PONTIFICIA UNIVERSIDAD CATÓLICA DEL PERÚ

ESCUELA DE POSGRADO

Jointly modelling of cluster dependent profiles of fractional and binary variables from a Bayesian point of view

TESIS PARA OPTAR POR EL GRADO DE MAGÍSTER EN ESTADÍSTICA

Presentado por:

Fernando Javier Cortés Tejada

Asesor: Cristian Luis Bayes Rodríguez

Miembros del jurado:
Dr. Cristian Luis Bayes Rodríguez
Dr. Luis Hilmar Valdivieso Serrano
Dr. Luis Enrique Benites Sánchez

Lima, Marzo 2020

Resumen

En la presente tesis se proponen modelos de clasificación basados en regresiones beta inflacionadas cero-uno con efectos mixtos para modelar perfiles longitudinales de variables fraccionarias mixtas y variables binarias de forma conjunta con formación de clústeres. Las distintas parametrizaciones de los modelos propuestos permiten modelar distintos efectos, como modelar directamente la media marginal a través de covariables e interpretar fácilmente su efecto sobre ella o modelar la media condicional y las probabilidades de inflación de forma separada. Además, se forman clústeres de grupos de individuos con perfiles longitudinales similares a través de una variable latente, asumiendo que las variables respuesta siguen un modelo de mixtura finita. Debido a la complejidad de los modelos, los parámetros se estiman desde un punto de vista bayesiano, a partir de simulaciones MCMC utilizando el software JAGS en R. Se prueban los modelos propuestos sobre diferentes bases de datos simulados para medir el desempeño de los mismos y se comparan con otros modelos a fin de verificar cual ajusta mejor los perfiles longitudinales de variables fraccionarias mixtas y variables binarias. Por último, se aplican los modelos propuestos a datos reales de un banco peruano, con información del ratio de uso de tarjetas de crédito en el periodo de un año, estado de default del cliente y otras covariables correspondientes al cliente poseedor de la tarjeta, con el objetivo de obtener clústeres de individuos con similar ratio de uso de tarjeta de crédito y relacionarlos con la probabilidad de caer en default que presenta cada grupo.

Palabras-clave: variables fraccionarias, inferencia bayesiana, modelo de regresión beta inflacionada, modelo de efectos mixtos, modelo de mixtura finita, modelo de clasificación, MCMC.

Abstract

The following thesis proposes classification models that consist of jointly fitting longitudinal profiles of mixed fractional and binary variables modelled by zero-one beta inflated mixed regressions with cluster formation. The distinct proposed parametrizations allow different effects to be modelled, such as modelling the marginal mean directly through independent variables and easily interpret its effect on it or modelling the conditional mean and the inflation probabilities separately. In addition, individuals with similar fractional longitudinal profiles are grouped into a cluster through a latent variable, assuming that the response variables follow a finite mixture model. Due to the complexity of the models, the parameters are estimated from a Bayesian point of view by simulating a MCMC using JAGS software in R. The proposed models are fitted in various simulated datasets and are compared against other models to measure performance in fitting fractional longitudinal profiles and binary variables. Finally, an application on real data is conducted, consisting on longitudinal information of credit card utilization ratio and default status as dependants variables and covariates corresponding to client information, aiming to obtain clusters of clients with similar behaviour in evolution of credit card utilization and relate them to their probability of default.
Keywords: fractional variables, Bayesian inference, beta inflated regression model, mixed effects model, finite mixture model, classification model, MCMC.

Table of Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1. Preliminary considerations

In the field of credit risk the main topic to be studied is the trend of clients that fail to pay their financial obligations or simply default. The probability of default (PD) is often modeled by a logistic regression, which assumes that observations are independent. It is also a matter of interest to study a client's behavior and/or evolution through time in other variables, for example their credit card utilization ratio (CCUR). This results in an interest of jointly modelling the time evolution of the credit card utilization ratio of a client and its probability of default.

The credit card utilization ratio is a fractional variable that can take values in the interval $[0, 1]$, this means it is a mixed random variable, with a discrete and a continuous component, since it can take with positive probability the values of 0 and 1. For example, if a client does not use its credit card, the utilization ratio will be 0. On the other hand, if a client use its entire credit card line, the utilization ratio will be 1. If it is not in either of these two cases, it must be lying in the continuous interval $(0, 1)$.

In order to model a fractional response variable Ferrari and Cribari-Neto (2004) proposed a beta regression model based on a reparametrization of the beta distribution and Figueroa-Zúñiga et al. (2013) extends the model adding mixed effects to it. However, these two models can only be used for response values in the open interval $(0, 1)$. Simpler attempts have been made to model a fractional mixed response variable, for example, a linear regression can be used, yet this would not be appropriate because the estimations could fall outside the interval $[0, 1]$. Alternatively, a transformation can be performed so the response values will belong to the closed interval, but the ease of interpretation of the regression parameters is reduced or even lost, and it also ignores the mixed nature of the variable.

In an effort to model a mixed fractional response variable, Ramalho and da Silva (2009) proposed a two-part model that consist of fitting first a multinomial model in order to estimate if the response variable is in the boundaries (zero or one) or lies in an open interval, then fitting another model for the open interval. Ospina and Ferrari (2010) proposed a zero-one beta inflated distribution or simply beta inflated distribution and in Ospina and Ferrari (2012) they propose the zero-or-one inflated beta regression model (inflation at either 0 or 1, but not both).

The formerly presented models where performed using frequentist procedures, Wieczorek and Hawala (2011) and Wieczorek et al. (2012) who introduces the zero-and-one inflated beta

regression model (inflation at both, 0 and 1) with estimations in the Bayesian framework and Liu and Kong (2015) builds an R package which allows the implementation of this type of models in a much easier way. Then, Bayes and Valdivieso (2016) proposed a reparametrization of the beta inflated distribution which allows to model the mean directly through independent variables and to easily interpret its effect on the mean. Finally, Fernandez et al. (2018) extends the latter model for longitudinal response variables and incorporates mixed effects to it.

Regarding cluster formation on longitudinal data, De la Cruz-Mesía et al. (2008) fits a non-fractional response variable assuming that it follows a mixture model of sigmoidal curves, using frequentist and Bayesian estimation procedures. Then, binary classification is incorporated to the latter idea in Gaskins et al. (2017) and De la Cruz et al. (2017), where the first models the longitudinal response by penalized splines and the second uses a Dirichlet process to clusterize the trajectories.

1.2. Objectives

The main objective of this thesis is to study, estimate and apply to real data, the classification beta inflated mixed regression models for longitudinal and binary response variables with cluster formation from a Bayesian point of view. Specifically:

- Investigate about the beta inflated mixed regression model and the finite mixture model in the literature.
- Study the properties of the beta inflated mixed regression model for longitudinal response variables with cluster formation.
- Implement a program on an open-source software (JAGS) for the estimation from a Bayesian point of view.
- Conduct a simulation study where the proposed models are compared with other models.
- Apply the proposed models to real data and compare the results against other models and the case without considering clusters.

1.3. Work organization

The thesis is organized as follows: in Chapter 2, the beta inflated distribution and their alternative parametrizations are presented, along with their properties, advantages and limitations. Chapter 3 presents the classification models consisting on different beta inflated mixed regressions for longitudinal response variables with cluster formation, the augmented likelihood function, the augmented posterior distribution, the chosen priors, the classification of a new subject and the model comparison criteria. Chapter 4 shows results obtained from a simulation study. Chapter 5 shows results obtained from the application of the models to real data. Finally, on Chapter 6 conclusions obtained from this work are discussed and suggestions for future studies are made.

Appendix A shows the tables corresponding to the simulation study, Appendix B shows the regressions structures for the zero-one inflated beta model on real data, Appendix C

shows the MCMC results of the application to real data and Appendix D shows the R code for the implementation of the different models used in this thesis in JAGS.

Chapter 2

The beta inflated distribution

This chapter introduces the beta distribution with its properties, its probability density function and an alternative parametrization. Then, extends it to the beta inflated distribution and presents two alternative parametrizations.

2.1. The beta distribution

The beta distribution is a continuous probability distribution with two parameters (α, β) which allows to model response variables restricted to the interval $(0, 1)$. The probability density function of a random variable $Y \sim Beta(\alpha, \beta)$ is given by

$$f_Y(y \mid \alpha, \beta) = \frac{\Gamma(\alpha + \beta)}{\Gamma(\alpha)\Gamma(\beta)} y^{\alpha-1}(1-y)^{\beta-1}, \quad 0 < y < 1,$$

where $\alpha > 0$ and $\beta > 0$. As Ferrari and Cribari-Neto (2004) states "the beta distribution, as is well known, is very flexible for modelling proportions since its density can have quite different shapes depending on the values of the two parameters that index the distribution". This flexibility can be seen in Figure 2.1, where J-shaped, inverted-J-shaped and bell-shaped probability density functions are presented, but it can also be U-shaped or uniform, depending on the combination of α and β.

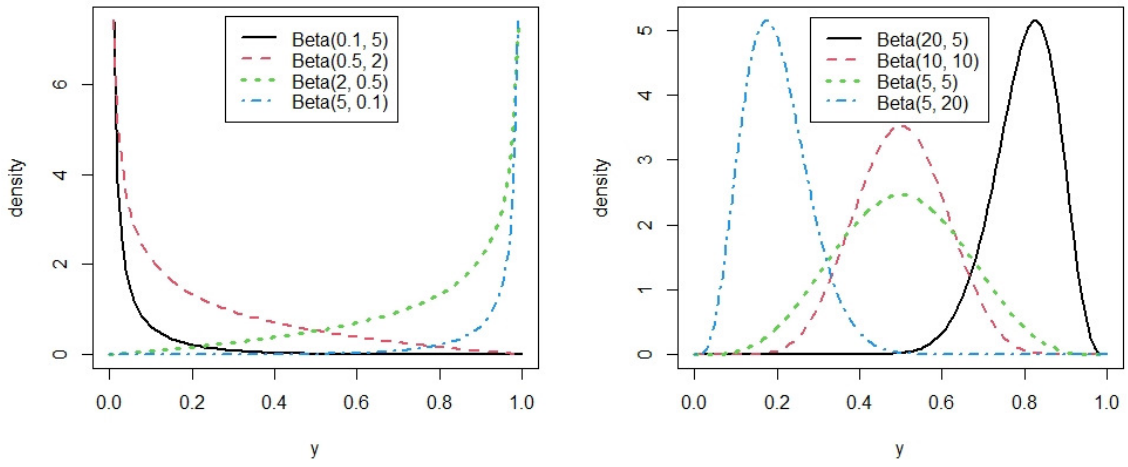


Figure 2.1: Probability density functions of the beta distribution for different parameter values

The mean and variance of the beta distribution are given by

$$\mathrm{E}(Y) = \frac{\alpha}{\alpha+\beta} \quad \text{and} \quad \mathrm{Var}(Y) = \frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}. \tag{2.1}$$

2.2. Reparametrized beta distribution

In order to propose a beta regression and to interpret directly the effect of covariates on the response variable mean, Ferrari and Cribari-Neto (2004) put forward a model considering the following reparametrization

$$\mu = \frac{\alpha}{\alpha+\beta} \quad \text{and} \quad \phi = \alpha+\beta,$$

by replacing this values in (2.1) the mean and variance under the new parametrization are

$$\mathrm{E}(Y) = \mu \quad \text{and} \quad \mathrm{Var}(Y) = \frac{V(\mu)}{1+\phi},$$

where $V(\mu) = \mu(1-\mu)$, so that μ is the mean of the response variable and ϕ can be interpreted as a precision parameter. The probability density function of Y will be written as

$$b(y \mid \mu, \phi) = \frac{\Gamma(\phi)}{\Gamma(\mu\phi)\Gamma((1-\mu)\phi)} y^{\mu\phi-1}(1-y)^{(1-\mu)\phi-1}, \quad 0 < y < 1, \tag{2.2}$$

where $0 < \mu < 1$ and $\phi > 0$, and we will use $b(\cdot|\mu, \phi)$ to refer to the beta distribution with the mean-precision parametrization.

2.3. The beta inflated distribution

The beta distribution is defined for the interval $(0, 1)$ but in some cases data contains values of zero and/or one. To solve this problem, Ospina and Ferrari (2010) proposed a beta inflated distribution (BEINF) extending (2.2) in which the probability density function of a random variable $Y \sim \mathrm{BEINF}(\delta_0, \delta_1, \mu, \phi)$ is given by

$$f_Y(y \mid \delta_0, \delta_1, \mu, \phi) = \begin{cases} \delta_0, & y = 0. \\ (1-\delta_0-\delta_1)b(y \mid \mu, \phi), & y \in (0,1). \\ \delta_1, & y = 1, \end{cases} \tag{2.3}$$

where $P(Y = 0) = \delta_0 \in (0, 1)$, $P(Y = 1) = \delta_1 \in (0, 1)$ with $\delta_0 + \delta_1 \leq 1$, $\mathrm{E}(Y \mid Y \in (0, 1)) = \mu$, $\phi > 0$ and $b(y \mid \mu, \phi)$ is the beta distribution presented in (2.2). Notice that under (2.3), Y can now take values in the interval $[0, 1]$ as shown in Figure 2.2. The mean and variance of $Y \sim \mathrm{BEINF}(\delta_0, \delta_1, \mu, \phi)$ are

$$\mathrm{E}(Y) = \delta_1 + (1-\delta_0-\delta_1)\mu \quad \text{and}$$
$$\mathrm{Var}(Y) = \delta_1(1-\delta_1) + (1-\delta_0-\delta_1)\left(\frac{V(\mu)}{1+\phi} + (\delta_0+\delta_1)\mu^2 - 2\mu\delta_1\right),$$

where $\mathrm{V}(\mu) = \mu(1-\mu)$.

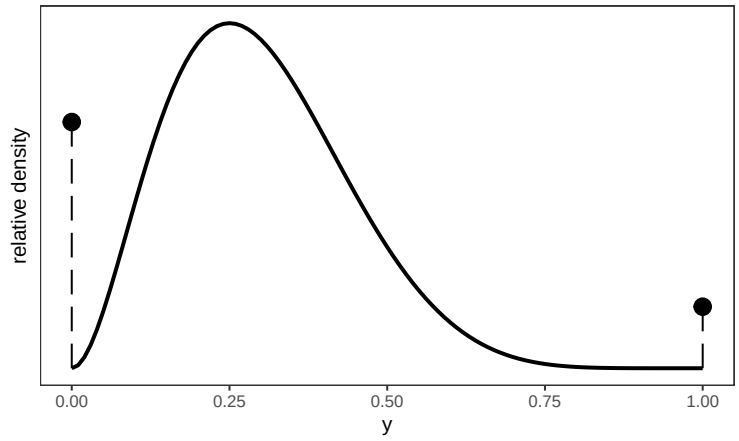


Figure 2.2: Beta inflated distribution.

2.4. Reparametrized beta inflated distributions

2.4.1. Zero-One Inflated Beta distribution (ZOIB)

An alternative parametrization of (2.3) was proposed by Liu and Kong (2015) which redefines the discrete probabilities as

$$\lambda_0 = \delta_0 \quad \text{and} \quad \lambda_1 = \frac{\delta_1}{1-\delta_0},$$

leading to the probability density function

$$f_Y(y \mid \lambda_0, \lambda_1, \mu, \phi) = \begin{cases} \lambda_0, & y = 0. \\ (1-\lambda_0)(1-\lambda_1) b(y \mid \mu, \phi), & y \in (0,1). \\ (1-\lambda_0)\lambda_1, & y = 1, \end{cases} \tag{2.4}$$

where $P(Y = 0) = \lambda_0$, $P(Y = 1) = (1-\lambda_0)\lambda_1$, $\mathrm{E}(Y \mid Y \in (0,1)) = \mu$, $\phi > 0$ and $b(y \mid \mu, \phi)$ is the beta distribution as presented in (2.2). This parametrization removes the constraint $\delta_0 + \delta_1 \leq 1$ of (2.3), which is a great advantage especially for computational simulation. From now on, (2.4) will be referred as ZOIB and will be used under the notation $Y \sim \mathrm{ZOIB}(\lambda_0, \lambda_1, \mu, \phi)$. Under this parametrization the mean and variance are

$$\mathrm{E}(Y) = (1-\lambda_0)\big(\lambda_1 + (1-\lambda_1)\mu\big) \quad \text{and}$$

$$\mathrm{Var}(Y) = (1-\lambda_0)\left[\lambda_0\lambda_1^2 + (1-\lambda_1)\left(\frac{V(\mu)}{1+\phi} + (\lambda_0 + (1-\lambda_0)\lambda_1)\mu^2 + \lambda_1\big(1 - 2\mu(1-\lambda_0)\big)\right)\right],$$

where $\mathrm{V}(\mu) = \mu(1-\mu)$.

2.4.2. Beta Inflated mean distribution (BIm)

In order to propose a beta inflated regression which allows to interpret the effect of regression parameters directly on the mean, Bayes and Valdivieso (2016) put forward a model considering the following reparametrization

$$\gamma = \delta_1 + (1 - \delta_0 - \delta_1)\mu, \qquad \alpha_0 = \frac{\delta_0}{1-\gamma} \quad \text{and} \quad \alpha_1 = \frac{\delta_1}{\gamma},$$

where $\delta_1 < \gamma < 1 - \delta_0$, so $\gamma \in (0,1)$, $\alpha_0 \in (0,1)$ and $\alpha_1 \in (0,1)$. From now on, this reparametrized beta inflated distribution will be referred as BIm and will be used under the notation $Y \sim \text{BIm}(\alpha_0, \alpha_1, \gamma, \phi)$. The mean and variance of this new beta inflated parametrization are

$$\text{E}(Y) = \gamma \quad \text{and}$$
$$\text{Var}(Y) = \frac{(1+\alpha_1\phi)}{1+\phi}\gamma + \left(\frac{(1-\alpha_1)^2\phi}{(1-\alpha_0(1-\gamma)-\alpha_1\gamma)(1+\phi)} - 1\right)\gamma^2.$$

The probability density function of Y under the new parametrization can be written as

$$f_Y(y \mid \alpha_0, \alpha_1, \gamma, \phi) = \begin{cases} \alpha_0(1-\gamma), & y = 0. \\ (1-\alpha_0(1-\gamma)-\alpha_1\gamma)b(y \mid \frac{\gamma(1-\alpha_1)}{1-\alpha_0(1-\gamma)-\alpha_1\gamma}, \phi), & y \in (0,1). \\ \alpha_1\gamma, & y = 1. \end{cases} \tag{2.5}$$

Notice that in (2.3) and (2.4) is easier to interpret the conditional mean $\text{E}(Y \mid Y \in (0,1))$ but not the marginal mean $\text{E}(Y)$. Because of this (2.5) gains relevance, allowing to model this marginal mean $\text{E}(Y) = \gamma$ directly.

Chapter 3

The classification beta inflated mixed regression model for longitudinal fractional and binary variables with cluster formation

This chapter presents the definition of the classification beta inflated mixed regression model for longitudinal fractional and binary variables with cluster formation, the augmented likelihood function, the augmented posterior distribution, the chosen priors for all the parameters, the classification of a new subject, the model comparison criteria and the label-switching problem.

3.1. Model definition

In order to model a mixed fractional variable, a beta inflated regression can be fitted using either ZOIB (2.4) or BIm (2.5) parametrizations (leaving BEINF (2.3) behind due to the constraint). It is unknown beforehand which of these two parametrizations will fit better a given response variable, because according to Carlin and Louis (2008) the deviance information criterion (DIC), used in this tesis as the model comparison criteria, is not invariant to parametrizations (more details about DIC can be found in section 3.4). For example, the parametrization of the ZOIB regression model proposed by Liu and Kong (2015) is like building a two-part model, fitting first a categorical model and then a beta regression. The categorical model consists of three categories, each one represents $\mathrm{P}(Y = 0)$, $\mathrm{P}(Y = 1)$ and $\mathrm{P}(Y \in (0, 1))$, respectively. The beta regression is only fitted to the third category. Therefore, the importance of this parametrization lies on the separation of parameter estimation, where the mean open interval (μ) does not interact with the discrete probabilities at all. On the other hand, the importance of the parametrization of the BIm regression model proposed by Bayes and Valdivieso (2016) lies on the easiness of covariate effects interpretation on the mean and the joint estimation of the discrete probabilities and the marginal mean (γ), which makes the beta distribution mean to be affected by the discrete probabilities and vice versa. If repeated measurements of a mixed fractional variable are performed for the same subject, a beta inflated regression with mixed effects has to be fitted as proposed by Fernandez (2017), Fernandez et al. (2018) and Di Brisco and Migliorati (2020).

On the other hand, binary variables are often modeled by logit or probit regressions. Regarding cluster formation, methods such as finite mixture models or Dirichlet process can be used. Gaskins et al. (2017) proposes to jointly model longitudinal and binary response variables with cluster formation induced by a Dirichlet process.

This thesis joins these ideas and proposes two classification beta inflated mixed regression models for longitudinal fractional and binary variables with cluster formation, using a latent variable instead of the Dirichlet process to model the clusters.

Let $Y_i = [Y_{i1}, Y_{i2}, ..., Y_{in_i}]^\top$, $i = 1, ..., n$ be n independent response mixed fractional vector variables each one observed n_i times and let D_i, $i = 1, ..., n$ be n independent binary response variables for the subjects in the study.

