Statistics and Applications {ISSN 2454-7395 (online)}
Special Issue in Memory of Prof. C R Rao
Volume 22, No. 3, 2024 (New Series), pp 285–309
http://www.ssca.org.in/journal

Mixtures of Linear Regressions with Measurement Error in the Response, with an Application to Gamma-Ray Burst Data

Xiaoqiong Fang[1], **Andy W. Chen**[2] **and Derek S. Young**[3]
[1] *Corporate & Investment Bank, J.P. Morgan*
Brooklyn, New York, USA
[2] *School of Business, Government, and Economics, Seattle Pacific University*
Seattle, Washington, USA
[3] *Dr. Bing Zhang Department of Statistics, University of Kentucky*
Lexington, Kentucky, USA

Received: 29 April 2024; Revised: 23 July 2024; Accepted: 30 August 2024

Abstract

Gamma-ray bursts are intense, energetic explosions of gamma rays that are usually accompanied by an afterglow, which is a longer-lived emission that is detected at longer wavelengths, like X-ray, infrared, and radio. Classic gamma-ray burst data is often analyzed using some sort of regression model (*e.g.*, linear, piecewise linear, or a broken-power law model) to relate the flux of the burst to the time since the event. While these models may provide good fits, there is also often a "flaring" phenomena that tends to noticeably deviate from the fitted model. One way we can characterize such a phenomena relative to the underlying general trend is through a mixture-of-regressions model. Some applications in astronomy, like color-luminosity relations for field galaxies, are known to have the variables in the models prone to both intrinsic scatter and measurement error. This assumption is also tenable for gamma-ray burst data where the variance of heteroscedastic measurement errors can be reasonably known. Thus, we introduce a mixture-of-linear-regressions model where the variance of the measurement error is roughly known. Estimation is accomplished using an expectation-maximization (EM) algorithm framework with a weighted least squares estimator that was developed for the non-mixture setting. The finite-sampling behavior of our proposed model's estimates is examined by a simulation study. We also demonstrate the efficacy of this approach on a dataset involving the flux measurements of gamma-ray bursts, where the variance of the measurement error for the flux measurements (the response) are known. Our results for this data problem are compared with estimates obtained using other traditional models, including the linear regression model and the mixture-of-linear-regressions model.

Key words: Astrostatistics; Bootstrap; EM algorithm; Finite mixture model; Intrinsic scatter; Weighted least squares.

AMS Subject Classifications: 62J05, 62P99

Corresponding Author: Derek S. Young
Email: derek.young@uky.edu

1. Introduction

Variability is an inherent part of the results of measurements and of the measurement process. *Measurement error models*, also called *errors-in-variables models*, account for the difference between a measured value of a quantity and its true value. The effect of such measurement error and how to incorporate it into a statistical model has been long investigated, with authoritative texts devoted to this topic, including Fuller (1987), Carroll *et al.* (2006b), and Buonaccorsi (2010). 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. Specifically, the text by Carroll *et al.* (2006b) covers measurement error in nonlinear models, with a special focus on bias reduction, also called *approximate consistency.* For linear regression models with measurement error in the predictors, it can cause an underestimate of the slope coefficients, known as *attenuation bias.* In nonlinear models, the direction of the bias is likely to be more complicated as treated in Carroll *et al.* (2006b). Such biases can of course lead to a loss of power as well as mask certain important features of the data.

The statistical analysis of data with measurement error has a long history, especially in econometrics, with Frisch (1935) being one of the earliest references. Measurement error models are also employed in other diverse research areas, including nutrition (Carroll *et al.*, 2006a; Murillo *et al.*, 2019), finance (Carmichael and Coën, 2008; Maddala and Nimalendran, 1996), and astrostatistics (Kelly, 2007, 2012). With respect to astronomical research, measurement error problems are widely employed due to the presence of *intrinsic scatter*, a type of measurement error regarding variations in the physical properties of astronomical sources that are not completely captured by the variables included in the (regression) model. Feigelson and Babu (1992) provided an early introduction to measurement error models for use in astronomical regressions. Morrison *et al.* (2000) studied galaxy formation with a large survey of stars in the Milky Way using star velocities, which contained heteroscedastic measurement errors. To verify 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) described a Bayesian method to account for measurement errors in linear regression of astronomical data. In another study, Andrae (2010) presented an overview of different methods for error estimation that are applicable to both model-based and model-independent parameter estimates in astronomy.

The focus of the present work will be on developing a model for gamma-ray bursts (GRBs), where we relate the flux of the burst to the time since the event. The flux measurement is prone to both intrinsic scatter and measurement error, where the variance of the measurement errors are available. Moreover, there is a "flaring" phenomena that tends to noticeably deviate from traditional models that are fit to the data; *e.g.*, linear regression models. We propose a novel mixture-of-linear-regressions model with measurement error in the response variable to characterize both the flaring phenomena relative and the underlying general trend, as well as incorporate the measurement error in the flux measurement.

In the non-mixture setting, many methods have been proposed for performing linear regression when intrinsic scatter and/or measurement error is present. Clutton-Brock (1967) proposed an *effective variance* method. Press *et al.* (1992) proposed a procedure for minimizing an *effective* χ^2-*statistic.* Stephens and Dellaportas (1992), Richardson and Gilks

(1993), Dellaportas and Stephens (1995), and Gustafson (2004) each developed Bayesian approaches for estimating measurement error models. Some methods specifically developed for and applied in astronomical research are the *bivariate correlated errors and intrinsic scatter* (BCES) estimator (Akritas and Bershady, 1996) and the FITEXY estimator (Press *et al.*, 1992).

Finite mixture models are used to characterize the presence of unobserved subpopulations (or latent classes) within an overall population. The theoretical, methodological, and computational developments concerning finite mixture models is expansive, and the application of such models have provided critical insights into problems spanning virtually every research discipline. We refer to the texts by Titterington *et al.* (1985), Lindsay (1995), McLachlan and Peel (2000), Frühwirth-Schnatter (2006), and Mengersen *et al.* (2011), as well as the numerous references therein. Mixture models have enjoyed a strong presence in a wide range of fields, spanning the biological, physical, and social sciences. In particular, they have been successfully used in agriculture, astrostatistics, bioinformatics, economics, engineering, marketing, healthcare, neuroscience, and psychology (McLachlan *et al.*, 2019). Some of the applications in astronomical research that use mixture models include classification of astronomical bodies, identification of contaminants in astronomical images, and clustering overlapping population of stars (Kuhn and Feigelson, 2019). These tasks are essential for the study of stars and planet formation as well as analyzing multi-band astronomical images (Feigelson *et al.*, 2021). There are also precedents with using mixture models in the analysis of GRBs. Tarnopolski (2019) analyzed different properties of GRBs from the Burst and Transient Source Experiment (BATSE) using mixtures of multivariate skewed distributions.

Research at the intersection of (finite) mixture models and measurement errors is fairly limited. Lindsay (1995) highlights examples where the joint distribution of observable variables (including the observed *surrogate variables*, which are the variables whose true values are subject to measurement error) has a mixture form. Richardson *et al.* (2002) provides a Bayesian treatment of mixture models in measurement error problems. For mixtures-of-linear-regressions models, measurement error has only been studied in the predictors. This model was introduced by Yao and Song (2015), who developed a deconvolution method to estimate the observed surrogates and employed a generalized expectation-maximization (GEM) algorithm (Dempster *et al.*, 1977) for performing maximum likelihood estimation. An extension of that work for the setting of mixtures of polynomial regressions was presented in Fang *et al.* (2023). The distinction with the contributions in the present paper is that we address the issue of measurement error in the response variable through a mixture structure.

This paper is organized as follows. In Section 2, we define the particular mixture model used in this study. The challenges with this model mostly concern estimation and inference, which are presented in Section 3. In particular, we extend the weighted least squares (WLS) estimator developed by Akritas and Bershady (1996), but in the context of our mixture model. In Section 4, we conduct a simulation study using our proposed algorithm. In Section 5, we perform a thorough analysis of a GRB dataset using our mixture model. We end with some concluding remarks in Section 6.

2. The model

We first consider the setup for the classic mixture-of-linear-regressions model. Suppose we have a random sample of response variables, $Y_1, \ldots, Y_n$, that are each measured with a vector of predictors, $\mathbf{X}_i = (1, X_{i,1}, \ldots, X_{i,p-1})^{\mathrm{T}}$, $p < n$, for $i = 1, \ldots, n$, such that the first entry is a 1 to accommodate an intercept. Let $\mathcal{Z}_i$ be a latent class variable with $\mathrm{P}(\mathcal{Z}_i = j|\mathbf{X}_i) = \lambda_j$ for $j = 1, \ldots, k$, where $\lambda_j > 0$ and $\sum_{j=1}^{k} \lambda_j = 1$. Given $\mathcal{Z}_i = j$, the relationship between a univariate observation Y_i and $\mathbf{X}_i$ is the linear regression model

$$Y_i = \mathbf{X}_i^{\mathrm{T}} \boldsymbol{\beta}_j + \epsilon_j. \tag{1}$$

Here, $\epsilon_j \sim \mathcal{N}(0, \sigma_j^2)$, where σ_j^2 is the error variance for class (component) j, and $\boldsymbol{\beta}_j = (\beta_{0,j}, \ldots, \beta_{p-1,j})^{\mathrm{T}}$ is the p-dimensional vector of regression coefficients. Therefore, unconditional on $\mathcal{Z}_i$, but conditional on $\mathbf{X}_i$, the Y_is follow the mixture distribution

$$Y_i \mid \mathbf{X}_i \sim \sum_{j=1}^{k} \lambda_j \mathcal{N}(\mathbf{X}_i^{\mathrm{T}} \boldsymbol{\beta}_j, \sigma_j^2). \tag{2}$$

Maximum likelihood estimation of mixtures of linear regressions is straightforward, and typically performed using an EM algorithm. Bayesian inference can easily be performed via classic MCMC algorithms. We refer to De Veaux (1989), Viele and Tong (2002), and Hurn *et al.* (2003) for sound treatments of both approaches, which can be implemented using, for example, the `R` package `mixtools` (Benaglia *et al.*, 2009).

