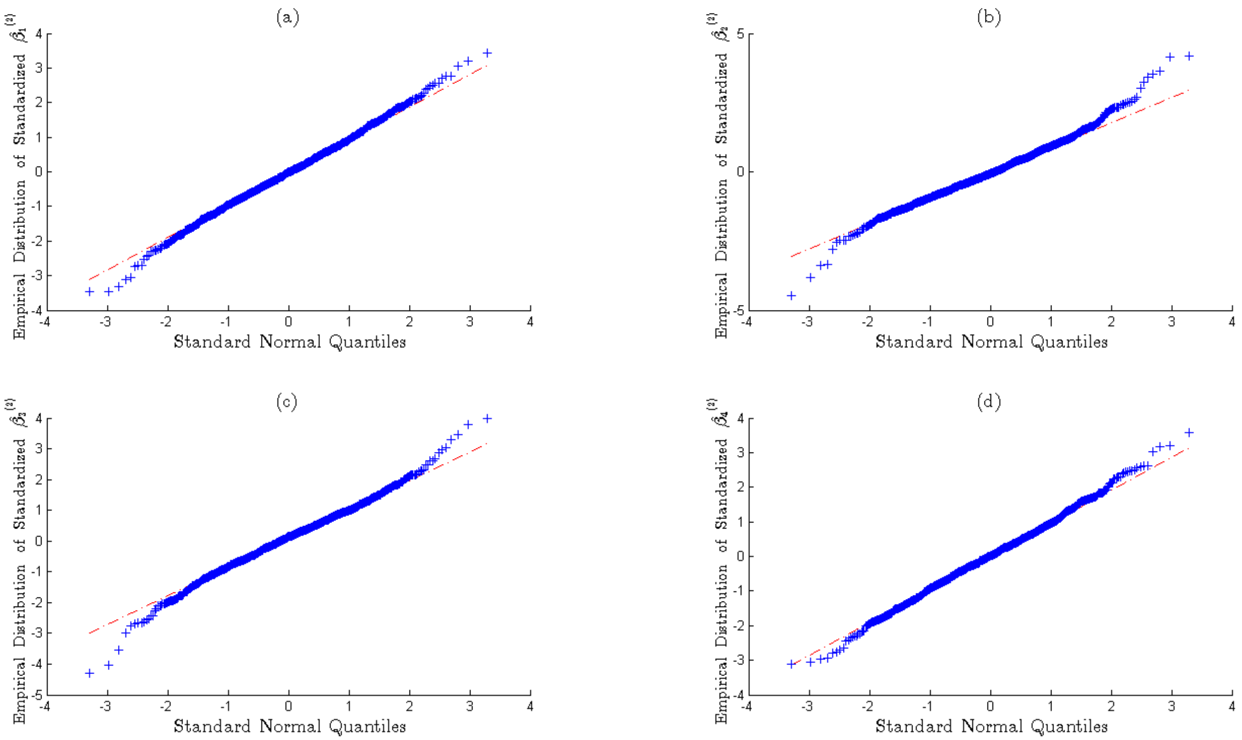

Figure 8.1: The Q-Q plots of the finite sample distributions of $\hat{\beta}_i^{(k)}$, where $k = 1$, $i = 1, \ldots, 4$ and $N = 600$.

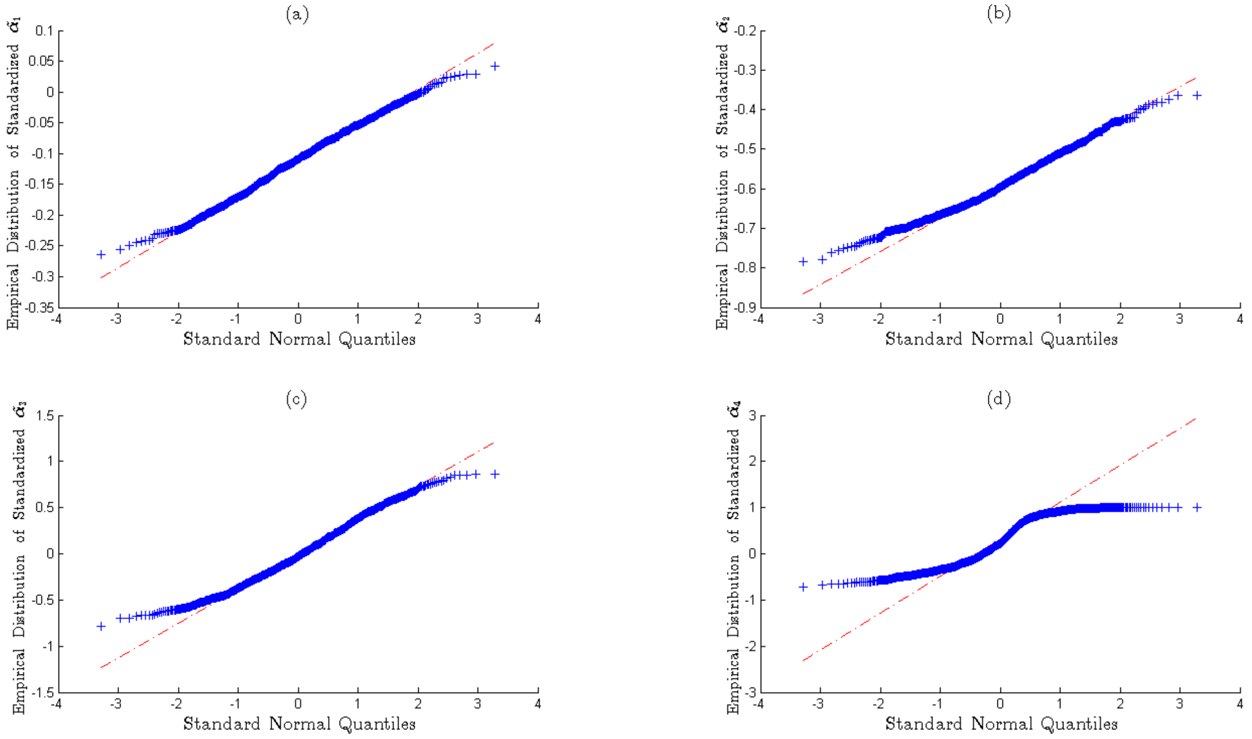

Figure 8.2: The Q-Q plots of the finite sample distributions of $\tilde{\alpha}_j$, where $j = 1, \ldots, 4$ and $N = 600$.

		$N = 200$		$N = 400$		$N = 600$	
	Parameter	K-S	p-value	K-S	p-value	K-S	p-value
$k = 1$	β_1	0.039	0.087	0.033	0.227	0.025	0.531
	β_2	0.045	0.034	0.035	0.180	0.037	0.119
	β_3	0.046	0.026	0.065	0.000	0.062	0.001
	β_4	0.023	0.665	0.021	0.783	0.029	0.365
$k = 2$	β_1	0.041	0.066	0.037	0.129	0.036	0.139
	β_2	0.027	0.429	0.040	0.082	0.021	0.756
	β_3	0.062	0.001	0.042	0.056	0.052	0.009
	β_4	0.022	0.700	0.035	0.174	0.031	0.288

Table 8.1: The Kolmogorov-Smirnov test on the normality of the standardized $\hat{\beta}_i^{(k)}$, where $k = 1, 2$, $i = 1, \ldots, 4$ and the sample size $N = 200, 400$ and 600. The K-S stands for the Kolmogorov-Smirnov test statistic.

8.4.3 The Convergence of $\tilde{\mathbf{\Gamma}}^{-1/2}$

Regularity Condition 8.D is a necessary condition for the asymptotic theorems of the ensemble inference; see Theorem 8.3.3. Without stronger conditions it is not possible to prove that the estimators of $\tilde{\mathbf{\Gamma}}$ follow this condition. Instead, we provide some empirical evidence that

$$\left\|\tilde{\mathbf{\Gamma}}^{-1/2} - (\mathbf{\Gamma}^*)^{-1/2}\right\|_2$$

converges at rate $J_N^{3/2} N^{-1/2}$.

Because $J_N = N^{1/5}$ in our simulation setup, we study the finite sample distribution of

$$N^{1/5}\left\|\tilde{\mathbf{\Gamma}}^{-1/2} - (\mathbf{\Gamma}^*)^{-1/2}\right\|_2$$

at sample size level $N = 200, 400$ and 600; see Figure 8.3. In Figure 8.3, we see that the finite sample distributions of $N^{1/5}\|\tilde{\mathbf{\Gamma}}^{-1/2} - (\mathbf{\Gamma}^*)^{-1/2}\|_2$ does not diverge with the increase of sample size N.

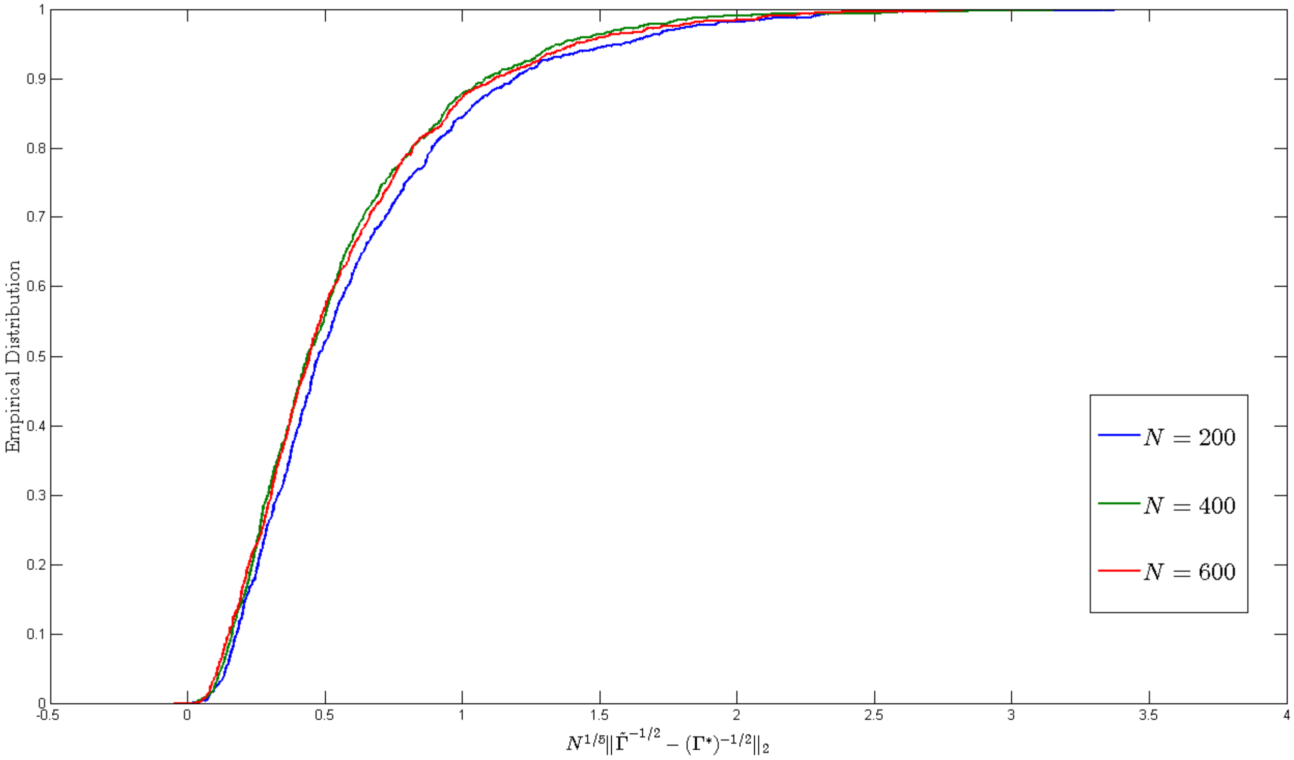

Figure 8.3: Plots of the finite sample distribution of $N^{1/5}\|\tilde{\mathbf{\Gamma}}^{-1/2} - (\mathbf{\Gamma}^{*})^{-1/2}\|_2$, when $N = 200, 400$ and 600.

