2018

# Mixtures-of-Regressions with Measurement Error

Xiaoqiong Fang
*University of Kentucky*, fxq824@gmail.com
Digital Object Identifier: https://doi.org/10.13023/etd.2018.489

STUDENT AGREEMENT:

I represent that my thesis or dissertation and abstract are my original work. Proper attribution has been given to all outside sources. I understand that I am solely responsible for obtaining any needed copyright permissions. I have obtained needed written permission statement(s) from the owner(s) of each third-party copyrighted matter to be included in my work, allowing electronic distribution (if such use is not permitted by the fair use doctrine) which will be submitted to UKnowledge as Additional File.

I hereby grant to The University of Kentucky and its agents the irrevocable, non-exclusive, and royalty-free license to archive and make accessible my work in whole or in part in all forms of media, now or hereafter known. I agree that the document mentioned above may be made available immediately for worldwide access unless an embargo applies.

I retain all other ownership rights to the copyright of my work. I also retain the right to use in future works (such as articles or books) all or part of my work. I understand that I am free to register the copyright to my work.

REVIEW, APPROVAL AND ACCEPTANCE

The document mentioned above has been reviewed and accepted by the student's advisor, on behalf of the advisory committee, and by the Director of Graduate Studies (DGS), on behalf of the program; we verify that this is the final, approved version of the student's thesis including all changes required by the advisory committee. The undersigned agree to abide by the statements above.

Xiaoqiong Fang, Student

Dr. Derek S. Young, Major Professor

Dr. Constance Wood, Director of Graduate Studies

Mixtures-of-Regressions with Measurement Error

---

DISSERTATION

---

A dissertation submitted in partial
fulfillment of the requirements for
the degree of Doctor of Philosophy
in the College of Arts and Sciences
at the University of Kentucky

By
Xiaoqiong Fang
Lexington, Kentucky

Director: Dr. Derek S. Young, Assistant Professor of Statistics
Lexington, Kentucky

2018

ABSTRACT OF DISSERTATION

Mixtures-of-Regressions with Measurement Error

Finite Mixture model has been studied for a long time, however, traditional methods assume that the variables are measured without error. Mixtures-of-regression model with measurement error imposes challenges to the statisticians, since both the mixture structure and the existence of measurement error can lead to inconsistent estimate for the regression coefficients. In order to solve the inconsistency, We propose series of methods to estimate the mixture likelihood of the mixtures-of-regressions model when there is measurement error, both in the responses and predictors. Different estimators of the parameters are derived and compared with respect to their relative efficiencies. The simulation results show that the proposed estimation methods work well and improve the estimating process.

KEYWORDS: mixtures-of-regression, measurement error, EM algorithm, Poisson regression

Author's signature: Xiaoqiong Fang

Date: December 10, 2018

Mixtures-of-Regressions with Measurement Error

By

Xiaoqiong Fang

Director of Dissertation: Derek S. Young

Director of Graduate Studies: Constance Wood

Date: December 10, 2018

# ACKNOWLEDGMENTS

I would like to express my sincere gratitude to my advisor Dr. Derek Young for his help with my Ph.D study and research, for his patience, motivation, and genius ideas. His guidance helped me in all the time of research and writing of this dissertation.

Besides my advisor, I would like to thank the committee members of my Ph.D. dissertation: Prof. Stromberg, Prof. Charnigo, Prof. Zhou, Prof. Bollinger and Dr. Ya Su, for their insightful encouragement, but also for the comments and suggestions which helped me to refine my research from various perspectives.

I would also like to thank my fellow doctoral students for their feedback, cooperation and of course friendship. I would like to thank my friends for accepting nothing less than excellence from me.

Last but not the least, I would like to thank my parents for supporting me spiritually throughout writing this dissertation and my my life in general.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

# Chapter 1 Introduction

Finite mixture models have been used for more than 100 years, and have seen a boost in their utility since the 1990s due to the substantial increase in computing power. The importance of mixture models is remarked by a number of books dedicated to the topic including Titterington et al. (1985) [110], McLachlan and Basford (1988) [76], Lindsay (1995) [72] and McLachlan and Peel (2000) [78]. The areas of application of mixture models range from biology and medicine to physics, economics and marketing. These models can be applied to characterize the presence of sub-populations within a broader population when knowledge about to which sub-population each observation belongs is unavailable, and also to provide approximations for multi-modal distributions.

Finite mixture models have been extended to mixtures of linear regression models (De Veaux (1989) [34]) as well as mixtures of generalized linear models (Wedel and DeSarbo (1995) [114]). Mixtures-of-experts models (Jacobs et al. (1991) [59]) and their generalization, hierarchical mixtures-of-expert models, (Jordan and Jacobs (1994) [62]) have been introduced to account for nonlinearities and other complexities in the data; Carvalho and Tanner (2009) [24] studied a class of hierarchical mixtures of Poisson experts to model nonlinear count time series; Hurn, Justel and Robert (2003) [56] showed how Bayesian inference for mixtures of regression models and their generalizations can be achieved by the specification of loss functions, which addresses the label switching problem when estimating mixture models.

Most of the existing inference procedures for mixtures-of-regressions models are limited to directly observed predictors. However, in actual problems, it is common to observe variables subject to measurement errors. *Measurement error models* (or *errors-in-variables models*) are regression models that account for measurement errors in the independent variables. The statistical analysis of errors-in-variables data has a long history, dating back to the days of econometrics as early as the 1930s (Frisch (1934) [44]). Methods of measurement error primarily aimed at linear models are

discussed by Fuller (1987) [1] and Cheng and Ness (1998) [27]. The popular book by Carroll et al. (2006) [94] covers nonlinear measurement error models, with a special focus on bias reduction (also called approximate consistency).

Measurement error might either be introduced by the measuring technique involving the subjective judgment by human action, or due to a more convenient substitution of the correct quantity. In the case when some variables have been measured with errors, estimation based on the standard assumption leads to inconsistent estimates, meaning that the parameter estimates do not tend to the true values even in very large samples. For simple linear regression with measurement error in the predictors, it can cause an underestimate of the coefficient, known as *attenuation bias*; in nonlinear models the direction of the bias is likely to be more complicated. The bias in parameter estimation for statistical modeling and analysis can lead to a loss of power, and mask certain features of the data.

Estimation of the mixtures-of-regressions model with measurement error has received limited attention in the literature. This dissertation will focus on estimation of various mixtures-of-regressions models where measurement error is assumed present.

### 1.1 Finite Mixture Models

Finite mixture models have long been used as a way to model a sample of observations that arise from a number of (usually) *a priori* known classes with unknown proportions. They provide a statistical model for a wide variety of random phenomena. Applications of mixture distributions can be found in various fields of statistical applications such as agriculture (Xu et al. (2010) [117]), biology (Bailey and Elkan (1994) [8]), economics (Liesenfeld (2001) [71]), medicine (Peng and Dear (2000) [90]) and genetics (Pagel and Meade (2004) [87]). Monographs concerning mixture modeling include Titterington et al. (1985) [110] and McLachlan and Peel (2000) [78].

Even if there is no realistic interpretation of the components of the mixture model, mixture distributions offer a very flexible modeling environment. We consider a parametric framework where the components are characterized by a particular parametric distribution. Within the family of mixture models, mixtures-of-linear-regressions

have also been studied. These arise when there appears to be multiple regression relationships, but no information about membership of the observations is available.

### 1.1.1 Mixtures of Linear Regression Models

Mixtures-of-linear-regressions models were introduced by Quandt and Ramsey (1978) [93] under the name of *switching regressions*. They used a technique based on a moment-generating function to estimate the parameters. Over the next 20 years, estimation of these models was mainly performed from a likelihood point of view. It is well known that mixture likelihoods are multimodal. Thus, the first step in an analysis is to identify as many local modes as possible. The standard approach to this problem is to use multiple random starts for an *Expectation-Maximization* (EM) algorithm. EM algorithm was first explained and given its name in a classic 1977 paper by Dempster, Laird and Rubin [36]. Then in 1989, De Veaux [34] developed the EM algorithm for fitting the two regression setting. Jones and McLachlan (1992) [61] applied mixtures of regressions in a data analysis and used the EM algorithm to fit these models. Turner (2000) [112] fitted a two-component mixture of one variable linear regression to a data set using the EM algorithm. Hawkins et al. (2001) [52] studied the problem of determining the number of components in a mixture of linear regression models using methods derived from the likelihood equation. Zhu and Zhang (2004) [123] established asymptotic theory for maximum likelihood estimators in mixtures-of-regression models; Young and Hunter (2010) [120] and Hunter and Young (2012) [55] developed semi-parametric mixtures-of-regressions models.

Suppose we have $n$ subjects with $m$ measurements, $\mathbf{Y}_i = (Y_{i1}, \cdots, Y_{im})$, on the $i$th subject for all $i = 1, \cdots, n$. Take $\mathbf{y}_1, \cdots, \mathbf{y}_n$ as realized values of the $\mathbf{Y}_i$s, which are independent and identically distributed (*i.i.d.*) according to a distribution $G$. In addition to this scenario, we also assume heterogeneity with respect to the response tendencies of the subjects. One way to account for this is by suggesting $k$ different classes with which the subjects could belong. For a fixed value $k \in \mathbb{N}$, we say the

distribution of $\mathbf{Y}_i$s has $k$-component mixture density

$$g_k(\mathbf{y}_i \mid \mathbf{x}, \boldsymbol{\psi}) = \sum_{j=1}^{k} \lambda_j f_j(\mathbf{y}_i \mid \mathbf{x}, \boldsymbol{\theta}_j) \tag{1.1}$$

$$\lambda_j > 0, \quad \sum_{j=1}^{k} \lambda_j = 1,$$

where $\lambda_j$ is the weight (or *mixing proportion*) for the $j$th component of the model, $\mathbf{y}_i$ is the dependent variable with conditional density $g_k$, $\mathbf{x}$ is a vector of independent (predictor) variables, $\boldsymbol{\theta}_j$ is the component specific parameter vector for the $j$th component density $f_j$, and $\boldsymbol{\psi} = (\lambda_1, \cdots, \lambda_{k-1}, \boldsymbol{\theta}_1, ..., \boldsymbol{\theta}_k)^T$ is the vector of all parameters.

The mixture density $g_k$ is parameterized by $\boldsymbol{\psi} \in \boldsymbol{\Psi}$ such that $\boldsymbol{\Psi}$ represents the specified parameter space for all unknown parameters in the mixture model. Note that

$$\boldsymbol{\Psi} = \left( \prod_{\boldsymbol{j=1}}^{\boldsymbol{k}} \boldsymbol{\Theta_j} \right) \times \Lambda_{k-1},$$

where $\boldsymbol{\Psi} \subset \mathbb{R}^r$ and $r = (\sum_{j=1}^{k} q_j) + (k-1)$ and $q_j$ is the dimension of the parameter in the $j$th component. We take $G$ as the corresponding $k-$component mixture distribution whose components are composed of the distributions $F_j$. For the scenarios presented in this dissertation, the $F_j$ differ only in $\boldsymbol{\theta}_j$, thus we take $f_j \equiv f$ and $q_j = q$, which yield $\boldsymbol{\Psi} = \boldsymbol{\Theta}^k \times \Lambda_{k-1}$ and $r = kq + k - 1$. Furthermore, we only consider the case where a vector of predictors, say $\mathbf{X}_i = (X_{i1}, \cdots, X_{ip})^T$ for $p < n$, is also observed with each response $\mathbf{Y}_i$. The goal is to describe the conditional distribution of $\mathbf{Y}_i \mid \mathbf{X}_i$ through a mixture of regressions. For the remainder of this dissertation, $\mathbf{Y}_i$ will be considered univariate, thus we will replace the boldface $\mathbf{Y}_i$ with $Y_i$.

### 1.1.2 Parameter Estimation

We will focus on estimating the parameters of the mixture model, $\boldsymbol{\psi}$, given observed data $\{(\mathbf{x}_1, y_1), \cdots, (\mathbf{x}_n, y_n)\}$ and the number of components $k$ in this subsection. An alternative method to maximum likelihood and EM, especially in the context of mixture models, is the *method of moments* approach. The method of moments dates back

to the origins of mixture models with Pearson's solution for identifying the parameters of a mixture of two univariate normals (Pearson (1894) [89]). In this approach, model parameters are chosen to specify a distribution whose $p$th order moments, for several values of $p$, are equal to the corresponding empirical moments observed in the data. Latter works of Belkin and Sinha (2010) [10], Kalai et al. (2010) [65], and Moitra and Valiant (2010) [81] can be thought of the modern implementations of the method of moments for mixture models. Unfortunately, this method often runs into trouble with large mixtures of high-dimensional distributions. This is because the equations determining the parameters are typically based on moments of order equal to the number of model parameters, and high-order moments are exceedingly difficult to estimate accurately due to their large variance.

Here we use some more efficient algorithms for estimating in the mixture setting. While Bayesian approaches are an active research area in their own right, we focus on likelihood method. We will provide a brief literature review on some of the available likelihood techniques, but only provide a complete description for one algorithm employed in this dissertation — the EM algorithm in the next subsection. We will also discuss some issues concerning estimation of the parameters.

### Likelihood Methods

Suppose the (observed) data consists of $n$ *i.i.d.* observations $(\mathbf{x}_1, y_1), \cdots, (\mathbf{x}_n, y_n)$ from a $k$-component mixture density given by (1.1). The associated *complete data* is denoted by $\mathbf{c} = (\mathbf{c}_1, \cdots, \mathbf{c}_n)$ with density $\mathbf{h}_{\boldsymbol{\psi}}(\mathbf{c}) = \prod_{i=1}^{n} h_{\boldsymbol{\psi}}(\mathbf{c}_i)$. In the model for complete data associated with model (1.1), each random vector $\mathbf{C}_i = (\mathbf{X}_i, \mathbf{Z}_i)$ where $\mathbf{Z}_i = (\mathcal{Z}_{ij}, j = 1, \cdots, k)$ and $\mathcal{Z}_{ij} \in \{0, 1\}$ is a Bernoulli random variable indicating that individual $i$ comes from component $j$. Since each individual comes from exactly one component, this implies $\sum_{j=1}^{k} \mathcal{Z}_{ij} = 1$, and

$$P(\mathcal{Z}_{ij} = 1) = \lambda_j, \quad (\mathbf{X}_i \mid \mathcal{Z}_{ij} = 1) \sim f_j, \quad j = 1, \cdots, k.$$

The complete data likelihood function for the parameters of a mixture model can

be written as

$$L_c(\boldsymbol{\psi} \mid \mathbf{y}) = \prod_{i=1}^{n} h_{\boldsymbol{\psi}}(\mathbf{c}_i \mid \boldsymbol{\psi}) = \prod_{i=1}^{n} \left[ \sum_{j=1}^{k} \mathbb{I}_{z_{ij}} \lambda_j f_j(y_i \mid \mathbf{x}_i, \boldsymbol{\theta}_j) \right]. \tag{1.2}$$

It is easy to check that the *maximum likelihood estimator* (MLE) $\hat{\boldsymbol{\psi}}_c$ can be achieved by maximizing $L(\boldsymbol{\psi})$. In dealing with likelihood methods, it is often easier to work with the log-likelihood

$$\ell_c(\boldsymbol{\psi}) = \log L_c(\boldsymbol{\psi} \mid \mathbf{y}) = \sum_{i=1}^{n} \log h_{\boldsymbol{\psi}}(\mathbf{c}_i \mid \boldsymbol{\psi}). \tag{1.3}$$

Then, an estimate $\hat{\boldsymbol{\psi}}_c$ (the MLE) of the complete data is provided by solving

$$S(\mathbf{y} \mid \boldsymbol{\psi}) = \frac{\partial \ell_c(\boldsymbol{\psi})}{\partial \boldsymbol{\psi}} = \mathbf{0}, \tag{1.4}$$

where $S(\mathbf{y} \mid \boldsymbol{\psi})$ is called the *score function*.

The corresponding *incomplete data* (observed data) log-likelihood is

$$\begin{aligned} \ell(\boldsymbol{\psi}) &= \sum_{i=1}^{n} \log g_k(y_i \mid \mathbf{x}_i, \boldsymbol{\psi}) \\ &= \sum_{i=1}^{n} \log \sum_{j=1}^{k} \lambda_j f_j(y_i \mid \mathbf{x}_i, \boldsymbol{\theta}_j). \end{aligned}$$

Note that this likelihood function has multiple modes. In fact, if $\mathbf{X}$ is one-dimensional, it has $k$ modes, but if $d > 1$, it can have more than $k$ modes (Carreira-Perpiñán and Williams (2003) [21]). Hence finding the global maximum will be difficult. One can use gradient based methods to find the MLE estimate. Taking the derivative with respect to the parameter of one component, say $\boldsymbol{\theta}_j$, and setting it equal to 0 yields:

$$\begin{aligned} \frac{\partial \ell}{\partial \boldsymbol{\theta}_j} &= \sum_{i=1}^{n} \frac{\partial}{\partial \boldsymbol{\theta}_j} \log g_k(y_i \mid \mathbf{x}_i, \boldsymbol{\psi}) \\ &= \sum_{i=1}^{n} \frac{\partial}{\partial \boldsymbol{\theta}_j} \log \sum_{j=1}^{k} \lambda_j f_j(y_i \mid \mathbf{x}_i, \boldsymbol{\theta}_j) \\ &= \sum_{i=1}^{n} \frac{1}{\sum_{j=1}^{k} \lambda_j f_j(y_i \mid \mathbf{x}_i, \boldsymbol{\theta}_j)} \lambda_j \frac{\partial}{\partial \boldsymbol{\theta}_j} f_j(y_i \mid \mathbf{x}_i, \boldsymbol{\theta}_j) \\ &= \sum_{i=1}^{n} \frac{\lambda_j f_j(y_i \mid \mathbf{x}_i, \boldsymbol{\theta}_j)}{\sum_{j=1}^{k} \lambda_j f_j(y_i \mid \mathbf{x}_i, \boldsymbol{\theta}_j)} \frac{1}{f_j(y_i \mid \mathbf{x}_i, \boldsymbol{\theta}_j)} \frac{\partial}{\partial \boldsymbol{\theta}_j} f_j(y_i \mid \mathbf{x}_i, \boldsymbol{\theta}_j) \\ &= \sum_{i=1}^{n} \frac{\lambda_j f_j(y_i \mid \mathbf{x}_i, \boldsymbol{\theta}_j)}{\sum_{j=1}^{k} \lambda_j f_j(y_i \mid \mathbf{x}_i, \boldsymbol{\theta}_j)} \frac{\partial}{\partial \boldsymbol{\theta}_j} \log f_j(y_i \mid \mathbf{x}_i, \boldsymbol{\theta}_j) = 0. \end{aligned}$$

If we just have a non-mixture parametric model, on the other hand, the derivative of the log-likelihood would be $\sum_{i=1}^{n} \frac{\partial}{\partial \boldsymbol{\theta}_j} \log f_j(y_i \mid \mathbf{x}_i, \boldsymbol{\theta}_j)$. So maximizing the likelihood for a mixture model is like doing a weighted likelihood maximization, where the weight of $\mathbf{x}_i$ depends on the following component membership probability:

$$p_{ij} = \frac{\lambda_j f_j(y_i \mid \mathbf{x}_i, \boldsymbol{\theta}_j)}{\sum_{j=1}^{k} \lambda_j f_j(y_i \mid \mathbf{x}_i, \boldsymbol{\theta}_j)}. \tag{1.5}$$

However, the likelihood equation will have multiple roots and, thus, result in local maxima. Moreover, the likelihood function may be unbounded, which becomes a considerable concern when implementing various algorithms. Focusing on local maxima on the interior of the parameter space helps circumvent this problem because under certain regularity conditions, there exists a strongly consistent sequence of roots to the likelihood equation that is asymptotically efficient (Ferguson (1996) [43]). In fact, a $\sqrt{n}$-consistent estimator can be constructed using the method of moments estimator mentioned earlier.

### Newton Method

An efficient way for solving (1.4) is to implement a Newton-type method. The Newton-Raphson method takes a linear Taylor series expansion about the current fit $\boldsymbol{\psi}^{(t)}$ for $\boldsymbol{\psi}$, which yields

$$S(\mathbf{y} \mid \boldsymbol{\psi}) \approx S(\mathbf{y} \mid \boldsymbol{\psi}^{(t)}) - \mathrm{I}(\boldsymbol{\psi} \mid \mathbf{y})(\boldsymbol{\psi} - \boldsymbol{\psi}^{(t)}), \tag{1.6}$$

where

$$\mathrm{I}(\boldsymbol{\psi} \mid \mathbf{y}) = -\frac{\partial S(\mathbf{y} \mid \boldsymbol{\psi})}{\partial \boldsymbol{\psi}^T} \tag{1.7}$$

is the negative of the Hessian of $\ell(\boldsymbol{\psi})$. Then, finding a zero for the right hand side of (1.6) yields the update

$$\boldsymbol{\psi}^{(t+1)} = \boldsymbol{\psi}^{(t)} + \left[\mathrm{I}(\boldsymbol{\psi}^{(t)} \mid \mathbf{y})\right]^{-1} S\left(\mathbf{y} \mid \boldsymbol{\psi}^{(t)}\right). \tag{1.8}$$

The Newton-Raphson method has the benefit of local quadratic convergence to a solution $\boldsymbol{\psi}^*$ of (1.4), but this convergence is not guaranteed. Aside from some other computational issues (McLachlan and Krishnan (1997) [79]), Newton-Raphson has

the benefit of providing, as an estimate of the variance-covariance matrix of the solution, the inverse of the observed information matrix, $[\mathrm{I}(\boldsymbol{\psi}^* \mid \mathbf{y})]^{-1}$. Thus, standard error estimates, confidence intervals, and inference procedures are readily available.

### 1.1.3 EM Algorithms

Newton-Raphson methods can provide relatively speedy convergence, but this convergence is not guaranteed and calculations like inverting the Hessian may be rather difficult to perform. As an alternative, EM algorithms are often preferred for finding MLEs of mixture models because of their simplicity. EM algorithms are commonly employed in the mixture modeling literature: Bailey and Elkan (1994) [8] fitted a mixture model by EM algorithm to discover motifs in bi-polymers; Ghahramani and Hinton (1996) [48] presented an exact EM algorithm for fitting the parameters of mixture of factor analyzers; Muthén and Shedden (1999) [83] discussed the estimation of parameters for an extended finite mixture model where the latent classes corresponding to the mixture components for one set of observed variables influence a second set of observed variables using EM algorithm; EM algorithms are also the primary method of estimation in the R package mixtools (Benaglia et al. (2009) [11]) In this dissertation, we focus on developing an EM algorithm for the mixture models, but it should be noted that this algorithm is one member in a much larger class of algorithms (McLachlan and Krishnan (1997) [80]).

The key insight behind EM is: if we knew the values of the $\mathcal{Z}_{ij}$s , then optimizing the (complete data) likelihood with respect to $\boldsymbol{\psi}$ would be easy. We could simply estimate $\lambda_j$ and $\boldsymbol{\theta}_j$ by applying the standard closed-form formula to all the data assigned to component $j$. Since we don't know the values of the $\mathcal{Z}_{ij}$s, we need to estimate them first and use them as substitutes for the true values. More precisely, we will optimize the expected complete data log likelihood instead of the actual complete data log likelihood. Since the estimates of the $\mathcal{Z}_{ij}$s depend on the parameters $\boldsymbol{\psi}$, we need to re-estimate them after each update to $\boldsymbol{\psi}$. This algorithm can be shown to monotonically increase a lower bound on the log likelihood, and hence it will converge. In more details, we can now construct an EM algorithm for mixtures-of-

linear-regressions models in Algorithm 1.1.

**Algorithm 1.1** (EM Algorithm)

(a) (*Expectation Step* (E-Step)) Given a fixed $\boldsymbol{\psi}^{(t)}$ at the $t$th iteration, $t = 0, 1, \cdots$, calculate

$$Q(\boldsymbol{\psi} \mid \boldsymbol{\psi}^{(t)}) := \mathbb{E}_{\boldsymbol{\psi}^{(t)}} \left[ \ell_c(\boldsymbol{\psi}) \mid \mathbf{C} = \mathbf{c}, \mathbf{Y} = \mathbf{y} \right].$$

(b) (*Maximization Step* (M-Step)) Find

$$\boldsymbol{\psi}^{(t+1)} = \arg\max_{\boldsymbol{\psi}} Q(\boldsymbol{\psi} \mid \boldsymbol{\psi}^{(t)}),$$

which implies

$$Q(\boldsymbol{\psi}^{(t+1)} \mid \boldsymbol{\psi}^{(t)}) \geq Q(\boldsymbol{\psi} \mid \boldsymbol{\psi}^{(t)})$$

for all $\boldsymbol{\psi} \in \boldsymbol{\Psi}$.

(c) Iterate until a *stopping criterion* is attained. The final estimate obtained will be denoted by $\hat{\boldsymbol{\psi}}$.

Because EM will only find a local minimum, good initialization is important. But how to do this is problem dependent. A general strategy is to try multiple restarts at random locations or to use a clustering algorithm.

Notice $\mathcal{Z}_{ij} \sim \text{Bern}(\lambda_j)$, where $\text{Bern}(\lambda_j)$ is the Bernoulli distribution with rate of success $\lambda_j$. Since $\mathbb{E}_{\boldsymbol{\psi}^{(t)}}$ is a linear functional, the expectation of $\mathcal{Z}_{ij}$ is the weight of observation $i$ belonging to the $j$th component,

$$\mathbb{E}_{\boldsymbol{\psi}}[\mathcal{Z}_{ij} \mid \mathbf{X} = \mathbf{x}, Y = y] = p_{ij}^{(t+1)} = \frac{\lambda_j^{(t)} f_j(y_i \mid \mathbf{x}_i, \boldsymbol{\theta}_j^{(t)})}{\sum_{j=1}^{k} \lambda_j^{(t)} f_j(y_i \mid \mathbf{x}_i, \boldsymbol{\theta}_j^{(t)})}.$$

The M step involves maximizing the expected complete data log likelihood $Q(\boldsymbol{\psi} \mid \boldsymbol{\psi}^{(t)})$:

$$\begin{aligned}
Q(\boldsymbol{\psi} \mid \boldsymbol{\psi}^{(t)}) &= \mathbb{E}_{\boldsymbol{\psi}^{(t)}} \left[ \ell_c(\boldsymbol{\psi}) \mid \mathbf{C} = \mathbf{c}, \mathbf{Y} = \mathbf{y} \right] \\
&= \mathbb{E}_{\boldsymbol{\psi}^{(t)}} \left[ \sum_{i=1}^{n} \log h_{\boldsymbol{\psi}}(\mathbf{C}_i \mid \boldsymbol{\psi}^{(t)}) \right] \\
&= \mathbb{E}_{\boldsymbol{\psi}^{(t)}} \left[ \sum_{i=1}^{n} \sum_{j=1}^{k} \log \left( \lambda_j f_j(y_i \mid \mathbf{x}_i, \boldsymbol{\theta}_j^{(t)}) \right) \mathbb{I}_{z_{ij}} \right] \\
&= \sum_{i=1}^{n} \sum_{j=1}^{k} \log \left( \lambda_j f_j(y_i \mid \mathbf{x}_i, \boldsymbol{\theta}_j^{(t)}) \right) p_{ij}^{(t+1)}.
\end{aligned}$$

Therefore, we can update $\lambda_j^{(t+1)}$ by

$$\lambda_j^{(t+1)} = \frac{1}{n}\sum_{i=1}^{n} p_{ij}^{(t+1)} \tag{1.9}$$

and $\boldsymbol{\theta}_j^{(t+1)}$ is the solution of

$$\sum_{i=1}^{n} p_{ij}^{(t+1)} \frac{\partial}{\partial \boldsymbol{\theta}_j} \log f_j(y_i \mid \mathbf{x}_i, \boldsymbol{\theta}_j^{(t)}) = 0. \tag{1.10}$$

As we can see, the structure for an EM algorithm is rather simple and programming is easy. We will stress some practical issues concerning implementation of Algorithm 1.1.

One issue is the selection of the *initial values* $\boldsymbol{\psi}^{(0)}$. Due to the multi-modality in the mixture likelihood, there are multiple local maxima, and in some cases a poor choice of $\boldsymbol{\psi}^{(0)}$ can lead to the sequence of EM algorithms diverging. Due to such features, it is recommended to start EM algorithm from different initial values. For reviews of possible options for starting values, see McLachlan and Krishman (1997) [79] or McLachlan and Peel (2000) [78].

Another issue concerns the stopping criterion. Usually an EM algorithm is run until

$$\ell(\boldsymbol{\psi}^{(t+1)}) - \ell(\boldsymbol{\psi}^{(t)}) < \epsilon, \tag{1.11}$$

or, when given a norm $\|\cdot\|$ on $\boldsymbol{\Psi}$, until

$$\|\boldsymbol{\psi}^{(t+1)} - \boldsymbol{\psi}^{(t)}\| < \epsilon$$

for some $\epsilon > 0$ chosen arbitrarily small. Schafer (1997) [100] discussed the stopping criterion

$$\frac{|\boldsymbol{\psi}_l^{(t+1)} - \boldsymbol{\psi}_l^{(t)}|}{\boldsymbol{\psi}_l^{(t)}} < \epsilon$$

for $l = 1, 2, \cdots, r$, though this method fails when $\boldsymbol{\psi}_l^{(t)} \approx 0$. Regardless, EM algorithms converge linearly, which can be very slow at times.

An inappropriate stopping criterion may cause one to claim convergence too soon. Certain methods, such as an Aitken-based acceleration technique, may be implemented to alleviate some of the difficulty with the slow rate of convergence (Lindsay (1995) [72]). We use the method in (1.11) as our criterion.

### 1.1.4 Selecting the Number of Components: Model Selection

Determining the number of components in finite mixture models is a very important problem. Chapter 6 of McLachlan and Peel (2000) [78] discusses many common approached. There is also the visualization tool called the *mixturegram*, which was recently introduced by Young et al (2018) [121]. Here, we discuss *information criteria* to assess the number of components for a mixture model.

An information criterion for model selection can be based on the bias-corrected log likelihood given by

$$\log L(\boldsymbol{\psi}) - b(F) \tag{1.12}$$

using an appropriate estimate of the bias term $b(F)$. The intent is to select the model (that is, the number of components in the present context) to maximize (1.12). In the literature, the information criteria are generally expressed in terms of twice the value of the difference, so that they are of the form

$$-2\log L(\boldsymbol{\psi}) + 2C, \tag{1.13}$$

where the first term on the right-hand side of (1.13) measures the lack of fit and the second term $C$ is the penalty term that measures the complexity of the model. The intent therefore is to choose a model that minimizes (1.13).

The four criteria we will compare here are Akaike's Information Criterion (AIC) of Akaike (1973) [5], the Bayesian Information Criterion (BIC) of Schwarz (1978) [102], the Integrated Completed Likelihood (ICL) of Biernacki et al. (2000) [12], and the consistent AIC (cAIC) of Bozdogan (1987) [14]. Given $\hat{\boldsymbol{\psi}}$, the MLE of $\boldsymbol{\psi}$ formed from the observed sample, the form of these criteria are, respectively,

$$\text{AIC} = -2\log L(\hat{\boldsymbol{\psi}}) + 2d \tag{1.14}$$

$$\text{BIC} = -2\log L(\hat{\boldsymbol{\psi}}) + d\log(n) \tag{1.15}$$

$$\text{ICL} = \text{BIC} + 2\left(-\sum_{i=1}^{n}\sum_{j=1}^{k}\hat{p}_{ij}\log\hat{p}_{ij}\right) \tag{1.16}$$

$$\text{cAIC} = -2\log L(\hat{\boldsymbol{\psi}}) + d(\log(n) + 1), \tag{1.17}$$

where $n$ is the number of observations, $d$ is the number of parameters in the mixture setting, and the $\hat{p}_{ij}$s are final posterior membership probabilities from an EM algorithm. These values are calculated for a reasonable range of components and then the minimum of these values (for each criterion) corresponds to the number of components selected by that criterion.

The four information criteria we employ in this dissertation are by no means an exhaustive collection of such information criteria. Indeed, they are some of the more common information criteria employed in the mixture literature. Beyond these, there are contemporary methods like the BIC for singular models that was introduced by Drton and Plummer (2017) [38]. This information criterion (which the authors termed sBIC) preserves the consistency properties of BIC, but is also demonstrated to show improved (frequentist) model selection properties.

### 1.1.5 Identifiability

In this subsection, we define identifiability for mixture distributions. This discussion and the definition of identifiability are adopted from McLachlan and Peel (2000) [78].

Let $\mathcal{F}_k$ denote a parametric family of $k$-component mixture densities as described in (1.1) and $\mathcal{F}$ the class of all such $\mathcal{F}_k$. So

$$\mathcal{F}_k = \{g_k(\mathbf{y} \mid \mathbf{x}, \boldsymbol{\psi}) : \boldsymbol{\psi} \in \boldsymbol{\Psi}\} \text{ and } \mathcal{F} = \bigcup_{k \in \mathbb{N}} \mathcal{F}_k.$$

Permuting the component labels of the mixture density results in $\mathcal{F}$ being nonidentifiable in $\boldsymbol{\Psi}$, where identifiability is defined as follows:

**Definition 1.1. (Identifiability)**

Consider

$$g_k(\mathbf{y}|\mathbf{x}, \boldsymbol{\psi}) = \sum_{j=1}^{k} \lambda_j f_j(\mathbf{y}|\mathbf{x}, \boldsymbol{\theta}_j)$$

and

$$g_{k^*}(\mathbf{y}|\mathbf{x}, \boldsymbol{\psi}^*) = \sum_{j=1}^{k^*} \lambda_j^* f_j(\mathbf{y}|\mathbf{x}, \boldsymbol{\theta}_j^*),$$

which are both members of the class $\mathcal{F}$.

$\mathcal{F}$ is said to be identifiable for $\boldsymbol{\psi} \in \boldsymbol{\Psi}$ if $g_k(\mathbf{y} \mid \mathbf{x}, \boldsymbol{\psi}) = g_{k^*}(\mathbf{y} \mid \mathbf{x}, \boldsymbol{\psi}^*)$ a.e. if and only if:

1. $k = k^*$;
2. under permutation of the component labels, $\lambda_j = \lambda_j^*$ and $f_j(\mathbf{y} \mid \mathbf{x}, \boldsymbol{\theta}_j) = f_j(\mathbf{y} \mid \mathbf{x}, \boldsymbol{\theta}_j^*)$ a.e. for all $j = 1, \cdots, k$;
3. $\lambda_j > 0$ and the $\boldsymbol{\theta}_j$ are distinct for all $j$.

Definition 1.1 states that no element of $\mathcal{F}$ can arise in two different ways except by trivial means, such as letting some $\lambda_j = 0$ or splitting a component by letting $\boldsymbol{\theta}_{j1} = \boldsymbol{\theta}_{j2}$.

### Label Switching

During the implementation of iterative methods in mixture modeling, such as the *parametric bootstrap* to obtain standard error estimates, we need to be cognizant of the solutions being calculated from one iteration to the next. This is because a given mixture component can't be extracted from the likelihood. This situation occurs because the component labels can't be distinguished from one another due to the nonidentifiability in $\boldsymbol{\psi}$ as established in Definition 1.1. Such a permutation of the component labels as in this definition is called *label switching*.

There are numerous methods in the literature for dealing with label switching (see Jasra et al. (2005) [60] for a review of some of these techniques). One of the easiest methods by Aitkin and Rubin (1985) [4] for dealing with this problem is by imposing artificial identifiability constraints on one of the parameters (such as $\lambda_1 \leq \lambda_2 \leq \cdots \leq \lambda_k$). Kim and Lindsay (2015) [67] utilized the notion of local identifiability, which guarantees the existence of the identifiable parameter region, to develop an empirical measure of the degree of local identifiability on the estimated parameters. However, it is not always possible to find such constraints and these choices of constraints depends heavily on the parameters (for instance, see McLachlan and Peel (2000) [78]

and Stephens (2000b) [108]). For example, consider fitting a mixture with $k = 2$ components with the mixing proportions close to 0.50. Imposing the identifiability constraint on $\lambda_i$s clearly influences the estimates of $\boldsymbol{\theta}_1$ and $\boldsymbol{\theta}_2$, thus creating a bias. Such a situation is highlighted in Celeux et al. (2000) [25] where they presented "disturbing" results when considering the various ordering constraints on a $k = 3$ component mixture of normals using a *Markov chain Monte Carlo* (MCMC) sampler. This identifiability can be imposed after the simulations have been completed, as Stephens (2000b) [108] demonstrated for a MCMC sample of size $N$ by relabeling the sample $(\boldsymbol{\Psi}^{(1)}, \boldsymbol{\Psi}^{(2)}, \cdots, \boldsymbol{\Psi}^{(N)})$ and applying permutations $\pi_1, \pi_2, \cdots, \pi_N$ such that the permuted sample $\left(\pi_1(\boldsymbol{\Psi}^{(1)}), \pi_2(\boldsymbol{\Psi}^{(2)}), \cdots, \pi_N(\boldsymbol{\Psi}^{(N)})\right)$ satisfies the identifiability constraints.

Strategies to handle the label switching problem have also been proposed in the Bayesian context. For example, Stephens (2000b) [108] proposed a class of relabeling algorithms that attempt to minimize the posterior loss under a class of loss functions. Chung et al. (2004) [28] proposed assigning as few as one observation to a component a priori, which effectively amounts to using data-dependent 3 priors where one or more observations are assigned to each component with certainty. Their strategy applies enough information to break the symmetry of the likelihood and flatten the posterior density over $k! - 1$ nuisance regions, which are the duplicate modes resulting from the permutations of the component labels.

Since there is not always a clear choice of labeling, Richardson and Green (1997) [96] stress post-processing the simulations under different permutations of the labels to determine an appropriate choice.

When the parameters are well-separated within the parameter space, identifiability constraints can be a very simple *post-hoc* method. Since this is the scenario for the examples we will present, identifiability constraints will be the method of choice for dealing with label switching in this dissertation. For the sake of completeness, we also discuss alternative methods for handling this issue.

First, consider bootstrapping in mixtures. McLachlan and Peel (2000) [78] point out that label switching can usually be avoided by setting the EM algorithm's starting

values to the maximum likelihood estimates, since EM algorithms are (generally) very dependent on the starting values.

Next, note that since the likelihood of a $k$-component mixture model is invariant under permutation of the component labels, it effectively has $k!$ modes. Label switching is often presented in the context of Bayesian mixture modeling since the posterior distribution will also have this property under a symmetric prior. The Bayesian method often involve a decision theoretic approach as implemented in Celeux et al. (2000) [25], Stephens (2000b) [108], Hurn et al. (2003) [56], and Jasra et al. (2005) [60].

