

IMPROVED EM-TYPE ALGORITHMS FOR FITTING MARGINAL ZERO-INFLATED
REGRESSION MODELS TO CLUSTERED DATA WITH EXCESS ZEROS

by

TAWANDA BENESI

(Under the direction of Daniel B. Hall)

ABSTRACT

Clustered data with excess zeros are becoming increasingly common in many research areas including health sciences, business, manufacturing and natural resource management. There is a need to improve existing methods and to create software routines that can handle such data. Marginal zero-inflated (ZI) regression models have been studied. They are two component mixtures of a degenerate component with a point mass of one at zero and a non-degenerate component suitable for the scale of the response. The EM algorithm is convenient for fitting mixture models with a closed form likelihood, however, marginal ZI regression models are partially specified models for dependent data that generally have an intractable likelihood function. EM-type algorithms that account for dependence in the data and extend beyond likelihood based formulation are desired. The expectation-solution (ES) algorithm is an EM-type algorithm that can be used to fit marginal ZI regression models to account for dependence in the data using GEEs. Nonetheless, the ES algorithm assumes a simplifying assumption of independence in the data to obtain a closed-form solution for the E-step and this is inefficient for clustered data. Marginal ZI regression models with multinomial mixing are proposed. This ensures we account for dependence in the mixing component and fit a fully parametric model instead of a GEE for the S-step of the mixing data. In addition, we propose the regression-solution (RS) algorithm to explicitly account for

dependence in the data when updating the “missing data in an EM-like formulation. The RS algorithm replaces the E-step of the ES algorithm by using an appropriate regression model to estimate the conditional expectation of the missing data given the observed data and current parameter estimates. In simulation studies, we demonstrate that models with a multinomial mixing structure fitted with the RS algorithm produce more efficient regression parameter estimators than those using the ES algorithm. We demonstrate with an example data set and introduce an R package, **margZIfit**, for implementing the model fitting algorithms.

INDEX WORDS: marginal models, zero inflation, clustered data, excess zeros,
 multinomial mixing, ES algorithm, RS algorithm, **margZIfit** package

IMPROVED EM-TYPE ALGORITHMS FOR FITTING MARGINAL ZERO-INFLATED
REGRESSION MODELS TO CLUSTERED DATA WITH EXCESS ZEROS

by

TAWANDA BENESI

B.S., University of Zimbabwe, Zimbabwe, 2010

M.S., University of Zimbabwe, Zimbabwe, 2014

M.S., University of Georgia, 2018

A Dissertation Submitted to the Graduate Faculty
of The University of Georgia in Partial Fulfillment
of the

Requirements for the Degree

DOCTOR OF PHILOSOPHY

ATHENS, GEORGIA

2020

©2020

Tawanda Benesi

All Rights Reserved

IMPROVED EM-TYPE ALGORITHMS FOR FITTING MARGINAL ZERO-INFLATED
REGRESSION MODELS TO CLUSTERED DATA WITH EXCESS ZEROS

by

TAWANDA BENESI

Major Professor: Daniel B. Hall

Committee: Nicole Lazar
Cheolwoo Park
Liang Liu
Jaxk Reeves

Electronic Version Approved:

Ron Walcott
Interim Dean of the Graduate School
The University of Georgia
May 2020

Dedicated to my wife, children and family ...

Acknowledgments

I thank God Almighty for his provision, protection, guidance and strength throughout this research. I give special thanks to my wife and family for their unwavering support and encouragement throughout my PhD program. Special mention goes to Professor Daniel B. Hall for guiding me relentlessly throughout this research. I have learned a lot from you as my major professor; to be selfless in helping and guiding others and to be always optimistic even when the challenges seem insurmountable. I thank my doctoral committee members for their contributions towards the completion of this manuscript. I thank the Department of Statistics for providing support throughout my academic program. Thank you all for making this research complete in many ways than one!

Contents

Acknowledgments	v
1 Introduction	1
1.1 Overview	1
1.2 Dissertation layout	5
1.3 Fitting algorithms for marginal ZI regression models	6
1.3.1 The ES-1 algorithm	7
1.3.2 The ES-2 algorithm	9
1.4 ZI regression models	11
1.5 Marginal ZI regression models for clustered data	13
2 Marginal zero-inflated regression models for clustered data with a multi- nomial mixing structure	17
2.1 Introduction	17
2.2 Background	20
2.2.1 Marginal ZI regression models for clustered data	21
2.2.2 Fitting algorithms for marginal ZI regression models	24
2.3 Marginal ZI regression models: multinomial mixing models	24
2.3.1 The EM algorithm for fitting marginal ZI regression models: multino- mial mixing models	27

2.3.2	The ES-1 algorithm for fitting marginal ZI regression models: multinomial mixing structure	29
2.3.3	The ES-2 algorithm for fitting marginal ZI regression models: multinomial mixing structure	30
2.4	Extensions to larger and/or unequal cluster sizes	33
2.4.1	A multinomial model for zero-inflation at each measurement occasion	34
2.4.2	A multinomial model for zero-inflation patterns based on the maximal cluster size	35
2.4.3	Multinomial models via log-linear models	37
2.5	Simulations	39
2.5.1	Other simulations	42
2.6	Example	43
2.7	Summary	45
3	A Regression-Solution algorithm for marginal zero-inflated regression mod- els with multinomial mixing	54
3.1	Introduction	54
3.2	Background	56
3.2.1	Marginal ZI regression models for clustered data with excess zeros assuming binary mixing	57
3.2.2	Marginal ZI regression models for clustered data with excess zeros: multinomial mixing	59
3.3	Regression solution (RS) algorithm for marginal ZI regression models	60
3.3.1	Linear discriminant analysis and multinomial logit models	61
3.3.2	The regression-solution (RS) algorithm	64
3.3.3	Extensions to larger and/or unbalanced clusters	68

3.4	Simulations	69
3.4.1	Comparing different R-step models	72
3.5	Example	72
3.6	Summary	75
3.6.1	Recommendations and future work	76
4	Estimation of marginal ZI regression models in R: The margZIfit package	84
4.1	Introduction	84
4.2	Background	86
4.2.1	Marginal ZI regression models: binary mixing structure	87
4.2.2	Fitting marginal ZI regression models via the ES algorithm	88
4.2.3	Marginal ZI regression models: multinomial mixing	89
4.2.4	Fitting marginal ZI regression models: multinomial mixing via the ES algorithm	90
4.2.5	Fitting marginal ZI regression models: multinomial mixing via the RS algorithm	91
4.3	Data handling and model description	92
4.3.1	Loading the package and dependencies	92
4.3.2	Data handling	92
4.3.3	Model description	94
4.4	Example	96
4.4.1	Marginal ZI regression models: binary mixing structure	97
4.4.2	Marginal ZI regression models: multinomial mixing structure (ES algorithm)	100
4.4.3	Marginal ZI regression models: multinomial mixing structure (RS algorithm)	104

4.5	Summary	108
5	Summary	111
5.1	Summary and conclusions	111
5.2	Recommendations and future work	112
5.3	Derivation of the RS algorithm for multinomial mixing models	121
5.4	Derivation of the complete data loglikelihood:	
	multinomial mixing	124
5.4.1	Estimating function for the ES-2 algorithm to marginal ZI regression	
	models for clustered data with excess zeros	125
5.4.2	Multinomial logit models score function	126
5.4.3	ES-2 algorithm for multinomial mixing models	128

List of Tables

2.1	A table showing all the possible patterns of zero inflation for the case $n_i = 2$	25
2.2	Unbalanced clusters example with just two clusters	34
2.3	A simulation study comparing the ratios of MSEs for the Binary Mixing and the Multinomial Mixing structures	47
2.4	A simulation study for comparing the Binary Mixing and the Multinomial Mixing structures (Correlation parameter $\rho = 0.3$)	47
2.5	A simulation study for comparing the Binary Mixing and the Multinomial Mixing structures (Test size (5%) and Coverage (95%): Correlation parameter $\rho = 0.3$)	48
2.6	A simulation study for comparing the Binary Mixing and the Multinomial Mixing structures (Correlation parameter $\rho = 0.8$)	49
2.7	A simulation study for comparing the Binary Mixing and the Multinomial Mixing Structures (Test size (5%) and Coverage(95%): Correlation parameter $\rho = 0.8$)	50
2.8	A simulation study for comparing the Binary Mixing and the Multinomial Mixing structures (Mis-specified Working Correlation with $\rho = 0.8$ and $K = 100$)	51
2.9	A simulation study for comparing the Binary Mixing and the Multinomial Mixing structures	51
2.10	A simulation study for comparing the computation times for the Binary Mixing and the Multinomial Mixing structures	52

2.11	Investigating zero-inflation in the data	52
2.12	Results of the hospital visits data using the ES-1 algorithms for multinomial and binary mixing models	52
2.13	Effects of time on hospital visits rates	53
3.1	Patterns of the mixing mechanism	64
3.2	A simulation study for comparing the <i>ES_BMX</i> , <i>ES_MMX</i> and <i>RS_MMX</i> (Correlation parameter $\rho = 0.3$)	78
3.3	A simulation study for comparing the <i>ES_BMX</i> , <i>ES_MMX</i> and <i>RS_MMX</i> (Test and Coverage: Correlation parameter $\rho = 0.3$)	79
3.4	A simulation study for comparing the <i>ES_BMX</i> , <i>ES_MMX</i> and <i>RS_MMX</i> (Correlation parameter $\rho = 0.8$)	80
3.5	A simulation study for comparing the <i>ES_BMX</i> , <i>ES_MMX</i> and <i>RS_MMX</i> (Test and Coverage: Correlation parameter $\rho = 0.8$)	81
3.6	A simulation study to compare inclusion/exclusion of linear/quadratic terms in the R-step of the RS algorithm	82
3.7	Hospital visits data set fitted using the <i>ES_BMX</i> , <i>ES_MMX</i> and <i>RS_MMX</i> .	82
3.8	Effects of time on hospital visits rates	83
3.9	Observed and predicted zeros	83
5.1	A simulation study comparing binary mixing and multinomial mixing struc- tures (correlation parameter $\rho = 0.25$)	117
5.2	A simulation study comparing binary mixing and multinomial mixing struc- tures (Test(5%) and Coverage(95%): Correlation parameter $\rho = 0.25$)	118
5.3	A simulation study comparing binary mixing and multinomial mixing struc- tures (correlation parameter $\rho = 0.75$)	119

- 5.4 A simulation study comparing binary mixing and multinomial mixing structures (Test size (5%) and Coverage(90%): correlation parameter $\rho = 0.75$) . 120

1. Introduction

1.1 Overview

Data with excess zeros (EZ) are often encountered in many research disciplines such as in biomedical research, business, manufacturing, epidemiology, fish and wildlife conservation, and psychology. Often, the occurrence of zeros exceeds what a standard distribution can predict, causing overdispersion. Zero-inflated (ZI) models are often the framework of choice for such data. ZI regression models mix a degenerate distribution with a point mass of one at zero and a standard distribution. The mixture form accounts for the excess zeros through the degenerate distribution and some zeros through the standard distribution. This differs from zero-altered models of Heilbron [1994] and Dobbie and Welsh [2001] which utilize a truncated distribution to model the zeros separately from the non-zero data. Examples of ZI models include the ZI Poisson (ZIP) models (Lambert [1992]), ZI negative binomial (ZINB) models for unbounded counts (Ridout et al. [2001]) and ZI binomial (ZIB) models for bounded counts (Hall [2000]).

EZ data are common in cross-sectional studies and are becoming increasingly observed in longitudinal studies where the data are measured repeatedly over time. Models that handle both correlated data and excess zeros are necessary for such studies. ZI mixed effects models with subject-specific random effects are a common class of models that can be used for correlated data with excess zeros. These models have been considered by many authors including Hall [2000] and Yau and Lee [2001].

The parameter estimates from ZI mixed effects models quantify the effects of covariates on the conditional means that depend on latent subject-level random effects yielding subject-specific interpretations. Often, such covariate effects are not the quantities of primary interest; instead population-level inferences may be desired.

Marginal and marginalized ZI regression models yield population-averaged interpretations rather than subject-specific interpretations. Marginal ZI regression models that specify the regression effects directly on the marginal means have been considered by Hall and Zhang [2004] who proposed a broad class of marginal ZI regression models for clustered data with excess zeros. Kassahun et al. [2014] considered marginalized zero-inflated models which specify the regression effects on the marginalized mean of the ZI mixed effects regression model and hurdle regression models which models the zeros and the counts separately through truncation. The aforementioned papers focus on ZI models in which the marginal mean of the non-degenerate component of the mixture is modeled in the context of longitudinal or clustered data. A separate vein of research focuses on ZI models in which the marginal mean response, averaged over the mixture, is modeled directly rather than modeling the mean of the non-degenerate component. Somewhat confusingly, such models are also known as both marginal (and, sometimes, marginalized) ZI regression models. For example, Martin and Hall [2017], Long et al. [2014] and Preisser et al. [2016] have investigated and applied marginalized regression models for data with excess zeros in cross-sectional studies. Martin and Hall (2018, *preprint*) extended marginalized regression models to the clustered data context. Although modeling the marginal mean, averaged over the mixture, has considerable appeal in some contexts, in the current work we focus on marginally specified ZI regression models in the tradition studied by Hall and Zhang.

The EM algorithm is a convenient method for fitting mixture models if the likelihood is available. However, marginal ZI regression models are mixture models that are partially specified because a suitable, tractable multivariate joint distribution is unavailable. The EM algorithm is then difficult to implement without simplifying assumptions and there is need to develop more general EM-type algorithms for these models.

The expectation-solution (ES) algorithm proposed by Rosen et al. [2000] is an extension of the EM algorithm to clustered data problems for mixtures of generalized linear models. The E-step imputes the “missing” data by taking the conditional expectation of the complete data loglikelihood function, (specified under an independence assumption) given the observed data and the current parameter estimates. The S-step updates parameter estimates using generalized estimating equations (GEEs) with working correlation matrices to account for the dependence in the data.

The E-step of the ES algorithm is computed under the simplifying assumption of independence in the data to obtain a closed form solution, but without such an assumption the computation does not in general, have a closed form solution. It is well known in the clustered data literature that methods that do not account for the dependence in the data will lead to inefficient parameter estimators. Therefore, there is a need to study and develop new EM-type algorithms that take into account the dependence in the data throughout the model fitting steps.

The main focus of this dissertation is to investigate and develop improved EM-type algorithms for fitting marginal zero-inflated (ZI) regression models to clustered data with excess zeros. Our goal is to develop a model framework that allows us to account for the dependence in the data more thoroughly, with the goal of improving estimation and inference on the regression parameters of central interest. In this dissertation we propose marginal ZI regression models for clustered data assuming a multinomial mixing structure for the patterns of zero-inflation. The multinomial mixing model assumes a multinomial distribution which

accounts for dependence in the mixing process automatically when updating regression parameters pertaining to that process rather than through the more ad hoc approach taken by Hall and Zhang [2004] who assumed binary mixing at each time point and introduced GEEs in the S step to account for dependence in these latent binary random variables.

The use of a multinomial mixing structure rather than GEEs to account for dependence in the mixing process is more appealing in part because it avoids the need to specify a working correlation structure for the mixing process. We study this approach, in conjunction with an independence-based E-step and an ES algorithm formulation in Chapter 2. However, we seek to drop the assumption of independence in the E-step as well, in the pursuit of more efficient estimation and inference. To this end, in Chapter 3 we propose the regression-solution (RS) algorithm for marginal ZI regression models assuming a multinomial mixing structure. In the RS algorithm, the E-step of the ES algorithm is replaced with an R-step which estimates the conditional expectation of the missing data given the observed data and current parameter estimates using an appropriate regression model. In particular, the regression-step fits a discrete-choice multinomial logit regression model to impute the missing data. The regression step ensures that we account for the dependence in the observed data when imputing the missing data, a feature ignored by the E-step of the ES algorithm. We expect the RS algorithm to produce estimators that have better statistical properties compared to the existing ES algorithm. In addition, we expect the RS algorithm to be computationally faster than the ES algorithm because at the S-step for the zero component we fit a multinomial logit model for the mixing patterns of zero inflation instead of a generalized estimating equation (GEE) with an *ad hoc* working correlation matrix for binary data.

Another aspect of the current work is to investigate an alternative implementation of our EM-like algorithms. In particular, we consider the approach of Elashoff and Ryan [2004] who proposed an approach whereby the E and M steps of the traditional EM algorithm can be unified into a single set of estimating equation. These authors also considered cases in

which the M-step of the traditional EM problem does not necessarily correspond to the maximization of an objective function. Thus, their work addressed a class of algorithms inclusive of the ES approach of Rosen et al. [2000] and, potentially, our RS algorithm. Confusingly, Elashoff and Ryan termed their methodology an ES algorithm. To avoid confusion, we distinguish the approaches of Rosen et al. [2000] as ES-1 and ES-2, respectively, and emphasize that they are wholly different in their purpose and implementation.

Elashoff and Ryan’s ES-2 approach offers two main advantages over a traditional EM implementation. The first is computational speed, which can potentially be achieved because of the unification of the E and M steps, which yields a singly iterative algorithm rather than the traditional EM approach, which is typically doubly-iterative (whenever the E and/or M steps are not available in closed form). The second advantage is that the ES-2 approach yields an estimate of the asymptotic covariance matrix of the parameter estimates at convergence, obviating a separate calculation of this quantity.

1.2 Dissertation layout

This dissertation comprises chapters. The first is a brief overview of the topics and the problem we are trying to solve and provides summaries of some relevant background material. The second chapter introduces marginal ZI regression models for clustered data with excess zeros assuming a multinomial mixing structure. The ES-1 algorithm (Hall and Zhang [2004]) and the ES-2 algorithm (Elashoff and Ryan [2004]) are developed for fitting the models. An example is presented to illustrate the methodology and simulations are provided to investigate its performance in practice. In chapter three we introduce the regression-solution (RS) algorithm for marginal ZI regression models assuming the multinomial mixing structure. Again, after describing the methodology in detail we illustrate it with an example and present simulations to evaluate its finite sample performance. Chapter four presents details of an

R package to implement the methodologies developed in this dissertation as well as the ES-1 approach previously developed by Hall and Zhang [2004]. Chapter five summarizes the results, gives some recommendations for the proposed model framework, and mentions some directions for future work.

1.3 Fitting algorithms for marginal ZI regression models

The EM algorithm is a convenient method for obtaining maximum likelihood estimates in the presence of “missing data”. Mixture models can be described more simply by assuming that each observed data point has a corresponding masked data point which specifies which component of the mixture each data point belongs to. ZI regression models fall under the mixture model formulation with two components, the degenerate distribution for the excess zeros and the standard distribution for the observed response. The EM algorithm relies heavily on the availability of a likelihood function; however, marginal ZI regression models are partially specified models for which a tractable likelihood function is unavailable. Many authors have proposed extensions of the EM algorithm for cases where the likelihood function is unavailable or intractable. These methods include the expectation-solution (ES) algorithm by Rosen et al. [2000], the projection-solution (PS) algorithm by Heyde and Morton [1996] and the expectation-solution (ES) algorithm by Elashoff and Ryan [2004] which is different in implementation to the ES algorithm by Rosen et al. [2000]. We shall consider the ES algorithm by Rosen et al. [2000] which hereafter will be called the ES-1 algorithm and the ES algorithm by Elashoff and Ryan [2004] which we shall call the ES-2 algorithm.

1.3.1 The ES-1 algorithm

Rosen et al. [2000] proposed the *expectation-solution* algorithm for clustered data when the true likelihood is intractable or is unavailable. The fitting algorithm is a generalization of the EM algorithm to so-called mixtures of experts, in generalized linear models for correlated data. Marginal ZI regression models are a special case of mixtures of experts with a two component mixture. At the E-step of the ES-1 algorithm, latent component membership data are imputed using the conditional expectation of the complete data *quasi-loglikelihood* given the observed data and current parameter estimates under the independence assumption in the data. The S-step relaxes the independence assumption and solves *generalized estimating equations* (GEEs) with working correlation matrices to account for dependence in the data. The ES-1 algorithm was implemented by Hall and Zhang [2004] for marginal ZI regression models for clustered data with excess zeros and by Hall and Shen [2015] for marginal models for clustered directional data.

The ES-1 algorithm as described in Rosen et al. [2000] is given below. Let the response random vector be denoted by $\mathbf{y}_i = (y_{i1}, \dots, y_{in_i})^T$ with mean vector $\boldsymbol{\mu}_i = (\mu_{i1}, \dots, \mu_{in_i})^T$ and an $n_i \times p$ covariate matrix $\mathbf{X}_i = (\mathbf{x}_{i1}, \dots, \mathbf{x}_{in_i})^T$, where $\mathbf{x}_{ij}$ is a p -dimensional covariate vector for the j^{th} subject in the i^{th} cluster for $i = 1, \dots, K$ and $j = 1, 2, \dots, n_i$. Each data point is selected from a set of R distribution components based on the covariate vector $\mathbf{x}_{ij}$, and the selected component distribution maps $\mathbf{x}_{ij}$ to the response y_{ij} . An $R \times 1$ vector $\boldsymbol{\omega}_r$ with a 1 at the r^{th} position and zeros elsewhere is used to label the r^{th} of the R components chosen from a multinomial distribution with probability $p(\boldsymbol{\omega}_r | \mathbf{x}_{ij}, \boldsymbol{\Psi}) = \pi_{rij}$, where $\boldsymbol{\Psi} = (\boldsymbol{\psi}_1, \dots, \boldsymbol{\psi}_{R-1})$ is the matrix of parameters underlying the multinomial distribution.

In addition, let $\mu_{rij} = f(\boldsymbol{\beta}_r^T \mathbf{x}_{ij})$ and $\boldsymbol{\mu}_{ri} = (\mu_{ri1}, \dots, \mu_{rin_i})^T$. Given $\boldsymbol{\omega}_r$, y_{ij} is generated with probability $p(y_{ij} | \mathbf{x}_{ij}, \boldsymbol{\beta}_r, \phi_r)$, where $\boldsymbol{\beta}_r$ and ϕ_r are the regression parameter vector and dispersion parameter, respectively.

E-step: Similar to the EM algorithm for mixtures, at the $(k + 1)^{th}$ step, the E-step corresponds to computing the membership probabilities given by:

$$u_{rij}^{(k+1)} = \frac{\pi_{rij}^{(k)} P(y_{ij} | \mathbf{x}_{ij}, \boldsymbol{\beta}_r^{(k)}, \phi_r^{(k)})}{\sum_{h=1}^R \pi_{rij}^{(k)} P(y_{ij} | \mathbf{x}_{ij}, \boldsymbol{\beta}_r^{(k)}, \phi_r^{(k)})}.$$

S-step: To update the parameter estimates, we solve a GEE (Liang and Zeger [1986]) with a working correlation matrix to account for the dependence in the data as follows:

$$\mathbf{U}^{(k)} = \sum_{i=1}^K \mathbf{D}_{ri}^T [\mathbf{V}_i(\alpha_r^{(k)}, \phi_r^{(k)})]^{-1} \mathbf{H}_{ri}^{(k)} (\mathbf{y}_i - \boldsymbol{\mu}_{ri}^{(k)}) = \mathbf{0}, \quad r = 1, \dots, R,$$

where $\mathbf{V}_i(\alpha_r^{(k)}, \phi_r^{(k)}) = \phi_r^{(k)} \mathbf{A}_{ri}^{1/2} \mathbf{R}_i(\alpha_r^{(k)}) \mathbf{A}_{ri}^{1/2}$, $\mathbf{D}_{ri} = \partial \boldsymbol{\mu}_{ri}^{(k)} / \partial \boldsymbol{\beta}_r^{T(k)}$, $\mathbf{H}_{ri}^{(k)} = (u_{ri1}^{(k)}, \dots, u_{rin_i}^{(k)})^T$ and $\mathbf{A}_{ri} = (\phi_r^{(k)} v(\mu_{ri1}^{(k)}), \dots, \phi_r^{(k)} v(\mu_{rin_i}^{(k)}))^T$. The $\pi_{rij}^{(k)}$ are obtained by fitting a multinomial logit model on the diagonal elements of $\mathbf{H}_{ri}$. The $\hat{\phi}_r$'s and $\hat{\alpha}_r$'s are $\sqrt{K}$ -consistent estimators, typically obtained from second-order estimating equations or a method of moments approach.

If $\mathbf{U}^{(k)}$ is unbiased, then the resulting parameter estimators are consistent and asymptotically normal for inference on the true parameters as the number of clusters increases, i.e., as $K \rightarrow \infty$. Under certain regularity conditions, the estimated asymptotic variance, $\text{avar}(\boldsymbol{\theta})$, is obtained from $\text{avar}(\hat{\boldsymbol{\theta}}) = \hat{I}^{-1} \hat{V}^{-1} \hat{I}^{-T}$ where

$$\hat{I} = \sum_{i=1}^K \sum_{j=1}^{n_i} \sum_{r=1}^R \nabla s_{ij}(\omega_r; \hat{\boldsymbol{\theta}}),$$

and

$$\hat{V} = \sum_{i=1}^K \left\{ \sum_{j=1}^{n_i} \sum_{r=1}^R s_{ij}(\omega_r; \hat{\boldsymbol{\theta}}) \right\} \left\{ \sum_{j=1}^{n_i} \sum_{r=1}^R s_{ij}(\omega_r; \hat{\boldsymbol{\theta}}) \right\}^T,$$

where $\boldsymbol{\theta}$ is the combined parameter vector with ES estimator $\hat{\boldsymbol{\theta}}$ and $s_{ij}(\omega_r; \hat{\boldsymbol{\theta}})$ contains the estimating functions of all the parameters involved in the fitting algorithm.

1.3.2 The ES-2 algorithm

Elashoff and Ryan [2004] proposed a more general algorithm for partially specified models which can be used for “missing” data problems involving any form of estimating equations. The algorithm is called the *expectation-solution* (ES) algorithm, but it is a wholly different methodology compared to the ES algorithm by Rosen et al. [2000]. The E-step for the ES-2 algorithm corresponds to when the complete data estimating equations are replaced by their corresponding conditional expected values, that is, we take a conditional expectation of the estimating function given the observed data and current parameter values and equate to zero. The S-step then solves the resulting imputed estimating equation.

Let us describe the ES-2 algorithm in more detail. Let the complete data be denoted by $\mathbf{z} = (\mathbf{y}^T, \mathbf{u}^T)^T$, where $\mathbf{y}$ is the observed data and $\mathbf{u}$ is the missing data. Suppose the complete data estimating function can be written or decomposed into the form:

$$\begin{aligned} \mathbf{U}^c(\boldsymbol{\theta}; \mathbf{z}) &= \mathbf{U}_1(\boldsymbol{\theta}; \mathbf{y}, \mathbf{S}(\mathbf{z})) = \sum_{j=1}^q \mathbf{a}_j(\boldsymbol{\theta}) S_j(\mathbf{z}) + \mathbf{b}_{\boldsymbol{\theta}}(\mathbf{y}) \\ &\equiv \mathbf{A}_{\boldsymbol{\theta}} \mathbf{S}(\mathbf{z}) + \mathbf{b}_{\boldsymbol{\theta}}(\mathbf{y}), \end{aligned} \tag{1.1}$$

where $\mathbf{a}_j(\boldsymbol{\theta})$ is a $p \times 1$ vector corresponding to the j^{th} column of a $p \times q$ matrix $\mathbf{A}_{\boldsymbol{\theta}}$ and $S_j(\mathbf{z})$ is the j^{th} element of a q -dimensional vector $\mathbf{S}(\mathbf{z})$ called the *complete data summary statistic*. The complete data summary statistic is in most cases of lesser dimension than the original

missing data. Note that the 1 subscript for $\mathbf{U}_1$ in equation (1.1) is introduced because we will introduce a second estimating function to accomplish the E-step. First, though note that the classical EM algorithm (generalized to this context where the M-step might involve the solution of an estimating equation rather than a true maximization step) proceeds by iterating two steps, as follows:

E-step: At the $(k+1)^{\text{th}}$ step, given the current estimate $\boldsymbol{\theta}^{(k)}$, the E-step can be done in terms of the “complete data summary statistic” $\mathbf{S}$ as follows

$$\hat{\mathbf{S}}^{(k+1)} \equiv h(\mathbf{y}; \hat{\boldsymbol{\theta}}^{(k)}) = \mathbb{E}[\mathbf{S}(\mathbf{z})|\mathbf{y}; \hat{\boldsymbol{\theta}}^{(k)}].$$

S-step: Replacing $\mathbf{S}(\mathbf{z})$ with $\hat{\mathbf{S}}^{(k+1)}$ leads to a new estimating equation $\mathbf{U}_1(\boldsymbol{\theta}; \mathbf{y}, \hat{\mathbf{S}}^{(k+1)}) = \mathbf{0}$, involving only the observed data $\mathbf{y}$ and which is solved to get the updated estimate $\hat{\boldsymbol{\theta}}^{(k+1)}$.

This approach can be very appealing for the following reasons. First, greater computational efficiency can be achieved because the estimating equation for the E-step need not be of the dimension of the entire missing data vector, but can be of lower dimension based on what they termed the “complete data summary statistic”, which is essentially a function of the missing data through which the E-step depends. Second, Elashoff and Ryan [2004] observed that the E-step can be accomplished as the solution of an estimating equation for $\mathbf{S}$ given by $U_2(\mathbf{S}; \hat{\boldsymbol{\theta}}^{(k)}, \mathbf{y}) = \mathbf{S} - h(\mathbf{y}; \hat{\boldsymbol{\theta}}^{(k)}) = \mathbf{0}$. This allows the algorithm to simultaneously compute the E-step and the S-step via a $(p+q) \times 1$ joint estimating function $\mathbf{U}(\boldsymbol{\delta}; \mathbf{y}) = \begin{pmatrix} \mathbf{U}_1 \\ \mathbf{U}_2 \end{pmatrix}$, where $\boldsymbol{\delta} = (\boldsymbol{\theta}^T, \mathbf{S}^T)^T$. Proceeding in this way obviates an alternating algorithm which makes the method more efficient. Third, by combining the estimating equations for the E-step and S-step via a joint estimating equation leads to a more convenient way of computing the standard errors for the parameter estimates. Specifically, under appropriate regularity

conditions, the algorithm produces a consistent estimator $\boldsymbol{\theta}_0$, the true value of the unknown parameter $\boldsymbol{\theta}$. Let $\mathbf{D}(\boldsymbol{\delta})$ be the $(p+q) \times (p+q)$ matrix of partial derivatives of $\mathbf{U}(\boldsymbol{\delta}; \mathbf{y})$ with respect to $\boldsymbol{\delta}^T$ and $\mathbf{V}$ be the covariance matrix of $\mathbf{U}(\boldsymbol{\delta}; \mathbf{y})$, then $\sqrt{n}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)$ is asymptotically normal with mean $\mathbf{0}$ and variance given by the $p \times p$ upper left-hand corner of $\boldsymbol{\Omega}^{-1} \boldsymbol{\Lambda} (\boldsymbol{\Omega}^{-1})^T$, where $(1/n)\mathbf{V} \xrightarrow{p} \boldsymbol{\Lambda}$ and $(1/n)\mathbf{D} \xrightarrow{p} \boldsymbol{\Omega}$ as $n \xrightarrow{p} \infty$. If we let $\mathbf{U}_i(\boldsymbol{\delta})$ be the i^{th} independent contribution to $\mathbf{U}(\boldsymbol{\delta}; \mathbf{y})$, then the asymptotic variance-covariance matrix of parameter estimates, $\text{avar}(\hat{\boldsymbol{\theta}})$, can be consistently estimated by the $p \times p$ upper left-hand corner of $\mathbf{D}^{-1}(\boldsymbol{\delta}) (\sum_i \mathbf{U}_i(\boldsymbol{\delta}) \mathbf{U}_i(\boldsymbol{\delta})^T) [\mathbf{D}^{-1}(\boldsymbol{\delta})]^T|_{\boldsymbol{\delta}=\hat{\boldsymbol{\delta}}}$, where $\mathbf{D}(\boldsymbol{\delta})$ is the $(p+q) \times (p+q)$ matrix of partial derivatives of $\mathbf{U}(\boldsymbol{\delta}; \mathbf{y})$ with respect to $\boldsymbol{\delta}^T$.

1.4 ZI regression models

Zero-inflated (ZI) regression models describe the relationship between explanatory variables and an outcome exhibiting excess zeros than are expected under the distribution underlying the outcome variable. Fitting a standard distribution to these data will lead to overdispersion. Regression is built in through the mean of the standard distribution F_2 , say, and, possibly, through the mixing probabilities. ZI models are presented in mixture form to account for excess zeros through a degenerate distribution with a point mass of one at zero and expected zeros through a standard distribution. If the standard distribution is the Poisson then we have the zero-inflated Poisson (ZIP) by Lambert [1992]. Other examples include zero inflated negative binomial (ZINB) models of Ridout et al. [2001]; and zero-inflated binomial (ZIB) models by Hall [2000].

In ZI models, the response variable is distributed as $Y_i \sim 0$ with probability π_i for the degenerate distribution and $Y_i \sim F_2$ with probability $1 - \pi_i$ for the standard probability distribution F_2 , which we have density/probability function that we denote as f_2 . For illustration, we adopt a ZI Poisson (ZIP) regression model by taking F_2 to be the Poisson(λ_i) distribution.

Suppose $\mathbf{y} = (y_1, \dots, y_K)^T$ is the vector of observed counts corresponding to independent random variables $Y_1, \dots, Y_K$ as defined above, where $f_2(y_i) = \frac{e^{-\lambda_i} \lambda_i^{y_i}}{y_i!}$. The parameters $\boldsymbol{\lambda} = (\lambda_1, \dots, \lambda_K)^T$ and $\boldsymbol{\pi} = (\pi_1, \dots, \pi_K)^T$ are assumed to depend on the covariates through GLM-type specifications as follows

$$\log(\boldsymbol{\lambda}) = \mathbf{B}\boldsymbol{\beta},$$

for the mean of the Poisson counts and

$$\text{logit}(\boldsymbol{\pi}) = \mathbf{G}\boldsymbol{\gamma},$$

for the mixing probabilities. Here, $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$ are unknown regression parameters, and $\mathbf{B}$ and $\mathbf{G}$ are respectively the $K \times p$ and $K \times q$ model matrices which contain the covariates used in the models.