Suppose now that we have additive measurement error in the response variable, which we can write using the following (additive) measurement error model:

$$Y_i^* = Y_i + \delta_i. \tag{3}$$

In the above, Y_i is the true response value, Y_i^* is the observed response variable (*i.e.*, the surrogate variable), and δ_i is the measurement error. The measurement error is assumed to be independent of the Y_i as well as to have zero mean and finite variance η_i^2. In classic measurement error models, including regression models where the measurement error occurs in the predictor, a stronger assumption of normality is usually imposed on the distribution of the δ_is. Regardless, the classic measurement error setting will seek out estimation of the variance, with such methods discussed in Carroll *et al.* (2006b). One may, however, have a known value of η_i^2s or be able to posit a good estimate. In the GRB data discussed, we can reasonably make this assumption through the reported errors in the flux measurement. Therefore, we consider the setting where we observe the following for the ith observation in the dataset:

$$(\mathbf{X}_i^{\mathrm{T}}, Y_i^*, \eta_i^2), \tag{4}$$

where the true response is assumed to arise from the mixture structure discussed above in (1) and (2).

In the non-mixture (*i.e.*, classic multiple linear regression) setting, we know that the ordinary least squares (OLS) estimator for $\boldsymbol{\beta}$ minimizes the residual sum of squares $\|\mathbf{Y} - \boldsymbol{\mathcal{X}}\boldsymbol{\beta}\|^2$, where $\mathbf{Y}$ is an n-dimensional vector consisting of the Y_is and $\boldsymbol{\mathcal{X}}$ is an $n \times p$ full-rank design matrix with ith row $\mathbf{X}_i^{\mathrm{T}}$. The OLS estimator is, thus, $\hat{\boldsymbol{\beta}}_{\mathrm{OLS}} = \left(\boldsymbol{\mathcal{X}}^{\mathrm{T}}\boldsymbol{\mathcal{X}}\right)^{-1} \boldsymbol{\mathcal{X}}^{\mathrm{T}}\mathbf{Y}$, which is also equal to the maximum likelihood estimator (MLE) in this setting. In the mixture setting, when performing maximum likelihood estimation via an EM algorithm, the MLE for the jth component's regression coefficient is calculated in the M-step at the tth iteration of the algorithm as $\hat{\boldsymbol{\beta}}_j^{(t+1)} = \left(\boldsymbol{\mathcal{X}}^{\mathrm{T}}\mathbf{W}_j^{(t)}\boldsymbol{\mathcal{X}}\right)^{-1}\boldsymbol{\mathcal{X}}^{\mathrm{T}}\mathbf{W}_j^{(t)}\mathbf{Y}$. In this expression, $\mathbf{W}_j^{(t)}$ is an $n \times n$ diagonal matrix with ith entry equal to the posterior membership probability of the ith observation belonging to component j, which is determined through an application of Bayes' rule in the E-step. Note that the form of the $\hat{\boldsymbol{\beta}}_j^{(t+1)}$ is that of a WLS estimator with weighting matrix $\mathbf{W}_j^{(t)}$. If there is measurement error in the predictors, as in the setting considered by Yao and Song (2015), or in the response, as in the present consideration, then the MLE just discussed will be biased. In our measurement error setting, we can modify the WLS estimator above to reflect the WLS approach developed in Akritas and Bershady (1996) for the non-mixture setting. This is the approach developed in the next section.

3. Estimating method

3.1. A WLS-based estimate

The model presented in the previous section has non-constant error variance (heteroscedasticity) for each observation. Though WLS was employed in the previous section during estimation of the mixture-of-regression coefficients, WLS is a classic framework for addressing heteroscedasticity. By design, WLS allows one to assign individual weights to the observations, thus removing, or at least improving, the effects of heteroscedasticity. WLS is an example of the broader class of generalized least squares estimators (Aitken, 1935). 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) 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

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

where $\epsilon_{i,k_i} \sim \mathcal{N}(0, \sigma_{k_i}^2)$. With this setting, we may develop a WLS-type approach while working under the assumption that the variance of ϵ_{i,k_i}^* is independent of Y_i^*; see Akritas and Bershady (1996). However, we need estimates of the variance of ϵ_{i,k_i}^*. Under our assumptions, we have

$$\mathrm{Var}(\epsilon_{i,k_i}^*) = \mathrm{Var}(\epsilon_{\cdot,k_i}) + \eta_i^2. \tag{5}$$

Since $\mathrm{Var}(\epsilon_{\cdot,k_i})$ is unknown, $\mathrm{Var}(\epsilon_{i,k_i}^*)$ is also unknown. We can extend the algorithm of Akritas and Bershady (1996) combined with estimates obtained via an EM algorithm to estimate $\mathrm{Var}(\epsilon_{\cdot,1}), \ldots, \mathrm{Var}(\epsilon_{\cdot,k})$; see Algorithm 1.

As shown in Algorithm 1, an EM algorithm is employed in Step (1), and then WLS is used to adjust the regression coefficients in Step (5). The difference between the WLS-

Algorithm 1 WLS-based Algorithm

(1) Given the observed data $\left\{(\mathbf{x}_1^{\mathrm{T}}, y_1^*), \ldots, (\mathbf{x}_n^{\mathrm{T}}, y_n^*)\right\}$ and $\eta_1^2, \ldots, \eta_n^2$, obtain the mixture-of-regressions coefficient estimates $(\hat{\boldsymbol{\beta}}_1^{\mathrm{T}}, \ldots, \hat{\boldsymbol{\beta}}_k^{\mathrm{T}})^{\mathrm{T}}$ using an EM algorithm.
(2) Calculate the residuals $R_{ij} = y_i^* - \mathbf{x}_i^{\mathrm{T}} \hat{\boldsymbol{\beta}}_j$, for $i = 1, \ldots, n$ and $j = 1, \ldots, k$.
(3) 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 (1).
(4) Obtain the estimates of $\mathrm{Var}(\epsilon_{\cdot,1}), \ldots, \mathrm{Var}(\epsilon_{\cdot,k})$ from

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

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

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

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

based estimators, $\widetilde{\boldsymbol{\beta}}_1, \ldots, \widetilde{\boldsymbol{\beta}}_k$, and the MLEs from the mixture-of-regressions EM algorithm, $\hat{\boldsymbol{\beta}}_1, \ldots, \hat{\boldsymbol{\beta}}_k$, will typically not be very large. The variance estimators from the classic mixture-of-regressions model will naturally be smaller than our corrected estimator, since the former excludes the variances from the response variable's measurement error. Notice in Step (3) that the weighted estimators of variances are obtained by subtracting the deviation of measurement error from the overall deviation. Thus, the value of $\left(R_{ij} - \bar{R}_{\cdot j}\right)^2 - \eta_i^2$ can be negative for some i or j, so we employ the usage of the hinge function for this difference; *i.e.*, $\left[\left(R_{ij} - \bar{R}_{\cdot j}\right)^2 - \eta_i^2\right]_+ = \left\{\left(R_{ij} - \bar{R}_{\cdot j}\right)^2 - \eta_i^2\right\} \vee 0$.

3.2. Asymptotic variance

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

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

The asymptotic variance of the MLEs obtained via an EM algorithm in Step (1) of Algorithm 1 can be obtained by the inverse of the information matrix $\mathcal{I}(\boldsymbol{\psi})$ that appears in the asymptotic result

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

However, likelihood functions for mixture models are often complicated, which translates to difficult calculations for the second derivatives of the likelihood function that comprise

$\mathcal{I}(\boldsymbol{\psi})$. Thus, other approaches are necessary (see Chapter 14 of Lange, 2010). For example, Efron and Hinkley (1978) suggested to use the observed Fisher information matrix instead. Later, Louis (1982) introduced a technique for computing the observed information by using calculations only done on the complete information when an EM algorithm is used.