Another procedure used within the Bayesian framework is by Chung et al. (2004) [28], who suggest assigning as few as one observation to a component a *priori*. This amounts to using data-dependent priors where one or more observations are assigned to each component with certainty. The point is to apply enough information to break the symmetry of the likelihood and flatten the posterior density over $k! - 1$ nuisance regions, which are the duplicate modes resulting from the permutations of the components. The posterior density in the sampler will now reflect a modified likelihood function which accommodates a density where one (or more) observations were assigned to each component. The major limitation of this approach is to what extent one is willing to accept preclassifying certain observations.

## 1.2 Measurement Error Models for Regression

Measurement error models are commonly used in making inference on the relationship of a response variable $Y$ and predictor variables when some of the variables may be measured with error. Fuller (1987) [1] and Cheng and Ness (1999) [27] discussed methods account for measurement errors primarily aimed at linear models. Reviews that center on the econometric literature are also available. Wansbeek & Meijer (2007) [113] focus primarily on linear models and make direct connections with latent variables and factor models. A broad review of nonlinear measurement error models with an emphasis on the use of auxiliary samples containing error-free covariates is provided by Chen et al. (2011) [26]. Schennach (2013) [101] reviewed

recent econometrics literature on measurement error in nonlinear models, especially regarding latent variables, factor models, and non separable error and providing more insight into the connection among different approaches.

In a simple linear bivariate regression, the presence of measurement error "attenuates" the relationship between the dependent variable and the mismeasured regressors. This means, if one neglects the presence of measurement error in the regression, the regression coefficients merely become less significantly different from zero, so that the resulting statistical inference is conservative, but otherwise valid.

However, this optimistic result fails to hold in general for multivariate linear regressions and for nonlinear specifications (Hausman (2001) [2], Hausman, Newey and Powell (1995) [51], Griliches and Ringstad (1970) [49]). To make matters worse, the standard instrumental variable approach, which is entirely adequate to correct for the endogeneity caused by measurement error in linear models, fails in nonlinear models (Amemiya (1985) [6]). These realizations have motivated the large and growing literature that aims to correct for the presence of measurement error in nonlinear models.

### 1.2.1 Classical Measurement Errors and Berkson Errors

Modeling the error caused by the measuring process has been long studied. The inspiration of measurement error model can date back to Pearson (1894) [89]. According to the introduction of measurement error models given in Carroll et al. (2006) [94], there are two types of models for the measurement error process:

- Error models, such as classical measurement error models. These models consider the conditional distribution of the observed variables measured with error given true variables.

- Regression calibration models, such as Berkson error models. These models consider the conditional distribution of the true variables given the observed variables measured with error.

There are two main consequences if the methods of estimation for the case without measurement error are misused when the measurement error is not negligible. One consequence is that the estimator will become biased. The other consequence is the loss of power in hypothesis testing. In this dissertation, we focus on the estimators for classical measurement errors in mixtures-of-regressions models such that these effects can be minimized.

### Classical Measurement Error Model

**Definition 1.2.** Consider the regression model of a response $Y$ on a $r$-dimensional predictor $\mathbf{X}$, when the predictor variable $\mathbf{X}$ or part of the $\mathbf{X}$ cannot be observed directly, but instead the surrogate, denoted by $\mathbf{W}$, of $\mathbf{X}$ is observed. The *classical measurement error* model can be defined as:

$$\mathbf{W} = \mathbf{X} + \mathbf{U}. \tag{1.18}$$

In this model, $\mathbf{W}$ is an unbiased measure of $\mathbf{X}$, so that $\mathbf{U}$ must have mean zero, that is, in symbols, $\mathbb{E}(\mathbf{U} \mid \mathbf{X}) = 0$. The error structure of $\mathbf{U}$ could be homoscedastic (constant variance) or heteroscedastic. Initially, we will consider the case that the measurement error structure is approximately normal with constant variance, so we can reasonably think that $\mathbf{U} \mid \mathbf{X} \sim N_r(0, \mathbf{\Sigma}_u)$, and later we may also discuss the case when the measurement $\mathbf{U}$ is not normally distributed.

### Berkson Measurement Error Model

What we see in the classical measurement error model (1.18) is that the observed predictor variable equals the true predictor variable plus (classical) measurement error. This, of course, means that the variability of the observed predictor variable will be greater than the variability of the true variable. In some situations, we do not only consider the classical measurement error, but also turn the issue around; namely, assume that the true predictor variable is equal to the estimated variable plus measurement error. In symbols, this is

$$\mathbf{X} = \mathbf{W} + \mathbf{U}, \tag{1.19}$$

where $\mathbb{E}(\mathbf{U}|\mathbf{W}) = 0$, so the true predictor variable has more variability than the estimated one; contrast with (1.18). Model (1.19) is called a *Berkson measurement error* model, which was first proposed by Berkson (1950) [63].

The major difference between classical and Berkson measurement error models is the dependence of the error and covariate. In the classical measurement error model, the error $\mathbf{U}$ is independent of the true covariate $\mathbf{X}$ and $\mathbb{E}(\mathbf{U}) = 0$, or no independence assumption but $\mathbb{E}(\mathbf{U} \mid \mathbf{X}) = 0$; while for Berkson measurement error model, the error $\mathbf{U}$ is independent of $\mathbf{W}$ and $\mathbb{E}(\mathbf{U}) = 0$, or no independence assumption but $\mathbb{E}(\mathbf{U} \mid \mathbf{W}) = 0$. Therefore, $\mathrm{Var}(\mathbf{W}) > \mathrm{Var}(\mathbf{X})$ for classical errors and $\mathrm{Var}(\mathbf{X}) > \mathrm{Var}(\mathbf{W})$ for Berkson errors. Nevertheless, the choice between the classical and Berkson measurement error models should depend on the background and interpretation of the data.

Testing for the presence of measurement error is mostly underdeveloped in the literature. One possibility is a nonparametric test developed in Wilhelm (2018) [115]. While focused on additive measurement error, a similar test could likely be developed for multiplicative measurement error structures.

### 1.2.2 Estimation Methods

A linear regression model with measurement error has been studied under the classical measurement error model in Fuller (1987) [1], where bias can be found on the parameter estimation. It has also been applied to epidemiology studies to correct the biased caused by measurement error (Wong et al. (2003) [84]). One straightforward way of estimating parametric models with measurement error is a likelihood-based approach. The likelihood function is constructed based on the specified parametric model and the chosen measurement error model. The estimators are obtained by maximizing the likelihood function through numerical techniques. There is some research considering this idea, e.g., Carroll et al. (1984) [22] for probit regression, Whittemore and Gong (1991) [99] for Poisson regression with misclassification model, Reeves et al. (1998) [64] for continuous and binary response models, and Yao and Song (2015) [119] for mixtures-of-regressions models. The likelihood method requires

stronger distributional assumptions and it can be computationally difficult, but it increases the efficiency of the estimators. Besides, likelihood-based tests and confidence intervals can be obtained with the fully-specified parametric model. However, the identifiability of the parametric model is one of the major concerns for likelihood function methods. If the model can not be identified, we most resort to using a semi-parametric or non-parametric model.

In nonlinear models with measurement error problems, there are extensive literature on various approaches developed by researchers; see the text by Carroll et. al (2006) [94]. One of the most common used approaches is the corrected score estimator. This method is based on the log-likelihood function (alternatively, the score function) of the error-free model, and then "corrected" for the measurement error. This approach has been promoted by Stefanski (1989) [106] and Nakamura (1990) [85], since it does not need to specify the distribution of the covariates $\mathbf{X}$, it is a so-called *functional method*.

Another type of approach is called the *structural method*. It works with the assumption that the distribution of $\mathbf{X}$ is known, possibly except for a finite number of unknown parameters. In this dissertation, we assume that $\mathbf{X}$ follows multivariate normal distribution, that is, $\mathbf{X} \sim N(\boldsymbol{\mu}_x, \Sigma_x)$. The idea is to set up unbiased estimating equations of observed data $(\mathbf{W}, Y)$ with the help of the conditional mean and possibly also the conditional variance of $Y$ given $\mathbf{X}$. We call the estimators originating from the solution to such estimating equations structural estimators, because in the theory of measurement error models, a model with a well-specified distribution of the covariates $\mathbf{X}$ is often called a structural model.

In both functional and structural case, the *simulation extrapolation* (SIMEX) estimator has become very popular. Those estimators are not consistent in general, although they often reduce the bias significantly, also see Carroll et al. (2006) [94].

In this subsection, we describe in details of an important example of the classical measurement error model — the *polynomial model*, where for simplicity the latent variable $X$ is scalar, and discuss some methods of consistent estimation in this model, particularly.

The polynomial model is given by

$$Y_i = \mathbf{X}_i^T \boldsymbol{\beta} + \epsilon_i = \eta_i + \epsilon_i,$$
$$W_i = X_i + U_i,$$

with $\mathbf{X}_i^T = \left(1, X_i, X_i^2, \cdots, X_i^k\right)$ and $\boldsymbol{\beta}^T = (\beta_0, \beta_1, \beta_2, \cdots, \beta_k)$ ,where $\epsilon_i \sim N(0, \sigma_\varepsilon^2)$ and $\epsilon_i$ is independent of $X_i$ for $i = 1, 2, \cdots, n$. This can be considered as measurement error model case with classical measurement error. The model requires we have some knowledge about the measurement error, so there are two possible situations: (a) the measurement error variance $\sigma_u^2$ is known, and (b) the ratio $\sigma_\epsilon^2/\sigma_u^2$ is known (see Shklyar (2008) [103]).

**Functional Method: Corrected Score**

If the variable $X$ were observable, we can estimate the unknown parameter $\boldsymbol{\beta}$ by the method of maximum likelihood. The corresponding likelihood-score function for $\boldsymbol{\beta}$ is given by

$$\psi(\boldsymbol{\beta} \mid y_i, x_i) = \frac{\partial \log f(y_i \mid x_i, \boldsymbol{\beta})}{\partial \boldsymbol{\beta}}.$$

We want to construct an unbiased estimating function for $\boldsymbol{\beta}$ in the observed variables. For this purpose, we need to find functions, say $\psi_C$ of $w_i$s and $\boldsymbol{\beta}$ such that

$$\mathbb{E}\left[\psi_{\mathrm{C}}(\boldsymbol{\beta} \mid y_i, w_i) \mid x_i\right] = \psi(\boldsymbol{\beta} \mid y_i, x_i).$$

Then $\psi_{\mathrm{C}}(\boldsymbol{\beta} \mid y_i, w_i)$ is called the corrected score function. The *corrected score* (CS) estimator $\hat{\boldsymbol{\beta}}_{\mathrm{C}}$ of $\boldsymbol{\beta}$ is the solution to

$$\sum_{i=1}^{n} \psi_{\mathrm{C}}(\boldsymbol{\beta} \mid y_i, w_i) = 0.$$

The corrected score function does not always exist. Stefanski (1989) [106] gives the conditions for their existence and shows how to find them if they exist. An alternative functional method, particularly adapted to generalized linear models, is the Conditional Score method, see Stefanski and Carroll (1987) [107].

**Structural Methods: Quasi-Likelihood and Maximum Likelihood**

The conditional mean and conditional variance of $y_i$ given $x_i$ are, respectively,

$$\mathbb{E}(y_i \mid x_i) = \mathbf{x}_i^T \boldsymbol{\beta} \equiv m^*(x_i \mid \boldsymbol{\beta}),$$
$$\mathrm{Var}(y_i \mid x_i) = \sigma_\epsilon^2 \equiv v^*(x_i).$$

Then the conditional mean and conditional variance of $y_i$ given the observable variables $w_i$ are

$$\mathbb{E}(y_i \mid w_i) = \mathbb{E}[m^*(x_i) \mid w_i] \equiv m(w_i \mid \boldsymbol{\beta}),$$
$$\mathrm{Var}(y_i \mid w_i) = \mathrm{Var}[m^*(x_i) \mid w_i] + \mathbb{E}[v^*(x_i) \mid w_i] \equiv v(w_i \mid \boldsymbol{\beta}).$$

For the *quasi-likelihood* (QL) estimator, we construct the quasi-score function

$$\psi_Q(\boldsymbol{\beta} \mid y_i, w_i) = [y - m(w_i \mid \boldsymbol{\beta})]\, v^{-1}(w_i \mid \boldsymbol{\beta})\, \frac{\partial m(w_i \mid \boldsymbol{\beta})}{\boldsymbol{\beta}}.$$

Here we drop the parameter $\sigma_\epsilon^2$ considering it to be known. Indeed, in order to compute $m$ and $v$, we need the conditional distribution of $x$ given $w$, which depends on the distribution of $x$ with its parameter. The *quasi-likelihood* (QL) estimator $\hat{\boldsymbol{\beta}}_Q$ of $\boldsymbol{\beta}$ is the solution to

$$\sum_{i=1}^{n} \psi_Q(\boldsymbol{\beta} \mid y_i, w_i) = 0.$$

The equation has a unique solution for large $n$, but it may have multiple roots if $n$ is not large. Heyde and Morton (1998) [54] develop methods to deal with this case.

Maximum likelihood is based on the joint density of $w, y$, thus while QL relies only on the error-free mean and variance functions, ML relies on the whole error-free model distribution. Therefore, ML is more sensitive than QL with respect to a potential model misspecification because QL is always consistent as long as the density of $x$ has been correctly specified. In addition, the likelihood function is generally much more difficult to compute than the quasi score function. This often justifies the use of the relatively less efficient QL instead of the more efficient ML method.

### 1.2.3 Simulation Extrapolation

SIMEX is a simulation-based method of estimating and reducing bias due to measurement error. SIMEX estimates are obtained by adding additional measurement error to the data in a resampling-like stage, establishing a trend of measurement error-induced bias versus the variance case of the added measurement error. It was first proposed by Cook and Stefanski (1994) [31] and further developed by Stefanski and Cook (1995) [105], Carroll and Stefanski (1995) [23] and Devanarayan and Stefanski (2002) [37]. It is a self-contained simulation study resulting in graphical displays that illustrate the effect of measurement error on parameter estimates and the need for bias correction.

#### SIMEX in Simple Linear Regression

We now describe the basic idea of SIMEX in the context of simple linear regression

$$Y = \beta_0 + \beta_x X + \epsilon$$

with classical measurement error

$$W = X + U,$$

where $U$ is independent of $(Y, X)$ and has mean zero and variance $\sigma_u^2$. The key idea of SIMEX is the fact that the effect of measurement error on an estimator can be determined experimentally via simulation.

First we get the ordinary least squares estimate of $\beta_x$ from the original data, denoted $\hat{\beta}_{x,\text{naive}}$, then we generate $M-1$ data sets, each with successively larger measurement error variances, say $(1+\zeta_m)\sigma_u^2$, where $0 = \zeta_1 < \zeta_2 < \cdots < \zeta_M$ are known. We can also get the least squares estimate of slope from the $m$th data set, called $\hat{\beta}_{x,m}$.

We can now formulate this setup as a nonlinear regression model, with data $\Big\{(\zeta_m, \hat{\beta}_{x,m}), m = 1, \cdots, M\Big\}$, where $\hat{\beta}_{x,m}$ is the response variable and $\zeta_m$ the predictor variable. Notice the mean function of this regression has the form

$$\mathbb{E}\left(\hat{\beta}_{x,m} \mid \zeta\right) = \mathcal{G}(\zeta) = \frac{\beta_x \sigma_x^2}{\sigma_x^2 + (1+\zeta)\sigma_u^2}, \quad \zeta \geq 0.$$

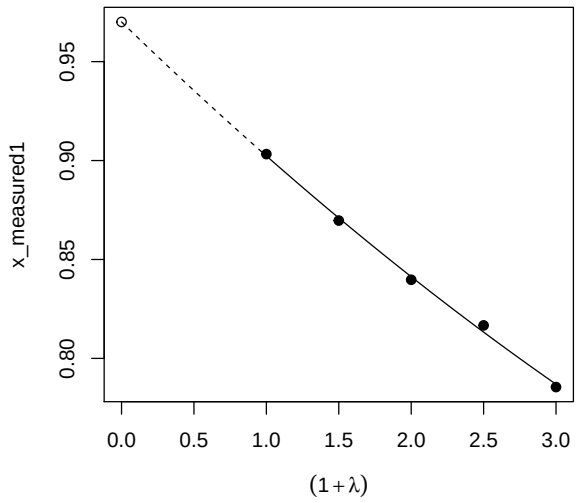


Figure 1.1: A SIMEX plot, where the x-axis is $\zeta$, and y-axis is the estimated coefficient. The SIMEX estimate is an extrapolation to $\zeta = -1$.The naïve estimate occurs at $\zeta = 0$.

Note that $\mathcal{G}(-1) = \beta_x$, that is the parameter of interest is obtained from $\mathcal{G}(\zeta)$ by extrapolation to $\zeta = -1$.

The steps of SIMEX can be written as follows:

**Algorithm 1.2** Simulation Extrapolation

- In the *simulation step*, additional independent measurement errors with variance $\zeta_m \sigma_u^2$ are generated and added to the original $\mathbf{W}$ data, creating data sets with successively larger measurement error variances. For the $m$th data set, the total measurement error variance is $(1 + \zeta_m)\, \sigma_u^2$.
- In the *estimation step*, estimates are obtained from each of the generated data sets.
- The simulation and estimation steps are repeated a large number of times, and the average value of the estimate for each group of data sets is estimated. These values are plotted against the $\zeta$ values and a regression technique is used to fit an extrapolant function to the averaged, error-contaminated estimates.
- Extrapolation to the ideal case of no measurement error ($\zeta = -1$) yields the SIMEX estimate.

### 1.2.4 Measurement Error in the Response and a WLS Estimate

Akritas and Bershady (1996) [46] discussed the problem of fitting regression models with data having heteroscedastic measurement errors of known standard deviation in the response. They defined a statistical model for data with astronomical (heteroscedastic) measurement errors which allows the possibility of correlated errors between both variables of interest, and the possibility that the size of the measurement error depends on the observation, and proposed a *weighted least squares* (WLS) estimator for estimating the model. This measurement error in the response model is practically useful for addressing other problems with such data, including intrinsic variance function estimation, goodness-of-fit, comparing $k$ multivariate samples.

Consider the linear model with $n$ pairs of observations, where the $i$th variables of interest $(\mathbf{X}_i, Y_i)$ follows

$$Y_i = \mathbf{X}_i^T \boldsymbol{\beta} + \epsilon_i \tag{1.20}$$

where $\epsilon_i \sim N(0, \sigma_i^2)$. We denote the observed data by $(\mathbf{X}_i, Y_i^*, \eta_i^2)$, where $\eta_i^2$ is the measurement error (for the response) provided by the researcher. Here, we don't assume measurement error in the predictor. The observed response is related to the unobserved response by

$$Y_i^* = Y_i + \delta_i \tag{1.21}$$

such that $\delta_i \sim N(0, \eta_i^2)$ is independent of $\epsilon_i$.

The method of WLS estimator applies when only the response variable is subject to measurement error and the size of the measurement error does not depend on the observation. The general idea of WLS is to weight the observations so that observations with a larger weight contribute more to the least squares fit. The regression parameter estimator with minimal variance is achieved by assigning weights inversely proportional to the variance of each term.

Relations (1.20) and (1.21) imply

$$\begin{aligned} Y_i^* &= Y_i + \delta_i \\ &= \mathbf{X}_i^T \boldsymbol{\beta} + \epsilon_i + \delta_i \\ &= \mathbf{X}_i^T \boldsymbol{\beta} + \epsilon_i^*, \end{aligned}$$

where $\epsilon_i^*$ has the set $\epsilon_i^* = \epsilon_i + \delta_i$. This is a valid setting for the application of WLS, provided that the variance of $\epsilon_i^*$ is independent of $Y_i^*$. To do so, however, we need to estimate the variance of $\epsilon_i^*$. Notice that

$$\mathrm{Var}(\epsilon_i^*) = \mathrm{Var}(\epsilon_i) + \eta_i^2. \tag{1.22}$$

Since $\mathrm{Var}(\epsilon_i)$ is unknown, $\mathrm{Var}(\epsilon_i^*)$ in also unknown. Using the results from an ordinary least square (OLS) estimator, Akritas and Bershady (1996) [46] extended it to WLS and estimate $\mathrm{Var}(\epsilon_1), \cdots, \mathrm{Var}(\epsilon_n)$.

---

**Algorithm 1.3** (WLS Algorithm)

---

(a) Obtain the regression coefficient estimator $\hat{\boldsymbol{\beta}}$ by a direct application of OLS to the observed data $(\mathbf{X}_1, Y_1^*), \cdots, (\mathbf{X}_n, Y_n^*)$.

(b) Calculate the residuals

$$R_i = Y_i^* - \mathbf{X}_i^T \hat{\boldsymbol{\beta}}$$

for $i = 1, \cdots, n$.

(c) Obtain the estimators of $Var(\epsilon_i)$ from

$$\hat{\mathrm{Var}}(\epsilon_i) = \sum_{i=1}^{n} (R_i - \bar{R})^2 - n^{-1} \sum_{i=1}^{n} \eta_i^2$$

where $\bar{R} = n^{-1} \sum_{i=1}^{n} R_i$.

---

Next, set $\hat{\mathrm{Var}}(\epsilon_i^*) = \hat{\sigma}_i^{*2} = \hat{\mathrm{Var}}(\epsilon_i) + \eta_i^2$ and let $\boldsymbol{A}$ be the $n \times n$ matrix with diagonal elements $\hat{\sigma}_i^{*2}$ and with all off-diagonal elements equal to zero. In terms of $\boldsymbol{A}$, a general formula for the WLS estimator is:

$$\hat{\boldsymbol{\beta}}_{WLS} = (\mathbf{X}^T \boldsymbol{A}^{-1} \mathbf{X})^{-1} \mathbf{X}^T \boldsymbol{A}^{-1} \mathbf{Y}^* \tag{1.23}$$

where $\mathbf{Y}^* = (Y_1^*, \cdots, Y_n^*)^T$.

### 1.2.5 Testing with Measurement Error in Predictor

As measurement error models have received increasing attention in the literature, model testing problems have also arisen. For example, we may be interested in testing whether a certain covariate needs to be included in the model, or may be interested in testing whether a certain parametric model is sufficient to describe the data. Although testing problems in measurement error models are important, it seems to be untouched except for some special cases. We briefly discuss hypothesis tests concerning regression parameters when $\mathbf{X}$ is measured with error.

Suppose the main problem of interest involves a response $Y$ and predictors $\mathbf{X}$. Consider the full model

$$\mathbb{E}\left(Y \mid \mathbf{X}_1, \mathbf{X}_2\right) = \beta_0 + \mathbf{X}_1^T \boldsymbol{\beta} + \mathbf{X}_2^T \boldsymbol{\gamma}, \tag{1.24}$$

we are interested in testing the null hypothesis

$$H_0 : \boldsymbol{\gamma} = 0. \tag{1.25}$$

However in the measurement error context, $\mathbf{X}_1$ and $\mathbf{X}_2$ can not be observed directly, and instead, the surrogate, call $\mathbf{W}_1$ and $\mathbf{W}_2$ are observed.

Hypothesis (1.25) can be tested using the following statistics:

$$\begin{aligned} \text{Likelihood Ratio:} \quad & \text{LR} = -2\left\{\ell(\widetilde{\boldsymbol{\beta}}) - \ell(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\gamma}})\right\}, \\ \text{Wald Test:} \quad & \text{WT} = n\hat{\boldsymbol{\gamma}}\hat{\mathbf{V}}^{-1}\hat{\boldsymbol{\gamma}}, \end{aligned}$$

where $\widetilde{\boldsymbol{\beta}}$ denotes the maximum likelihood estimator of $\boldsymbol{\beta}$ restricted to $H_0$ in (1.25), $\hat{\boldsymbol{\gamma}}$ and $\hat{\boldsymbol{\beta}}$ are the ML estimate of $\boldsymbol{\gamma}$ and $\boldsymbol{\beta}$, respectively, $\hat{\mathbf{V}}$ is a consistent estimate of its $\sqrt{n}$-asymptotic covariance matrix. Under some suitable regularity conditions, we have that under $H_0$ these statistics share the same asymptotic behavior; that is, when the sample size $n$ is large enough,

$$\begin{aligned} \text{LR} &\xrightarrow{\mathcal{D}} \chi^2_{p_\gamma}, \\ \text{WT} &\xrightarrow{\mathcal{D}} \chi^2_{p_\gamma}, \end{aligned}$$

where $p_\gamma$ is the dimension of $\boldsymbol{\gamma}$. Then the null hypothesis is rejected if LR or WT $> z^*$, where $z^*$ is a per-determined critical value.

In some simple cases, *likelihood ratio test* (LRT) or Wald-type test can be directly applied in measurement error models. However, there are some reasons to avoid the LRT or Wald-type test, and the reasons are primarily computational. The difficulty lies in solving the estimating equations, as it requires solving $\sqrt{p_\beta + p_\gamma}$ equations, where $p_\beta$ and $p_\gamma$ are the dimensions of $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$ respectively. Even in the simple case that $\boldsymbol{\gamma}$ is scalar, the increase in dimensionality can lead to difficult issues of computational stability.

There are some estimators proposed in the setting of functional measurement error models that are less computational and more theoretically-driven. Carroll, Hart and Ma (2011) [74] proposed a score-like local test and a series expansion based omnibus test in this context, where no likelihood function is available or calculated. All the tests are proposed in the semi-parametric model framework, based on a class of semi-parametric estimators developed by Tsiatis and Ma (2004) [111]. Based on Tsiatis and Ma (2004) [111], the estimating equations exist for $(\boldsymbol{\beta}, \boldsymbol{\gamma})$ and can be written as

$$0 = \sum_{i=1}^{n} \phi_\beta(\boldsymbol{W}_i, Y_i, \boldsymbol{\beta}, \boldsymbol{\gamma}),$$
$$0 = \sum_{i=1}^{n} \psi_\gamma(\boldsymbol{W}_i, Y_i, \boldsymbol{\beta}, \boldsymbol{\gamma}),$$

where $\phi_\beta$ and $\psi_\gamma$ have the same dimensions as $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$, respectively. They have used different symbols $\phi(\cdot)$ and $\psi(\cdot)$, because these estimating equations are not derivatives of some version of a profile likelihood, since no profile likelihood exists in this semi-parametric framework. Under $H_0$, the estimating equation can be simplified as

$$0 = \sum_{i=1}^{n} \phi_\beta(\boldsymbol{W}_i, Y_i, \boldsymbol{\beta}, \mathbf{0}),$$

and we call its root $\hat{\boldsymbol{\beta}}$. Then, in analogy with the score test, They proposed a test on the estimated score:

$$\hat{U} = \sqrt{n} \sum_{i=1}^{n} \psi_\gamma(\boldsymbol{W}_i, Y_i, \hat{\boldsymbol{\beta}}, \mathbf{0}).$$

The test statistic with significant level $\alpha$ proposed is to reject the hypothesis if $T = \hat{U}^T \hat{\Sigma}_0^{-1} \hat{U}$ exceeds the $1-\alpha$ quantile of the chi-squared distribution with $p_\gamma$ degrees of freedom. Of course, $T$ does not involve estimating $\gamma$. However, this test has asymptotic level $\alpha$ from standard Taylor series calculations, yielding the following result.

**Theorem 1.1.** *Under the null hypothesis,* $\sqrt{n}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) \xrightarrow{\mathcal{D}} N(0, \mathbf{V}_\beta)$ *and* $\hat{U} \xrightarrow{\mathcal{D}} N(0, \Sigma_0)$. *Hence* $T = \hat{U}^T \hat{\Sigma}_0^{-1} \hat{U}$ *is asymptotically chi-squared with* $p_\gamma$ *degrees of freedom.*

### 1.3 Mixtures of Regression with Measurement Errors

Research on mixtures of regression models primarily assumes directly observed predictors, and measurement error is often not taken into consideration. Yao and Song (2015) [119] developed a deconvolution method to get a consistent estimator for mixtures of linear regression model with measurement errors, and also proposed a generalized EM algorithm to find the estimator.

In this dissertation, we will discuss novel mixtures of regressions models with measurement errors. Here is an outline of the dissertation. In Chapter 2 we discuss the mixtures of linear regressions with measurement errors in the response, develop some estimating methods, and conduct simulations. In Chapter 3, we introduce the mixtures of linear regressions model with measurement errors in the predictor, develop some estimating methods, construct hypothesis test on polynomial regression with measurement error, and conduct simulations. Chapter 4 focuses on different estimating methods for mixtures of Poisson regressions with measurement errors and their applications. In Chapter 5, we present some concluding remarks and directions for future research.

# Chapter 2 Mixtures-of-Regressions with Measurement Error in the Response

*Measurement error* (ME) models, i.e. errors-in-variables models are an alternative to the classical model, which accounts for the difference between a measured value of a quantity and its true value. Variability is an inherent part of the results of measurements and of the measurement process. The effect of measurement error has been long investigated, details about this topic can be found in Fuller (1987) [1], Cheng and Van Ness (1999) [27] and Carroll et al. (2006) [94]. Some issues that arise due to the presence of measurement error include bias in parameter estimation for statistical models, loss of power, and masking the features of the data thus making graphical model analysis difficult.

Linear regression is one of the most common statistical techniques used in astronomical research. One of the interesting features of many astronomical data sets is the presence of *intrinsic scatter* in addition to heteroscedastic variances. Some of the most commonly used approaches in astronomy for regression in order to estimate the model parameters include *least square* (LS) fits, *weighted least squares* (WLS) methods, *maximum likelihood* (ML), survival analysis, and Bayesian methods.

In this chapter, we concentrate on the standard mixture of linear regression model, where the observed response includes measurement error with the variance roughly known, which does arise with astronomical data sets. We extend the WLS estimator discussed in Chapter 1, developed by Akritas and Bershady (1996) [46], but in the context of a mixture of linear regressions models.

## 2.1 Introduction

### 2.1.1 Measurement Error Problem in Linear Regression

Linear regression is commonly used in astronomical data analysis. While dealing with linear regression in astronomy, besides the regular random errors in the independent

and dependent variables, it is common to also have so-called intrinsic scatter on the regression line. In astronomy, intrinsic scatter is the variations in the physical properties of astronomical sources that are not completely captured by the variables included in the regression. It is important to also account for intrinsic scatter in the data analysis, since it has a non-negligible effect on the regression results. When the independent variable is measured with error, the *ordinary least squares* (OLS) estimate of the regression slope is biased toward zero (Fuller (1987) [1], Akritas and Bershady (1996) [46]).

Many methods have been proposed for performing linear regression when intrinsic scatter is present. Clutton-Brock (1967) [30] proposed a *effective variance* method; Press et al. (1992) [92] proposed a procedure of minimizing an "effective" $\chi^2$ statistic; Stephens and Dellaportas (1992) [33], Richardson and Gilks (1993) [95], Dellaportas and Stephens (1995) [35] and Gustafson (2004) [50] developed Bayesian approaches on estimating measurement error model; Schafer (1997) [100] assumed the probability distribution for the true independent variables and constructed the so-called structural equation models. Some of the methods applied in astronomy are the *bivariate correlated errors and intrinsic scatter* (BCES) estimator (Akritas and Bershady (1996) [46]) and the 'FITEXY' estimator (Press et al. (1992) [92]).

In this chapter, we consider the case where the observed response also include measurement error whose variance is roughly known, as is often the case with astronomical data sets. We extend the WLS estimator developed in Akritas and Bershady (1996) [46] to accommodate the mixture of regressions models we have discussed thus far.

### 2.1.2 Basic Model

Much of the literature on mixture models focuses on mixtures of normal distributions, which underlies the model we assume for this chapter.

For the observation $(\mathbf{X}_0, Y)$, let $\mathcal{Z}$ be a latent class variable with $P(\mathcal{Z} = j \mid \mathbf{X}_0) = \lambda_j > 0$, $\sum_{j=1}^{k} \lambda_j = 1$ for $j = 1, \cdots, k$ are mixing proportions. Given $\mathcal{Z} = j$, suppose

that

$$Y \quad = \quad \mathbf{X}_0^T \boldsymbol{\beta}_j + \epsilon_j,$$

where $\epsilon_j \sim N(0, \sigma_j^2)$. Then the response $Y$ has the form

$$Y \mid \mathbf{X}_0 \sim \sum_{j=1}^{k} \lambda_j N(\mathbf{X}_0^T \beta_j, \sigma_j^2).$$

Suppose we have $n$ observations. For the $i$th observation of interest $(\mathbf{X}_i, Y_i)$, conditional on component membership $k_i$, we have the following regression relationship:

$$Y_i = \mathbf{X}_i^T \boldsymbol{\beta}_{k_i} + \epsilon_{i,k_i}, \tag{2.1}$$

where $k_i \in \{1, \cdots, k\}$ and $\epsilon_{i,k_i} \sim N(0, \sigma_{k_i}^2)$. In the non-mixture setting, the OLS solution finds the values of $\boldsymbol{\beta}$s that minimize the *residual sum of squares* (RSS):

$$\text{RSS} = (\mathbf{Y} - \mathbf{X}\boldsymbol{\beta})^T (\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}).$$

where $\mathbf{Y} = (Y_1, \cdots, Y_n)^T$ is the vector of response variables $Y_i$s, and $\mathbf{X}$ is the matrix whose rows are $\mathbf{X}_1^T, \cdots, \mathbf{X}_n^T$ .

In the present setting, we can use an EM algorithm to perform the estimation, where OLS-type estimators appear in the M-step. However, the OLS slope is biased if there is measurement error in the independent variable. We introduce a WLS estimator for the case that only response variables $Y_1, \cdots, Y_n$ are observed with error.

## 2.2 Estimating Method

### 2.2.1 A WLS-based Estimate

We begin by denoting the observed data by

$$(\mathbf{X}_i^T, Y_i^*, \eta_i^2), \tag{2.2}$$

where $\eta_i^2$ is the variance of measurement error for the response, for which the researcher has a known, good estimate. Note that, we do not also assume measurement error in the predictor. For each component membership, the model is exactly the

same form as the one we discussed in Chapter 1 (see Subsection 1.2.4), where the observed response is related to the unobserved response by:

$$Y_i^* = Y_i + \delta_i \tag{2.3}$$

such that $\delta_i \sim N(0, \eta_i^2)$ is independent of $\epsilon_i$.

Clearly, the model has non-constant error variance (heteroscedasticity) for each observation. WLS is a commonly used technique for heteroscedasticity; by assigning individual weights to the observations the heteroscedasticity can be removed by design. WLS is an example of the broader class of generalized least squares estimators. The idea was first presented by Alexander Aitken (1935) [3]. The general idea of WLS is that less weight is given to those observations with a larger error variance, which forces the variance of the residuals to be constant.

Akritas and Bershady (1996) [46] note that the optimal weight for each observation comprises both the corresponding random error variance and the intrinsic scatter (measurement error) variance. However, in a mixture of regressions setting, we also need to account for the uncertainty of component membership, so we incorporate the unobserved $\mathcal{Z}_{ij}$s into our method.

Conditional on component membership $k_i$, we have

$$\begin{aligned} Y_i^* &= Y_i + \delta_i \\ &= \mathbf{X}_i^T \boldsymbol{\beta}_{k_i} + \epsilon_{i,k_i} + \delta_i \\ &= \mathbf{X}_i^T \boldsymbol{\beta}_{k_i} + \epsilon_{i,k_i}^*, \end{aligned}$$

where $\epsilon_{i,k_i} \sim N(0, \sigma_{k_i}^2)$. With this setting, we may develop a WLS-type approach while working under the assumption that $\epsilon_{i,k_i}^*$ is independent of $Y_i^*$. However, we need estimates of the variance of $\epsilon_{i,k_i}^*$. Under our assumptions, we have

$$\text{Var}(\epsilon_{i,k_i}^*) = \text{Var}(\epsilon_{\cdot,k_i}) + \eta_i^2. \tag{2.4}$$

Since $\text{Var}(\epsilon_{\cdot,k_i})$ is unknown, $\text{Var}(\epsilon_{i,k_i}^*)$ is also unknown. We can extend the algorithm of Akritas and Bershady (1996) [46] and use this extension within an EM algorithm to estimate $\text{Var}(\epsilon_{\cdot,1}), \cdots, \text{Var}(\epsilon_{\cdot,k})$; see Algorithm 2.1.

**Algorithm 2.1** WLS-based Algorithm

(a) Given the observed data $\left\{(\mathbf{x}_1^T, y_1^*), \cdots, (\mathbf{x}_n^T, y_n^*)\right\}$ and $\eta_1^2, \cdots, \eta_n^2$, obtain the regression coefficient estimates $(\hat{\boldsymbol{\beta}}_1^T, \cdots, \hat{\boldsymbol{\beta}}_k^T)^T$ using an EM algorithm for a mixture of linear regressions problem.

(b) Calculate the residuals

$$R_{ij} = y_i^* - \mathbf{x}_i^T \hat{\boldsymbol{\beta}}_j,$$

for $i = 1, \cdots, n$ and $j = 1, \cdots, k$.

(c) Calculate the weighted mean of the residuals for each component membership

$$\bar{R}_{\cdot j} = \frac{\sum_{i=1}^n \hat{p}_{ij} R_{ij}}{\sum_{i=1}^n \hat{p}_{ij}},$$

where $\hat{p}_{ij}$ are the final posterior membership probabilities from the EM algorithm in Step (a).

(d) Obtain the estimators of $\mathrm{Var}(\epsilon_{\cdot,1}), \cdots, \mathrm{Var}(\epsilon_{\cdot,k})$ from

$$\widehat{\mathrm{Var}}(\epsilon_{\cdot,j}) = \frac{\sum_{i=1}^n \hat{p}_{ij} \left[ \left(R_{ij} - \bar{R}_{\cdot j}\right)^2 - \eta_i^2 \right]}{\sum_{i=1}^n \hat{p}_{ij}}$$

(e) Set $\widehat{\mathrm{Var}}(\epsilon_{i,j}^*) = \hat{\sigma}_{ij}^{*2} = \widehat{\mathrm{Var}}(\epsilon_{\cdot,j}) + \eta_i^2$ and define $\boldsymbol{A}_j = \mathrm{diag}(\hat{\sigma}_{1j}^{*-2}\hat{p}_{1j}, \cdots, \hat{\sigma}_{nj}^{*-2}\hat{p}_{nj})$. Then, the WLS estimator based on the further weighting from the intrinsic scatter is

$$\widetilde{\boldsymbol{\beta}}_j = (\mathbf{X}^T \boldsymbol{A}_j \mathbf{X})^{-1} \mathbf{X}^T \boldsymbol{A}_j \mathbf{Y}^*,$$

for $j = 1, \cdots, k$, where $\mathbf{Y}^* = (Y_1^*, \cdots, Y_n^*)^T$ is the vector of observed response variables $Y_i^*$s.