8.4.4 The Type I Errors in the Ensemble Inference

Table 8.2 summarizes the Type I errors of the proposed test statistic ζ in Equation (8.10). Because $K_N = 2$ for each sample size level N and the dimension of the regression parameter p is 4, the degrees of freedom of the ensemble test statistics are calculated by $8 - J_N$, where $J_N = \lfloor 2N^{1/5} \rfloor$ and $N = 200, 300, 400, 500, 600$ and 700.

From Table 8.2, we see that the finite sample distributions of ζ is not appropriately approximated by its asymptotic distribution when the sample size is small ($N = 200$). The failure of this approximation can be explained by Figure 8.3, in which the term $N^{1/5}\|\tilde{\mathbf{\Gamma}}^{-1/2} - (\mathbf{\Gamma}^*)^{-1/2}\|_2$ is much larger in the cases where $N = 200$ than the ones in the other cases. Moreover, the large bias could be caused by the approximation quality, because the number of the generalized moments is 5 when $N = 200$.

On the other hand, when the sample size is large, the asymptotic distribution approximate the finite sample distribution appropriately.

N	J_N	d.f.	$\mathfrak{a} = 0.90$	$\mathfrak{a} = 0.95$	$\mathfrak{a} = 0.99$
200	5	3	0.806	0.867	0.943
300	6	3	0.879	0.933	0.981
400	6	2	0.884	0.932	0.981
500	6	2	0.901	0.952	0.984
600	7	1	0.886	0.942	0.988
700	7	1	0.887	0.936	0.982

Table 8.2: The simulation results of the Type I errors in the ensemble inferences. The d.f. is the degrees of freedom of the ensemble test statistic, and $\mathfrak{a}$ is the significant level.

8.4.5 The Power of the Ensemble Inference

Lastly, we study the powers that the null hypothesis

$$\mathrm{H}_0 : \quad \boldsymbol{\beta} = \mathbf{0} \in \mathbb{R}^4 \tag{8.20}$$

is rejected by the ensemble test statistic ζ. Table 8.3 summarizes the simulation results. From Table 8.3, we see that the proposed test statistic ζ has powers to reject the null hypothesis in Equation (8.20). When $N = 300$, 400 and 500, the power of ζ increases with the sample size N. There is a drop of power at $N = 600$, because the degrees of freedom of the χ^2 test statistic reduces to 1. Moreover, with the increase of the significant levels, the power decreases.

N	J_N	d.f.	$\mathfrak{a} = 0.90$	$\mathfrak{a} = 0.95$	$\mathfrak{a} = 0.99$
300	6	2	0.786	0.740	0.610
400	6	2	0.812	0.774	0.636
500	6	2	0.847	0.803	0.706
600	7	1	0.665	0.610	0.511
700	7	1	0.693	0.644	0.543

Table 8.3: The simulation results of the powers of the ensemble test statistics. The d.f. is the degrees of freedom of the ensemble test statistic, and $\mathfrak{a}$ is the significant level.

8.5 Application to the Epileptic Seizures Data

We have fitted the Epileptic Seizures Data in Section 6.5. In this section, we consider the simple hypothesis

$$\mathrm{H}_0: \quad \boldsymbol{\beta} = \mathbf{0} \in \mathbb{R}^5,$$

where $\boldsymbol{\beta} \in \mathbb{R}^5$ is the regression parameter in Model 6.C. Because there is only one generalized moment $\alpha_1 = \int_{\mathcal{B}} \exp(b) \mathrm{d}Q$ in the reparameterized model and the dimension of the regression parameter is 5, we set $K_N = 1$. Therefore, the ensemble test statistics follows a χ^2 distribution with degrees of freedom 4.

In this data set, for each $n \in \{1, \ldots, 59\}$, the visiting number of the n^{th} individual T_n is 4. Therefore, we may use a same 4×4 random correlation matrix for each

$n = 1, \ldots, 59$, following Algorithm 8.1. We repeat the ensemble inference 5000 times. In other words, we generate 5000 working correlation matrices from the Wishart distribution under in Algorithm 8.1, and compute 5000 ensemble test statistics and their p-values, each of which is associated with one generated working correlation matrix.

In Figure 8.4, we plot the histogram of these p-values. From this figure, we see that the ensemble test statistics do not have enough power to reject the null hypothesis. One possible reason is that the sample size is not large enough ($N = 59$).

8.6 Conclusion and Discussion

In this section, we propose using the ensemble inferences to construct a χ^2 distributed test statistic in the case that the true parameter is on the boundary of the parameter space. Although simulation evidences supports the idea that the finite sample distribution of the ensemble test statistics could be well approximated by the asymptotic distribution, there still exist many points that requires further investigation.

Firstly, we need to derive sufficient conditions for that the random correlation matrices generating process satisfies the regularity conditions listed in Section G.1. Using random correlation matrices aims to reduce the correlation between $\hat{\boldsymbol{\beta}}^{(k)}$, $k = 1, \ldots, K_N$, in the ensemble. However, the ensemble inference could be misleading if Regularity Condition 8.A-8.D are not satisfied by the random correlation matrices generating processes.

We also need to study the convergence rate of $\|\tilde{\boldsymbol{\Gamma}}^{-1/2} - (\boldsymbol{\Gamma}^*)^{-1/2}\|_2$, where $\tilde{\boldsymbol{\Gamma}}$ is obtained from the estimated covariance matrices $\tilde{\boldsymbol{\Sigma}}_n$ in Section 5.4. Although empirical evidence was provided in our simulation studies, computing the convergence rate of $\|\tilde{\boldsymbol{\Gamma}}^{-1/2} - (\boldsymbol{\Gamma}^*)^{-1/2}\|_2$ would complete the framework of the ensemble inference.

Different weighting matrices are also used in the method of the quadratic inference functions (QIF); [Qu et al., 2000]. It was shown that, because different weighting matrices are used, the QIF estimators are robust to the misspecification of the correla-

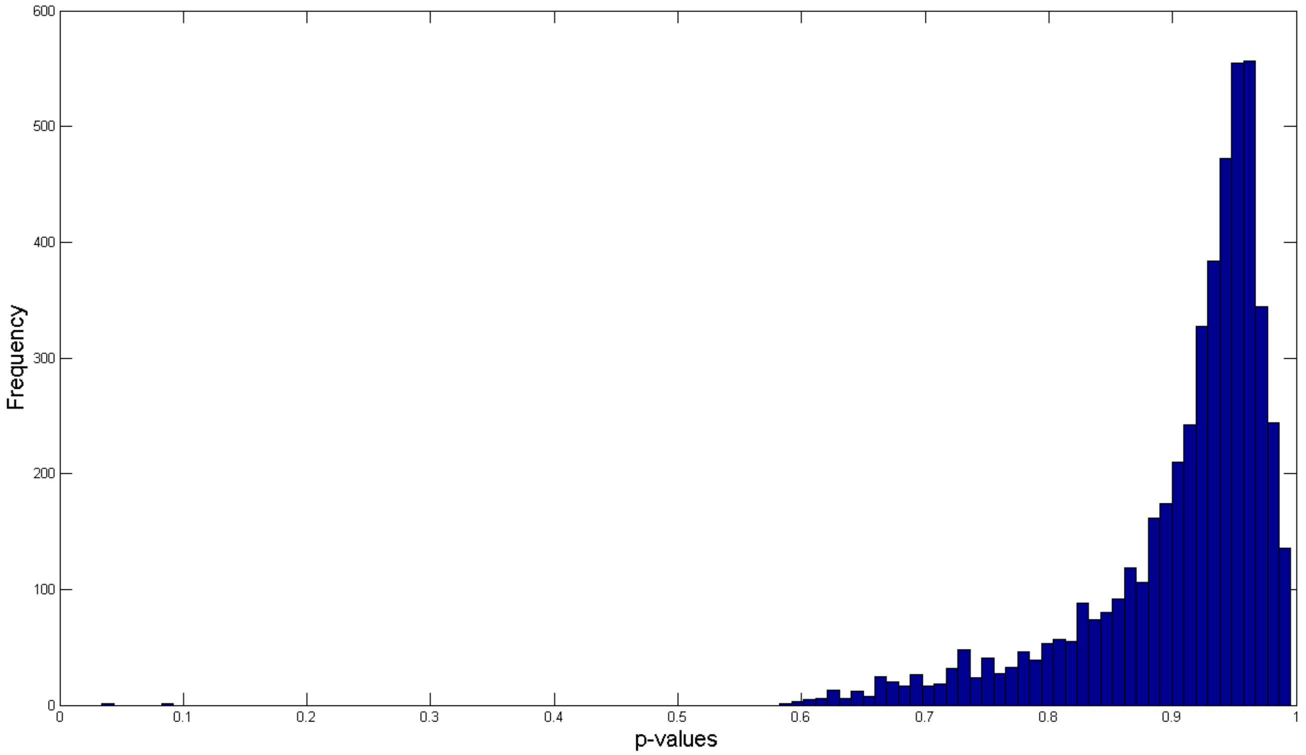

Figure 8.4: Plots of the 5000 p-values of the ensemble test statistics which are obtained from 5000 sets of random matrices.

tion structures; see [Qu and Song, 2004]. Note that different weighting matrices play important roles in both of the methods. We conjecture that the ensemble inferences are also robust to the misspecification of the correlation structure. It also worthwhile to design a framework that unify the QIF method with the ensemble inference.

The ensemble inference is closely related to Bayesian statistics. When the random matrices generating processes are given, we may consider them as a prior information to models and $\hat{\boldsymbol{\beta}}^{(k)}$, $k = 1, \ldots, K_N$, follows a posterior distribution. Therefore, it is also interesting to investigate the performances of the ensemble inference when different random matrices generating processes are used.

Lastly, we want to point out that the asymptotically χ^2 test statistic is not obtained for free. The power loses when we project the asymptotically normal statistics from $\mathbb{R}^{pK_N}$ to $\mathbb{R}^{pK_N - J_N - 1}$. With the increase of the number of the generalized moments in a model, the degrees of freedom of the ensemble test statistics decrease. And so the power of the proposed test statistic decreases. The power of the ensemble statistics could be increased by increasing the number of models K_N. However, we can not choose an arbitrary K_N, because we also need to control the convergence rate of $\tilde{\boldsymbol{A}}$, $\tilde{\boldsymbol{B}}$ and $\tilde{\boldsymbol{\Gamma}}^{-1/2}$, whose dimensions depend on K_N.