The proposed dependent variables are modelled by:

$$\begin{aligned} D_i \mid W_i = w_i &\sim \text{Bern}(\pi_i^{w_i}) \\ Y_{ij} \mid W_i = w_i &\sim \text{ZOIB}(\lambda_{0ij}^{w_i}, \lambda_{1ij}^{w_i}, \mu_{ij}^{w_i}, \phi^{w_i}) \\ W_i &\sim \text{Cat}(\boldsymbol{p}) \end{aligned} \tag{3.1}$$

and

$$\begin{aligned} D_i \mid W_i = w_i &\sim \text{Bern}(\pi_i^{w_i}) \\ Y_{ij} \mid W_i = w_i &\sim \text{BIm}(\alpha_{0ij}^{w_i}, \alpha_{1ij}^{w_i}, \gamma_{ij}^{w_i}, \phi^{w_i}) \\ W_i &\sim \text{Cat}(\boldsymbol{p}), \end{aligned} \tag{3.2}$$

where W_i is an unobserved indicator variable of the cluster w_i the subject i belongs to. W_i follows a categorical distribution, denoted as $W_i \sim \text{Cat}(\boldsymbol{p})$, with probability of belonging to each cluster $\boldsymbol{p} = [p_1, p_2, ..., p_K]^\top$ (also known as weights in the finite mixture context), given $\sum_{j=1}^{K} p_j = 1$ and the total number of clusters K; D_i is a binary variable that conditional on the cluster W_i follows a Bernoulli distribution, denoted as $D_i \mid W_i = w_i \sim \text{Bern}(\pi_i^{w_i})$, which takes the value of 1 with probability $\pi_i^{w_i}$; Y_{ij} is the fractional variable for subject i at time j with conditional mean $\mu_{ij}^{w_i}$ or marginal mean $\gamma_{ij}^{w_i}$; $\lambda_{0ij}^{w_i}$ and $\alpha_{0ij}^{w_i}$ are the parameters related to the probability that $Y_{ij} = 0$, λ_{1ij} and $\alpha_{1ij}^{w_i}$ are the parameters related to the probability that $Y_{ij} = 1$ and $\phi^{w_i} > 0$ is a cluster dependent precision parameter.

The proposed regression models for (3.1) are

$$\begin{aligned} g_\pi(\pi_i^{w_i}) &= \boldsymbol{z}_i^\top \boldsymbol{\beta}_\pi^{w_i} \\ g_{\lambda_0}(\lambda_{0ij}^{w_i}) &= \tilde{\boldsymbol{x}}_{ij}^\top \boldsymbol{\beta}_{\lambda_0}^{w_i} \\ g_{\lambda_1}(\lambda_{1ij}^{w_i}) &= \breve{\boldsymbol{x}}_{ij}^\top \boldsymbol{\beta}_{\lambda_1}^{w_i} \\ g_\mu(\mu_{ij}^{w_i}) &= \boldsymbol{x}_{ij}^\top \boldsymbol{\beta}_\mu^{w_i} + b_i \end{aligned}$$

and for (3.2) are

$$\begin{aligned} g_\pi(\pi_i^{w_i}) &= \boldsymbol{z}_i^\top \boldsymbol{\beta}_\pi^{w_i} \\ g_{\alpha_0}(\alpha_{0ij}^{w_i}) &= \tilde{\boldsymbol{x}}_{ij}^\top \boldsymbol{\beta}_{\alpha_0}^{w_i} \\ g_{\alpha_1}(\alpha_{1ij}^{w_i}) &= \breve{\boldsymbol{x}}_{ij}^\top \boldsymbol{\beta}_{\alpha_1}^{w_i} \\ g_\gamma(\gamma_{ij}^{w_i}) &= \boldsymbol{x}_{ij}^\top \boldsymbol{\beta}_\gamma^{w_i} + b_i, \end{aligned}$$

where $\boldsymbol{\beta}_\pi^{w_i}$ is the regression parameter vector (fixed effects) for the covariates $\boldsymbol{z_i}$ depending on the cluster w_i the subject i belongs to; $\boldsymbol{\beta}_{\lambda_0}^{w_i}$, $\boldsymbol{\beta}_{\lambda_1}^{w_i}$, $\boldsymbol{\beta}_\mu^{w_i}$, $\boldsymbol{\beta}_{\alpha_0}^{w_i}$, $\boldsymbol{\beta}_{\alpha_1}^{w_i}$ and $\boldsymbol{\beta}_\gamma^{w_i}$ are the regression parameter vectors (fixed effects) for the covariate vectors $\tilde{\boldsymbol{x}}_{ij}$, $\breve{\boldsymbol{x}}_{ij}$ and $\boldsymbol{x}_{ij}$, and b_i is a random

intercept (random effects) for the subject i. In this proposed models, the chosen link functions $g_\pi(.)$, $g_{\lambda_0}(.)$, $g_{\lambda_1}(.)$, $g_\mu(.)$, $g_{\alpha_0}(.)$, $g_{\alpha_1}(.)$ and $g_\gamma(.)$ are the logit function but other functions could be used.

The random intercepts are assumed to be independent and identically distributed with normal distribution:

$$b_i \sim \mathrm{N}(0, \sigma_b^2),$$

where σ_b^2 is the variance of the normal distribution.

To represent a finite mixture, a latent random variable W is used in (3.1) and (3.2). According to Frühwirth-Schnatter et al. (2019) this implies that first the group (cluster) w is drawn from $1, 2, ..., K$ with probabilities $p_1, p_2, ..., p_K$. Then, given the group w, the responses y and d are drawn from their respective distributions. In this case, conditional on W, Y follows the ZOIB or BIm distribution and D follows the Bernoulli distribution. The latent random variable W is unobserved but its inclusion to the model is important for modelling dependencies.

According to Gelman et al. (2013) the joint distribution of the observed data $\boldsymbol{Y}$ and $\boldsymbol{D}$, the unobserved indicators $\boldsymbol{W} = [w_1, ..., w_n]^\top$ and the random intercepts $\boldsymbol{b} = [b_1, ..., b_n]^\top$, conditional on the model parameters $\boldsymbol{\theta} = [(\boldsymbol{\theta}^w)^\top, \sigma_b^2, \boldsymbol{p}^\top]^\top$, where $\boldsymbol{\theta}^w = [\boldsymbol{\beta}_\pi^w, \boldsymbol{\beta}_{\alpha_0}^w, \boldsymbol{\beta}_{\alpha_1}^w, \boldsymbol{\beta}_\gamma^w, \phi^w]^\top$ or $\boldsymbol{\theta}^w = [\boldsymbol{\beta}_\pi^w, \boldsymbol{\beta}_{\lambda_0}^w, \boldsymbol{\beta}_{\lambda_1}^w, \boldsymbol{\beta}_\mu^w, \phi^w]^\top$ depending on the parametrization, $w = 1, ..., K$ and $\boldsymbol{p} = [p_1, p_2, ..., p_K]$, can be written as

$$\begin{aligned}
P(\boldsymbol{Y}, \boldsymbol{D}, \boldsymbol{W}, \boldsymbol{b} \mid \boldsymbol{\theta}) =& P(\boldsymbol{Y}, \boldsymbol{D}, \boldsymbol{b} \mid \boldsymbol{W}, \boldsymbol{\theta}) P(\boldsymbol{W} \mid \boldsymbol{\theta}) \\
=& \prod_{i=1}^{n} \prod_{k=1}^{K} \left[p_k f_{D_i}(d_i \mid \boldsymbol{\theta}) f_{Y_i}(\boldsymbol{y}_i \mid \boldsymbol{\theta}, b_i) \right]^{I(w_i = k)} \times \varphi(b_i \mid 0, \sigma_b^2) \\
=& \prod_{i=1}^{n} \prod_{k=1}^{K} \left[p_k f_{D_i}(d_i \mid \pi_i^{w_i}) \prod_{j=1}^{n_i} f_{Y_{ij}}(y_{ij} \mid \boldsymbol{\theta}, b_i) \right]^{I(w_i = k)} \times \varphi(b_i \mid 0, \sigma_b^2),
\end{aligned}$$

where Y and D, conditional on the cluster W, are assumed independent. Also $\varphi(\cdot \mid \mu, \sigma^2)$ is the probability density function of a normal distribution with mean μ and variance σ^2, the indicator function I is defined as

$$I(w_i = k) = \begin{cases} 1, & w_i = k \\ 0, & \text{otherwise} \end{cases}$$

and the Bernoulli probability mass function as

$$f_D(d \mid \pi) = \begin{cases} 1 - \pi, & d = 0. \\ \pi, & d = 1. \end{cases}$$

Finally, the augmented likelihood functions of models (3.1) and (3.2) are

$$L(\boldsymbol{\theta}, \boldsymbol{W}, \boldsymbol{b} \mid \boldsymbol{Y}, \boldsymbol{D}) = \prod_{i=1}^{n} \prod_{k=1}^{K} \left[p_k f_D(d_i \mid \pi_i^{w_i}) \prod_{j=1}^{n_i} g_Y(y_{ij} \mid \lambda_{0ij}^{w_i}, \lambda_{1ij}^{w_i}, \mu_{ij}^{w_i}, \phi^{w_i}, b_i) \right]^{I(w_i=k)} \times \varphi(b_i \mid 0, \sigma_b^2) \tag{3.3}$$

and

$$L(\boldsymbol{\theta}, \boldsymbol{W}, \boldsymbol{b} \mid \boldsymbol{Y}, \boldsymbol{D}) = \prod_{i=1}^{n} \prod_{k=1}^{K} \left[p_k f_D(d_i \mid \pi_i^{w_i}) \prod_{j=1}^{n_i} h_Y(y_{ij} \mid \alpha_{0ij}^{w_i}, \alpha_{1ij}^{w_i}, \gamma_{ij}^{w_i}, \phi^{w_i}, b_i) \right]^{I(w_i=k)} \times \varphi(b_i \mid 0, \sigma_b^2), \tag{3.4}$$

respectively, where $f_{D_i}(d_i \mid \pi_i^{w_i})$ is the probability mass function of the Bernoulli distribution with parameter $\pi_i^{w_i}$, $g_{Y_{ij}}(y_{ij} \mid \lambda_{0ij}^{w_i}, \lambda_{1ij}^{w_i}, \mu_{ij}^{w_i}, \phi^{w_i}, b_i)$ is the probability density function of the ZOIB distribution with parameters $\lambda_{0ij}^{w_i}, \lambda_{1ij}^{w_i}, \mu_{ij}^{w_i}, \phi^{w_i}$ and random intercept b_i, and $h_{Y_{ij}}(y_{ij} \mid \alpha_{0ij}^{w_i}, \alpha_{1ij}^{w_i}, \gamma_{ij}^{w_i}, \phi^{w_i}, b_i)$ is the probability density function of the BIm distribution with parameters $\alpha_{0ij}^{w_i}, \alpha_{1ij}^{w_i}, \gamma_{ij}^{w_i}, \phi^{w_i}$ and random intercept b_i, depending on the cluster w_i the subject i belongs to and with p_k as the probability of belonging to the cluster k.

3.2. Bayesian inference

The augmented posterior distribution of $\boldsymbol{\theta}$, $\boldsymbol{W}$ and $\boldsymbol{b}$ can be written as

$$P(\boldsymbol{\theta}, \boldsymbol{W}, \boldsymbol{B} \mid \boldsymbol{Y}, \boldsymbol{D}) \propto P(\boldsymbol{Y}, \boldsymbol{D}, \boldsymbol{W}, \boldsymbol{b} \mid \boldsymbol{\theta}) \times P(\boldsymbol{\theta}),$$

which can also be expressed as

$$P(\boldsymbol{\theta}, \boldsymbol{W}, \boldsymbol{b} \mid \boldsymbol{Y}, \boldsymbol{D}) \propto L(\boldsymbol{\theta}, \boldsymbol{W}, \boldsymbol{b} \mid \boldsymbol{Y}, \boldsymbol{D}) \times P(\boldsymbol{\theta}), \tag{3.5}$$

where L is the augmented likelihood function and $P(\boldsymbol{\theta})$ is the prior distribution of $\boldsymbol{\theta}$. In this thesis the parameters in $\boldsymbol{\theta}$ are considered independent, so the prior distribution is the following:

$$P(\boldsymbol{\theta}) = P(\sigma_b^2) \times P(\boldsymbol{p}) \times \prod_{w=1}^{K} P(\boldsymbol{\theta}^w),$$

where

$$P(\boldsymbol{\theta}^w) = P(\boldsymbol{\beta}_\pi^w) \times P(\boldsymbol{\beta}_{\lambda_0}^w) \times P(\boldsymbol{\beta}_{\lambda_1}^w) \times P(\boldsymbol{\beta}_\mu^w) \times P(\phi^w)$$

or

$$P(\boldsymbol{\theta}^w) = P(\boldsymbol{\beta}_\pi^w) \times P(\boldsymbol{\beta}_{\alpha_0}^w) \times P(\boldsymbol{\beta}_{\alpha_1}^w) \times P(\boldsymbol{\beta}_\gamma^w) \times P(\phi^w),$$

depending on the parametrization. For each fixed effect vector on each cluster, a multivariate normal distribution is proposed such that

$$
\begin{aligned}
\boldsymbol{\beta}_{\pi}^{w} &\sim \mathrm{N}_a(0, \boldsymbol{A^w}) \\
\boldsymbol{\beta}_{\lambda_0}^{w} &\sim \mathrm{N}_b(0, \boldsymbol{B^w}) \\
\boldsymbol{\beta}_{\lambda_1}^{w} &\sim \mathrm{N}_c(0, \boldsymbol{C^w}) \\
\boldsymbol{\beta}_{\mu}^{w} &\sim \mathrm{N}_d(0, \boldsymbol{D^w}) \\
\boldsymbol{\beta}_{\alpha_0}^{w} &\sim \mathrm{N}_e(0, \boldsymbol{E^w}) \\
\boldsymbol{\beta}_{\alpha_1}^{w} &\sim \mathrm{N}_f(0, \boldsymbol{F^w}) \\
\boldsymbol{\beta}_{\gamma}^{w} &\sim \mathrm{N}_g(0, \boldsymbol{G^w}),
\end{aligned}
$$

where $w = 1, ..., K$; a, b, c, d, e, f and g are the number of covariates including an intercept on the estimation of parameters π_i^w, λ_{0ij}^w, λ_{1ij}^w, μ_{ij}^w, α_{0ij}^w, α_{1ij}^w and γ_{ij}^w, respectively and $\boldsymbol{A^w}$, $\boldsymbol{B^w}$, $\boldsymbol{C^w}$, $\boldsymbol{D^w}$, $\boldsymbol{E^w}$, $\boldsymbol{F^w}$ and $\boldsymbol{G^w}$ are their corresponding covariance matrices.

For the random intercept variance and the precision parameter ϕ^w on each cluster, an inverse gamma distribution is proposed

$$
\begin{aligned}
\sigma_b^2 &\sim \text{Inv-Gamma}(l, m) \\
\phi^w &\sim \text{Inv-Gamma}(q^w, r^w).
\end{aligned}
$$

Finally, for the probability of cluster membership, a Dirichlet distribution is proposed

$$
\boldsymbol{p} \sim \mathrm{Dir}(\boldsymbol{u}),
$$

where $\boldsymbol{u} = [u_1, u_2, ..., u_K]^\top > 0$. For all these prior distributions, $\boldsymbol{A^w}$, $\boldsymbol{B^w}$, $\boldsymbol{C^w}$, $\boldsymbol{D^w}$, $\boldsymbol{E^w}$, $\boldsymbol{F^w}$, $\boldsymbol{G^w}$, a, b, c, d, e, f, g, l, m, q^w, r^w, $\boldsymbol{u}$ and K are specified hyperparameters. Note here that we have to choose K, if one knows the number of clusters beforehand, then K should be set to that value, but in most cases K is unknown, thus different values of K have to be tested. Further discussion about the choice of K are included in sections 4.3 and 5.3.

By replacing all the proposed prior distributions, the augmented posterior distributions for ZOIB and BIm parametrizations become:

$$
\begin{aligned}
P(\boldsymbol{\theta}, \boldsymbol{W}, \boldsymbol{b} \mid \boldsymbol{Y}, \boldsymbol{D}) \propto & \prod_{i=1}^{n} \prod_{k=1}^{K} \left[p_k f_{D_i}(d_i \mid \pi_i^{w_i}) \prod_{j=1}^{n_i} g_{Y_{ij}}(y_{ij} \mid \lambda_{0ij}^{w_i}, \lambda_{1ij}^{w_i}, \mu_{ij}^{w_i}, \phi^{w_i}, b_i) \right]^{I(w_i = k)} \\
& \times \varphi(b_i \mid 0, \sigma_b^2) \times IG(\sigma_b^2 \mid l, m) \times DIR(\boldsymbol{p} \mid \boldsymbol{u}) \\
& \times \prod_{w=1}^{K} \left[\varphi_a(\boldsymbol{\beta_\pi^w} \mid 0, \boldsymbol{A^w}) \times \varphi_b(\boldsymbol{\beta_{\lambda_0}^w} \mid 0, \boldsymbol{B^w}) \times \varphi_c(\boldsymbol{\beta_{\lambda_1}^w} \mid 0, \boldsymbol{C^w}) \right. \\
& \left. \times \varphi_d(\boldsymbol{\beta_\mu^w} \mid 0, \boldsymbol{D^w}) \times IG(\phi^w \mid q^w, r^w) \right]
\end{aligned}
\tag{3.6}
$$

and

$$
\begin{aligned}
P(\boldsymbol{\theta}, \boldsymbol{W}, \boldsymbol{b} \mid \boldsymbol{Y}, \boldsymbol{D}) \propto & \prod_{i=1}^{n} \prod_{k=1}^{K} \left[p_k f_{D_i}(d_i \mid \pi_i^{w_i}) \prod_{j=1}^{n_i} h_{Y_{ij}}(y_{ij} \mid \alpha_{0ij}^{w_i}, \alpha_{1ij}^{w_i}, \gamma_{ij}^{w_i}, \phi^{w_i}, b_i) \right]^{I(w_i=k)} \\
& \times \varphi(b_i \mid 0, \sigma_b^2) \times IG(\sigma_b^2 \mid l, m) \times DIR(\boldsymbol{p} \mid \boldsymbol{u}) \\
& \times \prod_{w=1}^{K} \left[\varphi_a(\boldsymbol{\beta}_{\boldsymbol{\pi}}^{\boldsymbol{w}} \mid 0, \boldsymbol{A}^{\boldsymbol{w}}) \times \varphi_e(\boldsymbol{\beta}_{\boldsymbol{\alpha_0}}^{\boldsymbol{w}} \mid 0, \boldsymbol{E}^{\boldsymbol{w}}) \times \varphi_f(\boldsymbol{\beta}_{\boldsymbol{\alpha_1}}^{\boldsymbol{w}} \mid 0, \boldsymbol{F}^{\boldsymbol{w}}) \right. \\
& \left. \times \varphi_g(\boldsymbol{\beta}_{\boldsymbol{\gamma}}^{\boldsymbol{w}} \mid 0, \boldsymbol{G}^{\boldsymbol{w}}) \times IG(\phi^w \mid q^w, r^w) \right],
\end{aligned}
\tag{3.7}
$$

respectively for each parametrization, where $\varphi_a(\cdot \mid \boldsymbol{\mu}, \boldsymbol{A})$ is the probability density function of a multivariate normal distribution with mean vector $\boldsymbol{\mu}$ and covariance square $a \times a$ matrix $\boldsymbol{A}$, $\text{IG}(\cdot \mid a, b)$ is the probability density function of an inverse gamma distribution with parameters a and b and $\text{DIR}(\cdot \mid \boldsymbol{a})$ is the probability density function of a Dirichlet distribution with vector parameter $\boldsymbol{a}$.

Is too complex to get samples from (3.6) or (3.7) with the Gibbs Algorithm because neither the augmented posterior distribution nor its conditional distributions are associable to any known statistical distribution. Thus, in order to sample from both augmented posterior distributions, according to Coro (2017), a Gibbs sampler that uses complex strategies such as slice sampling, adaptive rejection sampling and Metropolis-Hastings algorithm can be used. For this reason, JAGS (Just Another Gibbs Sampler) (Plummer (2012)), a software that uses the complex strategies mentioned before, will be used in this thesis throughout the R package RJAGS developed by Plummer et al. (2018).

3.3. Classification of new subjects

Assuming that for a new subject its fractional longitudinal trajectory $\boldsymbol{y}_{new}$ is known but its binary response variable D_{new} is yet unobserved, we can marginally calculate the new subject's probability $P(D_{new} = 1 \mid \boldsymbol{y}_{new}, \boldsymbol{\theta})$ over its cluster membership by using

$$
P(D_{new} = 1 \mid \boldsymbol{y}_{new}, \boldsymbol{\theta}) = \sum_{k=1}^{K} P(D_{new} = 1 \mid w_{new} = k, \boldsymbol{\theta}) P(w_{new} = k \mid \boldsymbol{y}_{new}, \boldsymbol{\theta}),
$$

due to the Law of total probability. Then, it can be expanded by Bayes's rule as

$$
\begin{aligned}
& P(D_{new} = 1 \mid \boldsymbol{y}_{new}, \boldsymbol{\theta}) = \\
& \qquad \frac{\sum_{k=1}^{K} P(D_{new} = 1 \mid w_{new} = k, \boldsymbol{\theta}) P(\boldsymbol{y}_{new} \mid w_{new} = k, \boldsymbol{\theta}) P(w_{new} = k \mid \boldsymbol{\theta})}{\sum_{w=1}^{K} P(\boldsymbol{y}_{new} \mid w_{new} = w, \boldsymbol{\theta}) P(w_{new} = w \mid \boldsymbol{\theta})}.
\end{aligned}
\tag{3.8}
$$

Notice that $P(\boldsymbol{y}_{new} \mid w_{new} = k, \boldsymbol{\theta})$ is the ZOIB or BIm distribution probability density function and it must be marginally calculated over its random effect, which can be obtained by the following integrals:

$$P(\boldsymbol{y}_{new} \mid w_{new} = k, \boldsymbol{\theta}) = \int_{-\infty}^{\infty} \prod_{j=1}^{n_{new}} g_{Y_{ij}}(y_{ij} \mid \lambda_{0ij}^k, \lambda_{1ij}^k, \mu_{ij}^k, \phi^k, b_i) \times \varphi(b_i \mid 0, \sigma_b^2)\, db_i \tag{3.9}$$

and

$$P(\boldsymbol{y}_{new} \mid w_{new} = k, \boldsymbol{\theta}) = \int_{-\infty}^{\infty} \prod_{j=1}^{n_{new}} h_{Y_{ij}}(y_{ij} \mid \alpha_{0ij}^k, \alpha_{1ij}^k, \gamma_{ij}^k, \phi^k, b_i) \times \varphi(b_i \mid 0, \sigma_b^2)\, db_i, \tag{3.10}$$

respectively, where $g_{Y_{ij}}(. \mid \lambda_{0ij}, \lambda_{1ij}, \mu_{ij}, \phi, b_i)$ is the ZOIB distribution probability density function with parameters $\lambda_{0ij}, \lambda_{1ij}, \mu_{ij}, \phi$ and random intercept b_i; $h_{Y_{ij}}(. \mid \alpha_{0ij}, \alpha_{1ij}, \gamma_{ij}, \phi, b_i)$ is the BIm distribution probability density function with parameters $\alpha_{0ij}, \alpha_{1ij}, \gamma_{ij}, \phi$ and random intercept b_i, and $\varphi(. \mid \mu, \sigma^2)$ is the normal probability density function with mean μ and variance σ^2.

