

**GENERALIZED LINEAR MIXED MODEL FOR FINITE NORMAL  
MIXTURES WITH APPLICATION TO  
TENDON FIBRILOGENESIS DATA**

---

A Dissertation  
Submitted to  
the Temple University Graduate Board

---

in Partial Fulfillment  
of the Requirements for the Degree  
DOCTOR OF PHILOSOPHY

---

by  
Tingting Zhan  
May, 2012

Examining Committee Members:

Dr. Boris Iglewicz. Department of Statistics, Temple University

Dr. Inna Chervoneva. Division of Biostatistics, Dept. of Pharmacology and Experimental  
Therapeutics, Thomas Jefferson University

Dr. Zhigen Zhao. Department of Statistics, Temple University

Dr. Mary D. Sammel. Dept. of Biostatistics & Epidemiology, University of Pennsylvania

©

Tingting Zhan

May, 2012

All Rights Reserved

**Abstract**

GENERALIZED LINEAR MIXED MODEL FOR FINITE NORMAL MIXTURES  
WITH APPLICATION TO  
TENDON FIBRILOGENESIS DATA

Tingting Zhan

DOCTOR OF PHILOSOPHY

Temple University, May, 2012

Chair: Dr. Boris Iglewicz. Department of Statistics, Temple University

We propose the generalized linear mixed model for finite normal mixtures (GLMFM), as well as the estimation procedures for the GLMFM model, which are widely applicable to the hierarchical dataset with small number of individual units and multi-modal distributions at the lowest level of clustering. The modeling task is two-fold: (a) to model the lowest-level cluster as a finite mixtures of the normal distribution; and (b) to model the properly transformed mixture proportions, means and standard deviations of the lowest-level cluster as a linear hierarchical structure. We propose the robust generalized weighted likelihood estimators and the new cubic-inverse weight for the estimation of the finite mixture model (Zhan et al., 2011). We propose two robust methods for estimating the GLMFM model, which accommodate the contaminations on all clustering levels, the standard-two-stage approach (Chervoneva et al., 2011, co-authored) and a robust joint estimation. Our research was motivated by the data obtained from the tendon fibril experiment reported in Zhang et al. (2006). Our statistical methodology is quite general and has potential application in

a variety of relatively complex statistical modeling situations.

## Acknowledgements

This research was supported by a Merck Quantitative Graduate Fellowship.

# Contents

<b>Abstract</b>	<b>iv</b>
<b>Acknowledgement</b>	<b>vi</b>
<b>List of Figures</b>	<b>x</b>
<b>List of Tables</b>	<b>xi</b>
<b>1 Introduction</b>	<b>1</b>
<b>2 Modeling framework and literature review</b>	<b>7</b>
2.1 Modeling estimation framework . . . . .	8
2.2 Finite normal mixture model and robust density estimation . . . . .	12
2.3 Standard mixed models based on mean structure . . . . .	18
2.4 Best linear unbiased predictor procedure for linear mixed model . . . . .	20
2.5 Mixed models without a mean structure . . . . .	22
<b>3 The robust estimation of finite mixture model</b>	<b>25</b>
3.1 Generalized weighted likelihood estimators . . . . .	26
3.2 Influence function and asymptotic normality . . . . .	29
3.3 The new truncated cubic-inverse weight . . . . .	31

3.3.1	Selection of smoothing bandwidth $h$ . . . . .	32
3.3.2	Selection of cubic coefficient $c$ . . . . .	34
3.4	The iterative algorithm . . . . .	35
3.5	The simulation studies . . . . .	38
3.6	The tendon fibril data example . . . . .	39
3.7	Summary of Chapter 3 . . . . .	40
<b>4</b>	<b>Robust estimation of generalized linear mixed model for finite normal mixtures</b>	<b>42</b>
4.1	STS approach to the GLMFM model . . . . .	43
4.2	Joint estimation of the GLMFM model . . . . .	44
4.3	Robust estimation of covariance parameters . . . . .	48
4.4	Inference for fixed effects estimates . . . . .	50
4.5	Simulation studies . . . . .	51
4.6	Tendon fibril data example . . . . .	54
4.6.1	A simplified two-stage model and STS approach . . . . .	54
4.6.2	GLMFM model and robust joint estimation . . . . .	57
4.7	Summary of Chapter 4 . . . . .	60
<b>5</b>	<b>Discussion and future work</b>	<b>65</b>
	<b>Bibliography</b>	<b>66</b>
<b>A</b>	<b>Derivation and proof</b>	<b>82</b>
A.1	Finite exponential-family mixtures . . . . .	82
A.1.1	General logit link for multinomial distribution . . . . .	83

A.1.2	Augmented distribution of finite mixture model . . . . .	83
A.1.3	EM algorithm . . . . .	85
A.1.4	Weighted score functions . . . . .	87
A.1.5	Fisher information and asymptotic normality . . . . .	89
A.2	Influence function of GWLE . . . . .	90

# List of Figures

1.1	Selected microscopic fields of the tendon fibril data . . . . .	3
2.1	Divergence G functions . . . . .	17
3.1	Weight functions . . . . .	29
3.2	Selection of $h$ . . . . .	34
3.3	Simulation scenarios . . . . .	38
3.4	Simulation results: box-plot of estimation errors . . . . .	39
3.5	The robust estimates of selected microscopic fields of the tendon fibril data	41
4.1	Joint estimation simulation results: parameter estimates under clean scenario	55
4.2	Joint estimation simulation results: parameter estimates under contaminated scenario . . . . .	56
4.3	Selected P3M microscopic fields: methods I vs. II . . . . .	58
4.4	Selected P3M microscopic fields: methods III vs. IV . . . . .	59

## List of Tables

- 4.1 The estimates of the population average of the normal mixture parameters for wild and mutant mice (in independent mixture parameters logit  $\pi$ 's,  $\mu$ 's and  $\log \sigma$ 's). The standard errors of the estimates are included in parenthesis. 62
- 4.2 The estimates of the population average of the normal mixture parameters for wild and mutant mice (in plain mixture parameters  $\pi$ 's,  $\mu$ 's and  $\sigma$ 's). The standard errors of the estimates are included in parenthesis. . . . . 63
- 4.3 The estimates of the genotype differences (using wild as the reference level) in each of the mixture parameters. The standard errors of the estimates are included in parenthesis. The p-value of the Wald test are included in square parenthesis. . . . . 64

# Chapter 1

## Introduction

This work is motivated by the tendon fibril genesis experiment first reported in Zhang et al. (2006), which focuses on the functional roles of a genetic mutation called "decorin-deficiency" in the development of the structural and functional properties of collagen tendon in newborn mice. We will use these data to motivate and illustrate the development and explanation of new statistical methodologies that can be useful in parameter estimation and analysis of such complex data sets.

The tendons are the tissues that connect the muscle to the bone for transmitting and withstanding the tissue loads. The tendons are composed primarily of aligned columns of fibroblasts, collagen fibrils grouped as fibers and an interfibrillar matrix. The tensile strength depends on the high collagen fiber content. The tissue-specific fibril genesis and extracellular matrix assembly is required for the development, growth and repair of the tendons, thus the studies of collagen fibril genesis afford better understanding of tendon development, growth, and maturation as well as the pathobiological changes associated with aging or injury and regeneration after wounding or surgical intervention. During the tendon development, collagen fibrils are initially assembled as immature fibril intermediates,

followed by linear and lateral growth of mature fibrils from the preformed intermediates.

Zhang et al. (2006) studied the mechanisms of the tendon extracellular matrix assembly that allow for the independent regulation of initial fibril assembly as well as growth in length and diameter by analyzing fibril diameters, with particular interest in the roles of decorin, lumican and fibromodulin in regulation of fibril genesis and the roles of type XIV collagen in regulation of growth. Decorin deficient mice demonstrate altered fibril structure and mechanical function in mature skin and tail tendons, with abnormal, irregularly contoured fibrils. The biologists are interested in further elucidating the developmental roles of decorin.

In the experiment reported in Zhang et al. (2006), the maturation of collagen fibril intermediates is disrupted using different genetic deficiencies, in order to illustrate the regulatory mechanisms. The primary covariates concerned are animal developmental stage (age as measured in postnatal days) and genotype. The primary response is the collagen fibril diameter measurements taken from experiment animals. In this study, 4-7 animals from each of the two mice strains, decorin deficient (DD) and wild type (WT), are sacrificed at different postnatal ages. Multiple fibril cross-sections are made from each animal, photographed under a transmission electron microscope and 5-6 negatives per animal are selected randomly. The fibril diameters in microscopic fields of defined size are digitized and measured (in unit of nm) by an image analysis system. Typically, each microscopic field of immature fibrils (post natal 1 month or younger) have around 200-500 diameter measurements, while mature fibrils have around 50-150 measurements per field. Figure 1.1(a), in Appendix ??, presents the scanned image of microscopic fields at different post natal ages. Figure 1.1(b) presents a few histograms of the fibril diameter measurement distributions on selected microscopic fields taken from postnatal 3-month animals. The Gaussian kernel density estimates of these distributions are plotted along with each histogram.

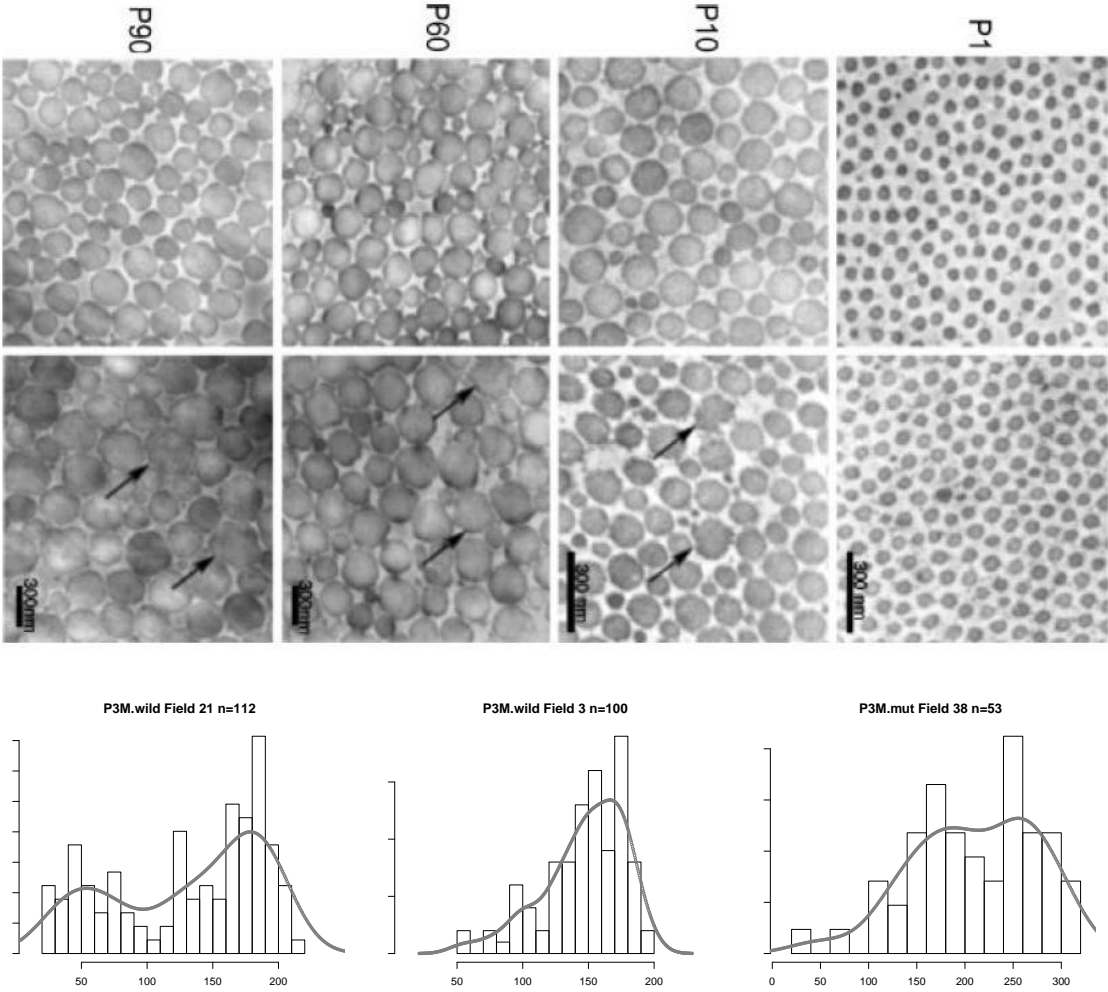


Figure 1.1: Top: scans of microscopic fields at different postnatal ages 1 day, 10 days, 2 months, 3 months; wild type (upper row) and decorin deficiency (lower row). Bottom: Selected microscopic fields of the fibril diameter measures from postnatal 3-month mice.

The data from the tendon fibril experiment serves an example of the various biological nano-structures of the hierarchically clustered data with large number of observations and multi-modal distributions at the lowest level of clustering. It is interesting to investigate the following two topics for the tendon fibril data,

- the mechanisms regulating the tendon fibril development, by decomposing the conditional distributions of the fibril diameters into sub-populations with different characteristics and functional roles, and
- the dependence of those sub-populations within each cluster

In Chapter 2, we propose the *generalized linear mixed model for finite normal mixtures* (GLMFM), which assumes a linear hierarchical relationship after proper transformation of the mixture proportions, means and standard deviations of the lowest-level sub-populations. We also provide a brief review of the related statistical methods in literature.

In Chapter 3 (published in Zhan et al., 2011), we investigate the robust estimation of the finite normal mixture model, in which the mixing proportions, location and spread of each sub-populations are of the primary interest. Zhang et al. (2006) suggested that the field-specific distributions be modeled as finite mixtures of the normal distributions: two-component for 4-day-old animals and three-component for 2 month or older. The fixed-number-of-components assumption is appropriate from the biological point of view, which is attributed to a common underlying distribution of the fibril diameters associated with each developmental age. The biological hypothesis is that the first component ( $\sim 35\text{nm}$ ), representing the initially assembled immature protofibrils or fibril intermediates, will remain the same in thickness during the maturation process, while the second component ( $\sim 50\text{nm}$ ) corresponds to the subpopulation of maturing fibrils, involving linear and lateral growth of

performed protofibrils. Even thicker fibrils ( $> 100\text{nm}$ ) emerge at later ages, representing the subpopulation of mature fibrils. On the other hand, the microscopic-field-specific distributions of the fibril diameters are often contaminated, shown as the heavy tails in Figure 1.1. These contaminations are either due to the genetic alterations (e.g. abnormally large fused fibrils) or the cross sections through the tapered ends of fibrils (smaller than diameter in the main cylindrical part). We develop robust estimation methods for the mixture model of exponential family distributions with fixed number of components, which is less sensitive to the contaminations and accommodate to the situation when the mixture components may exhibit moderate departures from the chosen model. We propose the family of generalized weighted likelihood estimators, which include many minimum divergence density estimators in literature. We also propose a new member of this family, named the *cubic-inverse weight*, with advantages in handling more complex contaminations. The proposed cubic-inverse-weighted likelihood estimator demonstrates desirable asymptotic properties and proves a good estimation method for the tendon fibril data.

In Chapter 4, we investigate the robust estimation of the GLMFM model. The clustered structure of the tendon fibril data enables the study of the functional roles of decorin on the subpopulation parameters of the field-specific distribution of the fibril diameter measurements, adjusted for the animal and field random effect. The between-animal and between-field variabilities are considered coming from the small variation of actually postnatal ages and the sampling variability due to shifts of cross-sections within tendons, respectively. The standard, i.e. the linear, generalized linear and nonlinear, and semi-parametric, mixed effects models in literature focus on analysis of mean structure as a function of the covariates, without identifying or addressing the properties such as mixing of sub-populations. The GLMFM model is an extension of the generalized linear mixed model to the conditional

finite normal mixture distribution. It is also an extension of the so-called finite mixture regression model to include all subpopulation parameters into the regression model. We propose a robust joint estimation for estimating the GLMFM model, which accommodates the contaminations on both animal-level and field-level.

In Chapter 5, we present a brief summary with discussions and outlines of the future work.

## Chapter 2

# Modeling framework and literature review

In this chapter, we formalize the statistical framework for modeling the tendon fibril data of Zhang et al. (2006) and provide a brief review of the related models as well as the non-robust and robust estimation methods in literature. In section 2.1, we introduce the generalized linear mixed model for finite normal mixtures (GLMFM) for the hierarchical data with multi-modal conditional distributions. We outline the proposed robust joint estimation methods for the GLMFM model. In section 2.2, we review the finite normal mixture model with fixed number of components, which is suitable for the skewed and multi-modal conditional distributions of the tendon fibril data. We concentrate on the corresponding density estimation techniques, both the maximum likelihood and the robust minimum divergence estimation. We also include a detailed review of the robust minimum divergence estimators. In section 2.3, we review the standard linear and nonlinear mixed models, as well as the non-robust and robust estimation methods. In section 2.4, we review the best linear unbiased predictor (BLUP) procedure for the linear mixed model. In section

2.5, we review the limited research in the family of non-mean-structure mixed model, to which the proposed GLMFM model belongs.

## 2.1 Modeling estimation framework

Consider the tendon fibril data (Zhang et al., 2006), let  $y_{ijk}$  be the  $k$ th ( $k = 1, \dots, n_{ij}$ ) fibril diameter measurement taken from the  $j$ th ( $j = 1, \dots, n_i$ ) microscopic field of the  $i$ th ( $i = 1, \dots, M$ ) animal. The sample sizes, in the postnatal three-month (P3M) data for example, the  $n_{ij}$ 's are around 150 fibrils, the majority of  $n_i$ 's range from 5 to 11 fields and  $M = 11$  animals. The covariates associated with animal  $i$  are the animal body weight  $m_i$ , the genotype  $g_i$  (0 for wild type and 1 for mutant) and the postnatal age  $t_i$ . No covariate is associated with the microscopic fields.

The fibril diameter measurements on field  $j$  of animal  $i$ ,  $\mathbf{y}_{ij} = \{y_{ijk} : k = 1, \dots, n_{ij}\}$ , are assumed to follow a  $S$ -component normal mixture distribution, where  $S$  is a pre-specified integer. The field-specific parameters include the component means  $\boldsymbol{\mu}_{ij} = \{\mu_{s,ij}\}$ , standard deviations  $\boldsymbol{\sigma}_{ij}^2 = \{\sigma_{s,ij}^2\}$  and mixing proportions  $\boldsymbol{\pi}_{ij} = \{\pi_{s,ij}\}$  for components  $s = 1, \dots, S$ . The mixing proportions  $\boldsymbol{\pi}_{ij}$  satisfies the constraint that  $\sum_{s=1}^S \pi_{s,ij} = 1, \forall i, j$ . The identifiability constraint of the mixture model is the increasing order of the component means  $\mu_{1,ij} < \dots < \mu_{S,ij}$ . Therefore, the joint distribution of the diameter measurements  $\mathbf{y}_{ij}$  on field  $j$  of animal  $i$  is

$$f(\mathbf{y}_{ij}; \boldsymbol{\eta}_{ij}(\boldsymbol{\mu}_{ij}, \boldsymbol{\sigma}_{ij}, \boldsymbol{\pi}_{ij})) = \prod_{k=1}^{n_{ij}} f(y_{ijk}; \boldsymbol{\eta}_{ij}) = \prod_{k=1}^{n_{ij}} \sum_{s=1}^S \pi_{s,ij} \phi(y_{ijk}; \mu_{s,ij}, \sigma_{s,ij}^2) \quad (2.1)$$

The model (2.1) is a density model in which no error term is involved, i.e. the measurement errors of  $y_{ijk}$  are considered negligible in comparison to other sources of variability. The vector  $\boldsymbol{\eta}_{ij}$  is a proper transformation of all independent parameters in  $(\boldsymbol{\mu}_{ij}, \boldsymbol{\sigma}_{ij}, \boldsymbol{\pi}_{ij})$ . For

example, a two-component mixture distribution may have the vector  $\boldsymbol{\eta}_{ij}$  as

$$\boldsymbol{\eta}_{ij} = \left( \text{logit } \pi_{2,ij}, \mu_{1,ij}, \mu_{2,ij}, \ln \sigma_{1,ij}, \ln \sigma_{2,ij} \right)' \quad (2.2)$$

The generalized logit transformation yields independent parameters from the multinomial distribution parameters. The transformation and its Jacobian matrix are derived in Appendix A.1.1.

The vector  $\boldsymbol{\eta}_{ij}$  plays a similar role as the linear predictor in (generalized) linear mixed model. We assume that the following linear relationship holds for all  $i$  and  $j$ ,

$$\boldsymbol{\eta}_{ij}(\boldsymbol{\mu}_{ij}, \boldsymbol{\sigma}_{ij}, \boldsymbol{\pi}_{ij}) = \boldsymbol{\eta}_{ij} \left( \boldsymbol{\beta}, \mathbf{b}_i^{(1)}, \mathbf{b}_{ij}^{(2)} \right) = \mathbf{A}_i \boldsymbol{\beta} + \mathbf{B}_i^{(1)} \mathbf{b}_i^{(1)} + \mathbf{B}_{ij}^{(2)} \mathbf{b}_{ij}^{(2)} \quad (2.3)$$

Let  $h$  be the length of vector  $\boldsymbol{\eta}_{ij}$ . In the linear relationship (2.3),  $\mathbf{A}_i$  is an  $h$  by  $a$  design matrix of fixed effect  $\boldsymbol{\beta}$ , which is a length- $a$  vector,  $\mathbf{B}_i^{(1)}$  is an  $h$  by  $b_1$  design matrix of length- $b_1$  animal random effects  $\mathbf{b}_i^{(1)} \sim N(\mathbf{0}, \Sigma_1(\boldsymbol{\theta}_1))$  and  $\mathbf{B}_{ij}^{(2)}$  is an  $h$  by  $b_2$  design matrix of length- $b_2$  field random effects  $\mathbf{b}_{ij}^{(2)} \sim N(\mathbf{0}, \Sigma_2(\boldsymbol{\theta}_2))$ . We assume that all random effects  $\mathbf{b}_i^{(1)}$ , and  $\mathbf{b}_{ij}^{(2)}$  given  $i$  fixed, are independent with each other. The following notations will be extensively used in the body and appendices of this work:

$\mathbf{e}_{s(S)}$  or simply  $e_s$ , the length- $S$  base vector with the  $s$ th element being 1

$\mathbf{1}_n$  the length- $n$  vector with all elements being 1

$\mathbf{I}_n$  the order- $n$  identity matrix

$\mathbb{1}_n = \mathbf{1}_n \mathbf{1}_n'$

$\phi(\cdot; \boldsymbol{\mu}, \boldsymbol{\Sigma})$  The (multivariate) normal density with mean  $\boldsymbol{\mu}$  and (co)variance  $\boldsymbol{\Sigma}$ .

The parameters  $\boldsymbol{\theta}_1$  and  $\boldsymbol{\theta}_2$  depend on our choice of covariance parametrization. The covariance structures of the random effects  $\mathbf{b}_i^{(1)}$  and  $\mathbf{b}_{ij}^{(2)}$  would explain the correlation among the elements of  $\boldsymbol{\eta}_{ij}$ . The multivariate-response linear model (2.3) can be transformed into a

univariate-response model by using the index of the multivariate response as a covariate for indicating the repeated measurements (Wright, 1998). The transformed univariate-response model has a balanced design, as the original multivariate responses  $\boldsymbol{\eta}_{ij}$ 's are of the same length across all  $j$  and  $i$ . The same transformation applies to the linear mixed model. Under such transformation, the error covariance and total covariance in linear mixed context, referring to the correlation structure among the repeated measurements, correspond to the correlation among elements of the multivariate-response  $\boldsymbol{\eta}_{ij}$ . We choose the simplified design of random effects so that  $\mathbf{B}_i^{(1)} = \mathbf{B}_{ij}^{(2)} = \mathbf{I}_h$ . The linear predictor (2.3) is equivalent to

$$\boldsymbol{\eta} = \boldsymbol{\eta}(\boldsymbol{\beta}, \mathbf{b}) = \mathbf{A}\boldsymbol{\beta} + \mathbf{B}_1\mathbf{b}_1 + \mathbf{B}_2\mathbf{b}_2 = \mathbf{A}\boldsymbol{\beta} + \left(\mathbf{B}_1, \mathbf{B}_2\right) \begin{pmatrix} \mathbf{b}_1 \\ \mathbf{b}_2 \end{pmatrix} = \mathbf{A}\boldsymbol{\beta} + \mathbf{B}\mathbf{b} \quad (2.5)$$

where, using  $\otimes$  to represent the Kronecker product,

$$\begin{aligned} \boldsymbol{\eta} &= (\boldsymbol{\eta}'_1, \dots, \boldsymbol{\eta}'_M)', \quad \text{where } \boldsymbol{\eta}_i = (\boldsymbol{\eta}'_{i1}, \dots, \boldsymbol{\eta}'_{in_i})' \\ \mathbf{b}_1 &= \left(\mathbf{b}_1^{(1),t}, \dots, \mathbf{b}_M^{(1),t}\right)' \\ \boldsymbol{\Sigma}_1 &= \text{var}(\mathbf{b}_1) = \mathbf{I}_M \otimes \boldsymbol{\Sigma}_1 \\ \mathbf{b}_2 &= \left(\mathbf{b}_1^{(2),t}, \dots, \mathbf{b}_M^{(2),t}\right)', \quad \text{where } \mathbf{b}_i^{(2)} = \left(\mathbf{b}_{i1}^{(2),t}, \dots, \mathbf{b}_{in_i}^{(2),t}\right)' \\ \boldsymbol{\Sigma}_2 &= \text{var}(\mathbf{b}_2) = \text{Diag}\{\mathbf{I}_{n_i} \otimes \boldsymbol{\Sigma}_2 : i = 1, \dots, M\} \\ \boldsymbol{\Sigma} &= \boldsymbol{\Sigma}(\boldsymbol{\theta}) = \text{var}(\mathbf{b}) = \text{Diag}\{\boldsymbol{\Sigma}_1, \boldsymbol{\Sigma}_2\}, \quad \boldsymbol{\theta} = (\boldsymbol{\theta}_1, \boldsymbol{\theta}_2) \\ \mathbf{A} &= \text{stack by rows}\{\mathbf{1}_{n_i} \otimes \mathbf{A}_i : i = 1, \dots, M\} \\ \mathbf{B}_1 &= \text{Diag}\{\mathbf{1}_{n_i} \otimes \mathbf{B}_i^{(1)} : i = 1, \dots, M\} \\ \mathbf{B}_2 &= \text{Diag}\{\text{Diag}\{\mathbf{B}_{ij}^{(2)} : j = 1, \dots, n_i\} : i = 1, \dots, M\} \end{aligned}$$

We introduce the hierarchical model (2.1); (2.5) as the generalized linear mixed model for finite normal mixtures (GLMFM). As the tendon fibril data contains contaminations in

all levels of the hierarchy, i.e. in both the finite mixture (2.1) and the distribution of random effects  $\mathbf{b}$  in (2.5), the estimation of the GLMFM model, as well as its robust counterpart, may be carried out using the following two kinds of approaches.