According to Algorithm 2.1, EM algorithm is applied only at Step (a), and WLS is only used to adjust the regression coefficients. Thus, the difference between the estimators $\widetilde{\boldsymbol{\beta}}_1, \cdots, \widetilde{\boldsymbol{\beta}}_k$ proposed by the WLS-based algorithm and the estimators from the simple mixtures-of-regressions $\hat{\boldsymbol{\beta}}_1, \cdots, \hat{\boldsymbol{\beta}}_k$ will typically not be very big. The way to correct the variances estimators suggest that the estimators of variances would be smaller than the estimators from the mixtures-of-regression, since it exclude the variances from measurement errors. Notice in Step (c), the weighted estimators of variances are obtained by subtracting the deviation of measurement error from the overall deviation, the value of $\left(R_{ij} - \bar{R}_{\cdot j}\right)^2 - \eta_i^2$ can be negative for some $i$ or $j$, which

is infrequent. We set the value to be 0 if that is the case.

### 2.2.2 Asymptotic Variance

Let $\boldsymbol{\psi}$ denote the vector of true unknown parameter values,

$$\boldsymbol{\psi} = \left(\lambda_1, \cdots, \lambda_{k-1}, \boldsymbol{\beta}_1^T, \cdots, \boldsymbol{\beta}_k^T, \sigma_1^2, \cdots, \sigma_k^2\right)^T,$$

the asymptotic variance of EM estimators, $\hat{\boldsymbol{\psi}}$, can be obtained by the inverse of the information matrix $I(\boldsymbol{\psi})$, which is the second derivatives of the likelihood function, that is

$$\sqrt{n}\left(\hat{\boldsymbol{\psi}} - \boldsymbol{\psi}\right) \xrightarrow{\mathcal{D}} N\left(0, I^{-1}(\boldsymbol{\psi})\right).$$

However, likelihood functions for mixture models are often complicated, thus other approaches are necessary. For example, Efron and Hinkley (1978) [39] suggested to use the observed Fisher information matrix instead. Later, Louis (1982) [73] introduced a technique for computing the observed information when an EM algorithm is used.

Suppose we have $n$ observations with a $k$-component mixture model, for $i = 1, \cdots, n$,

$$g_k(y_i \mid \mathbf{x}, \boldsymbol{\psi}) = \sum_{j=1}^{k} \lambda_j f_j(y_i \mid \mathbf{x}_i, \boldsymbol{\theta}_j)$$

where

$$f_j(y_i \mid \mathbf{x}_i, \boldsymbol{\theta}_j) = \frac{1}{\sigma_j}\phi\left(\frac{y_i - \mathbf{x}_i^T\boldsymbol{\beta}_j}{\sigma_j}\right)$$

is the probability density of the $i$th observation belonging to the $j$th component, $\boldsymbol{\theta}_j = \left(\boldsymbol{\beta}_j^T, \sigma_j\right)^T$ is the vector of parameters of $j$th component, and $\phi$ is the density of the standard normal distribution.

The idea here is to think of the complete data as consisting of $\mathbf{s} = \{(\mathbf{x}_i^T, y_i, \mathbf{z}_i^T), i = 1, \cdots, n\}$, where $\mathbf{z}_i = (z_{i1}, \cdots, z_{ik})^T$ is an indicator vector such that $\sum_{j=1}^{k} z_{ik} = 1$, represents which component of the mixture generated the observation $y_i$. In the current setting, $\mathbf{z}_i$ is unobserved and hence "missing", whereby the EM algorithm becomes applicable.

Let $\boldsymbol{\psi}$ be the complete parameter set of the model, consisting of the vectors of regression coefficients $\boldsymbol{\beta}_1, \cdots, \boldsymbol{\beta}_k$, the variances $\sigma_1^2, \cdots, \sigma_k^2$ and the mixing proportions

$\lambda_1, \cdots, \lambda_k$. Define

$$\mu\left(\mathbf{s} \mid \boldsymbol{\psi}\right) = \ell_{\mathrm{C}}(\boldsymbol{\psi}) = \sum_{i=1}^{n}\sum_{j=1}^{k} z_{ij}\left\{\log\lambda_j + \log f_j(y_i \mid \mathbf{x}_i, \boldsymbol{\theta}_j)\right\}$$

$$\mu^*\left((\mathbf{x}_i^T, y_i) \mid \boldsymbol{\psi}\right) = \ell_{\mathrm{O}}(\boldsymbol{\psi}) = \sum_{i=1}^{n}\sum_{j=1}^{k} \hat{p}_{ij}\left\{\log\lambda_j + \log f_j(y_i \mid \mathbf{x}_i, \boldsymbol{\theta}_j)\right\}$$

be the log-likelihood using the complete data and observed data, respectively. Here, $\hat{p}_{ij}$ is the conditional probability that the $i$th observation belongs to the $j$th component of the mixture, given that observation,

$$\hat{p}_{ij} = \frac{\lambda_j f_j(y_i \mid \mathbf{x}_i, \boldsymbol{\theta}_j)}{\sum_{s=1}^{k}\lambda_s f_s(y_i \mid \mathbf{x}_i, \boldsymbol{\theta}_j)}.$$

To compute the observed information in the EM algorithm, let $S(\mathbf{s} \mid \boldsymbol{\psi})$ and $S^*((\mathbf{x}_i^T, y_i) \mid \boldsymbol{\psi})$ be the gradient vectors of $\mu$ and $\mu^*$ respectively, and $B(\mathbf{s} \mid \boldsymbol{\psi})$ and $B^*((\mathbf{x}_i^T, y_i) \mid \boldsymbol{\psi})$ be the negatives of the associated second derivative matrices. Then by differentiation, the observed information matrix can be written as

$$I(\hat{\boldsymbol{\psi}}) = \mathbb{E}_{\boldsymbol{\psi}}\left\{B(\mathbf{s} \mid \boldsymbol{\psi})\right\} - \mathbb{E}_{\boldsymbol{\psi}}\left\{S(\mathbf{s} \mid \boldsymbol{\psi})S^T(\mathbf{s} \mid \boldsymbol{\psi})\right\} + S^*\left\{(\mathbf{x}_i^T, y_i) \mid \boldsymbol{\psi}\right\} S^{*T}\left\{(\mathbf{x}_i^T, y_i) \mid \boldsymbol{\psi}\right\}.$$

Thus, the asymptotic variance of the estimator $\boldsymbol{\psi}$ can be calculated based on $\mathrm{Var}(\hat{\boldsymbol{\psi}}) = 1/I(\hat{\boldsymbol{\psi}})$.

### 2.2.3 Bootstrap Estimator for the Standard Errors

Even when estimation of $\boldsymbol{\psi}$ is trivial, estimation of *standard errors* (SE) can be computationally burdensome, especially when measurement error is involved. One alternative strategy we can use is the parametric bootstrap (Efron and Tibshirani (1993) [40] and Davison and Hinkley (1997) [17]), which theoretically should provide similar estimates to the standard errors compared to the method involving the information matrix.

The development of this procedure has become especially useful for mixture settings. Feng and McCulloch (1994) [42] noted that the bootstrap is a preferred method for testing the number of components of a normal mixture with unequal variances;

Ciarlini, Regoliosi and Pavese (2004) [29] proposed a bootstrap algorithm for mixture models in inter-comparisons.

We now introduce an algorithm for a parametric bootstrapping in the mixtures-of-regressions model when accounting for measurement error in the response.

---

**Algorithm 2.2** Parametric Bootstrap for Standard Errors

---

(a) Find the maximum likelihood estimate $\hat{\boldsymbol{\psi}}_j = (\widetilde{\boldsymbol{\beta}}_j, \hat{\sigma}_j^2, \hat{\lambda}_j)^T$, $j = 1, \cdots, k$ by implementing Algorithm 2.1 based on the observed data $\{(\mathbf{x}_1, y_1^*), \cdots, (\mathbf{x}_n, y_n^*)\}$.

(b) Generate a bootstrap sample of size $n$ from

$$Y_i^{**} \sim \sum_{j=1}^{k} \hat{\lambda}_j N\left(\mathbf{x}^T \widetilde{\boldsymbol{\beta}}_j, \hat{\sigma}_j^2\right).$$

Call this sample $\{(\mathbf{x}_1, y_1^{**}), \cdots, (\mathbf{x}_n, y_n^{**})\}$.

(c) For each of $y_i^{**}$, record the "observed" response by

$$y_i^{***} = y_i^{**} + \delta_i.$$

(d) Find the estimate $\widetilde{\boldsymbol{\psi}}$ by implementing Algorithm 2.1 on $(\mathbf{x}_1, y_1^{***}), \cdots, (\mathbf{x}_n, y_n^{***})$.

(e) Repeat steps (b) - (d) $B$ times to generate the bootstrap sampling distribution $\widetilde{\boldsymbol{\psi}}^{(1)}, \widetilde{\boldsymbol{\psi}}^{(2)}, \cdots, \widetilde{\boldsymbol{\psi}}^{(B)}$.

---

After implementing Algorithm 2.2, the bootstrap variance-covariance matrix is easily computed as the sample variance-covariance matrix of the generated values $\widetilde{\boldsymbol{\psi}}^{(1)}, \widetilde{\boldsymbol{\psi}}^{(2)}, \cdots, \widetilde{\boldsymbol{\psi}}^{(B)}$. Thus, bootstrap standard errors are readily available. When performing a bootstrapping procedure in the mixture setting, one must be cognizant of the label switching problem described in Chapter 1, that is, we want to set the identifiability constraint for a particular data set before we conduct the data analysis, for example, set $\boldsymbol{\beta}_1 < \cdots < \boldsymbol{\beta}_k$, or $\sigma_1 < \cdots < \sigma_k$ according to the data.

## 2.3 Numerical Studies

In this section, the sampling behavior of the proposed estimates for our mixture-of-regression model with measurement error in the response is studied using *Monte*

*Carlo* (MC) simulation with different settings.

### 2.3.1 Simulated Data — Measurement Error in the Response

**Example 1: (Mixtures of Simple Linear Regressions)**

**I. 2-Component Mixtures**

We generate the *i.i.d.* data $(x_i, y_i, \eta_i)$, $i = 1, \cdots, n$ from the model

$$Y_i \sim \lambda N\left(\beta_{10} + \beta_{11}X_i, \sigma_1^2\right) + (1-\lambda)\, N\left(\beta_{20} + \beta_{21}X_i, \sigma_2^2\right),$$

$$Y_i^* = Y_i + \delta_i,$$

where $\delta_i \sim N(0, \eta_i^2)$, $\lambda = 0.5$ is the mixing proportion, $X_i \sim \text{Unif}(0,1)$, $\sigma_1^2 = 4$ and $\sigma_2^2 = 1$.

Let $\boldsymbol{\beta}_1 = (\beta_{10}, \beta_{11})^T, \boldsymbol{\beta}_2 = (\beta_{20}, \beta_{21})^T$. To study the effect of the measurement error $\delta_i$s on the proposed estimator, we consider the following three cases with different settings:

***Case I:*** Well-Separated Components

$$\boldsymbol{\beta}_1^T = (-10, 6), \quad \boldsymbol{\beta}_2^T = (10, 2).$$

***Case II:*** Moderately-Separated Components

$$\boldsymbol{\beta}_1^T = (5, 15), \quad \boldsymbol{\beta}_2^T = (25, -15).$$

***Case III:*** Heavily-Overlapping Components

$$\boldsymbol{\beta}_1^T = (5, 5), \quad \boldsymbol{\beta}_2^T = (15, -5).$$

For each simulation condition, we randomly generated $B = 1000$ data sets, each of size either $n = 100$ or 250. For each sample size, we generated a series of measurement error with either $\eta_i^2 \sim \text{Uniform}(0, 0.1)$ or $\eta_i^2 \sim \text{Uniform}(2, 6)$, where the former one causes a small discrepancy between original and observed data and the latter one

doubles the variability of the data points based on the variance of random error. For each MC sample, we add the measurement error with the same amount of standard deviation for all $i = 1, \cdots, n$.

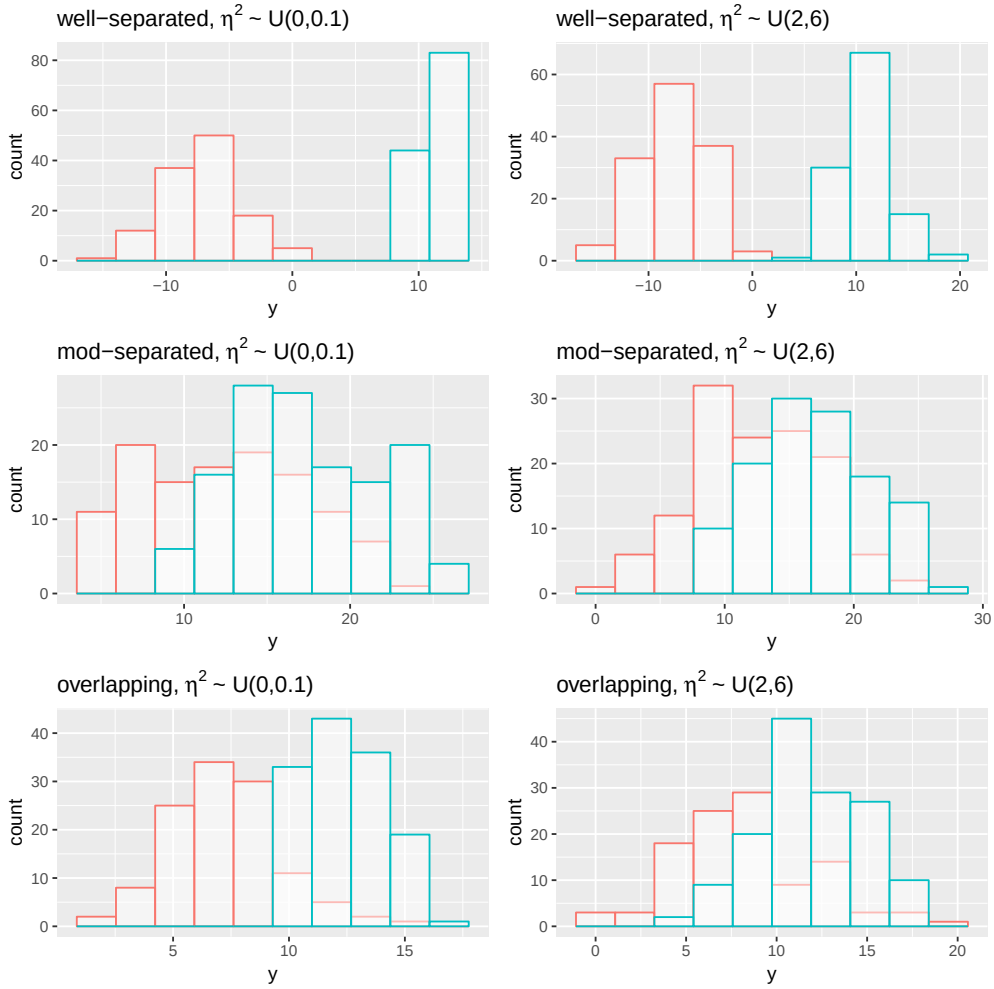


Figure 2.1: Histograms of observed response variables under different settings, with sample size $n = 250$. Note that the relationship conditioned on the predictors is not reflected in these histograms.

Figure 2.1 shows the histograms of observed responses $y^*$ under different circumstances. In the well-separated setting, there are two distinct modes representing two different components, while dealing with moderately-separated and overlapping settings, since the two components have increased mixing, it is harder to identify which component a certain data point belongs to. Note that this relationship conditioned on the predictors is not reflected in the figure. Besides, when increasing the variances

of measurement errors, the large variability of two components makes them closer to each other, which also leads to a harder time in identifying distinct components.

For each simulated data set, we estimate the mixture of regression parameters $\left(\boldsymbol{\beta}_1^T, \boldsymbol{\beta}_2^T, \sigma_1^2, \sigma_2^2\right)$ by the proposed method, and compare it with the so-called naïve method, which simply ignores the measurement error. The performance of the proposed method under different conditions is assessed by the *mean squared error* (MSE); i.e.,

$$\text{MSE}(\hat{\boldsymbol{\theta}}) = \frac{1}{B}\sum_{t=1}^{B}(\hat{\boldsymbol{\theta}}^{(t)} - \boldsymbol{\theta})^2$$

where $\hat{\boldsymbol{\theta}}^{(t)}$ is the estimate of the parameter $\boldsymbol{\theta}$ based on $t$th replication and $\boldsymbol{\theta}$ is the true value. The relative efficiency of MSE for the naïve method versus the proposed method is also recorded for all the parameters.

In Table 2.1 are the MSEs and relative efficiencies (in parentheses) for our simulated data sets. The values in the parentheses represent the relative efficiencies of MSEs for naïve versus proposed estimators. For example, 1.0552 means the MSE of $\beta_{21}$ of naïve method for moderately-separated component with measurement error $\text{U}(2, 6)$, with sample size 250, is $1.0552 \times$ MSE of proposed method for the same parameter. If the relative efficiency is greater than 1, it means the MSE of proposed method is smaller, which leads to a better performance of estimation. Label switching did not appear to be present since the identifiability constraint $\beta_{10} < \beta_{20}$ is met for all bootstrap estimates, even though it was never enforced.

Overall, the proposed method behaves better than the naïve method, since most of the relative efficiencies are greater than 1. For estimating the variances $\text{Var}\left(\epsilon_{\cdot,j}\right)$ with a larger value ($\sigma_1 = 2$ rather than $\sigma_2 = 1$), the average relative efficiency for the settings with measurement error $\text{U}(2, 6)$ is greater than 2. When measurement error is trivial, it is difficult to tell the difference between true and observed responses as the behaviors of the two methods are almost the same. We can conclude that our proposed method behaves better when the measurement error is larger, and accounting for measurement error in this setting is much more important. However, because our proposed method only deals with measurement error in the response after applying

Table 2.1: MSEs of estimators in 2-component mixture of simple linear regressions.

| $n$ | $\eta_i^2$ | $\beta_{10}$ | $\beta_{11}$ | $\beta_{20}$ | $\beta_{21}$ | $\sigma_1^2$ | $\sigma_2^2$ |
|---|---|---|---|---|---|---|---|
| **Well-Separated Components** | | | | | | | |
| 100 | U(0, 0.1) | 0.3531 | 1.0550 | 0.0801 | 0.2461 | 0.6722 | 0.0425 |
| | | (1.0002) | (1.0001) | (1.0019) | (1.0008) | (0.9843) | (1.0235) |
| 250 | | 0.1359 | 0.4356 | 0.0338 | 0.1000 | 0.2551 | 0.0177 |
| | | (1.0004) | (1.0003) | (1.0016) | (1.0025) | (0.9850) | (1.0895) |
| 100 | U(2, 6) | 0.6419 | 2.0757 | 0.3878 | 1.2180 | 8.2657 | 11.1670 |
| | | (1.0099) | (1.0121) | (1.0580) | (1.0551) | (1.8492) | (1.2782) |
| 250 | | 0.2442 | 0.7692 | 0.1616 | 0.4966 | 8.5673 | 12.1929 |
| | | (1.0171) | (1.0192) | (1.0499) | (1.0413) | (1.8948) | (1.2908) |
| **Moderately-Separated Components** | | | | | | | |
| 100 | U(0, 0.1) | 0.3684 | 1.1907 | 0.0943 | 0.3086 | 0.8366 | 0.0553 |
| | | (0.9994) | (0.9992) | (1.0020) | (1.0017) | (1.0389) | (1.0412) |
| 250 | | 0.1376 | 0.4311 | 0.0345 | 0.1184 | 0.3136 | 0.0234 |
| | | (1.0004) | (1.0022) | (1.0016) | (1.0032) | (1.8558) | (1.0260) |
| 100 | U(2, 6) | 0.8202 | 3.1092 | 0.4664 | 1.7427 | 7.7301 | 10.2705 |
| | | (1.0303) | (1.023) | (1.0611) | (1.0492) | (2.0686) | (1.2932) |
| 250 | | 0.2920 | 0.9428 | 0.1760 | 0.6098 | 7.9266 | 12.2029 |
| | | (1.0598) | (1.0514) | (1.0523) | **(1.0552)** | (2.1659) | (1.3049) |
| **Overlapping Components** | | | | | | | |
| 100 | U(0, 0.1) | 0.3920 | 1.3037 | 0.0988 | 0.4589 | 1.0774 | 0.0820 |
| | | (0.9990) | (0.9997) | (1.0027) | (1.0004) | (0.9799) | (0.9861) |
| 250 | | 0.1587 | 0.5338 | 0.0446 | 0.1836 | 0.3580 | 0.0319 |
| | | (0.9927) | (1.0026) | (0.9985) | (0.9916) | (0.9582) | (1.0240) |
| 100 | U(2, 6) | 1.3720 | 4.5647 | 0.8550 | 3.3583 | 7.0853 | 9.1205 |
| | | (1.6076) | (1.1515) | (1.4303) | (1.1468) | (2.9174) | (1.0341) |
| 250 | | 0.4532 | 1.8502 | 0.3732 | 1.6403 | 4.7926 | 11.0519 |
| | | (1.3647) | (0.9572) | (1.0541) | (0.8900) | (3.5687) | (1.3208) |

the EM algorithm to the whole mixture model, it is unable to also correct the mixing proportion, and also cannot improve the estimates of the regression parameters very much.

When the sample size increases from 100 to 250 the MSEs decrease, and our proposed method appears better than the naïve method. On the other side, if we expand the values of measurement error in the response, the MSEs become larger, however, the performance of proposed method according to the relative efficiencies is better

(with respect to the same sample size). It is reasonable to infer that, if we increase the measurement error, the estimators using our proposed method will not represent our true parameters as accurate as those with smaller measurement errors, but the performance of it will be much better than the naïve method, which simply ignores the measurement error term.

### II. 3-Component Mixtures

We next consider a 3-component mixture. We generate the *i.i.d.* data $(x_i, y_i, \eta_i)$, $i = 1, \cdots, n$ from the model

$$Y_i \sim \lambda_1 N\left(\beta_{10} + \beta_{11} X_i, \sigma_1^2\right) + \lambda_2 N\left(\beta_{20} + \beta_{21} X_i, \sigma_2^2\right) + \lambda_3 N\left(\beta_{30} + \beta_{31} X_i, \sigma_3^2\right),$$

$$Y_i^* = Y_i + \delta_i,$$

where $\delta_i \sim N(0, \eta_i^2)$, $\lambda_1 = \lambda_2 = \lambda_3 = \frac{1}{3}$ are the mixing proportions, $X_i \sim \text{Unif}(0, 1)$, $\sigma_1 = 2$, $\sigma_2 = 1$ and $\sigma_3 = 3$. Let $\boldsymbol{\beta}_1^T = (\beta_{10}, \beta_{11})$, $\boldsymbol{\beta}_2^T = (\beta_{20}, \beta_{21})$ and $\boldsymbol{\beta}_3^T = (\beta_{30}, \beta_{31})$. Again, we consider the following three cases with different settings:

***Case I:*** Well-Separated Components

$$\boldsymbol{\beta}_1^T = (-10, 6), \quad \boldsymbol{\beta}_2^T = (10, 2), \quad \boldsymbol{\beta}_3^T = (30, -5).$$

***Case II:*** Moderately-Separated Components

$$\boldsymbol{\beta}_1^T = (5, 15), \quad \boldsymbol{\beta}_2^T = (20, 20), \quad \boldsymbol{\beta}_3^T = (25, -15).$$

***Case III:*** Heavily-Overlapping Components

$$\boldsymbol{\beta}_1^T = (-10, 20), \quad \boldsymbol{\beta}_2^T = (5, 5), \quad \boldsymbol{\beta}_3^T = (15, -5).$$

For each simulation condition, we then randomly generated $B = 1000$ data sets, each of size either $n = 100$ or 250. For each sample size, we generated a series of measurement error with either $\eta_i^2 \sim \text{Uniform}(0, 0.5)$ or $\eta_i^2 \sim \text{Uniform}(5, 10)$. For

Table 2.2: MSEs of estimators in 3-component mixture of simple linear regressions.

| $n$ | $\eta_i^2$ | $\beta_{10}$ | $\beta_{11}$ | $\beta_{20}$ | $\beta_{21}$ | $\beta_{30}$ | $\beta_{31}$ | $\sigma_1^2$ | $\sigma_2^2$ | $\sigma_3^2$ |
|---|---|---|---|---|---|---|---|---|---|---|
| **Well-Separated Components** | | | | | | | | | | |
| 100 | U(0, .5) | 0.5330 | 1.5660 | 0.1870 | 0.4602 | 1.1617 | 3.5029 | 1.0515 | 6.1266 | 6.2885 |
| | | (1.0025) | (1.0012) | (1.0158) | (1.0089) | (0.9996) | (0.9982) | (0.9757) | (1.0225) | (0.9800) |
| 250 | | 0.2262 | 0.6790 | 0.0617 | 0.1904 | 0.4619 | 1.3618 | 0.5769 | 1.3600 | 3.0806 |
| | | (1.0030) | (1.0025) | (1.0071) | (1.0111) | (0.9987) | (0.9992) | (1.0280) | (1.0891) | (0.9848) |
| 100 | U(5, 10) | 2.2853 | 7.9456 | 2.2967 | 5.6084 | 2.8218 | 8.8450 | 41.2184 | 119.5947 | 49.2994 |
| | | (1.0224) | (1.0170) | (1.0354) | (1.0261) | (1.0474) | (1.0461) | (1.5465) | (1.2127) | (1.8582) |
| 250 | | 0.5122 | 1.6757 | 0.4544 | 1.4282 | 0.8378 | 2.7573 | 33.7626 | 53.1254 | 25.0650 |
| | | (1.0230) | (1.0188) | (1.0258) | (1.0275) | (1.0260) | (1.0323) | (1.5797) | (1.2139) | (2.2608) |
| **Moderately-Separated Components** | | | | | | | | | | |
| 100 | U(, 0.5) | 0.6619 | 2.5107 | 1.8705 | 4.6683 | 0.7329 | 2.0314 | 1.9033 | 61.7355 | 59.8475 |
| | | (0.9995) | (0.9969) | (1.0019) | (1.0037) | (0.9983) | (0.9998) | (0.9599) | (0.9631) | (1.0482) |
| 250 | | 0.2350 | 0.7756 | 0.5871 | 1.7277 | 0.1041 | 0.2834 | 0.8868 | 61.5826 | 64.6231 |
| | | (1.0031) | (1.0010) | (1.0009) | (0.9993) | (1.0072) | (1.0119) | (1.0031) | (0.9576) | (1.0485) |
| 100 | U(5, 10) | 6.1955 | 40.8465 | 7.4054 | 18.3020 | 11.4807 | 42.4403 | 51.5176 | 14.0030 | 167.5460 |
| | | (1.0728) | (1.0526) | (1.0209) | (1.0033) | (1.0613) | (1.0391) | (1.5418) | (2.2821) | (1.4550) |
| 250 | | 0.9832 | 5.4059 | 1.9183 | 4.3903 | 2.0748 | 5.7883 | 32.2413 | 5.4198 | 151.2731 |
| | | (1.0403) | (1.0278) | (0.9849) | (0.9899) | (1.0139) | (1.0287) | (1.6778) | (1.8687) | (1.4886) |
| **Overlapping Components** | | | | | | | | | | |
| 100 | U(, 0.5) | 2.0540 | 6.7647 | 1.8261 | 5.7137 | 0.2518 | 1.1633 | 12.227 | 6.7275 | 0.9974 |
| | | (0.9966) | (0.9952) | (0.9980) | (0.9902) | (1.0026) | (1.0309) | (0.9672) | (0.9896) | (1.1254) |
| 250 | | 0.5923 | 2.2360 | 0.3429 | 1.7953 | 0.0773 | 0.3423 | 3.8101 | 1.9859 | 0.6644 |
| | | (0.9976) | (0.9932) | (0.9970) | (0.9876) | (1.0037) | (0.9989) | (0.9477) | (0.9813) | (1.2213) |
| 100 | U(5, 10) | 10.0582 | 35.1593 | 24.5870 | 38.8456 | 7.3339 | 16.6268 | 49.3850 | 42.0632 | 71.0176 |
| | | (1.0882) | (1.0617) | (1.0170) | (1.0321) | (1.1401) | (1.1119) | (2.0085) | (1.6594) | (1.2376) |
| 250 | | 4.6846 | 10.0172 | 10.7153 | 18.6601 | 3.3252 | 6.3234 | 31.3635 | 36.5494 | 60.9078 |
| | | (1.0657) | (1.0444) | (1.0185) | (1.0413) | (1.1256) | (1.1043) | (2.2489) | (1.7373) | (1.2545) |

each MC sample, we add measurement error with the same standard deviation for all $i = 1, \cdots, n$.

For each simulated data set, we estimated the mixture of regression parameters $\left(\boldsymbol{\beta}_1^T, \boldsymbol{\beta}_2^T\right)$ by the proposed method, and also computed the relative efficiency of MSE for the naïve method versus the proposed method for all the parameters listed. In Table 2.2 are the MSEs and relative efficiencies (in parentheses) for our simulated data sets.

Label switching was present in this bootstrap sample for moderately-separated cases. This was diagnosed by first noting that the MSEs appeared to be fairly large for some parameters when measurement error is large. For example, MSE of $\beta_{21}$ for moderately-separated setting with $\eta_i^2 \sim \mathrm{U}(5, 10)$ and sample size $n = 100$ was first found to be 133.1943, much larger than expected. Since the values of $\beta_{20}$ and $\beta_{30}$ are close to each other, simply applying the identifiability constraint $\beta_{10} < \beta_{20} < \beta_{30}$ is not enough. To make the component distinct with each other, we then imposed the identifiability constraints of $\beta_{10} < \beta_{20}$ and $\beta_{21} > \beta_{31}$ in order to correct the label switching.

When number of components increases, we see the MSEs becomes much bigger, since the more components it has, the more complicated the model becomes. Thus, the estimation becomes more challenging. Besides the MSEs, we can obtain similar results from the 3-component mixtures, when we increase the sample size and decrease the variances of measurement error in the response, the MSEs of unknown parameters becomes smaller. Similarly, the relative efficiencies show that the case with larger sample size and bigger measurement error works better under our proposed method than the naïve method. For overlapping and moderately-separated cases, the MSEs are pretty large for certain parameters with large measurement error (with variances $\eta_i^2 \sim \mathrm{U}(5, 10)$). This is complicated by the fact that the three components have heavier mixing and sometimes it is difficult to distinguish different components, thus leading to greater uncertainty in the estimates.

### III. Summary

Figure 2.2 shows the scatter plots of all six settings discussed in Example 1. Different colors represent different components that the data point belongs to, and black lines are the estimated lines from our proposed method. According to the scatter plots, the proposed method fits pretty well to all settings, and based on the relative efficiencies reported, it slightly improves the performance of the estimates, compared to the naïve method. Although the improvement is not significant, it is

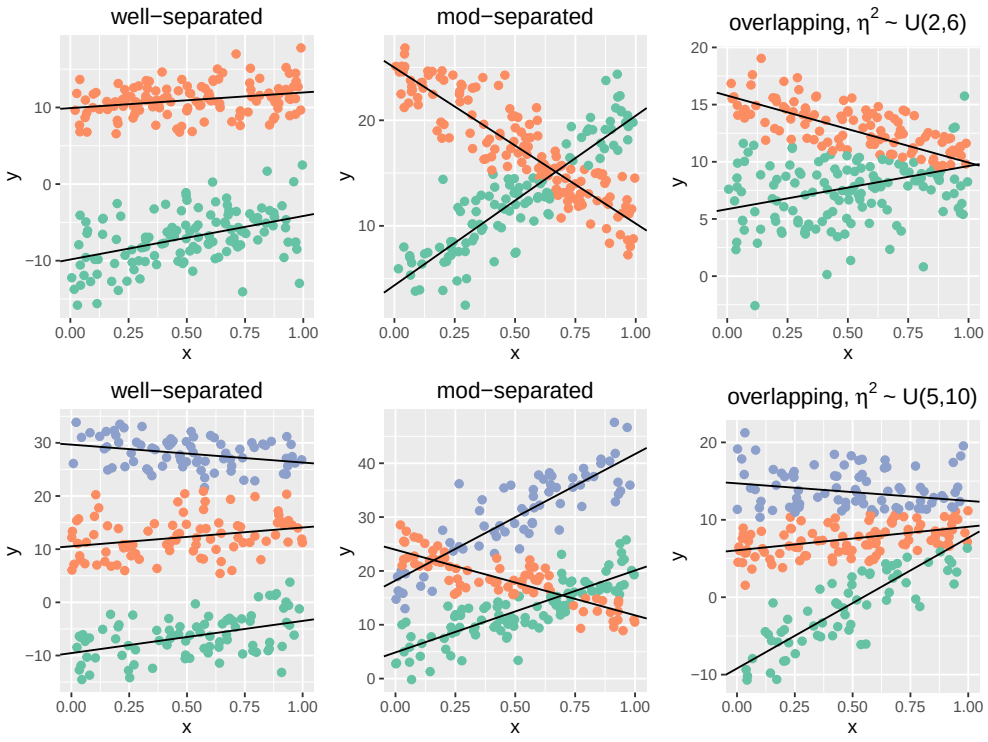


Figure 2.2: Scatter plots and fitted lines for six different settings, with sample size 250.

worth applying our new method, especially for some data sets that need improved estimates. The results show our proposed method is a reasonable way to incorporate measurement error in the response.

In general, well-separated behaves better than moderately-separated components and overlapping components, since the components in both moderately-separated and overlapping component models are harder to identify. Some data points from different components are mixed at certain values, especially when number of components become larger, there are more overlapping points. Meanwhile, for the same model with the same setting (well-separated, moderately-separated or overlapping), when we increase the sample size, MSEs will decrease; while when we increase the variances of measurement error, MSEs will increase, which makes sense for all different settings.

**Example 2: (Mixture of Multiple Linear Regressions)**

We next move on to multiple linear regressions with predictor variable is a 2-dimensional vector instead of scaler, and conduct the same simulations with two settings.

**I. 2-Component Mixtures**

Consider the data vector $\mathbf{x_i}^T = (x_{i1}, x_{i2})$, we generate the *i.i.d.* data $(\mathbf{x}_i^T, y_i, \eta_i)$, $i = 1, \cdots, n$ from the model

$$Y_i \sim \lambda N\left(\beta_{10} + \beta_{11}X_{i1} + \beta_{12}X_{i2}, \sigma_1^2\right) + (1-\lambda)\, N\left(\beta_{20} + \beta_{21}X_{i1} + \beta_{22}X_{i2}, \sigma_2^2\right),$$

$$Y_i^* = Y_i + \delta_i,$$

where $\delta_i \sim N(0, \eta_i^2)$ with either $\eta_i^2 \sim \mathrm{U}(0, 0.1)$ or $\eta_i^2 \sim \mathrm{U}(2, 6)$; $\lambda = 0.5$ is the mixing proportion, $X_{i1}, X_{i2} \sim \mathrm{Unif}(0, 1)$, $\sigma_1 = 2$ and $\sigma_2 = 1$.

Let $\boldsymbol{\beta}_1^T = (\beta_{10}, \beta_{11}, \beta_{12}), \boldsymbol{\beta}_2^T = (\beta_{20}, \beta_{21}, \beta_{22})$. To study the effect of measurement error $\delta_i$s on the proposed estimator, we consider the following three cases with different settings:

***Case I:*** Well-Separated Components

$$\boldsymbol{\beta}_1^T = (-10, 6, 4), \quad \boldsymbol{\beta}_2^T = (10, 2, 7).$$

***Case II:*** Moderately-Separated Components

$$\boldsymbol{\beta}_1^T = (5, 15, 10), \quad \boldsymbol{\beta}_2^T = (25, -15, -10).$$

***Case III:*** Overlapping Components

$$\boldsymbol{\beta}_1^T = (5, 5, 9), \quad \boldsymbol{\beta}_2^T = (15, -5, 3).$$

For each setting, we randomly generated $B = 1000$ data sets, each of size either $n = 100$ or 250. For each sample size, we generated a series of measurement error with either $\eta_i^2 \sim \mathrm{Uniform}(0, 0.1)$ or $\eta_i^2 \sim \mathrm{Uniform}(2, 6)$. For each data set, we add the measurement error with same amount of standard deviation for all $i = 1, \cdots, n$.

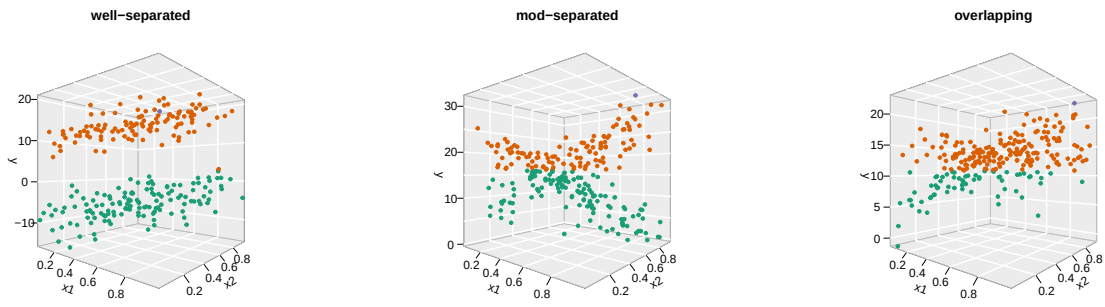


Figure 2.3: 3d scatter plots of 3 conditions with sample size $n = 250$ and measurement error $\eta_i^2 \sim \mathrm{U}(2, 6)$.

Figure 2.3 shows the 3d scatter plots under all three situations, different colors represent to which component each data point belongs. In well-separated case, two components are very well separated, makes it very easy to distinguish which component each point belongs to; for moderately-separated and overlapping cases, there are some areas that two components are mixing together and it is uncertain how to classify those points.

In Table 2.3, we report the MSEs and relative efficiencies (in parentheses) for our simulated data sets. Label switching did not appear to be present since the identifiability constraint $\beta_{10} < \beta_{20}$ is satisfied for all bootstrap estimates. The overall behaviors of the 2-component mixture of multiple linear regressions are similar to those of simple linear regressions, when we increase the sample size from 100 to 250, the MSEs become smaller and the relative efficiencies improved. Meanwhile, because we add a predictor $X_{i2}$, the models are more complicated than simple linear regressions, makes the estimation harder, especially when two components are overlapping. For example, with a overlapping component with large measurement errors (variances $\eta_i^2 \sim \mathrm{U}(2, 6)$) of sample size $n = 100$, the MSE of parameter for $X_{i2}$, call $\beta_{12}$ is 19.2855, much larger than the same setting with simple linear regressions. We can infer that if we keep increasing the dimension of predictor variables, the MSEs would be more and more difficult to capture the true parameters.