It is complex to solve these integrals analytically and computational limitations arise when performing integral numerical approximations due to the rigidity of the BIm distribution. Therefore, for approximating (3.9) and (3.10), n_b values of b_i are simulated from a normal distribution with mean $\mu = 0$ and variance $\sigma^2 = \sigma_b^2$, where the value of σ_b^2 varies at each iteration according to the sampled values of the MCMC. Then $\prod_{j=1}^{n_{new}} g_{Y_{ij}}(y_{ij} \mid \lambda_{0ij}^k, \lambda_{1ij}^k, \mu_{ij}^k, \phi^k, b_i)$ and $\prod_{j=1}^{n_{new}} h_{Y_{ij}}(y_{ij} \mid \alpha_{0ij}^k, \alpha_{1ij}^k, \gamma_{ij}^k, \phi^k, b_i)$ are evaluated at each simulated b_i value and the mean of the evaluated functions are computed.

Since $\boldsymbol{\theta}$ is not a fixed value, the default probabilities are approximated at each MCMC iteration m. Then, the mean is calculated as

$$P(D_{new} = 1 \mid \ Y_{new}, \boldsymbol{\theta}) = \frac{1}{M} \sum_{m=1}^{M} \frac{\sum_{k=1}^{K} P(D_{new} = 1 \mid w_{new} = k, \boldsymbol{\theta}_m) P(Y_{new} \mid w_{new} = k, \boldsymbol{\theta}_m) P(w_{new} = k \mid \boldsymbol{\theta}_m)}{\sum_{w=1}^{K} P(Y_{new} \mid w_{new} = w, \boldsymbol{\theta}_m) P(w_{new} = w \mid \boldsymbol{\theta}_m)}, \tag{3.11}$$

where M is the total number of MCMC samples.

3.4. Model comparison criteria

Comparisons of different criteria have to be made to determine which is the best model. Goodness of fit will determine which model fits better the given data and, since we are dealing with classification models, the prediction power will also be compared.

3.4.1. Goodness of fit

Most of the goodness of fit comparison criteria for models obtained by MCMC simulation are based on the concept of deviance $\mathcal{D}(.)$, which is defined as:

$$\mathcal{D}(\nu) = -2log\Big(L(\nu \mid y)\Big) + C, \tag{3.12}$$

where L is the likelihood function of the augmented posterior distribution, conditional on the

observed data, y and C is a constant related to the saturated model, but cancels out when doing model comparisons.

Using this deviance, Spiegelhalter et al. (2002) proposed a model comparison criteria named deviance information criterion (DIC), defined as:

$$\text{DIC} = \overline{\mathcal{D}(\nu)} + p_D, \tag{3.13}$$

where $p_D = \overline{\mathcal{D}(\nu)} - \mathcal{D}(\bar{\nu})$ is the number of effective parameters. For a MCMC, $\overline{\mathcal{D}(\nu)}$ is the mean of the deviance evaluated at each sampled value of ν and can be computed as $\overline{\mathcal{D}(\nu)} = (1/M)\sum_{m=1}^{M}\mathcal{D}(\nu_m)$, while $\mathcal{D}(\bar{\nu})$ is the deviance evaluated at the expectation of ν.

In this case, $\nu = [\boldsymbol{\theta}, \boldsymbol{W}, \boldsymbol{b}]$ and the observed data are $\boldsymbol{y}$ and $\boldsymbol{d}$, so the likelihoods in (3.12) are the ones defined in (3.3) and (3.4). The lower the model DIC is, the better fit it indicates, so the model with the lowest value on this criteria will be considered as the best.

3.4.2. Predictive power

For the predictive power of the model, the well known area under the receiver operating characteristic (ROC) curve (AUC) is used. This metric will measure how good is the model for making out-of-sample predictions.

3.5. Label-switching problem

As stated by Marin and Robert (2007), an important feature of a mixture model is that it is invariant under permutations of the indices of the components. This means that the component parameters are not identifiable, in the sense that labels can be exchanged between them without altering the result, for example, considering a two cluster case, the parameter vector $\boldsymbol{\theta}^\top = [(\boldsymbol{\theta}^1)^\top, (\boldsymbol{\theta}^2)^\top]^\top$, will lead to the same result if the values of the cluster parameter vectors $\boldsymbol{\theta}^1$ and $\boldsymbol{\theta}^2$ are exchanged, giving rise to the label-switching problem.

Marin et al. (2005) proposes a solution to overcome this problem when dealing with known number of components (which is our case). This solution consists on reordering the labels according to the permutation that gives the *maximum a posteriori* (MAP) in our MCMC sample. For example, given the total number of samples M of the MCMC, the MAP approximation will be given in the iteration m^* such that

$$m^* = \arg\max_{m=1,\dots,M}\big[P(\boldsymbol{\theta}_m, \boldsymbol{W}_m, \boldsymbol{b}_m \mid \boldsymbol{Y}, \boldsymbol{D})\big],$$

where $P(\boldsymbol{\theta}_m, \boldsymbol{W}_m, \boldsymbol{b}_m \mid \boldsymbol{Y}, \boldsymbol{D})$ is the posterior distribution evaluated with the m-sampled parameters and the observed data $\boldsymbol{Y}$ and $\boldsymbol{D}$. Thus, according to Marin and Robert (2007) the approximate MAP estimate will act as a *pivot* since it gives a good approximation to a mode of the posterior distribution and we can reorder the other iterations with respect to this mode.

This method is known as the pivotal reordering algorithm (PRA) and will be used to solve the identifiability problem throughout the R package *label.switching* created by Papastamoulis (2015).

Chapter 4

Simulation study

This chapter presents a simulation study for parameter recovery, goodness of fit, predictive power and cluster recovery. The classification beta inflated regressions with cluster formation presented in Chapter 3 compete against each other and against the BTran model, presented later in this chapter.

4.1. Generation of data

For this simulation study 50 datasets are generated, where each dataset consist of 400 subjects and each subject has 6 longitudinal observations, giving a total of 2400 observations (the same number as the application to real data in Chapter 5). Different combinations of number of subjects and longitudinal observations where also considered, the results can be found in Appendix A.

For the fixed effects associated with the mixed fractional response Y a tridimensional array $\boldsymbol{X}$ is constructed:

$$\boldsymbol{X} = \begin{bmatrix} \boldsymbol{x_{1,1}} & \boldsymbol{x_{1,2}} & \cdots & \boldsymbol{x_{1,6}} \\ \boldsymbol{x_{2,1}} & \boldsymbol{x_{2,2}} & \cdots & \boldsymbol{x_{2,6}} \\ \vdots & \vdots & \ddots & \vdots \\ \boldsymbol{x_{400,1}} & \boldsymbol{x_{400,2}} & \cdots & \boldsymbol{x_{400,6}} \end{bmatrix},$$

where each $\boldsymbol{x_{i,j}}$ is a vector of length 3 such that:

$$\boldsymbol{x_{i,j}} = \begin{bmatrix} 1 \\ j \\ x_{i,j,3} \end{bmatrix}.$$

The first element of $\boldsymbol{x_{i,j}}$ represents the intercept and is constant and equal to 1, the second element j represents the time of the measurement and the third element $x_{i,j,3}$ is sampled from a uniform distribution $x_{i,j,3} \sim U(-1.5, 1.5)$ and represents a uniform distributed covariate related to the subject i at time j. For the fixed effects associated with binary response D a matrix $\boldsymbol{Z}$ is constructed:

$$\boldsymbol{Z} = \begin{bmatrix} 1 & z_{1,2} \\ 1 & z_{2,2} \\ \vdots & \vdots \\ 1 & z_{400,2} \end{bmatrix},$$

where $z_{i,1}$ represents the intercept and is constant and equal to 1 and $z_{i,2}$ is sampled from a uniform distribution $z_{i,2} \sim U(-1.5, 1.5)$. Since the random effects are just intercepts, there is no need of a design matrix for them in this model.

4.2. Parameter recovery

For the parameter recovery in the simulation study, the number of clusters is set to 2, with probabilities of belonging to each cluster $p_1 = 0.65$ and $p_2 = 0.35$. So a vector $\boldsymbol{w} = [w_1, ..., w_{400}]$ is generated with 400 samples of $W \sim \text{Cat}(p_1, p_2)$. The parameter π^w is modelled depending on two covariates and λ_0^w, λ_1^w, μ^w, α_0^w, α_1^w and γ^w are modelled depending on three covariates each. The fixed effects coefficients and the variances for the random intercepts selected for this simulation study are shown on Table 4.1.

Parameter (Cluster 1)	**Value**	**Parameter (Cluster 2)**	**Value**
$\boldsymbol{\beta}_\pi^1$	[-0.50, 2.00]$^\top$	$\boldsymbol{\beta}_\pi^2$	[-0.50, 1.00]$^\top$
$\boldsymbol{\beta}_{\lambda_0}^1$	[-2.00, -0.70, 0.70]$^\top$	$\boldsymbol{\beta}_{\lambda_0}^2$	[-1.25, 1.00, -0.90]$^\top$
$\boldsymbol{\beta}_{\lambda_1}^1$	[-2.00, 0.70, 0.50]$^\top$	$\boldsymbol{\beta}_{\lambda_1}^2$	[-1.50, -0.70, 0.90]$^\top$
$\boldsymbol{\beta}_\mu^1$	[-0.75, 0.90, 0.80]$^\top$	$\boldsymbol{\beta}_\mu^2$	[-0.25, -0.80, 0.85]$^\top$
$\boldsymbol{\beta}_{\alpha_0}^1$	[-0.50, -0.80, 1.20]$^\top$	$\boldsymbol{\beta}_{\alpha_0}^2$	[-1.10, 0.85, -1.50]$^\top$
$\boldsymbol{\beta}_{\alpha_1}^1$	[-1.10, 0.50, 0.60]$^\top$	$\boldsymbol{\beta}_{\alpha_1}^2$	[-0.90, -0.50, 1.00]$^\top$
$\boldsymbol{\beta}_\gamma^1$	[-0.80, 0.90, 0.80]$^\top$	$\boldsymbol{\beta}_\gamma^2$	[-0.30, -0.80, 0.85]$^\top$

Table 4.1: Parameter values for the simulation study

The 400 random intercepts b_i are sampled from a normal distribution with mean 0 and variance $\sigma_b^2 = 0.5$. The precision parameters ϕ^w are set to 50 for both clusters, so we will use just ϕ to simplify notation. Finally, the 400 binary responses for the 400 subjects are sampled from

$$D_i \sim \text{Bern}(\pi_i^{w_i})$$

and 6 longitudinal observations for each of the 400 subjects are sampled from

$$Y_{ij} \sim \text{ZOIB}(\lambda_{0ij}^{w_i}, \lambda_{1ij}^{w_i}, \mu_{ij}^{w_i}, \phi)$$

and again from

$$Y_{ij} \sim \text{BIm}(\alpha_{0ij}^{w_i}, \alpha_{1ij}^{w_i}, \gamma_{ij}^{w_i}, \phi),$$

taking into account the cluster w_i of the *ith* element and where

$$
\begin{aligned}
\text{logit}(\pi_i^{w_i}) &= \boldsymbol{z_i}^\top \boldsymbol{\beta_\pi^{w_i}} \\
\text{logit}(\lambda_{0ij}^{w_i}) &= \boldsymbol{x_{ij}}^\top \boldsymbol{\beta_{\lambda_0}^{w_i}} \\
\text{logit}(\lambda_{1ij}^{w_i}) &= \boldsymbol{x_{ij}}^\top \boldsymbol{\beta_{\lambda_1}^{w_i}} \\
\text{logit}(\mu_{ij}^{w_i}) &= \boldsymbol{x_{ij}}^\top \boldsymbol{\beta_{\mu}^{w_i}} + b_i \\
\text{logit}(\alpha_{0ij}^{w_i}) &= \boldsymbol{x_{ij}}^\top \boldsymbol{\beta_{\alpha_0}^{w_i}} \\
\text{logit}(\alpha_{1ij}^{w_i}) &= \boldsymbol{x_{ij}}^\top \boldsymbol{\beta_{\alpha_1}^{w_i}} \\
\text{logit}(\gamma_{ij}^{w_i}) &= \boldsymbol{x_{ij}}^\top \boldsymbol{\beta_{\gamma}^{w_i}} + b_i.
\end{aligned}
$$

In order to visualize the simulated data, 200 subject responses are sampled from a random ZOIB dataset and a random BIm dataset and are plotted without cluster information in Figure 4.1. The ZOIB sampled responses are plotted by cluster in Figure 4.2 and the BIm sampled responses are plotted by cluster in Figure 4.3. The spaghetti plots in Figure 4.1 does not show a clear difference between the simulated datasets but shows us how the data looks like when is given to the model, with no easily identifiable pattern at first sight. The spaghetti plots in Figures 4.2 and 4.3 shows us how the models are expected to split the subjects into different clusters considering the similarities of their response variable y trajectory and its binary value d. Table 4.2 shows the frequency of zeros, ones and values in the open (0,1) interval of the simulated y responses by model and cluster.

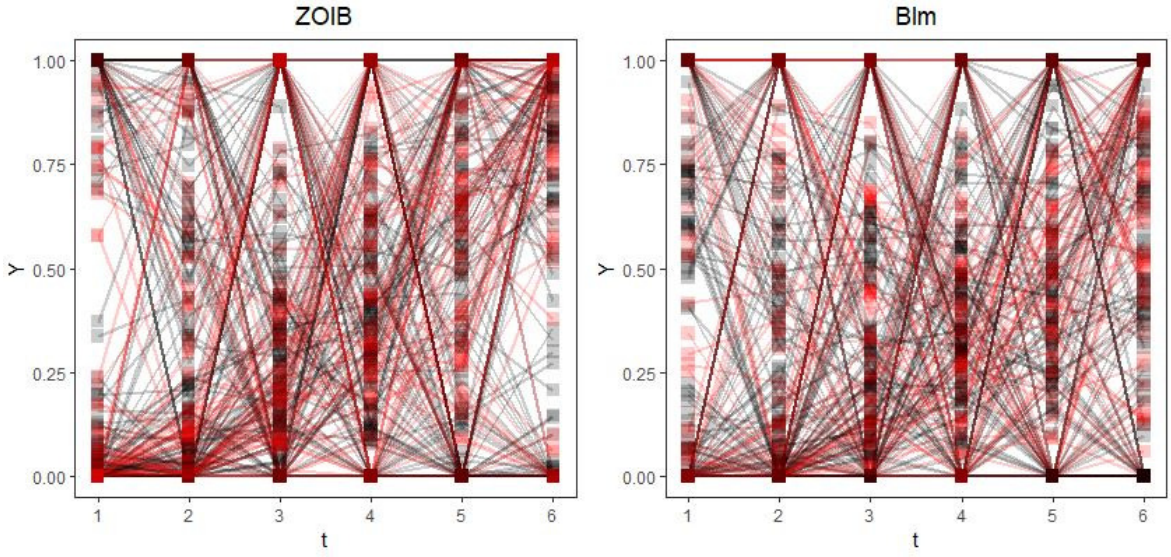


Figure 4.1: Spaghetti plots of longitudinal trajectory from a ZOIB and a BIm simulated dataset with 200 sampled subjects each. The red color represents subjects with $d = 1$, while black represents subjects with $d = 0$.

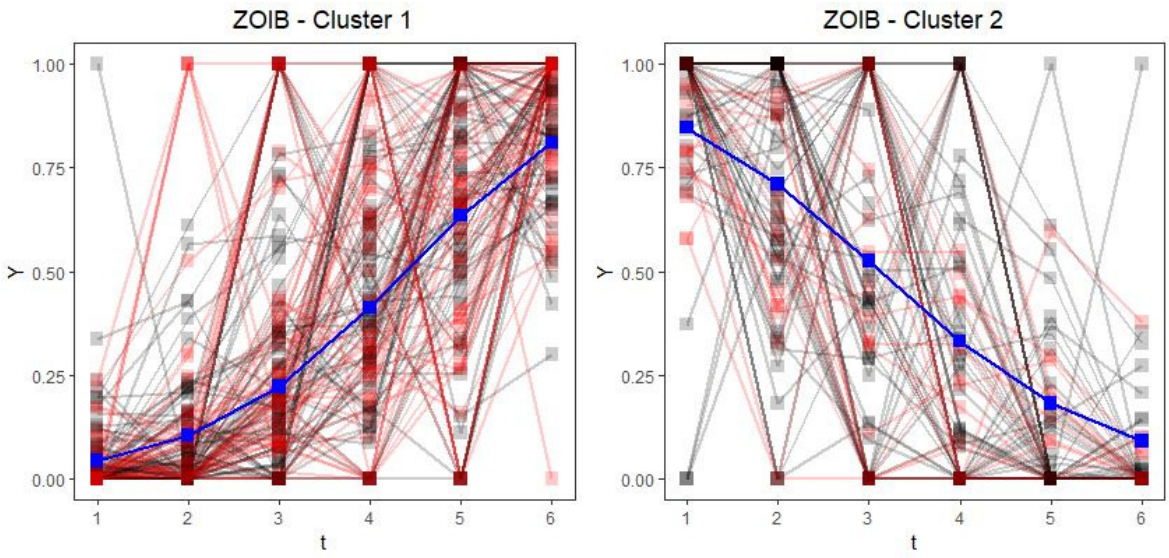


Figure 4.2: Spaghetti plots of longitudinal trajectory by cluster from a ZOIB simulated dataset with a total of 200 sampled subjects. The blue line is the conditional mean μ at each time. The red color represents subjects with $d = 1$, while black represents subjects with $d = 0$.

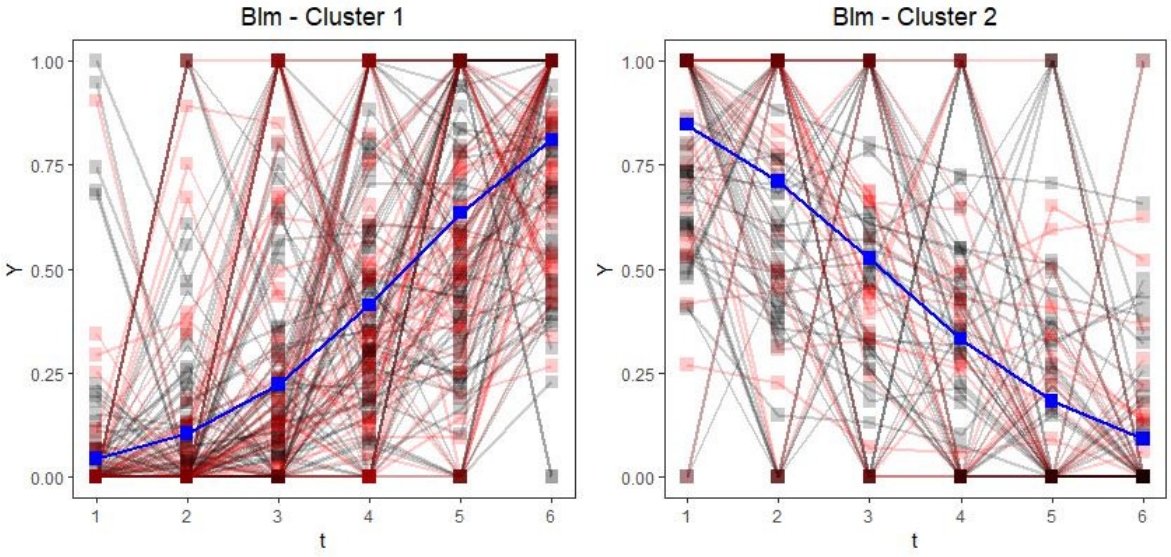


Figure 4.3: Spaghetti plots of longitudinal trajectory by cluster from a BIm simulated dataset with a total of 200 sampled subjects. The blue line is the mean γ at each time. The red color represents subjects with $d = 1$, while black represents subjects with $d = 0$.