Appendix: G

G.1 Regularity Conditions

In the ensemble inference, we need the following important assumptions. However, we could not prove that our proposed working correlation matrices generating process and estimated covariance matrices $\tilde{\boldsymbol{\Gamma}}$ could satisfies the given regularity conditions. Although empirical evidence in given in the simulation studies, further investigations are required.

Regularity Condition 8.A.
The $pK_N \times (J_N + 1)$ matrix $\boldsymbol{A}^*$ is full column rank, $pK_N > J_N + 1$. For each K_N, $\|\boldsymbol{A}^*\|_2$ is bounded. Here, for each matrix $\boldsymbol{A}$, $\|\boldsymbol{A}\|_2$ is the 2-norm of the matrix $\boldsymbol{A}$

which is defined as

$$\|\boldsymbol{A}\|_2 = \lambda_{\max}\left(\boldsymbol{A}^{\mathrm{T}}\boldsymbol{A}\right).$$

Moreover, for each $k \in \{1, \ldots, K_N\}$, the $p \times J_N + 1$ matrix

$$\frac{1}{N}\sum_{n=1}^{N} \boldsymbol{X}_n \boldsymbol{D}_n^* \boldsymbol{W}_n^{(k)} \boldsymbol{\Phi}_n^{\mathrm{T}}(\boldsymbol{\beta}^*),$$

converges to $\boldsymbol{A}^*$ element-wise at rate $J_N^{1/2} N^{-1/2}$.

Regularity Condition 8.B.

The $pK_N \times pK_N$ matrix $\boldsymbol{B}^*$ is full rank. For each K_N, $\|\boldsymbol{B}^*\|_2$ is bounded. Moreover, for each $k \in \{1, \ldots, K_N\}$, the $p \times p$ matrix and

$$\frac{1}{N}\sum_{n=1}^{N} \boldsymbol{X}_n \boldsymbol{D}_n^* \boldsymbol{W}_n^{(k)} \boldsymbol{D}_n^* \boldsymbol{X}_n^{\mathrm{T}}.$$

converges to $\boldsymbol{B}^*$ element-wise at rate $J_N^{1/2} N^{-1/2}$.

Regularity Condition 8.C.

The $pK_N \times pK_N$ covariance matrix $\boldsymbol{\Gamma}^*$ is full rank and its elements are bounded.

Regularity Condition 8.D.

The $pK_N \times pK_N$ matrix $\tilde{\boldsymbol{\Gamma}}^{-1/2}$ converges in probability to $(\boldsymbol{\Gamma}^*)^{-1/2}$ in the sense that

$$\left\|\tilde{\boldsymbol{\Gamma}}^{-1/2} - (\boldsymbol{\Gamma}^*)^{-1/2}\right\|_2^2 = O_p(J_N^3 N^{-1}),$$

as the sample size N goes to infinity.

G.2 Proof of Theorem 8.3.1

Proof. We aim to prove that, $\forall \epsilon > 0$, there exists a $C > 0$, depending on N_0, such that, for any $N \geq N_0$,

$$\begin{aligned}&\mathrm{pr}\left(\sup_{\|\boldsymbol{\beta}-\boldsymbol{\beta}^*\|_2 = C\Delta_N} (\boldsymbol{\beta}-\boldsymbol{\beta}^*)^{\mathrm{T}} \left(\frac{1}{N}\sum_{n=1}^{N} \boldsymbol{X}_n \tilde{\boldsymbol{D}}_n \tilde{\boldsymbol{W}}_n^{(k)} \boldsymbol{U}_n(\boldsymbol{\beta}, \tilde{\boldsymbol{\alpha}}_N)\right) < 0\right)\\ &\geq 1-\epsilon,\end{aligned}$$

where $\Delta_N = J_N^{1/2} N^{-1/2}$; see [Ortega and Rheinboldt, 1970, Theorem 6.3.4] and [Wang, 2011].

By Taylor's expansion, we have, for each $n \in \{1, \ldots, N\}$,

$$(\boldsymbol{\beta} - \boldsymbol{\beta}^*)^{\mathrm{T}} \boldsymbol{X}_n \tilde{\boldsymbol{D}}_n \tilde{\boldsymbol{W}}_n^{(k)} \boldsymbol{U}_n(\boldsymbol{\beta}, \tilde{\boldsymbol{\alpha}}_N) = I_{n1} + I_{n2},$$

where

$$I_{n1} = (\boldsymbol{\beta} - \boldsymbol{\beta}^*)^{\mathrm{T}} \boldsymbol{X}_n \tilde{\boldsymbol{D}}_n \tilde{\boldsymbol{W}}_n^{(k)} \boldsymbol{U}_n(\boldsymbol{\beta}^*, \tilde{\boldsymbol{\alpha}}_N)$$

and

$$I_{n2} = (\boldsymbol{\beta} - \boldsymbol{\beta}^*)^{\mathrm{T}} \boldsymbol{X}_n \tilde{\boldsymbol{D}}_n \tilde{\boldsymbol{W}}_n^{(k)} \boldsymbol{D}_n(\breve{\boldsymbol{\beta}}, \tilde{\boldsymbol{\alpha}}_N) \boldsymbol{X}_n^{\mathrm{T}} (\boldsymbol{\beta} - \boldsymbol{\beta}^*),$$

and $\breve{\boldsymbol{\beta}}$ is on the line segment between $\boldsymbol{\beta}$ and $\boldsymbol{\beta}^*$.

We write, for each $n \in \{1, \ldots, N\}$,

$$I_{n1} = I_{n11} + I_{n12} + I_{n13} + I_{n14},$$

where

$$\begin{aligned}
I_{n11} &= (\boldsymbol{\beta} - \boldsymbol{\beta}^*)^{\mathrm{T}} \boldsymbol{X}_n \boldsymbol{D}_n^* \boldsymbol{W}_n^{(k)} \boldsymbol{U}_n(\boldsymbol{\beta}^*, Q^*),\\
I_{n12} &= (\boldsymbol{\beta} - \boldsymbol{\beta}^*)^{\mathrm{T}} \boldsymbol{X}_n \boldsymbol{D}_n^* \boldsymbol{W}_n^{(k)} \left(\boldsymbol{U}_n(\boldsymbol{\beta}^*, \tilde{\boldsymbol{\alpha}}_N) - \boldsymbol{U}_n(\boldsymbol{\beta}^*, Q^*)\right),\\
I_{n13} &= (\boldsymbol{\beta} - \boldsymbol{\beta}^*)^{\mathrm{T}} \boldsymbol{X}_n \left(\tilde{\boldsymbol{D}}_n \tilde{\boldsymbol{W}}_n^{(k)} - \boldsymbol{D}_n^* \boldsymbol{W}_n^{(k)}\right) \boldsymbol{U}_n(\boldsymbol{\beta}^*, Q^*),
\end{aligned}$$

and

$$I_{n14} = (\boldsymbol{\beta} - \boldsymbol{\beta}^*)^{\mathrm{T}} \boldsymbol{X}_n \left(\tilde{\boldsymbol{D}}_n \tilde{\boldsymbol{W}}_n^{(k)} - \boldsymbol{D}_n^* \boldsymbol{W}_n^{(k)}\right) \left(\boldsymbol{U}_n(\boldsymbol{\beta}^*, \tilde{\boldsymbol{\alpha}}_N) - \boldsymbol{U}_n(\boldsymbol{\beta}^*, Q^*)\right).$$

By Regularity Condition 7.A-7.C, Lemma F.6 and 7.2.2, Theorem 7.4.1 and Equation (7.16), we have

$$\begin{aligned}
&|I_{n12}|^2\\
\le\ & \|\boldsymbol{\beta} - \boldsymbol{\beta}^*\|_2^2 \left\|\boldsymbol{U}_n(\boldsymbol{\beta}^*, \tilde{\boldsymbol{\alpha}}_N) - \boldsymbol{U}_n(\boldsymbol{\beta}^*, Q^*)\right\|_2^2 \lambda_{\max}\left(\boldsymbol{X}_n \left(\boldsymbol{D}_n^* \boldsymbol{W}_n^{(k)}\right)^2 \boldsymbol{X}_n^{\mathrm{T}}\right)\\
=\ & O_p(C^2 \Delta_N^4),
\end{aligned}$$

and

$$\begin{aligned}
&|I_{n13}|^2 \\
\leq\ & \|\boldsymbol{\beta}-\boldsymbol{\beta}^*\|_2^2 \, \|\boldsymbol{U}_n(\boldsymbol{\beta}^*, Q^*)\|_2^2 \, \lambda_{\max}\left(\boldsymbol{X}_n\boldsymbol{X}_n^{\mathrm{T}}\right) \\
& \times \lambda_{\max}\left(\left(\tilde{\boldsymbol{D}}_n\tilde{\boldsymbol{W}}_n^{(k)} - \boldsymbol{D}_n^*\boldsymbol{W}_n^{(k)}\right)^2\right) \\
=\ & O_p(C^2\Delta_N^4)
\end{aligned}$$

and

$$\begin{aligned}
&|I_{n14}|^2 \\
\leq\ & \|\boldsymbol{\beta}-\boldsymbol{\beta}^*\|_2^2 \, \|\boldsymbol{U}_n(\boldsymbol{\beta}^*, \tilde{\boldsymbol{\alpha}}_N) - \boldsymbol{U}_n(\boldsymbol{\beta}^*, Q^*)\|_2^2 \, \lambda_{\max}\left(\boldsymbol{X}_n\boldsymbol{X}_n^{\mathrm{T}}\right) \\
& \times \lambda_{\max}\left(\left(\tilde{\boldsymbol{D}}_n\tilde{\boldsymbol{W}}_n^{(k)} - \boldsymbol{D}_n^*\boldsymbol{W}_n^{(k)}\right)^2\right) \\
=\ & O_p(C^2\Delta_N^6).
\end{aligned}$$

By Lemma F.4 and the Cauchy-Schwarz inequality, we have

$$\begin{aligned}
\left|\frac{1}{N}\sum_{n=1}^{N} I_{n11}\right|^2 &\leq \|\boldsymbol{\beta}-\boldsymbol{\beta}^*\|_2^2 \times \left\|\frac{1}{N}\sum_{n=1}^{N} \boldsymbol{X}_n\boldsymbol{D}_n^*\boldsymbol{W}_n^{(k)}\boldsymbol{U}_n(\boldsymbol{\beta}^*, Q^*)\right\|_2^2 \\
&\leq O_p(C^2\Delta_N^4).
\end{aligned}$$

Therefore, we have

$$\frac{1}{N}\sum_{n=1}^{N} I_{n1} = O_p(C\Delta_N^2). \tag{G.1}$$

Next, we evaluate the asymptotic order of I_{n2}. We have

$$I_{n2} = I_{n21} + I_{n22},$$

where

$$I_{n21} = (\boldsymbol{\beta}-\boldsymbol{\beta}^*)^{\mathrm{T}} \boldsymbol{X}_n\boldsymbol{D}_n^*\boldsymbol{W}_n^{(k)}\boldsymbol{D}_n^*\boldsymbol{X}_n^{\mathrm{T}} (\boldsymbol{\beta}-\boldsymbol{\beta}^*),$$

and

$$I_{n22} = \left(\boldsymbol{\beta} - \boldsymbol{\beta}^*\right)^{\mathrm{T}} \boldsymbol{X}_n \left(\tilde{\boldsymbol{D}}_n \tilde{\boldsymbol{W}}_n^{(k)} \boldsymbol{D}_n(\breve{\boldsymbol{\beta}}, \tilde{\boldsymbol{\alpha}}_N) - \boldsymbol{D}_n^* \boldsymbol{W}_n^{(k)} \boldsymbol{D}_n^*\right) \boldsymbol{X}_n^{\mathrm{T}} \left(\boldsymbol{\beta} - \boldsymbol{\beta}^*\right)$$