Table 2.3: MSEs of estimators in 2-component mixture of multiple linear regressions.

| $n$ | $\eta_i^2$ | $\beta_{10}$ | $\beta_{11}$ | $\beta_{12}$ | $\beta_{20}$ | $\beta_{21}$ | $\beta_{22}$ | $\sigma_1^2$ | $\sigma_2^2$ |
|---|---|---|---|---|---|---|---|---|---|
| **Well-Separated Components** | | | | | | | | | |
| 100 | U(0, 0.1) | 0.5943 | 1.0542 | 0.9975 | 0.1654 | 0.2692 | 0.2721 | 0.6641 | 0.0429 |
| | | (0.9997) | (0.9998) | (0.9994) | (1.0005) | (0.9998) | (1.0009) | (0.9711) | (0.9570) |
| 250 | | 0.2344 | 0.3588 | 0.4091 | 0.0571 | 0.1029 | 0.1001 | 0.2772 | 0.0181 |
| | | (1.0000) | (0.9999) | (1.0000) | (1.0011) | (1.0025) | (0.9999) | (0.9924) | (1.0444) |
| 100 | U(2, 6) | 1.1410 | 1.8997 | 1.9631 | 0.7192 | 1.2127 | 1.2058 | 7.8854 | 11.2173 |
| | | (1.0242) | (1.0242) | (1.0200) | (1.0356) | (1.0453) | (1.0334) | (1.8798) | (1.2486) |
| 250 | | 0.4703 | 0.7942 | 0.7993 | 0.2633 | 0.4658 | 0.4882 | 8.5387 | 12.1649 |
| | | (1.0264) | (1.0361) | (1.0163) | (1.0345) | (1.0322) | (1.0419) | (1.8905) | (1.2733) |
| **Moderately-Separated Components** | | | | | | | | | |
| 100 | U(0, 0.1) | 0.6763 | 1.2041 | 1.2587 | 0.1686 | 0.3052 | 0.3084 | 0.8869 | 0.0652 |
| | | (1.0005) | (0.9991) | (0.9999) | (1.0002) | (0.9971) | (1.0026) | (0.9788) | (0.9522) |
| 250 | | 0.2414 | 0.4074 | 0.4098 | 0.0721 | 0.1136 | 0.1233 | 0.3040 | 0.0223 |
| | | (1.0003) | (1.0008) | (0.9994) | (0.9985) | (0.9973) | (1.0015) | (0.9714) | (0.9977) |
| 100 | U(2, 6) | 1.5240 | 2.9314 | 2.8395 | 0.9511 | 2.1858 | 1.6698 | 6.8091 | 10.6683 |
| | | (1.0258) | (1.0379) | (1.0185) | (1.0542) | (1.0472) | (1.0416) | (2.1127) | (1.2768) |
| 250 | | 0.5835 | 0.9993 | 0.9861 | 0.3567 | 0.5889 | 0.6688 | 7.0279 | 11.6471 |
| | | (1.0181) | (1.0142) | (1.0195) | (1.0337) | (1.0452) | (1.0421) | (2.1744) | (1.2959) |
| **Overlapping Components** | | | | | | | | | |
| 100 | U(0, 0.1) | 1.2866 | 2.3647 | 1.8994 | 0.4989 | 1.0341 | 0.7241 | 1.2633 | 0.2225 |
| | | (1.0030) | (1.0012) | (1.0024) | (1.0071) | (1.0004) | (1.0027) | (0.9695) | (0.9831) |
| 250 | | 0.3486 | 0.6162 | 0.5630 | 0.0847 | 0.1826 | 0.1721 | 0.3895 | 0.0461 |
| | | (1.0041) | (1.00021) | (1.0033) | (1.0082) | (1.0007) | (1.0029) | (0.9744) | (0.9672) |
| 100 | U(2, 6) | 10.2329 | 18.2687 | **19.2855** | 6.5878 | 12.7481 | 7.5360 | 6.6059 | 16.4143 |
| | | (1.0901) | (1.0874) | (1.1339) | (1.1815) | (1.1073) | (1.1758) | (2.4594) | (1.1897) |
| 250 | | 3.0658 | 4.1279 | 3.3197 | 1.9051 | 2.8471 | 1.9667 | 6.3793 | 12.4284 |
| | | (1.0561) | (1.0346) | (1.0758) | (1.0923) | (1.0537) | (1.0557) | (2.2934) | (1.2622) |

## II. 3-Component Mixtures

We next consider the 3-component mixtures. We generate the *i.i.d.* data $(x_i, y_i, \eta_i)$,

$i = 1, \cdots, n$ from the model

$$\begin{aligned} Y_i \sim & \lambda_1 N\left(\beta_{10} + \beta_{11} X_{i1} + \beta_{12} X_{i2}, \sigma_1^2\right) + \\ & \lambda_2 N\left(\beta_{20} + \beta_{21} X_{i1} + \beta_{22} X_{i2}, \sigma_2^2\right) + \\ & \lambda_3 N\left(\beta_{30} + \beta_{31} X_{i1} + \beta_{32} X_{i3}, \sigma_3^2\right), \end{aligned}$$

$$Y_i^* = Y_i + \delta_i,$$

where $\delta_i \sim N(0, \eta_i^2)$, $\lambda_1 = \lambda_2 = \lambda_3 = \frac{1}{3}$ are the mixing proportions, $X_i \sim \text{Unif}(0, 1)$, $\sigma_1 = 2$, $\sigma_2 = 1$ and $\sigma_3 = 3$.

Let $\boldsymbol{\beta}_1^T = (\beta_{10}, \beta_{11}, \beta_{12}), \boldsymbol{\beta}_2^T = (\beta_{20}, \beta_{21}, \beta_{22})$ and $\boldsymbol{\beta}_3^T = (\beta_{30}, \beta_{31}, \beta_{32})$. Again, we consider the following three cases with different settings:

***Case I:*** Well-Separated Components

$$\boldsymbol{\beta}_1^T = (-10, 6, 4), \quad \boldsymbol{\beta}_2^T = (10, 2, 7), \quad \boldsymbol{\beta}_3^T = (30, -5, 10)$$

***Case II:*** Moderately-Separated Components

$$\boldsymbol{\beta}_1^T = (5, 15, 10), \quad \boldsymbol{\beta}_2^T = (20, 20, 5), \quad \boldsymbol{\beta}_3^T = (25, -15, -10)$$

***Case III:*** Overlapping Components

$$\boldsymbol{\beta}_1^T = (5, 5, 9), \quad \boldsymbol{\beta}_2^T = (15, -5, 3), \quad \boldsymbol{\beta}_3^T = (-10, 20, 15)$$

In table 2.4, we report the MSEs and relative efficiencies (in parentheses) for our simulated data sets. Label switching did not appear to be present since the identifiability constraint $\beta_{30} < \beta_{10} < \beta_{20}$ is satisfied for all bootstrap estimates. Overall, the behavior of the method is similar to the 2-component mixtures, when we increase the sample size with large measurement error, it can improve the accuracy of variances for random errors. One thing to notice is, because the complexity of the model structure, there are some parameters that have large MSE values, for example, overlapping component with U$(5, 10)$ with sample size 100, the MSE for $\beta_{11}$ is 38.7232, which is not trivial. We can image, if there are more components, the estimating method can be more challenging.

Table 2.4: MSE of estimators in 3-component mixture of bivariate normals.

| $n$ | $\eta_i^2$ | $\beta_{10}$ | $\beta_{11}$ | $\beta_{12}$ | $\beta_{20}$ | $\beta_{21}$ | $\beta_{22}$ | $\beta_{30}$ | $\beta_{31}$ | $\beta_{32}$ | $\sigma_1^2$ | $\sigma_2^2$ | $\sigma_3^2$ |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **Well-Separated Components** | | | | | | | | | | | | | |
| 100 | U(0, .5) | 1.2136 | 1.9885 | 3.9334 | 0.3336 | 0.5340 | 0.5208 | 2.2203 | 3.8362 | 3.7758 | 1.0387 | 12.3354 | 5.5941 |
| | | (1.0076) | (1.0131) | (0.9976) | (1.0177) | (1.0233) | (1.0107) | (0.9989) | (0.9997) | (0.9989) | (0.9331) | (0.9881) | (0.9505) |
| 250 | | 0.3811 | 0.6459 | 0.6263 | 0.1039 | 0.1737 | 0.1823 | 0.8305 | 1.4460 | 1.3632 | 0.4005 | 0.0372 | 2.1773 |
| | | (1.0026) | (1.0021) | (1.0029) | (0.9925) | (0.9926) | (1.0119) | (1.0002) | (1.0000) | (0.9989) | (1.0085) | (1.8628) | (0.9591) |
| 100 | U(5, 10) | 3.2963 | 5.2986 | 5.0973 | 4.5584 | 8.1695 | 8.1475 | 5.7028 | 8.8767 | 12.3215 | 85.2660 | 158.6470 | 74.7738 |
| | | (1.0482) | (1.0178) | (1.0333) | (1.0294) | (1.0043) | (1.0053) | (1.0416) | (1.0286) | (1.0205) | (1.3193) | (1.2135) | (1.5328) |
| 250 | | 0.9410 | 1.7008 | 1.6351 | 0.7914 | 1.3628 | 1.3607 | 1.5043 | 2.6883 | 2.6383 | 34.6107 | 45.3006 | 24.9767 |
| | | (1.0164) | (1.0207) | (1.0115) | (1.0113) | (1.0046) | (1.0110) | (1.0186) | (1.0253) | (1.0261) | (1.5534) | (1.1457) | (2.2178) |
| **Moderately-Separated Components** | | | | | | | | | | | | | |
| 100 | U(0, .5) | 1.9663 | 4.8574 | 4.1719 | 1.2241 | 2.8285 | 2.2239 | 4.9887 | 7.7760 | 6.6005 | 7.5266 | 9.7945 | 12.0505 |
| | | (1.0006) | (1.0011) | (0.9981) | (1.0009) | (1.0036) | (1.0086) | (1.0005) | (0.9981) | (0.9991) | (0.9960) | (1.0163) | (0.9652) |
| 250 | | 0.4809 | 1.1818 | 0.9333 | 0.1374 | 0.2160 | 0.2007 | 1.4692 | 2.5039 | 2.1921 | 0.6890 | 0.0400 | 3.3639 |
| | | (0.9995) | (0.9986) | (0.9982) | (1.0164) | (1.0111) | (1.0111) | (1.0011) | (1.0003) | (0.9995) | (0.9518) | (1.8602) | (0.9606) |
| 100 | U(5, 10) | 12.9275 | 33.8055 | 22.2632 | 5.2212 | 15.3337 | 8.8258 | 18.1433 | 37.4492 | 25.1159 | 112.7497 | 70.9902 | 50.3589 |
| | | (1.0199) | (1.0141) | (1.0321) | (1.0872) | (1.0573) | (1.0569) | (1.0159) | (1.0131) | (1.0092) | (1.4687) | (1.2221) | (1.8285) |
| 250 | | 2.0909 | 4.3709 | 3.5039 | 1.2803 | 1.8202 | 1.7139 | 3.6181 | 6.3859 | 4.9530 | 37.6301 | 47.8919 | 23.1817 |
| | | (1.0224) | (1.0296) | (1.0131) | (1.0179) | (1.0284) | (1.0160) | (0.9911) | (0.9735) | (0.9864) | (1.6905) | (1.1922) | (2.4719) |
| **Overlapping Components** | | | | | | | | | | | | | |
| 100 | U(0, .5) | 10.3035 | 20.6498 | 15.4182 | 16.7390 | 21.5233 | 33.0813 | 3.3270 | 6.4835 | 4.3271 | 20.3703 | 8.4015 | 1.2845 |
| | | (1.0063) | (1.0067) | (0.9903) | (1.0017) | (0.9917) | (1.0050) | (0.9996) | (1.0079) | (1.0006) | (0.9868) | (1.0189) | (1.1515) |
| 250 | | 2.0177 | 3.6305 | 2.8213 | 1.6731 | 2.9781 | 2.3034 | 0.2443 | 0.5121 | 0.4392 | 5.4233 | 2.8357 | 0.1046 |
| | | (0.9972) | (1.0178) | (1.0030) | (0.9998) | (0.9955) | (0.9979) | (1.0065) | (1.0029) | (1.0073) | (0.9485) | (0.9773) | (1.4291) |
| 100 | U(5, 10) | 21.8372 | 38.7232 | 31.4980 | 40.1613 | 50.7183 | 46.2146 | 12.5149 | 26.3528 | 18.1346 | 29.0741 | 26.5541 | 47.2962 |
| | | (1.1114) | (1.0869) | (1.0810) | (1.0269) | (1.0467) | (1.0616) | (1.1389) | (1.1391) | (1.1859) | (2.4170) | (1.6763) | (0.8082) |
| 250 | | 11.8025 | 17.7110 | 15.0034 | 36.2944 | 43.8553 | 25.0780 | 9.3447 | 17.8059 | 10.7296 | 24.4411 | 31.1152 | 51.5546 |
| | | (1.0978) | (1.0866) | (1.0974) | (0.9999) | (1.0217) | (1.0725) | (1.1619) | (1.1165) | (1.1073) | (2.6009) | (1.7296) | (0.9340) |

### 2.3.2 Summary

Generally speaking, the MSEs of well-separated components are the smallest among three different types of components. When we assumed a smaller measurement error, the MSEs also seemed to be smaller, which makes sense because a smaller ME indicates smaller variability. Moreover, the sample sizes also affect the MSE; larger sample size leads to a larger MSE. Overall, two-component models behave much better than three-component models, for example, for a 3-component heavily overlapping mixture model with measurement error Unif$(5, 10)$ and sample size of 100, the MSEs of $\boldsymbol{\beta}_2^T = (15, -5, 3)$ are $(45.015, 65.014, 44.014)$, while the 2-component heavily overlapping mixture model with measurement error Unif$(2, 6)$ and sample size 100 for the same $\boldsymbol{\beta}_2^T$ has MSEs of $(9.127, 18.036, 14.975)$.

When dealing with more complicated models, the MSE of parameters sometimes seem to be quite large. The structure of the mixture model leads to some special problems, especially for overlapping models. Sometimes it is difficult to fit the correct model for every single estimating process. For practical purpose, we omit the extreme EM estimators from the output. For the 3-component simulated data sets with $B = 1000$, we trimmed $40(\approx 4\%)$ of the datasets that yield the largest deviations from the true parameter value for any single estimates from $\boldsymbol{\beta}$ vectors. After omitting the 'more extreme' simulated data sets, the MSE has much smaller value than before. This strategy has been employed for other simulations involving mixtures with complex structures; see, for example, Young (2014) [122].

## 2.4 Gamma-ray Burst Data — A Real Data Analysis

Measurement error problems are widely found in astronomical research, since it often has the feature of the presence of *intrinsic scatter*, a special type of measurement error for astronomical data sets. Morrison, Mateo, Olszewski, Harding et al. (2000) [82] studied galaxy formation with a large survey of stars in the Milky Way. The investigators were interested in the velocities of stars, such that the observed velocities involved heteroscedastic measurement errors. To verify the galaxy formation theories,

one can estimate the density function from contaminated data that are effective in unveiling the numbers of bumps or components. Kelly (2007) [66] described a Bayesian method to account for measurement errors in linear regression of astronomical data. Andrae (2010) [7] presented an overview of different methods for error estimation that are applicable to both model-based and model-independent parameter estimates in astronomy.

In this section, we discuss a special astronomy phenomena — *gamma-ray burst* (RGB) and how we can use our proposed method to deal with the GRB data with measurement error in the response.

### 2.4.1 Introduction

Gamma-ray bursts (GRBs) are extremely energetic explosions that occur at random times in distant galaxies. They are the brightest electromagnetic events known to occur in the universe. The bursts can last from ten milliseconds to several hours. These phenomena are still not entirely understood, but some theories suggest they arise during the birth of black holes or a massive super-giant's collapse. GRBs were first detected in 1967 by the Vela satellites, which had been designed to detect covert nuclear weapons tests. The launch of the *Swift* observatory (Gehrels et al. (2004) [47]) had brought the observations of GRBs to a new era. Swift provides rapid notification of GRB triggers to the ground using its sensitive Burst Alert Telescope (BAT; Barthelmy et al. (2006) [9]) and can make panchromatic observations of the burst and its afterglow by bringing its narrow-field X-Ray Telescope (XRT; Burrows et al. (2006) [18]) and Ultra Violet/Optical Telescope (UVOT; Roming et al. (2006a) [98]) to bear within about 1 minute of the burst going off.

There are copious data being collected on GRBs due to the launch of Swift observatory. In the next few Subsections, we analyze a typical GRB data set, representative of those that are widely discussed for astronomical researches.

### 2.4.2 Observations and Analysis

At 00:01:53.26 UT on May 25, 2005, the Swift Alert Telescope triggered and located on board GRB050525a[1]. GRB 050525a is the second most fluent GRB to have been observed by Swift and is the first bright low-red shift burst to have been observed since all Swift instruments have been operational. The X-ray decay 'light curve' (a time series) of GRB 050525a was obtained with the XRT on board the Swift satellite, it including both *photo-diode* (PD) mode ($T < 2000$s) and *photon-counting* (PC) mode ($T > 2000$s) data. The data was presented in Blustin et al. (2006) [58] [2] and reproduced in Figure (2.4). Like most of the astronomical data sets, the observation has suffered from the measurement error due to the detection technique being used.

This data set consists of $n = 63$ brightness measurements in the 0.4 – 4.5 $keV$ spectral band at times ranging from 2 minutes to 5 days after the burst. During this period, the brightness faded by a factor of 100,000. Due to the wide range in times and brightness, most analysis is done using logarithmic variables. The observations in the data set are: time of observation (in seconds), X-ray flux (in units of $10^{-11}$ $erg/cm^2/s$, $2-10$ $keV$), and measurement error of the flux based on detector signal-to-noise values.

The data and best-fit are shown in the plot below (Figure 2.4). Since the residuals suggest heteroscedastic variances of measurement error from the model, Blustin et al. (2006) [58] fit the data with a so-called 'broken power-law' model, which is typically a piece-wise linear regression with two temporal breaks. The power-law fit to the prebrightening PD mode data ($T < 280$s) extrapolates well to the prebreak PC mode data. They concluded that the brightening at about 280s in the PD mode data represents a flare in the X-ray flux, possibly similar to the sometimes much larger flares that are seen at early times in other bursts (Burrows et al. (2005) [18], Piro et al. (2005) [91]), and the flux returns to the preflare decay curve prior to the start of the PC data. So when they tried to analyze the data, they usually omitted the

[1] The gamma ray burst is named by "GRByymmdd", where a subsequent letter (i.e., a, b, c, etc.) denotes the observation on a day when multiple gamma ray bursts occurred.

[2] Available at http://arxiv.org/abs/astro-ph/0507515

'flaring' points (orange dots in the plot).

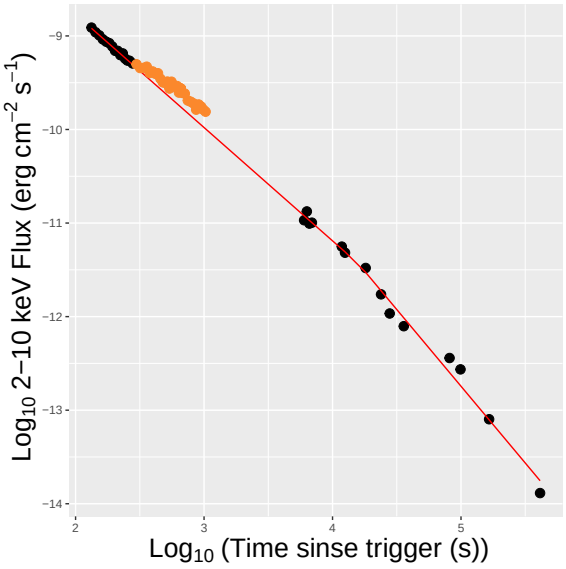


Figure 2.4: The GRB050525a data set with the best fit line from broken power-law model.

However, this approach suffers from losing some important information from the original data collected, as well as ignores the measurement error. In order to also capture the characteristic of the flaring part of this phenomena, we want to fit the data with a mixture of regression model, which can potentially identify separate regression models for the initial burst.

First, we want to assess the number of components for the mixture of regressions model to be fit to the GRB data. We consider $k = 1, 2, 3, 4$. The four model selection criteria discussed in Subsection 1.1.4 were used to assess these fits. The number of components is chosen based on the smallest penalized log likelihood value. This was repeated with $N = 100$ random starts, the scores from the best start are given in Table 2.5.

Among the model selection criteria, AIC typically overestimates while BIC, ICL, and cAIC are good indicators for the fit of a mixture model (Wedel and DeSarbo (1995) [114], McLachlan (1987) [77]). In this case, BIC, ICL, and cAIC all select $k = 2$ while AIC appears to overestimate by selecting $k = 4$. Based on this result, we proceed to fit a 2-component model with measurement error in the response.

Table 2.5: Various criteria for the determination of the number of components for the GRB data set. The bold values indicate the number of components chosen for that criterion.

| $k$ | AIC | BIC | cAIC | ICL |
|---|---|---|---|---|
| 1 | $-84.935$ | $-80.649$ | $-78.649$ | $-80.649$ |
| 2 | $-156.654$ | $\mathbf{-143.796}$ | $\mathbf{-137.796}$ | $\mathbf{-145.016}$ |
| 3 | $-130.872$ | $-109.440$ | $-99.440$ | $-111.137$ |
| 4 | $\mathbf{-158.57}$ | $-128.568$ | $-114.568$ | $-131.251$ |

The model incorporate the known measurement errors for the responses that we want to fit can be written as

$$y_i \sim \begin{cases} \mathbf{x}_i^T \boldsymbol{\beta}_1 + \epsilon_{i1}, & \text{with probability } \lambda \\ \mathbf{x}_i^T \boldsymbol{\beta}_2 + \epsilon_{i2}, & \text{with probability } 1-\lambda \end{cases} \tag{2.5}$$

$$y_i^* = y_i + \delta_i \tag{2.6}$$

where $\epsilon_{ij} \sim N(0, \sigma_j^2)$ are independent, $i = 1, \cdots, 63$ and $j = 1, 2$. $\mathbf{x}_i = (1, \log_{10}(t_i))$ where $t_i$ is the $i$th observation time since trigger (in seconds) and $y_i^*$ is logarithmic of the X-ray flux from $i$th measurement, $\log_{10}(f_i)$ and $\delta_i \sim N(0, \log_{10}^2(s_i))$, where $s_i$ is the known measurement error of the flux for the $i$th observation, and $\delta_i$ independent of $\epsilon_{ij}$.

For comparison, we also add the standard errors calculated by jackknife methods.

Table 2.6: Estimated SEs from parametric bootstrap and observed information matrix.

| Parameter | Bootstrap (SEs) | Jackknife SEs | Theoretical SEs |
|---|---|---|---|
| $\beta_{10}$ | $-6.782$ (2.438) | 0.086 | 0.209 |
| $\beta_{11}$ | $-1.007$ (0.912) | 0.032 | 0.049 |
| $\beta_{20}$ | $-5.286$ (3.561) | 0.113 | 0.147 |
| $\beta_{21}$ | $-1.552$ (1.178) | 0.040 | 0.022 |
| $\sigma_1$ | 0.792 (0.112) | 0.090 | 0.057 |
| $\sigma_2$ | 1.470 (0.600) | 0.296 | 0.413 |
| $\lambda$ | 0.601 (0.197) | 0.090 | 0.249 |

For the WLS estimates $\widetilde{\boldsymbol{\beta}}_j$ in the mixture of regressions setting, we obtain standard errors for the parameters using a parametric bootstrap with $B = 1000$, and

compare the result with variance estimates for the WLS estimators using the inverse of the observed information matrix; see Table 2.6. Based on the output, standard errors from parametric bootstrap are much larger than the inverse of observed information especially for the intercepts. For estimating mixtures-of-regressions using a resampling approach, there is usually more variability observed in the intercept estimates, thus making their standard errors of slopes much larger than expected. However, the standard errors for the variances ($\sigma_1$ and $\sigma_2$) and mixing proportion $\lambda$ are reasonable, as well as the intercepts $\beta_{11}$ and $\beta_{21}$. We can expect that if we keep increase the number of bootstrap samples $B$, the bootstrap standard errors should be closer and closer to the theoretical results, except for intercepts. However, in analysis procedure, slopes are always much more important than intercepts, as they contain more information about the data we investigated.

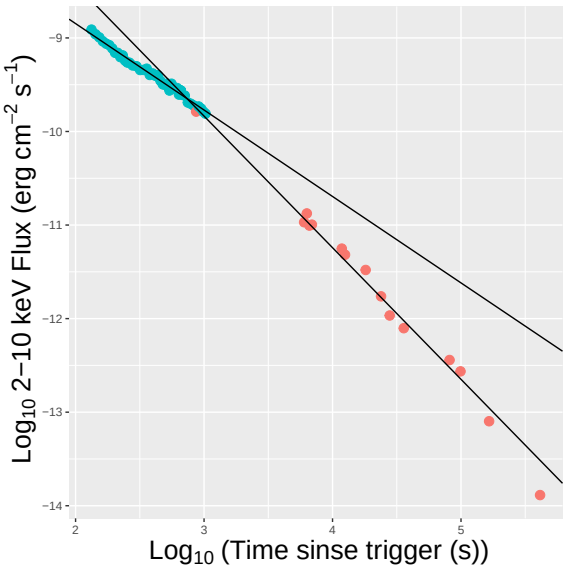


Figure 2.5: The GRB050525a data set with the estimated lines from a 2-component mixture of linear regressions model.

The estimated lines from the model is shown in Figure 2.5, different colors represent which component the plot is preferred. Based on the graph, there is clearly two distinct components: one with time $T < 2000$s, one with time $T > 2000$s. The result agrees with astronomers' assessment about PD mode and PC mode.

It is also worth investigating data within PD mode using our mixture model, since

it involves the "flaring" points as well as regular data points. We fit the data (time since trigger as predictor variable $x_i$ and X-ray flux as observed response variable $y_i^*$) with 2-component mixture model using our proposed method, the model we fit can be written as

$$y_i \sim \begin{cases} 59.023 - 0.047x_i + \epsilon_{i1}, & \text{with probability } 0.742 \\ 179.195 - 0.510x_i + \epsilon_{i2}, & \text{with probability } 0.258 \end{cases}$$

$$y_i^* = y_i + \delta_i$$

where $\epsilon_{i1} \sim N(0, 2.93^2)$ and $\epsilon_{i2} \sim N(0, 4.41^2)$ for $i = 1, \cdots, 63$.

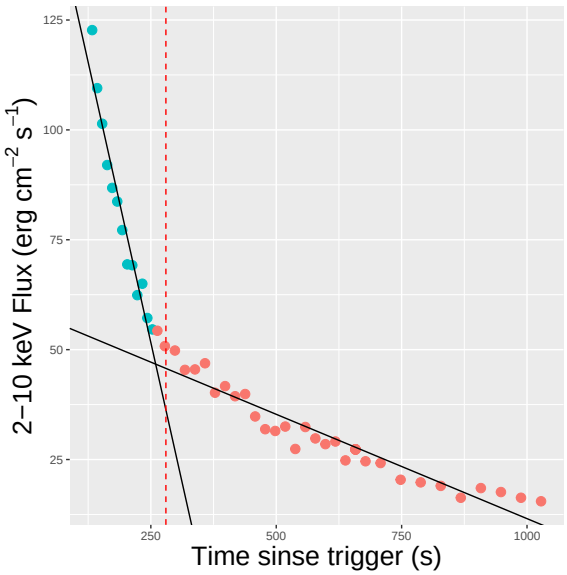


Figure 2.6: The GRB050525a data sets (PD mode) with the estimated lines from a 2-component mixture of linear regressions model.

Figure 2.6 shows the estimated lines from the 2-component mixture of linear regressions. The blue dots means those data points preferred first component, and red dots preferred the second, and the red dashed line is the break line of time before and after 280s. As we discussed before, data points with $T >$ 280s are considered as 'flaring' points, and the red dots agree with this assumption, and has a completely different component than those data sets before 280s.

## 2.5 Summary

In this chapter, we discussed the mixtures-of-regressions model with measurement error in the responses. We expand the weighted least squares method proposed by Akritas and Bershady to the mixture setting, and we use likelihood methods to compute the estimates for the parameters.

The measurement error in the response, also called intrinsic scatter in astronomy, is a problem often studied in astronomy. In this chapter, we conducted parameter estimation for a series of different settings of mixture models, including well-separated components, moderately-separated components and overlapping components. The results show our method can improve the performance of estimates, especially when measurement errors are large. A real data analysis from astronomy research is conducted and the results were evaluated. Notice for this particular model, the measurement error is considered to be a known value, which is the case for these type of gamma-fay burst data. The study of how to determine the values of the measurement error for general data problems is a separate topic that can be investigated with the help of subject matter experts.

# Chapter 3 Mixtures-of-Regressions with Measurement Error in the Predictors

Research on mixtures-of-regressions models is primarily limited to directly observed variables. However, the presence of measurement error imposes additional challenges for estimation. Mixture modeling and measurement error problems are each major areas of statistical research; however, there is limited work connecting them. One paper that does discuss mixture models in the presence of measurement errors in the predictors is Yao and Song (2015) [119]. In that paper, they consider the case when classical measurement error is present in the classic mixtures-of-regressions model. They then define the mixture likelihood and propose a generalized EM (GEM) algorithm for maximization, which provides a consistent estimate of parameters. In this section, we review this new estimation procedure accounting for the measurement error and focus on testing a specific type of model; i.e., testing for a higher-order polynomial term in one of the components, which is of practical interest. A simulation study and a real data application will be provided in Section 3.3 to illustrate the proposed estimation procedure.

## 3.1 Mixtures of Linear Regressions with Measurement Error in the Predictors

### 3.1.1 Introduction to the Method

Let $\mathcal{Z}$ be a latent class variable with $P(\mathcal{Z} = j \mid \mathbf{X} = \mathbf{x}) = \lambda_j$ for $j = 1, 2, \cdots, k$, where $\mathbf{X} = (1, X_1, \cdots, X_{p-1})^T$ is a $p$-dimensional vector of covariates, such that the first entry is a 1 to accommodate an intercept. Given $\mathcal{Z} = j$, the relationship between a uni-variate observation $Y$ and $\mathbf{X}$ is the linear regression model

$$Y = \mathbf{X}^T \boldsymbol{\beta}_j + \sigma_j \epsilon. \tag{3.1}$$

Here, $\boldsymbol{\beta}_j = (\beta_{0,j}, \cdots, \beta_{p-1,j})^T$ is the $p$-dimensional vector of regression coefficients, $\epsilon \sim N(0, 1)$, and $\sigma_j^2$ is the error variance for component $j$.

Suppose we observe the surrogate data $\mathbf{W}_1, \cdots, \mathbf{W}_n$ instead of $\mathbf{X}_1, \cdots, \mathbf{X}_n$ in the mixtures-of-linear-regressions model (3.1), where $\mathbf{X}_i = (1, X_{i1}, \cdots, X_{i,p-1})^T$ and the $\mathbf{W}_i$s are generated from an additive measurement error model $\mathbf{W}_i = \mathbf{X}_i + \mathbf{U}_i$. We further assume that the $\mathbf{X}_i$s are *i.i.d.* as $\mathbf{X}$, the error $\mathbf{U}_i$ is distributed as $N_p(0, \Sigma_{U_i})$, $i = 1, \cdots, n$, and the $\mathbf{X}_i$s and $\mathbf{U}_i$s are mutually independent.

The naïve maximum likelihood method for the model simply ignores the measurement error $\mathbf{U}$ and estimates

$$\boldsymbol{\psi} = \left(\boldsymbol{\beta}_1^T, \cdots, \boldsymbol{\beta}_k^T, \sigma_1, \cdots, \sigma_k, \lambda_1, \cdots, \lambda_{k-1}\right)$$

by maximizing the log-likelihood

$$\sum_{i=1}^{n} \log \left\{ \sum_{j=1}^{k} \frac{\lambda_j}{\sigma_j} \phi\left(\frac{y_i - \mathbf{w}_i^T \boldsymbol{\beta}_j}{\sigma_j}\right) \right\}, \tag{3.2}$$

where $\phi(\cdot)$ is the normal density for standard normal, $N(0, 1)$. Unfortunately, the naïve estimator, $\hat{\boldsymbol{\psi}}$, is not consistent, as the wrong model and likelihood function are used.

In order to incorporate measurement error in a mixtures-of-regressions setup, we need the correct conditional density of $Y$ given $\mathbf{W}$. Yao and Song (2015) [119] showed that, given $\mathcal{Z} = j$, the conditional density of $\mathbf{Y}_i$ given $\mathbf{W}_i = \mathbf{w}_i$ can be written as

$$\begin{aligned} f_j(y_i \mid \mathbf{w}_i, \boldsymbol{\theta}_j) &= \int_{\mathbb{R}^p} f(y_i \mid \mathbf{x}_i, \boldsymbol{\theta}_j) f(\mathbf{x}_i \mid \mathbf{w}_i) d\mathbf{x}_i \\ &= \frac{1}{\sigma_j} \int_{\mathbb{R}^p} \phi\left(\frac{y_i - \mathbf{w}_i^T \boldsymbol{\beta}_j}{\sigma_j}\right) f(\mathbf{x}_i \mid \mathbf{w}_i) d\mathbf{x}_i, \end{aligned} \tag{3.3}$$

where $\boldsymbol{\theta}_j = \left(\boldsymbol{\beta}_j^T, \sigma_j\right)^T$. Therefore, $Y \mid \mathbf{W} = \mathbf{w} \sim \sum_{j=1}^{k} \lambda_j f_j(y_i \mid \mathbf{w}, \boldsymbol{\theta}_j)$, and the log-likelihood function for $\boldsymbol{\psi}$ is

$$\ell(\boldsymbol{\psi}) = \log L(\boldsymbol{\psi}) = \sum_{i=1}^{n} \log \left\{ \sum_{j=1}^{k} \lambda_j f_j(y_i \mid \mathbf{w}_i, \boldsymbol{\theta}_j) \right\}. \tag{3.4}$$

Hence, we can estimate $\boldsymbol{\psi}$ by finding the maximizer of (3.4).

### 3.1.2 Estimation Algorithm

Maximizing (3.4) may be difficult as the integration is often not available in a closed form. One possibility is to evaluate it numerically, using numerical integration

(Spiegelman et al. (2000) [104]) or simulation-based methods. The latter is especially convenient for Bayesian estimation, where standard *Markov chain Monte Carlo* (MCMC) simulation methods are immediately applicable to measurement error problems. These are often formulated as algorithms where the values of $\mathbf{X}$ are regarded as missing data and the simulation involves imputing values for them. This approach has been considered for measurement error problems by, for example, Richardson and Gilks (1993) [95], Kuha (1997) [68] and Richardson et al. (2002) [97].

Besides that, Yao and Song (2015) [119] proposed the following GEM algorithm for maximization. Define the vector of component indicator $\mathbf{Z}_i = (\mathcal{Z}_{i1}, \cdots, \mathcal{Z}_{ik})^T$, where $\mathcal{Z}_{ij}$ is the indicator random variable

$$\mathcal{Z}_{ij} = \begin{cases} 1, & \text{if observation } (\mathbf{w_i}, y_i) \text{ is from the } j\text{th component;} \\ 0, & \text{otherwise.} \end{cases}$$

Then the complete log-likelihood function for $(\mathbf{w}_i^T, y_i, \mathbf{z}_i), i = 1 \cdots, n$ can be written as

$$\ell_c(\boldsymbol{\psi}) = \sum_{i=1}^{n} \sum_{j=1}^{k} z_{ij} \left\{ \log \lambda_j + \log f_j(y_i \mid \mathbf{w}_i, \boldsymbol{\theta}_j) \right\}.$$

Notice $Z_{ij} \sim \text{Bern}(\lambda_j)$, where $\text{Bern}(\lambda_j)$ is the Bernoulli distribution with rate of success $\lambda_j$. Since $\ell_c(\boldsymbol{\psi})$ is a linear function of $\mathcal{Z}_{ij}$s, in the $t$th iteration of E-step, the expectation of $\mathcal{Z}_{ij}$ is the weight of observation $i$ belonging to the $j$th component:

$$p_{ij}^{(t+1)} = \mathbb{E}\left[ \mathcal{Z}_{ij} \mid \boldsymbol{\psi}^{(t)}, \mathbf{y} \right] = \frac{\lambda_j^{(t)} f_j(y_i \mid \mathbf{w}_i, \boldsymbol{\theta}_j^{(t)})}{\sum_{j=1}^{k} \lambda_j^{(t)} f_j(y_i \mid \mathbf{w}_i, \boldsymbol{\theta}_j^{(t)})}, \tag{3.5}$$

for $i$ in $1, \cdots, n$, $j$ in $1, \cdots, k$.