The density for the k-component mixture-of-regressions model is

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

where

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

is the probability density of the ith observation belonging to the jth component. Here, $\boldsymbol{\theta}_j = \left(\boldsymbol{\beta}_j^{\mathrm{T}}, \sigma_j\right)^{\mathrm{T}}$ is the vector of parameters of the jth component and $\phi(\cdot)$ is the density of the standard normal distribution. We can, thus, write out the observed data loglikelihood as

$$\ell_{\mathrm{O}}(\boldsymbol{\psi}) = \sum_{i=1}^{n} \log\left\{\sum_{j=1}^{k} \lambda_j f(y_i \mid \mathbf{x}_i, \boldsymbol{\theta}_j)\right\},$$

which can be augmented with the vector of each observation's unobserved component membership – $\mathbf{z}_i = (z_{i1}, \ldots, z_{ik})^{\mathrm{T}}$ such that $z_{ij} = \mathtt{I}\{\text{observation } i \text{ belongs to component } j\}$ – to construct the complete data loglikelihood

$$\ell_{\mathrm{C}}(\boldsymbol{\psi}) = \sum_{i=1}^{n}\sum_{j=1}^{k} \mathbf{z}_{ij} \log\left\{\lambda_j f(y_i \mid \mathbf{x}_i, \boldsymbol{\theta}_j)\right\}.$$

The complete data is characterized through $\mathbf{s} = \{(\mathbf{x}_i^{\mathrm{T}}, y_i, \mathbf{z}_i^{\mathrm{T}}), i = 1, \ldots, n\}$. Since the $\mathbf{z}_i$ is unobserved, and hence "missing," use of an EM algorithm is appropriate. We forego stating the explicit E-step an M-step for this setting as it is quite standard in the mixture-of-regressions literature; see, for example, Benaglia *et al.* (2009).

To compute the observed information in the EM algorithm, let $S(\mathbf{s} \mid \boldsymbol{\psi})$ and $S((\mathbf{x}_i^{\mathrm{T}}, y_i) \mid \boldsymbol{\psi})$ be the complete data score function and observed data score function, respectively. Moreover, let $\mathcal{I}_{\mathbf{s}}(\boldsymbol{\psi})$ be the complete data information matrix; *i.e.*, the expected value of the negative of the Hessian of the complete data loglikelihood. Then, by differentiation, the observed data information matrix can be written as

$$\mathcal{I}(\hat{\boldsymbol{\psi}}) = \mathcal{I}_{\mathbf{s}}(\hat{\boldsymbol{\psi}}) - \left[\mathbb{E}_{\psi}\left\{S(\mathbf{s} \mid \boldsymbol{\psi})S^{\mathrm{T}}(\mathbf{s} \mid \boldsymbol{\psi})\right\} + S\left\{(\mathbf{x}_i^{T}, y_i) \mid \boldsymbol{\psi}\right\} S^{\mathrm{T}}\left\{(\mathbf{x}_i^{T}, y_i) \mid \boldsymbol{\psi}\right\}\right]\Bigg|_{\boldsymbol{\psi}=\hat{\boldsymbol{\psi}}}.$$

Thus, the asymptotic variance-covariance of the estimator $\hat{\boldsymbol{\psi}}$ can be calculated based on $\mathrm{Var}(\hat{\boldsymbol{\psi}}) = \mathcal{I}(\hat{\boldsymbol{\psi}})^{-1}$, and the estimated standard errors of the parameter estimates in $\hat{\boldsymbol{\psi}}$ are the square root of the diagonal entries of this matrix. Note that in the present setting, we are using the y_i^* in the role of the y_i that appear in the preceding formulas. Moreover, the MLE $\hat{\boldsymbol{\psi}}$ is actually based on the WLS estimators $\tilde{\boldsymbol{\beta}}_j$, $j = 1, \ldots, k$ in Step (5) of Algorithm 1, and not the $\hat{\boldsymbol{\beta}}_j$ calculated in Step (1); *i.e.*,

$$\hat{\boldsymbol{\psi}} = \left(\hat{\lambda}_1, \ldots, \hat{\lambda}_{k-1}, \tilde{\boldsymbol{\beta}}_1^{\mathrm{T}}, \ldots, \tilde{\boldsymbol{\beta}}_k^{\mathrm{T}}, \hat{\sigma}_1^2, \ldots, \hat{\sigma}_k^2\right)^{\mathrm{T}}.$$

3.3. Bootstrap estimator for the standard errors

Even when estimation of $\boldsymbol{\psi}$ is trivial, estimation of standard errors (SEs) can be computationally burdensome, especially when measurement error is involved. One alternative strategy is to use the parametric bootstrap (Efron and Tibshirani, 1993; Davison and Hinkley, 1997), which theoretically should provide similar estimates to the standard errors compared to the method involving the information matrix. This has become especially useful for standard error estimation in mixture settings, as noted in Chapter 2 of McLachlan and Peel (2000).

Algorithm 2 Parametric Bootstrap for Standard Errors

```
(1) Find $\hat{\boldsymbol{\psi}}$ by implementing Algorithm 1 using the observed data $\{(\mathbf{x}_1, y_1^*), \ldots, (\mathbf{x}_n, y_n^*)\}$.
(2) Generate a bootstrap sample $\{(\mathbf{x}_1, y_1^{**}), \ldots, (\mathbf{x}_n, y_n^{**})\}$, where each $y_i^{**}$ is a realization from the (conditional) mixture distribution $\sum_{j=1}^{k} \hat{\lambda}_j \mathcal{N}\left(\mathbf{x}_i^{\mathrm{T}} \widetilde{\boldsymbol{\beta}}_j, \hat{\sigma}_j^2\right)$.
(3) For each of $y_i^{**}$, generate the "observed" response by
        $y_i^{***} = y_i^{**} + \delta_i,$
    where $\delta_i \sim \mathcal{N}(0, \eta_i^2)$ is generated using the known variabilities $\eta_1^2, \ldots, \eta_n^2$.
(4) Find the estimate $\widetilde{\boldsymbol{\psi}}$ by implementing Algorithm 1 on $\{(\mathbf{x}_1, y_1^{***}), \ldots, (\mathbf{x}_n, y_n^{***})\}$.
(5) Repeat Steps (2) - (4) $B$ times to generate the bootstrap sampling distribution $\widetilde{\boldsymbol{\psi}}^{(1)}, \widetilde{\boldsymbol{\psi}}^{(2)}, \ldots, \widetilde{\boldsymbol{\psi}}^{(B)}$.
```

Algorithm 2 outlines a parametric bootstrap to estimate standard errors in our mixture-of-regressions model when specifying measurement error in the response. After implementing Algorithm 2, the bootstrap variance-covariance matrix is easily computed as the sample variance-covariance matrix of the generated values $\widetilde{\boldsymbol{\psi}}^{(1)}$, $\widetilde{\boldsymbol{\psi}}^{(2)}, \ldots, \widetilde{\boldsymbol{\psi}}^{(B)}$. 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, that is, we want to enforce a meaningful identifiability constraint for a particular analysis. For example, one could set $\beta_{11} < \ldots < \beta_{k1}$ (*i.e.*, a constraint on the slope for the first predictor in the model) or $\sigma_1 < \ldots < \sigma_k$. We will state the identifiability constraints used for our numerical work in the next section.

4. Numerical studies

We now study the finite sampling behavior of the proposed estimators for our mixture-of-regressions model with measurement error in the response. Our study considers mixtures of regressions with one or two predictors, as well as two or three components. The basic setting for our models involves *iid* data $(\mathbf{x}_i^{\mathrm{T}}, y_i, \eta_i)$, $i = 1, \ldots, n$ such that the response variable Y_i is drawn from the model

$$Y_i | \mathbf{X}_i = \mathbf{x}_i \sim \sum_{j=1}^{k} \lambda_j \mathcal{N}\left(\mathbf{x}_i^{\mathrm{T}} \boldsymbol{\beta}_j, \sigma_j^2\right),$$

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

where $\delta_i \sim \mathcal{N}(0, \eta_i^2)$ is the simulated measurement error in the response. To study the effect of the measurement error on the proposed estimator for mixtures of both simple and multiple linear regressions with different number of components, we consider the three component structures: well-separated (WS), moderately-separated (MS), and overlapping (OL). These three categorizations of separability were determined by considering component mean structures and error variances that yield varying degrees of overlap with the generated data. An explicit quantitative threshold was not employed to characterize if components are WS, MS, or OL, but rather a visual check on simulated datasets was employed to ascertain the appropriateness of the stated component structure. The 12 data-generating processes used to characterize these different structures are summarized in Table 1.