Model	Cluster	Value	Frequency %
ZOIB	1	0	18 %
		1	16 %
		(0,1)	66 %
	2	0	33 %
		1	21 %
		(0,1)	46 %
BIm	1	0	29 %
		1	15 %
		(0,1)	56 %
	2	0	26 %
		1	20 %
		(0,1)	54 %

Table 4.2: Frequency of zeros, ones and values in the open (0,1) interval of the simulated y responses

Regarding the estimation, the prior distributions are selected to be non-informative, thus all the fixed coefficients are set with the following priors

$$\begin{aligned}
\boldsymbol{\beta}_{\boldsymbol{\pi}}^{\boldsymbol{w}} &\sim \mathrm{N}_2(0, 10^4 I_2) \\
\boldsymbol{\beta}_{\boldsymbol{\lambda_0}}^{\boldsymbol{w}} &\sim \mathrm{N}_3(0, 10^4 I_3) \\
\boldsymbol{\beta}_{\boldsymbol{\lambda_1}}^{\boldsymbol{w}} &\sim \mathrm{N}_3(0, 10^4 I_3) \\
\boldsymbol{\beta}_{\boldsymbol{\mu}}^{\boldsymbol{w}} &\sim \mathrm{N}_3(0, 10^4 I_3) \\
\boldsymbol{\beta}_{\boldsymbol{\alpha_0}}^{\boldsymbol{w}} &\sim \mathrm{N}_3(0, 10^4 I_3) \\
\boldsymbol{\beta}_{\boldsymbol{\alpha_1}}^{\boldsymbol{w}} &\sim \mathrm{N}_3(0, 10^4 I_3) \\
\boldsymbol{\beta}_{\boldsymbol{\gamma}}^{\boldsymbol{w}} &\sim \mathrm{N}_3(0, 10^4 I_3),
\end{aligned}$$

the variance of the random intercepts and ϕ are set with

$$\begin{aligned}
\sigma_b^2 &\sim \text{Inv-Gamma}(0.001, 0.001) \\
\phi &\sim \text{Inv-Gamma}(0.001, 0.001),
\end{aligned}$$

and the probabilities of cluster membership $\boldsymbol{p}$ are set with

$$\boldsymbol{p} \sim \mathrm{Dir}(1, 1).$$

The estimation was performed using JAGS software in R through the package RJAGS, discarding the first 1000 iterations and sampling the next 3000 iterations considering a thinning interval equal to 6. The parameter recovery results for ZOIB and BIm models are shown in Tables 4.3 and 4.4, respectively, where it can be seen that all parameters' true values lie in

the credible intervals of their corresponding parameter estimation and that point estimation is really close to the real value.

Cluster	Parameter	True value	Mean	P2.5 %	P97.5 %
1	$\beta^1_{\lambda_0 1}$	-2.00	-2.01022	-2.16033	-1.82350
	$\beta^1_{\lambda_0 2}$	-0.70	-0.70251	-0.77599	-0.64327
	$\beta^1_{\lambda_0 3}$	0.70	0.70149	0.54973	0.82874
	$\beta^1_{\lambda_1 1}$	-2.00	-2.04531	-2.20819	-1.87471
	$\beta^1_{\lambda_1 2}$	0.70	0.70438	0.63239	0.78832
	$\beta^1_{\lambda_1 3}$	0.50	0.50431	0.38139	0.62899
	$\beta^1_{\mu 1}$	-0.75	-0.74537	-0.83261	-0.68272
	$\beta^1_{\mu 2}$	0.90	0.89989	0.88872	0.91368
	$\beta^1_{\mu 3}$	0.80	0.80244	0.78527	0.82065
	$\beta^1_{\pi 1}$	-0.55	-0.54264	-0.82723	-0.34246
	$\beta^1_{\pi 2}$	2.10	2.10126	1.61484	2.79065
	p_1	0.65	0.65937	0.62094	0.69313
	σ^2_b	0.50	0.49958	0.46174	0.52034
	ϕ	50.00	50.15587	46.17712	52.80843
2	$\beta^2_{\lambda_0 1}$	-1.25	-1.24258	-1.52979	-1.03545
	$\beta^2_{\lambda_0 2}$	1.00	1.00417	0.90143	1.11648
	$\beta^2_{\lambda_0 3}$	-0.90	-0.91976	-1.07734	-0.73711
	$\beta^2_{\lambda_1 1}$	-1.50	-1.52201	-1.87961	-1.30485
	$\beta^2_{\lambda_1 2}$	-0.70	-0.69285	-0.86827	-0.55123
	$\beta^2_{\lambda_1 3}$	0.90	0.93622	0.73304	1.21032
	$\beta^2_{\mu 1}$	-0.25	-0.26100	-0.37708	-0.11767
	$\beta^2_{\mu 2}$	-0.80	-0.80130	-0.82218	-0.77675
	$\beta^2_{\mu 3}$	0.85	0.85151	0.81910	0.88458
	$\beta^2_{\pi 1}$	-0.50	-0.51421	-0.82293	-0.25689
	$\beta^2_{\pi 2}$	1.00	0.96948	0.58919	1.29483
	p_2	0.35	0.34063	0.30687	0.37906
	σ^2_b	0.50	0.49958	0.46174	0.52034
	ϕ	50.00	50.15587	46.17712	52.80843

Table 4.3: Parameter recovery results for the ZOIB model.

Cluster	Parameter	True value	Mean	P2.5 %	P97.5 %
1	$\beta^1_{\alpha_0 1}$	0.50	-0.51784	-0.71421	-0.34385
	$\beta^1_{\alpha_0 2}$	-0.80	–0.80392	-0.88257	-0.73581
	$\beta^1_{\alpha_0 3}$	1.20	1.18153	1.13258	1.23817
	$\beta^1_{\alpha_1 1}$	-1.15	-1.14838	-1.31197	-0.98981
	$\beta^1_{\alpha_1 2}$	0.50	0.51228	0.43984	0.57756
	$\beta^1_{\alpha_1 3}$	0.60	0.60049	0.48017	0.74298
	$\beta^1_{\gamma 1}$	-0.80	-0.79418	-0.87281	-0.71445
	$\beta^1_{\gamma 2}$	0.90	0.89958	0.86483	0.93633
	$\beta^1_{\gamma 3}$	0.80	0.80662	0.76178	0.86482
	$\beta^1_{\pi 1}$	-0.50	-0.48314	-0.90162	-0.13216
	$\beta^1_{\pi 2}$	2.00	2.05952	1.70876	2.56960
	p_1	0.65	0.65007	0.60689	0.68786
	σ^2_b	0.50	0.48275	0.46068	0.50718
	ϕ	50.00	49.75871	46.19366	53.16776
2	$\beta^2_{\alpha_0 1}$	-1.10	-1.11347	-1.34095	-0.93378
	$\beta^2_{\alpha_0 2}$	0.85	0.85772	0.74302	0.96874
	$\beta^2_{\alpha_0 3}$	-1.50	-1.53203	-1.68264	-1.32170
	$\beta^2_{\alpha_1 1}$	-1.00	-0.95881	-1.24728	-0.74868
	$\beta^2_{\alpha_1 2}$	-0.50	-0.51789	-0.61756	-0.39027
	$\beta^2_{\alpha_1 3}$	1.00	1.05495	0.88531	1.24424
	$\beta^2_{\gamma 1}$	-0.30	-0.29168	-0.40016	-0.19491
	$\beta^2_{\gamma 2}$	-0.80	-0.80136	-0.85558	-0.75881
	$\beta^2_{\gamma 3}$	0.85	0.86699	0.74121	0.94757
	$\beta^2_{\pi 1}$	-0.45	-0.44694	-0.76779	-0.19615
	$\beta^2_{\pi 2}$	1.10	1.12601	0.79993	1.45874
	p_2	0.35	0.34993	0.31214	0.39311
	σ^2_b	0.50	0.48275	0.46068	0.50718
	ϕ	50.00	49.75871	46.19366	53.16776

Table 4.4: Parameter recovery results for the BIm model.

This parameter recovery exercise is not only aiming for the proposed models to correctly estimate the parameters but also to estimate them with less bias and variance than a simpli-

fied version of the model. For example, Smithson and Verkuilen (2006) proposed transforming the fractional response variable y as $y' = [y(n-1)+1/2]/n$ to constrain y' to the open interval $(0,1)$, where n is the size of the population. This transformation eliminates the need of zero-one inflation, thus the model can be reduced to

$$\begin{aligned} D_i \mid W_i = w_i &\sim \text{Bern}(\pi_i^{w_i}) \\ Y'_{ij} \mid W_i = w_i &\sim \text{Beta}(\mu_{ij}^{w_i}, \phi^{w_i}) \\ W_i &\sim \text{Cat}(\boldsymbol{p}), \end{aligned} \tag{4.1}$$

where

$$\begin{aligned} \text{logit}(\pi_i^{w_i}) &= \boldsymbol{z_i}^\top \boldsymbol{\beta_\pi^{w_i}} \\ \text{logit}(\mu_{ij}^{w_i}) &= \boldsymbol{x_{ij}}^\top \boldsymbol{\beta_\mu^{w_i}} + b_i. \end{aligned}$$

From now on (4.1) will be referred as BTran model.

The parameter recovery results on this simulation study for the BTran model applied on the ZOIB and BIm simulated datasets are shown in Tables 4.5 and 4.6, respectively, where it can be seen that no credible interval of the BTran estimation contains the true value, this indicates that BTran model induces high bias to the estimation. The precision parameter ϕ is the most affected parameter by this transformation. Since the inflation in 0 and 1 does not exists in the BTran model, this causes the beta distribution to become U-shaped (or L-shaped when inflation is just in 0 or J-shaped when inflation is just in 1) and the precision decreases drastically as can be seen in Tables 4.5 and 4.6, inducing bias in all other parameter estimations.

Another parameter recovery studies were conducted for different scenarios, varying the population size n, the number of longitudinal observations of each subject and the parameter true values. Results are shown in Appendix A in Tables A.1 and A.2. The results indicate that ZOIB and BIm estimations outperform BTran estimations in bias and length of the credible interval when dealing with parameters related to the fractional response variable Y, regardless of the population size. For the parameters related to the binary response variable D, there is not a clear difference in bias or length of credible intervals between the proposed models and the BTran model. We can conclude that doing the BTran transformation only affects the estimation of the parameters related to the fractional response Y inducing bias because of the change in the precision parameter ϕ.

4.3. Goodness of fit

Various ZOIB simulated datasets with different population size n were fitted by the ZOIB and BIm models. Table 4.7 shows that, for every value of n, ZOIB model outperformed BIm model in terms of DIC, indicating that ZOIB model fits better a ZOIB simulated dataset. In the same way, various BIm simulated datasets with different population size n were fitted by the BIm and ZOIB models. Table 4.8 shows that, for every value of n, BIm model outperformed ZOIB model in terms of DIC, indicating that BIm model fits better a BIm simulated dataset.

Cluster	Parameter	True value	Mean	P2.5 %	P97.5 %
1	$\beta^1_{\mu 1}$	-0.75	-0.25009	-0.29099	-0.21039
	$\beta^1_{\mu 2}$	0.90	0.57672	0.54712	0.61052
	$\beta^1_{\mu 3}$	0.80	0.15226	0.10857	0.19516
	$\beta^1_{\pi 1}$	-0.55	-0.54215	-0.83240	-0.34592
	$\beta^1_{\pi 2}$	2.10	2.09944	1.62850	2.77009
	p_1	0.65	0.65416	0.61486	0.68901
	σ^2_b	0.50	0.00192	0.00051	0.00591
	ϕ	50.00	1.08231	1.02166	1.16681
2	$\beta^2_{\mu 1}$	-0.25	-0.20746	-0.28400	-0.11026
	$\beta^2_{\mu 2}$	-0.80	-0.62638	-0.65811	-0.59183
	$\beta^2_{\mu 3}$	0.85	0.57959	0.53600	0.62670
	$\beta^2_{\pi 1}$	-0.50	-0.51508	-0.82826	-0.25068
	$\beta^2_{\pi 2}$	1.00	0.97894	0.58493	1.27076
	p_2	0.35	0.34584	0.31099	0.38514
	σ^2_b	0.50	0.00192	0.00051	0.00591
	ϕ	50.00	1.08231	1.02166	1.16681

Table 4.5: Parameter recovery results for the BTran model applied on the ZOIB datasets.

Cluster	Parameter	True value	Mean	P2.5 %	P97.5 %
1	$\beta^1_{\mu 1}$	-0.80	-0.39860	-0.44074	-0.34905
	$\beta^1_{\mu 2}$	0.90	0.56650	0.54105	0.60227
	$\beta^1_{\mu 3}$	0.80	0.25003	0.20158	0.29625
	$\beta^1_{\pi 1}$	-0.50	-0.49085	-0.90808	-0.15779
	$\beta^1_{\pi 2}$	2.00	2.07109	1.72739	2.54600
	p_1	0.65	0.64989	0.60285	0.68825
	σ^2_b	0.50	0.00854	0.00091	0.02932
	ϕ	50.00	1.04928	1.00410	1.11807
2	$\beta^2_{\mu 1}$	-0.30	-0.13482	-0.20560	-0.07330
	$\beta^2_{\mu 2}$	-0.80	-0.51788	-0.54762	-0.47545
	$\beta^2_{\mu 3}$	0.85	0.66568	0.58172	0.74795
	$\beta^2_{\pi 1}$	-0.45	-0.44103	-0.75414	-0.17656
	$\beta^2_{\pi 2}$	1.10	1.13380	0.80146	1.50725
	p_2	0.35	0.35011	0.31175	0.39715
	σ^2_b	0.50	0.00854	0.00091	0.02932
	ϕ	50.00	1.04928	1.00410	1.11807

Table 4.6: Parameter recovery results for the BTran model applied on the BIm datasets.

n	DIC (ZOIB)	DIC (BIm)
50	242	252
100	295	301
200	522	566
400	289	413

Table 4.7: DIC comparison between ZOIB and BIm models for different population sizes of ZOIB simulated dataset.

n	DIC (BIm)	DIC (ZOIB)
50	103	137
100	−254	−85
200	1	101
400	−1024	−43

Table 4.8: DIC comparison between BIm and ZOIB models for different population sizes of BIm simulated dataset.

These results might seem obvious but they show that each model outperforms the other in the scenario that corresponds to it. Also, if the distribution of the data is unknown beforehand, the best fit can help us understand if the discrete probabilities are related or not to the conditional mean.

4.4. Predictive Power

For the predictive power of the models, 50 datasets consisting on 50 subjects each with 6 longitudinal observations are generated from the true models in order to make out-of-sample predictions and obtain the area under the operating characteristic (ROC) curve (AUC) of each model.

Since binary response variables are often modeled by a logistic regression (LR), it is also included for the predictive power comparison. As stated in Chapter 1, the logistic regression model assumes that the observations are independent, but the longitudinal information Y is not independent since it is related to the same subject.

Morrison (2010) shows that incorporating trends of longitudinal information or time-series as covariates helps improving a classification model. Thus, in order to not lose information about the longitudinal trajectory of the subjects in the application of the LR model, two extra covariates related to Y have been included in the $\boldsymbol{Z}$ matrix. This covariates focus on recovering the trend of the longitudinal fractional variable Y in just one number, a simpler version of what Morrison (2010) proposed but with the same objective. The first extra covariate $z_{i,3}$ just considers the first ($t = 1$) and the final ($t = 6$) value of Y and consists on assigning the value of 1 if the final value is larger than the first, if the final value is equal to the first 0 is assigned and if the final value is lower than the first -1 is assigned, as can be seen in (4.2).

$$z_{i,3} = \begin{cases} 1, & y_{i,6} > y_{i,1}. \\ 0, & y_{i,6} = y_{i,1}. \\ -1, & y_{i,6} < y_{i,1}. \end{cases} \tag{4.2}$$

The second extra covariate $z_{i,4}$ is the difference between the number of months Y went up and the number of months Y went down, this is, if a subject reduces its Y value each month on the 6 periods, $z_{i,4}$ will take -5 as the corresponding value, as can be seen in (4.3).

$$z_{i,4} = \sum_{j=1}^{5} I(y_{i,j+1} > y_{i,j}) - \sum_{j=1}^{5} I(y_{i,j+1} < y_{i,j}), \tag{4.3}$$

where

$$I(y_{i,j+1} > y_{i,j}) = \begin{cases} 1, & y_{i,j+1} > y_{i,j} \\ 0, & \text{otherwise} \end{cases}$$

and

$$I(y_{i,j+1} < y_{i,j}) = \begin{cases} 1, & y_{i,j+1} < y_{i,j} \\ 0, & \text{otherwise.} \end{cases}$$

Finally, three logistic regressions are included, the first one does not considers the longitudinal Y information and is labelled as LR (i), the second one considers the covariates explained in (4.2) and (4.3) and is labelled as LR (ii) and the third one considers each longitudinal observation as a covariate and is labelled as LR (iii).

The probability of default is calculated using (3.11) and the results of predictive power simulation study are shown in Tables 4.9 and 4.10, where each AUC value within sample and out of sample achieved by each model is presented.

Model	**Within Sample**	**Out of Sample**	
	AUC	**Mean (AUC)**	**SD (AUC)**
ZOIB	0.8302	0.8569	0.0609
BIm	0.8302	0.8569	0.0609
BTran	0.8166	0.8569	0.0608
LR (i)	0.7114	0.5780	0.0576
LR (ii)	0.8693	0.7658	0.0644
LR (iii)	0.9355	0.7955	0.0556

Table 4.9: AUC comparison between ZOIB, BIm, BTran and logistic regression models for the ZOIB simuated datasets.

Model	Within Sample	Out of Sample	
	AUC	**Mean (AUC)**	**SD (AUC)**
BIm	0.8268	0.8468	0.0594
ZOIB	0.8149	0.8466	0.0602
BTran	0.8217	0.8459	0.0595
LR (i)	0.6248	0.5630	0.0456
LR (ii)	0.8183	0.7325	0.0885
LR (iii)	0.9457	0.7990	0.0705

Table 4.10: AUC comparison between BIm, ZOIB, BTran and logistic regression models for the BIm simulated datasets.

We can see in Tables 4.9 and 4.10 that the LR (iii) model outperforms the other models in terms of within sample AUC, but the out-of-sample AUC clearly shows it is an over-fitted model. There is not a clear difference in out-of-sample predictive power between the ZOIB, BIm and BTran models, but they outperform the three logistics regressions, showing that the logistic regression model is not competitive enough because of how the information of the longitudinal variable Y is included and that LR (i), the one that does not includes longitudinal information, clearly performs much worse than the others.

4.5. Cluster recovery

It is also a matter of interest to know if these implemented models can recover the cluster where each subject truly belongs. Table 4.11 shows the cluster accuracy.

Model	n	Cluster accuracy (ZOIB dataset)	Cluster accuracy (BIm dataset)
ZOIB	50	1.000	0.980
	100	1.000	1.000
	200	1.000	1.000
	400	1.000	0.983
BIm	50	1.000	0.980
	100	1.000	1.000
	200	1.000	1.000
	400	1.000	0.995
BTran	50	0.960	0.980
	100	0.990	0.950
	200	0.965	0.975
	400	0.958	0.938
LR (i)	50	0.760	0.760
	100	0.700	0.700
	200	0.775	0.775
	400	0.698	0.698
LR (ii)	50	0.660	0.520
	100	0.560	0.620
	200	0.525	0.510
	400	0.570	0.663
LR (iii)	50	0.580	0.620
	100	0.560	0.630
	200	0.630	0.500
	400	0.650	0.650

Table 4.11: Cluster accuracy comparison between ZOIB, BIm, BTran and logistic regression models for the ZOIB and BIm generated datasets.

Table 4.11 clearly shows that ZOIB, BIm and BTran models can correctly assign the subject to the real cluster with high accuracy (near or equal to 1), contrary to what is shown by the logistic regressions.

The code of the models (ZOIB, BIm, BTran and LR) written in JAGS can be found in Appendix D.

Chapter 5

Application to real data

This chapter presents the results obtained from the application of the proposed models to real data. The classification beta inflated mixed regression models with cluster formation presented in Chapter 3 are compared between them on a credit card (CC) portfolio of a Peruvian bank in order to classify the bank's clients according to its credit card utilization ratio and its probability of default. A brief discussion of the results is at the end of the chapter.

5.1. Data

The application dataset consists on a random selection of 100 default clients and 100 non-default clients from a Peruvian bank with one credit card with no change in its credit line between January 2017 and December 2017 (12 logitudinal observations per client). Table 5.1 shows how this dataset looks like, where the two dependent variables are the credit card utilization ratio (CCUR), a continuous variable bounded to the interval [0, 1] which represents the percentage of the credit line used rounded off to two decimal places (denoted as Y in previous sections), and Default, a binary variable that represents if the client fails to pay its debt (1) or not (0) within the next 12 months of the observation period (denoted as D in previous sections). The independent variables are the time, represented by Month-Year, a binary variable that indicates if the client had a Cash Advance in the month (1) or not (0), the number of months since the origination of the credit card and the credit line of the credit card in Peruvian currency *nuevos soles* (S/).

This information was obtained the last day of each month, from January 2017 to December 2017. Then, these clients were observed until December 2018 to see if they failed to pay its debt (default) or not. Table 5.2 shows the distribution of zeros, ones and values in the open (0,1) interval of the 2400 real CCUR responses by month, where it can be seen that the proportion of clients each month in the extremes (0 and 1) is significant.

Client	Month-Year	Cash Advance	Months with CC	Credit line (S/)	CCUR	Default
1	Jan-17	0	15	1000.00	0.11	0
1	Feb-17	1	16	1000.00	0.62	0
⋮	⋮	⋮	⋮	⋮	⋮	⋮
100	Nov-17	0	24	5500.00	0.55	0
100	Dec-17	0	25	5500.00	0.00	0
101	Jan-17	1	1	2350.00	0.14	1
101	Feb-17	0	2	2350.00	0.51	1
⋮	⋮	⋮	⋮	⋮	⋮	⋮
200	Nov-17	0	19	800.00	0.33	1
200	Dec-17	1	20	800.00	1.00	1

Table 5.1: Real dataset structure.