The first kind is the two-stage approaches which consider the GLMFM model as two separate models: the stage-1 subject-specific model (2.1) and the stage-2 linear mixed model (2.5). The stage-1 subject-specific estimates  $(\hat{\boldsymbol{\mu}}_{ij}, \hat{\boldsymbol{\sigma}}_{ij}, \hat{\boldsymbol{\pi}}_{ij})$  are passed into the stage-2 model as the pseudo-observations  $\hat{\boldsymbol{\eta}}_{ij} = \boldsymbol{\eta}_{ij}(\hat{\boldsymbol{\mu}}_{ij}, \hat{\boldsymbol{\sigma}}_{ij}, \hat{\boldsymbol{\pi}}_{ij})$ . Two major classes of the two-stage approaches are the standard-two-stage (STS) and global-two-stage (GTS) approaches (Davidian and Giltinan, 1993). The two-stage approaches allow the flexibility in choosing the robust or non-robust estimation methods independently for each stage. They also allow constructing the stage-2 model with only a subset of the stage-1 estimates. The intrinsic drawback of the two-stage approaches is that by estimating the stage-1 model separately for each subject, we fails to "borrow strength" from all subjects. The stage-1 estimates and the pseudo-observations  $\hat{\boldsymbol{\eta}}_{ij}$  will be very poor if there are not enough sample size on each subject. However, the two-stage approaches can serve as an exploratory analysis of screening potentially significant covariates and variance structure, as well as a method of obtaining the starting values of other more complicated algorithms.

The second kind is the joint estimation which is motivated by the drawbacks of the two-stage approach. We have an extensive literature on the joint estimation of standard mixed models with a mean structure, which is reviewed in section 2.3 and 2.4, however, we don't have a joint estimation method for the GLMFM model, not to mention the robust counterpart.

## 2.2 Finite normal mixture model and robust density estimation

The finite mixture model is often used to model the skewed or multi-modal densities, such as the microscopic-field-specific distributions of the tendon fibril diameter measurements of Zhang et al. (2006). This section reviews the robust density estimation techniques of the finite mixture model, which would be directly applicable if we choose to adopt the two-stage approaches for the tendon fibril data.

In the finite mixture model, it is often assumed that all of the mixing components come from the same family of distributions. Most frequently used is the finite normal mixture model. The  $S$ -component normal mixture density is

$$f_{\theta}(x) = \sum_{s=1}^S \pi_s \phi(x; \mu_s, \sigma_s^2), \quad \pi_s > 0, \quad \sum_s \pi_s = 1 \quad (2.7)$$

Model (2.7) is equivalent to the notation that  $x|z \sim \phi(x; z'\boldsymbol{\mu}, z'\boldsymbol{\sigma}^2)$  and  $z \sim \text{mnomial}(1; \boldsymbol{\pi})$ , where  $z$  is the latent component indicator variable taking values  $\mathbf{e}_s$ . The identifiability constraints must be specified, since any permutation of component index  $s$  gives the identical model fit. For univariate models, it could be a strictly increasing or decreasing order of mean parameters  $\mu_s$ . The choice of the number of components  $d$  is in hot debate (Roeder, 1994; Turner and West, 1993; West, 1997; Richardson and Green, 1997; Stephens, 2000; Ishwaran et al., 2001; Ishwaran and James, 2002; Chen and Khalili, 2009), however, we assume a fixed number of components  $d$  and focus on robust estimation of  $\mu_s$ ,  $\sigma_s$  and  $\pi_s$  without additional restriction or prior knowledge.

The maximum likelihood estimation (MLE) for finite mixture model are covered extensively in Titterington et al. (1985); McLachlan and Basford (1988); McLachlan and Peel (2000); Boldea and Magnus (2009). The EM algorithms for solving MLE of finite mix-

ture model includes MIXFIT (McLachlan and Peel, 1998) and EMMIX (McLachlan and Peel, 1999). Many applications could be found in Belin and Rubin (1995); Celeux et al. (2001); Pilla and Lindsay (2001). Woodward et al. (1984) provided ways to evaluate the amount of separation between components and to find starting values for iterative algorithms, which were adopted widely in later work. Bayesian estimations are also implemented for finite, especially normal, mixtures model (Diebolt and Robert, 1994; Vounatsou et al., 1998; Fruhwirth-Schnatter, 2001). Jasra et al. (2005) provides an extensive review of methods dealing with non-identifiability of the components under symmetric priors. Robust estimation is required for the finite mixture model under model misspecification (Gray, 1994) or in the presence of contamination, which often result in ML failures such as "single match" to one or few data points when the estimated variance approaches zero and log-likelihood becomes unbounded, if the component variances are allowed to differ in model (2.7) (Kiefer and Wolfowitz, 1956; Day, 1969; Kiefer, 1978; Hathaway, 1985). The robustness can be achieved by modifying the EM algorithms (DeVeaux and Krieger, 1990; Windham, 1995) or by using the robust minimum divergence estimators, which will be discussed in detail below. The choices of divergences in estimating finite normal mixture model include the  $L_2$  distance (Clarke and Heathcote, 1994; Scott, 2001, 2004), Hellinger distance (Cutler and Cordero-Braña, 1996), symmetric  $\chi^2$  distance (Markatou et al., 1998; Markatou, 2000) and density power divergence (Basu et al., 1998; Fujisawa and Eguchi, 2006). In the rest of this section, we give a brief review of the robust minimum divergence estimators from the literature.

Studies on robustness have been an expanding area since Andrews et al. (1972); Tukey (1977); Huber (1981); White (1982); Hoaglin et al. (1983); Hampel et al. (1985). Robust statistics focus on balancing these two fundamental but potentially competing ideals:

efficiency when the model precisely describes the data and robustness when a slightly misspecified model is used. Non-robust methods such as those based on maximum likelihood are often optimal in efficiency, but badly affected by contaminations. Let  $\mathcal{G}$  be the class of all distributions on a dominating measure such as the Lebesgue or the counting measure on sample space  $\mathcal{X}$ . Consider a general parametric model class  $\mathcal{F}_\Theta = \{F_\theta, \theta \in \Theta\} \subset \mathcal{G}$ , so that the unknown true distribution  $G$  is within the neighborhood of the class  $\mathcal{F}_\Theta$ . The estimation of the *density model* picks the parameter estimate  $\hat{\theta} \in \Theta$  so that  $F_{\hat{\theta}}$  is an appropriate estimate of the underlying distribution  $G$ . The density estimate  $\hat{\theta} = T(G)$  is called the minimum divergence estimate (MDE) if the estimation minimizes a certain divergence measure from the model to the underlying distribution  $D(F_\theta, G)$ , where  $T(\cdot)$  is the minimum divergence functional defined on  $\mathcal{G}$ . Given a random sample  $X_1, \dots, X_n$  with empirical distribution  $\hat{G}_n$  and density  $\hat{g}_n$ , the density estimate  $\hat{\theta}_n = T(\hat{G}_n)$  is often an *M-estimate* (Huber, 1981) solved from the estimating equation  $\sum_{i=1}^n \psi(x_i, \theta) = 0$ , where the estimating function  $\psi$  is the derivatives the target divergence measure  $D(F_\theta, \hat{G}_n)$ . The target divergence measure  $D$  may be defined based on the characteristic, moment generating, distribution or density functions, such as the Kolmogorov-Smirnov, Wolfowitz, Cramér-von Mises and squared  $L_2$  norm (Parr, 1981). The maximum likelihood estimator (MLE) belongs to the family of M-estimators, where the target function is the Kullback-Leibler divergence and the estimating function is  $\psi(x_i, \theta) = u_\theta(x_i) = \nabla \log f_\theta(x_i)$ . However, the general M-estimation theory does not require the existence of a target function. The M-estimator  $T(\cdot)$  is Fisher consistent for model class  $\mathcal{F}_\Theta$  when  $T(F_\theta) = \theta$ , if and only if the estimating function  $\psi$  is unbiased under  $\mathcal{F}_\Theta$ .

The *influence function* (IF or influence curve, Huber, 1981) reflects the robustness of estimator  $T(\cdot)$  under the gross-error model  $G_{\varepsilon, x_0} = (1 - \varepsilon)G + \varepsilon\Delta_{x_0}$ , where  $\Delta_{x_0}$  is the

distribution assigning mass 1 to point  $x_0$ ,

$$IF := T'(G, x_0) = \lim_{\varepsilon \downarrow 0} \frac{T(G_{\varepsilon, x_0}) - T(G)}{\varepsilon} = \left. \frac{\partial T(G_{\varepsilon, x_0})}{\partial \varepsilon} \right|_{\varepsilon=0} \quad (2.8)$$

The influence function (2.8) is the functional derivative of  $T(\cdot)$  at  $G$  in the direction of  $\Delta_{x_0} - G$ . The upper bound of the influence function, if it exists, is defined as gross-error sensitivity  $GES = \sup_{x \in \mathcal{X}} \|T'(G, x)\|$  (Hampel et al., 1985). The influence function at  $F_\theta$ ,  $T'(F_\theta, x_0)$ , can be calculated from the M-estimating function  $\psi$  and its derivatives (Hampel et al., 1985, Chapter 4) or by taking the derivative with respect to  $\varepsilon$  on the estimating equation  $\int \psi(x, T(F_{\theta, \varepsilon, x_0})) dF_{\theta, \varepsilon, x_0} = 0$ , and solving for  $\partial T(F_{\theta, \varepsilon, x_0})/\partial \varepsilon|_{\varepsilon=0}$  (Lindsay, 1994). The M-estimators are asymptotically normal with covariance matrix  $\mathbf{V}_{F_\theta} = E_\theta(T'(F_\theta, x)[T'(F_\theta, x)]^t)$ . The MLE is considered non-robust as the corresponding influence function  $T'(F_\theta, x_0) = I(\theta)^{-1}u_\theta(x_0)$ , where  $I(\theta)$  is the Fisher information matrix for model  $F_\theta$ , is unbounded when  $x_0 \rightarrow \infty$ . On the other hand, only the influence function of MLE achieves the Cramér-Rao lower bound for  $\mathbf{V}_{F_\theta}$ , i.e. the MLE is first order efficient. Robust estimators with bounded influence function sacrifice the first order efficiency (Hampel et al., 1985), such as the family of MDE's minimizing the density power divergence (Basu et al., 1998)

$$D_\beta(f_\theta, g) = \int [f_\theta^{1+\beta} - (1 + \beta^{-1})gf_\theta^\beta + \beta^{-1}g^{1+\beta}]dx, \quad \beta > 0 \quad (2.9)$$

The density power divergence family (2.9) include the Kullback-Leibler divergence ( $\beta \rightarrow 0$ ) and  $L_2$  distance ( $\beta = 1$ ). The minimum density power divergence estimators are not first order efficient when  $\beta > 0$  because of the bounded influence function (Jones et al., 2001).

However, the influence function does not tell the whole story of robustness for the gross-error model. Define the bias response curve  $\Delta T(\varepsilon, x_0) = T(F_{\theta, \varepsilon, x_0}) - T(F_\theta)$  (Lindsay, 1994) and the  $\varepsilon$ -influence curve  $\varepsilon IF(\varepsilon, x_0) = \varepsilon^{-1}\Delta T(\varepsilon, x_0)$  (Beran, 1977). The influence function

is either the first order coefficient in series expansion of bias response curve  $\Delta T(\varepsilon, x_0) \approx \varepsilon T'(F_\theta, x_0)$ , or the point-wise limit of  $\varepsilon$ -influence curve  $T'(F_\theta, x_0) = \lim_{\varepsilon \downarrow 0} \varepsilon IF(\varepsilon, x_0)$ . By studying the second or third order approximation of the bias response curve (Lindsay, 1994) or the boundedness of each  $\varepsilon$ -influence curves (Basu et al., 1998), robust estimators with the first-order efficiency are proposed. Beran (1977, 1978) addressed both robustness and full first order efficiency of minimum Hellinger distance estimator, as well as the non-uniform convergence of  $\varepsilon$ -influence curve to the influence function. Lindsay (1994); Basu and Lindsay (1994) addressed the inaccuracy of influence function as the first-order approximation to the bias response curve. In other words, an estimator could both have the same influence function as MLE so that to be first-order efficient, and be robust at the cost of the second-order efficiency (Rao, 1962). Define the smoothed model density  $f_\theta^*(t) = \int k(x; t, h) f_\theta(x) dx$  and the smoothed empirical density  $\hat{g}_n^*(t) = \int k(x; t, h) \hat{g}_n(x) dx$ , where  $k(x; t, h) = k((x - t)/h)$  is the common smoothing kernel. Larger value of the smoothing bandwidth  $h$  leads to faster convergence but less robustness (Basu and Lindsay, 2004). (see the choice of smoothing bandwidth  $h$  in Markatou, 2000). Define the smoothed Pearson residual  $\delta_n^*$  (Lindsay and Roeder, 1992)

$$\delta_n^*(x) = \hat{g}_n^*(x) / f_\theta^*(x) - 1 \quad (2.10)$$

Therefore, the contaminations are distinguished as outlier when  $\delta_n^* > 0$  and 'inlier' when  $-1 < \delta_n^* < 0$ , where less observations are present than the model predicts. Basu and Lindsay (1994) defined the family of  $\delta$ -divergences

$$D(f_\theta^*, \hat{g}_n^*) = \int f_\theta^*(x) \mathbf{G}(\delta_n^*(x)) dx \quad (2.11)$$

where  $\mathbf{G}$  is a thrice-differentiable function on  $[-1, \infty)$  with  $\mathbf{G}(0) = 0$ ,  $\mathbf{G}'(0) = 0$  and  $\mathbf{G}''(0) = 1$ . The family (2.11) includes the power class (Cressie and Read, 1984), blended-

$\chi^2$  and general negative exponential (Jeong and Sarkar, 2000) through the choice of function  $\mathbf{G}$ . For example, the symmetric- $\chi^2$  divergence (Markatou et al., 1998), which is a special case of the blended- $\chi^2$ , corresponds to  $\mathbf{G}(\delta) = \delta^2/(\delta + 2)$ ; the negative exponential divergence (Bhandari et al., 2006) corresponds to  $\mathbf{G}(\delta) = e^{-\delta} + \delta - 1$ . Figure 2.1 compares the functions  $\mathbf{G}(\cdot)$  corresponding to the Kullback-Leibler divergence with the Hellinger divergence, symmetric- $\chi^2$  divergence and negative exponential divergence, as the function of smoothed Pearson residuals  $\delta^*$  (section 2.2). The symmetric- $\chi^2$  and negative exponential divergences are either comparable or more robust than the Kullback-Leibler divergence, for both inliers ( $-1 < \delta^* < 0$ ) and outliers ( $\delta^* > 1$ ).

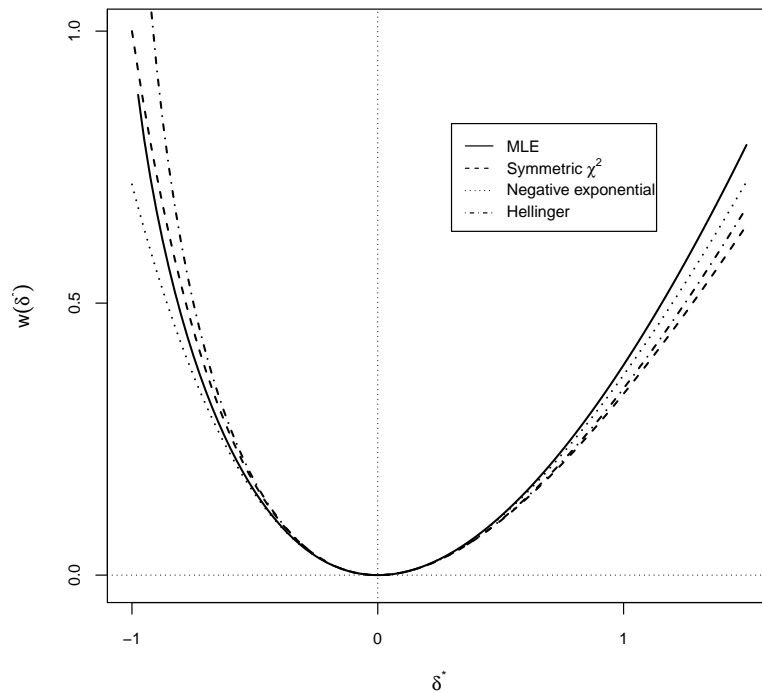


Figure 2.1: Divergence discriminating function  $\mathbf{G}(\delta_n^*)$

## 2.3 Standard mixed models based on mean structure

The mixed model is often referred to as the "hierarchical" model, as the subjects within the same hierarchical clusters are often correlated because of the random effects associated with each clusters. In this section we review the standard mixed models based on a mean structure and the corresponding estimation techniques in literature.

In this work, the term "standard mixed models" includes the linear, generalized linear and nonlinear mixed models. The observations within each cluster is modeled by a mean structure and a random error. These models are described using the following two-level data example. Let  $\mathbf{y}_i = \{y_{ij} : j = 1, \dots, n_i\}$  be the repeated observations from the  $i$ th individual. Let  $\boldsymbol{\mu}_{y_i}$  be the parameter vector specific to individual  $i$ . Most of the time, we do not need to fully specify the conditional distribution  $\mathbf{y}_i | \boldsymbol{\mu}_{y_i}$ , as the first two conditional moments are often sufficient for inference. In other words,  $\mathbf{y}_i | \boldsymbol{\mu}_{y_i}$  is often assumed to be multivariate normal with

$$\mathbb{E}(\mathbf{y}_i | \boldsymbol{\mu}_{y_i}) = \mathbf{f}_i(\boldsymbol{\mu}_{y_i}), \quad \text{cov}(\mathbf{y}_i | \boldsymbol{\mu}_{y_i}) = \mathbf{D}_i(\boldsymbol{\mu}_{y_i}, \boldsymbol{\xi}) \quad (2.12)$$

where  $\mathbf{f}_i$  is a length- $n_i$  vector-valued function specific for the individual  $i$  and  $\boldsymbol{\xi}$  is the population dispersion parameter. The within-individual heterogeneity is reflected in the covariance matrix  $\mathbf{D}_i(\boldsymbol{\mu}_i, \boldsymbol{\xi})$ , while the form of  $\mathbf{D}_i$  is assumed common across all individuals. Let  $\mathbf{x}_i$  be the vector of covariates for the  $i$ th individual. Let  $\boldsymbol{\beta}$  be the vector of fixed effect and  $\mathbf{b}_i$  be the vector of random effect specific to the  $i$ th individual, The distribution of random effect  $\mathbf{b}_i$  comes from a general distribution class  $\mathcal{H}$ . The *linear predictor*  $\boldsymbol{\eta}_i$  for the  $i$ th individual is

$$\boldsymbol{\eta}_i = \mathbf{d}(\mathbf{x}_i, \boldsymbol{\beta}, \mathbf{b}_i), \quad \mathbf{b}_i \sim h(\boldsymbol{\theta}) \in \mathcal{H} \quad (2.13)$$

where  $\mathbf{d}$  is a vector-valued function and very often the random effects follow a multivariate

normal distribution  $h = N(\mathbf{0}, \Sigma(\boldsymbol{\theta}))$ . By different parametrization of  $h$ , we may have fully parametric, nonparametric or semiparametric population models. The conditional mean  $\mathbf{f}_i(\boldsymbol{\mu}_i)$  is connected to the linear predictor through a link function  $\mathbf{g}_i(\cdot)$

$$\mathbf{g}_i(\mathbf{f}_i(\boldsymbol{\mu}_{y_i})) = \boldsymbol{\eta}_i \quad (2.14)$$

In Bayesian context, we need yet to specify hyperpriors, often non-informative, for parameters  $\boldsymbol{\beta}$ ,  $\boldsymbol{\xi}$  and  $\boldsymbol{\theta}$ .

The model (2.12); (2.13); (2.14) provides a general framework for the standard mixed models, including (1) linear mixed effect model (LME), where  $\mathbf{f}_i$  is linear with respect to  $\boldsymbol{\mu}_{y_i}$ ,  $\mathbf{g}_i$  is the identity function,  $\mathbf{d}$  is linear with respect to  $\boldsymbol{\beta}$  and  $\mathbf{b}_i$  and  $\mathbf{D}_i$  depends on  $i$  only through the dimension  $n_i$ ; (2) generalized linear mixed model or hierarchical generalized linear models (GLMM or HGLM, Lee and Nelder, 1996, 2001, 2003, 2006) where the within-subject observations  $\mathbf{y}_i|\boldsymbol{\mu}_i$  are i.i.d. following an exponential-family distributions as in the case of generalized linear model (GLM, McCullagh and Nelder, 1983),  $\mathbf{g}_i = \mathbf{g}$  is the corresponding (canonical) link function,  $\mathbf{d}$  is linear with respect to  $\boldsymbol{\beta}$  and  $\mathbf{v}(\mathbf{b}_i)$ , where  $\mathbf{v}(\cdot)$  is a strictly monotone function of the random effect  $\mathbf{b}_i$ . The distribution  $h$  is conjugate to the distribution of  $\mathbf{y}_i|\boldsymbol{\mu}_{y_i}$ . The usual LME is the normal-normal GLMM with identity link, where the first normal refers to the conditional model and the second refers to distribution of random effect; (3) nonlinear mixed model (NLMM, Lindstrom and Bates, 1990; Davidian and Giltinan, 1993, 1995, 2003) where  $\mathbf{f}_i$  is nonlinear with respect to  $\boldsymbol{\mu}_{y_i}$ , simple link function  $\boldsymbol{\mu}_{y_i} = \boldsymbol{\eta}_i$  and  $\mathbf{d}$  can be either linear or nonlinear.

Standard estimation techniques of LME and GLMM are extensively reviewed in Searle et al. (1992), such as MLE, generalized least squares (GLS) and restricted maximum likelihood estimator (REML, Patterson and Thompson, 1971; Harville, 1977). Robust estimation techniques in literature fall in the following categories: (i). Applying Huber (1981)'s weight-

ing functions to the ML and REML estimating equations to obtain the bounded-influence estimates (Welsh and Richardson, 1997; Richardson, 1997; Richardson and Welsh, 1995; Huggins, 1993); (ii). Construction of *pseudo-observations* which are less extreme by applying Huber or Tukey's weight functions, onto the extreme estimated residuals (Dueck and Lohr, 2005; Yau and Kuk, 2002; Stahel and Welsh, 1997; Rocke, 1983, 1991; Fellner, 1986; Bickel et al., 1976); (iii). Using a family of smoothed functionals with robust properties, such as the Fréchet differentiable functionals (Bednarski and Zontek, 1996); (iv). Uhlig's high-breakdown-point variance component estimator (Uhlig, 1997; Muller and Uhlig, 2001); (v). Bayesian robustness with respect to priors in nonparametric mixed model (Betrò et al., 2006); (vi). Incorporating a sub-model for the contaminations in random effects, such as adding random effects in dispersion of conditional model (Yun and Lee, 2006), using partially specified priors (Polasek and Pötzelberger, 1988; Pötzelberger and Polasek, 1991), or replacing the normal random effects by  $t$ -distribution (Wakefield et al., 1994), normal gross-error model and scale normal mixtures (Sharples, 1990; Gelman and King, 1990; Moreno and Pericchi, 1993; Datta and Lahiri, 1995; Muller and Rosner, 1997; Choy and Smith, 1997) or the family of exponential-power distributions (Box and Tiao, 1973; Solaro and Ferrari, 2007).

## 2.4 Best linear unbiased predictor procedure for linear mixed model

McGilchrist and Yau (1995) developed an algorithm for calculating the maximum likelihood estimator (MLE) and the restricted maximum likelihood estimator (REML) for the linear mixed model, through the help of the best linear unbiased predictors (BLUP, Hen-

derson et al., 1959; Henderson, 1975). Consider the standard normal-normal linear mixed model

$$\mathbf{y} = \mathbf{A}\boldsymbol{\beta} + \mathbf{B}\mathbf{b} + \boldsymbol{\varepsilon}, \quad \text{where } \mathbf{b} \sim N(\mathbf{0}, \boldsymbol{\Sigma}(\boldsymbol{\theta})) \quad \text{and} \quad \boldsymbol{\varepsilon} \sim N(\mathbf{0}, \mathbf{D}(\boldsymbol{\xi})) \quad (2.15)$$

The joint log-likelihood  $l(\boldsymbol{\beta}, \mathbf{b}, \boldsymbol{\theta}; \mathbf{y})$  of the LME model (2.15) is the sum of the conditional log-likelihood given random effects  $l_1(\boldsymbol{\beta}, \boldsymbol{\xi}|\mathbf{b}; \mathbf{y})$  and the log-likelihood of random effects  $l_2(\mathbf{b}, \boldsymbol{\theta})$ ,

$$\begin{aligned} l(\boldsymbol{\beta}, \mathbf{b}, \boldsymbol{\theta}; \mathbf{y}) &= l_1(\boldsymbol{\beta}, \boldsymbol{\xi}|\mathbf{b}; \mathbf{y}) + l_2(\mathbf{b}, \boldsymbol{\theta}) \\ &= \log \phi(\mathbf{y}; \mathbf{A}\boldsymbol{\beta} + \mathbf{B}\mathbf{b}, \mathbf{D}(\boldsymbol{\xi})) + \log \phi(\mathbf{b}; \mathbf{0}, \boldsymbol{\Sigma}(\boldsymbol{\theta})) \end{aligned} \quad (2.16)$$

The best linear unbiased predictors (BLUP)  $(\boldsymbol{\beta}_l, \mathbf{b}_l, \boldsymbol{\theta}_l, \boldsymbol{\xi}_l)$  are the realized values of the random effect  $\mathbf{b}$ , as well as the estimates of fixed effect  $\boldsymbol{\beta}$  and the covariance parameters of  $\boldsymbol{\theta}$  and  $\boldsymbol{\xi}$ , which maximizes the joint log-likelihood  $l(\boldsymbol{\beta}, \mathbf{b}, \boldsymbol{\xi}, \boldsymbol{\theta}; \mathbf{y})$ . The BLUP estimating equations respect to  $\boldsymbol{\beta}$  and  $\mathbf{b}$  boil down to the following *BLUP matrix equation* (Henderson et al., 1959), or the *normal equations*

$$\begin{pmatrix} \mathbf{A}'\mathbf{D}^{-1}(\boldsymbol{\xi})\mathbf{A} & \mathbf{A}'\mathbf{D}^{-1}(\boldsymbol{\xi})\mathbf{B} \\ \mathbf{B}'\mathbf{D}^{-1}(\boldsymbol{\xi})\mathbf{A} & \mathbf{B}'\mathbf{D}^{-1}(\boldsymbol{\xi})\mathbf{B} + \boldsymbol{\Sigma}^{-1}(\boldsymbol{\theta}) \end{pmatrix} \begin{pmatrix} \boldsymbol{\beta}_l \\ \mathbf{b}_l \end{pmatrix} = \begin{pmatrix} \mathbf{A}'\mathbf{D}^{-1}(\boldsymbol{\xi})\mathbf{y} \\ \mathbf{B}'\mathbf{D}^{-1}(\boldsymbol{\xi})\mathbf{y} \end{pmatrix} \quad (2.17)$$

The term "linear" indicates the linear relationship between the observations  $\mathbf{y}$  and the estimated BLUP of  $\mathbf{b}_l$  in (2.17). The BLUPs are not maximum likelihood estimators, because the joint log-likelihood  $l(\boldsymbol{\beta}, \mathbf{b}, \boldsymbol{\xi}, \boldsymbol{\theta}; \mathbf{y})$ , sometimes called the hierarchical *h*-likelihood (Lee and Nelder, 1996) or penalized log-likelihood (McGilchrist, 1994), is not a likelihood (Robinson, 1991). The BLUP procedure is an early and computationally expensive estimation method, as it requires the solving of random effects  $\mathbf{b}$ .