Let $u_i = 1$ if $Y_i \sim 0$ and $u_i = 0$ otherwise, then the complete data loglikelihood for a ZIP model is given by

$$\begin{aligned} l^c(\boldsymbol{\beta}, \boldsymbol{\gamma}; \mathbf{y}, \mathbf{u}) &= \sum_{i=1}^K \{u_i \log \pi_i(\boldsymbol{\gamma}) + (1 - u_i) \log [1 - \pi_i(\boldsymbol{\gamma})]\} + \\ &\quad \sum_{i=1}^K (1 - u_i) \log f_2(y_i; \lambda_i(\boldsymbol{\beta})) \\ &\equiv l^c(\boldsymbol{\gamma}; \mathbf{y}, \mathbf{u}) + l^c(\boldsymbol{\beta}; \mathbf{y}, \mathbf{u}), \end{aligned}$$

where $f_2(y_i; \lambda_i(\boldsymbol{\beta})) = e^{\lambda_i} \lambda_i^{y_i} / (y_i)!$ is the Poisson probability mass function (pmf). The EM algorithm proceeds as follows:

E-step: At the $(k+1)^{\text{th}}$ step of the EM algorithm, the E-step is computed by taking the expectation with respect to the conditional distribution of the complete data loglikelihood given the observed data $\mathbf{y}$ and the current parameter estimates $\boldsymbol{\beta}^{(k)}$ and $\boldsymbol{\gamma}^{(k)}$ which yields $l^c(\boldsymbol{\beta}, \boldsymbol{\gamma}; \mathbf{y}, \mathbf{u}^{(k+1)})$ where $\mathbf{u}^{(k+1)}$ has i th element

$$\begin{aligned} u_i^{(k+1)} &= P(u_i = 1 | y_i, \boldsymbol{\beta}^{(k)}, \boldsymbol{\gamma}^{(k)}) \\ &= I\{y_i = 0\} [1 + (1 - \pi_i(\boldsymbol{\gamma}^{(k)})) f_2(y_i; \boldsymbol{\beta}^{(k)}) / \pi_i(\boldsymbol{\gamma}^{(k)})]^{-1}. \end{aligned}$$

M-step: The $(k+1)^{\text{th}}$ step estimates $\boldsymbol{\gamma}^{(k+1)}$ and $\boldsymbol{\beta}^{(k+1)}$ are then conveniently obtained by fitting a binomial generalized linear model (GLM) to the imputed mixing data $\mathbf{u}^{(k+1)}$ and a weighted Poisson GLM for the observed counts. The convenience of performing the M step via separate regressions for $\boldsymbol{\gamma}$ and $\boldsymbol{\beta}$ follows from the decomposition of the expected complete data loglikelihood whereby these parameters appear in separate terms, each of which has the form of the loglikelihood for a standard (weighted) generalized linear model.

The EM algorithm iterates by updating the E-step weights, $\mathbf{u}^{(k+1)}$, at the $(k+1)^{\text{th}}$ step and fitting the two GLMs to update the parameter estimates at the M-step until convergence.

1.5 Marginal ZI regression models for clustered data

Extensions of ZI regression models to correlated data have been studied. One area of study involves extending ZIP and ZIB models to clustered data using cluster-specific random effects into the model (Hall [2000]). Such models are conditionally specified, given the values of random effects. Alternatively, marginally specified ZI regression models have been considered. For example, Moulton et al. [2002] treated the data as independent during model fitting pro-

cess and then used a robust sandwich estimate of the parameter variance-covariance matrix to adjust for within-cluster dependence in the data. Hall and Zhang [2004] implemented the ES-1 algorithm which incorporates generalized estimating equations (GEEs) with working correlation matrices to account for dependence in the data.

Hall and Zhang's approach is most easily understood as a modification of the EM algorithm as formulated under independence. Let $\mathbf{y}_i$ be the $n_i \times 1$ vector of responses for the i^{th} cluster, $i = 1, \dots, K$. Assume that the random variable Y_{ij} for the y_{ij} observation follows a marginal ZI distribution, that is, $Y_{ij} \sim 0$, with probability π_{ij} and $Y_{ij} \sim F_2$, with probability $1 - \pi_{ij}$, where $F_2(\mu_{ij})$ is a probability distribution with mean μ_{ij} . It is assumed that the means $\boldsymbol{\mu}_i = (\mu_{i1}, \dots, \mu_{in_i})^T$ and the mixing probabilities $\boldsymbol{\pi}_i = (\pi_{i1}, \dots, \pi_{in_i})^T$ are respectively related to covariates through the GLMs $g_\mu(\boldsymbol{\mu}_i) = \boldsymbol{\eta}_i = \mathbf{B}_i\boldsymbol{\beta}$ and $g_p(\boldsymbol{\pi}_i) = \mathbf{G}_i\boldsymbol{\gamma}$, where $g_\mu(\cdot)$ is a link function suitable for the scale of the response and $g_p(\cdot)$ is a link function for binary data such as the logit link, $\mathbf{B}_i$ and $\mathbf{G}_i$ are respectively the $n_i \times p$ and $n_i \times q$ model matrices. $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$ are the unknown regression parameter vectors.

Let $u_{ij} = 1$ if Y_{ij} is from the degenerate zero distribution and $u_{ij} = 0$ otherwise, for $i = 1, \dots, K$, and $j = 1, \dots, n_i$. Treating the u_{ij} 's as latent random variables, then under independence the 'complete data' loglikelihood based on $\mathbf{y} = (\mathbf{y}_1^T, \dots, \mathbf{y}_K^T)^T$ and $\mathbf{u} = (\mathbf{u}_1^T, \dots, \mathbf{u}_K^T)^T$ is given by

$$\begin{aligned} l^c(\boldsymbol{\beta}, \phi, \boldsymbol{\gamma}; \mathbf{y}, \mathbf{u}) &= \sum_{i,j} \{u_{ij} \log \pi_{ij}(\boldsymbol{\gamma}) + (1 - u_{ij}) \log [1 - \pi_{ij}(\boldsymbol{\gamma})]\} + \\ &\quad \sum_{i,j} (1 - u_{ij}) \log f_2(y_{ij}; \boldsymbol{\beta}, \phi) \\ &\equiv l^c(\boldsymbol{\gamma}; \mathbf{y}, \mathbf{u}) + l^c(\boldsymbol{\beta}, \phi; \mathbf{y}, \mathbf{u}). \end{aligned}$$

The ES-1 algorithm for marginal ZI regression models for clustered data with excess zeros is described in Hall and Zhang [2004] as follows:

E-step: At the $(k+1)^{\text{th}}$ step of the EM algorithm, the E-step is computed from $u_{ij}^{(k+1)} = I\{y_{ij} = 0\}[1 + \{1 - \pi_{ij}(\boldsymbol{\gamma}^{(k)})\}f_2(y_{ij}; \boldsymbol{\beta}^{(k)}, \phi^{(k)})/\pi_{ij}(\boldsymbol{\gamma}^{(k)})]^{-1}$ for $i = 1, 2, \dots, K$, $j = 1, 2, \dots, n_i$.

S-step: Instead of fitting two GLMs as in the classical EM algorithm, we solve respectively two GEEs with working correlation matrices to account for the within-cluster correlation. To estimate the correlation and scale parameters in addition to the regression parameters, Hall and Zhang used the extended GEEs Hall and Severini [1998], which leads to a set of expanded estimating equations at the S-step of the form

$$\sum_{i=1}^K \begin{pmatrix} \partial \boldsymbol{\mu}_i^T / \partial \boldsymbol{\beta} & \mathbf{0} \\ \mathbf{0} & \partial \boldsymbol{\sigma}_i^T / \partial \tilde{\boldsymbol{\rho}} \end{pmatrix} \begin{pmatrix} \mathbf{V}_{i11}^{-1} & \mathbf{0} \\ \mathbf{0} & \mathbf{V}_{i22}^{-1} \end{pmatrix} \mathbf{H}_i^{(k+1)} \begin{pmatrix} \mathbf{y}_i - \boldsymbol{\mu}_i \\ \mathbf{s}_i - \boldsymbol{\sigma}_i \end{pmatrix} = \mathbf{0},$$

and

$$\sum_{i=1}^K \begin{pmatrix} \partial \boldsymbol{\pi}_i^T / \partial \boldsymbol{\gamma} & \mathbf{0} \\ \mathbf{0} & \partial \boldsymbol{\tau}_i^T / \partial \boldsymbol{\delta} \end{pmatrix} \begin{pmatrix} \mathbf{W}_{i11}^{-1} & \mathbf{0} \\ \mathbf{0} & \mathbf{W}_{i22}^{-1} \end{pmatrix} \begin{pmatrix} \mathbf{u}_i^{(k+1)} - \boldsymbol{\pi}_i \\ \mathbf{t}_i - \boldsymbol{\tau}_i \end{pmatrix} = \mathbf{0},$$

where $\tilde{\boldsymbol{\rho}} = (\boldsymbol{\rho}^T, \phi)^T$, $\mathbf{V}_{i11} = \mathbf{D}_i^{1/2} \{\boldsymbol{\mu}_i(\boldsymbol{\beta})\} \mathbf{P}(\boldsymbol{\rho}) \mathbf{D}_i^{1/2} \{\boldsymbol{\mu}_i(\boldsymbol{\beta})\}$ for $\mathbf{D}_i \{\boldsymbol{\mu}_i(\boldsymbol{\beta})\} = \text{diag}\{\phi\nu(\mu_i(\boldsymbol{\beta}))/\omega_{i1}, \dots, \phi\nu(\mu_i(\boldsymbol{\beta}))/\omega_{in_i}\}$ and $\mathbf{P}(\boldsymbol{\rho})$ is a working correlation matrix with parameter $\boldsymbol{\rho}$, $\mathbf{s}_i = \text{vech}[(\mathbf{y}_i - \boldsymbol{\mu}_i)(\mathbf{y}_i - \boldsymbol{\mu}_i)^T]$, $\boldsymbol{\sigma}_i = E(\mathbf{s}_i) = \text{vech}(\mathbf{V}_{i11})$, $\mathbf{H}_i^{(k+1)} = \text{diag}[(\mathbf{j}_{n_i} - \mathbf{u}_i^{(k+1)}), \text{vech}[(\mathbf{j}_{n_i} - \mathbf{u}_i^{(k+1)})(\mathbf{j}_{n_i} - \mathbf{u}_i^{(k+1)})^T]]$ for $\mathbf{j}_{n_i}$ is a $n_i \times 1$ vector of ones. $\mathbf{W}_{i11} = \mathbf{A}_i^{1/2} \{\boldsymbol{\pi}_i(\boldsymbol{\gamma})\} \mathbf{R}(\boldsymbol{\delta}) \mathbf{A}_i^{1/2} \{\boldsymbol{\pi}_i(\boldsymbol{\gamma})\}$ for $\mathbf{A}_i \{\boldsymbol{\pi}_i(\boldsymbol{\gamma})\} = \text{diag}\{\pi_{i1}(1 - \pi_{i1}), \dots, \pi_{in_i}(1 - \pi_{in_i})\}$ and $\mathbf{R}(\boldsymbol{\delta})$ is a working correlation matrix with parameter $\boldsymbol{\delta}$, $\mathbf{t}_i = \text{vech}[(\mathbf{u}_i^{(k+1)} - \boldsymbol{\pi}_i)(\mathbf{u}_i^{(k+1)} - \boldsymbol{\pi}_i)^T]$, $\boldsymbol{\tau}_i = E(\mathbf{t}_i) = \text{vech}(\mathbf{W}_{i11})$.

The function $\text{vech}(\cdot)$ is a vector-valued function that takes a square matrix as an input and returns a vector of the lower triangular elements including the diagonal. $\mathbf{V}_{i22}$ and $\mathbf{W}_{i22}$ are weight matrices for the second moment estimating equations for $\boldsymbol{\rho}$ and $\boldsymbol{\delta}$ and Gaussian working structures are used. For example, $\mathbf{V}_{i22}$ has elements given by the relation $\text{cov}(s_{ijk}, s_{ilm}) = \sigma_{ijl}\sigma_{ikm} + \sigma_{ijm}\sigma_{ikl}$ where $s_{ijk} = (y_{ij} - \mu_{ij})(y_{ik} - \mu_{ik})$ and $\sigma_{ijk} = E(s_{ijk})$ are the elements of $\mathbf{s}_i$ and $\boldsymbol{\sigma}_i$, respectively.

Having described the scope of the dissertation and given the relevant background for our work, we now move on to develop the novel methodology, starting in Chapter 2, which introduces marginal ZI regression models for clustered data with a multinomial mixing structure. Chapter 3 discusses the RS algorithm for fitting marginal ZI regression models with a multinomial mixing structure. The `margZIfit` package in chapter 4 discusses the implementation of the methods described in this dissertation in R.

2. Marginal zero-inflated regression models for clustered data with a multinomial mixing structure

2.1 Introduction

There remains interest in many application areas such as business, epidemiology, biomedical health and many others, to analyze clustered data with excess zeros. Often, the occurrence of observed zeros tend to exceed what a standard probability distribution can predict. In such cases, zero-inflated (ZI) regression models can be used to account for the presence of excess zeros. ZI models are mixture models that comprise of a degenerate distribution with a point mass of one at zero and a standard distribution.

ZI mixed effects regression models with subject-specific random effects can be used to account for within-cluster association when subject-specific interpretations of the parameter estimates are desired (for example, Hall [2000] and Yau and Lee [2001]). Often however, our primary interest will be on population-averaged interpretations of the parameter estimates and, in that case, marginal regression models or marginalized regression models are desirable. Marginalized regression models provide inference regarding the marginal mean from a conditionally specified model by integrating over the random effects distribution. Examples of marginalized regression models for clustered data with excess zeros can be found in Kassahun et al. [2014] and Martin and Hall (2018, *preprint*). Marginal regression models assume regression effects directly on the marginal mean. Marginal ZI regression models for clustered data with excess zeros were studied by Hall and Zhang [2004] and others.

The expectation-maximization (EM) algorithm is a convenient method for finding parameter estimates in mixture models if the model is fully specified with a complete data loglikelihood in closed form. However, marginal ZI regression models are partially specified models with assumptions made on the means of the component distributions. The likelihood function for a marginal ZI regression model is generally intractable due to dependence in the data and the lack of flexible multivariate distributions. The EM algorithm is hard to implement without simplifying assumptions and alternative EM-type algorithms are needed to estimate the parameters.

Hall and Zhang [2004] applied the expectation-solution (ES) algorithm introduced by Rosen et al. [2000] for generalized regression models of mixtures of experts in clustered data problems to marginal ZI regression models for clustered data with excess zeros. The ES algorithm is an extension of the EM algorithm to clustered data problems and it alternates between an E-step and a S-step which are, respectively, analogous to the E-step and the M-step of the EM algorithm.

The ES algorithm makes a simplifying assumption of independence in the data to obtain a closed form solution of the E-step. The E-step imputes the missing data using the conditional expectation of the complete data quasi-loglikelihood given the observed data and current parameter estimates under independence. Instead of maximizing the imputed quasi-loglikelihood function to update the parameter estimates, the S-step relaxes the assumption of independence and solves generalized estimating equations (GEEs) with working correlation matrices for each component of the mixture to account for dependence in the data.

An obvious deficiency of this approach is the fact that the E-step of the ES algorithm is computed under independence. In clustered data literature it is well-known that ignoring within-cluster correlation typically leads to statistically inefficient parameter estimators. Our goal is to improve the E-step of the ES algorithm by accounting for dependence in the data.

We propose marginal ZI regression models for clustered data with excess zeros that assume a multinomial mixing structure on the patterns of zero-inflation. In contrast, Hall and Zhang assume a binary mixing structure by assuming independence in the data.

The multinomial mixing structure ensures that the E-step accounts for the dependence in the mixing mechanism by modeling it with a multivariate distribution with a non-diagonal covariance matrix. Furthermore, the S-step for the parameters of the mixing model avoids the use of GEEs and fits a fully parametric model. Use of a GEE with a working correlation matrix for the mixing data is both *ad hoc* and somewhat ill-suited to clustered binary data, for which correlations are not suitable measures of association.

The ES algorithm is used to fit marginal ZI regression models assuming a multinomial mixing structure to get a closed form solution of the E-step. We explore another EM-type algorithm which generalizes the EM algorithm in instances where the likelihood is intractable or unavailable. The algorithm is called the expectation-solution (ES) algorithm by Elashoff and Ryan [2004]. This is however a wholly different methodology compared to the ES algorithm by Rosen et al. [2000]. The ES algorithm by Rosen et al. [2000] will be called the ES-1 algorithm and that of Elashoff and Ryan [2004] will be called the ES-2 algorithm to distinguish between the two.

The ES-2 algorithm may have advantages compared to other fitting algorithms available for the following reasons. First, the E-step depends solely on a quantity involving the complete data which they called the “complete data summary statistic” and is generally of lower dimension than the actual “missing data”. Thus, greater computational efficiency will be expected because imputation is done on the conditional expectation of the complete data summary statistic given the observed data instead of the conditional expectation of the “missing data” given the observed data and current parameter estimates. Second, more computational efficiency can be achieved by casting the E-step as an estimating equation problem and combining the E-step and S-step into a single estimating equation to be solved

simultaneously. This reduces the usual ES approach from a doubly-iterative algorithm to a singly-iterative one. Third, the joint estimating equation for the E-step and the S-steps leads to a more convenient way of computing standard errors for the parameter estimates compared to other EM-type algorithms.

In section 2.2 we describe the background literature on marginal ZI regression models and EM-type algorithms in cases without a tractable likelihood function. We focus on applications to marginal ZI regression models for clustered data with excess zeros. Section 2.3 introduces marginal ZI regression models for clustered data with excess zeros assuming a multinomial mixing structure for the patterns of zero-inflation. The ES-1 and ES-2 algorithms are implemented to marginal ZI regression models for clustered data with excess zeros for both the binary mixing and the multinomial mixing structures. We present a simulation study in section 2.4 to assess and compare the statistical properties of the estimators for the multinomial and the binary mixing structures. In section 2.5 we apply our models to an example data set on the number of hospital visits. Lastly, we conclude and summarize our model results and discuss recommendations, limitations and future research.

2.2 Background

In this section we review marginal ZI regression models for clustered data with excess zeros. We discuss the fitting algorithms that can be used when the likelihood is intractable and limit our attention to the ES algorithm by Rosen et al. [2000] and the ES algorithm by Elashoff and Ryan [2004]. We implement the ES-1 and ES-2 algorithms to marginal ZI regression models for clustered data with excess zeros.

2.2.1 Marginal ZI regression models for clustered data

It is straight-forward to extend ordinary ZI regression models to marginal ZI regression models for clustered data with excess zeros. We let $\mathbf{y}_i$ be the $n_i \times 1$ vector of responses for the i^{th} cluster, $i = 1, \dots, K$. Assume that the random variable Y_{ij} for the y_{ij} observation follows a marginal ZI distribution, that is,

$$Y_{ij} \sim \begin{cases} 0, & \text{with probability } \pi_{ij} \\ F_2(\mu_{ij}), & \text{with probability } 1 - \pi_{ij}, \end{cases}$$

where $F_2(\mu_{ij})$ is a probability distribution function with mean μ_{ij} . It is assumed that the mean $\boldsymbol{\mu}_i = (\mu_{i1}, \dots, \mu_{in_i})^T$ and the mixing probabilities $\boldsymbol{\pi}_i = (\pi_{i1}, \dots, \pi_{in_i})^T$ are respectively related to covariates through a set of generalized linear models (GLMs)

$$g_\mu(\boldsymbol{\mu}_i) = \boldsymbol{\eta}_i = \mathbf{B}_i \boldsymbol{\beta},$$

for the non-degenerate response distribution and

$$g_p(\boldsymbol{\pi}_i) = \mathbf{G}_i \boldsymbol{\gamma},$$

for the degenerate component of the mixture. Here $g_\mu(\cdot)$ is a link function suitable for the scale of the response and $g_p(\cdot)$ is a link function for a binary outcome such as the logit link, $\mathbf{B}_i$ and $\mathbf{G}_i$ are respectively the $n_i \times p$ and $n_i \times q$ model matrices. $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$ are the unknown regression parameter vectors.

Let $u_{ij} = 1$ if Y_{ij} is from the degenerate zero distribution and 0 otherwise, for $i = 1, \dots, K$, and $j = 1, \dots, n_i$. Note that the u_{ij} 's are latent random variables. Under independence the

‘complete data’ loglikelihood based on $\mathbf{y} = (\mathbf{y}_1^T, \dots, \mathbf{y}_K^T)^T$ and $\mathbf{u} = (\mathbf{u}_1^T, \dots, \mathbf{u}_K^T)^T$ is given by

$$\begin{aligned} l^c(\boldsymbol{\beta}, \phi, \boldsymbol{\gamma}; \mathbf{y}, \mathbf{u}) &= \sum_{i,j} \{u_{ij} \log \pi_{ij}(\boldsymbol{\gamma}) + (1 - u_{ij}) \log [1 - \pi_{ij}(\boldsymbol{\gamma})]\} + \\ &\quad \sum_{i,j} (1 - u_{ij}) \log f_2(y_{ij}; \boldsymbol{\beta}, \phi) \\ &\equiv l^c(\boldsymbol{\gamma}; \mathbf{y}, \mathbf{u}) + l^c(\boldsymbol{\beta}, \phi; \mathbf{y}, \mathbf{u}). \end{aligned}$$

Hall and Zhang [2004] used the ES-1 algorithm to estimate the parameters. The E-step imputes the missing data $u_{ij}^{(k+1)} = E(u_{ij} | y_{ij}; \boldsymbol{\psi}^{(k)})$ where $\boldsymbol{\psi}^{(k)} = (\boldsymbol{\gamma}^{T(k)}, \boldsymbol{\beta}^{T(k)}, \boldsymbol{\rho}^{T(k)}, \boldsymbol{\delta}^{T(k)})^T$ is the vector of the current parameter estimates. The S-step updates the parameters by fitting a set of GEEs with working correlation matrices to account for dependence in the data. A binomial GEE is fitted for the degenerate component mixing data and a weighted GEE for example, a weighted Poisson GEE is fitted for the non-degenerate component data. To estimate the correlation parameters $\boldsymbol{\rho}$ and $\boldsymbol{\delta}$ in addition to the regression parameters $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$, Hall and Zhang [2004] used the extended GEEs (Hall and Severini [1998]).

Rosen et al. [2000] established that if the ES-1 algorithm converges, then the resulting parameter estimators solve an unbiased estimating equation; thus they are consistent and asymptotically normal under suitable regularity conditions. The asymptotic variance-covariance matrix can be estimated from the estimating functions used in the S-step.

In particular, if we let $\boldsymbol{\psi}$ be the combined parameter vector with corresponding ES estimator $\hat{\boldsymbol{\psi}}$, then the asymptotic variance-covariance of the parameter estimates, $\text{avar}(\hat{\boldsymbol{\psi}})$, is consistently estimated by $\hat{\boldsymbol{\xi}}^{-1} \hat{\boldsymbol{\zeta}} (\hat{\boldsymbol{\xi}}^{-1})^T$ where $\hat{\boldsymbol{\xi}} = \sum_{i=1}^K \sum_{j=1}^{n_i} \nabla s_{ij}(\hat{\boldsymbol{\psi}})$ and $\hat{\boldsymbol{\zeta}} = \sum_{i=1}^K \left\{ \sum_{j=1}^{n_i} s_{ij}(\hat{\boldsymbol{\psi}}) \right\} \left\{ \sum_{j=1}^{n_i} s_{ij}(\hat{\boldsymbol{\psi}}) \right\}^T$, where $s_{ij}(\hat{\boldsymbol{\psi}})$ is a $\dim(\boldsymbol{\psi}) \times 1$ vector containing the estimating functions of $\boldsymbol{\gamma}, \boldsymbol{\delta}, \boldsymbol{\beta}, \boldsymbol{\rho}$ and $\boldsymbol{\phi}$ with substituted final parameter estimates at the S-step. Numerical derivatives can also be used to obtain $\nabla s_{ij}(\hat{\boldsymbol{\psi}})$.

The ES-1 algorithm assumes independence to simplify the E-step and then relaxes independence at the S-step to update parameter estimates by solving a set of GEEs with working correlation matrices to account for dependence in the data. Proceeding by computing an E-step which does not account for the dependence in the data will produce inefficient parameter estimators, a phenomenon well known in clustered data literature. Essentially, instead of computing $E(\mathbf{u}_i | \mathbf{y}_i; \boldsymbol{\psi}^{(k+1)})$, we are computing $E(u_{ij} | y_{ij}; \boldsymbol{\psi}^{(k+1)})$ which ignores the correlation both within $\mathbf{u}_i$ and between $\mathbf{u}_i$ and $\mathbf{y}_i$.

Moreover, the S-step for the mixing data is computed using a GEE with a working correlation matrix that is not a natural measure of association for binary data. Binary data are highly constrained and odds ratios (ORs), for example, are more suitable than correlations. There is need to develop a model framework and estimation methodology that accounts properly for dependence in the data and is less reliant on ad hoc approaches that are poorly suited to the data structure. In the next section, we propose marginal ZI regression models for clustered with excess zeros assuming a multinomial mixing structure. We fit the models using the ES-1 and ES-2 algorithms. The multinomial mixing structure aims to account for the dependence in the mixing data using a multinomial distribution for the mixing patterns of zero-inflation. The multinomial distribution is natural for correlated categorical data.

2.2.2 Fitting algorithms for marginal ZI regression models

Marginal ZI regression models are mixture models with two components. The first component is a regression model assumed on the mean of a degenerate distribution with a point mass of one at zero to account for the excess zeros in the data. The second component is a regression model assumed on the mean of a standard non-degenerate distribution. The EM algorithm is a convenient fitting algorithm for finding the regression parameter estimates if the model is fully specified and having a tractable likelihood function.

Marginal ZI regression models are, however, partially specified models without a tractable likelihood function and hence there is need to develop EM-type algorithms that can handle missing data without a likelihood function. Several authors have proposed EM-type algorithms that can be used in situations where the likelihood is intractable or unavailable. Examples include the projection-solution (PS) algorithm (Heyde and Morton [1996]), the simulate-update (SU) algorithm (Satten and Datta [2000]), the approximate-update (AU) algorithm (Zhao et al. [2010]), the ES-1 algorithm (Rosen et al. [2000]), the ES-2 algorithm (Elashoff and Ryan [2004]) and many others. We shall focus our attention to the ES-1 algorithm of Rosen et al. [2000] and the ES-2 algorithm of Elashoff and Ryan [2004].

2.3 Marginal ZI regression models: multinomial mixing models

We shall assume an underlying latent multinomial distribution on the patterns of zero inflation for each cluster. This ensures that we account for the dependence in the mixing data using a multinomial distribution. We assume regression models on the means of the component mixtures and use the ES-1 or the ES-2 algorithms for finding the parameter estimates. Specifically, we assume a multinomial model for the mixing data and a GLM for

the non-degenerate data. The E-step will account for the dependence in the mixing data through the multinomial distribution and the S-step fits a more suitable multinomial model for binary data. The S-step for the mixing data replaces a GEE with an *ad hoc* working correlation matrix by the multinomial model. The model framework is described below.

For each response vector $\mathbf{Y}_i$, for cluster $i = 1, \dots, K$ of length n_i , let ω_i be a categorical random variable taking values in $\{1, 2, \dots, R_i\}$ where $R_i = 2^{n_i}$ is the total number of all possible patterns of zero inflation, $u_{i1}, \dots, u_{in_i}$ for the observation-specific indicators of zero-inflation, for cluster i . Let the vector $\tilde{\omega}_i = (\tilde{\omega}_{i1}, \dots, \tilde{\omega}_{iR_i-1})^T$ be the vector of indicators specifying the value of ω_i . That is, if $\tilde{\omega}_{ir}$ is the r th element of $\tilde{\omega}_i$, then $\tilde{\omega}_{ir} = I(\omega_i = r)$, $r = 1, \dots, R_i - 1$. For simplicity, we present the case $n_i = 2$ for all i (the bivariate case), but extensions to larger clusters and varying cluster sizes n_i is possible. In the bivariate case, the possible values ω_i can take are displayed in Table 2.1, along with the corresponding u_{ij} 's, multinomial probabilities, which we denote by $\boldsymbol{\nu}_i = (\nu_{i1}, \dots, \nu_{iR})^T$, and bivariate means for each mixing class of the patterns. A marginal ZI Poisson regression model is used for illustration.

Table 2.1: A table showing all the possible patterns of zero inflation for the case $n_i = 2$

Prob	u_{i1}	u_{i2}	ω_i	$\tilde{\omega}_i$	mean vector
ν_{i1}	0	0	1	$(1, 0, 0)^T$	$\boldsymbol{\mu}_{i1} = (\lambda_{i1}, \lambda_{i2})^T$
ν_{i2}	0	1	2	$(0, 1, 0)^T$	$\boldsymbol{\mu}_{i2} = (\lambda_{i1}, 0)^T$
ν_{i3}	1	0	3	$(0, 0, 1)^T$	$\boldsymbol{\mu}_{i3} = (0, \lambda_{i2})^T$
ν_{i4}	1	1	4	$(0, 0, 0)^T$	$\boldsymbol{\mu}_{i4} = (0, 0)^T$

In Table 2.1, there are four possible patterns of zero-inflation which the variable ω_i can take representing the mixing patterns of zero-inflation that can be observed by looking at the u_{ij} 's for each cluster. The $\boldsymbol{\mu}_{ir}$ are the bivariate means specifying whether the mean is from the degenerate or the non-degenerate distribution at each time point. Further, we assume the multinomial model

$$\mathbf{g}_\nu(\boldsymbol{\nu}_i) = \mathbf{G}_i^T \boldsymbol{\gamma}, \quad (2.1)$$

for the patterns in the mixing structure and a log-linear regression model

$$\log(\boldsymbol{\lambda}_i) = \mathbf{B}_i \boldsymbol{\beta}, \quad (2.2)$$

for the Poisson counts. Here $\mathbf{g}_\nu(\cdot)$ is a link function suitable for a multinomial response, for example, a multinomial logit model, and $\mathbf{G}_i$ is a design matrix, and the regression parameters $\boldsymbol{\gamma}$ and $\boldsymbol{\beta}$, and $\mathbf{B}_i$ is the design matrix for the Poisson component. This specification ensures the model accounts for the dependence in the mixing data at both the E-step and the S-step of the model fitting. This differs from Hall and Zhang [2004] who accounted for correlation in the mixing structure via GEEs rather than using a distributional assumption that implies dependence.

If we were to assume independence in the non-degenerate component of the distribution, then the EM algorithm could be used to compute the parameter estimates of our model because the model will be fully specified and have a closed form likelihood function. If, instead, we relax the independence assumption then, following Hall and Zhang [2004] the M-step of the EM algorithm can be replaced by a weighted GEE with a working correlation matrix to account for the dependence in the Poisson counts. We propose and explore such an approach in this chapter, which combines a multinomial mixing model; closed form E-step obtained under a simplifying assumption of independence in the non-degenerate distribution; multinomial logit M-step for $\boldsymbol{\gamma}$, the parameters of the mixing model; and a GEE-based S-step for $\boldsymbol{\beta}$, the parameters of the model for the non-degenerate distribution's mean.

Our approach can be seen as an incremental improvement on that of Hall and Zhang that avoids a binary mixing mechanism and ad hoc GEEs for γ in favor of a more fully specified mixing model and M-step for γ . While it is desirable to adopt a more rigorous approach that avoids the simplifying assumption of independence to obtain the E-step and the use of GEEs for β , a formidable obstacle is the specification of a joint multivariate discrete distribution for the cluster-level mixing random variables and the vector of non-degenerate counts. Such distributions are not readily available, unfortunately. Therefore, for now we explore the proposed methodology, with a further refinement that allows dependence to be more thoroughly accounted for in the E-step to be considered in Chapter 3.