In the M-step, we need to find $\boldsymbol{\psi}$ that maximizes

$$\begin{aligned} Q(\boldsymbol{\psi}) &= \mathbb{E}\left\{ \ell_c(\boldsymbol{\psi}) \mid \boldsymbol{\psi}^{(t)}, \mathbf{y} \right\} \\ &= \sum_{i=1}^{n} \sum_{j=1}^{k} p_{ij}^{(t+1)} \left\{ \log \lambda_j + \log f_j(y_i \mid \mathbf{w}_i, \boldsymbol{\theta}_j) \right\}. \end{aligned}$$

Through use of a Lagrange multiplier, it can be shown that the maximizer for $\lambda_j$ is

$$\lambda_j^{(t+1)} = \frac{1}{n} \sum_{i=1}^{n} p_{ij}^{(t+1)}. \tag{3.6}$$

The maximizer for $\boldsymbol{\theta}_j$ is

$$\boldsymbol{\theta}_j^{(t+1)} = \arg\max_{\boldsymbol{\theta}} \sum_{i=1}^{n} p_{ij}^{(t+1)} \log f_j(y_i \mid \mathbf{w}_i, \boldsymbol{\theta}_j). \tag{3.7}$$

Here, only only the $j$th component of the objective function contributes to the maximization process of the parameters from component $j$. Therefore, the maximizer for $\boldsymbol{\beta}_j$ is the solution of

$$\begin{aligned}
0 &= \frac{\partial Q(\boldsymbol{\psi})}{\partial \boldsymbol{\beta}_j} \\
&= \sum_{i=1}^{n} p_{ij}^{(t+1)} \frac{\partial \log f_j(y_i \mid \mathbf{w}_i, \boldsymbol{\theta}_j)}{\partial \boldsymbol{\beta}_j} \\
&= \sum_{i=1}^{n} p_{ij}^{(t+1)} \frac{\int \phi\left\{(y_i - \mathbf{x}^T\boldsymbol{\beta}_j)/\sigma_j\right\}(y_i - \mathbf{x}^T\boldsymbol{\beta}_j)\mathbf{x} f(\mathbf{x} \mid \mathbf{w}_i) d\mathbf{x}}{f_j(y_i|\mathbf{w}_i, \boldsymbol{\theta}_j)\sigma_j^3} \\
&\approx \sigma_j^{-2}\left\{\sum_{i=1}^{n} p_{ij}^{(t+1)} y_i \int \tau_{ij}^{(t+1)}(\mathbf{x})\mathbf{x} d\mathbf{x} - \left[\sum_{i=1}^{n} p_{ij}^{(t+1)} \int \tau_{ij}^{(t+1)}(\mathbf{x})\mathbf{x}\mathbf{x}^T d\mathbf{x}\right]\boldsymbol{\beta}_j\right\},
\end{aligned}$$

where

$$\tau_{ij}^{(t+1)}(\mathbf{x}) = f(\mathbf{x} \mid \boldsymbol{\theta}_j^{(t)}, y_i, \mathbf{w}_i) = \frac{\phi\{(y_i - \mathbf{x}^T\boldsymbol{\beta}_j^{(t)})/\sigma_j^{(t)}\} f(\mathbf{x} \mid \mathbf{w}_i)}{f_j(y_i|\mathbf{w}_i, \boldsymbol{\theta}_j^{(t)})\sigma_j^{(t)}} \tag{3.8}$$

is the conditional density of $\mathbf{x}$ given the $\mathbf{w}_i$, $y_i$, and the current estimate $\boldsymbol{\theta}_j^{(t)}$. The maximizer for $\sigma_j^2$ is the solution of

$$\begin{aligned}
0 &= \frac{\partial Q(\boldsymbol{\psi})}{\partial \sigma_j^2} = \sum_{i=1}^{n} p_{ij}^{(t+1)} \left[\frac{\int \phi\left\{(y_i - \mathbf{x}^T\boldsymbol{\beta}_j)/\sigma_j\right\}(y_i - \mathbf{x}^T\boldsymbol{\beta}_j)^2 f(\mathbf{x} \mid \mathbf{w}_i) d\mathbf{x}}{2\sigma_j^5 f_j(y_i \mid \mathbf{w}_i, \boldsymbol{\theta}_j)} - \frac{1}{2\sigma_j^2}\right] \\
&\approx (2\sigma_j^4)^{-1} \sum_{i=1}^{n} p_{ij}^{(t+1)} \left[\int \tau_{ij}^{(t+1)}(\mathbf{x})(y_i - \mathbf{x}^T\boldsymbol{\beta}_j^{(t+1)})^2 d\mathbf{x} - \sigma_j^2\right].
\end{aligned}$$

Based on the above approximations, we can update $\boldsymbol{\beta}_j$ and $\sigma_j$ by

$$\boldsymbol{\beta}_j^{(t+1)} = \left\{\sum_{i=1}^{n} p_{ij}^{(t+1)} \int \tau_{ij}^{(t+1)}(\mathbf{x})\mathbf{x}\mathbf{x}^T d\mathbf{x}\right\}^{-1} \left\{\sum_{i=1}^{n} p_{ij}^{(t+1)} y_i \int \tau_{ij}^{(t+1)}(\mathbf{x})\mathbf{x} d\mathbf{x}\right\} \tag{3.9}$$

and

$$\sigma_j^{(t+1)} = \left[\left\{\sum_{i=1}^{n} p_{ij}^{(t+1)}\right\}^{-1} \sum_{i=1}^{n} p_{ij}^{(t+1)} \int \tau_{ij}^{(t+1)}(\mathbf{x})(y_i - \mathbf{x}^T\boldsymbol{\beta}_j^{(t+1)})^2 d\mathbf{x}\right]^{1/2}, \tag{3.10}$$

respectively. If we assume the $\sigma_j$s are equal, that is, $\sigma_1 = \cdots = \sigma_k = \sigma$, then we can update $\sigma$ by

$$\sigma^{(t+1)} = \left[ n^{-1} \sum_{i=1}^{n} \sum_{j=1}^{k} p_{ij}^{(t+1)} \int \tau_{ij}^{(t+1)}(\mathbf{x})(y_i - \mathbf{x}^T \boldsymbol{\beta}_j^{(t+1)})^2 d\mathbf{x} \right]^{1/2}. \tag{3.11}$$

Based on the above, we next provide a compact description of the proposed GEM algorithm of Yao and Song (2015) [119] to estimate $\boldsymbol{\psi}$:

**Algorithm 3.1** GEM Algorithm

Starting with $\boldsymbol{\psi}^{(0)}$ at the $(t+1)$th iteration, $t = 0, 1, \cdots$,
(a) (E-Step) Calculate component membership probabilities $p_{ij}^{(t+1)}$s using (3.5).
(b) (M-Step) Update $\lambda_j$s, $\boldsymbol{\beta}_j$s and $\sigma_j$s based on (3.6), (3.9) and (3.10), respectively.
(c) Iterate until a specified stopping criterion is attained. Stopping criteria were discussed in Chapter 1 (see Subsection 1.1.3).

### 3.1.3 Estimating Variance of Measurement Errors

In the estimating process, we need to know the measurement error covariance matrix of the distribution of $\mathbf{U}_i$. The most common way of estimating it is based on partially replicated observations (Carroll et al. (2006) [94]). For this model, $J_i \geq 2$ replicate measurements are necessary for each subject in order to identify the error variances.

Following Carroll et al. (2006) [94], for each predictor value $i$, suppose the error model is

$$\mathbf{W}_{ih} = \mathbf{X}_i + \mathbf{U}_{ih},$$

where $\mathbf{U}_{ih}$, $h = 1, \cdots, J_i$, follows $N(0, \boldsymbol{\Sigma}_u)$, independent of $\mathbf{X}_i$ and $\mathbf{Y}_i$, with $\boldsymbol{\Sigma}_u$ unknown. With replicate measurements, the best measurement of $\mathbf{X}_i$ is the mean $\bar{\mathbf{W}}_{i\cdot} = J_i^{-1} \sum_{h=1}^{J_i} \mathbf{W}_{ih}$, where we define the so-called *naïve* estimation procedure as doing the usual, non-measurement-error analysis of data $\left\{ (\bar{\mathbf{W}}_{1\cdot}, Y_1), \cdots, (\bar{\mathbf{W}}_{n\cdot}, Y_n) \right\}$. Replication enables us to estimate the measurement error covariance matrix $\boldsymbol{\Sigma}_u$ by the usual components of variance analysis as follows:

$$\hat{\boldsymbol{\Sigma}}_u = \sum_{i=1}^{n} \sum_{h=1}^{J_i} (\mathbf{W}_{ih} - \bar{\mathbf{W}}_{i\cdot})(\mathbf{W}_{ih} - \bar{\mathbf{W}}_{i\cdot})^T / \sum_{i=1}^{n} (J_i - 1). \tag{3.12}$$

In linear regression, if there are no replicates ($J_i \equiv 1$) but an external estimate $\hat{\boldsymbol{\Sigma}}_u$ is available, or if there are exactly two replicates (2), in which case $\hat{\Sigma}_U$ is half the sample covariance matrix of the differences $\mathbf{W}_{i1} - \mathbf{W}_{i2}$, regression calibration reproduces the classical method-of-moments estimates.

When the number of replicates is not constant, the algorithm can be shown to produce consistent estimates in linear regression and (approximately) in logistic regression. For log-linear mean models, the intercept is biased, so one should add a dummy variable to the regression indicating whether or not an observation is replicated.

### 3.1.4 Model Selection Criteria

We've discussed information criteria in model selection in Chapter 1. We next expand to the setting with measurement error. Based on $\hat{\boldsymbol{\psi}}$ obtained using Algorithm 3.1, the observed log-likelihood function can be written as

$$\ell(\hat{\boldsymbol{\psi}}) = \sum_{i=1}^{n} \log \sum_{j=1}^{k} \left\{ \hat{\lambda} f_j(y_i \mid \mathbf{w}_i, \hat{\boldsymbol{\theta}}_j) \right\} \tag{3.13}$$

for $i = 1, \cdots, n$, $j = 1, \cdots, k$. Then, the four model selection criteria can be written as follows:

$$\text{AIC} = -2\ell(\hat{\boldsymbol{\psi}}) + 2d \tag{3.14}$$

$$\text{BIC} = -2\ell(\hat{\boldsymbol{\psi}}) + d\log(n) \tag{3.15}$$

$$\text{ICL} = \text{BIC} + 2\left( -\sum_{i=1}^{n}\sum_{j=1}^{k} \hat{p}_{ij} \log \hat{p}_{ij} \right) \tag{3.16}$$

$$\text{cAIC} = -2\ell(\hat{\boldsymbol{\psi}}) + d(\log(n) + 1), \tag{3.17}$$

where $d$ is the number of parameters in the mixture setting. These values can be calculated for a reasonable range of components and mixture settings, and the minimum of these values (for each criterion) corresponds to the model selected by that

criterion. Issues exist with the underlying asymptotic theory when dealing with the model selection problem for determining the number of components, which is due to the breakdown of the regularity conditions with a mixture setting. Regardless of the theoretical problems, model selection criteria typically perform well for determining the correct model.

## 3.2 Covariate Measurement Error in Mixtures of Quadratic Regression

Quadratic regression models are one of the simplest ways to explore the presence of nonlinearities. Suppose we are interested in the $k$-component mixtures of quadratic regression model for a response variable $Y_i$, $i = 1, \cdots, n$. Conditioning on $j$th component

$$Y_i = \beta_{0j} + \beta_{1j} X_i + \beta_{2j} X_i^2 + \sigma_i \epsilon \tag{3.18}$$

$$= \mathbf{X}_i^T \boldsymbol{\beta}_j + \sigma_i \epsilon \tag{3.19}$$

where $\boldsymbol{\beta}_j = (\beta_{0j}, \beta_{1j}, \beta_{2j})^T$ is the parameter vector of $j$th component, $\mathbf{X}_i = (1, X_i, X_i^2)$ is the vector of predictor on the $i$th observation, and $\sigma_i \epsilon$ is independent random variables with $\epsilon \sim N(0, 1)$.

Instead of observing $\mathbf{x}_i$ directly, we observe $\mathbf{W}_i = \mathbf{x}_i + \mathbf{u}_i$, with $\mathbb{E}(\mathbf{u}_i \mid \mathbf{x}_i) = \mathbf{0}$ and $\mathrm{Var}(\mathbf{u}_i \mid \mathbf{x}_i) = \boldsymbol{\Sigma}_u$. The fact that $\mathbb{E}(\mathbf{u}_i \mid \mathbf{x}_i) = \mathbf{0}$ means the measurement error is additive, and equivalently, $\mathbf{W}_i$ is unbiased for $\mathbf{x}_i$. We can derive the estimate the measurement error covariance matrix $\boldsymbol{\Sigma}_u$ using method in Subsection 3.1.3. In this situation, the curvature of the estimated function will be less steep than in the true model, and hence measurement error will tend to hide the presence of a nonlinearity. Kuha and Temple (2003) [69] examine the effects of measurement error on quadratic regressions and discuss ways to conduct a sensitivity analysis. These will be discussed below.

### 3.2.1 Estimating Methods

There are multiple ways to estimate the parameter $\boldsymbol{\beta}$ for measurement error in quadratic regression model with a non-mixture setting. For instance, Kuha and Tem-

ple (2003) [69] considered two types of adjusted estimators of $\boldsymbol{\beta}$, *regression calibration* (RC) estimators and method-of-moments (or *corrected score*) estimators.

In structural estimation, $X$ is regarded as a random variable. Suppose we fully specify the distribution of $\mathbf{X}$, as well as $Y$ given $\mathbf{X}$ and $\mathbf{W}$ given $\mathbf{X}$, and thus also for $\mathbf{X}$ given $\mathbf{W}$. The idea of simple regression calibration (SRC) is to replace the true predictors $\mathbf{X}$ by their means given the measured variable $\mathbf{W}$, and fit the original model for $Y$ given these conditional means. Expanded regression calibration (ERC) improves this approximation further by adding terms depending on the variance of $\mathbf{X}$ given $\mathbf{W}$.

In a non-structural setting, when there is no measurement error, estimates for the parameters $\boldsymbol{\theta}$ of a model for $Y$ given $\mathbf{X}$ are obtained by solving estimating equations $\sum_i \Psi(\boldsymbol{\theta} \mid Y_i, \mathbf{X}_i) = 0$ for some function $\Psi$. The score function $\Psi(\boldsymbol{\theta} \mid Y_i, \mathbf{X}_i) = \partial \log f(Y_i \mid \mathbf{X}_i, \boldsymbol{\theta})/\partial\boldsymbol{\theta}$ can be used to obtain maximum likelihood estimates. When $\mathbf{X}_i$ are measured with error, the estimating equations should depend only on $Y_i$ and $\mathbf{X}_i$. We can define corrected score functions $\Psi^*(\boldsymbol{\theta} \mid Y_i, \mathbf{X}_i)$ for which $\mathbb{E}\left[\Psi^*(\boldsymbol{\theta} \mid Y_i, \mathbf{W}_i) \mid \mathbf{X}_i\right] = \Psi(\boldsymbol{\theta} \mid Y_i, \mathbf{X}_i)$ for all $Y_i$, $\mathbf{X}_i$ and $\boldsymbol{\theta}$. Such $\Psi^*$ are then conditionally unbiased estimating functions for $\boldsymbol{\theta}$, and their solutions are consistent estimates. The method is functional, because the argument is conditional on $\mathbf{X}_i$. In the case which is relatively straightforward for additive measurement error model, especially when $\mathbf{u}_i$ is normally distributed, corrected score estimates are also known as method-of moments estimates. Here we consider the approach described in the previous section, and also incorporate the mixture setting introduced in Subsection 3.1.1. We further discuss a specific testing problem that can be used in a mixtures-of-regression model with measurement error in the predictor.

### 3.2.2 Bootstrap Estimator for the Standard Errors

Inferences in measurement error models can be challenging for a variety of reasons. While analytical standard errors are available for some methods, these usually involve some underlying assumptions. For instance, Wald-type confidence intervals based on these standard errors rely on approximate normality and unbiasedness of the estima-

tor. An additional concern is that the corrected estimators are always biased; rather, most are either consistent, or approximately consistent under appropriate conditions. A method that can deal with potential bias in either the corrected estimators or naïve estimators, which ignore the measurement error is desirable.

One way for mitigating the impact of these issues is the bootstrap method, which has received limited attention in the measurement error context. Similar to Chapter 2, we introduce an algorithm for a parametric bootstrap in the mixtures-of-regressions model when accounting for measurement error in the predictor.

---

**Algorithm 3.2** Parametric Bootstrap for Standard Errors

---

(a) Find the maximum likelihood estimate $\hat{\boldsymbol{\theta}}_j = (\hat{\boldsymbol{\beta}}_j, \hat{\sigma}_j^2, \hat{\lambda}_j)^T$, $j = 1, \cdots, k$ by implementing Algorithm 3.1 based on the observed data $\{(\mathbf{w}_1, y_1), \cdots, (\mathbf{w}_n, y_n)\}$.

(b) Generate a bootstrap sample of size $n$ from

$$Y_i^* \sim \sum_{j=1}^{k} \hat{\lambda}_j N\left(\mathbf{x}^T \hat{\boldsymbol{\beta}}_j, \hat{\sigma}_j^2\right),$$

with the observed surrogates. Call this bootstrap sample $\{(\mathbf{w}_1, y_1^*), \cdots, (\mathbf{w}_n, y_n^*)\}$.

(c) Find the estimate $\widetilde{\boldsymbol{\theta}}$ by implementing Algorithm 3.1 on $(\mathbf{w}_1, y_1^*), \cdots, (\mathbf{w}_n, y_n^*)$.

(d) Repeat steps (b) - (d) $B$ times to generate the bootstrap sampling distribution $\widetilde{\boldsymbol{\theta}}^{(1)}, \widetilde{\boldsymbol{\theta}}^{(2)}, \cdots, \widetilde{\boldsymbol{\theta}}^{(B)}$.

(e) Compute the standard error of $\hat{\boldsymbol{\theta}}^{(1)}, \hat{\boldsymbol{\theta}}^{(2)}, \cdots, \hat{\boldsymbol{\theta}}^{(B)}$.

---

One problem with this parametric bootstrap method is, in real data analysis, we don't know the true predictor values $\mathbf{x}$s, which means we can't compute bootstrap response $y^*$ based on this parametric approach. One solution is to consider a non-parametric, or semi-parametric alternative. Turner (2000) [112] introduced a non-parametric/semi-parametric bootstrap method, which requires re-sampling the residuals from the null model. Here, we expand Turner's model-based bootstrap method when correcting for additive measurement error in regression with replicate measures of the unobserved true values.

With this approach, we avoid the assumption of knowing the true predictor vari-

**Algorithm 3.3** Semi-parametric Bootstrap Standard Errors

(a) Fit the model to the observed data $\{(\mathbf{w}_1, y_1), \cdots, (\mathbf{w}_n, y_n)\}$, record the posterior membership probabilities $\hat{p}_{ij}$ and fitted residuals $r_{ij} = y_i - \hat{y}_{ij}$, where $\hat{y}_{ij}$ is the fitted response of $i$th observation for the $j$th component.

(b) Generate a semi-parametric bootstrap sample $\{(\mathbf{w}_1, y_1^*), \cdots, (\mathbf{w}_n, y_n^*)\}$ as follows:

1. Sample $n$ values with replacement from $\{1, 2, \cdots, n\}$, call these indices $i^*$;
2. For each $i^*$, generate $z_{i^*} \sim \text{Multinomial}\left(1, \hat{p}_{i^*\cdot}\right)$, $z_{i^*} = \{1, \cdots, n\}$ represents which component $i^*$ belongs to;
3. For each $\mathbf{w}_i$, $i = 1, \cdots, n$, select which component to generate from based on $z_{i^*}$;
4. Generate a residual $r^*_{i,z_i^*} = r_{i^*,z_i^*}$.
5. Define $y_i^* = \hat{y}_{i,z_{i^*}} + r^*_{i,z_i^*}$.

(c) Find the estimate $\hat{\boldsymbol{\theta}}$ by implementing Algorithm 3.1 on $(\mathbf{w}_1, y_1^*), \cdots, (\mathbf{w}_n, y_n^*)$.

(e) Repeat steps (b) and (c) $B$ times to generate the bootstrap sampling distribution $\hat{\boldsymbol{\theta}}^{(1)}, \hat{\boldsymbol{\theta}}^{(2)}, \cdots, \hat{\boldsymbol{\theta}}^{(B)}$.

(f) Compute the standard error of $\hat{\boldsymbol{\theta}}^{(1)}, \hat{\boldsymbol{\theta}}^{(2)}, \cdots, \hat{\boldsymbol{\theta}}^{(B)}$.

ables, however, the semi-parametric method may lead to instability when estimating the parameters. We will see some examples of both approaches later to see highlight the advantages and disadvantages of both methods.

### 3.2.3 Likelihood Ratio Test

Consider the two-component mixture model

$$Y_i \sim \lambda N\left(\mathbf{x}_i^T \boldsymbol{\beta_1}, \sigma_1^2\right) + (1-\lambda) N\left(\mathbf{x}_i^T \boldsymbol{\beta_2}, \sigma_2^2\right),$$

where $\boldsymbol{\beta}_1 = (\beta_{10}, \beta_{11}, \cdots, \beta_{1p}, 0)^T$ is a $(p+1)$-dimensional parameter vector, $\boldsymbol{\beta}_2 = (\beta_{20}, \beta_{21}, \cdots, \beta_{2p}, \beta_{2,p+1})^T$ is a $(p+2)$-dimensional parameter vector, $U_i \sim N(0, \sigma_u^2)$, the distribution of $u_i$s is known, and $\mathbf{x}_i = (1, x_i, x_i^2, \cdots, x_i^p)$ such that $x_i$ is predictor variable for $i$th observation. For each predictor variable, consider the measurement error model $w_i = x_i + u_i$. The observed variables are $\mathbf{w}_i = (1, w_i, w_i^2, \cdots, w_i^p)$.

We are interested in testing for a quadratic effect; i.e., if it is appropriate to keep $\beta_{2,p+1}$ term in the model. The hypothesis test of interest is, thus,

$$\begin{aligned} H_0 &: \beta_{2,p+1} = 0 \\ H_1 &: \beta_{2,p+1} \neq 0. \end{aligned} \tag{3.20}$$

Here, we might consider constructing the traditional likelihood ratio test (LRT) statistic. Given $\mathcal{Z} = j$, $j = 1, 2$, the conditional density of $Y$ given $\mathbf{W} = \mathbf{w}$ is

$$f_j(y \mid \mathbf{w}, \boldsymbol{\theta}_j) = \frac{1}{\sigma_j} \int \phi\left\{(y - \mathbf{x}^T \boldsymbol{\beta}_j)/\sigma_j\right\} f(\mathbf{x} \mid \mathbf{w}) d\mathbf{x},$$

where $\boldsymbol{\theta}_j = (\boldsymbol{\beta}_j^T, \sigma_j)^T$. Therefore $Y \mid \mathbf{W} \sim \lambda f_1(y \mid \mathbf{w}, \boldsymbol{\theta}_1) + (1 - \lambda) f_2(y \mid \mathbf{w}, \boldsymbol{\theta}_2)$, and the likelihood function for the parameter vector $\boldsymbol{\theta}^T = (\boldsymbol{\theta}_1^T, \boldsymbol{\theta}_2^T)$ is

$$L(\boldsymbol{\theta}) = \prod_{i=1}^{n} \left\{\lambda f_1(y \mid \mathbf{w}, \boldsymbol{\theta}_1) + (1 - \lambda) f_2(y \mid \mathbf{w}, \boldsymbol{\theta}_2)\right\}.$$

Note that under either hypothesis, the distribution of the data is fully specified. Let $\boldsymbol{\theta}_0$ is the parameter space under null hypothesis and $\boldsymbol{\theta}_A$ the parameter space under alternative hypothesis. Then the likelihood ratio test based on the likelihood ratio, can be written as

$$\Lambda(y) = \frac{L(\boldsymbol{\theta}_0)}{L(\boldsymbol{\theta}_A)}.$$

The test statistic

$$-2\log(\Lambda) = 2\left\{\log L(\hat{\boldsymbol{\theta}}_A) - \log L(\hat{\boldsymbol{\theta}}_0)\right\} \tag{3.21}$$

for a nested model will be asymptotically chi-squared with degrees of freedom equal to the difference in dimensionality of $\boldsymbol{\theta}_1$ and $\boldsymbol{\theta}_0$, when $H_0$ is true. This means we can compute the likelihood ratio $\Lambda$ for the data and compare $-2\log(\Lambda)$ to the $\chi^2$ value corresponding to a desired statistical significance as an approximate statistical test.

The LRT statistic is straightforward, however, for mixture models we have to also consider the asymptotic condition for different settings of parameters. Another way to approach the test in (3.20) is to bootstrap (parametrically or semi-parametrically)

the LRT statistic as proposed in McLachlan (1987) [77]. The algorithm is an attempt to approximate the null distribution of the LRT statistic values given in (3.21), thus avoiding the regularity conditions for asymptotic theory. The algorithm is as follows:

---

**Algorithm 3.4** Parametric Bootstrap Likelihood Ratio Test (BLRT)

---

(a) Fit both the null model and alternative model to the observed data $\{(\mathbf{w}_1, y_1), \cdots, (\mathbf{w}_n, y_n)\}$, which leads to the estimates $\hat{\boldsymbol{\theta}}_0$ and $\hat{\boldsymbol{\theta}}_1$, respectively.

(b) Calculate the (observed) log-likelihood ratio statistic in Equation (3.21). Denote this value by $\Lambda_{obs}$.

(c) Simulate a data set of size $n$ from the null distribution ($\beta_{2,p+1} = 0$). Call this sample $\{(\mathbf{w}_1, y_1^*), \cdots, (\mathbf{w}_n, y_n^*)\}$.

(d) Fit both the null model and alternative model to the simulated data and calculate the corresponding bootstrap log-likelihood ratio statistic.

(e) Repeat steps (c) and (d) $B$ times to generate the bootstrap sampling distribution of likelihood ratio statistic $\Lambda^{(1)}, \Lambda^{(2)}, \cdots, \Lambda^{(B)}$.

(e) Compute the bootstrap $p$-values as

$$p_B = \sum_{i=1}^{B} I\left\{\Lambda^{(i)} \geq \Lambda_{obs}\right\}.$$

---

We then obtain $p_B$ for this test and if it is lower than some significance level $\alpha$ (say, 0.05), we claim statistical significance in favor of $H_1$.

## 3.3 Numerical Studies

In this section, the finite sampling behavior of the proposed mixture-of-regression estimates with measurement error is studied using Monte Carlo simulation with different settings as well as a real data example.

### 3.3.1 Simulated Example

We are interested in assessing the presence of a quadratic effect for one of the components in a mixture of regressions setting. Suppose our data are assumed to follow the 2 - component mixture-of-regressions model

$$y_i \sim \lambda N(\mathbf{x}_{i(1)}^T \boldsymbol{\beta}_1, \sigma_1^2) + (1-\lambda) N(\mathbf{x}_{i(2)}^T \boldsymbol{\beta}_2, \sigma_2^2),$$

where $\mathbf{x}_{i(1)} = (1, x_i)^T$ and $\mathbf{x}_{i(2)} = (1, x_i, x_i^2)^T$. Instead of observing $x_i$s directly, the surrogate, $w_i$, is given by the classical measurement error model

$$w_i = x_i + u_i,$$

where $u_i$ and $x_i$ are independent, and $u_i$ follows a normal distribution $N(0, \sigma_u^2)$.

We consider three different simulation conditions: well-separated components, moderately-separated components and overlapping components. For each simulation condition, we randomly generated $B = 1000$ datasets, each of size $n = 200$ and 350, estimated the corresponding model (mixture of linear regression vs. mixture of one linear and on quadratic regression) using an EM algorithm. Then the output of the EM algorithm is used to calculate the four model selection criteria discussed in Subsection 3.1.4. We then report the percentage of times each model selection criterion selected the appropriate model for our 1000 simulated data sets.

In order to avoid the possible bias created by different starting values among replications or label switching issues, see, for example, Stephens (2000b) [108], we use the true initial values for parameters in the GEM algorithm, which follows Bordes, Chauveau and Vandekerkhove's work in 2007 [13].

In order to estimate the variance of measurement error $\sigma_u^2$, for each predictor value $x_i$, we randomly generate $r = 3$ different measurement errors $u_{i1}, u_{i2}$ and $u_{i3}$ and compute the estimated variance based on the method discussed in Subsection 3.1.3. The observed predictor $\bar{w}_i$ is given by the average of three observations, that is,

$$\bar{w}_i = \frac{1}{3}\left[(x_i + u_{i1}) + (x_i + u_{i2}) + (x_i + u_{i3})\right].$$

### I. Well-separated

We generated the i.i.d. data $(w_i, y_i)$, $i = 1, \cdots, n$ from the model

$$Y_i \sim 0.5N\left(10 - 3x_i, \sigma_1^2\right) + 0.5N\left(-4 + x_i + 3x_i^2, \sigma_2^2\right)$$

$$w_i = x_i + u_i,$$

where $x_i \sim N(1, 1)$, $u_i \sim N(0, 0.01)$ for $i = 1, 2, \cdots n$, $\sigma_1 = 1$ and $\sigma_2 = 2$.

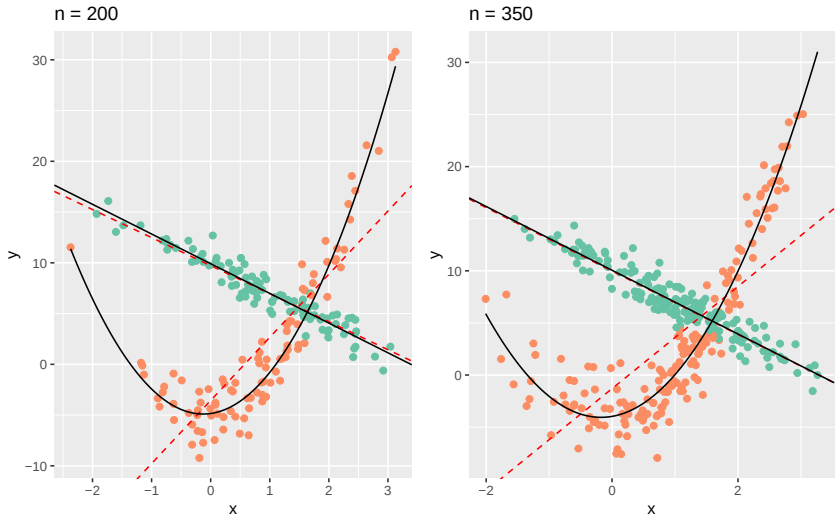


Figure 3.1: Scatterplots and fitted lines for well-separated case.

Figure 3.1 are the scatterplots and fitted lines of well-separated components case, when sample size is 200 and 350. The red dashed lines are the estimated lines when using a mixture of simple linear regressions, while black solid lines are the estimated lines for the model such that one component has a quadratic term. Different colors represent the different components the data points belongs to, according to mixtures-of-regressions with quadratic term in one component. Based on the graph, we can see a clear curve to one of the component (orange dots), and the black solid lines can represent the behavior of data set much better than the red dashed lines — since the two components are well-separated, they fit well for both linear and quadratic settings. On the other hand, when sample size increases, more data points make it easier to fit the model, and should have a better performance comparing to the smaller sample size case.

### II. Moderately-separated

We next generated the i.i.d. data $(w_i, y_i)$, $i = 1, \cdots, n$ from the model

$$Y_i \sim 0.2N\left(5 - x_i, \sigma_1^2\right) + 0.8N\left(-3 + 2x_i + x_i^2, \sigma_2^2\right)$$

$$w_i = x_i + u_i,$$

where $x_i \sim N(-1, 1)$, $u_i \sim N(0, 0.01)$, $\sigma_1 = 1$ and $\sigma_2 = 2$.

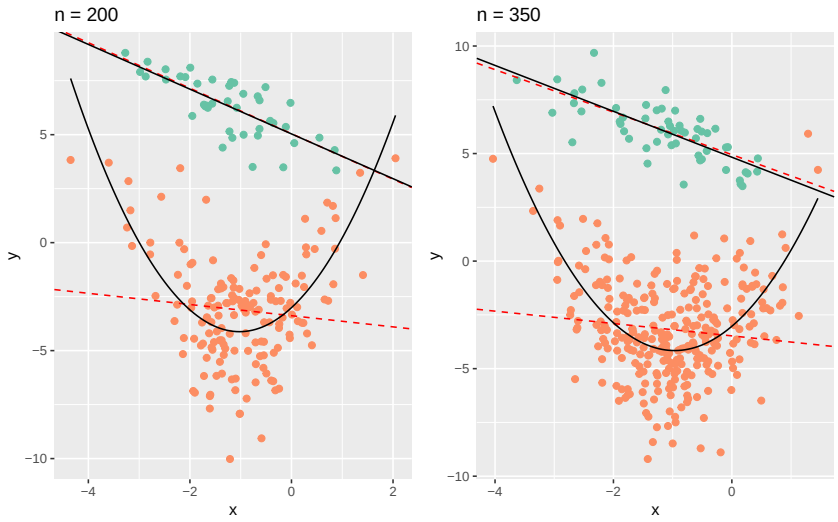


Figure 3.2: Scatterplots and fitted lines for moderately-separated case.

Similarly, Figure 3.2 are the scatterplots and fitted lines of moderately-separated components case, when sample size is 200 and 350. The red dashed lines are the estimated lines when using a mixture of simple linear regressions, while black solid lines are the estimated lines for the model such that one component has a quadratic term. Unlike the well-separated case, moderately-separated components have some dataset mixing present, making it more difficult to determine which component each data point belongs. For example, when sample size $n = 350$, there are two data points on the top right, which are supposed to belong ton the linear component, however, they are labeled as quadratic component, based on our estimating method.

According to Figure 3.2, it can be seen that one of the component is linear (the top data points), and both methods (linear or quadratic component) can capture

the characteristic of this component pretty well; for the other component, which is supposed to be quadratic, the linear method fail to predict the behavior, while the quadratic method behaves much better, if not perfectly. We can also see, when increasing the sample size of the data for some case, the quadratic characteristic can be reduced, and makes it harder to detect the quadratic in the data. It is always a challenge to determine whether there is a quadratic term in the model.

### III. Overlapping

Finally, we generated the i.i.d. data $(w_i, y_i)$, $i = 1, \cdots, n$ from the model

$$Y_i \sim 0.7N\left(5 + x_i, \sigma_1^2\right) + 0.3N\left(1 + 2x_i + x_i^2, \sigma_2^2\right)$$

$$w_i = x_i + u_i,$$

where $x_i \sim N(0, 1)$, $u_i \sim N(0, 0.01)$, $\sigma_1 = 1$ and $\sigma_2 = 2$.

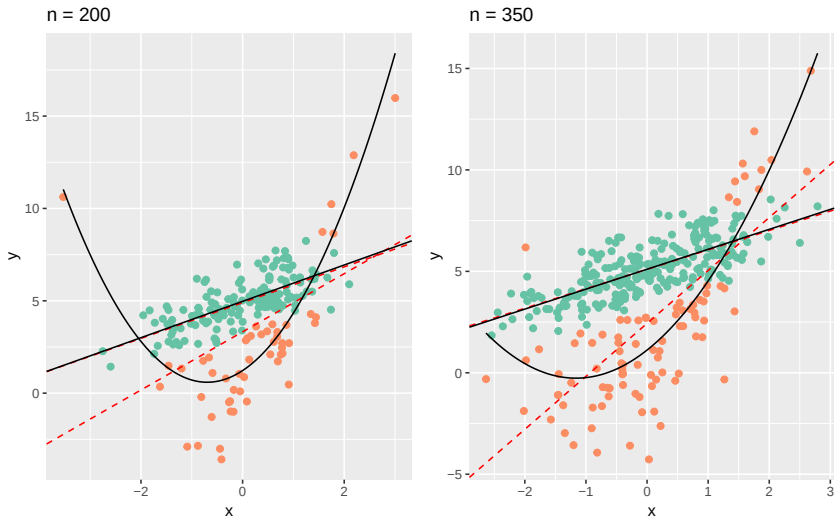


Figure 3.3: Scatterplots and fitted lines for overlapping case.

Figure 3.3 are the scatterplots and fitted lines of overlapping components case, when sample size is 200 and 350. Since the two components are heavily overlapping, the estimation is much more difficult compared to the well-separated and moderately-separated cases. From the scatter plots, the overlapping structure hides the characteristic of quadratic form, and it is harder to determine whether quadratic is the

better choice for this setting. To see the performance of quadratic versus the linear component, we access the model selection criteria. For each simulation situation, we computed the values of the four model selection criteria: AIC, BIC, ICL and cAIC, and recorded the times each situation selected the correct model structure.

Table 3.1: Percentage of times each model selection criterion selected the correct model.

| $n$ | AIC | BIC | ICL | cAIC |
|---|---|---|---|---|
| Well-Separated Components | | | | |
| 200 | 100% | 100% | 100% | 100% |
| 350 | 100% | 100% | 100% | 100% |
| Moderately-Separated Components | | | | |
| 200 | 99.8% | 99.7% | 99.7% | 99.7% |
| 350 | 100% | 100% | 100% | 100% |
| Overlapping Components | | | | |
| 200 | 95% | 83.6% | 77.3% | 83.7% |
| 350 | 99.9% | 97% | 95.5% | 97% |

Table 3.1 shows the percentage of times each model selection criterion chose the correct model. Overall, the model selection criteria performed well with all three model settings and selects the correct model a reasonable percentage of the time for all the cases, thus suggesting the use of model selection criteria for problems like this is a viable strategy.

Among the model selection criteria, AIC typically overestimates while BIC, ICL and cAIC are good indicators for the fit of a mixture model (Wedel and DeSarbo (1994) [114] and McLachlan (1987) [77]). Overall, the performance with a data set having a larger sample size ($n = 350$) appears better than that for a smaller sample size ($n = 200$), when well-separated components and moderately-separated components capture the correct model 100% of the time and overlapping components more than 95% of the time, although it seems hard to identify the components from the scatter plots of overlapping components.

The model selection criteria also have good performance for well-separated components and moderately-separated model with a smaller sample size ($n = 200$), with a 100% and about 99.7% chance of selecting the correct model, respectively. However,

if the sample size is not large enough, the performance of model selection criteria is not as good for overlapping components model, as to be expected. It can only choose the correct model for as low as 77.3% of the time, with ICL and around 83% for both BIC and cAIC. It suggests that when doing the model selection procedure with not so well-separated data, one needs to heavily scrutinize the results, and possibly investigate other techniques for determining the best model to use.

### 3.3.2 MSE and Relative Efficiency

According to the model selection criteria, it is appropriate to use the true model from which we simulated. To test the performance of the estimation method for the presence of a quadratic term, we estimated the mixture of regression parameters $\left(\boldsymbol{\beta}_1^T, \boldsymbol{\beta}_2^T, \lambda, \sigma_1, \sigma_2\right)$ by the proposed method for each simulated data set, and compared the results with the so-called 'naïve' method, which simply ignore the measurement error. The performance of the proposed method under different conditions is assessed by the *mean squared error* (MSE); i.e.,

$$\text{MSE}(\hat{\boldsymbol{\theta}}) = \frac{1}{B}\sum_{t=1}^{B}(\hat{\boldsymbol{\theta}}^{(t)} - \boldsymbol{\theta})^2$$

where $\hat{\boldsymbol{\theta}}^{(t)}$ is the estimate of the parameter $\boldsymbol{\theta}$ based on $t$th replication and $\boldsymbol{\theta}$ is the true value. The relative efficiency of the MSE for the naïve method versus the proposed method is also recorded for all the parameters.

In order to better see the difference between the naïve method and proposed method, for each simulated data set we added a larger amount of measurement error ($u \sim N(0, 0.5^2)$). Table 3.2 shows the MSEs and relative efficiencies (in parentheses) for our simulated data sets.

The MSE measures the accuracy of the method for estimating the unknown parameter; the smaller the value, the better the performance. When the ratio of the MSE is greater than 1, it means our proposed method behaves better than the naïve method. From Table 3.2, we can see that the proposed method, which incorporates the measurement error, for most parameters of different cases, has a relative efficiency greater than 1. This implies it works relatively better than the naïve method.