McGilchrist and Yau (1995) proposed an iterative algorithm for calculating the maximum and restricted maximum likelihood estimates from the BLUPs for linear mixed model. Each

iteration of the algorithm contains two steps. First, given the starting values  $(\boldsymbol{\beta}_0, \mathbf{b}_0, \boldsymbol{\xi}_0, \boldsymbol{\theta}_0)$ , solve the BLUP's  $(\hat{\boldsymbol{\beta}}_l(\boldsymbol{\theta}_0, \boldsymbol{\xi}_0), \hat{\mathbf{b}}_l(\boldsymbol{\theta}_0, \boldsymbol{\xi}_0))$  using the BLUP matrix equation (2.17). Second, solve  $(\hat{\boldsymbol{\xi}}, \hat{\boldsymbol{\theta}})$  from  $(\hat{\boldsymbol{\beta}}_l(\boldsymbol{\theta}_0, \boldsymbol{\xi}_0), \hat{\mathbf{b}}_l(\boldsymbol{\theta}_0, \boldsymbol{\xi}_0))$  using the ML (or REML) estimating equations for covariance parameters (section 7, McGilchrist and Yau, 1995). Let the new starting values be  $(\hat{\boldsymbol{\beta}}_l(\boldsymbol{\theta}_0, \boldsymbol{\xi}_0), \hat{\mathbf{b}}_l(\boldsymbol{\theta}_0, \boldsymbol{\xi}_0), \hat{\boldsymbol{\xi}}, \hat{\boldsymbol{\theta}})$  and repeat the iteration until all parameters converge. The final estimates  $(\hat{\boldsymbol{\beta}}_l(\hat{\boldsymbol{\xi}}, \hat{\boldsymbol{\theta}}), \hat{\mathbf{b}}_l(\hat{\boldsymbol{\xi}}, \hat{\boldsymbol{\theta}}))$  are the ML (or REML) estimates of the standard LME model (2.15), i.e. they maximize the ML (or REML) likelihood in section 5 and 6 of McGilchrist and Yau (1995). McGilchrist and Yau (1995) also demonstrates that the MLE and REML estimates, as well as their information matrix, only depend on the BLUP's and the second order derivatives of the log-likelihoods  $l_1$  and  $l_2$ .

## 2.5 Mixed models without a mean structure

In this section, we provide an overview of the family of the non-mean-structure mixed model, which does not belong to the standard mixed models framework of (2.12); (2.13); (2.14). In the standard mixed models (2.12), the conditional distribution  $\mathbf{y}_i | \boldsymbol{\mu}_{y_i}$  is assumed to be multivariate normal through the inverse link function of the linear predictor  $\mathbf{g}^{-1}(\boldsymbol{\eta}_i)$ . On the other hand, the non-mean-structure mixed model deals with the situation when the conditional distributions  $\mathbf{y}_i | \boldsymbol{\mu}_{y_i}$  cannot be assumed multivariate normal, but has to be written as a density model. Consider

$$y_{ij} \sim f(y_{ij}; \boldsymbol{\mu}_{y_i}) \tag{2.18a}$$

$$\boldsymbol{\mu}_{y_i} = \mathbf{g}^{-1}(\boldsymbol{\eta}_i), \quad \mathbf{g} \text{ is an arbitrary link function} \tag{2.18b}$$

$$\boldsymbol{\eta}_i = \mathbf{d}(\mathbf{x}_i, \boldsymbol{\beta}, \mathbf{b}_i), \quad \mathbf{b}_i \sim h(\boldsymbol{\theta}) \in \mathcal{H} \tag{2.18c}$$

The proposed GLMFM model belongs to the non-mean-structure mixed model (2.18), where (2.18a) is the finite mixture density with parameters  $\boldsymbol{\mu}_{y_i}$  and (2.18c) is the linear predictor with  $h$  being multivariate normal.

The non-mean-structure mixed model (2.18) is more than a generalization of the standard mixed models (2.12); (2.13); (2.14) in the following aspects. First, the non-mean-structure mixed model includes both the mean and covariance structures, while in standard mixed models, only the mean structure of the repeated measurements are taken into account in the link function and linear predictor. Specifically, in the GLMFM model we are interested in the overall properties of the mixture distribution, thus all independent parameters from the mixture would be of interest, transformed and passed into the linear predictor  $\boldsymbol{\eta}$ . Second, in standard mixed models, the random errors are naturally the variances or "uncertainty" of the conditional distribution, whether it is normal or a general exponential family distribution. In non-mean-structure mixed model this no longer holds as the variances parameters are included in the linear predictor. Instead, the random error terms come from *estimating* these mixture parameters. Luckily we may make use of the asymptotic normality of finite mixture estimates, so as to assume a multivariate-normal error term. Third, the concept of the "canonical link" no longer exists in the non-mean-structure mixed model. The choice of transformation (2.18b) depends more on the distribution assumptions, such as the choice of log transformation for standard deviations and general logit transformation for mixing proportions in the GLMFM model.

The examples of non-mean-structure mixed model (2.18) are very few in literature. These are often analyzed under the Bayesian framework, with the help of the latent multinomial indicator variables  $z_{ij}$  associated with observations  $y_{ij}$  (Gelman et al., 2004; Yau et al., 2003; Scaccia and Green, 2003; Skates et al., 2001; Pauler and Laird, 2000; Thompson et al., 1998).

The tendon fibril data inspires to construct an hierarchical model with little constraint on mixture parameters and versatile enough to depict the dependency and correlations of those parameters. The proposed GLMFM model which emphasizes the correlation among the finite mixture parameters, together with reasonable robust estimation techniques, are not discussed in literature.

## Chapter 3

# The robust estimation of finite mixture model

The family of weighted likelihood estimators largely overlaps with the minimum divergence estimators. They are robust to data contaminations compared to maximum likelihood estimators. In this chapter, we define the class of the generalized weighted likelihood estimators (GWLE) and introduce a new truncated cubic-inverse weight. In section 3.1, we unify previously considered minimum divergence estimators as the GWLE's. In section 3.2, we provide their influence function and discuss the efficiency requirements. In section 3.3, we introduce the GWLE with a new truncated cubic-inverse weight, which is both first and second order efficient and with stronger robustness than those previously reported weights. We also discuss new ways of selecting the smoothing bandwidth and weighted starting values for the iterative algorithm. In section 3.4, we consider the density estimation of the finite mixture of exponential family distributions. We describe a simple iterative algorithm for computing such GWLE. In section 3.5, the advantage of the truncated cubic-inverse weight is illustrated in a simulation study of three-components normal mixtures model with large

overlaps and heavy contaminations. In section 3.6, we apply the GWLE to the analysis of the tendon fibril data. The content of this chapter has been published in Zhan et al. (2011).

### 3.1 Generalized weighted likelihood estimators

Let  $X_1, \dots, X_n$  be a random sample and  $\hat{G}_n$  be the corresponding empirical distribution. Let  $\mathcal{F}_\Theta = \{F_\theta : \theta \in \Theta\}$  be the fitted model family. Consider an estimator  $T(\cdot)$  which is the root of estimating equation  $\sum_i \tilde{\psi}(x_i, \theta) = 0$ ,

$$\tilde{\psi}(x, \theta) = w(x; F_\theta, \hat{G}_n) \mathbf{u}_\theta(x) - \mathbf{a}(\theta), \quad \mathbf{a}(\theta) = \mathbb{E}_\theta (w(x) \mathbf{u}_\theta(x)) \quad (3.1)$$

In the estimating function (3.1),  $\mathbf{u}_\theta$  is the vector of score functions and  $\mathbf{a}(\theta)$  is the bias adjustment term to ensure that (3.1) is unbiased under  $F_\theta$ . Kuk (1995) has provided an iterative algorithm to calculate  $\mathbf{a}(\theta)$  when the integration in (3.1) is numerically difficult. The term  $w(x_i; F_\theta, \hat{G}_n)$  is the weight assigned to observation  $x_i$ , as an adjustment to the potential contaminations. For example, the MLE corresponds to the ordinary score function  $\tilde{\psi}(x, \theta) = \mathbf{u}_\theta(x)$ , which assigns weight  $w(x) = 1$  to all observations, and the log-likelihood target function, which is the empirical version of the Kullback-Leibler divergence (Kullback and Leibler, 1951; Kullback, 1959). Generally, the estimating function  $\tilde{\psi}(x, \theta)$  does not satisfy the definition of M-estimators, where the estimating function  $\psi(x_i, \theta)$  contains only the observation  $x_i$  and parameter  $\theta$ . On the other hand,  $\tilde{\psi}(x, \theta)$  in (3.1) may contain the whole random sample  $\hat{G}_n$  through the weight  $w(x; F_\theta, \hat{G}_n)$ . Therefore we must be cautious when deriving the inferential properties comparable to M-estimators. We call the solution of estimating function (3.1) the generalized weighted likelihood estimator (GWLE) with weights  $w$ . We provide an explanation of the weights for some previously reported estimators, and discuss the possibility of introducing more general weights combining these

advantages.

The minimum density power divergence (Basu et al., 1998) estimator (Dens.Pow) is an M-estimator with corresponding GWLE terms

$$w(x, F_\theta) = f_\theta^\beta(x); \quad \mathbf{a}(\theta) = \int \mathbf{u}_\theta(x) f_\theta^{1+\beta}(x) dx \quad (3.2)$$

with  $\beta \in [0, 1]$ . In equation (3.2) the weight  $w(x_i) = f_\theta^\beta(x_i)$  depends on the concordance between the single observation  $x_i$  and the fitted model. High leverage points are down-weighted at a scale less extreme than local density  $f_\theta$ . The flaws of this scheme are obvious. First, no distinction between outliers and inliers can be made; sample points with more or less observations than they should will receive the same weights. Second, outlier contaminations are down-weighted only if they occur at high leverage sample points; they are not going to be noticed if they occur at the “middle” of the data.

The estimating equation of Lindsay’s family, as the derivative of the  $\delta$ -divergence (Basu and Lindsay, 1994), is  $\int w(\delta_n^*) \cdot \nabla \ln f_\theta^* \cdot \hat{g}_n^* dx = 0$ . Since the presence of smoothed empirical density  $\hat{g}_n^*$  disallows writing the integration into a summation, we substitute the smoothed densities  $f_\theta^*$  and  $\hat{g}_n^*$  by the original densities  $f_\theta$  and  $\hat{g}_n$ , respectively, while keeping the smoothed Pearson residual  $\delta_n^*$  intact (Basu and Lindsay, 2004). We end up with the estimating equation  $\int w(\delta_n^*) \cdot \nabla \ln f_\theta \cdot \hat{g}_n dx = 0$ , which is equivalent to GWLE with  $\mathbf{a}(\theta) = 0$  and weights  $w(x; F_\theta, \hat{G}_n) = w(\delta_n^*(x))$ . For the minimum symmetric  $\chi^2$  divergence estimator (Sym.Chi), the weight term is

$$w(\delta_n^*(x)) = 1 - \delta_n^{*2} / (\delta_n^* + 2)^2 \quad (3.3)$$

and for minimum negative exponential divergence estimator (Neg.Exp),

$$w(\delta_n^*(x)) = \left( e - (2 + \delta_n^*) e^{-\delta_n^*} \right) / (\delta_n^* + 1) \quad (3.4)$$

The estimators from Lindsay's family of estimating equations are not M-estimators, as the estimating equations contain the empirical density  $\hat{g}_n$  in the  $\delta$ -weights  $w(\delta_n^*(x))$ . Compared to the weight in (3.2), the weights  $w(x_i) = w(\delta_n^*(x_i))$  allow the distinction of outliers and inliers by concordance between the whole sample and the fitted model. The weight of Lindsay's estimators has an one-to-one correspondence to certain divergence measure from  $F_\theta$  to  $\hat{G}_n$  or  $G$ . This weight is rescaled to  $w(0) = 1$  for zero residual, which corresponds to zero divergence and concordance, as a reference for comparing different weight functions. This task is not feasible for weights depending on  $f_\theta$  such as (3.2). Figure 3.1 shows the weights Sym.Chi (3.3) and Neg.Exp (3.4) together with the constant weight for MLE, as well as the weight for the popular minimum Hellinger distance estimator  $w(\delta^*) = (2\sqrt{1 + \delta^*} - 1)/(\delta^* + 1)$ . Both (3.3) and (3.4) provide more down-weight for outliers ( $\delta^* > 0$ ) than minimum Hellinger distance estimator. The difference is that Sym.Chi (3.3) represents the intuitive idea of a unimodal weight on interval  $(0, 1]$  with  $w(-1) = w(\infty) \rightarrow 0$  for extreme inlier or outlier; while Neg.Exp (3.4) assigns weights larger than 1 for inliers, as a compensation of lack of observed data. Lindsay (1994) showed that Neg.Exp shrinks the residual adjust function for both outliers and inliers, and is second order efficient, which favors the latter interpretation. It's also noticed that this difference is only observed for significant inliers, which is not a phenomena frequently encountered in real data.

A heuristic way of understanding the GWLE estimating equation is described as follow: in the kernel part of (3.1), i.e.  $w(x_i)\mathbf{u}_\theta(x_i) = w(x_i)\partial \ln f(x_i|\theta)/\partial \theta = \partial[\ln f(x_i|\theta)^{w(x_i)}]/\partial \theta$ , although the weight function  $w(x_i)$  contains parameter  $\theta$ , the differential does not involve  $w(x_i)$ . Thus we may treat the weight  $w(x_i)$  as a constant and the notation  $f(x_i|\theta)^{w(x_i)}$  implies the fact that we are assuming that we observed data  $x_i$  for  $w(x_i)$  times, which is a way to "dilute" outliers and "condense" inliers. In order to take advantage of the weights

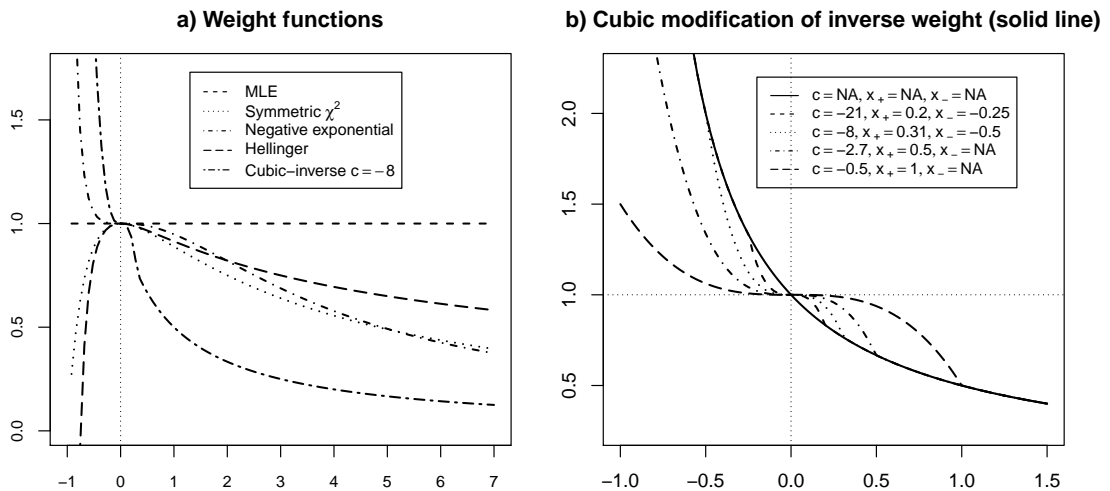


Figure 3.1: Weight functions depending on smoothed Pearson residual  $\delta_n^*$

(3.2), (3.3) and (3.4), we propose a generalized weight of the form

$$w(x_i; F_\theta, \hat{G}_n) = w(x_i; \delta_n^*, f_\theta) \quad (3.5)$$

with normal smoothing kernel  $k(x; t, h) = \phi(x; t, h)$ , which allows simultaneous down weighing of the outliers, inliers and low-density points. In the rest of this work, we refer to the estimator with weight (3.5) as our GWLE.

## 3.2 Influence function and asymptotic normality

**Theorem 3.1** (Influence function and asymptotic covariance). *Let  $G$  be the true distribution and  $F_\theta$  be the model distribution.  $T(\cdot)$  is the GWLE with weights (3.5). Let  $\theta = T(G)$ , we have*

(i). *Under mild regularity conditions (Basu and Lindsay, 1994), the sequence of estimators  $T(\hat{G}_n)$  exists.*

(ii). The influence function of  $T$  is

$$\text{IF}(x_0, G, F_\theta) = T'_h(x_0, G) = -\text{DEN}_{h,g}^{-1} \text{NUM}_{h,g,x_0} \quad (3.6)$$

$$\begin{aligned} \text{NUM}_{h,g,x_0} = & \int \frac{k(x; x_0, h) - g^*(x)}{f_\theta^*(x)} \cdot w'_\delta \mathbf{u}_\theta (g - f_\theta) dx \\ & + w(x_0, \delta^*, f_\theta) \mathbf{u}_\theta(x_0) - \int w(x, \delta^*, f_\theta) \mathbf{u}_\theta g dx \end{aligned} \quad (3.7)$$

$$\begin{aligned} \text{DEN}_{h,g} = & \int [(\delta^* + 1) w'_\delta \mathbf{u}_\theta^* - f_\theta w'_f \mathbf{u}_\theta] \mathbf{u}_\theta^t (f_\theta - g) dx \\ & + \int w(x, \delta^*, f_\theta) \nabla \mathbf{u}_\theta^t g dx - \int w(x, \delta^*, f_\theta) \nabla^2 f_\theta dx \end{aligned}$$

where  $\mathbf{u}_\theta^*(x) = \partial \log f_\theta^*(x) / \partial \theta$ . The notation  $(\cdot)^t$  denotes vector transposition. The partial derivatives are  $w'_\delta = \partial w(x, \delta^*, f_\theta) / \partial \delta^*$  and  $w'_f = \partial w(x, \delta^*, f_\theta) / \partial f_\theta$ .

(iii). The asymptotic distribution of  $\sqrt{n}(T(\hat{G}_n) - T(G))$  is multivariate normal with mean 0 and covariance matrix  $\mathbf{V}_G = \int T'_h(x, G)[T'_h(x, G)]^t dG(x)$ , and

$$\hat{\mathbf{V}}_{\hat{G}_n} = \text{DEN}_{h,\hat{g}_n}^{-1} \left( \frac{1}{n} \sum_{i=1}^n \text{NUM}_{h,\hat{g}_n,x_i} \text{NUM}_{h,\hat{g}_n,x_i}^t \right) \left( \text{DEN}_{h,\hat{g}_n}^{-1} \right)^t \quad (3.8)$$

*Proof of Theorem 3.1.* The proof is straightforward but tedious differentiation of  $\tilde{\psi}$  (Appendix A.2) □

**Theorem 3.2** (Efficiency requirement). *The GWLE  $T(\cdot)$  with weight (3.5) is first order efficient if*

$$w'_\delta(\delta^*, f_\theta)|_{\delta^*=0} = 0$$

and is second order efficient (Rao, 1962) if

$$w''_\delta(\delta^*, f_\theta)|_{\delta^*=0} = 0$$

where  $w''_\delta$  is the second order derivative.

*Proof of Theorem 3.2.* The proof is a quick derivation from Remark E of Lindsay (1994), where the author listed the efficiency requirements on residual adjustment function. As

there exists a one-to-one correspondence between residual adjustment function and weight function, the equivalent requirement on weight function is obtained.  $\square$

*Note on Theorem 3.2.* Theorem 3.2 implies that an estimator which is second order efficient assigns weights larger than 1 to inliers, as does the negative exponential weight (3.4) shown in Figure 3.1.  $\square$

### 3.3 The new truncated cubic-inverse weight

We introduce a new truncated cubic-inverse weight in this section. This new weight not only satisfies the efficiency requirements in Theorem 3.2. but also has better empirical robustness properties than the previous weights (3.2), (3.3) and (3.4), which will be shown later using simulation. First, we define the *inverse weight*

$$w_1(\delta_n^*(x)) = (\delta_n^*(x) + 1)^{-1} = f_\theta^*(x)/\hat{g}_n^*(x) \quad (3.9)$$

A heuristic explanation of the advantage of inverse weight is given below. Consider the estimating function

$$\int_x w_1(\delta_n^*(x)) \tilde{\mathbf{u}}_\theta(x) \hat{g}_n^*(x) dx = \int \tilde{\mathbf{u}}_\theta(x) f_\theta^*(x) dx \quad (3.10)$$

where  $\tilde{\mathbf{u}}_\theta(x) = \int k(t; x, h) \mathbf{u}_\theta(t) dt$ . On the other hand, let estimator  $MLE^*$  be the solution of estimating equation  $\sum_i \mathbf{u}_\theta^*(x_i) = \int \mathbf{u}_\theta^*(x) d\hat{G}_n(x) = 0$ , where  $\mathbf{u}_\theta^*(x) = \int k(t; x, h) \tilde{\mathbf{u}}_\theta(t) dt$  (Basu and Lindsay, 1994). For any random sample  $\hat{g}_n$  coming from underlying density  $g = f_\theta$ , the asymptotic limit of  $MLE^*$  estimating equations is (3.10),

$$\lim_{n \rightarrow \infty} \int_x \mathbf{u}_\theta^*(x) d\hat{G}_n(x) = \int \mathbf{u}_\theta^*(x) dF_\theta(x) = \int \tilde{\mathbf{u}}_\theta(x) f_\theta^*(x) dx$$

Given the full efficiency of  $MLE^*$  under transparent kernels, such as a normal kernel  $k(x; t, h)$  for normal model (Basu and Lindsay, 1994), the estimator as solution to (3.10) is

also fully efficient. The estimating equation (3.10) is simplified by removing the smoothing kernel on score function  $\mathbf{u}_\theta$  and  $\hat{g}_n$  (Basu and Lindsay, 2004),

$$\int_x w_1(\delta_n^*(x)) \tilde{\mathbf{u}}_\theta(x) \hat{g}_n^*(x) dx \approx \int_x w_1(\delta_n^*(x)) \mathbf{u}_\theta(x) \hat{g}_n(x) dx = 0 \quad (3.11)$$

which is equivalent to WLE with the inverse weight  $w_1$  in (3.9).

However, the simplification (3.11) no longer keeps the efficiency of the estimator. We restore the efficiency (according to Theorem 3.2) by replacing the inverse weight  $w_1$  by a cubic curve at the neighborhood of  $\delta^* = 0$ . Define the *cubic-inverse weight*

$$w_2(\delta_n^*(x), c, x_+, x_-) = \begin{cases} w_1(\delta_n^*(x)) & \text{if } \delta_n^*(x) \notin [x_-, x_+] \\ c[\delta_n^*(x)]^3 + 1 & \text{if } \delta_n^*(x) \in [x_-, x_+] \end{cases} \quad (3.12)$$

Any one of the three parameters  $(c, x_+, x_-)$  determines the other two by solving the continuity equations at the positive real root  $\delta^* = x_+$ . The negative real root  $x_-$  may not exist, in which case we let  $w_2 = w_1$  when  $\delta^* \geq x_+$  and be the cubic curve when  $\delta^* < x_+$ . Figure 3.1(b) gives some examples of weight  $w_2$  with different  $(c, x_+, x_-)$ , in which the solid line represents the inverse weight  $w_1$ .

### 3.3.1 Selection of smoothing bandwidth $h$

The choice of bandwidth  $h$  in the smoothing kernel  $k(x; t, h)$  plays an important role in determining the cubic coefficient  $c$  and the appropriate right-truncation point. Let  $g$  be an arbitrary continuous density and  $X_1, \dots, X_n$  be a random sample with empirical density  $\hat{g}_n$ . Define the nonparametric Pearson residual  $\tilde{\delta}_n^*$  calculated by smoothed empirical densities at a narrow bandwidth  $h_1$  and a wide bandwidth  $h_2 = 2h_1$ ,

$$\tilde{\delta}_n^*(x) = \hat{g}_1^*(x) / \hat{g}_2^*(x) - 1 \quad (3.13)$$

where  $\hat{g}_i^*(x) = \int k(x; t, h_i) \hat{g}_n(t) dt$  for  $i = 1, 2$ . Our criterion of choosing  $h = h_1$  is that the distribution of  $\tilde{\delta}_n^*(x)$  does not depend heavily on density  $g$  nor sample size  $n$ , since the subsequent choice of cubic coefficient  $c$  and truncation threshold  $l_2$  will be decided by the approximated distribution of  $\tilde{\delta}_n^*(x)$ .

It has been suggested that for the  $k$ -component normal mixtures model, one may use the smoothing bandwidth  $h^2 = \kappa \sum_{i=1}^k \hat{p}_i \hat{\sigma}_i^2$  in the iteration algorithm, where  $\kappa$  takes values roughly in range  $(.001, .05)$  based on the criterion of average down-weight (Markatou, 2000). However, this choice of bandwidth does not satisfy our criterion; as it's difficult to generalize it to an arbitrary density  $g$ , and unrealistic to use the same bandwidth for both large and small sample size  $n$ . Therefore, we propose a new choice of bandwidth

$$h = \text{mad}(x)/\sqrt{n} \tag{3.14}$$

where MAD stands for median-absolute-deviation. We compare the bandwidth (3.14) with the standard R (R Development Core Team, 2011) functions of bandwidth selection for Gaussian kernels `bw.*`. The choices in R function include replacing the `*` by `nrd0` or `nrd` representing the "rule-of-thumb" choice (Silverman, 1986; Scott, 1992); `ucv` or `bcv` representing unbiased or biased cross-validation (Scott and Terrell, 1987); and `SJ` representing the method using pilot estimation of derivatives (Sheather and Jones, 1991). All these choices contain different powers of the sample size  $n$ . In Figure 3.2, the nonparametric Pearson residuals  $\tilde{\delta}_n^*$  for various densities  $g$ 's are calculated and we provide the boxplots of  $\tilde{\delta}_n^*$  under sample size  $n = 500$  and  $100$ , with  $y$ -axis range  $(-0.6, 0.6)$  to magnify the boxes and exclude the points outside the whiskers. The six sets of box-plots, from left to right, are from density  $g$  consisting of `Unif(0, 1)`, `Beta(5, 3)`, `N(0, 1)`, `t4`, `F7,10` and our finite normal mixture density in Figure 3.3(a) of next section, respectively. Within each set of box-plots, the five parallel boxes are, from left to right, using bandwidth (3.14), "nrd", "ucv", "bcv"

and "SJ", respectively.

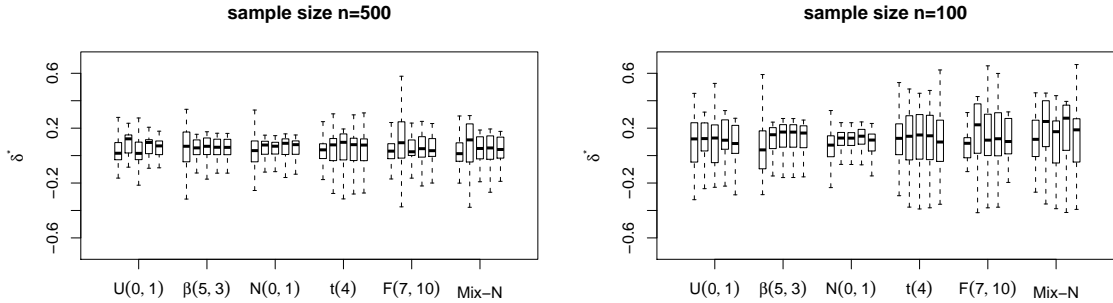


Figure 3.2: Distribution of  $\delta^*$  for 6 generating distributions. For each distribution the 5 boxplots represent smoothing bandwidths  $h$  for (3.14), `nrd`, `ucw`, `bcw`, and `SJ`.

The bandwidth (3.14), which is the first box-plot of each set, best satisfies our criterion that the distribution of  $\tilde{\delta}_n^*$  is closer to symmetry around 0 and stays roughly the same across different models and sample sizes. The other choices, however, have varying performances under our criterion for different sample sizes, and are overall not as good as bandwidth (3.14). Thus we suggest that bandwidth (3.14) is preferred under our criterion regardless of the underlying distribution and sample size. Nonetheless, the choice of  $h$  is still wide open for different weights, and researchers may choose other  $h$  for their own smoothing problems.