2.3.1 The EM algorithm for fitting marginal ZI regression models: multinomial mixing models

Define $Y_{ij} = (1 - U_{ij})Z_{ij}$ where Z_{ij} are independent Poisson random variables with mean λ_{ij} and $U_{ij} = 1$ if $Y_{ij} \sim 0$, and zero otherwise. Note that $Y_{ij} \sim 0$ means that Y_{ij} follows a zero/degenerate distribution. Assuming independence in the Poisson counts, we have the following mixing structure

$$\begin{aligned}\omega_i = 1 &\Rightarrow (Y_{i1}, Y_{i2}) = (Z_{i1}, Z_{i2}), \\ \omega_i = 2 &\Rightarrow (Y_{i1}, Y_{i2}) = (Z_{i1}, 0), \\ \omega_i = 3 &\Rightarrow (Y_{i1}, Y_{i2}) = (0, Z_{i2}), \text{ and} \\ \omega_i = 4 &\Rightarrow (Y_{i1}, Y_{i2}) = (0, 0).\end{aligned}$$

Here, 0 represents that $Y_{ij} \sim 0$, that is, the response variable Y_{ij} follows a zero degenerate distribution. Let $\tilde{\omega}_{ir} = 1$ if $\omega_i = r$ and 0 otherwise and let $P(\tilde{\omega}_{ir} = 1) = \nu_{ir}$ which implies that $P(y_{ij}|\tilde{\omega}_{ir} = 1) = \prod_{j=1}^{n_i} f(y_{ij}; \mu_{ir}(\beta))$. The observed data loglikelihood function is given

by:

$$l(\boldsymbol{\psi}; y_{ij}) = \sum_{i=1}^K \log \left\{ \sum_{r=1}^{R_i} \nu_{ir}(\boldsymbol{\gamma}) \prod_{j=1}^{n_i} f(y_{ij}; \boldsymbol{\mu}_{ir}(\boldsymbol{\beta})) \right\}, \quad (2.3)$$

where $\boldsymbol{\psi} = (\boldsymbol{\gamma}^T, \boldsymbol{\beta}^T)^T$ is vector of regression parameters. The loglikelihood function in equation (2.3) can be maximized using numerical optimization methods. Alternatively, we can use the EM algorithm by introducing the latent random variables $\{\tilde{\omega}_{ir}\}$. Then the ‘complete data’ loglikelihood function is given by:

$$\begin{aligned} l^c(\boldsymbol{\psi}; \mathbf{y}, \tilde{\boldsymbol{\omega}}) &= \sum_{i=1}^K \sum_{r=1}^{R_i} \tilde{\omega}_{ir} \log \nu_{ir}(\boldsymbol{\gamma}) + \sum_{i=1}^K \sum_{r=1}^{R_i} \sum_{j=1}^{n_i} \tilde{\omega}_{ir} \log f(y_{ij} | \boldsymbol{\mu}_{ir}(\boldsymbol{\beta})) \\ &= \sum_{i=1}^K \sum_{r=1}^{R_i} \tilde{\omega}_{ir} \log \nu_{ir}(\boldsymbol{\gamma}) + \sum_{i=1}^K \sum_{j=1}^{n_i} (1 - u_{ij}) \log f(y_{ij} | \boldsymbol{\mu}_{ir}(\boldsymbol{\beta})) \\ &= l^c(\boldsymbol{\gamma}; \mathbf{y}, \tilde{\boldsymbol{\omega}}) + l^c(\boldsymbol{\beta}; \mathbf{y}, \mathbf{u}(\tilde{\boldsymbol{\omega}})), \end{aligned} \quad (2.4)$$

where $\mathbf{u}$ has elements u_{ij} which are computed from the $\tilde{\boldsymbol{\omega}}$ as described below. For simplicity, assume $n_i = n$ for all i so that $R_i = R = 2^n$ for all i . Then ω_i takes values in $\{1, \dots, R\}$ and for $r \in \{1, \dots, R\}$ let $d_1(r), d_2(r), \dots, d_n(r)$, $d_j(r) \in \{0, 1\}$ for all j , denote the binary representation of $r - 1$. For example, for $n = 3$ and $r = 1$, $d_1(r), d_2(r), \dots, d_n(r) = 0, 0, 0$; for $n = 3$ and $r = 8$, $d_1(r), d_2(r), \dots, d_n(r) = 1, 1, 1$. Then the u_{ij} ’s can be computed from $\tilde{\boldsymbol{\omega}}_i$ via

$$u_{ij} = \sum_{r: d_j(r)=1} \tilde{\omega}_{ir}, \quad j = 1, \dots, n. \quad (2.5)$$

The EM algorithm alternates between an E-step and a M-step and proceeds as follows:

E-step: At the $(k+1)^{th}$ step compute the conditional multinomial mixing weights $\tilde{\omega}_{ir}^{(k+1)}$ obtained by taking the conditional expectation of the loglikelihood in equation(2.4) given

the observed data and current parameter estimates to get:

$$\tilde{\omega}_{ir}^{(k+1)} = P(\tilde{\omega}_{ir} = 1 | y_{ij}; \boldsymbol{\psi}^{(k)}) = \frac{\nu_{ir}^{(k)} \prod_{j=1}^{n_i} f(y_{ij}; \boldsymbol{\psi}^{(k)}) I(y_{ij} = 0)}{\sum_{l=1}^R \nu_{il}^{(k)} \prod_{j=1}^{n_i} f(y_{ij}; \boldsymbol{\psi}^{(k)})} \quad (2.6)$$

M-step: The M-step is the maximization of the imputed complete data loglikelihood which is equivalent to fitting a multinomial logit on the imputed multinomial mixing weights $\tilde{\omega}^{(k+1)}$ to update $\boldsymbol{\gamma}$, and a weighted generalized linear model for $\boldsymbol{\beta}$ with weights $\{1 - u_{ij}^{(k+1)}\}$. Here, $u_{ij}^{(k+1)}$ is calculated from the imputed multinomial mixing weights via equation (2.5).

2.3.2 The ES-1 algorithm for fitting marginal ZI regression models: multinomial mixing structure

For clustered data, the EM is modified to account for dependence in the data. The E-step is computed similarly to the E-step of the EM algorithm described above by taking the conditional expectation of the ‘complete data’ quasi-loglikelihood (2.4) given the observed data and current parameter estimates. We relax the assumption of independence and update the parameter estimates by solving a weighted GEE with a working correlation matrix for the Poisson counts and fitting a multinomial logit model for the imputed mixing data. Correlation parameters can be estimated using the method of moments or by extended GEEs of Hall and Severini [1998]. The algorithm is described as follows:

E-step: At the $(k + 1)^{th}$ step compute the conditional mixing probabilities $\tilde{\omega}_{ir}^{(k+1)}$ as shown in equation (2.6).

S-step: Following Hall and Zhang [2004], we relax the assumption of independence on the Poisson counts and account for within-cluster dependence using a weighted GEE with a working correlation matrix to update the parameter estimates. The weights are computed as shown in equation (2.5). The multinomial logit regression model is used to update parameter estimates of the mixing data.

2.3.3 The ES-2 algorithm for fitting marginal ZI regression models: multinomial mixing structure

The ES-2 algorithm can be very useful for fitting marginal ZI regression models for clustered data with excess zeros assuming a binary mixing or a multinomial mixing structure. In both cases we formulate an expanded complete data estimating equation $\mathbf{U}_1^c(\boldsymbol{\psi}; \mathbf{z}) = 0$ involving the GEEs for the degenerate part, the non-degenerate part and the correlation parameters. Here, $\boldsymbol{\psi}$ is a vector with all the regression and correlation parameters and $\mathbf{z}$ is the complete data vector. The complete data estimating functions are decomposed into two terms with one involving the complete data $\mathbf{z}$ via what is known as the “complete data summary statistic” $\mathbf{S}(\mathbf{z})$ and the other depending only on the observed data $\mathbf{y}$ as follows $\mathbf{U}_1^c(\boldsymbol{\psi}; \mathbf{z}) = \mathbf{A}_\psi \mathbf{S}(\mathbf{z}) + \mathbf{b}_\psi(\mathbf{y})$. The complete data summary statistic is in most cases of lesser dimension than the original missing data vector which can be helpful in achieving computational efficiency.

The algorithm in its simple form alternates between an E-step and a S-step until convergence. The E-step imputes the missing data by computing the conditional expectation of the *complete data summary statistic* given the observed data and current parameter estimates at the $(k + 1)^{th}$ step, that is, compute $\hat{\mathbf{S}}^{(k+1)} \equiv h(\mathbf{y}; \hat{\boldsymbol{\psi}}^{(k)}) = E[\mathbf{S}(\mathbf{z})|\mathbf{y}; \hat{\boldsymbol{\psi}}^{(k)}]$. The S-step involves updating the parameter estimates, $\hat{\boldsymbol{\psi}}^{(k+1)}$, by solving the imputed estimating equation $\mathbf{U}_1^c(\mathbf{y}, \hat{\mathbf{S}}^{(k+1)}; \boldsymbol{\psi}) = \mathbf{0}$ which involves only the observed data $\mathbf{y}$.

Alternatively, we can reformulate the E-step and write it as an estimating equation $U_2(\mathbf{y}, \mathbf{S}; \hat{\boldsymbol{\psi}}^{(k)}) = \mathbf{S} - h(\mathbf{y}; \hat{\boldsymbol{\psi}}^{(k)}) = \mathbf{0}$ where we treat $\mathbf{S}$ as an unknown quantity to be estimated together with $\boldsymbol{\psi}$. We then formulate a joint expanded estimating function $\mathbf{U}(\mathbf{y}; \boldsymbol{\Psi}^{(k)}) = \begin{pmatrix} \mathbf{U}_1(\mathbf{y}; \boldsymbol{\Psi}^{(k)}) \\ \mathbf{U}_2(\mathbf{y}; \boldsymbol{\Psi}^{(k)}) \end{pmatrix}$, where $\boldsymbol{\Psi} = (\boldsymbol{\psi}^T, \mathbf{S}^T)^T$. Instead of alternating between an E-step and S-step, the joint expanded estimating equation can be solved simultaneously for $\boldsymbol{\Psi}$ using a multivariate optimizer of choice. From this estimating equation, we can easily obtain the variance-covariance matrix of the parameter estimates, $\text{avar}(\hat{\boldsymbol{\Psi}})$. If we let $U_i(\boldsymbol{\Psi})$ be the i^{th} independent contribution to $\mathbf{U}(\mathbf{y}; \boldsymbol{\Psi})$, then $\text{avar}(\hat{\boldsymbol{\Psi}})$ can be consistently estimated by the $p \times p$ upper left-hand corner of $\mathbf{D}^{-1}(\mathbf{y}; \boldsymbol{\Psi}) \left(\sum_i U_i(\boldsymbol{\Psi}) U_i(\boldsymbol{\Psi})^T \right) [\mathbf{D}^{-1}(\mathbf{y}; \boldsymbol{\Psi})]^T|_{\boldsymbol{\Psi}=\hat{\boldsymbol{\Psi}}}$, where $\mathbf{D}(\mathbf{y}; \boldsymbol{\Psi})$ is the $(p+q) \times (p+q)$ matrix of partial derivatives of $\mathbf{U}(\mathbf{y}; \boldsymbol{\Psi})$ with respect to $\boldsymbol{\Psi}^T$. Under appropriate regularity conditions the ES-2 algorithm produces a consistent estimator of $\boldsymbol{\psi}$, the true value of the unknown parameter.

For ZI regression models with a binary mixing structure, the complete data estimating equation is given by

$$\mathbf{U}_1^c(\boldsymbol{\psi}; \mathbf{z}) = \sum_i^K \begin{bmatrix} \partial \boldsymbol{\mu}_i^T(\boldsymbol{\beta}) / \partial \boldsymbol{\beta} \mathbf{V}_{i11}^{-1} \mathbf{H}_{i1} [\mathbf{y}_i - \boldsymbol{\mu}_i(\boldsymbol{\beta})] \\ \partial \boldsymbol{\pi}_i^T(\boldsymbol{\gamma}) / \partial \boldsymbol{\gamma} \mathbf{W}_{i11}^{-1} [\mathbf{u}_i - \boldsymbol{\pi}_i(\boldsymbol{\gamma})] \\ \partial \boldsymbol{\sigma}_i^T(\boldsymbol{\rho}, \phi) / \partial \tilde{\boldsymbol{\rho}} \mathbf{H}_{i2} [\mathbf{s}_i - \boldsymbol{\sigma}_i(\boldsymbol{\rho}, \phi)] \\ \partial \boldsymbol{\tau}_i^T(\boldsymbol{\delta}) / \partial \boldsymbol{\delta} [\mathbf{t}_i - \boldsymbol{\tau}_i(\boldsymbol{\delta})] \end{bmatrix} = \mathbf{0},$$

and the complete data summary statistic $\mathbf{S}$ is given by

$$\mathbf{S} = \sum_i^K \mathbf{S}_i = \sum_i^K \begin{bmatrix} \partial \boldsymbol{\mu}_i^T(\boldsymbol{\beta}) / \partial \boldsymbol{\beta} \mathbf{V}_{i11}^{-1} \mathbf{u}_i [\boldsymbol{\mu}_i(\boldsymbol{\beta}) - \mathbf{y}_i] \\ \partial \boldsymbol{\pi}_i^T(\boldsymbol{\gamma}) / \partial \boldsymbol{\gamma} \mathbf{W}_{i11}^{-1} \mathbf{u}_i \\ \partial \boldsymbol{\sigma}_i^T(\boldsymbol{\rho}, \phi) / \partial \tilde{\boldsymbol{\rho}} \mathbf{H}_{i2} [\mathbf{s}_i - \boldsymbol{\sigma}_i(\boldsymbol{\rho}, \phi)] \\ \partial \boldsymbol{\tau}_i^T(\boldsymbol{\delta}) / \partial \boldsymbol{\delta} \mathbf{t}_i \end{bmatrix}.$$

Here, $\partial\boldsymbol{\mu}_i^T(\boldsymbol{\beta})/\partial\boldsymbol{\beta} \mathbf{V}_{i11}^{-1}\mathbf{H}_{i1}[\mathbf{y}_i - \boldsymbol{\mu}_i(\boldsymbol{\beta})]$ and $\partial\boldsymbol{\pi}_i^T(\boldsymbol{\gamma})/\partial\boldsymbol{\gamma} \mathbf{W}_{i11}^{-1}[\mathbf{u}_i - \boldsymbol{\pi}_i(\boldsymbol{\gamma})]$ are the estimating functions for $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$ the regression parameters, respectively. The estimating functions for the correlation parameters $\boldsymbol{\rho}$ and $\boldsymbol{\delta}$, and the dispersion parameter ϕ are given by $\partial\boldsymbol{\sigma}_i^T(\boldsymbol{\rho}, \phi)/\partial\tilde{\boldsymbol{\rho}} \mathbf{H}_{i2}[\mathbf{s}_i - \boldsymbol{\sigma}_i(\boldsymbol{\rho}, \phi)]$, and $\partial\boldsymbol{\tau}_i^T(\boldsymbol{\delta})/\partial\boldsymbol{\delta} [\mathbf{t}_i - \boldsymbol{\tau}_i(\boldsymbol{\delta})]$, respectively. The variance-covariance matrices are $\mathbf{V}_{i11}$ and $\mathbf{W}_{i11}$ for the non-degenerate and the zero components, respectively. The matrices $\mathbf{V}_{i11}$ and $\mathbf{W}_{i11}$ account for within-cluster association by plugging in the working correlation matrices $\mathbf{P}(\boldsymbol{\rho})$ and $\mathbf{R}(\boldsymbol{\delta})$, respectively, to the diagonal variance covariance matrices which imply independence. The terms $\mathbf{s}_i = \text{vech}[(\mathbf{y}_i - \boldsymbol{\mu}_i)(\mathbf{y}_i - \boldsymbol{\mu}_i)^T]$ and $\mathbf{t}_i = \text{vech}[(\mathbf{u}_i - \boldsymbol{\pi}_i)(\mathbf{u}_i - \boldsymbol{\pi}_i)^T]$ are for the estimation of second moments parameters with $\boldsymbol{\sigma}_i = E(\mathbf{s}_i) = \text{vech}(\mathbf{V}_{i11})$ and $\boldsymbol{\tau}_i = E(\mathbf{t}_i) = \text{vech}(\mathbf{W}_{i11})$, for the non-degenerate and zero component parts, respectively. The weight matrices $\mathbf{H}_{i1} = \text{diag}[(\mathbf{j}_{n_i} - \mathbf{u}_i)]$ and $\mathbf{H}_{i2} = \text{vech}[(\mathbf{j}_{n_i} - \mathbf{u}_i)(\mathbf{j}_{n_i} - \mathbf{u}_i)^T]$ are the weights for zero-inflation for the first and second moments, respectively. The function $\text{vech}(\cdot)$ is a vector-valued function that takes a square matrix as an input and returns a vector of the lower triangular elements including the diagonal.

For ZI regression models with a multinomial mixing structure, the complete data estimating equation is given by

$$\mathbf{U}_1^c(\boldsymbol{\beta}, \boldsymbol{\gamma}; \mathbf{z}) = \sum_i^K \begin{bmatrix} \partial\boldsymbol{\mu}_i^T(\boldsymbol{\beta})/\partial\boldsymbol{\beta} \mathbf{V}_{i11}^{-1}\mathbf{H}_{i1}[\mathbf{y}_i - \boldsymbol{\mu}_i(\boldsymbol{\beta})] \\ \mathbf{G}_i^T[\tilde{\boldsymbol{\omega}}_i - \boldsymbol{\omega}_i(\boldsymbol{\gamma})] \\ \partial\boldsymbol{\sigma}_i^T(\boldsymbol{\rho}, \phi)/\partial\tilde{\boldsymbol{\rho}} \mathbf{H}_{i2}[\mathbf{s}_i - \boldsymbol{\sigma}_i(\boldsymbol{\rho}, \phi)] \end{bmatrix} = \mathbf{0},$$

and the complete data summary statistic $\mathbf{S}$ is given by

$$\mathbf{S} = \sum_i^K \mathbf{S}_i = \sum_i^K \begin{bmatrix} \partial\boldsymbol{\mu}_i^T(\boldsymbol{\beta})/\partial\boldsymbol{\beta} \mathbf{V}_{i11}^{-1}\mathbf{u}_i[\boldsymbol{\mu}_i(\boldsymbol{\beta}) - \mathbf{y}_i] \\ \mathbf{G}_i^T \tilde{\boldsymbol{\omega}}_i \\ \partial\boldsymbol{\sigma}_i^T(\boldsymbol{\rho}, \phi)/\partial\tilde{\boldsymbol{\rho}} \mathbf{H}_{i2}[\mathbf{s}_i - \boldsymbol{\sigma}_i(\boldsymbol{\rho}, \phi)] \end{bmatrix}.$$

Here, $\partial \boldsymbol{\mu}_i^T(\boldsymbol{\beta})/\partial \boldsymbol{\beta} \mathbf{V}_{i11}^{-1} \mathbf{H}_{i1}[\mathbf{y}_i - \boldsymbol{\mu}_i(\boldsymbol{\beta})]$ is the estimating function for $\boldsymbol{\beta}$, $\mathbf{G}_i^T[\tilde{\boldsymbol{\omega}}_i - \boldsymbol{\omega}_i(\boldsymbol{\gamma})]$ is the score function for $\boldsymbol{\gamma}$ and $\partial \boldsymbol{\sigma}_i^T(\boldsymbol{\rho}, \phi)/\partial \tilde{\boldsymbol{\rho}} \mathbf{H}_{i2}[\mathbf{s}_i - \boldsymbol{\sigma}_i(\boldsymbol{\rho}, \phi)]$ is the estimating function for $\boldsymbol{\rho}$ and ϕ , the correlation and dispersion parameters, respectively.

The terms $\boldsymbol{\sigma}_i$, $\mathbf{V}_{i11}$, $\mathbf{H}_{i2}$, $\mathbf{H}_{i2}$ are defined as before. The vector $\mathbf{u}_i$ is for the vector of latent indicators of zero-inflation for cluster i . The matrix $\mathbf{G}_i$ is the design matrix for the multinomial mixing model with corresponding matrix of probabilities $\boldsymbol{\omega}_i$.

2.4 Extensions to larger and/or unequal cluster sizes

It is straight-forward to extend marginal ZI regression models to larger cluster sizes. However, the larger the cluster sizes imply an increase in the number of parameters to estimate. In addition, we often observe unequal cluster sizes because, for example, subjects may enter the study at different occasions or may be lost to follow-up. In the case of non-constant cluster sizes, it is necessary to modify the model framework to model functions of the multinomial probability vector that have consistent meaning across subjects with differing cluster sizes. Moreover, a different model framework can also be useful to reduce the dimension of the $\boldsymbol{\gamma}$ parameter vector. While there are various models for $\{\boldsymbol{\nu}_i\}$ that could achieve these goals and which might be appealing in particular applications, we focus on just three such approaches: (1) modeling the log-odds of zero inflation at each measurement occasion, (2) a model that assumes a constant probability vector for the maximal cluster size with probability vectors at smaller cluster sizes equal to the corresponding marginal probability vector obtained by summing the maximal one in a suitable manner, and (3) a proposal to model each subject's probability vector via loglinear models for the $2 \times 2 \times \cdots \times 2$ table corresponding to zero-inflation at time 1 $\times$ zero inflation at time 2, $\dots$, $\times$ zero inflation at time n_i . The above model frameworks are suitable or preferable in one application or another, but they all can be fit similarly with the ES-1 and ES-2 algorithms.

2.4.1 A multinomial model for zero-inflation at each measurement occasion

For each cluster i , let n_i , R_i , and $\boldsymbol{\nu}_i$ be the cluster size, the number of zero-inflation mixing patterns, and the multinomial probability vector for all possible patterns of zero-inflation, respectively. Let $\mathbf{A}_i$ denote the $R_i \times n_i$ matrix of indicators for all possible patterns of zero-inflation; that is, the r th row of $\mathbf{A}_i$ has ones in the j^{th} column if the r^{th} pattern of zero inflation involves an excess zero at time j and zero otherwise. Note that $\mathbf{A}_i^T \boldsymbol{\nu}_i$ is therefore an $n_i \times 1$ vector of marginal probabilities of zero-inflation at times $1, \dots, n_i$. For example, suppose that, as in Table 2.2, we have data from just two clusters in which we have $n_1 = 2$ and $n_2 = 3$ observations, respectively.

Table 2.2: Unbalanced clusters example with just two clusters

cluster i	n_i	R_i	$\boldsymbol{\nu}_i$	$\mathbf{A}_i$
1	$n_1 = 2$	$R_1 = 4$	$\boldsymbol{\nu}_1 = (\nu_{11}, \dots, \nu_{14})$	$\mathbf{A}_1$
2	$n_2 = 3$	$R_2 = 8$	$\boldsymbol{\nu}_2 = (\nu_{21}, \dots, \nu_{28})$	$\mathbf{A}_2$

Here, the matrices $\mathbf{A}_1$ and $\mathbf{A}_2$ are given by

$$\mathbf{A}_1 = \begin{pmatrix} 0 & 0 \\ 0 & 1 \\ 1 & 0 \\ 1 & 1 \end{pmatrix} \quad \text{and} \quad \mathbf{A}_2 = \begin{pmatrix} 0 & 0 & 0 \\ 0 & 0 & 1 \\ 0 & 1 & 0 \\ 0 & 1 & 1 \\ 1 & 0 & 0 \\ 1 & 0 & 1 \\ 1 & 1 & 0 \\ 1 & 1 & 1 \end{pmatrix}.$$

In the unbalanced case and/or when the cluster size is large, a suitable and more convenient model for the mixing process would be one that models the log odds of zero inflation at each time point. For example, we consider models of the form

$$\text{logit}\{(\mathbf{A}_i^T \boldsymbol{\nu}_i)_j\} = \mathbf{g}_{ij}^T \boldsymbol{\gamma}, \quad i = 1, \dots, K, \quad j = 1, \dots, n_i,$$

where, for example $\mathbf{g}_{ij}^T \boldsymbol{\gamma} = \gamma_j$ or $\mathbf{g}_{ij}^T \boldsymbol{\gamma} = \gamma_0 + \gamma_1 t_j$ for t_j the time of observation j .

2.4.2 A multinomial model for zero-inflation patterns based on the maximal cluster size

Alternatively, a second approach is to model each subject's probability vector $\boldsymbol{\nu}_i$ in terms of a multinomial model for zero-inflation pattern probabilities at the maximal cluster size. Such an approach is motivated as an extension of model (2.1) for non-constant n_i and is appropriate if we wish to assume a constant multinomial probability vector across subjects with the same dimensional response vector. Let $n^* = \max(n_i)$ denote the maximal cluster size and let $\boldsymbol{\zeta}$ denote the corresponding multinomial probability vector for the possible patterns of zero-inflation at that cluster size. That is, $\boldsymbol{\zeta} = \boldsymbol{\nu}_i$ for i , such that $n_i = n^*$. We assume that the probability vector for zero-inflation patterns is constant over all subjects of a given cluster size and that, for any cluster size less than n^* , such probability vectors are simply marginal probabilities determined by $\boldsymbol{\zeta}$. We model $\boldsymbol{\zeta}$ with a multinomial logit model of the form

$$\log \left(\frac{\zeta_r}{\zeta_{R^*}} \right) = \gamma_r, \quad r = 1, \dots, R^*,$$

where $R^* = 2^{n^*}$. Let $\mathcal{A}_{n^*-\ell} = \{i : n_i = n^* - \ell\}$, $\ell = 0, \dots, n^* - 1$. Then $\boldsymbol{\nu}_i = \mathbf{A}_{n^*-\ell}\boldsymbol{\zeta}$, for $i \in \mathcal{A}_{n^*-\ell}$, $\ell = 0, \dots, n^* - 1$, where $\mathbf{A}_{n^*} = \mathbf{I}_{R^*}$ and

$$\mathbf{A}_{n^*-\ell} = \prod_{l=1}^{\ell} (\mathbf{I}_{R^*/2^l} \otimes \mathbf{j}_2^T) \mathbf{I}_{R^*}, \quad \ell = 1, \dots, n^* - 1.$$

That is, the model takes the form

$$\begin{aligned} \log \left(\frac{\nu_{ir}}{\nu_{iR^*}} \right) &= \gamma_r, \quad r = 1, \dots, R^*, \quad \text{for } i \in \mathcal{A}_{n^*} \\ \boldsymbol{\nu}_i &= \mathbf{A}_{n^*-1}\boldsymbol{\zeta}, \quad \text{for } i \in \mathcal{A}_{n^*-1} \\ &\vdots \\ \boldsymbol{\nu}_i &= \mathbf{A}_1\boldsymbol{\zeta}, \quad \text{for } i \in \mathcal{A}_1. \end{aligned}$$

To illustrate, consider again the example of Table 2.2 in which we have two clusters with $n_1 = 2$ and $n_2 = 3$. In this case $n^* = n_2 = 3$, $R^* = 2^3 = 8$, $\mathbf{A}_{n^*} = \mathbf{A}_3 = \mathbf{I}_8$, and $\mathbf{A}_{n^*-1} = \mathbf{A}_2 = (\mathbf{I}_4 \otimes \mathbf{j}_2^T) \mathbf{I}_8$. The probability vector for the $R^* = 8$ possible mixing patterns when $n_i = 3$ is $\boldsymbol{\zeta}$ which we model as

$$\log \left(\frac{\zeta_r}{\zeta_7} \right) = \gamma_r, \quad r = 1, \dots, 7.$$

Finally,

$$\begin{aligned}\boldsymbol{\nu}_1 &= \mathbf{A}_2 \boldsymbol{\zeta} = (\mathbf{I}_4 \otimes \mathbf{j}_2^T) \mathbf{I}_8 \boldsymbol{\zeta} = \begin{pmatrix} \zeta_1 + \zeta_2 \\ \zeta_3 + \zeta_4 \\ \zeta_5 + \zeta_6 \\ \zeta_7 + \zeta_8 \end{pmatrix} \\ \boldsymbol{\nu}_2 &= \mathbf{A}_3 \boldsymbol{\zeta} = \mathbf{I}_8 \boldsymbol{\zeta} = \boldsymbol{\zeta}.\end{aligned}$$

Note that the model can be thought of as based on a saturated multinomial model for a 2^{n^*} contingency table for all $i \in \mathcal{A}_{n^*}$, where the probability vectors for clusters with smaller cluster size are just those from the corresponding marginal tables. For example, in the example, the model for subject 2 can be thought of as a saturated multinomial model for a $2 \times 2 \times 2$ table (zero-inflation at time 1 $\times$ zero-inflation at time 2 $\times$ zero-inflation at time 3), and the probability vector for subject 1 is just the marginal probability vector from the corresponding table of zero inflation at time 1 $\times$ zero inflation at time 2. Here, we have assumed that unequal cluster sizes arise from monotone missing patterns consistent with attrition. That is, the model is suitable assuming a subject with cluster size n_i has measurements at the first n_i of n^* max possible measurement occasions. The model could be adapted easily for non-monotone patterns of missingness, however, by defining the $\mathbf{A}_{n_i}$ matrices suitably to form the marginal probability vectors correctly for any given pattern of missing measurement occasions.

2.4.3 Multinomial models via log-linear models

For Poisson, multinomial and product-multinomial sampling, log-linear models are a natural framework (Agresti [1990]). Log-linear models are flexible and more general than logit models for the analysis of count categorical data. Some log-linear models have a direct cor-

response to logit models. Interaction terms in log-linear models are used to model the associations in the effects under study. A log-linear model is equivalent to a logit model if all the possible associations among the predictors are included. Log-linear models are flexible in specifying joint independence, conditional independence and homogeneous association models.

The multinomial mixing structure introduced in section 2.3, will result in a multinomial logit model for the S-step that has a growing number of parameters with increasing cluster sizes. Log-linear models may be useful in this situation to flexibly control the model complexity compared to a corresponding multinomial logit model. This corresponds to fitting a log-linear model to the $2 \times 2 \times \cdots \times 2$ table corresponding to zero-inflation at time $j = 1 \times \cdots \times$ zero-inflation at time $j = n_i$. For the case with $t = 2$ measurement occasions, the saturated log-linear model is given by:

$$\log(\tilde{\omega}_{ijk}) = \lambda + \lambda_i^S + \lambda_j^{T_1} + \lambda_k^{T_2} + \lambda_{ij}^{ST_1} + \lambda_{ik}^{ST_2} + \lambda_{jk}^{T_1T_2} + \lambda_{ijk}^{ST_1T_2},$$

where the λ 's represent the main effects for subject i (λ_i^S), measurement occasions T_1 ($\lambda_j^{T_1}$), and T_2 ($\lambda_k^{T_2}$), respectively, two-way and three-way interaction terms for $i = 1, \dots, K$ clusters and $j, k = 1, 2$ measurement occasions. A more parsimonious model will assume the conditional independence model of the form:

$$\log(\tilde{\omega}_{ijk}) = \lambda + \lambda_i^S + \lambda_j^{T_1} + \lambda_k^{T_2} + \lambda_{ij}^{ST_1} + \lambda_{ik}^{ST_2} + \lambda_{jk}^{T_1T_2}.$$

If the subject effects are linear in the covariates $\mathbf{x}_i$, then we can allow the arbitrary subject effects to depend upon covariates to get an even more parsimonious model of the form:

$$\log(\tilde{\omega}_{ijk}) = \alpha + \lambda_i^S + \lambda_j^{T_1} + \lambda_k^{T_2} + \mathbf{x}_i^T \theta_{1j} + \mathbf{x}_i^T \Theta_{1k} + \lambda_{jk}^{T_1T_2},$$

where $\alpha = \lambda + \theta_0 + \Theta_0$. The subject effects λ_i^S remain in the model to ensure normalization and correspondence to a multinomial logit model. The above concepts can be extended to any number of time points.

Finally, note that regardless of the specific model chosen to accommodate large and/or unequal cluster sizes, the model fitting details remain the same as in the balanced case.

2.5 Simulations

In this section we shall investigate and compare the performance of the parameter estimators computed from marginal ZI regression models under different assumptions and conditions. We consider marginal ZI regression models with both a binary mixing (BMX) and a multinomial mixing (MMX) structure. The ES-1 algorithm is used to fit the models to the simulated data. Correlated Poisson counts, $\mathbf{z}_i$, were generated using a multivariate Gaussian copula from an R package called `simstudy` package (Goldfeld [2019]). A multinomial random variable, $\boldsymbol{\omega}_i$, for the mixing patterns of zero inflation is generated from the multinomial distribution. For each subject we generated a three dimensional vector $\mathbf{z}_i$ which follows a correlated Poisson($\boldsymbol{\mu}_i$) where $\mu_{ij} = \exp(\beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2})$ for $i = 1, 2, \dots, K$ and $j = 1, 2, 3$. The covariates X_{i1} and X_{i2} were generated as standard normal random variables. We fixed the regression parameters for the Poisson mean to be $\boldsymbol{\beta} = (2.12, 0.43, -0.49)^T$. Covariates can be added via a multinomial model for the mixing patterns of zero-inflation, but for simplicity, we consider a multinomial logit model of the form $\log(\nu_r/\nu_R) = \gamma_r, r = 1, 2, \dots, R - 1$ and fix the intercept to $\boldsymbol{\gamma} = (1, 1, 1, 1, 1, 1)^T$. This implies a constant zero-inflation probability of $\pi = 0.457103$ at each time point. Excess zeros are introduced into the correlated Poisson counts according to the mixing pattern determined by the simulated classes from the multinomial distribution.

That is, we fix the nonzero counts and force zeros on counts which the mixing classes presume an excess zero. For example, in the case with cluster size $n = 2$, we generate the counts $\mathbf{z}_i = (z_{i1}, z_{i2})$ and a mixing pattern of, say $\omega_i = 2$ from a multinomial model which has a mixing pattern $(0, 1)$, this then implies that at time point 2 we have an excess zero, i.e., $\mathbf{y}_i = (z_{i1}, 0)$.

One thousand (1000) data sets each with clusters of size three are randomly generated for each of the possible combinations of varying sample sizes ($K = 50$ and $K = 200$), varying working correlation structures (Exchangeable and AR(1)), and two levels of correlation strengths ($\rho = 0.3$ and $\rho = 0.8$).

The simulation results show tables for each combination of scenarios described above. For each simulation scenario, results are presented showing the asymptotic properties of the regression estimators and their respective asymptotic coverage and test size. The columns labeled $\overline{\text{ASE}}$ are averaged asymptotic standard errors, the columns labeled “Mean” and “SD” are averaged Monte Carlo results for the estimator’s mean and standard deviation, respectively, and the columns labeled “MSE” are the corresponding mean-squared errors for the estimators. We present results for the regression parameters of the non-degenerate component, β and ρ for both approaches. The regression parameters for the binary and the multinomial mixing models have different interpretations and therefore not comparable. For this reason, we present the estimated zero-inflation probabilities at each time point, the $\hat{\pi}_{ij}$ ’s. We do not provide measures for the averaged asymptotic SE ($\overline{\text{ASE}}$) for π_{ij} because they are not of interest in this study, however, such a result can be obtained using the *delta method*. The results are presented below.

The ratios of mean squared errors with the BMX model on the divisor, are given in Table 2.3. It appears that there are no strong gains in efficiency between the ES-1 algorithm fitted assuming multinomial mixing and the ES-1 algorithm fitted assuming binary mixing. Table 2.4 shows the properties of the estimators considered at the correlation strength of $\rho = 0.3$.

As expected the results improve with increasing number of clusters. The results for the multinomial mixing and the binary mixing structures do not indicate any strong differences in the properties of the estimators considered. The Monte Carlo standard deviations (SD) are quite close to the averaged asymptotic standard errors ($\overline{ASE}$) from the fitted models and hence the sandwich estimator is a reasonable estimator of the variance-covariance matrix.

Table 2.5 shows the test size and the coverage for both methods at the correlation strength of $\rho = 0.3$. The results show that the nominal test size is sometimes larger than expected and lower coverage rates than expected, especially for small clusters. This is a result of the $\overline{ASE}$'s which are slightly smaller than the Monte Carlo standard deviations (SD) from Table 2.4, which leads to tighter confidence bounds. The results show slightly better test and coverage for the multinomial mixing structure versus the binary mixing structure in comparing the correlation parameter.

Table 2.7 summarizes the test size and coverage properties at correlation strength of $\rho = 0.8$. A similar phenomenon observed in Table 2.5 is apparent in this case as well. The results improve with increasing number of clusters and multinomial mixing models often show slightly better results compared to binary mixing models for the correlation parameter.

Table 2.6 shows the results for both methods computed at a correlation strength of $\rho = 0.8$. The results for both fitting algorithms show an improvement in statistical properties of the estimators as the number of clusters increases. There appears to be no strong differences between the properties of the estimators in both approaches. The averaged asymptotic standard errors ($\overline{ASE}$) from the sandwich estimator are comparably close to the averaged Monte Carlo standard deviations (SD) which indicates that the sandwich estimator is reasonable. Table 2.8 summarizes the effects of fitting an incorrectly specified working correlation matrix for the multinomial mixing and the binary mixing models. The data generating process is similar to the one described above except that we fixed the correlation strength to

be $\rho = 0.8$, the number of clusters is taken to be $K = 100$ and 500 data sets are used to compare the estimators. Both approaches show that we get better estimates if the working correlation structure is correctly specified. This is particularly observable in the correlation parameter $\rho = 0.8$.