By Regularity Condition 7.A and Lemma 7.2.1, we have

$$\begin{aligned} |I_{n21}| \leq & \ \|\boldsymbol{\beta} - \boldsymbol{\beta}^*\|_2^2 \, \lambda_{\max}\left(\boldsymbol{X}_n \boldsymbol{D}_n^* \boldsymbol{W}_n^{(k)} \boldsymbol{D}_n^* \boldsymbol{X}_n^{\mathrm{T}}\right) \\ = & \ O_p(C^2 \Delta_N^2). \end{aligned}$$

By Theorem 7.4.1, Lemma 7.2.2 and F.6, Regularity Condition 7.A and Equation (7.16), we have

$$\begin{aligned} & |I_{n22}| \\ \leq & \ \|\boldsymbol{\beta} - \boldsymbol{\beta}^*\|_2^2 \, \lambda_{\max}\left(\boldsymbol{X}_n \boldsymbol{X}_n^{\mathrm{T}}\right) \lambda_{\max}\left(\tilde{\boldsymbol{D}}_n \tilde{\boldsymbol{W}}_n^{(k)} \boldsymbol{D}_n(\breve{\boldsymbol{\beta}}, \tilde{\boldsymbol{\alpha}}_N) - \boldsymbol{D}_n^* \boldsymbol{W}_n^{(k)} \boldsymbol{D}_n^*\right), \\ = & \ O_p(C^2 \Delta_N^3). \end{aligned}$$

Therefore, we have

$$\frac{1}{N} \sum_{n=1}^{N} I_{n2} = O_p(C^2 \Delta_N^2). \tag{G.2}$$

By Equation (G.1) and (G.2), we have

$$\left(\boldsymbol{\beta} - \boldsymbol{\beta}^*\right)^{\mathrm{T}} \boldsymbol{X}_n \tilde{\boldsymbol{D}}_n \tilde{\boldsymbol{W}}_n^{(k)} \boldsymbol{U}_n(\boldsymbol{\beta}, \tilde{\boldsymbol{\alpha}}_N)$$

is dominated by

$$\frac{1}{N} \sum_{n=1}^{N} I_{n21} = \frac{1}{N} \sum_{n=1}^{N} \left(\boldsymbol{\beta} - \boldsymbol{\beta}^*\right)^{\mathrm{T}} \boldsymbol{X}_n \boldsymbol{D}_n^* \boldsymbol{W}_n^{(k)} \boldsymbol{D}_n^* \boldsymbol{X}_n^{\mathrm{T}} \left(\boldsymbol{\beta} - \boldsymbol{\beta}^*\right) > 0$$

by allowing C to be large enough. It follows that

$$\left(\boldsymbol{\beta} - \boldsymbol{\beta}^*\right)^{\mathrm{T}} \boldsymbol{X}_n \tilde{\boldsymbol{D}}_n \tilde{\boldsymbol{W}}_n^{(k)} \boldsymbol{U}_n(\boldsymbol{\beta}, \tilde{\boldsymbol{\alpha}}_N)$$

converges to a positive number in probability. □

G.3 Proof of Lemma G.1

Lemma G.1.

Assume that Regularity Condition 7.A-7.I are satisfied. Further assume that, for each $k \in \{1, \ldots, K_N\}$, $(\hat{\boldsymbol{\beta}}^{(k)}, \tilde{\boldsymbol{\alpha}}_N)$ and the initial estimator $(\tilde{\boldsymbol{\beta}}, \tilde{\boldsymbol{\alpha}}_N)$ converges to $(\boldsymbol{\beta}^*, \boldsymbol{\alpha}_N^*)$ in the sense that

$$\|\tilde{\boldsymbol{\alpha}}_N - \boldsymbol{\alpha}_N^*\|_2^2 + \left\|\hat{\boldsymbol{\beta}}^{(k)} - \boldsymbol{\beta}^*\right\|_2^2 = O_p(J_N N^{-1})$$

and

$$\|\tilde{\boldsymbol{\alpha}}_N - \boldsymbol{\alpha}_N^*\|_2^2 + \left\|\tilde{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\right\|_2^2 = O_p(J_N N^{-1}),$$

as the sample size N goes to infinity. If $J_N N^{-1/2} = o(1)$ as the sample size N goes to infinity, then, for each $k \in \{1, \ldots, K_N\}$,

$$\begin{aligned} & N^{1/2}\left(\boldsymbol{A}_{(k)}^* \left(\tilde{\boldsymbol{\alpha}}_N - \boldsymbol{\alpha}_N^*\right) + \boldsymbol{B}_{(k)}^* \left(\hat{\boldsymbol{\beta}}^{(k)} - \boldsymbol{\beta}^*\right)\right) \\ = {} & N^{1/2} \sum_{n=1}^{N} \boldsymbol{X}_n \boldsymbol{D}_n^* \boldsymbol{W}_n^{(k)} \boldsymbol{U}_n(\boldsymbol{\beta}^*, Q^*) + o_p(J_N N^{-1/2}). \end{aligned}$$

Proof. For each n, we have

$$\begin{aligned} & \boldsymbol{X}_n \tilde{\boldsymbol{D}}_n \tilde{\boldsymbol{W}}_n^{(k)} \boldsymbol{U}_n(\hat{\boldsymbol{\beta}}^{(k)}, \tilde{\boldsymbol{\alpha}}_N) \\ = {} & I_{n1} + I_{n2} + I_{n3} + I_{n4}, \end{aligned}$$

where

$$\begin{aligned} I_{n1} &= \boldsymbol{X}_n \boldsymbol{D}_n^* \boldsymbol{W}_n^{(k)} \boldsymbol{U}_n(\boldsymbol{\beta}^*, Q^*), \\ I_{n2} &= \boldsymbol{X}_n \boldsymbol{D}_n^* \boldsymbol{W}_n^{(k)} \left(\boldsymbol{U}_n(\hat{\boldsymbol{\beta}}^{(k)}, \tilde{\boldsymbol{\alpha}}_N) - \boldsymbol{U}_n(\boldsymbol{\beta}^*, Q^*)\right), \\ I_{n3} &= \boldsymbol{X}_n \left(\tilde{\boldsymbol{D}}_n \tilde{\boldsymbol{W}}_n^{(k)} - \boldsymbol{D}_n^* \boldsymbol{W}_n^{(k)}\right) \boldsymbol{U}_n(\boldsymbol{\beta}^*, Q^*) \end{aligned}$$

and

$$I_{n4} = \boldsymbol{X}_n \left(\tilde{\boldsymbol{D}}_n \tilde{\boldsymbol{W}}_n^{(k)} - \boldsymbol{D}_n^* \boldsymbol{W}_n^{(k)}\right) \left(\boldsymbol{U}_n(\hat{\boldsymbol{\beta}}^{(k)}, \tilde{\boldsymbol{\alpha}}_N) - \boldsymbol{U}_n(\boldsymbol{\beta}^*, Q^*)\right).$$

By Lemma F.7, we have

$$-\frac{1}{N}\sum_{n=1}^{N} I_{n2} = \boldsymbol{A}^*_{(k)}(\tilde{\boldsymbol{\alpha}} - \boldsymbol{\alpha}^*) + \boldsymbol{B}^*_{(k)}(\tilde{\boldsymbol{\beta}} - \boldsymbol{\beta}^*) + o_p(J_N N^{-1})$$

By Lemma F.8 and F.9, we have

$$\frac{1}{N}\sum_{n=1}^{N} I_{n3} = O_p(J_N N^{-1}).$$

By Regularity Condition 7.A, Theorem 7.4.1 and Lemma 7.2.2 and F.6, we have,

$$\begin{aligned}
&|I_{n4}|^2 \\
=& \left\|\boldsymbol{U}_n(\hat{\boldsymbol{\beta}}^{(k)}, \tilde{\boldsymbol{\alpha}}_N) - \boldsymbol{U}_n(\boldsymbol{\beta}^*, Q^*)\right\|_2^2 \lambda_{\max}\left(\boldsymbol{X}_n \boldsymbol{X}_n^{\mathrm{T}}\right) \\
&\times \lambda_{\max}\left(\left(\tilde{\boldsymbol{D}}_n \tilde{\boldsymbol{W}}_n^{(k)} - \boldsymbol{D}_n^* \boldsymbol{W}_n^{(k)}\right)^2\right) \\
\leq& O_p(J_N^2 N^{-2}).
\end{aligned}$$

Because $J_N N^{-1/2} = o(1)$, we have

$$\begin{aligned}
&N^{1/2}\left(\boldsymbol{A}^*_{(k)}(\tilde{\boldsymbol{\alpha}} - \boldsymbol{\alpha}^*) + \boldsymbol{B}^*_{(k)}(\boldsymbol{\beta}^{(k)} - \boldsymbol{\beta}^*)\right) + o_p(J_N N^{-1/2}) \\
=& N^{-1/2}\sum_{n=1}^{N} I_{n1}.
\end{aligned}$$

□

G.4 Proof of Theorem 8.3.2

Proof. According to Lemma G.1, we have, for each $k \in \{1, \ldots, K_N\}$,

$$\begin{aligned}
&N^{1/2}\left(\boldsymbol{A}^* \boldsymbol{v}_{\boldsymbol{\alpha}} + \boldsymbol{B}^* \boldsymbol{v}_{\boldsymbol{\beta}}\right) \\
=& \begin{bmatrix} \frac{1}{N}\sum_{n=1}^{N} \boldsymbol{X}_n \boldsymbol{D}_n^* \boldsymbol{W}_n^{(1)} \boldsymbol{U}_n(\boldsymbol{\beta}^*, Q^*) \\ \vdots \\ \frac{1}{N}\sum_{n=1}^{N} \boldsymbol{X}_n \boldsymbol{D}_n^* \boldsymbol{W}_n^{(K_N)} \boldsymbol{U}_n(\boldsymbol{\beta}^*, Q^*) \end{bmatrix} + o_p(J_N N^{-1/2}),
\end{aligned}$$

as N goes to infinity. By Lemma F.12, it is known that, for each $k \in \{1, \ldots, K_N\}$,

$$\frac{1}{N}\sum_{n=1}^{N} \boldsymbol{X}_n \boldsymbol{D}_n^* \boldsymbol{W}_n^{(k)} \boldsymbol{U}_n(\boldsymbol{\beta}^*, Q^*) \in \mathbb{R}^p$$

converges in distribution to a multivariate normal. It follows that

$$\begin{bmatrix} \frac{1}{N}\sum_{n=1}^{N} \boldsymbol{X}_n \boldsymbol{D}_n^* \boldsymbol{W}_n^{(1)} \boldsymbol{U}_n(\boldsymbol{\beta}^*, Q^*) \\ \vdots \\ \frac{1}{N}\sum_{n=1}^{N} \boldsymbol{X}_n \boldsymbol{D}_n^* \boldsymbol{W}_n^{(K_N)} \boldsymbol{U}_n(\boldsymbol{\beta}^*, Q^*) \end{bmatrix} \in \mathbb{R}^{pK_N}$$

converges in distribution to a multivariate normal in $\mathbb{R}^{pK_N}$ converges in distribution to a multivariate normal with mean zero and covariance matrix $\boldsymbol{\Gamma}^*$, which is defined in Equation (8.13). Here we complete the proof. □

G.5 Proof of Lemma G.2

Lemma G.2.