Table 1: The 12 models used for the simulation study

Model	**Structure**	$\boldsymbol{\beta}_1^{\mathrm{T}}$	$\boldsymbol{\beta}_2^{\mathrm{T}}$	$\boldsymbol{\beta}_3^{\mathrm{T}}$	σ_1^2	σ_2^2	σ_3^2	λ_1	λ_2
Mixtures of Simple Linear Regressions									
$M1$	WS	$(-10, 6)$	$(10, 2)$	—	4	1	—	1/2	—
$M2$	MS	$(5, 15)$	$(25, -15)$	—	4	1	—	1/2	—
$M3$	OL	$(5, 5)$	$(15, -5)$	—	4	1	—	1/2	—
$M4$	WS	$(-10, 6)$	$(10, 2)$	$(30, -5)$	4	1	9	1/3	1/3
$M5$	MS	$(5, 15)$	$(20, 20)$	$(25, -15)$	4	1	9	1/3	1/3
$M6$	OL	$(-10, 20)$	$(5, 5)$	$(15, -5)$	4	1	9	1/3	1/3
Mixtures of Multiple Linear Regressions									
$M7$	WS	$(-10, 6, 4)$	$(10, 2, 7)$	—	4	1	—	1/2	—
$M8$	MS	$(5, 15, 10)$	$(25, -15, -10)$	—	4	1	—	1/2	—
$M9$	OL	$(5, 5, 9)$	$(15, -5, 3)$	—	4	1	—	1/2	—
$M10$	WS	$(-10, 6, 4)$	$(10, 2, 7)$	$(30, -5, 10)$	4	1	9	1/3	1/3
$M11$	MS	$(5, 15, 10)$	$(20, 20, 5)$	$(25, -15, -10)$	4	1	9	1/3	1/3
$M12$	OL	$(5, 5, 9)$	$(15, -5, 3)$	$(-10, 20, 15)$	4	1	9	1/3	1/3

For each simulation condition, we randomly generated $B = 1000$ datasets for the sample sizes $n \in \{100, 250\}$. For each sample size, we generated the predictor variables as $X_{ij} \sim \mathcal{U}(0, 1)$, while different measurement errors for the response were considered for each mixture-of-regressions setting. The Monte Carlo samples for the 2-component mixtures of regressions were generated under the two conditions of $\eta_i^2 \sim \mathcal{U}(0, 0.1)$ and $\eta_i^2 \sim \mathcal{U}(2, 6)$. The Monte Carlo samples for the 3-component mixtures of regressions were generated under the two conditions of $\eta_i^2 \sim \mathcal{U}(0, 0.5)$ and $\eta_i^2 \sim \mathcal{U}(5, 10)$.

For each simulated dataset, we estimate the parameters $(\boldsymbol{\beta}_1^{\mathrm{T}}, \ldots, \boldsymbol{\beta}_k^{\mathrm{T}}, \sigma_1^2, \ldots, \sigma_k^2)$ using Algorithm 1, and compare them with the estimates obtained via the "naïve" method, which simply ignores the measurement error; *i.e.*, estimation of the classic mixtures-of-regressions model without measurement error in the response. The performance of the proposed method under different conditions is assessed by calculating the mean squared error (MSE),

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

where $\hat{\boldsymbol{\theta}}^{(t)}$ is the estimate of the parameter $\boldsymbol{\theta}$ based on the tth Monte Carlo sample and $\boldsymbol{\theta}$ is the true value. The relative efficiencies based on the MSEs for the naïve method versus the proposed method are also calculated for all of the parameters.

4.1. Results for mixtures of simple linear regressions

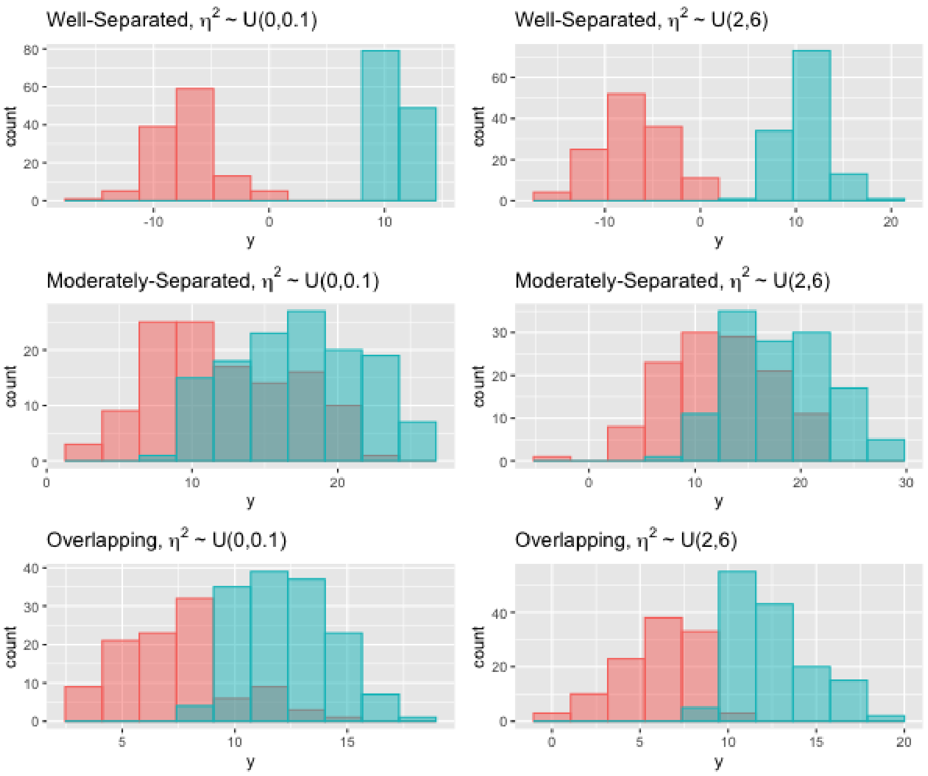


Figure 1: Histograms of observed response variables for 2-component mixtures of simple regression under different settings, with sample size $n = 250$

We first discuss the numerical results obtained for the 2-component mixtures where we have a single predictor. In particular, we first focus on models $M1$, $M2$, and $M3$ in Table 1. Figure 1 shows the histograms of observed responses y^* under different circumstances. Even though these are histograms of the unconditional distribution of the response with measurement error, it still gives an indication about the degree of separability that was incorporated in the mixtures-of-regressions structure. In the WS setting, there are two distinct regression relationships corresponding to the two different components. For the MS and OL settings, the two components have a greater degree of mixing, thus it is harder to identify to which component a certain data point belongs. Regardless, increasing the variance of the measurement errors forces the two components to be closer to each other, which compounds the ability to identify the distinct components.

Table 2 gives the MSEs and relative efficiencies (in parentheses) for the simulated datasets from models $M1$, $M2$, and $M3$. The values in the parentheses represent the relative efficiencies of MSEs for the naïve versus the proposed estimators. For example, the boldface value of 1.0552 means the MSE when estimating β_{21} using the naïve method is 1.0552 times the MSE when estimating the parameter using our proposed method. If the relative efficiency

Table 2: The MSEs and relative efficiencies (in parentheses) of the naïve estimators versus the proposed estimators for 2-component mixtures of simple linear regressions; models $M1$, $M2$, and $M3$

n	η_i^2	β_{10}	β_{11}	β_{20}	β_{21}	σ_1^2	σ_2^2
Well-Separated Components							
100	$\mathcal{U}(0, 0.1)$	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	$\mathcal{U}(2, 6)$	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	$\mathcal{U}(0, 0.1)$	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	$\mathcal{U}(2, 6)$	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	$\mathcal{U}(0, 0.1)$	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	$\mathcal{U}(2, 6)$	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)

is greater than 1, it means the MSE of proposed method is smaller, which leads to greater precision of the estimator. We note that label switching did not appear to be present since a check on the estimates of β_{10} and β_{20} showed that $\hat{\beta}_{10} < \hat{\beta}_{20}$ was met for each sample. Thus, no identifiability constraint had to be enforced for this set of simulations.

Overall, the proposed method appears to behave better than the naïve method with respect to their relative efficiencies since they are greater than 1. For estimating the variances $\text{Var}\,(\epsilon_{\cdot,j})$ when a larger value is used (*i.e.*, when $\sigma_1 = 2$ rather than $\sigma_2 = 1$), the average relative efficiency for the settings with measurement error $\mathcal{U}(2, 6)$ is greater than 2. When the measurement error is trivial, this translates to the behaviors of both methods being nearly the same. Thus, we can conclude that our proposed method behaves better when the measurement error is larger, which accounting for measurement error in such a circumstance is likely of greater importance. Note that because our proposed method only accommodates measurement error in the response after obtaining the maximum likelihood estimates via an

EM algorithm, there is no adjustment to the mixing proportion estimates; *i.e.*, $\hat{\lambda}$ is the same under both methods and, thus, the relative efficiency is necessarily 1.

When the sample size increases from 100 to 250, the MSEs decrease. Moreover, our proposed method shows improvement over the naïve method. If we expand the values of measurement error in the response, the MSEs become larger, however, the performance of the proposed method according to the relative efficiencies is better for 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 accurately 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.

Table 3: The MSEs and relative efficiencies (in parentheses) of the naïve estimators versus the proposed estimators for 3-component mixtures of simple linear regressions; models $M4$, $M5$, and $M6$

n	η_i^2	β_{10}	β_{11}	β_{20}	β_{21}	β_{30}	β_{31}	σ_1^2	σ_2^2	σ_3^2
Well-Separated Components										
100	$\mathcal{U}(0, 0.5)$	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	$\mathcal{U}(5, 10)$	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	$\mathcal{U}(0, 0.5)$	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	$\mathcal{U}(5, 10)$	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	$\mathcal{U}(0, 0.5)$	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	$\mathcal{U}(5, 10)$	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)

In Table 3 we report the MSEs and relative efficiencies (in parentheses) for our simulated datasets from the 3-component setting. The models for this part of our discussion are $M4$, $M5$, and $M6$ in Table 1. Label switching was present when comparing the bootstrap samples for the moderately-separated cases. This was diagnosed by first noting that the MSEs appeared to be fairly large for some parameters when the measurement error is large. For example, the MSE of β_{21} for the moderately-separated setting with $\eta_i^2 \sim \mathcal{U}(5, 10)$ and sample size $n = 100$ was first found to be 133.1943, a value much larger than expected. Since the values of β_{20} and β_{30} are close to each other, simply using the identifiability constraint

$\beta_{10} < \beta_{20} < \beta_{30}$ is not enough. To make the components distinct with each other and correct the label switching in the simulation, we imposed the identifiability constraint of β_{10} being the smallest estimated intercept of the three components and $\beta_{21} > \beta_{31}$.