Month	Frequency of values		
	0	1	(0,1)
Jan-17	33.5 %	9.5 %	57.0 %
Feb-17	31.0 %	9.5 %	59.5 %
Mar-17	34.0 %	6.5 %	59.5 %
Apr-17	35. %	8.5 %	56.5 %
May-17	25.5 %	7.5 %	67.5 %
Jun-17	25.0 %	9.5 %	65.5 %
Jul-17	26.5 %	11.0 %	62.5 %
Aug-17	31.0 %	9.5 %	59.5 %
Sep-17	29.0 %	6.0 %	65.0 %
Oct-17	27.5 %	8.5 %	64.0 %
Nov-17	28.0 %	11.5 %	60.5 %
Dec-17	29.5 %	8.5 %	62.0 %

Table 5.2: Frequency of zeros, ones and values in the open (0,1) interval of the real $CCUR$ responses by month.

In order to explore the variables in the dataset, a summary of the mean and median value of each variable grouped by the default status of the clients is shown in Table 5.3. An early interpretation of the variables can be made with the information provided by Table 5.3, for example, a client that defaults has in average 2.5 times more months with cash advance than a client that does not default, this is an expected behavior in banking because cash advance is an expensive transaction and it is usually made by clients with need of liquidity; as for

the number of months with CC, the clients that do not default have an 'older' credit card because they must have a longer credit history being a good payer; as for the credit line, the clients that do not default have a larger credit line because it is expected that a less risky client can handle a higher debt, therefore the bank offers them a larger line and finally; for the credit card utilization ratio, the clients that default have a greather CCUR because it is expected for this clients to expend more of their credit lines as they start having financial problems.

Variable		**Default**	
		0	1
Number of months with cash advance	Mean	0.06	0.16
	Median	0.00	0.00
Number of months with CC	Mean	55	42
	Median	40	35
Credit line (S/)	Mean	10,373	7,345
	Median	5,850	3,300
CCUR	Mean	0.30	0.35
	Median	0.11	0.18

Table 5.3: Mean and median of each independent variable grouped by default status.

It is expected that this relationship between dependent and independent variables shown in Table 5.3 will be manifested in the regression fixed effects associated with each covariate presented in the following sections.

5.2. Model Structure

For the fixed effects associated with the mixed fractional response *CCUR*, the credit card utilization ratio, we consider the covariates time (j), $j = 1, ..., 12$, a binary variable that indicates if the client had a cash advance in the month (1) or not (0) (*cash_adv_month*), the mean number of years the client had its credit card (*mean_age_cc*) and the credit line of the credit card (*credit_line*) in thousands. The covariates (*cash_adv_month*) and (*credit_line*) can be obtained directly from Table 5.1 while the covariate (*mean_age_cc*) is computed by client as the mean of the column *Months with CC* from Table 5.1 divided by 12 (to take it to the dimension of years).

For the fixed effects associated with the default response D we consider the covariate number of months with cash advances made during the twelve months of observation (*cash_adv_total*), the mean number of years the client had its credit card (*mean_age_cc*) and the credit line of the credit card (*credit_line*) in thousands. The covariate (*credit_line*) can be obtained directly from Table 5.1, the covariate (*mean_age_cc*) is computed by client as the mean of the column *Months with CC* from Table 5.1 divided by 12 (to take it to the dimension of years) and the covariate (*cash_adv_total*) is calculated as the sum by client of

the column *Cash Advance* from Table 5.1.

The regression structure presented on (5.1) allows us to model the probability of default for each client (π) and the parameters associated with the credit card utilization ratio ($\alpha_0, \alpha_1, \gamma$) for each client at each time.

$$\begin{aligned}
\text{logit}(\pi_i^{w_i}) =& \beta_{\pi 0}^{w_i} + \beta_{\pi 1}^{w_i} \cdot cash_adv_total \\
&+ \beta_{\pi 2}^{w_i} \cdot mean_age_cc + \beta_{\pi 3}^{w_i} \cdot credit_line \\
\text{logit}(\alpha_{0ij}^{w_i}) =& \beta_{\alpha_0 0}^{w_i} + \beta_{\alpha_0 1}^{w_i} \cdot j + \beta_{\alpha_0 2}^{w_i} \cdot cash_adv_month \\
&+ \beta_{\alpha_0 3}^{w_i} \cdot mean_age_cc + \beta_{\alpha_0 4}^{w_i} \cdot credit_line \\
\text{logit}(\alpha_{1ij}^{w_i}) =& \beta_{\alpha_1 0}^{w_i} + \beta_{\alpha_1 1}^{w_i} \cdot j + \beta_{\alpha_1 2}^{w_i} \cdot cash_adv_month \\
&+ \beta_{\alpha_1 3}^{w_i} \cdot mean_age_cc + \beta_{\alpha_1 4}^{w_i} \cdot credit_line \\
\text{logit}(\gamma_{ij}^{w_i}) =& \beta_{\gamma 0}^{w_i} + \beta_{\gamma 1}^{w_i} \cdot j + \beta_{\gamma 2}^{w_i} \cdot cash_adv_month \\
&+ \beta_{\gamma 3}^{w_i} \cdot mean_age_cc + \beta_{\gamma 4}^{w_i} \cdot credit_line + b_i.
\end{aligned} \tag{5.1}$$

Considering just one cluster, without loss of generality, the regression structure (5.1) assumes that CCUR mean (γ) can only rise ($\beta_{\gamma 1} > 0$), fall ($\beta_{\gamma 1} < 0$) or stay constant ($\beta_{\gamma 1} = 0$) in time, with an extra fluctuation given by $\beta_{\gamma 2}$ (expected to be positive). However, this is not necessary the behavior of the clients, in Perú its common to have extra expenses on February and March due to the beginning of the academic year and on December due to Christmas celebration. This particular effects cannot be taken in consideration by (5.1), thus two additional regression structures are considered: dummy variables for each time period and polynomial spline basis covariates (*spline_basis*) with 5 degrees of freedom, shown in equations (5.2) and (5.3) respectively.

The incorporation of dummy variables act as a modifier of the intercept $\beta_{\gamma 0}$ at each time j, allowing us to recover any situational effect to the mean γ in a specific month.

$$
\begin{aligned}
\text{logit}(\pi_i^{w_i}) =& \beta_{\pi 0}^{w_i} + \beta_{\pi 1}^{w_i} \cdot cash\_adv\_total \\
&+ \beta_{\pi 2}^{w_i} \cdot mean\_age\_cc + \beta_{\pi 3}^{w_i} \cdot credit\_line \\
\text{logit}(\alpha_{0ij}^{w_i}) =& \beta_{\alpha 0 0}^{w_i} + \beta_{\alpha 0 1}^{w_i} \cdot I(j=1) + \beta_{\alpha 0 2}^{w_i} \cdot I(j=2) \\
&+ \beta_{\alpha 0 3}^{w_i} \cdot I(j=3) + \beta_{\alpha 0 4}^{w_i} \cdot I(j=4) + \beta_{\alpha 0 5}^{w_i} \cdot I(j=5) \\
&+ \beta_{\alpha 0 6}^{w_i} \cdot I(j=6) + \beta_{\alpha 0 7}^{w_i} \cdot I(j=7) + \beta_{\alpha 0 8}^{w_i} \cdot I(j=8) \\
&+ \beta_{\alpha 0 9}^{w_i} \cdot I(j=9) + \beta_{\alpha 0 10}^{w_i} \cdot I(j=10) + \beta_{\alpha 0 11}^{w_i} \cdot I(j=11) \\
&+ \beta_{\alpha 0 12}^{w_i} \cdot cash\_adv\_month + \beta_{\alpha 0 13}^{w_i} \cdot mean\_age\_cc \\
&+ \beta_{\alpha 0 14}^{w_i} \cdot credit\_line \\
\text{logit}(\alpha_{1ij}^{w_i}) =& \beta_{\alpha 1 0}^{w_i} + \beta_{\alpha 1 1}^{w_i} \cdot I(j=1) + \beta_{\alpha 1 2}^{w_i} \cdot I(j=2) \\
&+ \beta_{\alpha 1 3}^{w_i} \cdot I(j=3) + \beta_{\alpha 1 4}^{w_i} \cdot I(j=4) + \beta_{\alpha 1 5}^{w_i} \cdot I(j=5) \\
&+ \beta_{\alpha 1 6}^{w_i} \cdot I(j=6) + \beta_{\alpha 1 7}^{w_i} \cdot I(j=7) + \beta_{\alpha 1 8}^{w_i} \cdot I(j=8) \\
&+ \beta_{\alpha 1 9}^{w_i} \cdot I(j=9) + \beta_{\alpha 1 10}^{w_i} \cdot I(j=10) + \beta_{\alpha 1 11}^{w_i} \cdot I(j=11) \\
&+ \beta_{\alpha 1 12}^{w_i} \cdot cash\_adv\_month + \beta_{\alpha 1 13}^{w_i} \cdot mean\_age\_cc \\
&+ \beta_{\alpha 1 14}^{w_i} \cdot credit\_line \\
\text{logit}(\gamma_{ij}^{w_i}) =& \beta_{\gamma 0}^{w_i} + \beta_{\gamma 1}^{w_i} \cdot I(j=1) + \beta_{\gamma 2}^{w_i} \cdot I(j=2) \\
&+ \beta_{\gamma 3}^{w_i} \cdot I(j=3) + \beta_{\gamma 4}^{w_i} \cdot I(j=4) + \beta_{\gamma 5}^{w_i} \cdot I(j=5) \\
&+ \beta_{\gamma 6}^{w_i} \cdot I(j=6) + \beta_{\gamma 7}^{w_i} \cdot I(j=7) + \beta_{\gamma 8}^{w_i} \cdot I(j=8) \\
&+ \beta_{\gamma 9}^{w_i} \cdot I(j=9) + \beta_{\gamma 10}^{w_i} \cdot I(j=10) + \beta_{\gamma 11}^{w_i} \cdot I(j=11) \\
&+ \beta_{\gamma 12}^{w_i} \cdot cash\_adv\_month + \beta_{\gamma 13}^{w_i} \cdot mean\_age\_cc \\
&+ \beta_{\gamma 14}^{w_i} \cdot credit\_line + b_i,
\end{aligned}
\tag{5.2}
$$

where

$$
I(j=k) = \begin{cases} 1, & j = k \\ 0, & \text{otherwise,} \end{cases}
$$

and j can take the values $j = 1, ..., 12$.

On the other hand, the incorporation of spline basis covariates replaces the time j variable but adds a 5 degree polynomial flexibility to it, allowing the CCUR mean γ to have curvature with up to 4 turning points.

$$
\begin{aligned}
\text{logit}(\pi_i^{w_i}) =& \beta_{\pi 0}^{w_i} + \beta_{\pi 1}^{w_i} \cdot cash\_adv\_total \\
&+ \beta_{\pi 2}^{w_i} \cdot mean\_age\_cc + \beta_{\pi 3}^{w_i} \cdot credit\_line \\
\text{logit}(\alpha_{0ij}^{w_i}) =& \beta_{\alpha_0 0}^{w_i} + \beta_{\alpha_0 1}^{w_i} \cdot spline\_basis_{1j} + \beta_{\alpha_0 2}^{w_i} \cdot spline\_basis_{2j} \\
&+ \beta_{\alpha_0 3}^{w_i} \cdot spline\_basis_{3j} + \beta_{\alpha_0 4}^{w_i} \cdot spline\_basis_{4j} \\
&+ \beta_{\alpha_0 5}^{w_i} \cdot spline\_basis_{5j} + \beta_{\alpha_0 6}^{w_i} \cdot cash\_adv\_month \\
&+ \beta_{\alpha_0 7}^{w_i} \cdot mean\_age\_cc + \beta_{\alpha_0 8}^{w_i} \cdot credit\_line \\
\text{logit}(\alpha_{1ij}^{w_i}) =& \beta_{\alpha_1 0}^{w_i} + \beta_{\alpha_1 1}^{w_i} \cdot spline\_basis_{1j} + \beta_{\alpha_1 2}^{w_i} \cdot spline\_basis_{2j} \\
&+ \beta_{\alpha_1 3}^{w_i} \cdot spline\_basis_{3j} + \beta_{\alpha_1 4}^{w_i} \cdot spline\_basis_{4j} \\
&+ \beta_{\alpha_1 5}^{w_i} \cdot spline\_basis_{5j} + \beta_{\alpha_1 6}^{w_i} \cdot cash\_adv\_month \\
&+ \beta_{\alpha_1 7}^{w_i} \cdot mean\_age\_cc + \beta_{\alpha_1 8}^{w_i} \cdot credit\_line \\
\text{logit}(\gamma_{ij}^{w_i}) =& \beta_{\gamma 0}^{w_i} + \beta_{\gamma 1}^{w_i} \cdot spline\_basis_{1j} + \beta_{\gamma 2}^{w_i} \cdot spline\_basis_{2j} \\
&+ \beta_{\gamma 3}^{w_i} \cdot spline\_basis_{3j} + \beta_{\gamma 4}^{w_i} \cdot spline\_basis_{4j} \\
&+ \beta_{\gamma 5}^{w_i} \cdot spline\_basis_{5j} + \beta_{\gamma 6}^{w_i} \cdot cash\_adv\_month \\
&+ \beta_{\gamma 7}^{w_i} \cdot mean\_age\_cc + \beta_{\gamma 8}^{w_i} \cdot credit\_line + b_i.
\end{aligned}
\tag{5.3}
$$

where j can take the values $j = 1, ..., 12$.

From now on we will refer to (5.1) as linear regression, (5.2) as dummy regression and (5.3) as spline regression. These 3 regression structures are also replicated on the ZOIB model for the parameters π, λ_0, λ_1 and μ. The complete regression structures for the ZOIB model can be seen in Appendix B.

5.3. Results

We partition the real data into two 50/50 sets using stratified sampling based on the default status to have a training and a test dataset with equal number of default subjects. The BIm and ZOIB models presented in Chapter 3 are applied to the training dataset in order to obtain the best fit by comparing the DIC of each model. Then, the trained models are applied to the test dataset for making out-of-sample predictions in order to obtain the best predictive power according to the AUC achieved by each model.

5.3.1. Training set

The BIm and ZOIB models are applied to the training dataset using JAGS software in R through the package RJAGS, discarting the first 1000 iterations and sampling the next 2000 iterations considering a thinning interval equal to 5. Multiple values of cluster number (K) and different regression structures (linear (5.1), dummy (5.2) and spline (5.3)) were used. The number of clusters (K) was iterated up to 10 but just 5 or 6 clusters had a subject in it, thus we set the maximum of clusters for this analysis to 6.

Tables 5.4 and 5.5 show the DIC obtained for each K and each regression structure of the BIm and ZOIB models respectively. These results have been plotted in Figure 5.1 to facilitate their interpretation. Figures 5.2, 5.3 and 5.4 show the estimated trajectories by cluster for the best fit for every regression structure of the BIm model and Figures 5.5, 5.6 and 5.7

for the ZOIB model. Tables 5.6 and 5.7 show the parameter estimation for the best model and Figure 5.8 show the cluster belonging probability for each subject, assigned by the best model.

K	DIC (BIm linear)	DIC (BIm dummy)	DIC (BIm spline)
1	1957	1974	1956
2	**1621**	**1652**	**1603**
3	1914	2295	1993
4	1624	3167	2063
5	1900	3037	2317
6	1920	3262	2123

Table 5.4: DIC comparison between different number of clusters (K) and regression structures for the BIm model. Bold numbers indicate the best fit for each type of regression.

K	DIC (ZOIB linear)	DIC (ZOIB dummy)	DIC (ZOIB spline)
1	2026	2042	2022
2	**1714**	**1796**	**1687**
3	1995	2432	2533
4	1736	3372	2598
5	1820	2802	2180
6	2494	2428	2489

Table 5.5: DIC comparison between different number of clusters (K) and regression structures for the ZOIB model. Bold numbers indicate the best fit for each type of regression.

As can be seen in Tables 5.4 and 5.5 the lowest DIC is obtained by the BIm model using the spline regression structure (5.3) with 2 clusters, thus this gives the best fit for the real dataset. When looking at Tables 5.4 and 5.5 and Figure 5.1 some inferences arise directly, for example, considering $K = 1$ (no clusters) is not the best option in any model or regression structure, indicating that considering a finite mixture was a correct choice; in all regression structures considering $K \leq 4$, the BIm model outperforms the ZOIB model; for both models in every regression structure, considering $K = 2$ gives the best fit according to DIC, indicating that the real dataset presents two groups quite different from each other.

These results does not mean that considering other values of $K = 3, 4, 5, 6$ are a wrong choice, actually the result could have been for example 3 clusters, representing 3 different trajectories, one with increasing CCUR, another with decreasing CCUR and the last one with constant CCUR over time. However, according to DIC, this fit is not as good as considering $K = 2$.

We can see in Figure 5.2 the trajectories obtained by the application of BIm linear regression with 2 clusters. It is clearly seen that the signs of β_γ are opposite between the clusters, where β_γ is negative in cluster 1 and positive in cluster 2. Both clusters present a PD close to 0.50, therefore they can be labelled as medium risk trajectories, with cluster 1 representing slightly less risk than cluster 2 because clients that reduce their debt over time

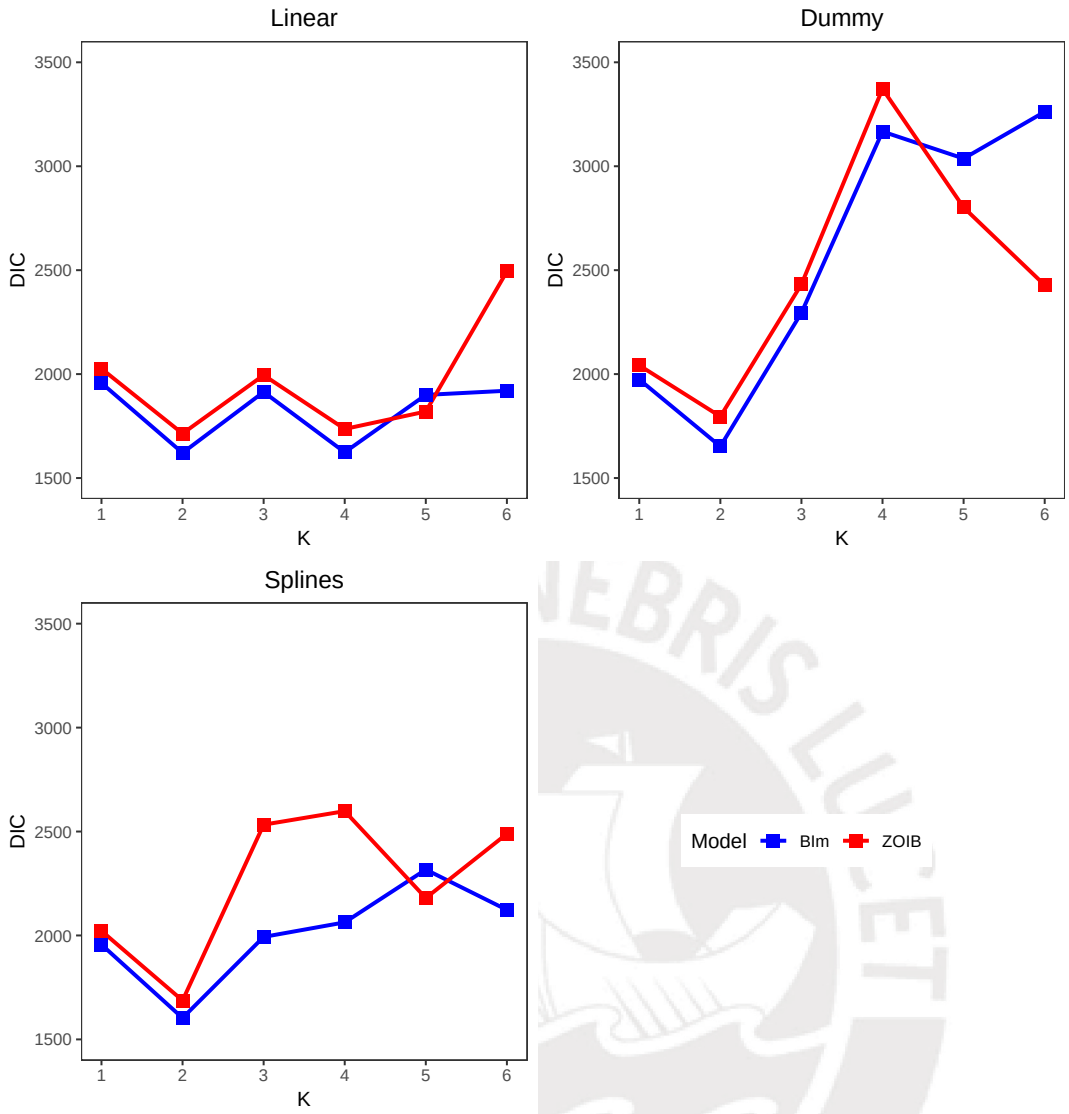


Figure 5.1: DIC comparison between different number of clusters (K) and regressions for the BIm and ZOIB model.

are less likely to default.

In Figure 5.3 we can see the trajectories obtained by the application of BIm dummy regression with 2 clusters. Due to the nature of this regression structure, we do not necessarily see a monotonous increase or decrease of the CCUR mean γ. This happens to cluster 1, where γ decreases every month from January to November but increases in December, inferring that it is due to additional Christmas expenses. Cluster 1 presents a decreasing trend of γ and a PD of 0.44, thus this cluster can be labelled as low risk trajectory. On the other hand, cluster 2 presents a stable trajectory the first 4 months but then the rise in CCUR begins and has a PD of 0.56, thus this cluster can be labelled as high risk trajectory.