Assume that Regularity Condition 8.A and 8.B are satisfied. Further assume that, as the sample size N goes to infinity, $J_N^2 N^{-1} = o(1)$, $K_N = O(J_N)$, and for each $k \in \{1, \ldots, K_N\}$, the initial estimator $(\tilde{\boldsymbol{\beta}}, \tilde{\boldsymbol{\alpha}}_N)$ and $(\hat{\boldsymbol{\beta}}^{(k)}, \tilde{\boldsymbol{\alpha}}_N)$ converges to $(\boldsymbol{\beta}^*, \boldsymbol{\alpha}_N^*)$ in the sense that

$$\|\tilde{\boldsymbol{\alpha}}_N - \boldsymbol{\alpha}_N^*\|_2^2 + \left\|\hat{\boldsymbol{\beta}}^{(k)} - \boldsymbol{\beta}^*\right\|_2^2 = O_p(J_N N^{-1})$$

and

$$\|\tilde{\boldsymbol{\alpha}}_N - \boldsymbol{\alpha}_N^*\|_2^2 + \left\|\tilde{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\right\|_2^2 = O_p(J_N N^{-1}).$$

Then, $\tilde{\boldsymbol{A}}\boldsymbol{v}_{\boldsymbol{\alpha}} + \tilde{\boldsymbol{B}}\boldsymbol{v}_{\boldsymbol{\beta}}$ converge in probability to $\boldsymbol{A}^*\boldsymbol{v}_{\boldsymbol{\alpha}} - \boldsymbol{B}^*\boldsymbol{v}_{\boldsymbol{\beta}}$ in the sense that

$$\left\|\tilde{\boldsymbol{A}}\boldsymbol{v}_{\boldsymbol{\alpha}} + \tilde{\boldsymbol{B}}\boldsymbol{v}_{\boldsymbol{\beta}} - \boldsymbol{A}^*\boldsymbol{v}_{\boldsymbol{\alpha}} - \boldsymbol{B}^*\boldsymbol{v}_{\boldsymbol{\beta}}\right\|_2 = O_p(J_N^2 N^{-1}),$$

as the sample size N goes to infinity, where $\tilde{\boldsymbol{A}}$, $\tilde{\boldsymbol{B}}$, $\boldsymbol{A}^*$, $\boldsymbol{B}^*$, $\boldsymbol{v}_{\boldsymbol{\alpha}}$ and $\boldsymbol{v}_{\boldsymbol{\beta}}$ are defined in Equation (8.6), (8.9), (8.15), (8.16) , (8.12) and (8.11).

Proof. By the triangle inequality we have

$$\left\|\tilde{\boldsymbol{A}}\boldsymbol{v}_{\boldsymbol{\alpha}} + \tilde{\boldsymbol{B}}\boldsymbol{v}_{\boldsymbol{\beta}} - \boldsymbol{A}^*\boldsymbol{v}_{\boldsymbol{\alpha}} - \boldsymbol{B}^*\boldsymbol{v}_{\boldsymbol{\beta}}\right\|_2 \le \left\|\tilde{\boldsymbol{A}}\boldsymbol{v}_{\boldsymbol{\alpha}} - \boldsymbol{A}^*\boldsymbol{v}_{\boldsymbol{\alpha}}\right\|_2 + \left\|\tilde{\boldsymbol{B}}\boldsymbol{v}_{\boldsymbol{\beta}} - \boldsymbol{B}^*\boldsymbol{v}_{\boldsymbol{\beta}}\right\|_2 .$$

We consider the asymptotic order of the two terms on the right hand side of the above inequality.

For each $k \in \{1, \ldots, K_N\}$, we have that $\tilde{\boldsymbol{A}}_{(k)}$ and $\tilde{\boldsymbol{B}}_{(k)}$ converges in probability to $\boldsymbol{A}^*_{(k)}$ and $\boldsymbol{B}^*_{(k)}$ element-wise at rate $J_N^{-1/2}N^{1/2}$, by Theorem 7.4.1, Lemma F.6 and Regularity Condition 8.A and 8.B. It follows that

$$\begin{aligned}
\left\|\tilde{\boldsymbol{A}}\boldsymbol{v}_{\boldsymbol{\alpha}} - \boldsymbol{A}^*\boldsymbol{v}_{\boldsymbol{\alpha}}\right\|_2^2 &= \sum_{k=1}^{K_N} \left\|\tilde{\boldsymbol{A}}_{(k)}\boldsymbol{v}_{\boldsymbol{\alpha}} - \boldsymbol{A}^*_{(k)}\boldsymbol{v}_{\boldsymbol{\alpha}}\right\|_2^2 \\
&\le \sum_{k=1}^{K_N} \|\boldsymbol{v}_{\boldsymbol{\alpha}}\|_2^2 \left\|\tilde{\boldsymbol{A}}_{(k)} - \tilde{\boldsymbol{A}}^*_{(k)}\right\|_2^2 \\
&\le \sum_{k=1}^{K_N} \|\boldsymbol{v}_{\boldsymbol{\alpha}}\|_2^2 \times pJ_N \times \left\|\tilde{\boldsymbol{A}}_{(k)} - \tilde{\boldsymbol{A}}^*_{(k)}\right\|_\infty^2 \\
&= K_N O_p(J_N N^{-1}) O(J_N) O(J_N N^{-1}) \\
&= O_p(J_N^4 N^{-2}).
\end{aligned}$$

On the other hand,

$$\begin{aligned}
\left\|\tilde{\boldsymbol{B}}\boldsymbol{v}_{\boldsymbol{\beta}} - \boldsymbol{B}^*\boldsymbol{v}_{\boldsymbol{\beta}}\right\|_2 &= \sum_{k=1}^{K_N} \left\|\tilde{\boldsymbol{B}}_{(k)}\boldsymbol{v}_{\boldsymbol{\beta}} - \boldsymbol{B}^*_{(k)}\boldsymbol{v}_{\boldsymbol{\beta}}\right\|_2^2 \\
&\le \sum_{k=1}^{K_N} \left\|\hat{\boldsymbol{\beta}}^{(k)} - \boldsymbol{\beta}^*\right\|_2^2 \left\|\tilde{\boldsymbol{B}}_{(k)} - \tilde{\boldsymbol{B}}^*_{(k)}\right\|_2^2 \\
&\le \sum_{k=1}^{K_N} \left\|\hat{\boldsymbol{\beta}}^{(k)} - \boldsymbol{\beta}^*\right\|_2^2 \times p^2 \times \left\|\tilde{\boldsymbol{B}}_{(k)} - \tilde{\boldsymbol{B}}^*_{(k)}\right\|_\infty^2 \\
&= K_N O_p(J_N N^{-1}) O(1) O(J_N N^{-1}) \\
&= O_p(J_N^3 N^{-2}).
\end{aligned}$$

In sum, we have

$$\left\|\tilde{\boldsymbol{A}}\boldsymbol{v}_{\boldsymbol{\alpha}} + \tilde{\boldsymbol{B}}\boldsymbol{v}_{\boldsymbol{\beta}} - \boldsymbol{A}^*\boldsymbol{v}_{\boldsymbol{\alpha}} - \boldsymbol{B}^*\boldsymbol{v}_{\boldsymbol{\beta}}\right\|_2 = O_p(J_N^2 N^{-1}).$$

□

G.6 Proof of Theorem 8.3.3

Proof. Consider the expansion

$$N^{1/2}\tilde{\mathbf{\Gamma}}^{-1/2}\left(\tilde{\boldsymbol{A}}\boldsymbol{v}_{\boldsymbol{\alpha}}+\tilde{\boldsymbol{B}}\boldsymbol{v}_{\boldsymbol{\beta}}\right)=I_1+I_2+I_3+I_4$$

where

$$\begin{aligned}
I_1 &= N^{1/2}\left(\mathbf{\Gamma}^*\right)^{-1/2}\left(\boldsymbol{A}^*\boldsymbol{v}_{\boldsymbol{\alpha}}+\boldsymbol{B}^*\boldsymbol{v}_{\boldsymbol{\beta}}\right),\\
I_2 &= N^{1/2}\left(\tilde{\mathbf{\Gamma}}^{-1/2}-\left(\mathbf{\Gamma}^*\right)^{-1/2}\right)\left(\boldsymbol{A}^*\boldsymbol{v}_{\boldsymbol{\alpha}}+\boldsymbol{B}^*\boldsymbol{v}_{\boldsymbol{\beta}}\right),\\
I_3 &= N^{1/2}\left(\mathbf{\Gamma}^*\right)^{-1/2}\left(\left(\tilde{\boldsymbol{A}}-\boldsymbol{A}^*\right)\boldsymbol{v}_{\boldsymbol{\alpha}}+\left(\tilde{\boldsymbol{B}}-\boldsymbol{B}^*\right)\boldsymbol{v}_{\boldsymbol{\beta}}\right)
\end{aligned}$$

and

$$I_4=N^{1/2}\left(\tilde{\mathbf{\Gamma}}^{-1/2}-\left(\mathbf{\Gamma}^*\right)^{-1/2}\right)\left(\left(\tilde{\boldsymbol{A}}-\boldsymbol{A}^*\right)\boldsymbol{v}_{\boldsymbol{\alpha}}+\left(\tilde{\boldsymbol{B}}-\boldsymbol{B}^*\right)\boldsymbol{v}_{\boldsymbol{\beta}}\right).$$

By Theorem 8.3.2, we know that I_1 converges in distribution to a standard multivariate normal. So, it is sufficient to show that each element of $I_2+I_3+I_4$ is $o_p(1)$, as N goes to infinity.