When the number of components increase, the MSEs become noticeably larger since the model is growing in complexity. With a heavier-parameterized model, the estimation becomes more challenging. When we increase the sample size and decrease the variance of the measurement errors in the response, the MSEs of the unknown parameters becomes smaller. Similarly, the relative efficiencies show that for the case with larger sample size and bigger measurement error, our proposed method performs better than naïve method. For overlapping and moderately-separated cases, the MSEs are fairly large for certain parameters with large measurement error (*e.g.*, with variances $\eta_i^2 \sim \mathcal{U}(5, 10)$), since the three components are subject to heavy mixing and it becomes difficult to consistently distinguish different components, thus leading to greater uncertainty in the estimators.

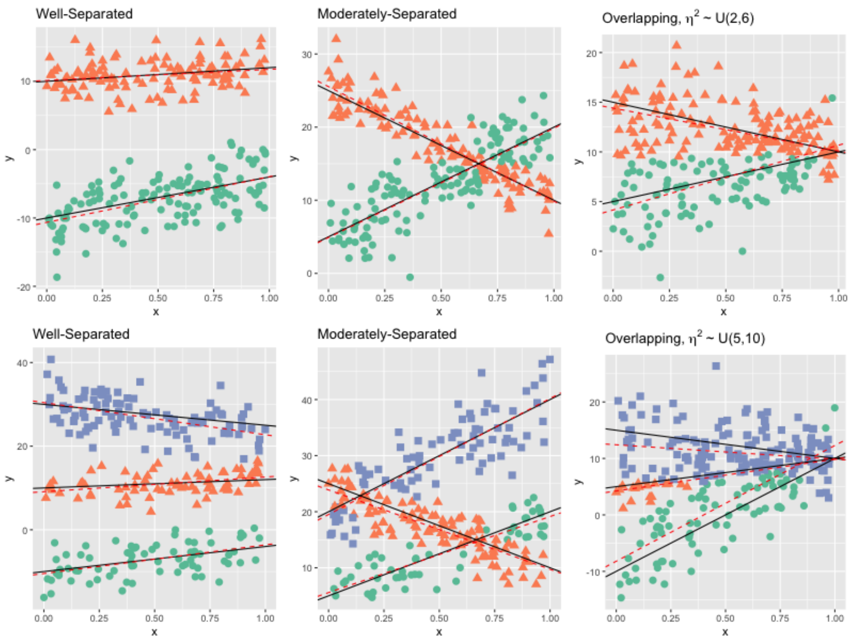


Figure 2: Scatter plots for datasets generated from each of the models $M1 - M6$, inclusive (sample size $n = 250$), where dashed red lines are the estimates obtained using Algorithm 1 and solid black lines are the lines based on the true parameters

Figure 2 shows scatterplots of datasets generated from each of the six settings (models $M1 - M6$) for mixtures of simple linear regressions with measurement error in the response. Different colors and shapes indicate from which component each observation was generated. The dashed red lines are the estimates obtained from our proposed method outlined in Algorithm 1. The solid black lines are the lines based on the true parameters. According to the scatterplots, the proposed method fits well in all settings as the dashed red lines (estimates) are similar to the solid black lines (truth). Moreover, based on the relative efficiencies reported earlier, it improves the performance of estimating parameters when compared to the naïve method. Overall, these results are consistent with demonstrating the efficacy of our proposed method as a way to incorporate measurement error in the response when the underlying data come from a mixture-of-regressions setting.

4.2. Results for mixtures of multiple linear regressions

We next consider the 2-component mixtures of multiple linear regressions with measurement errors, which correspond to the models $M7$, $M8$, and $M9$ in Table 1. Figure 3 shows 3d scatterplots of data simulated from each of these models, where different colors represent to which component each data point belongs. In the well-separated case, the two components are very well-separated, thus making it very easy to distinguish to which component each point belongs. For the moderately-separated and overlapping cases, there are some areas where the two components are mixing, which is where we would expect to have the greatest uncertainty as to how to classify those observations if we were estimating the underlying model.

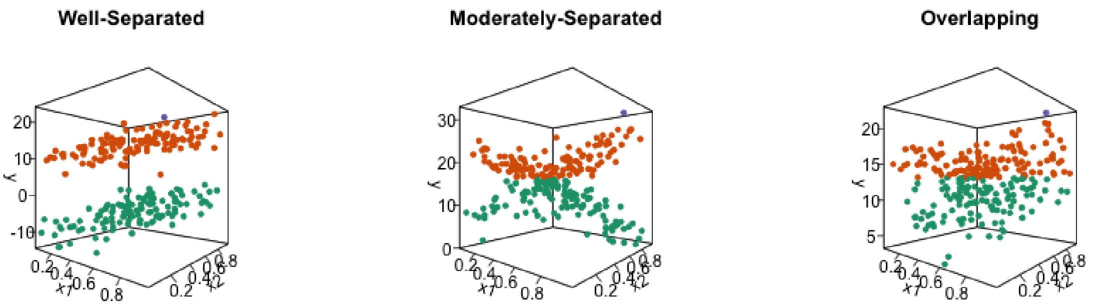


Figure 3: 3d scatterplots of the three different component structures for the 2-component mixtures of multiple linear regressions with sample size $n = 250$ and measurement error $\eta_i^2 \sim \mathcal{U}(2, 6)$ for the response

In Table 4, we report the MSEs and relative efficiencies (in parentheses) for our simulated datasets from the models $M7$, $M8$, and $M9$. Label switching did not appear to be present since the identifiability constraint $\beta_{10} < \beta_{20}$ is satisfied for all bootstrap estimates. The overall behavior of these three 2-component mixtures of multiple linear regressions are similar to those of the 2-component mixtures of simple linear regressions. When we increase the sample size from 100 to 250, the MSEs become smaller and the relative efficiencies improve. Meanwhile, because we add the predictor X_{i2}, the models are more parameterized than when the components are simple linear regressions, thus making the estimation more challenging, especially when the components are overlapping. For example, with an overlapping component, with large measurement errors (variances $\eta_i^2 \sim \mathcal{U}(2, 6)$), and with a sample size of $n = 100$, the boldface value in Table 4 is the MSE of the slope parameter for X_{i2}, β_{12}. This value of 19.2855 is a value much larger than the corresponding setting with simple linear regression components. Naturally, when increasing the number of predictor variables in settings with overlapping components, the increase in the MSEs reflect the greater difficulty in being able to estimate the true parameters.

Table 4: The MSEs and relative efficiencies (in parentheses) of the naïve estimators versus the proposed estimators for 2-component mixtures of multiple linear regressions; models $M7$, $M8$, and $M9$

n	η_i^2	β_{10}	β_{11}	β_{12}	β_{20}	β_{21}	β_{22}	σ_1^2	σ_2^2
Well-Separated Components									
100	$\mathcal{U}(0, 0.1)$	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	$\mathcal{U}(2, 6)$	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	$\mathcal{U}(0, 0.1)$	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	$\mathcal{U}(2, 6)$	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	$\mathcal{U}(0, 0.1)$	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	$\mathcal{U}(2, 6)$	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)

Finally, in Table 5, we report the MSEs and relative efficiencies (in parentheses) for our simulated datasets from the models $M10$, $M11$, and $M12$. The overall behavior of these three 3-component mixtures of multiple linear regressions are similar to those of the 3-component mixtures of simple linear regressions. When we increase the sample size from 100 to 250, the MSEs become markedly smaller and the relative efficiencies improve. Meanwhile, adding the predictor X_{i2} creates heavier-parameterized model than when the components are simple linear regressions, thus making the estimation more challenging. This, again, is especially the case when the components are overlapping. Naturally, when increasing the number of predictor variables in settings with overlapping components, the increase in the MSEs reflect the greater difficulty in being able to precisely estimate the true parameters.