We can see in Figure 5.4 the trajectories obtained by the application of BIm spline regression with 2 clusters. Cluster 1 presents a decreasing trend of γ and a PD of 0.42, thus this cluster can be labelled as low risk trajectory. Cluster 2 presents an increasing trend of γ and a PD of 0.58, thus this cluster can be labelled as high risk trajectory. This is the

regression with the largest difference of PD between their clusters.

The interpretation of the trajectories generated by the ZOIB model, shown in Figures 5.5, 5.6 and 5.7, are very similar, where the Cluster 1 of each regression structure presents a decreasing trend of μ and the Cluster 2 of each regression structure presents an increasing trend of μ, leading to have low and high risk trajectories, respectively.

A fact is that all clusters presented in Figures 5.2, 5.3, 5.4, 5.5, 5.6 and 5.7 have similar grouping of clients regardless of the regression structure or the model. Therefore, we can infer that the nature of the data has a structure of two clusters. This cluster formation favors the creation of groups of clients with different patterns in their CCUR and facilitates the inference of possible default cases in spite of the regression structure or the model.

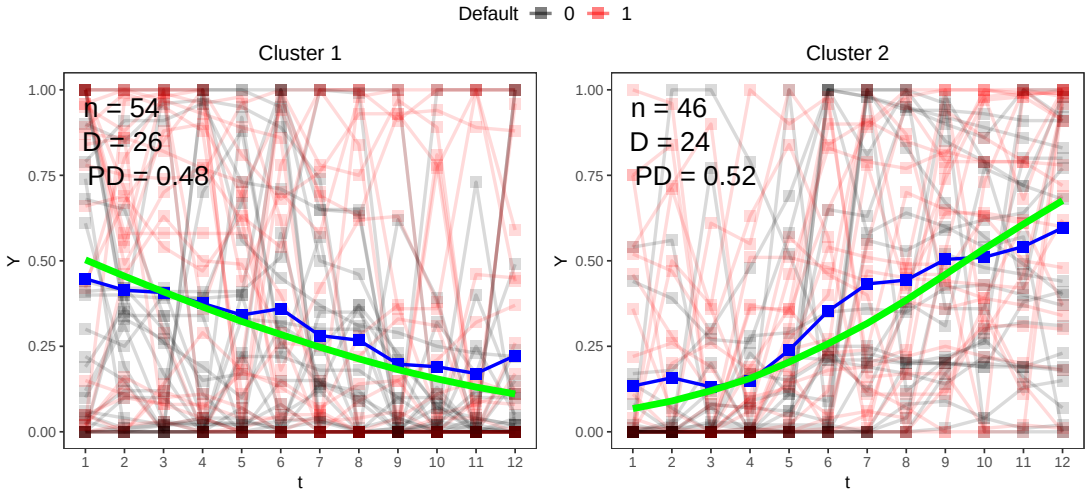


Figure 5.2: Longitudinal trajectory by cluster using BIm linear regression on real training dataset. The blue line is the real mean γ and the green line is its estimation. n is the number of clients in the cluster, D is the number of default clients and PD is the ratio between default and total clients.

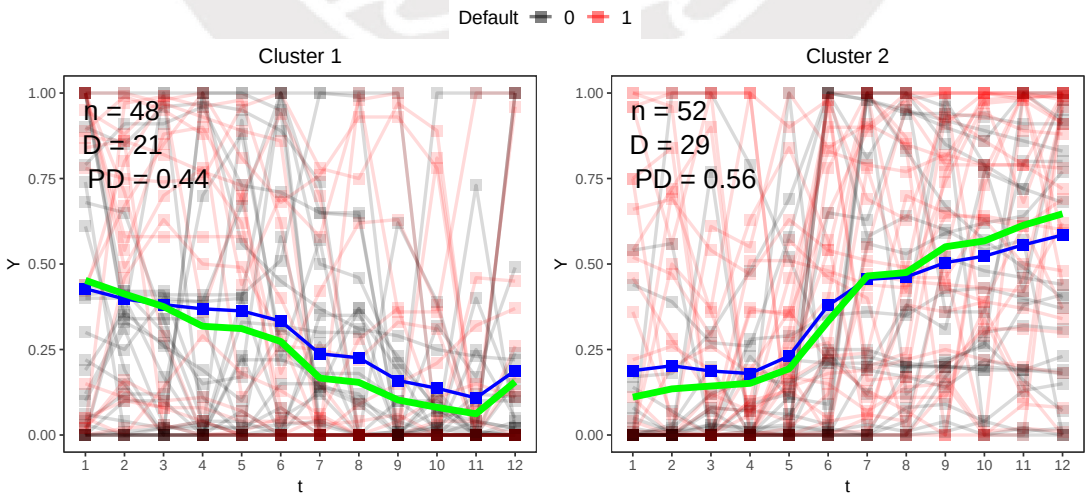


Figure 5.3: Longitudinal trajectory by cluster using BIm dummy regression on real training dataset. The blue line is the real mean γ and the green line is its estimation. n is the number of clients in the cluster, D is the number of default clients and PD is the ratio between default and total clients.

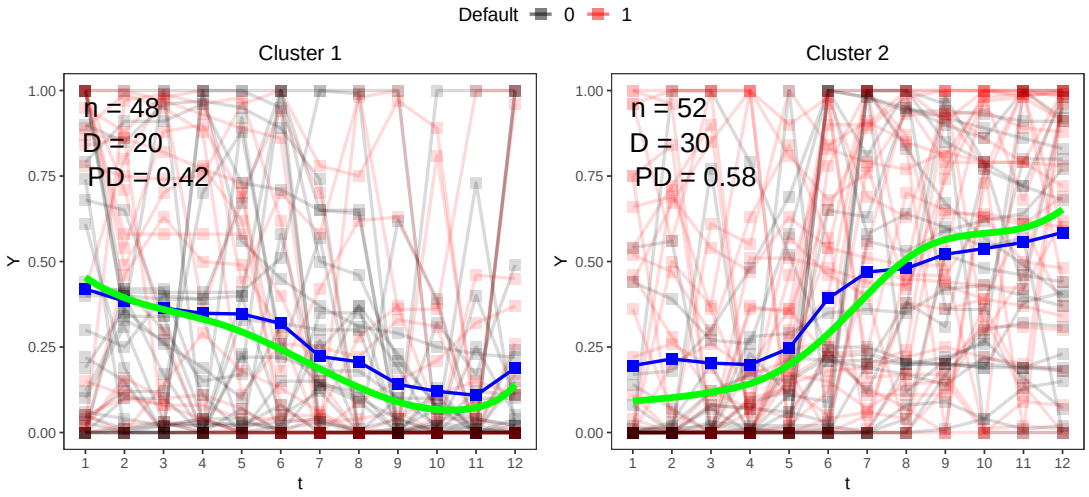


Figure 5.4: Longitudinal trajectory by cluster using BIm spline regression on real training dataset. The blue line is the real mean γ and the green line is its estimation. n is the number of clients in the cluster, D is the number of default clients and PD is the ratio between default and total clients.

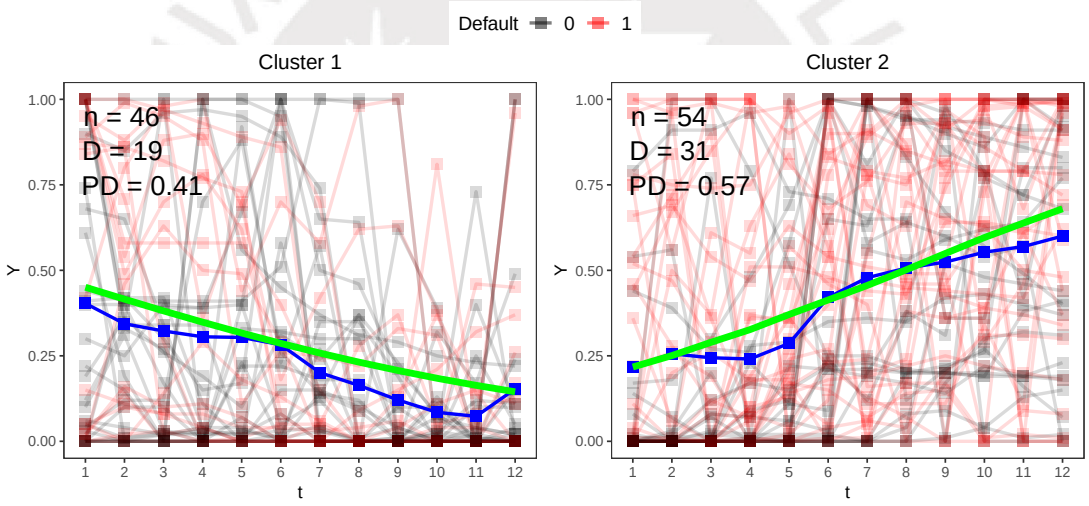


Figure 5.5: Longitudinal trajectory by cluster using ZOIB linear regression on real training dataset. The blue line is the real mean γ and the green line is the conditional mean μ estimation. n is the number of clients in the cluster, D is the number of default clients and PD is the ratio between default and total clients.

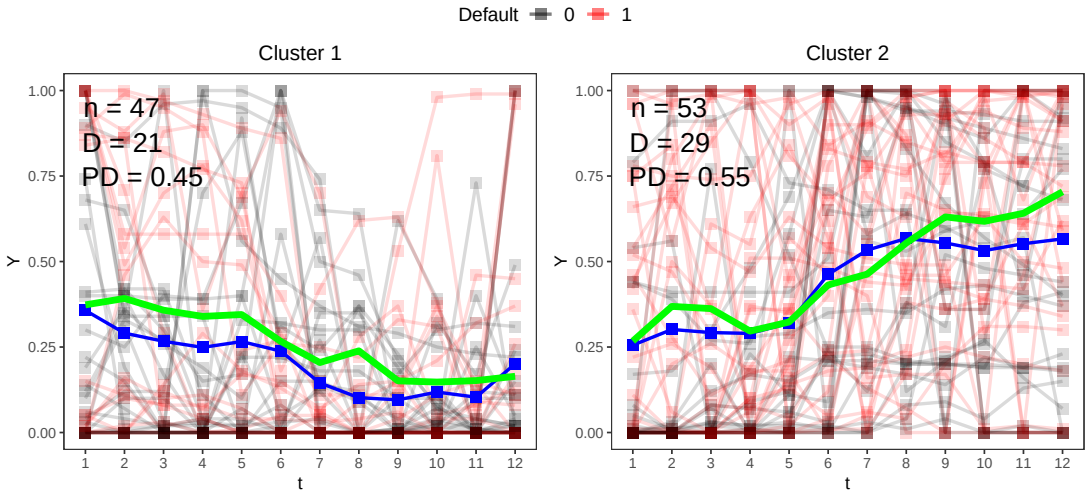


Figure 5.6: Longitudinal trajectory by cluster using ZOIB dummy regression on real training dataset. The blue line is the real mean γ and the green line is the conditional mean μ estimation. n is the number of clients in the cluster, D is the number of default clients and PD is the ratio between default and total clients.

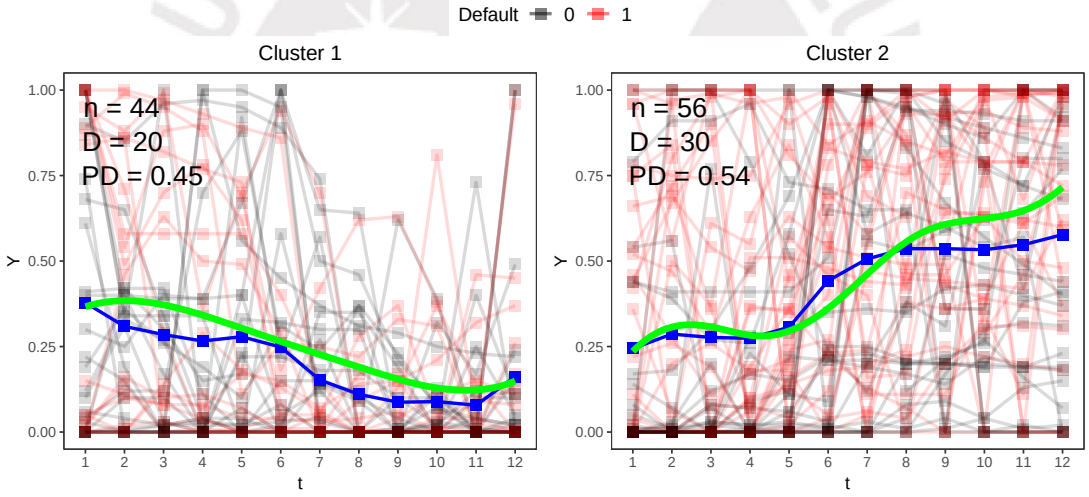


Figure 5.7: Longitudinal trajectory by cluster using ZOIB spline regression on real training dataset. The blue line is the real mean γ and the green line is the conditional mean μ estimation. n is the number of clients in the cluster, D is the number of default clients and PD is the ratio between default and total clients.

Regarding the interpretation of fixed effects for the BIm spline regression with 2 clusters shown in Tables 5.6 and 5.7, results help us have a better understanding of the clients' behavior from a bank perspective. In cluster 1, the low risk trajectory, fixed effect $\beta_{\gamma 8}$ = -0.05136 related to the covariate *credit_line* indicates that for each extra S/ 1,000 of credit line, the odds of γ would decrease 5.01 % and; the fixed effect $\beta_{\pi 1} = 0.31941$ related to the covariate *cash_adv_total* indicates that the odds of π, the default probability of a client, is 1.37632 times greater for every month that the client had a cash advance. In cluster 2, the high risk trajectory, fixed effect $\beta_{\gamma 6} = 0.38196$ related to covariate *cash_adv_month* indicates

that the odds of the credit card utilization ratio would be 1.46515 times greater the months the client had a cash advance; fixed effect $\beta_{\gamma 7}$ = -0.12429 related to covariate *mean_age_cc* indicates that the odds of the mean CCUR would be 11.69 % less for every extra year the credit card has; fixed effect $\beta_{\gamma 8}$ = -0.04865 related to the covariate *credit_line* indicates that for each extra S/ 1,000 of credit line the odds of γ would decrease 4.75 % and; fixed effect $\beta_{\pi 2}$ = -0.33823 related to covariate *mean_age_cc* indicates that the odds of π, the default probability of a client, would be 28.70 % less for every extra year the credit card has.

The banking interpretation of these effects is somewhat expected, for $\beta_{\gamma 8}$ from both clusters, the CCUR would decrease with more credit line because increasing the credit line of a client does not means he will expend more, so the numerator would stay the same and the denominator would increase, therefore γ will decrease, especially in lower risk trajectories; for the fixed effect $\beta_{\gamma 7}$ related to *mean_age_cc* in cluster 2 (high risk trajectory), it indicates that clients that 'survived' another year tend to reduce their expenses in order not to default; for the fixed effects related to cash advance, $\beta_{\pi 1}$ from cluster 1 and $\beta_{\gamma 6}$ from cluster 2, as it was expected, having a cash advance will increase γ and since it is an expensive product and it is taken by the clients in need of liquidity, it increases their PD and; for $\beta_{\pi 2}$ from cluster 2, clients that have 'survived' longer with a credit card are expected to have a better credit history, therefore less probability of default.

Fixed effects that contain the value of 0 in their credibility interval are considered not significant.

Every MCMC considered for this analysis converged. The plots of the sampled chains for the parameters interpreted before are shown in Appendix C.

Cluster	Parameter	Mean	SD	P 2.5 %	P 97.5 %
1	$\beta_{\alpha_0 0}$	-0.32002	0.40713	-1.11080	0.46618
	$\beta_{\alpha_0 1}$	-0.14484	0.81040	-1.74932	1.45029
	$\beta_{\alpha_0 2}$	-0.23972	0.76397	-1.85995	1.25905
	$\beta_{\alpha_0 3}$	1.57724	0.78398	0.00780	3.09146
	$\beta_{\alpha_0 4}$	1.20899	0.65005	-0.09735	2.48987
	$\beta_{\alpha_0 5}$	1.30029	0.54273	0.25273	2.42177
	$\beta_{\alpha_0 6}$	-0.39945	0.40251	-1.17497	0.36580
	$\beta_{\alpha_0 7}$	-0.12390	0.04236	-0.20673	-0.04235
	$\beta_{\alpha_0 8}$	0.01571	0.01244	-0.01002	0.04011
	$\beta_{\alpha_1 0}$	0.83665	0.52749	-0.19611	1.81105
	$\beta_{\alpha_1 1}$	-3.70876	1.20795	-6.19746	-1.53007
	$\beta_{\alpha_1 2}$	-1.25200	1.12798	-3.38714	0.92021
	$\beta_{\alpha_1 3}$	-1.79139	1.23783	-4.19404	0.59789
	$\beta_{\alpha_1 4}$	-2.94768	1.58037	-6.41901	0.06058
	$\beta_{\alpha_1 5}$	-0.20309	0.75527	-1.60576	1.31421
	$\beta_{\alpha_1 6}$	-3.54251	1.38064	-6.82240	-1.32560
	$\beta_{\alpha_1 7}$	-0.04886	0.07908	-0.20565	0.09879
	$\beta_{\alpha_1 8}$	-0.01167	0.02646	-0.07049	0.03294
	$\beta_{\gamma 0}$	-0.31355	0.39312	-1.05879	0.46270
	$\beta_{\gamma 1}$	-0.31072	0.38815	-1.10491	0.42903
	$\beta_{\gamma 2}$	-0.39955	0.38796	-1.15465	0.37164
	$\beta_{\gamma 3}$	-1.61252	0.49156	-2.54813	-0.59761
	$\beta_{\gamma 4}$	-3.02455	0.50462	-3.98798	-2.05014
	$\beta_{\gamma 5}$	-1.43247	0.45450	-2.20696	-0.45790
	$\beta_{\gamma 6}$	-0.04010	0.29566	-0.59399	0.56029
	$\beta_{\gamma 7}$	0.10956	0.08233	-0.04762	0.27634
	$\beta_{\gamma 8}$	-0.05136	0.02212	-0.09265	-0.00704
	$\beta_{\pi 0}$	-0.89510	0.59863	-2.10945	0.23702
	$\beta_{\pi 1}$	0.31941	0.16577	0.04006	0.67030
	$\beta_{\pi 2}$	0.10806	0.13945	-0.15523	0.40578
	$\beta_{\pi 3}$	-0.02350	0.04436	-0.11843	0.05496
	σ^2_γ	1.49301	0.24604	1.09189	2.03902
	ϕ	8.79251	0.25989	8.30086	9.28398
	p	0.48228	0.05227	0.38100	0.58500

Table 5.6: Cluster 1 estimated posterior distribution of parameters from BIm spline regression with 2 clusters applied to training real dataset.

Cluster	Parameter	Mean	SD	P 2.5 %	P 97.5 %
2	$\beta_{\alpha_0 0}$	0.90668	0.33978	0.23703	1.56249
	$\beta_{\alpha_0 1}$	0.67013	0.65031	-0.65322	1.90928
	$\beta_{\alpha_0 2}$	-1.48781	0.65369	-2.79686	-0.22800
	$\beta_{\alpha_0 3}$	-1.52614	0.89328	-3.28116	0.29069
	$\beta_{\alpha_0 4}$	-3.59298	1.17398	-6.10044	-1.52632
	$\beta_{\alpha_0 5}$	-1.97249	0.69692	-3.53480	-0.73355
	$\beta_{\alpha_0 6}$	-0.42760	0.29033	-1.00511	0.09860
	$\beta_{\alpha_0 7}$	-0.02214	0.03075	-0.08292	0.03627
	$\beta_{\alpha_0 8}$	-0.01202	0.01221	-0.03777	0.01184
	$\beta_{\alpha_1 0}$	-1.49633	0.99093	-3.79334	0.14332
	$\beta_{\alpha_1 1}$	1.40082	1.65537	-1.55733	5.01656
	$\beta_{\alpha_1 2}$	1.10775	1.19707	-1.02801	3.67774
	$\beta_{\alpha_1 3}$	1.38801	1.55572	-1.36584	4.60945
	$\beta_{\alpha_1 4}$	0.77707	1.16891	-1.36610	3.22134
	$\beta_{\alpha_1 5}$	1.20902	1.08745	-0.67417	3.63494
	$\beta_{\alpha_1 6}$	-0.52323	0.45723	-1.36978	0.42053
	$\beta_{\alpha_1 7}$	-0.18428	0.08873	-0.35041	0.00168
	$\beta_{\alpha_1 8}$	-0.02912	0.02645	-0.08499	0.01860
	$\beta_{\gamma 0}$	-1.34257	0.41426	-2.21761	-0.49850
	$\beta_{\gamma 1}$	0.07096	0.49007	-0.90617	1.00192
	$\beta_{\gamma 2}$	0.29416	0.40348	-0.49455	1.08543
	$\beta_{\gamma 3}$	2.99830	0.44714	2.12514	3.90884
	$\beta_{\gamma 4}$	2.55073	0.34821	1.88222	3.20699
	$\beta_{\gamma 5}$	2.92105	0.30829	2.33351	3.50310
	$\beta_{\gamma 6}$	0.38196	0.17012	0.04857	0.71490
	$\beta_{\gamma 7}$	-0.12429	0.05490	-0.23033	-0.00985
	$\beta_{\gamma 8}$	-0.04865	0.01977	-0.08394	-0.00743
	$\beta_{\pi 0}$	1.93620	0.68767	0.70420	3.29830
	$\beta_{\pi 1}$	0.08794	0.12234	-0.11911	0.36781
	$\beta_{\pi 2}$	-0.33823	0.14898	-0.66040	-0.08390
	$\beta_{\pi 3}$	-0.06208	0.04039	-0.14796	0.00871
	σ_γ^2	1.49301	0.24604	1.09189	2.03902
	ϕ	8.79251	0.25989	8.30086	9.28398
	p	0.51772	0.05227	0.41500	0.61900

Table 5.7: Cluster 2 estimated posterior distribution of parameters from BIm spline regression with 2 clusters applied to training real dataset.