2.5.1 Other simulations

The number of parameters to estimate in the multinomial mixing framework increases with an increase in the number of time points. We can impose constraints or use other modeling frameworks where we can control the number of parameters, for example, fitting a log-linear model as an equivalent to the multinomial logit model for the mixing data. We can then control model complexity by imposing some structure or constraints on the model parameters. We investigated the case where the number of parameters for the multinomial mixing and the binary mixing structures are equivalent. The simulation setup is similar to the one described above except that we use clusters of size two and add an additional working correlation parameter in the binary mixing structure of the mixing data. We consider correlation strengths of $\rho = 0.25$ and $\rho = 0.75$. Most of the results are in Appendix A and show that there are no strikingly different results between the two approaches and the results are consistent with what we have seen in the simulations above. Table 2.9 summarizes the ratios of MSEs with the BMX MSEs on the divisor and Table 2.10 summarizes the comparison in computational times for both approaches.

Ratios of mean squared errors from the simulation study are given in Table 2.9. It appears that there are no clear gains in efficiency between the ES-1 assuming the multinomial mixing structure and the ES-1 assuming a binary mixing structure.

Table 2.10 shows that the multinomial mixing models perform computationally faster on average than binary mixing models when the number of parameters to be estimated are the same. It appears that computation times tend to decrease on average with an increase in correlation strength as well as the standard deviations of the computation time.

2.6 Example

In this section, we analyze a data set on the number of hospital visits per quarter during a one year period for $K = 73$ children. These data were considered by Karim and Zeger [1988] to investigate the number of hospital visits, the response variable. The potential predictors are the sex of the child, the maternal smoking status of the mother, and the baseline age of the child in months, and the effects of time (quarter). The data has complete cases from each child and altogether there are $73 \times 4 = 292$ observations.

The results in Table 2.11 are obtained using an *R* package called the **performance** package (Ludecke [2019]). The *check_zeroinflation* function in the package, allows the user to pass a model object specified without assuming zero-inflation, for example, a Poisson GLM, and outputs the predicted and observed zeros together with a message for whether there is probable zero-inflation. Table 2.11 shows the results from such an implementation by supplying a Poisson GLM object that does not assume zero-inflation at each quarter and overall. The results indicate that the models underestimate the observed zeros in the data because all the ratios in the table are far off from one. This suggests probable zero-inflation in the data. In this example we consider a marginal Poisson ZI regression model with multi-way interactions for the model effects.

We assume an exchangeable working correlation matrix for the Poisson component. The working correlation matrix is assumed to be exchangeable for the excess zero component of the binary mixing model. The models are fitted using both the ES-1 and the ES-2 algorithms for binary mixing and multinomial mixing models. The results are given in Table 2.12.

The results of the ES-2 algorithm fit are omitted from Table 2.12 because they are similar to the results using the ES-1 algorithm. The ES-2 algorithm (11.78s) computed unexpectedly longer compared to the ES-1 algorithm (2.05s) for the binary mixing model. However, the gains in speed are noticeable in the multinomial mixing model with the ES-1 taking 135.19s compared to 24.41s for the ES-2 algorithm. Multinomial mixing models, in general, would take longer to fit than binary mixing models because the dimension of the γ parameter in such models is larger. The results for the parameter estimates are slightly different between the multinomial mixing model and the binary mixing model, but the parameter estimates are consistent in terms of magnitude and sign.

The results in Table 2.12 show the parameter estimates for the non-degenerate Poisson component. The results show the group effects for gender and smoking status, together with three-way and two-way interactions involving the effect of time on a continuous scale. Results also show the estimated zero-inflation probabilities at each time point. Estimates of the association parameters for the exchangeable working correlation matrices for both components where applicable are also presented.

The interaction term is significant which allows us to infer how the effects of time relate to the combinations of levels for gender and smoking. We can let the zero state represent children who are healthy and do not have any chance of visiting the hospital and the non-zero state represent children with health complication who may visit the hospital. Using the model *MMX*, the exponent of the intercept (1.92168) in the Poisson component represent the incidence ratio at quarter 1 for males whose mothers smoke and are prone to health complications that may require a visit to the hospital. Table 2.13 gives the rate ratios per

unit increase in time/quarter in each of the combination of levels for gender and smoking. Among children with health complications that may require a visit to the hospital, there is a 13.06% decrease in the hospital visitation rate per each quarter for male children whose mothers smoke (rate ratio = 0.8694), a 21.41% decrease per each quarter for females whose mothers are smokers (rate ratio = 0.7859). Similarly, we can infer that among children with health complications leading to possible hospitalization, the rates for hospital visits for children whose mothers are non-smokers decreases over time for males (rate ratio = 0.6991), that is a 30.09% decrease per each quarter, but increases for females (rate ratio = 1.62), that is a 62% increase per each quarter.

2.7 Summary

In this chapter we proposed marginal ZI regression models for clustered data with excess zeros assuming a multinomial mixing structure. The multinomial mixing structure ensures the model accounts for the dependence in the mixing data at the E-step. This differs from Hall and Zhang [2004] who assumed a binary mixing structure by assuming independence wholly at the E-step. The ES-1 and ES-2 algorithms were used to fit multinomial mixing models and binary mixing models for clustered data with excess zeros.

Simulations show that there are no major differences in the properties of the parameter estimators for binary mixing models and multinomial mixing models. This is expected since the ES algorithm simplifies the E-step by assuming independence in the response data. Future research should focus on developing EM-type algorithms that can effectively account for the dependence in the data. Marginal ZI regression models assuming a multinomial mixing structure showed to be computationally faster on average than marginal ZI regression models assuming a binary mixing structure.

We implemented marginal ZI regression models assuming both multinomial and binary mixing structures to the hospital visits data set. We used both the ES-1 and the ES-2 algorithms to fit the models to data. The ES-1 and ES-2 algorithms had roughly equivalent parameter estimates for both the binary mixing and the multinomial mixing models. The computational time was surprisingly higher for the ES-2 algorithm compared to the ES-1 algorithm for the binary mixing model, however, the ES-2 performed as expected for the multinomial mixing model. The ES-2 algorithm on the other hand is convenient in estimating the variance-covariance matrix of the parameter estimates.

The parameters of the degenerate component in marginal ZI regression models with multinomial mixing structure grow with increasing number of time points or measurement occasions in the data. Future research work should concentrate on imposing constraints on the parameter estimates or to use alternative log-linear models that constrain the model parameters more flexibly than multinomial logit models. With large clusters, there are chances that we might not observe all the possible mixing patterns in the data for multinomial mixing models. The number of mixing classes can be restricted to only those that are observed in the data. However, this alters the number of classes and consequently the interpretation of model parameters in the degenerate component. The model for the non-degenerate component and the fitted zero-inflation probabilities are not altered. When that occurs, it is recommended to focus on interpreting the regression effects of the non-degenerate component and the fitted zero-inflation probabilities rather than the effects of the zero component to avoid misleading conclusions. Multinomial mixing models are useful if the primary interest is on the regression effects in the non-degenerate component of the model.

Table 2.3: A simulation study comparing the ratios of MSEs for the Binary Mixing and the Multinomial Mixing structures

Correlation Structure	MSE Ratios $K = 50$		MSE Ratios $K = 200$	
	$\rho = 0.3$	$\rho = 0.8$	$\rho = 0.3$	$\rho = 0.8$
Exchangeable				
β_0	1.0000	0.9999	1.0000	0.9999
β_1	1.0000	0.9999	1.0000	0.9999
β_2	1.0000	1.0005	1.0000	1.0002
ρ	1.0073	1.0073	1.0085	1.0000
π	1.0000	1.0000	1.0000	1.0000
AR(1)				
β_0	1.0000	1.0000	0.9999	1.0000
β_1	1.0000	1.0000	1.0000	1.0000
β_2	1.0001	1.0000	1.0000	1.0000
ρ	1.0011	1.0064	1.0072	0.9919
π	1.0000	1.0000	1.0000	1.0000

¹. MSE ratios with *Binary mixing* on the divisor.

Table 2.4: A simulation study for comparing the Binary Mixing and the Multinomial Mixing structures (Correlation parameter $\rho = 0.3$)

Correlation Structure	Mean	Binary Mixing			Mean	Multinomial Mixing		
		ASE	SD	MSE		ASE	SD	MSE
<i>Exchangeable: K = 50</i>								
β_0	2.1136	0.0462	0.0489	0.0024	2.1136	0.0462	0.0489	0.0024
β_1	0.4298	0.0391	0.0433	0.0019	0.4298	0.0390	0.0433	0.0019
β_2	−0.4906	0.0385	0.0434	0.0019	−0.4906	0.0385	0.0434	0.0019
ρ	0.3245	0.2219	0.2612	0.0688	0.3277	0.2233	0.2618	0.0693
π	0.4559		0.0363	0.0013	0.4558		0.0364	0.0013
<i>AR(1): K = 50</i>								
β_0	2.1168	0.0459	0.0471	0.0021	2.1168	0.0459	0.0472	0.0021
β_1	0.4314	0.0381	0.0411	0.0017	0.4314	0.0381	0.0412	0.0017
β_2	−0.4910	0.0385	0.0414	0.0017	−0.4911	0.0385	0.0413	0.0017
ρ	0.3219	0.2619	0.3063	0.0943	0.3253	0.2577	0.3062	0.0944
π	0.4576		0.0397	0.0016	0.4576		0.0397	0.0016
<i>Exchangeable: K = 200</i>								
β_0	2.1187	0.0235	0.0243	0.0006	2.1187	0.0235	0.0243	0.0006
β_1	0.4311	0.0197	0.0200	0.0004	0.4311	0.0197	0.0200	0.0004
β_2	−0.4904	0.0195	0.0197	0.0004	−0.4904	0.0195	0.0197	0.0004
ρ	0.3052	0.1057	0.1084	0.0118	0.3078	0.1061	0.1087	0.0119
π	0.4574		0.0180	0.0003	0.4574		0.0180	0.0003
<i>AR(1): K = 200</i>								
β_0	2.1187	0.0231	0.0242	0.0006	2.1186	0.0231	0.0242	0.0006
β_1	0.4305	0.0190	0.0202	0.0004	0.4305	0.019	0.0201	0.0004
β_2	−0.4908	0.0192	0.0201	0.0004	−0.4908	0.0192	0.0201	0.0004
ρ	0.3032	0.1132	0.1175	0.0138	0.3058	0.1136	0.1177	0.0139
π	0.4570		0.0182	0.0003	0.4570		0.0182	0.0003

Table 2.5: A simulation study for comparing the Binary Mixing and the Multinomial Mixing structures (Test size (5%) and Coverage (95%): Correlation parameter $\rho = 0.3$)

Correlation Structure	Binary Mixing		Multinomial Mixing	
	Test Size	Coverage	Test Size	Coverage
<i>Exchangeable: $K = 50$</i>				
β_0	0.0670	0.9330	0.0660	0.9340
β_1	0.0840	0.9160	0.0840	0.9160
β_2	0.1020	0.8980	0.0970	0.9030
ρ	0.1510	0.8490	0.1470	0.8530
<i>AR(1): $K = 50$</i>				
β_0	0.0560	0.9440	0.0560	0.9440
β_1	0.0770	0.9230	0.0810	0.9190
β_2	0.0760	0.9240	0.0770	0.9230
ρ	0.1390	0.8610	0.1340	0.8660
<i>Exchangeable: $K = 200$</i>				
β_0	0.0580	0.9420	0.0580	0.9420
β_1	0.0550	0.9450	0.0550	0.9450
β_2	0.0580	0.9420	0.0580	0.9420
ρ	0.0840	0.9160	0.0810	0.9190
<i>AR(1): $K = 200$</i>				
β_0	0.0650	0.9350	0.0640	0.9360
β_1	0.0710	0.9290	0.0720	0.9280
β_2	0.0570	0.9430	0.0600	0.9400
ρ	0.0690	0.9310	0.0670	0.9330

Table 2.6: A simulation study for comparing the Binary Mixing and the Multinomial Mixing structures (Correlation parameter $\rho = 0.8$)

Correlation Structure	Mean	Binary Mixing			Mean	Multinomial Mixing		
		ASE	SD	MSE		ASE	SD	MSE
<i>Exchangeable: K = 50</i>								
β_0	2.1150	0.0539	0.0575	0.0033	2.1149	0.0539	0.0575	0.0033
β_1	0.4282	0.0448	0.0538	0.0029	0.4283	0.0448	0.0538	0.0029
β_2	−0.4910	0.0449	0.0507	0.0026	−0.4912	0.0449	0.0507	0.0026
ρ	0.8217	0.1939	0.2017	0.0412	0.8250	0.1948	0.2022	0.0415
π	0.4566		0.0362	0.0013	0.4564		0.0361	0.0013
<i>AR(1): K = 50</i>								
β_0	2.1165	0.0551	0.0584	0.0034	2.1164	0.0552	0.0585	0.0034
β_1	0.4295	0.0462	0.0529	0.0028	0.4295	0.0462	0.0530	0.0028
β_2	−0.4908	0.0460	0.0519	0.0027	−0.4909	0.0461	0.0519	0.0027
ρ	0.8271	0.2024	0.2145	0.0467	0.8304	0.2032	0.2147	0.0470
π	0.4557		0.0370	0.0014	0.4556		0.0370	0.0014
<i>Exchangeable: K = 200</i>								
β_0	2.1178	0.0277	0.0273	0.0008	2.1177	0.0277	0.0273	0.0008
β_1	0.4302	0.0228	0.0239	0.0006	0.4303	0.0228	0.0239	0.0006
β_2	−0.4910	0.0227	0.0236	0.0006	−0.4911	0.0227	0.0236	0.0006
ρ	0.7957	0.1041	0.1062	0.0113	0.7985	0.1043	0.1063	0.0113
π	0.4562		0.0184	0.0003	0.4561		0.0183	0.0003
<i>AR(1): K = 200</i>								
β_0	2.1206	0.0281	0.0275	0.0008	2.1205	0.0281	0.0275	0.0008
β_1	0.4305	0.0233	0.0246	0.0006	0.4306	0.0233	0.0246	0.0006
β_2	−0.4891	0.0232	0.0244	0.0006	−0.4892	0.0232	0.0244	0.0006
ρ	0.7957	0.1099	0.1111	0.0124	0.7987	0.1102	0.1111	0.0123
π	0.4568		0.0188	0.0004	0.4568		0.0188	0.0004

Table 2.7: A simulation study for comparing the Binary Mixing and the Multinomial Mixing Structures (Test size (5%) and Coverage(95%): Correlation parameter $\rho = 0.8$)

Correlation Structure	Binary Mixing		Multinomial Mixing	
	Test Size	Coverage	Test Size	Coverage
<i>Exchangeable: $K = 50$</i>				
β_0	0.0640	0.9360	0.0660	0.9340
β_1	0.1120	0.8880	0.1130	0.8870
β_2	0.1100	0.8900	0.1090	0.8910
ρ	0.0940	0.9060	0.0930	0.9070
<i>AR(1): $K = 50$</i>				
β_0	0.0840	0.9160	0.0830	0.9170
β_1	0.1020	0.8980	0.1010	0.8990
β_2	0.1010	0.8990	0.0980	0.9020
ρ	0.1250	0.8750	0.1230	0.8770
<i>Exchangeable: $K = 200$</i>				
β_0	0.0500	0.9500	0.0510	0.9490
β_1	0.0590	0.9410	0.0610	0.9390
β_2	0.0590	0.9410	0.0590	0.9410
ρ	0.0740	0.9260	0.0730	0.9270
<i>AR(1): $K = 200$</i>				
β_0	0.0540	0.9460	0.0530	0.9470
β_1	0.0650	0.9350	0.0640	0.9360
β_2	0.0700	0.9300	0.0690	0.9310
ρ	0.0800	0.9200	0.0760	0.9240

Table 2.8: A simulation study for comparing the Binary Mixing and the Multinomial Mixing structures (Mis-specified Working Correlation with $\rho = 0.8$ and $K = 100$)

Correlation Structure	Binary Mixing				Multinomial Mixing			
	Mean	ASE	SD	MSE	Mean	ASE	SD	MSE
<i>Exchangeable fit Exchangeable</i>								
β_0	2.1180	0.0389	0.0396	0.0016	2.1178	0.0389	0.0396	0.0016
β_1	0.4297	0.0322	0.0361	0.0013	0.4298	0.0322	0.0360	0.0013
β_2	−0.4888	0.0328	0.0330	0.0011	−0.4889	0.0328	0.0331	0.0011
ρ	0.8078	0.1423	0.1455	0.0212	0.8109	0.1427	0.1455	0.0213
π	0.4561		0.0256	0.0007	0.4560		0.0256	0.0007
<i>Exchangeable fit AR(1)</i>								
β_0	2.1170	0.0407	0.0402	0.0016	2.1168	0.0407	0.0401	0.0016
β_1	0.4305	0.0331	0.0348	0.0012	0.4306	0.0331	0.0348	0.0012
β_2	−0.4918	0.0337	0.0333	0.0011	−0.4919	0.0337	0.0333	0.0011
ρ	0.7932	0.1545	0.1576		0.7964	0.1548	0.1582	
π	0.4571		0.0273	0.0007	0.4570		0.0274	0.0008
<i>AR(1) fit AR(1)</i>								
β_0	2.1195	0.0386	0.0383	0.0015	2.1193	0.0386	0.0382	0.0015
β_1	0.4310	0.0320	0.0331	0.0011	0.4311	0.0320	0.0331	0.0011
β_2	−0.4891	0.0314	0.0327	0.0011	−0.4893	0.0313	0.0328	0.0011
ρ	0.7987	0.1426	0.1496	0.0247	0.8018	0.1428	0.1497	0.0245
π	0.4571		0.0264	0.0007	0.4569		0.0264	0.0007
<i>AR(1) fit Exchangeable</i>								
β_0	2.1180	0.0396	0.0421	0.0018	2.1178	0.0397	0.0421	0.0018
β_1	0.4294	0.0326	0.0358	0.0013	0.4296	0.0327	0.0358	0.0013
β_2	−0.4892	0.0328	0.0346	0.0011	−0.4894	0.0328	0.0347	0.0012
ρ	0.7514	0.1497	0.1642		0.7541	0.1500	0.1646	
π	0.4567		0.0262	0.0007	0.4565		0.0262	0.0007

Table 2.9: A simulation study for comparing the Binary Mixing and the Multinomial Mixing structures

Correlation Structure	Ratios $K = 50$		Ratios $K = 200$	
	$\rho = 0.25$	$\rho = 0.75$	$\rho = 0.25$	$\rho = 0.75$
<i>Exchangeable</i>				
β_0	1.0044	0.9999	0.9983	0.9999
β_1	1.0016	0.9999	0.9987	0.9999
β_2	1.0012	1.0001	0.9998	1.0001
ρ	0.9946	1.0161	1.0140	1.0032
π_1	0.9996	1.0008	1.0008	1.0008
π_2	1.0013	0.9996	0.9998	0.9994
<i>AR(1)</i>				
β_0	0.9999	0.9980	0.9999	0.9983
β_1	1.0002	0.9975	1.0001	0.9977
β_2	1.0001	1.0006	1.0000	0.9980
ρ	1.0069	1.0226	1.0163	1.0053
π_1	0.9985	0.9982	1.0025	1.0013
π_2	0.9987	0.9995	0.9989	0.9983

¹. MSE ratios with *Binary mixing* divisor.

Table 2.10: A simulation study for comparing the computation times for the Binary Mixing and the Multinomial Mixing structures

Sample size	Correlation Structure	Correlation level	Avg. Time: BMX	SD. Time: BMX	Avg. Time: MMX	SD. Time: MMX
$K = 50$	Exchangeable	$\rho = 0.25$	5.4565	4.3627	4.6574	6.0546
	AR(1)		4.9475	3.7643	4.1951	6.6613
	Exchangeable	$\rho = 0.75$	3.8301	2.8067	2.8679	2.9055
	AR(1)		3.6296	2.1220	2.5929	1.8281
$K = 200$	Exchangeable	$\rho = 0.25$	16.7680	11.8907	11.4932	10.6999
	AR(1)		15.6380	10.6255	10.5402	9.4970
	Exchangeable	$\rho = 0.75$	12.3298	2.7428	7.4553	2.0480
	AR(1)		12.4367	10.3198	7.3299	1.9227

Table 2.11: Investigating zero-inflation in the data

Time	Observed	Predicted	Ratio
Overall	154	114	0.74
Quarter 1	25	18	0.72
Quarter 2	39	30	0.77
Quarter 3	38	27	0.71
Quarter 4	52	46	0.88

¹. Ratio = Predicted/Observed.

Table 2.12: Results of the hospital visits data using the ES-1 algorithms for multinomial and binary mixing models

Coefficient	BMX(ES-1)				MMX (ES-1)			
	Estimate	SE	Z-value	P-value	Estimate	SE	Z-value	P-value
<i>Poisson component:</i>								
(Intercept)	0.5785	0.2107	2.7457	0.0060	0.6532	0.1835	3.5592	0.0004
SexF	0.1425	0.3062	0.4656	0.6415	0.0977	0.2857	0.3420	0.7323
SmokeNo	-0.0359	0.3194	-0.1125	0.9104	-0.0222	0.3039	-0.0730	0.9418
time	-0.1172	0.1103	-1.0622	0.2882	-0.1400	0.0985	-1.4211	0.1553
SexF:SmokeNo	-0.3691	0.4534	-0.8140	0.4157	-0.4617	0.4554	-1.0137	0.3107
SexF:time	-0.1025	0.1504	-0.6812	0.4957	-0.1009	0.1454	-0.6941	0.4876
SmokeNo:time	-0.1761	0.1790	-0.9837	0.3253	-0.2180	0.1585	-1.3755	0.1690
SexF:SmokeNo:time	0.8518	0.2321	3.6696	0.0002	0.9413	0.2219	4.2428	0.0000
count correlation	0.1183	0.0790		0.0657	0.0658			
<i>ZI probabilities:</i>								
π_1	0.2083				0.2500			
π_2	0.4108				0.4250			
π_3	0.3623				0.3852			
π_4	0.5846				0.5623			
zero correlation	0.4141	0.1121						

Table 2.13: Effects of time on hospital visits rates

Smoke	Sex	
	Male	Female
Yes	0.8694 ($n = 39$)	0.7859 ($n = 14$)
No	0.6991 ($n = 14$)	1.6200 ($n = 6$)

3. A Regression-Solution algorithm for marginal zero-inflated regression models with multinomial mixing

3.1 Introduction

Marginal zero-inflated (ZI) regression models are useful in many applications that involve clustered data with excess zeros. ZI models are two component mixtures of a degenerate distribution with a point mass of one at zero and a standard distribution. In some applications, the data are observed as bounded or unbounded counts with excess zeros and measured repeatedly over time for the response variable. The expectation-maximization (EM) algorithm, if the likelihood is available and tractable, is a convenient way of computing the model parameter estimates because of the mixture form of ZI models. In some correlated data problems, however, it is difficult to construct a tractable likelihood function and the EM algorithm will be challenging to implement. Marginal ZI regression models are partially specified models and the likelihood function is generally intractable. There is need to develop more general EM-type algorithms that account for the dependence in the data.

The expectation-solution (ES) algorithm is an EM-type algorithm that was proposed by Rosen et al. [2000] for mixtures of experts in generalized linear models. Hall and Zhang [2004] applied the ES algorithm to marginal ZI regression models for clustered data with excess zeros. The E-step is computed from the conditional expectation of the complete data quasi-loglikelihood given the observed data and current parameter estimates under independence. The S-step relaxes the independence assumption and updates parameter estimates by solving a set of GEEs with working correlation matrices to account for the dependence

in the data. The E-step ignores within-cluster dependence in the data and it is well known in clustered data literature that estimation methods that do not account for the correlation in the data are inefficient and associated methods of inference are invalid. Our goal is to develop EM-type algorithms that account for the dependence in the data for both imputing the missing data and updating the parameter estimates.

Benesi and Hall [2020a] introduced marginal ZI regression models for clustered data with excess zeros assuming a multinomial mixing (MMX) structure. The multinomial mixing models and the binary mixing models (Hall and Zhang [2004]) use the ES algorithm for fitting the models. In both model frameworks, the E-step assumes independence to get a closed form solution. Multinomial mixing models, however, differ from binary mixing models because the solution step for the degenerate component is performed by fitting a fully parametric multinomial model instead of an *ad hoc* GEE with a working correlation matrix that may not be suitable for correlated binary data.

However, independence is assumed for the non-degenerate data when computing the E-step for multinomial mixing models. The E-step imputes the missing data by computing the conditional expectation of the complete data quasi-loglikelihood function given the observed data and current parameter estimates. The independence assumption is relaxed at the solution step by using a weighted GEE with a working correlation matrix for the non-degenerate component to account for the dependence in the data. In simulation studies they conducted, multinomial mixing models did not show much gain in efficiency of the parameter estimators compared to binary mixing models. The ES algorithm for multinomial mixing models ignores dependence in the non-degenerate data when computing the E-step and hence little or no efficiency is gained.

For clustered data, the E-step is a conditional expectation of the complete data loglikelihood function given the observed data and current parameter estimates. That computation requires a joint multivariate distribution for the random vector of the latent indicators of

zero-inflation and the random vector of the response. This joint multivariate distribution is usually intractable and not easy to specify, let alone work with. In this paper, we shall introduce a more flexible EM-type algorithm for fitting marginal zero-inflated regression models assuming a multinomial mixing structure.

Specifically, we propose the regression-solution (RS) algorithm which alternates between a regression step (R-step) and a solution step (S-step). The R-step imputes the missing data by estimating the conditional expectation of the missing data given the observed data and current parameter estimates using a regression model. The S-step is similar to the S-step of the ES algorithm for multinomial mixing models. The regression step directly models the dependence in the data, unlike the E-step of the ES algorithm. We expect to obtain parameter estimators which are efficient compared to the ES algorithm. The algorithm is motivated by the correspondence between *linear discriminant analysis* and multinomial logit models.

In section 3.2 we shall briefly explore the background literature on marginal ZI regression models. Section 3.3 describes the ideas behind the regression-solution (RS) algorithm and its extensions to the unbalanced cluster sizes. In section 3.4 we shall conduct a simulation study to compare the statistical properties of the parameter estimators produced by the RS algorithm and the ES algorithm. Section 3.5 implements the RS algorithm and other algorithms to a longitudinal data set containing counts of hospital visits. Lastly, we conclude by summarizing results, providing recommendations, and discussing future research work.

3.2 Background

This section briefly describes marginal ZI regression models for clustered data with excess zeros as considered by Hall and Zhang [2004] and by Benesi and Hall [2020a]. The ES algorithm (Rosen et al. [2000]) is used to fit the models to data. A brief description of the

relationship between multinomial logit models and linear discriminant analysis is covered. In the subsequent section we will use this relationship to establish the regression-solution algorithm for fitting marginal ZI regression models with multinomial mixing structure.

3.2.1 Marginal ZI regression models for clustered data with excess zeros assuming binary mixing

For each cluster i , $i = 1, \dots, K$, assume the data consists of a $n_i \times 1$ vector $\mathbf{y}_i = (y_{i1}, \dots, y_{in_i})$ for the response and a corresponding vector $\mathbf{u}_i = (u_{i1}, \dots, u_{in_i})$ for the latent indicators of zero-inflation. Assume that the random variables Y_{ij} for the y_{ij} observation follows a marginal ZI distribution, that is, $Y_{ij} \sim 0$ with probability π_{ij} and $Y_{ij} \sim F_2$ with probability $1 - \pi_{ij}$. Note that $Y_{ij} \sim 0$ means that the response Y_{ij} follows a zero degenerate distribution. The distribution F_2 can be a member of the exponential dispersion family, but that is not necessary. The mean of F_2 , $\boldsymbol{\mu}_i = (\mu_{i1}, \dots, \mu_{in_i})^T$, and the mixing probabilities $\boldsymbol{\pi}_i = (\pi_{i1}, \dots, \pi_{in_i})^T$ are respectively related to covariates via generalized linear models (GLMs)

$$g_\mu(\boldsymbol{\mu}_i) = \mathbf{B}_i \boldsymbol{\beta},$$

for F_2 , the non-degenerate distribution, and

$$g_\pi(\boldsymbol{\pi}_i) = \mathbf{G}_i \boldsymbol{\gamma},$$

for the mixing mechanism. Here, $g_\mu(\cdot)$ is a link function suitable for the scale of the response (e.g., $\log(\cdot)$ for unbounded counts), $g_\pi(\cdot)$ is a link function for binary data such as logit, $\mathbf{B}_i$ and $\mathbf{G}_i$ are respectively the $n_i \times p$ and $n_i \times q$ model matrices, and $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$ are unknown regression parameter vectors.

Assuming independence both within and across clusters, the ‘complete data’ loglikelihood is given by

$$\begin{aligned}
l^c(\boldsymbol{\beta}, \phi, \boldsymbol{\gamma}; \mathbf{y}, \mathbf{u}) &= \sum_{i,j} \{u_{ij} \log \pi_{ij}(\boldsymbol{\gamma}) + (1 - u_{ij}) \log [1 - \pi_{ij}(\boldsymbol{\gamma})]\} + \\
&\quad \sum_{i,j} (1 - u_{ij}) \log f_2(y_{ij}; \boldsymbol{\beta}, \phi) \\
&\equiv l^c(\boldsymbol{\gamma}; \mathbf{y}, \mathbf{u}) + l^c(\boldsymbol{\beta}, \phi; \mathbf{y}, \mathbf{u}).
\end{aligned} \tag{3.1}$$

Hall and Zhang [2004] used the ES-1 algorithm to obtain parameter estimates. The E-step in this approach is the conditional expectation of the complete data loglikelihood (3.1) given the observed and current parameter estimates, under an assumption of within-cluster independence. This reduces to imputing the missing data using $u_{ij}^{(k+1)} = E(u_{ij} | y_{ij}; \boldsymbol{\gamma}^{(k)}, \boldsymbol{\beta}^{(k)})$. The S-step is achieved by relaxing the independence assumption and fitting a set of GEEs with working correlation matrices to account for within cluster association. A “quasi-binomial” fitted with GEEs is used to estimate $\boldsymbol{\gamma}$ and a weighted GEE, e.g., using a Poisson mean and variance function for unbounded count data, is used to estimate $\boldsymbol{\beta}$. The correlation parameters $\boldsymbol{\rho}$ and $\boldsymbol{\delta}$ are estimated simultaneously with regression parameters $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$ using extended GEEs of Hall and Severini [1998].

Rosen et al. [2000] established that if the ES-1 algorithm converges, then the resulting parameter estimates solve an unbiased estimating equation. Therefore it can be shown that they are consistent and asymptotically normal under suitable regularity conditions. The variance-covariance matrix can be estimated from the estimating functions used at the S-step.

If we let $\boldsymbol{\psi} = (\boldsymbol{\gamma}^T, \boldsymbol{\beta}^T, \boldsymbol{\rho}^T, \boldsymbol{\delta}^T)^T$ be the combined parameter vector with corresponding ES estimator $\hat{\boldsymbol{\psi}}$, then the asymptotic variance-covariance matrix, $\text{avar}(\hat{\boldsymbol{\psi}})$, is consistently estimated by $\hat{\boldsymbol{\xi}}^{-1} \hat{\boldsymbol{\zeta}} (\hat{\boldsymbol{\xi}}^{-1})^T$ where $\hat{\boldsymbol{\xi}} = \sum_{i=1}^n \sum_{j=1}^{n_i} \nabla s_{ij}(\hat{\boldsymbol{\psi}})$ and $\hat{\boldsymbol{\zeta}} = \sum_{i=1}^n \left\{ \sum_{j=1}^{n_i} s_{ij}(\hat{\boldsymbol{\psi}}) \right\} \left\{ \sum_{j=1}^{n_i} s_{ij}(\hat{\boldsymbol{\psi}}) \right\}^T$, where $s_{ij}(\hat{\boldsymbol{\psi}})$ is a $\dim(\boldsymbol{\psi}) \times 1$ vector containing the estimating functions of $\boldsymbol{\gamma}, \boldsymbol{\delta}, \boldsymbol{\beta}, \boldsymbol{\rho}$ and $\boldsymbol{\phi}$ used in the S-step. Numerical derivatives can be used to obtain $\nabla s_{ij}(\hat{\boldsymbol{\psi}})$.

3.2.2 Marginal ZI regression models for clustered data with excess zeros: multinomial mixing

The multinomial mixing structure assumes a multinomial random variable on the mixing patterns of zero-inflation indicators instead of treating the indicators as binary random variables at each time point/measurement occasion. Within-cluster association is accounted for by the multivariate structure of the multinomial distribution. For each cluster i , the data consists of a $n_i \times 1$ vector $\mathbf{y}_i = (y_{i1}, \dots, y_{in_i})$ of responses and a corresponding categorical random variable $\omega_i \in \{1, 2, \dots, R_i\}$ where $R_i = 2^{n_i}$, for the class of the ZI mixing patterns, $i = 1, \dots, K$. The means of the non-degenerate distribution, $\boldsymbol{\mu}_i = (\mu_{i1}, \dots, \mu_{in_i})^T$, and the multinomial mixing probabilities $\boldsymbol{\nu}_i = (\nu_{i1}, \dots, \nu_{iR_i})$ are related to covariates via GLM-like specifications,

$$g_{\mu}(\boldsymbol{\mu}_i) = \mathbf{B}_i \boldsymbol{\beta},$$

for the non-degenerate distribution and

$$\mathbf{g}_{\nu}(\boldsymbol{\nu}_i) = \mathbf{G}_i^T \boldsymbol{\gamma},$$

for the mixing patterns of zero-inflation. Here, $\mathbf{g}_\nu(\cdot)$ is a vector-valued link function suitable for a multinomial response, for example, a generalized logit link function, $\mathbf{G}_i$ is a design matrix, and $\boldsymbol{\gamma}$ is the regression parameter.