Table 3.2: Ratio of the MSEs of naïve method to proposed estimators.

| $n$ | $\beta_{10}$ | $\beta_{10}$ | $\beta_{20}$ | $\beta_{21}$ | $\beta_{22}$ | $\lambda$ | $\sigma_1$ | $\sigma_2$ |
|---|---|---|---|---|---|---|---|---|
| Well-Separated Components | | | | | | | | |
| 200 | 0.505 | 0.206 | 1.810 | 0.749 | 0.9169 | 0.003 | 0.447 | 2.221 |
| | (2.611) | (3.769) | (1.638) | (1.844) | (2.006) | (1.301) | (1.741) | (1.706) |
| 350 | 0.576 | 0.245 | 2.130 | 0.354 | 0.387 | 0.001 | 0.286 | 3.616 |
| | (1.510) | (2.575) | (1.309) | (1.438) | (2.651) | (3.575) | (2.724) | (1.676) |
| Moderately-Separated Components | | | | | | | | |
| 200 | 2.747 | 0.061 | 1.500 | 18.357 | 1.763 | 0.640 | 0.971 | 0.089 |
| | (1.201) | (2.260) | (1.314) | (0.877) | (1.194) | (1.002) | (0.628) | (0.888) |
| 350 | 1.013 | 0.029 | 1.650 | 19.786 | 1.533 | 0.669 | 1.598 | 0.043 |
| | (1.354) | (1.588) | (1.179) | (0.878) | (1.239) | (0.999) | (0.852) | (1.583) |
| Overlapping Components | | | | | | | | |
| 200 | 1.309 | 0.028 | 7.050 | 0.455 | 0.184 | 0.056 | 0.100 | 0.590 |
| | (0.971) | (1.264) | (0.962) | (1.375) | (1.493) | (1.006) | (1.004) | (0.955) |
| 350 | 1.043 | 0.009 | 3.233 | 0.368 | 0.108 | 0.014 | 0.021 | 1.058 |
| | (0.939) | (2.320) | (1.027) | (1.367) | (1.374) | (1.037) | (1.334) | (0.865) |

As we can see, when the two components are well-separated, the proposed method is much better in estimating $\boldsymbol{\beta}$s, and has a relative smaller MSE for all parameters; for moderately-separated case, most parameters behave pretty well, except for $\beta_{21}$, which has a relatively larger MSE, and the relative efficiencies for this parameter and also the variance terms. For the overlapping components case, the MSEs are also not large, but the difference between the naïve method and proposed is not much. Overall, when the structure of the model becomes more complicated, it gets harder to estimate the parameters, and for some scenarios, it may cause inaccurate estimates for some of the parameters.

On the other hand, the naïve method and the proposed method have almost the same ability to distinguish different components, as the relative efficiencies of $\lambda$s are alway close to 1. When the sample size increases (from 200 to 350), the structure of the data becomes more complicated, and the estimating process becomes more difficult, which leads to the increasing of MSEs and decreasing of relative efficiencies. We can conclude that, for more complex data set (e.g., overlapping components case),

the task of identifying the correct components becomes harder.

Overall, our proposed method behaves better than the naïve method, for most of the cases, because of the accuracy of correctly estimating all the parameters.

### 3.4 NO data — A Real Data Analysis

Brinkman (1981) [16] studied the usefulness of pure ethanol as a spark-ignition engine fuel, which at the time was being considered for use in the U.S. and elsewhere. Efficiency and exhaust emissions with ethanol were quantified using a single-cylinder engine at compression ratios from 7.5 to 18, at equivalence ratios (the richness of the air-ethanol mix in an engine) from 1.2 (rich) to the lean limit, and at maximum brake torque (MBT) spark timing. Results were compared to those with gasoline at 7.5 compression ratio. With ethanol, compared to gasoline at the same compression ratio, engine thermal efficiency increased 3 percent, peak nitrogen oxide emissions decreased 40 percent, unburned fuel and carbon monoxide emissions were similar, and aldehyde emissions increased 110 to 360 percent. Increasing compression ratio from 7.5 to 18 with ethanol increased efficiency 18 percent, peak nitrogen oxide emissions 30 percent, unburned fuel emissions 25 to 200 percent, and aldehyde emissions 50 to 140 percent. Regression analysis indicated that the increased aldehyde emissions at high CR's may result from reduced exhaust temperatures. As exhaust temperature decreases, oxidation of aldehydes in the exhaust system decreases.

The data set describes the equivalence ratio, that is, against the peak nitric oxide emissions, while using pure ethanol as a spark-ignition engine fuel (Hurvich, Simonoff and Tsai (1998) [57]). Figure 3.4 shows the scatter plot of the 88 data points.

The scatter plot clearly indicates two different nitric oxide concentration dependencies, which means we can consider it as a mixture model problem. These data were analyzed using a mixture of linear regression in Hurn et al. (2003) [56]. However, the appropriateness of one component appears to have a nonlinear pattern (top data points), which could be captured using a quadratic effect. Thus we want to test the appropriateness of a mixtures-of-regressions model with one of the components to be quadratic.

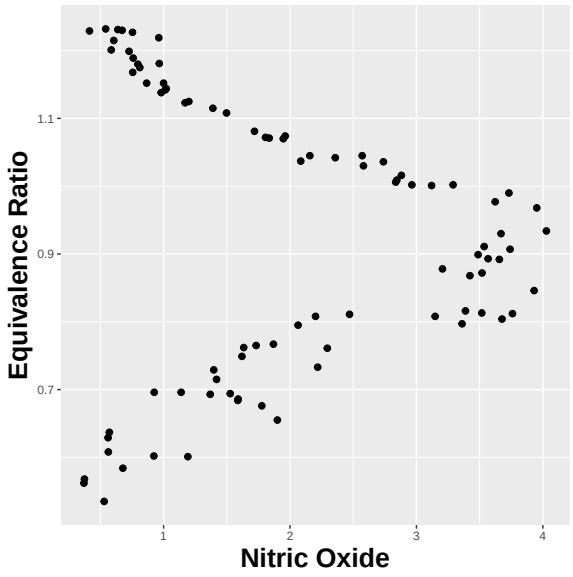


Figure 3.4: Equivalence ratio against exhaust nitric oxide concentration (*Source: Hurvich et al., 1998*).

To address the impact of measurement error, we add a measurement error term $u \sim N(0, 0.01)$ to the predictor $x$ (NO), and denote $w = x + u$ as the surrogate of predictor variable. This strategy was employed for the real data analysis in Yao and Song (2015)'s [119] work. We fit the data $(w, y)$ using both mixture of simple linear regressions model and a model with one quadratic component.

### 3.4.1 Parameter Estimation

To see the performance of the method, we proposed the semi-parametric bootstrap algorithm described in Subsection 3.2.2 with $B = 500$ bootstrap samples, which we assume the model is the one with quadratic term in one component; we output the estimation results for the NO data with both models, as well as the bootstrap standard errors (SE) of the estimation of 500 bootstrap samples (in parentheses) for each estimator.

Since the original data set has only a small amount of curvature, the result (model with quadratic term) depends heavily on starting values; that is, our estimation method can only perform well when we have a good starting value. We report our results of mixture-of-linear-regression and one term with quadratic term, both with

or without these informed *starting values* (SV); the results can be found in Table 3.3.

Table 3.3: Estimates for the NO data for a 2-component mixture model with both models.

| Parameter | Linear | Quadratic (without SV) | Quadratic (with SV) |
|---|---|---|---|
| $\lambda$ | 0.487<br>(0.068) | 0.486<br>(0.133) | 0.487<br>(0.096) |
| $\boldsymbol{\beta}_1$ | $(0.580, 0.078)^T$<br>$((0.016), (0.009))^T$ | $(0.600, 0.078)^T$<br>$((0.082), (0.050))^T$ | $(0.592, 0.076)$<br>$((0.063), (0.016))^T$ |
| $\boldsymbol{\beta}_2$ | $(1.239, -0.080)$<br>$((0.030), (0.008))^T$ | $(1.262, -0.133, 0.021)^T$<br>$((0.086), (0.116), (0.108))^T$ | $(1.270, -0.127, 0.011)$<br>$((0.055), (0.029), (0.005))^T$ |
| $\sigma_1$ | 0.045<br>(0.006) | 0.080<br>(0.093) | 0.078<br>(0.094) |
| $\sigma_2$ | 0.024<br>(0.017) | 0.031<br>(0.042) | 0.025<br>(0.030) |

The estimated values are calculated by the average of estimated values of 500 bootstrap samples, with different estimating methods. As we can see, when we specify the starting values for quadratic model fitting, the performance of estimation method is much better. The reason behind it is that, our proposed method is very sensitive to the starting values, and if we don't specify the starting values, sometimes the final result is quite variable. Hence, it is usually very important to choose a set of different starting values for the data set, and select the one with best results. The choice of starting values is always an ongoing topic of research, but we will not expand further on this topic for this dissertation.

Figure 3.3 shows the estimated regression lines for both models. The black solid lines are the estimated lines for the model that one component has a quadratic term (with starting values) and the red dashed lines are for the mixture of simple linear regressions. Different colors of data points represent different component those points belongs to according to the proposed method, assuming one component is quadratic.

From the output and figure, it is not absolutely clear which model is more appropriate. In fact, according to the bootstrap sample SEs, it seems like the model with quadratic term has a relative bigger SEs than the mixtures of linear regression model for all the parameters; what's more, if we see the average estimation values for those

500 bootstrap samples without starting values, the estimators for quadratic model is a little bit "off". One of the reasons for this situation, is possibly because our data set does not have a heavily quadratic trend for the component, and it leads to greater variability when estimating the parameters for some bootstrap samples.

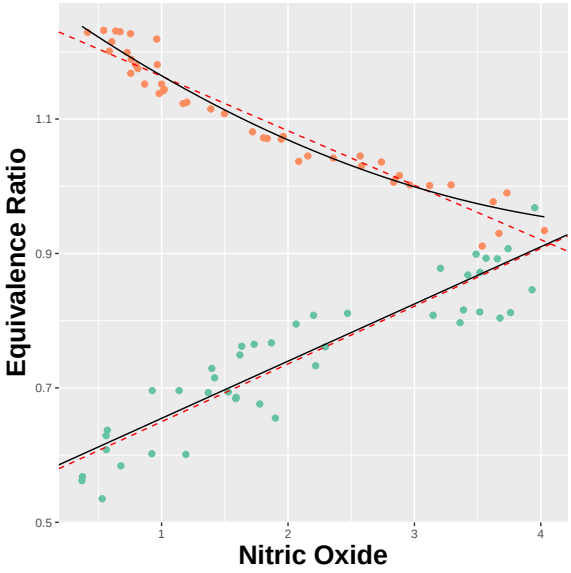


Figure 3.5: Estimated regression lines for both models.

In order to justify our decision, we apply model selection criteria, as well as a parametric bootstrap procedure here, where we compare the results from two different methods. The values of four model selection criteria discussed in Subsection 3.1.4 were calculated for both models, the appropriate model is chosen to correspond to the smallest penalized value. We report the result in Table 3.4.

Table 3.4: Various criteria for the determination of appropriate models for the NO data.

| | AIC | BIC | ICL | cAIC |
|---|---|---|---|---|
| Linear | −54.68 | −37.33 | −35.95 | −30.33 |
| Quadratic | **−64.45** | **−44.63** | **−43.25** | **−36.63** |

From the original data points we can see that the quadratic trend is limited, but there is a slight distinction from the linear trend. This can also be confirmed by the values of model selection criteria. For example, the cAIC values for linear and quadratic setting are −30.33 and −36.63, respectively, which is not necessarily a

substantial difference. Because of the existence of measurement error, the quadratic trend is more challenging to detect. But we believe this is only due to the structure of the original data set, since the model selection criteria perform well when we have some data set with a more clear trend of higher order terms.

According to the original scatter plot, the data with the larger responses seems to have a quadratic term. All the results above suggest it is more appropriate to use the model with a quadratic term for one of the components. The bold values in the table indicate the model chosen for that criterion, according to the table, all four criteria prefer quadratic model, which suggests a potential quadratic term in one component.

### 3.4.2 Likelihood Ratio Test simulation

To test whether we should keep the quadratic term in the model, it is also suggested to do a likelihood ratio test. Consider the data set from the previous subsection, assume we have the two-component mixture model

$$Y_i \sim \lambda N\left((1, x_i)\boldsymbol{\beta}_1, \sigma_1^2\right) + (1-\lambda) N\left((1, x_i, x_i^2)\boldsymbol{\beta}_2, \sigma_2^2\right),$$

where $\boldsymbol{\beta}_1 = (\beta_{10}, \beta_{11})^T$ and $\boldsymbol{\beta}_2 = (\beta_{20}, \beta_{21}, \beta_{22})^T$ for $i = 1, \cdots, n$. We want to test

$$H_0 : \beta_{22} = 0$$
$$H_1 : \beta_{22} \neq 0.$$

To see the behavior of likelihood ratio test, we also performed a bootstrap LRT with $B = 500$ bootstrap samples. Consider the two-component mixture model

$$Y_i \sim \lambda N\left((1, x_i)\boldsymbol{\beta}_1, \sigma_1^2\right) + (1-\lambda) N\left((1, x_i, x_i^2)\boldsymbol{\beta}_2, \sigma_2^2\right),$$

where $\boldsymbol{\beta}_1 = (\beta_{10}, \beta_{11})^T$ and $\boldsymbol{\beta}_2 = (\beta_{20}, \beta_{21}, \beta_{22})^T$ for $i = 1, \cdots, 88$. To incorporate the measurement error, we added a measurement error, $u_i \sim N(0, 1)$ to each predictor and indicate the observed $w = x + u$ the surrogate as $w_i$. Thus, we observed $\{(w_1, y_1), \cdots, (w_{88}, y_{88})\}$. In order to improve the performance of our likelihood ratio test, we also specified the starting values for this test in order to get a more accurate result.

Figure 3.6 shows all the LRT statistics of 500 bootstrap samples. Because of the variability with the estimating method, there are some statistics with very extreme values. On the other hand, we know the LRT statistics should have positive values for all sample statistics. However, in this data, the curvature is not very strong, when we add the measurement error to the original predictor variables, the property of quadratic curve may be reduced and it makes even more challenging to identify which one (quadratic or linear) is better, sometimes it leads to the likelihood value for alternative hypothesis is smaller than that for null hypothesis. We have conducted the same test using a simulated data set with a more strong curvature, the result shows that all the test statistics are positive and follows the $\chi^2$ distribution well, it shows that the most possible reason that we have negative test statistics is the structure of original data is not very suitable when dealing with the case with measurement error.

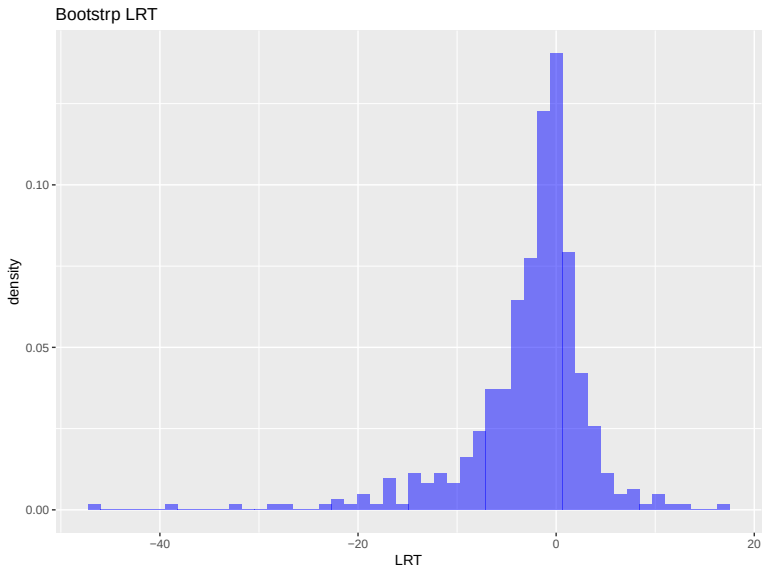


Figure 3.6: Likelihood ratio test statistics of 500 bootstrap samples.

Since the distribution of test statistics have so many negative values, the $\chi^2$ distribution is clearly not appropriate. Thus, we standardized the LRT by subtracting the mean of those statistics and dividing the standard deviation of the statistics,

$$T = \frac{\text{LRT} - \text{mean(LRT)}}{\text{sd(LRT)}}.$$

The distribution of the test statistics $T$ should follow approximately a standard normal distribution. The histogram of the test statistics $T$ is given in Figure 3.7, and the red curve is the density of standard normal distribution. From the histogram, there is clearly more peakedness in the distribution of the standardized test statistics relative to the standard normal density. Regardless, we proceed to compute the observed LRT, $T_{\text{obs}} = 11.775$, and then obtain the bootstrap p-value $p_B$ for this test by

$$p_B = \frac{1}{500} \sum_{i=1}^{500} I\left\{ \mid T^{(i)} \mid \geq \Lambda_{obs} \right\} \approx 0,$$

which is lower than the significance level $\alpha = 0.05$. We can claim the result is statistically significant and reject $H_0$, which is consistent with the assumption we made.

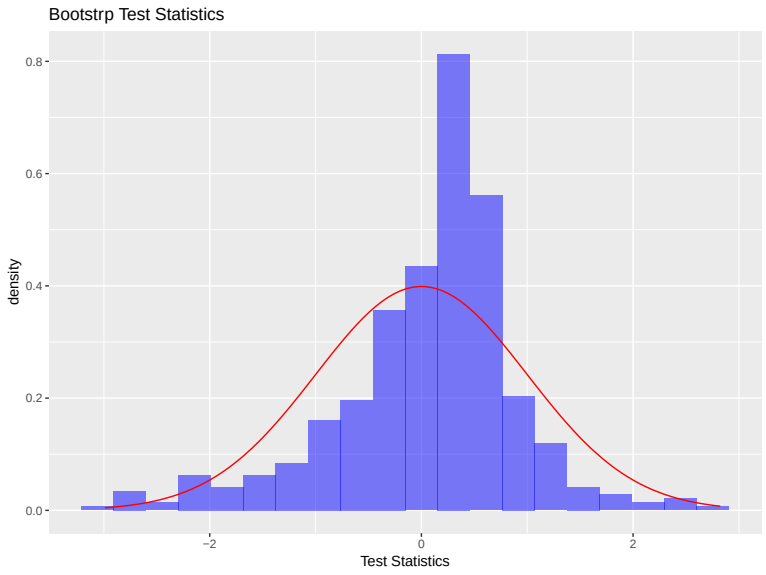


Figure 3.7: Bootstrap distribution of likelihood ratio test (LRT) statistics.

Given these results, we proceed to fit a 2 - component model with one component having a quadratic term. In other words, the model we fit the model

$$y_i \sim \begin{cases} (1, x_i)\boldsymbol{\beta}_1 + \epsilon_{i,1} & \text{with probability } \lambda; \\ (1, x_i, x_i^2)\boldsymbol{\beta}_2 + \epsilon_{i,2} & \text{with probability } 1 - \lambda, \end{cases} \tag{3.22}$$

where the $\epsilon_{i,j} \sim N(0, \sigma_j^2)$ are independent, $i = 1, 2, \cdots, 88$ and $j = 1, 2$.

## 3.5 Summary

In this chapter, we discussed the mixtures-of-regression model with measurement error in the predictors. We compute the conditional density of $Y \mid W$ following Yao and Song's paper, then found the parameters of interest by maximizing the likelihood function. Because of the existence of measurement error, the original estimates are biased, and the conditional density can correct the bias towards the measurement error, thus leading to better performance of the estimates.

We conducted a series of simulation studies to test the possible case when one of the components has a quadratic term in the parameter. The presence of measurement error can complicate the ability to estimate the effect of curvature, but the proposed method demonstrated smaller MSEs in estimating the parameters. We also conducted the bootstrap likelihood ratio test to test the quadratic term for a real data set, although the data itself only demonstrated moderate curvature in one of the components the appropriateness. We also showed that the proposed test is appropriate for detecting the presence of a quadratic effect in a mixtures-of-regressions model when measurement error is present.

# Chapter 4 Mixtures-of-Poisson Regressions with Measurement Error in the Predictors

Count data are frequently encountered in diverse areas, such as ecology, economics, and finance. One of the classical models for analyzing count data is the Poisson regression model. Poisson regression has a wide range of applications: Zou (2004) [124] developed a modified Poisson regression approach for epidemiologic and medical studies with binary data; Faria and F. Gonçalves (2013) [41] analyzed the financial data modeling by Poisson mixture regression.

However, most of the methods have been applied to the models under the assumption that predictors are measured without measurement error, which leads to biased estimation. In this chapter, we first introduce the Poisson regression model with measurement error in the predictors, and then generalize to the mixture setting. We then use the proposed model to analyze data regarding clandestine drug lab seizures in the states of Kentucky, Illinois, and Louisiana.

## 4.1 Poisson Regression with Measurement Error in Predictors

### 4.1.1 Introduction

Poisson regression is one of the most widely used models when dealing with data where the response is a count. Many statistical inferences and analyses have been discussed, for example, Frome et al. (1973) [45] applied Poisson regression model to analyze the rate collected by epidemiologic follow-up studies; Cameron and Trivedi (1998) [19] introduced regression analysis of count data, including Poisson regression; Winkelmann (2008) [116] discussed the Poisson regression in econometric analysis of count data. Moreover, Poisson regression is part of the broader class of generalized linear models, which has an extensive body of literature; see, for example McCullagh and Nelder (1987) [75] and Nelder and Wedderburn (1972) [86].

One of the model assumptions for the Poisson regression model is that the variance

of the response variable equals its mean, both conditional upon the predictor variables; a characteristic known as *equi-dispersion*. However, in many practical settings, it has been found that the conditional variance is greater than its conditional mean, a phenomenon called *over-dispersion*, which may lead to a possible loss of efficiency (Cox (1983) [32]). An alternative model that takes into account over-dispersion is the *negative binomial* (NB) model. Various inference considerations for the negative binomial model have been addressed using, for example, likelihood methods (Lawless (1987) [70]), weighted least squares (Breslow (1984) [15]), and quasi-likelihood (McCullagh and Nelder (1987) [75]).

There are many sources that could lead to over-dispersion, for example, the lack of covariates, the non-independence of the data set, or an excess frequency of zeroes (zero-inflation). Another possible source is measurement error in the predictors, which is the focus of our present research. In this chapter, we will first introduce Poisson regression with measurement error, and then show how it can lead to over-dispersion in the observed data, thus causing inconsistent estimates. Finally, we will propose a method for estimating a Poisson regression model with measurement error, and apply it to both simulated data and real data sets.

### 4.1.2 Poisson Regression with Additive Measurement Error

To simplify our discussion, here, we restrict our attention to the case of just one covariate $X$; further research can be done by expanding this result to multiple regression analysis. Suppose $Y$ is a Poisson random variable distributed with parameter $\theta_x$, which is a function of $X$ and the unknown parameter $\boldsymbol{\beta} = (\beta_0, \beta_1)^T$; i.e.,

$$Y \mid X \sim \text{Poi}(\theta_X) \tag{4.1}$$

where $\theta_X = \exp(\beta_0 + \beta_1 X)$. Assume we have $n$ independent observations $(x_i, y_i)$. The probability mass function of $y_i \mid x_i$ can be written as

$$f(y_i \mid x_i, \theta_{x_i}) = e^{-\theta_{x_i}} \left(\theta_{x_i}\right)^{y_i} / y_i!. \tag{4.2}$$

The log-likelihood function of $\boldsymbol{\beta}$ is then

$$\ell(\boldsymbol{\beta}) = \sum_{i=1}^{n} \left[ y_i \log \theta_{x_i}(\boldsymbol{\beta}) - \theta_{x_i}(\boldsymbol{\beta}) - \log\left(y_i!\right) \right], \tag{4.3}$$

where $\theta_{x_i}(\boldsymbol{\beta}) = \exp(\beta_0 + \beta_1 x_i)$. Therefore, the MLE of the unknown parameter $\boldsymbol{\beta}$ can then be obtained by solving the estimating equation

$$S_n(\boldsymbol{\beta} \mid \mathbf{y}, \mathbf{x}) = \frac{1}{n} \sum_{i=1}^{n} \left[ y_i - \exp\left(\beta_0 + \beta_1 x_i\right) \right] (1, x_i)^T = 0, \tag{4.4}$$

where $\mathbf{x} = (x_1, x_2, \cdots, x_n)^T$ is the vector of predictor variables, and $\mathbf{y} = (y_1, y_2, \cdots, y_n)^T$ is the vector of response variables

Within the framework where measurement error is present, we observe the surrogate $w_1, \cdots, w_n$ instead of the true predictors $x_1, \cdots, x_n$ where

$$\mathrm{w}_i = x_i + u_i$$

and the corresponding vector of surrogates $\mathbf{w} = (w_1, w_2, \cdots, w_n)^T$. Here $u_i$ is the measurement error, which is assumed to be independent of $(x_i, y_i)$ and often assumed to be normally distributed:

$$u_i \sim N(0, \sigma_u^2).$$

Furthermore, its variance $\sigma_u^2 < \infty$ is assumed to be known. In the presence of measurement error, the estimators for the Poisson regression model could be biased. We now introduce some existing methods that applied for Poisson regression model with measurement error in the predictors.

### 4.1.3 Existing Estimators

Poisson regression models with measurement errors in the predictors, has been investigated in the past; see Carroll et al. (1995) [23]. There are some general approaches in addressing measurement error problems in GLMs. In this section, we will consider two types of adjusted estimators; *structural estimator* (Thamerus (1998) [109]) and *corrected score* (CS) *estimator* (Stefanski (1989) [106] and Nakamura (1990) [85]). Both of the methods can be applied to a wide variety of regression models with

covariate measurement error, including the Poisson regression model. For example, Patriota et al. (2009) [88] used the method of moment (corrected score) method for a heteroscedastic structural measurement error model with epidemiological data sets while Cao and Zhu (2011) [20] discussed the structural method for measurement error model under heavy-tailed distributions. The following describes the two methods and how they are applied under a Poisson regression model.

**Functional Method: Corrected Score Estimator**

When there is no measurement error, estimates for the parameter $\boldsymbol{\beta}$ are obtained by solving the estimating equation (4.4). If we replace the unobservable variables $x_i$ by the observable surrogates $w_i$, we now arrive at the so-called "naïve" estimator, which is found by maximizing

$$\ell_{\text{naïve}}(\boldsymbol{\beta}) = \sum_{i=1}^{n} \left[ y_i \log \theta_{w_i}(\boldsymbol{\beta}) - \theta_{w_i}(\boldsymbol{\beta}) - \log\left(y_i!\right) \right],$$

where $\theta_{w_i}(\boldsymbol{\beta}) = \exp(\beta_0 + \beta_1 w_i)$.

The resulting estimator $\hat{\boldsymbol{\beta}}_{\text{naïve}}$ would be the MLE if $w_i$s were measured without errors, i.e., if $w_i = x_i$ for all $i$. In this case, $\hat{\boldsymbol{\beta}}_{\text{naïve}}$ would be consistent. But as $x_i$ has been replaced by $w_i$, the $\hat{\boldsymbol{\beta}}_{\text{naïve}}$ is inconsistent. To construct a consistent estimator, we have to correct for the measurement error.

The idea underlying the corrected score estimator is that the conditional distribution of the corrected estimate with respect to $w_i$ given the true independent variables $x_i$ and the dependent variable $y_i$ is centered around the ML estimator, which is consistent to the true value of the parameter of interest. That is, we utilize a 'corrected' log-likelihood function $\ell_{\text{CS}}(\boldsymbol{\beta}, w_i, y_i)$, such that

$$\mathbb{E}\left[\ell_{\text{CS}}(\boldsymbol{\beta}, w_i, y_i)\right] = \ell(\boldsymbol{\beta}, x_i, y_i) = y_i \log \theta_{x_i}(\boldsymbol{\beta}) - \theta_{x_i}(\boldsymbol{\beta}) - \log\left(y_i!\right).$$

Such a function is given by

$$\ell_{\text{CS}}(\boldsymbol{\beta}, x_i, y_i) = y_i \log \theta_{w_i}(\boldsymbol{\beta}) - \exp\left(-\frac{1}{2}\beta_1^2 \sigma_u^2\right) \theta_{w_i}(\boldsymbol{\beta}) - \log\left(y_i!\right),$$

since

$$\mathbb{E}\left[\log \theta_{w_i}(\boldsymbol{\beta}) \mid x_i\right] = \mathbb{E}\left[\beta_0 + \beta_1 w_i \mid x_i\right] = \beta_0 + \beta_1 x_i = \log \theta_{x_i}(\boldsymbol{\beta})$$

and

$$\mathbb{E}\left[\theta_{w_i}(\boldsymbol{\beta}) \mid x_i\right] = \exp\left(\beta_0 + \beta_1 x_i\right) \mathbb{E}\left[\exp(\beta_1 u_i)\right] = \theta_{x_i}(\boldsymbol{\beta}) \exp\left(\frac{1}{2}\beta_1^2 \sigma_u^2\right).$$

Hence, the corresponding corrected criterion function is given by

$$\ell_{\mathrm{CS}}(\boldsymbol{\beta}) = \sum_{i=1}^{n} \left[ y_i \log \theta_{w_i}(\boldsymbol{\beta}) - \exp\left(-\frac{1}{2}\beta_1^2 \sigma_u^2\right) \theta_{w_i}(\boldsymbol{\beta}) - \log\left(y_i!\right) \right]. \tag{4.5}$$

We can now define the new corrected score function, $S_n^{\mathrm{CS}}(\boldsymbol{\beta} \mid \mathbf{y}, \mathbf{w})$, by taking the derivative of $\ell_{\mathrm{CS}}(\boldsymbol{\beta})$, which is unbiased for $S_n(\boldsymbol{\beta} \mid \mathbf{y}, \mathbf{x})$; i.e.,

$$\mathbb{E}\left[S_n^{\mathrm{CS}}(\boldsymbol{\beta} \mid \mathbf{y}, \mathbf{w}) \mid (\mathbf{y}, \mathbf{x})\right] = S_n(\boldsymbol{\beta} \mid \mathbf{y}, \mathbf{x}).$$

Thus, the score function of the corrected score estimator is given by

$$S_n^{\mathrm{CS}}(\boldsymbol{\beta}|\mathbf{y}, \mathbf{w}) = \frac{1}{n} \sum_{i=1}^{n} \left[ (y_i - \phi_{\mathrm{cs}} \theta_{w_i})(1, w_i)^T + \phi_{\mathrm{cs}} \theta_{w_i} \sigma_u^2 \beta_1 \left(0, 1\right)^T \right], \tag{4.6}$$

where $\phi_{\mathrm{CS}} = \exp\left(-\frac{1}{2}\beta_1^2 \sigma_u^2\right)$, $\theta_{w_i} = \exp(\beta_0 + \beta_1 w_i)$. Setting (4.6) equal to zero and solving for $\boldsymbol{\beta}$ yields the solution $\hat{\boldsymbol{\beta}}_{\mathrm{CS}}$, which is called the corrected score estimator.

According to the theory of quasi-score estimators (Heyde (1997) [53]), this estimator is strongly consistent, and $\sqrt{n}(\hat{\boldsymbol{\beta}}_{\mathrm{CS}} - \hat{\boldsymbol{\beta}})$ converges in distribution to $N(0, \Sigma_{\mathrm{CS}})$, where $\Sigma_{\mathrm{CS}}$ is given by $A^{-1}BA^{-1}$ with

$$A = -\mathbb{E}\frac{\partial S_n^{\mathrm{CS}}}{\partial \boldsymbol{\beta}^T},$$

$$B = \mathrm{cov}\left\{S_n^{\mathrm{CS}}\right\}.$$

**Structural Method: Structural Estimator**

The corrected score estimator is constructed without using the distribution of $X$. There is, however, a completely different approach to the construction of consistent estimators. In structural estimation, $X$ is regarded as a random variable. Suppose now that we fully specify a distribution for $Y \mid X$ and $X \mid W$, and thus for $X \mid W$. Estimators originating as the solution to such estimating equations are called structural estimators (Carroll et al. (1995) [23]). The idea of structural estimators

is to substitute the unobserved $\theta_X$ — which is the mean of $Y$, with the conditional mean of $Y \mid W$.

We base our investigation on the full understanding of the structure of the predictor variable, and assume that

$$x_i \sim N(\mu_x, \sigma_x^2).$$

We also assume that the $(y_i, x_i, u_i)$, $i = 1, \cdots, n$ are i.i.d. Then $x_i$ given $w_i$ is also normal with mean and variance, respectively,

$$\mathbb{E}(x_i \mid w_i) = [\rho/(1+\rho)]\,\mu_x + [1/(1+\rho)]\,w_i \equiv m(w_i), \tag{4.7}$$

$$\mathrm{Var}(x_i \mid w_i) = \rho(1+\rho)^{-2}\sigma_w^2 \equiv \tau^2(w_i), \tag{4.8}$$

where $\sigma_w^2 = \mathrm{Var}(w_i) = \sigma_x^2 + \sigma_u^2 = (1+\rho)\sigma_x^2$ and $\rho = \sigma_u^2/\sigma_x^2$. Even when $x_i$ is not normal, (4.7) is the best *linear* approximation of $\mathbb{E}(x_i \mid w_i)$ (Carroll et al. (1995) [23]).

The Poisson regression model can be written as a mean-variance model in $x_i$:

$$\mathbb{E}(y_i \mid x_i) = \exp(\beta_0 + \beta_1 x_i).$$

Recall for $(x_i, y_i)$, the log-likelihood function of the Poisson regression without measurement error is given by (4.3). Since the mean $\theta_{x_i} = \mathbb{E}(y_i|x_i) = \exp(\beta_0 + \beta_1 x_i)$ is not observable, we replaced it by the conditional mean of $y_i$ on the observed $w_i$,

$$\mathbb{E}\left(y_i \mid w_i\right) = \exp\left[\beta_0 + \beta_1 m(w_i) + \frac{1}{2}\beta_1^2\tau^2(w_i)\right] \equiv \eta_i. \tag{4.9}$$

Then,

$$\ell_{\mathrm{S}}(\boldsymbol{\beta} \mid \mathbf{y}, \mathbf{w}) = \sum_{i=1}^{n} \left[y_i \log(\eta_i) - \eta_i - \log\left(y_i!\right)\right], \tag{4.10}$$

where $\boldsymbol{\beta} = (\beta_0, \beta_1)^T$ can be used similar to the likelihood function, and obtain a consistent structural estimator when $\ell_{\mathrm{S}}(\boldsymbol{\beta} \mid \mathbf{y}, \mathbf{w})$ is maximized.

Taking the derivative of (4.10) with respect to $\boldsymbol{\beta}$, we get the score function of the structural estimator $\hat{\boldsymbol{\beta}}_{\mathrm{S}}$:

$$\begin{aligned} S_n^{\mathrm{S}}(\boldsymbol{\beta}) &= \frac{1}{n}\sum_{i=1}^{n} \frac{y_i - \eta_i}{\eta_i}\frac{\partial \eta_i}{\partial \boldsymbol{\beta}} \\ &= \frac{1}{n}\sum_{i=1}^{n} (y_i - \eta_i)(1, m(w_i) + \beta_1\tau^2(w_i))^T. \end{aligned} \tag{4.11}$$

By setting the above equal to zero and numerically solving for $\boldsymbol{\beta}$, we can obtain the structural estimator $\hat{\boldsymbol{\beta}}_{\mathrm{S}}$.

### 4.1.4 Approximated Maximum Likelihood Estimator for a Small Measurement Error

Both the structural method and the corrected score method focus on adjusting the expectation of the likelihood, instead of working with the true density $f(\mathbf{y} \mid \mathbf{w})$. Similar to Chapter 2 for linear regression models with measurement error in the predictors, as discussed by Yao and Song (2015) [119], we now come up with the idea to compute the conditional density function of $Y \mid W$ by the integral, and then estimate the parameters using the true density function.

Suppose all $\mathbf{w}, \mathbf{x}$ and $\mathbf{u}$ are normally distributed. The density function of the observation $(y_i, w_i)$ is then given by a combination of a Poisson and normal distribution

$$\begin{aligned} f(y_i \mid w_i) &= \int_{-\infty}^{\infty} f(y \mid x_i) f(x_i \mid w_i) dx \\ &= \frac{1}{\sqrt{2\pi\tau^2(w_i)}} \frac{1}{y_i!} \int_{-\infty}^{\infty} \exp\left[-e^{\beta_0+\beta_1 x_i} + y_i\left(\beta_0 + \beta_1 x_i\right)\right] \exp\left\{-\frac{(x_i - m(w_i))^2}{2\tau^2(w_i)}\right\} dx \end{aligned} \tag{4.12}$$

where $x_i \mid w_i \overset{iid}{\sim} N(m(w_i), \tau^2(w_i))$ for $i = 1, \cdots, n$.

We need to find $\boldsymbol{\beta} = (\beta_0, \beta_1)^T$ by maximizing the log-likelihood function

$$\ell(\boldsymbol{\beta}) = \sum_{i=1}^{n} \log f(y_i \mid w_i).$$

Therefore, $\boldsymbol{\beta}$ can be obtained by the solution of its first derivatives:

$$\begin{aligned}
\mathbf{0} &= \frac{\partial \ell(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}} \\
&= \sum_{i=1}^{n} \frac{\partial \log f(y_i \mid w_i)}{\partial \boldsymbol{\beta}} \\
&= \sum_{i=1}^{n} \frac{\frac{\partial}{\partial \boldsymbol{\beta}} f(y_i \mid w_i)}{f(y_i \mid w_i)} \\
&\approx \sum_{i=1}^{n} \left\{ \frac{y_i \int x_i \exp\left[-e^{\beta_0+\beta_1 x_i} + y_i(\beta_0+\beta_1 x_i)\right] S_i dx}{f(y_i \mid w_i, \theta_{x_i})} - \right. \\
&\quad \left. \frac{\int x_i \exp\left[-e^{\beta_0+\beta_1 x_i} + (y_i+1)(\beta_0+\beta_1 x_i)\right] S_i dx}{f(y_i \mid w_i, \theta_{x_i})} \right\}
\end{aligned}$$

where $S_i = \exp\left[-\frac{(x-m(w_i))^2}{2\tau^2(w_i)}\right]$, and $f(y_i \mid w_i)$ is given by (4.12). However, the numerical solution for this equation is challenging, as it involves evaluating the value of the integrals numerically for each iteration. Moreover, the initial value for the EM algorithm can also cause a problem, as the integrals may not converge after the first few iterations.