4.3. Summary of simulation results

The combination of simulation conditions we considered in this section is fairly broad in ascertaining the applicability and robustness of our method. The conditions considered are more extensive relative to the most closely-related works of Yao and Song (2015) and Fang *et al.* (2023). The former only considered a two-component mixture structure ($k = 2$) in a

Table 5: The MSEs and relative efficiencies (in parentheses) of the naïve estimators versus the proposed estimators for 3-component mixtures of multiple linear regressions; models $M10$, $M11$, and $M12$

n	η_i^2	β_{10}	β_{11}	β_{12}	β_{20}	β_{21}	β_{22}	β_{30}	β_{31}	β_{32}	σ_1^2	σ_2^2	σ_3^2
						Well-Separated Components							
100	$\mathcal{U}(0, 0.5)$	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	$\mathcal{U}(5, 10)$	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	$\mathcal{U}(0, 0.5)$	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	$\mathcal{U}(5, 10)$	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	$\mathcal{U}(0, 0.5)$	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	$\mathcal{U}(5, 10)$	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)

single predictor. The latter considered two-component mixture structures ($k = 2$), but where the components could be linear, quadratic, or cubic functions of a single predictor. In our simulation work, we considered two-component and three-component mixtures ($k = 2, 3$), each with one or two predictors. The parameters for the underlying regression components are then selected to be well-separated, moderately-separated, or overlapping, yielding the 12 models in Table 1. Moreover, we considered two measurement error structures and two sample sizes, further demonstrating the performance of our methods on a variety of models.

In general, the results reported in this section are consistent with results typically seen in simulations involving mixtures. When the components are well-separated, the results tend to be more stable compared to moderately-separated and overlapping settings. This, of course, follows from the variables in both moderately-separated and overlapping component models being harder to identify. Meanwhile, for the same model with the same component setting (*i.e.*, well-separated, moderately-separated, or overlapping), an increase in the sample size yields a decrease in the MSE, while an increase in the the variances of the measurement error increases the MSE.

Generally speaking, the MSEs of well-separated components are the smallest among the three different types of component settings. When we assumed a smaller measurement error, the MSEs are almost unanimously smaller, which makes sense due to smaller measurement error infusing smaller variability in the response. Overall, 2-component models had better results than the three-component models. For example, for a 3-component heavily-overlapping mixture model with measurement error $\mathcal{U}(5, 10)$ and sample size of 100, the MSEs of $\boldsymbol{\beta}_2^{\mathrm{T}} = (15, -5, 3)$ are $(40.1613, 50.7183, 46.2146)$ (see Table 4), while the 2-component heavily overlapping mixture model with measurement error $\mathcal{U}(2, 6)$ and sample size 100 for the same $\boldsymbol{\beta}_2^{\mathrm{T}}$ has MSEs of $(6.5878, 12.7481, 7.5360)$ (see Table 5).

Our routine developed to estimate the mixture models under consideration do occasionally encounter some numerical issues, especially for the 3-component overlapping models. Sometimes, bad solutions (*i.e.*, estimates that are clearly far away from the true parameter values) were obtained. This would occasionally occur even after starting the algorithm from multiple random starting values. For practical purpose, in the 3-component simulated datasets 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 those results, the MSEs were much more consistent with what was observed under the other conditions. This strategy has been employed for other simulations involving mixtures with complex structures; see, for example, Young (2014).

5. Example: Gamma-ray burst data

GRBs are key observations in gamma-ray astronomy, as they are extremely energetic explosions that occur at random times in distant galaxies. Since the Big Bang, they are considered the brightest electromagnetic events known to occur in the universe. The bursts can last from ten milliseconds to several hours. These phenomena are still the subject of intense research, but some theories suggest they arise during the birth of black holes or a massive super-giant's collapse. See the review article by Piran *et al.* (2013).

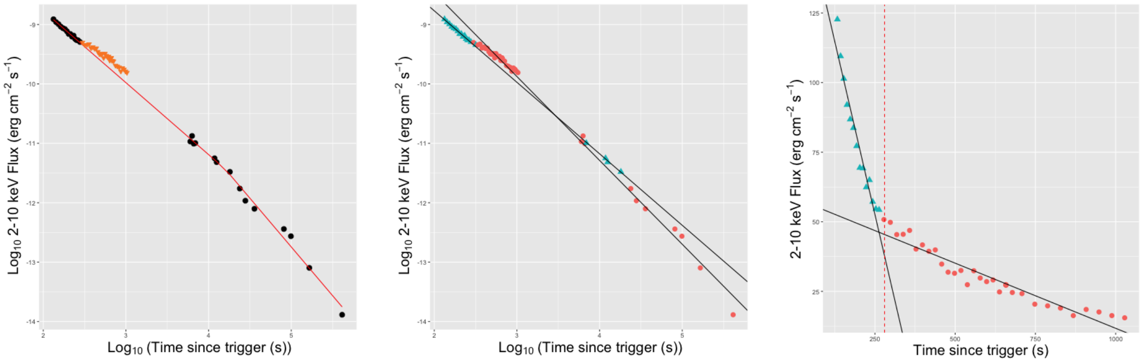
The launch of the Swift observatory (Gehrels *et al.*, 2004) modernized how we observe GRBs. The Swift observatory, which has collected and made available copious amounts of GRB data, provides rapid notification of GRB triggers to the ground using a highly-sensitive Burst Alert Telescope (BAT; Barthelmy, 2004). It also makes panchromatic observations of the burst and its afterglow. On May 25th, 2005, the Swift BAT was triggered and located GRB050525a[1] (Blustin *et al.*, 2006), the significance being that this was the first bright, low-redshift burst to have been observed using the observatory. The X-ray decay 'light curve' of GRB050525a that was obtained includes both *photo-diode* (PD) mode ($T < 2000$s) and *photon-counting* (PC) mode ($T > 2000$s) data. The data are plotted in Figure (4(a)), and like many astronomical datasets, the GRB observations suffer from measurement error due to the detection technique used.

The GRB050525a dataset 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 dataset are: time since trigger (in seconds), X-ray flux (in units of 10^{-11} erg/cm^2/s, $2 - 10$ keV), and the variability of the measurement error of the flux based on detector signal-to-noise values.

Blustin *et al.* (2006) fit the data with a power-law model; *i.e.*, a linear regression model. However, they note systematic deviations of the residuals at certain time points, which they attempt to capture using temporal breaks, resulting in what they call a broken power-law model; *i.e.*, a piecewise linear regression model. The data and best-fit line using a single breakpoint are shown in Figure 4(a). Blustin *et al.* (2006) note that the power-law fit

[1] The naming convention for GRBs is "GRByymmdd", where a subsequent letter (*i.e.*, a, b, c, *etc.*) denotes the observation on a day when multiple GRBs occurred.

of the pre-brightening PD mode data ($T < 280$s) extrapolates well to the pre-break 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. The authors further note that the flux returns to the pre-flare decay curve prior to the start of the PC data.



(a) (b) (c)

Figure 4: Scatterplots of the GRB050525a data with (a) the best-fit line from a broken power-law model, (b) the estimated 2-component measurement error model fit, and (c) the estimated 2-component measurement error model fit on the PD mode data

Blustin *et al.* (2006) do not directly model the flaring points in their modeling. The flaring points are denoted by orange dots in Figure 4(a). In order to also capture the characteristic of the flaring part of this phenomena, we fit the data with a mixture-of-linear-regressions model, which can potentially identify separate regression models for the initial burst. Moreover, we can incorporate the reported variability of the measurement error of the flux through the model we developed in Section 2.

While we hypothesize that separate regression models could be appropriate for the initial burst and the remaining flux measurements, we will proceed to assess the number of components for the proposed mixture-of-linear-regressions model. We consider $k = 1, 2, 3, 4$ and select the best model according to results using the following model selection criteria: Akaike's information criterion (AIC; Akaike, 1973), the Bayesian information criterion (BIC; Schwarz, 1978), the Integrated Completed Likelihood criterion (ICL; Biernacki *et al.*, 2000), and the consistent AIC (cAIC; Bozdogan, 1987). The number of components is chosen based on the smallest respective model selection value. This was repeated with $N = 100$ random starts, where the scores from the best start are given in Table 6. 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; McLachlan and Peel, 2000). In this case, BIC, ICL, and cAIC all select $k = 2$ while AIC appears to overestimate by selecting $k = 4$. We also compare the model selection results (AIC, BIC, and cAIC) to the simple linear regression (SLR) fit[2] with no measurement error. Each of these is just slightly larger than the $k = 1$ fit, indicating that including the measurement error in the estimation provides

[2] Note that ICL, which is a penalized form of BIC, is not calculated for the SLR or the $k = 1$ fit. ICL and its variants are designed to identify the number of components in a model-based clustering framework,

a slight improvement over the traditional SLR fit. Regardless, based on these results we proceed to use the fit for the 2-component model with measurement error in the response.

Table 6: Model selection criteria for determining the number of components for the GRB dataset, where bold values indicate the number of components chosen under that criterion

k	AIC	BIC	cAIC	ICL
1	−84.935	−80.649	−78.649	—
2	−156.654	**−143.796**	**−137.796**	**−145.016**
3	−130.872	−109.440	−99.440	−111.137
4	**−158.57**	−128.568	−114.568	−131.251
SLR	−82.944	−76.515	−73.515	—

The model with known measurement errors in the responses that we fit is written as

$$\begin{aligned} y_i &\sim \begin{cases} \mathbf{x}_i^T \boldsymbol{\beta}_1 + \epsilon_{i1}, & \text{with probability } \lambda \\ \mathbf{x}_i^T \boldsymbol{\beta}_2 + \epsilon_{i2}, & \text{with probability } 1-\lambda, \end{cases} \\ y_i^* &= y_i + \delta_i, \end{aligned} \tag{6}$$

where $\epsilon_{ij} \sim \mathcal{N}(0, \sigma_j^2)$ are independent, $i = 1, \ldots, 63$, and $j = 1, 2$, $\mathbf{x}_i = (1, \log_{10}(t_i))$, t_i is the ith observation time since trigger (in seconds), y_i^* is the logarithm (base 10) of the X-ray flux from the ith measurement, $\delta_i \sim \mathcal{N}(0, \eta_i^2)$, $\eta_i^2 = \log_{10}^2(s_i)$, s_i is the reported variability for the measurement error of the flux for the ith observation, and δ_i is independent of ϵ_{ij}.