Benesi and Hall [2020a] used the ES algorithm to obtain the regression parameter estimates. The E-step computes the conditional expectation of the complete data loglikelihood under the simplifying assumption of independence conditional on the observed data and current parameter estimates. This reduces to imputing a latent multinomial random vector that specifies the mixing pattern underlying the data for each cluster. The S-step solves a weighted GEE with a working correlation matrix using the response data and a multinomial logit model using the imputed mixing data. The ES algorithm depends on the simplifying assumption of independence to compute the E-step. To be done properly, the E-step requires a joint multivariate distribution for the random vector of responses and the multinomial mixing random variables. However, a suitable and tractable joint distribution is not readily available. The regression-solution algorithm, introduced below, will circumvent the need for a joint distribution by using a suitably chosen regression model to estimate the conditional expectation at the E-step.

3.3 Regression solution (RS) algorithm for marginal ZI regression models

In this section we shall develop the regression-solution (RS) algorithm, focusing on clusters of size $n_i = 2, \forall i$ to simplify notation. However, extensions to cluster sizes of more than $n_i = 2$ follows naturally, and we will also briefly discuss extensions to that context and to the case where the number of observations within a cluster are unbalanced. First, we discuss the relationship between classical linear discriminant analysis and multinomial logit models before linking these ideas to the RS algorithm.

3.3.1 Linear discriminant analysis and multinomial logit models

Before we describe the RS algorithm in detail, we present a quick overview of the relationship between linear discriminant analysis (LDA) and multinomial logit models. The setting of LDA is one in which we assume that we have multivariate data from several groups (or classes) and we wish to use those data to develop a function, which could be used as a classifier, as is the focus in classification analysis, to discriminate between the groups. Classical LDA assumes that we observe data on a random vector $\mathbf{y}$ which is, for each class, multivariate normal with mean that differs from class to class. In the two class case, under an assumption that the covariance matrix of $\mathbf{y}$ is constant across classes, this yields a linear discriminant function which has a form similar to that of a logistic regression of the class labels on $\mathbf{y}$. Thus, there is a close relationship between classical LDA and so-called logistic discriminant analysis where $\mathbf{y}$ plays the role of a covariate vector.

The latter approach has the advantage that it does not depend upon any distributional assumption on $\mathbf{y}$ and is, therefore, more robust to non-normality of $\mathbf{y}$ and more general in its domain of application. It is well-known that classical discriminant analysis typically outperforms logistic discriminant analysis when normality assumptions hold, but not by a large margin, and the logistic approach offers greater flexibility and robustness.

Suppose we have a single covariate X , say a bio-marker for cancer, and we want to classify subjects into two groups: (1) Diseased and (2) Non-diseased. Let

$$\begin{aligned} X &\sim N(\mu_0, \sigma^2) = f_0, \text{ for Diseased } (D), \text{ and} \\ X &\sim N(\mu_1, \sigma^2) = f_1, \text{ for Non-diseased } (\bar{D}). \end{aligned}$$

Suppose we have prior probabilities π that a subject is diseased, and $1 - \pi$ that a subject is non-diseased. Let D be the random variable that a subject is diseased. We want $Pr(D|X = x)$ which can be used as the basis for classification via, for example, the rule: if $Pr(D|X = x) > Pr(\bar{D}|X = x)$, then classify a subject in the diseased class ($D = 1$). By the Bayes' Theorem we have

$$\begin{aligned} Pr(D|X = x) &= \frac{f_1(x)Pr(D)}{f_1(x)Pr(D) + f_0(x)Pr(\bar{D})} \\ &= \frac{\frac{\pi}{1-\pi} \exp\{\frac{\mu_0^2 - \mu_1^2}{2\sigma^2} + \frac{\mu_1 - \mu_0}{\sigma^2}x\}}{\frac{\pi}{1-\pi} \exp\{\frac{\mu_0^2 - \mu_1^2}{2\sigma^2} + \frac{\mu_1 - \mu_0}{\sigma^2}x\} + 1}. \end{aligned}$$

This expression implies a logistic model for the conditional probability of diseased having the form:

$$\text{logit}\{Pr(D|X = x)\} - \log\left(\frac{\pi}{1-\pi}\right) = \alpha + \beta x, \quad (3.2)$$

where $\log\left(\frac{\pi}{1-\pi}\right)$ is an offset. This is a conditional likelihood version of the full likelihood model that assumes that X is normal in each of the two groups (classical discriminant analysis model), and we have

$$\alpha = \frac{\mu_0^2 - \mu_1^2}{2\sigma^2}, \quad \beta = \frac{\mu_1 - \mu_0}{\sigma^2}$$

The model in equation 3.2 is known as *logistic discriminant analysis*, and, while it can be derived from classical normal-theory model, its use and validity does not depend upon that derivation. Classical discriminant analysis requires normality of the covariates and hence is more restrictive than the logistic model. However, when the assumptions underlying classical discriminant analysis hold, it can be substantially more powerful than logistic discriminant analysis.

The close connection between classical discriminant analysis and logistic regression extends to the case of more than two groups, but then the relevant logistic model becomes a generalized logit model based on a multinomial distributional assumption for the groups rather than a binomial model with an ordinary logit link. Classical linear discriminant analysis under normality yields an expression for the conditional mean of a group indicator given the data (which is, the best predictor of group membership) that has a form very similar to a multinomial logit model with covariate $\mathbf{y} - \boldsymbol{\mu}$. Such a regression can be seen as a natural alternative in the best linear predictor suitable for a categorical response; as is well known, the best linear predictor of a random variable Y_0 given a random vector $\mathbf{Y}$ with realized value $\mathbf{y}$ takes the form of a linear predictor in $\mathbf{y}$ minus its mean. Moreover, under an assumption of unequal covariance of $\mathbf{y}$ across groups, LDA becomes QDA, in which the discriminant function, which corresponds to the linear predictor of our multinomial logit model, becomes quadratic in $\mathbf{y}$ minus its mean.

Now for our purposes, suppose $\mathbf{y}_i$ is multivariate normal; that is, $\mathbf{y}_i|\omega_i = r \sim N(\boldsymbol{\mu}_i, \boldsymbol{\Sigma})$, for $r = 1, \dots, R$, $R = 2^n$. Here, ω_i is a latent variable representing the mixing patterns of zero-inflation as defined below and $\mathbf{y}_i$ is the observed response for cluster i . Then the conditional expectation of $\omega_i = r$ given $\mathbf{y}_i$ is given by

$$\begin{aligned} E(\tilde{\omega}_{ir}|\mathbf{y}_i; \boldsymbol{\psi}) &= P(\tilde{\omega}_{ir} = 1|\mathbf{y}_i; \boldsymbol{\psi}) = \frac{f(\mathbf{y}_i; \boldsymbol{\mu}_{ir}, \boldsymbol{\Sigma})P(\tilde{\omega}_{ir} = 1)}{\sum_{h=1}^R f(\mathbf{y}_i; \boldsymbol{\mu}_{ih}, \boldsymbol{\Sigma})P(\tilde{\omega}_{ih} = 1)} \\ &= \frac{\nu_{ir}f(\mathbf{y}_i|\omega_i = r)}{\sum_{l=1}^R \nu_{il}f(\mathbf{y}_i|\omega_i = l)}. \end{aligned}$$

This leads to an equivalent multinomial logit model of the form

$$\log \left(\frac{\tilde{\omega}_{ir}}{\tilde{\omega}_{iR}} \right) = \log \left(\frac{\nu_{ir}}{\nu_{iR}} \right) + \mathbf{G}_i^T \boldsymbol{\gamma}, \quad r = 1, \dots, R-1,$$

where $\mathbf{G}_i$ has terms that depend on $(\mathbf{y}_i - \boldsymbol{\mu}_i)$, the residuals.

3.3.2 The regression-solution (RS) algorithm

The RS algorithm is used to estimate marginal ZI regression models assuming a multinomial mixing structure. For each response vector $\mathbf{Y}_i$, for cluster $i = 1, \dots, K$ of length n_i , let ω_i be a categorical random variable taking values in $\{1, 2, \dots, R_i\}$ where $R_i = 2^{n_i}$ is the total number of all possible patterns of zero inflation, $u_{i1}, \dots, u_{in_i}$ for the observation-specific indicators of zero-inflation, for cluster i . Let the vector $\tilde{\boldsymbol{\omega}}_i = (\tilde{\omega}_{i1}, \dots, \tilde{\omega}_{iR_i-1})^T$ be the vector of indicators specifying the value of ω_i . That is, if $\tilde{\omega}_{ir}$ is the r th element of $\tilde{\boldsymbol{\omega}}_i$, then $\tilde{\omega}_{ir} = I(\omega_i = r)$, $r = 1, \dots, R_i - 1$. For simplicity, we present the case $n_i = 2$ for all i (the bivariate case), but extensions to larger clusters and varying cluster sizes n_i is possible.

In the bivariate case, the possible values ω_i can take are displayed in Table 3.1, along with the corresponding u_{ij} 's, multinomial probabilities, which we denote by $\boldsymbol{\nu}_i = (\nu_{i1}, \dots, \nu_{iR})^T$, and bivariate means for each pattern of the mixing class. A marginal ZIP regression model is used for illustration, but the approach applies to other ZI models such as ZINB and ZIB regression.

Here, ω_i represents the category for the pattern of zero inflation. Similar to Benesi and

Table 3.1: Patterns of the mixing mechanism

Prob	u_{i1}	u_{i2}	ω_i	$\tilde{\boldsymbol{\omega}}_i$	mean vector
ν_{i1}	0	0	1	$(1, 0, 0)^T$	$\boldsymbol{\mu}_{i1} = (\lambda_{i1}, \lambda_{i2})^T$
ν_{i2}	0	1	2	$(0, 1, 0)^T$	$\boldsymbol{\mu}_{i2} = (\lambda_{i1}, 0)^T$
ν_{i3}	1	0	3	$(0, 0, 1)^T$	$\boldsymbol{\mu}_{i3} = (0, \lambda_{i2})^T$
ν_{i4}	1	1	4	$(0, 0, 0)^T$	$\boldsymbol{\mu}_{i4} = (0, 0)^T$

Hall [2020a], we assume the generalized logit model given by

$$\log(\nu_{ir}/\nu_{iR}) = \mathbf{G}_i \boldsymbol{\gamma}, \quad r = 1, 2, \dots, R-1, \quad (3.3)$$

for the mixing data and a GLM for the Poisson counts with a log-link function given by

$$\log(\boldsymbol{\lambda}_i) = \mathbf{B}_i \boldsymbol{\beta}.$$

Let $\boldsymbol{\psi} = (\boldsymbol{\gamma}^T, \boldsymbol{\beta}^T)^T$ be the vector of regression parameters for the zero component and the Poisson component. In addition, let $\tilde{\omega}_{ir} = 1$ if $\omega_i = r$ and 0 otherwise. Define $Y_{ij} = (1 - U_{ij})Z_{ij}$ where Z_{ij} are independent Poisson random variables with mean λ_{ij} and $U_{ij} = 1$ if $Y_{ij} \sim 0$, and zero otherwise. Assuming independence, we have

$$\begin{aligned} \omega_i = 1 &\Rightarrow (Y_{i1}, Y_{i2}) = (Z_{i1}, Z_{i2}), \\ \omega_i = 2 &\Rightarrow (Y_{i1}, Y_{i2}) = (Z_{i1}, 0), \\ \omega_i = 3 &\Rightarrow (Y_{i1}, Y_{i2}) = (0, Z_{i2}), \text{ and} \\ \omega_i = 4 &\Rightarrow (Y_{i1}, Y_{i2}) = (0, 0). \end{aligned}$$

Then the complete data loglikelihood function is as follows:

$$\begin{aligned} l^c(\boldsymbol{\psi}; \mathbf{y}, \tilde{\boldsymbol{\omega}}) &= \sum_{i=1}^K \sum_{r=1}^R \tilde{\omega}_{ir} \log \nu_{ir}(\boldsymbol{\gamma}) + \sum_{i=1}^K \sum_{r=1}^R \sum_{j=1}^{n_i} \tilde{\omega}_{ir} \log f(y_{ij} | \boldsymbol{\mu}_{ir}(\boldsymbol{\beta})) \\ &= \sum_{i=1}^K \sum_{r=1}^R \tilde{\omega}_{ir} \log \nu_{ir}(\boldsymbol{\gamma}) + \sum_{i=1}^K \sum_{j=1}^{n_i} (1 - u_{ij}) \log f(y_{ij} | \boldsymbol{\mu}_{ir}(\boldsymbol{\beta})) \\ &= l^c(\boldsymbol{\gamma}; \mathbf{y}, \tilde{\boldsymbol{\omega}}) + l^c(\boldsymbol{\beta}; \mathbf{y}, \mathbf{u}(\tilde{\boldsymbol{\omega}})). \end{aligned} \quad (3.4)$$

The correlation in the mixing data is accounted for by the multinomial distribution. Benesi and Hall [2020a] computed the E-step as the conditional expectation of the complete data quasi-loglikelihood in equation (3.4) given the observed data and current parameter estimates to get the multinomial mixing weights. The S-step for the non-degenerate data remains the same as in Hall and Zhang [2004] and we fit a multinomial logit model for the multinomial mixing weights for the mixing data.

Hall and Zhang [2004] and Benesi and Hall [2020a] assumed independence to get closed form solutions of the E-step using the Bayes' formula. The E-step, in general, does not have a closed form solution if we allow dependence in the data because a joint multivariate distribution for the mixing data and the response data is intractable. The RS algorithm replaces the E-step of the ES algorithm with a regression step that estimates the more complex conditional expectation to account for the dependence in the data. The RS algorithm is described below as follows.

R-step: At the $(k+1)^{th}$ step, $\tilde{\omega}_{ir}^{(k+1)} \equiv E(\tilde{\omega}_{ir} | \mathbf{y}_i; \boldsymbol{\psi}^{(k)})$ is obtained via a generalized logit multinomial regression model of the form

$$\tilde{\omega}_{ir}^{(k+1)} = \frac{\exp(o_{ir}^{(k)} + \eta_{ir}^{(k)})}{\sum_l \exp(o_{il}^{(k)} + \eta_{il}^{(k)})}, \quad r = 1, \dots, R-1, \quad (3.5)$$

where $o_{ir}^{(k)} = \log(\nu_{ir}^{(k)} / \nu_{iR}^{(k)})$ is an offset, $\eta_{ir}^{(k)}$ is a linear predictor with covariates formed from the residuals, $(\mathbf{y}_i - \boldsymbol{\mu}_i)$, their quadratic terms and cross product terms. We make the convention that $o_{iR}^{(k)} + \eta_{iR}^{(k)} = 0$ for the last category for the sake of identifiability. Such a model yields the conditional log odds, $\log(\tilde{\omega}_{ir}^{(k+1)} / \tilde{\omega}_{iR}^{(k+1)})$, of the r th pattern of zero inflation by updating the unconditional log odds $o_{ir}^{(k)}$ via a regression on the residual terms. The specific form of the linear predictor $\eta_{ir}^{(k)}$ in equation (3.5) is motivated by the connections between multinomial logit models and classical linear discriminant analysis. The derivation

is given in Appendix B for the case with $n_i = 2, \forall i$ observations in each cluster and can be generalized to more than two clusters and unbalanced cases. Note that the sum on the denominator of model (3.5) is restricted to only the possible patterns of the mixing structure for the i^{th} cluster because the presence of observed nonzero response values precludes certain categories from occurring. This is similar to a multinomial discrete choice model in which the multinomial categories correspond to a choice set that may vary across subjects. In our context, only the observed zeros are ambiguous concerning the component from which they emanate and hence a given pattern of nonzero responses restricts the choice-set as we know certainly that any non-zero response is not from the zero component. So the pattern of positive values in $\mathbf{y}_i$ determine the choice-set, which can be accommodated by taking the denominator of equation (3.5) to be a sum over the categories that are possible for ω_i given the pattern of non-zero values in $\mathbf{y}_i$. Models of this form with non-constant choice sets arise in discrete-choice modeling and can be implemented in multinomial logit software such as the `mlogit` package (Croissant [2019]) in R. The resulting regression model for estimating the E-step is referred to as the discrete-choice multinomial logit model.

S-step: The S-step for $\boldsymbol{\gamma}$ is obtained by fitting the discrete-choice multinomial model in equation (3.3) using the imputed “missing data”, $\tilde{\boldsymbol{\omega}}^{(k)}$. The S-step for $\boldsymbol{\beta}$ is obtained by solving a weighted GEE with a working correlation matrix to account for the dependence in the response data. Extended GEEs (Hall and Severini [1998]) are used to solve the joint estimating equations for the regression parameter vector, $\boldsymbol{\beta}$ and the association parameter vector, $\boldsymbol{\rho}$. The u_{ij} ’s are calculated from the imputed multinomial mixing weights from the R-step as follows:

$$u_{ij} = \sum_{r: d_j(r)=1} \tilde{\omega}_{ir}, \quad j = 1, \dots, n,$$

where $d_1(r), d_2(r), \dots, d_n(r), d_j(r) \in \{0, 1\}$ for all j is the binary representation of $r-1$ for $r \in \{1, \dots, R\}$. For example, for $n = 3$ and $r = 1$, $d_1(r), d_2(r), \dots, d_n(r) = 0, 0, 0$ and $u_{ij} = 0$; for $n = 3$ and $r = 8$, $d_1(r), d_2(r), \dots, d_n(r) = 1, 1, 1$. With this notation, $u_{i1} = \tilde{\omega}_{i5} + \tilde{\omega}_{i6} + \tilde{\omega}_{i7} + \tilde{\omega}_{i8}$, $u_{i2} = \tilde{\omega}_{i3} + \tilde{\omega}_{i4} + \tilde{\omega}_{i7} + \tilde{\omega}_{i8}$, and $u_{i3} = \tilde{\omega}_{i2} + \tilde{\omega}_{i4} + \tilde{\omega}_{i6} + \tilde{\omega}_{i8}$.

3.3.3 Extensions to larger and/or unbalanced clusters

The RS algorithm can be extended to larger cluster sizes and unbalanced cluster sizes. Unequal clusters result when subjects may enter the study at different times or may be lost to follow-up. [Benesi and Hall, 2020a] considered three approaches that are useful in some applications which are: (1) modeling the log-odds of zero inflation at each time point, (2) a multinomial model that models probabilities of the smaller clusters as the collapsed sums of the probability vector of the largest cluster which is assumed to be constant, and (3) using log-linear models for the $2 \times 2 \times \dots \times 2$ table corresponding to zero-inflation at time $1 \times$ zero inflation at time $2 \times \dots \times$ zero inflation at time n_i . While some of these approaches may be applicable for the RS algorithm, the form of the R-step needs to accommodate unbalanced clusters and larger clusters. The extent to which such an R-step is more efficient than an E step under independence would depend on the form of that regression model and is a topic for future investigation.

The R step for a framework that takes approach (1) would be a logit model for the conditional zero-inflation mixing patterns at each measurement occasion at which we observe $y_{ij} = 0$, that is,

$$(\mathbf{A}_i^T \tilde{\boldsymbol{\omega}}_i^{(k+1)})_j = \frac{\exp(o_{ij}^{(k)} + \eta_{ij}^{(k)})}{1 + \exp(o_{ij}^{(k)} + \eta_{ij}^{(k)})}, \quad (3.6)$$

where $\mathbf{A}_i$ is an $R_i \times n_i$ matrix of indicators for zero inflation at each time point, $o_{ij}^{(k)} = (\mathbf{A}_i^T \boldsymbol{\nu}_i^{(k)})_j / (1 - (\mathbf{A}_i^T \boldsymbol{\nu}_i^{(k)})_j)$ is an offset, and $\eta_{ij}^{(k)}$ is a linear predictor with terms involving $\mathbf{y}_i - \boldsymbol{\mu}_i$. The dimension of the covariate vector has to be the same for all i , and hence the covariates would have to include $x_{i1} = y_{i1} - \mu_{i1}, \dots, x_{iM} = y_{iM} - \mu_{iM}$ for $M = \max(n_i)$, where x_{ik} is set to zero when k is a time point not observed for cluster i . Similarly, covariates would have to be included for all potential quadratic and cross-product terms based on $\max(n_i)$. Model (3.6) updates the unconditional log odds for the mixing patterns at each measurement occasion $o_{ij}^{(k)}$ to obtain the conditional log odds given the residual terms and current parameter estimates.

3.4 Simulations

In this section we will conduct a simulation study to investigate the properties of the estimators from marginal ZI regression models assuming a multinomial mixing structure using the RS algorithm (*RS_MMX*) and the ES algorithm (*ES_MMX*), and the ES algorithm (*ES_BMX*) for binary mixing models. We generate correlated count data, $\mathbf{z}_i$, using a multivariate Gaussian copula method implemented in the *simstudy* package (Goldfeld [2019]) in R. Multinomial mixing random variables, ω_i , are generated to introduce excess zeros in the response variable $\mathbf{y}_i$. That is, for each subject we generate $\mathbf{z}_i \sim \text{Poisson}(\boldsymbol{\mu}_i)$ where

$\mu_{ij} = \exp(\beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2})$ for $i = 1, 2, \dots, K$ and $j = 1, 2, 3$. The covariates X_{i1} and X_{i2} are generated from standard normal random variables. Excess zeros are introduced into the correlated Poisson counts as determined by the class of the mixing pattern generated from the multinomial distribution. That is, the generated class for the mixing pattern for cluster i determines where to fix the zeros, for example, in the case with cluster size $n = 2$, with generated counts $\mathbf{z}_i = (z_{i1}, z_{i2})$, and a generated mixing pattern of, say $\omega_i = 3$, which corresponds to zero-inflation at time point 1 will result in $\mathbf{y}_i = (0, z_{i2})$. Covariates can be added via a multinomial logit model, but for simplicity, we assume an intercept only multinomial logit model $\log(\nu_r/\nu_R) = \gamma_r, r = 1, 2, \dots, R - 1$, in the simulations. The three fitting algorithms *RS_MMX*, *ES_MMX* and *ES_BMX* were each successively fit to 500 data sets that were randomly generated as described above and under all the possible combinations of sample sizes ($K = 100$ and $K = 200$), the correlation structures (Exchangeable and $AR(1)$) and the correlation strengths ($\rho = 0.3$ and $\rho = 0.8$). The true regression parameters used to generate the data were fixed to $\boldsymbol{\beta} = (1.5, 0.43, -0.49)^T$ for the Poisson mean and $\boldsymbol{\gamma} = (1, 1, 1, 1, 1, 1, 1)^T$ for the multinomial mixing component. The mixing parameters imply a constant zero-inflation probability of approximately $\pi = 0.4571$ at each time point.

Table 3.2 shows the averaged Monte Carlo statistical properties of the estimators for the regression parameters and the zero-inflation probabilities fitted under the correlation strength of $\rho = 0.3$. The statistical properties improve for larger clusters as expected. The Monte Carlo standard deviations (SD) for the estimators are quite close to the averaged asymptotic standard errors ($\overline{ASE}$) indicating that the sandwich estimator for the variance-covariance matrix is a reasonable estimator of the variance covariance matrix for the parameter estimates. The ratios of MSEs are almost entirely below one for the *RS_MMX* versus the *ES_BMX* and the *ES_MMX*.

Consistent with what Benesi and Hall [2020a] observed, there are no strong differences in efficiency between the *ES_BMX* and the *ES_MMX*. We can conclude that as expected, the RS algorithm produces more efficient estimators compared to the other methods.

Table 3.3 shows the result for test size (5%) and coverage (95%) for the three fitting algorithms using a correlation strength of $\rho = 0.3$. The results indicate the three methods roughly perform about the same in terms of test size and coverage of the true parameters. The results show that test sizes are greater than the nominal size of 5% and low coverage rates than expected for all the methods. From Table 3.2 we observe that the average $\overline{ASE}$ s are slight underestimates of the average Monte Carlo SDs and this results in tighter confidence intervals with under-coverage.

Table 3.4 shows the averaged Monte Carlo statistical properties of the estimators for the regression parameters and the zero-inflation probabilities fitted under a correlation strength of $\rho = 0.8$. The statistical properties are better for larger clusters compared to smaller clusters. The sandwich estimator is a reasonable estimator of the variance covariance matrix of the parameter estimators as the Monte Carlo standard deviations (SD) for the estimators are quite close to the averaged asymptotic standard errors ($\overline{ASE}$). The RS algorithm (*RS_MMX*) produces more efficient estimators than the other two algorithms as indicated by the ratios of the MSEs that are almost entirely less than one versus the *ES_BMX*.

Table 3.5 shows the 95% coverage and the 5% test size results for the three fitting algorithms using a correlation strength of $\rho = 0.8$. Similar to Table 3.3, the test size and the coverage from all methods are roughly the same with often exceeding test sizes than the nominal 5% and low coverage than expected. The results show improvement with increasing number of clusters.

3.4.1 Comparing different R-step models

Let $r_{ij} = y_{ij} - \mu_{ij}$ be the residual for the j^{th} observation in the i^{th} cluster. The RS algorithm derived in Appendix B shows that the regression model depends on the residual terms that have squared terms, r_{ij}^2 and the cross-product terms, $r_{ij}r_{ik}$. There are no linear terms r_{ij} in the derivation, however, we expect linear terms to be in the model because generally, the form of $E(X|Y)$ depends on the means (linear), standard deviations (squared) and correlations (cross products) of the random variables. It is worthy to explore the exclusion and inclusion of the linear terms and the quadratic terms while fixing the cross product terms in the regression step of the RS algorithm. The simulation results are shown in Table 3.6.

The ratios of MSEs in Table 3.6 show that the inclusion of both linear and quadratic terms results in slightly improved regression estimators compared to excluding any of these terms. Estimators from the models with linear terms only are slightly better than those from models with quadratics terms only.

3.5 Example

In this section we analyze the number of hospital visits data from Karim and Zeger [1988], and were considered by Benesi and Hall [2020a]. The data have complete cases obtained from each of the $K = 73$ children followed for four quarters of the year, making $73 \times 4 = 292$ observations. The potential predictors are gender of the child, the maternal smoking status of the mother, the baseline age of the child in months, and the effects of time (quarter). The data exhibit presence of excess zeros as shown by Benesi and Hall [2020a], who showed that a regular Poisson GLM was under-fitting the observed zeros at each time point and overall. The marginal ZI Poisson model with binary and multinomial mixing structures were considered. The models were fitted using *ES_BMX*, *ES_MMX* and *RS_MMX* methods.

We assumed an exchangeable working correlation matrix for the non-degenerate part and an exchangeable working correlation for the degenerate part of the binary mixing model. Choice of the “best” working correlation matrix is a common question that arise in the analysis of clustered data with GEEs. The regression estimates are the same regardless of which correlation structure is chosen, but relatively large standard errors can impact inferences when the correlation structure is misspecified (see Benesi and Hall [2020a], Section 2.5). If the number of clusters is large, however, there is less cost in using a misspecified correlation structure. Often, it is reasonable to assume a correlation structure based on the assumptions made on the data generating process, for example, if observations further away in time are assumed to be less correlated, then an $AR(1)$ correlation structure is reasonable. If there is a reason to believe that correlation within a cluster remains the same over time, then an exchangeable correlation structure is reasonable. Advanced methods such as those investigated by Goshio [2014] can be used if there is no other way to choose or assume a working structure, for example, the QIC which stands for quasi-likelihood under the independence model criterion (Pan [2001]). The QIC behaves like the akaike information criterion (AIC) for likelihood based models and is used to select between competing models.

The fitted results are summarized in Table 3.7 and show that the three fitting algorithms agree in making inferences about the regression parameters. This shows that the RS algorithm is capable of producing reasonable results for marginal ZI regression models in a real-world data setting. The results show parameter estimates for the non-degenerate Poisson component that are group effects for gender and smoking status, together with three-way and two-way interactions involving the effect of time on a continuous scale, coded for convenience as 0, 1, 2 and 3 for quarters 1, 2, 3 and 4, respectively. We are mainly interested in the regression effects of the Poisson counts, but the table also shows the estimated zero-inflation probabilities at each time point. The table also gives estimates of the association parameters for the exchangeable working correlation matrices for both components

where applicable. The computation times measured by the elapsed time in *R* version 3.5.1 using the function *system.time* are respectively, 22.03s, 135.19s and 39.15s for *ES_BMX*, *ES_MMX* and *RS_MMX*. These times were measured on a Dell laptop with an Intel(R) Core(TM) i7-7700HQ quad core CPU @ 2.80GHz with a 64-bit Windows 10 Pro operating system and 16GB of RAM. Multinomial mixing models take more time to compute than the binary mixing model in this example. This can mostly be attributed to the larger number of parameters to be estimated in multinomial mixing models compared to the binary mixing model. The computation time for the *ES_MMX* model is unusually high.

Interaction terms allow us to test for many effects among the groups and how they relate to the effect of time. Suppose we let the zero state represent children who are healthy and do not have any chance of visiting the hospital and the non-zero state represent children with health complications who may visit the hospital. Using the model fitted by *RS_MMX*, the exponent of the intercept (1.7069) in the Poisson component represent the incidence ratio in time/quarter 1 for males whose mothers smoke and are prone to health complications that may require a visit to the hospital. The three-way interaction term is significant which implies that the effects of time vary across the levels of gender and smoking group. Table 3.8 gives the rate ratios per unit increase in time/quarter in the each of the combination of levels for gender and smoking. We can deduce that among children with health complications that may require a visit to the hospital, there is a 9.24% decrease in hospital visitation rate per each quarter for male children whose mothers are smokers (rate ratio = 0.9076), a 19.44% decrease in visitation rate per each quarter for females whose mothers are smokers (rate ratio = 0.8056). Similarly, among children with health complications leading to possible hospitalization, the rates for hospital visits for children whose mothers are non-smokers decreases over time for males (rate ratio = 0.7209), that is a 27.91% decrease in visitation rate per each quarter, but this increases for females (rate ratio = 1.594), that is a 59.4% increase per each quarter. However, there are only $n = 6$ female children whose mothers are

smokers and, on average, about 2 of the 6 would be healthy at each time point requiring no hospitalization. So the increase in the mean intensity over time in this group (NS, F), is due, on average, to about 4 individual sequences. From the fitted models we can predict the excess zeros and the systematic zeros at each time point and overall. We can then compare these predicted values to the observed zeros in the data to assess how well the models fit the data. Table 3.9 shows the predicted zeros and the observed zeros from each of the three methods used for model fitting in this example. The result indicates that multinomial mixing models fit with either the ES or RS algorithm tend to slightly overestimate the observed zeros when compared to the binary mixing model in this data set.

3.6 Summary

The RS algorithm is an algorithm for fitting marginal ZI regression models for clustered data with excess zeros assuming a multinomial mixing structure. The regression step estimates the conditional expectation of the missing data given the observed data and current parameter estimates using a regression model. The S-step solves a weighted GEE for the non-degenerate part and a multinomial logit model for the mixing component. The RS algorithm is more general than the ES algorithm previously used to estimate these models. In simulation studies, the RS algorithm provided efficient parameter estimators compared to the ES algorithm for binary and multinomial mixing models.

The RS algorithm and the ES algorithm were implemented to the number of hospital visits data set. Results show the feasibility of implementing the RS algorithm to a real world data set and the parameters were comparable to those fitted with the ES-1 algorithm. The example shows that we can extend the RS algorithm and the ES algorithm for multinomial mixing models to clusters with more than two time points in a straight-forward manner.

3.6.1 Recommendations and future work

The challenge of assuming a multinomial mixing structure is that the model parameters in the mixing component grow with increasing number of time points. Moreover, real data sets have unbalanced cluster sizes due to subjects dropping off the study or joining later in the study. Methods for handling larger and/or unbalanced clusters have been discussed by Benesi and Hall [2020a] when using the ES-1 algorithm with multinomial mixing models. Though these approaches may be applicable to the RS algorithm, unlike the E-step of the ES-1 algorithm, the R-step is affected by large and/or unbalanced clusters. The regression model for the R-step has to be framed to accommodate large and/or unbalanced clusters. The performance of such a model versus the E-step under independence is a subject for future research.

An important question that arise in the analysis of marginal ZI models is on how to select the best model and/or correlation structure. Loglikelihood based approaches are used to select models when the likelihood is available, however, marginal ZI models are partially specified and the likelihood is unavailable. A method that extend likelihood based approaches in selecting models have been studied by Pan [2001] using the so-called quasi-likelihood information criterion (QIC) which behaves like the akaike information criterion (AIC) for likelihood based models. In the context of marginal ZI regression models, Zhang [2004] in his unpublished PhD dissertation, proposed a quasi-likelihood based information criterion similar in spirit to Pan [2001] and a more general approach that accounts for the working correlation structure. Zhang [2004] admit that despite working well in selecting covariates in the model, his approach showed no incremental benefit compared to the QIC. Thus, there is a need for developing model selection approaches that will work more generally including in multinomial mixing models.

Multinomial mixing models are recommended if the primary interest is to quantify the regression effects in the non-degenerate component. Though regression effects can be added to the mixing component of the model, if one of the mixing patterns in the mixing structure is not observed, care should be taken in interpreting the regression parameters from the S-step. The model restricts the number of the possible mixing patterns of zero-inflation to only those that are observed in the data. This alters the levels in the mixing classes and hence the model interpretation. The regression parameters for the non-degenerate component and the fitted ZI probabilities are, however, not altered. It is recommended to check if all the possible mixing patterns are observed in the data and rather quantify the marginal zero-inflation probabilities at each time point rather than the regression effects from the model for mixing patterns.