Let $\boldsymbol{u}_i\in\mathbb{R}^{pK_N}$ whose i^{th} element is one and the rests are zeros. We have

$$\begin{aligned}
&\left|\boldsymbol{u}_i^{\mathrm{T}}\left(\tilde{\mathbf{\Gamma}}^{-1/2}-\left(\mathbf{\Gamma}^*\right)^{-1/2}\right)\left(\boldsymbol{A}^*\boldsymbol{v}_{\boldsymbol{\alpha}}+\boldsymbol{B}^*\boldsymbol{v}_{\boldsymbol{\beta}}\right)\right|\\
\leq&\left|\boldsymbol{u}_i^{\mathrm{T}}\left(\tilde{\mathbf{\Gamma}}^{-1/2}-\left(\mathbf{\Gamma}^*\right)^{-1/2}\right)\boldsymbol{A}^*\boldsymbol{v}_{\boldsymbol{\alpha}}\right|+\left|\boldsymbol{u}_i^{\mathrm{T}}\left(\tilde{\mathbf{\Gamma}}^{-1/2}-\left(\mathbf{\Gamma}^*\right)^{-1/2}\right)\boldsymbol{B}^*\boldsymbol{v}_{\boldsymbol{\beta}}\right|\\
\leq&\left(\boldsymbol{u}_i^{\mathrm{T}}\left(\tilde{\mathbf{\Gamma}}^{-1/2}-\left(\mathbf{\Gamma}^*\right)^{-1/2}\right)^2\boldsymbol{u}_i\right)^{1/2}\times\left(\left\|\boldsymbol{v}_{\boldsymbol{\alpha}}\right\|_2\left\|\boldsymbol{A}^*\right\|_2+\left\|\boldsymbol{v}_{\boldsymbol{\beta}}\right\|_2\left\|\boldsymbol{B}^*\right\|_2\right)\\
\leq&\left\|\tilde{\mathbf{\Gamma}}^{-1/2}-\left(\mathbf{\Gamma}^*\right)^{-1/2}\right\|_2\times\left(\left\|\boldsymbol{v}_{\boldsymbol{\alpha}}\right\|_2\left\|\boldsymbol{A}^*\right\|_2+\left\|\boldsymbol{v}_{\boldsymbol{\beta}}\right\|_2\left\|\boldsymbol{B}^*\right\|_2\right)\\
=&O_p(J_N^{3/2}N^{-1/2})\times O_p(J_N^{1/2}N^{-1/2})\\
=&O_p(J_N^2N^{-1}),
\end{aligned}$$

by Regularity Condition 8.A, 8.B and 8.D. Therefore, each element of I_2 is $O_p(J_N^2N^{-1/2})=o_p(1)$.

Consider the asymptotic order of I_3. We have

$$\begin{aligned}
&\left|\boldsymbol{u}_i^{\mathrm{T}}\left(\boldsymbol{\Gamma}^*\right)^{-1/2}\left(\left(\tilde{\boldsymbol{A}}-\boldsymbol{A}^*\right)\boldsymbol{v}_{\boldsymbol{\alpha}}+\left(\tilde{\boldsymbol{B}}-\boldsymbol{B}^*\right)\boldsymbol{v}_{\boldsymbol{\beta}}\right)\right| \\
\leq\ &\left(\boldsymbol{u}_i^{\mathrm{T}}\left(\boldsymbol{\Gamma}^*\right)^{-1}\boldsymbol{u}_i\right)^{1/2}\left\|\left(\tilde{\boldsymbol{A}}-\boldsymbol{A}^*\right)\boldsymbol{v}_{\boldsymbol{\alpha}}+\left(\tilde{\boldsymbol{B}}-\boldsymbol{B}^*\right)\boldsymbol{v}_{\boldsymbol{\beta}}\right\|_2 \\
=\ &O_p(J_N^2N^{-1})
\end{aligned}$$

by Lemma G.2 and Regularity Condition 8.C. Therefore, each element of I_3 is $O_p(J_N^2N^{-1/2})=o_p(1)$.

Lastly, consider the asymptotic order of I_4. We have

$$\begin{aligned}
&\left|\boldsymbol{u}_i^{\mathrm{T}}\left(\tilde{\boldsymbol{\Gamma}}^{-1/2}-\left(\boldsymbol{\Gamma}^*\right)^{-1/2}\right)\left(\left(\tilde{\boldsymbol{A}}-\boldsymbol{A}^*\right)\boldsymbol{v}_{\boldsymbol{\alpha}}+\left(\tilde{\boldsymbol{B}}-\boldsymbol{B}^*\right)\boldsymbol{v}_{\boldsymbol{\beta}}\right)\right| \\
\leq\ &\left(\boldsymbol{u}_i^{\mathrm{T}}\left(\tilde{\boldsymbol{\Gamma}}^{-1/2}-\left(\boldsymbol{\Gamma}^*\right)^{-1/2}\right)^2\boldsymbol{u}_i\right)^{1/2}\times\left\|\left(\tilde{\boldsymbol{A}}-\boldsymbol{A}^*\right)\boldsymbol{v}_{\boldsymbol{\alpha}}+\left(\tilde{\boldsymbol{B}}-\boldsymbol{B}^*\right)\boldsymbol{v}_{\boldsymbol{\beta}}\right\|_2 \\
=\ &O_p(J_N^4N^{-2}).
\end{aligned}$$

Therefore, each element of I_4 is also $o_p(1)$. □

Chapter 9

Concluding Remarks and Future Work

9.1 Concluding Remarks

In this thesis, the following major contributions are made:

1. A new reparameterization-approximation procedure for non-parametric mixture (or mixed-effects) models is proposed in Chapter 2.

2. Two new important properties of the generalized moment spaces, the positive representation and the gradient characterization, are derived in Chapter 3.

3. The generalized method of moments is proposed as a new estimation method for mixture models in Chapter 4;

4. The generalized method of moments is proposed as a new estimation method for mixed-effects models with univariate random effects in Chapter 5.

5. The method proposed in Chapter 5 is extended to a Poisson regression model with random intercept and slope in Chapter 6.

6. Some asymptotic results for the generalized method of moments for mixed-effects models with univariate random effects are established in Chapter 7.

7. The ensemble inference idea is used to construct an asymptotically χ^2 test statistic in Chapter 8.

We have, in this thesis, stated that

> "*Under the circumstances considered in this thesis, most of the difficulties in estimating and undertaking statistical inference with the mixing (or random effects) distribution Q could be prevented or solved by the methods proposed in this thesis, if the model $h_{\mathrm{Mix}}(s;Q)$, defined in Equation (2.1), can be reparameterized in the generalized moments of Q.*"

In the rest of this section, we revisit these difficulties, which have been discussed in Chapter 1, and discuss how the methods proposed in this thesis overcome them.

9.1.1 Identifiability

Let $h(s;\boldsymbol{m})$ be the model obtained through the reparameterization-approximation procedure proposed in Chapter 2, where $\boldsymbol{m} \in \mathbb{R}^{J_N}$ are the generalized moments of Q and J_N is an integer that diverges with the sample size N; see $f_{\mathrm{spec}}(x;\boldsymbol{m})$ in Chapter 4 and $\boldsymbol{U}_n(\boldsymbol{\beta},\boldsymbol{\alpha})$ in Chapter 5 for examples.

Note that, while a probability measure can uniquely determine its generalized moments, the converse is not true. In other words, the condition that $h(s;\boldsymbol{m})$ is identifiable by the generalized moments $\boldsymbol{m} \in \mathbb{R}^{J_N}$ is a weaker condition comparing to that $h_{\mathrm{Mix}}(s;Q)$ is identifiable by its mixing distribution Q.

Moreover, because the model $h(s;\boldsymbol{m})$ are constructed in an embedding affine space, the generalized moment vectors $\boldsymbol{m}$ are always linear in $h(s;\boldsymbol{m})$; see $f_{\mathrm{spec}}(x;\boldsymbol{m})$ and $\boldsymbol{U}_n(\boldsymbol{\beta},\boldsymbol{\alpha})$ for examples. Therefore, the identifiability of $h(s;\boldsymbol{m})$ can be easily shown by the linearly independent results in linear algebra.

9.1.2 Determining the Number of Generalized Moments

Using $h(s;\boldsymbol{m})$ may introduce extra bias to the estimators from modelling. However, the asymptotic orders of the approximation residuals could be characterized by using J_N and the smoothness of the kernel function $h(s;\theta)$ as a function of $\theta \in \Theta$; see Corollary 2.3.1, 2.4.1 and 2.4.2. In other words, J_N itself can be chosen to control the asymptotic orders of the approximation residuals.

On the other hand, the asymptotic orders of the approximation residuals need to be determined for inference purposes. For example, to compute the convergence rate of the GMM estimators in Chapter 7, it is only required that, for each n,

$$|\boldsymbol{U}_n(\boldsymbol{\beta}^*, Q^*) - \boldsymbol{U}_n(\boldsymbol{\beta}^*, \boldsymbol{\alpha}_N^*)| = o(J_N^{1/2} N^{-1/2});$$

see the proof of Theorem 7.3.1. However, the above condition is not sufficient enough to derive the asymptotic normality results in the GMM; see Theorem 7.5.1. Instead, Regularity Condition 7.B is needed to ensure that, for each n,

$$\sqrt{N}\,(\boldsymbol{U}_n(\boldsymbol{\beta}^*, Q^*) - \boldsymbol{U}_n(\boldsymbol{\beta}^*, \boldsymbol{\alpha}_N^*)) = o(J_N N^{-1/2}),$$

as $J_N N^{-1/2} = o(1)$.

In sum, given the sample size N and the smoothness of $h(s;\theta)$, the J_N could be chosen that not only controls the approximation residuals but also satisfies the purpose of statistical inferences.

9.1.3 Dimension of the Parameter Space

When mixing (or random effects) distributions are non-parametric, in the previous literature [Lindsay, 1995] and [Sutradhar and Godambe, 1997], models are always embedded in a space whose dimension diverges with N at rate $O(N)$. For example, in the NPMLE [Lindsay, 1995], the likelihood functions are embedded in a space whose dimension is a half of the distinct number of observed sample, which is $O(N)$ when the sample space is not discrete and finite. Another example is predicating the unobserved random effects in a mixed-effects model. Because the random effects depends on each individual, the number of the random effects to be predicated is the sample size N; see the UMM in [Sutradhar and Godambe, 1997]. Therefore, there may not be enough information from data for fitting models and making inferences with out further assumptions.

Reparameterized in the generalized moments, the model $h(s;\boldsymbol{m})$ is embedded in a space whose dimension is J_N. Although J_N also diverges with the sample size N, it is usually at a much lower rate. For examples, $J_N^{(2r+2)}N^{-1} = O(1)$ in Theorem 4.5.1 and $J_N^2 N^{-1} = o(1)$ in Theorem 7.3.1, where r is an arbitrary positive integer. The divergence of J_N may slow the convergence rate in the GMM; see Theorem 4.5.1 and 7.3.1 for examples. However, the lower order of J_N ensures that the parameters in $h(s;\boldsymbol{m})$ can be estimated consistently, under mild regularity conditions; see Theorem 4.5.1 and 7.3.1. Furthermore, the lower order of J_N also allow us to use extra information for statistical inference. The test statistics ζ in the ensemble inference asymptotically follows a $\chi^2_{pK_N - J_N}$ distribution, where $pK_N > J_N$. If J_N is small, ζ could have larger degrees of freedom and thus more power to reject the null hypothesis.

9.1.4 Geometric Properties of the Parameter Space

It is always not easy to estimate and make inference with a finite mixture model. One of the reasons is the complexity of the parameter space of a finite mixture model. As discussed in Section 1.2, the parameter space may include singularities and boundaries.

Comparing to the parameter space of a finite mixture model, the generalized moment space, as the parameter space of $h(s;\boldsymbol{m})$, is much better behaved. Firstly, with the compactness assumption on the set of the mixing parameter Θ, the generalized moment space is compact. Secondly, the generalized moment spaces have boundaries, but the geometric properties on the boundaries are well studied; see the positive representation and the gradient characterization in Chapter 3. Lastly but most importantly, the generalized moment space is a convex set. Therefore, by choosing a convex function of $\boldsymbol{m}$ as an objective function to minimize over the generalized moment space, we can construct the estimators of $\boldsymbol{m}$ which are the solutions of convex optimization problems; see the optimization problem (4.3) and (5.11) for examples. Moreover, the computational speeds for the proposed estimators are stable and fast by using the gradient-based computational algorithms; see Algorithm 4.1 and 5.1.