Table 7: Parameter estimates, estimated SEs from the parametric bootstrap, and the estimated SEs using the observed information matrix

Parameter	Estimates	Bootstrap SEs	Theoretical SEs
β_{10}	−6.782	2.438	0.209
β_{11}	−1.007	0.912	0.049
β_{20}	−5.286	3.561	0.147
β_{21}	−1.552	1.178	0.022
σ_1	0.792	0.112	0.057
σ_2	1.470	0.600	0.413
λ	0.601	0.197	0.249

For the WLS estimate $\widetilde{\boldsymbol{\beta}}_j$ in our mixture-of-regressions setting, we obtain standard errors for the parameters using a parametric bootstrap with $B = 1000$. We then compare the result with variance estimates for the WLS estimators using the inverse of the observed information matrix (see Table 7). Based on the output, the standard errors from the parametric bootstrap are much larger than the inverse of observed information, especially for the

which is achieved through the estimated mean entropy that is used as the penalty term (Biernacki *et al.*, 2000; Baudry *et al.*, 2010; Bertoletti *et al.*, 2015). As noted in Bertoletti *et al.* (2015), “the ICL tends to be less prone to discriminate overlapping groups, essentially becoming an efficient model-based criterion that can be used to outline the clustering structure in the data.”

intercepts. However, the standard errors for the variances, σ_1 and σ_2, and mixing proportion λ are reasonable, as well as the intercepts β_{11} and β_{21}.

The lines from the estimated model are shown in Figure 4(b), where each color represents the component based on the largest posterior membership probability. Based on this figure there are clearly two distinct components: one with time $T < 2000$s and the other 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. The data within PD mode consists of the first $n = 49$ data points. We fit the non-log-transformed data (time since trigger as predictor variable x_i and X-ray flux as observed response variable y_i^*) with a 2-component mixture model using our proposed method. The fit for the model in (6) is

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

where $\epsilon_{i1} \sim \mathcal{N}(0, 2.93^2)$ and $\epsilon_{i2} \sim \mathcal{N}(0, 4.41^2)$ for $i = 1, \ldots, 49$. The estimated regression lines from this fit are overlaid on the scatterplot of the PD mode data in Figure 4(c). Based on the calculated posterior membership probabilities, the blue triangles are those observations assigned to the first component and the red bullets are those observations assigned to the second component. While our fit identified two clear components, the clusterings are clearly affected by the time since trigger variable. Such a clustering affected by the predictor variable is called *assignment dependence*, and is treated extensively by Hennig (2000). Such a feature can be incorporated via the use of cluster-weighted models (see Gershenfeld, 1997; Ingrassia *et al.*, 2012, 2014). While our model is not a cluster-weighted model, we do note what it is identifying in this particular part of our analysis. Referring again to Figure 4(c), the red vertical dashed line is the break line of time before and after $T = 280$s. As discussed, data points with $T > 280$s are considered as flaring points, and those points classified to the second component give strong evidence in favor of this flaring assumption as they have a noticeably different linear structure than those datasets before 280s. Thus, the fit from our proposed mixture model gives evidence to the presence of a structural changepoint at this time of $T = 280$s.

6. Conclusion

Measurement error in a response variable is considered as intrinsic scatter when incorporated as part of astronomical regression models. In this paper, we discussed a mixture-of-regressions model where measurement error is treated in the response. We extended the WLS method proposed by Akritas and Bershady (1996) to the mixture setting, and used likelihood methods to compute the estimates of the parameters. Our proposed model differs from the mixture-of-regressions model introduced by Yao and Song (2015), who modeled measurement error in the predictors.

We conducted extensive simulation studies to characterize the performance of our WLS-based algorithm to reflect weighting from the intrinsic scatter. The simulation study included combinations of 2-component and 3-component models having either one or two predictors, various degrees of separability between the components, and difference amounts of variability assumed for the measurement error. The overall results show that our method can improve the performance of estimates, especially when the measurement error is not too large. It is often the case that proposed numerical procedures for measurement error models perform best when the measurement error is not too large. Moreover, mixture models with well-separated components tend to do better in terms of their MSE and relative efficiencies when compared to the naïve estimators that do not reflect the measurement error. Again, numerical procedures for finite mixture models tend to do better under model settings with well-separated components.

Our model was motivated as a way to analyze GRB data, for which we do have a reliable estimate for the variability of the measurement error in the response variable. In particular, a 2-component mixture-of-regressions model is tenable since it can be used to characterize those flux measurements that are likely to be occurring during the flaring portion of the GRB's X-ray decay. Our model was able to make use of all of the reported data, and provided a more nuanced view of these GRB data.

There are various considerations for future research to expand on the work presented in this paper. For example, a more formal inference framework could be implemented for determining the number of components. While we just applied model selection criteria in our paper, one could proceed to perform (nonparametric) bootstrapping (McLachlan, 1987). Moreover, one could investigate bootstrapping for developing certain goodness-of-fit tests of our proposed model, some of which have appealing asymptotic properties (Babu and Rao, 2004).

Another possibility is to consider more flexibility to our general model. For example, one might assume something other than Gaussian components for the mixture structure used in this paper to achieve greater flexibility in the modeling process. Moreover, modifications to Algorithm 1 could be investigated to handle different assumptions on the measurement error δ_i. For example, one obvious setting is where the η_i are unknown, which is likely to be the more common situation encountered in practice. Another possibility is that the measurement error could also be conditioned on component membership k_i, resulting in η_i being replaced by η_{k_i} in the variance in (5). However, such an assumption surely has added identifiability issues that would require further constraints in order to perform estimation.

Another direction is how clustering can be affected by the predictor variable, which is a limitation with our work that we briefly mentioned at the end of Section 5. In the analysis of the PD mode data of the GRB, the predictor would be time since trigger. Expanding our proposed mixture-of-regressions model to also incorporate such assignment dependence would be a more flexible generalization. A cluster-weighted model could be a viable extension to our approach as it could provide a reasonable mechanism to handle measurement errors in both the response and predictor variables.

Acknowledgements

The authors would first like to thank an anonymous reviewer for helpful comments and suggestions about this manuscript. The authors are also grateful for the invitation to contribute to this special issue in memory of Prof. C. R. Rao. While a PhD student at Penn State University, the third author of this paper recalls many memories of Prof. Rao in the Department of Statistics, especially with how he engaged with graduate students. One

personal memory was when the author presented their dissertation work during the event held for awarding the C. R. and Bhargavi Rao Prize. After the talk, Prof. Rao approached them, gave them praise for their research, and asked for a copy of their dissertation. Such a gesture was quite impactful for a statistician-in-training.

References

Aitken, A. C. (1935). On least squares and linear combination of observations. *Proceedings of the Royal Society of Edinburgh*, **55**, 42–48.

Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. In Petrov, B. N. and Csaki, F., editors, *Second International Symposium on Information Theory*, pages 267–281. Akademiai Kiado, Budapest.

Akritas, M. G. and Bershady, M. A. (1996). Linear regression for astronomical data with measurement errors and intrinsic scatter. *The Astrophysical Journal*, **470**, 706–714.

Andrae, R. (2010). Error estimation in astronomy: A guide. `https://doi.org/10.48550/arXiv.1009.2755`. 1–23.

Babu, G. J. and Rao, C. R. (2004). Goodness-of-fit tests when parameters are estimated. *Sankhyā: The Indian Journal of Statistics*, **66**, 63–74.

Barthelmy, S. D. (2004). Burst Alert Telescope (BAT) on the Swift MIDEX mission. In *Proc. SPIE 5165, X-Ray and Gamma-Ray Instrumentation for Astronomy XIII*, pages 1–15.

Baudry, J.-P., Raftery, A. E., Celeux, G., Lo, K., and Gottardo, R. (2010). Combining mixture components for clustering. *Journal of Computational and Graphical Statistics*, **9**, 332–353.

Benaglia, T., Chauveau, D., Hunter, D. R., and Young, D. S. (2009). mixtools: An R package for analyzing finite mixture models. *Journal of Statistical Software*, **32**, 1–29.

Bertoletti, M., Friel, N., and Rastelli, R. (2015). Choosing the number of clusters in a finite mixture model using an exact integrated completed likelihood criterion. *Metron*, **73**, 177–199.

Biernacki, C., Celeux, G., and Govaert, G. (2000). Assessing a mixture model for clustering with the integrated completed likelihood. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **22**, 719–725.