Table 3.2: A simulation study for comparing the *ES_BMX*, *ES_MMX* and *RS_MMX* (Correlation parameter $\rho = 0.3$)

Correlation Structure	<i>ES_BMX</i>				<i>ES_MMX</i>				<i>RS_MMX</i>			
	Mean	SD	<i>ASE</i>	MSE	Mean	SD	<i>ASE</i>	MSE	Mean	SD	<i>ASE</i>	MSE
<i>Exchangeable: K = 100</i>												
$\beta_0 = 1.5000$	1.4978	0.0479	0.0464	0.0023 (1.0000)	1.4973	0.0481	0.0464	0.0023 (1.0111)	1.5014	0.0418	0.0415	0.0017 (0.7619)
$\beta_1 = 0.4300$	0.4285	0.0411	0.0379	0.0017 (1.0000)	0.4288	0.0413	0.0379	0.0017 (1.0109)	0.4267	0.0391	0.0354	0.0015 (0.9120)
$\beta_2 = -0.4900$	-0.4904	0.0399	0.0384	0.0016 (1.0000)	-0.4907	0.0399	0.0384	0.0016 (1.0020)	-0.4879	0.0371	0.0356	0.0014 (0.8687)
$\rho = 0.3000$	0.3011	0.1834	0.1555	0.0336 (1.0000)	0.2510	0.1541	0.1324	0.0262 (0.7774)	0.2648	0.1572	0.1369	0.0260 (0.7716)
$\pi_1 = 0.4571$	0.4542	0.0519		0.0027 (1.0000)	0.4538	0.052		0.0027 (1.0040)	0.4565	0.0514		0.0026 (0.9779)
$\pi_2 = 0.4571$	0.4542	0.0519		0.0027 (1.0000)	0.4538	0.052		0.0027 (1.0040)	0.4563	0.0513		0.0026 (0.9746)
$\pi_3 = 0.4571$	0.4536	0.0516		0.0027 (1.0000)	0.4538	0.052		0.0027 (1.0153)	0.4564	0.0510		0.0026 (0.9737)
<i>AR(1): K = 100</i>												
$\beta_0 = 1.5000$	1.4992	0.0469	0.0459	0.0022 (1.0000)	1.4989	0.0470	0.0459	0.0022 (1.0082)	1.5035	0.0414	0.0409	0.0017 (0.7858)
$\beta_1 = 0.4300$	0.4298	0.0392	0.0370	0.0015 (1.0000)	0.4300	0.0392	0.0371	0.0015 (0.9981)	0.4275	0.0372	0.0345	0.0014 (0.9021)
$\beta_2 = -0.4900$	-0.4883	0.0396	0.0375	0.0016 (1.0000)	-0.4885	0.0398	0.0375	0.0016 (1.0092)	-0.4857	0.0369	0.0347	0.0014 (0.8801)
$\rho = 0.3000$	0.2986	0.1915	0.1780	0.2881 (1.0000)	0.3118	0.1931	0.1789	0.2757 (0.9569)	0.3188	0.1968	0.1828	0.2703 (0.9383)
$\pi_1 = 0.4571$	0.4545	0.0517		0.0027 (1.0000)	0.4541	0.0518		0.0027 (1.0048)	0.4573	0.0514		0.0026 (0.9831)
$\pi_2 = 0.4571$	0.4545	0.0517		0.0027 (1.0000)	0.4541	0.0518		0.0027 (1.0048)	0.4575	0.0514		0.0026 (0.9853)
$\pi_3 = 0.4571$	0.4541	0.0518		0.0027 (1.0000)	0.4541	0.0518		0.0027 (1.0015)	0.4571	0.0508		0.0026 (0.9602)
<i>Exchangeable: K = 200</i>												
$\beta_0 = 1.5000$	1.5009	0.0335	0.0332	0.0011 (1.0000)	1.5007	0.0335	0.0332	0.0011 (1.0007)	1.5044	0.0296	0.0297	0.0009 (0.7970)
$\beta_1 = 0.4300$	0.4291	0.0284	0.0274	0.0008 (1.0000)	0.4293	0.0283	0.0274	0.0008 (0.9949)	0.4273	0.0268	0.0256	0.0007 (0.9014)
$\beta_2 = -0.4900$	-0.4899	0.0301	0.0273	0.0009 (1.0000)	-0.4901	0.0301	0.0273	0.0009 (1.0011)	-0.4878	0.0284	0.0254	0.0008 (0.8922)
$\rho = 0.3000$	0.2909	0.1107	0.1088	0.0123 (1.0000)	0.3024	0.1123	0.1101	0.0126 (1.0229)	0.3081	0.1156	0.1119	0.0134 (1.0883)
$\pi_1 = 0.4571$	0.4582	0.0364		0.0013 (1.0000)	0.4580	0.0364		0.0013 (0.9994)	0.4608	0.0360		0.0013 (0.9891)
$\pi_2 = 0.4571$	0.4582	0.0364		0.0013 (1.0000)	0.4545	0.0373		0.0014 (1.0529)	0.4574	0.0369		0.0014 (1.0294)
$\pi_3 = 0.4571$	0.4574	0.0361		0.0013 (1.0000)	0.4573	0.0360		0.0013 (0.9960)	0.4601	0.0358		0.0013 (0.9927)
<i>AR(1): K = 200</i>												
$\beta_0 = 1.5000$	1.4998	0.0339	0.0326	0.0011 (1.0000)	1.4995	0.0338	0.0325	0.0011 (0.9950)	1.5033	0.0300	0.0291	0.0009 (0.7946)
$\beta_1 = 0.4300$	0.4282	0.0280	0.0264	0.0008 (1.0000)	0.4283	0.0280	0.0264	0.0008 (0.9983)	0.4264	0.0266	0.0247	0.0007 (0.9143)
$\beta_2 = -0.4900$	-0.4900	0.0280	0.0267	0.0008 (1.0000)	-0.4902	0.0280	0.0267	0.0008 (1.0027)	-0.4879	0.0264	0.0248	0.0007 (0.8961)
$\rho = 0.3000$	0.2793	0.1279	0.1198	0.0168 (1.0000)	0.2912	0.1277	0.1213	0.0164 (0.9750)	0.2953	0.1309	0.1236	0.0172 (1.0222)
$\pi_1 = 0.4571$	0.4590	0.0392		0.0015 (1.0000)	0.4588	0.0392		0.0015 (1.0010)	0.4617	0.0387		0.0015 (0.9899)
$\pi_2 = 0.4571$	0.4590	0.0392		0.0015 (1.0000)	0.4556	0.0356		0.0013 (0.8246)	0.4585	0.0351		0.0012 (0.8018)
$\pi_3 = 0.4571$	0.4574	0.0358		0.0013 (1.0000)	0.4572	0.0358		0.0013 (0.9982)	0.4601	0.0355		0.0013 (0.9885)

¹ The values in parentheses are ratios of MSEs with *ES_BMX* at the bottom.

Table 3.3: A simulation study for comparing the ES_BMX , ES_MMX and RS_MMX (Test and Coverage: Correlation parameter $\rho = 0.3$)

Correlation Structure	ES_BMX		ES_MMX		RS_MMX	
	Test Size	Coverage	Test Size	Coverage	Test Size	Coverage
<i>Exchangeable: $K = 100$</i>						
β_0	0.060	0.940	0.060	0.940	0.058	0.942
β_1	0.084	0.916	0.088	0.912	0.096	0.904
β_2	0.080	0.920	0.078	0.922	0.090	0.910
ρ	0.104	0.896	0.102	0.898	0.095	0.905
<i>AR(1): $K = 100$</i>						
β_0	0.058	0.942	0.060	0.940	0.062	0.938
β_1	0.096	0.904	0.094	0.906	0.098	0.902
β_2	0.066	0.934	0.070	0.930	0.082	0.918
ρ	0.095	0.905	0.095	0.905	0.093	0.907
<i>Exchangeable: $K = 200$</i>						
β_0	0.054	0.946	0.054	0.946	0.058	0.942
β_1	0.070	0.930	0.070	0.930	0.078	0.922
β_2	0.072	0.928	0.068	0.932	0.122	0.878
ρ	0.092	0.908	0.092	0.908	0.091	0.909
<i>AR(1): $K = 200$</i>						
β_0	0.062	0.938	0.066	0.934	0.060	0.940
β_1	0.060	0.940	0.060	0.940	0.068	0.932
β_2	0.066	0.934	0.062	0.938	0.090	0.910
ρ	0.086	0.914	0.078	0.922	0.075	0.925

Table 3.4: A simulation study for comparing the *ES_BMX*, *ES_MMX* and *RS_MMX* (Correlation parameter $\rho = 0.8$)

Correlation Structure	Mean	<i>ES_BMX</i>			Mean	<i>ES_MMX</i>			Mean	<i>RS_MMX</i>			
		SD	<i>ASE</i>	MSE		SD	<i>ASE</i>	MSE		SD	<i>ASE</i>	MSE	
<i>Exchangeable: K = 100</i>													
$\beta_0 = 1.5000$	1.4961	0.0594	0.0546	0.0035 (1.0000)	1.4952	0.0594	0.0546	0.0035 (0.9996)	1.5008	0.0526	0.0485	0.0028 (0.7801)	
$\beta_1 = 0.4300$	0.4310	0.0462	0.0443	0.0021 (1.0000)	0.4315	0.0461	0.0443	0.0021 (0.9946)	0.4286	0.0436	0.0413	0.0019 (0.8894)	
$\beta_2 = -0.49$	-0.4930	0.0489	0.0442	0.0024 (1.0000)	-0.4936	0.0488	0.0442	0.0024 (0.9989)	-0.4901	0.0458	0.0409	0.0021 (0.8754)	
$\rho = 0.8000$	0.7647	0.1441	0.1413	0.0220 (1.0000)	0.7746	0.1452	0.1424	0.0217 (0.9870)	0.7912	0.1448	0.1508	0.0210 (0.9555)	
$\pi_1 = 0.4571$	0.4543	0.0521	0.0027 (1.0000)		0.4535	0.0521	0.0027 (1.0043)		0.4574	0.0515	0.0027 (0.9757)		
$\pi_2 = 0.4571$	0.4543	0.0521	0.0027 (1.0000)		0.4535	0.0521	0.0027 (1.0043)		0.4557	0.0520	0.0027 (0.9956)		
$\pi_3 = 0.4571$	0.4533	0.0531	0.0028 (1.0000)		0.4535	0.0521	0.0027 (0.9625)		0.4565	0.0522	0.0027 (0.9613)		
<i>AR(1): K = 100</i>													
$\beta_0 = 1.5000$	1.494	0.0602	0.0558	0.0037 (1.0000)	1.4930	0.0601	0.0559	0.0037 (1.0006)	1.4987	0.0534	0.0499	0.0029 (0.7801)	
$\beta_1 = 0.4300$	0.4330	0.0489	0.0438	0.0024 (1.0000)	0.4336	0.0489	0.0438	0.0024 (1.0025)	0.4304	0.0465	0.0409	0.0022 (0.9003)	
$\beta_2 = -0.4900$	-0.4910	0.0501	0.0446	0.0025 (1.0000)	-0.4916	0.0498	0.0446	0.0025 (0.9876)	-0.4883	0.0471	0.0413	0.0022 (0.8826)	
$\rho = 0.8000$	0.758	0.1585	0.1545	0.0269 (1.0000)	0.7671	0.1588	0.1559	0.0263 (0.9776)	0.7838	0.1557	0.1599	0.0245 (0.9111)	
$\pi_1 = 0.4571$	0.4542	0.0520	0.0027 (1.0000)		0.4535	0.052	0.0027 (1.0027)		0.4571	0.0515	0.0027 (0.9793)		
$\pi_2 = 0.4571$	0.4542	0.0520	0.0027 (1.0000)		0.4535	0.052	0.0027 (1.0027)		0.4553	0.0519	0.0027 (0.9961)		
$\pi_3 = 0.4571$	0.4540	0.0528	0.0028 (1.0000)		0.4535	0.052	0.0027 (0.9694)		0.4571	0.0518	0.0027 (0.9596)		
<i>Exchangeable: K = 200</i>													
$\beta_0 = 1.5000$	1.5010	0.0370	0.0388	0.0014 (1.0000)	1.5002	0.0369	0.0388	0.0014 (0.9984)	1.5039	0.0329	0.0346	0.0011 (0.8040)	
$\beta_1 = 0.4300$	0.4276	0.0317	0.0316	0.0010 (1.0000)	0.4281	0.0318	0.0316	0.0010 (0.9984)	0.4261	0.0303	0.0296	0.0009 (0.9188)	
$\beta_2 = -0.4900$	-0.4894	0.0315	0.0316	0.0010 (1.0000)	-0.4899	0.0313	0.0315	0.0010 (0.9896)	-0.4877	0.0297	0.0293	0.0009 (0.8935)	
$\rho = 0.8000$	0.7587	0.1120	0.1035	0.0143 (1.0000)	0.7681	0.1128	0.1040	0.0137 (0.9628)	0.7782	0.1119	0.1033	0.0130 (0.9112)	
$\pi_1 = 0.4571$	0.4572	0.0359	0.0013 (1.0000)		0.4566	0.0359	0.0013 (1.0036)		0.4595	0.0356	0.0013 (0.9897)		
$\pi_2 = 0.4571$	0.4572	0.0359	0.0013 (1.0000)		0.4558	0.0367	0.0013 (1.0471)		0.4588	0.0364	0.0013 (1.0314)		
$\pi_3 = 0.4571$	0.4569	0.0370	0.0014 (1.0000)		0.4563	0.0371	0.0014 (1.0073)		0.4592	0.0368	0.0014 (0.9926)		
<i>AR(1): K = 200</i>													
$\beta_0 = 1.5000$	1.5005	0.0424	0.0394	0.0018 (1.0000)	1.4996	0.0423	0.0394	0.0018 (0.9971)	1.5036	0.0377	0.0352	0.0014 (0.7978)	
$\beta_1 = 0.4300$	0.4305	0.0322	0.0318	0.0010 (1.0000)	0.4310	0.0321	0.0318	0.0010 (0.9961)	0.4289	0.0305	0.0298	0.0009 (0.8982)	
$\beta_2 = -0.4900$	-0.4883	0.0352	0.0321	0.0012 (1.0000)	-0.4890	0.0352	0.0321	0.0012 (0.9983)	-0.4866	0.0331	0.0298	0.0011 (0.8941)	
$\rho = 0.8000$	0.7632	0.1098	0.1120	0.0134 (1.0000)	0.7727	0.1104	0.1126	0.0129 (0.9650)	0.7834	0.1092	0.1115	0.0122 (0.9090)	
$\pi_1 = 0.4571$	0.4590	0.0366	0.0013 (1.0000)		0.4584	0.0367	0.0013 (1.0008)		0.4614	0.0362	0.0013 (0.9846)		
$\pi_2 = 0.4571$	0.4590	0.0366	0.0013 (1.0000)		0.4562	0.0366	0.0013 (0.9932)		0.4593	0.0361	0.0013 (0.9709)		
$\pi_3 = 0.4571$	0.4548	0.0384	0.0015 (1.0000)		0.4541	0.0384	0.0015 (1.0024)		0.4572	0.0381	0.0014 (0.9805)		

¹ The values in parentheses are ratios of MSEs with *ES_BMX* at the bottom.

Table 3.5: A simulation study for comparing the ES_BMX , ES_MMX and RS_MMX (Test and Coverage: Correlation parameter $\rho = 0.8$)

Correlation Structure	ES_BMX		ES_MMX		RS_MMX	
	Test Size	Coverage	Test Size	Coverage	Test Size	Coverage
<i>Exchangeable: $K = 100$</i>						
β_0	0.072	0.928	0.074	0.926	0.074	0.926
β_1	0.076	0.924	0.070	0.930	0.074	0.926
β_2	0.068	0.932	0.068	0.932	0.146	0.854
ρ	0.104	0.896	0.103	0.897	0.102	0.898
<i>AR(1): $K = 100$</i>						
β_0	0.062	0.938	0.060	0.940	0.058	0.942
β_1	0.098	0.902	0.098	0.902	0.104	0.896
β_2	0.104	0.896	0.058	0.942	0.134	0.866
ρ	0.092	0.908	0.092	0.908	0.090	0.910
<i>Exchangeable: $K = 200$</i>						
β_0	0.032	0.968	0.036	0.964	0.034	0.966
β_1	0.048	0.952	0.048	0.952	0.052	0.948
β_2	0.046	0.954	0.048	0.952	0.046	0.954
ρ	0.098	0.902	0.086	0.914	0.082	0.918
<i>AR(1): $K = 200$</i>						
β_0	0.074	0.926	0.072	0.928	0.070	0.930
β_1	0.058	0.942	0.056	0.944	0.062	0.938
β_2	0.082	0.918	0.078	0.922	0.057	0.943
ρ	0.072	0.928	0.062	0.938	0.064	0.936

Table 3.6: A simulation study to compare inclusion/exclusion of linear/quadratic terms in the R-step of the RS algorithm

Correlation Structure	Mean	Both Terms			Mean	Linear Terms Only			Mean	Quadratic Terms Only		
		SD	SE	MSE		SD	SE	MSE		SD	SE	MSE
<i>Exchangeable: n = 100</i>												
$\beta_0 = 1.5000$	1.5031	0.0508	0.0487	0.0026 (0.9982)	1.5002	0.0510	0.0478	0.0026 (1.0000)	1.5026	0.0509	0.0488	0.0026 (1.0011)
$\beta_1 = 0.4300$	0.4287	0.0477	0.0409	0.0023 (0.9967)	0.4269	0.0481	0.0404	0.0023 (1.0000)	0.4290	0.0477	0.0409	0.0023 (1.0175)
$\beta_2 = -0.4900$	-0.4881	0.0452	0.0412	0.0019 (0.9958)	-0.4860	0.0457	0.0408	0.0021 (1.0000)	-0.4884	0.0453	0.0412	0.0021 (1.0260)
$\rho = 0.8$	0.8009	0.1500	0.1624	0.0225 (1.0002)	0.7458	0.1529	0.1639	0.0263 (1.0000)	0.8003	0.1500	0.1861	0.0225 (1.1696)
$\pi_1 = 0.4571$	0.4578	0.0531		0.0028 (1.0000)	0.4578	0.0531		0.0028 (1.0000)	0.4578	0.0531		0.0028 (1.0000)
$\pi_2 = 0.4571$	0.4603	0.0531		0.0028 (1.0000)	0.4603	0.0531		0.0028 (1.0000)	0.4603	0.0531		0.0028 (1.0001)
$\pi_3 = 0.4571$	0.4589	0.0508		0.0026 (1.0000)	0.4589	0.0508		0.0026 (1.0000)	0.4589	0.0508		0.0026 (1.0000)

¹ The values in parentheses are ratios of MSEs with Linear Terms Only MSE at the bottom.

Table 3.7: Hospital visits data set fitted using the *ES_BMX*, *ES_MMX* and *RS_MMX*

Coefficient	BMX				MXIND				MXMB			
	Estimate	SE	Z-value	P-value	Estimate	SE	Z-value	P-value	Estimate	SE	Z-value	P-value
<i>Poisson component:</i>												
(Intercept)	0.5785	0.2107	2.7457	0.0060	0.6532	0.1835	3.5592	0.0004	0.5347	0.1458	3.6675	0.0002
SexF	0.1425	0.3062	0.4656	0.6415	0.0977	0.2857	0.3420	0.7323	0.2252	0.2546	0.8846	0.3764
SmokeNo	-0.0359	0.3194	-0.1125	0.9104	-0.0222	0.3039	-0.0730	0.9418	0.1200	0.2428	0.4942	0.6211
time	-0.1172	0.1103	-1.0622	0.2882	-0.1400	0.0985	-1.4211	0.1553	-0.0970	0.0736	-1.3168	0.1879
SexF:SmokeNo	-0.3691	0.4534	-0.8140	0.4157	-0.4617	0.4554	-1.0137	0.3107	-0.5670	0.3825	-1.4822	0.1383
SexF:time	-0.1025	0.1504	-0.6812	0.4957	-0.1009	0.1454	-0.6941	0.4876	-0.1192	0.1235	-0.9648	0.3346
SmokeNo:time	-0.1761	0.1790	-0.9837	0.3253	-0.2180	0.1585	-1.3755	0.1690	-0.2304	0.1238	-1.8612	0.0627
SexF:SmokeNo:time	0.8518	0.2321	3.6696	0.0002	0.9413	0.2219	4.2428	0.0000	0.9127	0.1760	5.1873	0.0000
count correlation	0.1183	0.0790			0.0657	0.0658			0.1241	0.0713		
<i>ZI probabilities:</i>												
π_1	0.2083				0.2500				0.2319			
π_2	0.4108				0.4250				0.4329			
π_3	0.3623				0.3852				0.3853			
π_4	0.5846				0.5623				0.6414			
zero correlation	0.4141	0.1121										

Table 3.8: Effects of time on hospital visits rates

Smoke	Sex	
	Male	Female
Yes	0.9076 ($n = 39$)	0.8056 ($n = 14$)
No	0.7209 ($n = 14$)	1.594 ($n = 6$)

Table 3.9: Observed and predicted zeros

Type	Time point				Overall
	1	2	3	4	
Observed	25	39	38	52	154
Predicted (ES_BMX)	25.00	38.97	38.44	52.16	154.57
Predicted (ES_MMX)	26.61	39.34	39.61	51.30	156.87
Predicted (RS_MMX)	26.38	40.09	39.43	54.82	160.71

4. Estimation of marginal ZI regression models in R: The `margZIfit` package

4.1 Introduction

Marginal ZI regression models for clustered data with excess zeros have been studied, but there is limited availability of software packages to implement the model framework. Clustered data with excess zeros are becoming increasingly common in many research areas including health sciences, business, agriculture and manufacturing. ZI regression models assume that the data are generated from a mixture distribution consisting of a degenerate distribution with a point mass of one at zero and a non-degenerate distribution. The EM algorithm is convenient for estimating the model parameter estimates because of the mixture formulation of the model.

In longitudinal studies, or more generally clustered data problems with excess zeros, there is need to develop a model framework that accounts for the association in the data and the excess zeros. ZI mixed effects regression models with random effects that account for cluster-specific associations have been proposed. These models result in parameter estimates that have cluster-specific interpretations. Often, however, we might be interested in population-averaged interpretations. Marginalized ZI regression models with population-averaged interpretations of the model coefficients have been studied and marginal ZI regression models are another framework. Marginal ZI regression models for clustered data with excess zeros have been studied by Hall and Zhang [2004] and others.

Marginal ZI regression models are partially specified models with assumptions made on the marginal means of the component distributions and explanatory variables are added via GLMs for each component. Generally, the likelihood function for marginal ZI regression models is intractable and the EM algorithm is not viable in this case. EM-type algorithms for situations in which the likelihood function is intractable or unavailable have been studied. These include the expectation-solution (ES) algorithm (Rosen et al. [2000]), the expectation-solution (ES) algorithm (Elashoff and Ryan [2004]), the projection-solution (PS) algorithm (Heyde and Morton [1996]) and many others. Hall and Zhang [2004] have extended the ES algorithm of Rosen et al. [2000] to marginal ZI regression models for clustered with excess zeros.

The ES algorithm imputes the missing data assuming independence at the E-step and then relaxes the assumption of independence at the S-step by solving GEEs with working correlation matrices to account for the dependence in the data. The E-step assumes independence for correlated data and this is inefficient. Following Hall and Zhang [2004], Benesi and Hall [2020a] proposed marginal ZI regression models for clustered data with excess zeros assuming a multinomial mixing structure to account for the dependence in the mixing data at both the E-step and S-step of the ES algorithm. However, independence is assumed in the response data and simulations show that there is no much gain in efficiency compared to marginal ZI models wholly assuming independence at the E-step (binary mixing). On another related piece of work, Benesi and Hall [2020b] proposed the regression-solution (RS) algorithm for fitting marginal ZI regression models for clustered data with excess zeros assuming a multinomial mixing structure. The RS algorithm ensures that dependence in the data is accounted for when imputing the data using an appropriate regression model that estimates the multivariate E-step required for clustered data.

In this paper we aim to present a unified implementation of marginal ZI regression models for clustered data with excess zeros in R . The `margZIfit` package is developed to implement three methods for fitting marginal ZI regression models for clustered data with excess zeros. These are the ES algorithm assuming a binary mixing structure (Hall and Zhang [2004]), the ES algorithm assuming a multinomial mixing structure (Benesi and Hall [2020a]) and the RS algorithm assuming a multinomial mixing structure (Benesi and Hall [2020b]). The `margZIfit` function of the `margZIfit` package flexibly implements all the three fitting algorithms in one function.

Section 4.2 gives a brief background on marginal ZI regression models for clustered data with excess zeros. Section 4.3 describes the handling of data, the model fitting function and the interface in detail. In section 4.4 we look at an example by implementing the model fitting algorithms to the given data. Section 4.5 is the summary.

4.2 Background

In this section we shall briefly describe marginal ZI regression models for clustered data with excess zeros and the fitting algorithms implemented by the `margZIfit` function. Our package will fit marginal ZI regression models using the ES algorithm (*ES_BMX*), the ES algorithm assuming multinomial mixing (*ES_MMX*) and the RS algorithm (*RS_MMX*) using the `margZIfit` function.

4.2.1 Marginal ZI regression models: binary mixing structure

Marginal ZI regression models for clustered data with excess zeros are models which are partially specified by assuming a regression model on the marginal means of the mixture distributions. Hall and Zhang [2004] presented the model framework as follows:

Let $\mathbf{y}_i$ be the $n_i \times 1$ vector of responses for the i^{th} cluster, $i = 1, \dots, K$. Assume that the random variable Y_{ij} for the y_{ij} observation follows a marginal ZI distribution, that is, $Y_{ij} \sim 0$, with probability π_{ij} and $Y_{ij} \sim F_2$, with probability $1 - \pi_{ij}$ and mean μ_{ij} . The parameter $\boldsymbol{\mu}_i = (\mu_{i1}, \dots, \mu_{in_i})^T$ and the mixing probabilities $\boldsymbol{\pi}_i = (\pi_{i1}, \dots, \pi_{in_i})^T$ are assumed to be related to covariates through the GLMs

$$g_\mu(\boldsymbol{\mu}_i) = \mathbf{B}_i \boldsymbol{\beta}, \text{ and} \quad (4.1)$$

$$g_p(\boldsymbol{\pi}_i) = \mathbf{G}_i \boldsymbol{\gamma}, \text{ respectively.} \quad (4.2)$$

Here $g_\mu(\cdot)$ is a link function suitable for the response and $g_p(\cdot)$ is a link function such as the logit link function, $\mathbf{B}_i$ and $\mathbf{G}_i$ are respectively the $n_i \times p$ and $n_i \times q$ model matrices. The unknown regression parameters are $\boldsymbol{\beta}$ for the nonzero component and $\boldsymbol{\gamma}$ for the zero component.

If we let $u_{ij} = 1$ if Y_{ij} is from the zero component and $u_{ij} = 0$ otherwise, for $i = 1, \dots, K$, and $j = 1, \dots, n_i$ be the latent indicators of zero-inflation, then under independence within and between the clusters, the ‘complete data’ loglikelihood based on $\mathbf{y} = (\mathbf{y}_1^T, \dots, \mathbf{y}_K^T)^T$ and

$\mathbf{u} = (\mathbf{u}_1^T, \dots, \mathbf{u}_K^T)^T$ is given by

$$\begin{aligned}
l^c(\boldsymbol{\beta}, \phi, \boldsymbol{\gamma}; \mathbf{y}, \mathbf{u}) &= \sum_{i,j} \{u_{ij} \log \pi_{ij}(\boldsymbol{\gamma}) + (1 - u_{ij}) \log [1 - \pi_{ij}(\boldsymbol{\gamma})]\} + \\
&\quad \sum_{i,j} (1 - u_{ij}) \log f_2(y_{ij}; \boldsymbol{\beta}, \phi) \\
&\equiv l^c(\boldsymbol{\gamma}; \mathbf{y}, \mathbf{u}) + l^c(\boldsymbol{\beta}, \phi; \mathbf{y}, \mathbf{u}).
\end{aligned} \tag{4.3}$$

Equation (4.3) can be used to obtain parameter estimates assuming independence between and within clusters. Next, a review of the ES algorithm that modifies the EM algorithm to account for within cluster dependence is briefly described.

4.2.2 Fitting marginal ZI regression models via the ES algorithm

The ES algorithm described by Hall and Zhang [2004] is summarized as follows:

E-step: The E-step is computed as the conditional expectation of the complete data loglikelihood in equation (4.3) given the observed data $\mathbf{y}$ and current parameter estimates. It can be shown that at the $(k + 1)^{\text{th}}$ step of the ES algorithm, the E-step is given by

$$u_{ij}^{(k+1)} = \text{I}\{y_{ij} = 0\} [1 + \{1 - \pi_{ij}(\boldsymbol{\gamma}^{(k)})\} f_2(y_{ij}; \boldsymbol{\beta}^{(k)}, \phi^{(k)}) / \pi_{ij}(\boldsymbol{\gamma}^{(k)})]^{-1},$$

for $i = 1, 2, \dots, K$, $j = 1, 2, \dots, n_i$.

S-step: The classical EM algorithm will proceed by maximizing the imputed complete data loglikelihood and update the $(k + 1)^{\text{th}}$ step estimates for the model. This has been shown to be equivalent to fitting a weighted GLM for the nonzero component and a binomial GLM for the zero component (Hall and Zhang [2004]).

Instead of fitting a pair of GLMs and ignore within-cluster dependence, Hall and Zhang [2004] relaxed the independence assumption and solved a weighted GEE for the non-degenerate component and a binomial GEE for the degenerate component both with working correlation matrices to account for the within-cluster association. Extended GEEs (Hall and Severini [1998]) are used to solve a set of expanded estimating equations at the S-step to estimate the correlation and scale parameters in addition to the regression parameters.

4.2.3 Marginal ZI regression models: multinomial mixing

Hall and Zhang [2004] assumed independence in the mixing structure when computing the E-step. Benesi and Hall [2020a] modified the model framework by assuming that the mixing patterns of zero inflation can be modeled by an underlying multinomial distribution for the patterns of zero-inflation. A multinomial distribution ensures that we account for the dependence in the mixing data and that we have a fully parametric model for the S-step instead of a GEE. Instead of assuming a GLM of the form given in equation (4.2), we assume a multinomial logit model of the form

$$\mathbf{g}_\nu(\boldsymbol{\nu}_i) = \mathbf{G}_i\boldsymbol{\gamma}, \quad (4.4)$$

for the patterns of zero-inflation in the mixing data and a GLM given in equation (4.1). Assuming independence in the non-degenerate component we have the ‘complete data’ log-

likelihood given by

$$\begin{aligned}
l^c(\boldsymbol{\psi}; \mathbf{y}, \tilde{\boldsymbol{\omega}}) &= \sum_{i=1}^K \sum_{r=1}^{R_i} \tilde{\omega}_{ir} \log \nu_{ir}(\boldsymbol{\gamma}) + \sum_{i=1}^K \sum_{r=1}^{R_i} \sum_{j=1}^{n_i} \tilde{\omega}_{ir} \log f(y_{ij} | \boldsymbol{\mu}_{ir}(\boldsymbol{\beta})) \\
&= \sum_{i=1}^K \sum_{r=1}^{R_i} \tilde{\omega}_{ir} \log \nu_{ir}(\boldsymbol{\gamma}) + \sum_{i=1}^K \sum_{j=1}^{n_i} (1 - u_{ij}) \log f(y_{ij} | \boldsymbol{\mu}_{ir}(\boldsymbol{\beta})) \\
&= l^c(\boldsymbol{\gamma}; \tilde{\boldsymbol{\omega}}, \mathbf{y}) + l^c(\boldsymbol{\beta}; \mathbf{y}, \mathbf{u}(\tilde{\boldsymbol{\omega}})).
\end{aligned} \tag{4.5}$$

The likelihood in equation (4.5) can be used to obtain parameter estimates using the EM algorithm assuming independence, or the ES algorithm (Rosen et al. [2000]), or the RS algorithm (Benesi and Hall [2020b]).

4.2.4 Fitting marginal ZI regression models: multinomial mixing via the ES algorithm

E-step: The E-step is computed at each iteration as the conditional expectation of the complete data loglikelihood in equation (4.5) given the observed data $\mathbf{y}$ and current parameter estimates. This can be simplified by computing the conditional multinomial mixing weights $\tilde{\omega}_{ir}^{(k+1)}$ at the $(k+1)^{th}$ step as follows:

$$\tilde{\omega}_{ir}^{(k+1)} = P(\tilde{\omega}_{ir} = 1 | y_{ij}; \boldsymbol{\psi}^{(k)}) = \frac{\nu_{ir}^{(k)} \prod_{j=1}^{n_i} f(y_{ij}; \boldsymbol{\psi}^{(k)}) I(y_{ij} = 0)}{\sum_{l=1}^T \nu_{il}^{(k)} \prod_{j=1}^{n_i} f(y_{ij}; \boldsymbol{\psi}^{(k)})}.$$

S-step: By following Hall and Zhang [2004], we relax the independence assumption and update the parameter estimates by solving an extended weighted GEE with a working correlation matrix to account for the within-cluster dependence for the non-degenerate data and a multinomial logit model to update the parameter estimates for the mixing data.

4.2.5 Fitting marginal ZI regression models: multinomial mixing via the RS algorithm

Following Benesi and Hall [2020a], Benesi and Hall [2020b] proposed the regression-solution (RS) algorithm for fitting marginal ZI regression models for clustered data with excess zeros assuming a multinomial mixing structure. Benesi and Hall [2020b] assumed the GLMs given in equation (4.4) and equation (4.2) for the degenerate component and the non-degenerate component, respectively. The complete data loglikelihood under independence is similar to the one given in equation (4.5). Hall and Zhang [2004] and Benesi and Hall [2020a] assumed independence to obtain closed form solutions for the E-step. Generally, the E-step calculation does not have a closed form solution because the joint multivariate distribution required is intractable. The RS algorithm replaces the E-step of the ES algorithm with a regression step that estimates the conditional expectation of the missing data given the observed data and current parameter estimates. The RS algorithm is described below as follows:

R-step: At the $(k+1)^{th}$ step, compute the conditional mixing weights given the observed data and current parameter estimates, $\tilde{\omega}_{ir}^{(k+1)} \equiv E(\tilde{\omega}_{ir} | \mathbf{y}_i; \boldsymbol{\psi}^{(k)})$, via a regression model on the residuals (Benesi and Hall [2020b]).

S-step: By following Benesi and Hall [2020a], the parameter estimates are updated by fitting a multinomial logit regression model for the mixing data and solving an extended weighted GEE with a working correlation matrix to account for the within-cluster association for the response data.

4.3 Data handling and model description

In this section we present the `margZIfit` package and the `margZIfit` function that can be used to fit marginal ZI regression models for clustered data as described in the previous section. The `margZIfit` function requires data to be in a certain format and the section on data handling outlines the details. We conclude by describing the interface of the `margZIfit` function by exploring the various required, fixed and optional arguments passed to the function.

4.3.1 Loading the package and dependencies

The `margZIfit` package can be downloaded [here](#). After downloading, the package is loaded using the command:

```
> library(margZIfit)
```

The package depends on the following packages: `purrr`, `matrixcalc`, `mlogit`, `R.utils`, `numDeriv` and `gamlss`.