### 3.3.2 Selection of cubic coefficient $c$

The cubic coefficient  $c$  is chosen such that within the inter-quartile range of the smoothed Pearson residuals (3.13), the cubic-inverse weight  $w_2(\delta_n^*)$  falls in the interval  $1 \pm \Delta w$ . Figure 3.2 shows that when using  $h = \text{MAD}(x)/\sqrt{n}$ , the inter-quartile range of nonparametric Pearson residuals  $\tilde{\delta}_n^*$  is safely covered by the interval  $(-0.2, 0.2)$ ; thus the cubic coefficient  $c$  is determined by solving  $0.2^3|c| \leq \Delta w$ . Let  $\Delta w = 0.1$ , then  $|c| \leq 12.5$ . On the other hand, we place another restraint that  $|c| \geq 8$  to avoid too much cubic modification. As

shown in Figure 3.1(b), different  $c$  values in the range  $(-12.5, -8)$  hardly affect the shape of cubic-inverse curve, thus we may pick our choice as we like. The weight  $w_3$  with  $c = -8$  is added to Figure 3.1(a) for comparison.

A refinement to the cubic-inverse weight  $w_2$  is to truncate the weight when  $\delta^*$  exceeds a threshold  $l_2$ , since the smoothed Pearson residuals for most continuous densities are right-skewed with a very heavy tail, which is not shown in Figure 3.2. In order to further reduce the influence of extreme outliers, we iteratively assign weight zero to sample points with Pearson residuals  $\delta_n^*$  greater than a threshold  $l_2$ . Define the *truncated cubic-inverse* weight

$$w_3(x) = w_2(\delta_n^*(x)) \cdot \mathbf{I}_{\{\delta_n^*(x) < l_2\}} \quad (3.15)$$

One possible choice is to let  $l_2$  be the 95% percentile of  $\delta_n^*$  calculated at each iteration step.

### 3.4 The iterative algorithm

In this section we briefly outline the steps for obtaining the truncated cubic-inverse weight (3.15) and provide an iterative algorithm for solving GWLE (3.1) under the finite mixture model of exponential-family distributions

$$f(x_i; \boldsymbol{\xi}, \mathbf{p}) = \sum_{s=1}^d \pi_s \phi(x; \boldsymbol{\xi}_s) \quad (3.16)$$

where  $\boldsymbol{\xi}_s = (\xi_{s1}, \dots, \xi_{sQ})^t$  are the canonical parameters and  $\boldsymbol{\eta}_s = (\eta_{s1}, \dots, \eta_{sQ})^t$  the mean parameters of the  $s$ th component. Let  $\mathbf{T}_X = (\mathbf{T}_{X,1}, \dots, \mathbf{T}_{X,Q})^t$  be the sufficient statistics of  $X$ . Further discussion of the mixture model (3.16) can be found in Appendix A.1.2. In this work, we do not step away from the discussion of the selection of the number of mixture components, which itself constitutes a major topic of interest in this field (see Turner and West, 1993; Roeder, 1994; West, 1997; Richardson and Green, 1997; Stephens,

2000; Ishwaran et al., 2001; Ishwaran and James, 2002). Instead, we assume that the number of components  $k$  is fixed and known in our problem.

First of all, the choice of starting value is critical for most iterative algorithms. We obtain a tentative partition of the data into  $k$  groups through one of the existing methods such as  $k$ -means and trimmed  $k$ -means (Cuesta-Albertos et al., 1997), robust clustering (Woodward et al., 1984) or the watershed algorithm (Vincent and Soille, 1991). All of these partitions have been developed under certain robustness considerations and produce their own corresponding starting values. However, for our iterative algorithm with updating steps (3.17a) and (3.17b), we can greatly reduce the number of iterations if we use these partitions to produce our "weighted starting values", which is described in detail below. Within each group, we calculate the nonparametric smoothed Pearson residual (3.13) and the corresponding weight  $\tilde{w}_i = w(\tilde{\delta}_n^*(x_i))$ . We use weight  $\tilde{w}_i$  to obtain the starting values  $\boldsymbol{\pi}^{(0)}$ , as the weighted proportion of the sample size in each group  $\pi_s^{(0)} = \sum_{x_i \in \mathbb{S}} \tilde{w}_i / \sum \tilde{w}_i$ , where  $\mathbb{S}$  represents the set of random sample in group  $s$ , and the weighted moments of the random sample within each group. As every exponential family distribution has a one-to-one correspondence to its first few moments, we could get the starting values  $\boldsymbol{\xi}_s^{(0)}$  through simple transformation of these weighted moments. Specifically, when the model (3.16) is a mixture of normals, the starting  $\mu_0$ 's are the weighted medians and  $\sigma_0$ 's are the weighted median-absolute-deviation (MAD)'s of each group. The steps of solving model (3.16) is given below,

1. Let  $\boldsymbol{\theta}^{(\gamma-1)} = (p^{(\gamma-1)}, \boldsymbol{\eta}^{(\gamma-1)})$  or  $\boldsymbol{\theta}^{(\gamma-1)} = (p^{(\gamma-1)}, \boldsymbol{\xi}^{(\gamma-1)})$  be the estimates from  $(\gamma - 1)$ th iteration.
2. Calculate Pearson residual  $\delta_n^*$  (2.10) with  $\boldsymbol{\theta}^{(\gamma-1)}$ .

3. Let  $w^{(\gamma-1)}(x_i) = w(x_i, \delta_n^*, f_{\boldsymbol{\theta}^{(\gamma-1)}})$ . For truncated cubic-inverse weight (3.15), choose  $c$  and  $l_2$  as suggested in previous discussion. Let

$$p_w^{(\gamma-1)}(s; x_i) = w^{(\gamma-1)}(x_i) \pi_s \phi(x_i; \boldsymbol{\xi}_s^{(\gamma-1)}) / f(x_i; \boldsymbol{\theta}^{(\gamma-1)}), \quad s = 1, \dots, k$$

4. Obtain the bias adjustment term  $\mathbf{a}(\boldsymbol{\xi}^{(\gamma-1)}, \mathbf{p}^{(\gamma-1)})$  by numeric integration. Let  $\mathbf{a}_{:\xi_{sq}}$  and  $\mathbf{a}_{:p_s}$  be the elements of  $\mathbf{a}$  corresponding to  $\xi_{sq}$  and  $p_s$ .

5. Update  $p_s$  and  $\eta_{sq}$ ,  $s = 1, \dots, k$ ,  $q = 1, \dots, Q$ , by

$$p_s^{(\gamma)} = \frac{1}{n} \sum_{i=1}^n p_w^{(\gamma-1)}(s; x_i) - p_s \mathbf{a}_{:p_s}(\boldsymbol{\xi}^{(\gamma-1)}, \mathbf{p}^{(\gamma-1)}) \quad (3.17a)$$

$$\eta_{sq}^{(\gamma)} = \left[ \sum_{i=1}^n p_w^{(\gamma-1)}(s; x_i) \right]^{-1} \left( \sum_{i=1}^n \mathbf{T}_{x_i, q} p_w^{(\gamma-1)}(s; x_i) - n \mathbf{a}_{:\xi_{sq}}(\boldsymbol{\xi}^{(\gamma-1)}, \mathbf{p}^{(\gamma-1)}) \right) \quad (3.17b)$$

The derivation of the updating steps (3.17) are presented in Appendix A.1.3 and A.1.4. The corresponding updating steps for finite normal mixtures model are included in this algorithm with  $\boldsymbol{\eta}_s = (\mu_s, \mu_s^2 + \sigma_s^2)^t$  and  $\mathbf{T}_x = (x, x^2)^t$ ; this special case together with density power weight (3.2) is discussed in Fujisawa and Eguchi (2006).

The convergence of the series of estimates  $\hat{\boldsymbol{\theta}}$ 's is equivalent to the convergence of the series of weights  $w$ , since they depend on each other iteratively, i.e.  $\boldsymbol{\theta}^{(\gamma-1)} \Rightarrow w^{(\gamma)} \Rightarrow \boldsymbol{\theta}^{(\gamma)} \Rightarrow w^{(\gamma+1)} \Rightarrow \boldsymbol{\theta}^{(\gamma+1)}$ . The convergence criterion based on weights  $w$  rather than estimates  $\boldsymbol{\theta}$  is more reliable and not specific to particular model. Let  $w^{\text{new}}(x) = w(x, \delta_n^*, f_{\boldsymbol{\theta}^{\text{new}}})$ . Similar to  $R^2$  in regression, the criterion is set to be

$$\frac{n^{-1} \|w^{\text{new}} - w\|^2}{\text{var}(w^{\text{new}})} < \varepsilon \quad (3.18)$$

where function  $\text{var}()$  calculates the sample variance. We use  $\varepsilon = 1\%$  in the simulation studies.

### 3.5 The simulation studies

The simulation study is carried out on a scenario of three components normal mixtures sketched in Figure 3.3 with sample size  $n = 400$ . We generate 1000 data sets from each scenarios: a clean density (a)  $.2N(-10, 3) + .5N(0, 5) + .3N(15, 4)$ ; and a contaminated density (a1) where outliers of 5% are added at  $N(-18, 4)$  and the third component is replaced by a shifted and re-scaled F-distribution with the same mode as the original normal density. The overlaps (Woodward et al., 1984) between adjacent components in scenario (a), the areas highlighted, are 7.2% and 3.7%, respectively.

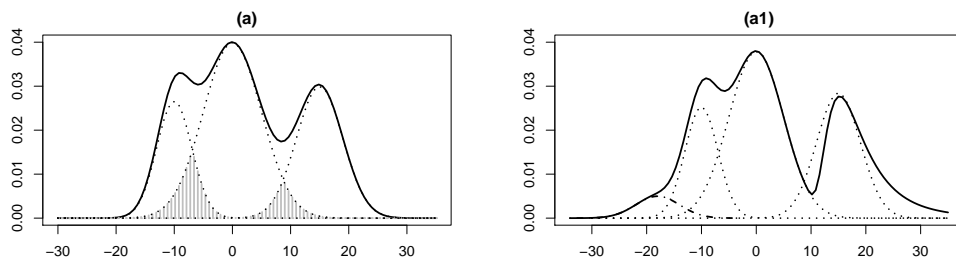


Figure 3.3: (a) clean data; (a1) contaminated data.

The estimators compared are the MLE and the GWLEs with weights of density power (3.2), symmetric  $\chi^2$  (3.3), negative exponential (3.4), cubic-inverse (3.12) and truncated cubic-inverse (3.15). The starting partitions are obtained from robust clustering (Woodward et al., 1984). The tuning parameter for density power weight (3.2) is selected from  $\beta = .15, .20, .25, .30$  by minimizing Cramer-von Mises divergence through cross-validation (Fujisawa and Eguchi, 2006). The kernel smoothing bandwidth is  $h = \text{mad}(X)/\sqrt{n}$ . The truncated cubic-inverse weight (3.15) has the cubic coefficient  $c = -8$ . The convergence criterion for MLE is a set of pre-specified thresholds on the  $L_2$  norms of  $\|p^{\text{new}} - p\|^2$ ,  $\|\mu^{\text{new}} - \mu\|^2$  and  $\|\sigma^{\text{new}} - \sigma\|^2$ ; while for all GWLE estimators we use criterion (3.18). Figure 3.4 shows

boxplots of errors of different estimators for  $\mu$ 's,  $\sigma$ 's and  $p$ 's, where the scale of  $p$ 's are multiplied by a factor of 10 in order to enlarge the boxplots. The truncated cubic-inverse weight (3.15) generally retains efficiency under the uncontaminated scenario (lower panel of Figure 3.4) and provides the overall best estimation under the contaminated scenario (upper panel of Figure 3.4), in terms of smaller bias and mean squared error. This advantage is especially obvious when estimating the scale parameters  $\sigma_1$  and  $\sigma_3$ , where either inliers or outliers are present.

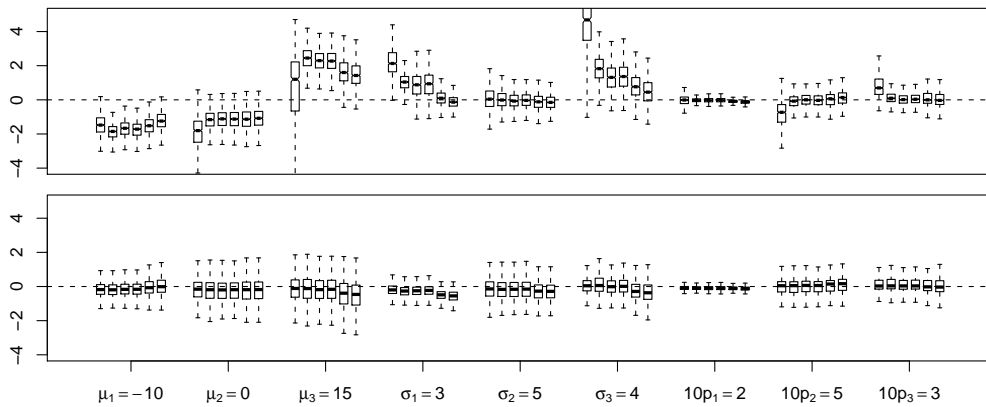


Figure 3.4: Box-plots of estimation errors of contaminated (a1) at top vs. clean (a) at bottom, with sample size 400. From left to right: MLE, density power, symmetric  $\chi^2$ , negative exponential, cubic-inverse, truncated cubic-inverse.

### 3.6 The tendon fibril data example

We apply the GWLE's with various robust weights to the tendon fibril data. It was suggested (Zhang et al., 2006) that three-components finite-normal-mixture model is appropriate for 2 month or older data, which provides insight into the mechanisms of collagen fibrillogenesis.

The MLE and robust GWLE estimates with various weights are calculated, using the

microscopic fields from the postnatal three-month mice (P3M) as the data example. In Figure 3.5, the estimates from selected microscopic fields (which are previously shown in Fig 1.1) are plotted as the density curves against the histograms and the Gaussian smooth curves of the fibril measures, with MLE, negative-exponential weight  $\text{Neg.Exp}(3.4)$  and cubic-inverse weight  $\text{Cub.Inv}(3.12)$  from the top to bottom.

The first of the three selected fields shows an "almost clean" scenario, in which the MLE and all robust GWLE estimates give similar results. The second and third fields contain small clusters of outliers appearing to the left of the data, which is a common contamination present in the tendon fibril data. The cubic-inverse weights  $\text{Cub.Inv}(3.12)$  (as well as the truncated cubic-inverse weights (3.15), which is not shown in Figure 3.5) is much less affected by the outliers. We prefer the (truncated-)cubic-inverse weights over the previously robust GWLE's from the literature based on the simulation results presented in section 3.5.

### 3.7 Summary of Chapter 3

In this chapter, we approached the first task in the analysis of the tendon fibril data, presented in Chapter 1, by developing the robust generalized weighted likelihood estimators with (truncated-)cubic-inverse weights. We also made numerical improvements to the EM-like algorithm in selection of the starting values and the specification of convergence criterion. Our approaches are more appropriate for the kind of contamination presented in the tendon fibril data, compared to previous robust GWLE's in literature. This work has been published in Zhan et al. (2011).

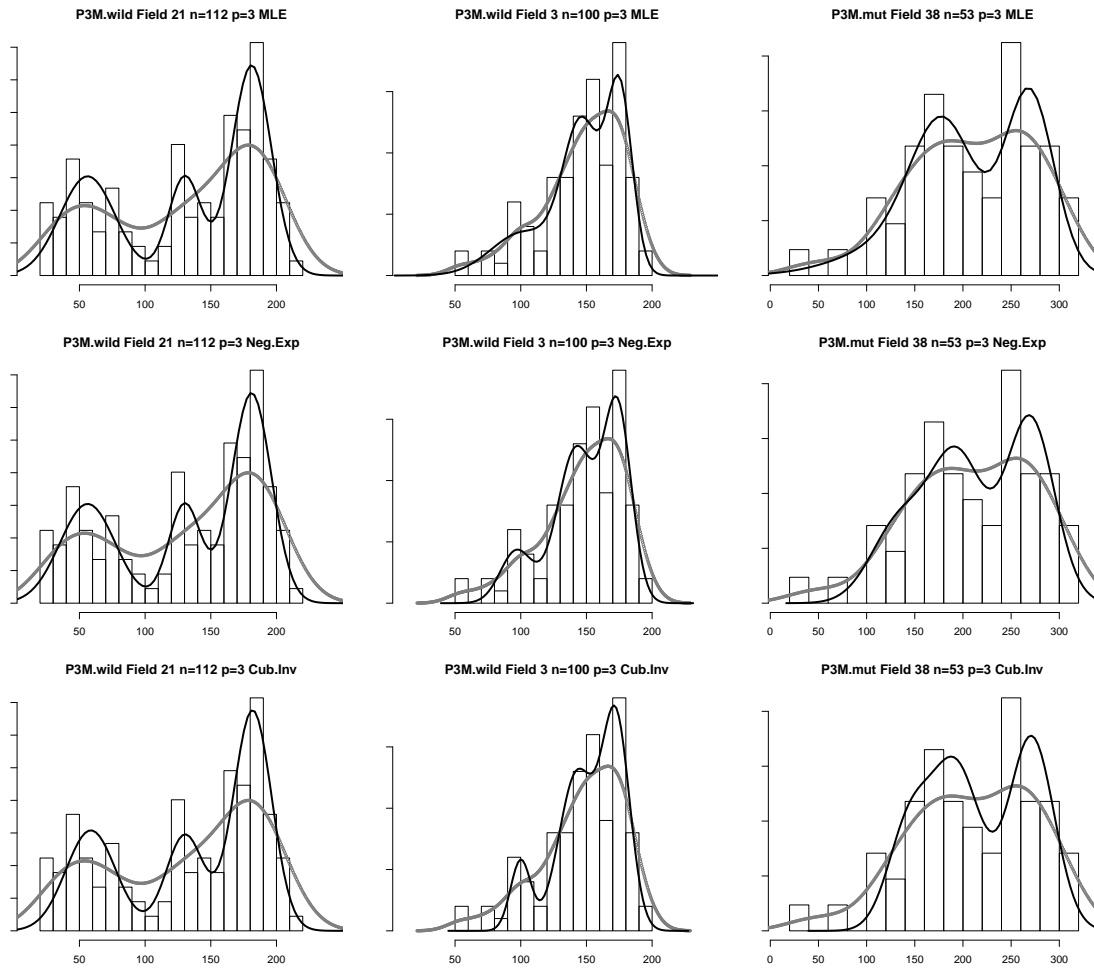


Figure 3.5: The robust estimates of selected microscopic fields (previously shown in Fig 1.1) of the fibril diameter measures from postnatal 3-month mice. Upper row: MLE estimates. Middle row: negative-exponential estimates (Bhandari et al., 2006). Lower row: cubic-inverse estimates (Zhan et al., 2011).

## Chapter 4

# Robust estimation of generalized linear mixed model for finite normal mixtures

The generalized linear mixed model for finite normal mixtures (GLMFM) assumes a linear relationship (after proper transformation) between the predictors and the mixture proportions, means and standard deviations of the lowest-level distributions of the response. We develop a robust joint estimation for fitting the GLMFM model to the hierarchical data with potential contaminations present in all clustering levels.

This chapter is organized as follow. In section 4.1, we review the standard-two-stage (STS) approach to the GLMFM model (Chervoneva et al., 2011), and use the STS estimates as the starting values of our approach. In section 4.2, we introduce the joint estimation to the GLMFM model. In section 4.4, we derive the asymptotic distribution of the fixed effect estimates and provide the empirical formulas. In section 4.3, we discuss the robust joint estimation, which is summarized in Algorithm 4.1, and some computational simplifications.

In section 4.5, we describe the simulation study to illustrate the advantages of robust joint estimation over the STS approach. In section 4.6, we apply the robust joint estimation to the tendon fibril data and compare the results with previous studies.

## 4.1 STS approach to the GLMFM model

The STS approach treats the joint GLMFM model (2.1); (2.5) as two separate models: the stage-1 subject-specific model (2.1) and the stage-2 linear mixed model (2.5). In the first stage, we obtain the mixture parameters estimates  $(\hat{\boldsymbol{\mu}}_{ij}, \hat{\boldsymbol{\sigma}}_{ij}, \hat{\boldsymbol{\pi}}_{ij})$  from each animal  $i$  and field  $j$ , using MLE or one of the robust estimators such as Dens.Pow or Sym.Chi. An exploratory analysis of the stage-1 subject-specific estimates is performed to screen for potentially outliers, the significant covariates such as animal genotype and body weight, as well as the variance structure of the animal and/or field random effects. Then the stage-1 estimates are transformed, as necessary, so that the response vector  $\hat{\boldsymbol{\eta}}_{ij} = \boldsymbol{\eta}_{ij}(\hat{\boldsymbol{\mu}}_{ij}, \hat{\boldsymbol{\sigma}}_{ij}, \hat{\boldsymbol{\pi}}_{ij})$  has a linear relationship with the covariates and the normal distribution assumption is appropriate for the error terms and random effects. In the second stage, the LME model (2.5) is fitted to the linear response vector  $\hat{\boldsymbol{\eta}} = \{\hat{\boldsymbol{\eta}}_{ij} : i, j\}$  using either the REML estimators (Harville, 1977) or a robust version (Welsh and Richardson, 1997). The population parameter estimates from the stage-2 model (2.5) are reported as the STS estimates of the GLMFM model.

The STS approach allows for the flexibility in choosing the robust or non-robust estimation methods independently for each stage, as well as the possibility of constructing the stage-2 model based on only a subset of the stage-1 parameter estimates such as  $\hat{\boldsymbol{\eta}}_{ij} = \boldsymbol{\eta}_{ij}(\hat{\boldsymbol{\mu}}_{ij}, \hat{\boldsymbol{\sigma}}_{ij})$ . However, the STS approach uses the subject-specific estimates  $\hat{\boldsymbol{\eta}}_{ij}$ , which may be very poor if there are not enough observations on each subject. In this work,

we use the STS estimates  $(\hat{\beta}_{\text{STS}}, \hat{\theta}_{\text{STS}})$  as the starting values of our joint estimation procedure. The implementation and results from using the STS approach for the tendon fibril data are described in section 4.6.1.

## 4.2 Joint estimation of the GLMFM model

The joint estimation of the GLMFM model (2.1); (2.5) is an iterative procedure, which involves an alteration between calculating the BLUP-like estimates of the fixed and random effect and estimating the covariance parameters. We describe a joint estimation procedure and show that it reduces to the procedure described in sections 8 and 9 of McGilchrist and Yau (1995) under certain conditions.

Given the GLMFM model (2.1); (2.5), we consider a conditional divergence measure  $l_1(\beta|\mathbf{b}; \mathbf{y})$  which is based on the observations  $\mathbf{y}$  given the random effects  $\mathbf{b}$ , and a divergence measure  $l_2(\mathbf{b}, \theta)$  of the random effects  $\mathbf{b}$ . For the standard linear mixed model, the conditional divergence  $l_1$  is the conditional log-likelihood and the random effect divergence  $l_2$  is the log-likelihood of the multivariate normal distribution of the random effect. Therefore, the summation  $l_1 + l_2$  is the joint log-likelihood  $l$  of the observations  $\mathbf{y}$  and the random effects  $\mathbf{b}$ , i.e. the objective function (2.16) of the BLUP procedure. In this work, we choose different divergence measures for  $l_1$  and  $l_2$  for more complicated model structures as well as for robust purposes. The tendon fibril data are substantially contaminated in the distribution of fibril diameter measurements on each microscopic field. Lindsay (1994)'s divergences (2.11) are good target functions for the robust estimation of the finite normal mixture model (Zhan et al., 2011). Therefore, we choose the family of Lindsay (1994)'s divergences as the

conditional divergence  $l_1^D$  on each microscopic field,

$$l_1^D(\boldsymbol{\beta}|\mathbf{b}; \mathbf{y}_{ij}) = \text{Lindsay.Div}\left(f^*(\mathbf{y}_{ij}; \boldsymbol{\eta}_{ij}(\boldsymbol{\beta}, \mathbf{b})), \hat{g}^*(\mathbf{y}_{ij})\right) \quad \text{as in (2.11)} \quad (4.1)$$

where  $f^*(\cdot)$  and  $\hat{g}^*(\cdot)$  are the same-kernel-smoothed conditional model density (2.1) and conditional empirical density, respectively. We define the negative value of the second order derivative of the conditional divergence  $l_1^D(\boldsymbol{\beta}|\mathbf{b}; \mathbf{y}_{ij})$  with respect to  $\boldsymbol{\eta}_{ij}$ , which would be the Hessian matrix if  $l_1^D$  were the log-likelihood, as

$$\mathbf{H}_{ij}(\boldsymbol{\beta}, \mathbf{b}) = \mathbf{H}_{ij}\left(\boldsymbol{\eta}_{ij}(\boldsymbol{\beta}, \mathbf{b})\right) = -\frac{\partial^2 l_1^D(\boldsymbol{\beta}|\mathbf{b}; \mathbf{y}_{ij})}{\partial \boldsymbol{\eta}_{ij} \partial \boldsymbol{\eta}_{ij}'} \quad (4.2)$$

The derivation of equation (4.2) is given in Appendix A.1.5. We define the cross product of the first order derivative, which would be the cross product of the score function if  $l_1^D$  were the log-likelihood, as

$$\mathbf{J}_{ij}(\boldsymbol{\beta}, \mathbf{b}) = \mathbf{J}_{ij}\left(\boldsymbol{\eta}_{ij}(\boldsymbol{\beta}, \mathbf{b})\right) = \left(\frac{\partial l_1^D(\boldsymbol{\beta}|\mathbf{b}; \mathbf{y}_{ij})}{\partial \boldsymbol{\eta}_{ij}}\right) \left(\frac{\partial l_1^D(\boldsymbol{\beta}|\mathbf{b}; \mathbf{y}_{ij})}{\partial \boldsymbol{\eta}_{ij}}\right)' \quad (4.3)$$

Let the total conditional divergence of all microscopic fields and animals be  $l_1^D(\boldsymbol{\beta}|\mathbf{b}; \mathbf{y}) = \sum_{i,j} l_1^D(\boldsymbol{\beta}|\mathbf{b}; \mathbf{y}_{ij})$ . In this work, the random effect divergence  $l_2$  remains the negative of the multivariate normal log-likelihood. Define the total divergence measure  $l^D(\boldsymbol{\beta}, \mathbf{b}, \boldsymbol{\theta}; \mathbf{y})$  for the GLMFM model (2.1); (2.5),

$$l^D(\boldsymbol{\beta}, \mathbf{b}, \boldsymbol{\theta}; \mathbf{y}) = l_1^D(\boldsymbol{\beta}|\mathbf{b}; \mathbf{y}) + l_2(\mathbf{b}, \boldsymbol{\theta}) \quad (4.4)$$

The parameters of the GLMFM model are obtained by minimizing the total divergence measure (4.4)

$$(\boldsymbol{\beta}_l(\boldsymbol{\theta}), \mathbf{b}_l(\boldsymbol{\theta})) = \arg_{(\boldsymbol{\beta}, \mathbf{b})} \min l^D(\boldsymbol{\beta}, \mathbf{b}, \boldsymbol{\theta}(\boldsymbol{\beta}, \mathbf{b}); \mathbf{y}) \quad (4.5)$$

The term "BLUP" is no longer appropriate, as the linear relationship (2.17) between the BLUP's  $(\boldsymbol{\beta}_l(\boldsymbol{\theta}), \mathbf{b}_l(\boldsymbol{\theta}))$  and the observations  $\mathbf{y}$  in standard linear mixed model is voided.

Nevertheless, our approach carries over the idea of minimizing the total divergence measure based on the observed data  $\mathbf{y}$  and the unobserved random effects  $\mathbf{b}$  in the BLUP procedure. We describe the following two steps in each iteration of the joint estimation.