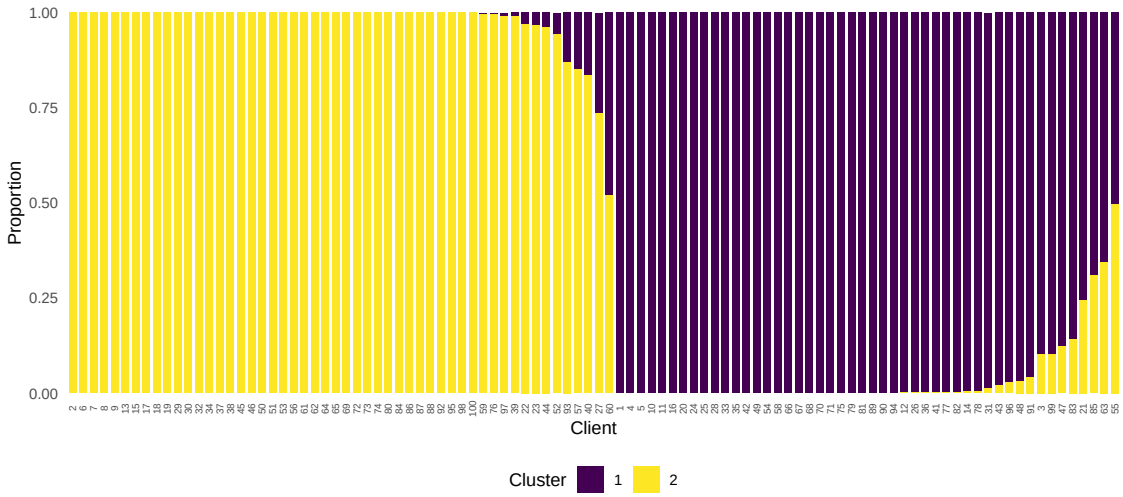


Figure 5.8: Probability of belonging to the different clusters for each client assigned by the BIm spline regression with 2 clusters.

In Figure 5.8 we can see the proportion of times a subject was assigned to cluster 1 or cluster 2 at each MCMC iteration. Clients 55 and 60 clearly stand out for their proximity to the probability of 0.50, this means the model assigned these clients half of the time to cluster 1 and the other half to cluster 2. To verify if the model is properly assigning the cluster respective cluster, we plotted the longitudinal trajectories of these two clients in Figure 5.9, where we can see that there is not a clear cluster belonging for those clients and that it is expected for the model to assign them almost randomly to clusters 1 or 2.

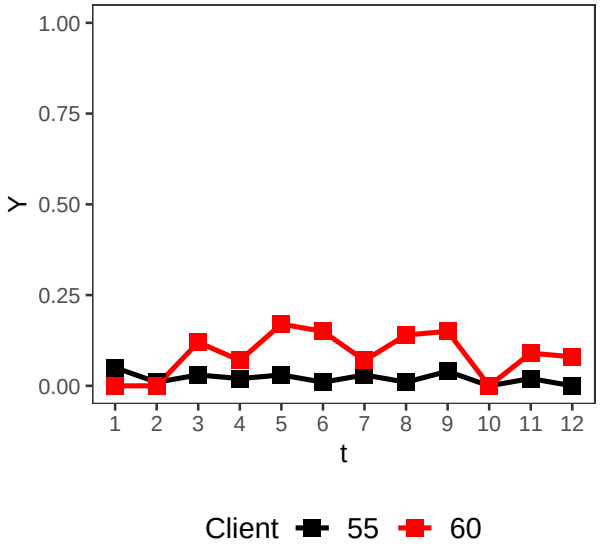


Figure 5.9: Longitudinal trajectory of the clients with less precision in their cluster belonging probability.

5.3.2. Test set

Since binary response variables are often modeled by a logistic regression (LR), it is also included for the predictive power comparison. As stated in Chapter 1, the logistic regression model assumes that the observations are independent, but the longitudinal information CCUR is not independent since it is related to the same subject.

Morrison (2010) shows that incorporating trends of longitudinal information or time-series as covariates helps improving a classification model. Thus, in order to not lose information about the longitudinal trajectory of the subjects in the application of the LR model, two extra covariates related to CCUR have been included for modelling π. This covariates focus on recovering the trend of the longitudinal fractional variable CCUR in just one number, a simpler version of what Morrison (2010) proposed but with the same objective. The first extra covariate *ccur_trend* just considers the first ($t = 1$) and the final ($t = 12$) value of CCUR and consists on assigning the value of 1 if the final value is larger than the first, if the final value is equal to the first 0 is assigned and if the final value is lower than the first -1 is assigned, as can be seen in (5.4).

$$ccur\_trend_i = \begin{cases} 1, & \text{CCUR}_{i,12} > \text{CCUR}_{i,1}. \\ 0, & \text{CCUR}_{i,12} = \text{CCUR}_{i,1}. \\ -1, & \text{CCUR}_{i,12} < \text{CCUR}_{i,1}, \end{cases} \tag{5.4}$$

The second extra covariate *ccur_up_down* is the difference between the number of months CCUR went up and the number months CCUR went down, this is, if a subject reduces its CCUR value each month on the 12 periods, *ccur_up_down* will take -11 as the corresponding value, as can be seen in (5.5).

$$ccur\_up\_down_i = \sum_{j=1}^{11} I(\text{CCUR}_{i,j+1} > \text{CCUR}_{i,j}) - \sum_{j=1}^{11} I(\text{CCUR}_{i,j+1} < \text{CCUR}_{i,j}), \tag{5.5}$$

where

$$I(\text{CCUR}_{i,j+1} > \text{CCUR}_{i,j}) = \begin{cases} 1, & \text{CCUR}_{i,j+1} > \text{CCUR}_{i,j} \\ 0, & \text{otherwise} \end{cases}$$

and

$$I(\text{CCUR}_{i,j+1} < \text{CCUR}_{i,j}) = \begin{cases} 1, & \text{CCUR}_{i,j+1} < \text{CCUR}_{i,j} \\ 0, & \text{otherwise.} \end{cases}$$

Finally, three logistic regressions are included, the first one does not considers the longitudinal CCUR information and is labelled as LR (i), the second one considers the covariates explained in (5.4) and (5.5) and is labelled as LR (ii) and the third one considers each longitudinal observation as a covariate and is labelled as LR (iii). The probability of default π will be modelled by LR (i) as

$$\begin{aligned}\text{logit}(\pi_i^{w_i}) =& \beta_{\pi 0}^{w_i} + \beta_{\pi 1}^{w_i} \cdot cash\_adv\_total \\ &+ \beta_{\pi 2}^{w_i} \cdot mean\_age\_cc + \beta_{\pi 3}^{w_i} \cdot credit\_line,\end{aligned} \tag{5.6}$$

by LR (ii) as

$$\begin{aligned}\text{logit}(\pi_i^{w_i}) =& \beta_{\pi 0}^{w_i} + \beta_{\pi 1}^{w_i} \cdot cash\_adv\_total \\ &+ \beta_{\pi 2}^{w_i} \cdot mean\_age\_cc + \beta_{\pi 3}^{w_i} \cdot credit\_line \\ &+ \beta_{\pi 4}^{w_i} \cdot ccur\_trend + \beta_{\pi 5}^{w_i} \cdot ccur\_up\_down,\end{aligned} \tag{5.7}$$

and by LR (iii) as

$$\begin{aligned}\text{logit}(\pi_i^{w_i}) =& \beta_{\pi 0}^{w_i} + \beta_{\pi 1}^{w_i} \cdot cash\_adv\_total \\ &+ \beta_{\pi 2}^{w_i} \cdot mean\_age\_cc + \beta_{\pi 3}^{w_i} \cdot credit\_line \\ &+ \beta_{\pi 4}^{w_i} \cdot ccur_1 + \beta_{\pi 5}^{w_i} \cdot ccur_2 + \beta_{\pi 6}^{w_i} \cdot ccur_3 \\ &+ \beta_{\pi 7}^{w_i} \cdot ccur_4 + \beta_{\pi 8}^{w_i} \cdot ccur_5 + \beta_{\pi 9}^{w_i} \cdot ccur_6 \\ &+ \beta_{\pi 10}^{w_i} \cdot ccur_7 + \beta_{\pi 11}^{w_i} \cdot ccur_8 + \beta_{\pi 12}^{w_i} \cdot ccur_9 \\ &+ \beta_{\pi 13}^{w_i} \cdot ccur_{10} + \beta_{\pi 14}^{w_i} \cdot ccur_{11} + \beta_{\pi 15}^{w_i} \cdot ccur_{12}.\end{aligned} \tag{5.8}$$

For the BIm and ZOIB models the best fit of each regression structure is applied to the test dataset using (3.11) in order to make out-of-sample predictions of default status. Results are shown in Table 5.8 where the logistic regressions (5.6), (5.7) and (5.8) were included in the estimation.

Model	Regression	K	AUC	
			Within Sample	Out of Sample
BIm	Linear	2	0.7112	0.6128
	Dummy	2	0.7528	0.6036
	Splines	2	0.7548	0.6272
ZOIB	Linear	2	0.7512	0.6224
	Dummy	2	0.7856	0.6388
	Splines	2	0.7832	0.6308
LR	(i)	1	0.6792	0.6032
		2	0.5406	0.5280
	(ii)	1	0.6900	0.6076
		2	0.8270	0.5190
	(iii)	1	0.7988	0.5448
		2	0.5484	0.5784

Table 5.8: AUC comparison between BIm, ZOIB and LR models.

Table 5.8 shows that ZOIB dummy regression with 2 clusters is the model with the best prediction power in the test set. Regarding the BIm model, the BIm spline regression with 2 clusters is the one that outperformed the other regression structures. LR (i) and LR (ii) both with 1 cluster performed as well as BIm linear and dummy regression in terms of out-of-sample AUC. BIm spline regression and all ZOIB regressions outperformed the LR models, therefore the classic LR models might not be the best at predicting new clients probability of default when dealing with this type of problem that includes longitudinal data.

Figure 5.10 shows the predicted probability of default assigned by the best model, sorted in ascending order. This Figure shows that the predicted default probability for the subjects that default is greater than for the ones that do not default, therefore the model can be used to make predictions.

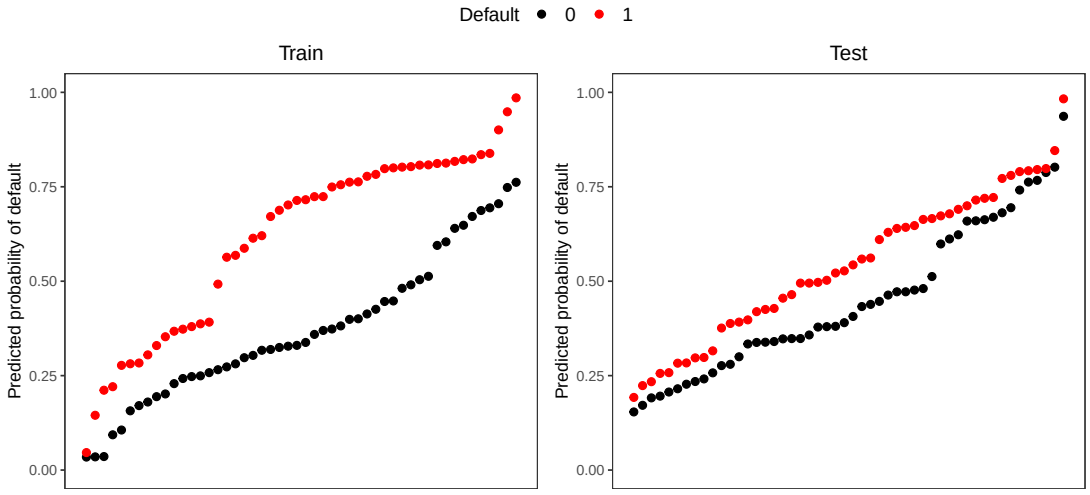


Figure 5.10: Predicted probability of default in the real dataset by the ZOIB dummy regression with 2 clusters. Red dots represent clients that default, while black represents subjects that do not default.

5.3.3. Discussion

It seems contradictory that the best fit is achieved by the BIm spline model and the best prediction power by the ZOIB dummy model. However, this may be due to the nature of the comparison criterion, in this case the deviance information criterion (DIC).

The way the DIC is computed, according to (3.13), means that the likelihood is evaluated with the same data with which the posterior distribution of the parameters was estimated. Therefore, DIC tends to choose over-fitted models.

Since prediction power is measured in an out-of-sample dataset, an over-fitted model is not expected to perform as well as a slightly more general (and therefore less over-fitted) one. And so we end in the well-known bias-variance dilemma.

In this particular case, BIm models achieve better fits than ZOIB ones because, as stated in Chapter 4, the distribution of the credit card training dataset must (without knowing it beforehand) be BIm-like. This indicates that the discrete probabilities are related to the conditional mean, thus modelling the marginal mean with the BIm parametrization allows λ_0 and λ_1 to have effects on μ and vice versa. However, as this distribution increases the over-fit of the model, it causes its predictive power to decrease and, therefore, the ZOIB models gain more relevance.

The final decision on which model to choose will depend mainly on the application it will be given. For example, for a credit risk analyst, the model that will work best for him is the one with the best predictive power because with it he could reduce the delinquency of the portfolio. On the other hand, a sales analyst would be interested in the model with the best fit because he would understand better the reasons why a customer uses more or less his credit card line and could offer a more suitable product for him.

It is left for future studies to better understand this phenomenon and to test both models on different datasets.

Chapter 6

Conclusions

6.1. Conclusions

In this thesis we introduced the classification beta inflated mixed regression models with cluster formation. In these models we used two distinct parametrizations, one of which has the advantage of modelling directly the effects of covariates in the mean of a fractional response variable (BIm) and the other for modelling the conditional mean and the inflation probabilities separately (ZOIB). The objective was to jointly model fractional and binary response variables using covariates and build different clusters with similar default probabilities and trajectories of credit card utilization ratio. This cluster formation facilitates interpretation of various client behavior profiles just by looking at the trajectories and favors the inference of a default status.

A simulation study with synthetic data was conducted. This study showed that ignoring the mixed nature of the fractional response variable using the BTran model induced bias and loss of precision in the estimation of the true parameter value, making the usage of zero-one inflated models more suitable for this type of variables. For goodness of fit on various population sizes, the ZOIB model outperformed the BIm model in terms of DIC when the simulated data was generated from a ZOIB model and vice versa, the BIm model outperformed the ZOIB model in terms of DIC when the simulated data was generated from a BIm model. Also, the ZOIB, BIm, BTran and LR models competed in the simulation study in terms of prediction power where the results indicated that was no clear winner but ZOIB, BIm and BTran models outperformed the LR model in out-of-sample prediction power.

An application to real data of BIm and ZOIB models was conducted. The real data consisted on bank clients' longitudinal credit card utilization ratio and default status within the next 12 months of observation as response variables. In order to achieve the best fit three types of regression structures were applied: linear, dummy and spline regression. The application of the models to the training dataset showed that the best fit was achieved by the BIm model with the spline regression structure considering 2 clusters, where the first cluster represents a low risk trajectory and the second a high risk trajectory. Also, when comparing the models goodness of fit, information criteria indicates that BIm model outperforms ZOIB model in the three regression structures when the number of clusters is less than 5. The best fit for both models in every regression structure is achieved when considering 2 clusters, indicating that the nature of the data has a structure of two clusters, thus considering a finite mixture was the right choice. The best fit of each structure and each model were applied to the

test dataset in order to get out-of-sample predictions where the best model for prediction was the ZOIB dummy regression model with 2 clusters. BIm and ZOIB regressions outperformed the LR models in out-of-sample prediction power. All these results will be useful for banking purposes in order to identify possible default clients earlier, take preventive actions and understand their behavior.

6.2. Suggestions for future studies

For future studies it would be interesting to introduce a parametrization of the BIm distribution that involves the dispersion parameter ϕ or to model it through covariates. Also, testing the model against the flexible beta (FB) regression proposed by Migliorati et al. (2018) which is robust for bi-modal fractional distributions. Other distributions The Tobit model restricted to [0, 1] would be challenging to the BIm model in terms of goodness of fit due to fewer parameter estimation. To reduce the number of dummy variables in the dummy regression an effect fusion as proposed by Malsiner-Walli et al. (2018) can be used. For business purposes, applying the model to a dataset that contains information of the whole Peruvian financial system could help identifying possible new good clients.

Appendix A

Simulation Study results

The following Tables show the results of the simulation study for parameter recovery, where the ZOIB and BIm models are compared against the BTran model.

Cluster	Parameter	True value	Model	Mean	P2.5 %	P97.5 %
1	$\beta^1_{\mu 1}$	-0.75	ZOIB	-0.77995	-0.81403	-0.73010
			BTran	-0.41898	-0.46811	-0.35116
	$\beta^1_{\mu 2}$	0.90	ZOIB	0.89809	0.88233	0.91739
			BTran	0.77466	0.71837	0.86752
	$\beta^1_{\mu 3}$	0.80	ZOIB	0.80057	0.77725	0.83262
			BTran	0.40118	0.29424	0.46205
	$\beta^1_{\pi 1}$	-0.55	ZOIB	-0.54008	-0.79316	-0.31414
			BTran	-0.53666	-0.80926	-0.32723
	$\beta^1_{\pi 2}$	2.00	ZOIB	2.03313	1.70657	2.33667
			BTran	1.98287	1.58181	2.35490
2	$\beta^2_{\mu 1}$	-0.30	ZOIB	-0.27768	-0.37340	-0.20813
			BTran	-0.23132	-0.33676	-0.10588
	$\beta^2_{\mu 2}$	0.80	ZOIB	-0.79808	-0.82637	-0.76526
			BTran	-1.21108	-1.41376	-1.00557
	$\beta^2_{\mu 3}$	0.85	ZOIB	0.84585	0.80600	0.87268
			BTran	0.79595	0.62980	0.93919
	$\beta^2_{\pi 1}$	-0.50	ZOIB	-0.49866	-0.71633	-0.29507
			BTran	-0.49758	-0.70999	-0.29865
	$\beta^2_{\pi 2}$	1.00	ZOIB	1.04452	0.72449	1.37151
			BTran	0.94214	0.56256	1.28802

Table A.1: Parameter estimation comparison between ZOIB model and Beta Transformed (BTran) model for the ZOIB simulated datasets of population size $n = 800$ and 3 longitudinal observations.

Cluster	Parameter	True value	Model	Mean	P2.5 %	P97.5 %
1	$\beta^1_{\gamma 1}$	-0.75	BIm	-0.77328	-0.81949	-0.73186
			BTran	-0.50042	-0.54727	-0.44967
	$\beta^1_{\gamma 2}$	1.00	BIm	0.98273	0.93262	1.03448
			BTran	0.96754	0.89352	1.04003
	$\beta^1_{\gamma 3}$	0.60	BIm	0.59144	0.55879	0.63185
			BTran	0.28882	0.22413	0.35665
	$\beta^1_{\pi 1}$	-0.50	BIm	-0.49605	-0.73518	-0.20805
			BTran	-0.49748	-0.76345	-0.15940
	$\beta^1_{\pi 2}$	2.00	BIm	2.03781	1.77738	2.33268
			BTran	2.13022	1.79256	2.46965
2	$\beta^2_{\gamma 1}$	0.55	BIm	0.54021	0.46778	0.62992
			BTran	0.24232	0.15460	0.30955
	$\beta^2_{\gamma 2}$	-1.00	BIm	-0.99147	-1.03907	-0.94478
			BTran	-0.85981	-0.97740	-0.77866
	$\beta^2_{\gamma 3}$	0.90	BIm	0.88804	0.83720	0.93639
			BTran	0.81775	0.74502	0.90851
	$\beta^2_{\pi 1}$	-0.50	BIm	-0.51234	-0.70677	-0.32643
			BTran	-0.51864	-0.70494	-0.24220
	$\beta^2_{\pi 2}$	1.00	BIm	1.05931	0.82401	1.31543
			BTran	1.03834	0.78517	1.30782

Table A.2: Parameter estimation comparison between BIm model and Beta Transformed (BTran) model for the BIm simulated datasets of population size $n = 800$ and 3 longitudinal observations.