As we have seen in Theorem 7.5.1, the boundaries of the generalized method of moments still affect the asymptotic normality in the GMM. However, the boundary issue is potentially solved by using the ensemble idea at the cost of losing powers to reject the null hypothesis; see Chapter 8.

9.2 Future Work

There are many possible future research directions given in this thesis. In this section, we discuss the following two.

9.2.1 From Univariate to Multivariate

Most of the models considered in this thesis have a univariate mixing parameter (or random effects). The mixture (or mixed-effects) models with a multivariate mixing parameter (or random effects) have much wider applications, because they are more flexible. One of the examples is the Poisson regression model with random intercept and slope, which has been seen in Chapter 6. Also see [Karlis and Meligkotsidou, 2007] for finite mixture of multivariate Poisson distributions and [Chen and Tan, 2009] for finite mixture of multivariate normal distributions. We have discussed some possible extensions for mixed-effects models; see Section 2.4.3 and Chapter 6. However, there requires much more future work.

Firstly, we need to define the generalized moments of multivariate random variables. Because the Chebyshev system for multivariate functions is not well-defined, the generalized moments of multivariate random variables will be defined in a wider class of system; see the tensor product basis given in Section 2.4.3. Next, we need to investigate the geometric properties of the generalized moment space of multivariate random variables. The new generalized moment space is convex and has two important geometric properties, the positive representation and the gradient characterization, can be preserved, because both are directly from the convexity. However, it can be expected that both of the geometric properties become more complex, because of the increase in the dimension.

When the multivariate mixture models are considered, it is challenging to construct the generalized moment conditions in Definition 4.2. It is possible that the generalized moment conditions may not exist without extra conditions.

It might be straightforward to define the GMM for mixed-effects models with multivariate random effects based on Definition 5.4.1. However, the challenges exist in computing the estimators. The major reason is that the geometric properties of generalized moment space of multivariate random variables is unclear. Moreover, it could be even more difficult to predicate multivariate random effects.

9.2.2 The Families of Weighting Matrices

In the literature of the GMM (or GEE), choosing appropriate weighting (or working correlation) matrices has attracted much interest; see [Liang and Zeger, 1986], [Mátyás, 1999] and [Thall and Vail, 1990]. Under the regularity conditions listed in [Mátyás, 1999, Section 1.3], the inverse of the covariance matrix of the generalized moment conditions is optimal in the sense that the MSE of the resulting GMM estimator is minimized. Although the regularity conditions in [Mátyás, 1999] fail in this thesis, we still observe the efficiency gain in the GMM estimators when the weighting matrices are the inverse of the covariance matrix of the generalized moment conditions; see Section 4.6, 5.7 and 6.4.

It turns out that, weighting (or working correlation) matrices can be carefully designed either for efficiency or robustness, when our model is correctly specified. For example, the weighting matrix is design for the robustness of the GMM for mixture models in Section 4.6. Another example is the modelling of correlation matrix in the GMM for mixed-effects model; see Section 5.6. From the simulation studies in Section 6.4, we see that the GMM estimator could gain huge increase in the efficiency when the working correlation matrix are correctly specified.

Therefore, it becomes important to model the correlation structures of the generalized moment conditions (or estimating functions); see [Liang and Zeger, 1986] and [Thall and Vail, 1990]. However, modelling and validating the correlation structures require further studies. Using different weighting (or working correlation) matrices is one possible alternative way; see the QIF in [Qu et al., 2000] and the ensemble inference in Chapter 8. However, the way that a family of weighting (or correlation) matrices systematically affect the estimation or statistical inference still remains as a mystery.

Bibliography

M. Aitkin. A general maximum likelihood analysis of variance components in generalized linear models. *Biometrics*, 55(1):117–128, 1999.

A. Antoniou and W. Lu. *Practical Optimization: Algorithms and Engineering Applications*. Springer-Verlag, New York, 2007.

O.E. Barndorff-Nielsen and B. Jørgensen. Some parametric models on the simplex. *Journal of Multivariate Analysis*, 39(1):106–116, 1991.

H. Bauer. *Probability theory, volume 23 of de Gruyter Studies in Mathematics*. Walter de Gruyter & Co., Berlin, 1996.

D. Böhning. A review of reliable maximum likelihood algorithms for semiparametric mixture models. *Journal of Statistical Planning and Inference*, 47(1-2):5–28, 1995.

D. Böhning and V. Patilea. Asymptotic normality in mixtures of power series distributions. *Scandinavian Journal of Statistics*, 32(1):115–131, 2005.

D. Böhning, P. Schlattmann, and B.G. Lindsay. Computer-assisted analysis of mixtures (C.A.MAN): statistical algorithms. *Biometrics*, 48(1):283–303, 1992.

D. Böhning, E. Dietz, R. Schaub, P. Schlattmann, and B.G. Lindsay. The distribution of the likelihood ratio for mixtures of densities from the one-parameter exponential family. *Annals of the Institute of Statistical Mathematics*, 46(2):373–388, 1994.

J.P. Boyd. Asymptotic coefficients of hermite function series. *Journal of Computational Physics*, 54(3):382–410, 1984.

J.P. Boyd. *Chebyshev and Fourier spectral methods.* Dover Publications, Toronto, 2001.

L. Breiman. Random forests. *Machine learning*, 45(1):5–32, 2001.

N.E. Breslow and D.G. Clayton. Approximate inference in generalized linear mixed models. *Journal of the American Statistical Association*, 88(421):9–25, 1993.

G. Celeux, M. Hurn, and C.P. Robert. Computational and inferential difficulties with mixture posterior distributions. *Journal of the American Statistical Association*, 95(451), 2000.

H. Chen, J. Chen, and J.D. Kalbfleisch. Testing for a finite mixture model with two components. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 66(1):95–115, 2004.

J. Chen. Optimal rate of convergence for finite mixture models. *The Annals of Statistics*, 23(1):221–233, 1995.

J. Chen. Penalized likelihood-ratio test for finite mixture models with multinomial observations. *Canadian Journal of Statistics*, 26(4):583–599, 1998.

J. Chen and J.D. Kalbfleisch. Penalized minimum-distance estimates in finite mixture models. *Canadian Journal of Statistics*, 24(2):167–175, 1996.

J. Chen and A. Khalili. Order selection in finite mixture models. *Journal of American Statistical Association*, 103(484):1674–1683, 2008.

J. Chen and X. Tan. Inference for multivariate normal mixtures. *Journal of Multivariate Analysis*, 100(7):1367–1383, 2009.

S. Chen, L. Peng, and Y. Qin. Effects of data dimension on empirical likelihood. *Biometrika*, 96(3):711–722, 2009.

X. Chen and K. Zhou. On the probabilistic characterization of model uncertainty and robustness. In *Decision and Control, 1997., Proceedings of the 36th IEEE Conference on*, volume 4, pages 3816–3821. IEEE, 1997.

R.C.H. Cheng and L. Traylor. Non-regular maximum likelihood problems. *Journal of the Royal Statistical Society. Series B (Methodological)*, 57(1):3–44, 1995.

H. Chernoff. On the distribution of the likelihood ratio. *The Annals of Mathematical Statistics*, 25(3):573–578, 1954.

A. Cutler and O.I. Cordero-Braña. Minimum hellinger distance estimation for finite mixture models. *Journal of the American Statistical Association*, 91(436):1716–1723, 1996.

L. Debnath and P. Mikusiński. *Introduction to Hilbert spaces with applications.* Academic Press, San Diego, 1999.

F.R. Deutsch. *Best approximation in inner product spaces.* Springer, New York, 2001.

J. Diebolt and C.P. Robert. Estimation of finite mixture distributions through Bayesian sampling. *Journal of the Royal Statistical Society. Series B (Methodological)*, 56(2):363–375, 1994.

P. Diggle. *Analysis of Longitudinal Data.* Oxford University Press, Oxford, 2002.

M.D. Escobar and M. West. Bayesian density estimation and inference using mixtures. *Journal of the american statistical association*, 90(430):577–588, 1995.

B.S. Everitt, S. Landau, and M. Leese. *Cluster analysis.* John Wiley & Sons, Chichester, 5th edition, 2011.

W. Feller. On a general class of "contagious" distributions. *The Annals of mathematical statistics*, 14(4):389–400, 1943.

G.M. Fitzmaurice, N.M. Laird, and J.H. Ware. *Applied longitudinal analysis.* John Wiley & Sons, New York, 2012.

Y. Freund and R.E. Schapire. A decision-theoretic generalization of on-line learning and an application to boosting. *Journal of Computer and System Sciences*, 55(1): 119–139, 1997.

W.D. Furman and B.G. Lindsay. Measuring the relative effectiveness of moment estimators as starting values in maximizing likelihoods. *Computational Statistics & Data Analysis*, 17(5):493–507, 1994.

A.R. Gallant and D.W. Nychka. Semi-nonparametric maximum likelihood estimation. *Econometrica: Journal of the Econometric Society*, 55(2):363–390, 1987.

L. Gan and J. Jiang. A test for global maximum. *Journal of the American Statistical Association*, 94(447):847–854, 1999.

A. Gelman, J.B. Carlin, H.S. Stern, and D.B. Rubin. *Bayesian data analysis.* CRC Press, London, 2014.

C.W. Ha. Eigenvalues of differentiable positive definite kernels. *SIAM Journal on Mathematical Analysis*, 17(2):415–419, 1986.

A.R. Hall. *Generalized method of moments.* Oxford University Press, Oxford, 2005.

F.R. Hampel. The influence curve and its role in robust estimation. *Journal of the American Statistical Association*, 69(346):383–393, 1974.

L.P. Hansen. Large sample properties of generalized method of moments estimators. *Econometrica*, 50(4):1029–1054, 1982.

P.J. Heagerty and B.F. Kurland. Misspecified maximum likelihood estimates and generalised linear mixed models. *Biometrika*, 88(4):973–985, 2001.

L. Horváth and P. Kokoszka. *Inference for functional data with applications.* Springer, New York, 2012.

L.F. James, C.E. Priebe, and D.J. Marchette. Consistent estimation of mixture complexity. *Annals of Statistics*, 29(5):1281–1296, 2001.

A. Jasra, C.C. Holmes, and D.A. Stephens. Markov Chain Monte Carlo methods and the label switching problem in bayesian mixture modeling. *Statistical Science*, 20 (1):50–67, 2005.

J. Jiang. Conditional inference about generalized linear mixed models. *The Annals of Statistics*, 27(6):1974–2007, 1999.

S. Karlin and W.J. Studden. *Tchebycheff Systems: with Applications in Analysis and Statistics*. Interscience Publishers, New York, 1966.

D. Karlis and L. Meligkotsidou. Finite mixtures of multivariate poisson distributions with application. *Journal of Statistical Planning and Inference*, 137(6):1942–1960, 2007.

C. Keribin. Consistent estimation of the order of mixture models. *Sankhyā: The Indian Journal of Statistics, Series A*, 62(1):49–66, 2000.

J. Kiefer and J. Wolfowitz. Consistency of the maximum likelihood estimator in the presence of infinitely many incidental parameters. *The Annals of Mathematical Statistics*, 27(4):887–906, 1956.