Yang (2012) [118] described a approximation method of the density function when the conditional variance of the surrogate $w$ is small. The density function of observations can be expressed by the form of an expectation

$$\begin{aligned}
f(y \mid w) &= \int f(y \mid x) f(x \mid w) dx \\
&= \mathbb{E}_{x|w} f_0(x),
\end{aligned}$$

where $x \mid w \sim N(m(w), \tau^2(w))$ and $f_0(x) = f(y \mid x) = \frac{1}{y!} \exp\left[y(\beta_0+\beta_1 x) - \exp(\beta_0+\beta_1 x)\right]$. By taking the Taylor expansion of $f_0(x)$ on $x = \mathbb{E}_{x|w}(x \mid w) = m(w)$, the density function can be written as

$$\begin{aligned}
f(y \mid w, \theta_x) &= \mathbb{E}_{x|w} f_0(x) \\
&= \mathbb{E}_{x|w} \left[ f_0(m(w)) + \sum_{t=1}^{\infty} \frac{1}{t!} f_0^{(t)}(m(w))(x-m(w))^t \right] \\
&= f_0(m(w)) + \sum_{t=1}^{\infty} \frac{1}{t!} f_0^{(t)}(m(w)) M_t, \qquad (4.13)
\end{aligned}$$

where $f_0^{(t)}(\cdot)$ is the $t$th derivative of $f_0(\cdot)$, and $M_t = \mathbb{E}_{x|w}(x - m(w))^t$ is the $t$th moment of $x \mid w$. Under the assumption that $x$ follows a normal distribution, the $t$th moment $M_t$ is

$$M_t = \begin{cases} 0 & t \text{ is odd,} \\ \tau^t \cdot (t-1) \cdot (t-3) \cdots 3 \cdot 1 & t \text{ is even.} \end{cases}$$

We then plug the expression of moments into (4.13);

$$\begin{aligned} f(y \mid w) &= f_0(m(w)) + \sum_{t=1}^{\infty} \frac{1}{t!} f_0^{(t)}(m(w)) M_t \\ &= f_0(m(w)) + \sum_{t \text{ is even}}^{\infty} \tau^t(w)(t-1)!! \frac{1}{t!} f_0^{(t)}(m(w)) \\ &= f_0(m(w)) \left[ 1 + \sum_{s=1}^{\infty} \tau^{2s}(w) \frac{1}{2^s s!} h_s(m(w)) \right], \end{aligned} \tag{4.14}$$

where

$$\begin{aligned} h_0(x) &= \beta_1 \left[ y - \exp(\beta_0 + \beta_1 x) \right] \\ h_1(x) &= h_0^2(x) + h_0^{(1)}(x) \\ h_s(x) &= h_1(x) h_{s-1}(x) + 2 h_0(x) h_{s-1}(x) + h_{s-1}^{(2)}(x), \text{ for } s > 1 \end{aligned}$$

and $h_s^{(i)}(\cdot)$ is the $i$th derivative of the function $h_s(\cdot)$.

Consider the case where $\tau^2(w)$ is small, so the proportion of variability explained by measurement error is relatively small. Taking the expansion in (4.14) up to the term of $s = 1$, we can approximate the density function by

$$\begin{aligned} f(y \mid w) &= f_0(m(w)) \left[ 1 + \frac{1}{2} \tau^2(w) h_1(m(w)) \right] + O(\tau^4(w)) \\ &\approx f_0(m(w)) \left\{ 1 + \frac{1}{2} \tau^2(w) \cdot \beta_1^2 \left[ (y - \theta_{m(w)})^2 - \theta_{m(w)} \right] \right\}, \end{aligned} \tag{4.15}$$

with the order $O(\tau^4(w))$ (Yang (2012) [118]), where $\theta_{m(w)} = \exp(\beta_0 + \beta_1 m(w))$ and $m(\cdot)$ is defined in (4.7).

Now we can estimate the parameters $\boldsymbol{\beta} = (\beta_0, \beta_1)^T$ by solving the first derivative

of the log-likelihood function of the approximated density function (4.15),

$$\begin{aligned}\ell(\boldsymbol{\beta}) &= \sum_{i=1}^{n} \log f_0(m(w_i)) \left\{ 1 + \frac{1}{2}\tau^2(w_i) \cdot \beta_1^2 \left[ (y - \theta_{m(w_i)})^2 - \theta_{m(w_i)} \right] \right\} \\ &= \sum_{i=1}^{n} \left[ -\log y_i! + \left( y_i \theta_{m(w_i)} - \theta_{m(w_i)} \right) + \log A_i \right],\end{aligned}$$

where $A_i = 1 + \frac{1}{2}\tau^2(w_i) \cdot \beta_1^2 \left[ (y_i - \theta_{m(w_i)})^2 - \theta_{m(w_i)} \right]$. So the estimating equations can be written as

$$\begin{aligned}0 &= \frac{\partial}{\partial \beta_0} \ell(\boldsymbol{\beta}) \\ &= \sum_{i=1}^{n} \left[ \frac{\partial}{\partial \beta_0} \left( y_i \theta_{m(w_i)} - \theta_{m(w_i)} \right) + \frac{1}{A_i} \frac{\partial}{\partial \beta_0} A_i \right] \\ &= \sum_{i=1}^{n} \left[ \left( y_i - \theta_{m(w_i)} \right) - \frac{1}{2A_i} \tau^2(w_i) \cdot \beta_1^2 \theta_{m(w_i)} \left( 2(y_i - \theta_{m(w_i)}) + 1 \right) \right],\end{aligned}$$

$$\begin{aligned}0 &= \frac{\partial}{\partial \beta_1} l(\boldsymbol{\beta}) \\ &= \sum_{i=1}^{n} \left[ \frac{\partial}{\partial \beta_1} \left( y_i \theta_{m(w_i)} - \theta_{m(w_i)} \right) + \frac{1}{A_i} \frac{\partial}{\partial \beta_1} A_i \right] \\ &= \sum_{i=1}^{n} [ \left( y_i - \theta_{m(w_i)} \right) m(w_i) - \frac{1}{2A_i} \tau^2(w_i) \cdot \beta_1^2 \theta_{m(w_i)} \left( 2(y_i - \theta_{m(w_i)}) + 1 \right) m(w_i) \\ &\quad + \frac{1}{A_i} \tau^2(w_i) \cdot \beta_1 \left( (y_i - \theta_{m(w_i)})^2 - \theta_{m(w_i)} \right) ].\end{aligned}$$

The estimator we find based on approximated density functions is called the *approximated maximum likelihood estimator* (AMLE) by Yang (2012) [118].

## 4.2 Mixtures of Poisson Regression with Measurement Errors

Based on the estimating methods discussed in the previous Section, we now expand the model to the mixture setting, and discuss how we can estimate mixtures of Poisson regression model using the previous estimators.

### 4.2.1 Mixtures of Poisson Regression

Suppose $\mathbf{Z}$ is the $n \times k$ random indicator matrix whose $(i,j)$th element $z_{ij}$ equals 1 when $y_i$ is from the $j$th component, zero otherwise, with $P(z_{ij} = 1) = \lambda_j$, for $j = 1, 2, \cdots, k$, where $\sum_{j=1}^{k} \lambda_j = 1$. Let $\mathbf{y} = (y_1, \cdots, y_n)^T$ be the vector of responses, $\mathbf{x} = (x_1, x_2, \cdots, x_n)^T$ be the predictor vector. Suppose $\boldsymbol{\beta}_j = (\beta_{0j}, \beta_{1j})^T$ is the unknown parameter for the $j$th component, and $\boldsymbol{\lambda} = (\lambda_1, \cdots, \lambda_k)^T$ is the vector of mixing proportions. Given $z_{ij} = 1$, the mixtures of Poisson regressions model can be written as

$$y_i \mid x_i \sim Poi(\theta_{ij}), \tag{4.16}$$

where $\theta_{ij} = \exp(\beta_{0j} + \beta_{1j} x_i)$. Thus, the probability mass function of $y_i$ belonging to the $j$th component can be written as

$$f(y_i \mid x_i, \theta_{ij}) = e^{-\theta_{ij}} \left(\theta_{ij}\right)^{y_i} / y_i!.$$

The complete data set, is given by $\{\mathbf{x}, \mathbf{y}, \mathbf{z}\}$.

### 4.2.2 Poisson mixture regression model with measurement error

Let $\mathbf{u} = (u_1, u_2, \cdots, u_n)^T$ be the $n$-dimensional measurement error vector that satisfies $\mathbb{E}(\mathbf{u}) = \mathbf{0}$. Suppose we observe the surrogate data $w_1, \cdots, w_n$ instead of $x_1, \cdots, x_n$ in the mixture of Poisson regression model (4.16), where the $x_i$s and the $w_i$s are related by the classical measurement error model:

$$w_i = x_i + u_i.$$

To make all the methods work properly, we further assume that the $x_i$s, are *independent and identically-distributed* (i.i.d.) as $X \sim N(\mu_x, \sigma_x^2)$, the error $u_i$ is distributed as $N(0, \sigma_u^2)$, $i = 1, \cdots, n$, and the $x_i$s and $u_i$s are mutually independent. So the $k$ - component mixture of Poisson regression model with measurement error can be written as

$$f(y_i \mid w_i) = \sum_{j=1}^{k} \lambda_j \exp\left[-e^{(\beta_{0j} + \beta_{1j} w_i)}\right] \left[e^{(\beta_{0j} + \beta_{1j} w_i)}\right]^{y_i} / y_i!.$$

Thus, the likelihood function of parameters $(\boldsymbol{\beta}, \boldsymbol{\lambda})$ is

$$L(\boldsymbol{\beta}, \boldsymbol{\lambda}) = \prod_{i=1}^{n} \left\{ \sum_{j=1}^{k} \lambda_j \cdot \exp\left[-e^{(\beta_{0j}+\beta_{1j}w_i)}\right] \left[e^{(\beta_{0j}+\beta_{1j}w_i)}\right]^{y_i} / y_i! \right\}.$$

The next Subsections discuss some methods that can be used for estimating the mixtures-of-Poisson regression model, when measurement error is also addressed.

### 4.2.3 Corrected Score Estimator

Like the Poisson regression without the mixture setting, the 'naïve' estimators for mixture of Poisson regressions model are found by maximizing the log-likelihood function

$$\begin{aligned}\ell_{\text{naive}}(\boldsymbol{\beta}, \boldsymbol{\lambda}) &= \log L(\boldsymbol{\beta}, \boldsymbol{\lambda}) \\ &= \sum_{i=1}^{n}\sum_{j=1}^{k} z_{ij} \log(\lambda_j) + \sum_{i=1}^{n}\sum_{j=1}^{k} z_{ij} \left[y_i \log \theta_{w_i}(\boldsymbol{\beta}_j) - \theta_{w_i}(\boldsymbol{\beta}_j) - \log(y_i!)\right],\end{aligned}$$

where $\theta_{w_i}(\boldsymbol{\beta}_j) = \exp(\beta_{0j} + \beta_{1j} w_i)$. Because of the existence of measurement error, the naïve estimator is biased, so we need to 'correct' for the measurement error in order to get the unbiased estimator.

To incorporate the corrected score method in a mixture setting, similar to the non-mixture setting, we substitute the log-likelihood function with our corrected criterion log-likelihood function:

$$\ell_{\text{cor}}(\boldsymbol{\beta}, \boldsymbol{\lambda}) = \sum_{i=1}^{n}\sum_{j=1}^{k} z_{ij} \log(\lambda_j) + \sum_{i=1}^{n}\sum_{j=1}^{k} z_{ij} \left[y_i \log \theta_{w_i}(\boldsymbol{\beta}_j) - \exp\left(-\frac{1}{2}\beta_{1j}^2 \sigma_u^2\right) \theta_{w_i}(\boldsymbol{\beta}_j) - \log(y_i!)\right]. \tag{4.17}$$

We then define the new corrected score function — $S_n^{cor}(\boldsymbol{\beta}, \boldsymbol{\lambda} \mid \mathbf{y}, \mathbf{w})$, by taking the derivative to $\ell_{\text{cor}}(\boldsymbol{\beta}, \boldsymbol{\lambda})$, which is unbiased for $S_n(\boldsymbol{\beta}, \boldsymbol{\lambda} \mid \mathbf{y}, \mathbf{x})$, i.e.,

$$\mathbb{E}\left[S_n^{cor}(\boldsymbol{\beta}, \boldsymbol{\lambda} \mid \mathbf{y}, \mathbf{w}) \mid (\mathbf{y}, \mathbf{x})\right] = S_n(\boldsymbol{\beta}, \boldsymbol{\lambda} \mid \mathbf{y}, \mathbf{x}).$$

Thus, the score function of the corrected score estimator $\boldsymbol{\beta}$ is given by

$$S_n^{cor}(\boldsymbol{\beta}, \boldsymbol{\lambda} \mid \mathbf{y}, \mathbf{w}) = \sum_{i=1}^{n}\sum_{j=1}^{k} z_{ij} \left[(y_i - \phi_{\text{cs}}^{(j)} \theta_{w_i}^{(j)})(1, w_i)^T + \phi_{\text{cs}}^{(j)} \theta_{w_i}^{(j)} \sigma_u^2 \beta_{1j} (0, 1)^T\right], \tag{4.18}$$

where $\phi_{\text{cs}}^{(j)} = \exp\left(-\frac{1}{2}\beta_{1j}^2\sigma_u^2\right)$, $\theta_{w_i}^{(j)} = \exp(\beta_{0j} + \beta_{1j}w_i)$. We then set (4.18) equal to zero for the parameter estimation, $S_n^{cor}(\hat{\boldsymbol{\beta}} \mid \mathbf{y}, \mathbf{w}) = 0$. We call this estimator the corrected score estimator.

### 4.2.4 Structural Estimator

For the structural estimation, suppose we fully know the distribution of the predictors $X$ and measurement error $U$. The joint probability density function of the data $(\mathbf{y}, \mathbf{w}, \mathbf{z})$ is

$$f_{\text{joint}}(\mathbf{y}, \mathbf{w}, \mathbf{z}) = \prod_{i=1}^{n}\prod_{j=1}^{k} \lambda_j^{z_{ij}} \left[e^{-\eta_{ij}} \left(\eta_{ij}\right)^{y_i} / y_i!\right]^{z_{ij}},$$

where $\eta_{ij} = \mathbb{E}\left(y_i \mid w_i, z_{ij}\right) = \exp\left[\beta_{0j} + \beta_{1j}m(w_i) + \frac{1}{2}\beta_1^2\tau^2(w_i)\right]$, while $m(w_i)$ and $\tau^2(w_i)$ are the conditional expectation and variance of $x_i$ given $w_i$, respectively.

Like the non-mixture setting, the conditional structural log-likelihood function can be written as

$$\ell_s(\boldsymbol{\beta}, \boldsymbol{\lambda} \mid \mathbf{y}, \mathbf{w}, \mathbf{z}) = \sum_{i=1}^{n}\sum_{j=1}^{k} z_{ij}\log(\lambda_j) + \sum_{i=1}^{n}\sum_{j=1}^{k} z_{ij}\left[y_i\log(\eta_{ij}) - \eta_{ij} - \log\left(y_i!\right)\right].$$

Taking the partial derivative of $\ell_s(\boldsymbol{\beta}, \boldsymbol{\lambda} \mid \mathbf{y}, \mathbf{w}, \mathbf{z})$ with respect to $\lambda_j$, we can obtain the structural estimator $\hat{\lambda}_j$ by setting the score function to 0; i.e.,

$$S_n^s(\lambda_j) = \sum_{i=1}^{n}\left(\frac{z_{ij}}{\lambda_j} - \frac{z_{ik}}{\lambda_k}\right) = 0, \tag{4.19}$$

for $j = 1, \cdots, k-1$. The above yields $\hat{\lambda}_j = \frac{1}{n}\sum_{i=1}^{n} z_{ij}$.

Taking the partial derivative with respect to $\boldsymbol{\beta}_j$, we get the score function of the structural estimator $\hat{\boldsymbol{\beta}}_j$:

$$\begin{aligned} S_n^s(\boldsymbol{\beta}_j) &= \sum_{i=1}^{n}\sum_{j=1}^{k} z_{ij}\frac{y_i - \eta_{ij}}{\eta_{ij}}\frac{\partial\eta_i}{\partial\boldsymbol{\beta}_j} \\ &= \sum_{i=1}^{n}\sum_{j=1}^{k} z_{ij}(y_i - \eta_{ij})(1, m(w_i) + \beta_{1j}\tau^2(w_i))^T. \end{aligned} \tag{4.20}$$

By setting the score function equal to zero, we can numerically solve for the structural estimator $\hat{\boldsymbol{\beta}}_{\boldsymbol{s}j}$ for $j = 1, \cdots, k$. We call these estimators the structural estimator.

### 4.2.5 Approximated Maximum Likelihood Estimator

The most straightforward way to estimate the parameters is to maximize the true log-likelihood function based on the conditional density of $Y \mid W$. In Subsection 4.1.4, we introduced the approximated maximum likelihood for the non-mixture setting. We now expand it to the mixture model.

Given $\mathcal{Z} = j$, the conditional density of $Y_i$ given $W_i = w_i$ can be given by

$$\begin{aligned} f_j(y_i \mid w_i, \theta_{jx_i}) &= \int_{-\infty}^{\infty} f(y \mid x_i, \theta_{jx_i}) f(x_i \mid w_i) dx_i \\ &= \frac{1}{\sqrt{2\pi\tau^2(w_i)}} \int_{-\infty}^{\infty} \exp\left[-e^{\beta_{0j}+\beta_{1j}x_i} + y_i(\beta_{0j} + \beta_{1j}x_i)\right] \\ &\quad \exp\left\{-\frac{(x_i - m(w_i))^2}{2\tau^2(w_i)}\right\} dx_i. \end{aligned}$$

From Subsection 4.1.4, we can find the explicit expression of the density by the approximation method when measurement error is small. Therefore, the conditional density of the observed data is given by $\mathbf{Y} \mid \mathbf{W} = \mathbf{w} \sim \sum_{j=1}^{k} \lambda_j f_j(y_i \mid w, \theta_{jx_i})$, and the log-likelihood function for $\boldsymbol{\beta} = (\boldsymbol{\beta}_1^T, \cdots, \boldsymbol{\beta}_k^T)$ is

$$\ell_{\text{AMLE}}(\boldsymbol{\beta}) = \sum_{i=1}^{n} \log\left\{\sum_{j=1}^{k} \lambda_j f_j(y_i \mid w_i, \theta_{jx_i})\right\}. \tag{4.21}$$

Once we have the expression of the log-likelihood function, we can obtain the approximated maximum likelihood estimator $\hat{\boldsymbol{\beta}}_{\text{AMLE}}$, by finding the maximizer of the log-likelihood function given above in (4.21).

One thing to notice is that this estimator can only be used when measurement error is small. When we have large measurement error, the approximating condition cannot be reached, so the estimation may not be appropriate.

### 4.2.6 EM Algorithm

Define the vector of component indicators $\mathbf{Z}_i = (\mathcal{Z}_{i1}, \cdots, \mathcal{Z}_{ik})^T$, where $\mathcal{Z}_{ij}$ is the indicator random variable

$$\mathcal{Z}_{ij} = \begin{cases} 1, & \text{if observation } (\mathbf{w}_i, \mathbf{y}_i) \text{ is from the } j\text{th component;} \\ 0, & \text{otherwise.} \end{cases}$$

Because of the mixture setting, the complete data $\{\mathbf{w}, \mathbf{y}, \mathbf{z}\}$ cannot be obtained directly. Thus, it is suggested to use an EM algorithm to find the maximum of the log-likelihood functions proposed above for estimating the parameters.

Let $\boldsymbol{\psi} = (\boldsymbol{\beta}, \boldsymbol{\lambda})^T$ be the vector of parameters, we propose the following generalized EM algorithm in order to solve for the maximum of the log-likelihood.

---

**Algorithm 4.1** EM Algorithm for Mixtures-of-Poisson Regression

---

(a) Set the starting values of parameters as $\boldsymbol{\psi}^{(0)} = (\boldsymbol{\beta}^{(0)}, \boldsymbol{\lambda}^{(0)})^T$.

(b) **(E-Step)** Calculate component membership probabilities $p_{ij}^{(t+1)}$s by the expectation of $\mathcal{Z}_{ij}$, the weight of observation $i$ belonging to the $j$th component:

$$p_{ij}^{(t+1)} = \mathbb{E}[\mathcal{Z}_{ij} \mid \boldsymbol{\psi}^{(t)}, y] = \frac{\lambda_j^{(t)} f_j(y_i \mid w_i, z_j{}^{(t)})}{\sum_{j=1}^{k} \lambda_j^{(t)} f_j(y_i \mid w_i, z_j{}^{(t)})},$$

for $i$ in $1, \cdots, n$, $j$ in $1, \cdots, k$.

(c) **(M-Step)** The maximizer for $\lambda_j$ can be calculated by

$$\lambda_j^{(t+1)} = \frac{1}{n} \sum_{i=1}^{n} p_{ij}^{(t+1)}.$$

The maximizer for $\boldsymbol{\beta}_j$ is

$$\boldsymbol{\beta}_j^{(t+1)} = \arg\max_{\boldsymbol{\beta}} \sum_{i=1}^{n} p_{ij}^{(t+1)} \log f_j(y_i \mid w_i, z_j{}^{(t)}).$$

Therefore, the maximizer for $\boldsymbol{\beta}_j$ is the solution of

$$0 \;=\; \frac{\partial \ell(\boldsymbol{\beta}, \boldsymbol{\lambda})}{\partial \boldsymbol{\beta}_j} = \sum_{i=1}^{n} p_{ij}^{(t+1)} \frac{\partial \log f_j(y_i \mid w_i, z_j{}^{(t)})}{\partial \boldsymbol{\beta}_j}.$$

(c) Iterate until a *stopping criterion* is attained. The final estimate obtained will be denoted by $\hat{\boldsymbol{\psi}} = \left(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\lambda}}\right)^T$.

---

For different estimating methods, the conditional density function $f_j(y_i \mid w_i, z_j{}^{(t)})$ are different, according to the corresponding log-likelihood function. As we know, the likelihood function is obtained by the product of density functions. We can then get the single conditional density from the formulas given in the previous Subsections.

When using different log-likelihood functions in an EM algorithm, we can calculate the corrected score estimator, structural estimator, or the AMLE discussed in Subsections 4.2.3, 4.2.4 and 4.2.5, respectively.

## 4.3 Numerical Studies and Real Data Analyses

To see the behavior of our proposed methods, we conduct various simulation studies in this section.

### 4.3.1 Simulated Data — number of components

Like mixtures-of-linear-regressions, the first thing we want to identify is the correct number of components for the model. To test whether mixtures-of-Poisson regression model can correctly select the number of components, we first conduct a simulation study to compare the performance of different methods for determining the correct number of components using different model selection criteria, including AIC, BIC, cAIC and ICL.

Consider the simulated data with the 2-component mixture of Poisson regression model

$$y_i \sim \lambda e^{-\theta_{i1}} (\theta_{i1})^{y_i} / y_i! + (1-\lambda) e^{-\theta_{i2}} (\theta_{i2})^{y_i} / y_i!.$$

where $\theta_{ij} = \exp(\mathbf{x}_i \boldsymbol{\beta}_j)$ for $j = 1, 2$ with $\mathbf{x}_i = (1, x_i)^T$ and the explanatory variable $X$ is drawn from $N(\mu, \sigma^2)$. Instead of observing the $x_i$s directly, the surrogate, $w_i$, is given by the classical measurement error model

$$w_i = x_i + u_i,$$

where $u_i$ and $x_i$ are independent, $u_i$ follows a normal distribution with mean 0, and $U \sim N(0, \sigma_u^2)$ for $i = 1, \cdots, n$.

We consider a sample with sample size $n = 200$ for 1000 replications, with two different settings: well-separated and moderately-separated cases. We will generate $X \sim N(5, 1)$ with $\lambda = 0.3$, for well-separated component, $\boldsymbol{\beta}_1 = (2, 0.6)^T$ and $\boldsymbol{\beta}_2 = (0.7, 0.3)^T$ and for moderately-separated components, $\boldsymbol{\beta}_1 = (1, 0.5)^T$ and

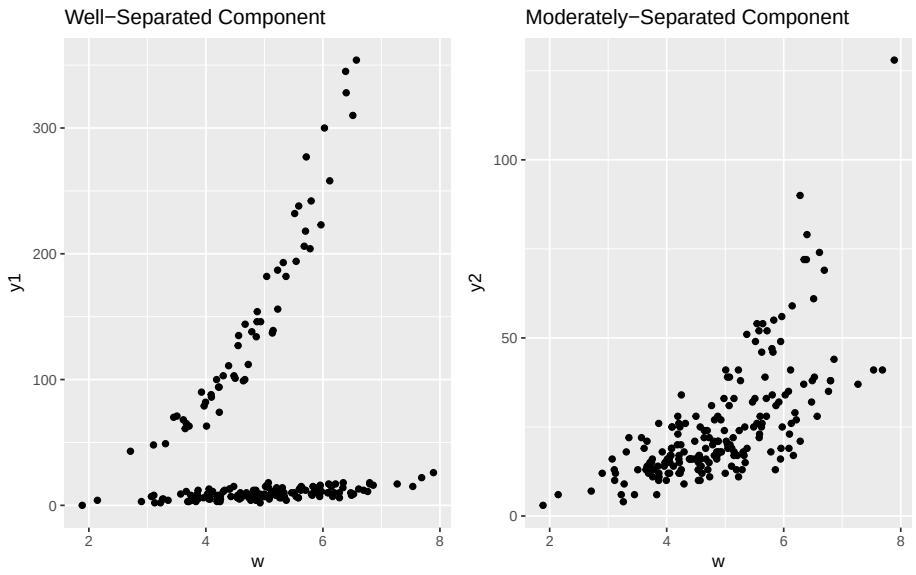


Figure 4.1: Scatterplots of simulated data from different settings.

$\boldsymbol{\beta}_2 = (1.2, 0.35)^T$. For each setting, we also add different amounts of measurement error, both $\sigma_u = 0.1$ and 0.5. Figure 4.1 shows the scatterplots of the simulated data points from different settings. The well-separated component model has two distinct components that can be easily separated; while the moderately-separated component model has two components with some part overlapped.

To see the behavior of model selection criteria for different methods under different circumstances, we fit the data with simple a Poisson regression as well as mixtures-of-Poisson regression with 2, 3 and 4 components, using the different estimators. The percentage of 1000 replications selecting the correct model is calculated for all criteria. Table 4.1 gives the results of the different methods and their performance.

Different estimating methods all do a good job in correctly identifying the correct number of components. When measurement error is small ($\sigma_u = 0.1$), the methods can always select the correct number of components; when we increase the measurement error, the model has more instability, and makes the ability to discern different components not as accurate. However, for most of the cases with a reasonable amount of measurement error, our methods can do a good job in selecting the appropriate model. Hence, we can move forward to estimate the parameters under the appropriate assumption for the number of components.

Table 4.1: Percentage of times different methods selected the correct model.

| Method | $\sigma_u$ | AIC | BIC | ICL | cAIC |
|---|---|---|---|---|---|
| **Well-Separated Components** | | | | | |
| CS | 0.1 | 100% | 100% | 100% | 100% |
| Structural | | 100% | 100% | 100% | 100% |
| CS | 0.5 | 99.1% | 99.4% | 99.4% | 99.4% |
| Structural | | 93.3% | 94.3% | 94.3% | 94.4% |
| **Moderately-Separated Components** | | | | | |
| CS | 0.1 | 100% | 100% | 100% | 100% |
| Structural | | 99.8% | 100% | 100% | 100% |
| CS | 0.5 | 96.6% | 97.5% | 97.5% | 97.6% |
| Structural | | 96.7% | 98.5% | 98.5% | 98.9% |

### 4.3.2 Simulated Data — estimators using different methods

We next assess the performance of estimating parameters using the different methods. We compare the MSEs of the parameters from our methods with the values from the naïve method, which is the setting where we simply ignore measurement errors.

Suppose the response variable follows a 2 - component mixture of Poisson regression. We generate the i.i.d data $(x_i, y_i, \eta_i)$, $i = 1, \cdots, n$ from the model

$$Y_i \sim \lambda_1 \text{Poi}\left\{\exp\left(\beta_{10} + \beta_{11}\text{X}_\text{i}\right)\right\} + \lambda_2 \text{Poi}\left\{\exp\left(\beta_{20} + \beta_{21} X_i\right)\right\}$$

$$W_i = X_i + U_i,$$

where $U_i \sim N(0, 1)$, $\lambda_1 = \lambda_2 = 0.5$ are the mixing proportions, $X_i \sim N(5, 1)$. Let $\boldsymbol{\beta}_1^T = (\beta_{10}, \beta_{11}), \boldsymbol{\beta}_2^T = (\beta_{20}, \beta_{21})$. Assume we know the correct number of components. We fit the simulated data set using different methods and record the values of estimators. To study the effect of the measurement error $u_i$s on the proposed estimator, we consider the following two cases, apply them to different methods and compare the behaviors of each method.

#### Case I: Well-separated Components

For the well-separated components case, suppose we have the parameters with values

$$\boldsymbol{\beta}_1^T = (1, 0.6), \boldsymbol{\beta}_2^T = (0.7, 0.5).$$

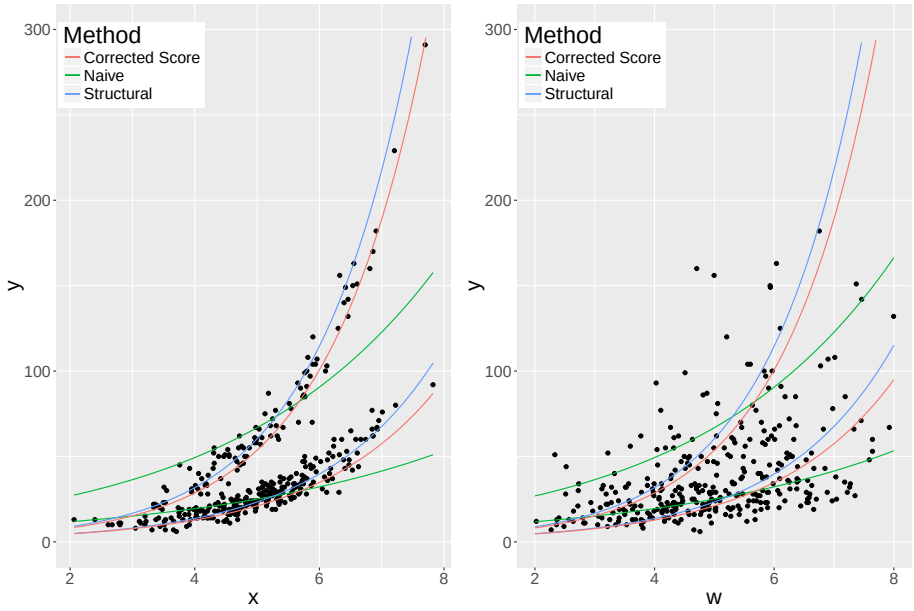


Figure 4.2: Scatterplots and fitted lines for well-separated case with different methods.

The left-hand side scatter plot in Figure 4.2 shows the relationship between the true predictors and the response variables, with sample size $n = 350$. Since this is a well-separated case, we can see two distinct component, which have noticeably different curvature. Because of the existence of measurement error, it is impossible for us to observe the true predictors $\mathbf{x}$ directly; instead, we observe the surrogate $\mathbf{w} = \mathbf{x} + \mathbf{u}$. The right-hand side scatterplot shows the relationship between the observed surrogate and the response variable. As we can see, the measurement error makes it more challenging to distinguish different components.

To see how the measurement error could affect the regression process, we fit the observed data set $\{\mathbf{w}, \mathbf{y}\}$ using the naïve method, the structural method, and the corrected score method, the corresponding fitted lines are shown in Figure 4.2. Green lines are the fitted lines from the naïve method. Comparing to the other two methods, it performs worse when trying to correctly represent the true model with predictor variables $\mathbf{x}$. The two proposed methods, on the other hand, both capture the true model pretty well, according to the scatterplot.

### Case II: Moderately-Separated Components

Now we consider a moderately-separated components case. We modify the parameter values to

$$\boldsymbol{\beta}_1^T = (1, 0.5), \boldsymbol{\beta}_2^T = (1.2, 0.35).$$

According to the scatterplots, the difference between this case and the previous one is the structure of the data points. For this case, all the data points demonstrate heavy mixing, thus making the original data set more challenging for estimating a model.

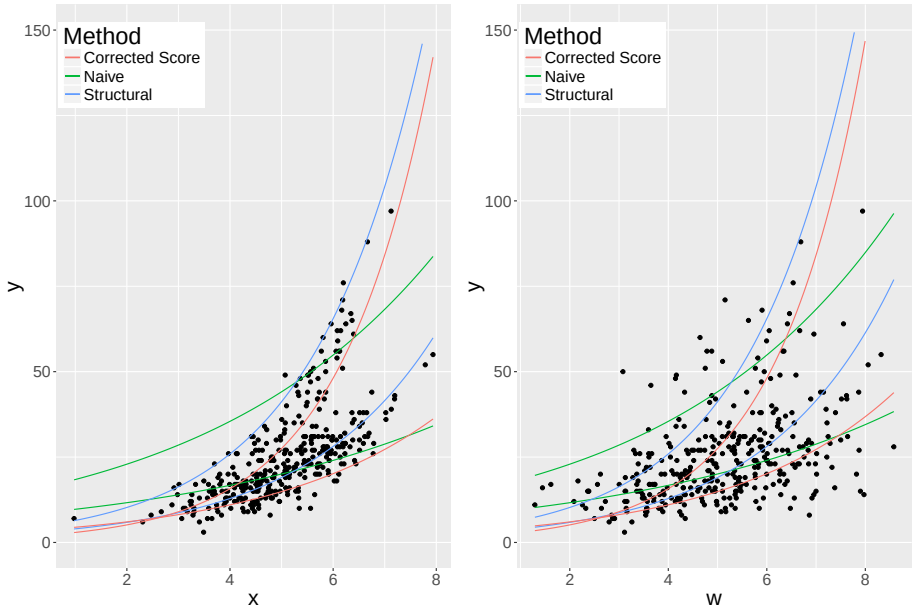


Figure 4.3: Scatterplots and fitted lines for moderately-separated case with different methods.

Similarly, Figure 4.3 shows the scatterplots and fitted lines for the moderately-separated components case. Like the well-separated case, the methods being investigated have a better performance in estimating the parameter values; however, it seems like there are some discrepancies between the two proposed methods. To investigate the differences, we conduct a thorough simulation study to see the behavior of each method and compare them using some standard.

**Simulation and Results**

For each simulation condition, we randomly generated $B = 1000$ datasets, each of size $n = 200$ and 350. For each simulated dataset, we estimated the mixture of regression parameters $\boldsymbol{\beta}_1$, $\boldsymbol{\beta}_2$ by both the structural method and the corrected score method. The accuracy of the proposed method under different conditions is assessed by the *mean squared error* (MSE).

To compare the performance of the estimation methods versus the naïve method, we report the relative efficiency of the MSEs between the naïve method and the proposed methods (corrected score and structural). This calculation involves simply taking the ratio of the MSEs of the naïve method to that of the proposed estimators for our simulated datasets.

Table 4.2: The MSEs and relative efficiencies of naïve method vs. proposed methods.

| $n$ | Method | $\beta_{10}$ | $\beta_{11}$ | $\beta_{20}$ | $\beta_{21}$ |
|---|---|---|---|---|---|
| **Well-Separated Components** | | | | | |
| 200 | Naïve vs. CS | 0.793 (2.744) | 0.053 (1.838) | 0.249 (13.786) | 0.005 (13.354) |
| | Naïve vs. Structural | 0.482 (4.724) | 0.016 (5.386) | 0.715 (3.894) | 0.009 (6.889) |
| 350 | Naïve vs. CS | 0.797 (3.039) | 0.053 (1.829) | 0.249 (11.316) | 0.005 (12.049) |
| | Naïve vs. Structural | 0.280 (8.809) | 0.010 (9.343) | 0.545 (4.942) | 0.005 (12.027) |
| **Moderately-Separated Components** | | | | | |
| 200 | Naïve vs. CS | 0.461 (4.545) | 0.053 (1.340) | 0.410 (2.921) | 0.015 (1.941) |
| | Naïve vs. Structural | 0.378 (5.687) | 0.015 (4.718) | 0.360 (3.128) | 0.006 (4.269) |
| 350 | Naïve vs. CS | 0.466 (4.445) | 0.051 (1.443) | 0.400 (3.148) | 0.015 (1.963) |
| | Naïve vs. Structural | 0.242 (9.220) | 0.008 (8.300) | 0.199 (5.030) | 0.003 (7.965) |

Table 4.2 shows the MSEs of the proposed methods, as well as the relative efficiencies (in parentheses). From the output, both the corrected score method and

the structural method perform better than simply ignoring the measurement error, which is consistent with the scatterplots.

Overall, the corrected score method is more sensitive to the structure of the data. When the components demonstrate heavier mixing, the corrected score has less power than the structural method, which has the assumption of the distribution of the variables. The structural method appears more stable, despite the structure of the data set. When increasing the sample size, the behavior of the corrected score method has little improvement, however, the structural method has a much smaller MSE as the sample size increases.

The corrected score method appears more accurate when dealing with case with smaller parameter values, and the structural method depends heavily on the distribution of the variables. There appears to always be a trade-off to determine which method is more appropriate under different circumstances. When we have a small sample size with a relatively well-separated data set, it is better to use the corrected score method. When we have a larger sample size with more complex structure, the structural method with the assumption of the distribution appears to be more appropriate.

### 4.3.3 Approximated Maximum Likelihood Estimator

Unlike the structural estimator and the corrected score estimator, the approximated maximum likelihood estimator is more sensitive to the measurement error; it can only be applied when measurement error is not too large. When the measurement error is not too large, the behavior of the AMLE may be unstable compared to the naïve estimator. To see how the AMLE performs when measurement error is incorporated, we also simulate data from two different settings with different sample sizes, and add a small amount of measurement error to the true predictors. Like the previous simulation study, we simulate data from two-component mixtures-of-Poisson regressions with two settings: well-separated and moderately-separated cases. We generate the predictors $X \sim N(10, 3)$ with $\lambda = 0.3$, for well-separated component, $\boldsymbol{\beta}_1 = (1, 0.16)^T$ and $\boldsymbol{\beta}_2 = (2, -0.2)^T$, and for moderately-separated components,

$\boldsymbol{\beta}_1 = (1, 0.16)^T$ and $\boldsymbol{\beta}_2 = (3, -0.11)^T$. For each data set generated, we add a small amount of measurement error, $\sigma_u = 0.25$. To see how the sample size may affect the estimating process, we consider different sample sizes, both $n = 200$ or 350.

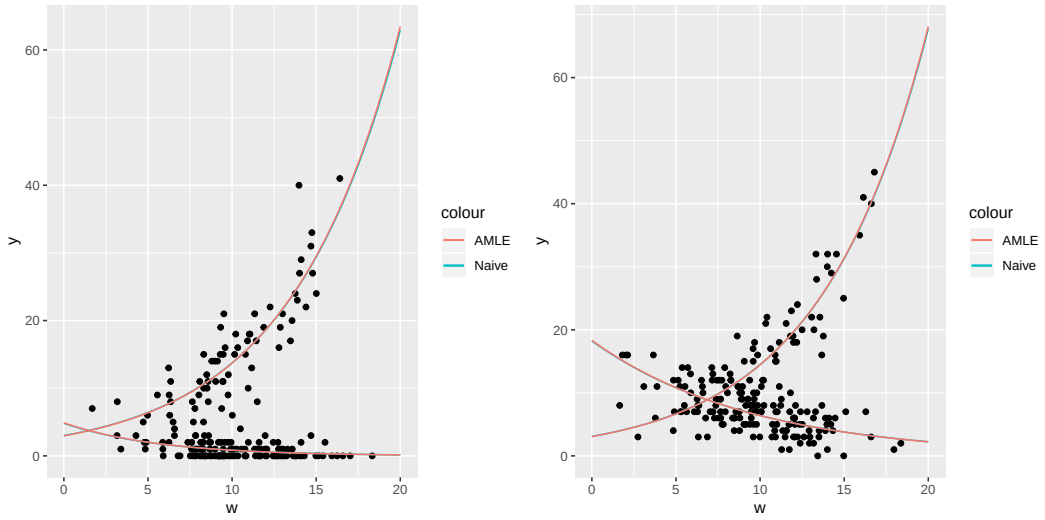


Figure 4.4: Scatterplots and fitted lines from both AMLE method and naïve method, under different settings.