Appendix B

Regression Structures

The ZOIB linear regression:

$$
\begin{aligned}
\text{logit}(\pi_i^{w_i}) =& \beta_{\pi 0}^{w_i} + \beta_{\pi 1}^{w_i} \cdot cash\_adv\_total + \beta_{\pi 2}^{w_i} \cdot mean\_age\_cc + \beta_{\pi 3}^{w_i} \cdot credit\_line \\
\text{logit}(\lambda_{0ij}^{w_i}) =& \beta_{\lambda_0 0}^{w_i} + \beta_{\lambda_0 1}^{w_i} \cdot j + \beta_{\lambda_0 2}^{w_i} \cdot cash\_adv\_month + \beta_{\lambda_0 3}^{w_i} \cdot mean\_age\_cc \\
& + \beta_{\lambda_0 4}^{w_i} \cdot credit\_line \\
\text{logit}(\lambda_{1ij}^{w_i}) =& \beta_{\lambda_1 0}^{w_i} + \beta_{\lambda_1 1}^{w_i} \cdot j + \beta_{\lambda_1 2}^{w_i} \cdot cash\_adv\_month + \beta_{\lambda_1 3}^{w_i} \cdot mean\_age\_cc \\
& + \beta_{\lambda_1 4}^{w_i} \cdot credit\_line \\
\text{logit}(\mu_{ij}^{w_i}) =& \beta_{\mu 0}^{w_i} + \beta_{\mu 1}^{w_i} \cdot j + \beta_{\mu 2}^{w_i} \cdot cash\_adv\_month + \beta_{\mu 3}^{w_i} \cdot mean\_age\_cc \\
& + \beta_{\mu 4}^{w_i} \cdot credit\_line + b_i.
\end{aligned}
$$

The ZOIB dummy regression:

$$
\begin{aligned}
\text{logit}(\pi_i^{w_i}) =& \beta_{\pi 0}^{w_i} + \beta_{\pi 1}^{w_i} \cdot cash\_adv\_total + \beta_{\pi 2}^{w_i} \cdot mean\_age\_cc + \beta_{\pi 3}^{w_i} \cdot credit\_line \\
\text{logit}(\lambda_{0ij}^{w_i}) =& \beta_{\lambda_0 0}^{w_i} + \beta_{\lambda_0 1}^{w_i} \cdot I(j=1) + \beta_{\lambda_0 2}^{w_i} \cdot I(j=2) + \beta_{\lambda_0 3}^{w_i} \cdot I(j=3) + \beta_{\lambda_0 4}^{w_i} \cdot I(j=4) \\
& + \beta_{\lambda_0 5}^{w_i} \cdot I(j=5) + \beta_{\lambda_0 6}^{w_i} \cdot I(j=6) + \beta_{\lambda_0 7}^{w_i} \cdot I(j=7) + \beta_{\lambda_0 8}^{w_i} \cdot I(j=8) \\
& + \beta_{\lambda_0 9}^{w_i} \cdot I(j=9) + \beta_{\lambda_0 10}^{w_i} \cdot I(j=10) + \beta_{\lambda_0 11}^{w_i} \cdot I(j=11) \\
& + \beta_{\lambda_0 12}^{w_i} \cdot cash\_adv\_month + \beta_{\lambda_0 13}^{w_i} \cdot mean\_age\_cc + \beta_{\lambda_0 14}^{w_i} \cdot credit\_line \\
\text{logit}(\lambda_{1ij}^{w_i}) =& \beta_{\lambda_1 0}^{w_i} + \beta_{\lambda_1 1}^{w_i} \cdot I(j=1) + \beta_{\lambda_1 2}^{w_i} \cdot I(j=2) + \beta_{\lambda_1 3}^{w_i} \cdot I(j=3) + \beta_{\lambda_1 4}^{w_i} \cdot I(j=4) \\
& + \beta_{\lambda_1 5}^{w_i} \cdot I(j=5) + \beta_{\lambda_1 6}^{w_i} \cdot I(j=6) + \beta_{\lambda_1 7}^{w_i} \cdot I(j=7) + \beta_{\lambda_1 8}^{w_i} \cdot I(j=8) \\
& + \beta_{\lambda_1 9}^{w_i} \cdot I(j=9) + \beta_{\lambda_1 10}^{w_i} \cdot I(j=10) + \beta_{\lambda_1 11}^{w_i} \cdot I(j=11) \\
& + \beta_{\lambda_1 12}^{w_i} \cdot cash\_adv\_month + \beta_{\lambda_1 13}^{w_i} \cdot mean\_age\_cc + \beta_{\lambda_1 14}^{w_i} \cdot credit\_line \\
\text{logit}(\mu_{ij}^{w_i}) =& \beta_{\mu 0}^{w_i} + \beta_{\mu 1}^{w_i} \cdot I(j=1) + \beta_{\mu 2}^{w_i} \cdot I(j=2) + \beta_{\mu 3}^{w_i} \cdot I(j=3) + \beta_{\mu 4}^{w_i} \cdot I(j=4) \\
& + \beta_{\mu 5}^{w_i} \cdot I(j=5) + \beta_{\mu 6}^{w_i} \cdot I(j=6) + \beta_{\mu 7}^{w_i} \cdot I(j=7) + \beta_{\mu 8}^{w_i} \cdot I(j=8) \\
& + \beta_{\mu 9}^{w_i} \cdot I(j=9) + \beta_{\mu 10}^{w_i} \cdot I(j=10) + \beta_{\mu 11}^{w_i} \cdot I(j=11) \\
& + \beta_{\mu 12}^{w_i} \cdot cash\_adv\_month + \beta_{\mu 13}^{w_i} \cdot mean\_age\_cc + \beta_{\mu 14}^{w_i} \cdot credit\_line + b_i.
\end{aligned}
$$

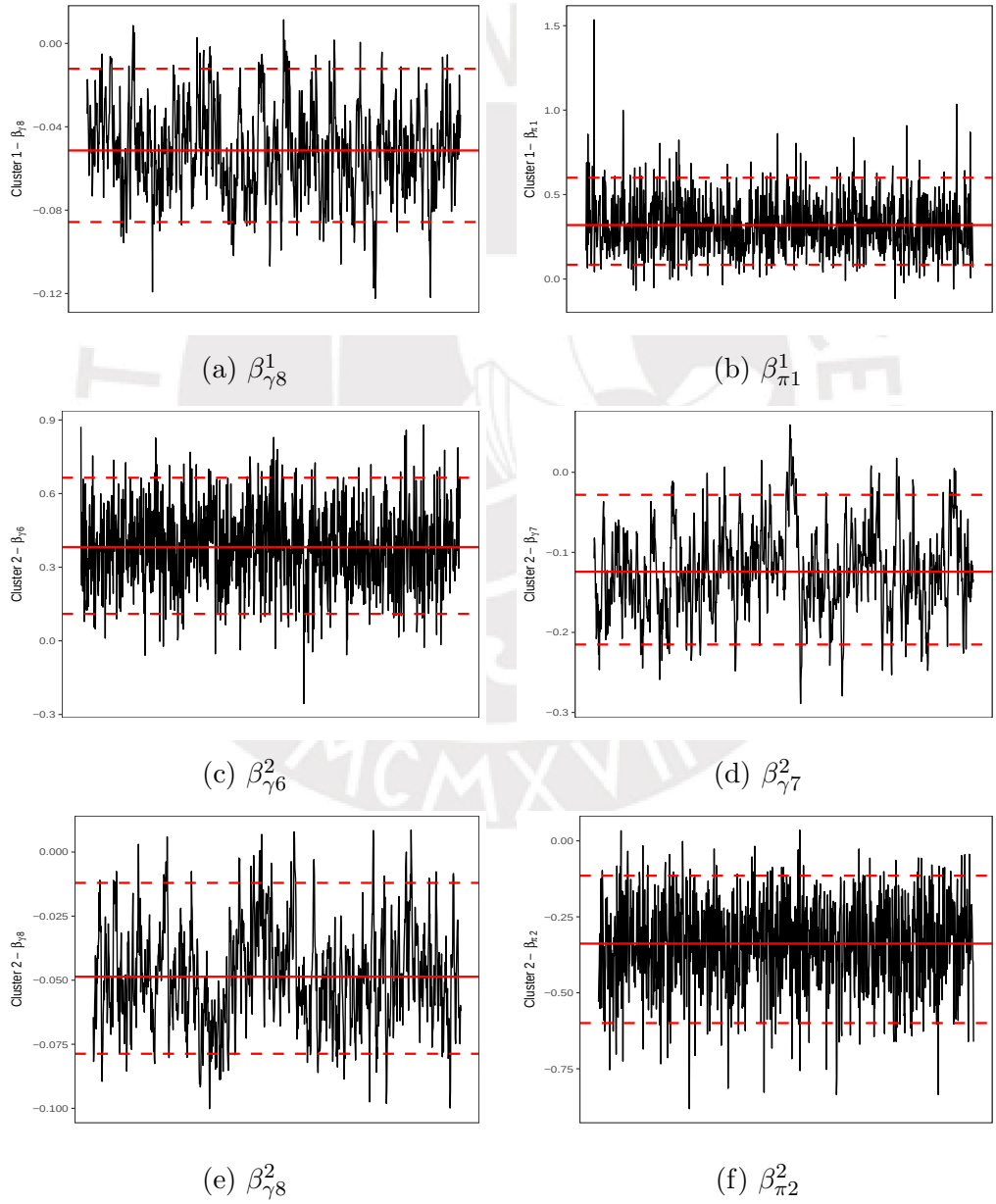
The ZOIB spline regression:

$$
\begin{aligned}
\text{logit}(\pi_i^{w_i}) =& \beta_{\pi 0}^{w_i} + \beta_{\pi 1}^{w_i} \cdot cash\_adv\_total + \beta_{\pi 2}^{w_i} \cdot mean\_age\_cc + \beta_{\pi 3}^{w_i} \cdot credit\_line \\
\text{logit}(\lambda_{0ij}^{w_i}) =& \beta_{\lambda_0 0}^{w_i} + \beta_{\lambda_0 1}^{w_i} \cdot spline\_basis_{1j} + \beta_{\lambda_0 2}^{w_i} \cdot spline\_basis_{2j} + \beta_{\lambda_0 3}^{w_i} \cdot spline\_basis_{3j} \\
& + \beta_{\lambda_0 4}^{w_i} \cdot spline\_basis_{4j} + \beta_{\lambda_0 5}^{w_i} \cdot spline\_basis_{5j} + \beta_{\lambda_0 6}^{w_i} \cdot cash\_adv\_month \\
& + \beta_{\lambda_0 7}^{w_i} \cdot mean\_age\_cc + \beta_{\lambda_0 8}^{w_i} \cdot credit\_line \\
\text{logit}(\lambda_{1ij}^{w_i}) =& \beta_{\lambda_1 0}^{w_i} + \beta_{\lambda_1 1}^{w_i} \cdot spline\_basis_{1j} + \beta_{\lambda_1 2}^{w_i} \cdot spline\_basis_{2j} + \beta_{\lambda_1 3}^{w_i} \cdot spline\_basis_{3j} \\
& + \beta_{\lambda_1 4}^{w_i} \cdot spline\_basis_{4j} + \beta_{\lambda_1 5}^{w_i} \cdot spline\_basis_{5j} + \beta_{\lambda_1 6}^{w_i} \cdot cash\_adv\_month \\
& + \beta_{\lambda_1 7}^{w_i} \cdot mean\_age\_cc + \beta_{\lambda_1 8}^{w_i} \cdot credit\_line \\
\text{logit}(\mu_{ij}^{w_i}) =& \beta_{\mu 0}^{w_i} + \beta_{\mu 1}^{w_i} \cdot spline\_basis_{1j} + \beta_{\mu 2}^{w_i} \cdot spline\_basis_{2j} + \beta_{\mu 3}^{w_i} \cdot spline\_basis_{3j} \\
& + \beta_{\mu 4}^{w_i} \cdot spline\_basis_{4j} + \beta_{\mu 5}^{w_i} \cdot spline\_basis_{5j} + \beta_{\mu 6}^{w_i} \cdot cash\_adv\_month \\
& + \beta_{\mu 7}^{w_i} \cdot mean\_age\_cc + \beta_{\mu 8}^{w_i} \cdot credit\_line + b_i.
\end{aligned}
$$

Appendix C

Application to real data results

The following Figure shows the MCMC of the application to real data using BIm spline model with two clusters.



(a) $\beta^1_{\gamma 8}$ (b) $\beta^1_{\pi 1}$

(c) $\beta^2_{\gamma 6}$ (d) $\beta^2_{\gamma 7}$

(e) $\beta^2_{\gamma 8}$ (f) $\beta^2_{\pi 2}$

Figure C.1: The solid red line is the mean and the dashed red lines are the 5 % and 95 % respectively

Appendix D

Model Implementation - RJAGS code

The JAGS code for the Beta Inflated mean mixed regression model (BIm), the Beta Transformed mixed regression model (BTran) and the Zero One Inflated Beta mixed regression model (ZOIB), all three with cluster formation and binary classification.

•BIm model:

```
model{
  #Likelihood
  for(i in 1:n){
    #Latent
    w[i] ~ dcat(p)

    #Random intercept
    b[i] ~ dnorm(0, omega_b)

    #Probability of default
    logit(pi[i]) <- betapi[ , w[i]] %*% Z[i, ]
    d[i] ~ dbern(pi[i])

    #Alpha0 and Alpha1
    for(j in 1:t){
      logit(alpha0[i, j]) <- beta0[ , w[i]] %*% X[i, j, ]
      logit(alpha1[i, j]) <- beta1[ , w[i]] %*% X[i, j, ]
      logit(gamma[i, j]) <- betag[ , w[i]] %*% X[i, j, ] + b[i]
      d0[i, j] <- alpha0[i, j]*(1-gamma[i, j])
      d1[i, j] <- alpha1[i, j]*gamma[i, j]
      z[i, j] ~ dcat(c(d0[i, j], d1[i, j], 1-d0[i, j]-d1[i, j]))
    }

    #Gamma
    for(j in ts[i, 1:open[i]]){
      y[i, j] ~ dbeta(mu[i, j]*phi, (1-mu[i, j])*phi)
      mu[i, j] <- gamma[i, j]*(1-alpha1[i, j])/
                  (1-alpha0[i, j]*(1-gamma[i, j])-alpha1[i, j]*gamma[i, j])
    }
  }
```

```
  #Priori
  p ~ ddirch(rep(1, k))
  invphi ~  dgamma(0.0001, 0.0001)
  phi <- 1/invphi
  omega_b ~  dgamma(0.0001, 0.0001)
  sigma_b <- 1/omega_b
  for(j in 1:k){
    for(i in 1:ncov){
      beta0[i, j] ~ dnorm(0.0, 0.001)
      beta1[i, j] ~ dnorm(0.0, 0.001)
      betag[i, j] ~ dnorm(0.0, 0.001)
    }
    for(i in 1:(ncov-1)){betapi[i, j] ~ dnorm(0.0, 0.001)}
  }
}
```

•BTran model:

```
model{
  #Likelihood
  for(i in 1:n){
    #Latent
    w[i] ~ dcat(p)

    #Random intercept
    b[i] ~ dnorm(0, omega_b)

    #Probability of default
    logit(pi[i]) <- betapi[ , w[i]] %*% Z[i, ]
    d[i] ~ dbern(pi[i])

    #Mu
    for(j in 1:t){
      y[i, j] ~ dbeta(mu[i, j]*phi, (1-mu[i, j])*phi)
      logit(mu[i, j]) <- betamu[ , w[i]] %*% X[i, j, ] + b[i]
    }
  }

  #Priori
  p ~ ddirch(rep(1, k))
  invphi ~  dgamma(50, 2500)
  phi <- 1/invphi
  omega_b ~  dgamma(0.0001, 0.0001)
  sigma_b <- 1/omega_b
  for(j in 1:k){
    for(i in 1:ncov){betamu[i, j] ~ dnorm(0.0, 0.001)}
    for(i in 1:(ncov-1)){betapi[i, j] ~ dnorm(0.0, 0.001)}
  }
}
```

•ZOIB model:

```
model{
  #Likelihood
  for(i in 1:n){
    #Latent
    w[i] ~ dcat(p)

    #Random intercept
    b[i] ~ dnorm(0, omega_b)

    #Probability of default
    logit(pi[i]) <- betapi[ , w[i]] %*% Z[i, ]
    d[i] ~ dbern(pi[i])

    #P and Q
    for(j in 1:t){
      logit(pb[i, j]) <- betapb[ , w[i]] %*% X[i, j, ]
      logit(qb[i, j]) <- betaqb[ , w[i]] %*% X[i, j, ]
      z[i, j] ~ dcat(c(pb[i, j], (1-pb[i, j])*qb[i, j],
                      (1-pb[i,j])*(1-qb[i, j])))
    }

    #Mu
    for(j in ts[i, 1:open[i]]){
      y[i, j] ~ dbeta(mu[i, j]*phi, (1-mu[i, j])*phi)
      logit(mu[i, j]) <- betamu[ , w[i]] %*% X[i, j, ] + b[i]
    }
  }

  #Priori
  p ~ ddirch(rep(1, k))
  invphi ~  dgamma(0.0001, 0.0001)
  phi <- 1/invphi
  omega_b ~  dgamma(0.0001, 0.0001)
  sigma_b <- 1/omega_b
  for(j in 1:k){
    for(i in 1:ncov){
      betapb[i, j] ~ dnorm(0.0, 0.001)
      betaqb[i, j] ~ dnorm(0.0, 0.001)
      betamu[i, j] ~ dnorm(0.0, 0.001)
    }
    for(i in 1:(ncov-1)){betapi[i, j] ~ dnorm(0.0, 0.001)}
  }
}
```

•LR model:

```
model{
  #Likelihood
```

```
  for(i in 1:n){
    #Latent
    w[i] ~ dcat(p)

    #Probability of default
    logit(pi[i]) <- betapi[ , w[i]] %*% Z[i, ]
    d[i] ~ dbern(pi[i])
  }

  #Priori
  p ~ ddirch(rep(1, k))
  for(j in 1:k){
    for(i in 1:(ncov-1)){betapi[i, j] ~ dnorm(0.0, 0.001)}
  }
}
```

Bibliography

Bayes, C. L. and Valdivieso, L. (2016). A beta inflated mean regression model for fractional response variables, *Journal of Applied Statistics* **43**: 1814–1830.

Carlin, B. P. and Louis, T. A. (2008). *Bayesian methods for data analysis*, CRC Press.

Coro, G. (2017). Gibbs sampling with jags: Behind the scenes.

De la Cruz-Mesía, R., Quintana, F. A. and Marshall, G. (2008). Model-based clustering for longitudinal data, *Computational Statistics & Data Analysis* **52**: 1441–1457.

De la Cruz, R., Fuentes, C., Meza, C., Lee, D.-J. and Arribas-Gil, A. (2017). Predicting pregnancy outcomes using longitudinal information: a penalized splines mixed-effects model approach, *Statistics in Medicine* **36**(13): 2120–2134.

Di Brisco, A. M. and Migliorati, S. (2020). A new mixed-effects mixture model for constrained longitudinal data, *Statistics in medicine* **39**(2): 129–145.

Fernandez, R. (2017). *A beta inflated mean regression model with mixed effects for fractional response variables*, Master's thesis, PUCP.

Fernandez, R., Bayes, C. L. and Valdivieso, L. (2018). A beta-inflated mean regression model with mixed effects for fractional response variables, *Journal of Statistical Computation and Simulation* **88**(10): 1936–1957.

Ferrari, S. and Cribari-Neto, F. (2004). Beta regression for modelling rates and proportions, *Journal of Applied Statistics* **31**: 799–815.

Figueroa-Zúñiga, J. I., Arellano-Valle, R. B. and Ferrari, S. L. (2013). Mixed beta regression: A bayesian perspective, *Computational Statistics and Data Analysis* **61**: 137 – 147.

Frühwirth-Schnatter, S., Celeux, G. and Robert, C. P. (2019). *Handbook of Mixture Analysis*, Chapman & Hall/CRC Handbooks of Modern Statistical Methods, first edn, Chapman and Hall/CRC.

Gaskins, J. T., Fuentes, C. and De la Cruz, R. (2017). A bayesian nonparametric model for predicting pregnancy outcomes using longitudinal profiles.

Gelman, A., Carlin, J. B. and Stern, H. S. (2013). *Bayesian Data Analysis*, Chapman & Hall/CRC Texts in Statistical Science, third edn, Chapman and Hall/CRC.

Liu, F. and Kong, Y. (2015). zoib: An R Package for Bayesian Inference for Beta Regression and Zero/One Inflated Beta Regression, *The R Journal* **7**(2): 34–51.

Malsiner-Walli, G., Pauger, D. and Wagner, H. (2018). Effect fusion using model-based clustering, *Statistical Modelling* **18**(2): 175–196.

Marin, J.-M., Mengersen, K. and Robert, C. P. (2005). Bayesian modelling and inference on mixtures of distributions, *Handbook of statistics* **25**: 459–507.

Marin, J.-M. and Robert, C. (2007). *Bayesian core: a practical approach to computational Bayesian statistics*, Springer Science & Business Media.

Migliorati, S., Di Brisco, A. M. and Ongaro, A. (2018). A new regression model for bounded responses, *Bayesian Anal.* **13**(3): 845–872.

Morrison, J. S. (2010). Marrying credit scoring and time-series data, *The RMA Journal* .

Ospina, R. and Ferrari, S. L. (2010). Inflated beta distributions, *Statistical papers* **51**: 111–126.

Ospina, R. and Ferrari, S. L. (2012). A general class of zero-or-one inflated beta regression models, *Computational Statistics and Data Analysis* **56**(6): 1609 – 1623.

Papastamoulis, P. (2015). label. switching: An r package for dealing with the label switching problem in mcmc outputs, *arXiv preprint arXiv:1503.02271* .

Plummer, M. (2012). Jags: Just another gibbs sampler, *Astrophysics Source Code Library* .

Plummer, M., Stukalov, A., Denwood, M. and Plummer, M. M. (2018). Package 'rjags', *update* **16**: 1.

Ramalho, J. J. and da Silva, J. V. (2009). A two-part fractional regression model for the financial leverage decisions of micro, small, medium and large firms, *Quantitative Finance* **9**(5): 621–636.

Smithson, M. and Verkuilen, J. (2006). A better lemon squeezer? maximum-likelihood regression with beta-distributed dependent variables, *Psychological methods* **11**: 54.

Spiegelhalter, D. J., Best, N. G., Carlin, B. P. and van der Linde, A. (2002). Bayesian measures of model complexity and fit, *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **64**: 583–639.

Wieczorek, J. and Hawala, S. (2011). A bayesian zero-one inflated beta model for estimating poverty in us counties.

Wieczorek, J., Nugent, C. and Hawala, S. (2012). A bayesian zero-one inflated beta model for small area shrinkage estimation.