N. Laird. Nonparametric maximum likelihood estimation of a mixing distribution. *Journal of the American Statistical Association*, 73(364):805–811, 1978.

C. Lam and J. Fan. Profile-kernel likelihood inference with diverging number of parameters. *Annals of statistics*, 36(5):2232–2260, 2008.

D. Lambert and L. Tierney. Asymptotic properties of maximum likelihood estimates in the mixed poisson model. *The Annals of Statistics*, 12(4):1388–1399, 1984.

N.N. Lebedev. *Special functions and their applications*. Dover Publications, New York, 1972.

Y. Lee and J.A. Nelder. Conditional and marginal models: another view. *Statistical Science*, 19(2):219–238, 2004.

B.G. Leroux. Consistent estimation of a mixing distribution. *The Annals of Statistics*, 20(3):1350–1360, 1992.

P. Li and J. Chen. Testing the order of a finite mixture. *Journal of the American Statistical Association*, 105(491):1084–1092, 2010.

P. Li, J. Chen, and P. Marriott. Non-finite Fisher information and homogeneity: an EM approach. *Biometrika*, 96(2):411–426, 2009.

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

X. Lin and N.E. Breslow. Bias correction in generalized linear mixed models with multiple components of dispersion. *Journal of the American Statistical Association*, 91(435):1007–1016, 1996.

B.G. Lindsay. Nuisance parameters, mixture models, and the efficiency of partial likelihood estimators. *Philosophical Transactions of the Royal Society of London. Series A, Mathematical and Physical Sciences*, 296(1427):639–662, 1980.

B.G. Lindsay. The geometry of mixture likelihoods: a general theory. *The Annals of Statistics*, 11(1):86–94, 1983.

B.G. Lindsay. On the determinants of moment matrices. *The Annals of Statistics*, 17 (2):711–721, 1989a.

B.G. Lindsay. Moment matrices: applications in mixtures. *The Annals of Statistics*, 17(2):722–740, 1989b.

B.G. Lindsay. Mixture models: theory, geometry and applications. In *NSF-CBMS regional conference series in probability and statistics*, volume 5, Hayward, 1995. Institute of Mathematical Statistics.

B.G. Lindsay and K. Roeder. Uniqueness of estimation and identifiability in mixture models. *Canadian Journal of Statistics*, 21(2):139–147, 1993.

P. Marriott. On the local geometry of mixture models. *Biometrika*, 89(1):77–93, 2002.

P. Marriott. Extending local mixture models. *Annals of the Institute of Statistical Mathematics*, 59(1):95–110, 2007.

J.C. Mason and D.C. Handscomb. *Chebyshev polynomials.* CRC Press, Boca Raton, 2002.

L. Mátyás. *Generalized Method of Moments Estimation.* Cambridge University Press, Cambridge, 1999.

C.E. McCulloch and J.M. Neuhaus. *Generalized Linear Mixed Models.* John Wiley & Sons, New York, 2005.

G.J. McLachlan and K.E. Basford. *Mixture models: Inference and applications to clustering.* M. Dekker, New York, 1988.

G.J. McLachlan and D. Peel. *Finite Mixture Models.* John Wiley & Sons, New York, 2000.

K. Mengersen, C. Robert, and M. Titterington. *Mixtures: Estimation and Applications.* John Wiley & Sons, Chichester, 2011.

C.N. Morris. Natural exponential families with quadratic variance functions. *The Annals of Statistics*, 10(1):65–80, 1982.

C.N. Morris. Natural exponential families with quadratic variance functions: statistical theory. *The Annals of Statistics*, 11(2):515–529, 1983.

J.M. Neuhaus, W.W. Hauck, and J.D. Kalbfleisch. The effects of mixture distribution misspecification when fitting mixed-effects logistic models. *Biometrika*, 79(4):755–762, 1992.

J. Neyman and E.L. Scott. Consistent estimates based on partially consistent observations. *Econometrica*, 16(1):1–32, 1948.

S.A. Orszag and C.M. Bender. *Advanced mathematical methods for scientists and engineers.* Springer, New York, 1999.

J.M. Ortega and W.C. Rheinboldt. *Iterative solution of nonlinear equations in several variables.* Academic Press, San Diego, 1970.

K. Pearson. Mathematical Contributions to the Theory of Evolution. V. On the Reconstruction of the Stature of Prehistoric Races.[Abstract]. *Proceedings of the Royal Society of London*, 63:417–420, 1898.

A. Pinkus. Spectral properties of totally positive kernels and matrices. *Total positivity and its applications*, 359:1–35, 1996.

Z. Qiu, X.K. Song, and M. Tan. Simplex mixed-effects models for longitudinal proportional data. *Scandinavian Journal of Statistics*, 35(4):577–596, 2008.

A. Qu and X.K. Song. Assessing robustness of generalised estimating equations and quadratic inference functions. *Biometrika*, 91(2):447–459, 2004.

A. Qu, B.G. Lindsay, and B. Li. Improving generalised estimating equations using quadratic inference functions. *Biometrika*, 87(4):823–836, 2000.

S. Ray and B.G. Lindsay. Model selection in high dimensions: a quadratic-risk-based approach. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 70(1):95–118, 2007.

J.B. Reade. Eigenvalues of positive definite kernels. *SIAM Journal on Mathematical Analysis*, 14(1):152–157, 1983.

R.A. Redner and H.F. Walker. Mixture densities, maximum likelihood and the em algorithm. *SIAM review*, 26(2):195–239, 1984.

S. Richardson and P.J. Green. On bayesian analysis of mixtures with an unknown number of components (with discussion). *Journal of the Royal Statistical Society: series B (statistical methodology)*, 59(4):731–792, 1997.

C.P. Robert and G. Casella. *Monte Carlo Statistical Methods.* Springer, Berlin Heidelberg, 2004.

T. Ryden. Estimating the order of hidden markov models. *Statistics: A Journal of Theoretical and Applied Statistics*, 26(4):345–354, 1995.

P. Schlattmann. *Medical Applications of Finite Mixture Models.* Springer, Berlin, 2009.

A. Shapiro. Asymptotic distribution of test statistics in the analysis of moment structures under inequality constraints. *Biometrika*, 72(1):133–144, 1985.

M.J. Silvapulle and P.K. Sen. *Constrained statistical inference: Order, inequality, and shape constraints*. John Wiley & Sons, New York, 2005.

C.G. Small. *Expansions and asymptotics for statistics*. CRC Press, Boca Raton, 2010.

X.K. Song. *Correlated data analysis: modeling, analytics, and applications*. Springer Science & Business Media, New York, 2007.

X.K. Song and M. Tan. Marginal models for longitudinal continuous proportional data. *Biometrics*, 56(2):496–502, 2000.

M. Sperrin, T. Jaki, and E. Wit. Probabilistic relabelling strategies for the label switching problem in bayesian mixture models. *Statistics and Computing*, 20(3): 357–366, 2010.

M. Stephens. Dealing with label switching in mixture models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 62(4):795–809, 2000.

B.C. Sutradhar and V.P. Godambe. On estimating function approach in the generalized linear mixed model. *Lecture Notes-Monograph Series*, pages 193–213, 1997.

B.C. Sutradhar and R.P. Rao. On joint estimation of regression and overdispersion parameters in generalized linear models for longitudinal data. *Journal of multivariate analysis*, 56(1):90–119, 1996.

G.M. Tallis. The identifiability of mixtures of distributions. *Journal of Applied Probability*, 6(2):389–398, 1969.

G.M. Tallis and P. Chesson. Identifiability of mixtures. *Journal of the Australian Mathematical Society*, 32(03):339–348, 1982.

H. Teicher. On the mixture of distributions. *The Annals of Mathematical Statistics*, 31(1):55–73, 1960.

H. Teicher. Identifiability of mixtures. *The Annals of Mathematical Statistics*, 32(1): 244–248, 1961.

H. Teicher. Identifiability of finite mixtures. *The Annals of Mathematical Statistics*, 34(4):1265–1269, 1963.

P.F. Thall and S.C. Vail. Some covariance models for longitudinal count data with overdispersion. *Biometrics*, 46(3):657–671, 1990.

D.M. Titterington, A.F.M. Smith, and U.E. Makov. *Statistical Analysis of Finite Mixture Distributions*. John Wiley & Sons, Chichester, 1985.

J.W. Tukey. A survey of sampling from contaminated distributions. In *Contributions to probability and statistics*, volume 2, pages 448–485, 1960.

S. Van De Geer. Asymptotic normality in mixture models. *ESAIM: Probability and Statistics*, 1:17–33, 1997.

E.F. Vonesh, H. Wang, L. Nie, and D. Majumdar. Conditional second-order generalized estimating equations for generalized linear and nonlinear mixed-effects models. *Journal of the American Statistical Association*, 97(457):271–283, 2002.

P.W. Vos. Geometry of f-divergence. *Annals of the Institute of Statistical Mathematics*, 43(3):515–537, 1991.

G.G. Walter and X. Shen. *Wavelets and other Orthogonal Systems*. CRC Press, Boca Raton, 2rd edition, 2001.

L. Wang. GEE analysis of clustered binary data with diverging number of covariates. *The Annals of Statistics*, 39(1):389–417, 2011.

P. Wang, G.F. Tsai, and A. Qu. Conditional inference functions for mixed-effects models with unspecified random-effects distribution. *Journal of the American Statistical Association*, 107(498):725–736, 2012.

Y. Wang. On fast computation of the non-parametric maximum likelihood estimate of a mixing distribution. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 69(2):185–198, 2007.

Y. Wang. Maximum likelihood computation for fitting semiparametric mixture models. *Statistics and Computing*, 20(1):75–86, 2010.

S.S. Wilks. The large-sample distribution of the likelihood ratio for testing composite hypotheses. *The Annals of Mathematical Statistics*, 9(1):60–62, 1938.

M.J. Woo and T.N. Sriram. Robust estimation of mixture complexity. *Journal of the American Statistical Association*, 101(476):1475–1486, 2006.

G.R. Wood. Binomial mixtures: geometric estimation of the mixing distribution. *The Annals of Statistics*, 27(5):1706–1721, 1999.

H. Wu and J.T. Zhang. *Nonparametric regression methods for longitudinal data analysis: mixed-effects modeling approaches*. John Wiley & Sons, New York, 2006.

S.J. Yakowitz and J.D. Spragins. On the identifiability of finite mixtures. *The Annals of Mathematical Statistics*, 39(1):209–214, 1968.

S.L. Zeger, K.Y. Liang, and P.S. Albert. Models for longitudinal data: a generalized estimating equation approach. *Biometrics*, 44(4):1049–1060, 1988.

J. Zhang. Divergence function, duality, and convex analysis. *Neural Computation*, 16 (1):159–195, 2004.

J. Zhang. Referential duality and representational duality on statistical manifolds. In *Proceedings of the Second International Symposium on Information Geometry and Its Applications, Tokyo*, pages 58–67, 2005.

J. Zhang. Nonparametric information geometry: From divergence function to referential-representational biduality on statistical manifolds. *Entropy*, 15(12): 5384–5418, 2013.

J. Zhang and P. Hasto. Statistical manifold as an affine space: A functional equation approach. *Journal of Mathematical Psychology*, 50(1):60–65, 2006.

M. Zhu. Kernels and ensembles. *The American Statistician*, 62(2), 2008.