The first step is that for given starting values  $(\boldsymbol{\beta}_0, \mathbf{b}_0, \boldsymbol{\theta}_0)$ , solve for the global minima  $(\hat{\boldsymbol{\beta}}_l(\boldsymbol{\theta}_0), \hat{\mathbf{b}}_l(\boldsymbol{\theta}_0))$  of the total divergence measure  $l^D(\boldsymbol{\beta}, \mathbf{b}, \boldsymbol{\theta}_0; \mathbf{y})$ . The minimization is accomplished by the Newton-Raphson iteration with the updating step

$$\begin{pmatrix} \hat{\boldsymbol{\beta}}_l(\boldsymbol{\theta}_0) \\ \hat{\mathbf{b}}_l(\boldsymbol{\theta}_0) \end{pmatrix} := \begin{pmatrix} \boldsymbol{\beta}_0 \\ \mathbf{b}_0 \end{pmatrix} - \begin{pmatrix} \mathbf{A}'\hat{\mathbf{H}}_{\boldsymbol{\theta}_0}\mathbf{A} & \mathbf{A}'\hat{\mathbf{H}}_{\boldsymbol{\theta}_0}\mathbf{B} \\ \mathbf{B}'\hat{\mathbf{H}}_{\boldsymbol{\theta}_0}\mathbf{A} & \mathbf{B}'\hat{\mathbf{H}}_{\boldsymbol{\theta}_0}\mathbf{B} + \boldsymbol{\Sigma}^{-1}(\boldsymbol{\theta}_0) \end{pmatrix}^{-1} \begin{pmatrix} \partial l^D / \partial \boldsymbol{\beta}_0 \\ \partial l^D / \partial \mathbf{b}_0 \end{pmatrix} \quad (4.6)$$

where the block-diagonal matrix  $\hat{\mathbf{H}}_{\boldsymbol{\theta}_0} = \text{Diag}\{\hat{\mathbf{H}}_{ij, \boldsymbol{\theta}_0}\} = \text{Diag}\{\mathbf{H}_{ij}(\hat{\boldsymbol{\beta}}_l(\boldsymbol{\theta}_0), \hat{\mathbf{b}}_l(\boldsymbol{\theta}_0))\}$ .

The second step is to update the covariance parameter  $\hat{\boldsymbol{\theta}}$  conditional on the current estimates  $(\hat{\boldsymbol{\beta}}_l(\boldsymbol{\theta}_0), \hat{\mathbf{b}}_l(\boldsymbol{\theta}_0))$ . This is accomplished through the construction of a pseudo normal-normal linear mixed model. The rationale is that the total divergence measure  $l^D(\boldsymbol{\beta}, \mathbf{b}, \boldsymbol{\theta}_0; \mathbf{y})$  can be approximated by a multivariate normal distribution (i.e. quadratically) in the neighborhood of its global maxima  $(\hat{\boldsymbol{\beta}}_l(\boldsymbol{\theta}_0), \hat{\mathbf{b}}_l(\boldsymbol{\theta}_0))$ . Since we assume that the divergence  $l_2$  of the random effects is multivariate normal, it follows that the conditional divergence  $l_1^D$  is quadratically approximated in the neighborhood of the global maxima  $(\hat{\boldsymbol{\beta}}_l(\boldsymbol{\theta}_0), \hat{\mathbf{b}}_l(\boldsymbol{\theta}_0))$  by

$$l_1^D(\boldsymbol{\beta}|\mathbf{b}; \mathbf{y}) \approx c - \frac{1}{2} \begin{pmatrix} \hat{\boldsymbol{\beta}}_l(\boldsymbol{\theta}_0) - \boldsymbol{\beta} \\ \hat{\mathbf{b}}_l(\boldsymbol{\theta}_0) - \mathbf{b} \end{pmatrix}' \begin{pmatrix} \mathbf{A}' \\ \mathbf{B}' \end{pmatrix} \hat{\mathbf{D}}_{\boldsymbol{\theta}_0}^{-1}(\mathbf{A}, \mathbf{B}) \begin{pmatrix} \hat{\boldsymbol{\beta}}_l(\boldsymbol{\theta}_0) - \boldsymbol{\beta} \\ \hat{\mathbf{b}}_l(\boldsymbol{\theta}_0) - \mathbf{b} \end{pmatrix} \quad (4.7)$$

The block-diagonal matrix  $\hat{\mathbf{D}}_{\boldsymbol{\theta}_0} = \text{Diag}\{\hat{\mathbf{D}}_{ij, \boldsymbol{\theta}_0}\}$  is the asymptotic error covariance matrix of the estimates  $\boldsymbol{\eta}(\hat{\boldsymbol{\beta}}_l(\boldsymbol{\theta}_0), \hat{\mathbf{b}}_l(\boldsymbol{\theta}_0))$ , in the sandwich form

$$\hat{\mathbf{D}}_{ij, \boldsymbol{\theta}_0} = n_{ij}^{-1} \cdot \left( \text{E} \left[ \hat{\mathbf{H}}_{ij, \boldsymbol{\theta}_0} \right] \right)^{-1} \cdot \text{E} \left[ \hat{\mathbf{J}}_{ij, \boldsymbol{\theta}_0} \right] \cdot \left( \text{E} \left[ \hat{\mathbf{H}}_{ij, \boldsymbol{\theta}_0} \right] \right)^{-1} \quad (4.8)$$

where  $\hat{\mathbf{J}}_{ij, \boldsymbol{\theta}_0} = \mathbf{J}_{ij}(\hat{\boldsymbol{\beta}}_l(\boldsymbol{\theta}_0), \hat{\mathbf{b}}_l(\boldsymbol{\theta}_0))$ . The expectation in equation (4.8) is taken with respect to the finite normal distribution of the fibril diameters  $\mathbf{y}_{ij}$ . If we define the pseudo-observations

by

$$\hat{\mathbf{y}}_l(\boldsymbol{\theta}_0) = \mathbf{A}\hat{\boldsymbol{\beta}}_l(\boldsymbol{\theta}_0) + \mathbf{B}\hat{\mathbf{b}}_l(\boldsymbol{\theta}_0) \quad (4.9)$$

the quadratic term in equation (4.7) becomes  $(\hat{\mathbf{y}}_l(\boldsymbol{\theta}_0) - \mathbf{A}\boldsymbol{\beta} - \mathbf{B}\mathbf{b})' \hat{\mathbf{D}}_{\boldsymbol{\theta}_0}^{-1} (\hat{\mathbf{y}}_l(\boldsymbol{\theta}_0) - \mathbf{A}\boldsymbol{\beta} - \mathbf{B}\mathbf{b})$ ,

which implies that  $\hat{\mathbf{y}}_l(\boldsymbol{\theta}_0)$  satisfies an approximate LME model

$$\hat{\mathbf{y}}_l(\boldsymbol{\theta}_0) = \mathbf{A}\boldsymbol{\beta} + \mathbf{B}\mathbf{b} + \boldsymbol{\varepsilon}, \quad \text{where } \mathbf{b} \sim N(\mathbf{0}, \boldsymbol{\Sigma}(\boldsymbol{\theta})) \text{ and } \boldsymbol{\varepsilon} \sim N(\mathbf{0}, \hat{\mathbf{D}}_{\boldsymbol{\theta}_0}) \quad (4.10)$$

We can use the pseudo-model (4.10) to obtain the estimates of  $\hat{\boldsymbol{\theta}}$ , holding the pseudo-covariance matrix  $\hat{\mathbf{D}}_{\boldsymbol{\theta}_0}$  of the random error fixed in the estimation. We let the new starting values be  $(\hat{\boldsymbol{\beta}}_l(\boldsymbol{\theta}_0), \hat{\mathbf{b}}_l(\boldsymbol{\theta}_0), \hat{\boldsymbol{\theta}})$  and repeat the iteration until all parameters converge.

Our joint estimation procedure should not be confused with the non-iterative global-two-stage (GTS) approach (Chervoneva et al., 2006; Yeap and Davidian, 2001; Davidian and Giltinan, 1993). The GTS approach calculates the subject-specific estimates, which essentially utilizes the quadratic approximation based on the conditional log-likelihood  $l_1(\boldsymbol{\beta}|\mathbf{b}; \mathbf{y})$  alone at the neighborhood of its global maximum  $(\hat{\boldsymbol{\beta}}_{l_1}, \hat{\mathbf{b}}_{l_1})$ . The corresponding quadratic term  $(\hat{\mathbf{y}}_{l_1} - \mathbf{A}\boldsymbol{\beta} - \mathbf{B}\mathbf{b})' \hat{\mathbf{D}}_{l_1}^{-1} (\hat{\mathbf{y}}_{l_1} - \mathbf{A}\boldsymbol{\beta} - \mathbf{B}\mathbf{b})$ , where  $\hat{\mathbf{y}}_{l_1} = \mathbf{A}\hat{\boldsymbol{\beta}}_{l_1} + \mathbf{B}\hat{\mathbf{b}}_{l_1}$  and  $\hat{\mathbf{D}}_{l_1}^{-1} = -\partial^2 l_1 / \partial \boldsymbol{\eta} \partial \boldsymbol{\eta}' |_{\hat{\boldsymbol{\beta}}_{l_1}, \hat{\mathbf{b}}_{l_1}}$ , suggest the pseudo-model  $\hat{\mathbf{y}}_{l_1} = \mathbf{A}\boldsymbol{\beta} + \mathbf{B}\mathbf{b} + \boldsymbol{\varepsilon}$  with  $\mathbf{b} \sim N(\mathbf{0}, \boldsymbol{\Sigma}(\boldsymbol{\theta}))$  and  $\boldsymbol{\varepsilon} \sim N(\mathbf{0}, \hat{\mathbf{D}}_{l_1})$ . The ML or REML estimates  $(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\theta}})$  of the pseudo-model, holding matrix  $\hat{\mathbf{D}}_{l_1}$  fixed in estimation, are reported. In the GTS approach, the quadratic approximation does not involve the covariance parameter  $\boldsymbol{\theta}$ . The maxima  $(\hat{\boldsymbol{\beta}}_{l_1}, \hat{\mathbf{b}}_{l_1})$ , as well as the pseudo-observations  $\hat{\mathbf{y}}_{l_1}$ , are essentially the subject-specific estimates of the conditional model (2.1). Therefore, the GTS approach is indeed an "inference based on individual estimates" (Davidian and Giltinan, 1995) which fails to account for the correlation of the random effect.

### 4.3 Robust estimation of covariance parameters

The pseudo-model (4.10) is difficult to solve by standard software for solving linear mixed models, such as the `lme()` function in R, because the block-diagonal elements of the random error matrix  $\hat{\mathbf{D}}_{\theta_0}$  are different for each animal and microscopic fields. On the other hand, by using negative multivariate normal log-likelihood as the random effect divergence  $l_2$  in the total divergence measure breakdown (4.4), the joint estimation may be subject to the contaminations in the field random effects.

We address these two issues simultaneously by modifying the random effects  $\hat{\mathbf{b}}_l(\theta_0)$  before passing them into the pseudo-observations (4.9). The total variance of the pseudo-observation (4.9) is  $\text{var}(\hat{\mathbf{y}}_l(\theta_0)) = \mathbf{B}\boldsymbol{\Sigma}(\theta_0)\mathbf{B}' + \hat{\mathbf{D}}_{\theta_0}$ , which implies that

$$\text{var}(\hat{\mathbf{b}}_l(\theta_0)) = \boldsymbol{\Sigma}(\theta_0) + (\mathbf{B}'\mathbf{B})^{-1}\mathbf{B}'\hat{\mathbf{D}}_{\theta_0}\mathbf{B}(\mathbf{B}'\mathbf{B})^{-1}$$

Therefore, we re-scale the random effect estimates  $\hat{\mathbf{b}}_l(\theta_0)$  to have

$$\hat{\mathbf{b}}_l^*(\theta_0) = \boldsymbol{\Sigma}^{1/2}(\theta_0) \cdot \left( \boldsymbol{\Sigma}(\theta_0) + (\mathbf{B}'\mathbf{B})^{-1}\mathbf{B}'\hat{\mathbf{D}}_{\theta_0}\mathbf{B}(\mathbf{B}'\mathbf{B})^{-1} \right)^{-1/2} \cdot \hat{\mathbf{b}}_l(\theta_0) \quad (4.11)$$

so that  $\text{var}(\hat{\mathbf{b}}_l^*(\theta_0)) = \boldsymbol{\Sigma}(\theta_0)$ , and define the corresponding pseudo-observation as  $\hat{\mathbf{y}}_l^*(\theta_0) = \mathbf{A}\hat{\boldsymbol{\beta}}_l(\theta_0) + \mathbf{B}\hat{\mathbf{b}}_l^*(\theta_0)$ . Then the pseudo-model (4.10) is replaced by

$$\hat{\mathbf{y}}_l^*(\theta_0) = \mathbf{A}\boldsymbol{\beta} + \mathbf{B}\mathbf{b}, \quad \text{where } \mathbf{b} \sim N(\mathbf{0}, \boldsymbol{\Sigma}(\theta)) \quad (4.12)$$

and  $\mathbf{B}\mathbf{b}$  is regarded as the combined random error term. In other words, the re-scaled random effect  $\hat{\mathbf{b}}_l^*(\theta_0)$  is adjusted for the estimation errors of  $\boldsymbol{\varepsilon}$ , so that we may eliminate the matrix  $\hat{\mathbf{D}}_{\theta_0}$  from the pseudo-model.

We further modify the re-scaled random effect (4.11) to down weigh the potential outliers in the field random effect. We choose not to change the random effect divergence  $l_2$  into robust divergence measures, because of small sample size of both animals and fields. We

apply the Huber's weight function or Tukey's bi-weight function  $\psi_c$  with tuning constant  $c$  to the re-scaled random effects, similar to Dueck and Lohr (2005),

$$\hat{\mathbf{b}}_l^{**}(\boldsymbol{\theta}_0) = \boldsymbol{\Sigma}^{1/2}(\boldsymbol{\theta}_0) \cdot \mathbf{K}_c \cdot \psi_c\left(\left(\boldsymbol{\Sigma}(\boldsymbol{\theta}_0) + (\mathbf{B}'\mathbf{B})^{-1}\mathbf{B}'\hat{\mathbf{D}}_{\boldsymbol{\theta}_0}\mathbf{B}(\mathbf{B}'\mathbf{B})^{-1}\right)^{-1/2} \cdot \hat{\mathbf{b}}_l(\boldsymbol{\theta}_0)\right) \quad (4.13)$$

where  $\mathbf{K}_c$  is a pre-specified constant to correct the downwardly biased estimated variance calculated from the set of Winsorized standardized residuals, i.e. so that when  $Z \sim N(0, 1)$  and  $Y \sim \psi_c(Z)$ ,  $\text{var}(K_c Y) = \text{var}(Z)$ . The closed form of  $\mathbf{K}_c$  for Huber's weight function is available in Dueck and Lohr (2005). An estimate of  $\mathbf{K}_c$  for Tukey's bi-weight function can be obtained through Monte-Carlo method. The Huber tuning parameter  $c = 1.345$  (Huber, 1981; Rey, 1983), or the Tukey's biweight tuning parameter  $c = 4.685$  (Goodall, 1983), corresponds to the 95% asymptotic efficiency on the standard normal distribution. The corresponding pseudo-observation and pseudo-model are  $\hat{\mathbf{y}}_l^{**}(\boldsymbol{\theta}_0) = \mathbf{A}\hat{\boldsymbol{\beta}}_l(\boldsymbol{\theta}_0) + \mathbf{B}\hat{\mathbf{b}}_l^{**}(\boldsymbol{\theta}_0)$  and

$$\hat{\mathbf{y}}_l^{**}(\boldsymbol{\theta}_0) = \mathbf{A}\boldsymbol{\beta} + \mathbf{B}\mathbf{b}, \quad \text{where } \mathbf{b} \sim N(\mathbf{0}, \boldsymbol{\Sigma}(\boldsymbol{\theta})) \quad (4.14)$$

The robust joint estimation using a robustified pseudo-model is summarized in the following algorithm.

**Algorithm 4.1.** *The robust joint estimation for the GLMFM model (2.1); (2.5), based on the chosen definition of the conditional divergence (4.1), is an iterative procedure with the following steps,*

1. Obtain the stage-1 subject-specific estimates  $\hat{\boldsymbol{\eta}}_{ij}$ 's from the finite normal mixture model (2.1). Substitute  $\hat{\boldsymbol{\eta}}_{ij}$ 's into the linear mixed model (2.5). Record the STS estimates (or predictors) as the starting values  $(\boldsymbol{\beta}_0, \mathbf{b}_0, \boldsymbol{\theta}_0) = (\hat{\boldsymbol{\beta}}_{STS}, \hat{\mathbf{b}}_{STS}, \hat{\boldsymbol{\theta}}_{STS})$ .
2. For given estimates of  $\boldsymbol{\theta}_0$ , solve for the minima  $(\hat{\boldsymbol{\beta}}_l(\boldsymbol{\theta}_0), \hat{\mathbf{b}}_l(\boldsymbol{\theta}_0))$  of the total divergence  $l(\boldsymbol{\beta}, \mathbf{b}, \boldsymbol{\theta}_0; \mathbf{y})$  by the Newton-Raphson iteration (4.6).

3. Define the pseudo-observations using the rescaled and robustified random effects (4.13) and fit the pseudo-model (4.14). Record the covariance parameter estimates  $\hat{\boldsymbol{\theta}}$ .
4. Let the new starting values be  $(\hat{\boldsymbol{\beta}}_l(\boldsymbol{\theta}_0), \hat{\mathbf{b}}_l(\boldsymbol{\theta}_0), \hat{\boldsymbol{\theta}})$  and repeat the iteration until all parameters converge.

The final estimates  $(\hat{\boldsymbol{\beta}}, \hat{\mathbf{b}}, \hat{\boldsymbol{\theta}})$  are reported as the robust joint estimation for the GLMFM model.

#### 4.4 Inference for fixed effects estimates

Let  $M$  be the total number of independent units in the GLMFM model, either the animals or the fields if there is no animal random effect, and  $\boldsymbol{\beta}_0$  be the true value of the fixed effects parameters. The robust joint estimation of the fixed effect  $\hat{\boldsymbol{\beta}}_{(M)}$ , based on the  $M$  independent units, is the solutions of the generalized estimating equation

$$\frac{\partial}{\partial \boldsymbol{\beta}} l^D(\boldsymbol{\beta}, \mathbf{b}, \boldsymbol{\theta}; \mathbf{y}) = \frac{\partial}{\partial \boldsymbol{\beta}} l_1^D(\boldsymbol{\beta} | \mathbf{b}; \mathbf{y}) = 0 \quad (4.15)$$

Let  $l_{1i}^D(\boldsymbol{\beta} | \mathbf{b}_i, \mathbf{y}_i)$  be the conditional divergence based on the  $i$ th independent unit, the estimating equation (4.15) is also the sum of independent score functions

$$\frac{\partial}{\partial \boldsymbol{\beta}} l_1^D(\boldsymbol{\beta} | \mathbf{b}; \mathbf{y}) = \sum_{i=1}^M \frac{\partial}{\partial \boldsymbol{\beta}} l_{1i}^D(\boldsymbol{\beta} | \mathbf{b}_i, \mathbf{y}_i)$$

Following the general theory of GEE, the Proposition 5.6 and Theorem 5.14 of Shao (2003) imply the consistency and the asymptotic normality of the fixed effects estimator  $\hat{\boldsymbol{\beta}}_{(M)}$ , that is

$$(\Sigma_{(M)}(\boldsymbol{\beta}_0))^{-1/2} \left( \hat{\boldsymbol{\beta}}_{(M)} - \boldsymbol{\beta}_0 \right) \rightarrow N(0, \mathbf{I}) \quad \text{when } M \rightarrow \infty \quad (4.16)$$

The asymptotic covariance matrix  $\Sigma_{(M)}(\boldsymbol{\beta}_0)$  has the sandwich form

$$\Sigma_{(M)}(\boldsymbol{\beta}_0) = [\mathbf{H}_{(M)}(\boldsymbol{\beta}_0)]^{-1} \mathbf{J}_{(M)}(\boldsymbol{\beta}_0) [\mathbf{H}_{(M)}(\boldsymbol{\beta}_0)]^{-1} \quad (4.17)$$

where

$$\begin{aligned}\mathbf{H}_{(M)}(\boldsymbol{\beta}_0) &= -\sum_{i=1}^M \mathbb{E} \left[ \frac{\partial^2 l_{1i}^D(\boldsymbol{\beta}|\mathbf{b}_i, \mathbf{y}_i)}{\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}'} \Big|_{\boldsymbol{\beta}=\boldsymbol{\beta}_0} \right] = -\sum_{i=1}^M \mathbf{A}'_i \mathbb{E} \left[ \frac{\partial^2 l_{1i}^D(\boldsymbol{\beta}|\mathbf{b}_i, \mathbf{y}_i)}{\partial \boldsymbol{\eta}_i \partial \boldsymbol{\eta}'_i} \Big|_{\boldsymbol{\beta}=\boldsymbol{\beta}_0} \right] \mathbf{A}_i \\ \mathbf{J}_{(M)}(\boldsymbol{\beta}_0) &= \sum_{i=1}^M \mathbf{A}'_i \mathbb{E} \left[ \frac{\partial l_{1i}^D(\boldsymbol{\beta}|\mathbf{b}_i, \mathbf{y}_i)}{\partial \boldsymbol{\eta}_i} \frac{\partial l_{1i}^D(\boldsymbol{\beta}|\mathbf{b}_i, \mathbf{y}_i)}{\partial \boldsymbol{\eta}'_i} \Big|_{\boldsymbol{\beta}=\boldsymbol{\beta}_0} \right] \mathbf{A}_i\end{aligned}$$

Given the final estimates  $(\hat{\boldsymbol{\beta}}, \hat{\mathbf{b}}, \hat{\boldsymbol{\theta}})$ , the finite sample approximation to the asymptotic covariance matrix (4.17) may be computed as

$$\hat{\boldsymbol{\Sigma}}_{\boldsymbol{\beta}} = \hat{\mathbf{H}}_{\boldsymbol{\beta}}^{-1} \hat{\mathbf{J}}_{\boldsymbol{\beta}} \hat{\mathbf{H}}_{\boldsymbol{\beta}}^{-1} \quad (4.19)$$

where

$$\begin{aligned}\hat{\mathbf{H}}_{\boldsymbol{\beta}} &= \mathbf{A}' \text{Diag} \left\{ \mathbf{H}_{ij}(\hat{\boldsymbol{\beta}}, \hat{\mathbf{b}}) \right\} \mathbf{A} \\ \hat{\mathbf{J}}_{\boldsymbol{\beta}} &= \mathbf{A}' \text{Diag} \left\{ \mathbf{J}_{ij}(\hat{\boldsymbol{\beta}}, \hat{\mathbf{b}}) \right\} \mathbf{A}\end{aligned}$$

with  $\mathbf{H}_{ij}$  and  $\mathbf{J}_{ij}$  defined in (4.2) and (4.3). For a general linear hypothesis  $H_0 : \mathbf{G}\boldsymbol{\beta} = \mathbf{0}$ , where  $\mathbf{G}$  is the contrast matrix, the Wald test statistic is

$$(\mathbf{G}\hat{\boldsymbol{\beta}})' (\mathbf{G}\hat{\boldsymbol{\Sigma}}_{\boldsymbol{\beta}}\mathbf{G}')^{-1} (\mathbf{G}\hat{\boldsymbol{\beta}}) \stackrel{H_0}{\sim} \chi_{\text{rank}(\mathbf{G})}^2 \quad (4.21)$$

## 4.5 Simulation studies

The simulation study was carried out to evaluate and compare the performance of the proposed estimation methods for the GLMFM model. We simulated 6 animals in each of the two different genotypes, with 5 microscopic fields per animal and 100 fibril diameter measures per microscopic field, which was similar in structure to the P4 tendon fibril data. For each microscopic field  $j$  from animal  $i$ , the conditional distribution of fibril diameters was either an uncontaminated or contaminated two-component normal mixture distribution. The uncontaminated data were generated as 100 realizations of the random variable with

density function  $f(x) = \pi_{1,ij}\phi(x, \mu_{1,ij}, \sigma_{1,ij}^2) + \pi_{2,ij}\phi(x, \mu_{2,ij}, \sigma_{2,ij}^2)$ . The vector of the mixing parameters  $\boldsymbol{\eta}_{ij} = (\text{logit } \pi_{2,ij}, \mu_{1,ij}, \mu_{2,ij}, \ln \sigma_{1,ij}, \ln \sigma_{2,ij})'$  are generated from the LME model

$$\boldsymbol{\eta}_{ij} = \boldsymbol{\beta}_{g_i} + \mathbf{b}_{ij}^{(2)}$$

which had the field random effects  $\mathbf{b}_{ij}^{(2)}$  only. We assumed no animal random effect because of the limited number of animals. The fixed effect parameters  $\boldsymbol{\beta}_g$  took values

$$\boldsymbol{\beta}_{\text{wild}} = (\log(0.75/0.25), -10, 5, \log 2.2, \log 2.7)'$$

$$\boldsymbol{\beta}_{\text{mut}} = (\log(0.6/0.4), -7, 10, \log 2.4, \log 2.9)'$$

and the field random effects  $\mathbf{b}_{ij}^{(2)}$  followed a common compound-symmetry covariance structure for both genotypes

$$\text{cov}(\mathbf{b}_{ij}^{(2)}) = (0.3, 1.5, 1.7, 0.3, 0.5) \begin{pmatrix} 1 & 0.3 & 0.3 & 0.3 & 0.3 \\ 0.3 & 1 & 0.3 & 0.3 & 0.3 \\ 0.3 & 0.3 & 1 & 0.3 & 0.3 \\ 0.3 & 0.3 & 0.3 & 1 & 0.3 \\ 0.3 & 0.3 & 0.3 & 0.3 & 1 \end{pmatrix} \begin{pmatrix} 0.3 \\ 1.5 \\ 1.7 \\ 0.3 \\ 0.5 \end{pmatrix}$$

The standard deviations and correlations of  $\mathbf{b}_{ij}^{(2)}$  are chosen as such so that the normal mixture distributions have meaningful overlap. We simulated the data with contaminations in both levels of the clustering, within and between microscopic fields. The between-field contamination was that for 10% of all the microscopic fields, the mean parameter of the first component was shifted towards left by a constant so that  $\mu_{1,ij}^* = \mu_{1,ij} - 5$ . The within-field contamination was a 5% normal outlier  $N(\mu_{2,ij} + 4 \times \sigma_{2,ij}, 0.7 \times \sigma_{2,ij})$  on the right side of the mixture, so that the second mixture component had a thicker tail than the standard normal distribution.

We compare the following four methods under the clean and contaminated scenarios:

- I. the starting values from the STS approach, using the MLE in the stage-1 estimation and the non-robust REML as the stage-2 estimator;
- II. the non-robust joint estimation, using MLE as the stage-1 estimator and the re-scaled random effects (4.11) in the pseudo-observations;
- III. the starting values from the STS approach, using the robust GWLE estimator Sym.Chi in the stage-1 estimation and the non-robust REML as the stage-2 estimator; and
- IV. the robust joint estimation, using the robust GWLE estimator Sym.Chi in stage-1 estimation and the re-scaled and robustified random effects (4.13), using Tukey's bi-weight function with tuning parameter  $c = 4.7$  in the pseudo-observations.

We have a fifth method serving as the "gold standard" under the clean scenario, which substitutes the known simulated mixture parameters  $\text{logit } \pi_{2,ij}$ ,  $\mu_{1,ij}$ ,  $\mu_{2,ij}$ ,  $\log \sigma_{1,ij}$ ,  $\log \sigma_{2,ij}$  from each animal  $i$  and field  $j$  into the pseudo-model (4.12) and calculates the fixed effect estimate  $\tilde{\beta}$  and the covariance estimate  $\tilde{\theta}$ . The estimates  $(\tilde{\beta}, \tilde{\theta})$  are considered to be the "best-possible" estimates, should we be able to restore the value of the true mixture parameters in the stage-1 estimation. Therefore, it provides more insight under the clean scenario if we compare the methods I, II, III and IV with the "gold standard".