4.3.2 Data handling

The `margZIfit` function requires the data to be in *long format* allowing the user to supply an *id* variable for each cluster and a *time* or measurement occasion variable. The user must transform the data from *wide format* to *long format* using *R* functions such as the `reshape`. We illustrate with an example data set for the number of hospital visits. The data is available in *wide format* and needs to be converted to *long format* using the `reshape` function before it is passed to the `margZIfit` function. The data is available with the `margZIfit` package and can be loaded into *R* using the `data` function as follows:

```

> #Load the data set wide format
> data(nHosp)
> names(nHosp) <-c("ID","Age","Sex","Smoke","Qrt1","Qrt2",
"Qrt3","Qrt4") #colnames
> dim(nHosp) #dim wide format data
[1] 73 8
> head(nHosp) #View the wide data
  ID Age Sex Smoke Qrt1 Qrt2 Qrt3 Qrt4
1  1  63  M   Yes    5    0    0    1
2  2   8  M   Yes    3    4    2    2
3  3  31  M   Yes    0    0    1    2
4  4  33  M   Yes    2    0    1    1
5  5  24  F   Yes    1    0    0    0
6  6  34  F   Yes    2    0    2    0

```

The data has $K = 73$ subjects and 8 columns including the ID variable. The response variable is a vector of counts contained in the time dependent variables `Qrt1`, `Qrt2`, `Qrt3` and `Qrt4`, representing each quarter of the year. We can convert the data into long format as shown below.

```

> # Transform data to Long Format on chosen time points
> nHospLong <- reshape(nHosp, varying = c("Qrt1","Qrt2","Qrt3","Qrt4"),
v.names = "nHospCount", direction = "long")
> dim(nHospLong) # dim long format
[1] 292  7
> head(nHospLong) # view long format data
      ID Age Sex Smoke time nHospCount id
1.1  1  63  M   Yes   1         5  1
2.1  2   8  M   Yes   1         3  2
3.1  3  31  M   Yes   1         0  3
4.1  4  33  M   Yes   1         2  4
5.1  5  24  F   Yes   1         1  5
6.1  6  34  F   Yes   1         2  6

```

Now, as expected we have $73 \times 4 = 292$ observations with the response variable in the column *nHospCount* and additional *time* and *id* variables indicating the measurement occasion and the cluster ID, respectively. The *id* or *ID* variable is passed to the *margZIfit* function with argument *idvar* and the *time* variable is passed with the *timevar* argument.

4.3.3 Model description

The *margZIfit* function uses a **formula** and **data** interface as arguments to specify the model and the user supplied data. The formula interface used by the *margZIfit* function differs from standard *R* formula and has similar interface as the *zeroinfl* function of the **pcsl** package in *R* for fitting ZI regression models for count data in cross-sectional studies. The **formula** argument of the *margZIfit* function is specified with two parts separated with a bar (*|*). The first part of the **formula** argument specifies the regression model for the non-degenerate

component and the second part specifies the regression model for the zero component of the model. The `data` argument allows the user to supply a *long format* data set with an *id* and *time* variables specified via the `idvar` and `timevar` arguments, respectively. A help page that describes the *margZIfit* function is provided at the end of this chapter.

The additional arguments are `method`, `errDist`, `mulink`, `plink`, `nophi`, `corstruct` and `mixcorstr`. The mandatory arguments to the function are `method`, `errDist` and `mulink` which are passed to the function as *R* strings. The `method` argument specifies one of the three fitting algorithms `ES_BinMix`, `ES_MultiMix` and `RS_MultiMix` for the ES algorithm assuming a binary mixing structure, the ES algorithm assuming a multinomial mixing structure and the RS algorithm assuming a multinomial mixing structure, respectively. The default `method` for fitting is `ES_BinMix`. The mandatory argument `errDist` specifies the underlying distribution for the non-degenerate component. Currently, the function only supports the “poisson” and the “binomial” distributions and others will be added with updates. The default `errDist` for is “poisson”.

The argument `mulink` specifies the desired link function for the `errDist` distribution selected. The list of supported `mulink` functions are given in the model help page. The default `mulink` is “log” for the the default `errDist`. The `plink` argument is the link function for the zero component and it is fixed to a default value of “logit”, however, the user can specify the “probit” link function if so desired. The argument `nophi` is a logical argument which specifies whether a dispersion parameter ϕ is desired. If `nophi` is `TRUE`, the default value, then the dispersion parameter ϕ is fixed to 1; Otherwise it is taken to be a parameter to be estimated. The arguments `corstruct` and `mixcorstr` describe the correlation structures for the non-degenerate distribution and the mixing distribution, respectively. The list of correlation structures that are supported are given in the model description above. Both `corstruct` and `mixcorstr` are fixed to the default value “indep”.

The function *margZIfit* creates an object which can be passed to other generic functions such as the *summary*, *predict*, *coef*, *resid* and *fitted* functions. In the next section we shall present an example in which we use the *margZIfit* function to fit models to an example dataset.

4.4 Example

The hospital visits data set consists of $K = 73$ children followed over time to observe the number of hospital visits made during the four quarters of the year (Karim and Zeger [1988]). For each child, the response variable are counts for the number of hospital visits the child made over the year. There is evidence of excess zeros in the data for the predicted zeros from a regular Poisson GLM were fewer than the observed zeros in the data (Benesi and Hall [2020a]). The potential predictors are gender, smoking status, and time. Multi-way interactions are included to model more complex relationships among the predictors. We shall use the *margZIfit* function to fit our data with the three fitting algorithms available. The response variable are counts in each cluster and hence we assume the `errDist` to be the Poisson distribution with a corresponding link function `mulink="log"`. The link function for the degenerate component, `plink`, will be fixed to "logit" the default value (note that this argument need not be specified for multinomial mixing models). The working correlation matrices for both components of the model are assumed to be exchangeable and specified as `construct="exch"` and `mixcorstr="exch"` (note that the `mixcorstr` argument need not be specified for multinomial mixing models).

4.4.1 Marginal ZI regression models: binary mixing structure

We shall fit our data by specifying `method="ESBinMix"` which performs the ES algorithm used by Hall and Zhang [2004]. The function call is shown below:

```
> fit <- margZIfit(nHospCount ~ Sex*Smoke *time |factor(time),
                  data = nHospLong, idvar = "id", timevar = "time",
                  method = "ESBinMix", mulink = "log", errDist = "poisson",
                  corstruct = "exch", mixcorstr = "exch", nophi = TRUE)
```

The object `fit` can be supplied to the generic function *summary* to summarize the model fit. This is shown below as follows:

Call:

```
margZIfit.formula(formula = nHospCount ~ Sex * Smoke * time |
                  factor(time), data = nHospLong, idvar = "id", timevar = "time",
                  method = "ESBinMix", mulink = "log", errDist = "poisson",
                  corstruct = "exch", mixcorstr = "exch", nophi = TRUE)
```

Count model coefficients

	Estimate	StdErr	z.value	p.value
(Intercept)	0.578526	0.210700	2.7457	0.0060375 **
SexF	0.142536	0.306162	0.4656	0.6415325
SmokeNo	-0.035937	0.319401	-0.1125	0.9104148
time	-0.117197	0.110338	-1.0622	0.2881606
SexF:SmokeNo	-0.369061	0.453413	-0.8140	0.4156662

```

SexF:time      -0.102482  0.150433 -0.6812 0.4957149
SmokeNo:time   -0.176069  0.178983 -0.9837 0.3252524
SexF:SmokeNo:time 0.851841  0.232135  3.6696 0.0002429 ***

```

```

Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1 1

```

Zero-Inflation model coefficients

	Estimate	StdErr	z.value	p.value
(Intercept)	0.34171	0.36755	0.9297	0.35254
factor(time)0	-1.67718	0.67693	-2.4776	0.01323 *
factor(time)1	-0.70238	0.51386	-1.3669	0.17166
factor(time)2	-0.90720	0.40345	-2.2486	0.02454 *

```

Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1 1

```

Correlation coefficients

	Estimate	StdErr
count_correlation	0.11825	0.0790
zero_correlation	0.41414	0.1121

The number of iterations: 17

Convergence: TRUE

Dispersion parameter taken to be 1

The *summary* function is divided into four parts which are the *call* to the function, the *count model coefficients* for summarizing the regression estimates for the non-degenerate component, the *zero inflation model coefficients* for the zero component regression estimates, the *correlation coefficients* for the estimated correlation parameters and the last part provides a summary of the model fit such as the number of iterations, convergence and the dispersion parameter. The user can compute other quantities from the *margZIfit* object such as the probabilities of zero-inflation. Below is an outline of how the user can obtain the probabilities of zero-inflation.

```
> # Computing ZI probabilities from model coefficients
> zero_fit <- fit$zero_coefficients
> pi_hat <- c(expit(zero_fit[1]+zero_fit[-1]),expit(zero_fit[1]))
> names(pi_hat) <- c("pi1_hat","pi2_hat","pi3_hat","pi4_hat")
> pi_hat
      pi1_hat      pi2_hat      pi3_hat      pi4_hat
0.2082559 0.4107961 0.3622773 0.5846052
```

From the above output, the function *expit* is the inverse logit function that is available in the *margZIfit* package. The user can use other available inverse logit functions available in *R*. The ZI probabilities are computed for each time point in the model we have formulated.

4.4.2 Marginal ZI regression models: multinomial mixing structure (ES algorithm)

The *margZIfit* function can fit marginal ZI regression models assuming multinomial mixing using the ES algorithm by simply specifying `method="ESMultiMix"` argument and other required arguments. To fit a marginal ZI regression model with multinomial mixing, the syntax to call the function is given below:

```
> fit2 <- margZIfit(nHospCount ~ Sex*Smoke*time | 1,
                    data = nHospLong, idvar = "id", timevar = "time",
                    method = "ESMultiMix", mulink = "log", errDist = "poisson",
                    corstruct = "exch", nophi = TRUE)
```

The object `fit2` can similarly be summarized using the *summary* function as follows:

```
> summary(fit2)
```

Call:

```
margZIfit.formula(formula = nHospCount ~ Sex * Smoke * time |
  1, data = nHospLong, idvar = "id", timevar = "time",
  method = "ESMultiMix", mulink = "log", errDist = "poisson",
  corstruct = "exch", nophi = TRUE)
```

Count model coefficients

	Estimate	StdErr	z.value	p.value
(Intercept)	0.653211	0.183527	3.5592	0.000372 ***
SexF	0.097717	0.285694	0.3420	0.732327
SmokeNo	-0.022171	0.303900	-0.0730	0.941843

time	-0.139962	0.098491	-1.4211	0.155297
SexF:SmokeNo	-0.461673	0.455445	-1.0137	0.310739
SexF:time	-0.100949	0.145443	-0.6941	0.487633
SmokeNo:time	-0.218034	0.158514	-1.3755	0.168979
SexF:SmokeNo:time	0.941327	0.221866	4.2428	2.208e-05 ***

Signif. codes: 0 *** 0.001 ** 0.01 * 0.05 . 0.1 1

Zero-Inflation model coefficients

	Estimate	StdErr	z.value	p.value
1:(intercept)	1.311344	0.888694	1.4756	0.1401
2:(intercept)	0.690911	1.066824	0.6476	0.5172
3:(intercept)	-20.036990	0.861947	-23.2462	< 2.2e-16 ***
4:(intercept)	0.551937	0.944348	0.5845	0.5589
5:(intercept)	0.878162	0.940036	0.9342	0.3502
6:(intercept)	-0.038189	1.224173	-0.0312	0.9751
7:(intercept)	-19.551280	2.455111	-7.9635	1.672e-15 ***
8:(intercept)	0.604690	1.013539	0.5966	0.5508
9:(intercept)	0.252315	1.031900	0.2445	0.8068
10:(intercept)	-1.213295	1.664336	-0.7290	0.4660
11:(intercept)	-0.407424	1.442444	-0.2825	0.7776
12:(intercept)	-19.761260	2.154800	-9.1708	< 2.2e-16 ***
13:(intercept)	-19.761676	1.955789	-10.1042	< 2.2e-16 ***
14:(intercept)	-0.036629	1.247356	-0.0294	0.9766

Signif. codes: 0 *** 0.001 ** 0.01 * 0.05 . 0.1 1

Correlation coefficients

```
Estimate StdErr
count_correlation 0.065679 0.0658
```

The number of iterations: 155

Convergence: TRUE

Dispersion parameter taken to be 1

The output is similar to the one described above except that the zero-inflation parameters are from a multinomial logit model. It is important to note that the zero-inflation part for multinomial mixing models can currently fit an *intercept only* multinomial logit model. Future releases will aim to include covariate effects and the use of parsimonious log-linear models instead of multinomial logit models. In this example, we have four time points which corresponds to $R = 16$ different time points. We should expect $R - 1 = 15$ intercepts, but we've got 14 intercepts in the output. This happens because one of the patterns of zero-inflation is not observed in the data and is eliminated to avoid model fitting problems. This can be shown below:

```
> # Possible patterns of ZI inflation
> Y0 <- as.numeric(nHospLong$nHospCount==0)
> # Create a matrix of indicators
> Y0Mat <- matrix(Y0, nrow = 73, ncol = 4, byrow = FALSE)
> # Check counts in each pattern
> (Y0Tab <- ftable(as.data.frame(Y0Mat), col.vars = 1:4))

V1 0 1
```

```

V2  0          1          0          1
V3  0      1      0      1      0      1      0      1
V4  0  1  0  1  0  1  0  1  0  1  0  1  0  1  0  1

```

```

7  8  1  7  6  7  2 10  0  5  1  5  1  1  3  9

```

From the above output we observe that the 9th class which corresponds to observing a zero at time point one followed by nonzero counts is not an observable pattern in the data. This class can be eliminated and proceed with fitting the model. This alters the model interpretation of the zero component as the classes have changed, however, the quantities of interest are often the zero inflation probabilities and the regression effects of the non-degenerate model, are not affected. With a little bit of work, we can get the ZI probabilities as shown in the code below:

```

> # ZI probability at each time point
> # Identify unobserved classes in data
> Noclass <- which(Y0Tab==0)
> R <- 2^4 #The number of classes
> # Adjust the number of classes
> len_Noclass <- length(Noclass); R2 <- R-len_Noclass
> # Extract the model coefficient
> zero_fit2 <- fit2$zero_coefficients
> # Compute the inverse generalized logit
> piMat_fit2 <- getmultip(zero_fit2)
> piMat_fit2 <- c(piMat_fit2, 1-sum(piMat_fit2))
> # Initialize the ZI probs at each time
> names(piMat_fit2) <- intToBin(0:(R-1))[-Noclass] # 1 denotes an excess zero
> probMat <- matrix(0, nrow = 1 ,ncol=4) # matrix of zero inflation probs at each time p

```

```

> for(h in 1:4){
+   probMat[,h] <- sum(piMat_fit2[substr(names(piMat_fit2),h,h)=="1"])
+ }
>
> # ZI probs at each time point
> colnames(probMat) <- c("pi1_hat","pi2_hat","pi3_hat","pi4_hat")
> probMat
      pi1_hat  pi2_hat  pi3_hat  pi4_hat
[1,] 0.249968 0.4249801 0.3852443 0.5622558

```

To get the zero-inflation probabilities, we begin by eliminating the patterns of zero-inflation that are not observed in the data. We then transform the model coefficients for the zero component using the *getmultip* function available in the **margZIfit** package. The *getmultip* function computes the inverse generalized logits of the coefficients. We then compute the marginal probabilities by summing the probabilities respectively for each time point to get the ZI probabilities.

4.4.3 Marginal ZI regression models: multinomial mixing structure (RS algorithm)

The *margZIfit* function can fit marginal ZI regression models with multinomial mixing structure using the RS algorithm by specifying the argument `method="RSMultiMix"`. The call to the function is shown below:

```

> fit3 <- margZIfit(nHospCount ~ Sex  Smoke  time |1,
      data = nHospLong, idvar = "id", timevar = "time",
      method = "RSMultiMix", mulink = "log", errDist = "poisson",
      corstruct = "exch", nophi = TRUE)

```

Similarly, the object `fit3` can be summarized using the *summary* function as follows:

```
> summary(fit3)
```

Call:

```
margZIfit.formula(formula = nHospCount ~ Sex * Smoke * time |
  1, data = nHospLong, idvar = "id", timevar = "time",
  method = "RSMultiMix", mulink = "log", errDist = "poisson",
  corstruct = "exch", nophi = TRUE)
```

Count model coefficients

	Estimate	StdErr	z.value	p.value	
(Intercept)	0.534651	0.145780	3.6675	0.0002449	***
SexF	0.225178	0.254564	0.8846	0.3763936	
SmokeNo	0.120022	0.242850	0.4942	0.6211472	
time	-0.096970	0.073642	-1.3168	0.1879122	
SexF:SmokeNo	-0.567003	0.382548	-1.4822	0.1382935	
SexF:time	-0.119197	0.123546	-0.9648	0.3346421	
SmokeNo:time	-0.230350	0.123767	-1.8612	0.0627216	.
SexF:SmokeNo:time	0.912723	0.175952	5.1873	2.133e-07	***

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Zero-Inflation model coefficients

	Estimate	StdErr	z.value	p.value
1:(intercept)	0.883997	0.484419	1.8249	0.0680221

```

2:(intercept)    0.565115   0.497483   1.1360  0.2559775
3:(intercept)   -1.992454   0.540013  -3.6896  0.0002246 ***
4:(intercept)    0.226020   0.524283   0.4311  0.6663938
5:(intercept)    0.432027   0.494189   0.8742  0.3820016
6:(intercept)    0.088675   0.514287   0.1724  0.8631048
7:(intercept)   -1.273503   0.605962  -2.1016  0.0355866 *
8:(intercept)    0.325073   0.494817   0.6570  0.5112094
9:(intercept)   -0.025968   0.523762  -0.0496  0.9604573
10:(intercept)  -2.868064   0.493169  -5.8156  6.042e-09 ***
11:(intercept)  -0.378918   0.502731  -0.7537  0.4510178
12:(intercept)  -3.959820   0.433348  -9.1377 < 2.2e-16 ***
13:(intercept)  -2.402066   0.591593  -4.0603  4.900e-05 ***
14:(intercept)  -1.870312   0.318805  -5.8666  4.447e-09 ***

```

Signif. codes: 0 *** 0.001 ** 0.01 * 0.05 . 0.1 1

Correlation coefficients

Estimate StdErr

count_correlation 0.12414 0.0713

The number of iterations: 9

Convergence: TRUE

Dispersion parameter taken to be 1

The output is similar to the others shown above. Again we notice that the zero-inflation coefficients have a missing unobserved class. Similarly, we can get the probabilities of zero-inflation as outlined in the previous subsection. Below we won't show the operations of eliminating the unobserved classes and proceed as if the user has done so. The operations are similar to the previous case and proceeds as follows:

```
> #Zero-inflation probabilities
> zero_fit3 <- fit3$zero_coefficients
> piMat_fit3 <- getmultip(zero_fit3)
> piMat_fit3 <- c(piMat_fit3, 1-sum(piMat_fit3))
> names(piMat_fit3) <- intToBin(0:(R-1))[-Noclass] # 1 denotes an excess zero
> # matrix of zero inflation probs at each time point
> probMat <- matrix(0, nrow = 1 ,ncol=4)
> for(h in 1:4){
+   probMat[,h] <- sum(piMat_fit3[substr(names(piMat_fit3),h,h)=="1"])
+ }
> # Overall ZI probability
> colnames(probMat) <- c("pi1_hat","pi2_hat","pi3_hat","pi4_hat")
> probMat
      pi1_hat  pi2_hat  pi3_hat  pi4_hat
[1,] 0.2319234 0.4328549 0.3852684 0.6413792
```

4.5 Summary

In this chapter we have shown the functionality and interface of the *margZIfit* function for fitting marginal ZI regression models for clustered data with excess zeros. The *margZIfit* function can fit the ES algorithm for binary mixing models as well as the ES and RS algorithms for multinomial mixing models in one function. We applied the *margZIfit* function to an example data set to demonstrate its functionality and how to use the resulting model object to compute zero inflation probabilities. The *margZIfit* package comes along with other useful functions, such as the *expit* for computing the inverse logit and the *getmultip* for computing the generalized inverse logit.

The *margZIfit* function implements the intercept only model for the zero component of the multinomial mixing models. Future work will involve extending the functionality of the function by adding more models, link functions and modeling parsimonious multinomial mixing models with covariate effects using flexible log-linear models. The user should take caution when interpreting the zero component model coefficients of the multinomial mixing models. The model interpretation may change and becomes altered if some patterns of ZI classes are not observed in the data. However, the non-degenerate model coefficients and the fitted ZI probabilities are not altered. It is recommended that the user check if all the classes of the patterns of ZI are observed in the data and rather quantify the fitted ZI probabilities.

Formula interface for fitting marginal ZI regression models

Description

Formula interface for fitting marginal ZI regression models

Usage

```
## S3 method for class 'formula'
margZIfit(formula, data, idvar, timevar, method, mulink,
  plink = "logit", errDist, mcorstruct = "indep", mixcorstr = "indep", nophi =
  TRUE)
```

Arguments

<code>formula</code>	an object of class <code>Formula</code> with two parts, the first part for the non-degenerate, second part degenerate e.g. $y \sim 1 + a \mid 1$.
<code>data</code>	a data set to be used.
<code>idvar</code>	a cluster specific ID variable.
<code>timevar</code>	a time or measurement occasion variable.
<code>method</code>	a method for fitting, one of "ES_BinMix" (ES binary mixing), "ES_MultiMix" (ES multinomial mixing), "RS_MultiMix" (RS multinomial mixing). Default = "ES_BInMix"
<code>mulink</code>	a link function for the non-degenerate component one of "log", "sqrt", "logit", "probit", "identity", "cloglog", "inverse". Default = "log".
<code>plink</code>	a link function for the degenerate component one of "logit" and "probit". Default = "logit".
<code>errDist</code>	a distribution for the non-degenerate component one of "poisson" or "binomial". Default = "poisson"
<code>corstruct</code>	a correlation structure for the non-degenerate component one of "indep", "exch", "ar1", "unspec", "onedep", "toep3". Default is "indep".
<code>mixcorstr</code>	a correlation structure for the degenerate component one of "indep", "exch", "ar1", "unspec", "onedep", "toep3". Default is "indep".
<code>nophi</code>	a logical for whether the dispersion parameter is to be estimated or not, default = FALSE.

Value

An object of class `margZIfit`.

<code>iterations</code>	number of iterations.
<code>converge</code>	a logical of whether convergence was attained.
<code>count_coefficients</code>	regression estimates for non-degenerate component
<code>zero_coefficients</code>	regression estimates for the degenerate component
<code>count_correlation</code>	estimate(s) of the working correlation structure for the non-degenerate component.
<code>correlation_coefficients</code>	estimate(s) of the working correlation structure (all components).
<code>correlation_StdErr</code>	standard errors for all the correlation parameters
<code>zero_correlation</code>	estimate(s) of the working correlation structure for the degenerate component.
<code>phi</code>	estimate of the dispersion parameter.
<code>coefficients</code>	estimates of all model parameters.
<code>VarCov</code>	the variance covariance matrix of parameter estimates.
<code>count_StdErr_fit</code>	standard errors for the non-degenerate model parameter estimates.
<code>zero_StdErr_fit</code>	standard errors for the degenerate model parameter estimates.

[Package *margZIfit* version 0.1.0]

5. Summary

5.1 Summary and conclusions

This dissertation develops and investigates EM-type algorithms that are used for fitting marginal ZI regression models for clustered data with excess zeros. The EM algorithm can be used to estimate the parameter estimates if the likelihood is tractable. The likelihood function for marginal ZI regression models is generally intractable and hence there is need to find alternative EM-type algorithms to fit the models. The expectation-solution (ES) algorithm is an EM-type algorithm for fitting models for clustered data. To update the parameter estimates, the ES algorithm modifies the M-step of the EM algorithm by fitting GEEs with working correlation matrices to account for the dependence in the data instead of maximizing the imputed complete data loglikelihood function. The E-step of the ES algorithm is computed under the assumption of independence to obtain a closed form solution and for clustered data this is known to be inefficient.

Marginal ZI regression models can be modified by assuming a multinomial mixing structure instead of the binary mixing structure (independence), to account for the dependence in the mixing data. In this framework, the complete data quasi-loglikelihood is obtained by assuming independence in the response data and the ES algorithm can be used to estimate the model parameter estimates. However, simulations showed that there is no much gain in efficiency between the multinomial mixing models and the binary mixing models. The regression-solution algorithm is proposed and it is robust than the ES algorithm in accounting for the dependence in the data for multinomial mixing models. The RS algo-

rithm estimates the complex conditional expectation of the missing data given the observed data and current parameter estimates at the E-step using an appropriate regression model. Overall, simulations show that the RS algorithm provides estimators which are more efficient than the ES algorithm used in fitting the multinomial and binary mixing models.

Lastly, we developed the `margZIfit` package in *R* for fitting marginal ZI regression models using the ES algorithm for binary mixing models and the ES and RS algorithms for multinomial mixing models. The `margZIfit` function is very flexible and with an easy to use interface. The `margZIfit` function fits marginal ZI regression models to data using all the three fitting algorithms in one function. We illustrated the functionality of the `margZIfit` function using a clustered data with excess zeros for the number of hospital visits of subjects who were followed in each of the four quarters in a year. The three fitting algorithms were used to find the parameters estimates and the inferences on the non-degenerate regression model were roughly similar for all the three methods.

5.2 Recommendations and future work

The number of parameters that need to be estimated in multinomial mixing models are generally greater than that from binary mixing models. That is, the number of regression parameters for the zero component of the multinomial mixing models increases with increasing number of time points or measurement occasions. Future research direction should focus on developing more parsimonious multinomial logit models by either imposing constraints on the parameter estimates or using corresponding log-linear models that constrain the model parameters more flexibly. Some log-linear models are equivalent to multinomial logit models and are very flexible in restricting the model complexity than multinomial logit models.

Benesi and Hall [2020a] and Benesi and Hall [2020b] have briefly described on the possible regression models that can be formulated for multinomial mixing models when the data is

unbalanced or the cluster sizes are too large. Regression models are suggested for the S-step as well as the R-step of the RS algorithm. The success of such approaches relies on the form of the regression that will be used, and how such a model can be determined and improved is an area for further investigation.

Bibliography

- A. Agresti. *Categorical Data Analysis*. Wiley series in probability and mathematical statistics, 1990.
- T. Benesi and D. B. Hall. Marginal zero-inflated regression models for clustered data with multinomial mixing structure. *Working paper*, 2020a.
- T. Benesi and D. B. Hall. A Regression-Solution algorithm for marginal zero-inflated models with multinomial mixing. *Working paper*, 2020b.
- Yves Croissant. *Multinomial logit models*, 2019. URL <https://cran.r-project.org/web/packages/mlogit/mlogit.pdf>.
- M. J. Dobbie and A. H. Welsh. Modelling correlated zero-inflated count data. *Australian and New Zealand Journal of Statistics*, 43:431–44, 2001.
- M. Elashoff and L. Ryan. An EM algorithm for estimating equations. *Journal of Computational and Graphical Statistics*, 13:48–65, 2004.
- Keith Goldfeld. *Simulation of study data*, 2019. URL <https://cran.r-project.org/web/packages/simstudy/simstudy.pdf>.
- Masahiko Goshō. Criteria to select a working correlation structure for Generalized Estimating Equations methods in SAS. *Journal of Statistics Software*, 57, 2014.
- D. B. Hall. Zero-inflated poisson and binomial regression with random effects: a case study. *Biometrics*, 56:1030–1039, 2000.

- D. B. Hall and T. Severini. Extended generalized estimating equations for clustered data. *Journal of the American Statistical Association*, 93:1365–75, 1998.
- D. B. Hall and J. Shen. Marginal projected multivariate linear models for clustered angular data. *Australian and New Zealand Journal of Statistics*, 57:241–257, 2015.
- D. B. Hall and Z. Zhang. Marginal models for zero inflated clustered data. *Statistical Modeling*, 4:161–180, 2004.
- D. C. Heilbron. Zero-altered and other regression models for count data with added zeros. *Biometrical Journal*, 36:531–47, 1994.
- C. C. Heyde and R. Morton. Quasi-likelihood and generalizing the EM algorithm. *Journal of the Royal Statistical society*, B:317–327, 1996.
- M. R. Karim and S. L. Zeger. GEE: A SAS macro for longitudinal data analysis. *Technical report no. 674: The John Hopkins University*, 1988.
- W. Kassahun, T. Neyens, G. Molenberghs, C. Faes, and G. Verbeke. Marginalized multilevel hurdle and zero-inflated models for overdispersed and correlated count data with excess zeros. *Stat Med*, 33(25):4402–4419, 2014. doi: 10.1002.6237.
- D. Lambert. Zero-inflated poisson regression, with an application to defects in manufacturing. *Technometrics*, 34:1–14, 1992.
- K. Y. Liang and S. Zeger. Longitudinal data analysis using generalised linear models. *Biometrika*, 73:13–22, 1986.
- D. L. Long, J. S. Preisser, A. H. Herring, and C. E. Golin. A marginalized zero-inflated poisson regression model with overall exposure effects. *Stat. Med.*, 33:5151–5165, 2014.

- Daniel Ludecke. *Assessment of regression models performance*, 2019. URL <https://cran.r-project.org/web/packages/performance/performance.pdf>.
- J. Martin and D. B. Hall. Marginal zero-inflated regression models for count data. *Journal of Applied Statistics*, 44:10:1807–1826, 2017. doi: 10.1080/02664763.2016.1225018.
- L. H. Moulton, F. C. Curriero, and P. F. Barroso. Mixture models for quantitative HIV RNA data. *Statistical Methods in Medical Research*, 11:317–25, 2002.
- W. Pan. Akaike’s information criterion in Generalized Estimating Equations. *Biometrics*, 57(1):120–125, 2001.
- J. S. Preisser, K. Das, D. L. Long, and K. Divaris. Marginalized zero-inflated negative binomial regression with application to dental caries. *Stat. Med.*, 35:1722–1735, 2016.
- M. Ridout, J. Hinde, and C. G. B. Demetrio. A score test for testing a zero-inflated poisson regression model against zero-inflated negative binomial alternatives. *Biometrics*, 57:219–23, 2001.
- O. Rosen, W. Jiang, and M. A. Tanner. Mixtures of marginal models. *Biometrika*, 87:391–404, 2000.
- G. A. Satten and S. Datta. The SU algorithm for missing data problems. *Computational Statistics*, 15:243–277, 2000.
- K. K. W. Yau and A. H. Lee. Zero-inflated poisson regression with random effects to evaluate an occupations injury prevention programme. *Statistics in Medicine*, 20:2907–20, 2001.
- Z. Zhang. *Marginal models for zero-inflated clustered data*. PhD thesis, University of Georgia, 2004.
- H. Zhao, W. Ma, and J. Guo. The AU algorithm for estimating equations in the presence missing data. *Statistics & Probability Letters*, 80:639–647, 2010.

Appendix A

Table 5.1: A simulation study comparing binary mixing and multinomial mixing structures (correlation parameter $\rho = 0.25$)

Correlation Structure	Binary Mixing				Multinomial Mixing			
	Mean	SE	SD	MSE	Mean	SE	SD	MSE
<i>Exchangeable: n = 50</i>								
β_0	2.1163	0.0503	0.0518	0.0027	2.1161	0.0504	0.0519	0.0027
β_1	0.4322	0.0417	0.0481	0.0023	0.4324	0.0418	0.0481	0.0023
β_2	-0.4903	0.0420	0.0484	0.0023	-0.4904	0.0420	0.0484	0.0023
ρ	0.2371	0.3159	0.4192	0.1759	0.2495	0.3187	0.4183	0.1749
π_1	0.4036		0.0701	0.0049	0.4034		0.0701	0.0049
π_2	0.4045		0.0743	0.0055	0.4043		0.0743	0.0055
<i>AR(1): n = 50</i>								
β_0	2.1171	0.0504	0.0552	0.0031	2.1170	0.0504	0.0551	0.0030
β_1	0.4302	0.0421	0.0486	0.0024	0.4303	0.0422	0.0485	0.0024
β_2	-0.4896	0.0426	0.0485	0.0024	-0.4897	0.0426	0.0485	0.0023
ρ	0.2636	0.3279	0.4142	0.1718	0.2732	0.3353	0.4153	0.1730
π_1	0.4047		0.0695	0.0048	0.4046		0.0694	0.0048
π_2	0.4040		0.0730	0.0053	0.4040		0.0730	0.0053
<i>Exchangeable: n = 200</i>								
β_0	2.1201	0.0257	0.0260	0.0007	2.1201	0.0257	0.0260	0.0007
β_1	0.4304	0.0215	0.0226	0.0005	0.4305	0.0215	0.0226	0.0005
β_2	-0.4907	0.0215	0.0222	0.0005	-0.4907	0.0215	0.0222	0.0005
ρ	0.2457	0.1506	0.1619	0.0262	0.2518	0.1516	0.1630	0.0266
π_1	0.4066		0.0348	0.0012	0.4065		0.0348	0.0012
π_2	0.4038		0.0338	0.0011	0.4037		0.0338	0.0011
<i>AR(1): n = 200</i>								
β_0	2.1205	0.0257	0.0257	0.0007	2.1204	0.0257	0.0257	0.0007
β_1	0.4294	0.0215	0.0226	0.0005	0.4294	0.0215	0.0227	0.0005
β_2	-0.4905	0.0214	0.0213	0.0005	-0.4906	0.0214	0.0213	0.0005
ρ	0.2556	0.1502	0.1585	0.0252	0.2618	0.1510	0.1595	0.0256
π_1	0.4076		0.0357	0.0013	0.4076		0.0357	0.0013
π_2	0.4045		0.0334	0.0011	0.4045		0.0334	0.0011

Table 5.1 shows the various properties of the estimators considered at the low level of correlation $\rho = 0.25$. As expected the results improve with increasing sample size. The results for the multinomial mixing and the binary mixing do not indicate any strong differences in the properties of the estimators considered. The Monte Carlo standard deviations (SD) are quite close to the average standard errors (SE) from the model fitting and hence

the sandwich estimator for the variance-covariance matrix is a reasonable estimator of the variance covariance matrix. The AR(1) and Exchangeable results are quite close because this is expected in the case of two time points.