Figure 4.4 shows the scatter plots of the simulated data points from different settings, with fitted lines from both the AMLE method and the naïve method. The left-hand figure gives the well-separated component model, which has two distinct components can be easily separated; while the moderately-separated component model (right-hand side) has two components with several sections where the data from the different components overlap.

Since the measurement error is relatively small ($\sigma_u = 0.25$) compared to the standard deviation of the true predictors ($\sigma_x = 3$), the difference between the naïve method and the AMLE method is relatively small. Based on the scatterplots, the curves plotted from the two estimation methods look similar, which means they have returned similar results for both cases.

When measurement error is small, the effect of the surrogate becomes trivial, however, for some cases that requires an accurate result, we still want to take measurement error into consideration. To see the performance of the AMLE over the naïve method, we also conduct a simulation study, with replicates $B = 1000$. For

each replicate, we generate the sample from the previous setting, and similarly, compare the MSEs of the naïve method over the AMLE method by reporting the relative efficiency.

Table 4.3: MSEs and relative efficiency, naïve vs. AMLE.

| $n$ | $\beta_{10}$ | $\beta_{11}$ | $\beta_{20}$ | $\beta_{21}$ |
|---|---|---|---|---|
| **Well-Separated Components** | | | | |
| 200 | 0.022 | $1.62 \times 10^{-4}$ | 1.042 | $8.53 \times 10^{-4}$ |
| | (1.007) | (1.006) | (1.019) | (1.008) |
| 350 | 0.014 | $9.96 \times 10^{-5}$ | 1.018 | $4.93 \times 10^{-4}$ |
| | (1.001) | (1.001) | (1.019) | (1.006) |
| **Moderately-Separated Components** | | | | |
| 200 | 0.035 | $2.32 \times 10^{-4}$ | 3.957 | $1.48 \times 10^{-4}$ |
| | (1.002) | (1.003) | (1.005) | (1.005) |
| 350 | 0.017 | $1.11 \times 10^{-4}$ | 3.961 | $8.13 \times 10^{-5}$ |
| | (1.005) | (1.006) | (1.005) | (1.001) |

Table 4.3 shows the MSEs and relative efficiencies of the naïve method versus the AMLE method. When we increase the sample size for both cases, the MSEs become smaller, as more data points provide more information about the structure of the data. Overall, the estimators have very small MSEs, which means the estimators perform very well in estimating those parameter values.

The relative efficiencies are relatively close to 1, which means the AMLE has little difference in estimating the parameters. One of the main reasons is that, the standard deviation of the measurement error is relatively small, and this results in only moderate differences in the estimates. For example, the MSE of the well-separated components case with sample size $n = 350$ for $\beta_{11}$ is $9.96 \times 10^{-5}$, which is already quite small. However, if measurement error is known to be present, then it should be taken into consideration during estimation.

### 4.3.4 The Relationship between Pseudoephedrine Sales and Methamphetamine Labs — A Real Data Analysis

Illicit production of methamphetamine from the *precursor chemical pseudoephedrine* (PSE) in clandestine laboratories (labs) has produced health risks to society. Although states and the federal government have taken varying regulatory approaches to control access to PSE, the illicit production of methamphetamine in clandestine labs continues. The total number of domestic labs seized in the United States (US) peaked in 2010 at 15,217 labs while the number of seized labs declined to 12,409 in 2013, and 9,306 in 2014.

A previous study has shown a strong statistical relationship between the sale of PSE (grams/100 residents) in community pharmacies and methamphetamine lab incidents reported in Kentucky in 2010, with counties recording larger sales of PSE significantly associated with greater numbers of clandestine labs. The response variable is lab count. The sale of PSE is a value that possibly suffers from measurement error, so we can investigate Poisson regression modeling in the presence of measurement error. To do the analysis, we utilize the data sets with PSE sales and methamphetamine lab incidents in Kentucky, Illinois, and Louisiana in 2012.

Figure 4.5 shows the scatterplot of the data. There appears to be multiple relationships that could underlie these data, that is, we can fit the model with the mixtures-of-Poisson regression model. The PSE sales can be considered as a variable suffers from measurement error, so we also want to incorporate measurement error in the data analysis.

To see the impact of measurement error, we add a measurement error $N(0, 5)$ to the predictor $X$, which is PSE in this case consistent with the type of analysis performed in Yao and Song (2015) [119]. Denote by $W$ the surrogate of PSE, $Y$ the corresponding response – lab count. Firstly, we want to determine the appropriate number of components for this model. Similar to the simulation study, we fit the data with simple Poisson regression, as well as mixtures-of-Poisson regression with 2, 3 and 4 components, using both corrected score method and structural method.

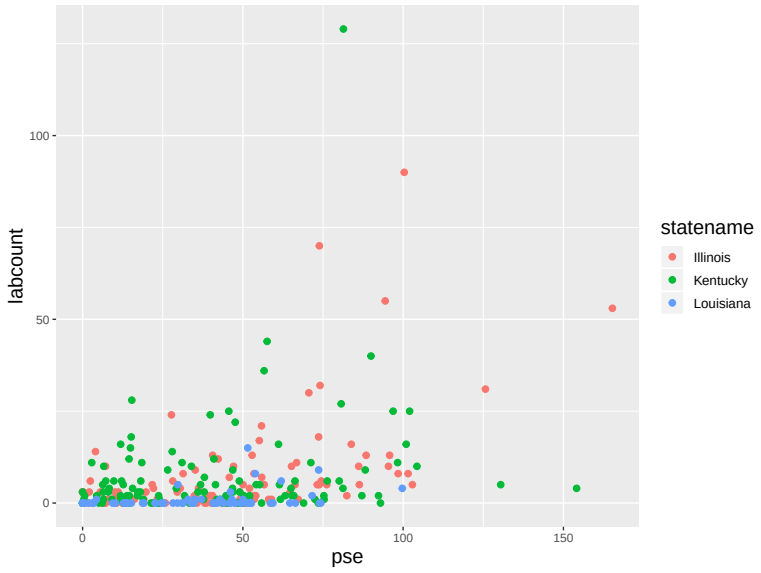


Figure 4.5: The scatterplot of the original PSE sales data and the fitted regression lines by different methods when the measurement error is added.

Table 4.4 shows the corresponding values for different model selection criteria, the bold values are the model selected.

Table 4.4: Values of model selection criteria calculated by different estimating methods.

| Method | $k$ | AIC | BIC | ICL | cAIC |
|---|---|---|---|---|---|
| Corrected Score | 1 | 3563.600 | 3570.912 | 3570.912 | 3572.912 |
| | 2 | 1889.156 | 1907.436 | 1908.443 | 1912.436 |
| | 3 | **1689.965** | **1719.213** | **1721.071** | **1727.213** |
| | 4 | 1746.237 | 1786.453 | 1788.833 | 1797.453 |
| Structural | 1 | 3580.441 | 3587.753 | 3587.753 | 3589.753 |
| | 2 | 1915.970 | 1934.235 | 1935.210 | 1939.235 |
| | 3 | **1702.499** | **1731.746** | **1733.596** | **1739.746** |
| | 4 | 1709.682 | 1749.898 | 1751.760 | 1760.898 |

The bold values represent the selected number of components, based on each model selection criteria. For all the different criteria, they select the same number of components: $k = 3$. While the corrected score has a relatively smaller value for $k = 3$, the structural method does not have such a big difference between 3 and 4 components. Meanwhile, we can look at the fitted lines of all different number of components, and the graph also shows that 3 components appears reasonable. Hence,

we will focus on the mixtures-of-Poisson regression model with 3 components in the next step.

We fit the data $(W, Y)$ with 3-component mixtures-of-Poisson regression model using both the naïve method, which ignores the measurement error, and the proposed methods, corrected score and structural estimators. For comparison, we also add an oracle method which uses the $(X, Y)$ directly. We plot the fits in Figure 4.6, as expected, the three components fit well with the data and can reflect some properties of the data set. Because the measurement error is not too large, according to the scatterplot, all the four methods have relatively close results, while the regression lines estimated by the new methods are closer to the oracle method, and the naïve estimate has some bias for both of the fitted lines.

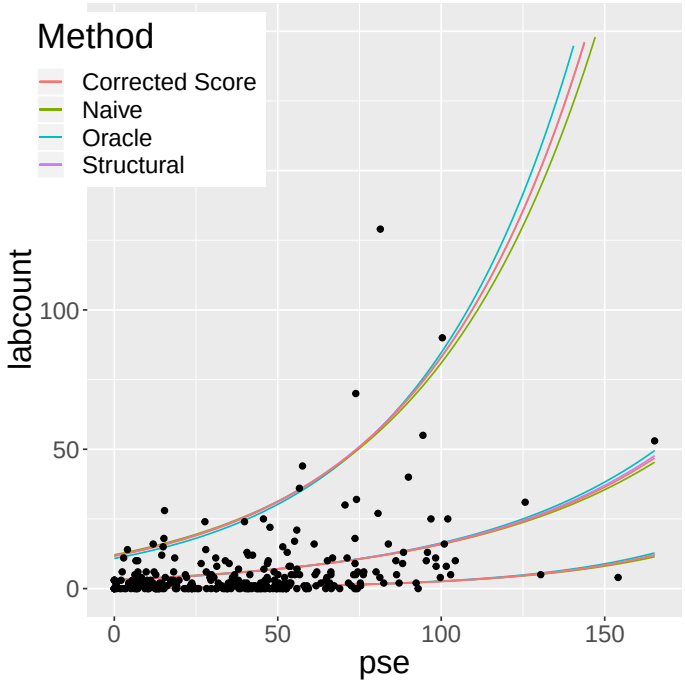


Figure 4.6: Scatterplot of meth lab data and the fitted lines from different methods.

Table 4.5 reports the mixtures-of-Poisson regressions parameter estimates for all methods described above, as well as the bootstrap SEs (in parentheses). The oracle method is the closest to the true model, and both the structural and corrected score methods have relatively similar results compared to the oracle method. The standard errors of all the parameters are quite small, which also shows that our methods

perform well in the setting above. Because we choose a measurement error that is not very large, the differences between the naïve method and the proposed methods are subtle. However, we can still see some discrepancies between the estimates, for example the slope of the third component, $\beta_{31}$, is 0.020 in the oracle method; however, the naïve method has a value of 0.018, which is a little smaller. The differences between the intercepts are slightly larger.

Table 4.5: Regression parameter estimates for the meth lab data with measurement error.

| | $\beta_{10}$ | $\beta_{11}$ | $\beta_{20}$ | $\beta_{21}$ | $\beta_{30}$ | $\beta_{31}$ |
|---|---|---|---|---|---|---|
| Oracle | −1.428 | 0.024 | 1.100 | 0.016 | 2.372 | 0.020 |
| | (0.232) | (0.002) | (0.104) | (0.001) | (0.104) | (0.001) |
| Naïve | −1.441 | 0.023 | 1.103 | 0.016 | 2.511 | 0.018 |
| | (0.244) | (0.002) | (0.096) | (0.001) | (0.095) | (0.001) |
| Structural | −1.444 | 0.023 | 1.015 | 0.017 | 2.309 | 0.021 |
| | (0.237) | (0.002) | (0.101) | (0.001) | (0.101) | (0.001) |
| Corrected Score | −1.541 | 0.022 | 0.895 | 0.018 | 2.292 | 0.020 |
| | (0.269) | (0.003) | (0.108) | (0.001) | (0.099) | (0.001) |

The original data points are collected from three different states. We fit the model with 3-component mixtures-of-Poisson regressions, however, the three components cannot be interpreted simply by three different states. Based on the data, there are some states that basically follow the component with smaller outcomes, for example, Louisiana, and some states have more data points with larger number of lab counts. There are some potential reasons that may cause the data set to have different components:

- 3 different components could be characterizing trends in counties that have varying degrees of public policy on how to address the problem of meth labs.
- These could also represent the availability of support by local law enforcement. For example, the curve capturing the larger number of lab counts could be counties that generally have lower number of law enforcement officials to actively capture the presence of meth labs. Thus, manufacturers of meth would be naturally drawn to such a region.

Another thing that can be addressed is, although in our model we assume each component has measurement errors with the same standard deviations. In reality, different components and even different subtle of the data (e.g., representing different counties) may have different amounts of measurement error. This will make the model even more complex. To maximize the utility of such a mixture model, one would have to think of how best to characterize such scenarios before analyzing the real data set.

## 4.4 Summary

In this chapter, we focused on the mixtures-of-Poisson regression models with measurement error in the predictors. We expand the classical measurement error methods — corrected score method and structural method — into the mixture setting. We also developed a density-based algorithm to compute the likelihood function, which is called the approximated maximum likelihood method. We conducted a series of simulations to see the effect of measurement error in the model, and identified that the AMLE performs well only when the measurement error is small, while the corrected score estimator and the structural estimator have better performance under a broader range of conditions.

The real data analysis demonstrated the relationship between the sales of PSE and the number of lab seizures. According to the model selection criteria, it is appropriate to use a 3-component mixture model, and we identified several possible reasons that might result in this structure. The PSE sales, which is measured by electronic devices, may suffer from measurement error, hence, we also add a fixed amount of variability to each predictor variable. We futher compared the behaviors of different methods under this situation. The results of the data analysis shows our methods perform quite well under the existence of measurement error. This underscores the importance of accounting for measurement error in such a real data analysis.

# Chapter 5 Summary and Future Work

In this Chapter, we briefly summarize the content in this dissertation, and also discuss some additional problems that can be solved in the future.

## 5.1 Summary

In this dissertation, we considered the case when mixtures-of-regression models are observed with measurement error. The measurement error, an inaccuracy introduced to the observed data set, may leads to the inconsistency of the estimator. We discuss several different mixtures-of-regression models that may suffer from measurement error, introduce and develop some methods in estimating the parameters, and also use them for some real data applications.

First, we discuss the mixtures-of-linear-regression model with measurement errors in the response. We extend a weighted least squares estimator developed by Akritas and Bershady to the mixture setting to adjust for the measurement error, and compute the asymptotic variance of the estimators by Fisher information.

In the second part, we discuss the mixtures-of-linear-regression model with measurement errors in the predictors. By incorporate measurement error in the predictor, we built the new conditional density function, under certain assumption. Using this estimating method, we construct hypothesis test on polynomial regression with measurement error, and compute the standard errors of estimators by bootstrap method. The performance of the method is tested by series of simulation studies and a real data analysis.

Finally, we discussed different estimating methods in Poisson regression with measurement errors, including structural method, corrected score method and approximated likelihood estimation method. We expand these estimating methods to the mixture setting, and compared each method by a series of tests. To test the performance of the estimating methods, we generated data points with different setting, fit

the data using different method and compare them using different metric. We also used them in a real data analysis. The results showed we should carefully choose the appropriate method when dealing with different problems.

## 5.2 Future Work

There are many different models, tests, and analyses that can be explored for future research. Future work can be done by generalizing our analysis to some other models, identifying new and relevant tests for certain types of model structures. There are some ideas that I identified, during the development of the methods in the previous chapters. The following ideas could be some directions that can be done in later research:

1. It could be interesting to consider different regression models having measurement error present, for example, mixtures of logistic regression or mixtures of negative binomial regression. Once we have gained the knowledge of the basic mechanism of the model, we can easily solve for these different types of models.

2. We can do further research by accessing measurement error in the count response, in addition to the measurement error in the continuous response variables.

3. In Chapter 3, we highlighted a semi-parametric efficient score method for measurement error in the predictor problem, however, due to the lack of time, we are unable to expand the theoretical result to the mixture setting. We hope future work can be done with this problem, as it is an interesting topic for later researchers.

4. Additional theoretical work can be done, for example, under mixture-of-regression setting, some regularization rules no longer hold. We can try to prove the consistency of the model under the mixture setting, given that it is already proved in the non-mixture setting.

Regarding the real data analyses we presented, there are always different ways to interpret the same data set. In this dissertation, due to the complexity with addressing measurement error in a data set, some estimating methods cannot be used appropriately under certain conditions. We hope researchers find the utility of our work and that these methods and tools help inform applications across numerous different fields.

## References


[1] Wayne A. Fuller. *Measurement error models*. John Wiley & Sons, Inc., 1987.

[2] Jerry A. Hausman. Mismeasured variables in econometric analysis: Problems from the right and problems from the left. *Journal of Economic Perspectives*, 15:57–67, 11 2001.

[3] A. C. Aitken. On least squares and linear combination of observations. *Proceedings of the Royal Society of Edinburgh*, 55:42–48, 1935.

[4] M Aitkin and D B. Rubin. Estimation and hypothesis testing in finite mixture models. *Journal of the Royal Statistical Society. Series B*, 47, 01 1985.

[5] H. Akaike. Information theory and an extension of the maximum likelihood principle. *Proceedings of the 2nd International Symposium on Information Theory*, pages 267–281, 1973.

[6] T. Amemiya. *Advanced Econometrics*. Harvard University Press, 1985.

[7] Rene Andrae. Error estimation in astronomy: A guide. 09 2010.

[8] Timothy L. Bailey and Charles Elkan. Fitting a mixture model by expectation maximization to discover motifs in biopolymers. *Proceedings. International Conference on Intelligent Systems for Molecular Biology*, 2:28–36, 1994.

[9] Scott D. Barthelmy. Burst alert telescope (bat) on the swift midex mission. *Proc.SPIE*, 5165:5165–5165, 2006.

[10] Mikhail Belkin and Kaushik Sinha. Polynomial learning of distribution families. *Proceedings of the 2010 IEEE 51st Annual Symposium on Foundations of Computer Science*, pages 103–112, 2010.

[11] T. Benaglia, D. Chauveau, D. R. Hunter, and D. S. Young. mixtools : An r package for analyzing mixture models. *Journal of Statistical Software*, 32(6): 1–29, 2009.

[12] Christophe Biernacki, Gilles Celeux, and Gérard Govaert. Assessing a mixture model for clustering with the integrated completed likelihood. *IEEE Trans. Pattern Anal. Mach. Intell.*, 22(7):719–725, July 2000.

[13] Laurent Bordes, Celine Delmas, and Pierre Vandekerkhove. Semiparametric estimation of a two-component mixture model where one component is known. *Scandinavian Journal of Statistics*, 33(4):733–752, 2006.

[14] Hamparsum Bozdogan. Model selection and akaike's information criterion (aic): The general theory and its analytical extensions. *Psychometrika*, 52(3):345–370, Sep 1987.

[15] N. E. Breslow. Extra-poisson variation in log-linear models. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, 33(1):38–44, 1984.

[16] Norman D. Brinkman. Ethanol fuel-a single-cylinder engine study of efficiency and exhaust emissions. 02 1981.

[17] Anthony C. Davison and D Hinkley. Bootstrap methods and their application. *Journal of the American Statistical Association*, 94, 01 1997.

[18] David C. Morris, David N. Burrows, Joanne Hill, Jamie Kennea, Judith Racusin, Paul Wood, C Mangels, R Klar, Lorella Angelini, and Francesca Tamburelli. Temperature dependent calibration products of the swift x-ray telescope. *Proceedings of SPIE - The International Society for Optical Engineering*, 5898, 08 2006.

[19] A Cameron and Pravin K. Trivedi. *Essentials of Count Data Regression*. 01 1999.

[20] C. Cao and X. Zhu. A structural errors-in-variables model with heteroscedastic measurement errors under heavy-tailed distributions. *2011 Fourth International Conference on Information and Computing*, pages 461–463, April 2011.

[21] Miguel Á. Carreira-Perpiñán and Christopher K. I. Williams. On the number of modes of a gaussian mixture. *Proceedings of the 4th International Conference on Scale Space Methods in Computer Vision*, pages 625–640, 2003.

[22] Raymond J. Carroll, Clifford H. Spiegelman, K. K. Gordon Lan, Kent T. Bailey, and Robert D. Abbott. On errors-in-variables for binary regression models. *Biometrika*, 71(1):19–25, 1984.

[23] R.J. Carroll, D. Ruppert, and L.A. Stefanski. *Measurement Error in Nonlinear Models*. Chapman & Hall/CRC Monographs on Statistics & Applied Probability. Taylor & Francis, 1995.

[24] Alexandre X. Carvalho and Martin A. Tanner. Modeling nonlinear time series with local mixtures of generalized linear models. *Canadian Journal of Statistics*, 33(1):97–113, 2009.

[25] Gilles Celeux, Merrilee Hurn, and Christian P. Robert. Computational and inferential difficulties with mixture posterior distributions. *Journal of the American Statistical Association*, 95(451):957–970, 2000.

[26] Xiaohong Chen, Han Hong, and Denis Nekipelov. Nonlinear models of measurement errors. *Journal of Economic Literature*, 49(4):901–937, 2011.

[27] C.L. Cheng and J.W. Van Ness. *Statistical Regression with Measurement Error*. Wiley, 1998.

[28] Hwan Chung, Eric Loken, and Joseph L Schafer. Difficulties in drawing inferences with finite-mixture models. *The American Statistician*, 58(2):152–158, 2004.

[29] Patrizia Ciarlini, Maurice Cox, Franco Pavese, and Giuseppe Regoliosi. The use of a mixture of probability distributions in temperature interlaboratory comparisons. *Metrologia*, 41:116–121, 06 2004.

[30] M. Clutton-Brock. Likelihood distributions for estimating functions when both variables are subject to error. *Technometrics*, 9(2):261–269, 1967.

[31] J. R. Cook and L. A. Stefanski. Simulation-extrapolation estimation in parametric measurement error models. *Journal of the American Statistical Association*, 89(428):1314–1328, 1994.

[32] D. R. Cox. Some remarks on overdispersion. *Biometrika*, 70(1):269–274, 1983.

[33] P. Dellaportas D. A. Stephens. Bayesian inference of generalized linear models with covariate measurement errors. *Bayesian Statistics*, pages 813–820, 1992.

[34] R. D. de Veaux. Mixtures of linear regressions. *Comput. Stat. Data Anal.*, 8 (3):227–245, November 1989.

[35] Petros Dellaportas and David A. Stephens. Bayesian analysis of errors-in-variables regression models. *Biometrics*, 51(3):1085–1095, 1995.

[36] A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society. Series B (Methodological)*, 39(1):1–38, 1977.

[37] Viswanath Devanarayan and Leonard A. Stefanski. Empirical simulation extrapolation for measurement error models with replicate measurements. *Statistics & Probability Letters*, 59(3):219 – 225, 2002.

[38] Mathias Drton and Martyn Plummer. A bayesian information criterion for singular models. *International Agency for Research on Cancer, Lyon, France*, 2013.

[39] Bradley Efron and David V. Hinkley. Assessing the accuracy of the maximum likelihood estimator: Observed versus expected fisher information. *Biometrika*, 65(3):457–482, 1978.

[40] Bradley Efron and Robert J. Tibshirani. An introduction to the bootstrap. *Journal of the American Statistical Association*, 89:436, 01 1993.

[41] S. Faria and F. Gonï¿œalves. Financial data modeling by poisson mixture regression. *Journal of Applied Statistics*, 40(10):2150–2162, 2013.

[42] Z. D. Feng and C. E. McCulloch. Using bootstrap likelihood ratios in finite mixture models. *Journal of the Royal Statistical Society. Series B (Methodological)*, 58(3):609–617, 1994.

[43] Thomas S Ferguson. *A Course in Large Sample Theory*. Taylor I& Francis, 1996.

[44] R. Frisch, David F. Hendry, and Mary S. Morgan. *Statistical Confluence Analysis by Means of Complete Regression Systems (University Institute of Economics, Oslo, 1934, pp. 5?8)*. Cambridge University Press, 1995.

[45] Edward L. Frome, Michael H. Kutner, and John J. Beauchamp. Regression analysis of poisson-distributed data. *Journal of the American Statistical Association*, 68(344):935–940, 1973.

[46] Michael G. Akritas and Matthew A. Bershady. Linear regression for astronomical data with measurement errors and intrinsic scatter. *The Astrophysical Journal*, 470, 05 1996.

[47] N Gehrels, G Chincarini, P Giommi, K O. Mason, J Nousek, AA Wells, Nicholas White, Scott Barthelmy, D N. Burrows, Lynn Cominsky, K C. Hurley, F E. Marshall, P Mï¿œszï¿œros, P W. A. Roming, L Angelini, Louis Barbier, T Belloni, S Campana, P Caraveo, and W W. Zhang. The swift gamma?ray burst mission. *The Astrophysical Journal*, 611:1005–1020, 08 2004.

[48] Zoubin Ghahramani and Geoffrey E. Hinton. Parameter estimation for linear dynamical systems. 1996.

[49] Zvi Griliches and Vidar Ringstad. Error-in-the-variables bias in nonlinear contexts. *Econometrica*, 38(2):368–370, 1970.

[50] P. Gustafson. *Measurement Error and Misclassification in Statistics and Epidemiology: Impacts and Bayesian Adjustments*. 2004.

[51] J.A. Hausman, W.K. Newey, and J.L. Powell. Nonlinear errors in variables estimation of some engel curves. *Journal of Econometrics*, 65(1):205 – 233.

[52] Dollena S. Hawkins, David M. Allen, and Arnold J. Stromberg. Determining the number of components in mixtures of linear models. *Comput. Stat. Data Anal.*, 38(1):15–48, November 2001. ISSN 0167-9473.

[53] C C Heyde. *Quasi-likelihood and Its Application: A General Approach to Optimal Parameter Estimation*. New York: Springer, 1997.

[54] C. C. Heyde and R. Morton. Multiple roots in general estimating equations. *Biometrika*, 85(4):954–959, 1998.

[55] D. R. Hunter and D. S. Young. Semiparametric mixtures of regressions. *Journal of Nonparametric Statistics*, 24(1):19–38, 2012.

[56] Merrilee Hurn, Ana Justel, and Christian P Robert. Estimating mixtures of regressions. *Journal of Computational and Graphical Statistics*, 12(1):55–79, 2003.

[57] Clifford M. Hurvich, Jeffrey S. Simonoff, and Chih-Ling Tsai. Smoothing parameter selection in nonparametric regression using an improved akaike information criterion. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 60(2):271–293, 1998.

[58] A J. Blustin, D Band, Scott Barthelmy, P Boyd, M Capalbi, S T. Holland, F E. Marshall, K O. Mason, M Perri, T Poole, Peter Roming, Siv Rosï¿œn,

P Schady, M Still, Bryna Zhang, L Angelini, Louis Barbier, A Beardmore, Alice Breeveld, and Nicholas White. Swift panchromatic observations of the bright gamma-ray burst grb 050525a. *The Astrophysical Journal*, 637, 02 2006.

[59] Robert Jacobs, Michael Jordan, Steven J. Nowlan, and Geoffrey E. Hinton. Adaptive mixture of local expert. *Neural Computation*, 3:78–88, 02 1991.

[60] A. Jasra, C. C. Holmes, and D. A. Stephens. Markov chain monte carlo methods and the label switching problem in bayesian mixture modeling. *Statistical Science*, 20(1):50–67, 02 2005.

[61] P.N. Jones and G. J. McLachlan. Fitting finite mixture models in a regression context. *Australian Journal of Statistics*, 34(2):233–240, 1992.

[62] Michael I. Jordan and Robert A. Jacobs. Hierarchical mixtures of experts and the em algorithm. *Neural Computation*, 6:181–214, 1994.

[63] D.Sc. Joseph Berkson M.D. Are there two regressions? *Journal of the American Statistical Association*, 45(250):164–180, 1950.

[64] G K Reeves, David Cox, S C Darby, and Elise Whitley. Some aspects of measurement error in explanatory variables for continuous and binary regression models. *Statistics in medicine*, 17:2157–77, 11 1998.

[65] Adam Tauman Kalai, Ehud Kalai, Ehud Lehrer, and Dov Samet. A commitment folk theorem. *Games and Economic Behavior*, 69(1):127 – 137, 2010.

[66] B.C. Kelly. Some aspects of measurement error in linear regression of astronomical data. 665:1489–1506, 2007.

[67] Daeyoung Kim and Bruce Lindsay. *Annals of the Institute of Statistical Mathematics*, 67(4):745–772, 2015.

[68] Jouni Kuha. Estimation by dada augmentation in regression models with continuous and discrete covariates measured with error. *Statistics in Medicine*, 16 (2):189–201.

[69] Jouni Kuha and Jonathan Temple. Covariate measurement error in quadratic regression. *International Statistical Review*, 71, 02 2003.

[70] Jerald F. Lawless. Negative binomial and mixed poisson regression. *Canadian Journal of Statistics*, 15(3):209–225, 1987.

[71] Roman Liesenfeld. A generalized bivariate mixture model for stock price volatility and trading volume. *Journal of Econometrics*, 104(1):141 – 178, 2001.

[72] B.G. Lindsay. *Mixture Models: Theory, Geometry, and Applications*. Conference Board of the Mathematical Sciences: NSF-CBMS regional conference series in probability and statistics. Institute of Mathematical Statistics, 1995.

[73] Thomas A. Louis. Finding the observed information matrix when using the em algorithm. *Journal of the Royal Statistical Society. Series B (Methodological)*, 44(2):226–233, 1982.

[74] Yanyuan Ma, Jeffrey D. Hart, Ryan Janicki, and Raymond J. Carroll. Local and omnibus goodness-of-fit tests in classical measurement error models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 73(1):81–98, 2011.

[75] Peter McCullagh and John A Nelder. *Generalized linear models*, volume 28. 03 1986.

[76] G Mclachlan and K Basford. *Mixture Models: Inference and Applications to Clustering*. Marcel Dekker, New York, 1988.

[77] G. J. McLachlan. On bootstrapping the likelihood ratio test stastistic for the number of components in a normal mixture. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, 36(3):318–324, 1987.

[78] Geoffrey McLachlan and David Peel. *Finite Mixture Models*. John Wiley & Sons, Inc., 2000.

[79] Geoffrey J. McLachlan and T. Krishnan. *The EM algorithm and extensions.* Wiley-Interscience Hoboken, N.J, 1997.

[80] Geoffrey J. McLachlan and Thriyambakam Krishnan. *Wiley Series in Probability and Statistics.* John Wiley & Sons, Inc., 1997.

[81] Ankur Moitra and Gregory Valiant. Settling the polynomial learnability of mixtures of gaussians, 2010.

[82] Heather L. Morrison, Edward W. Olszewski, Mario Mateo, John E. Norris, Harding. Paul, R. C. Dohm-Palmer, and Kenneth C. Freeman. Mapping the galactic halo. iv. finding distant giants reliably with the washington system. *The American Astronomical Society*, 121:37–40, 2000.

[83] Bengt Muthén and Kerby Shedden. Finite mixture modeling with mixture outcomes using the em algorithm. *Biometrics*, 55:463–9, 07 1999.

[84] Wong MY, Day NE, Luan JA, Chan KP, and Wareham NJ. The detection of gene-environment interaction for continuous traits: should we deal with measurement error by bigger studies or better measurement? *Int J Epidemiol*, 32: 51–57, 2003.

[85] Tsuyoshi Nakamura. Corrected score function for errors-in-variables models: Methodology and application to generalized linear models. *Biometrika*, 77(1): 127–137, 1990.

[86] J. A. Nelder and R. W. M. Wedderburn. Generalized linear models. *Journal of the Royal Statistical Society. Series A (General)*, 135(3):370–384, 1972.

[87] Mark Pagel and Andrew Meade. A phylogenetic mixture model for detecting pattern-heterogeneity in gene sequence or character-state data. *Systematic Biology*, 53(4):571–581, 2004.

[88] A.G. Patriota, H. Bolfarine, and M. de Castro. A heteroscedastic structural errors-in-variables model with equation error. *Statistical Methodology*, 6(4): 408–423, 2009.

[89] Karl Pearson. Contributions to the mathematical theory of evolution. *Philosophical Transactions of the Royal Society of London A: Mathematical, Physical and Engineering Sciences*, 185:71–110, 1894.

[90] Yingwei Peng and Keith B. G. Dear. A nonparametric mixture model for cure rate estimation. *Biometrics*, 56(1):237–243, 2000.

[91] A. L Piro. Shear waves and giant-flare oscillations from soft gamma-ray repeaters. *The Astrophysical Journal*, 634:153–156, 2005.

[92] William H. Press, Saul A. Teukolsky, William T. Vetterling, and Brian P. Flannery. *Numerical Recipes in n FORTRAN: The Art of Scientific Computing*. Cambridge University Press, 1992.

[93] Richard E. Quandt and James B. Ramsey. Estimating mixtures of normal distributions and switching regressions. *Journal of the American Statistical Association*, 73(364):730–738, 1978.

[94] Carroll R., Ruppert D., Stefanski L., Crainiceanu C., Reid N., Tibshirani R., Louis T., Tong H., Keiding N., Murphy S., and Isham V. *Measurement Error in Nonlinear Models: A Modern Perspective*. New York: Chapman and Hall/CRC, 2006.

[95] Sylvia Richardson and Walter R. Gilks. A bayesian approach to measurement error problems in epidemiology using conditional independence models. *American Journal of Epidemiology*, 138(6):430–442, 1993.

[96] Sylvia. Richardson and Peter J. Green. On bayesian analysis of mixtures with an unknown number of components (with discussion). *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 59(4):731–792, 1997. doi: 10.1111/1467-9868.00095.

[97] Sylvia Richardson, Laurent Leblond, Isabelle Jaussent, and Peter J. Green. Mixture models in measurement error problems, with reference to epidemio-

logical studies. *Journal of the Royal Statistical Society. Series A (Statistics in Society)*, 165(3):549–566, 2002.

[98] Peter Roming, Thomas E. Kennedy, Keith O. Mason, J Nousek, Lindy Ahr, Richard E. Bingham, Patrick S. Broos, Mary J. Carter, Barry K. Hancock, Howard E. Huckle, Sally D. Hunsberger, Hajime Kawakami, Ronnie Killough, T Scott Koch, Michael K. Mclelland, Kelly Smith, Philip J. Smith, Juan Soto, P Boyd, and Joseph Stock. The swift ultra-violet/optical telescope. *Space Science Reviews*, 120, 07 2006.

[99] A S Whittemore and Gail Gong. Poisson regression with misclassified counts: Application to cervical cancer mortality rates. *Journal of the Royal Statistical Society. Series C, Applied statistics*, 40:81–93, 02 1991.

[100] J.L. Schafer. *Analysis of Incomplete Multivariate Data*. Chapman & Hall/CRC Monographs on Statistics & Applied Probability. CRC Press, 1997.

[101] Susanne M. Schennach. Measurement error in nonlinear models – a review. *Advances in Economics and Econometrics: Tenth World Congress*, 3:296?337, 2013.

[102] Gideon Schwarz. Estimating the dimension of a model. *The Annals of Statistics*, 6(2):461–464, 1978.

[103] Sergiy Shklyar. Consistency of an estimator of the parameters of a polynomial regression with a known variance relation for errors in the measurement of the regressor and the echo. *Teoriya Jmovirnostej ta Matematychna Statystyka*, 76, 01 2008.

[104] Donna Spiegelman, Bernard Rosner, and Roger Logan. Estimation and inference for logistic regression with covariate misclassification and measurement error in main study/validation study designs. *Journal of the American Statistical Association*, 95(449):51–61, 2000.

[105] L. A. Stefanski and J. R. Cook. Simulation-extrapolation: The measurement error jackknife. *Journal of the American Statistical Association*, 90(432):1247–1256, 1995.

[106] Leonard A. Stefanski. Unbiased estimation of a nonlinear function of a normal mean with application to measurement error models. *Communications in Statistics, Series A*, pages 4335–4358, 1989.

[107] Leonard A. Stefanski and Raymond J. Carroll. Conditional scores and optimal scores for generalized linear measurement- error models. *Biometrika*, 74(4): 703–716, 1987.

[108] Matthew Stephens. Dealing with label switching in mixture models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 62(4):795–809, 2000.

[109] Markus Thamerus. *Different Nonlinear Regression Models with Incorrectly Observed Covariates*, pages 31–44. Physica-Verlag HD, 1998.

[110] D. M. Titterington, Adrian F. M. Smith, and U. E. Makov. *Statistical analysis of finite mixture distributions*. Chichester: Wiley, 1985.

[111] Anastasios A. Tsiatis and Yanyuan Ma. Locally efficient semiparametric estimators for functional measurement error models. *Biometrika*, 91(4):835–848, 2004.

[112] T. Rolf Turner. Estimating the propagation rate of a viral infection of potato plants via mixtures of regressions. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 49(3):371–384, 2000.

[113] Tom Wansbeek and Erik Meijer. Measurement error and latent variables, 11 2007.

[114] Michel Wedel and Wayne S. DeSarbo. A mixture likelihood approach for generalized linear models. *Journal of Classification*, 12(1):21–55, Mar 1995.

[115] Daniel Wilhelm. Testing for the presence of measurement error. *cemmap working paper CWP45/18*, 2018.

[116] Rainer Winkelmann. *Econometric Analysis of Count Data*. Springer Publishing Company, Incorporated, 5th edition, 2008. ISBN 3540776486, 9783540776482.

[117] Ling Xu, Timothy Hanson, Edward J. Bedrick, and Carla Restrepo. Hypothesis tests on mixture model components with applications inï¿œecologyï¿œandï¿œagriculture. *Journal of Agricultural, Biological, and Environmental Statistics*, 15(3):308–326, Sep 2010.

[118] Yingsi Yang. Poisson regression with measurement error in covariate. 2012.

[119] Weixin Yao and Weixing Song. Mixtures of linear regression with measurement errors. *Communications in Statistics - Theory and Methods*, 44(8):1602–1614, 2015.

[120] D. S. Young and D. R. Hunter. Mixtures of regressions with predictor-dependent mixing proportions. *Computational Statistics and Data Analysis*, 54(10):2253–2266, 2010.

[121] D. S. Young, C. Ke, and X. Zeng. The mixturegram: A visualization tool for assessing the number of components in finite mixture models. *Journal of Computational and Graphical Statistics*, 2018.

[122] Derek S. Young. Mixtures of regressions with changepoints. *Statistics and Computing*, 24(2), 2014.

[123] Hong-Tu Zhu and Heping Zhang. Hypothesis testing in mixture regression models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 66(1):3–16, 2004.

[124] Guangyong Zou. A modified poisson regression approach to prospective studies with binary data. *American Journal of Epidemiology*, 159(7):702–706, 2004.

## Vita

### Education

- Master of Science in Statistics, University of Kentucky, 2013 - 2015
- Bachelor of Science in Statistics, Wuhan University, 2009 - 2013

### Professional Position

- Research Assistant, University of Kentucky, 2016 - 2018
- Teaching Assistant, University of Kentucky, 2013 - 2015