The simulation results are summarized in Figures 4.1 and 4.2, which are the boxplots of the parameter estimates of the fixed effects as well as the standard deviations of the field random effects, under the clean and contaminated scenarios, respectively. Under the clean scenario in Figure 4.1, both the non-robust joint estimation II and the robust joint estimation IV improve their corresponding STS approaches I and III, respectively, and they perform almost as good as the "gold standard". The non-robust joint estimation II

has slightly better efficiency than the robust joint estimation IV. Under the contaminated scenario in Figure 4.2, the robust joint estimation IV gives much more accurate estimates of the standard deviations of the field random effects, especially on the first component mean, than the non-robust joint estimation II.

## 4.6 Tendon fibril data example

### 4.6.1 A simplified two-stage model and STS approach

In this section, we use a simplified two-stage model for the postnatal-4-day (P4) subset of the tendon fibril data and obtain the corresponding STS estimates. We limit the vector of the mixing parameters to the mean and standard deviations of the two-component normal mixture distribution  $\boldsymbol{\eta}_{ij}^* = (\mu_{1,ij}, \mu_{2,ij}, \ln \sigma_{1,ij}, \ln \sigma_{2,ij})'$ . In the second stage linear mixed model, we have the animal genotype  $g_i$  and body weight  $m_i$  as the covariates for the fixed effect  $\boldsymbol{\beta}$ , and both the animal and field random effect

$$\boldsymbol{\eta}_{ij}^* = \boldsymbol{\beta}_{g_i} + \boldsymbol{\beta}_m m_i + \mathbf{b}_i^{(1)} + \mathbf{b}_{ij}^{(2)} \quad (4.23)$$

where  $\boldsymbol{\beta}_{g_i} = (\mu_{1g_i}, \mu_{2g_i}, \ln \sigma_{1g_i}, \ln \sigma_{2g_i})'$ ,  $g_i = 1, 2$  and  $\boldsymbol{\beta}_m = (\beta_{m\mu_1}, \beta_{m\mu_2}, \beta_{m\sigma_1}, \beta_{m\sigma_2})'$ . The exploratory analysis suggested that the initial covariance structure include uncorrelated animal random effects and unstructured field random effects, as it is not feasible to estimate unstructured covariance matrix for the animal random effects based on merely 12 animals.

In the STS approach, we compare MLE, Dens.Pow and Sym.Chi as the stage-1 estimates, using REML as the common stage-2 estimate. The result shows that (a). the estimates of fixed effects  $\boldsymbol{\beta}_{g_i}$ ,  $\boldsymbol{\beta}_m$  and covariance parameters were similar for all stage-1 estimates, but their standard errors were lower by using the Dens.Pow or Sym.Chi estimates other than MLE; (b). the robust stage-1 estimates Dens.Pow or Sym.Chi detected the genotype differ-

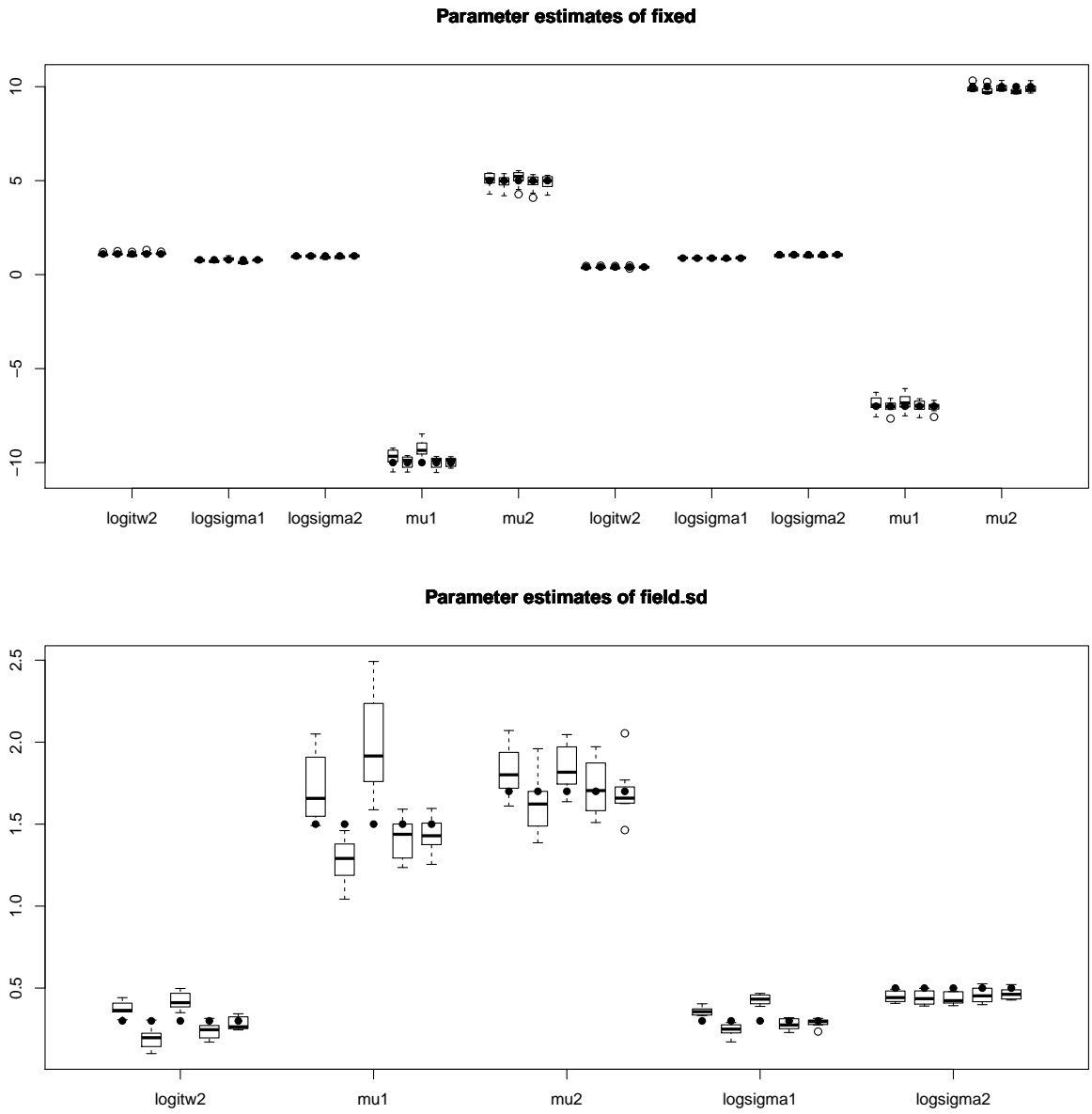


Figure 4.1: The box-plots of the parameter estimates of the fixed effects  $\beta_{\text{wild}}$  and  $\beta_{\text{mut}}$  (top) and the field random effect covariance parameters (bottom) under the clean scenario. The adjacent boxplots represents the methods I, II, III, IV and the "gold standard", from left to right. The true parameter values are plotted as the thick dot.

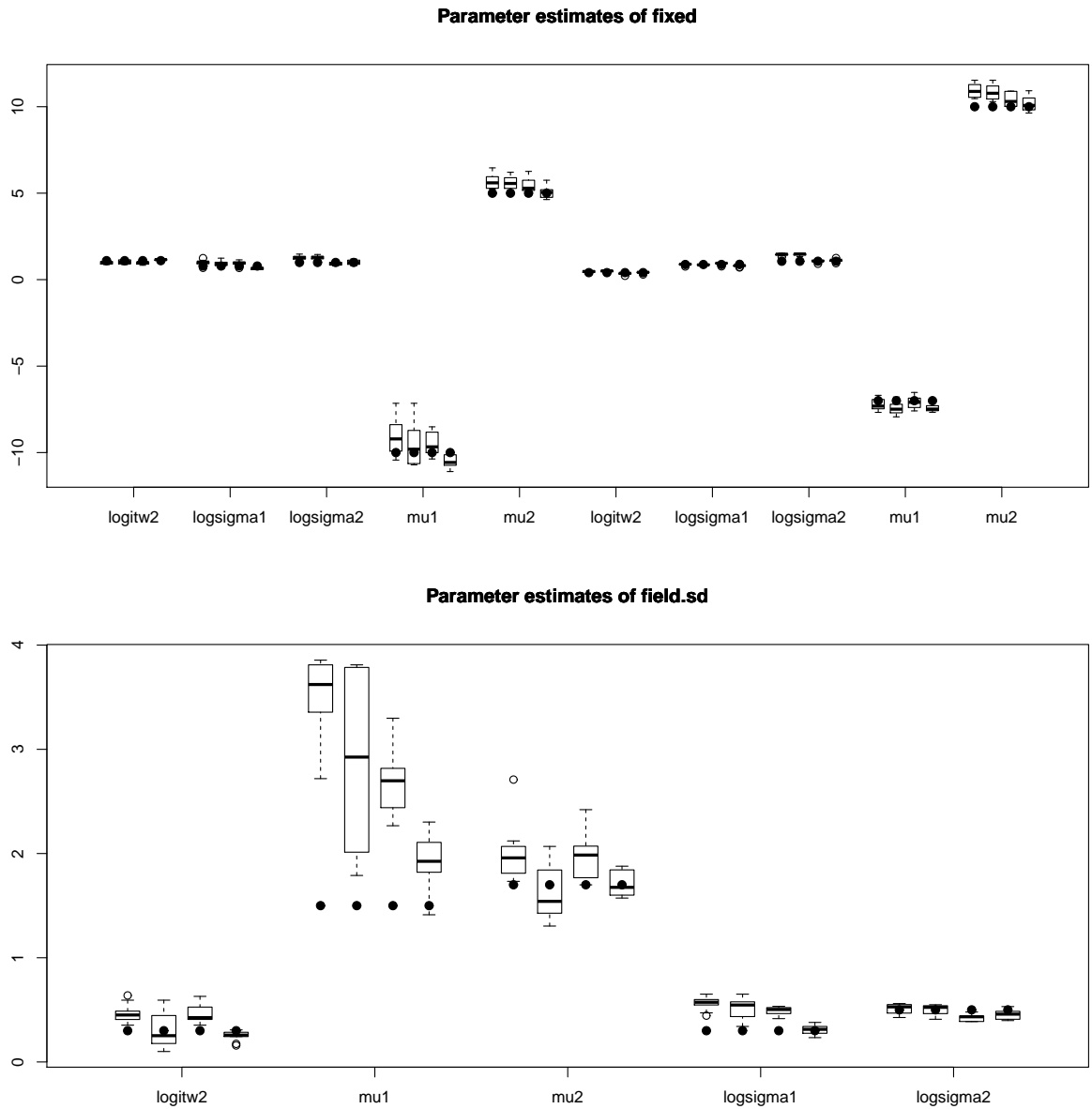


Figure 4.2: The box-plots of the parameter estimates of the fixed effects  $\beta_{\text{wild}}$  and  $\beta_{\text{mut}}$  (top) and the field random effect covariance parameters (bottom) under the fibril- and field-contaminated scenario. The adjacent boxplots represents the methods I, II, III and IV, from left to right. The true parameter values are plotted as the thick dot.

ence in both  $\mu_1$  and  $\mu_2$ , while MLE detected the genotype difference only in  $\mu_1$ . Therefore, the robust stage-1 estimates such as Dens.Pow and Sym.Chi, instead of MLE, improved the precision of population parameters estimation in stage-2, which allowed the identification of the biologically important difference between the second component means as significant, in addition to the significance of the difference between the first component means. These findings support the biological hypothesis that decorin deficiency further manifests in the process of fibril growth. This simplified two-stage model as well as the STS approach has been published in the co-authored work Chervoneva et al. (2011).

#### 4.6.2 GLMFM model and robust joint estimation

We consider the full GLMFM model and compare the two-stage estimation vs. the joint estimation, either non-robust or robust, for the postnatal-3-month (P3M) tendon fibril data. The GLMFM model assumes a three-component normal mixture distribution on each microscopic field and the vectors of transformed mixture parameters

$$\boldsymbol{\eta}_{ij} = \left( \text{logit } \pi_{2,ij}, \text{logit } \pi_{3,ij}, \mu_{1,ij}, \mu_{2,ij}, \mu_{3,ij}, \ln \sigma_{1,ij}, \ln \sigma_{2,ij}, \ln \sigma_{3,ij} \right)'$$

where  $\text{logit } \pi_{2,ij} = \log(\pi_{2,ij}/\pi_{1,ij})$  and  $\text{logit } \pi_{3,ij} = \log(\pi_{3,ij}/\pi_{1,ij})$ , are assumed to follow the linear mixed model with fixed effects of the animal genotype  $g_i$  and body weight  $m_i$ ,

$$\boldsymbol{\eta}_{ij} = \boldsymbol{\beta}_{g_i} + \boldsymbol{\beta}_m m_i + \mathbf{b}_{ij}^{(2)} \tag{4.24}$$

where the field random effects  $\mathbf{b}_{ij}^{(2)}$  follow a common unstructured covariance structure  $\Sigma_2(\boldsymbol{\theta}_2)$  for both genotypes.

We apply the following four estimation methods to our GLMFM model: the maximum-likelihood two-stage approach I, the maximum-likelihood joint estimation II, the robust two-stage approach III and the robust joint estimation IV. The standard errors of the fixed

effect parameters ( $\beta_{g_i}, \beta_m$ ) are calculated and the tests on genotype difference are carried out. Figures 4.3 and 4.4 present the estimates from all four methods for three sample microscopic fields.

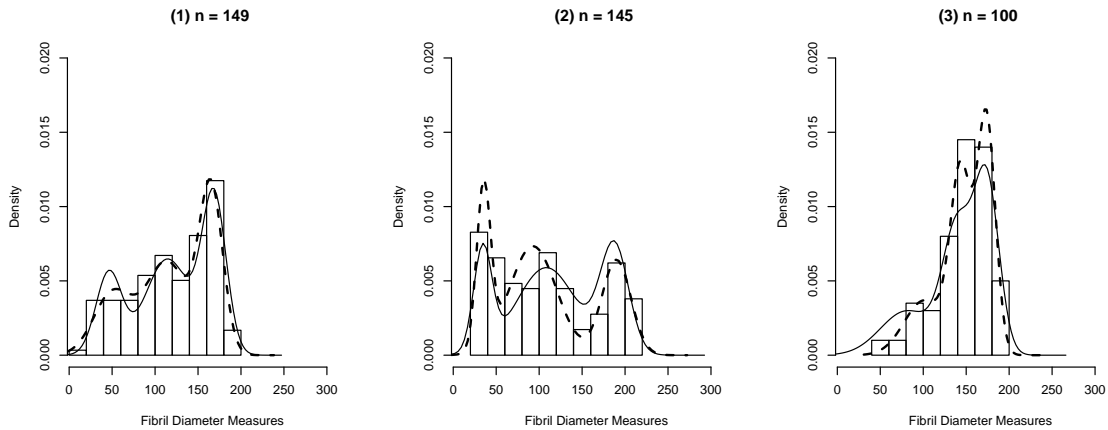


Figure 4.3: Maximum-likelihood estimates of 3-component normal mixtures for selected microscopic fields from P3M tendon fibril data. Dashed lines corresponds to two-stage estimation result (method I) and solid line corresponds to the joint estimation result (method II)

Table 4.1, in Appendix ??, presents the population average of the transformed normal mixture parameters for both the wild and mutant mice, obtained from each of the four methods, as well as the standard errors of these estimates. Table 4.2 is the conventional normal mixture parameters in mixing proportions, means and standard deviations, as well as the corresponding standard errors, obtained from applying  $\delta$ -method on Table 4.1 results. The population averages are calculated by  $\beta_{\text{wild}} + \beta_m \text{mean}\{m_i : \text{wild}\}$  for wild mice and  $\beta_{\text{mut}} + \beta_m \text{mean}\{m_i : \text{mut}\}$  for mutant mice. All four estimation methods show that the mutation leads to both larger means and larger standard deviations of all three normal components. The standard errors of the estimates are generally lower using the joint estimations (methods II and IV) than using the corresponding two-stage approaches (methods I and III). The maximum-likelihood and robust two-stage approaches (method I vs. method

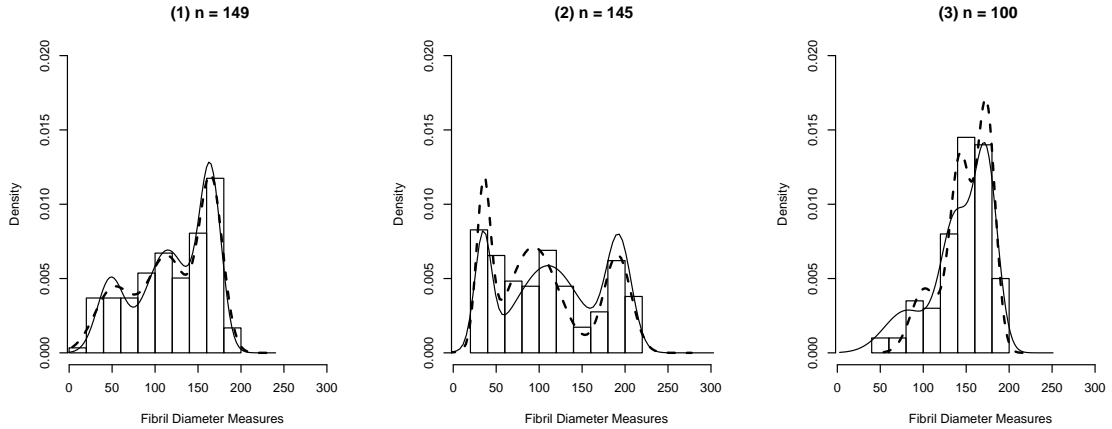


Figure 4.4: Robust estimates of 3-component normal mixtures for selected microscopic fields from P3M tendon fibril data. Dashed lines corresponds to two-stage estimation result (method III) and solid line corresponds to the joint estimation result (method IV)

III) give similar estimates for all mixing parameters of either genotype, except  $\mu_1$  for wild and mutant mice, due to the fact that the robust approach III downweights outliers from measuring tapered ends and results in larger  $\mu_1$  estimates, 55.75nm for wild and 67.12 for mutant, compared to 52.32nm for wild and 63.14nm for mutant. Similarly, compared to the maximum-likelihood joint estimation II, the robust joint estimation IV has larger estimates of  $\mu_1$  (50.65 vs. 48.40nm for wild and 60.58 vs. 57.64nm for mutant) and smaller estimates of  $\mu_3$  (174.67 vs. 178.70nm for wild and 199.92 vs. 203.85nm for mutant). Thus, the robust joint estimation IV also reduces the bias due to the large outliers corresponding to the fused fibrils.

Table 4.3 presents the estimates of the genotype differences in each of the mixture parameters, the standard errors of these estimates and the corresponding  $p$  values from Wald test. The point estimates of the genotype differences are similar for all four methods, but the corresponding standard errors from the joint estimations (methods II and IV) are either of the same magnitude or much smaller than the standard errors from the two-stage ap-

proaches (methods I and III). Therefore, the significance of these differences are of the same magnitude or larger based on the joint estimations than the two-stage approaches. Among the eight normal mixture parameters, the largest decrease in the standard errors, resulting from using the joint estimations rather than two-stage approaches, happens for  $\mu_1$ ,  $\mu_2$ ,  $\mu_3$ ,  $\log \sigma_1$  and  $\log \sigma_3$ . The joint estimations, but not the two-stage approaches, identify the genotype differences in  $\mu_1$  as significant, so that all mean parameters of the three normal components have significant genotype differences. The joint estimations results indicate significantly larger spread of the first and third normal components in mutant mice, while two-stage approaches attribute larger spread to the second component. On the other hand, the test results of based on the robust joint estimation IV are generally more conservative than the non-robust joint estimation II, with larger p-values due to larger estimated standard error.

## 4.7 Summary of Chapter 4

In this chapter, we approached the second task in the analysis of the tendon fibril data, presented in Chapter 1, by developing the robust joint estimation for the GLMFM model. The robust joint estimation improved greatly compared to the previous two-stage approaches.

By applying the robust joint estimation to the tendon fibril data, we may answer the biological question of the genotype effect of decorin deficiency by performing Wald tests on the population parameters of each genotype. The analysis shows that the three-normal-component assumption for the multi-modal microscopic-field-specific distribution of the fibril diameter measures from postnatal-3-month mice are adequate, as the three components are statistically significant from each other, for both genotypes. The mutant (decorin

deficient) mice have significantly thicker fibrils represented by larger means of the normal components, and slightly wider spread of fibrils represented by the standard deviations.

Table 4.1: The estimates of the population average of the normal mixture parameters for wild and mutant mice (in independent mixture parameters  $\text{logit } \pi$ 's,  $\mu$ 's and  $\text{log } \sigma$ 's). The standard errors of the estimates are included in parenthesis.

	Method I	Method II	Method III	Method IV
logitw2.wild	0.58 (0.09)	0.91 (0.09)	0.50 (0.08)	0.81 (0.11)
logitw3.wild	0.77 (0.07)	0.78 (0.06)	0.77 (0.06)	0.88 (0.06)
mu1.wild	52.32 (4.31)	48.41 (0.66)	55.75 (4.12)	50.61 (0.83)
mu2.wild	120.22 (5.10)	120.92 (1.42)	120.12 (5.04)	118.37 (1.27)
mu3.wild	176.41 (5.31)	178.70 (1.07)	175.54 (5.30)	174.73 (0.94)
logsigma1.wild	2.78 (0.11)	2.69 (0.05)	2.94 (0.08)	2.77 (0.04)
logsigma2.wild	3.06 (0.06)	3.22 (0.06)	3.01 (0.05)	3.15 (0.07)
logsigma3.wild	2.95 (0.06)	2.92 (0.04)	2.98 (0.05)	2.97 (0.03)
logitw2.mut	0.43 (0.09)	0.67 (0.11)	0.34 (0.08)	0.57 (0.23)
logitw3.mut	0.79 (0.07)	0.85 (0.07)	0.78 (0.05)	0.83 (0.07)
mu1.mut	63.14 (4.20)	57.65 (0.64)	67.12 (4.01)	60.57 (1.07)
mu2.mut	140.35 (4.96)	139.55 (2.33)	141.79 (4.91)	139.45 (3.22)
mu3.mut	202.49 (5.18)	203.86 (1.42)	202.23 (5.16)	199.92 (1.87)
logsigma1.mut	3.01 (0.10)	2.87 (0.05)	3.13 (0.08)	2.95 (0.06)
logsigma2.mut	3.20 (0.05)	3.32 (0.08)	3.13 (0.05)	3.24 (0.15)
logsigma3.mut	3.07 (0.06)	3.06 (0.04)	3.07 (0.05)	3.10 (0.05)

Table 4.2: The estimates of the population average of the normal mixture parameters for wild and mutant mice (in plain mixture parameters  $\pi$ 's,  $\mu$ 's and  $\sigma$ 's). The standard errors of the estimates are included in parenthesis.

	Method I	Method II	Method III	Method IV
wild.w1	0.20 (0.01)	0.18 (0.01)	0.21 (0.01)	0.18 (0.01)
wild.w2	0.36 (0.02)	0.44 (0.02)	0.34 (0.02)	0.40 (0.03)
wild.w3	0.44 (0.02)	0.39 (0.02)	0.45 (0.02)	0.43 (0.02)
wild.mu1	52.32 (4.31)	48.41 (0.66)	55.75 (4.12)	50.61 (0.83)
wild.mu2	120.22 (5.10)	120.92 (1.42)	120.12 (5.04)	118.37 (1.27)
wild.mu3	176.41 (5.31)	178.70 (1.07)	175.54 (5.30)	174.73 (0.94)
wild.sigma1	16.12 (1.71)	14.77 (0.67)	18.89 (1.56)	16.02 (0.69)
wild.sigma2	21.34 (1.20)	24.91 (1.49)	20.25 (1.06)	23.29 (1.57)
wild.sigma3	19.20 (1.09)	18.58 (0.67)	19.64 (1.07)	19.49 (0.56)
mut.w1	0.21 (0.01)	0.19 (0.01)	0.22 (0.01)	0.20 (0.02)
mut.w2	0.32 (0.02)	0.37 (0.03)	0.31 (0.02)	0.35 (0.06)
mut.w3	0.47 (0.02)	0.44 (0.03)	0.47 (0.02)	0.45 (0.05)
mut.mu1	63.14 (4.20)	57.65 (0.64)	67.12 (4.01)	60.57 (1.07)
mut.mu2	140.35 (4.96)	139.55 (2.33)	141.79 (4.91)	139.45 (3.22)
mut.mu3	202.49 (5.18)	203.86 (1.42)	202.23 (5.16)	199.92 (1.87)
mut.sigma1	20.19 (2.09)	17.64 (0.83)	22.92 (1.85)	19.13 (1.06)
mut.sigma2	24.65 (1.35)	27.52 (2.19)	22.86 (1.16)	25.46 (3.93)
mut.sigma3	21.49 (1.19)	21.24 (0.80)	21.51 (1.14)	22.26 (1.18)

Table 4.3: The estimates of the genotype differences (using wild as the reference level) in each of the mixture parameters. The standard errors of the estimates are included in parenthesis. The p-value of the Wald test are included in square parenthesis.

	Method I		Method II		Method III		Method IV	
logitw2	-0.17	(0.13) [p=0.211]	-0.27	(0.14) [p=0.054]	-0.17	(0.12) [p=0.143]	-0.26	(0.26) [p=0.325]
logitw3	0.02	(0.10) [p=0.821]	0.06	(0.10) [p=0.511]	-0.00	(0.08) [p=0.996]	-0.05	(0.09) [p=0.596]
mu1	9.96	(6.10) [p=0.102]	8.48	(0.92) [p=0.000]	9.96	(5.83) [p=0.087]	9.05	(1.37) [p=0.000]
mu2	19.48	(7.21) [p=0.007]	18.09	(2.73) [p=0.000]	20.97	(7.13) [p=0.003]	20.90	(3.50) [p=0.000]
mu3	26.28	(7.52) [p=0.000]	25.42	(1.80) [p=0.000]	26.97	(7.49) [p=0.000]	25.77	(2.15) [p=0.000]
logsigma1	0.20	(0.15) [p=0.173]	0.16	(0.07) [p=0.017]	0.14	(0.12) [p=0.231]	0.15	(0.07) [p=0.037]
logsigma2	0.17	(0.08) [p=0.032]	0.12	(0.10) [p=0.226]	0.15	(0.07) [p=0.039]	0.12	(0.17) [p=0.500]
logsigma3	0.11	(0.08) [p=0.161]	0.13	(0.05) [p=0.013]	0.09	(0.08) [p=0.244]	0.13	(0.06) [p=0.035]

## Chapter 5

# Discussion and future work

The theoretical work presented in this dissertation is motivated by the tendon fibril data and has potential application in a variety of relatively complex statistical modeling situations. The tendon fibril data has small number of individual units and multi-modal distributions at the lowest level of clustering. The biological question we are trying to answer is the genotype effect of decorin deficiency on the distribution of the fibril diameter measurements. We approached this question by constructing the generalized linear mixed model for finite normal mixtures (GLMFM) and developing robust estimation techniques. The biological conclusion is that the fibrils from postnatal-3-month mice constitute three normally distributed components, the mutant (decorin deficient) mice have significantly thicker fibrils represented by larger means of the normal components and slightly wider spread of fibrils represented by the standard deviations.