Blustin, A. J., Band, D., Barthelmy, S., Boyd, P., Capalbi, M., Holland, S. T., Marshall, F. E., Mason, K. O., Perri, M., Poole, T., and 55 others (2006). Swift panchromatic observations of the bright gamma-ray burst GRB 050525a. *The Astrophysical Journal*, **637**, 901–913.

Bozdogan, H. (1987). Model selection and Akaike's information criterion (AIC): The general theory and its analytical extensions. *Psychometrika*, **52**, 345–370.

Buonaccorsi, J. P. (2010). *Measurement Error: Models, Methods, and Applications.* Chapman and Hall/CRC, New York, NY.

Carmichael, B. and Coën, A. (2008). Asset pricing with errors-in-variables˙ *Journal of Empirical Finance*, **15**, 778–788.

Carroll, R. J., Midthune, D., Freedman, L. S., and Kipnis, V. (2006a). Seemingly unrelated measurement error models, with application to nutritional epidemiology. *Biometrics*, **62**, 75–84.

Carroll, R. J., Ruppert, D., Stefanski, L. A., and Crainiceanu, C. (2006b). *Measurement Error in Nonlinear Models: A Modern Perspective.* Chapman and Hall/CRC, New York, NY, 2^{nd} edition.

Clutton-Brock, M. (1967). Likelihood distributions for estimating functions when both variables are subject to error. *Technometrics*, **9**, 261–269.

Davison, A. C. and Hinkley, D. (1997). *Bootstrap Methods and Their Application.* Cambridge University Press, New York, NY.

De Veaux, R. D. (1989). Mixtures of linear regressions. *Computational Statistics and Data Analysis*, **8**, 227–245.

Dellaportas, P. and Stephens, D. A. (1995). Bayesian analysis of errors-in-variables regression models. *Biometrics*, **51**, 1085–1095.

Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society, Series B (Statistical Methodology)*, **39**, 1–38.

Efron, B. and Hinkley, D. V. (1978). Assessing the accuracy of the maximum likelihood estimator: Observed versus expected Fisher information. *Biometrika*, **65**, 457–482.

Efron, B. and Tibshirani, R. J. (1993). *An Introduction To The Bootstrap.* Chapman and Hall/CRC, New York, NY.

Fang, X., Chen, A. W., and Young, D. S. (2023). Predictors with measurement error in mixtures of polynomial regressions. *Computational Statistics*, **38**, 373–401.

Feigelson, E. D. and Babu, G. J. (1992). Linear regression in astronomy. II. *The Astrophysical Journal*, **397**, 55–67.

Feigelson, E. D., de Souza, R. S., Ishida, E. E. O., and Babu, G. J. (2021). 21st century statistical and computational challenges in astrophysics. *Annual Review of Statistics and Its Application*, **8**, 493–517.

Frisch, R. (1935). Statistical confluence analysis by means of complete regression systems. *The Economic Journal*, **45**, 741–742.

Frühwirth-Schnatter, S. (2006). *Finite Mixture and Markov Switching Models.* Springer, New York, NY.

Fuller, W. A. (1987). *Measurement Error Models.* John Wiley & Sons, Inc., New York, NY.

Gehrels, N., Chincarini, G., Giommi, P., Mason, K., Nousek, J., Wells, A. A., White, N. E., Barthelmy, S. D., Burrows, D. N., Cominsky, L. R., and Hurley, K. C. (2004). The Swift gamma-ray burst mission. *The Astrophysical Journal*, **611**, 1005–1020.

Gershenfeld, N. (1997). Nonliner inference and cluster-weighted modeling. *Annals of the New York Academy of Sciences*, **808**, 18–24.

Gustafson, P. (2004). *Measurement Error and Misclassification in Statistics and Epidemiology: Impacts and Bayesian Adjustments.* Chapman and Hall/CRC.

Hennig, C. (2000). Identifiablity of models for clusterwise linear regression. *Journal of Classification*, **17**, 273–296.

Hurn, M., Justel, A., and Robert, C. P. (2003). Estimating mixtures of regressions. *Journal of Computational and Graphical Statistics*, **12**, 55–79.

Ingrassia, S., Minotti, S. C., and Punzo, A. (2014). Model-based clustering via linear cluster-weighted models. *Computational Statistics and Data Analysis*, **71**, 159–182.

Ingrassia, S., Minotti, S. C., and Vittadini, G. (2012). Local statistical modeling via a cluster-weighted approach with elliptical distributions. *Journal of Classification*, **29**, 363–401.

Kelly, B. C. (2007). Some aspects of measurement error in linear regression of astronomical data. *The Astrophysical Journal*, **665**, 1489–1506.

Kelly, B. C. (2012). Measurement error models in astronomy. In Feigelson, E. D. and Babu, G. J., editors, *Statistical Challenges in Modern Astronomy V*, pages 147–162. Springer-Verlag, New York, NY.

Kuhn, M. A. and Feigelson, E. D. (2019). Mixture models in astronomy. In Fruhwirth-Schnatter, S., Celeux, G., and Robert, C. P., editors, *Handbook of Mixture Analysis*, chapter 19, pages 463–483. CRC Press.

Lange, K. (2010). *Numerical Analysis for Statisticians*. Springer, New York, NY, 2$^{\text{nd}}$ edition.

Lindsay, B. G. (1995). *Mixture Models: Theory, Geometry and Applications*, volume 5 of *NSF-CBMS Regional Conference Series in Probability and Statistics*. Institute of Mathematical Statistics and the American Statistical Association.

Louis, T. A. (1982). Finding the observed information matrix when using the EM algorithm. *Journal of the Royal Statistical Society. Series B (Methodological)*, **44**, 226–233.

Maddala, G. S. and Nimalendran, M. (1996). 17 errors-in-variables problems in financial models. In Maddala, G. S. and Rao, C. R., editors, *Handbook of Statistics, Volume 14: Statistical Methods in Finance*, pages 507–528. North Holland - Elsevier, Amsterdam, Netherlands.

McLachlan, G. J. (1987). 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**, 318–324.

McLachlan, G. J., Lee, S. X., and Rathnayake, S. I. (2019). Finite mixture models. *Annual Review of Statistics and Its Application*, **6**, 355–378.

McLachlan, G. J. and Peel, D. (2000). *Finite Mixture Models*. Wiley, New York, NY.

Mengersen, K. L., Robert, C. P., and Titterington, D. M., editors (2011). *Mixtures: Estimation and Applications*, West Sussex, England. Wiley.

Morrison, H. L., Olszewski, E. W., Mateo, M., Norris, J. E., Harding, P., Dohm-Palmer, R. C., and Freeman, K. C. (2000). Mapping the galactic halo. IV. Finding distant giants reliably with the Washington System. *The American Astronomical Society*, **121**, 37–40.

Murillo, A. L., Affuso, O., Peterson, C. M., Li, P., Wiener, H. W., Tekwe, C. D., and Allison, D. B. (2019). Illustration of measurement error models for reducing bias in nutrition and obesity research using 2-d body composition data. *Obesity*, **27**, 489–495.

Piran, T., Bromberg, O., Nakar, E., and Sari, R. (2013). The long, the short and the weak: The origin of gamma-ray bursts. *Philosophical Transactions of the Royal Society A*, **371**, 1–10.

Press, W. H., Teukolsky, S. A., Vetterling, W. T., and Flannery, B. P. (1992). *Numerical Recipes in FORTRAN: The Art of Scientific Computing*. Cambridge University Press.

Richardson, S. and Gilks, W. R. (1993). A Bayesian approach to measurement error problems in epidemiology using conditional independence models. *American Journal of Epidemiology*, **138**, 430–442.

Richardson, S., Leblond, L., Jaussent, I., and Green, P. J. (2002). Mixture Models in Measurement Error Problems, with Reference to Epidemiological Studies. *Journal of the Royal Statistical Society, Series A (Statistics in Society)*, **165**, 549–566.

Schwarz, G. (1978). Estimating the dimension of a model. *The Annals of Statistics*, **6**, 461–464.

Stephens, D. A. and Dellaportas, P. (1992). Bayesian inference of generalized linear models with covariate measurement errors. In Bernardo, J. M., Berger, J. O., Dawid, A. P., and Smith, A. F. M., editors, *Bayesian Statistics 4*, pages 813–820. Oxford University Press, Oxford, UK.

Tarnopolski, M. (2019). Multivariate analysis of BATSE gamma-ray burst properties using skewed distributions. *The Astrophysical Journal*, **887**, 1–9.

Titterington, D. M., Smith, A. F. M., and Makov, U. E. (1985). *Statistical Analysis of Finite Mixture Distributions*. Wiley, New York, NY.

Viele, K. and Tong, B. (2002). Modeling with mixtures of linear regressions. *Statistics and Computing*, **12**, 315–330.

Wedel, M. and DeSarbo, W. S. (1995). A mixture likelihood approach for generalized linear models. *Journal of Classification*, **12**, 21–55.

Yao, W. and Song, W. (2015). Mixtures of linear regression with measurement errors. *Communications in Statistics - Theory and Methods*, **44**, 1602–1614.

Young, D. S. (2014). Mixtures of regressions with changepoints. *Statistics and Computing*, **24**, 265–281.