Table 5.2: A simulation study comparing binary mixing and multinomial mixing structures (Test(5%) and Coverage(95%): Correlation parameter $\rho = 0.25$)

Correlation Structure	Binary Mixing		Multinomial Mixing	
	Test Size	Coverage	Test Size	Coverage
<i>Exchangeable: n = 50</i>				
β_0	0.0600	0.9400	0.0600	0.9400
β_1	0.1010	0.8990	0.0100	0.9000
β_2	0.1010	0.8990	0.1010	0.8990
ρ	0.1660	0.8340	0.1630	0.8370
<i>AR(1): n = 50</i>				
β_0	0.0780	0.9220	0.0780	0.9220
β_1	0.1060	0.8940	0.1050	0.8950
β_2	0.1010	0.8990	0.0980	0.9020
ρ	0.1640	0.8360	0.1610	0.8390
<i>Exchangeable: n = 200</i>				
β_0	0.0540	0.9460	0.0560	0.9440
β_1	0.0700	0.9300	0.0700	0.9300
β_2	0.0630	0.9370	0.0630	0.9370
ρ	0.094	0.9060	0.0870	0.9130
<i>AR(1): n = 200</i>				
β_0	0.0530	0.9470	0.0510	0.9490
β_1	0.0740	0.9260	0.0740	0.9260
β_2	0.0590	0.9410	0.0570	0.9430
ρ	0.0840	0.9160	0.0810	0.9190

Table 5.2 shows the test size and the coverage of both methods at the low level of correlation $\rho = 0.3$. The results as expected show improvement with increasing sample size especially for the correlation parameter and the results are slightly better for the multinomial mixing mechanism compared to the binary mixing mechanism.

Table 5.3: A simulation study comparing binary mixing and multinomial mixing structures (correlation parameter $\rho = 0.75$)

Correlation Structure	Binary Mixing				Multinomial Mixing			
	Mean	SE	SD	MSE	Mean	SE	SD	MSE
<i>Exchangeable: n = 50</i>								
β_0	2.1157	0.0556	0.0570	0.0033	2.1156	0.0556	0.0569	0.0033
β_1	0.4289	0.0461	0.0556	0.0031	0.4289	0.0461	0.0556	0.0031
β_2	-0.4922	0.0468	0.0513	0.0026	-0.4923	0.0468	0.0512	0.0026
ρ	0.7722	0.2483	0.2577	0.0669	0.7819	0.2497	0.2586	0.0680
π_1	0.4057		0.0701	0.0049	0.4056		0.0702	0.0049
π_2	0.4036		0.0696	0.0049	0.4035		0.0696	0.0049
<i>AR(1): n = 50</i>								
β_0	2.1151	0.0555	0.0601	0.0036	2.1151	0.0555	0.0600	0.0036
β_1	0.4305	0.0463	0.0535	0.0029	0.4305	0.0463	0.0534	0.0029
β_2	-0.4902	0.0466	0.0524	0.0027	-0.4903	0.0466	0.0524	0.0028
ρ	0.7706	0.2480	0.2630	0.0696	0.7802	0.2492	0.2651	0.0712
π_1	0.4076		0.0701	0.0049	0.4076		0.0700	0.0049
π_2	0.4061		0.0703	0.0049	0.4061		0.0703	0.0049
<i>Exchangeable: n = 200</i>								
β_0	2.1203	0.0285	0.0288	0.0008	2.1202	0.0285	0.0287	0.0008
β_1	0.4298	0.0233	0.0237	0.0006	0.4298	0.0233	0.0237	0.0006
β_2	-0.4896	0.0233	0.0236	0.0006	-0.4896	0.0233	0.0236	0.0006
ρ	0.7403	0.1315	0.1316	0.0174	0.7482	0.1321	0.1322	0.0175
π_1	0.4073		0.0351	0.0012	0.4072		0.0352	0.0012
π_2	0.4051		0.0350	0.0012	0.4051		0.0350	0.0012
<i>AR(1): n = 200</i>								
β_0	2.1187	0.0284	0.0291	0.0008	2.1186	0.0284	0.0291	0.0008
β_1	0.4307	0.0234	0.0242	0.0006	0.4307	0.0234	0.0242	0.0006
β_2	-0.4907	0.0233	0.0249	0.0006	-0.4907	0.0233	0.0248	0.0006
ρ	0.7398	0.1311	0.1298	0.0170	0.7477	0.1314	0.1306	0.0171
π_1	0.4066		0.0355	0.0013	0.4066		0.0356	0.0013
π_2	0.4066		0.0351	0.0012	0.4066		0.0350	0.0012

Table 5.3 shows the results from both methods computed at a high level of correlation, $\rho = 0.75$. The results for both approaches show an improvement in the properties of the estimators as the sample sizes increase. As in the low level of correlation results, there also appears to be no strong differences between the properties of the estimators for both the approaches and the average standard errors (SE) from the sandwich estimator are comparably close to the Monte Carlo standard deviations (SD) indicating the estimator is reasonable. Again, the AR(1) and the Exchangeable results are quite close as expected in the two time points case.

Table 5.4: A simulation study comparing binary mixing and multinomial mixing structures (Test size (5%) and Coverage(90%): correlation parameter $\rho = 0.75$)

Correlation Structure	Binary Mixing		Multinomial Mixing	
	Test Size	Coverage	Test Size	Coverage
<i>Exchangeable: n = 50</i>				
β_0	0.0650	0.9350	0.0640	0.9360
β_1	0.1170	0.8830	0.1160	0.8840
β_2	0.0790	0.9210	0.0790	0.9210
ρ	0.1080	0.8920	0.1090	0.8910
<i>AR(1): n = 50</i>				
β_0	0.0850	0.9150	0.0850	0.9150
β_1	0.1100	0.8900	0.1120	0.8880
β_2	0.1020	0.8980	0.0990	0.9010
ρ	0.1310	0.8690	0.1320	0.8680
<i>Exchangeable: n = 200</i>				
β_0	0.0540	0.9460	0.0530	0.9470
β_1	0.0640	0.9360	0.0650	0.9350
β_2	0.0630	0.9370	0.0640	0.9360
ρ	0.0700	0.9300	0.0670	0.9330
<i>AR(1): n = 200</i>				
β_0	0.0630	0.9370	0.0620	0.9380
β_1	0.0650	0.9350	0.0620	0.9380
β_2	0.0620	0.9380	0.0620	0.9380
ρ	0.0670	0.9330	0.0670	0.9330

Table 5.4 for the high correlation level also shows similar results to what appeared in Table 5.2, that is, the results improve with increasing sample size and the multinomial mixing approach often shows slightly better results compared to the binary mixing.

Appendix B

5.3 Derivation of the RS algorithm for multinomial mixing models

Suppose $\mathbf{y}_i$ is multivariate normal for each pattern of zero inflation: $\mathbf{y}_i|\omega_i = r \sim N(\boldsymbol{\mu}_{ir}, \boldsymbol{\Sigma})$, for $r = 1, \dots, R$. Then the conditional expectation of the latent variables given the observed response is given by:

$$\begin{aligned} E(\tilde{\omega}_{ir}|\mathbf{y}_i; \boldsymbol{\psi}) &= P(\tilde{\omega}_{ir} = 1|\mathbf{y}_i; \boldsymbol{\psi}) = \frac{f(\mathbf{y}_i; \boldsymbol{\mu}_{ir}, \boldsymbol{\Sigma})P(\tilde{\omega}_{ir} = 1)}{\sum_{h=1}^R f(\mathbf{y}_i; \boldsymbol{\mu}_{ih}, \boldsymbol{\Sigma})P(\tilde{\omega}_{ih} = 1)} \\ &= \frac{\frac{\nu_{ir}}{\nu_{iR}} \exp\{\mathbf{y}_i \boldsymbol{\Sigma}^{-1} \boldsymbol{\mu}_{ir} - \frac{1}{2} \boldsymbol{\mu}_{ir}^T \boldsymbol{\Sigma}^{-1} \boldsymbol{\mu}_{ir}\}}{1 + \sum_{h=1}^{R-1} \frac{\nu_{ih}}{\nu_{iR}} \exp\{\mathbf{y}_i \boldsymbol{\Sigma}^{-1} \boldsymbol{\mu}_{ih} - \frac{1}{2} \boldsymbol{\mu}_{ih}^T \boldsymbol{\Sigma}^{-1} \boldsymbol{\mu}_{ih}\}}. \end{aligned} \quad (5.1)$$

Equation (5.1) corresponds to a logistic form for $E(\tilde{\omega}_{ir}|\mathbf{y}_i; \boldsymbol{\psi})$ of the form

$$\log \left(\frac{E(\tilde{\omega}_{ir}|\mathbf{y}_i; \boldsymbol{\psi})}{E(\tilde{\omega}_{iR}|\mathbf{y}_i; \boldsymbol{\psi})} \right) = \log(\nu_{ir}/\nu_{iR}) + \alpha_{ir} + \mathbf{y}_i^T \boldsymbol{\Psi}_{ir}, \forall r = 1, 2, \dots, R,$$

where $\alpha_{ir} = -\frac{1}{2} \boldsymbol{\mu}_{ir}^T \boldsymbol{\Sigma}^{-1} \boldsymbol{\mu}_{ir}$ and $\boldsymbol{\Psi}_{ir} = \boldsymbol{\Sigma}^{-1} \boldsymbol{\mu}_{ir}$. Note that $\alpha_{ir} + \mathbf{y}_i^T \boldsymbol{\Psi}_{ir}$ is a linear function of $\text{vech}(\boldsymbol{\Sigma}^{-1})$ allowing us to write these terms as $\mathbf{x}_{ir} \boldsymbol{\kappa}$ where $\mathbf{x}_{ir} = -\frac{1}{2} \text{vec}^T(\boldsymbol{\mu}_{ir}(\boldsymbol{\mu}_{ir}^T - 2\mathbf{y}_i^T)) \mathbf{D}$ and $\boldsymbol{\kappa} = \text{vech}(\boldsymbol{\Sigma}^{-1})$ where $\mathbf{D}$ is the *elimination matrix* that transforms *vec* to *vech*. Based upon these arguments, we take the linear predictor in equation (3.8) to be of the form $\mathbf{x}_{ir}^{(k)} \boldsymbol{\kappa}$ where $\mathbf{x}_{ir}^{(k)}$ is $\mathbf{x}_{ir}$ evaluated at $\boldsymbol{\mu}_{ir} = \boldsymbol{\mu}_{ir}^{(k)}$.

$$E(\tilde{\omega}_{ir}|\mathbf{y}_i, \boldsymbol{\psi}^{(k)}) = \frac{\nu_{ir}f(\mathbf{y}_i|\omega_i = r)}{\sum_{l=1}^R \nu_{il}f(\mathbf{y}_i|\omega_i = l)}$$

Suppose $\mathbf{y}_i = \begin{pmatrix} y_{i1} \\ y_{i2} \end{pmatrix} | \omega_i = r \sim N(\boldsymbol{\mu}_i, \boldsymbol{\Sigma})$. With $t_i = 2, \forall i$, then $r = 1, 2, 3, R = 4$. Let $\boldsymbol{\mu}_i = (\mu_{i1}, \mu_{i2})^T$ and $\boldsymbol{\Sigma} = \begin{pmatrix} \sigma_{11}^2 & \rho\sigma_{11}\sigma_{22} \\ \rho\sigma_{11}\sigma_{22} & \sigma_{22}^2 \end{pmatrix}$.

Then we have:

$$f(\mathbf{y}_i|\omega_i = 1) = [(2\pi)^2|\boldsymbol{\Sigma}|]^{-1/2} \exp \left\{ -1/2 (\mathbf{y}_i^T \boldsymbol{\Sigma}^{-1} \mathbf{y}_i - 2\mathbf{y}_i^T \boldsymbol{\Sigma}^{-1} \boldsymbol{\mu}_i + \boldsymbol{\mu}_i^T \boldsymbol{\Sigma}^{-1} \boldsymbol{\mu}_i) \right\}$$

$$f(\mathbf{y}_i|\omega_i = 2) = [(2\pi)\sigma_{11}^2]^{-1/2} \exp \left\{ \frac{-1}{2\sigma_{11}^2} (y_{i1}^2 - 2y_{i1}\mu_{i1} - \mu_{i1}^2) \right\} I\{y_{i2} = 0\}$$

$$\sim N(\mathbf{e}_1^T \boldsymbol{\mu}_i, \mathbf{e}_1^T \boldsymbol{\Sigma} \mathbf{e}_1) I\{y_{i2} = 0\}, \text{ where } \mathbf{e}_1^T = (1, 0)$$

$$f(\mathbf{y}_i|\omega_i = 3) = [(2\pi)\sigma_{22}^2]^{-1/2} \exp \left\{ \frac{-1}{2\sigma_{22}^2} (y_{i2}^2 - 2y_{i2}\mu_{i2} - \mu_{i2}^2) \right\} I\{y_{i1} = 0\}$$

$$\sim N(\mathbf{e}_2^T \boldsymbol{\mu}_i, \mathbf{e}_2^T \boldsymbol{\Sigma} \mathbf{e}_2) I\{y_{i1} = 0\}, \text{ where } \mathbf{e}_2^T = (0, 1)$$

$$f(\mathbf{y}_i|\omega_i = 4) = I(\{y_{i1} = y_{i2} = 0\})$$

Succinctly we can write the mixture as:

$$f(\mathbf{y}_i|\omega_i = r) = [(2\pi)|\mathbf{A}_r^T \boldsymbol{\Sigma} \mathbf{A}_r|]^{-1/2} \times \exp \left\{ -1/2 (\mathbf{y}_i - \boldsymbol{\mu}_i)^T \mathbf{A}_r^T (\mathbf{A}_r \boldsymbol{\Sigma}^{-1} \mathbf{A}_r^T \mathbf{A}_r)^{-1} (\mathbf{y}_i - \boldsymbol{\mu}_i) \right\} \times a_r,$$

for $r = 1, 2, 3$

and $I(y_{i1} = y_{i2} = 0)$, for $r = 4$.

$$\text{where } a_r = \begin{cases} 1, & r = 1 \\ I(y_{i1} = 0), & r = 2 \\ I(y_{i2} = 0), & r = 3 \end{cases}$$

Then we combine to get:

$$\begin{aligned} \tilde{\omega}_{ir}^{(k+1)} &= \text{E} \left(\tilde{\omega}_{ir} = 1 | \mathbf{y}_i, \boldsymbol{\psi}^{(k)} \right) \\ &= \frac{\frac{\nu_r}{\nu_4} a_r [2\pi \mathbf{A}_r \boldsymbol{\Sigma} \mathbf{A}_r^T]^{(-1/2)} \exp \left\{ -1/2 (\mathbf{y}_i - \boldsymbol{\mu}_i)^T \mathbf{A}_r^T (\mathbf{A}_r \boldsymbol{\Sigma}^{-1} \mathbf{A}_r^T)^{-1} \mathbf{A}_r (\mathbf{y}_i - \boldsymbol{\mu}_i) \right\}}{a_4 + \sum_{l=1}^3 \frac{\nu_l}{\nu_4} a_l [2\pi \mathbf{A}_l \boldsymbol{\Sigma} \mathbf{A}_l^T]^{(-1/2)} \exp \left\{ -1/2 (\mathbf{y}_i - \boldsymbol{\mu}_i)^T \mathbf{A}_l^T (\mathbf{A}_l \boldsymbol{\Sigma}^{-1} \mathbf{A}_l^T)^{-1} \mathbf{A}_l (\mathbf{y}_i - \boldsymbol{\mu}_i) \right\}} \end{aligned}$$

This leads to an equivalent multinomial logit model:

$$\frac{\nu_r}{\nu_4} + \alpha_r + \mathbf{x}_{ir}^T \boldsymbol{\gamma}$$

where $\mathbf{x}_{ir}^T \boldsymbol{\gamma}$ is derived from $-1/2 (\mathbf{y}_i - \boldsymbol{\mu}_i)^T \mathbf{A}_r^T (\mathbf{A}_r \boldsymbol{\Sigma}^{-1} \mathbf{A}_r^T)^{-1} \mathbf{A}_r (\mathbf{y}_i - \boldsymbol{\mu}_i)$ with terms of the form $-1/2(y_{i1} - \mu_{i1})^2 \times \sigma_{11}^{-1}$, $-1/2(y_{i2} - \mu_{i2})^2 \times \sigma_{22}^{-1}$ and $-1/2(y_{i1} - \mu_{i1})(y_{i2} - \mu_{i2}) \times f(\sigma_{11}^{-1}, \sigma_{22}^{-1}, \sigma_{12})$. The unknown terms are will be regression coefficients $\boldsymbol{\gamma}$ which corresponds to the unknowns σ_{11}^{-1} , $\sigma_{12}^{-1} = \sigma_{21}^{-1}$ and σ_{22}^{-1} . $\alpha_r = \log [2\pi |\mathbf{A}_r \boldsymbol{\Sigma} \mathbf{A}_r^T|]^{(-1/2)}$.

5.4 Derivation of the complete data loglikelihood: multinomial mixing

Suppose $n_i = 2, \forall i$ and the $\omega \in \{1, 2, 3, 4\}$ and we have the indicator vectors

$$\begin{aligned} l^c(\boldsymbol{\psi}; \mathbf{y}, \tilde{\boldsymbol{\omega}}) &= \log f(\mathbf{y}, \tilde{\boldsymbol{\omega}}; \boldsymbol{\psi}) \\ &= \sum_i \log f(\mathbf{y}_i; \tilde{\boldsymbol{\omega}}_i; \boldsymbol{\psi}) \\ &= \sum_i \log f(\tilde{\boldsymbol{\omega}}_i) + \sum_i \log f(\mathbf{y}_i | \tilde{\boldsymbol{\omega}}, \boldsymbol{\psi}) \end{aligned}$$

Where

$$\begin{aligned} f(\mathbf{y}_i | \tilde{\boldsymbol{\omega}}, \boldsymbol{\psi}) &= [f_2(y_{i1})f_2(y_{i2})]^{\tilde{\omega}_{i1}} + [f_2(y_{i1})I(y_{i2} = 0)]^{\tilde{\omega}_{i2}} + \\ &\quad [I(y_{i1} = 0)f_2(y_{i2})]^{\tilde{\omega}_{i3}} + [I(y_{i1} = y_{i2} = 0)]^{1 - \sum_r \tilde{\omega}_{ir}} \end{aligned}$$

$$\begin{aligned} \Rightarrow \sum_i \log f(\mathbf{y}_i | \tilde{\boldsymbol{\omega}}, \boldsymbol{\psi}) &= (\tilde{\omega}_{i1} + \tilde{\omega}_{i2}) \log f_2(y_{i1}) + (\tilde{\omega}_{i1} + \tilde{\omega}_{i3}) \log f_2(y_{i2}) + \\ &\quad (1 - (\tilde{\omega}_{i3} + \tilde{\omega}_{i4})) \log I(y_{i1} = 0) + (1 - (\tilde{\omega}_{i2} + \tilde{\omega}_{i4})) \log I(y_{i2} = 0) \\ &= (1 - u_{i1}) \log f_2(y_{i1}) + (1 - u_{i2}) \log f_2(y_{i2}) \end{aligned}$$

5.4.1 Estimating function for the ES-2 algorithm to marginal ZI regression models for clustered data with excess zeros

$$\begin{aligned}
\mathbf{U}_1 &= \sum_i^K \begin{bmatrix} \mathbf{B}_i^T \partial \boldsymbol{\mu}_i^T / \partial \boldsymbol{\eta}_{1i} \mathbf{V}_{i11}^{-1} \mathbf{U}_{i1}(\mathbf{y}_i - \boldsymbol{\mu}_i) \\ \mathbf{G}_i^T \partial \mathbf{p}_i^T / \partial \boldsymbol{\eta}_{2i} \mathbf{W}_{i11}^{-1}(\mathbf{u}_i - \mathbf{p}_i) \\ \partial \boldsymbol{\sigma}_i^T / \partial \tilde{\boldsymbol{\rho}} \tilde{\mathbf{U}}_{i2}(\mathbf{s}_i - \boldsymbol{\sigma}_i) \\ \partial \boldsymbol{\tau}_i^T / \partial \boldsymbol{\delta}(\mathbf{t}_i - \boldsymbol{\tau}_i) \end{bmatrix} \\
&= \sum_i^K \left(\begin{bmatrix} \mathbf{B}_i^T \partial \boldsymbol{\mu}_i^T / \partial \boldsymbol{\eta}_{1i} \mathbf{V}_{i11}^{-1} \tilde{\mathbf{U}}_i(\boldsymbol{\mu}_i) - \mathbf{y}_i \\ \mathbf{G}_i^T \partial \mathbf{p}_i^T / \partial \boldsymbol{\eta}_{2i} \mathbf{W}_{i11}^{-1} \mathbf{u}_i \\ \partial \boldsymbol{\sigma}_i^T / \partial \tilde{\boldsymbol{\rho}} \tilde{\mathbf{U}}_{i2}(\mathbf{s}_i - \boldsymbol{\sigma}_i) \\ \partial \boldsymbol{\tau}_i^T / \partial \boldsymbol{\delta} \mathbf{t}_i \end{bmatrix} - \begin{bmatrix} \mathbf{B}_i^T \partial \boldsymbol{\mu}_i^T / \partial \boldsymbol{\eta}_{1i} \mathbf{V}_{i11}^{-1}(\boldsymbol{\mu}_i - \mathbf{y}_i) \\ \mathbf{G}_i^T \partial \mathbf{p}_i^T / \partial \boldsymbol{\eta}_{2i} \mathbf{W}_{i11}^{-1} \mathbf{p}_i \\ \mathbf{0} \\ \partial \boldsymbol{\tau}_i^T / \partial \boldsymbol{\delta} \boldsymbol{\tau}_i \end{bmatrix} \right) \\
&= \mathbf{S} - \sum_i^K \begin{bmatrix} \mathbf{B}_i^T \partial \boldsymbol{\mu}_i^T / \partial \boldsymbol{\eta}_{1i} \mathbf{V}_{i11}^{-1}(\boldsymbol{\mu}_i - \mathbf{y}_i) \\ \mathbf{G}_i^T \partial \mathbf{p}_i^T / \partial \boldsymbol{\eta}_{2i} \mathbf{W}_{i11}^{-1} \mathbf{p}_i \\ \mathbf{0} \\ \partial \boldsymbol{\tau}_i^T / \partial \boldsymbol{\delta} \boldsymbol{\tau}_i \end{bmatrix},
\end{aligned}$$

$$\text{where } \mathbf{S} = \sum_i^K \mathbf{S}_i = \begin{bmatrix} \mathbf{B}_i^T \partial \boldsymbol{\mu}_i^T / \partial \boldsymbol{\eta}_{1i} \mathbf{V}_{i11}^{-1} \tilde{\mathbf{U}}_i(\boldsymbol{\mu}_i) - \mathbf{y}_i \\ \mathbf{G}_i^T \partial \mathbf{p}_i^T / \partial \boldsymbol{\eta}_{2i} \mathbf{W}_{i11}^{-1} \mathbf{u}_i \\ \partial \boldsymbol{\sigma}_i^T / \partial \tilde{\boldsymbol{\rho}} \tilde{\mathbf{U}}_{i2}^{(r)}(\mathbf{s}_i - \boldsymbol{\sigma}_i) \\ \partial \boldsymbol{\tau}_i^T / \partial \boldsymbol{\delta} \mathbf{t}_i \end{bmatrix}.$$

$$\begin{aligned}
E(\mathbf{S}_i|\mathbf{Y}_i) &= \sum_i^K \mathbf{S}_i \\
&= \sum_i^K \begin{bmatrix} \mathbf{B}_i^T \partial \boldsymbol{\mu}_i^T / \partial \boldsymbol{\eta}_{1i} \mathbf{V}_{i11}^{-1} \tilde{\mathbf{U}}_{i1}^{(r)}(\boldsymbol{\mu}_i) - \mathbf{y}_i \\ \mathbf{G}_i^T \partial \mathbf{p}_i^T / \partial \boldsymbol{\eta}_{2i} \mathbf{W}_{i11}^{-1} \mathbf{u}_i \\ \partial \boldsymbol{\sigma}_i^T / \partial \tilde{\boldsymbol{\rho}} \tilde{\mathbf{U}}_{i2}^{(r)}(\mathbf{s}_i - \boldsymbol{\sigma}_i) \\ \partial \boldsymbol{\tau}_i^T / \partial \tilde{\boldsymbol{\delta}} \tilde{\mathbf{U}}_{i3}^{(r)} \end{bmatrix},
\end{aligned}$$

where $\tilde{\mathbf{U}}_{i3}^{(r)} = E(\mathbf{t}_i|\mathbf{Y}_i)$. We can estimate $\tilde{\mathbf{U}}_{i2}^{(r)} = \text{vech}\{\mathbf{j}_{n_i}\mathbf{j}_{n_i}^T - \mathbf{j}_{n_i}(\mathbf{u}_i^{(r)})^T - \mathbf{u}_i^{(r)}\mathbf{j}_{n_i} + \mathbf{u}_i^{(r)}(\mathbf{u}_i^{(r)})^T\}$ and $\tilde{\mathbf{U}}_{i3}^{(r)} = \text{vech}\{\mathbf{W}_{i11} + \mathbf{u}_i^{(r)}(\mathbf{u}_i^{(r)})^T - \mathbf{u}_i^{(r)}\mathbf{p}_i^T - \mathbf{p}_i(\mathbf{u}_i^{(r)})^T + \mathbf{p}_i\mathbf{p}_i^T\}$.

$$\mathbf{U}_2 = \mathbf{S} - E(\mathbf{S}|\mathbf{Y})$$

.

We can solve simultaneously:

$$\mathbf{U} = \begin{pmatrix} \mathbf{U}_1 \\ \mathbf{U}_2 \end{pmatrix}$$

5.4.2 Multinomial logit models score function

Let $\tilde{\omega}_{ir} = 1$ if $\omega_i = r$ and 0 otherwise. $P(\tilde{\omega}_{ir} = 1) = P(\omega_i = r) = \nu_{ir}, r = 1, 2, \dots, R$. Assume the logit regression model

$$\log\left(\frac{\nu_{ir}}{\nu_{iR}}\right) = \mathbf{G}_i^T \boldsymbol{\gamma}_r, r = 1, 2, \dots, R - 1$$

Where the last category is used as the baseline-category. Re-writing we have

$$\nu_{ir} = \frac{\exp(\mathbf{G}_i^T \boldsymbol{\gamma}_r)}{1 + \sum_{h=1}^{R-1} \exp(\mathbf{G}_i^T \boldsymbol{\gamma}_h)}, r = 1, 2, \dots, R-1,$$

where $\sum_{r=1}^R \nu_{ir} = 1$ and $\nu_{iR} = 1 - \sum_{r=1}^{R-1} \nu_{ir}$. Hence

$$\nu_{iR} = \frac{1}{1 + \sum_{h=1}^{R-1} \exp(\mathbf{G}_i^T \boldsymbol{\gamma}_h)}.$$

The kernel of the likelihood function is given by

$$\begin{aligned} L(\boldsymbol{\gamma}; \boldsymbol{\Omega}) &\simeq \prod_{i=1}^K \prod_{r=1}^R \nu_{ir}^{\tilde{\omega}_{ir}} \\ &= \prod_{i=1}^K \prod_{r=1}^{R-1} \nu_{ir}^{\tilde{\omega}_{ir}} \nu_{iR}^{1 - \sum_{r=1}^{R-1} \tilde{\omega}_{ir}} \\ &= \prod_{i=1}^K \prod_{r=1}^{R-1} \nu_{ir}^{\tilde{\omega}_{ir}} \frac{\nu_{iR}}{\nu_{iR}^{\sum_{r=1}^{R-1} \tilde{\omega}_{ir}}} \\ &= \prod_{i=1}^K \prod_{r=1}^{R-1} \nu_{ir}^{\tilde{\omega}_{ir}} \frac{\nu_{iR}}{\prod_{r=1}^{R-1} \nu_{iR}^{\tilde{\omega}_{ir}}} \\ &= \prod_{i=1}^K \prod_{r=1}^{R-1} \left(\frac{\nu_{ir}}{\nu_{iR}} \right)^{\tilde{\omega}_{ir}} \nu_{iR} \\ &= \prod_{i=1}^K \prod_{r=1}^{R-1} \exp(\tilde{\omega}_{ir} \mathbf{G}_i^T \boldsymbol{\gamma}_r) \left(\frac{1}{1 + \sum_{h=1}^{R-1} \exp(\mathbf{G}_i^T \boldsymbol{\gamma}_h)} \right). \end{aligned}$$

The loglikelihood is given by

$$l(\boldsymbol{\gamma}; \boldsymbol{\Omega}) = \sum_{i=1}^K \sum_{r=1}^R \left\{ \tilde{\omega}_{ir} \mathbf{G}_i^T \boldsymbol{\gamma}_r - \log \left(1 + \sum_{h=1}^{R-1} \exp(\mathbf{G}_i^T \boldsymbol{\gamma}_h) \right) \right\}$$

The score function $U(\boldsymbol{\gamma})$ is given by

$$\begin{aligned}
U(\boldsymbol{\gamma}) &= \frac{\partial l(\boldsymbol{\gamma}; \boldsymbol{\Omega})}{\partial \boldsymbol{\gamma}_r} \\
&= \sum_{i=1}^K \sum_{r=1}^R \left\{ \tilde{\omega}_{ir} \mathbf{G}_i^T - \frac{\exp(\mathbf{G}_i^T \boldsymbol{\gamma}_r)}{1 + \sum_{h=1}^{R-1} \exp(\mathbf{G}_i^T \boldsymbol{\gamma}_h)} \mathbf{G}_i^T \right\} \\
&= \sum_{i=1}^K \sum_{r=1}^R \{ \tilde{\omega}_{ir} \mathbf{G}_i^T - \nu_{ir} \mathbf{G}_i^T \} \\
&= \sum_{i=1}^K \sum_{r=1}^R (\tilde{\omega}_{ir} - \nu_{ir}) \mathbf{G}_i^T \\
&= \mathbf{G}^T (\boldsymbol{\Omega} - \boldsymbol{\Pi}),
\end{aligned}$$

where $\boldsymbol{\Omega}$ is the $K \times (R-1)$ matrix with elements $\tilde{\omega}_{ir}$, $\boldsymbol{\Pi}$ is the $K \times (R-1)$ matrix with elements ν_{ir} , $\mathbf{G}$ is the $K \times p$ and $\boldsymbol{\gamma}$ is the $p \times (R-1)$ matrix of parameters.

5.4.3 ES-2 algorithm for multinomial mixing models

The ES-2 algorithm as described in section 2.3 can be applied to the ZI regression models with particular application to ZIP regression models. For each subject i the observed data vector be given by $\mathbf{y}_i$ and the missing data vector be given by $\tilde{\boldsymbol{\omega}}_i$. Then the complete data is given by $(\mathbf{y}_i^T, \tilde{\boldsymbol{\omega}}_i^T)^T$. First we construct the expanded ‘complete data’ estimating equation for all the parameters as follows

$$\mathbf{U}_1 = \sum_i^K \begin{bmatrix} \mathbf{B}_i^T \partial \boldsymbol{\mu}_i^T(\boldsymbol{\beta}) / \partial \boldsymbol{\eta}_{1i} \mathbf{V}_{i11}^{-1} \mathbf{H}_{i1} [\mathbf{y}_i - \boldsymbol{\mu}_i(\boldsymbol{\beta})] \\ \mathbf{G}_i^T [\tilde{\boldsymbol{\omega}}_i - \boldsymbol{\omega}_i(\boldsymbol{\gamma})] \\ \partial \boldsymbol{\sigma}_i^T(\tilde{\boldsymbol{\rho}}) / \partial \tilde{\boldsymbol{\rho}} \mathbf{H}_{i2} [\mathbf{s}_i - \boldsymbol{\sigma}_i(\tilde{\boldsymbol{\rho}})] \end{bmatrix}.$$

where $\boldsymbol{\eta}_{1i} = \mathbf{B}_i^T \boldsymbol{\beta}$ and $\boldsymbol{\eta}_{2i} = \mathbf{G}_i^T \boldsymbol{\gamma}$. For the ZIP regression models, the log-link is often used

for η_{1i} and the logit-link is used for η_{2i} for the degenerate part.

Next, we can the decompose the above function to conform with Equation ?? . This lead to the following decomposition

$$\begin{aligned}
\mathbf{U}_1 &= \sum_i^K \begin{bmatrix} \mathbf{B}_i^T \partial \boldsymbol{\mu}_i^T(\boldsymbol{\beta}) / \partial \boldsymbol{\eta}_{1i} \mathbf{V}_{i11}^{-1} \mathbf{u}_i[\boldsymbol{\mu}_i(\boldsymbol{\beta}) - \mathbf{y}_i] \\ \mathbf{G}_i^T \tilde{\boldsymbol{\omega}}_i \\ \partial \boldsymbol{\sigma}_i^T(\tilde{\boldsymbol{\rho}}) / \partial \tilde{\boldsymbol{\rho}} \quad \mathbf{H}_{i2}[\mathbf{s}_i - \boldsymbol{\sigma}_i(\tilde{\boldsymbol{\rho}})] \end{bmatrix} - \\
&\sum_i^K \begin{bmatrix} \mathbf{B}_i^T \partial \boldsymbol{\mu}_i^T(\boldsymbol{\beta}) / \partial \boldsymbol{\eta}_{1i} \mathbf{V}_{i11}^{-1} [\boldsymbol{\mu}_i(\boldsymbol{\beta}) - \mathbf{y}_i] \\ \mathbf{G}_i^T \boldsymbol{\nu}_i \\ \mathbf{0} \end{bmatrix} \\
&= \mathbf{S} - \sum_i^K \begin{bmatrix} \mathbf{B}_i^T \partial \boldsymbol{\mu}_i^T(\boldsymbol{\beta}) / \partial \boldsymbol{\eta}_{1i} \mathbf{V}_{i11}^{-1} [\boldsymbol{\mu}_i(\boldsymbol{\beta}) - \mathbf{y}_i] \\ \mathbf{G}_i^T \boldsymbol{\nu}_i \\ \mathbf{0} \end{bmatrix}.
\end{aligned}$$

In this case, the “complete data summary statistic is given by

$$\mathbf{S} = \sum_i^K \mathbf{S}_i = \sum_i^K \begin{bmatrix} \mathbf{B}_i^T \partial \boldsymbol{\mu}_i^T(\boldsymbol{\beta}) / \partial \boldsymbol{\eta}_{1i} \mathbf{V}_{i11}^{-1} \mathbf{u}_i[\boldsymbol{\mu}_i(\boldsymbol{\beta}) - \mathbf{y}_i] \\ \mathbf{G}_i^T \tilde{\boldsymbol{\omega}}_i \\ \partial \boldsymbol{\sigma}_i^T(\tilde{\boldsymbol{\rho}}) / \partial \tilde{\boldsymbol{\rho}} \quad \mathbf{H}_{i2}[\mathbf{s}_i - \boldsymbol{\sigma}_i(\tilde{\boldsymbol{\rho}})] \end{bmatrix}$$