In Chapter 3, we developed the family general weighted likelihood estimators and also developed the robust (truncated-)cubic-inverse weights for the finite mixture model of normal distributions. We applied our robust estimation methods to the mixture model of exponential family distributions with fixed number of components. Our robust weight is

equally efficient but more robust to potential outliers compared to the other robust weights found in the literature. In Chapter 4, we developed the standard-two-stage approaches and the robust joint estimation for the GLMFM model. We also developed the asymptotic covariance matrix for the fixed effect estimates, which is useful to construct the Wald test statistics for the evaluation of comparison of covariates on means. We effectively applied our method to the P3M tendon fibril data and evaluated the genotype effect of decorin deficiency in the mixture parameters.

The following future improvements could be made on both the estimation approach and the statistical model. The robust estimation of the covariance parameters in the GLMFM model may be accomplished using the estimating equations of Richardson and Welsh (1995). Furthermore, we may adopt a full Bayesian approach to the GLMFM model. The GLMFM model itself may be extended to a general non-linear mixed effects models with flexible conditional and random effects distributions. The random effects distributions may also be represented by a mixture of normal components (Beal and Sheiner, 1988) or by a member of the class of smooth densities defined in Gallant and Nychka (1987), as developed by Davidian and Gallant (1993).

The future work may be incorporating variable number of components, so that we may analyze the tendon fibril data over time. In the tendon fibril data, P4 microscopic fields generally have two components and P3M three, while the postnatal-10-day (P10) fields have either two or three components. By making the number of components an additional parameter in the finite mixture, we may be able to analyze P4, P10 and P3M data together and have more interesting findings.

# Bibliography

- Andrews, D. F., Bickel, P. J., Hampel, F. R., Huber, P. J., Rogers, W. H., and Tukey, J. W. (1972). *Robust Estimates of Location: Survey and Advances*. Princeton University Press, Princeton, N.J.
- Basu, A., Harris, I. R., Hjort, N. L., and Jones, M. C. (1998). Robust and efficient estimation by minimising a density power divergence. *Biometrika*, 85(3):549–559.
- Basu, A. and Lindsay, B. G. (1994). Minimum disparity estimation for continuous models: efficiency, distributions and robustness. *Annals of the Institute of Statistical Mathematics*, 46(4):683–705.
- Basu, A. and Lindsay, B. G. (2004). The iteratively reweighted estimating equation in minimum distance problems. *Computational Statistics & Data Analysis*, 45(2):105–124.
- Bednarski, T. and Zontek, S. (1996). Robust estimation of parameters in a mixed unbalanced model. *The Annals of Statistics*, 24(4):1493–1510.
- Belin, T. R. and Rubin, D. B. (1995). A method for calibrating false-match rates in record linkage. *Journal of the American Statistical Association*, 90(430):694–707.
- Beran, R. (1977). Minimum Hellinger distance estimates for parametric models. *The Annals of Statistics*, 5(3):445–463.

- Beran, R. (1978). An efficient and robust adaptive estimator of location. *The Annals of Statistics*, 6(2):292–313.
- Betrò, B., Bodini, A., and Guglielmi, A. (2006). Generalized moment theory and Bayesian robustness analysis for hierarchical mixture models. *Annals of the Institute of Statistical Mathematics*, 58(4):721–738.
- Bhandari, S. K., Basu, A., and Sarkar, S. (2006). Robust inference in parametric models using the family of generalized negative exponential dispatches. *Australian & New Zealand Journal of Statistics*, 48(1):95–114.
- Bickel, P. J., Holm, S., Rosén, B., Spjøtvoll, E., Lauritzen, S., Johansen, S., and Barndorff-Nielsen, O. (1976). Another look at robustness: A review of reviews and some new developments [with discussion and reply]. *Scandinavian Journal of Statistics*, 3(4):145–168.
- Boldea, O. and Magnus, J. R. (2009). Maximum likelihood estimation of the multivariate normal mixture model. *Journal of the American Statistical Association*, 104(488):1539–1549.
- Box, G. E. P. and Tiao, G. C. (1973). *Bayesian Inference in Statistical Analysis*. Addison-Wesley Publishing Co., Reading, Mass.-London-Don Mills, Ont. Addison-Wesley Series in Behavioral Science: Quantitative Methods.
- Celeux, G., Chrétien, S., Forbes, F., and Mkhadri, A. (2001). A component-wise EM algorithm for mixtures. *Journal of Computational and Graphical Statistics*, 10(4):697–712.

- Chen, J. and Khalili, A. (2009). Order selection in finite mixture models with a nonsmooth penalty. *Journal of the American Statistical Association*, 104(485):187–196.
- Chervoneva, I., Iglewicz, B., and Hyslop, T. (2006). A general approach for two-stage analysis of multilevel clustered non-Gaussian data. *Biometrics*, 62(3):752–759.
- Chervoneva, I., Zhan, T., Iglewicz, B., Hauck, W. W., and Birk, D. E. (2011). Two-stage hierarchical modeling for analysis of subpopulations in conditional distributions. *Journal of Applied Statistics*, 0(0):1–16. To appear. Available online since 14 Jul 2011.
- Choy, S. T. B. and Smith, A. F. M. (1997). Hierarchical models with scale mixtures of normal distributions. *Test*, 6(1):205–221.
- Clarke, B. R. and Heathcote, C. R. (1994). Robust estimation of  $k$ -component univariate normal mixtures. *Annals of the Institute of Statistical Mathematics*, 46(1):83–93.
- Cressie, N. and Read, T. R. C. (1984). Multinomial goodness-of-fit tests. *Journal of the Royal Statistical Society. Series B. Methodological*, 46(3):440–464.
- Cuesta-Albertos, J. A., Gordaliza, A., and Matrán, C. (1997). Trimmed  $k$ -means: an attempt to robustify quantizers. *The Annals of Statistics*, 25(2):553–576.
- Cutler, A. and Cordero-Braña, O. I. (1996). Minimum Hellinger distance estimation for finite mixture models. *Journal of the American Statistical Association*, 91(436):1716–1723.
- Datta, G. S. and Lahiri, P. (1995). Robust hierarchical Bayes estimation of small area characteristics in the presence of covariates and outliers. *Journal of Multivariate Analysis*, 54(2):310 – 328.

- Davidian, M. and Giltinan, D. M. (1993). Some general estimation methods for nonlinear mixed-effects model. *Journal of Biopharmaceutical Statistics*, 3(1):23–55.
- Davidian, M. and Giltinan, D. M. (1995). *Nonlinear Models for Repeated Measurement Data*. Number 62 in Monographs on Statistics and Applied Probability. Chapman & Hall Ltd.
- Davidian, M. and Giltinan, D. M. (2003). Nonlinear models for repeated measurement data: An overview and update. *Journal of Agricultural, Biological, and Environmental Statistics*, 8(4):387–419.
- Day, N. E. (1969). Estimating the components of a mixture of normal distributions. *Biometrika*, 56(3):463–474.
- DeVeaux, R. D. and Krieger, A. M. (1990). Robust estimation of a normal mixture. *Statistics and Probability Letters*, 10:1–7.
- Diebolt, J. and Robert, C. P. (1994). Estimation of finite mixture distributions through bayesian sampling. *Journal of the Royal Statistical Society. Series B (Methodological)*, 56(2):363–375.
- Dueck, A. and Lohr, S. (2005). Robust estimation of multivariate covariance components. *Biometrics*, 61(1):162–169.
- Fellner, W. H. (1986). Robust estimation of variance components. *Technometrics*, 28(1):51–60.
- Fruhwrith-Schnatter, S. (2001). Markov Chain Monte Carlo estimation of classical and dynamic switching and mixture models. *Journal of the American Statistical Association*, 96(453):194–209.

- Fujisawa, H. and Eguchi, S. (2006). Robust estimation in the normal mixture model. *Journal of Statistical Planning and Inference*, 136(11):3989–4011.
- Gelman, A., Carlin, J. B., Stern, H. S., and Rubin, D. B. (2004). *Bayesian Data Analysis*. Texts in Statistical Science Series. Chapman & Hall/CRC, Boca Raton, FL, second edition.
- Gelman, A. and King, G. (1990). Estimating the electoral consequences of legislative redistricting. *Journal of the American Statistical Association*, 85(410):274–282.
- Goodall, C. (1983). M-estimators of location: An outline of the theory. In Hoaglin, D. C., Mosteller, F., and Tukey, J. W., editors, *Understanding Robust and Exploratory Data Analysis*, Wiley Classics Library, pages 339–403. Wiley-Interscience, New York.
- Gray, G. (1994). Bias in misspecified mixtures. *Biometrics*, 50(2):457–470.
- Hampel, F. R., Ronchetti, E. M., Rousseeuw, P. J., and Stahel, W. A. (1985). *Robust Statistics: The Approach based on Influence Functions*. Wiley Series in Probability and Statistics.
- Harville, D. A. (1977). Maximum likelihood approaches to variance component estimation and to related problems [with a comment by J. N. K. Rao and a reply by the author]. *Journal of the American Statistical Association*, 72(358):320–340.
- Hathaway, R. J. (1985). A constrained formulation of maximum-likelihood estimation for normal mixture distributions. *The Annals of Statistics*, 13(2):795–800.
- Henderson, C. R. (1975). Best linear unbiased estimation and prediction under a selection model. *Biometrics*, 31(2):423–447.

- Henderson, C. R., Kempthorne, O., Searle, S. R., and von Krosigk, C. M. (1959). The estimation of environmental and genetic trends from records subject to culling. *Biometrics*, 15(2):192–218.
- Hoaglin, D. C., Mosteller, F., and Tukey, J. W. (1983). *Understanding Robust and Exploratory Data Analysis*. Wiley, 2 edition.
- Huber, P. J. (1981). *Robust Statistics*. John Wiley & Sons Inc., New York. Wiley Series in Probability and Mathematical Statistics.
- Huggins, R. M. (1993). A robust approach to the analysis of repeated measures. *Biometrics*, 49(3):715–720.
- Ishwaran, H. and James, L. F. (2002). Approximate Dirichlet process computing in finite normal mixtures: Smoothing and prior information. *Journal of Computational and Graphical Statistics*, 11(3):508–532.
- Ishwaran, H., James, L. F., and Sun, J. (2001). Bayesian model selection in finite mixtures by marginal density decompositions. *Journal of the American Statistical Association*, 96(456):1316–1332.
- Jasra, A., Holmes, C. C., and Stephens, D. A. (2005). Markov Chain Monte Carlo methods and the label switching problem in bayesian mixture modeling. *Statistical Science*, 20(1):50–67.
- Jeong, D.-B. and Sarkar, S. (2000). Negative exponential disparity based family of goodness-of-fit tests for multinomial models. *Journal of Statistical Computation and Simulation*, 65(1):43–61.

- Jones, M. C., Hjort, N. L., Harris, I. R., and Basu, A. (2001). A comparison of related density-based minimum divergence estimators. *Biometrika*, 88(3):865–873.
- Kiefer, J. and Wolfowitz, J. (1956). Consistency of the maximum likelihood estimator in the presence of infinitely many incidental parameters. *The Annals of Mathematical Statistics*, 27(4):887–906.
- Kiefer, N. M. (1978). Discrete parameter variation: Efficient estimation of a switching regression model. *Econometrica*, 46(2):427–434.
- Kuk, A. Y. C. (1995). Asymptotically unbiased estimation in generalized linear models with random effects. *Journal of the Royal Statistical Society. Series B. Methodological*, 57(2):395–407.
- Kullback, S. (1959). *Information Theory and Statistics*. John Wiley and Sons, Inc., New York.
- Kullback, S. and Leibler, R. A. (1951). On information and sufficiency. *The Annals of Mathematical Statistics*, 22(1):79–86.
- Lee, Y. and Nelder, J. A. (1996). Hierarchical generalized linear models. *Journal of the Royal Statistical Society. Series B (Methodological)*, 58(4):619–678.
- Lee, Y. and Nelder, J. A. (2001). Hierarchical generalised linear models: A synthesis of generalised linear models, random-effect models and structured dispersions. *Biometrika*, 88(4):987–1006.
- Lee, Y. and Nelder, J. A. (2003). Extended-REML estimators. *Journal of Applied Statistics*, 30(8):845–856.

- Lee, Y. and Nelder, J. A. (2006). Double hierarchical generalized linear models (with discussion). *Journal of the Royal Statistical Society, Series C: Applied statistics*, 55(2):139–185.
- Lindsay, B. G. (1994). Efficiency versus robustness: the case for minimum Hellinger distance and related methods. *The Annals of Statistics*, 22(2):1081–1114.
- Lindsay, B. G. and Roeder, K. (1992). Residual diagnostics for mixture models. *Journal of the American Statistical Association*, 87(419):785–794.
- Lindstrom, M. J. and Bates, D. M. (1990). Nonlinear mixed effects models for repeated measures data. *Biometrics*, 46(3):673–687.
- Markatou, M. (2000). Mixture models, robustness, and the weighted likelihood methodology. *Biometrics*, 56(2):483–486.
- Markatou, M., Basu, A., and Lindsay, B. G. (1998). Weighted likelihood equations with bootstrap root search. *Journal of the American Statistical Association*, 93(442):740–750.
- McCullagh, P. and Nelder, J. A. (1983). *Generalized Linear Models*. Monographs on Statistics and Applied Probability. Chapman & Hall Ltd, London.
- McGilchrist, C. and Yau, K. (1995). The derivation of BLUP, ML, REML estimation methods for generalised linear mixed models. *Communications in Statistics - Theory and Methods*, 24(12):2963–2980.
- McGilchrist, C. A. (1994). Estimation in generalized mixed models. *Journal of the Royal Statistical Society. Series B (Methodological)*, 56(1):61–69.
- McLachlan, G. J. and Basford, K. E. (1988). *Mixture Models: Inference and Applications to Clustering*, volume 84 of *Statistics: Textbooks and Monographs*. Marcel Dekker Inc., New York.

- McLachlan, G. J. and Krishnan, T. (2008). *The EM Algorithm and Extensions*. John Wiley & Sons, Inc., 2 edition.
- McLachlan, G. J. and Peel, D. (1998). MIXFIT: An algorithm for the automatic fitting and testing of normal mixture models. In *ICPR '98: Proceedings of the 14th International Conference on Pattern Recognition*, pages 553–557, Washington, DC, USA. IEEE Computer Society.
- McLachlan, G. J. and Peel, D. (1999). The EMMIX algorithm for the fitting of normal and  $t$ -components. *Journal of Statistical Software*, 4(2).
- McLachlan, G. J. and Peel, D. (2000). *Finite Mixture Models*. Wiley Series in Probability and Statistics: Applied Probability and Statistics. Wiley-Interscience, New York.
- Moreno, E. and Pericchi, L. R. (1993). Bayesian robustness for hierarchical  $\varepsilon$ -contamination models. *Journal of Statistical Planning and Inference*, 37(2):159 – 167.
- Muller, C. H. and Uhlig, S. (2001). Estimation of variance components with high breakdown point and high efficiency. *Biometrika*, 88(2):353–366.
- Muller, P. and Rosner, G. L. (1997). A Bayesian population model with hierarchical mixture priors applied to blood count data. *Journal of the American Statistical Association*, 92(440):1279–1292.
- Parr, W. C. (1981). Minimum distance estimation: A bibliography. *Communications in Statistics - Theory and Methods*, 10(12):1205–1224.
- Patterson, H. D. and Thompson, R. (1971). Recovery of inter-block information when block sizes are unequal. *Biometrika*, 58(3):545–554.

- Pauler, D. K. and Laird, N. M. (2000). A mixture model for longitudinal data with application to assessment of noncompliance. *Biometrics*, 56(2):464–472.
- Pilla, R. S. and Lindsay, B. G. (2001). Alternative EM methods for nonparametric finite mixture models. *Biometrika*, 88(2):535–550.
- Polasek, W. and Pötzelberger, K. (1988). Robust Bayesian analysis in hierarchical models. In *Bayesian statistics, 3 (Valencia, 1987)*, Oxford Sci. Publ., pages 377–394. Oxford Univ. Press, New York.
- Pötzelberger, K. and Polasek, W. (1991). Robust HPD regions in Bayesian regression models. *Econometrica*, 59(6):1581–1589.
- R Development Core Team (2011). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0.
- Rao, C. R. (1962). Efficient estimates and optimum inference procedures in large samples. *Journal of the Royal Statistical Society. Series B (Methodological)*, 24(1):46–72.
- Rey, W. J. J. (1983). *Introduction to Robust and Quasi-Robust Statistical Methods*. Universitext (1979). Springer-Verlag.
- Richardson, A. M. (1997). Bounded influence estimation in the mixed linear model. *Journal of the American Statistical Association*, 92(437):154–161.
- Richardson, A. M. and Welsh, A. H. (1995). Robust restricted maximum likelihood in mixed linear models. *Biometrics*, 51(4):1429–1439.
- Richardson, S. and Green, P. J. (1997). On Bayesian analysis of mixtures with an unknown number of components. *Journal of the Royal Statistical Society. Series B (Methodological)*, 59(4):731–792.

- Robinson, G. K. (1991). That BLUP is a good thing: The estimation of random effects. *Statistical Science*, 6(1):15–32.
- Rocke, D. M. (1983). Robust statistical analysis of interlaboratory studies. *Biometrika*, 70(2):421–431.
- Rocke, D. M. (1991). Robustness and balance in the mixed model. *Biometrics*, 47(1):303–309.
- Roeder, K. (1994). A graphical technique for determining the number of components in a mixture of normals. *Journal of the American Statistical Association*, 89(426):487–495.
- Scaccia, L. and Green, P. J. (2003). Bayesian growth curves using normal mixtures with nonparametric weights. *Journal of Computational and Graphical Statistics*, 12(2):308–331.
- Scott, D. W. (1992). *Multivariate Density Estimation: Theory, Practice, and Visualization*. Wiley.
- Scott, D. W. (2001). Parametric statistical modeling by minimum integrated square error. *Technometrics*, 43(3):274–285.
- Scott, D. W. (2004). Partial mixture estimation and outlier detection in data and regression. In *Theory and Applications of Recent Robust Methods*, Statistics for Industry and Technology, pages 297–306. Birkhäuser Verlag, Basel.
- Scott, D. W. and Terrell, G. R. (1987). Biased and unbiased cross-validation in density estimation. *Journal of the American Statistical Association*, 82(400):1131–1146.
- Searle, S. R., Casella, G., and McCulloch, C. E. (1992). *Variance Components*. Wiley

- Series in Probability and Mathematical Statistics: Applied Probability and Statistics.  
John Wiley & Sons Inc., New York. A Wiley-Interscience Publication.
- Shao, J. (2003). *Mathematical Statistics*. Springer Texts in Statistics. Springer.
- Sharples, L. D. (1990). Identification and accommodation of outliers in general hierarchical models. *Biometrika*, 77(3):445–453.
- Sheather, S. J. and Jones, M. C. (1991). A reliable data-based bandwidth selection method for kernel density estimation. *Journal of the Royal Statistical Society. Series B (Methodological)*, 53(3):683–690.
- Silverman, B. W. (1986). *Density Estimation*. London: Chapman and Hall.
- Skates, S. J., Pauler, D. K., and Jacobs, I. J. (2001). Screening based on the risk of cancer calculation from Bayesian hierarchical changepoint and mixture models of longitudinal markers. *Journal of the American Statistical Association*, 96(454):429–439.
- Solaro, N. and Ferrari, P. A. (2007). Robustness of parameter estimation procedures in multilevel models when random effects are MEP distributed. *Statistical Methods and Applications*, 16(1):51–67.
- Stahel, W. A. and Welsh, A. (1997). Approaches to robust estimation in the simplest variance components model. *Journal of Statistical Planning and Inference*, 57(2):295 – 319. Robust Statistics and Data Analysis, Part II.
- Stephens, M. (2000). Bayesian analysis of mixture models with an unknown number of components - an alternative to reversible jump methods. *The Annals of Statistics*, 28(1):40–74.
- Thompson, T. J., Smith, P. J., and Boyle, J. P. (1998). Finite mixture models with con-

- comitant information: Assessing diagnostic criteria for diabetes. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, 47(3):393–404.
- Titterton, D. M., Smith, A. F. M., and Makov, U. E. (1985). *Statistical Analysis of Finite Mixture Distributions*. Wiley.
- Tukey, J. W. (1977). *Exploratory Data Analysis*. Addison-Wesley, 8 edition.
- Turner, D. A. and West, M. (1993). Bayesian analysis of mixtures applied to post-synaptic potential fluctuations. *Journal of Neuroscience Methods*, 47(1-2):1 – 21.
- Uhlig, S. (1997). *Industrial Statistics. Aims and Computational Aspects*, chapter Robust estimation of variance components with high breakdown point in the 1-way random effect model, pages 65–73. Heidelberg: Physica-Verlag.
- Vincent, L. and Soille, P. (1991). Watersheds in digital spaces: An efficient algorithm based on immersion simulations. *IEEE PAMI*, 13(6):583–598.
- Vounatsou, P., Smith, T., and Smith, A. F. M. (1998). Bayesian analysis of two-component mixture distributions applied to estimating malaria attributable fractions. *Applied Statistics*, 47(4):575–587.
- Wakefield, J. C., Smith, A. F. M., Racine-Poon, A., and Gelfand, A. E. (1994). Bayesian analysis of linear and non-linear population models by using the Gibbs sampler. *Applied Statistics*, 43(1):201–221.
- Welsh, A. H. and Richardson, A. M. (1997). Approaches to the robust estimation of mixed models. In *Robust Inference*, volume 15 of *Handbook of Statistics*, pages 343–384. North-Holland, Amsterdam.

- West, M. (1997). Hierarchical mixture models in neurological transmission analysis. *Journal of the American Statistical Association*, 92(438):587–606.
- White, H. (1982). Maximum likelihood estimation of misspecified models. *Econometrica*, 50(1):1–25.
- Windham, M. P. (1995). Robustifying model fitting. *Journal of the Royal Statistical Society, Series B (Methodological)*, 57(3):599–609.
- Woodward, W. A., Parr, W. C., Schucany, W. R., and Lindsey, H. (1984). A comparison of minimum distance and maximum likelihood estimation of a mixture proportion. *Journal of the American Statistical Association*, 79(387):590–598.
- Wright, S. P. (1998). Multivariate analysis using the MIXED procedure. In *Statistics, Data Analysis, and Modeling*, volume 23. SAS Users Group International.
- Yau, K. K. W. and Kuk, A. Y. C. (2002). Robust estimation in generalized linear mixed models. *Journal of the Royal Statistical Society. Series B (Statistical Methodology)*, 64(1):101–117.
- Yau, K. K. W., Lee, A. H., and Ng, A. S. K. (2003). Finite mixture regression model with random effects: application to neonatal hospital length of stay. *Computational Statistics & Data Analysis*, 41(3-4):359 – 366.
- Yeap, B. Y. and Davidian, M. (2001). Robust two-stage estimation in hierarchical nonlinear models. *Biometrics*, 57(1):266–272.
- Yun, S. and Lee, Y. (2006). Robust estimation in mixed linear models with non-monotone missingness. *Statistics in Medicine*, 25(22):3877–3892.

Zhan, T., Chevoneva, I., and Iglewicz, B. (2011). Generalized weighted likelihood density estimators with application to finite mixture of exponential family distributions. *Computational Statistics & Data Analysis*, 55(1):457 – 465.

Zhang, G., Ezura, Y., Chervoneva, I., Robinson, P. S., Beason, D. P., Carine, E. T., Soslowsky, L. J., Iozzo, R. V., and Birk1, D. E. (2006). Decorin regulates assembly of collagen fibrils and acquisition of biomechanical properties during tendon development. *Journal of Cellular Biochemistry*, 98:1436–1449.

# Appendix A

## Derivation and proof

### A.1 Finite exponential-family mixtures

We give a few important results about the finite mixture model of exponential-family distributions. In section A.1.1, we give the Jacobian matrix of the general logit transformations. In section A.1.2, we prove that the augmented (by the latent component indicator variable) distribution of the finite exponential-family mixtures model belongs to the exponential family. In section A.1.3, we develop the EM algorithm from the derivatives of conditional expectation of complete log-likelihood. In section A.1.4, we give the estimating equations in the form of weighted score functions. We also give the robust EM updating steps for the finite exponential-family mixtures model. In section A.1.5, we give the Fisher information matrix and the asymptotic covariance matrix of the finite exponential-family mixtures model.

### A.1.1 General logit link for multinomial distribution

In this section, we discuss the general logit parametrization of the multinomial probabilities and provide the Jacobian matrix of this transformation. Consider the multinomial distribution with parameters  $\tilde{\boldsymbol{\pi}} = (\tilde{\pi}_1, \dots, \tilde{\pi}_d)'$ , which satisfy the restriction  $\mathbf{1}'\tilde{\boldsymbol{\pi}} = 1$ . However, instead of working with the restricted parameters  $\tilde{\boldsymbol{\pi}}$ , we consider the unrestricted set of mixing proportions  $\boldsymbol{\pi} = (\pi_1, \dots, \pi_d)'$ , which satisfy that  $\tilde{\pi}_t = \pi_t / \sum_s \pi_s$ . The unrestricted parametrization  $\boldsymbol{\pi}$  is not unique, but it is symmetric and of  $d$  degree-of-freedom. Define the natural-logs of the unrestricted proportions  $\boldsymbol{\tau} = \log \boldsymbol{\pi}$ , which are not unique neither. The general logits of the mixing proportions are

$$\tilde{\boldsymbol{\tau}} = (0, \tilde{\tau}_2, \dots, \tilde{\tau}_d)', \text{ where } \tilde{\tau}_1 = 0 \text{ and } \tilde{\tau}_s = \tau_s - \tau_1 \text{ for } s = 2, \dots, d \quad (\text{A.1})$$

The general logits (A.1) is unique and the relationship between the restricted mixing proportion  $\tilde{\boldsymbol{\pi}}$  and the general logits  $\tilde{\boldsymbol{\tau}}$  is

$$\tilde{\pi}_t = \exp \tilde{\tau}_t / \sum_s \exp \tilde{\tau}_s \quad (\text{A.2})$$

The Jacobian matrix elements of the transformation from  $\tilde{\boldsymbol{\pi}}$  to  $\tilde{\boldsymbol{\tau}}$  are

$$\frac{\partial \tilde{\pi}_t}{\partial \tilde{\tau}_t} = \frac{\exp \tilde{\tau}_t}{\sum_s \exp \tilde{\tau}_s} \left( 1 - \frac{\exp \tilde{\tau}_t}{\sum_s \exp \tilde{\tau}_s} \right) = \tilde{\pi}_t (1 - \tilde{\pi}_t), \quad t \neq 1 \quad (\text{A.3a})$$

$$\frac{\partial \tilde{\pi}_u}{\partial \tilde{\tau}_t} = - \left( \frac{\exp \tilde{\tau}_u}{\sum_s \exp \tilde{\tau}_s} \right) \left( \frac{\exp \tilde{\tau}_t}{\sum_s \exp \tilde{\tau}_s} \right) = -\tilde{\pi}_u \tilde{\pi}_t, \quad t \neq u, t \neq 1, u \neq 1 \quad (\text{A.3b})$$

$$\frac{\partial \tilde{\pi}_1}{\partial \tilde{\tau}_t} = \frac{\partial (\sum_s \exp \tilde{\tau}_s)^{-1}}{\partial \tilde{\tau}_t} = - \left( \frac{\exp \tilde{\tau}_t}{\sum_s \exp \tilde{\tau}_s} \right) \left( \frac{1}{\sum_s \exp \tilde{\tau}_s} \right) = -\tilde{\pi}_t \tilde{\pi}_1, \quad t \neq 1 \quad (\text{A.3c})$$

### A.1.2 Augmented distribution of finite mixture model

Consider a  $k$ -component mixtures model, where each component comes from a same member distribution  $\phi(\cdot)$  of the general exponential family. Let  $Y$  be the observed values.

The marginal likelihood is

$$f_Y(\Xi, \boldsymbol{\pi}) = \sum_{s=1}^k \tilde{\pi}_s \cdot \phi(y; \boldsymbol{\xi}_s) \quad (\text{A.4})$$

where  $\tilde{\pi}_s$ 's are mixing proportions summed up to 1,  $\boldsymbol{\xi}_s = (\xi_{s1}, \dots, \xi_{sQ})^t$  are the canonical parameters for the  $s$ th component, and  $\Xi$  is the  $k$  by  $Q$  canonical parameters matrix such that  $\Xi = (\boldsymbol{\xi}_1, \dots, \boldsymbol{\xi}_k)^t$ . Let  $Z$  be the corresponding latent component indicator taking values  $\mathbf{e}_s$ ,  $s = 1, \dots, k$ , following multinomial distribution with restricted parameters  $\tilde{\boldsymbol{\pi}}$ . The base vector  $\mathbf{e}_s$  is a length- $k$  vector whose  $j$ th element is 1 and others are 0. Thus the finite mixture model (A.4) can be written in the missing data form as

$$f_{Y|Z=\mathbf{e}_s}(\Xi) = \phi(y; \boldsymbol{\xi}_s); \quad Z \sim \text{multinomial}(\tilde{\boldsymbol{\pi}}) \quad (\text{A.5})$$

Let  $\boldsymbol{\tau}$  be the natural logs of  $\boldsymbol{\pi}$ , the marginal log-likelihood of  $Z$  is

$$\log f_{Z=\mathbf{e}_s}(\boldsymbol{\tau}) = \tau_Z - \log \left( \sum_s \exp \tau_s \right) = \mathbf{e}_s^t \boldsymbol{\tau} - C(\boldsymbol{\tau}) \quad (\text{A.6})$$

where  $C(\boldsymbol{\tau}) = \log[\sum_s \exp(\tau_s)]$  is the normalization factor of the multinomial distribution. For each component distribution ( $Y|Z = \mathbf{e}_s$ )  $\sim \phi(y; \boldsymbol{\xi}_s)$ , let the mean parameters be  $\boldsymbol{\eta}_s = (\eta_{s1}, \dots, \eta_{sQ})^t$  and the normalization factor be  $b(\boldsymbol{\xi}_s) = b(\xi_{s1}, \dots, \xi_{sQ})$ . Let  $\mathbf{T}_Y = (T_{Y,1}, \dots, T_{Y,Q})^t$  be the vector of sufficient statistics based on random sample  $Y$ . Define the length- $k$  normalization vector  $b(\Xi) = \{b(\boldsymbol{\xi}_s)\}$ ,  $s = 1, \dots, k$ . The log-likelihood of  $Y|Z$  is

$$\log f_{Y|Z=\mathbf{e}_s}(\Xi) = \mathbf{e}_s^t \cdot \Xi \cdot \mathbf{T}_Y - \mathbf{e}_s^t b(\Xi) \quad (\text{A.7})$$

Therefore, the log-likelihood based on complete data  $(Y, Z)$

$$\log f_{Y,Z=\mathbf{e}_s}(\Xi, \boldsymbol{\tau}) = (\mathbf{e}_s^t [\boldsymbol{\tau} - b(\Xi)] + \mathbf{e}_s^t \cdot \Xi \cdot \mathbf{T}_Y) - C(\boldsymbol{\tau}) \quad (\text{A.8})$$

belongs to the exponential family with canonical parameters  $\Xi$  and a new length- $k$  parameter  $\xi^{(\tau)} = \tau - b(\Xi)$  with elements  $\xi_s^{(\tau)} = \tau_s - b(\xi_s)$ . If we let  $\xi^{(q)} = (\xi_{1q}, \dots, \xi_{kq})^t$  be the  $q$ th column of  $\Xi$ , the complete log-likelihood (A.8) is equivalent to

$$\log f_{Y,Z=e_s}(\xi^{(\tau)}, \Xi) = \left( e_s^t \xi^{(\tau)} + \sum_{q=1}^Q T_{Y,q} e_s^t \xi^{(q)} \right) - B(\xi^{(\tau)}, \Xi) \quad (\text{A.9})$$

thus it's clear that the complete log-likelihood (A.9) belongs to the exponential family with canonical parameters  $\xi^{(\tau)}$  and  $\xi^{(q)}$ ,  $q = 1, \dots, Q$ . Based on a random sample of size  $n$ , the corresponding sufficient statistics are

$$T^{(\tau)}(Y, Z) = \sum_{i=1}^n e_{s_i} \quad (\text{A.10a})$$

$$T^{(q)}(Y, Z) = \sum_{i=1}^n T_{Y_i,q} e_{s_i}, \quad q = 1, \dots, Q \quad (\text{A.10b})$$

On the other hand, the normalization factor in (A.9) is indeed the normalization factor of multinomial distribution,

$$B(\xi^{(\tau)}, \Xi) = C(\tau) = \log \left( \sum_s \exp \left( \xi_s^{(\tau)} + b(\xi_s) \right) \right) \quad (\text{A.11})$$

thus the corresponding mean parameters are

$$\mathbb{E} \left( T^{(\tau)} \right) = \partial B / \partial \xi^{(\tau)} = n(\tilde{\pi}_1, \dots, \tilde{\pi}_k)^t \quad (\text{A.12a})$$

$$\mathbb{E} \left( T^{(q)} \right) = \partial B / \partial \xi^{(q)} = n(\tilde{\pi}_1 \eta_{1q}, \dots, \tilde{\pi}_k \eta_{kq})^t, \quad q = 1, \dots, Q \quad (\text{A.12b})$$

For example, the sufficient statistics of the augmented finite normal mixtures model are

$$T^{(\tau)}(Y, Z) = \sum_i e_{s_i}, \quad T^{(1)}(Y, Z) = \sum_i y_i e_{s_i} \quad \text{and} \quad T^{(2)}(Y, Z) = \sum_i y_i^2 e_{s_i}.$$

### A.1.3 EM algorithm

If we consider the component indicator variable  $z$  as missing data, the finite mixture model (A.5) can be estimated by EM algorithm. The updating steps of EM algorithm can

be obtained in two ways. One is from the standard expectation-maximization approach, the other is solving the so-called Sundberg formulas (McLachlan and Krishnan, 2008) of conditional and unconditional expectation of sufficient statistics. In this section we give the formulas needed in both of these approaches, as these formulas themselves will be used later in this work. For simplicity, we use  $[\cdot]_{s=1,\dots,k}$  to represent a length- $k$  vector with corresponding elements, and  $\langle \cdot \rangle_{s=1,\dots,k}$  to represent a diagonal matrix of order  $k$  with corresponding diagonal elements.

The first approach is the standard expectation-maximization procedure, calculating the conditional expectation of the complete log-likelihood (A.9) and maximizing it by having it's derivatives equal to 0. The posterior (or conditional) expectation  $\mathbf{p}_i = (p_{1,i}, \dots, p_{k,i})^t$  of missing data  $z_i$ , given observed data  $y_i$ , is

$$p_{s,i} = \Pr(Z_i = e_s | y_i, \Xi, \boldsymbol{\pi}) = \frac{\pi_s \phi(y_i, \boldsymbol{\xi}_s)}{\sum_{t=1}^k \pi_t \phi(y_i, \boldsymbol{\xi}_t)} = \frac{\tilde{\pi}_s \phi(y_i; \boldsymbol{\xi}_s)}{f(y_i; \Xi, \boldsymbol{\pi})} \quad (\text{A.13})$$

The conditional expectation of the complete log-likelihood (A.9) is

$$\mathbb{E}_{\mathbf{z}|\mathbf{y}}(\log f_{Y,Z=e_s}) = \sum_{s=1}^d p_s \left( \boldsymbol{\xi}_s^{(\tau)} + \sum_{q=1}^Q T_{Y,q} \boldsymbol{\xi}_{sq} \right) - B(\boldsymbol{\xi}^{(\tau)}, \Xi) \quad (\text{A.14})$$

and the derivatives of conditional expectation (A.14), with respect to canonical parameters  $\boldsymbol{\xi}^{(\tau)}$  and  $\Xi$ , are

$$\frac{\partial \mathbb{E}_{\mathbf{z}|\mathbf{y}}(\log f_{Y,Z=e_s})}{\partial \boldsymbol{\xi}^{(\tau)}} = \left[ p_s \right]_s - \tilde{\boldsymbol{\pi}} \quad (\text{A.15a})$$

$$\frac{\partial \mathbb{E}_{\mathbf{z}|\mathbf{y}}(\log f_{Y,Z=e_s})}{\partial \boldsymbol{\xi}^{(q)}} = \left[ p_s T_{Y,q} - \tilde{\pi}_s \eta_{sq} \right]_s, \quad q = 1, \dots, Q \quad (\text{A.15b})$$

The second approach solves equations between conditional and unconditional expectations, also known as the Sundberg formulas

$$\mathbb{E}(T(Y, Z) | Y, \Xi, \boldsymbol{\pi}) = \mathbb{E}(T(Y, Z) | \Xi, \boldsymbol{\pi})$$

Given a random sample of size  $n$ , the left-hand-side conditional expectations are

$$\mathbb{E}(T^{(\tau)}|y, \Xi, \boldsymbol{\pi}) = \sum_{i=1}^n \sum_{s=1}^k \mathbf{e}_s \cdot p_{s,i}; \quad (\text{A.16a})$$

$$\mathbb{E}(T^{(q)}|y, \Xi, \boldsymbol{\pi}) = \sum_{i=1}^n \sum_{s=1}^k \mathbf{e}_s \cdot T_{y_i,q} p_{s,i}, \quad q = 1, \dots, Q \quad (\text{A.16b})$$

The right-hand-side unconditional expectations are

$$\mathbb{E}(T^{(\tau)}|\Xi, \boldsymbol{\pi}) = n \sum_{s=1}^k \mathbf{e}_s \cdot \tilde{\pi}_s \quad (\text{A.17a})$$

$$\mathbb{E}(T^{(q)}|\Xi, \boldsymbol{\pi}) = \mathbb{E}[\mathbb{E}(T^{(q)}(Y, Z)|\boldsymbol{\pi})|\Xi] = n \sum_{s=1}^k \mathbf{e}_s \cdot \tilde{\pi}_s \eta_{sq}, \quad q = 1, \dots, Q \quad (\text{A.17b})$$

Either the derivatives (A.15) of conditional expectation, or the Sundberg formula (A.16) with (A.17), give us following updating equations of ordinary EM algorithm

$$\tilde{\pi}_s^{\text{new}} = \frac{1}{n} \sum_{i=1}^n p_{s,i}, \quad s = 1, \dots, k \quad (\text{A.18a})$$

$$\eta_{sq}^{\text{new}} = \left( \frac{1}{n} \sum_{i=1}^n T_{y_i,q} p_{s,i} \right) / \tilde{\pi}_s^{\text{new}}, \quad s = 1, \dots, k, \quad q = 1, \dots, Q \quad (\text{A.18b})$$

#### A.1.4 Weighted score functions

The score functions are the derivatives of the marginal log-likelihood (A.4) based on observed data  $y$ . Since the marginal distribution (A.4) does not belongs to the exponential family distributions, we take these derivatives with respect to mixing proportions  $\boldsymbol{\pi}$  and canonical parameters  $\boldsymbol{\xi}$ 's of the subpopulations. Given random sample  $Y_1, \dots, Y_n$ , the vectors of score functions, as the derivatives of  $\log f_Y(\Xi, \boldsymbol{\pi})$ , are

$$\mathbf{u}(\mathbf{y}; \boldsymbol{\pi}) = \sum_{i=1}^n \left\langle \frac{1}{\pi_s} \right\rangle_s \left[ \frac{\tilde{\pi}_s \phi(y_i; \boldsymbol{\xi}_s)}{f(y_i; \Xi, \boldsymbol{\pi})} - \tilde{\pi}_s \right]_s = \sum_{i=1}^n \left\langle \frac{1}{\pi_s} \right\rangle_s \left( [p_{s,i}]_s - \tilde{\pi} \right) \quad (\text{A.19a})$$

$$\mathbf{u}(\mathbf{y}; \boldsymbol{\xi}^{(q)}) = \sum_{i=1}^n \left[ \frac{\tilde{\pi}_s \phi(y_i; \boldsymbol{\xi}_s)}{f(y_i; \Xi, \boldsymbol{\pi})} \cdot (T_{y_i,q} - \eta_{sq}) \right]_s = \sum_{i=1}^n [p_{s,i} \cdot (T_{y_i,q} - \eta_{sq})]_s \quad (\text{A.19b})$$

Denote the elements of these score functions as  $\mathbf{u}(\mathbf{y}; \boldsymbol{\pi}) = \{u_s(\mathbf{y}; \boldsymbol{\pi})\}$  and  $\mathbf{u}(\mathbf{y}; \boldsymbol{\xi}^{(q)}) = \{u_s(\mathbf{y}; \boldsymbol{\xi}^{(q)})\}$ , for  $s = 1, \dots, d$ . The updating steps for the iterative solution of score func-

tions (A.19) are equivalent to the EM updating steps (A.18). The weighted score functions discussed in Chapter 3 apply weights  $w(y_i)$  and the bias adjustment term  $\mathbf{a}(\Xi, \boldsymbol{\pi})$  onto the score functions (A.19),

$$\mathbf{u}^*(\mathbf{y}; \boldsymbol{\pi}) = \sum_{i=1}^n w(y_i) \left\langle \frac{1}{\pi_s} \right\rangle_s \left( \left[ p_{s,i} \right]_s - \tilde{\boldsymbol{\pi}} \right) - n\mathbf{a}(\Xi, \boldsymbol{\pi})_{:\boldsymbol{\pi}} \quad (\text{A.20a})$$

$$\mathbf{u}^*(\mathbf{y}; \boldsymbol{\xi}^{(q)}) = \sum_{i=1}^n w(y_i) \left[ p_{s,i} \cdot (T_{y_i,q} - \eta_{sq}) \right]_s - n\mathbf{a}(\Xi, \boldsymbol{\pi})_{:\boldsymbol{\xi}^{(q)}}, \quad q = 1, 2 \quad (\text{A.20b})$$

The corresponding weighted likelihood estimating equations let (A.20) equal to  $\mathbf{0}$  and element-wise they are

$$0 = \sum_{i=1}^n w(y_i) \left( p_{s,i} - \tilde{\pi}_s \right) - n\pi_s \mathbf{a}(\Xi, \boldsymbol{\pi})_{:\pi_s} \approx \sum_{i=1}^n w(y_i) p_{s,i} - n\tilde{\pi}_s - n\pi_s \mathbf{a}(\Xi, \boldsymbol{\pi})_{:\pi_s} \quad (\text{A.21a})$$

$$0 = \sum_{i=1}^n w(y_i) \cdot p_{s,i} \cdot (T_{y_i,q} - \eta_{sq}) - n\mathbf{a}(\Xi, \boldsymbol{\pi})_{:\boldsymbol{\xi}_{sq}} \quad (\text{A.21b})$$

The approximation in (A.21a) follows from the fact that  $w(y_i) \approx 1$  for all  $y_i$ 's. The WLE updating equations discussed in Chapter 3 provides an iterative solution to (A.21). For the special case of finite normal mixtures, these updating equations reduce to those in Fujisawa and Eguchi (2006).

We have shown in Appendix section A.1.3 and A.1.4 that although the derivatives of conditional expectations (A.15) are not equivalent to the score functions (A.19), the derivatives of marginal log-likelihood, the iterative updating steps for solving these two sets of equations are essentially the same for the finite exponential-family mixture model. We use this fact to formulate the robust estimation of LME with conditional finite normal mixtures model discussed in Chapter 4.

The basic idea is to apply weight  $w(y_i)$  on derivatives (A.15), just as we did for weighted score equations in Chapter 3, and these "weighted derivatives" will be used as building blocks in the robustified BLUP procedure. We have shown that for weighted likelihood estimators

which belong to Lindsay (1994)'s family, the bias adjustment term  $\mathbf{a}(\Xi, \boldsymbol{\pi}) = 0$ . Thus the weighted derivatives modified from (A.15), based on  $(y_i, z_i)$ ,  $i = 1, \dots, n$ , are

$$\left( \frac{\partial E_{\mathbf{z}|\mathbf{y}}(\log f_{Y,Z=e_s})}{\partial \boldsymbol{\xi}^{(\tau)}} \right)^* = \sum_{i=1}^n w(y_i) \left( [p_{s,i}]_s - \tilde{\boldsymbol{\pi}} \right) \approx \sum_{i=1}^n w(y_i) [p_{s,i}]_s - n\tilde{\boldsymbol{\pi}} \quad (\text{A.22a})$$

$$\left( \frac{\partial E_{\mathbf{z}|\mathbf{y}}(\log f_{Y,Z=e_s})}{\partial \boldsymbol{\xi}^{(q)}} \right)^* = \sum_{i=1}^n w(y_i) [p_s T_{Y,q} - \tilde{\boldsymbol{\pi}}_s \eta_{sq}]_s, \quad q = 1, \dots, Q \quad (\text{A.22b})$$

Thus in the robust BLUP procedure of Chapter 4, we do not look for a robust  $l_1^*$  to substitute the conditional loglikelihood  $l_1$ , instead we substitute the derivatives  $\partial l_1 / \partial (\text{canonical par.})$  with (A.22), as only the derivatives are what we concern about.

### A.1.5 Fisher information and asymptotic normality

The Fisher information of the finite mixture density (A.4) and the variance of the score functions will not be equivalent even when all weights  $w_i = 1$ . The calculation requires the "full rank" score function with respect to canonical parameter  $\boldsymbol{\xi}$ 's of the mixing components and the general logit parameters  $\tilde{\boldsymbol{\tau}}$ . The score function (A.19a) with respect to mixing proportions  $\boldsymbol{\pi}$  is not "full rank", as  $\boldsymbol{\pi}$ 's are not independent. Let

$$\mathbf{u}(\mathbf{y}; \tilde{\boldsymbol{\tau}}) = \frac{\partial \tilde{\boldsymbol{\pi}}}{\partial \tilde{\boldsymbol{\tau}}} \cdot \mathbf{u}(\mathbf{y}; \tilde{\boldsymbol{\pi}}) = (\text{A.3}) \cdot (\text{A.19a}) \quad (\text{A.23})$$

and the "full-rank" score function be  $\mathbf{u}(\mathbf{y}; \tilde{\boldsymbol{\tau}}, \boldsymbol{\xi}) = (\mathbf{u}^t(\mathbf{y}; \tilde{\boldsymbol{\tau}}), \mathbf{u}^t(\mathbf{y}; \boldsymbol{\xi}^{(1)}), \mathbf{u}^t(\mathbf{y}; \boldsymbol{\xi}^{(2)}))^t$ . The variance of the score function is

$$\mathcal{I}_1(\tilde{\boldsymbol{\tau}}, \boldsymbol{\xi}) = E \left( \mathbf{u}(\mathbf{y}; \tilde{\boldsymbol{\tau}}, \boldsymbol{\xi}) \mathbf{u}^t(\mathbf{y}; \tilde{\boldsymbol{\tau}}, \boldsymbol{\xi}) \right) \quad (\text{A.24})$$

The second method calculates the expected second order derivative of the log-likelihood  $l = \log f_Y(\Xi, \boldsymbol{\pi})$ . Let  $\phi_s = \phi(y, \boldsymbol{\xi}_s)$ , the second order derivatives with respect to  $\boldsymbol{\xi}$ 's and  $\boldsymbol{\pi}$

are

$$\begin{aligned}\frac{\partial^2 l}{\partial \pi_t \partial \pi_u} &= \left( \frac{1}{\sum \pi_s} \right)^2 \left( 1 - \frac{\phi_t \phi_u}{f^2} \right) \\ \frac{\partial^2 l}{\partial \pi_t \partial \xi_{uq}} &= -\frac{\pi_u \phi_u \phi_t (T_q - \eta_{uq})}{(\sum \pi_s \phi_s)^2}, \quad t \neq u \\ \frac{\partial^2 l}{\partial \pi_t \partial \xi_{tq}} &= \frac{\phi_t (T_q - \eta_{tq}) (\sum \pi_s \phi_s - \pi_t \phi_t)}{(\sum \pi_s \phi_s)^2} \\ \frac{\partial^2 l}{\partial \xi_{tq_1} \partial \xi_{uq_2}} &= -\frac{\pi_t \pi_u \phi_t \phi_u (T_{q_1} - \eta_{tq_1}) (T_{q_2} - \eta_{uq_2})}{(\sum \pi_s \phi_s)^2}, \quad t = u \text{ and } t \neq u\end{aligned}$$

If we use  $\tilde{\pi}$  instead of  $\pi$  and define  $\bar{\phi} = (\phi_1, \dots, \phi_d)'$ ,  $\bar{\phi}_{\tilde{\pi}} = (\tilde{\pi}_1 \phi_1, \dots, \tilde{\pi}_d \phi_d)'$  and  $\bar{\mathbf{T}}_q^\eta = (T_q - \eta_{1q}, \dots, T_q - \eta_{dq})'$ , these derivatives take matrix form

$$\begin{aligned}\frac{\partial^2 l}{\partial \tilde{\pi} \partial \tilde{\pi}'} &= \mathbf{1}_d - (f^{-1} \bar{\phi}) (f^{-1} \bar{\phi})' \\ \frac{\partial^2 l}{\partial \tilde{\pi} \partial (\boldsymbol{\xi}^{(q)})'} &= \left( (f^{-1} \bar{\phi}) (\bar{\mathbf{T}}_q^\eta)' \right) \otimes \left( \mathbf{I}_d - \mathbf{1}_d (f^{-1} \bar{\phi}_{\tilde{\pi}})' \right) \\ \frac{\partial^2 l}{\partial \boldsymbol{\xi}^{(q_1)} \partial (\boldsymbol{\xi}^{(q_2)})'} &= - \left( \bar{\mathbf{T}}_{q_1}^\eta (\bar{\mathbf{T}}_{q_2}^\eta)' \right) \otimes \left( (f^{-1} \bar{\phi}_{\tilde{\pi}}) (f^{-1} \bar{\phi}_{\tilde{\pi}})' \right)\end{aligned}$$

where  $\otimes$  denote the element-wise multiplication of vectors or matrices. The Fisher information matrix is

$$\mathcal{I}_2(\tilde{\tau}, \boldsymbol{\xi}) = \text{Diag} \left\{ \frac{\partial \tilde{\pi}}{\partial \tilde{\tau}}, \mathbf{I}_d, \mathbf{I}_d \right\} \text{E} \left( -\frac{\partial^2 l}{\partial (\tilde{\pi}, \boldsymbol{\xi}^{(1)}, \boldsymbol{\xi}^{(2)})^2} \right) \text{Diag} \left\{ \frac{\partial \tilde{\pi}}{\partial \tilde{\tau}}, \mathbf{I}_d, \mathbf{I}_d \right\} \quad (\text{A.27})$$

The Cramer-Rao lower bound for the asymptotic covariance matrix is

$$\left( \mathcal{I}_2(\tilde{\tau}, \boldsymbol{\xi}) \right)^{-1} \mathcal{I}_1(\tilde{\tau}, \boldsymbol{\xi}) \left( \mathcal{I}_2(\tilde{\tau}, \boldsymbol{\xi}) \right)^{-1} \quad (\text{A.28})$$

## A.2 Influence function of GWLE

*Proof of Theorem 3.1.* Given a fixed distribution  $G$ , contaminated  $G_{\varepsilon, x_0} = (1-\varepsilon)G + \varepsilon\Delta(x_0)$  and model family  $\mathcal{F}_\Theta$ , the estimating equation (3.1) for GWLE  $T(\cdot)$  with weight  $w(x, \delta_\varepsilon^*, f_\theta)$  is equivalent to

$$\int w(x) \mathbf{u}_\theta(x) \cdot (g_{\varepsilon, x_0}(x) - f_\theta(x)) dx = 0 \quad (\text{A.29})$$

where  $\theta = T(G_{\varepsilon, x_0})$  and  $k(x; t, h)$ -smoothed Pearson residual  $\delta_\varepsilon^*(x) = g_\varepsilon^*(x)/f_\theta^*(x) - 1$ . The general theory about influence function of M-estimator (Hampel et al., 1985) is not applicable. Here let  $(\cdot)^t$  denote transposition of vectors and take the derivative of equation (A.29) with respect to  $\varepsilon$ ,

$$\begin{aligned} 0 = & \int \left( \frac{\partial \delta_\varepsilon^*}{\partial \varepsilon} \frac{\partial w(\delta^*, f_\theta)}{\partial \delta_\varepsilon^*} + \frac{\partial \theta^t}{\partial \varepsilon} \frac{\partial f_\theta}{\partial \theta} \frac{\partial w(\delta^*, f_\theta)}{\partial f_\theta} \right) \mathbf{u}_\theta^t (g_{\varepsilon, x_0} - f_\theta) dx \\ & + \int w(\delta^*, f_\theta) \frac{\partial \theta^t}{\partial \varepsilon} \frac{\partial \mathbf{u}_\theta^t}{\partial \theta} (g_{\varepsilon, x_0} - f_\theta) dx \\ & + \int w(\delta^*, f_\theta) \mathbf{u}_\theta^t \left( (\chi_{x_0} - g) - \frac{\partial \theta^t}{\partial \varepsilon} \frac{\partial f_\theta}{\partial \theta} \right) dx \end{aligned} \quad (\text{A.30})$$

where

$$\left. \frac{\partial \theta}{\partial \varepsilon} \right|_{\varepsilon=0} = \text{IF}(x_0), \quad \frac{\partial g_{\varepsilon, x_0}^*(x)}{\partial \varepsilon} = k(x; x_0, h) - g^*(x)$$

and  $\mathbf{u}_\theta^*(x) = [f_\theta^*(x)]^{-1} \partial f_\theta^*(x)/\partial \theta$ . Therefore

$$\left. \frac{\partial \delta^*}{\partial \varepsilon} \right|_{\varepsilon=0} = \frac{k(x; x_0, h) - g^*(x)}{f_\theta^*(x)} - \frac{g^*(x)}{f_\theta^*(x)} \cdot \text{IF}^t \cdot \mathbf{u}_\theta^*(x) \quad (\text{A.31})$$

Substitute (A.31) into (A.30) and evaluate at  $\varepsilon = 0$ ,

$$\begin{aligned} 0 = & \int \left( \frac{k(x; x_0, h) - g^*(x)}{f_\theta^*(x)} - (\delta^* + 1) \text{IF}^t \mathbf{u}_\theta^* \right) w'_\delta \mathbf{u}_\theta^t \cdot (g - f_\theta) dx \\ & + \int \text{IF}^t f_\theta w'_f \mathbf{u}_\theta \mathbf{u}_\theta^t \cdot (g - f_\theta) dx + \int \text{IF}^t w \frac{\partial \mathbf{u}_\theta^t}{\partial \theta} \cdot (g - f_\theta) dx \\ & + w(x_0) \mathbf{u}_\theta^t(x_0) - \int w \mathbf{u}_\theta^t g dx - \int \text{IF}^t f_\theta w \mathbf{u}_\theta \mathbf{u}_\theta^t dx \end{aligned} \quad (\text{A.32})$$

and substitute  $f_\theta \cdot (\partial \mathbf{u}_\theta^t / \partial \theta + \mathbf{u}_\theta \mathbf{u}_\theta^t) = \partial (f_\theta \mathbf{u}_\theta)^t / \partial \theta = \partial^2 f_\theta / \partial \theta^2$  into (A.32), one could solve for influence functions (2.8). Note this is influence function of the WLE from (3.1), which is an approximation of minimum disparity estimator of Lindsay (1994); for influence function of the latter, refer to Basu and Lindsay (1994). The proof is applicable to empirical distribution  $\hat{G}_n$  and contaminated  $\hat{G}_{n, \varepsilon, x}$ .  $\square$